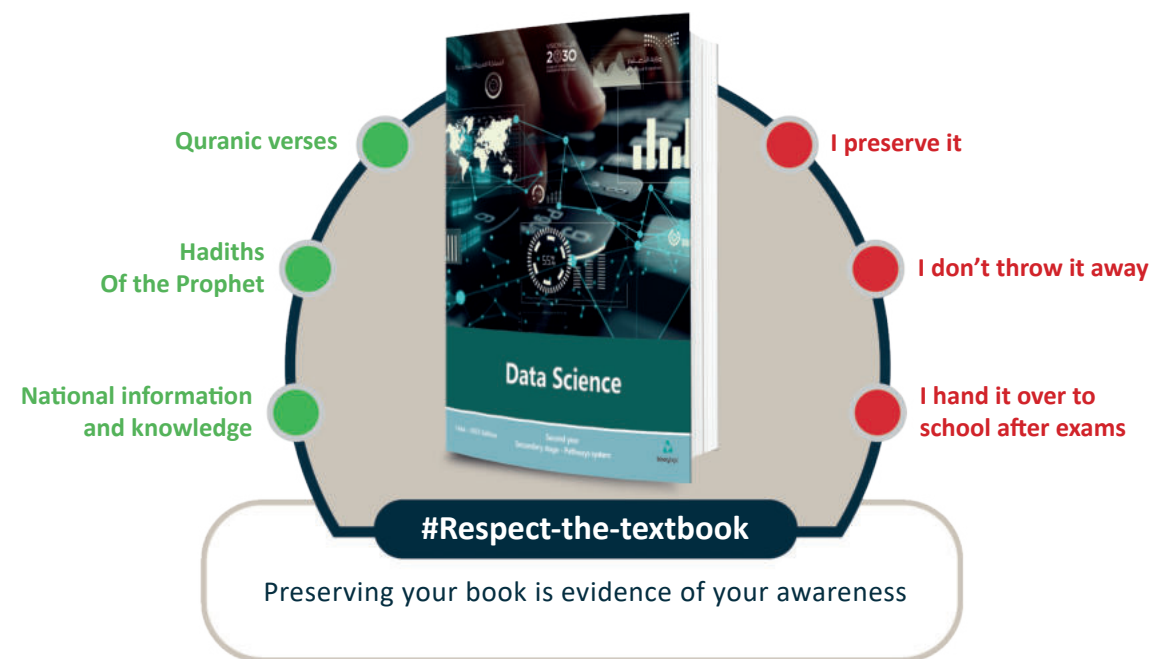


Dear student

There are significant efforts behind the completion of this book, in the process of its preparation, review, and development, and there are funds spent to print it and deliver it to you, to support your learning and your scientific and moral advancement, so, try to be faithful to this effort, appreciating it by preserving your book.



Data Science



قررت وزارة التعليم تدريس
هذا الكتاب وطبعه على نفقتها



وزارة التعليم
Ministry of Education

المملكة العربية السعودية

Data Science

Secondary stage - Pathways system

Second year



وزارة التعليم

Ministry of Education

The book is distributed freely and cannot be sold.

2023 - 1445

1445 - 2023 Edition

Publisher: Tatweer Company for Educational Services

Published under a special agreement between Binary Logic SA and Tatweer Education Services Company (Contract No. 0003/2022) for use only in the Kingdom of Saudi Arabia

Copyright © 2023 Binary Logic SA

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without permission in writing from the publishers.

Please note: This book contains links to websites that are not maintained by Binary Logic. Although we make every effort to ensure these links are accurate, up-to-date and appropriate, Binary Logic cannot take responsibility for the content of any external websites.

Trademark notice: Product or corporate names mentioned herein may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe. Binary Logic disclaims any affiliation, sponsorship, or endorsement by the respective trademark owners. Excel is a registered trademark of Microsoft Corporation. Tinkercad is a registered trademark of Autodesk Inc. "Python" and the Python logos are registered trademarks of Python Software Foundation. Jupyter is a registered trademark of Project Jupyter. PyCharm is a trademark of JetBrains s.r.o. MultisimLive is a trademark of National Instruments Corporation. CupCarbon is a registered trademark of CupCarbon. Arduino is a registered trademark of Arduino SA. Micro:bit is a registered trademark of Micro:bit Educational Foundation.

The above companies or organizations do not sponsor, authorize, or endorse this book.

The publisher has made every effort to trace all copyright holders, but if they have inadvertently overlooked any they will be pleased to make the necessary arrangements at the first opportunity.

©Ministry of Education, 2023

King Fahd National Library Cataloging-in-Publication Data

Ministry of Education

Data Science - Secondary Education - Pathways System / Second Year

Ministry of Education - Riyadh, 2023.

205p.; 210*25.5cm

ISBN: 978-603-511-234-5

1- Data science

2- Curriculum

I-Title

004 dc 1443/12720

L.D. no.: 1443/12720

ISBN: 978-603-511-234-5

Educational Support Materials at “iEN Ethraia Platform”



ien.edu.sa

Dear students, parents and anyone interested in education, we welcome your communication to improve our textbooks. Your suggestions are our top priorities.



fb.ien.edu.sa

Dear teachers and educational supervisors, we appreciate your participation in developing the new textbooks. Your input will have a definite impact on supporting and improving the educational process for our students.



fb.ien.edu.sa/BE

وزارة التعليم

Ministry of Education

2023 - 1445

Introduction:

The progress and development of countries is measured by the ability to invest in education, and the extent to which their educational system responds to the requirements and changes of the generations. In the interest of the Ministry of Education sustaining the development of its educational systems, and in response to the vision of the Kingdom of Saudi Arabia 2030, the Ministry of Education has taken the initiative to adopt the “Secondary Education Pathways” system to bring about an effective and comprehensive change in high school.

The secondary education pathways system provides a distinguished and modern educational model for high school in the Kingdom of Saudi Arabia, which efficiently contributes to:

- Strengthening the values of belonging to our homeland “the Kingdom of Saudi Arabia” and loyalty to its wise leadership “may God protect him” based on a pure belief supported by the tolerant teachings of Islam.
- Strengthening the values of citizenship by focusing on them in school subjects and activities, in line with the demands of sustainable development, and the development plans in the Kingdom of Saudi Arabia that emphasize the consolidation of both values and identity, based on the teachings of Islam and its moderation.
- Qualifying students in line with future specializations in universities or the required jobs; ensuring the consistency of education outputs with the labor market requirements.
- Enabling students to pursue education in their preferred path at early stages, according to their interests and abilities.
- Enabling students to join specific scientific and administrative disciplines related to the labor market and future jobs.
- Participation of students in an enjoyable and encouraging learning environment in school based on a constructive philosophy and applied practices within an active learning environment.
- Delivering students through an integrated educational journey from the primary level to the end of the high school and facilitating their transition process to post-general education.
- Providing students with technical and personal skills that help them deal with life and respond to the requirements of their level.
- Expanding opportunities for graduate students through various options in addition to universities, such as: obtaining professional certificates, joining applied faculties, and earning job diplomas.

The pathways system consists of nine semesters that are taught over three years, including a common first year in which students receive lessons in various scientific and humanities fields, followed by two specialized years, in which students study a general path and four specialized paths consistent with their interests and abilities, which are: the Rightful path, Business Administration path, Computer Science and Engineering path, Health and Life path, which makes this system the best for students in terms of:

- The existence of new study subjects that match the requirements of the Fourth Industrial Revolution and development plans, and the Kingdom’s Vision 2030, which aims to develop higher-order thinking, problem-solving, and research skills.
- Elective field programs that are consistent with the needs of the labor market and students’ interests, as they enable students to join a specific elective field according to a specific job skill.
- Scale as it ensures the achievement of students’ efficiency and effectiveness, and helps them identify their tendencies and interests, and reveal their strengths, which enhances their chances of success in the future.
- Volunteer work designed specifically for students in line with the philosophy of activities in schools, and is one of the graduation requirements; which helps to promote human values, and build society (its development and cohesion).
- Bridging which enables students to move from one path to another according to specific mechanisms.
- Proficiency classes through which skills are developed and the achievement level improved, by providing enrichment and remedial mastery classes.

- The options of integrated learning and distance learning, which are built in the paths system based on flexibility, convenience, interaction and effectiveness.
- The graduation project that helps students integrate theoretical experiences with applied practices.
- Professional and skill certificates granted to students after completing specific tasks, and certain tests compatible with specialized organizations.

Accordingly, the computer science and engineering path as one of the updated paths at the secondary level contributes to achieving best practices by investing in human capital, and transforming the student into a participating and productive individual for science and knowledge, while providing him with the skills and experience necessary to complete his studies in fields that meet his interests and abilities, or to join the labor market.

Data science is one of the main subjects in the course of computer science and engineering that contributes toward clarifying the nature of data and methods of analyzing it, which helps in understanding reality, making informed decisions and making useful predictions for the future in several areas of life. The course aims to introduce students to the importance of data, methods of collecting and evaluating it, how to benefit from data in solving life problems, and its role in decision-making at the personal and societal levels, with an introduction to policies and legislation related to the safe and ethical use of data. The course also focuses on enhancing computational thinking skills by dealing with data as a basic available resource that can be benefited from, also showing the importance of big data, methods of analysis, classification, characteristics, sources, techniques, applications, and areas of utilization in the educational and economic fields, artificial intelligence and machine learning and their role in the data system. This course also includes practical work from what students learn; To solve realistic problems at the students' level, under the guidance and supervision of the teacher.

The Data Science book is characterized by modern methods, in which there are elements of attraction and suspense, which make students welcome learning and interacting with it, through the various exercises and activities. This book also emphasizes important aspects in teaching and learning data science, which are:

- The close relationship between content, situations, and life problems.
- The diversity of ways to present content in an attractive and interesting way.
- Highlighting the role of the learner in the teaching and learning processes.
- Attention to the coherence of its content, which makes it integrated.
- Attention to employ appropriate techniques in different situations.
- Attention to the employment of various methods in evaluating students with respect to their individual differences.

To keep with global developments in this field, the Data Science book will provide the teacher with an integrated set of diverse educational materials that consider the individual differences between students, in addition to educational software and websites, which provide students with the opportunity to engage in modern technologies and practice-based communication; This confirms its role in the teaching and learning process.

As we present this book to our dear students, we hope that it will catch their interest, meet their requirements, and make their learning of this syllabus more enjoyable and useful.

God grants success



وزارة التعليم

Ministry of Education

2023 - 1445

Contents

1. Introduction to Data Science / 8

Lesson 1	Data, Information and Knowledge.....	9
	Exercises	17
Lesson 2	Working with Data	21
	Exercises	29
Lesson 3	Data Science Fundamentals.....	34
	Exercises	39
Project	43

2. Data Collection and Validation / 46

Lesson 1	Data Collection	47
	Exercises	53
Lesson 2	Data Types	57
	Exercises	62
Lesson 3	Data Entry Validation	65
	Exercises	89
Project	92

3. Exploratory Data Analysis / 94

Lesson 1	Data Analysis	95
	Exercises	105
Lesson 2	Python Libraries for Data Analysis	108
	Exercises	127
Lesson 3	Data Visualization	130
	Exercises	139
Project	142

4. Predictive Data Modeling and Forecasting / 144

Lesson 1	Predictive Data Modeling	145
	Exercises	157
Lesson 2	Forecasting	160
	Exercises	182
Lesson 3	Optimization	185
	Exercises	202
Project	205



بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



وزارة التعليم

Ministry of Education

2023 - 1445

1. Introduction to Data Science

In this unit, students will obtain basic knowledge about Data Science.

More specifically, students will learn what data, information and knowledge are, as well as the difference between them. Special mention will be made of the topic of the Data Science Life Cycle, as well as dealing with big data. Data governance and policies will also be discussed.

Finally, students will learn about the Data Science fundamentals, also focusing on the career opportunities that Data Science offers.

Learning Objectives

In this unit, you will learn to:

- > Define Data Science.
- > Differentiate between data, information and knowledge.
- > Recognize the differences between Data Science and Business Intelligence.
- > Examine the convergence of Data Science and Artificial Intelligence.
- > Identify the stages of the Data Science Life Cycle.
- > Describe what Big Data is.
- > Identify the characteristics of Big Data.
- > Categorize Big Data technologies.
- > Define what data governance is.
- > Identify data governance principles.
- > Discuss the skills and tools Data Science requires.
- > Identify professions related to Data Science.
- > Understand the importance of Data Science online communities.

Python programming prerequisite

The Data Science and Engineering curricula in the pathways system require knowledge of Python programming basics. You can scan the QR code on the right to access Python introductory content. To find out what topics are available and for quick access to each unit, you can see pages 208-209.





Data Science

The importance of Data Science lies in the fact that data has become an essential part of industry, because companies require data to function, grow and improve their businesses. Data helps companies in making proper decisions through data-driven approaches that analyze a large amount of data to derive meaningful insights.

Data Science application areas:

- Commercial and industrial applications.
- Healthcare, bioinformatics and natural sciences.
- Digital economy, social media and social networks analysis.
- Smart homes, smart cities, smart transportation.
- Education, e-learning, entertainment.
- Energy, sustainability and climate.

Data and Information

We are surrounded daily by data. We receive information from television, newspapers, books and the Web. But what is the difference between data and information?

Data is a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process. For example, Figure 1.1 shows a collection of a student's personal data. When data is processed, organized, structured or presented in a given context so as to make it useful, it is called information.

For example, Figure 1.2 provides organized information about a student. On this student card, you can see information such as the name, home address, telephone, email and date of birth.



Ministry of Education
Figure 1.1: Unstructured data
2023 - 1445

STUDENT CARD

Name: Mohammad
Home address: 14 Bader street
Telephone: 05*** ** *
Email: mohammedsa.bl@outlook.com
Date of birth: 16th April

Figure 1.2: Information

Data Science

Data Science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

Example

Estishraf, a National Information Center (NIC) online platform, applies advanced data science technologies to its database to generate valuable insights in more than 50 decision-making scenarios.

Data

The representation of facts or ideas in a suitable format for storage, processing, or transmission.

Information

A set of processed, organized, and structured data that provides context and enables decision making processes.

Raw Data and Information

Raw data is data that has just been collected from various sources and has not yet been processed for use. Data usually refers to raw data. Once the data has been analysed, it is considered information. Let's think about some examples:

- > The number "8122001" is considered raw data because it is a value with no contextual meaning. Now, if this value is presented as: "8/12/2001, your date of birth" then this is information as it provides knowledge about a certain matter.
- > Each student's test score is one piece of data. The average score of a class or of the entire school is information that can be derived from the data.

Information for Further Processing

Data or information from different sources can also be combined together to create more powerful datasets. This process is called data blending. For example, you can combine information from the marketing and sales departments to understand which marketing campaigns were more successful and profitable for a group of products.

Table 1.1: Differences between data and information

Data	Information
Unstructured.	Has a logical structure.
Presented in the form of numbers, figures, or statistics.	Presented through reports, graphs, or plots.
No dependencies.	Dependent on data.
Derived from user or computer system inputs.	Derived from data processing.

Knowledge

Knowledge is our understanding of the world. In other words, it is the appropriate collection of information in a way that makes it useful. We can say that when a person understands some information about something, then they have knowledge about it. Information becomes knowledge when critical thinking, evaluation, structure, or organization is applied.

Let's look at the example in Figure 1.3: The data you can see at the bottom is a list of words having no context. Now, if we organize this data, we can provide information. Let's suppose that this is a list of the sales of ice cream flavors from yesterday. A bit of analysis is useful to glean more information. For example, the most popular flavor of ice cream sold yesterday was chocolate.

The knowledge is that the shop manager can see that chocolate is the most popular ice cream flavor. The next time he places an order, he will ask for five times as much chocolate ice cream as mocha ice cream.

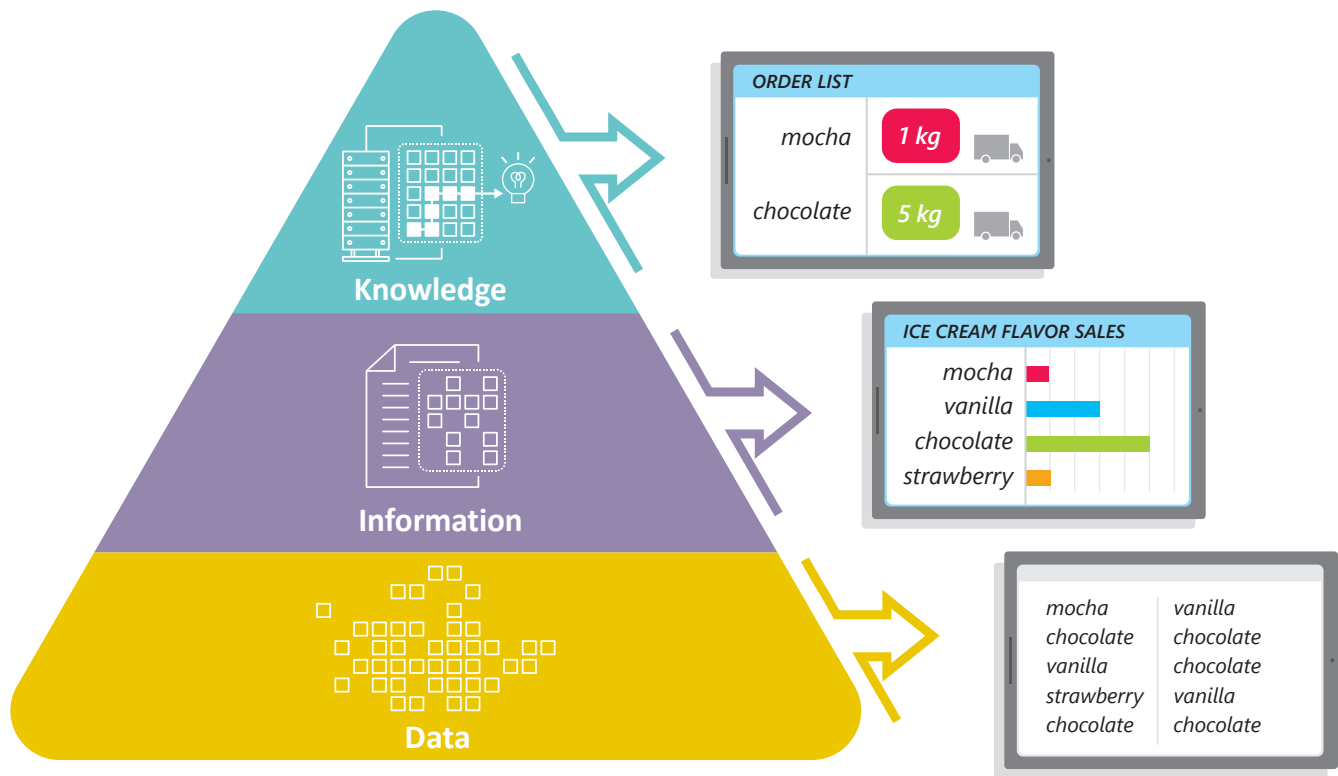


Figure 1.3: The Data - Information - Knowledge pyramid

Table 1.2: Differences between information and knowledge

	Information	Knowledge
Meaning	A refined form of processed data.	Relevant information that leads to conclusions.
Predictability	Not sufficient to make predictions.	Provides the ability to predict or make decisions.
Transfer	Can be transferred easily through verbal, written or electronic means.	Requires learning of the subject.
Outcome	The outcome is understanding.	The outcome is comprehension.
Objective	Answers the questions of who, when, what, and where.	Answers the questions of how and why.

Data Science versus Business Intelligence

Data is everywhere around us, and it is used, processed and analyzed in every field today. At the same time, data is constantly evolving and is used in several business applications, like Business Intelligence. Business Intelligence is a technology-driven process that analyzes data, providing important information that helps executives and managers make careful business decisions. While both Data Science and Business Intelligence involve data, they are different from one another.

Data Science is much more complex compared to Business Intelligence. The scope of Business Intelligence is limited to the business domain. In Business Intelligence, past data is analyzed by developing dashboards, creating business insights, organizing data and extracting information that would help the businesses to grow, with the final goal being the understanding of the current trends of the business. However, in Data Science, we use data to make future predictions and forecast the growth of the business, using a wide array of complex statistical algorithms and predictive models.

Additionally, Business Intelligence tools are limited to analyzing organizational information and setting up business strategies. On the other hand, the tools of a data scientist involve complex algorithmic models, data processing and even big data tools.

Business Intelligence

A data-driven system that incorporates data collection, data storage, data analysis, and data visualization to support decision making.

Table 1.3: Differences between Data Science and Business Intelligence

	Data Science	Business Intelligence
Scope	Data is used to make future forecasts for the development of the business.	Past data is analyzed to understand the current trends of the business.
Tools	It includes complex algorithmic models, data processing, and even big data tools.	The tools are limited to analyzing management information and overseeing business strategies.
Data types	It works with structured data, but mainly deals with unstructured and semi-structured data.	It works with structured data that is typically data warehoused or stored in data silos.
Complexity	It has more complexity compared to business intelligence.	It is much simpler compared to data science.
Flexibility	It is much more flexible as data sources can be added as required.	It is less flexible as data sources must be pre-designed.

Data Science and Artificial Intelligence

Data science has already been defined, and you are aware that Artificial Intelligence (AI) is another field that deals with massive amounts of data. These two technologies can be used independently to solve difficult challenges and they can also converge and complement one another.

Data science processes historical data using computational tools to describe situations (descriptive analysis), predict results (predictive analysis), and provide recommended solutions to problems (prescriptive analysis). The most commonly used tools are statistical and management tools, which enable the analysis of historical data. On the other hand, AI employs a variety of techniques to mimic the way people think, decide, and solve problems.

Rather than focusing on computation, the emphasis when working with AI tools is on knowledge and intelligence as critical elements for solving problems. Additionally, AI is concerned with cognitive computing. This distinction is less obvious in practice because sophisticated data science projects often include machine learning (an AI discipline) to facilitate data analysis in both prediction and prescription.

Data science and machine learning provide significant contributions to many organizations when used independently. However, traditional data analysis techniques are unsuitable when working with incomplete or inaccurate data, or the business or scientific contexts are changing so quickly that accurate data becomes obsolete very quickly. Similarly, machine learning technologies require a relatively significant amount of data.

Therefore, the next generation of data science tools and business intelligence platforms use machine learning to conduct, for example, pattern recognition to discover hidden patterns and visualize crucial insights. In addition, machine learning and deep learning support data science with more accurate predictions. The availability of large datasets and the reduced cost of processing on the cloud empower machine learning with capabilities not possible in the past. When data science and AI are combined, they create synergies that provide significantly superior results and lead to better and faster decisions.

Artificial Intelligence (AI)

A computer science field that focuses on building systems capable of performing tasks that usually require human intelligence, such as learning, reasoning, problem-solving, language and perception.

Example

Saudi Aramco has created a new "Corporate Digital Factory Department" supported by data scientists and machine learning experts who seek out operational challenges and develop intelligent solutions to help improve business performance. The company is actively promoting AI-inspired solutions to utilize billions of data points collected by geologists and petroleum engineers over the decades.

As Aramco has always been an early adopter of AI technologies, data science and machine learning tools are used to improve the performance of reservoirs deep below the surface. Advanced AI techniques optimize field development plans and well trajectories, leading to cost reduction and improvement of the environmental impact. The company's geologists have deployed AI tools to study the data collected faster and more efficiently than ever. This process improves the understanding of the petro-physical properties of the terrain to be explored and drilled and enhances decision making.

Data Science Life Cycle

Through their experience working in data science projects, data scientists and data professionals follow specific steps to implement each new project successfully. This process, called the Data Science Life Cycle, has five distinct stages. This model has numerous variations that extend the stages to cover special projects, such as AI and machine learning projects, or to represent the internal processes of specific organizations.

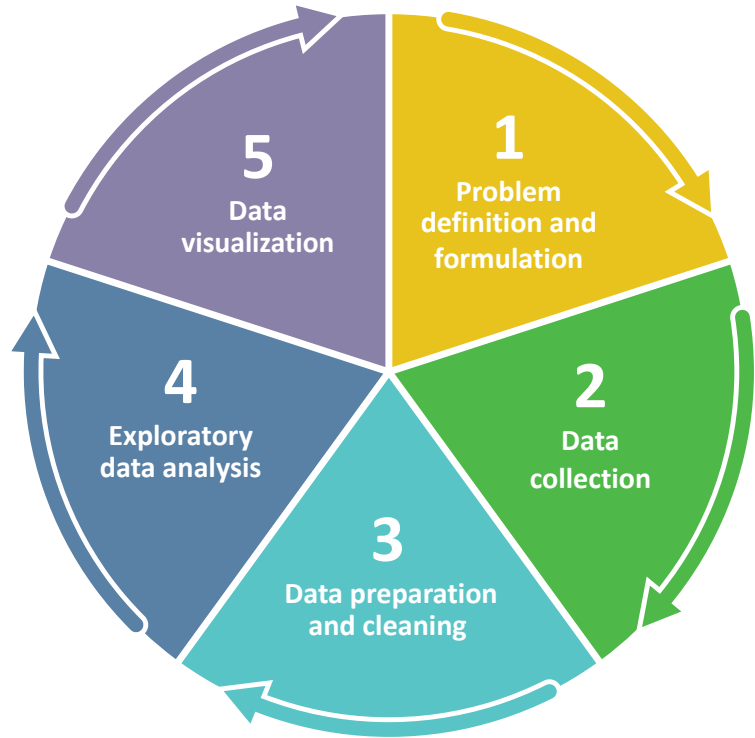


Figure 1.4: The Data Science Life Cycle stages

1. Problem Definition and Formulation

In order to design and create a solution for a Data Science problem, we first need to understand what the problem itself is. A thorough analysis of the problem, its environment and the variables that affect it are crucial for developing the solution. The understanding that we have of a problem can greatly improve or hinder the development of its solution because it directly correlates with our approach to that solution. The next objective is to define the goal we want from that solution. A dataset always contains the same data, but the answers we want to derive can vary.

Problem definition and formulation

Understanding the objectives and requirements of a business or scientific problem and converting this knowledge into a data analysis problem.

Table 1.4: The most common types of data analysis

Regression analysis	Get the quantities or qualities that exist in the dataset
Classification analysis	Organize the data into categories
Clustering analysis	Organize the data into groupings
Anomaly detection analysis	Find oddities or rarities in the data
Recommendation engines	Give an informed decision on a specific question

2. Data Collection

After we have set our objectives, we need the dataset itself. Besides manual entry of data, the most common way is data mining or data gathering. In this stage, enough data must be collected for further processing. The data itself can come from a variety of sources. Environmental sensors or mobile applications and web platforms continuously generate data. This data is automatically stored in databases.

Data Collection

The process of gathering and measuring data, including data acquisition, data labeling, and data improvement.

Table 1.5: The most common data storage formats

Formatted files	JSON, XML, CSV, Spreadsheet XLS
Relational Databases	Microsoft SQL Server, Oracle Database, Oracle MySQL
Non-Relational (NoSQL) Databases	MongoDB, Azure Cosmos DB, AWS DynamoDB
Graph Databases	Neo4j, AWS Neptune, Dgraph
Time-Series Databases	InfluxDB, AWS Timescale

3. Data Preparation and Cleaning

Data cleaning, or data wrangling, is one of the most important stages in the Data Science Life Cycle. The data scientist must clean and prepare the collected data from the data mining stage to ensure they are suitable for the subsequent analysis stage. When we combine multiple data sources, there are many chances for data to be duplicated or mixed up, and these issues will need to be fixed. If there are corrupted or incorrectly formatted data, duplicate or false data, or just incomplete data, the insights derived in the analysis stage will be false, and it will be very difficult to deduct whether the problem with the false insights originates from errors in the analysis steps or uncleaned data. This is why taking the time and the effort to clean and validate the data thoroughly before analyzing it is highly important for the entire process.

Data Cleaning

The multistage process of reviewing and correcting data to ensure it is in a standardized format, including handling missing values, smoothing noisy data, and resolving inconsistencies and duplicates.

4. Exploratory Data Analysis

We have collected and thoroughly cleaned our data, and now it is time to analyze the dataset we have gathered and derive the desired answers to our questions. Data analysis is performed with data analysis tools or programming code and the relevant code libraries. It can start with a relatively simple analysis of one or more variables and expand to more sophisticated processes involving advanced statistics.

Nowadays, the most prominent method of analyzing a dataset is Machine Learning. To analyze data with Machine Learning, we need to follow specific steps. We first need to define the Machine Learning (ML) model. We do this by first specifying what the input and output values are. The next step is to construct the analysis algorithm itself. This is a complicated process, and specialist data scientists and machine learning engineers are sometimes used solely for this task. After the algorithm is completed, it is time to train and test the model. When the training and testing phases are completed, we can then use the production data and finally generate the answers we want.

Exploratory data analysis

The approach to analyzing datasets to summarize their main characteristics, often using visual methods.

5. Data Visualization

The analyzed data are usually tables of new data that are useful in the experienced eyes of data analysts. Working with a visual representation of the analysis helps to derive better insights. Graphs, plots and charts, or even maps, along with formatted reports, provide an efficient way to see and understand trends and patterns in data. When working with massive amounts of information, visualization of the results is essential to make data-driven decisions.

Data Visualization

A graphical representation of information that highlights patterns and trends in data and aids the reader in gaining quick insights.

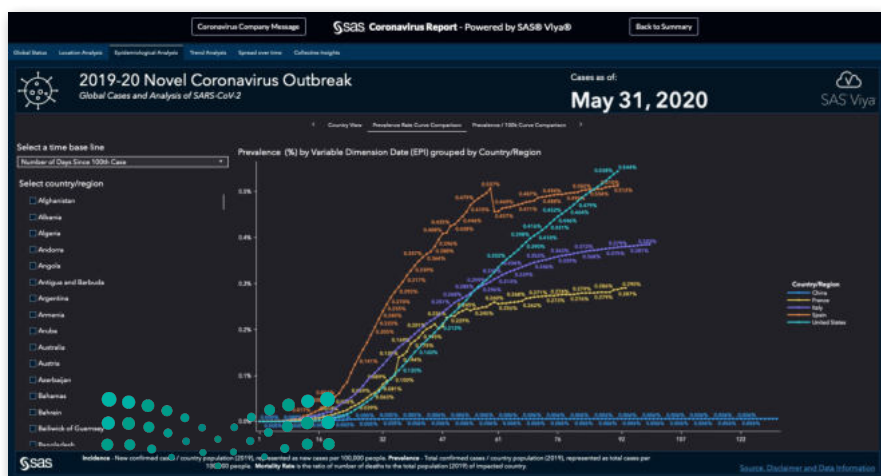


Figure 1.5: COVID-19 outbreak analysis with SAS Visual Analytics. © 2022 SAS Institute Inc.

Exercises

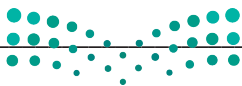
1

Read the sentences and tick ✓ True or False.	True	False
1. Data Science is a multidisciplinary field that focuses on extracting meaningful information from data.	<input type="radio"/>	<input type="radio"/>
2. When data is processed, organized, structured or presented in a given context so as to make it useful, it is called knowledge.	<input type="radio"/>	<input type="radio"/>
3. Information is obtained from data analysis.	<input type="radio"/>	<input type="radio"/>
4. Knowledge is the appropriate collection of data in a way that makes it useful.	<input type="radio"/>	<input type="radio"/>
5. Graphs and charts provide information.	<input type="radio"/>	<input type="radio"/>
6. Forecasts are considered knowledge.	<input type="radio"/>	<input type="radio"/>
7. Data Science, Artificial Intelligence and Business Intelligence are three fields that coexist independently.	<input type="radio"/>	<input type="radio"/>
8. Working with a visual representation of the analysis helps to derive better insights, so as to acquire better knowledge.	<input type="radio"/>	<input type="radio"/>
9. Recommendation engines and Regression analysis are part of the data storage procedure.	<input type="radio"/>	<input type="radio"/>
10. Time-Series Databases and Non-Relational (NoSQL) Databases are part of the data collection procedure.	<input type="radio"/>	<input type="radio"/>

2 Create a list of data and then convert the data into meaningful information. How does the computer convert data into information?

3 Mention three basic differences between Data Science and Artificial Intelligence. Justify your answers providing examples.

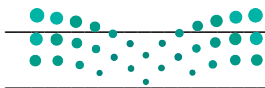
4 Contrast and compare Data Science and Business Intelligence. If you owned a trading company, in which of the two fields would you invest?



5 How effective is the convergence of Data Science and Artificial Intelligence? Search the internet and find two successful examples.

6 Explain what Data Science is and identify three applications in everyday life for health, business and entertainment. Why is Data Science so important for these applications?

7 Compare and contrast sets of unprocessed and processed data that describe the annual grades and performance of a student. What insights can you get from datasets like this? Can you predict the academic performance of the student at university?



8 Find more information on Saudi Aramco's "Corporate Digital Factory Department" and identify three examples of the use of AI in data mining. What do you think about its effect on their operational practices?

9 Search on the internet for Data Science life cycle models that describe the key stages mentioned in this lesson in more detail. Select one of them, identify the additional stages and briefly explain them.





What is Big Data?

The term "Big Data" refers to data that is either too large or complex to process using typical methods. Due to the fact that this amount of data is too large for typical computing systems to manage, the storing and the processing of these huge datasets is considered a challenge. Furthermore, data collection might be so rapid that storage requirements are extremely high.

Characteristics of Big Data

There are five key concepts that help us to classify any data under the term of "Big Data": the Variety, the Value, the Volume, the Veracity and the Velocity. Data is considered "Big" when it comes in large volumes, at a very fast rate, with great Variety, and is accurate and useful. Data must fulfill all these "Vs" in order to be considered "Big Data".

Variety

Variety refers to the many different types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semistructured data types (such as text, audio, and video) require additional preprocessing to derive meaning and support metadata information. Without the metadata, it will be impossible to know what is stored and how it can be processed.

Big Data

A large dataset that requires scalable technologies for storage, processing, management, and analysis due to its characteristics of volume, variety, velocity, veracity and value.



Figure 1.6: The Big Data characteristics - The 5 Vs

Value

Just because we collected lots of data, this does not mean it is of any value, we have to garner some insights out of it. Value refers to how useful the data is in decision making. We need to extract the value of the big data using proper analytics.

Volume

Because large volumes of low-density, unstructured data must be handled, the amount of data is a critical aspect in big data. This can be unvalued data like clickstreams on a website or mobile app, or sensor-enabled IoT devices. It might be tens of terabytes of data at times, and hundreds of petabytes at other times.

Veracity

Data veracity has to do with how accurate or truthful a dataset may be. It's not just the quality of the data itself but how trustworthy the data source, type, and processing is.

Velocity

The rate at which data is captured and stored is referred to as velocity. Most of the internet-connected smart devices (IoT devices) and mobile devices work in real-time or near real-time, requiring instant data collection, transmission and storage.

Technologies that Enable the Management of Big Data

Businesses use computer systems and databases to keep records of transactions such as order processing, payments, customer tracking, and cost management. Furthermore, a company will require a reporting system to provide information that will help it run more efficiently and help executives make more informed and, hopefully, better decisions.

Furthermore, an e-shop, for example, will need to enhance the buying experience and ensure that the website visitors become customers or that an existing customer will return to buy again. By analyzing all the data captured during the e-shop browsing on the web or through a mobile app, the company can find out where its visitors place their cursors, which parts of the website they stare at the most and how long they hover over a product before making a click for more information or an actual purchase. Tiny details are becoming a huge amount of data waiting to be analyzed and become valuable insights. This information will drive changes in the website's layout, price reductions or increases, and product campaigns on social media to influence buying behaviors.

Companies require new technologies and tools to manage and analyze big data to extract business value. The required data must be gathered from internal sources such as sales, manufacturing, and accounting, and external sources such as demographic and competition data to extract concise, reliable information about the company's current state and market dynamics. Modern infrastructure for business intelligence has an array of tools to store and process data to obtain useful information from big data. These technologies include data warehouses, data lakes and in-memory computing.



Data Warehouse

As the most traditional tool to analyze corporate data, a data warehouse refers to the database that stores current and historical data originating from many core operational transaction systems (sales, customer support, manufacturing) and makes data available to a company's decision makers. This data is combined with data from external sources to transform incomplete data to structured data before being stored in the data warehouse. A data warehouse system also provides a range of ad hoc and standardized query analysis and graphical reporting tools.

In-Memory Computing

This is a way of facilitating big data analysis, because it relies primarily on the computer's main memory (RAM) for data storage. Users access data stored in system primary memory, thereby eliminating bottlenecks from retrieving and reading data that are present in a traditional, disk-based database and dramatically shortening query response times. Very large quantities of RAM on cloud servers facilitate this method.

Data Lake

A data lake is a repository, usually in the cloud, to store huge amounts of raw and unprocessed data. It uses a flat URL structure to support both structured data (such as databases) and unstructured data (such as emails and documents).

The distinction between these three technologies is important because they serve different purposes and require different handling to be properly optimized. They do not work all together but, depending on the type of company, one of the three is chosen – a data lake may work well for one company, while a data warehouse will be a better fit for another.

Mining Big Data

Big data is being continuously collected by sensors and by applications in our environments and applications that we use personally. But collecting the data is only the first step in the process referred to as Knowledge Discovery. Knowledge discovery refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data and identifying relationships. The additional steps in the knowledge discovery process, such as data cleansing, data integration, data transformation, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data (see Table 1.6).

Data Mining

Analysis of large pools of data to find patterns and rules that can be used to guide decision making and predict future behavior.

Some of the main tasks accomplished by data mining are:

Analyzing data to discover patterns and trends.

Formulating predictions for different dataset inputs.

Classifying, clustering or forecasting the different values of the dataset.

Facilitating decision recommendations.

Table 1.6: Steps of Knowledge Discovery

Data cleaning	Clean corrupt data, irregularities, false data types, etc.
Data integration	Data mining occurs from data that originates from multiple sources. These data sources need to be merged into a single dataset.
Data selection	Selecting the part of the dataset that we want to use for the data mining process. It is important to select the dataset that is most representative of your goals because data mining is a time consuming task.
Data transformation	Preparing and formatting raw datasets is necessary because data mining processes need their inputs to have a specific format in order to analyze them.
Data mining	The actual process of analyzing the data and extracting the desired results from the analysis through patterns.
Pattern evaluation	Decoding the patterns that were generated by the data mining steps and deciding which are beneficial for each specific goal.
Knowledge representation	Visualizing the generated results with clear and concise reports, graphs and plots.

Big Data and Cloud Storage

There are two options when storing big data: cloud storage and on-premises storage. In the beginning, the development of big data applications usually required keeping data in on-premises storage, which means inside expensive, local data warehouses with complex software installed. However, the following developments spelled the end of this way of thinking, introducing cloud storage as the optimum solution to big data storage:

- (a) The widespread availability of high-speed broadband which facilitates the movement of data from one place to another. Data produced locally need no longer be analyzed locally. It can be moved to the cloud for analysis.
- (b) Nowadays, the majority of applications are cloud-based, meaning that more data is being produced and stored in the cloud. Increasing numbers of entrepreneurs are building new big data analytics to help companies analyze cloud-based data such as e-commerce transactions and web application performance data.

The biggest benefit of the cloud is versatility. Cloud based storage services include big data storage and backup systems.

For big data storage, there are a lot of options available offered by service providers such as Amazon, Microsoft and Google. All of them provide data security and privacy as well as scalability and cost efficiency.

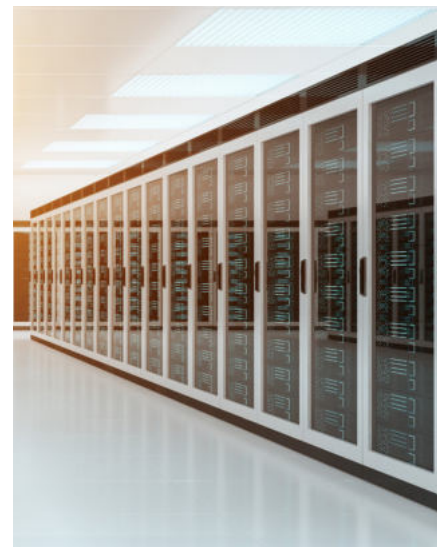


Figure 1.7: A data center providing cloud storage

By using cloud backup for big data, enterprises can utilize services from data centers that span multiple geographic locations, ensuring high availability and easy data recovery. Using the cloud, backed up data can be replicated over multiple data centers in different regions of the world. This way, the backups aren't kept at a single location. There is another layer of security to the backup. Service providers ensure that the data being backed up to the cloud is protected via advanced encryption techniques before, after and during transit.

As mentioned earlier, big data handling requires storage capacity and processing power. In terms of storage capacity, the cloud fulfills this role. Enterprises can acquire storage services that facilitate simplified scalability. And these services are also capable of meeting the computation requirements of big data. Actually, experts recommend cloud powered data analytics for big data analysis based on the almost infinite computing capabilities of the cloud.

Pros and Cons of Big Data Cloud Storage

The combination of big data analytics and cloud computing can generate opportunities not feasible before. Apart from the advantages, the data scientist needs to be aware of the challenges.

Table 1.7: Big data cloud storage advantages and disadvantages

Advantages	Disadvantages
Large volumes of structured and unstructured data require increased transmission bandwidth and storage. The cloud provides readily-available infrastructure and the ability to scale up to handle any amount of data traffic and storage requirements.	Less direct control over data security. Data breaches could lead to serious penalties under the latest data privacy regulations.
Storing big data in the cloud eliminates the need to maintain expensive on-premises hardware, software and specialized staff. The pay-as-you-go cloud computing model is more cost-efficient, reducing the waste of resources.	The cloud service provider can raise the rates of their cloud infrastructure anytime. The company that consumes these services may become locked in a business relationship that is not cost-efficient.
The company focuses on the analytics process rather than the infrastructure management, reflecting positively on the business's culture, performance, and competitive advantage.	Storing big data in the cloud means data availability depends on network connectivity. Also, the issue of latency in the cloud environment spills over into the speed of capturing, processing and storing data.

Data Governance and Policies

The policies, processes, and organizational structures define the decision rights and accountabilities to support data management. Data governance includes internal policies and procedures controlling the management of data.

Data governance assists private enterprises or state and non-profit organizations in working with high-quality data management processes through all data life cycle phases. These effective policies and procedures lead to improved business or organizational outcomes. Enterprises and organizations currently collect vast amounts of internal and external data, and data governance is necessary to use that data effectively, manage risks, and reduce costs.

Data governance ensures that data is:

Secure

Trustworthy

Documented

Managed

Audited

The Importance of Data Governance

Data inconsistencies in various systems within an organization may not be resolved without proper data governance. In sales and customer service systems, for example, customer names may be listed differently. This could make data integration more challenging and affect the accuracy of business intelligence and reporting. Furthermore, data errors may not be detected and corrected, compromising the integrity of data.

More importantly, organizations that must comply with new data privacy and protection legislation, such as the European Union's GDPR and the California Consumer Privacy Act (CCPA), may encounter difficulties or even penalties as a result of poor data governance.

In the Kingdom of Saudi Arabia, the new Personal Data Protection Law (PDPL) regulates the processing of personal data. The PDPL is the first state data privacy legislation in Saudi Arabia covering all industries and types of organizations. The National Data Management Office (NDMO) supervises and enforces the new regulations. The PDPL also applies to foreign organizations operating in Saudi Arabia and processing the personal data of Saudi residents, especially regarding genetic, health, credit and financial data.

There are special types of data, such as financial or health data, that require careful handling.

Health data is usually well governed from the time of data collection up to reporting and dissemination of information. All stakeholders fully understand the privacy risk and the constraints set by legislation, therefore a well-defined data governance framework, in a hospital, for example, is valuable.

Data Governance Framework Components

The policies, guidelines, processes, organizational structures, and technology implemented as part of a governance program make up a data governance framework. The framework also specifies the program's mission, goals, how success will be measured, and accountability for the functions that will be included in the program. The governance framework of an organization should be established and disseminated internally to explain how the program will work so that everyone engaged has a clear understanding from the start.

Data Governance Standards

ISO, the International Standards Organization, has developed a standard, ISO/IEC 38505, to apply IT governance principles to the data governance requirements.

Table 1.8: The six data governance principles

Responsibility	Assign to personnel
Strategy	Align with the mission and vision of the organization
Acquisition	Align with the organizational requirements
Conformance	Ensure compliance with legislation, internal policies and business ethics
Performance	Meet the requirements of the organization
Human behavior	Encourage people to get involved

Saudi Data Management Standards

Similar to the ISO/IEC 38505 data governance requirements, the National Data Management Office (NDMO) developed the National Data Management and Personal Data Protection Standards. The NDMO is responsible for implementing the standards and policies, governance mechanisms and controls for data and artificial intelligence and monitoring compliance by organizations and companies. The standards apply to all data regardless of form or type, including paper records, digital data, voice recordings, photos, videos, handwritten documents, or any other recorded data.



<https://sdaia.gov.sa/ndmo>



Figure 1.8: Sample pages from the NDMO Data Management and Personal Data Protection Standards. © Saudi Data & AI Authority

Example

A Saudi telecom company created its corporate analytics and data division to help meet its objective of introducing data governance and management best practices. The company's pillars of data governance are people, processes and technologies, with the initiative addressing all pillars for a successful digital transformation. The company now seeks to adopt innovative data governance solutions that leverage artificial intelligence and extend the data governance concept to "analytics governance". The goal is to achieve positive business change through well-defined data workflows and requirements.

Data Governance versus Data Management

It is critical to recognize that data governance is a component of overall data management. Data governance without actual implementation is just paperwork. Data governance establishes all policies and processes, whereas data management implements them to compile data and use it for decision-making. To draw an analogy, data governance is designing the plan of a new building, whereas data management is the act of building it. Furthermore, while you could build a house without a plan, it would be less efficient and effective, with a high risk of structural failures.

Data Governance Challenges

Cloud data and big data are two common data governance concerns that organizations encounter. Cloud services and big data systems introduce new governance requirements. Traditionally, data governance programs have focused on structured data stored in the data center. They now have to cope with the usual mix of structured, unstructured, and semi-structured data seen in big data environments, as well as the privacy threats associated with cloud data platforms.

Who is Responsible?

The data governance process involves a variety of people in most organizations. End-users familiar with relevant data in an organization's systems are included, as are business executives, data management specialists, and IT personnel. The key persons are the Chief Information Officer (CIO) or Chief Data Officer (CDO) and the Data Governance Manager (DGM).

The CIO is usually a senior executive in charge of the data governance program. The CIO's responsibilities include obtaining approval, funding and staffing for the program, taking the lead in its establishment, evaluating its development, and acting as its internal advocate.

Depending on the organization's size, a dedicated DGM may be appointed to lead and coordinate the process, hold meetings and training sessions, track KPIs (key performance indicators), and manage internal communications for the initiative. The DGM works with data owners and data stewards who ensure that the data governance policies and rules are enforced, and that end-users follow them.

Data management

Data management is the creation and implementation of architectures, policies, and procedures that manage an organization's full data life cycle needs.

Data Owner

An individual or people who are accountable for particular data.

Data Steward

A data management role that includes implementing and maintaining data governance policies within an organization.



Exercises

1

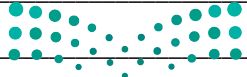
Read the sentences and tick ✓ True or False.	True	False
1. Big data refers to data that is either too large or complex to process using typical methods.	<input type="radio"/>	<input type="radio"/>
2. Some of the five technologies that enable the management of big data are the Velocity, the Veracity and the Data Warehouse.	<input type="radio"/>	<input type="radio"/>
3. Knowledge discovery is a simple process that doesn't require any specific steps.	<input type="radio"/>	<input type="radio"/>
4. Cloud storage is the only storage method provided for such a big amount of data, as it is big data.	<input type="radio"/>	<input type="radio"/>
5. Faster scalability and lower cost of analytics are some of the many advantages of big data cloud storage.	<input type="radio"/>	<input type="radio"/>
6. A data warehouse is a repository, usually in the cloud, to store huge amounts of raw and unprocessed data.	<input type="radio"/>	<input type="radio"/>
7. In-memory computing is a way of facilitating big data analysis, because it relies primarily on the computer's main memory (RAM) for data storage.	<input type="radio"/>	<input type="radio"/>
8. A data lake refers to the database that stores current and historical data originated from many core operational transaction systems.	<input type="radio"/>	<input type="radio"/>
9. Data selection involves selecting the part of the dataset that we want to use for the knowledge discovery process.	<input type="radio"/>	<input type="radio"/>
10. Knowledge representation is the process of extracting the data from the analysis through patterns.	<input type="radio"/>	<input type="radio"/>



2 Give three examples of how big data can help businesses.

3 Search the internet in order to find today's most popular cloud computing service providers in the global market, which are used to store and process big data.

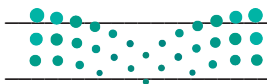
4 Explain in a few sentences how the cloud helps us deal with the problem of storing the huge amount of data that big data represents.



5 Big data is a recent development in the history of computing. Can you identify two factors that enabled this sudden growth of data collection?

6 Compare the three big data storage technologies. If you developed an application that requires very fast access to data, which one would you choose?


7 Why is pattern evaluation important for data mining?



8 Explain how scalability works in cloud data storage. Find two cloud data storage services on the internet.

9 What is the purpose of data governance? Is data governance a synonym of data management?


10 Search on the internet for information about the health data management regulations or laws in Saudi Arabia. What would be the consequences of a data leak from a health care facility?



11 Create a report on climate change by comparing historical weather data of two countries. Where will you search for information on the internet? Explain the factors behind your decision.

12 What privacy concerns can you think of when a big enterprise organization deals with big data?

13 Can you think of how much a social network you have joined knows about your family and friends? Provide a short list of information.





Mathematics Needed to Become a Data Scientist

Data science algorithms, as well as implementing analyses and discovering insights from data, require mathematical knowledge. While mathematics isn't the only tool required for a data scientist, it is one of the most significant. One of the most critical elements in a data science project workflow is identifying and comprehending business challenges and turning them into mathematical ones.

Linear Algebra

Linear algebra is concerned with matrix and vector operations. This is very important because in data science models and algorithms, all the numbers and information are converted into matrices. Another technique linear algebra is used for is dimensionality reduction which is necessary for processing large datasets. Computer vision and natural language processing (NLP) are also data science fields that rely heavily on linear algebra. All the numbers and information are converted into matrices in data science models and algorithms.

Discrete Mathematics

Discrete mathematics specializes in logic and deduction methods, which are paramount aspects of algorithm design and are the basis for data science. Another very important field of discrete mathematics is graph theory. Graphs are used for modeling very complex networks such as gene regulatory networks. Their study in data science is valuable for the advancement of fields such as precision medicine, systems biology and many more.

Probability and Statistics

When the data from an analysis gets generated, a data scientist needs practical statistical and probabilistic knowledge to be able to understand and interpret that data. Measures such as the variance, correlation and standard deviation are used extensively by data scientists to gather insight into the underlying relationships of the features of a dataset.

Calculus

Visualizing the results from a data analysis is critical to provide insightful information through the generation of plots and graphs. Calculus is an integral part of the algorithms used for the complex arithmetic operations required in this process. Properties such as partial derivatives, linear regression and gradient descent are used extensively in optimization and loss calculation.

Python for Data Science

Data Science professionals generally prefer using Python for their Data Science projects. It is a high-level, object-oriented programming language that has an easy learning curve. It is easy to begin working on a project, as you can start by writing simple structured code or design and implement a solution with Object Oriented Programming (OOP) principles.

The use of Application Programming Interfaces (APIs) and library modules provides access to powerful functionalities that are easy to use. There are numerous Python libraries that are used by professionals in various enterprises covering a wide variety of needs: data mining, data preparation and analysis, data processing, predictive modeling, data visualization and reporting. Going beyond traditional data science applications, Python libraries support machine learning and advanced artificial intelligence requirements.

Python

A high-level and general-purpose programming language which has gained increasing popularity in data science and machine learning.

Introduction to Jupyter

Python scripts can be written in an Integrated Development Environment (IDE) such as Visual Studio Code or JetBrains PyCharm, or they can be written in Jupyter Notebook. Jupyter Notebook is an open-source web application which is used to develop and present data science projects with Python. The interactive environment enables data scientists to create "notebooks". A notebook integrates Python code and its output into a single document that combines visualizations, narrative text, mathematical equations, and other data visualizations. After Jupyter is installed, it runs in a web browser either online or on a personal computer.

Besides Python, Jupyter Notebook supports over 100 programming languages (called "kernels" in the Jupyter ecosystem) including Java, R, Julia, MATLAB, Octave, Scheme, Processing, Scala, and many more. Out of the box, Jupyter will only run the IPython kernel, but additional kernels may be installed.

We will use Jupyter Notebook for Exploratory Data Analysis later in this book. The latest web-based application for Jupyter is JupyterLab, and all notebook documents work the same in both environments.

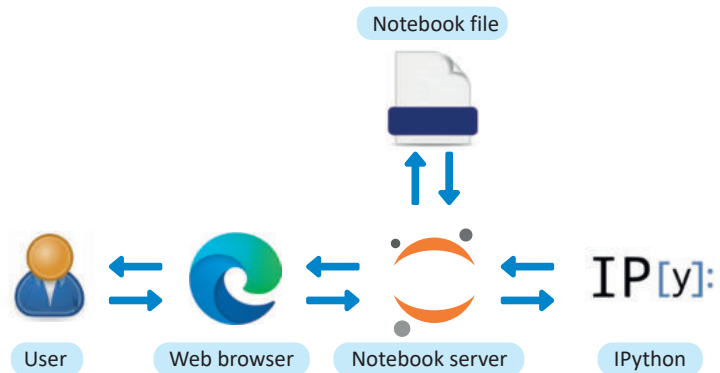


Figure 1.9: Jupyter Notebook architecture

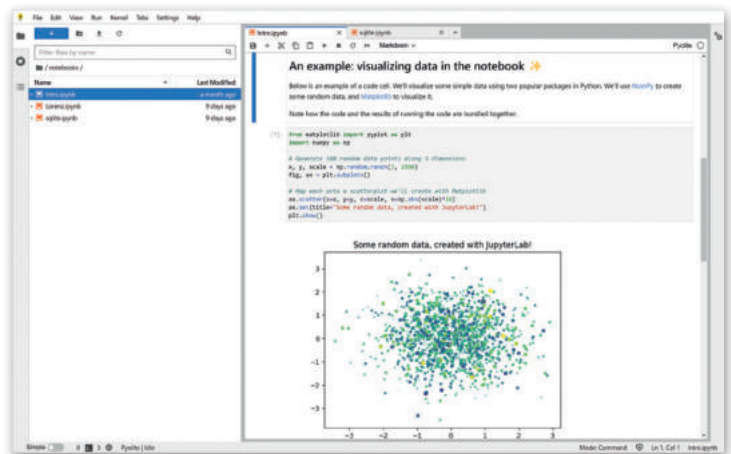


Figure 1.10: Jupyter Notebook sample screenshot

Tools for Data Science

Data science is a complex process which requires a lot of steps in order to create a data science solution. For each step of the process there exist numerous tools for accomplishing the desired task. Table 1.9 presents the most popular tools for each data science step:

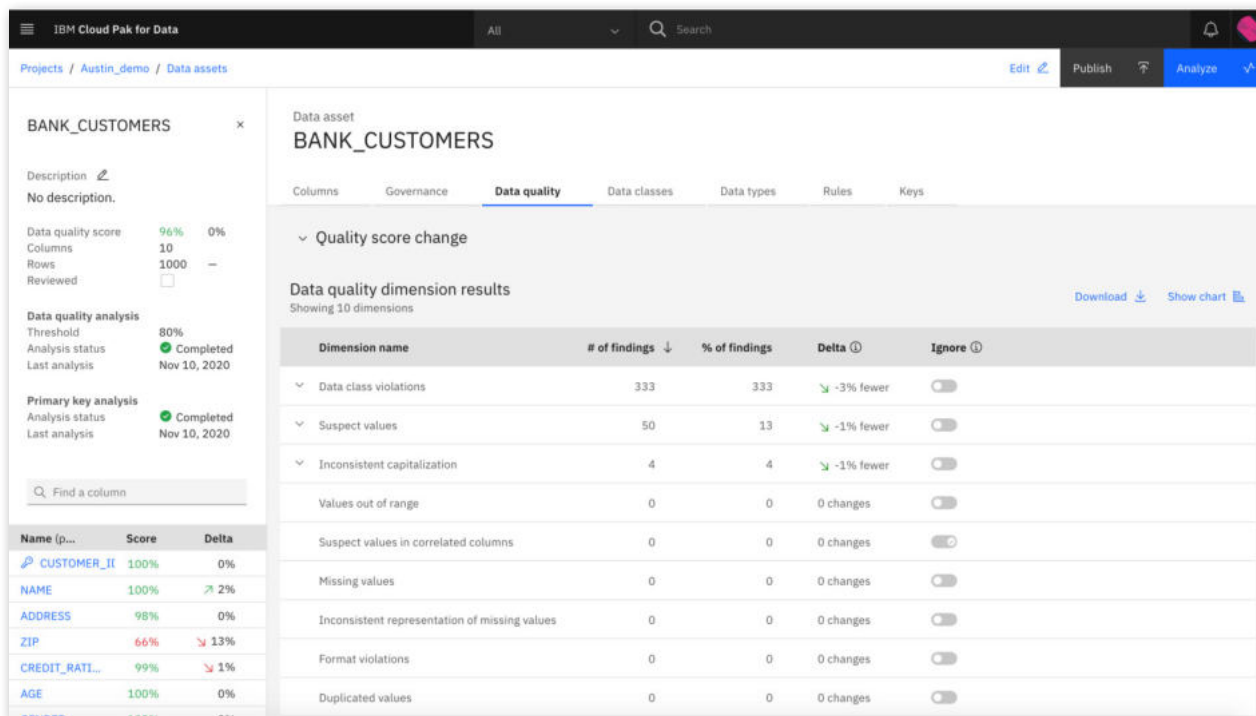


Figure 1.11: IBM Cloud Pak for Data sample screenshot

Table 1.9: Popular tools for data science steps

	Purpose	Software tools
Data Storage	The databases where the data is stored	MySQL, SQL Server, MongoDB, Neo4j
Data Transformation	Tools that query the data that we want to analyze	Python, SQL, Apache TinkerPop
Modeling	Converting the queried data into models that are appropriate for analysis	Pandas, NumPy, Apache Spark
Analysis	The process that generates the desired insights	Tensorflow, PyTorch, IBM Watson, AWS Sagemaker
Visualization	Visualizing the results in the optimal format	Matplotlib, D3.js, R

Data Science jobs

Data Science is one of the fastest growing and most in-demand computer-related fields today. The Misk Foundation has published a Saudi Jobs Market report focusing on current in-demand job roles, and Data Science career opportunities look particularly promising, especially for careers that support the goals of Saudi Vision 2030.

Table 1.10: Professions related to Data Science

Data Scientist	Their job is to find, process and analyze data for companies and organizations. They take raw and unprocessed data and extract insights and patterns from the data that help companies and organizations analyze their performance and make mission critical decisions.
Machine Learning Engineer	They are responsible for implementing Machine Learning (ML) solutions and systems for the appropriate applications. They need to be knowledgeable in software engineering and statistics in order to be able to test their solutions and judge the correctness of the produced ML models.
Machine Learning Specialist	While ML engineers are concerned with the application of ML models, ML specialists focus on the mathematics of the specific algorithms that produce the models that engineers are then able to utilize.
Applications Architect	They design the information systems for organizations and companies.
Enterprise Architect	They combine business and technical knowledge, and they are in constant communication between stakeholders and technical departments. They are tasked with translating business and organization data needs into technological specifications and solutions which they forward to the technical teams.
Data Architect	They are responsible for the storage and flow of information in a company or organization. They work with data scientists and engineers to build the appropriate data pipelines for dataset input, analysis and results output.
Data Engineer	Data engineers assist the data architects in building the digital framework for data capture, storage and processing, which both data scientists and analysts will use for their work.
Infrastructure Architect	Their role is to manage the infrastructure where data is stored and processed. They need to take into consideration factors such as data privacy, protection and infrastructure performance on the servers where the data analysis takes place. Data science projects are continuously becoming more complex, so infrastructure architects need to make sure that the data processing is completed within the appropriate timelines.
Data Analyst	Data analysts are the professionals that take the insights from the processed datasets and generate reports, visualizations and various other analytics that are aligned with the original objectives of the data science project.

Data Science Online Communities

Data scientists want to stay in touch with their peers in the field or in similar professions to learn new ideas and approaches because Data Science methodologies and technologies are always changing. Only online resources can aid data scientists in keeping up the pace. The need for a community of Data Science experts to support this work has sparked a variety of online fora and groups. Data scientists can connect and efficiently evolve the field by participating in Data Science online communities. The most prominent communities are mentioned below but this is an area where new communities may emerge and become successful.

Kaggle

Kaggle, a Google subsidiary, is the largest data science community with millions of active members and a wide range of resources. Data scientists can find public datasets, educational resources and cloud-based workbenches to support their data analysis work. <https://www.kaggle.com/>

IBM Data Community

IBM Data Community is an online forum with blogs dedicated to data science. It hosts research papers, webcasts and presentations that are updated as the field evolves. <https://community.ibm.com/community/user/ai-datascience>

There are more online communities, some of them supported by governments, some run by volunteers. Some are more focused on the community side with face-to-face meetings, while others are focused on the code required for data science projects.

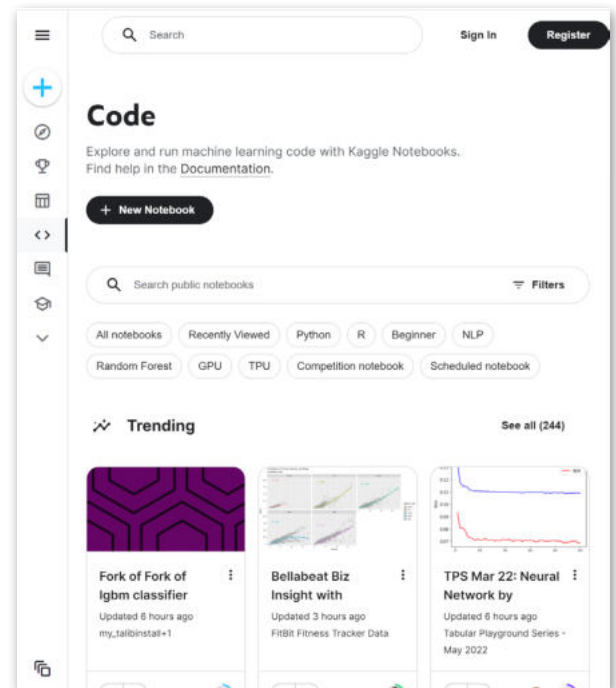


Figure 1.12: Kaggle.com home page

Table 1.11: Online communities

Data Science Central	https://www.datasciencecentral.com/
Stack Exchange	https://datascience.stackexchange.com/
Data Science Society	https://dssberkeley.com/
Driven Data	https://www.drivendata.org/
Data Community DC	https://www.datacommunitydc.org/
Reddit	https://www.reddit.com/r/datascience/

Remember to always check the online reputation of the content contributor before using a dataset, code or tools. Check for the permissions of use for each dataset and try to download software tools directly from their developers' repositories.

Exercises

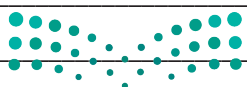
1

Read the sentences and tick ✓ True or False.	True	False
1. In machine learning models and algorithms, all the numbers and information are converted into matrices.	<input type="radio"/>	<input type="radio"/>
2. When the data from an analysis gets generated, a data scientist needs practical statistical and probabilistic knowledge to be able to understand and interpret that data.	<input type="radio"/>	<input type="radio"/>
3. Discrete mathematics specializes in logic and deduction methods which are paramount aspects of algorithm design which is the basis for machine learning.	<input type="radio"/>	<input type="radio"/>
4. Some online communities are supported by governments and some run by volunteers.	<input type="radio"/>	<input type="radio"/>
5. An Enterprise Architect is the person who designs the information systems for organizations and companies.	<input type="radio"/>	<input type="radio"/>
6. A Data Scientist is a professional that takes the insights from the processed datasets and generates reports, visualizations and various other analytics that are aligned with the original objectives of the data science project.	<input type="radio"/>	<input type="radio"/>
7. A Data Analyst is a professional who is responsible for the storage and flow of information in a company or organization. He works with data scientists and engineers to build the appropriate data pipelines for dataset input, analysis and results output.	<input type="radio"/>	<input type="radio"/>

2 Explain how Python can help a Data Science professional.

3 Explain how Jupyter can help a Data Science professional.

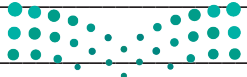
4 Mention the most important tools for Data Science. How exactly do they contribute to each Data Science step?



5 Why is understanding statistics a fundamental skill for a data scientist? Can you think of an example involving data analysis?

6 Python is a versatile programming language. Is it enough for data science projects?

7 On the internet, find three Python libraries that are very popular among data scientists. Briefly explain why.

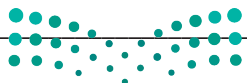


8 Compare and contrast an IDE and Jupyter Notebook. What are the main ways they differ?

9 You are learning to become a data scientist and have mastered Python coding. What other tools will you need for your data science toolkit?

10 In this lesson there is a list of professions related to data science. Which one would you prefer to follow and why? What challenges do you think you would face in this profession?

11 Visit an online data science community and search for a simple self-study training course to enhance your knowledge of data science. Evaluate how appropriate the course is to your level of knowledge.



Project

1

Social networks accumulate vast amounts of information every day. Identify three daily routines that produce private data useful to these organizations.

2

More specifically, you should think about:
What types of data are collected?
Is all this data available to the public?

3

Prepare a presentation about the privacy concerns related to social networks and how a user can be protected. What are the best practices to avoid your data becoming useful information that can be exploited by others?



Wrap up

Now you have learned:

- > **what Data Science is.**
- > **the difference between data, information and knowledge.**
- > **how Data Science is differentiated from Business Intelligence and Artificial Intelligence.**
- > **about the Data Science Life Cycle.**
- > **what big data is.**
- > **how Python or other tools can contribute to Data Science.**

KEY TERMS

Artificial Intelligence	Data Mining	In-Memory Computing
Big Data	Data Preparation	Jupyter Notebook
Business Intelligence	Data Science	Knowledge
Cloud Storage	Data Science Life Cycle	Python
Data	Data Scientist	Raw Data
Data Analysis	Data Visualization	Value
Data Analyst	Data Warehouse	Variety
Data Cleaning	Exploratory Data Analysis	Velocity
Data Collection	Information	Veracity
Data Lake		Volume

2. Data Collection and Validation

In this unit, students will obtain basic knowledge about data collection and validation.

More specifically, students will learn what data collection is, as well as the different types of data and data sources. Special mention will be made on the topic of data coding, focusing on its advantages and disadvantages.

Finally, students will learn about the data validation procedure, focusing on its respective types.



Learning Objectives

In this unit you will learn to:

- > Define what data collection is.
- > Classify the data sources.
- > Describe the attributes of information quality.
- > Understand the concept of open data platforms.
- > Recognize the importance of legal permissions for data collection.
- > Identify the different data types.
- > Define what data coding is.
- > Understand the process of data validation.
- > Categorize the data entry validation types.



Lesson 1

Data Collection

Link to digital lesson



www.ien.edu.sa

Data Collection

The most important stage of research is the stage of data collection, which is the process of collecting facts, numbers and words of the target variables. Data collection can be carried out using various devices such as sensors and data recorders. It requires a deep understanding of the under study parameters, as well as planning, and diligent work in order to obtain good quality data. Good quality data enables the proper analysis process in order to perform the tasks effectively and further extract meaningful information about the phenomenon under study.

Data collection methods vary depending on the type of data, but the process of verifying the stages of data collection in an accurate and truthful manner always remains important.

Data Collection:

The process of gathering and measuring data, including data acquisition, data labeling, and data improvement.



Figure 2.1: Engineer collecting weather data

Example

Knowing the weather is one of the most important areas relating to travel. Several devices can be used to measure weather-related factors, including temperature sensors, anemometers and hygrometers. The data collected from these devices are temperature values, wind speed values and the concentration of water vapor in the air.

Sources of Data

There are two main classifications of information sources: primary data sources and secondary data sources.

Primary Data Source

A primary data source contains data that has only been collected. It can be collected from sensors, data recorders or even from questionnaires. For example, a temperature sensor that collects air temperature data is considered a primary data source. Another example is a wind speed sensor that measures wind speed. A questionnaire given to clients about the nature of the weather they prefer for foreign trips is also a primary data source.

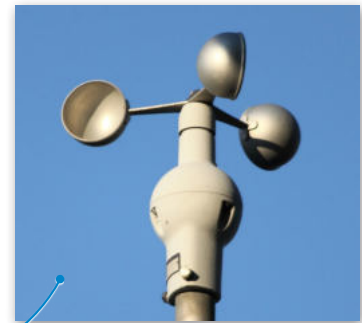


Figure 2.2: Anemometer

The anemometer is a wind speed sensor that measures wind speed. The wind generated by the airflow drives the top three wind cups to rotate, and the central axis drives an electric generator. The output of the generator operates an electric meter that is calibrated in wind speed.

Secondary Data Source

This type of data is generated when we use a primary data source in order to produce other data. For example, we can use air temperature and wind speed data from two different sensors, in order to produce data for another parameter, called wind-chill temperature. Wind-chill temperature can be found by multiplying the wind speed by 0.7 and then subtracting that value from the air temperature (wind-chill formula). In other words, first we can use the temperature and the wind sensors as primary data sources in order to collect temperature and wind speed data, and then a researcher can use the wind-chill temperature formula as a secondary data source in order to get wind-chill temperature data.

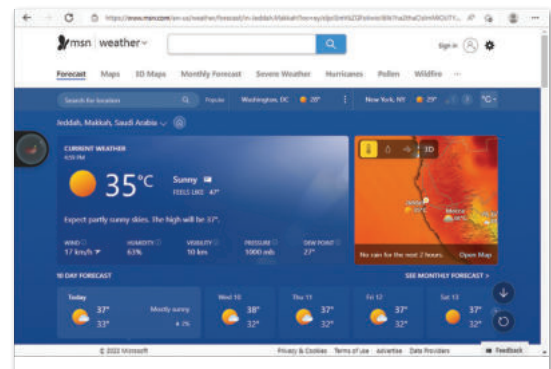


Figure 2.3: Weather forecast website

Table 2.1: Differences between primary and secondary data sources

	Primary data sources	Secondary data sources
Originality	Collected directly from the original sources.	Not original data because someone has already collected them.
Type of form	Are in raw and unorganized form.	Are in organized and processed form.
Accuracy	More accurate as it is current data.	Less accurate as it relates to the past.
Source	Collected through sensors, questionnaires, interviews, experiments, etc.	Collected from books, journals, documents, web pages, blogs, etc.
Cost	Expensive procedure and more time consuming.	Less expensive procedure and less time consuming.

Internal and External Data Sources

Data sources can be categorized into internal and external sources. Internal data sources reflect those data that are under the control of the business while external data, on the other hand, are any data generated outside the walls of the business. For example, data collected from a sensor that belongs to a "university" or to a science institution is considered internal data, while data collected from other institutions, individuals or from sources outside the specific university is considered external data in respect to that university.

Information Quality

When data are processed, organized, structured or presented in a given context so as to make them useful, they are called Information. The value of information for a given use is characterized as "Information Quality" and it is an important factor that expresses the extent to which information can be used in making decisions. With the increase in data collection and preservation, the quality of information resulting from its processing has become of great and increasing importance. Ensuring the quality of information helps to accurately determine the actual needs to implement projects, as well as to direct services effectively, and increase efficiency in each working day. In comparison, inaccurate information can cause business disruptions, reduce efficiency, and lead to delays in completing projects. The quality of information can be checked by specific criteria which are called quality attributes and they are shown in the following figure:

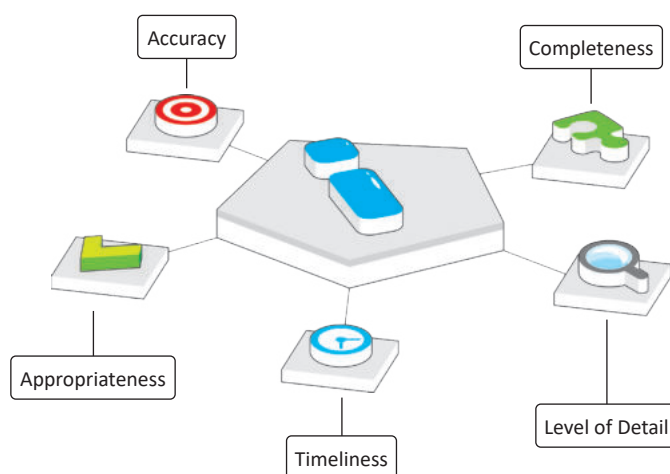


Figure 2.4: Attributes of Information Quality

Here are some questions that can help you check if information is accurate:

Can facts, statistics or other information be verified by other sources?

Can the experiment be replicated and does it have the same results?

Where is the information from?

Why was the information generated?

...Based on your knowledge, does the information seem accurate?

Does the information include misspelled words, misplaced characters and are the quotations cited correctly?

Before the collection of any kind of information through a website and before we proceed with the next data science step (which is the step of knowledge), we must verify the quality of the information that we are about to obtain from the site. If the information is not reliable, we definitely can't proceed to extract knowledge, so this means that the information must be checked by following these five quality attributes:

The importance of these five quality attributes of information is that they help us to check the reliability of the information that we find on a website.

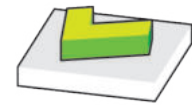
1. Accuracy

Accuracy is the correctness of the information. Information must be accurate in order to be considered of good quality. The higher the accuracy of the information, the better is its quality. Arithmetical and grammatical errors are the most common causes of inaccurate information.



2. Appropriateness

Information must be related to your topic or your research question. If there is additional and unsuitable information, the user will have to waste time searching through the data to find what is required. The more suitable the information to the receiver, the better its quality.



3. Timeliness

The date information was published or produced tells you how current it is and how up to date it is with the topic you are researching. You must always check that you have the most recent version of the information and that it is from original research.



4. Level of detail

The quality of information is also determined by the level of detail. If you give either too little or too much information, you will reduce its quality. Having too much detail makes the information overwhelming and can make it difficult to find the exact information required, while not having enough detail makes the information difficult to understand. The right amount of information is a key attribute of quality.



5. Completeness

It is the measure of comprehensiveness, that is required to ensure that the information provided gives the complete picture of reality and not just a part of the picture. Not having all the information required means you will not be able to make accurate decisions and the information cannot be used properly which also



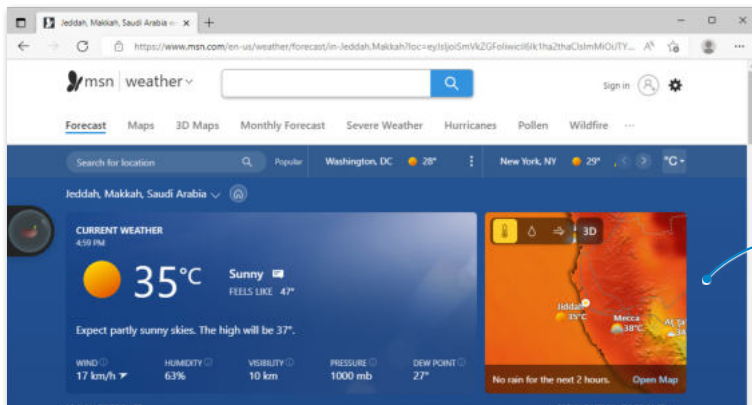
Important issues for the information timeline are the following:

Check the dates of the sources used.

Check the history of keywords for intellectual rights, such as registered trademarks, copyrights, patents, and trade secrets.

Check the history of revisions or editing of the information.

Check the date of publication.



The MSN Weather website is a great example of finding information that meets the specific five attributes of quality of information described above.

Figure 2.5: Example of a source of information

Open Data Platforms

Open data platforms are platforms which support users in accessing collections of open data. Typical open data platforms present the data of the organization which hosts the platform. State governments or non-profit organizations host open data platforms which allow access by the general public to data. More specifically, they continuously collect and organize data from a variety of public sectors. These datasets can be utilized without any financial or technical constraints. Open data can be reused and redistributed, while taking into account the requirements posed by the data License. They can also be used by citizens of other countries as well. Enterprises may also provide open data through their corporate social responsibility programs. Some of the common uses of open data platforms include:

- > *Transparency for government budgeting and spending on state services.*
- > *Performance statistics for government agencies.*
- > *Data from various public sectors e.g. education, healthcare or transportation which can be used for research that provides insights into the functioning of the country.*
- > *These datasets can be integrated into other applications.*

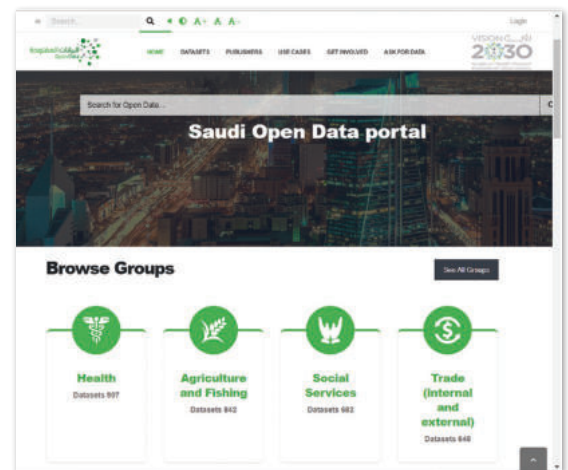


Figure 2.6: Saudi open data platform

In Saudi Arabia, the state open data platform can be found at the address: <https://data.gov.sa/>

Data Privacy

Any data that is related to a person and that can identify him or her is called personal data. For example, a name and a surname, a telephone number, an identity number, etc. are all personal data. Nowadays with so many people communicating online, there are many dangers, so it is important to protect ourselves. Data privacy ensures the ability of a person to determine for themselves when, how, and to what extent personal information about them is shared with or communicated to others.

Legal Permissions to Collect and Use Data

Collecting and using data for a research project requires legal permissions. Due to this fact, an Institutional Review Board (IRB) reviews proposals before a research project begins to determine if it follows ethical principles and legal regulations. The legal permissions can fluctuate depending on numerous variables. The two main factors to take into consideration are the location in which the data is stored and the location of the end users that consume it. Companies and organizations need to ensure that the services that collect and consume data are legally aligned with the laws of their respective countries.

Example

Data which are hosted on Open Data Platform of Saudi Arabia must be used by the visitors under the terms of the Open Data License (<https://data.gov.sa/ar/policies>).

Targeted Research and Data Comparison

Targeted research is used when we want to focus on specific issues that have emerged from our primary research. For example, if we used temperature and wind values to predict weather in a city and then we observed that specific parts of this city have recorded extreme temperature values, this means that we have to conduct targeted research into these parts, in order to assess what other parameters apart from temperature are affecting the area.

Data comparison is carried out when we have more than one dataset with registered data from the same area and from similar time periods. For example, we may have a dataset of temperature values recorded for the city of Jeddah in March 2021 and another set recorded in March 2022. Having these two data sets, we can easily perform data comparison in order to detect temperature variations or changes through the years.

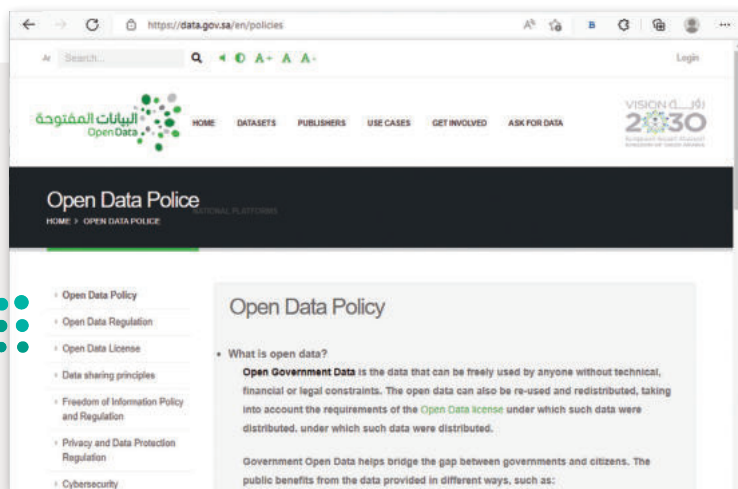


Figure 2.7: Saudi open data policies

Exercises

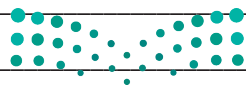
1

Read the sentences and tick ✓ True or False.	True	False
1. Data collection is the process of gathering and measuring data.	<input type="radio"/>	<input type="radio"/>
2. There are two main classifications of data collection sources: primary and secondary.	<input type="radio"/>	<input type="radio"/>
3. The date that the information was published is an important parameter of information quality.	<input type="radio"/>	<input type="radio"/>
4. Appropriateness means that the more irrelevant the information is to what is being searched for, the worse its quality.	<input type="radio"/>	<input type="radio"/>
5. Levels of detail and accuracy are considered quality standards of information.	<input type="radio"/>	<input type="radio"/>
6. The five quality attributes help us check the reliability of information.	<input type="radio"/>	<input type="radio"/>
7. The Government has no authority on open data platforms.	<input type="radio"/>	<input type="radio"/>
8. The legal permissions to collect and use data can fluctuate depending on numerous variables.	<input type="radio"/>	<input type="radio"/>
9. Targeted research is used when we want to focus on specific issues that have emerged from our primary research.	<input type="radio"/>	<input type="radio"/>
10. Data comparison can be done when we have more than one dataset with registered data from the same area and from similar time periods.	<input type="radio"/>	<input type="radio"/>

2 Briefly explain what primary and secondary data sources are.

3 Briefly describe each quality attribute which can be used to check the quality of information.

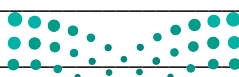
4 Give an example of targeted research and data comparison.



5 Give examples and compare the primary and secondary weather data sources.

6 Visit the data.gov.sa open data platform and search for the information on the data usage permissions. Are there any exceptions?

7 Search on the internet for open data platforms in other countries. Can you find personal information on these platforms?



8 Select two websites on the internet, one state and one private. Compare the quality of the information between them based on the five attributes.





Numerical Data and Categorical Data

Now that we've learnt what data is, let's look at its different types. Data can take different forms, for example, it may be the number of visitors to an event, the duration of that visit, and so on. In a research study, there are mainly two types of data: Numerical data and Categorical data.

Numerical Data

Numerical data consists of measurable facts, for example, the number of events that take place in a city is a type of numerical data. Numerical data can be discrete or continuous.

Discrete Data

Discrete data represents countable items and can only take certain values, e.g., the number of students in a class.

Continuous Data

Continuous data, represents data measurement and can take any value, e.g., a person's height.

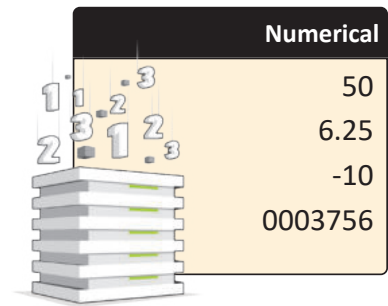


Figure 2.8: Numerical data

Categorical Data

Categorical data is a data type that is not quantitative. Categorical data refers to a data type that can be stored and identified based on the names or labels given to them. Categorical data can be nominal or ordinal.

Nominal Data

Nominal data is defined as data that is used for naming or labeling variables, without any quantitative value and without placing it in some sort of order. For example, the results of a test could be each classified nominally as a "pass" or "fail".

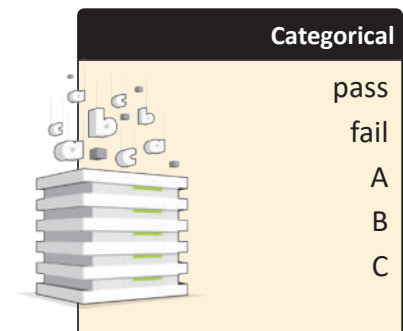


Figure 2.9: Categorical data

Ordinal Data

Ordinal data is a type of categorical data with an order. Ordinal data groups data according to some sort of ranking system. For example, test results could be grouped in descending order by grade: A, B, C, D and E.

Graphical, Video and Audio Data Types

Although the data is usually in alphanumeric form (text, numbers and symbols), it may consist of images, audio clips, or video clips. Here are some other types of data:

Graphical Data

Graphical data consists of charts, graphs, etc. For example, a set of pictures of the attractions of a particular region, or a graph of the number of visitors to a tourist place in Saudi Arabia.

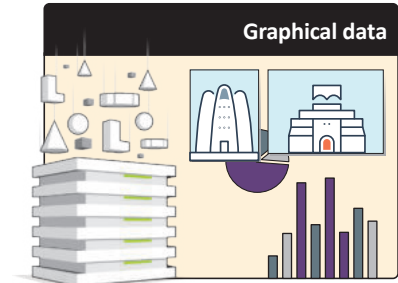


Figure 2.10: Graphical data

Image and Video Data

A digital image can be a photo or illustration represented by pixels or vectors. Video data consists of a series of moving images and audio, such as a TV advertisement for a tourism campaign, a video about the Boulevard in Saudi Arabia, a video streaming from the KSA Qur'an TV channel during the Hajj, etc.

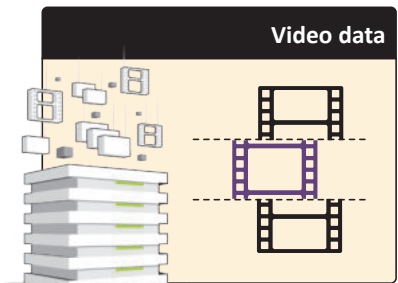


Figure 2.11: Video data

Audio Data

The audio data consists of different sounds and sound effects, such as informational audio recordings of different museums and tourist places in the Kingdom of Saudi Arabia.

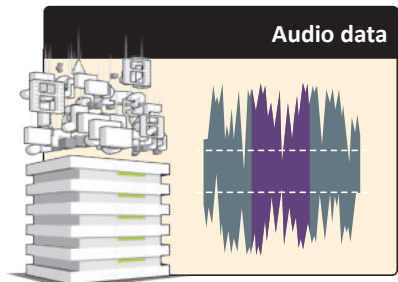


Figure 2.12: Audio data

Static and Dynamic Data

Sometimes data changes after being recorded and sometimes it doesn't. Due to this, data can be characterized as static or dynamic:

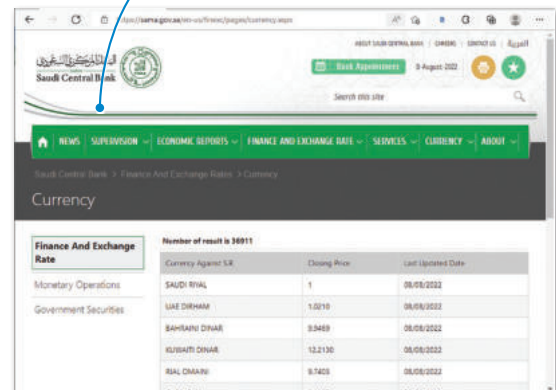
Static Data

Static data is data that does not change after being recorded. An example of static data, is a magazine with tourist places in Saudi Arabia. Once the magazine has been printed, the information in it cannot be changed.

Dynamic Data

Dynamic data is data that may change after it is recorded and has to be continually updated. An example of dynamic data can be a website with tourist places in Saudi Arabia, as it can be updated when needed.

Example of dynamic data.

A screenshot of the Saudi Central Bank website. The page title is "Currency" and it shows a table of "Finance And Exchange Rate" data. The table has columns for "Currency Against S.R.", "Closing Price", and "Last Updated Date". The data includes rows for "SAUDI RIYAL", "UAE DIRHAM", "BAHRAINI DINAR", "KUWAITI DINAR", and "REAL DINARI".

Finance And Exchange Rate	Number of result is 56911		
Monetary Operations	Currency Against S.R.		
	SAUDI RIYAL	1	08/08/2022
Government Securities	UAE DIRHAM	1.0019	08/08/2022
	BAHRAINI DINAR	3.3469	08/08/2022
	KUWAITI DINAR	12.2130	08/08/2022
	REAL DINARI	9.7425	08/08/2022
	QATAR RIAL	3.6730	08/08/2022

Figure 2.13: Saudi Central Bank website

Data Coding

The data obtained from experiments or through surveys is in its raw form and often requires coding. This process allows people to organize and arrange data in a specific way, and by using different codes such as numbers, letters, or short words, the meanings and contexts of sentences can be described as well as of entire phrases or paragraphs. Let's look at some examples from everyday life where codes are used to represent data.

Airport Codes

The International Air Transport Association (IATA) has set a three-letter code defining many airports and metropolitan areas around the world. We can search for air tickets online using this code. Also, this code is displayed on baggage tags attached at airport check-in desks, to provide safety in case of losing our baggage.

Currency Codes

Every country around the world has its own currency, and currency symbols are used instead of the currency name as customary abbreviations for financial transactions.



Figure 2.14: Currency symbols

Table 2.2: Airport codes

Code	Airport
DMM	King Fahd International Airport
JED	King Abdulaziz International Airport
RUH	King Khaled International Airport

Table 2.3: Currency codes

Code	Currency
SAR	Saudi riyal
USD	US dollar
EUR	Euro

Table 2.4: Advantages of data coding

Data entry is faster	It is easier to type RUH than King Khaled International Airport.
Takes up less space	It is difficult to write the full name of the country on a car number plate or on the back of public transportation vehicles such as taxis and buses, but with international vehicle registration codes, this is no longer a problem.
Speeds up data searches	Each region has its own code. This code is used to search for an address by area code, street number, and building, and is used by the Post Office to facilitate mail distribution.

Table 2.5: Disadvantages of data coding

Ambiguous meaning of the data.	It can be difficult to distinguish between similar codes.
Coding can be difficult to understand.	The code can be difficult to interpret or remember its meaning.
The codes used may be exhausted.	The number of items to be coded may be too large, for example, that a set of letters is not enough to code them, then numbers and letters are combined or long numbers are used, which complicates the coding process, such as coding consumer products in stores.

Barcodes

We see barcodes all around us on a daily basis, for example on electronic tickets, products in supermarkets, etc. A barcode is a label with thin, black lines across it, along with a variation of different numbers. These are used to help organize and index information or tag product prices.



Figure 2.15: Example of barcode structure

ISBN (International Standard Book Number)

On the back of most books (e.g. a guidebook), and above the barcode is a number called the ISBN. This unique number is used by publishers, libraries, and bookstores to identify book titles and editions. ISBN numbers are 13 digits long. Each ISBN consists of 5 consecutive groups of numbers: 2023 - 1445

Table 2.6: 13-digit ISBN structure

Prefix number	It consists of 3 digits, and it's either 978 or 979. For Saudi Arabia's books the prefix element is 978.
Registration group number	It consists of 1 or up to 5 digits, and it is used as an identifier for the particular country, geographical region, or language area participating in the ISBN system. For Saudi Arabia's books, the registration group element is 603 or 9960.
Registrant number	It consists of up to 7 digits, and it is used to identify the particular publisher or imprint.
Publication number	It consists of up to 6 digits, and it is used to identify the particular edition and format of a specific title.
Check digit	It consists of one single digit, it is always placed at the end, and it is used for validation of the rest of the numbers.

Example

Here is an example of an ISBN code, where each item denotes a specific piece of information about the guide book.

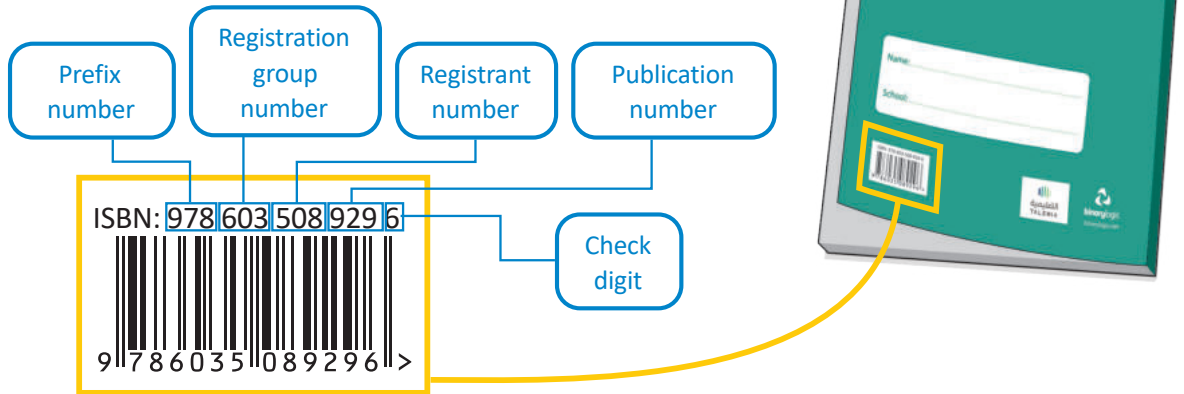


Figure 2.16: Example of 13-digit ISBN

QR Codes

The Quick Response Code is the second generation of bar code (barcode), which consists of black boxes, and contains more information. It may refer to electronic content such as websites, videos or digital files, and this code can be read using smartphone cameras.



This QR code points to the website in the link <https://visitsaudi.com>

Figure 2.17: Example of QR code

Exercises

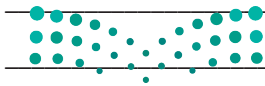
1

Read the sentences and tick ✓ True or False.	True	False
1. Categorical data is a data type that is quantitative.	<input type="radio"/>	<input type="radio"/>
2. Ordinal data is defined as data that is used for naming or labeling variables.	<input type="radio"/>	<input type="radio"/>
3. Discrete data represents countable items and can only take certain values.	<input type="radio"/>	<input type="radio"/>
4. Airport codes and currency symbols are examples of data coding.	<input type="radio"/>	<input type="radio"/>
5. Fixed data is data that may change after it is recorded and has to be continually updated.	<input type="radio"/>	<input type="radio"/>
6. Dynamic data is data that does not change after being recorded.	<input type="radio"/>	<input type="radio"/>
7. Data coding is often performed on data in its raw form, obtained from experiments or through surveys.	<input type="radio"/>	<input type="radio"/>
8. An ISBN consists of 10 consecutive groups of numbers.	<input type="radio"/>	<input type="radio"/>
9. A Barcode is a label with thin, black lines across it, along with a variation of different numbers.	<input type="radio"/>	<input type="radio"/>
10. The Quick Response Code consists of black boxes that contain information.	<input type="radio"/>	<input type="radio"/>

2 Briefly explain what Static and Dynamic data are.

3 Give some examples of products that have QR codes or Barcodes on them.

4 Briefly explain what data coding is.



5 Find a website that creates free QR codes and generate the QR codes for the home page and a specific web page of a website of your choice. Do you notice the differences in the black boxes of each QR code?

6 The International Organization for Standardization maintains the official list of country codes through ISO 3166 standard. Find the two-letter country codes of the GCC countries. Can you give examples of usage of these codes?

7 Find the ISBN barcode of this book. Can you identify the country's and the publisher's numbers?





Data Entry Validation

The main concept of data entry validation refers to any activity whose purpose is to verify that the entered data comes from a set of approved values, conforms to the accepted rules for the data, and that data can also be followed by some corrective processes and actions. If the data complies with the rules it will be accepted, otherwise it will be rejected.

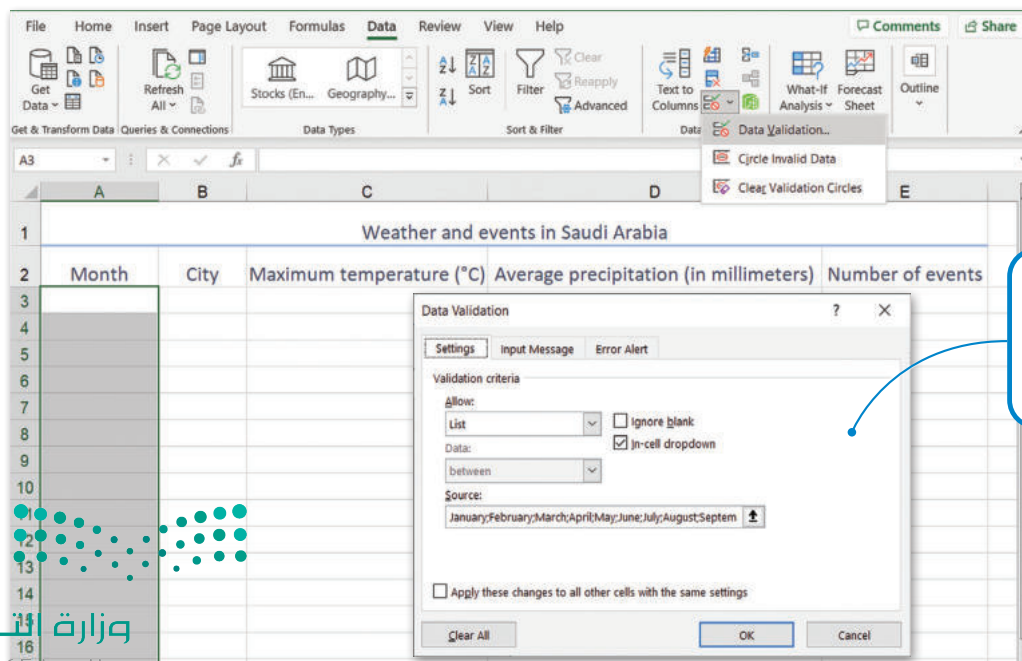
An acceptable range of recorded air temperature values from a temperature sensor could be between -89.2 degrees Celsius (lowest temperature) and 58 degrees Celsius (highest temperature). A temperature sensor shouldn't register air temperature values of 100 °C. The appearance of such data registration in the sensor's registered results indicates a malfunction of the sensor and the value should be rejected.

Data validation:

The process of ensuring the accuracy and the quality of data and it is implemented by building several checks to ensure the logical consistency of input and stored data.

Types of Data Entry Validation

There are many types of validation that we can use to validate the entered data, and there are many applications that can help us implement data entry validation rules, for example, Microsoft Excel. Most data validation procedures will perform one or more of the following checks in order to ensure that the data is correct before storing it. Common types of data validation checks are: The LookUp check, The Presence check, The Length check, The Range check, The Format check, The Type check.



This is an example of data validation in Excel.

Table 2.7: Validation types

LookUp check	<p>The LookUp validation check helps reduce errors, by using a limited list of predefined values.</p> <p>Instead of typing the name of an airport, an airline employee can select the airport from a list with the airport codes. Besides having fewer data entry errors, the process will also be faster.</p>
Presence check	<p>The Presence validation check makes the entry mandatory in the cell, ensuring that it is not left blank.</p> <p>Important data must be entered; otherwise, the data integrity is compromised. For example, the fields for a person’s name and surname cannot be empty.</p>
Length check	<p>The Length validation check aims to ensure that characters and symbols are entered within a specific length range.</p> <p>For example, ISBNs, currency symbols or ISO country codes have a fixed length of 13, 3 and 2 digits or characters, respectively.</p>
Range check	<p>The Range validation check is used to ensure that entered numbers fall within a certain range. This includes the set of two limits: a minimum and a maximum value.</p> <p>If you want to enter a person’s age, the system will have to accept only positive numbers and an upper limit, such as 140. Anything else beyond this range is invalid data.</p>
Format check	<p>The Format validation check ensures that the data is entered in a predefined format. Any other format being entered in the cell will not be allowed.</p> <p>These checks are useful when the data fields are used for zip codes or telephone numbers. In both cases, the system will expect a very specific data format.</p>
Type check	<p>The Type validation check will ensure that users enter the correct type of value in a given field.</p> <p>For example, if a data field is characterized as a number field, you will not be able to store a text value.</p>

Check Digit

The check digit is used for data validation on fixed length numbers. It can be one or two redundant digits used extensively in banking applications where bank accounts and cheque identification numbers that are entered manually need a simple error detection check. An algorithm calculates the check digit from the other digits of the number entered and compares it with the digits typed. If there is a mistyped or missing digit, the system will display a data validation warning. ISBNs, ticket numbers and a wide range of barcodes include a check digit. In recent years, where data is mainly entered via scanners, cameras and automated processes, the importance of the check digit is decreasing.



Figure 2.19: Check digit in ISBN

Data Validation Example

Let's suppose that some students are working as travel agents, and they want to create a tourist campaign for the cities of Jeddah and Riyadh, where important events will take place throughout the year. A crucial factor of the tourist campaign organization is to follow the weather conditions in each city, in order to inform tourists on how to prepare for these conditions so they can enjoy the event as much as possible.

So, as travel agents, the students have already visited the site of the National Center for Meteorology (<https://ncm.gov.sa>) and they have downloaded temperature and precipitation data for the cities of Jeddah and Riyadh as illustrated in the table below. Now, they have to open a validation program like Excel, for example, create five columns labeled months, cities, maximum temperature in degrees Celsius, average precipitation in millimeters and Number of events, program the data validation in each column and then in the cells of each column enter the data obtained from the National Center for Meteorology.

Table 2.8: Weather and events in Saudi Arabia

Excel rows / Excel columns	A	B	C	D	E
2	Month	City	Maximum temperature (°C)	Average precipitation (in millimeters)	Number of events
3	January	Jeddah	28.8	12.50	2
4	January	Riyadh	20.7	14.80	5
5	February	Jeddah	29.8	3.30	1
6	February	Riyadh	23.7	8.30	8
7	March	Jeddah	25.5	2.60	1
8	March	Riyadh	28	19.90	7
9	April	Riyadh	33.6	23.70	1
10	May	Jeddah	30.7	0.10	1
11	May	Riyadh	39.5	5.60	1
12	June	Jeddah	38.2	0.00	1
13	July	Jeddah	39.4	0.40	2
14	September	Riyadh	32.8	0.00	4
15	October	Riyadh	27.5	1.50	4
16	November	Jeddah	27.6	27.10	1
17	November	Riyadh	20.4	20.00	5

A summary of the data validation procedure that will be followed is shown in the following graphic:

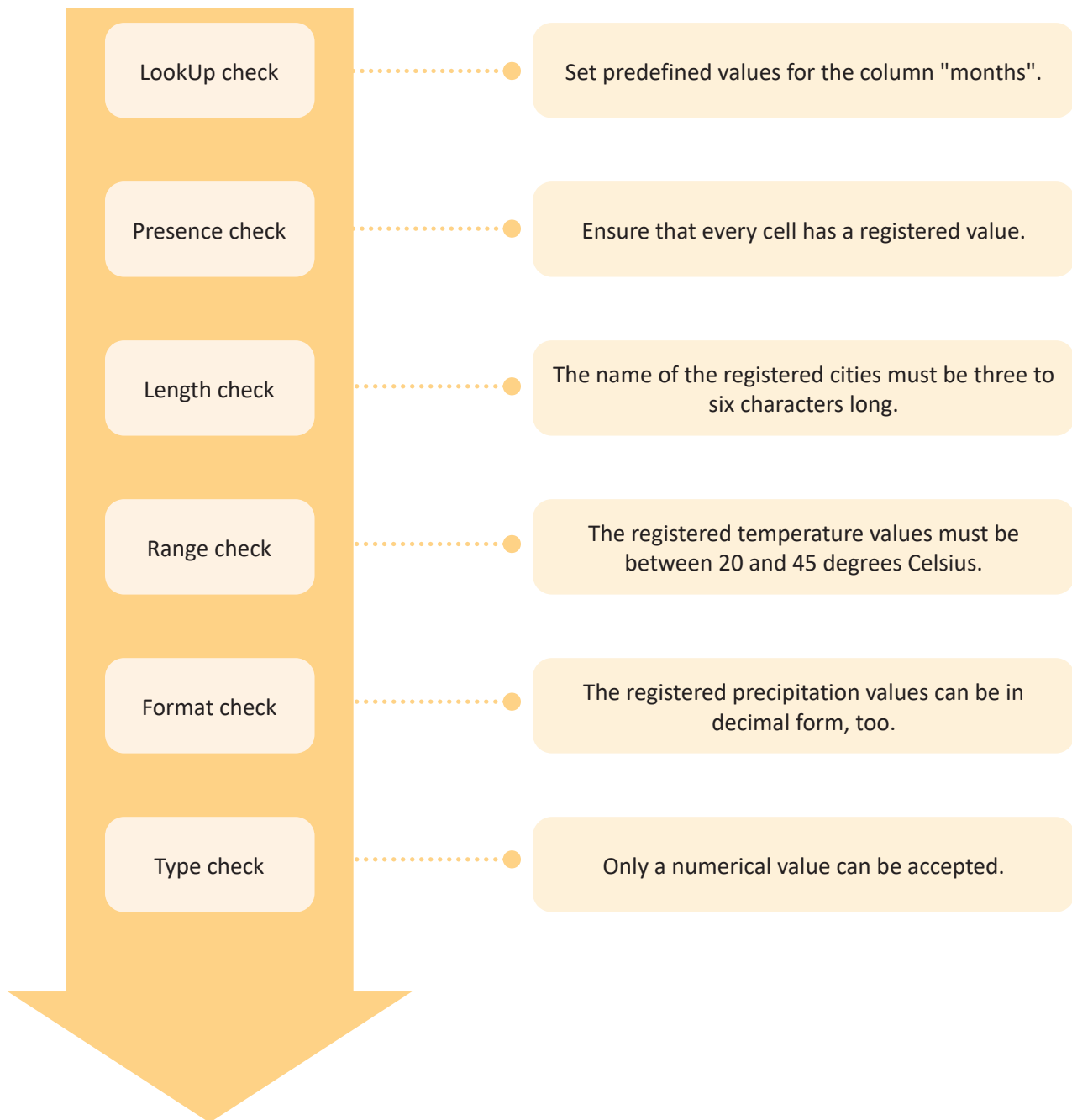


Figure 2.20: Data validation steps



LookUp and Presence Validation Check in Excel

By performing the LookUp check in the first column labeled "Months", we set all the names of the months as predefined values in order that the user can enter the data by just selecting each month from the predefined list. In the validation check, we set the program not to accept empty cells, which means cells with no value.

To start Data Validation in Excel:

- > Go to the "Weather and Events in Saudi Arabia" Excel Sheet. ①
- > Choose cells from A3 to A17. ②
- > Choose the **Data** tab. ③
- > In the **Data Tools** group ④, click **Data Validation**. ⑤
- > The **Data Validation** window appears. ⑥

To select a group of cells, you can select the first cell then use **SHIFT + arrow keys**.

The screenshot shows the Excel interface with the following elements:

- Worksheet:** "Weather and events in Saudi Arabia" with columns: Month, City, Maximum temperature (°C), Average precipitation (in millimeters), Number of events.
- Selection:** Cells A3 to A17 are selected in the "Month" column.
- Data Tab:** The "Data" tab is active, showing the "Data Tools" group.
- Data Validation Dialog:** Opened over the selected range. The "Settings" tab is active, showing "Validation criteria" set to "Any value" and "Ignore blank" checked.

To apply LookUp and Presence check in Excel:

- > In the **Data Validation** window **1**, choose the **Settings** tab. **2**
- > In the **Allow** box, choose **List**. **3**
- > In the **Source** box, type the months. **4**
- > Uncheck the **Ignore blank** option. **5**

It is very important to prevent the user from entering the wrong type of data. For this purpose, we will set invalid input and error messages in order to inform the user during the process of entering the data.

To enter the list of months in Arabic, you need to write the first month name in Arabic then change the language to English to enter ":" then change it back to Arabic to enter the next month and so on.

To set an invalid Input Message:

- > In the **Data Validation** window **1**, choose the **Input Message** tab. **2**
- > In the **Title** box, write "Data entry instruction". **3**
- > In the **Input message** box, write "Choose one of the months from the list". **4**

To set an Error message:

- > In the **Data Validation** window **1**, choose the **Error Alert** tab. **2**
- > In the **Style** box, choose **Stop**. **3**
- > In the **Title** box, write "Invalid Input". **4**
- > In the **Error message** box, write "You must choose one of the months from the list". **5**
- > Click **OK**. **6**

The Presence validation check is achieved when we uncheck the "Ignore blank" option.

The LookUp validation check is achieved when we add the names of the months in the Source: box.

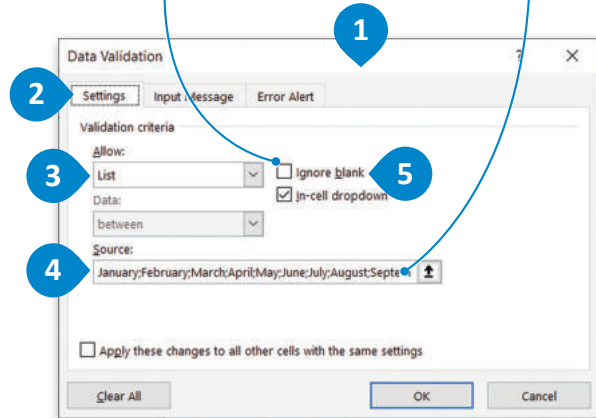


Figure 2.22: Apply lookup and presence check in Excel

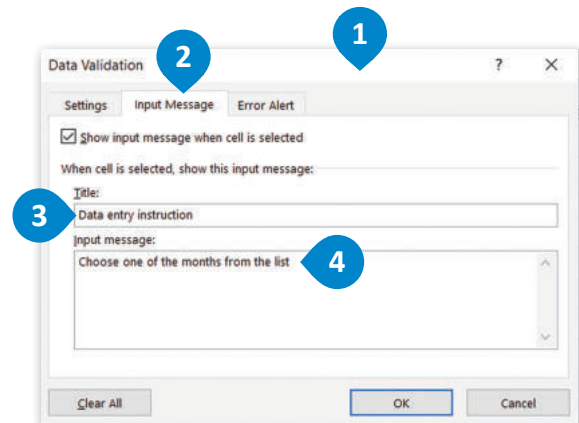


Figure 2.23: Set an invalid input message

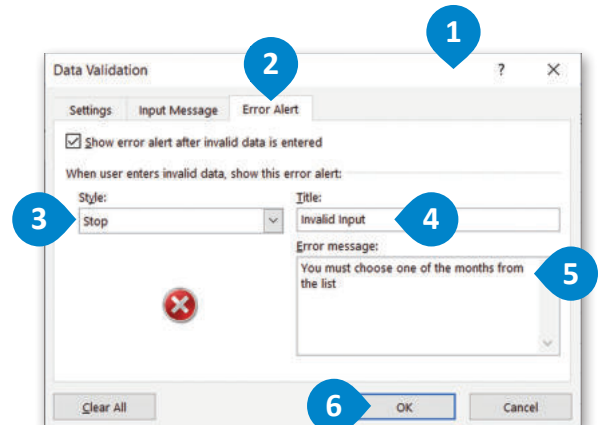


Figure 2.24: Set an error message

Now that we've finished with the LookUp validation check and the Presence validation check, we can start entering the data. For every cell (from A3 to A17) that we enter a value, we can see that a drop down list appears, showing us the predefined values for the month's column. We can choose one of the months from the drop down menu or we can just type the month in the cell. While we enter the month values, we can see that the input message "Data entry instruction" is displayed. Also, if we enter a value by mistake that is different from the predefined ones, the error message that we've previously set will appear on the screen.

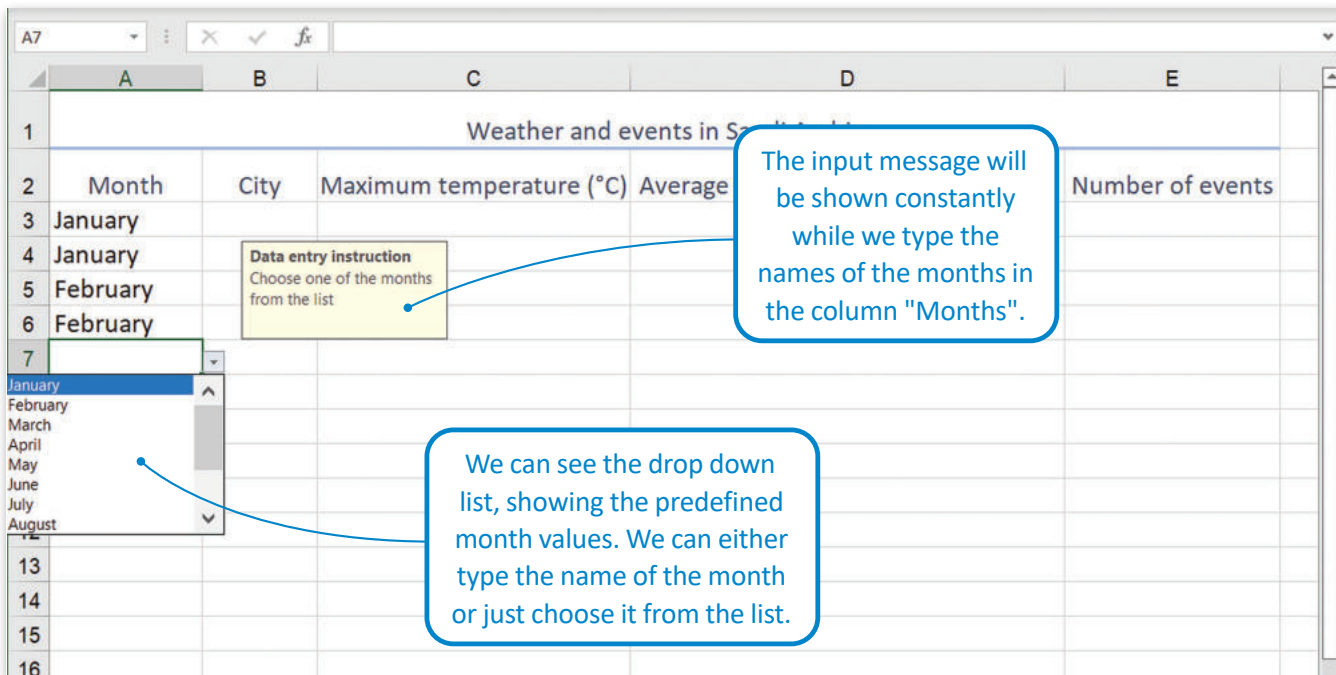


Figure 2.25: Drop-down list of the months

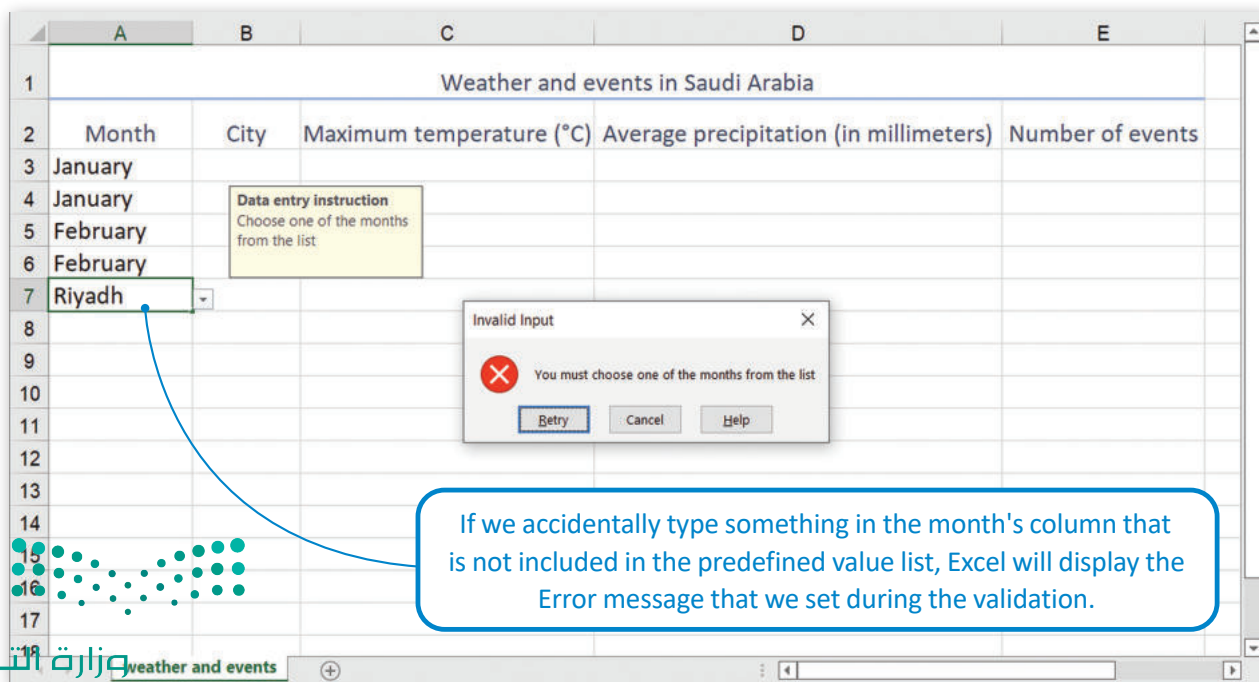


Figure 2.26: Error message of lookup validation

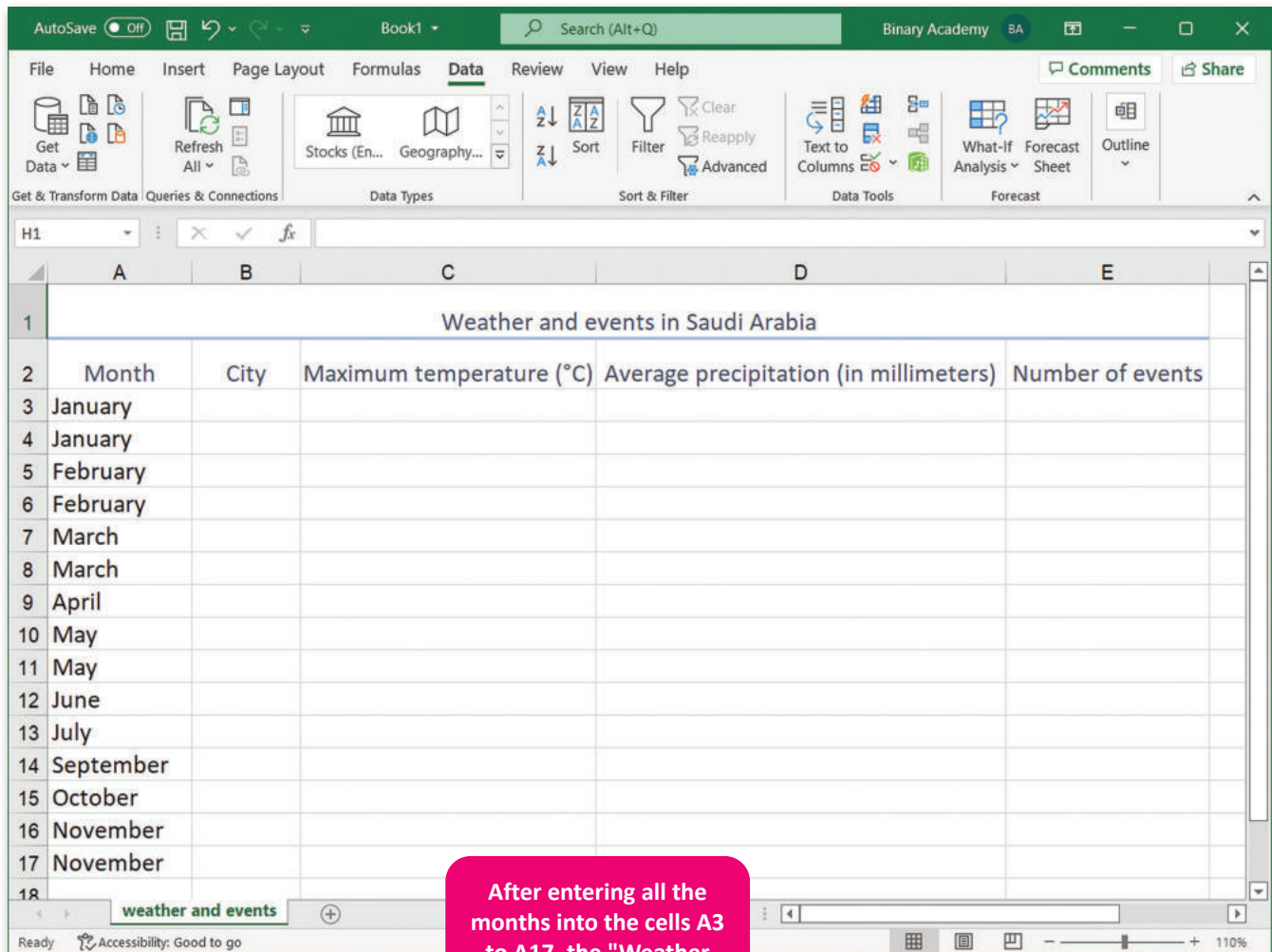


Figure 2.27: Validated data table

After entering all the months into the cells A3 to A17, the "Weather and Events in Saudi Arabia" Excel Sheet will look like this.



Length Validation Check in Excel

Now we will continue with the second column, which is the "City" column. Before entering the names of the cities, we will perform the length validation check, so that we can only be able to enter values from 3 to 6 characters in length.

To start the validation process:

- > Go to the "Weather and Events in Saudi Arabia" Excel Sheet. **1**
- > Choose cells from **B3** to **B17**. **2**
- > Choose the **Data** tab. **3**
- > In the **Data Tools** group **4**, click **Data Validation**. **5**
- > The **Data Validation** window appears. **6**

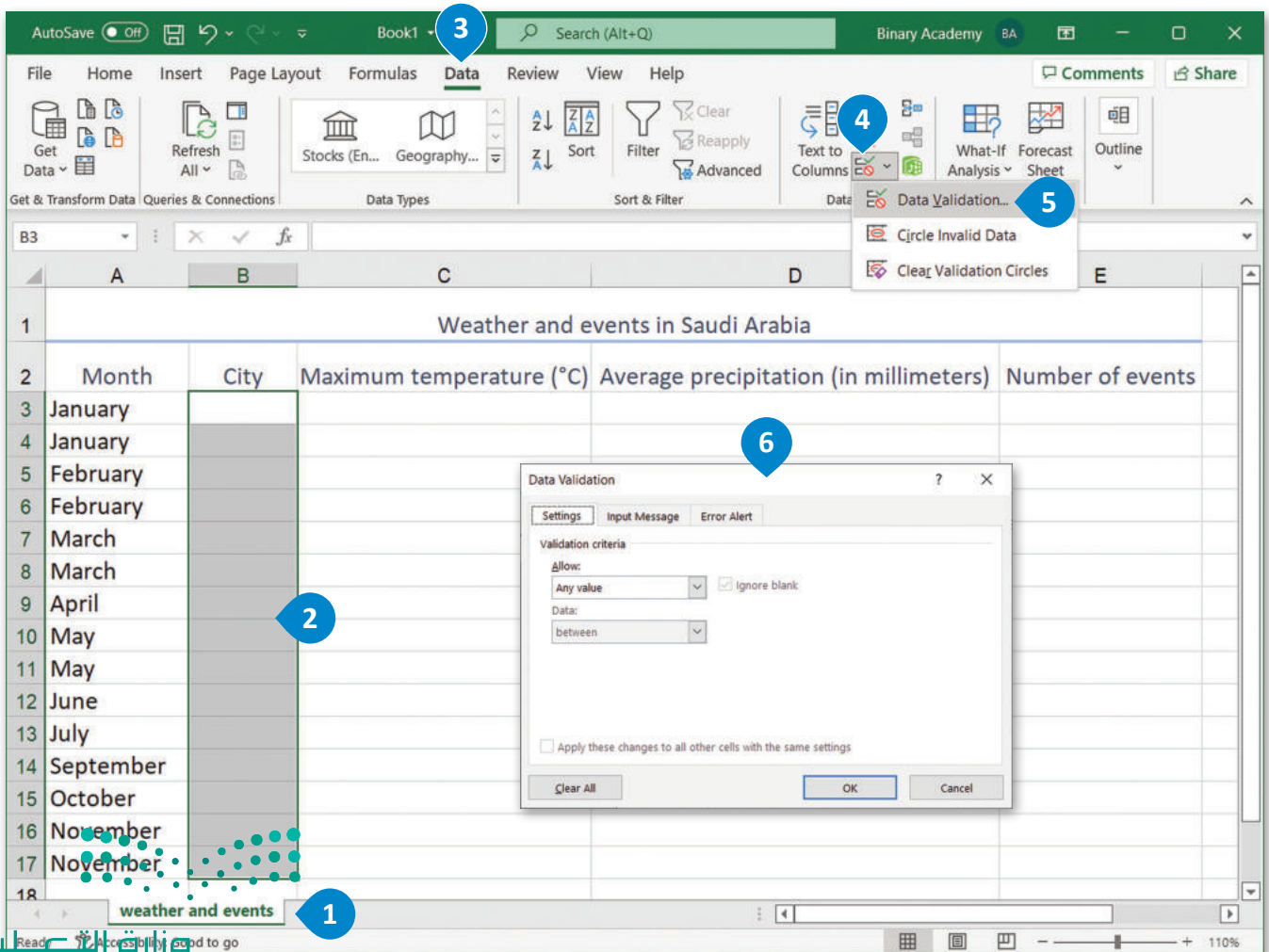


Figure 2.28 Start the validation process

To apply Length validation check in Excel:

- > In the **Data Validation** window **1** , choose the **Settings** tab. **2**
- > In the **Allow** box, choose **Text length**. **3**
- > In the **Data** box, choose **between**. **4**
- > In the **Minimum** box, type **3** and in the **Maximum** box, type **6**. **5**
- > Uncheck the **Ignore blank** option. **6**

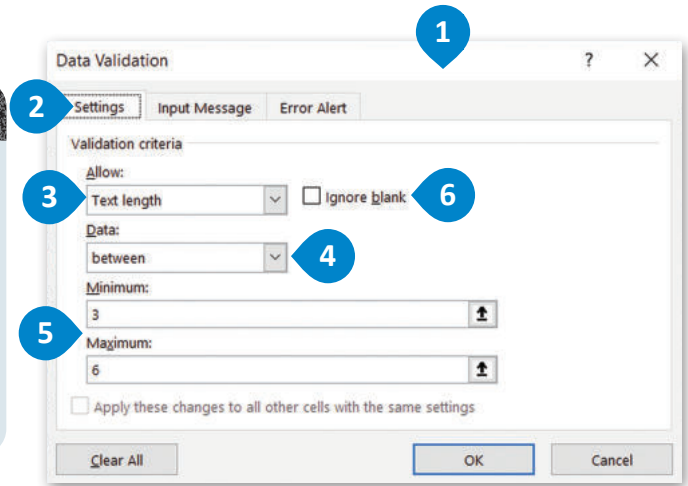


Figure 2.29: Apply length validation check in Excel

To set an invalid Input Message:

- > In the **Data Validation** window **1** , choose the **Input Message** tab. **2**
- > In the **Title** box, write "Data entry instruction". **3**
- > In the **Input message** box, write "Enter a city name between 3 and 6 characters". **4**

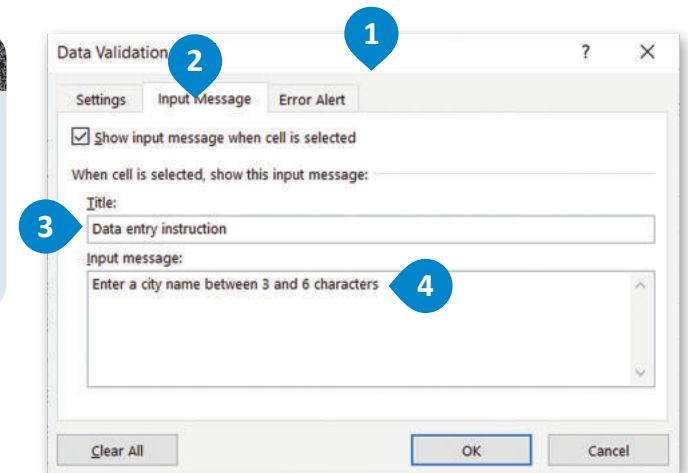
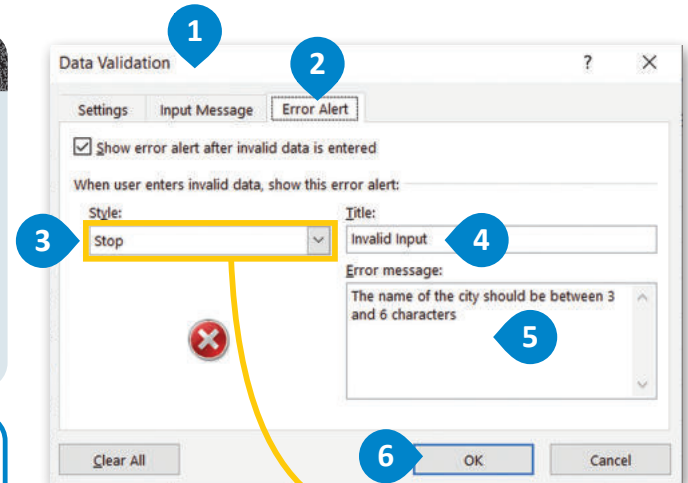


Figure 2.30: Set an invalid input message

To set an Error message:

- > In the **Data Validation** window **1** , choose the **Error Alert** tab. **2**
- > In the **Style** box choose **Stop**. **3**
- > In the **Title** box write "Invalid Input". **4**
- > In the **Error message** box, write "The name of the city should be between 3 and 6 characters". **5**
- > Click **OK**. **6**



The "Warning" style discourages the entry of invalid data. The error message icon is a yellow triangle with black exclamation mark.

The "Information" style announces the entry of invalid data. The error message icon is a white speech bubble, with blue lower-case "i".

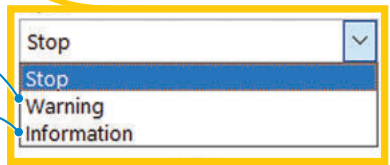


Figure 2.31: Set an error message

Now that we've finished with the Length validation check, we can start entering the data of cities. For every cell (from B3 to B17) that we enter a value, we are allowed to write a city that consists of 3 and up to 6 characters. The input message is constantly shown and if we accidentally enter a value that is less than 3 characters or more than 6 characters, the error message that we've previously set will appear on the screen.

If we accidentally type a value in the cities' column that does not meet the criteria that we've already set, Excel will display the Error message that we've set during the validation.

The input message will always appear while we type the names of the Cities in the column "City".

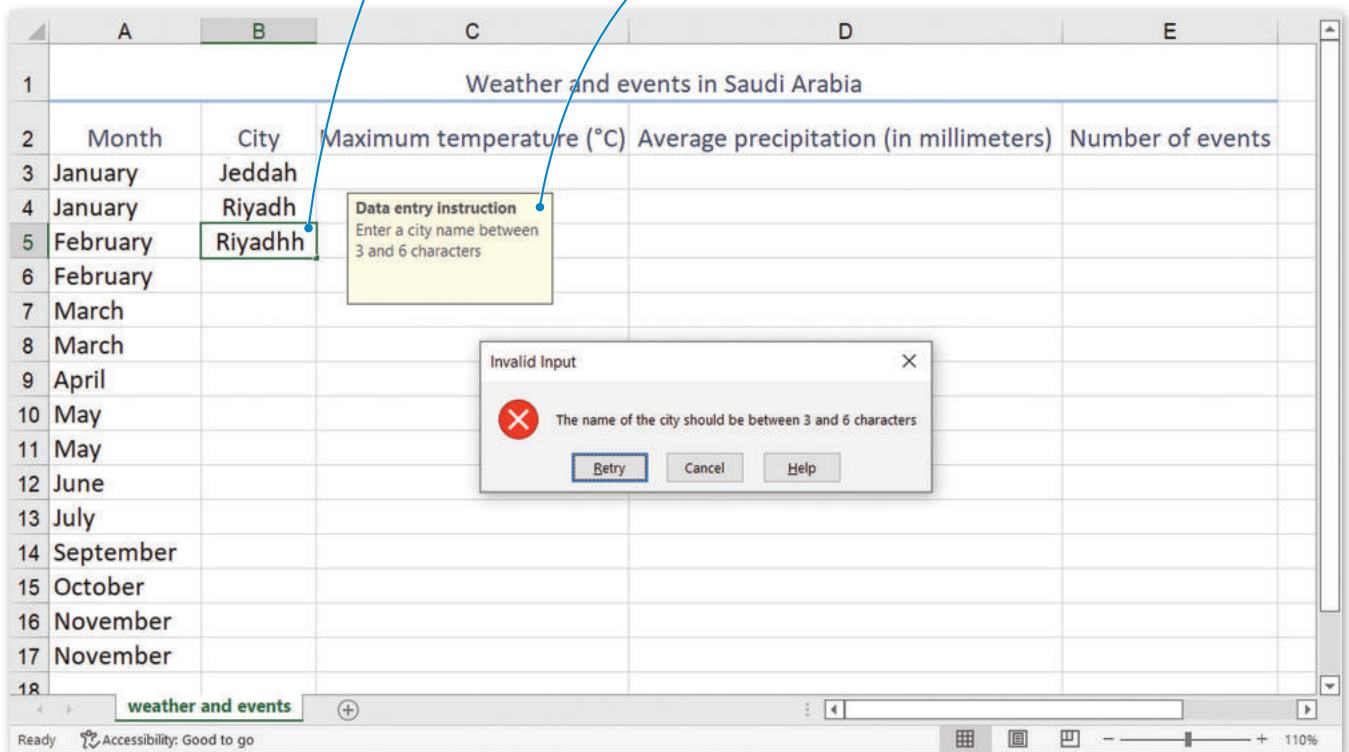


Figure 2.32: Input and error message of length validation



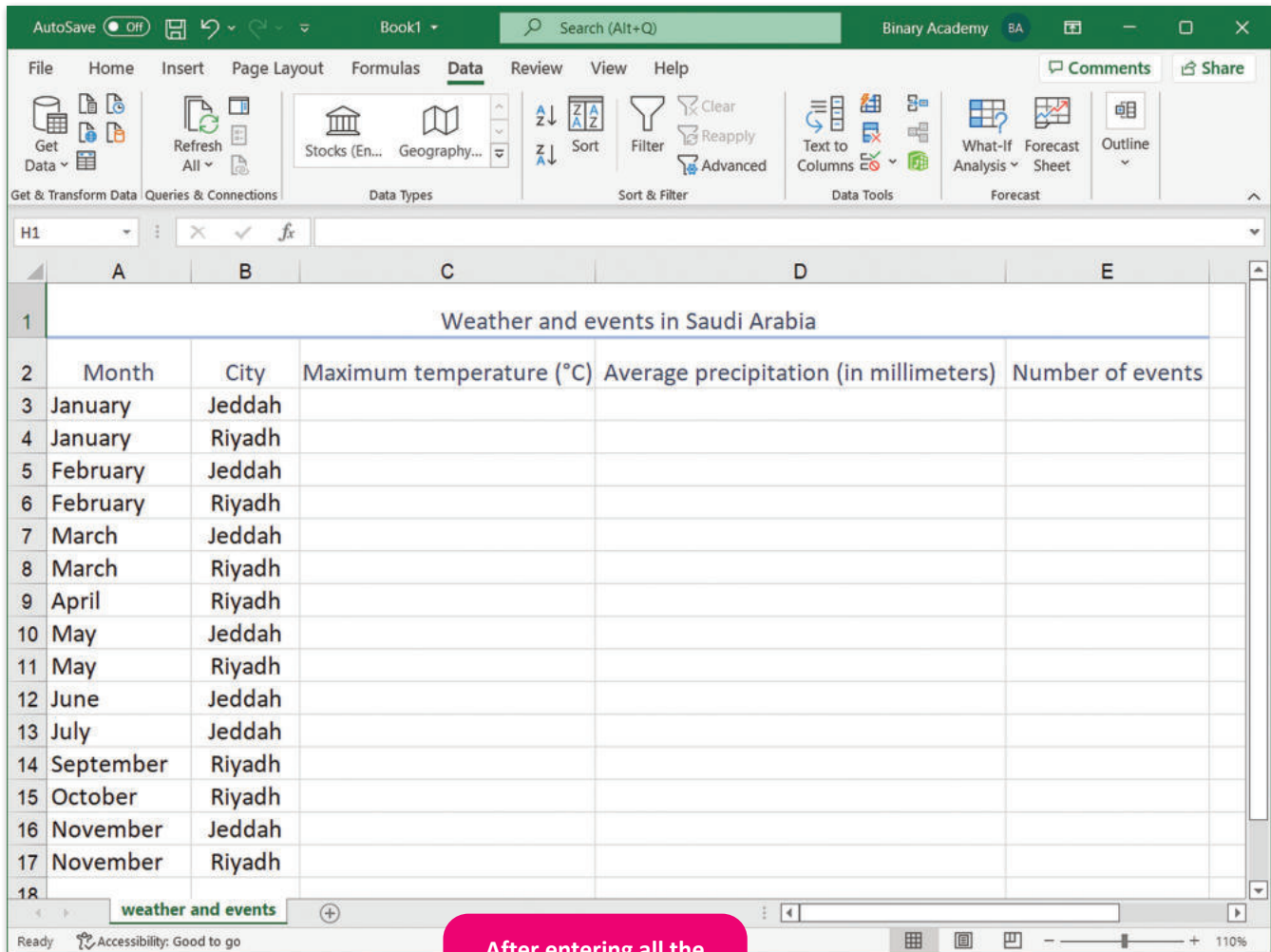


Figure 2.33: Validated data table

After entering all the cities into the cells B3 to B17, the "Weather and Events in Saudi Arabia" Excel Sheet will look like this.



Range Validation Check in Excel

Now we will continue with the third column, which is the "Temperature maximum (°C)" column. Before entering the temperature values, we will perform the Range validation check, so that we can only be able to enter temperature values from 20 to 45 degrees Celsius.

To apply Range validation check in Excel:

- > Go to the "Weather and Events in Saudi Arabia" Excel Sheet. **1**
- > Choose cell **C3**. **2**
- > From the **Data** tab, in the **Data Tools** group, click **Data Validation**. **3**
- > In the **Data Validation** window, choose the **Settings** tab. **4**
- > In the **Allow** box, choose **Custom**. **5**
- > In the **Formula** box, type **=AND(C3>20;C3<45)**. **6**
- > Uncheck the **Ignore blank** option **7** and click **OK**. **8**
- > Use the Autofill tool to apply the validation to cells **C4** through **C17**. **9**

The "**=AND(C3>20;C3<45)**" formula means that the value that we will enter in the cell C3 must be greater than 20 degrees Celsius and less than 45 degrees Celsius.

Month	City	Maximum temperature (°C)	Average precipitation (in millimeters)	Number of events
January	Jeddah			
January	Riyadh			
February	Jeddah			
February	Riyadh			
March	Jeddah			
March	Riyadh			
April	Riyadh			
May	Jeddah			
May	Riyadh			
June	Jeddah			
July	Jeddah			
September	Riyadh			
October	Riyadh			
November	Jeddah			
November	Riyadh			

1	Weather and events in Saudi Arabia				
2	Month	City	Maximum temperature (°C)	Average precipitation (in millimeters)	Number of events
3	January	Jeddah			
4	January	Riyadh			
5	February	Jeddah			
6	February	Riyadh			
7	March	Jeddah			
8	March	Riyadh			
9	April	Riyadh			
10	May	Jeddah			
11	May	Riyadh			
12	June	Jeddah			
13	July	Jeddah			
14	September	Riyadh			
15	October	Riyadh			
16	November	Jeddah			
17	November	Riyadh			

Figure 2.34: Apply range validation in Excel

To set an invalid Input Message:

- > In the **Data Validation** window **1**, choose the **Input Message** tab. **2**
- > In the **Title** box, write "Data entry instruction". **3**
- > In the **Input message** box, write "Temperature data must be within a specific range of values". **4**

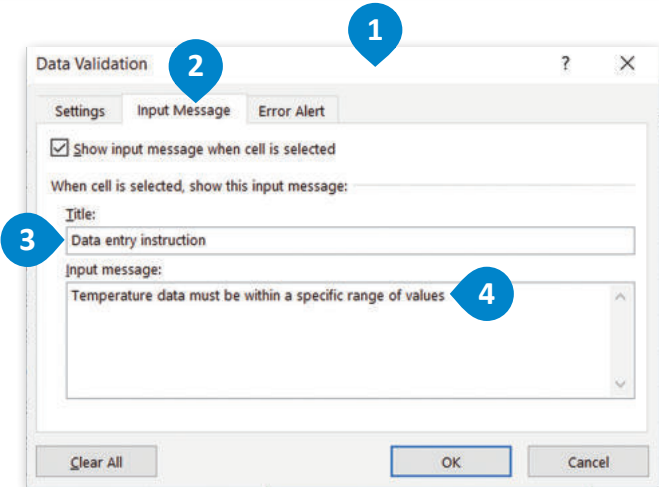


Figure 2.35: Set an invalid input message

To set an Error message:

- > In the **Data Validation** window **1**, choose the **Error Alert** tab. **2**
- > In the **Style** box choose **Stop**. **3**
- > In the **Title** box write "Invalid Input". **4**
- > In the **Error message** box, write "Temperature value must be between 20 and 45 degrees Celsius". **5**
- > Click **OK**. **6**

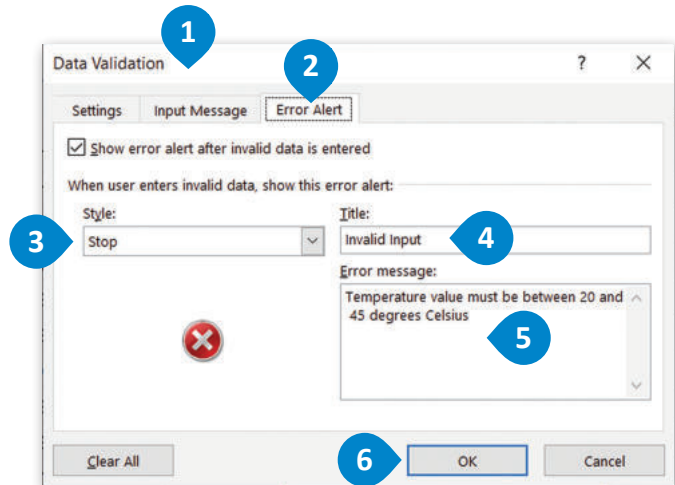


Figure 2.36: Set an error message

Now that we've finished with the Range validation check, we can start entering the data of maximum temperature (°C) values. For every cell (from C3 to C17) we are allowed to enter a temperature value that is between the range of 20 degrees Celsius to 45 degrees Celsius. The input message is shown constantly inside the cells and if we accidentally enter a value that is less than 20 degrees Celsius or greater than 45 degrees Celsius, the error message that we previously set will appear on the screen.

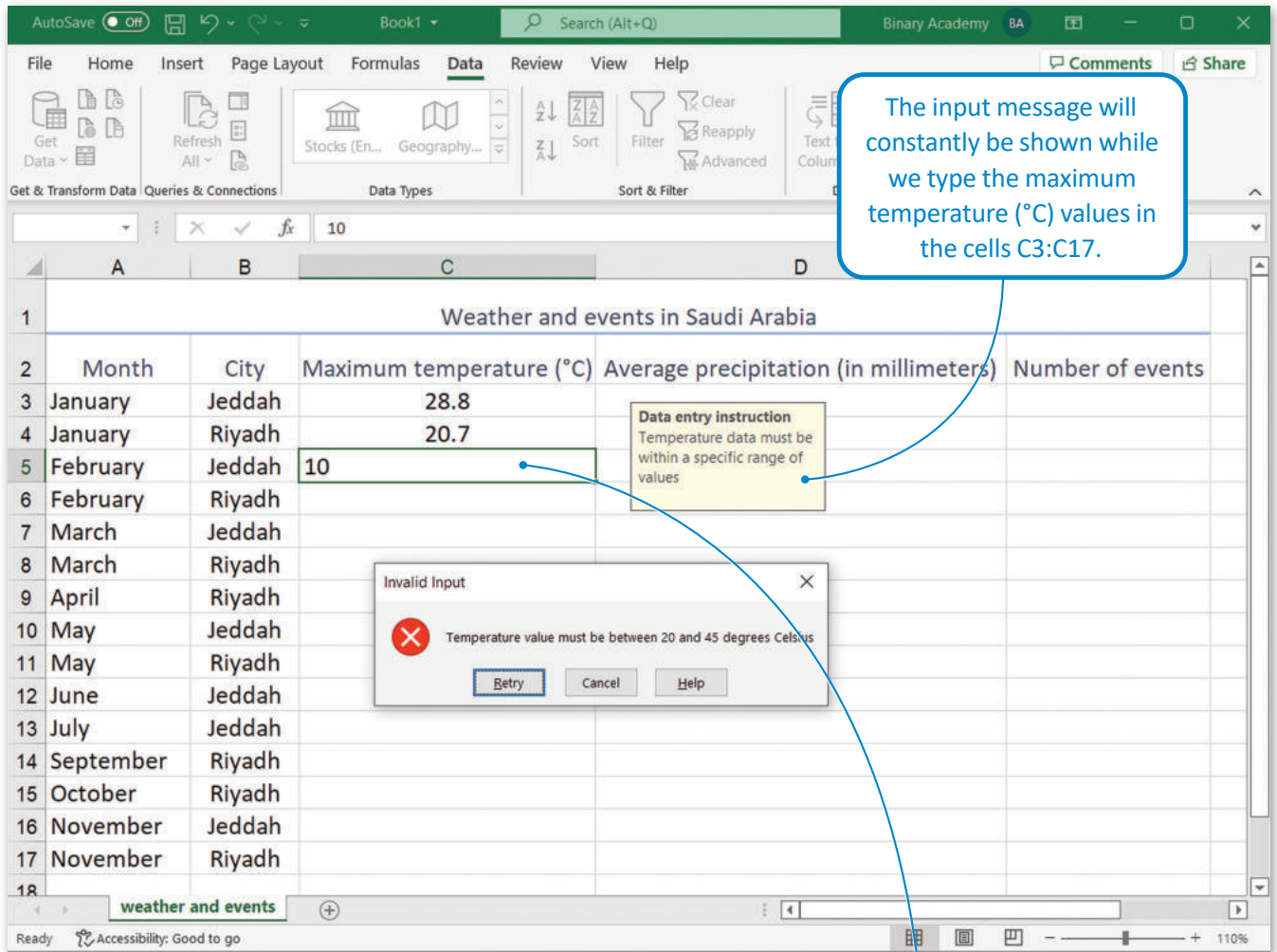


Figure 2.37: Input and error message of range validation

If we accidentally type a value in the maximum temperature column that does not meet the criteria that we already set, Excel will display the Error message that we've set during the validation.



Weather and events in Saudi Arabia				
Month	City	Maximum temperature (°C)	Average precipitation (in millimeters)	Number of events
January	Jeddah	28.8		
January	Riyadh	20.7		
February	Jeddah	29.8		
February	Riyadh	23.7		
March	Jeddah	25.5		
March	Riyadh	28.0		
April	Riyadh	33.6		
May	Jeddah	30.7		
May	Riyadh	39.5		
June	Jeddah	38.2		
July	Jeddah	39.4		
September	Riyadh	32.8		
October	Riyadh	27.5		
November	Jeddah	27.6		
November	Riyadh	20.4		

Figure 2.38: Validated data table

After entering all the temperature values into the cells C3 to C17, the "Weather and Events in Saudi Arabia" Excel Sheet will look like this.



Format Validation Check in Excel

Now we will continue with the fourth column, which is the "Precipitation average (mm)" column. Before entering the precipitation values, we will perform the Format validation check, so that we can enter, not only integer values, but also decimals. It is a procedure that will ask for a minimum and a maximum value to be set, so we will set the minimum precipitation average value equal to 0 and the maximum precipitation average value equal to 30.

To start the validation process:

- > Go to the "Weather and Events in Saudi Arabia" Excel Sheet. ①
- > Choose cells from D3 to D17. ②
- > Choose the **Data** tab. ③
- > In the **Data Tools** group ④, click **Data Validation**. ⑤
- > The **Data Validation** window appears. ⑥

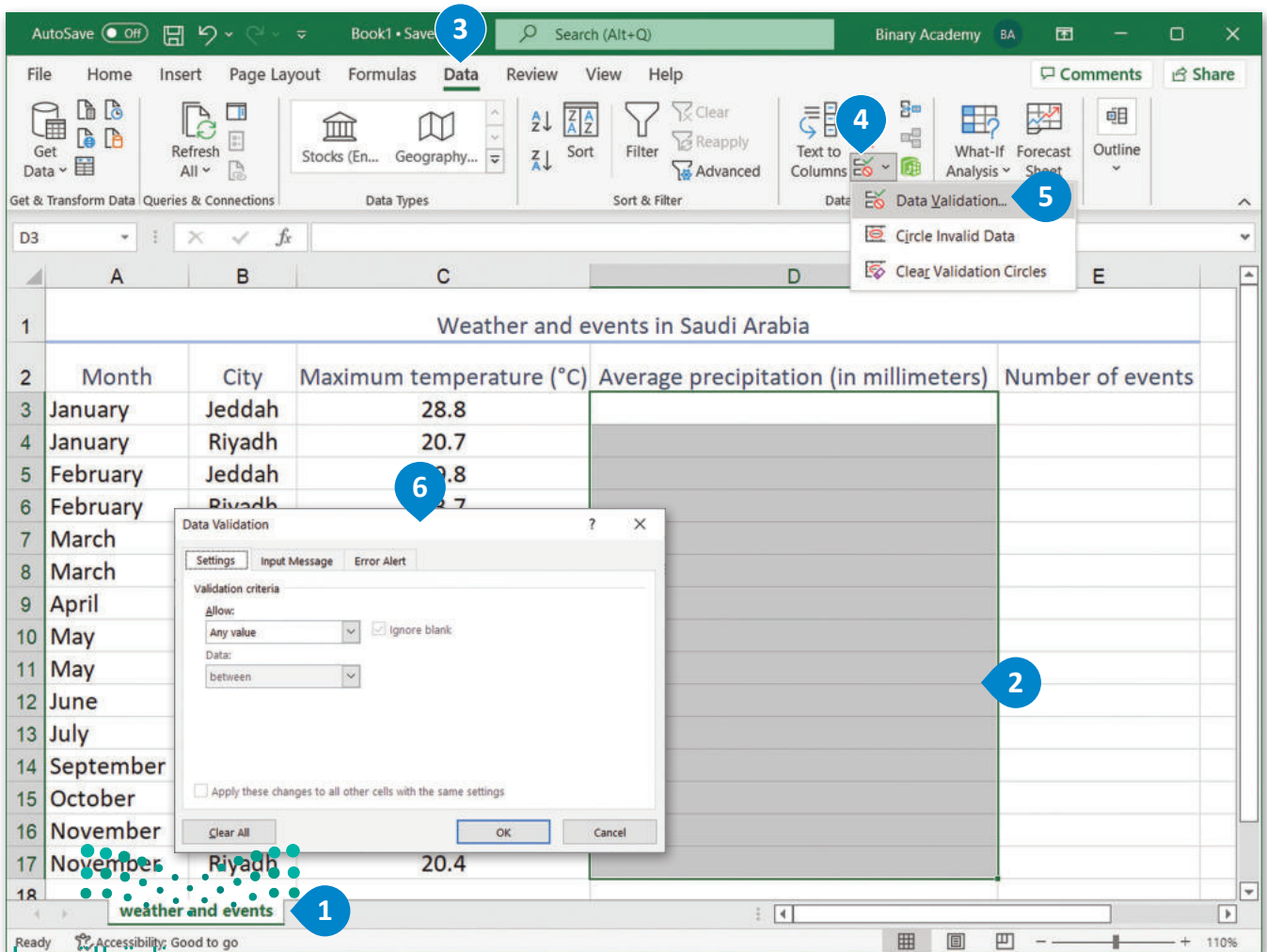


Figure 2.39: Start data validation in Excel
Ministry of Education

To apply Format validation check in Excel:

- > In the **Data Validation** window **1** , choose the **Settings** tab. **2**
- > In the **Allow** box, choose **Decimal**. **3**
- > In the **Data** box, choose **between**. **4**
- > In the **Minimum** box, type **0** and in the **Maximum** box, type **30**. **5**
- > Uncheck the **Ignore blank** option. **6**

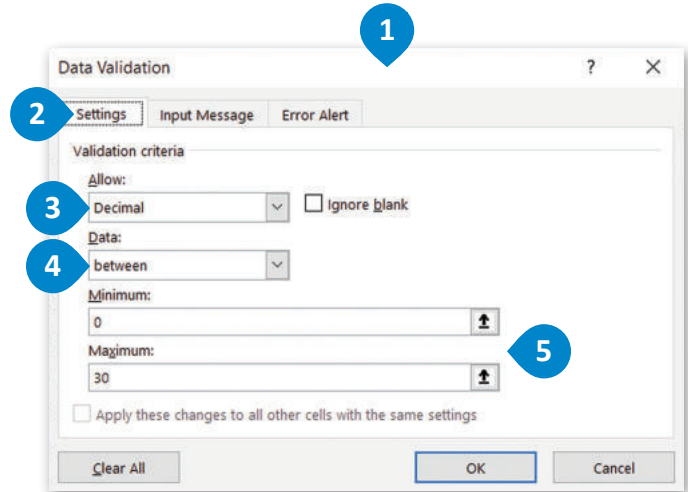


Figure 2.40: Apply format validation in Excel

To set an invalid Input Message:

- > In the **Data Validation** window **1** , choose the **Input Message** tab. **2**
- > In the **Title** box, write "Data entry instruction". **3**
- > In the **Input message:** box, write "Precipitation values must be in decimal form". **4**

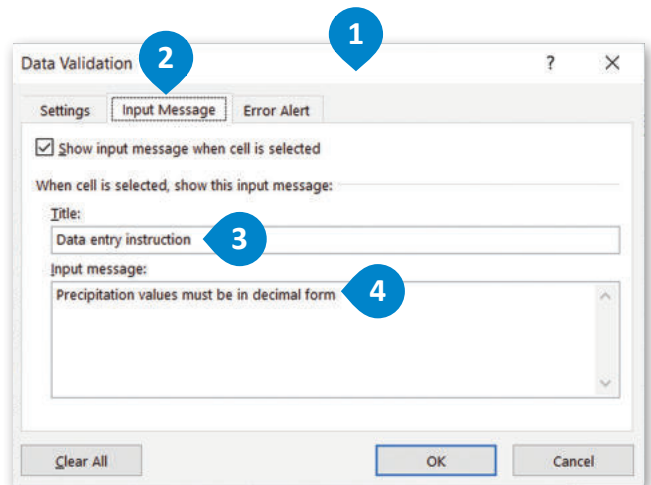


Figure 2.41: Set an invalid input message

To set an Error message:

- > In the **Data Validation** window **1** , choose the **Error Alert** tab. **2**
- > In the **Style** box, choose **Stop**. **3**
- > In the **Title** box, write "Invalid Input". **4**
- > In the **Error message** box, write "Precipitation values are not in decimal form". **5**
- > Click **OK**. **6**

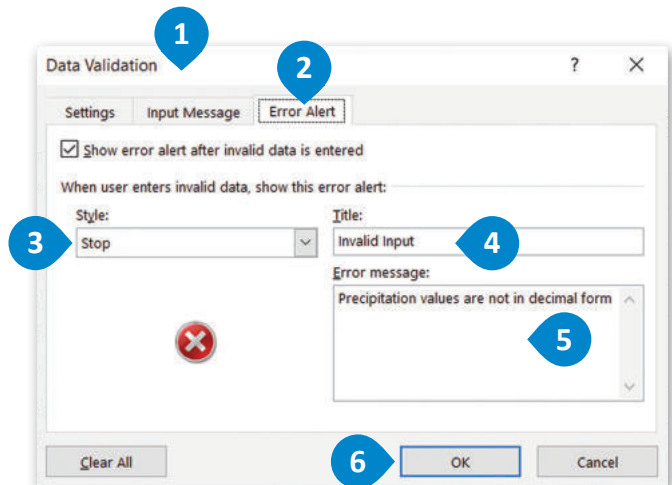


Figure 2.42: Set an error message

Now that we've finished with the Format validation check, we can start entering the data of precipitation. For every cell (from D3 to D17) we enter a value, we are allowed to enter a precipitation value that is in decimal form and between the range of 0 mm to 30 mm. The input message is constantly shown and if we accidentally enter a value that is less than 0 mm or greater than 30 mm, the error message that we previously set, will appear on the screen.

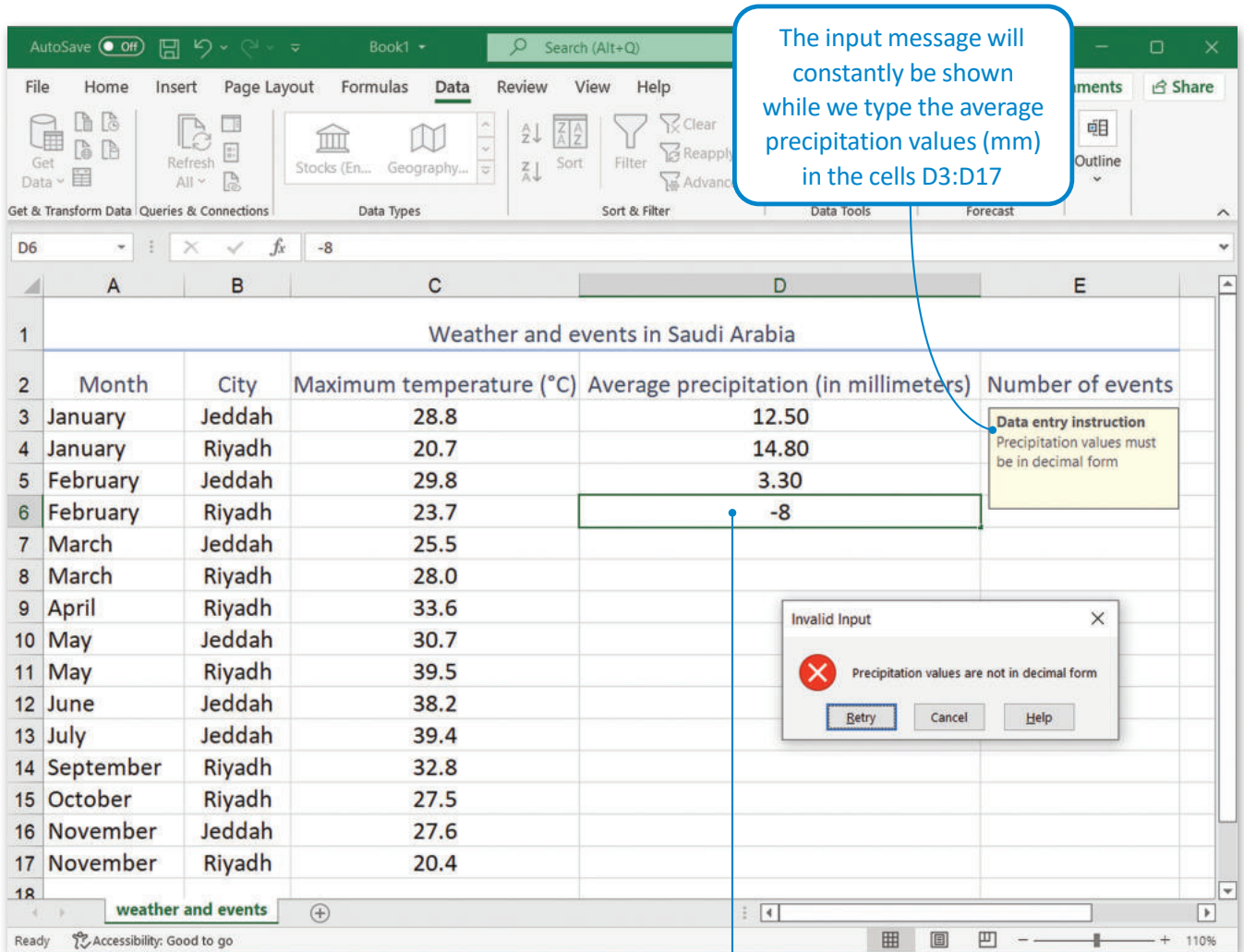


Figure 2.43: Input and error message of range validation

If we accidentally type a value in the precipitation's column that does not meet the criteria that we've already set, Excel will display the Error message that we've set during the validation.



Weather and events in Saudi Arabia				
Month	City	Maximum temperature (°C)	Average precipitation (in millimeters)	Number of events
January	Jeddah	28.8	12.50	
January	Riyadh	20.7	14.80	
February	Jeddah	29.8	3.30	
February	Riyadh	23.7	8.30	
March	Jeddah	25.5	2.60	
March	Riyadh	28.0	19.90	
April	Riyadh	33.6	23.70	
May	Jeddah	30.7	0.10	
May	Riyadh	39.5	5.60	
June	Jeddah	38.2	0.00	
July	Jeddah	39.4	0.40	
September	Riyadh	32.8	0.00	
October	Riyadh	27.5	1.50	
November	Jeddah	27.6	27.10	
November	Riyadh	20.4	20.00	

Figure 2.44: Validated data table

After entering all the precipitation values into the cells D3 to D17, the "Weather and Events in Saudi Arabia" Excel Sheet will look like this.



Type Validation Check in Excel

Now we will continue with the fifth column, which is the "Number of events" column. Before entering the number of events for each city, we will perform the Type validation check, so that we cannot enter negative values. It is a procedure that demands a minimum value to be set, so we will set a minimum value equal to 1 because, apart from negative values, we also don't want the event values to be equal to zero.

To start the validation process:

- > Go to the "Weather and Events in Saudi Arabia" Excel Sheet. ①
- > Choose cells from E3 to E17. ②
- > Choose the **Data** tab. ③
- > In the **Data Tools** group ④, click **Data Validation**. ⑤
- > The **Data Validation** window appears. ⑥

The screenshot shows the Excel interface with the 'Data' tab selected. The 'Data Validation' dialog box is open, showing the 'Settings' tab. The 'Validation criteria' section is set to 'Any value' and 'Ignore blank' is checked. The 'Data' section is set to 'between'. The spreadsheet shows a table with columns for Month, City, Maximum temperature, Average precipitation, and Number of events. The 'Number of events' column (E3:E17) is highlighted. The 'Data Validation' dialog box is open, showing the 'Settings' tab. The 'Validation criteria' section is set to 'Any value' and 'Ignore blank' is checked. The 'Data' section is set to 'between'. The spreadsheet shows a table with columns for Month, City, Maximum temperature, Average precipitation, and Number of events. The 'Number of events' column (E3:E17) is highlighted.

Month	City	Maximum temperature (°C)	Average precipitation (in millimeters)	Number of events
January	Jeddah	28.8	12.50	
January	Riyadh	20.7	14.80	
February	Jeddah	20.8	3.30	
February	Riyadh	22.7	8.30	
March			2.60	
March			19.90	
April			23.70	
May			0.10	
May			5.60	
June			0.00	
July			0.40	
September			0.00	
October			1.50	
November			27.10	
November	Riyadh	20.4	20.00	

Figure 2.45: Start data validation in Excel
Ministry of Education

To apply Type validation check in Excel:

- > In the **Data Validation** window **1**, choose the **Settings** tab. **2**
- > In the **Allow** box, choose **Whole number**. **3**
- > In the **Data** box, choose **greater than or equal to**. **4**
- > In the **Minimum** box, type **1**. **5**
- > Uncheck the **Ignore blank** option. **6**

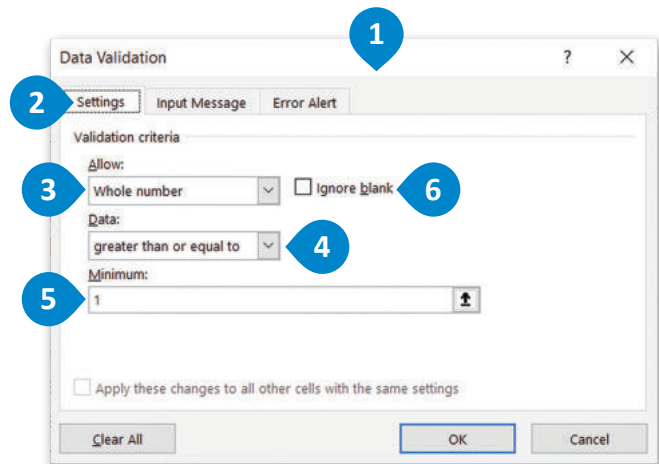


Figure 2.46: Apply type validation in Excel

To set an invalid Input Message:

- > In the **Data Validation** window **1**, choose the **Input Message** tab. **2**
- > In the **Title** box, write "Data entry instruction". **3**
- > In the **Input message** box, write "Enter a non-negative integer number". **4**

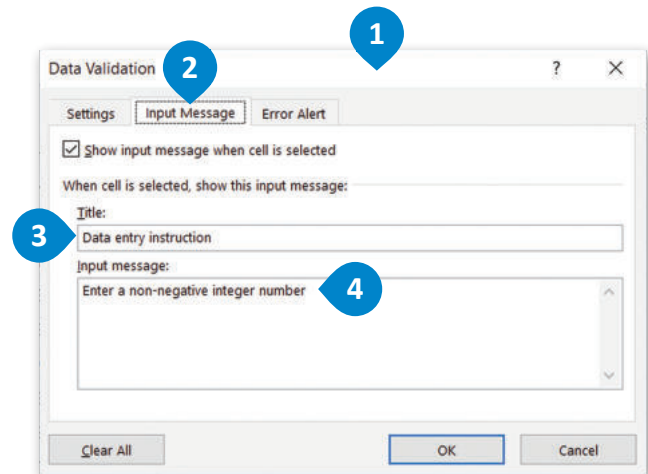


Figure 2.47: Set an invalid input message

To set an Error message:

- > In the **Data Validation** window **1**, choose the **Error Alert** tab. **2**
- > In the **Style** box, choose **Stop**. **3**
- > In the **Title** box, write "Invalid Input". **4**
- > In the **Error message** box, write "The number of events can't be negative". **5**
- > Click **OK**. **6**

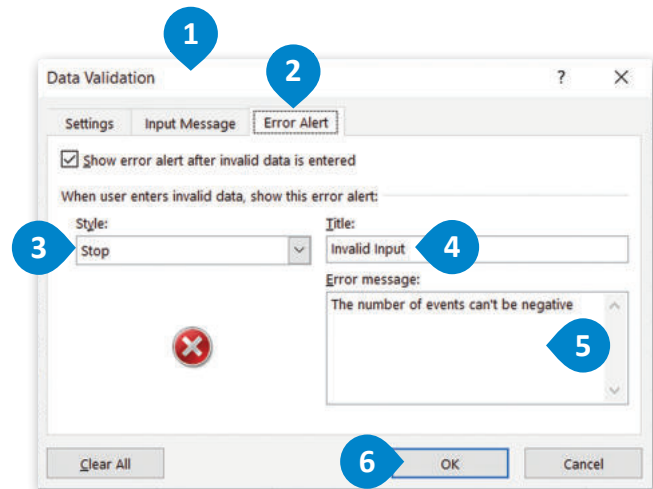


Figure 2.48: Set an error message

Now that we've finished with the Type validation check, we can start entering the data of events (based on the table 2.8). For every cell (from E3 to E17) we enter a value, we are allowed to enter a number that is equal to or greater than 1. The input message is shown constantly and if we accidentally enter a value that is less than 1, the error message that we previously set will appear on the screen.

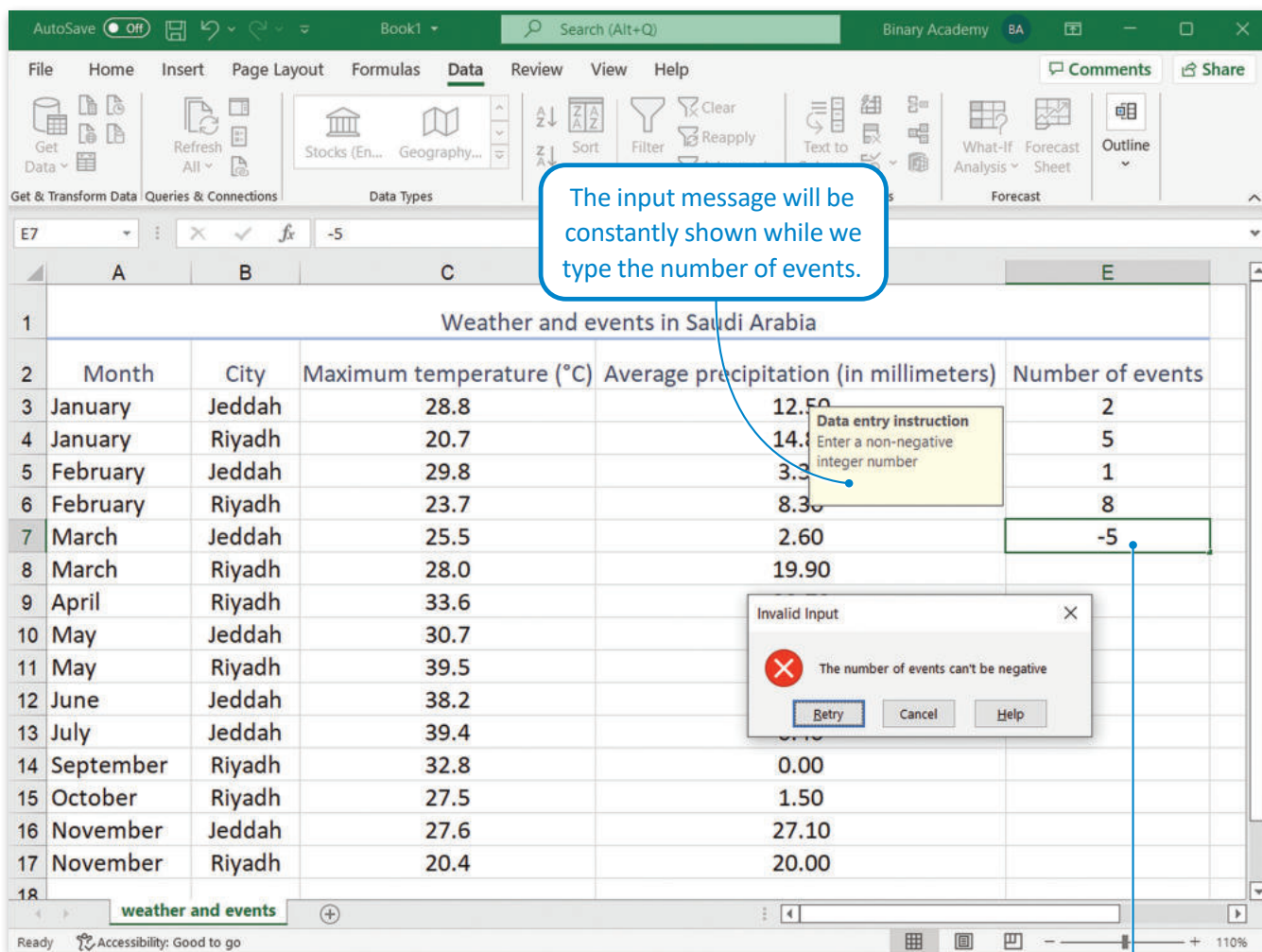


Figure 2.49: Input and error message of type validation

If we accidentally type a value in the events column that does not meet the criteria that we've already set, Excel will display the Error message that we set during the validation.



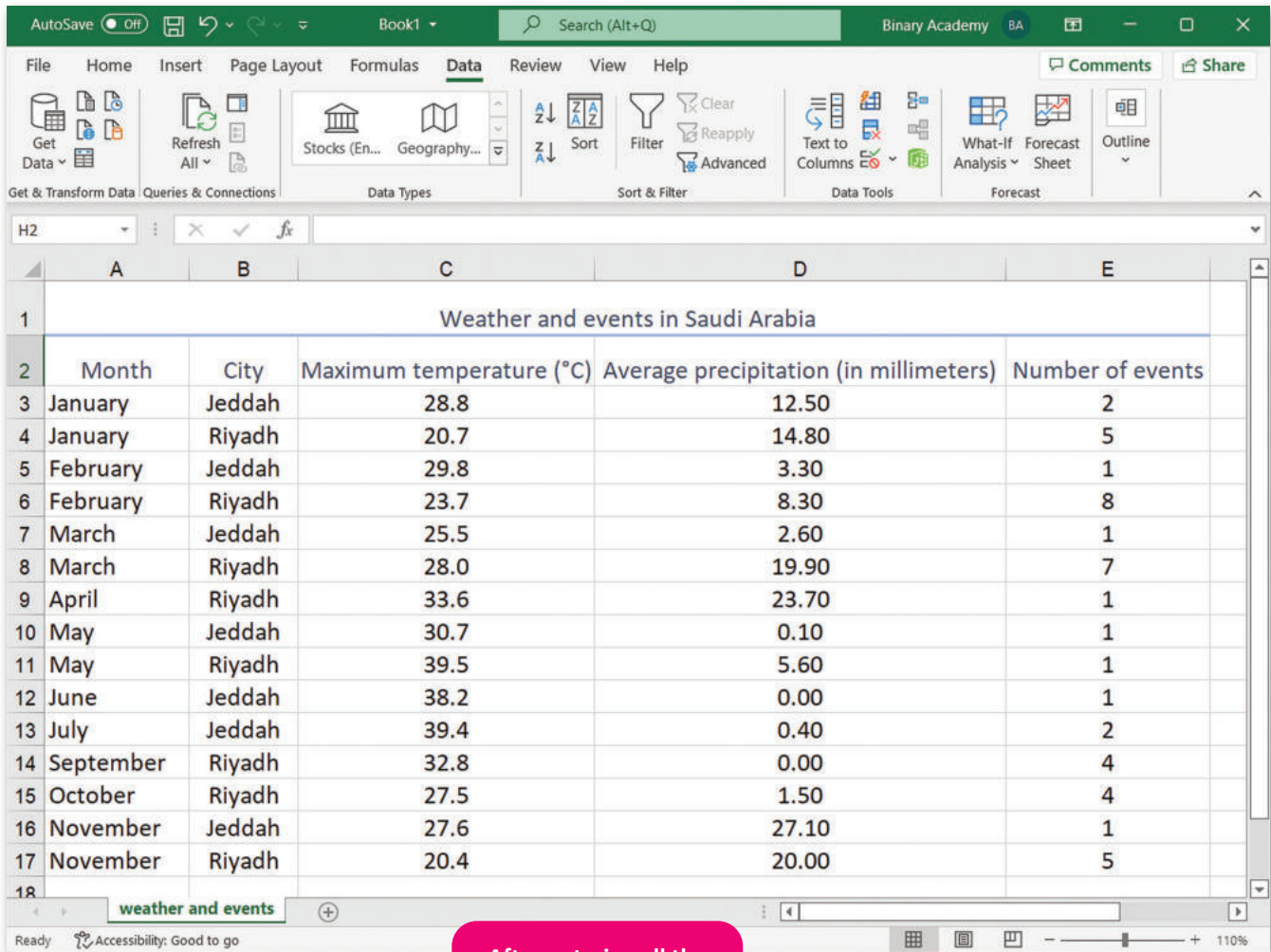


Figure 2.50: Validated data table

After entering all the events into the cells E3 to E17, the "Weather and Events in Saudi Arabia" Excel Sheet will look like this.



Exercises

1

Read the sentences and tick ✓ True or False.	True	False
1. Data validation refers to the procedure that automatically deletes any raw data that does not meet certain criteria.	<input type="radio"/>	<input type="radio"/>
2. There are only five types of data entry validation.	<input type="radio"/>	<input type="radio"/>
3. The Presence check helps us reduce errors by using a limited list of predefined values.	<input type="radio"/>	<input type="radio"/>
4. The LookUp check aims to ensure that characters and symbols have specific lengths.	<input type="radio"/>	<input type="radio"/>
5. The Range check is used to ensure that the entered numbers fall within a certain range.	<input type="radio"/>	<input type="radio"/>
6. The Format check ensures that data has a specific format.	<input type="radio"/>	<input type="radio"/>
7. The Type check helps us reduce language errors.	<input type="radio"/>	<input type="radio"/>
8. The Check digit is used if we want to ensure that a set of numbers is entered correctly.	<input type="radio"/>	<input type="radio"/>
9. Microsoft Excel is the only tool for data validation.	<input type="radio"/>	<input type="radio"/>
10. The data validation can be performed after we enter the values in a data validation program.	<input type="radio"/>	<input type="radio"/>

4 Create an address book table of your friends' information which will include the fields: Name, Telephone, Home address, Email address, Birthday, Hobby. Next, write down what type of data validation should be performed on each field.

5 Compare the following: (a) Length check vs Range check, (b) Format check vs Type check. Give examples of the use of each validation type.



Project

1

Let's suppose that you are working as a health researcher and you want to make a report about the problem of diabetes in your country. After collecting the data, explain how the validation checks will be performed on the data.

2

More specifically, you have to answer questions like:

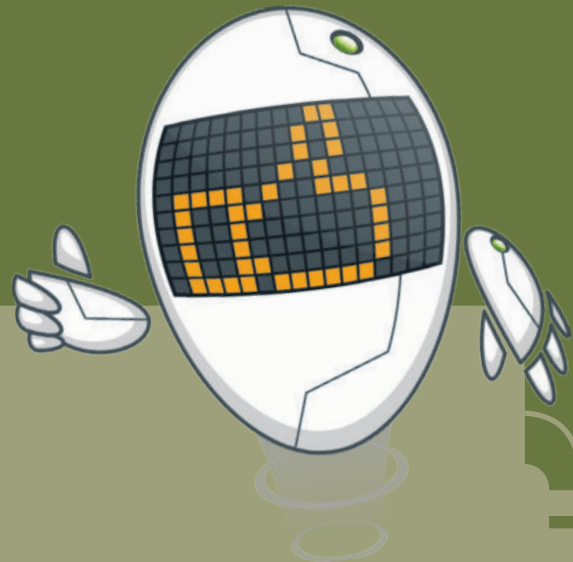
- What kind of columns will you create?
- Which of the six validation checks will be performed in each column and why?

3

Create a presentation in PowerPoint, explaining the steps of your validation procedure.



Wrap up



Now you have learned:

- > what data collection is.
- > the different types of data.
- > how data is coded.
- > how data is validated.
- > how to check the quality of information.
- > how to distinguish primary data sources from secondary data sources.

KEY TERMS

Accuracy	Data Types	Level of Detail
Appropriateness	Data Validation	LookUp Check
Audio Data	Fixed Data	Open Data Platforms
Barcode	Format Check	Presence Check
Check Digit	Graphical Data	QR Code
Completeness	Information Quality	Range Check
Data	ISBN	Type Check
Data Coding	Legal Permissions	Variable Data
Data Collection	Length Check	Video Data

3. Exploratory Data Analysis

In the previous units, you learned what data is, the different types of data and how to properly collect data. In this unit, you will learn about how to explore and analyze the data you have in order to better understand the data.



Learning Objectives

In this unit, you will learn to:

- > Categorize the types of data analysis.
- > Define what exploratory data analysis is.
- > Categorize the types of exploratory data analysis.
- > List the stages of the exploratory data analysis process.
- > Define what a programming library is.
- > Develop a data analysis program using programming libraries.
- > Use data preparation and cleaning techniques in a dataset.
- > Discuss the importance of data visualization.
- > Generate different types of charts using Python libraries.

Lesson 1

Data Analysis

Link to digital lesson



www.ien.edu.sa

What is Data Analysis

We analyze many things in our everyday life, for example, when we think about what happened last time we did something and what will happen if we choose to make that particular decision again. This is nothing but analyzing our past or future and making decisions based on our analysis.

Data analysis is defined as the process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions and support decision-making.

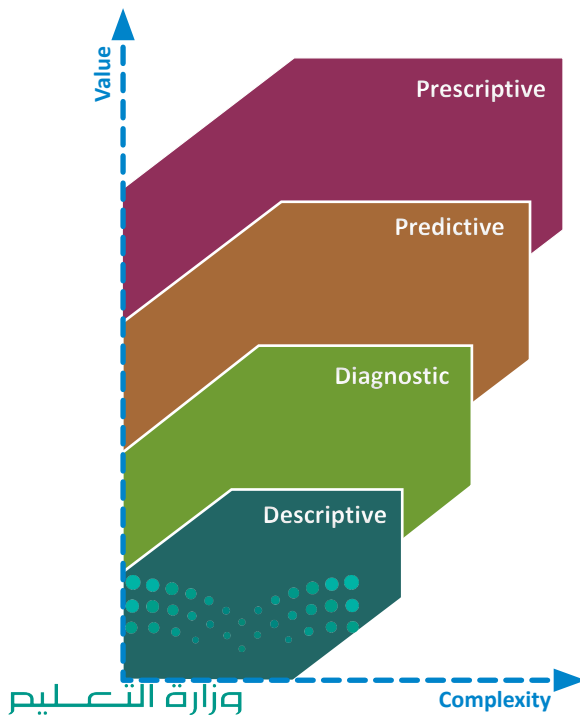
Data Analysis

A systematic examination of data through measurement, and visualization.

Types of Data Analysis

Depending on the reason you want to analyze data and the specific problem you are trying to solve, you might choose different types of analysis.

- > *Prescriptive Analysis*
- > *Predictive Analysis*
- > *Diagnostic Analysis*
- > *Descriptive Analysis*



■ Descriptive Analysis

Descriptive analysis is concerned with what has happened. It is often known as descriptive analytics or descriptive statistics and it is the act of describing or summarizing a set of data using statistical techniques. Its popularity as one of the key forms of data analysis stems from its capacity to provide accessible insights from otherwise uninterpreted data. Descriptive analytics does not make predictions about the future.

■ Diagnostic Analysis

Diagnostic data analysis is concerned with why something happened. It usually follows descriptive analysis, and it is the process by which analysts try to understand the cause of the trends and patterns that have been observed.

وزارة التعليم
Ministry of Education
Figure 3.1. Types of Data Analysis
2023 - 1445

■ Predictive Analysis

Predictive data analysis is concerned with trying to predict future outcomes based on previously discovered trends and historical data, by using modeling techniques and statistics. Predictive analysis has been used in many different cases, such as weather forecasting, insurance policies and more.

Predictive Analysis

The practice of using historical data combined with mathematical models to predict future outcomes or unknown events.

■ Prescriptive Analysis

The final stage of data analysis is prescriptive analysis, which is concerned with trying to find the optimal course of action. Based on the discoveries of the previous analysis stages, the goal of prescriptive analytics is to provide recommendations for future action. This type of analysis is especially useful in the healthcare sector where safe recommendations are needed.

Predictive and prescriptive analyses are more complex than descriptive and diagnostic ones, but they bring more added value and insights to a project.

Data Analysis Process

The data analysis process involves gathering information, processing it and exploring the data. Based on the results, you can take decisions or draw conclusions.

The steps of the data analysis process are:

- > **Data Preparation and Cleaning:** This process is where you remove white spaces, duplicate records, and basic errors. Data cleaning is mandatory before sending the information on for analysis.
- > **Exploratory Data Analysis:** In this step, you use data analysis software and other tools to help you interpret and understand the data and draw conclusions.
- > **Data Visualization:** Data visualization is the graphical representation of information and data. Data visualizations make data easier for the human brain to understand and analyze. By using visual elements like charts, graphs, and maps, data visualization makes data more accessible, understandable and usable.



Figure 3.2: Data Science Life Cycle

What is Exploratory Data Analysis

Generally, it is good practice to try to understand the data and gather as much insight as possible before you proceed to the modeling stage. Exploratory data analysis (EDA) is a way of making sense of the data, performing initial investigations and summarizing their main characteristics. The main goals of EDA are to discover trends, patterns and new features in the data. You can also spot anomalies in a dataset, test your initial hypothesis and get a better understanding of dataset variables and the relationships between them. EDA can also help you identify obvious errors and ensure that the results of a specific task are valid and applicable to any desired outcome. Because deriving insights by looking at plain numbers can be tedious, boring and even overwhelming, EDA has been developed as an aid in this process. All these are being achieved with the help of statistical summary, graphical representations and data visualization methods. Once EDA is completed and you have drawn enough insights from the data, then you can use these features to carry out more sophisticated data analyses such as machine learning.

Exploratory Data Analysis

The approach to analyzing datasets by summarizing their main characteristics, often with visual methods.

Types of Exploratory Data Analysis

Exploratory data analysis is generally cross classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate). Univariate analysis means that the effect of only one independent variable is analyzed, while multivariate analysis which is more common in big projects, analyzes the effect of more than one independent variable.

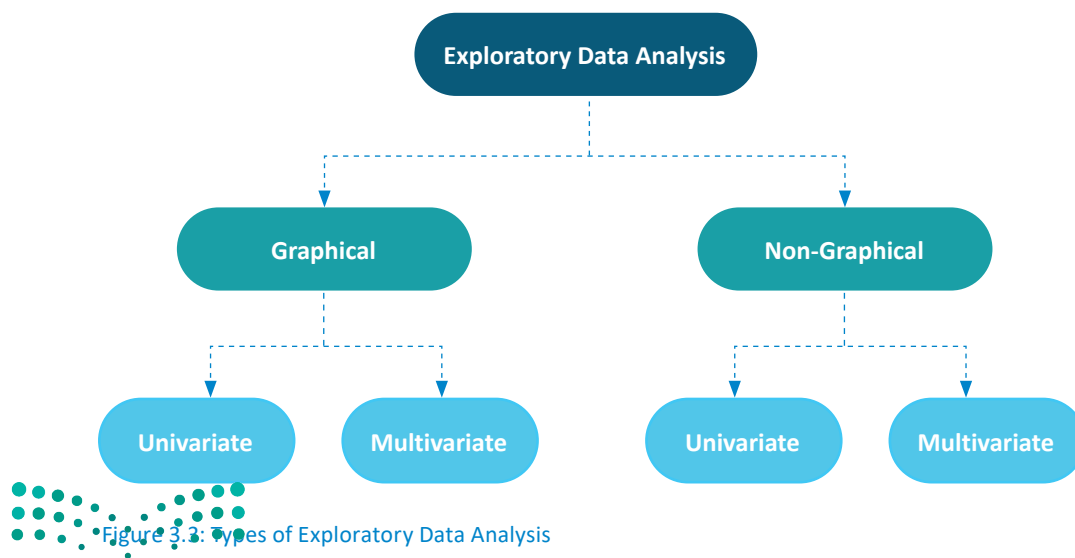


Figure 3.3: Types of Exploratory Data Analysis

Non-Graphical Analysis

Univariate Non-Graphical Analysis

An example of a univariate non-graphical analysis could be the effect age has on the probability of developing some types of disease such as Alzheimer's. This analysis is univariate because only the effect of age is being measured. It is also non-graphical because no visualization techniques are used.

Multivariate Non-Graphical Analysis

If in the previous example you took into account the effects of diet, mental exercise, and also heredity, this analysis would be a multivariate non-graphical analysis.

Graphical Analysis

Univariate Graphical Analysis

An example of a univariate graphical analysis is shown in Figure 3.4. It is a bar chart of candy bars in which each bar represents the percentage of sugar that the candy bar contains. This is a univariate graphical analysis because only one variable is taken into consideration, and it is shown graphically.

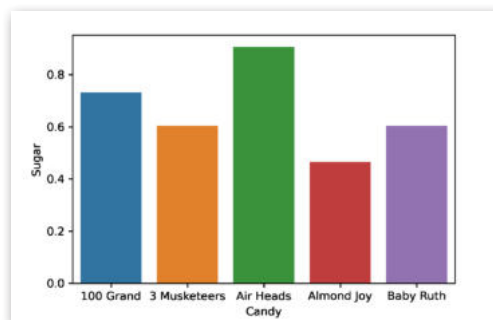


Figure 3.4: Univariate graphical analysis

Multivariate Graphical Analysis

An example of a multivariate graphical analysis is shown in Figure 3.5. It is a scatter plot of candy bars in which the x-axis is the sugar content, the y-axis is the price, and it is also color coded based on whether the candy has chocolate or not. You will learn about scatter plots and other types of data visualization later in this unit. This is a multivariate graphical analysis because three variables are taken into consideration, and their relationship is shown graphically.

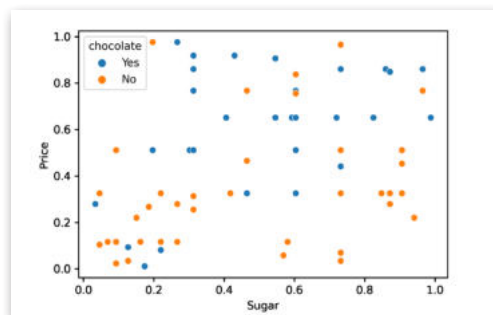


Figure 3.5: Multivariate graphical analysis

Data Analysis Tools

There are many tools we can use to process, manipulate and analyze the relationships and correlations between datasets, and these tools also help us identify patterns and trends for interpretation. To choose a data analytics tool, you must first understand your needs. The most popular and widely used analytical tool in almost all industries is Excel. In addition to spreadsheet programs, data analysis can also be conducted in specialized programming languages and environments. The most popular environments are Jupyter Notebook, RStudio and MATLAB.

In this unit, you will use Jupyter Notebook as a data analysis tool.

وزارة التعليم

Ministry of Education

2023 - 1445

Data Analysis with Python

As we mentioned earlier, Python can be used in Data Analysis. It is one of the most commonly used languages for Data Science projects by both data scientists and software developers. It can be used to forecast results, automate jobs, streamline operations, and provide business intelligence. To perform a data analysis with Python, you can use Python libraries.



Python Libraries/Modules

A library is typically a collection of books or a location where many books are kept for later use. In programming, a library is a collection of pre-written code and subroutines that a program can use. It is designed to help both the programmer and the programming language compiler to create a program. In order to use a library, you have to include it in your code. To use a library in Python, you have to use the command "import" and the name of the library.

A library in programming languages such as Python is a collection of precompiled code routines that can be utilized later in a program for specific, well-defined operations. Compared to other programming languages, a library does not pertain to any specific context in Python. A library may contain documentation, configuration data, message templates, classes, and values, among other things.

In Python, a "library" loosely describes a collection of core modules. It contains code bundles that can be reused across several programs. It simplifies and accelerates Python programming for developers because they don't have to rewrite the same code for different programs. Machine learning, data science, data visualization, and other industries rely heavily on Python libraries.

Table 3.1: Advantages and disadvantages of using code libraries

 Pros	 Cons
Fast to set up and use in your code.	If you need changes, it is very difficult or impossible to implement them.
Usually bug-free and work as expected. No debugging and testing are required.	You do not know if the library will be supported for as long as your code is in use.
Usually optimized and fast code.	
No need to learn complex algorithms to implement them.	

Python Standard Library

Python's Standard Library is a collection of the language's syntax, tokens, and semantics. It's included in the standard Python distribution. It includes modules for things like I/O (Input/Output) and other basic functions. The standard library is built around more than 200 core modules. More functionality can then be added by importing any of the thousands of other available libraries. This enormous functionality is what makes Python so popular.

Python Libraries for Data Science

Although you can work with data in plain Python, there are several open-source libraries that make data science projects considerably easier. Some of the libraries used for different tasks in data science are shown in the table.

Table 3.2: Python libraries for data science

Data science tasks	Libraries
Data mining	Scrapy, BeautifulSoup, Requests
Data processing/ Scientific computing	NumPy, SciPy, pandas, TensorFlow, Keras, scikit-learn, PyBrain, PyTorch, OpenCV, Mahotas
Data visualization	Matplotlib, seaborn, Altair, Bokeh, plotly

In this unit you will use:

- > *NumPy for numerical and mathematical operations.*
- > *Pandas library for data handling and manipulation.*
- > *Matplotlib library for data visualization.*

Jupyter Notebook is not a full IDE for Python but is optimized for data science projects.

Jupyter Notebook

In this unit, you will use Jupyter Notebook as a data analysis tool. Jupyter Notebook is an online web application to create and share computational documents. Each document, called a notebook, includes your code, comments, raw and processed data, and data visualizations. The data can be stored in an external file or integrated into the notebook. The environment supports not only Python but other programming languages as well. Furthermore, through Jupyter Notebook you can create interactive output such as HTML or videos.

In this unit, you will use the online version of the Jupyter Notebook. The easiest way to install it locally is through Anaconda, an open-source distribution platform, which is free for students and hobbyists. Download and install Anaconda from here: <https://www.anaconda.com/products/distribution>. Python and Jupyter Notebook will be installed automatically.



To open Jupyter Notebook:

- > Click **Start** 1, click **Anacoda3**. 2
- > Select **Jupyter Notebook**. 3
- > Jupyter's Notebook home page opens in the browser.

Jupyter's Notebook
home page

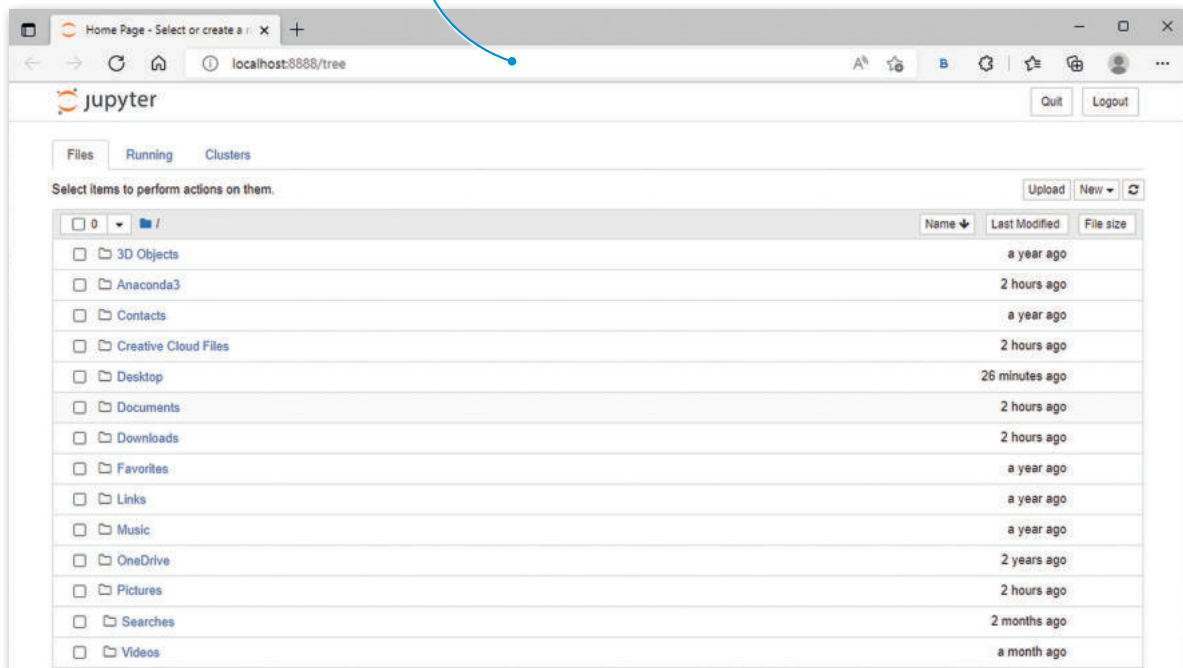
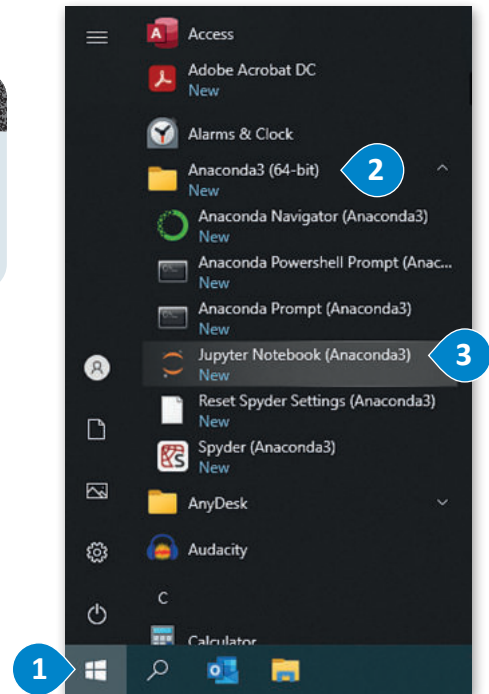


Figure 3.6: Jupyter Notebook's home page

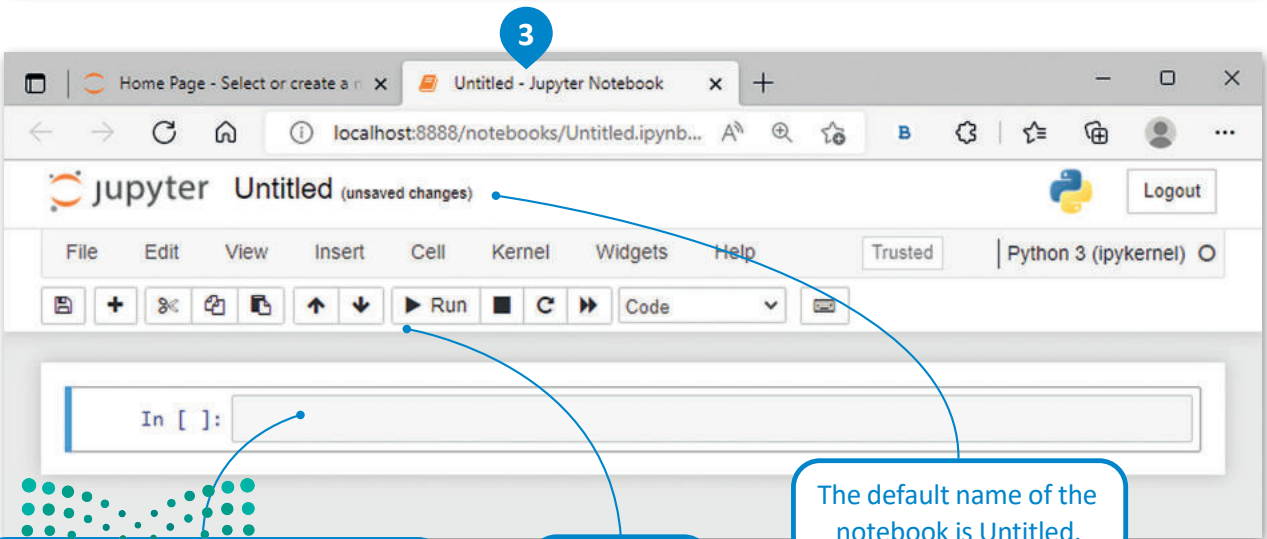
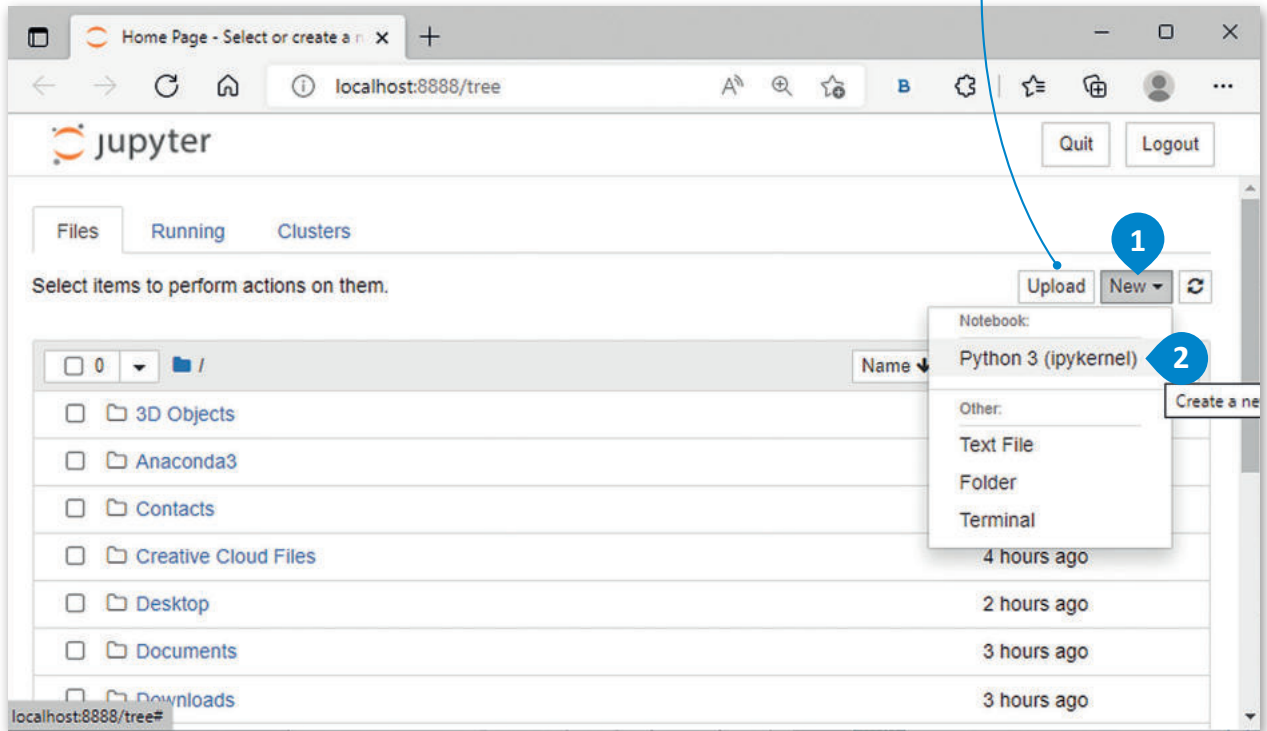
HISTORY

The American mathematician John Tukey defined data analysis in 1961 as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data".

To create a new Jupyter Notebook:

- > At the top right corner of your screen, click **New**. 1
- > Select **Python 3 (ipykernel)**. 2
- > Your Notebook opens in a new tab in your browser. 3

You can Upload a notebook from your computer.



Code cell. You can type text, a math expression or a Python command.

Notebook toolbar.

The default name of the notebook is Untitled.

Figure 3.7: Create a new Jupyter Notebook

Now that your notebook is ready, it's time to write and run your first program in Jupyter Notebook.

To create a program in Jupyter Notebook:

- > Type the commands inside the code cell. 1
- > Click the **Run** button. 2
- > The result is displayed under the commands. 3

You can run your program by pressing **Shift + Enter**.

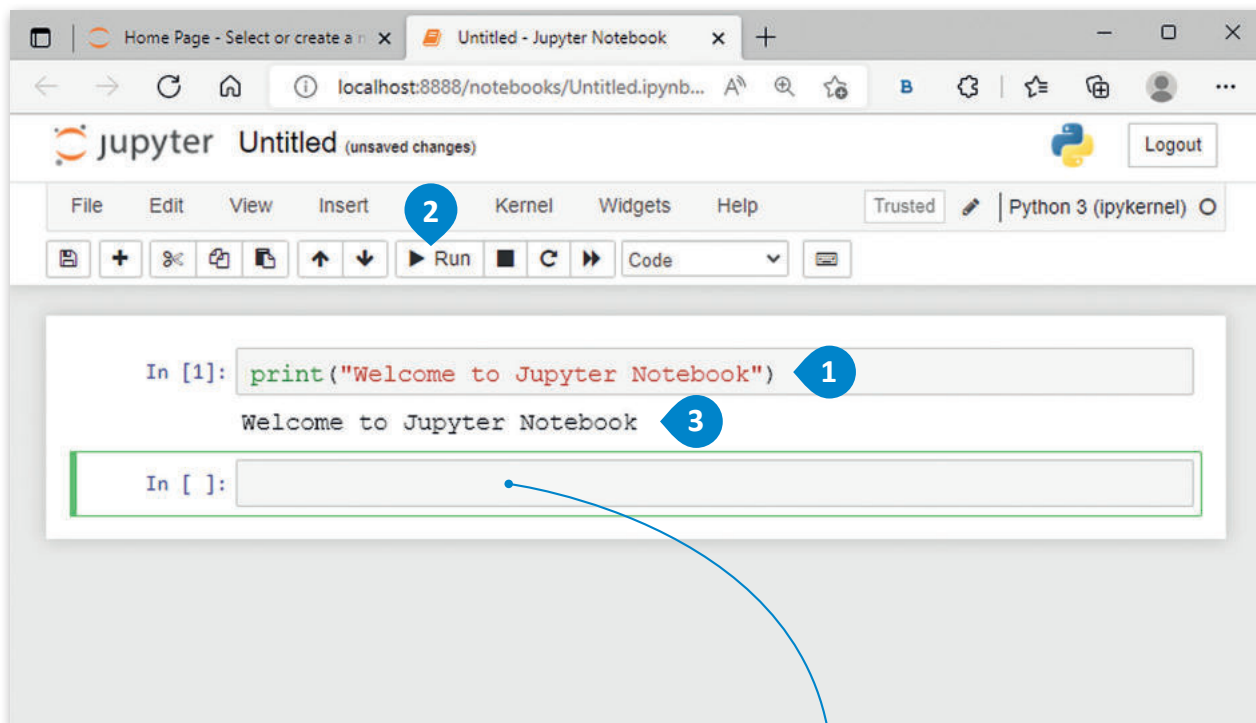


Figure 3.8: Create a program in Jupyter Notebook

When you run your program, a new code cell is automatically added.

You can have as many different cells as you need in the same Notebook. Each cell contains its own code.

INFORMATION

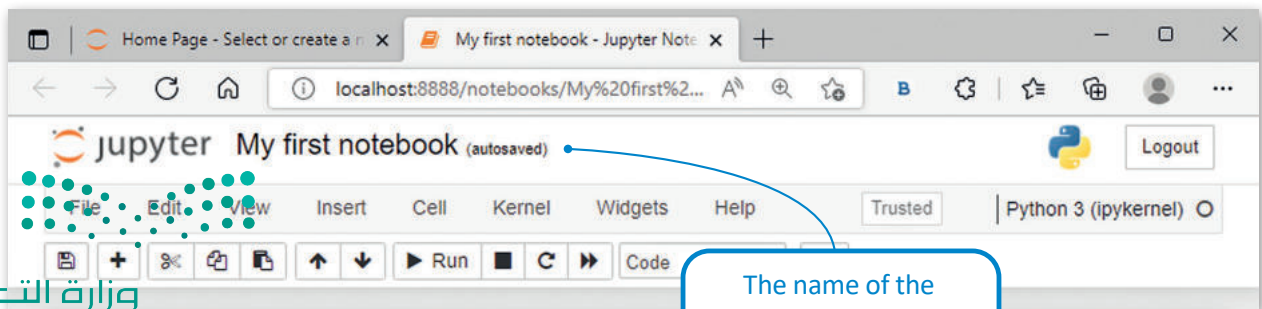
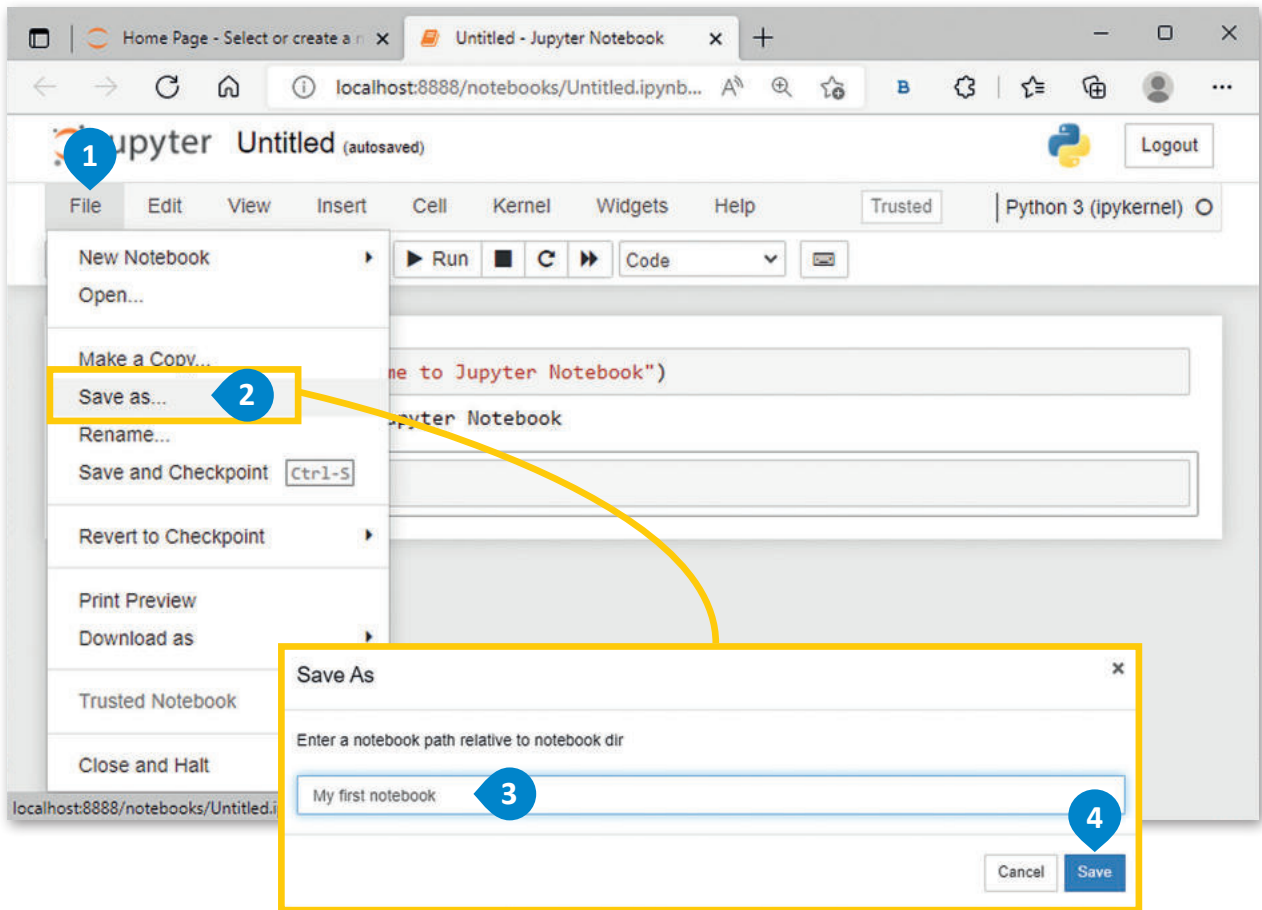
Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are Julia, Python and R.

It's time to save your Notebook.

To save your Notebook:

- > Click **File**. 1
- > Select **Save as....** 2
- > Type a name for your Notebook. 3
- > Press **Save**. 4

When you are working, the Notebook is autosaved.



وزارة التعليم

Ministry of Education

2023 - 1445

Exercises

1

Read the sentences and tick ✓ True or False.	True	False
1. Descriptive data analysis is performed if you want to find out why something happened.	<input type="radio"/>	<input type="radio"/>
2. Diagnostic data analysis provides more added value than prescriptive data analysis.	<input type="radio"/>	<input type="radio"/>
3. Predictive data analysis uses already discovered trends to predict future outcomes.	<input type="radio"/>	<input type="radio"/>
4. Prescriptive data analysis is the easiest type of data analysis.	<input type="radio"/>	<input type="radio"/>
5. Exploratory data analysis always involves a graphical representation of data.	<input type="radio"/>	<input type="radio"/>
6. With EDA, you can spot anomalies in the dataset.	<input type="radio"/>	<input type="radio"/>
7. A multivariate data analysis takes into consideration more than one independent variable.	<input type="radio"/>	<input type="radio"/>
8. Python libraries contain bundles of code that simplify many programming tasks.	<input type="radio"/>	<input type="radio"/>
9. A Python library cannot contain configuration data or message templates.	<input type="radio"/>	<input type="radio"/>
10. Matplotlib is a Python library used to create charts and graphs.	<input type="radio"/>	<input type="radio"/>

2 Compare predictive and prescriptive data analysis. What are the differences? Give an example of each type of analysis.

3 Give two examples of problems that require a univariate analysis and two examples of problems that require a multivariate analysis. Can you identify the increased complexity?

4 Compare the pros and cons of using Python libraries instead of writing your own code. Which approach would you choose?




5 You are a data analyst for a company that wants to know how its expenses are distributed in different areas. Which type of data analysis will you apply and why?

6 What is the main advantage of using Jupyter Notebook?

7 Create a new notebook in Jupyter:

> Print the message "This is my first notebook".

> Save the notebook with a name of your choice.





In the previous lesson, we discussed how Python uses libraries in order to handle data. In this lesson, you will learn how to use some of these libraries in your Jupyter Notebook.

NumPy Library

NumPy stands for Numerical Python. It is a popular library for working with numerical data in Python. NumPy can be used to perform a wide variety of mathematical operations on arrays.

Table 3.3: NumPy library methods

Methods	Meaning
<code>add(arr1,arr2,...)</code>	Adds arrays.
<code>multiply(arr1,arr2,...)</code>	Multiplies arrays.
<code>absolute(arr)</code>	Returns absolute value of each element in an input array.
<code>maximum(arr1,arr2,...)</code>	Returns the maximum value in the input arrays.

Method

A method is a function which is associated with an object. It is defined inside a class body. For example, `np.add(arr1, arr2)`.

Let's start by creating a simple list in your Jupyter Notebook. This is your list.

```
myList = [-3,-2,-1,0,1,2,3,4,5,5,5,6,7,8]

print(type(myList))
print(myList)

<class 'list'>
[-3, -2, -1, 0, 1, 2, 3, 4, 5, 5, 5, 6, 7, 8]
```

Figure 3.10: Creating a list in Jupyter Notebook

Array

Array is a data type which can hold a fixed number of values of the same data type.

Let's use the NumPy library. In this code, you will use the **absolute()** method to print the absolute values of the list.

When you use a function from the library, you type the name of the library or the name of the function.

```
import numpy as np

a = np.absolute(myList)
print(a)

[3 2 1 0 1 2 3 4 5 5 5 6 7 8]
```

Figure 3.11: Use of NumPy library

When you are using a library, you give it a name in order to use its functions in your code.

Pandas Library

Pandas library takes data and creates a Python object. It creates two main types of object:

> A *Series* is a one-dimensional labeled array capable of holding any data type (integers, strings, floating point numbers, Python objects, etc.).

> A *DataFrame* is a two-dimensional data structure which looks very similar to a table in a spreadsheet.

Each object has its own methods and attributes. You can either create a Series or a DataFrame from scratch (from lists, dictionaries, etc.) or you can import data from data sources, such as Excel, CSV, SQL, JSON and more.

Table 3.4: Differences between Pandas and NumPy libraries

	Pandas	NumPy
Types of data	works with the tabular data	works with numerical data
Types of objects	Series, DataFrame	Arrays
Performance	handles hundreds of thousands of data items	handles better 50K rows or less.
Memory utilization	consumes more memory	consumes less memory
Usage	data analysis and visualization	calculations

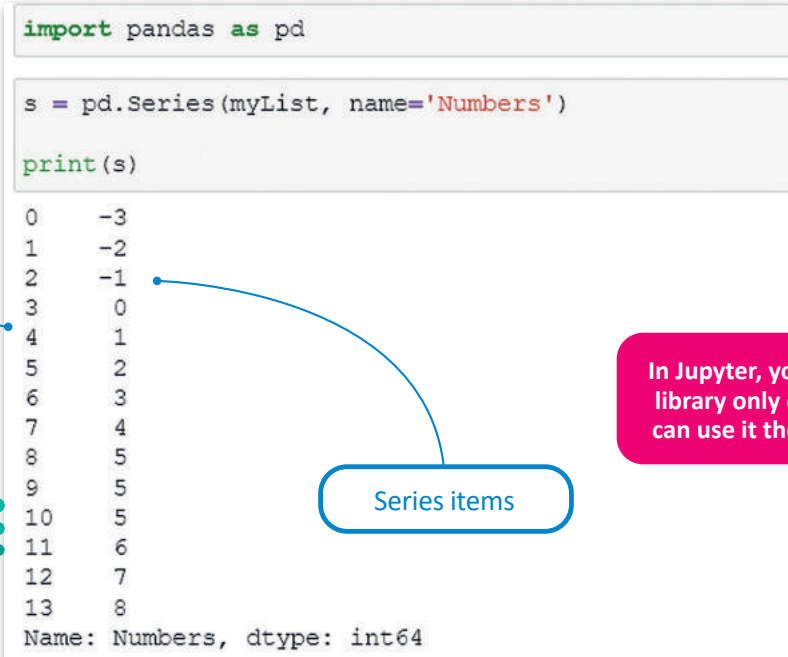
Series Object

Now, you are going to transform your list into a Series object. To do this, you have to include the Pandas library in your notebook. As you already know, to use a library in Python, you add the word `import` and the name of the library.

```
import pandas as pd

s = pd.Series(myList, name='Numbers')

print(s)
```



Index

Series items

In Jupyter, you have to import a library only once and then you can use it the whole notebook.

```
0    -3
1    -2
2    -1
3     0
4     1
5     2
6     3
7     4
8     5
9     5
10    5
11    6
12    7
13    8
Name: Numbers, dtype: int64
```

Figure 3.12: Transform a list into a Series object

Attributes of a Series Object

In table 3.5 some of the most common attributes of Series object are presented.

Table 3.5: Attributes of a Series Object

Attribute	Meaning
name	Returns the name of the Series.
size	Returns the size of the Series.
is_unique	Returns True if the values of the Series object are unique, else it returns False.
hasnans	Returns True if the given Series object has missing values in it, else it return False.

Attribute

A value associated with an object which is referenced by name using dotted expressions. For example, if an object **student** has an attribute **grade** it would be referenced as **student.grade**.

In computing, NaN stands for Not a Number.

Let's see some of these attributes of the Series object.

```
# What is the name of the Series?  
print("The name of the series is:", s.name)
```

```
The name of the series is: Numbers
```

```
# Print Series size  
print("Size of the series is:", s.size)
```

```
Size of the series is: 14
```

```
print("Are the elements of this series unique?", s.is_unique)
```

```
Are the elements of this series unique? False
```

```
# Check if there are empty rows in the Series (nan = Not A Number)  
print("Are there empty values in the series?", s.hasnans)
```

```
Are there empty values in the series? False
```


DataFrame Object

The most popular and widely used analytical tool is Excel. You can work with excel files in Jupyter using the Pandas library. To open an Excel file in Jupyter, you need these files (the Excel and the Notebook) to be in the same folder.

OS Library

To find information about your current file, you can use the OS library. The OS library in Python provides functions for creating and removing a directory (folder), fetching its contents, changing and identifying the current directory, etc.

```
import os
os.getcwd()
```

'C:\\Users\\Documents\\Jupyter examples'

getcwd stands for get current working directory.

Figure 3.14: OS library

This is your Excel file.

	A	B	C	D	E	F
1	Administrative District	Educational Stage	School Type	Total Number of Students	Total Number of Teachers	Total Number of Administrative Staff
2	Al Bahah	Primary School	Special Education	0	0	0
3	Al Bahah	Primary School	Day School	7	0	0
4	Al Bahah	Secondary School	Day School	1	0	1
5	Al Bahah	Kindergarten	Special Education	1	0	0
6	Al Bahah	Secondary School	Special Education	1	0	0
7	Al Bahah	Primary School	Special Education	2	3	0
8	Al Bahah	High School	Special Education	2	3	0
9	Al Bahah	Secondary School	Special Education	2	2	0
10	Al Bahah	Primary School	Special Education	2	4	0
11	Al Bahah	High School	Special Education	2	0	0
12	Al Bahah	Secondary School	Special Education	2	5	0
13	Al Bahah	High School	Special Education	3	3	0
14	Al Bahah	High School	Day School	3	1	1
15	Al Bahah	High School	Day School	3	0	0
16	Al Bahah	Primary School	Special Education	3	2	0
17	Al Bahah	Primary School	Special Education	3	2	0
18	Al Bahah	High School	Special Education	4	1	0
19	Al Bahah	High School	Special Education	4	2	0
20	Al Bahah	Primary School	Day School	5	2	0
21	Al Bahah	Secondary School	Special Education	5	4	0
22	Al Bahah	High School	Special Education	5	8	0
23	Al Bahah	Secondary School	Special Education	5	1	0
24	Al Bahah	Secondary School	Special Education	5	2	0

Figure 3.15: saudischools.xlsx Excel file

The dataset that you will use in this lesson is provided by the Ministry of Education through the Saudi Open Data Platform (<https://data.gov.sa/>). You can use the data of the Excel file for the purpose of this lesson following the Open Data Policies (<https://data.gov.sa/ar/policies>).

Now, you are going to transform this Excel file to a DataFrame in order to manipulate its data.

```
data = pd.read_excel('saudischools - EN.xlsx')
```

data

	Administrative District	Educational Stage	School Type	Total Number of Students	Total Number of Teachers	Total Number of Administrative Staff
0	Mecca	Primary School	Day School	36416.0	2025	946.0
1	Riyadh	Primary School	Day School	35570.0	1684	1080.0
2	Riyadh	Primary School	Day School	34668.0	1843	1152.0
3	Riyadh	Primary School	Day School	32883.0	1445	128.0
4	Tabuk	Primary School	Day School	32465.0	1959	1057.0
...
5594	Riyadh	Secondary School	Special Education	0.0	0	0.0
5595	Riyadh	Secondary School	Special Education	0.0	0	0.0
5596	Riyadh	Secondary School	Special Education	0.0	0	0.0
5597	Riyadh	Secondary School	Special Education	0.0	0	0.0
5598	Riyadh	Secondary School	Special Education	0.0	0	0.0

5599 rows × 6 columns

In order to open an Excel file in Jupyter Notebook, both the Excel file and the Jupyter Notebook file should be saved in the same folder.

Figure 3.16: Create a DataFrame

If the Excel file has multiple sheets, you can read a specific sheet. The Pandas read_excel method takes an argument called sheet_name that tells Pandas which sheet to read in the data from the Excel file. If you don't specify the sheet, it reads the first one.

Attributes of a DataFrame Object

The following table presents some of the most common attributes of DataFrames.

Attribute	Meaning
shape	Returns the dimensions of the DataFrame.
size	Returns the total number of elements in the DataFrame (n × m).
dtypes	Returns the type of value in each column.
columns	Returns the names of the DataFrame columns.
axes	Returns the number of rows and the names of the columns.

```

# Printing the table dimensions
data.shape

(5599, 6)

# Return the total number of elements in the array (n * m)
data.size

33594

# Return the type of the value of each column
data.dtypes

Administrative District      object
Educational Stage           object
School Type                 object
Total Number of Students    float64
Total Number of Teachers    int64
Total Number of Administrative Staff float64
dtype: object

# Return the range of rows and the column names
data.axes

[RangeIndex(start=0, stop=5599, step=1),
 Index(['Administrative District', 'Educational Stage', 'School Type',
       'Total Number of Students', 'Total Number of Teachers',
       'Total Number of Administrative Staff'],
      dtype='object')]

```

You can add comments to your code by using a hash (#) at the start of the line. Comments are statements that are not executed, rather they make the code more readable.

Figure 3.17: Use of the attributes of a DataFrame Object

In Pandas library, the object is usually a string.data type.

Table 3.7: Pandas dtype mapping

Pandas dtype	Python type
object	str or mixed
int64	int
float64	float
bool	bool
datetime64	NA
timedelta[ns]	NA
category	NA

Indexing

An index is a list of integers or labels you use to uniquely identify rows or columns. In Pandas, indexing involves picking specific rows and columns of data from a DataFrame. You can select all rows and some columns, some rows and all columns, or a subset of rows and columns. Subset Selection is another term for indexing. Let's see some examples of methods you can use for indexing.

Table 3.8: Indexing methods

Method	Meaning
head()	Returns the first elements of the object.
tail()	Returns the last elements of the object.
value_counts()	Returns unique values and their counts.
idxmax()	Returns the index of the maximum element.
idxmin()	Returns the index of the minimum element.

Using Indexing in a Series Object

Let's apply these indexing methods to the Series object you have created. First print the Series object, to remember its contents.

```
print(s)
0    -3
1    -2
2    -1
3     0
4     1
5     2
6     3
7     4
8     5
9     5
10    5
11    6
12    7
13    8
Name: Numbers, dtype: int64
```

```
x=4
print("the value of the index",x, "is",s[x])
```

the value of the index 4 is 1

```
# Return the first 2 rows of the series
s.head(2)
```

```
0    -3
1    -2
Name: Numbers, dtype: int64
```

```
# Return the last rows of the series
s.tail()
```

```
9     5
10    5
11    6
12    7
13    8
Name: Numbers, dtype: int64
```

```
# Return a count of the unique values of the series
s.value_counts()
```

```
5     3
-3    1
-2    1
-1    1
0     1
1     1
2     1
3     1
4     1
6     1
7     1
8     1
Name: Numbers, dtype: int64
```

By default, if no argument is given, the head() and tail() methods will return 5 elements.

Figure 3.18: Using Indexing in a Series Object



Using Indexing in a DataFrame Object

```
# Printing the first 10 rows of the table
data.head(10)
```

	Administrative District	Educational Stage	School Type	Total Number of Students	Total Number of Teachers	Total Number of Administrative Staff
0	Mecca	Primary School	Day School	36416.0	2025	946.0
1	Riyadh	Primary School	Day School	35570.0	1684	1080.0
2	Riyadh	Primary School	Day School	34668.0	1843	1152.0
3	Riyadh	Primary School	Day School	32883.0	1445	128.0
4	Tabuk	Primary School	Day School	32465.0	1959	1057.0
5	Mecca	Primary School	Day School	32429.0	1661	637.0
6	Riyadh	Primary School	Day School	31026.0	1691	960.0
7	Tabuk	Primary School	Day School	30078.0	1836	66.0
8	Mecca	Primary School	Day School	29004.0	1402	730.0
9	Eastern Province	Primary School	Day School	28948.0	1092	481.0

```
data.tail()
```

	Administrative District	Educational Stage	School Type	Total Number of Students	Total Number of Teachers	Total Number of Administrative Staff
5594	Riyadh	Secondary School	Special Education	0.0	0	0.0
5595	Riyadh	Secondary School	Special Education	0.0	0	0.0
5596	Riyadh	Secondary School	Special Education	0.0	0	0.0
5597	Riyadh	Secondary School	Special Education	0.0	0	0.0
5598	Riyadh	Secondary School	Special Education	0.0	0	0.0



وزارة التعليم

Ministry of Education

2023 - 1445

```
# Accessing the DataFrame attribute 'columns' to print the names of
# the table's columns
for col in data.columns:
    print(col)
```

```
Administrative District
Educational Stage
School Type
Total Number of Students
Total Number of Teachers
Total Number of Administrative Staff
```

Prints the names of the columns of the DataFrame.

The describe() method is used to view some basic statistical details.

```
data.describe()
```

	Total Number of Students	Total Number of Teachers	Total Number of Administrative Staff
count	5598.000000	5599.000000	5598.000000
mean	1109.923901	89.373460	19.455341
std	2950.764755	192.541064	66.794989
min	-12.000000	-586.000000	-2.000000
25%	31.000000	4.000000	0.000000
50%	136.000000	17.000000	1.000000
75%	808.000000	82.000000	10.000000
max	36416.000000	2090.000000	1152.000000

Figure 3.19: Using Indexing in a DataFrame Object



Filtering Data or Subset Selection

Sometimes you don't need the whole dataset. You need to isolate some specific data. To do this, you need to add filters. There are many ways to select a subset of a DataFrame or a Series. A very simple way is with Boolean indexing, but the more robust way is by using the **loc** and **iloc** methods. First you will learn Boolean indexing and then the **loc** and **iloc** method.

Data filtering

Data filtering is the process of choosing a smaller part of your dataset and using that subset for viewing or analysis.

Boolean Indexing

Boolean indexing is a type of indexing which uses the actual values of the dataset.

In Boolean indexing, you need to use the Boolean operators. Boolean operators are written differently in Jupyter than in Python.

Let's see some examples with the Series object.

Table 3.9: Boolean operators in Jupyter

Python	Jupyter
and	&
or	
not	~

```
# Return the elements of the series that satisfy the expression s>0
s[s > 0]
```

```
4    1
5    2
6    3
7    4
8    5
9    5
10   5
11   6
12   7
13   8
Name: Numbers, dtype: int64
```

```
s[(s < -1) | (s > 6)]
```

```
0    -3
1    -2
12    7
13    8
Name: Numbers, dtype: int64
```

```
# Printing not(s<0) => (s>=0)
s[~(s < 0)]
```

```
3    0
4    1
5    2
6    3
7    4
8    5
9    5
10   5
11   6
12   7
13   8
Name: Numbers, dtype: int64
```


Indexing with Loc and Iloc Methods

In Pandas library, loc and iloc are two commonly used methods for indexing.

> **loc**, selects rows and columns with specific labels (the names of the columns).

> **iloc**, selects rows and columns at specific integer positions (the numbers of the rows and the columns).

Let's see some examples with the DataFrame object using the **loc()** method.

In the following example, you will use the **loc()** method to print the first five rows of two specific columns.

```
# Choosing the first 5 rows of the columns 'Administrative District' and 'Educational Stage'  
data.loc[:4,['Administrative District','Educational Stage']]
```

	Administrative District	Educational Stage
0	Mecca	Primary School
1	Riyadh	Primary School
2	Riyadh	Primary School
3	Riyadh	Primary School
4	Tabuk	Primary School

Figure 3.21: Print the first five rows of two specific columns

In this example, you will print the rows of the DataFrame that have a specific value in a specific column.

```
# Print the rows of the DataFrame that have a specific value in a specific column  
data.loc[data['Administrative District'].isin(['Riyadh','Al Bahah'])]
```

	Administrative District	Educational Stage	School Type	Total Number of Students	Total Number of Teachers	Total Number of Administrative Staff
1	Riyadh	Primary School	Day School	35570.0	1684	1080.0
2	Riyadh	Primary School	Day School	34668.0	1843	1152.0
3	Riyadh	Primary School	Day School	32883.0	1445	128.0
6	Riyadh	Primary School	Day School	31026.0	1691	960.0
10	Riyadh	Primary School	Day School	28727.0	1520	835.0
...
5594	Riyadh	Secondary School	Special Education	0.0	0	0.0
5595	Riyadh	Secondary School	Special Education	0.0	0	0.0
5596	Riyadh	Secondary School	Special Education	0.0	0	0.0
5597	Riyadh	Secondary School	Special Education	0.0	0	0.0
5598	Riyadh	Secondary School	Special Education	0.0	0	0.0

وزارة التعليم
1318 rows x 6 columns
Ministry of Education

Figure 3.25: Print the rows of the DataFrame that have a specific value in a specific column

In this example, you will create a new DataFrame named `studentsReg`. This DataFrame will have two columns, one column for Region and another for Number of Students.

```
# Create a DataFrame called studentsReg with two columns 'Administrative District'
# and 'Total Number of Students'

studentsReg = data.loc[:,['Administrative District','Total Number of Students']]
studentsReg
```

	Administrative District	Total Number of Students
0	Mecca	36416.0
1	Riyadh	35570.0
2	Riyadh	34668.0
3	Riyadh	32883.0
4	Tabuk	32465.0
...
5594	Riyadh	0.0
5595	Riyadh	0.0
5596	Riyadh	0.0
5597	Riyadh	0.0
5598	Riyadh	0.0

5599 rows × 2 columns

Figure 3.23: Create a new DataFrame named `studentsReg`

Now, you will use the `iloc()` method to select all the elements from the 1st row of the DataFrame.

```
# Print all the elements from the [row] of the table
data.iloc[0]
```

```
Administrative District      Mecca
Educational Stage           Primary School
School Type                 Day School
Total Number of Students    36416.0
Total Number of Teachers     2025
Total Number of Administrative Staff  946.0
Name: 0, dtype: object
```

Figure 3.24: Print the elements of the 1st row of the DataFrame

Remember, indexing in Python starts from 0.



Now for these examples, you will print specific elements of the DataFrame.

```
# Print the element in the [row,col] position of the table
data.iloc[0,3]

36416.0

# Print the elements [start:end , start:end], the end is not included
# This example prints the elements of the 2nd and 3rd row,
# but only in the the 0th, 1st and 2nd column
data.iloc[1:3, 0:3]
```

	Administrative District	Educational Stage	School Type
1	Riyadh	Primary School	Day School
2	Riyadh	Primary School	Day School

Figure 3.25: Print specific elements of the DataFrame

Prints the elements in the 2nd and 3rd rows, but only from the 0th, 1st and 2nd column.

And in this example, you will use a **for** loop to print the first 10 rows of the 1st column of the studentsReg DataFrame.

```
for i in range (10):
    print(studentsReg.iloc[i][1])

36416.0
35570.0
34668.0
32883.0
32465.0
32429.0
31026.0
30078.0
29004.0
28948.0
```

Figure 3.26: Print elements of the DataFrame



Grouping and Aggregating

The process of putting a dataset's elements in groups based on some criteria and applying functions to these groups is called grouping. In Pandas library, this action is performed with the **df.groupby()** method. As an example, imagine you have a dataset with the top basketball scorers of all time. If you want to see how many players in this dataset played for a certain team, you can group this dataset by the "Team" column and perform the `sum()` function on the grouped data.

Aggregate function

A function that makes mathematical calculations with the values of multiple rows which are grouped together, and as a result returns a single, summary value.

The most common aggregate functions are `sum`, `count`, `max`, `min` and `mean`.

Table 3.10: Aggregate functions

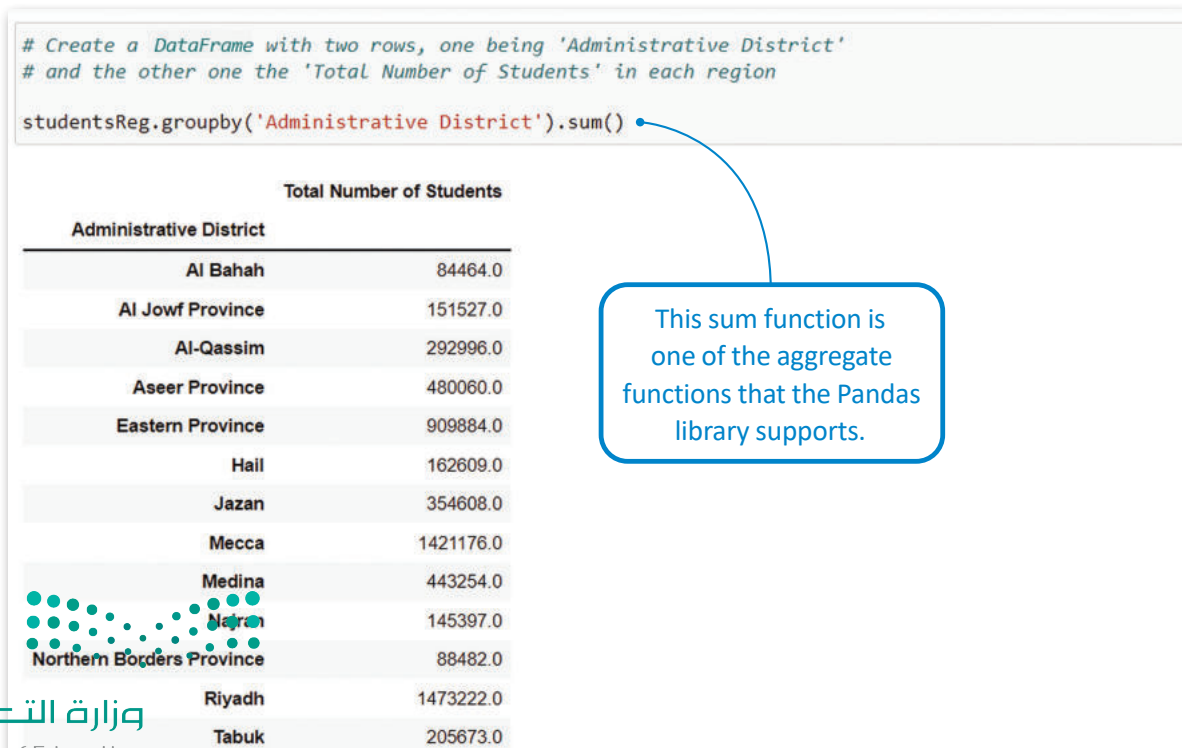
Function	Meaning
<code>sum</code>	Calculates the sum of a list of numbers.
<code>max</code>	Returns the maximum number of a list of numbers.
<code>min</code>	Returns the minimum number of a list of numbers.
<code>mean</code>	Calculates the average of a list of numbers.

Groupby Method

Using the **groupby()** method you can split your data into different groups. This can help you to perform calculations for better data analysis.

Let's see some examples of the **groupby()** method in the new DataFrame you have created.

In this example, you group the students according to their region and you calculate the sum of the students in each region.



In this example you group the students according to two criteria, their region and the level, and you calculate the sum of the students in each region.

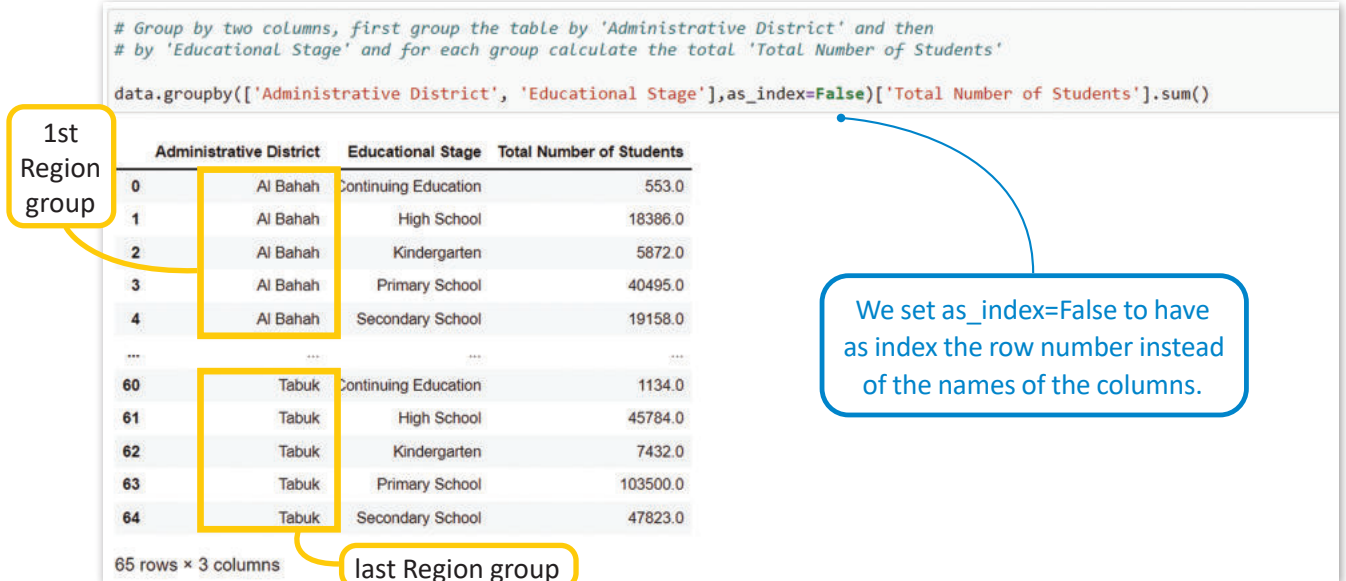
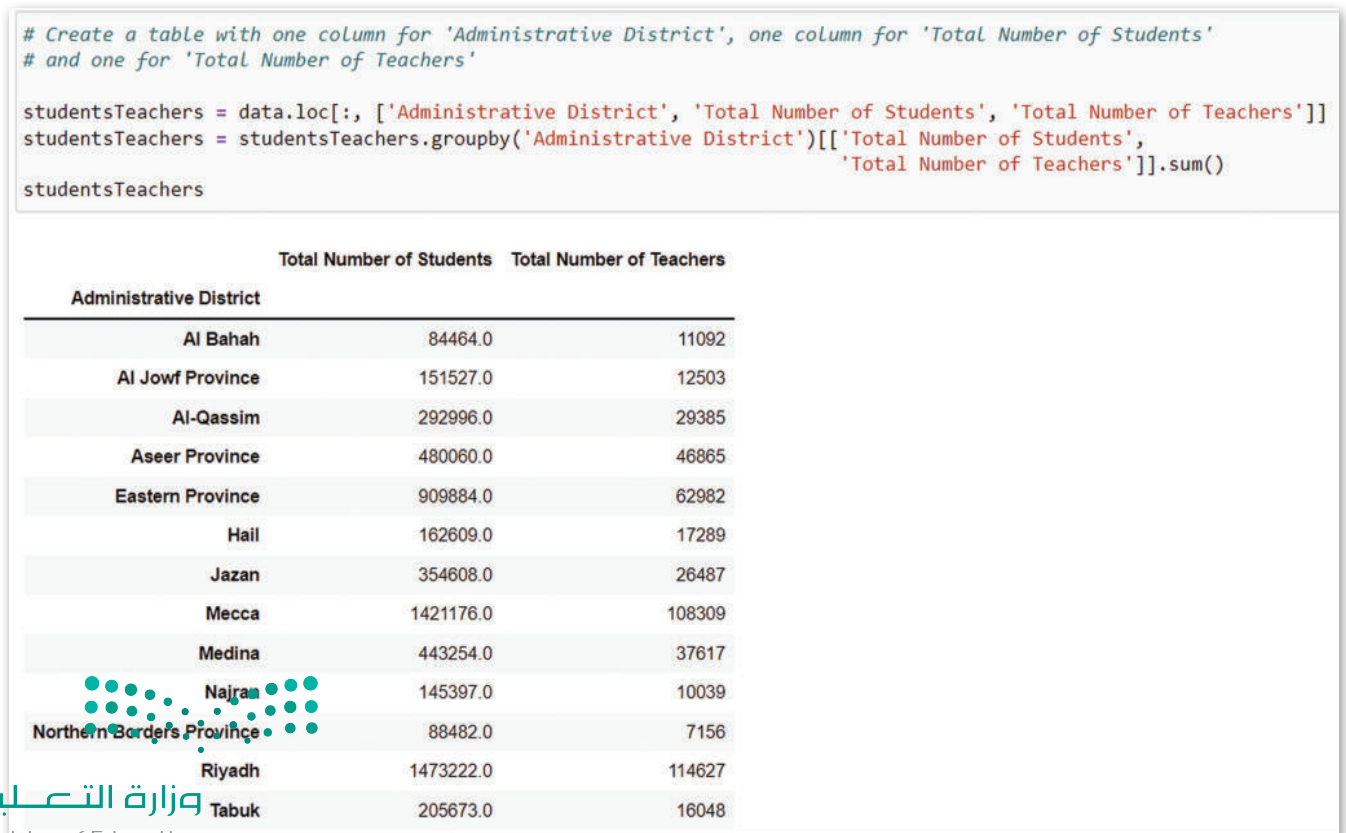


Figure 3.28: Use of the groupby() method to group the DataFrame by multiple columns

In this example you create a new w for Region, Number of students and Number of teachers. Then you group by the Region and calculate the sum of the students and the sum of the teachers in each region.



Data Cleaning

Before starting your data analysis, it is very important that the data you are going to analyze are correct. This means that duplicate, corrupted or inaccurate data must be removed from your dataset. If this data remains in your dataset, the results of the data analysis will not be correct.

Data cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data from a dataset.

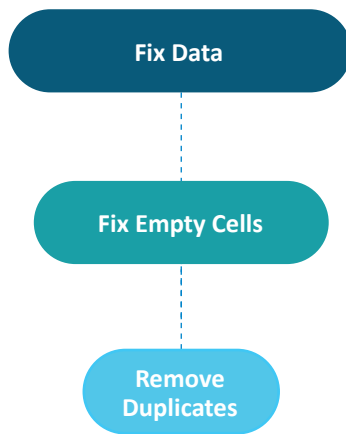


Figure 3.30: Data cleaning process

Table 3.11: Data cleaning methods

Method	Meaning
<code>df.duplicated()</code>	Returns a boolean value for each row that contains duplicated data.
<code>df.value_counts()</code>	Returns the unique values in a dataset.
<code>df.isnull()</code>	Returns a boolean value for each empty cell of the dataset.
<code>df.dropna()</code>	Deletes the empty rows.

Duplicated Data

To check if your dataset contains duplicated data you use the `df.duplicated()` method. It returns a boolean value for every row. It returns:

> *True for duplicated data*

> *False for non-duplicated data*

Let's see how you can handle duplicated rows in the dataset.

```
dup = data.duplicated()

# To see how many duplicated rows there are in the table
dup.value_counts()

False    5428
True      171
dtype: int64
```

Number of duplicates

Figure 3.31: Use of the `df.duplicated()` method

In your dataset there are 171 duplicated rows.

To delete these rows you use the **drop_duplicates()** method. This method deletes the duplicated rows.

After deleting the duplicates, you have to refresh your dataset to check that the duplicates have been removed.

```
# Now remove duplicated rows from the table
data = data.drop_duplicates()

dup = data.duplicated()

dup.value_counts()

False    5428
dtype: int64
```




Figure 3.32: Use of the drop_duplicates() method

Empty Cells

To check if your dataset has missing values you use the **data.isnull()** method. This method returns a boolean value for each empty cell of the dataset:

> True for empty

> False for not empty

Let's see how you can count the empty cells in a dataset.

In this example you count the empty cells per column.

```
# Get the number of missing data points per column
missing_values_count = data.isnull().sum()
missing_values_count
```

Administrative District	1
Educational Stage	2
School Type	2
Total Number of Students	1
Total Number of Teachers	0
Total Number of Administrative Staff	1

dtype: int64



Figure 3.33: Count the empty cells per column

You can see the number of empty cells in each column.

To delete these rows you use the **dropna()** method. This method deletes the rows which contain missing values. After deleting these rows you have to refresh your dataset to check that they have been removed.



```
# Delete the rows containing missing values
data = data.dropna()
```

```
missing_values_count = data.isnull().sum()
missing_values_count
```

```
Administrative District      0
Educational Stage           0
School Type                  0
Total Number of Students    0
Total Number of Teachers    0
Total Number of Administrative Staff 0
dtype: int64
```

No empty cells

Figure 3.34: Delete the rows containing missing values

Wrong Data

Sometimes your dataset may contain wrong data. For example, in our dataset we cannot have negative numbers in the number of students column. To check if your dataset contains wrong data you need to write code specific to your dataset.

In this example you will check for negative numbers in the columns of the dataset.

What type of data could be considered wrong depends on the dataset. You have to decide what to do with this "wrong" data. You might want to delete it or replace it with other values.

```
# Check if there are negative elements in the columns that have numbers
data[data['Total Number of Students']<0].nunique()
```

```
Administrative District      1
Educational Stage           1
School Type                  1
Total Number of Students    1
Total Number of Teachers    1
Total Number of Administrative Staff 1
dtype: int64
```

```
data[data['Total Number of Teachers']<0].nunique()
```

```
Administrative District      1
Educational Stage           1
School Type                  1
Total Number of Students    1
Total Number of Teachers    1
Total Number of Administrative Staff 1
dtype: int64
```

```
data[data['Total Number of Administrative Staff']<0].nunique()
```

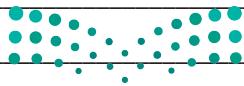
```
Administrative District      1
Educational Stage           1
School Type                  1
Total Number of Students    1
Total Number of Teachers    1
Total Number of Administrative Staff 1
dtype: int64
```


Exercises

1 What is the difference between Series and DataFrame objects?

2 Describe the difference between data indexing and data filtering.

3 Explain the importance of data cleaning before starting data analysis.



4 Import the random library and use the random.randrange() function to print a random number between 1 and 100.

5 Open a new Jupyter Notebook, import the Excel file with the name "tourist-indicators.xlsx".

- > Load the sheet "I3" in a Pandas DataFrame.
- > Print the shape of the DataFrame.
- > Print the types of data stored in each column of the dataset.
- > Print the names of the columns of the dataset.

6 Use the dataset you imported in the previous exercise and:

- > create a new Series object containing the number of inbound tourists from GCC countries.
- > find the maximum and the minimum number of tourists and in which rows of the Series these values occur.
- > check the Series for inappropriate and missing values, and if there are any, remove those rows.
- > print the number of tourists in descending order but only for values greater than 300.

- 7 Open the sheet "I3" from the file "tourist-indicators.xlsx" and read it to a new DataFrame. Then:
- > inspect the whole dataset for missing values and duplicates.
 - > print the number of missing values and the number of duplicated rows.
 - > remove duplicated rows and rows with missing values.
 - > group the DataFrame by month and find out the month with the most visitors for every region.



Lesson 3

Data Visualization

Link to digital lesson



www.iien.edu.sa

As we have mentioned before, data visualization is the graphical representation of information and data. Data visualizations make data easier for the human brain to understand and analyze. By using visual elements like charts, graphs, and maps, you make the data more accessible, understandable and usable.

In this lesson, you are going to use Jupyter to visualize your data. Jupyter supports data visualization in combination with Python libraries.

Types of Data Visualization

The most common types of data visualization are:

- > *charts (line chart, bar chart, pie chart)*
- > *graphs*
- > *plots*
- > *histograms*
- > *tables*
- > *maps*

Each type of visualization represents the data differently. You should choose the visualization according to what you want to learn from your report.



وزارة التعليم

Ministry of Education

2023 - 1445

Figure 3.36: Board showing the most common types of data visualization

Charts

Line Chart

A line chart or line graph is a data visualization technique where each value of an independent variable is plotted individually and these values are connected with straight lines. The horizontal axis is usually a continuous variable, such as time, and the vertical axis shows the values of the independent variable. One advantage of line graphs is their simplicity for visualizing the change of a variable over time. This can help in detecting trends and patterns. You can plot multiple lines on the same graph and compare the progress of more than one independent variable over the same time period.

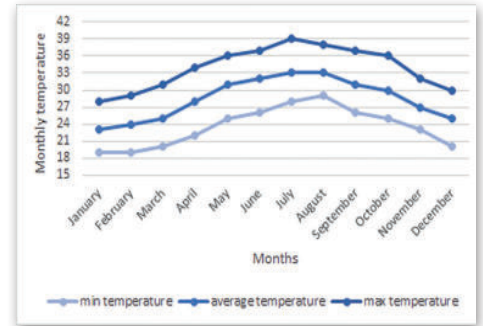


Figure 3.37: Line chart showing the annual min, max and average temperatures recorded in Abha

Bar Chart

Bar charts are figures with the elements of a categorical variable on the x axis and rectangular bars whose height illustrates the values of those elements. Bar charts can either be vertical or horizontal. Vertical bar charts are usually called column charts. There are many types of bar chart such as grouped bar charts, stacked bar charts, bar charts with error bars and more.

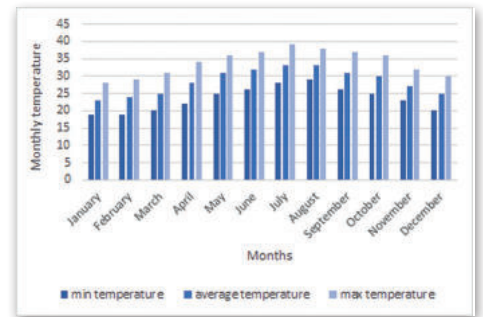


Figure 3.38: Bar chart showing the annual min, max and average temperatures recorded in Abha

Scatter Plot

A scatter plot is a way to visualize data by using dots to represent the values of different variables. These dots are "scattered" on the figure, hence the name scatter plot. Their positions on the x and y axes represent their x and y values. You can use different colors to draw the dots, with each color representing a particular variable. When the values of the variables studied are discrete, a scatter plot is more suitable than a line chart. Line charts are more applicable for representing variables whose values show continuous change. There are different types of scatter plot based on the correlation between the variables (positive, negative, null).

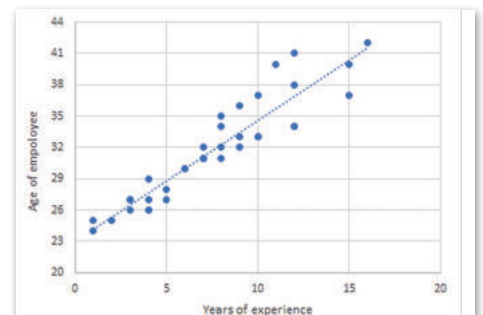


Figure 3.39: Scatter plot showing positive correlation between years of experience and age of employee

Pie Chart

Pie charts are circular charts that look like pies divided into slices that represent the proportional values of variables in a specific category. Each slice of the pie chart represents a different category. There are many types of pie chart, such as doughnut charts, half-doughnut pie charts, multilayered pie charts and more.

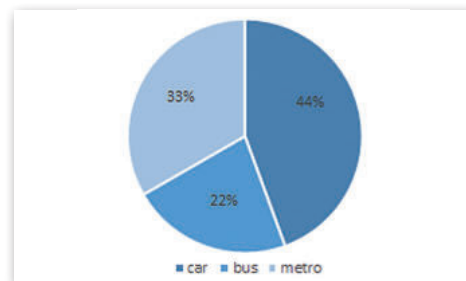


Figure 3.40: Pie chart showing preferred means of transport as a percentage

Histogram

Histograms are one of the first visualization techniques developed in the field of mathematical statistics. They are similar to bar charts but histograms show the frequency of numerical data while bar charts compare categories of data. To create a histogram, the data are grouped into ranges which are then plotted as bars connected to each other. The height of the bars shows how many values are in each range.

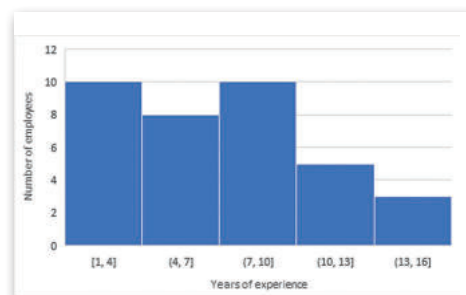


Figure 3.41: Histogram showing the distribution of the years of experience among employees

Categorical data are discrete variables. They can have a certain number of values, for example the number of students in each region in SA. Continuous data can have any value between a minimum and a maximum value, for example time or temperature.

Matplotlib Library

In order to visualize your data, you need to import a new library called **Matplotlib**. This library contains some ready to use methods that you can use to make your diagram more understandable. You can see these methods in table 3.12. Using this library, you can present your data in any diagram or chart you want. In this lesson, you will use these methods to create diagrams based on your DataFrame.

Table 3.12: Methods of Matplotlib library

Method	Meaning
bar()	Creates a bar chart
pie()	Creates a pie chart
set_title()	Sets the title of the chart
set_ylabel()	Sets the label of the y axis
set_xlabel()	Sets the label of the x axis
show()	Creates the chart



Bar Chart

It is time to create your first chart in Jupyter!

Start by importing the libraries you are going to use.

```
import matplotlib.pyplot as plt
import numpy as np
```

Figure 3.42: Import libraries

The next step is to create the dataset that you are going to use.

From the dataset that you used in the previous lesson, group your data by region and get the mean() of students, teachers and administrators.

Then you sort this DataFrame by the students column.

```
groupsB = data.groupby(['Administrative District'],as_index=False)
[['Total Number of Students','Total Number of Teachers','Total Number of Administrative Staff']].mean().round(0)

groupsB = groupsB.sort_values(by=['Total Number of Students'],ascending=False)
groupsB
```

	Administrative District	Total Number of Students	Total Number of Teachers	Total Number of Administrative Staff
4	Eastern Province	1582.0	110.0	22.0
7	Mecca	1378.0	105.0	20.0
11	Riyadh	1312.0	102.0	27.0
8	Medina	1148.0	97.0	17.0
12	Tabuk	1088.0	85.0	20.0
9	Najran	1054.0	73.0	16.0
6	Jazan	956.0	71.0	17.0
1	Al Jowf Province	953.0	79.0	19.0
10	Northern Borders Province	756.0	61.0	8.0
3	Aseer Province	741.0	72.0	17.0
2	Al-Qassim	708.0	71.0	15.0
5	Hall	648.0	69.0	15.0
0	Al Bahah	433.0	57.0	10.0

Sort the data in descending order.

Figure 3.43: Create the dataset

You are going to select and use only the first five rows of your dataset, to create a clearer bar chart.

```
reg = groupsB.iloc[:5,0].tolist()
studentsH = groupsB.iloc[:5,1].tolist()
teachersH = groupsB.iloc[:5,2].tolist()
adminsH = groupsB.iloc[:5,3].tolist()
print(reg)
print(studentsH)
print(teachersH)
print(adminsH)
```

```
['Eastern Province', 'Mecca', 'Riyadh', 'Medina', 'Tabuk']
[1627.0, 1419.0, 1367.0, 1182.0, 1094.0]
[113.0, 108.0, 107.0, 100.0, 85.0]
[22.0, 21.0, 28.0, 18.0, 20.0]
```


The code to create your diagram.

```
# the Label Locations
x = np.arange(len(reg))
```

The x coordinates of the bars

```
# the width of the bars
width = 0.5
```

```
# This is a Matplotlib built-in style.
plt.style.use('fivethirtyeight')

fig, ax = plt.subplots(figsize=(10,6))

myLabel = 'Total Students'

rects1 = ax.bar(x, studentsH, width, label=myLabel)

# Add some text for Labels, title and custom x-axis tick labels, etc.
regionsLabel = 'Administrative Districts'
meanLabel = 'Average number'
title = 'Total students, teachers and administrators, top 5 regions'

ax.set_xlabel(regionsLabel)
ax.set_ylabel(meanLabel)
ax.set_title(title)

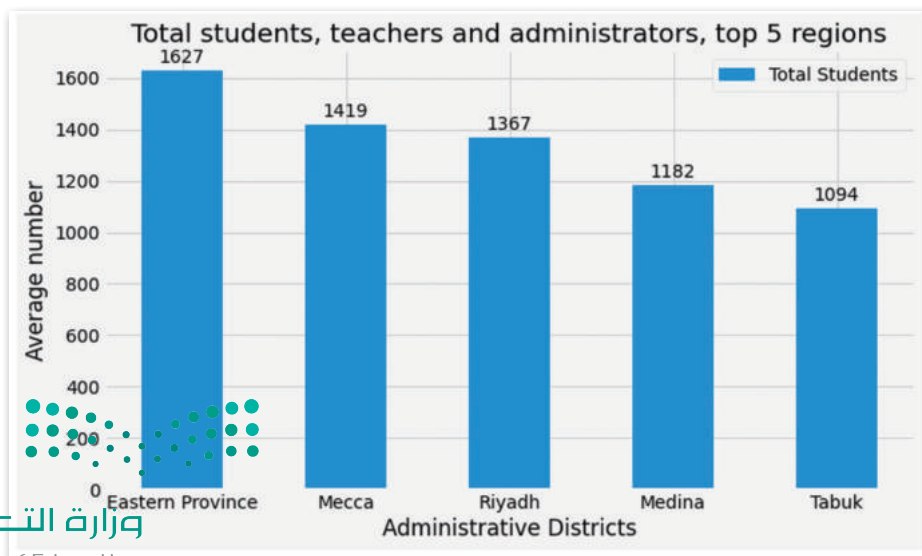
ax.set_xticks(x, reg)
ax.legend()

ax.bar_label(rects1, padding=3)
fig.tight_layout()
```

Set the label to the legend of the diagram.

Figure 3.45: Create a bar chart

Your bar chart is ready!



وزارة التعليم

Ministry of Education

2023 - 1445

Figure 3.46: Bar chart

Now let's say you want to plot the number of students, teachers and administrators on the same bar chart. This is called a grouped bar chart and you need to place the bars correctly depending on the bar width.

```
fig, ax = plt.subplots(figsize=(10,6))

studentsLabel = 'Total Students'
teachersLabel = 'Total Teachers'
adminsLabel = 'Total Administrators'

rects1 = ax.bar(x - width/3, studentsH, width, label=studentsLabel)
rects2 = ax.bar(x, teachersH, width, label=teachersLabel)
rects3 = ax.bar(x + width/3, adminsH, width, label=adminsLabel)

# Add some text for labels, title and custom x-axis tick labels, etc.

regionsLabel = 'Administrative Districts'
meanLabel = 'Average Number'
title = 'Total students, teachers and administrators, top 5 regions'

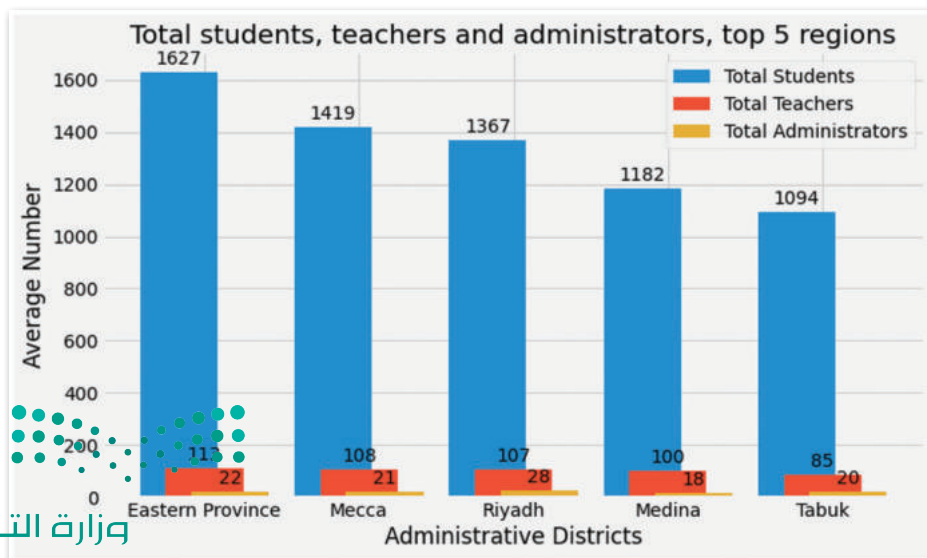
ax.set_xlabel(regionsLabel)
ax.set_ylabel(meanLabel)
ax.set_title(title)

ax.set_xticks(x, reg)
ax.legend()

ax.bar_label(rects1, padding=3)
ax.bar_label(rects2, padding=3)
ax.bar_label(rects3, padding=3)

fig.tight_layout()
```

Figure 3.47: Create a grouped bar chart



وزارة التعليم

Ministry of Education

Figure 3.48: Grouped bar chart

2023 - 1445

Pie Chart

Let's see how you can create a pie chart in Jupyter Notebook.

You will create a new DataFrame named groupsP. From the dataset that you have already used in the previous lesson, group your data by Stage and get the mean() of students, teachers and administrators. Then you will sort this DataFrame by the mean number of administrators.

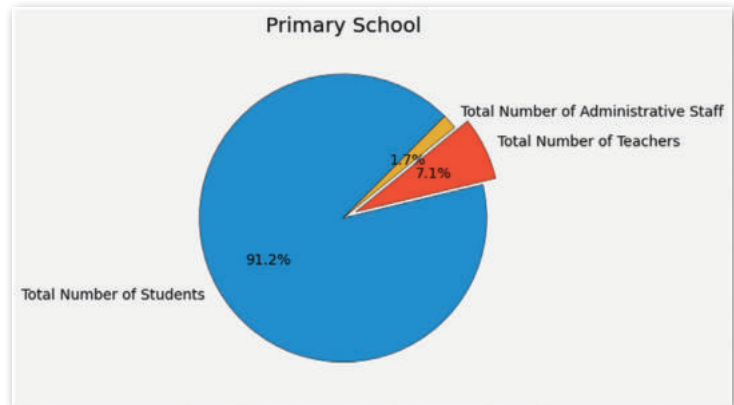


Figure 3.49: Pie chart

```
groupsP = data.groupby(['Educational Stage'],as_index=False) [['Total Number of Students',  
                    'Total Number of Teachers',  
                    'Total Number of Administrative Staff']].mean().round(0)  
  
# Sorting the values of the DataFrame  
groupsP = groupsP.sort_values(by=['Total Number of Administrative Staff'],ascending=False)  
groupsP
```

	Educational Stage	Total Number of Students	Total Number of Teachers	Total Number of Administrative Staff
3	Primary School	1915.0	149.0	35.0
2	Kindergarten	587.0	53.0	26.0
4	Secondary School	899.0	74.0	15.0
1	High School	969.0	83.0	11.0
0	Continuing Education	123.0	0.0	1.0

Figure 3.50: Create a new DataFrame

Now let's create a pie chart showing the proportions of students, teachers and administrators for one region. First, you need to create a list containing the slices of the pie chart. In your example, the slices will be a list containing the numbers of students, teachers and administrators for the District in the first row.

```
fig, ax = plt.subplots(figsize=(10,6), subplot_kw=dict(aspect="equal"))  
  
# create the lists of the slices  
slices = groupsP.iloc[0,1:].tolist()  
my_labels = groupsP.columns[1:].tolist()  
  
# array that specifies the fraction of the radius with which to offset each wedge  
explode = [0,0.1,0]  
  
# create the pie chart  
ax.pie(slices, labels=my_labels, explode=explode, shadow=False, startangle=45, autopct='%1.1f%%',  
       wedgeprops={'edgecolor':'black'})  
  
title = groupsP.iloc[0,0]  
ax.set_title(title);
```

Figure 3.51: Create a pie chart

Properties for the
appearance of the pie chart.

To show the percentage
of each slice.

Now you will create a figure with more than one pie chart.

```
fig, ([ax1,ax2],[ax3,ax4]) = plt.subplots(2,2,figsize=(16,10), subplot_kw=dict(aspect="equal"))

# First pie chart
slices = groupsP.iloc[0,1:].tolist()
my_labels = groupsP.columns[1:].tolist()

ax1.pie(slices, labels=my_labels, shadow=False, startangle=45, autopct='%1.1f%%',
        wedgeprops={'edgecolor':'black'})

title1 = groupsP.iloc[0,0]
ax1.set_title(title1)

# Second pie chart
slices = groupsP.iloc[1,1:].tolist()
my_labels = groupsP.columns[1:].tolist()

ax2.pie(slices, labels=my_labels, shadow=False, startangle=45, autopct='%1.1f%%',
        wedgeprops={'edgecolor':'black'})

title2 = groupsP.iloc[1,0]
ax2.set_title(title2)

# Third pie chart
slices = groupsP.iloc[2,1:].tolist()
my_labels = groupsP.columns[1:].tolist()

ax3.pie(slices, labels=my_labels, shadow=False, startangle=45, autopct='%1.1f%%',
        wedgeprops={'edgecolor':'black'})

title3 = groupsP.iloc[2,0]
ax3.set_title(title3)

# Fourth pie chart
slices = groupsP.iloc[3,1:].tolist()
my_labels = groupsP.columns[1:].tolist()

ax4.pie(slices, labels=my_labels, shadow=False, startangle=45, autopct='%1.1f%%',
        wedgeprops={'edgecolor':'black'})

title4 = groupsP.iloc[3,0]
ax4.set_title(title4);
```

Figure 3.52: Create four pie charts



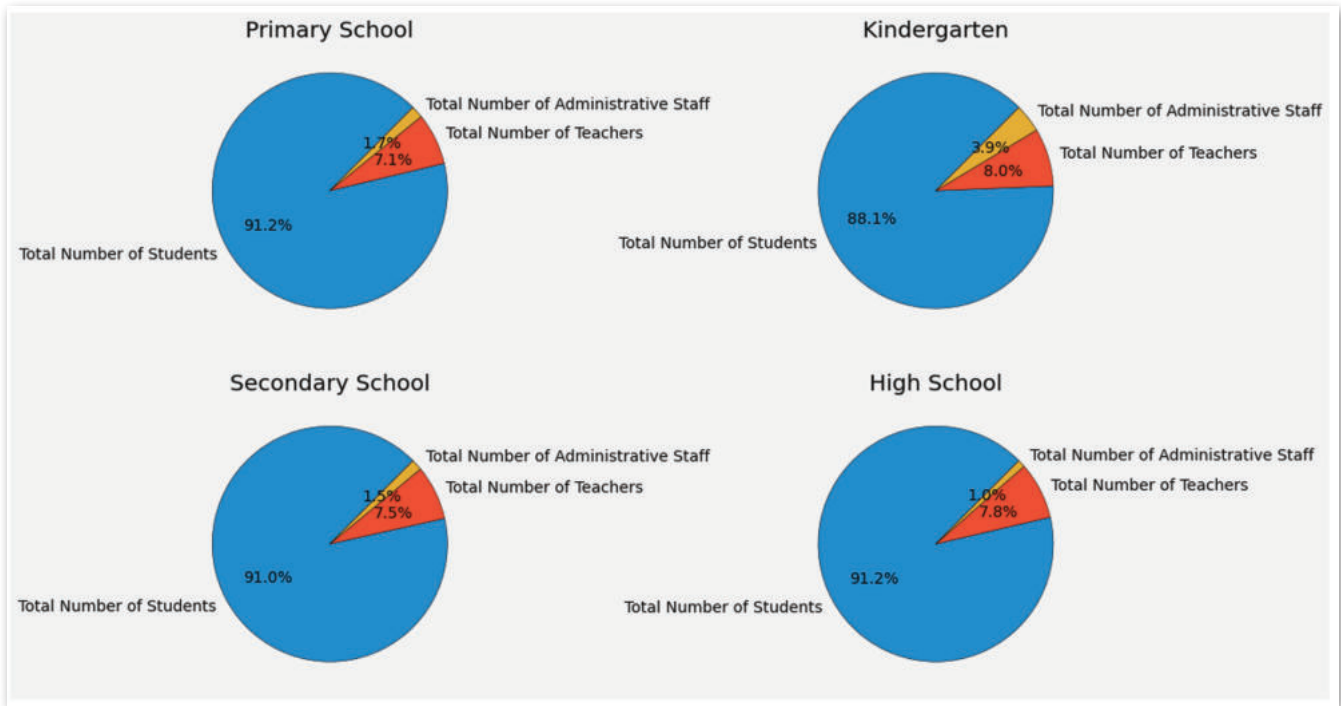


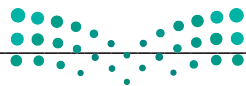
Figure 3.53: Pie charts

Exercises

- 1 Discuss the importance of data visualization as a stage in exploratory data analysis. Illustrate its importance with two examples.

- 2 Compare the main characteristics of line and bar charts. Find two examples of datasets and select the most appropriate chart for each one.

- 3 Identify the main difference between the scatter plot and the other charts. Give an example of the use of a scatter plot.



4 Name some Python libraries that can be used for applying data visualization techniques. What must you do to start using them in Jupyter Notebook?

5 You want to figure out how many tourists are visiting KSA per month for one region in the dataset.

> What kind of chart would be the most appropriate to use? (Justify your answer.)

> Choose any region from the dataset and, using the Matplotlib library, create the kind of chart you think is the most appropriate.

> Based on the chart you created, figure out which month had the most visitors for the region you chose.

6 You want to compare the number of tourists visiting KSA from 3 parts of the world, Europe, Asia, and the Middle East, for the months of October to January.

> What kind of chart would be the most appropriate to use? (Justify your answer.)

> Create the kind of chart you think is the most appropriate.

> Based on the chart you created, figure out for every month from October to January which part of the world the most tourists came from.

7 You want to find out the month with the most visitors and then figure out the percentage of visitors to each region for this month.

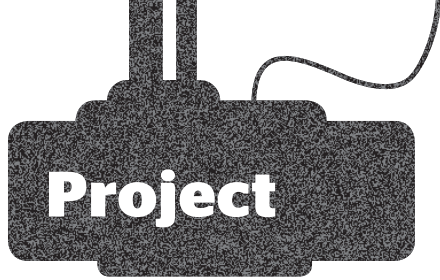
> Create a DataFrame to determine which month has the greatest number of visitors.

> What kind of chart would be the most appropriate in this case? Justify your answer.

> Create the kind of chart you think is the most appropriate.

> Based on the chart you created, which region has the highest percentage of visitors and which one has the lowest percentage of visitors?





You want to figure out which the most preferred way for tourists to visit the KSA is. You have an excel file called "tourist-indicators.xlsx" which contains information about the number of tourists arriving in KSA by air, land and sea per month.

1

Open the file "tourist-indicators.xlsx".

2

Load the "I7" sheet into a new DataFrame using the Pandas library.

3

Find out how many tourists arrived by air, land and sea for each month.

4

Compare the mean number of tourists arriving in KSA by air and by land for the months of January, February and March using the appropriate visualization technique.

5

What are the percentages of each arrival mode for the 3 months with the lowest total number of visitors?

To answer the question, you need to create a new column in your DataFrame with the total number of visitors per month.

6

What kind of graph will be more helpful to answer this question? Justify your answer.

Wrap up



Now you have learned:

- > the steps of the data analysis process.
- > how to use Jupyter Notebook as a data analysis tool.
- > how to use Pandas library to create statistics.
- > the importance of data visualization.
- > how to use matplotlib.pyplot library to graphically represent data.
- > how to create bar charts and pie charts in Jupyter Notebook.

KEY TERMS

Attribute	Filtering	Non-Graphical Analysis
Data Cleaning	Function	Predictive Analysis
Data Visualization	Graphical Analysis	Prescriptive Analysis
DataFrame	Grouping	Programming Library
Descriptive Analysis	Indexing	Series Object
Diagnostic Analysis	Method	Univariate
Exploratory Data Analysis	Multivariate	



4. Predictive data modeling and forecasting

In this unit, students will obtain basic knowledge about predictive data modeling and forecasting.

More specifically, students will learn what predictive modeling is and the different types of predictive models and their applications.

Additionally, students will learn what forecasting is, as well as the different ways of illustrating the obtained results of a forecast. Special mention will be made of the optimization problems, focusing on how to formulate a problem and seek possible solutions, using Excel Solver. Finally, students will learn how to assess the obtained results, focusing on optimal conclusions for future actions.

Learning Objectives

In this unit, you will learn to:

- > Define what predictive modeling is.
- > Describe the predictive modeling categories.
- > Understand the process of predictive modeling.
- > Recognize the pros and cons of predictive modeling.
- > Define what forecasting is.
- > Define the steps of forecasting.
- > Make a forecast in Microsoft Excel.
- > Understand the concept of the confidence interval.
- > Categorize the different forecast charts.
- > Define what an optimization model is.
- > Understand the process of optimization using Excel Solver.
- > Assess the optimization results and determine future actions.

Lesson 1

Predictive Data Modeling

Link to digital lesson



www.iien.edu.sa

When conducting predictive analysis, organizations may employ Predictive Modeling to help them make better business decisions. They can use predictive models to understand their consumer bases, potential sales prospects, or account-related security issues.

What is Predictive Modeling?

Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining and machine learning. Companies employ predictive analytics to find patterns in this data to identify risks and opportunities. The National Meteorological Service collects daily data on variables such as temperature, humidity, etc. to be able to predict the weather in the coming days.

Predictive models are widely employed in the healthcare industry to improve diagnostic methods and effectively treat terminal or chronically ill patients. Human resources departments and companies use predictive models to hire staff, and banks use them to detect fraud.

Predictive Modeling

A statistical technique in which a company uses past results and data to predict future events.

Example

When the coronavirus disease (COVID-19) became a pandemic affecting all countries worldwide, health officials relied on data scientists to model the epidemiological behavior of the disease and predict infection and mortality rates. With these models as tools, health professionals and medical researchers could develop methods to control the disease and minimize the effects.

Researchers from King Saud University in Saudi Arabia with the collaboration of other universities conducted a study to predict the COVID-19 disease spreading in Saudi Arabia, as well as to gain insight into the dynamic behavior of the infection using predictive models and simulations. The scientists used real data from the Saudi Ministry of Health to feed their models of the epidemic and generate a prediction of infections. This prediction assisted in the decision making of the Saudi Arabian government allowing them to take effective control and prevention measures such as travel restrictions and the closure of schools and mosques. These measures had the maximum impact on delaying the epidemic peak and slowing down the infection rate.

As the days were passing and real data were available, the disease spreading prediction model could be evaluated by comparing predicted and actual infections. The number of newly confirmed cases was decreasing as the measures, such as lockdown and travel limitations, were being implemented. Figure 4.1 shows that the researchers' predictions were very close to what actually happened. The bars show the cumulative actual infections, and the line shows the predicted infections. The chart also shows the dates when the restrictions were imposed.

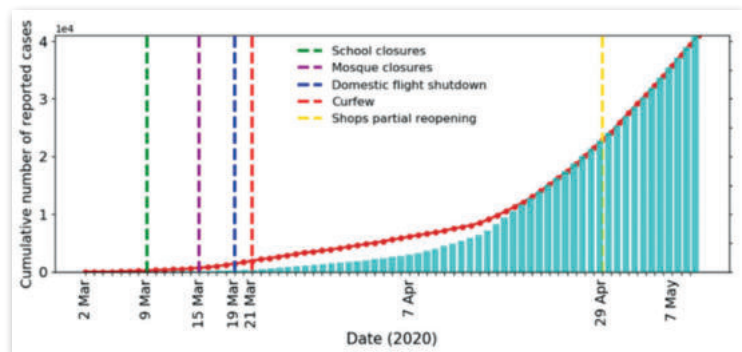


Figure 4.1: Evaluation of the predictive model with actual and simulated cumulative number of recorded cases per day

Predictive Modeling Categories

In prediction, the task of the learner is to approximate the function that maps the input variables to the outputs (predictions) in the training data. However, the configuration (form and parameters) of the function is undetermined. Once this functional relationship is obtained, one can exploit it to predict the values of the outputs, based on measurements from the respective input variables. There are two categories of Predictive Modeling: A model with a set number of parameters is a parametric model, whereas a model without a set number of parameters is referred to as non-parametric.

1. Parametric Models

Assumptions are an essential part of any data model, they improve predictions and make the model easier to understand. A parametric model makes specific assumptions about the form of the mapping function and assumes a set of parameters of predetermined size, independent of the number of training examples. Thus, a parametric model summarizes training data through this set of parameters.

Parameter

A parameter can be described as a configuration variable that is intrinsic to the model.

2. Non-Parametric Models

Non-Parametric Machine Learning models do not make strong assumptions about the mapping function. Such models can pick up any functional form from training data. Non-Parametric models are therefore an excellent choice for analyzing large volumes of data about which you have no prior knowledge.

Analytics professionals frequently feed predictive models with data from the following sources:

Transactional data

Customer data

Medical data

Financial data

Demographic information

Geographic data

Digital marketing data

Web traffic statistics

Table 4.1: Comparison of parametric and non-parametric models

Criteria	Parametric	Non-Parametric
Training data	Parametric models require less training data than non-parametric ones.	Non-parametric models require far more data than parametric ones to estimate the mapping function.
Training speed	Parametric models are computationally faster and can be trained faster because they have fewer parameters to train.	Non-parametric models take longer to train because there are more complex relationships to be estimated during the training process.
Fit	Methods of parametric models do not offer the best fit for data. They are not likely to perfectly match the mapping function.	Non-parametric models may provide more accurate predictions because they fit the data better than parametric models, but these algorithms are more prone to overfitting.
Complexity	Methods of parametric models are simple to interpret and understand.	Methods of non-parametric models are more complex and harder to interpret and understand.

Predictive Modeling Tasks

The most basic and widely used models of Predictive Modeling are Classification and Regression:

1. Classification

A classification model assesses the input values of a variable and then tries to classify them into a group, making the output data. Therefore, the variable to be predicted has discrete values. For example, it could be a simple yes or no answer to a question. The classification model is often used in retail and finance because it quickly collects information and puts it into groups to answer questions.

2. Regression

A regression model tries to find mathematical rules that connect two variables so it can predict the one variable if it knows the other. The input variable is called the independent variable and the output variable is the dependent variable. This model predicts the dependent variable values using the independent variables. The graph showing this connection is normally a straight line (linear regression) that is closest to all the independent data points. As an example, a regression model can predict how long a person will stay in a hospital when the person first goes into the hospital (number of days or dependent variable), given a parameter like the person's heart beat (independent variable).

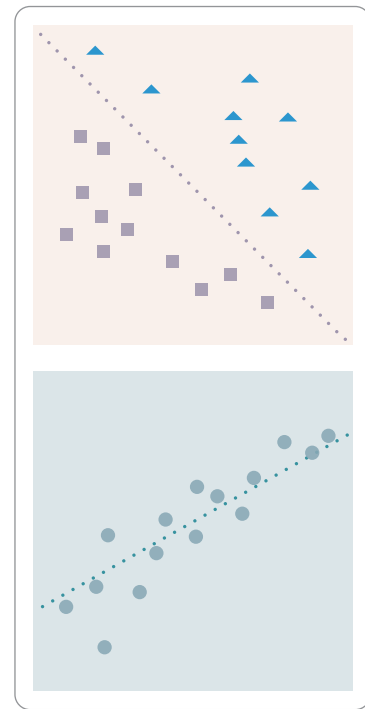


Figure 4.2: Classification vs Regression example. In classification, the dotted line represents a linear boundary that separates the two classes, while in regression, the dotted line models the linear relationship between the two variables.

Table 4.2: Comparison between Classification and Regression

Classification	Regression
Classification is the problem of predicting a discrete class label output (meaning that the output variable must be a whole number).	Regression is the problem of predicting a continuous quantity output (meaning that the output variable must be a continuous value or a real number).
The classification algorithm is used to map the input value (x) with the discrete output variable (y).	The regression algorithm is used to map the input value (x) with the continuous output variable (y).

Other common tasks of Predictive Modeling are:

3. Forecasting

Forecasting models generate numerical responses and make estimates based on the analysis of historical data. Investment companies use them to predict closing values of stocks on a daily basis or on a long-term basis. They are characterized by their versatility and, for this reason, they are the most common prediction models.

4. Clustering

A clustering model categorizes data based on similar characteristics and then uses each group's data to determine large-scale outcomes for each cluster. It operates through two types of clustering: the hard clustering (which categorizes data by determining whether each point belongs to a particular cluster entirely) and the soft clustering (which assigns a probability to each data point). Businesses can use a clustering model to determine marketing strategies for specific consumer groups.

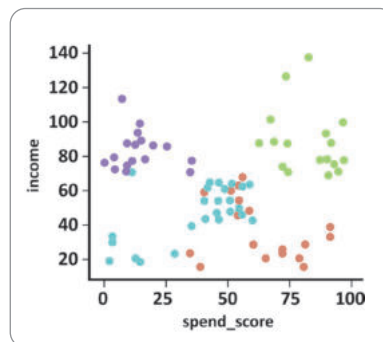


Figure 4.3: A clustering example with four clusters based on two characteristics, income and spending score

5. Outlier Detection

An outlier is an unusual or outlying data value in a dataset. Outlier detection models can examine specific instances of unusual data and connections to other categories and numbers.

6. Time Series

Time series models use past trends and data points from a specific time sequence as input factors in a dataset, in order to predict future trends or occurrences. They can forecast multiple trends and projects simultaneously, or they can concentrate on a single project. Time series models can also analyze external factors, such as seasonal variations, that may influence future trends. For example, an electronic manufacturing company can use a time series model to analyze processing times over the last year. The model can then forecast the monthly average processing time.

More advanced Predictive Modeling methods are used in more complex problems.

Predictive Modeling Methods

Decision trees

Gradient boosted

General linear models

Neural networks

Prophet models

The Predictive Modeling Process

Predictive Modeling involves the execution of algorithms on datasets to create predictions. This is an iterative process in which the model is trained, validated and refined in order to obtain the information best suited to an organization's needs. The basic steps of a typical Predictive Modeling procedure are the following:

1. Data Collection and Cleaning

Data is collected from all sources to extract the necessary information and cleaned with operations that eliminate noisy data to obtain accurate estimates. Transaction and customer assistance data, survey and economic data, demographic and geographical data, machine and web-generated data, etc., are all included.

2. Data Transformation

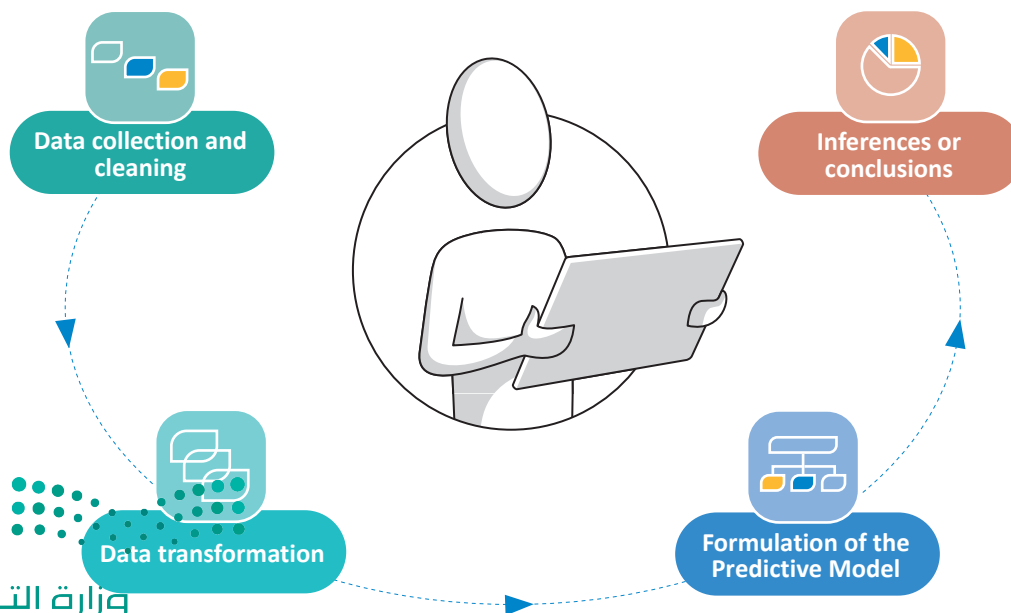
To obtain normalized data, data must be transformed using precise processing. Values are scaled to a given range and extraneous elements are removed using correlation analysis to obtain the final data.

3. Formulation of the Predictive Model

The formulation of a Predictive Model frequently involves selecting the proper prediction methods according to the required task. For example, for a classification task, a decision tree may be selected, while for a regression task, a gradient boosted model may be considered. During this process, training and test data are identified. The algorithm for the selected method is trained using the available training data. The resulting model is then applied to test data to determine the model's performance.

4. Inferences or Conclusions

Finally, inferences are exported from the model and conclusions are drawn that help answer the business questions.



وزارة التعليم

Ministry of Education

2023 - 1445

Figure 4.4: The workflow of the predictive modeling process

Practical Classification Example

The objective of this example is to show how you can build a predictive model in the context of data science. Imagine you are working on a project whose goal is to inspect concrete buildings for cracks. Because this process can be dangerous and difficult for humans, buildings can be really tall, you need to build a machine learning model that can look at a picture of concrete and classify it as positive if there is a crack and negative if there is not. This model could then be integrated with a drone which would perform the inspection much more safely.

To train a model, you need data. Once you have obtained the data you need to separate them into two basic categories or classes. One class will be images of concrete that has cracks and the other class will be images of concrete that doesn't have cracks.

Additionally, you must split this image dataset into two separate datasets.

> *A training dataset which includes the images that you will use to train the machine learning model.*

> *A test dataset which includes images the model hasn't seen, and with these images you will test and evaluate the model's performance.*

In both training and test datasets there must be images of both classes.

To train a model to classify concrete images, you will use an online tool called Teachable Machine which is available at <https://teachablemachine.withgoogle.com> where you will upload images from the folder Images for classification on your computer.

To create and train a model:

- > Open a browser and go to <https://teachablemachine.withgoogle.com>. **1**
- > Click **Get Started**. **2**
- > Click **Image Project**. **3**
- > Click **Standard image model**. **4**
- > Rename **Class 1** to **Positive** and **Class 2** to **Negative**. **5**
- > Click **Upload** for the positive class. **6**
- > Click **Choose images from your files, or drag & drop here** **7** to select and upload the training set of images that have cracks in the concrete from the **Positive** subfolder of the **Images for classification** folder, in **Documents**.
- > Repeat the process to select and upload the training set of images that do not have cracks in the concrete from the **Negative** subfolder of the **Images for classification** folder, in **Documents**. **8**
- > Click **Train model**. **9**



Teachable Machine

https://teachablemachine.withgoogle.com

1

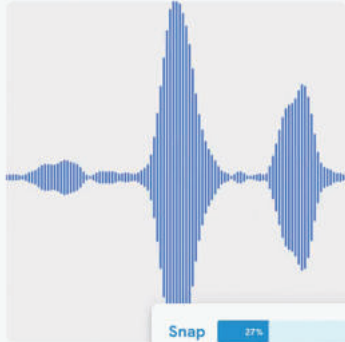
About FAQ Get Started

Teachable Machine

Train a computer to recognize your own images, sounds, & poses.

A fast, easy way to create machine learning models for your sites, apps, and more – no expertise or coding required.

Get Started 2



Snap 27%

Clap 65%

ml ml p5.js Coral node.js ARDUINO

Teachable Machine

https://teachablemachine.withgoogle.com/train

Teachable Machine

New Project

Open an existing project from Drive.

Open an existing project from a file.

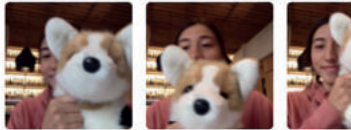
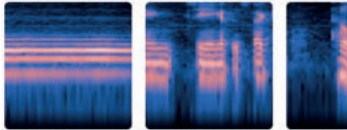



Image Project

Teach based on images, from files or your webcam.



Audio Project

Teach based on one-second-long sounds, from files or your microphone.



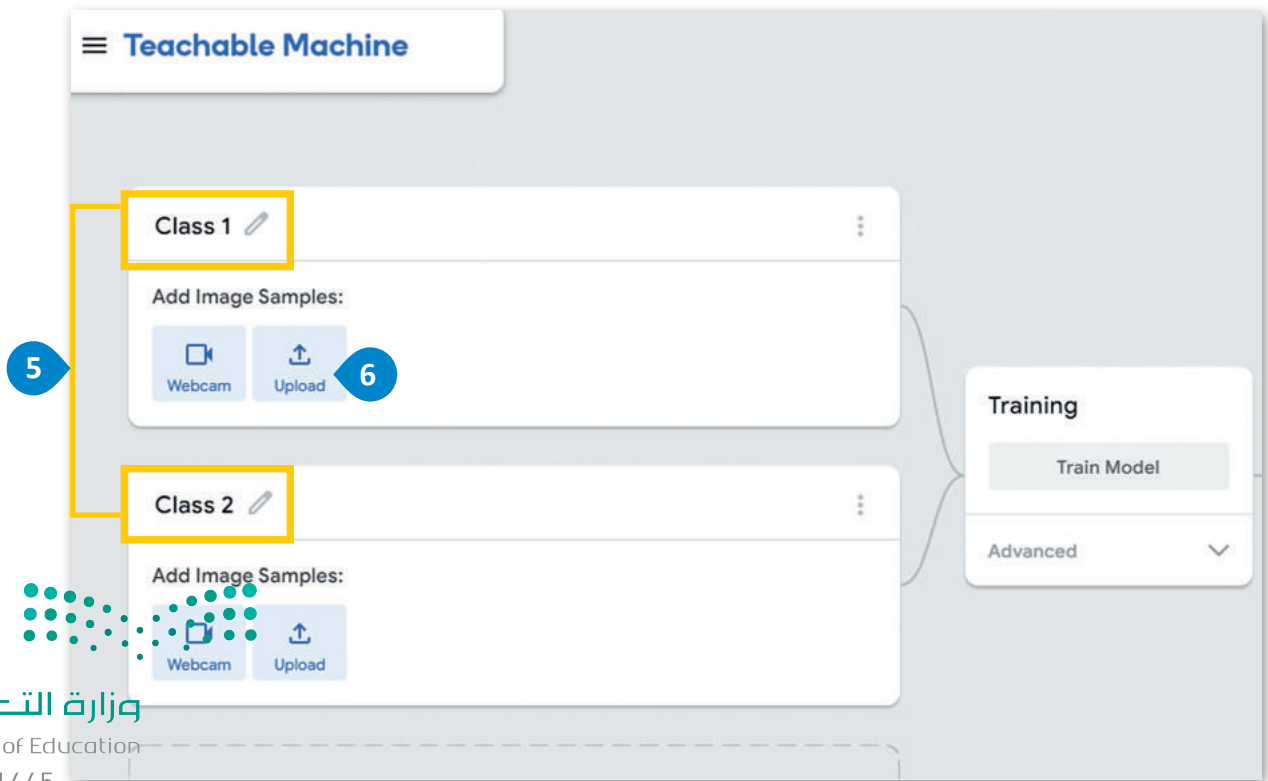
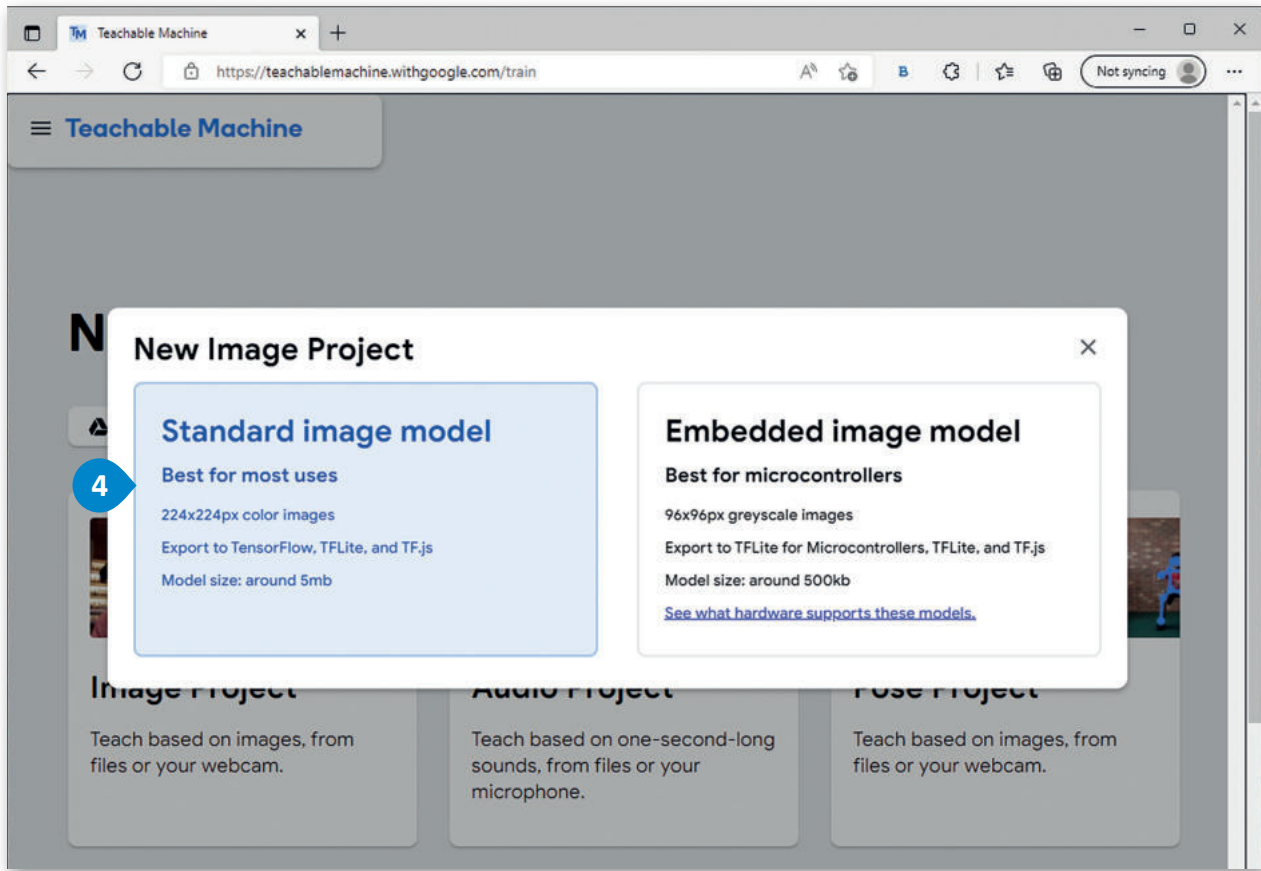
Pose Project

Teach based on images, from files or your webcam.

3

وزارة التعليم
Ministry of Education
2023 - 1445

English release-2-4-4 - 2.4.4#95c54c



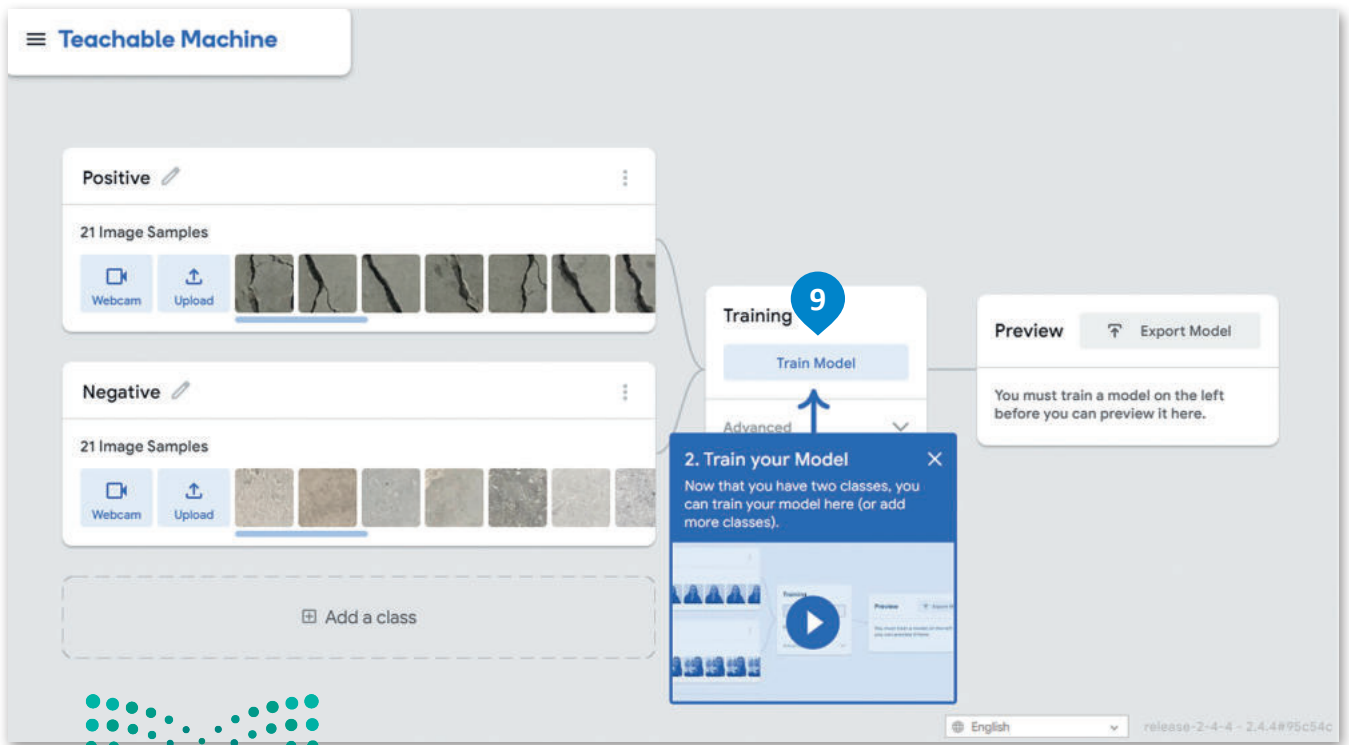
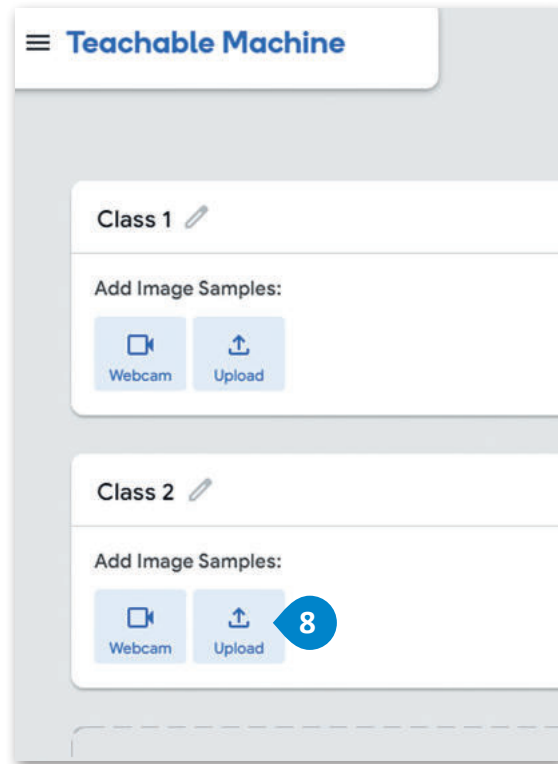
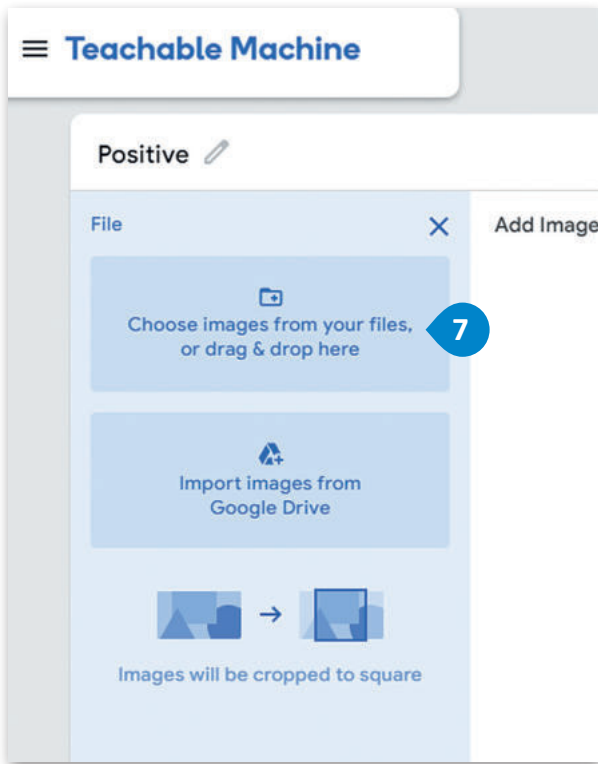


Figure 4.5: Create and train a model

وزارة التعليم

Ministry of Education
2023-1445

Once the training process is finished, you can test the model by giving it an image from the test dataset, either from the Positive class or the Negative class and evaluate the output.

To test and evaluate a model:

- > Click **Choose images from your files, or drag & drop here.** 1
- > Select and upload an image that has cracks in the concrete, from the **Test** subfolder of the **Images for classification**, in **Documents.** 2

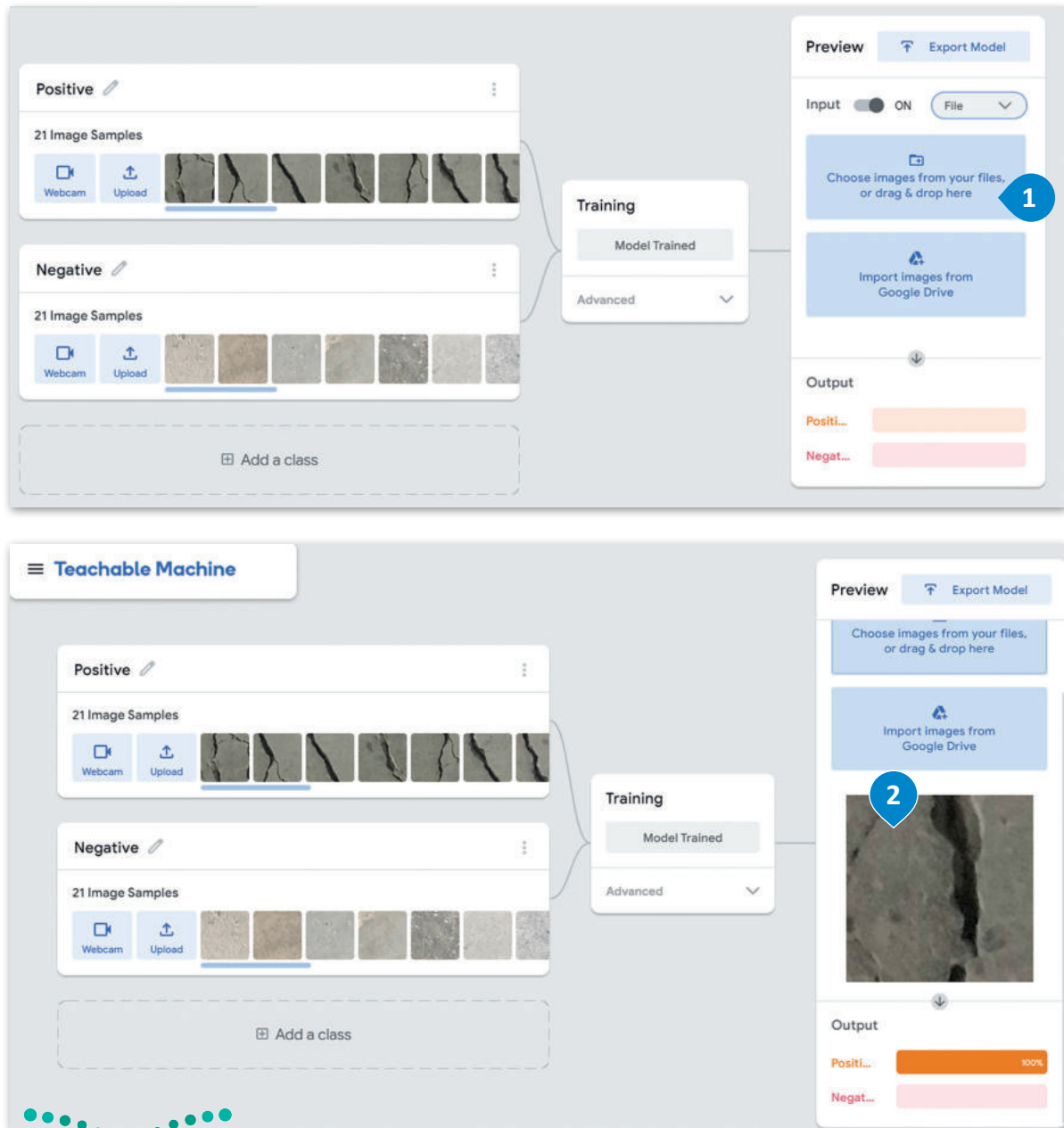


Figure 4.6: Test and evaluate a model

As you can see, the model correctly classified the image in the Positive class with 100% certainty, which is as expected because the concrete in the image you uploaded has a crack. You should repeat the last two steps to upload a different image and evaluate the model again.

Benefits and Limitations of Predictive Modeling

Benefits of Predictive Modeling:

Improves marketing, sales, and customer service strategies.

Improves knowledge of the competition and employment of strategies to gain a competitive advantage.

Enhances current products or services.

Improves recognition of consumer requirements.

Provides forecasts for external factors that may have an impact on productivity or workflow.

Improves recognition of financial risks.

Provides inventory forecasting or resource management procedures.

Predicts future of trends.

Supports workforce planning and churn analysis.

Limitations of Predictive Modeling:

Security and privacy of data.

Handling large volume of data.

Management of data.

Adapting models to new business problems.

Predictive Modeling tools

Modern Predictive Modeling tools provide all-in-one platforms that support algorithm development, data analysis and the output of reliable results. These tools are used by businesses and research organizations to produce accurate and comprehensive conclusions that can lead to effective decision-making.

Available tools:

H2O Driverless AI

IBM Watson Studio

RapidMiner Studio

SAP Analytics Cloud



IBM SPSS

Oracle DataScience

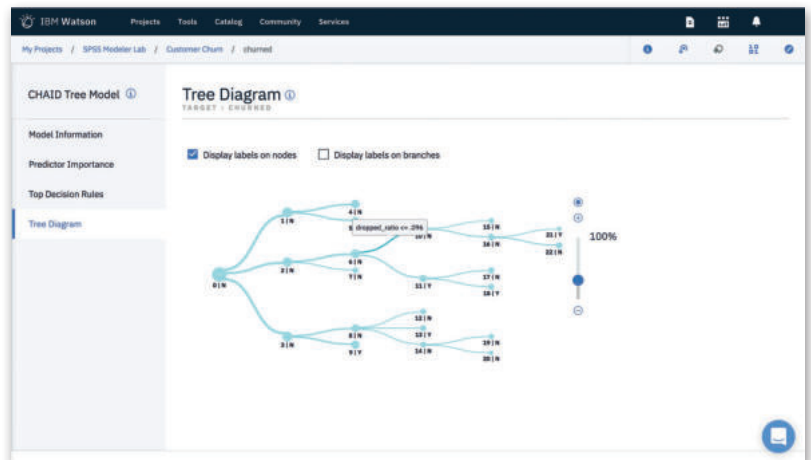


Figure 4.7 The workflow of Data Analysis and Transformation

Table 4.3: Applications of Predictive Modeling

Application	Description
Sales	Predictive analysis can decide a company's future in terms of sales and profits, by detecting anomalies in past data. Modeling can show where the sales department is lagging, resulting in improved company performance in targeted areas or demographics.
Marketing	Based on past data, marketing promotes a specific service or commodity to a group of target customers, by predicting and forecasting their reactions and requirements. Historical data is gathered and analyzed for this reason, in order to predict outcomes and types of services that a customer may desire.
Social media	Social media is an essential resource of unstructured, heterogeneous and massive data, where millions of people interact daily. For this reason, social media modeling and analytics are among the most widely used applications of Predictive Modeling, allowing organizations to detect customer activity and compute future outcomes accordingly.
Risk Assessment	It is commonly used in financial institutions and fraud detection cases where it is necessary to assess the type of risk that a person is exposed to. Predictive analytics tools can assist an organization in conducting a risk assessment and determining the degree of risk or profit that the future offers.
Quality Enhancement	Quality enhancement involves using customer feedback on a product or service to develop proposals for improving product or service quality. It is also used for testing the proposed changes in order to predict how they will perform in the market.



Exercises

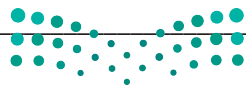
1

Read the sentences and tick ✓ True or False.	True	False
1. Companies employ predictive analytics to find patterns in this data to identify risks and opportunities.	<input type="checkbox"/>	<input type="checkbox"/>
2. As you push towards higher accuracy, models become more complex and harder to interpret.	<input type="checkbox"/>	<input type="checkbox"/>
3. Complex variables, e.g. human behavior, are one of the reasons that a model may fail.	<input type="checkbox"/>	<input type="checkbox"/>
4. One of the requirements of an effective predictive model is to begin the process with relevant data.	<input type="checkbox"/>	<input type="checkbox"/>
5. A challenge of Predictive Modeling is the recognition of financial risks.	<input type="checkbox"/>	<input type="checkbox"/>
6. The forecast model can't handle more than one variable at the same time.	<input type="checkbox"/>	<input type="checkbox"/>
7. The outlier model can be useful in detecting fraudulent transactions and unusual behavior.	<input type="checkbox"/>	<input type="checkbox"/>
8. The time series model can analyze external factors such as seasonality that can influence future trends.	<input type="checkbox"/>	<input type="checkbox"/>
9. A parameter can be described as a configuration variable that is intrinsic to the model.	<input type="checkbox"/>	<input type="checkbox"/>
10. Forecast models use past trends and data points from a specific time sequence as input factors in a dataset, in order to predict future trends or occurrences.	<input type="checkbox"/>	<input type="checkbox"/>

2 Briefly explain what predictive modeling is, use online research and give an example.

3 Briefly explain how to get started in creating a predictive model.

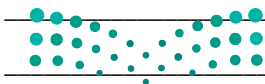
4 Describe where predictive modeling can be applied in the real world.



5 You want to build a predictive model for traffic accidents, and you need data for your model. Search on the Saudi Open Data Platform (<https://data.gov.sa>) to find the correct datasets. How many years of data, and what kind of data do you need?

6 Imagine you want to build a classification model to classify images of cars, planes and ships. Describe this process step by step, from gathering the data to training the model.

7 Perform online research to find examples of privacy and ethical concerns in predictive modeling. For example, can companies base their HR operations on prediction models using employees' health data?



Lesson 2

Forecasting

Link to digital lesson



www.ien.edu.sa

Forecasting is an estimation of future events made by incorporating and casting forward data related to the past in a pre-determined and systematic manner. A common example of forecasting is estimation of future sales or income, where the data of the previous sales or income is used as a reference for how the future ones will fare. However, forecasting can be applied in many other cases, such as: how much the population will grow next year, how many tourists will visit the Kingdom of Saudi Arabia, etc.

The terms forecasting and prediction are similar but not identical. Prediction is the process of creating a model to guess or estimate the outcome of the unseen data based on the values of the present variables, while forecasting is the process of estimating the value of a variable at some time in the future based on the previous values of the same variable given in a fixed order of time. This means that forecasting is just a prediction based on time, it implies time series and the future, whereas prediction considers features other than time. Any time you predict the future, it is a forecast. All forecasts are predictions, but not all predictions are forecasts, as when you would use regression to explain the relationship between two variables.

Forecasting

The process of making estimations of future events based on past data.

With these definitions, we can now appreciate why weather forecasting is not called weather prediction: weather forecasting predicts the weather in the future using temporal information. For example, if there is a downpour at the moment, what is the likelihood that it will still be raining in five minutes? Independent of all other features that influence the weather (e.g. atmospheric pressure and temperature), the likelihood that it will still be raining in five minutes will be high because it is raining at the moment.

By using forecasting techniques, decisions can be adjusted in order to help a company or an organization to achieve its goals. For this reason, certain steps must be followed in the forecasting procedure:

Table 4.4: Forecasting application steps

	Description
Step 1:	Determine and then obtain the data we want to use for our analysis.
Step 2:	Use a software tool to set up the dataset.
Step 3:	Set the time series we want to forecast.
Step 4:	Create the forecast.
Step 5:	Graph the data.
Step 6:	Analyze the results.

Forecasting in Excel

There are various software tools that we can use to create a model that analyzes past data in order to forecast future data. One of them is Microsoft Excel. Excel uses past time-based data in order to create a forecast. In this lesson, we will learn how to predict future tourist visits data by using Excel as an ICT tool. More specifically, we will use the forecast method in Excel to predict the 2023 tourist visits by month in the Kingdom of Saudi Arabia, based on past tourist visits data (2019).

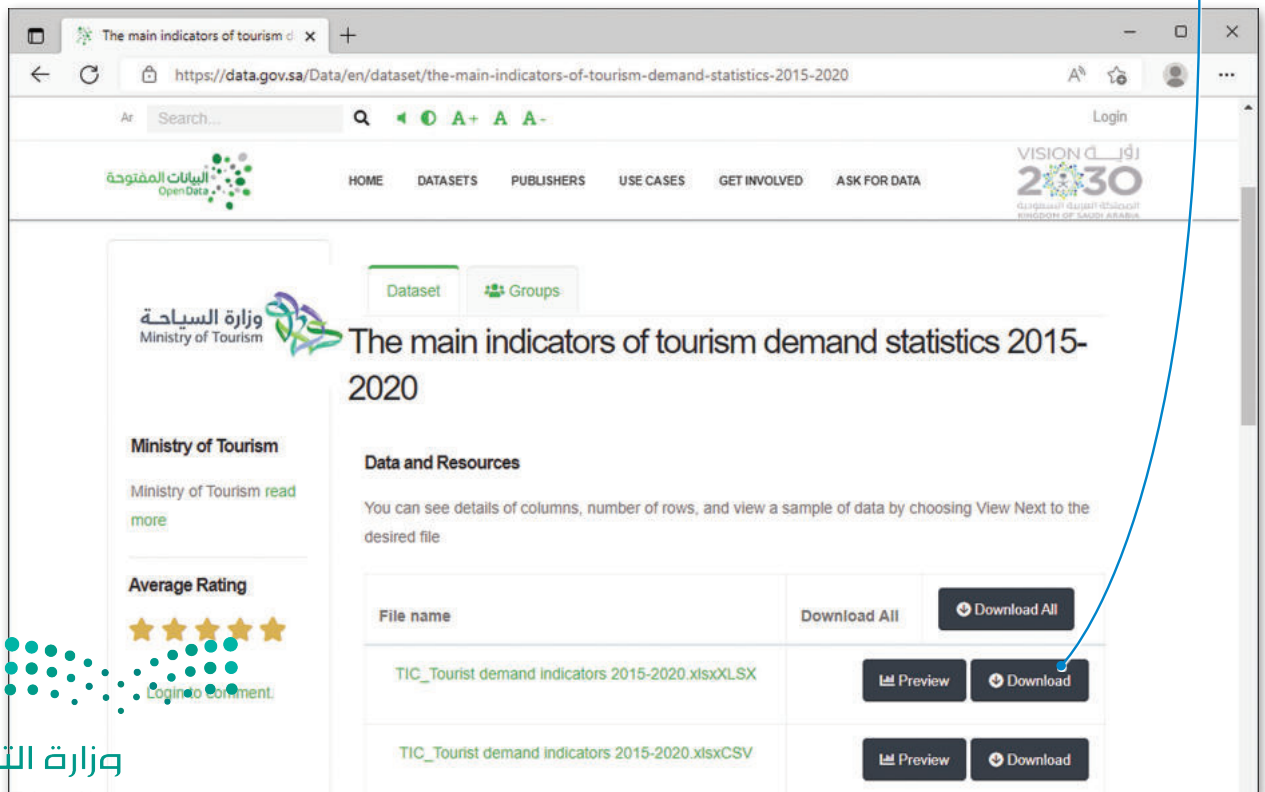
Step 1: Obtain the Data

First of all, we have to obtain the data required for our forecast analysis. Our goal is to predict monthly tourist visits for the year 2023, so what we need is the monthly tourist visits data of the previous year. For this purpose, we will obtain this data from the Tourism Intelligence Center of the Ministry of Tourism through the Saudi Open Data Platform (<https://data.gov.sa>). More specifically, we will obtain monthly tourist visits data for the year 2019 through the following link: <https://od.data.gov.sa/Data/en/dataset/the-main-indicators-of-tourism-demand-statistics-2015-2020> Although there is data regarding the years 2020 and 2021, the numbers are not useful due to the COVID-19 pandemic. For this reason, we will use the 2019 data for our forecasting calculations.



Scan the QR code to download the data file.

This is the data of the tourist visits to the Kingdom of Saudi Arabia for the year 2019, provided by the Tourism Intelligence Center (Ministry of Tourism).



The screenshot shows the Saudi Open Data Platform interface. The browser address bar displays the URL: <https://data.gov.sa/Data/en/dataset/the-main-indicators-of-tourism-demand-statistics-2015-2020>. The page title is "The main indicators of tourism demand statistics 2015-2020". The Ministry of Tourism logo is visible on the left. The "Data and Resources" section contains a table with the following data:

File name	Download All
TIC_Tourist demand Indicators 2015-2020.xlsxXLSX	Download All
TIC_Tourist demand Indicators 2015-2020.xlsxCSV	Download All

To download the data:

- > Click on the **download** button. **1**
- > In the **Downloads** window, click **Open File**. **2**
- > In the **I1** sheet select cells **C59:C70** **3** and paste them in cells **A2:A13** of a new Excel file. **4**
- > Type "Month" in the **A1** cell and add "2019" to every month value. **5**
- > In the **I1** sheet of the downloaded Excel file, select cells **D59:D70** **6** and paste them in cells **B2:B13** of your new Excel file. **7**
- > Type "Tourist visits" in the **B1** cell. **8**

The screenshot shows a web browser window displaying the data.gov.sa website. The URL is <https://data.gov.sa/Data/en/dataset/the-main-indicators-of-tourism-demand-statistics-2015-2020>. The page title is "The main indicators of tourism demand statistics 2015-2020". The page content includes a sidebar for the Ministry of Tourism, a "Data and Resources" section, and a table of files. A "Downloads" window is open, showing the file "tic_tourist-demand-indicators-2015-2020.xlsx" with an "Open file" button. A blue circle with the number "2" is placed over the "Open file" button. Another blue circle with the number "1" is placed over the "Download" button in the table below.

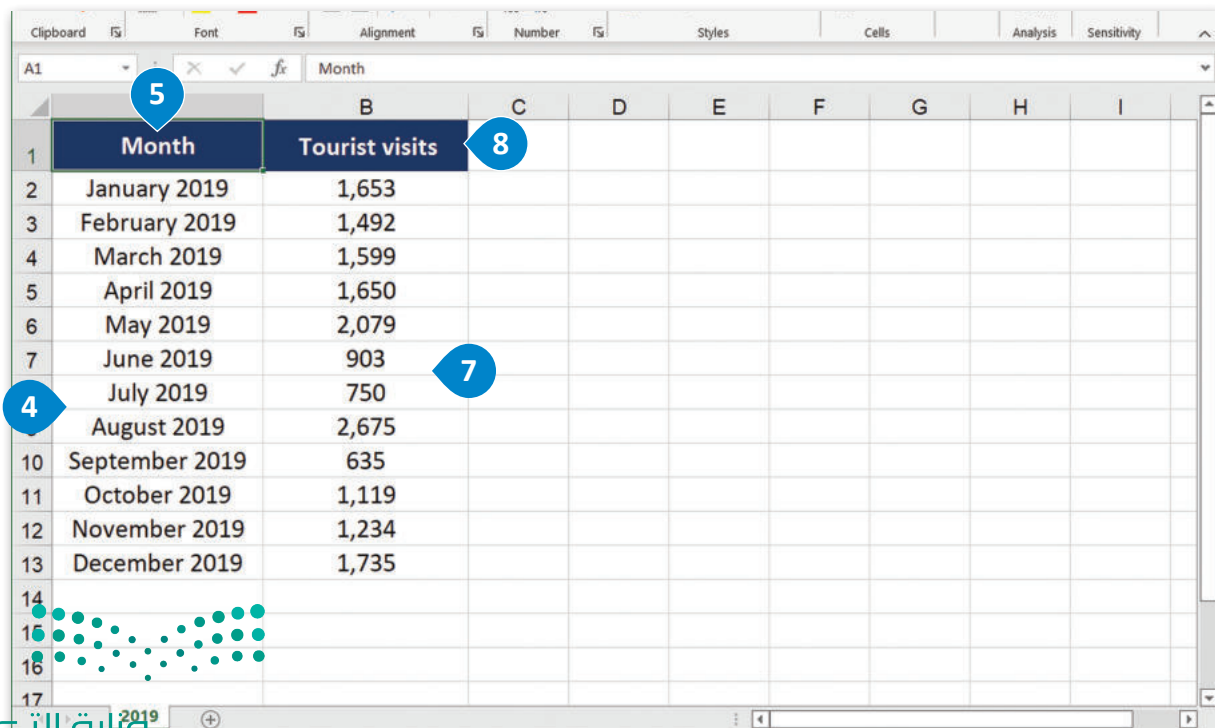
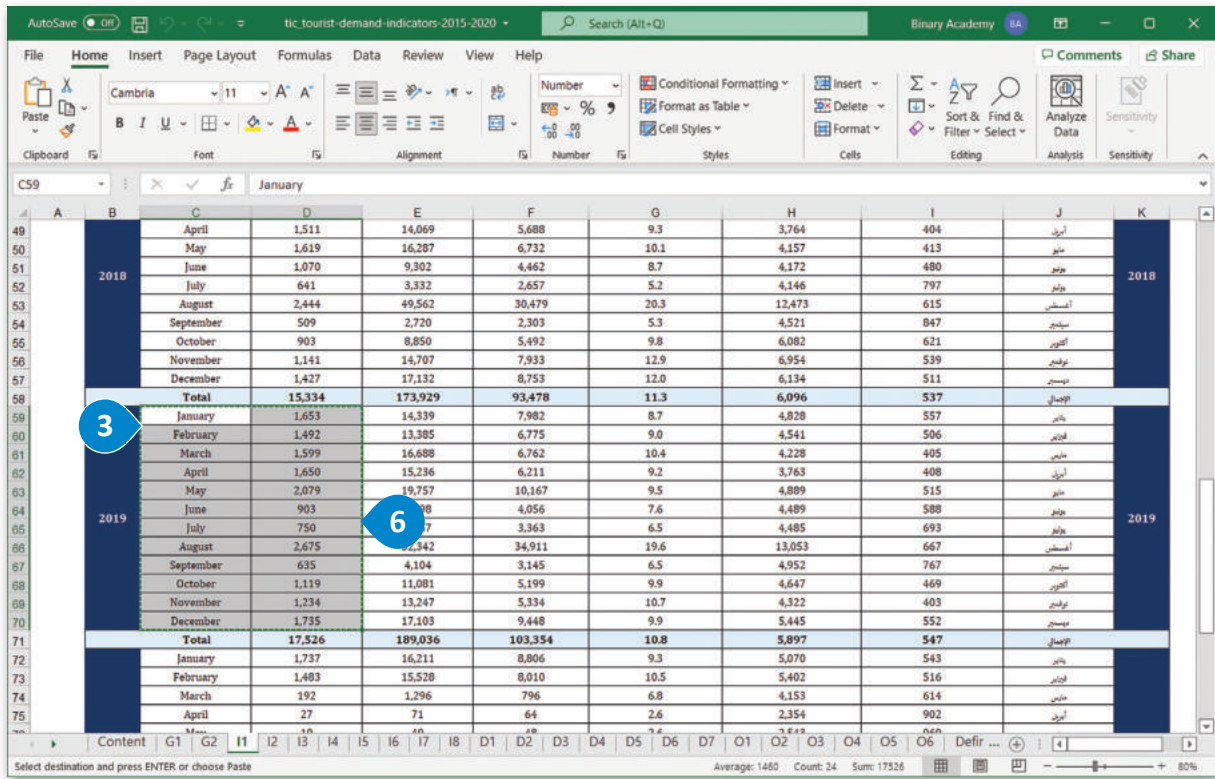
File name	Download All
TIC_Tourist demand indicators 2015-2020.xlsxXLSX	Preview Download
TIC_Tourist demand indicators 2015-2020.xlsxCSV	Preview Download



وزارة التعليم

Ministry of Education

2023 - 1445



Step 2: Use a Forecasting Tool

Once we have obtained the monthly tourist visits data for the year 2019, we have to import them into a forecasting software tool. We will use Microsoft Excel as a forecasting software tool and create two columns in a new sheet called "2019". One with the months and a second one with the tourist visits values for every month of 2019.

Step 3: Set the Time Series

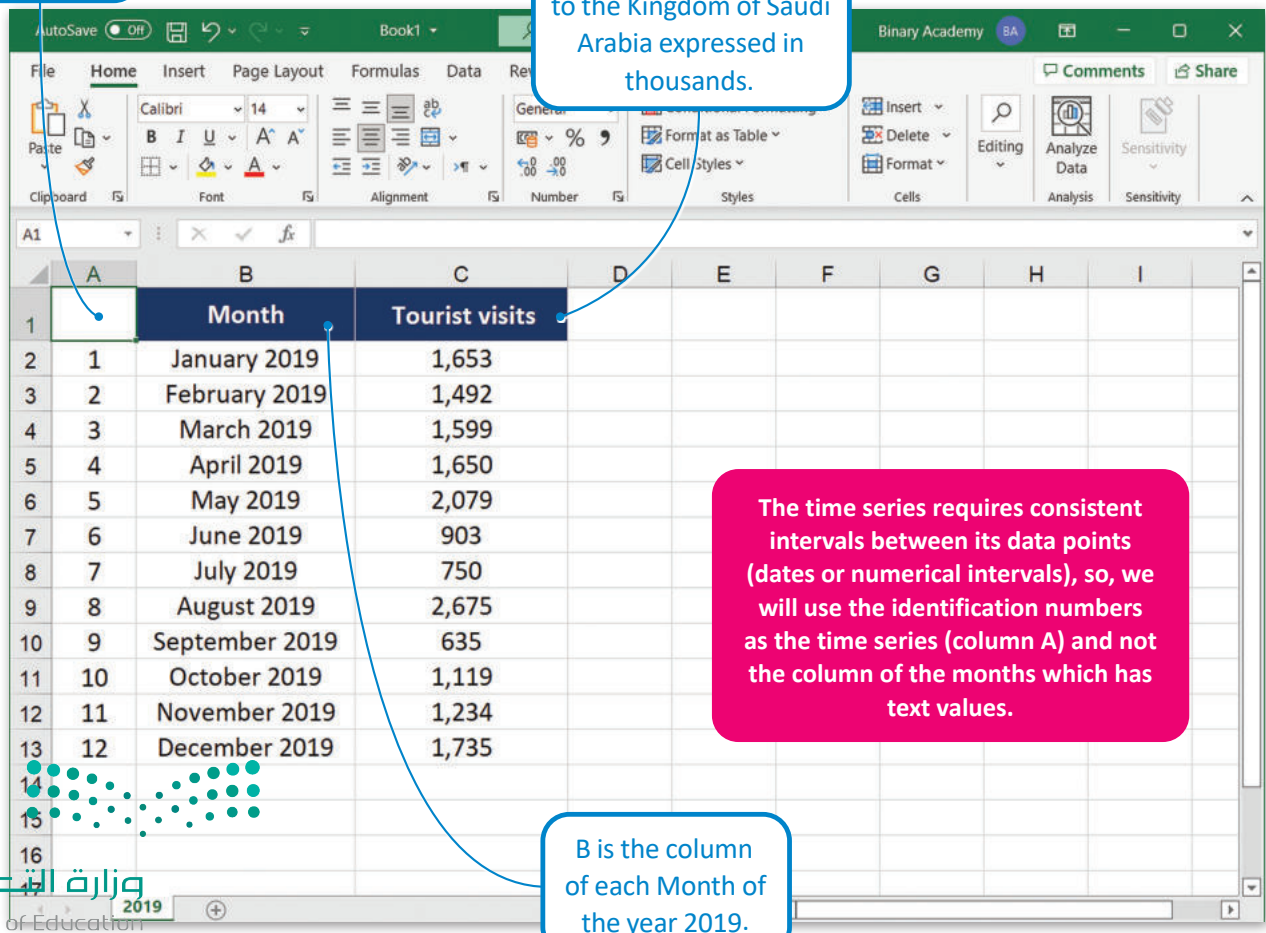
Microsoft Excel needs a column that contains a time series with numerical values (numbers or dates) in order to create a forecast. The reason that we can't use the months column as a time series column is because it contains text values. So, to solve this problem, we will create a column that contains the numbers from 1 to 12, called Identification numbers (Figure 4.10).

Time series

A set of regular time-ordered observations of a quantitative characteristic of an individual or collective phenomenon taken at successive periods of time.

Identification numbers.

Number of tourist visits to the Kingdom of Saudi Arabia expressed in thousands.



The time series requires consistent intervals between its data points (dates or numerical intervals), so, we will use the identification numbers as the time series (column A) and not the column of the months which has text values.

B is the column of each Month of the year 2019.

Figure 4.10: Excel Sheet with data

Step 4: Create the Forecast

Based on the monthly tourist visits data for the year 2019, we will use the Forecast sheet option in the Data tab of Microsoft Excel, in order to create the forecast.

To create a forecast:

- > Click on the cell **A1**. **1**
- > In the **Data** tab **2**, in the **Forecast** group, click **Forecast sheet**. **3**
- > A preview of the Forecast Worksheet will appear. **4**
- > Choose the **Line chart**. **5**
- > Set Forecast end to **24**. **6**
- > Click **Options** **7** for any changes to the additional forecast settings.
- > Click **Create**. **8**
- > Excel will create another Sheet with the forecast values. **9**

The screenshot shows the 'Create Forecast Worksheet' dialog box in Microsoft Excel. The dialog box is titled 'Create Forecast Worksheet' and contains the following elements:

- A line chart showing historical data for 'Tourist visits' from January to December 2019, with a forecast for the next 12 months (months 13 to 24). The chart includes a lower confidence bound and an upper confidence bound.
- A 'Forecast End' field set to 24.
- An 'Options' section with a radio button selected for 'Line chart' and another for 'Column chart'.
- 'Create' and 'Cancel' buttons at the bottom.

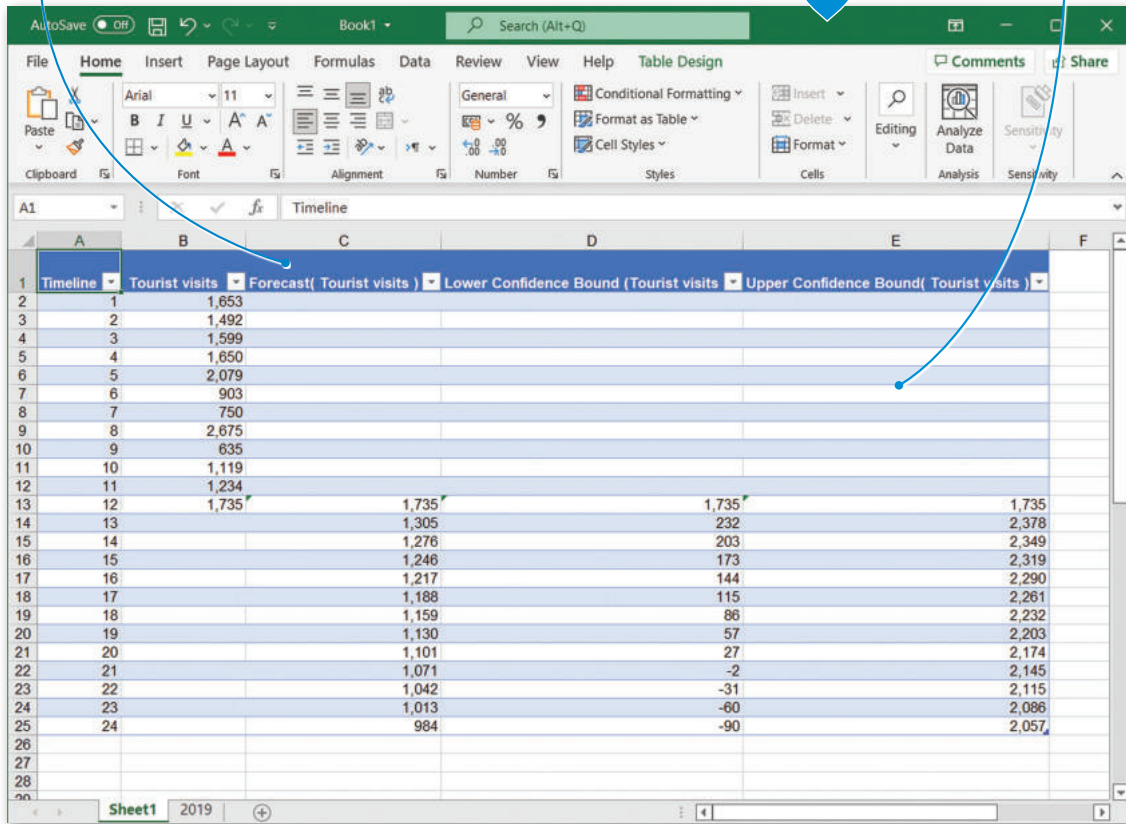
Numbered callouts (1-9) point to various elements in the Excel interface: 1 points to cell A1, 2 to the Data tab, 3 to the Forecast sheet button, 4 to the chart preview, 5 to the chart type options, 6 to the Forecast End field, 7 to the Options button, 8 to the Create button, and 9 to the new forecast sheet.

As Forecast End, we set the number 24 because we have data for 12 months and we want a prediction for the next 12 months (12 + 12 = 24).

There are two options to display the forecast, either the Line chart or the Column chart.

Column C contains the predicted values.

Columns D and E display the uncertainty of the forecast.



When we create a forecast, Excel creates a new worksheet that contains both a table of the past data values and the predicted (future) data values. The amount of uncertainty is displayed too, with the upper and lower confidence bounds. Excel also creates the chart that we've chosen to express this data.

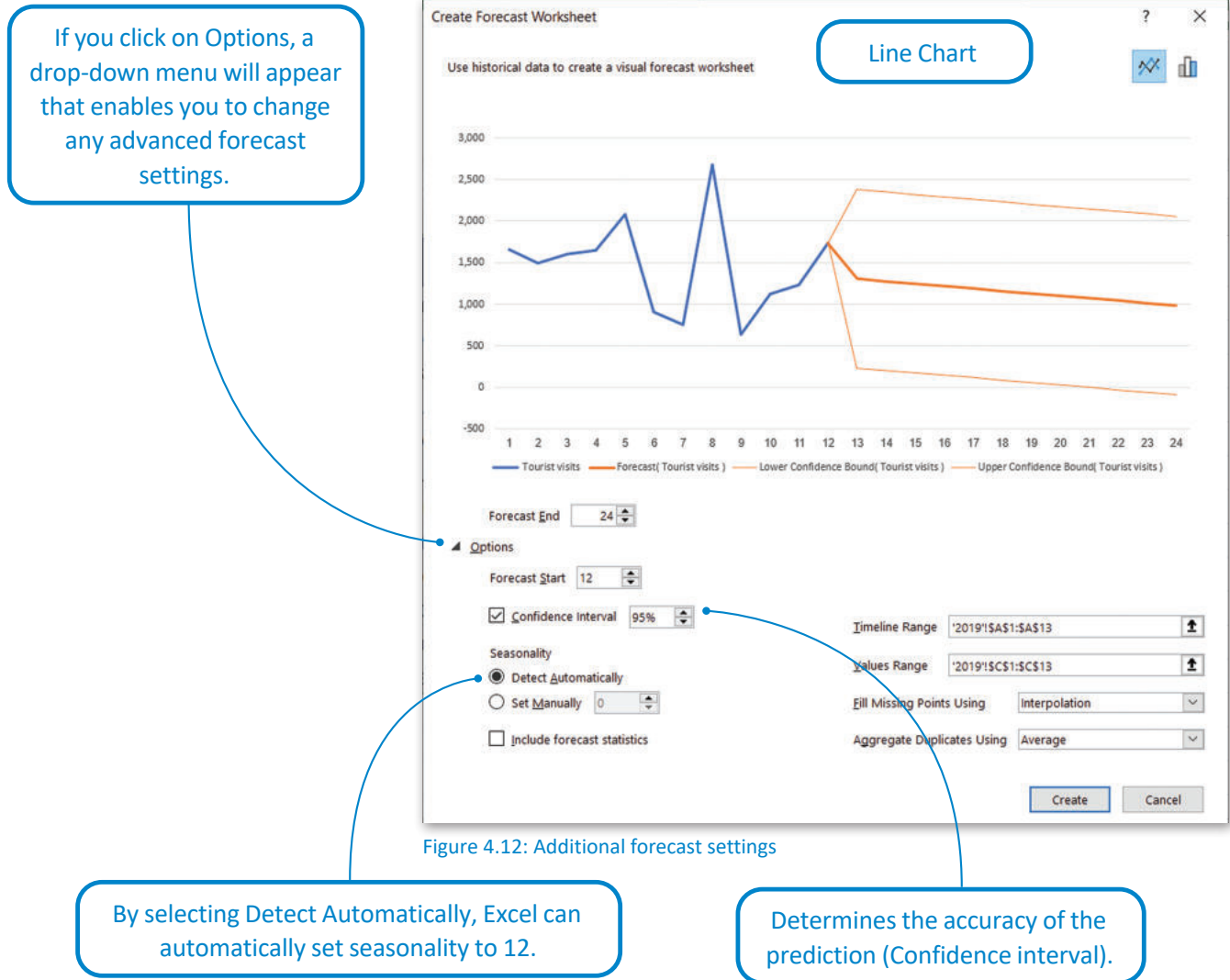
The graphical representation of our forecast.



Figure 4.11: Create a forecast

Additional Forecast Settings

The model uses Excel in order to forecast future data values based on existing values (past data) by using linear regression. Linear regression is a model that detects the relationship between two continuous variables, in order to predict the values of a dependent variable based on the values of an independent variable (in our case, the dependent variable is the tourist visits and the independent is the time/months). Linear regression is a basic and commonly used type of predictive analysis, because it allows us to summarize and study relationships between these two continuous (quantitative) variables.



Even though linear regression is the most common and the most reliable way of modeling prediction, it lacks qualitative factors. For example, in our case some qualitative factors could be the opinions of tourists, their judgment and their available free-time that influence them when it comes to choosing a place to visit on vacation. The forecast function based on linear regression may work sometimes, but the lack of a qualitative baseline is one of the key reasons that most forecasts are significantly off from real observations, a fact that could impact predictions in a negative way.

Confidence Interval

All predictions have an amount of uncertainty in them. They are not "real" values measured and obtained from research, they are estimated values, which means that they are values that do not really exist.

So, when we "guess" the value of a parameter, this means that our "guess" might prove wrong in the future. The confidence interval comes in to explain this "wrong guessing" by giving us, not only a single predicted value, but a range of predicted values. This range is determined by the lower confidence bound and the upper confidence bound, meaning that, even if our "guessing" proves wrong, the actual value that we will get should not be lower than the lower confidence bound value or greater than the upper confidence bound.

In statistics, this is called "Confidence Interval (CI)" and it is defined as a range of estimated values for an unknown parameter. A confidence interval is the mean of your estimate \pm the variation in that estimate. It is calculated at a specific confidence level, usually equal to 95%. The confidence level means that the real value has 95% chance of falling within the range of predicted values between the lower confidence bound and the upper confidence bound.

Let's take as an example the prediction that the forecast gives us for the month of January 2023. Based on the forecast formula, Excel gives us an estimated value of tourist visits for January 2023 equal to 1,305 thousand. It also gives us a lower confidence bound equal to 232 thousand and an upper confidence bound equal to 2,378 thousand. The confidence interval consists of all the values between 232 thousand and 2,378 thousand. The confidence level of the forecast method in Excel is predefined and equal to 95%, so the value of the future tourist visits for January of 2023 has 95% chance of being between the values of 232 thousand and 2,378 thousand. Now, let's suppose that in the future, the tourist visits of January 2023 prove to be equal to 1,000 thousand. This would mean that the forecast prediction was totally right, because the value 1,000 thousand may not be equal to the value 1,305 thousand that the forecast predicted, but it falls within the range of values 232 thousand and 2,378 thousand (confidence interval).

Confidence Interval

An interval which has a known and controlled probability (generally 95% or 99%) to contain the true value.

Timeline	Tourist visits	Forecast (Tourist visits)	Lower Confidence Bound (Tourist visits)	Upper Confidence Bound (Tourist visits)
1	1,853			
2	1,492			
3	1,599			
4	1,650			
5	2,079			
6	903			
7	750			
8	2,675			
9	635			
10	1,119			
11	1,234			
12	1,735	1,735	1,735	1,735
13	1,735	1,735	1,735	1,735
14	1,305	1,305	232	2,378
15	1,276	1,276	203	2,349
16	1,246	1,246	173	2,319
17	1,217	1,217	144	2,290
18	1,188	1,188	115	2,261
19	1,159	1,159	86	2,232
20	1,130	1,130	57	2,203
21	1,101	1,101	27	2,174
22	1,071	1,071	-2	2,145
23	1,042	1,042	-31	2,115
24	1,013	1,013	-60	2,086
25	984	984	-90	2,057

13	1,735	1,735	1,735	1,735
14	1,305	1,305	232	2,378
15	1,276	1,276	203	2,349
16	1,246	1,246	173	2,319

Step 5: Graph the Data

As mentioned before, the forecast can be displayed in two different charts, the line chart or the column chart:

Line Chart

Line charts are commonly used to display change over time as a series of data points connected by straight lines. The line chart helps to determine the relationship between two sets of values (for example the set of values for time and the set of values for tourist visits), with one data set always being dependent on the other set (for example, the tourist visits are dependent on time).

Benefits of line charts:

They allow a quick analysis of data.

They allow us to easily observe changes over a certain period of time.

They are adequate for data sets with up to 50 data values.

They help in making predictions about the results of data not yet recorded.

Column Chart

Column charts are used for displaying the quantities of data collected through questionnaires and interviews, such as age groups, items of products sold, etc. They can also be used for data such as tourist visits per month, but only if the number of values in the data set is not large.

Benefits of column charts:

They are great when comparison between data sets is needed.

They can summarize a large amount of data in a visual, easily interpretable form.

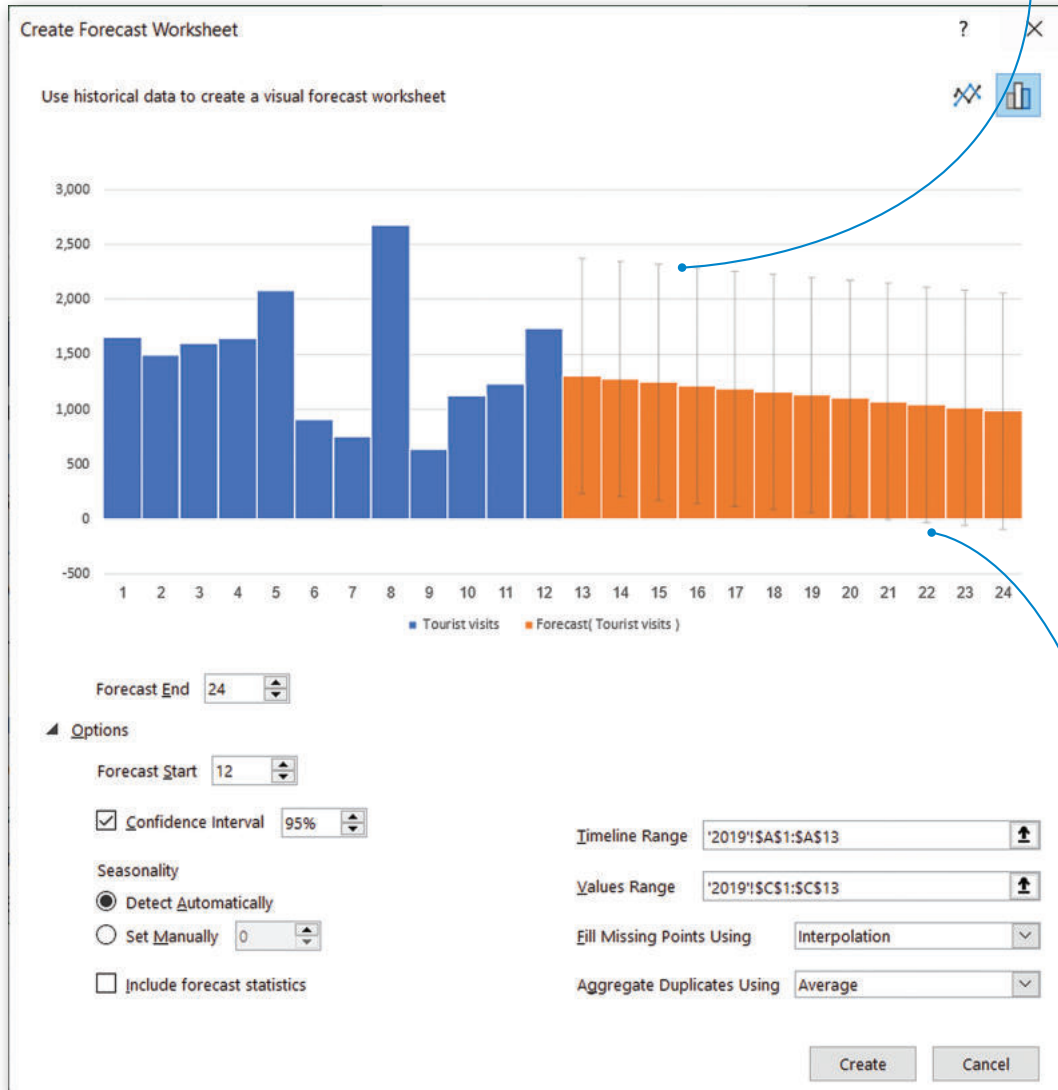
They make trends easier to observe.

They help in studying patterns over a long period of time.



The Forecast Sheet procedure gives us the opportunity to select between a Line Chart and a Column Chart. In our case, we chose the Line Chart for a more suitable visual representation of the information.

The top edges of all the gray lines indicate the upper confidence bound values.



The lower edges of all the gray lines indicate the lower confidence bound values.

Figure 4.14: Column chart

Customize the Graphics

In the new Excel sheet that contains the forecast values, the lower and upper confidence bounds columns are in a fixed form that Excel generates on its own. We can change the names of the columns by simply editing them.

To change the names of the columns:

- > In the **Forecast** sheet, click cell **D1**. **1**
- > Select the words written in the cell "(Tourist visits)" **2**, delete them and press **Enter**.
- > Do the same in the cell **E1**. **3**
- > The changes will be applied in the **Forecast** sheet **4** and in the **line chart** also. **5**

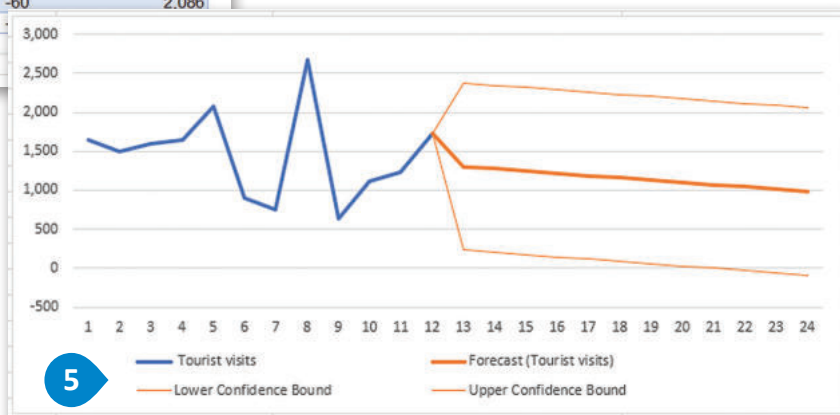
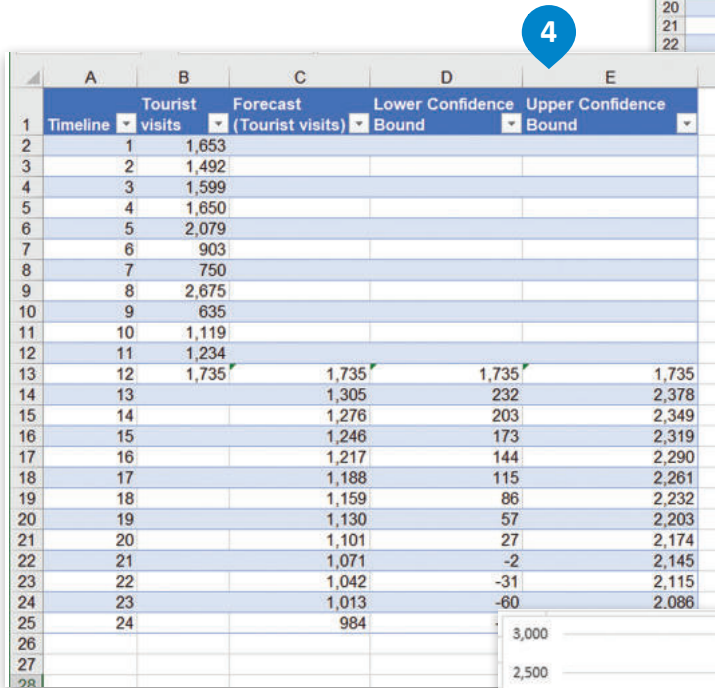
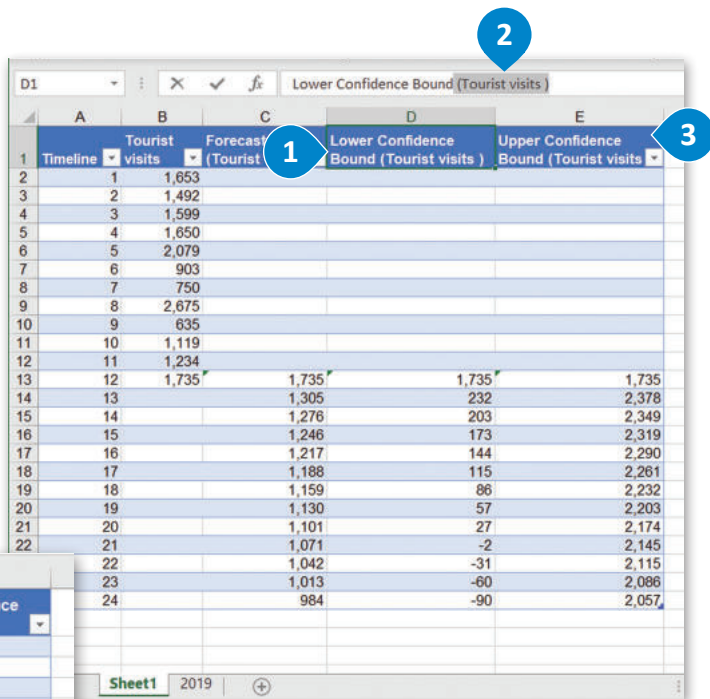


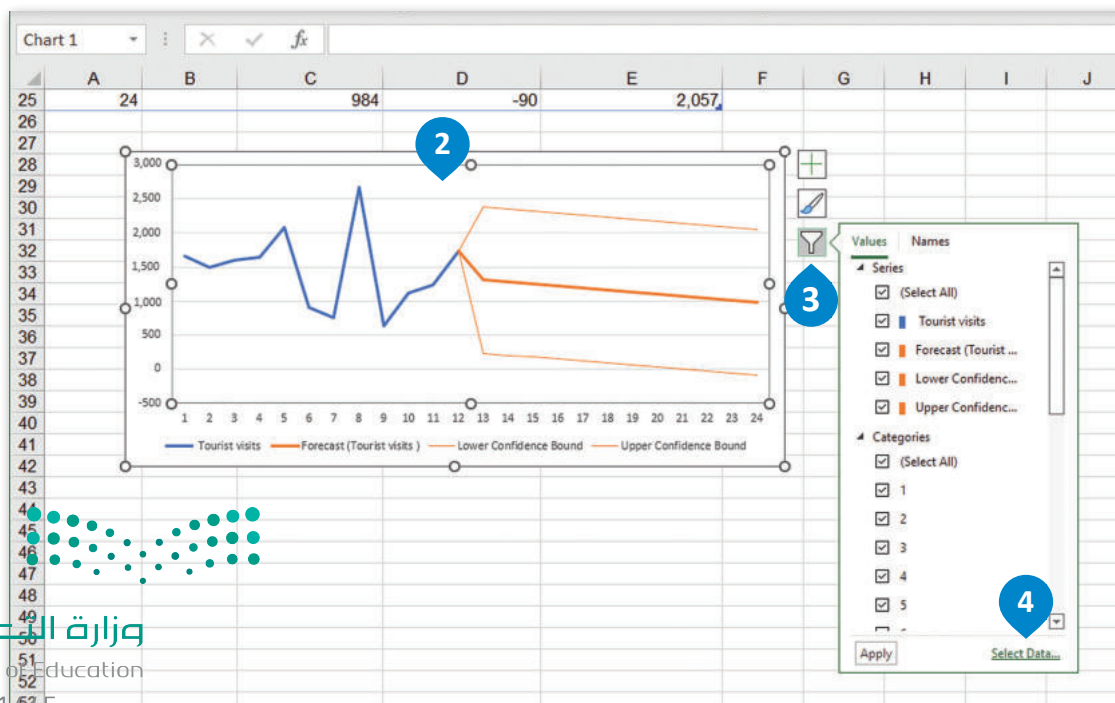
Figure 4.15: Change the names of the columns

As we can see, in the line chart that we created the horizontal axis labels are not the appropriate ones. Instead of having the names of the months, our chart has the identification numbers as labels on its horizontal axis. In order to fix this, we have to do some further editing to our chart, by changing this data series.

To change the data series:

- > In the **2019** sheet, at the end of column B, add the months of 2023. **1**
- > In the sheet that contains the generated forecast values and the chart, click inside the chart **2** and then click the **Chart Filters** icon. **3**
- > Click **Select Data**. **4**
- > In the **Select Data Source** window, on the Horizontal (Category) Axis Labels, click **Edit**. **5**
- > On the **2019** sheet, select the months. **6**
- > In the **Axis Labels** window, click **OK**. **7**
- > In the **Select Data Source** window, click **OK**. **8**
- > The months will appear as labels on the horizontal axis. **9**

A	B	C
	Month	Tourist visits
1	January 2019	1,653
2	February 2019	1,492
3	March 2019	1,599
4	April 2019	1,650
5	May 2019	2,079
6	June 2019	903
7	July 2019	750
8	August 2019	2,675
9	September 2019	635
10	October 2019	1,119
11	November 2019	1,234
12	December 2019	1,735
13	January 2023	
14	February 2023	
15	March 2023	
16	April 2023	
17	May 2023	
18	June 2023	
19	July 2023	
20	August 2023	
21	September 2023	
22	October 2023	
23	November 2023	
24	December 2023	



A	B	C	D	E	F
	Month	Tourist visits			
1	January 2019	1,653			
2	February 2019	1,492			
3	March 2019	1,599			
4	April 2019	1,650			
5	May 2019	2,079			
6	June 2019	903			
7	July 2019	750			
8	August 2019	2,675			
9	September 2019	635			
10	October 2019	1,119			
11	November 2019	1,234			
12	December 2019	1,735			
13	January 2023				
14	February 2023				
15	March 2023				
16	April 2023				
17	May 2023				
18	June 2023				
19	July 2023				
20	August 2023				
21	September 2023				
22	October 2023				
23	November 2023				
24	December 2023				

Axis Labels

Axis label range: = January 2019; ...

OK Cancel

Select Data Source

Chart data range:

The data range is too complex to be displayed. If a new range is selected, it will replace all of the series in the Series panel.

Switch Row/Column

Legend Entries (Series)

- Tourist visits
- Forecast (Tourist visits)
- Lower Confidence Bound
- Upper Confidence Bound

Horizontal (Category) Axis Labels

- 1
- 2
- 3
- 4
- 5

Hidden and Empty Cells

OK Cancel

Select Data Source

Chart data range:

The data range is too complex to be displayed. If a new range is selected, it will replace all of the series in the Series panel.

Switch Row/Column

Legend Entries (Series)

- Tourist visits
- Forecast (Tourist visits)
- Lower Confidence Bound
- Upper Confidence Bound

Horizontal (Category) Axis Labels

- January 2019
- February 2019
- March 2019
- April 2019
- May 2019

Hidden and Empty Cells

OK Cancel

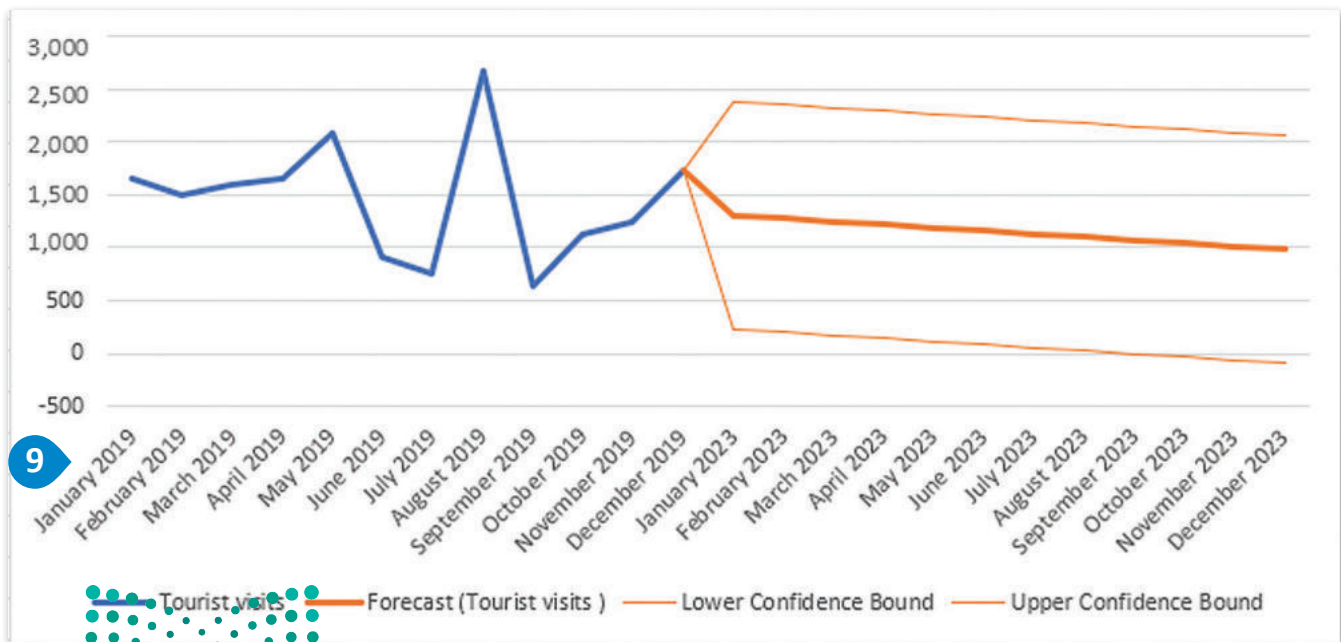
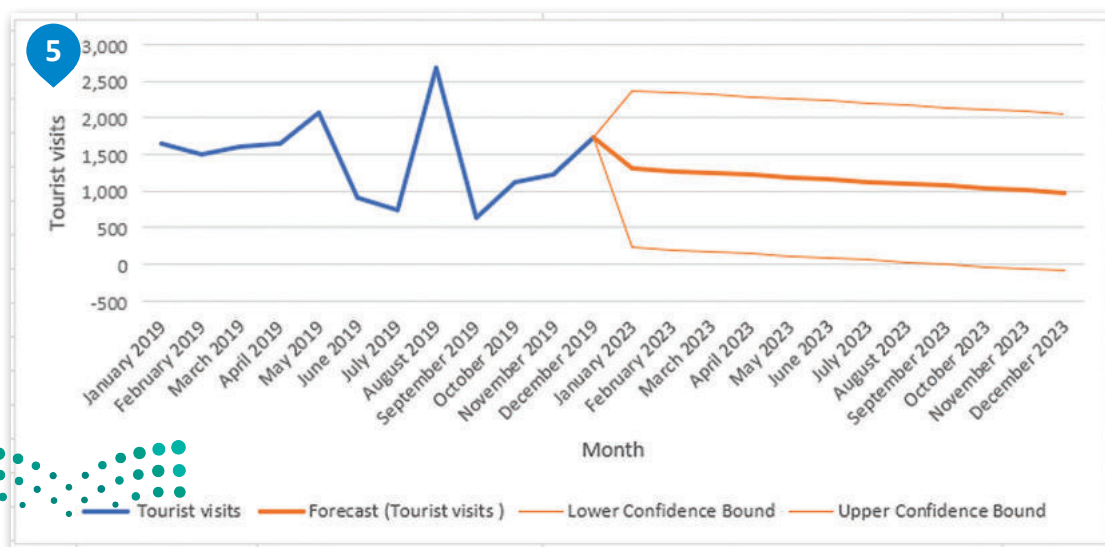
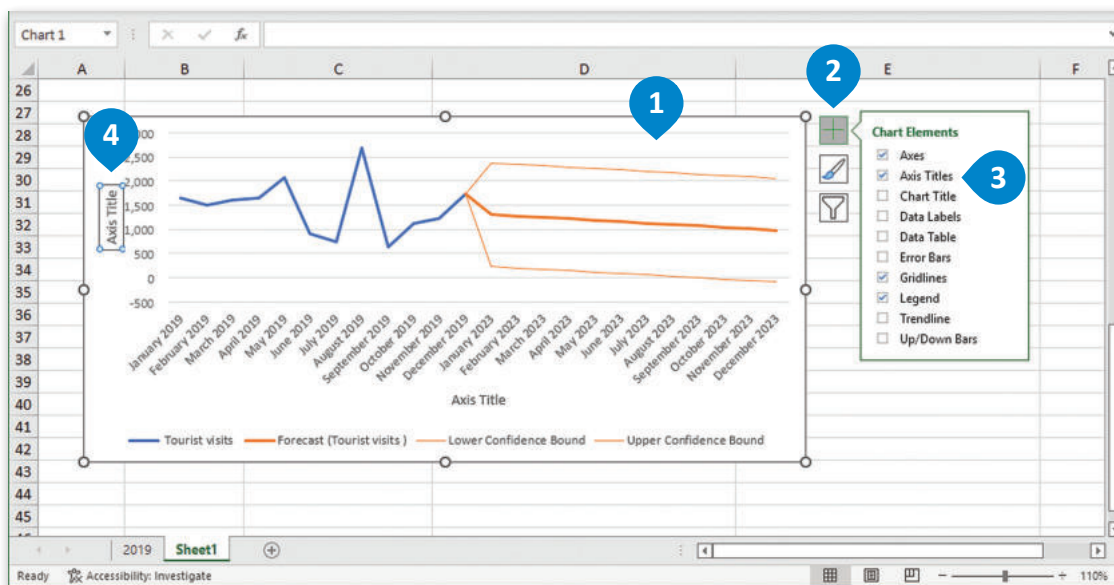


Figure 4.16: Change the data series

Further customizing the line chart, we can add x and y axis labels.

To change the label names:

- > Select the line chart, **1** and click the + button. **2**
- > Select the **Axis Titles** option. **3**
- > In the **label boxes** that appear, click each one and type the correct axis label names. **4**
- > The correct labels will appear in the line chart. **5**



Step 6: Analyze the Data

Sometimes, we want to see how different the predicted values are from the initial ones in order to better understand the phenomenon under study and come to the right conclusions. For example, in our case study we would like to know in which months of the year 2023 the tourist visits will increase and in which months not. Based on this information, we could organize, for example, certain advertising strategies in order to increase the tourist visits. In order to get this kind of information, we will subtract the predicted tourist visits data values from the past tourist visits data values, getting as results the forecast difference data values.

To prepare a new sheet:

- > In the new **Sheet 2**, create a column called **Months**. **1**
- > From the **2019** sheet, copy the 12 tourist visits values **2** and paste them in **Sheet 2**, in a column called **Tourist visits 2019**. **3**
- > From **Sheet 1**, select the 12 predicted tourist visits values **4** and paste them as values in **Sheet 2**, in a column called **Tourist visits 2023**. **5**

Data Analysis

A systematic examination of data through samples, measurement, and visualization.

Months	Tourist visits 2019	Tourist visits 2023	Forecast difference
January			
February			
March			
April			
May			
June			
July			
August			
September			
October			
November			
December			

	A	B	C
1		Month	Tourist visits
2	1	January 2019	1,653
3	2	February 2019	1,492
4	3	March 2019	1,599
5	4	April 2019	1,650
6	5	May 2019	2,079
7	6	June 2019	903
8	7	July 2019	750
9	8	August 2019	2,675
10	9	September 2019	635
11	10	October 2019	1,119
12	11	November 2019	1,234
13	12	December 2019	1,735
14	13	January 2023	
15	14	February 2023	
16	15	March 2023	
17	16	April 2023	
18	17	May 2023	
19	18	June 2023	
20	19	July 2023	

2

	B	C	
	Tourist visits 2019	Tourist visits 2023	Forecast
	1,653	1,305	
	1,492	1,276	
	1,599	1,246	
	1,650	1,217	
	2,079	1,188	
	903	1,159	
	750	1,130	
	2,675	1,101	
	635	1,071	
	1,119	1,042	
	1,234	1,013	
	1,735	984	

3

5

	A	B	C	D
1	Timeline	Tourist visits	Forecast (Tourist visits)	Lower Confidence Bound
2	1	1,653		
3	2	1,492		
4	3	1,599		
5	4	1,650		
6	5	2,079		
7	6	903		
8	7	750		
9	8	2,675		
10	9	635		
11	10	1,119		
12	11	1,234		
13	12	1,735	1,735	1,735
14	13		1,305	23
15	14		1,276	20
16	15		1,246	17
17	16		1,217	14
18	17		1,188	11
19	18		1,159	8
20	19		1,130	5
21	20		1,101	2
22	21		1,071	-1
23	22		1,042	-3
24	23		1,013	-6
25	24		984	-9

4

To calculate the forecast difference:

- > In **Sheet 2**, create a new column called **Forecast difference**. **1**
- > In the cell **D2** type the formula: **=C2-B2**. **2**
- > Copy the formula from **D2** to **D13** to generate the rest of the values. **3**

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F
1	Months	Tourist visits 2019	Tourist visits 2023	Forecast difference		
2	January	1,653	1,305	=C2-B2		
3	February	1,492	1,276			
4	March	1,599	1,246			
5	April	1,650	1,217			
6	May	2,079	1,188			
7	June	903	1,159			
8	July	750	1,130			
9	August	2,675	1,101			
10	September	635	1,071			
11	October	1,119	1,042			
12	November	1,234	1,013			
13	December	1,735	984			
14						
15						
16						

The screenshot shows the completed Excel spreadsheet with the following data:

C	D	E	F	C
Tourist visits 2023	Forecast difference			
1,305	-349			
1,276	-216			
1,246	-353			
1,217	-433			
1,188	-891			
1,159	256			
1,130	380			
1,101	-1,574			
1,071	436			
1,042	-77			
1,013	-221			
984	-751			



Figure 4.19: Calculate the forecast difference

Creating a Clustered Column Chart

Apart from the forecast chart that we've already created, now we are ready to start creating two more charts that we are going to use for our analysis. More specifically, we will create:

- > a chart that visualizes the comparison between the past tourist visits data and the forecast tourist visits data values.
- > a chart that visualizes the forecast difference between past tourist visits data and forecast tourist visits data values.

To create a clustered column chart:

- > Select columns **A, B** and **C**. **1**
- > In the **Insert** tab, click **Recommended Charts**. **2**
- > Choose the **Clustered Column** chart. **3**
- > Click **OK**. **4**

The screenshot shows an Excel spreadsheet with the following data:

Months	Tourist visits 2019	Tourist visits 2023	Forecast difference
January	1,653	1,305	-349
February	1,492	1,276	-216
March	1,599	1,246	-353
April	1,650	1,217	-433
May	2,079	1,188	-891
June	903	1,159	256
July	750	1,130	380
August	2,675	1,101	-1,574
September	635	1,071	436
October	1,119	1,042	-77
November	1,234	1,013	-221
December	1,735	984	-751

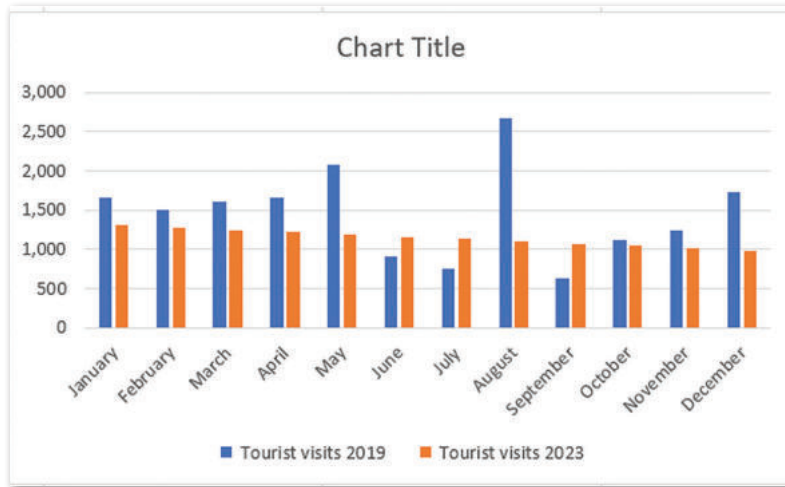
The 'Insert > Recommended Charts' dialog box is open, showing the 'Clustered Column' chart type selected. The dialog box includes a preview of the chart and a description: "A clustered column chart is used to compare values across a few categories. Use it when the order of categories is not important."



وزارة التعليم

Ministry of Education
Figure 4.20: Create a clustered column chart
2023 - 1445

The chart that visualizes the comparison of the past tourist visits data and the forecast tourist visits data values is the following. For a more comprehensive visualization of the information, we can change the bounds or the units of the vertical axis.



Compared with the first chart, this chart has smaller major units equal to 250 (instead of 500 in the initial chart).



Figure 4.21: Clustered column chart

When it comes to visualizing information in Excel, it is very important to choose the right charts, so the audience can easily read and understand them. For this purpose, we usually choose charts that Excel recommends us as suitable for our type of data. If we choose a chart that doesn't belong to the recommended ones, it is very likely that the audience of our presentation won't understand the chart. An example of an unsuitable chart for our case study is the chart in Figure 4.22.

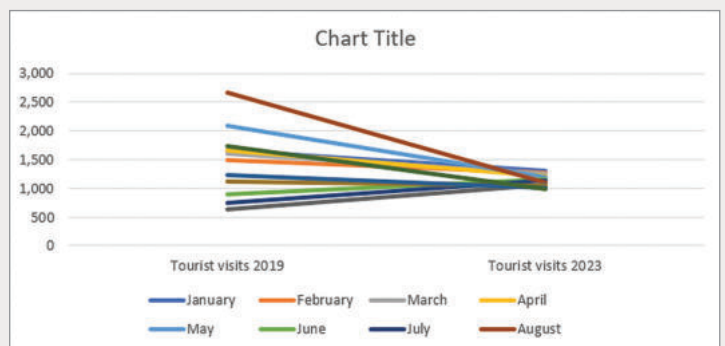
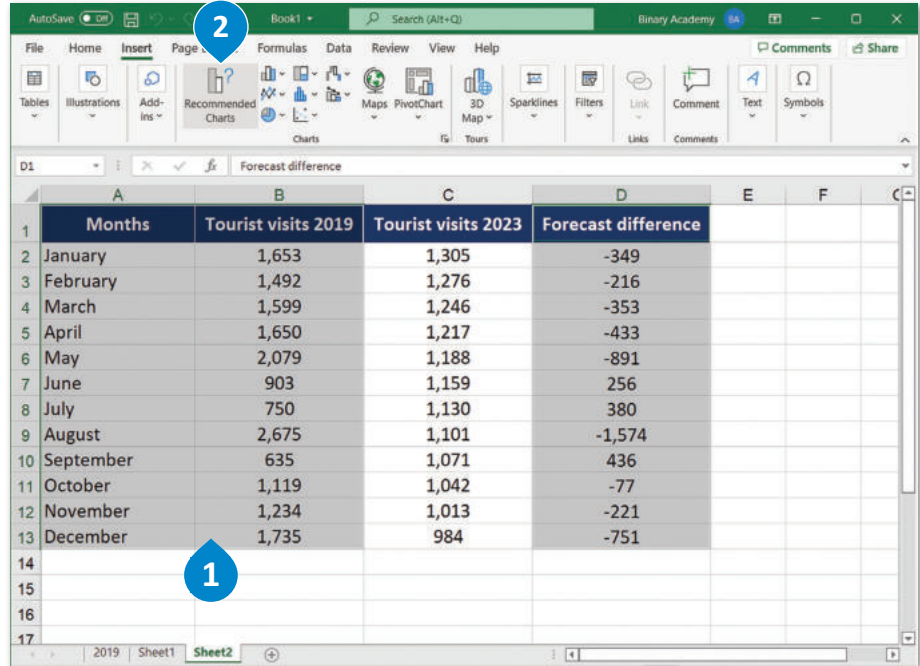


Figure 4.22: Confusing line chart

Stacked Column Chart

To create a stacked column chart:

- > Select columns A, B and D. **1**
- > On the Insert tab, click **Recommended Charts**. **2**
- > Choose the **Stacked Column** chart. **3**
- > Click **OK**. **4**



A Stacked Column chart is used to compare parts of a whole. We use it to show how segments of a whole change over time.

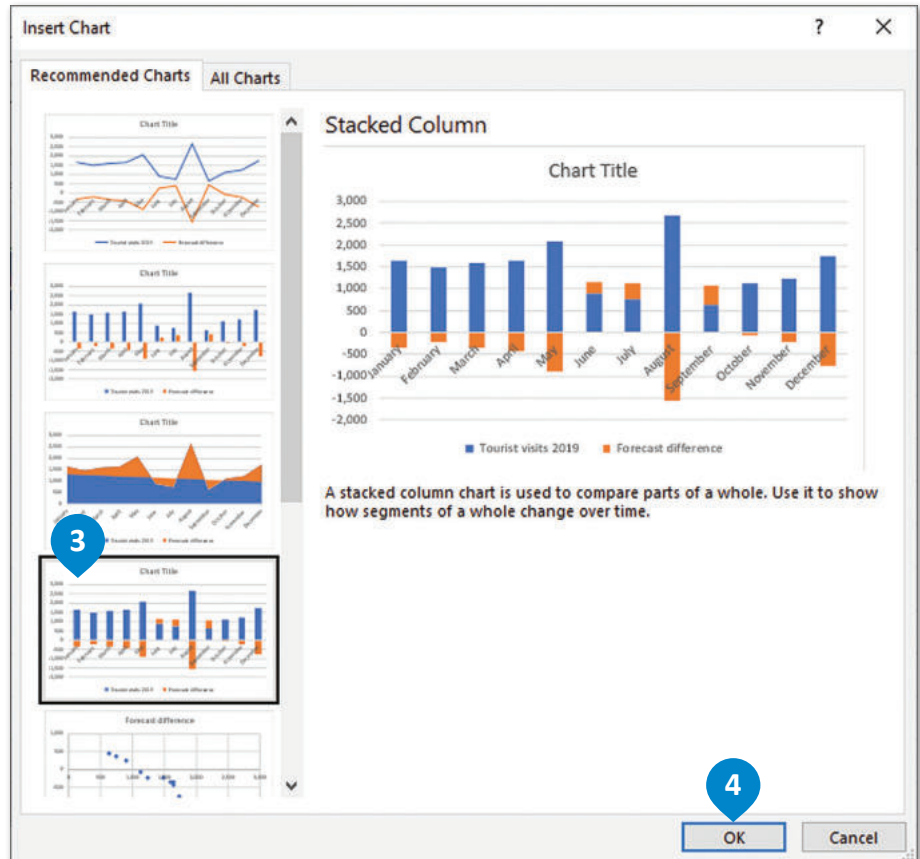


Figure 4.23: Create a stacked column chart



The chart that visualizes the forecast difference between past tourist visits data and forecast tourist visits data values is the following (Figure 4.24). For a more comprehensive visualization of the information, we can change the bounds or the units of the vertical axis.

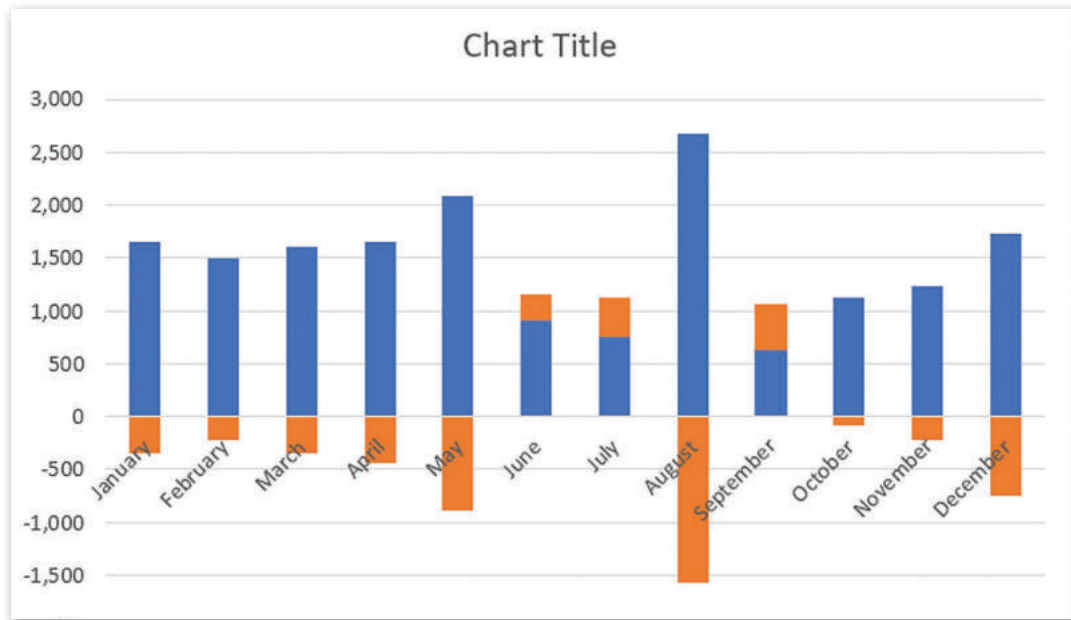


Figure 4.24: Stacked Column chart

It is not always the case that the recommended charts are suitable for the visualization of our information. For example, in our case Excel recommends us to use a Funnel chart (Figure 4.25) or a Scatter chart (Figure 4.26), but if we look carefully at them, we will realize that they won't be easily read or understood.

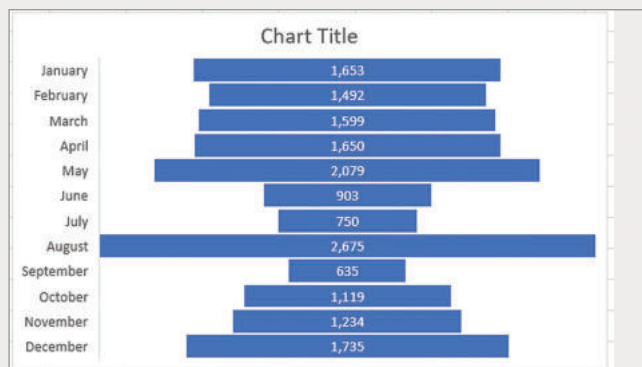


Figure 4.25 Funnel chart

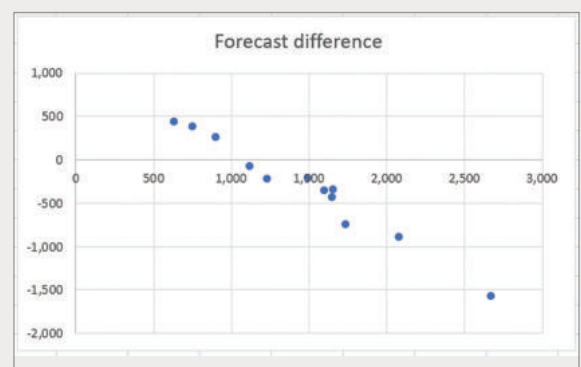


Figure 4.26 Scatter chart

Also, we must take into consideration that not all the charts are appropriate for all kinds of audience. Some types of charts (and even different software tools), we would choose to present our information to a scientist or marketing manager, and other types of charts to the owner of the company.

Exercises

1

Read the sentences and tick ✓ True or False.	True	False
1. Forecasting is the only method for predicting future data based on past data.	<input type="radio"/>	<input type="radio"/>
2. The key to making a good forecast is to define the steps clearly before executing the forecasting procedure.	<input type="radio"/>	<input type="radio"/>
3. It is not essential to follow the six forecast steps in a specific order.	<input type="radio"/>	<input type="radio"/>
4. The confidence interval gives us information about the uncertainty of the prediction.	<input type="radio"/>	<input type="radio"/>
5. A prediction with a confidence interval equal to 95% is more accurate than one with a confidence interval of 75%.	<input type="radio"/>	<input type="radio"/>
6. Forecast and Prediction refer to the same procedure.	<input type="radio"/>	<input type="radio"/>
7. A line chart is always preferable to a column chart for the visualization of a forecast.	<input type="radio"/>	<input type="radio"/>
8. It is preferable to choose the charts that Excel recommends as suitable for our type of data.	<input type="radio"/>	<input type="radio"/>
9. The choice of the right chart for the visualization of data depends on who the information is for.	<input type="radio"/>	<input type="radio"/>
10. Lower and Upper confidence bounds define the range of accepted values.	<input type="radio"/>	<input type="radio"/>

Lesson 3

Optimization

Link to digital lesson



www.ien.edu.sa

Dealing with Optimization Problems

Optimization problems are real-world problems that arise in many areas such as mathematics, engineering, science, business and economics. In these problems, we are trying to find a way (the optimal or the most efficient way) of using limited resources to achieve an objective in a given situation. Our objective might be maximizing profit, minimizing cost, minimizing the total distance traveled or minimizing the total time to complete a project. In other cases, it might be maximizing tourist visits in a country, choosing the optimal budget for an advertising campaign, designing the best work schedule for employees, minimizing delivery costs, and so on.

Optimization

The process of choosing the best element from a set of available alternatives under some constraints.

What is Excel Solver?

Microsoft Excel Solver is an add-in program, used for the simulation and optimization of various business and engineering models. It belongs to a special set of Excel commands often referred to as "What-If Analysis tools", and is used for finding the best solutions for a model that consists of multiple inputs.

The most common use of Excel Solver is to determine a value for one cell (called the objective cell) by changing the values in certain other cells (called variable cells), with or without the use of constraints. It is useful for solving linear programming problems (also known as linear optimization problems) and therefore is sometimes called the Linear Programming Solver.

Excel Solver is a powerful tool for dealing with optimization problems, because we can use spreadsheets to insert the decision variables and the constraints of a model, and execute the objective function that describes it.

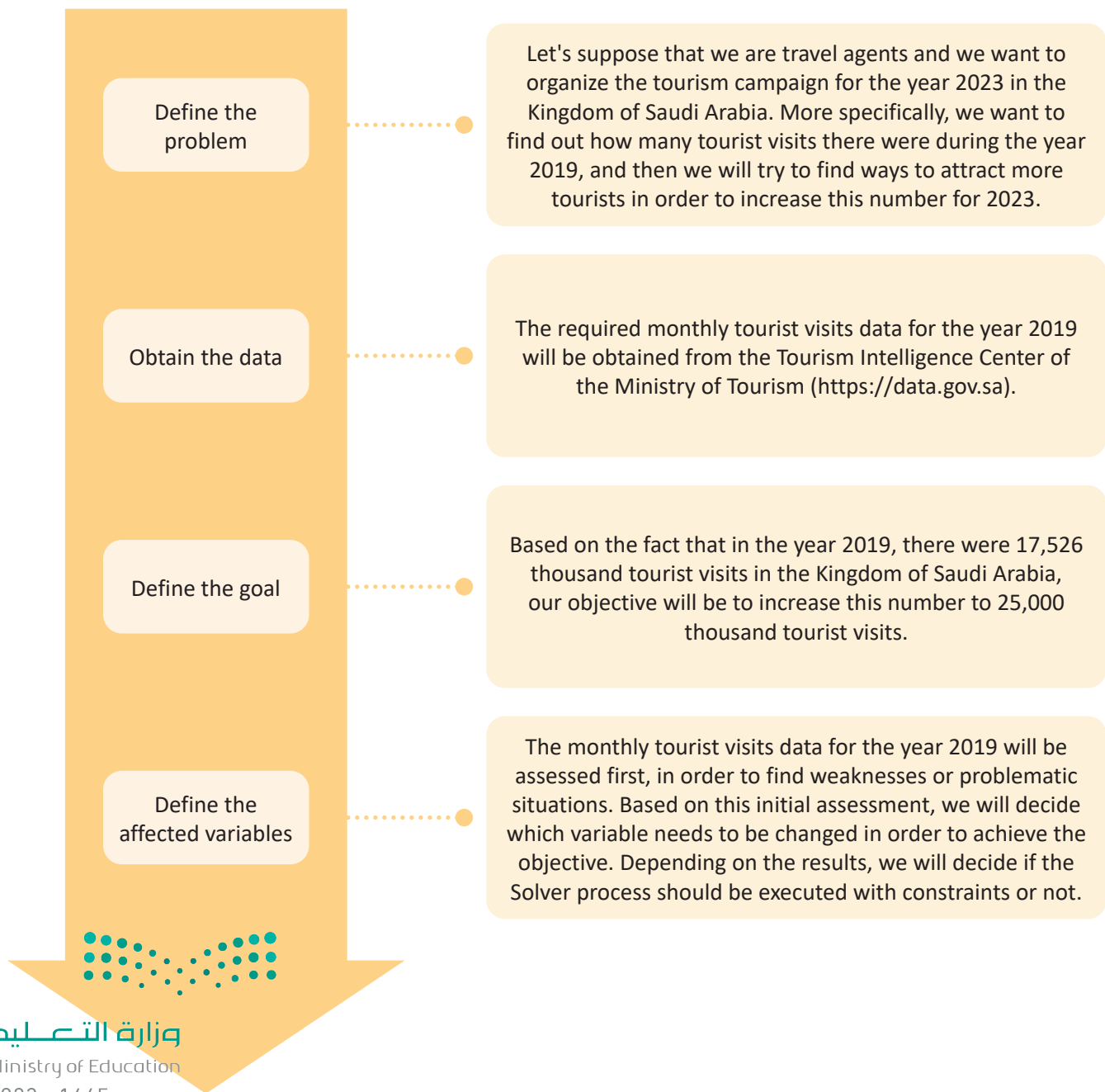
When a model has two decision variables, graphical methods can be used to solve the model. Very few real world problems involve only two variables. For problems with more than two variables, we need to use complex techniques and calculations to find the optimal solution. The spreadsheet and Solver approach makes solving optimization problems a simple task and it is convenient for all users, regardless of their mathematical background.

It is crucial to understand that changes may be made to the Solver parameters or to the performed method at any time. Depending on the results that the Solver tool gives us, we can reassess the problem and decide if we will need to execute a Solver process with constraints. Solver results are not just numbers, they are values with a specific meaning for the situation under study, so the data scientist or business analyst must critically evaluate these results and take further actions if the results are not meaningful or satisfactory.

Formulating the Problem

Before running the Microsoft Excel Solver add-in, we must formulate the problem (model) in a worksheet. This model represents the problem we want to solve.

In the previous lesson, we obtained monthly tourist visits data for the year 2019 from the Tourism Intelligence Center of the Ministry of tourism (<https://data.gov.sa>). In this lesson, we will use the same data in order to obtain specific results for the organization of the Tourism campaign for the year 2023 in the Kingdom of Saudi Arabia, using Excel Solver. To achieve this, we first have to formulate the problem and then specify the type of information that we want to obtain from the Excel Solver:

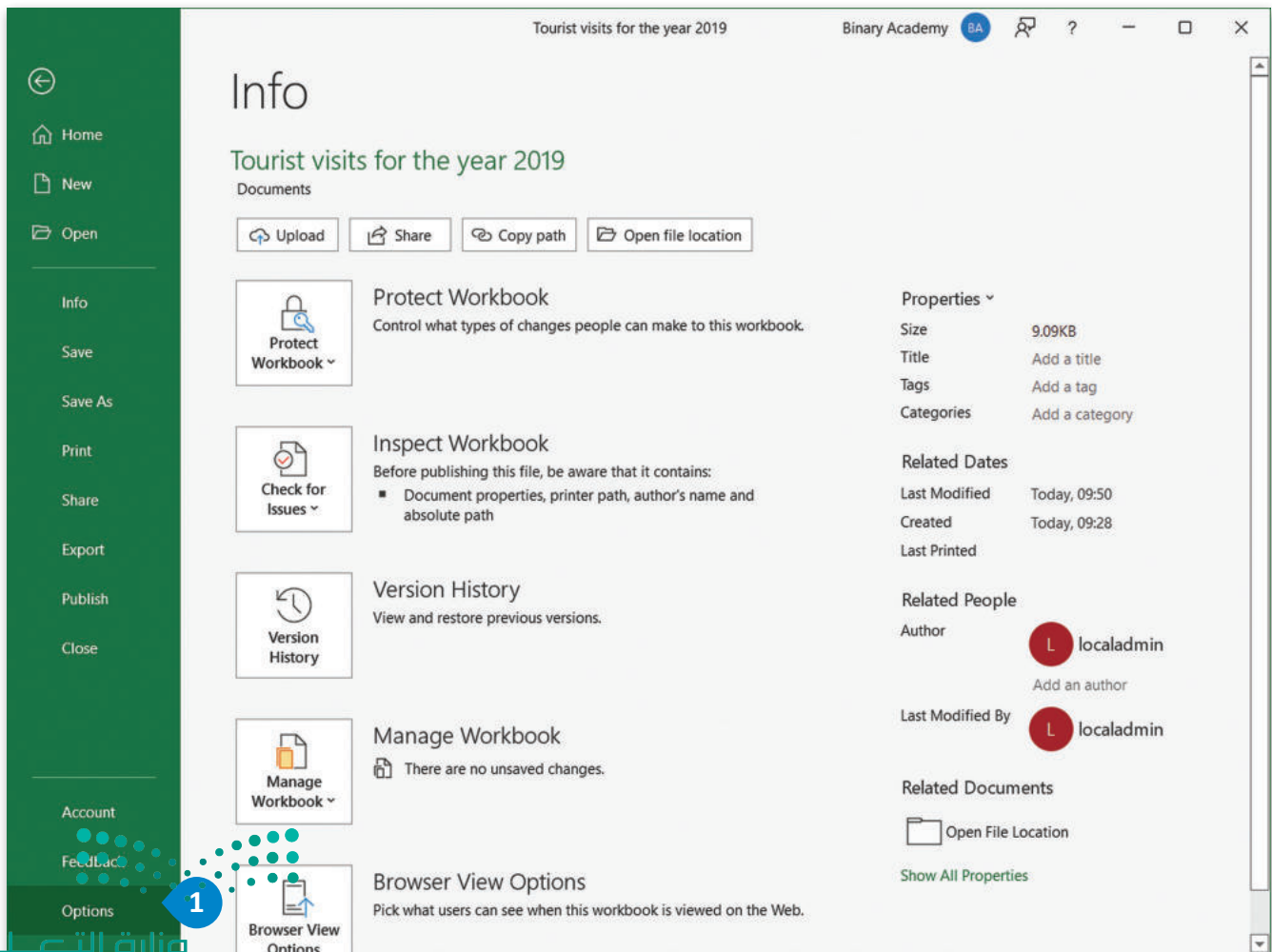


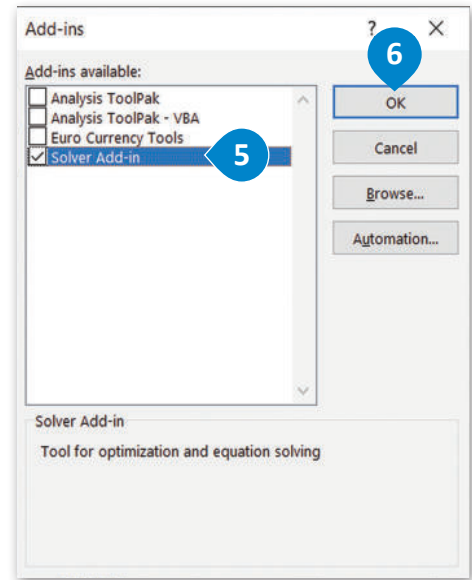
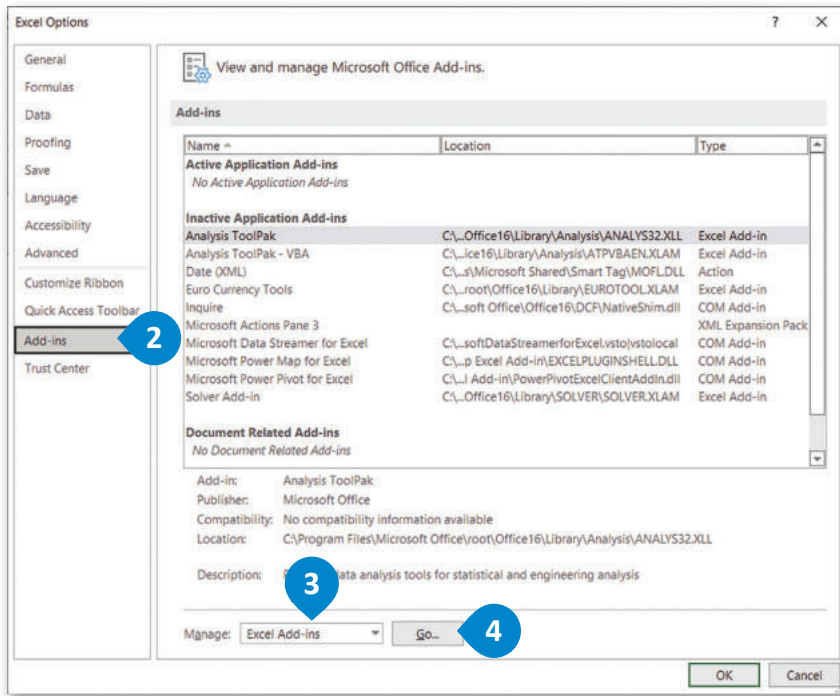
The Excel Solver Add-In

Now that the problem is well formulated, we can open Excel and perform Solver. The first thing that we must do is to activate the tool. The Solver Add-In is not automatically activated on installation of Microsoft Office, so we will activate it from the Options window.

To activate Solver Add-In:

- > In the **File** tab, click **Options**. ①
- > In the **Excel Options** window, click **Add-ins**. ②
- > In the **Manage** box, select **Excel Add-ins** ③ and then click **Go**. ④
- > In the **Excel Add-Ins** window, check **Solver Add-in**. ⑤
- > Click **OK**. ⑥
- > The Solver button will appear. ⑦





AutoSave Off Tourist visits for the year 2019 - Saved Binary Academy

	A	B	C	D	E	F	G	H	I
1		Month	Tourist visits						
2	1	January 2019	1,653						
3	2	February 2019	1,492						
4	3	March 2019	1,599						
5	4	April 2019	1,650						
6	5	May 2019	2,079						
7	6	June 2019	903						
8	7	July 2019	750						
9	8	August 2019	2,675						
10	9	September 2019	635						
11	10	October 2019	1,119						
12	11	November 2019	1,234						
13	12	December 2019	1,735						
14									
15									
16									
17									

Using Solver

First of all, we must calculate the total amount of tourist visits of 2019 estimated in thousands. To achieve this, we will use the SUM function in Excel, by selecting the cells that contain all the monthly tourist visits values.

To calculate Total SUM:

- > Open the "Tourist visits for the year 2019" file in Excel. 1
- > In cell B14 type "Total". 2
- > In cell C14 type =SUM(C2:C13) 3 and press Enter.
- > The Total SUM will appear in the cell. 4

The figure consists of two side-by-side screenshots of an Excel spreadsheet titled "Tourist visits". The spreadsheet has three columns: A, B, and C. Column A contains row numbers 1 through 12. Column B contains the months from January 2019 to December 2019. Column C contains the number of tourist visits for each month. The data is as follows:

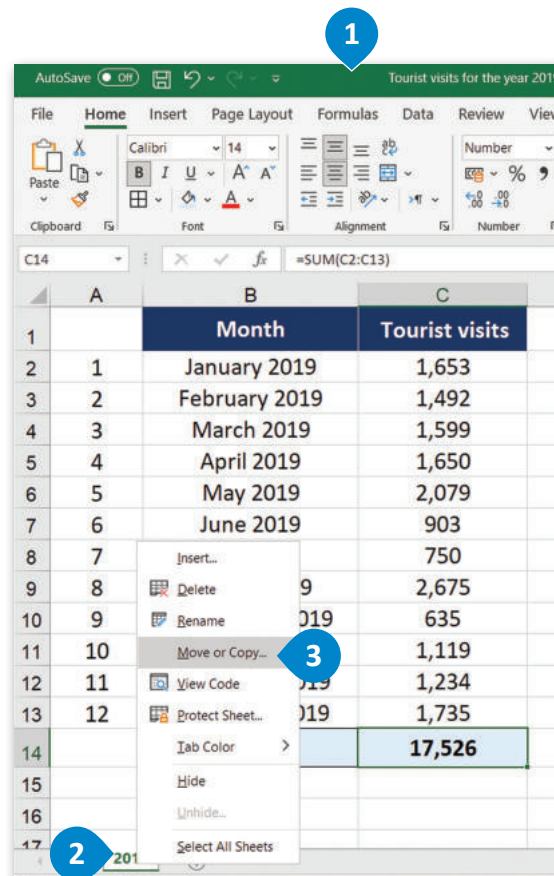
	Month	Tourist visits
1	January 2019	1,653
2	February 2019	1,492
3	March 2019	1,599
4	April 2019	1,650
5	May 2019	2,079
6	June 2019	903
7	July 2019	750
8	August 2019	2,675
9	September 2019	635
10	October 2019	1,119
11	November 2019	1,234
12	December 2019	1,735
	Total	17,526

The first screenshot shows the formula bar containing the formula =SUM(C2:C13) and the formula being entered in cell C14. The second screenshot shows the final result, 17,526, in cell C14.

Then, we must open the Excel file that contains the tourist visits data for the year 2019 in a sheet called "2019". We will also create a new sheet called "Solver" with the same data as sheet "2019". We do this because when Excel executes the Solver function, it permanently changes the values in our data, without providing the "undo" option. It is therefore necessary to preserve a sheet with the original data (sheet "2019" in our case) because if the Solver results are not satisfactory, we will not be able to get the original values back. The easiest way to copy the values of sheet "2019" is to make a copy of the sheet itself.

To copy a sheet in Excel:

- > Open the "Tourist Visits for the year 2019" file in Excel. 1
- > Right-click the 2019 sheet 2 and the select **Move or Copy**. 3
- > In the Move or Copy window select **2019** 4 and choose the **Create a copy** option. 5
- > Click **OK**. 6
- > The new sheet has been created. 7



	A	B	C
1		Month	Tourist visits
2	1	January 2019	1,653
3	2	February 2019	1,492
4	3	March 2019	1,599
5	4	April 2019	1,650
6	5	May 2019	2,079
7	6	June 2019	903
8	7	July 2019	750
9	8	August 2019	2,675
10	9	September 2019	635
11	10	October 2019	1,119
12	11	November 2019	1,234
13	12	December 2019	1,735
14		Total	17,526

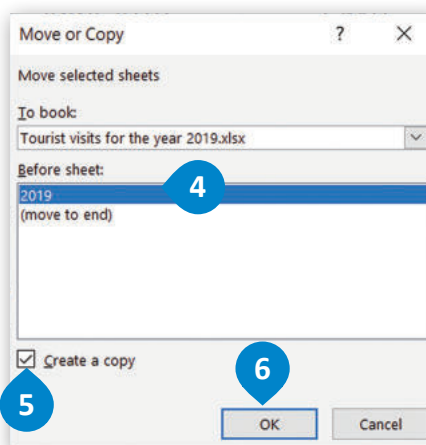


Figure 4.29: Copy a Sheet in Excel

Find the Problematic Cell Values

As we mentioned above, the monthly tourist visits data for the year 2019 will be assessed in order to find problematic values (if there are any) and then to decide which variables will need to be changed in order to achieve the objective. If we take a closer look at the "2019" Excel sheet, we observe that the monthly tourist visits values for June, July and September are significantly lower than those for the other months of the year. This observation allows us to suggest that, in order to achieve the goal of total tourist visits in 2023, it is not necessary to improve the values for the months with high numbers of tourist visits. To achieve our objective, we will only try to increase the number of tourist visits in June, July and September.

In the Solver function parameters, the Objective cell will therefore be the total tourist visits and the Variable cells will be the tourist visits in the months of June, July and September. Specific constraints will not be added. The value of the Objective cell will be set to 25,000 thousand.

To use Solver with no constraints:

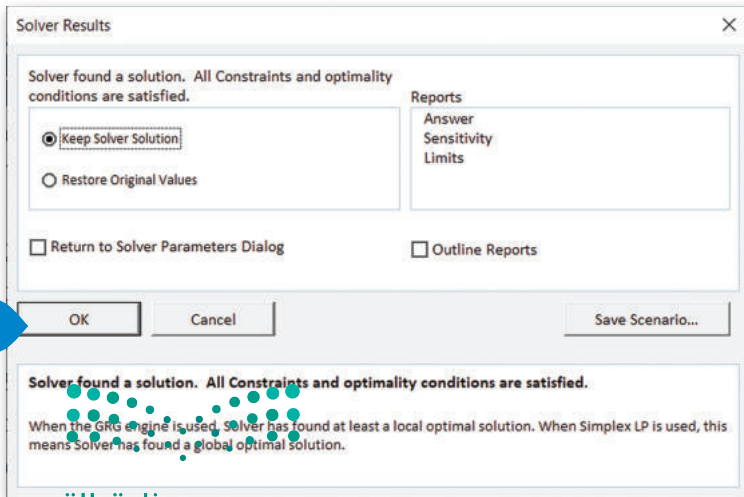
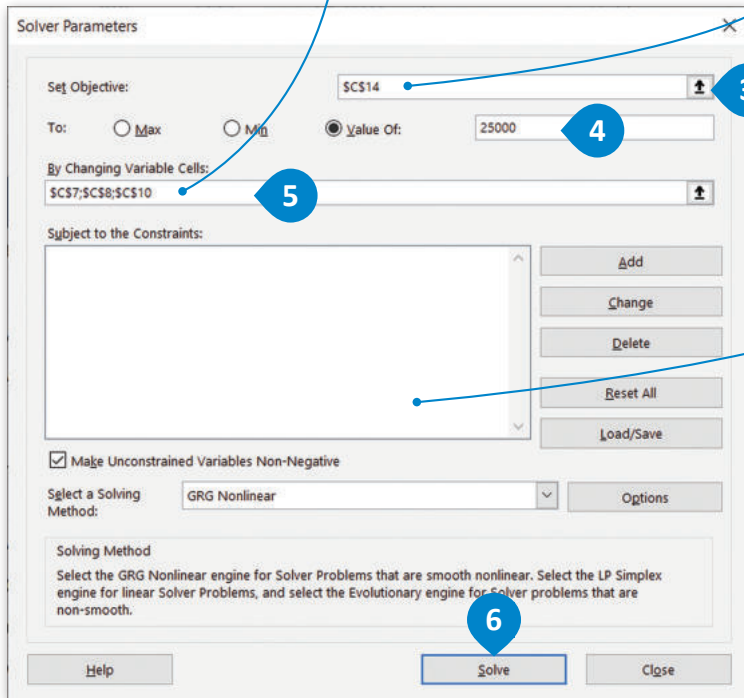
- > In the **Data** tab, **1** click the **Solver** button. **2**
- > In the **Set Objective** field choose cell **C14**. **3**
- > Select **Value of: 25000**. **4**
- > In the **By Changing Variable Cells** field, choose the cells **C7;C8;C10**. **5**
- > Click **Solve**. **6**
- > In the **Solver Results** window, click **OK**. **7**
- > Changes will appear in the selected cells. **8**

	A	B	C	D	E	F	G	H	I
1		Month	Tourist visits						
2	1	January 2019	1,653						
3	2	February 2019	1,492						
4	3	March 2019	1,599						
5	4	April 2019	1,650						
6	5	May 2019	2,079						
7	6	June 2019	903						
8	7	July 2019	750						
9	8	August 2019	2,675						
10	9	September 2019	635						
11	10	October 2019	1,119						
12	11	November 2019	1,234						
13	12	December 2019	1,735						
14		Total	17,526						

Variable cells are the cells in your worksheet whose values will change. These are the decision variables that will be adjusted until the optimum solution is found.

The objective cell is the target cell in your worksheet whose value is to be maximized, minimized, or made to reach a particular value. This is the cell that contains the objective function (the formula).

In this list, constraints can be added. These are limits that we impose on changes to the values in certain cells.



Month	Tourist visits
January 2019	1,653
February 2019	1,492
March 2019	1,599
April 2019	1,650
May 2019	2,079
June 2019	4,327
July 2019	3,109
August 2019	2,675
September 2019	2,327
October 2019	1,119
November 2019	1,234
December 2019	1,735
Total	25,000

Figure 4.30: Use Solver with no constraints

Assess the Results

Now that we've executed the Solver function, we are ready to take a look at its results. As mentioned at the beginning of the lesson, Solver results are not just numbers, they are values with a specific meaning for the given situation under study. As a data scientist, we must assess these results to decide if further actions are needed. So, first of all, in the Solver Sheet we will create a five column table (Identification number, Month, 2019 monthly tourist visits, Solver results 2023 and Difference) in order to easily compare the before and after for the Solver process.

To calculate the difference:

- > In the **2019** sheet, copy value cells **C1:C14**. **1**
- > In the **Solver** sheet, select column **D**, **2** and right click it.
- > Paste the values. **3**
- > Change the column titles and delete "2019" from all the months of column **B**. **4**
- > Add a column called "Difference". **5**
- > In cell **E2**, type **=C2-D2**. **6**
- > Perform the function on all the cells from **E2** to **E14**, **7** and press **Enter**.
- > The **Solver** sheet is now ready to assess the results. **8**

Month	Tourist visits
January 2019	1,653
February 2019	1,492
March 2019	1,599
April 2019	1,650
May 2019	2,079
June 2019	903
July 2019	750
August 2019	2,675
September 2019	635
October 2019	1,119
November 2019	1,234
December 2019	1,735
Total	17,526

Month	Tourist visits
January 2019	1,653
February 2019	1,492
March 2019	1,599
April 2019	1,650
May 2019	2,079
June 2019	4,327
July 2019	3,109
August 2019	2,675
September 2019	2,327
October 2019	1,119
November 2019	1,234
December 2019	1,735
Total	25,000

AutoSave OFF | Tourist visits for the year 2019 | Binary Academy

File Home Insert Page Layout Formulas **Data** Review View Help

Get Data Refresh All Stocks (En... Geography... Sort Filter Clear Reapply Text to Columns What-If Analysis Forecast Sheet Outline Solver

Get & Transform Data Queries & Conne... Data Types Sort & Filter Data Tools Forecast Analyze

A1

	A	B	C	D	E	F	G	H
1		Month	Solver results for 2023	Monthly tourist visits for 2019	Difference			
2	1	January	1,653	1,653				
3	2	February	1,492	1,492				
4	3	March	1,599	1,599				
5	4	April	1,650	1,650				
6	5	May	2,079	2,079				
7	6	June	4,327	903				
8	7	July	3,109	750				
9	8	August	2,675	2,675				
10	9	September	2,327	635				
11	10	October	1,119	1,119				
12	11	November	1,234	1,234				
13	12	December	1,735	1,735				
14		Total	25,000	17,526				

2019 Solver

Home Insert Page Layout Formulas **Data** Review View Help

Transform Data Queries & Conne... Data Types Sort & Filter Data Tools Forecast Analyze

=C2-D2

	A	B	C	D	E	F	G	H
		Month	Solver results for 2023	Monthly tourist visits for 2019	Difference			
1		January	1,653	1,653	=C2-D2			
2		February	1,492	1,492				
3		March	1,599	1,599				
4		April	1,650	1,650				
5		May	2,079	2,079				
6		June	4,327	903				
7		July	3,109	750				
8		August	2,675	2,675				
9		September	2,327	635				
10		October	1,119	1,119				
11		November	1,234	1,234				
12		December	1,735	1,735				
		Total	25,000	17,526				

2019 Solver

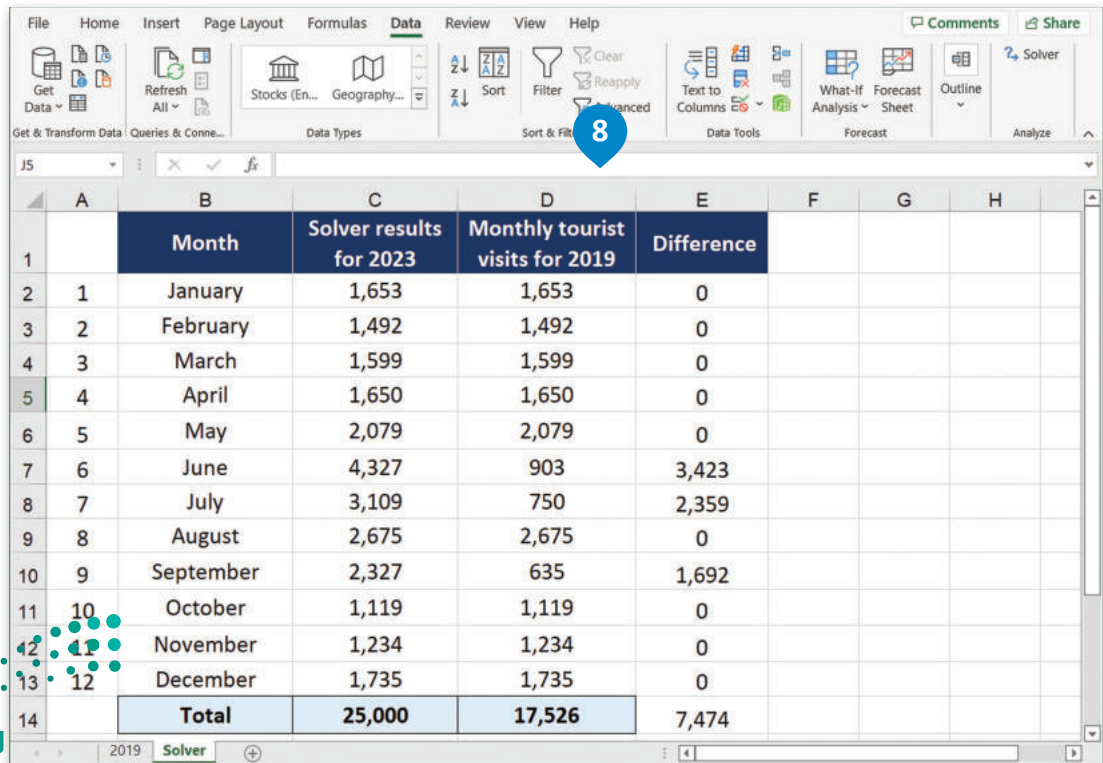
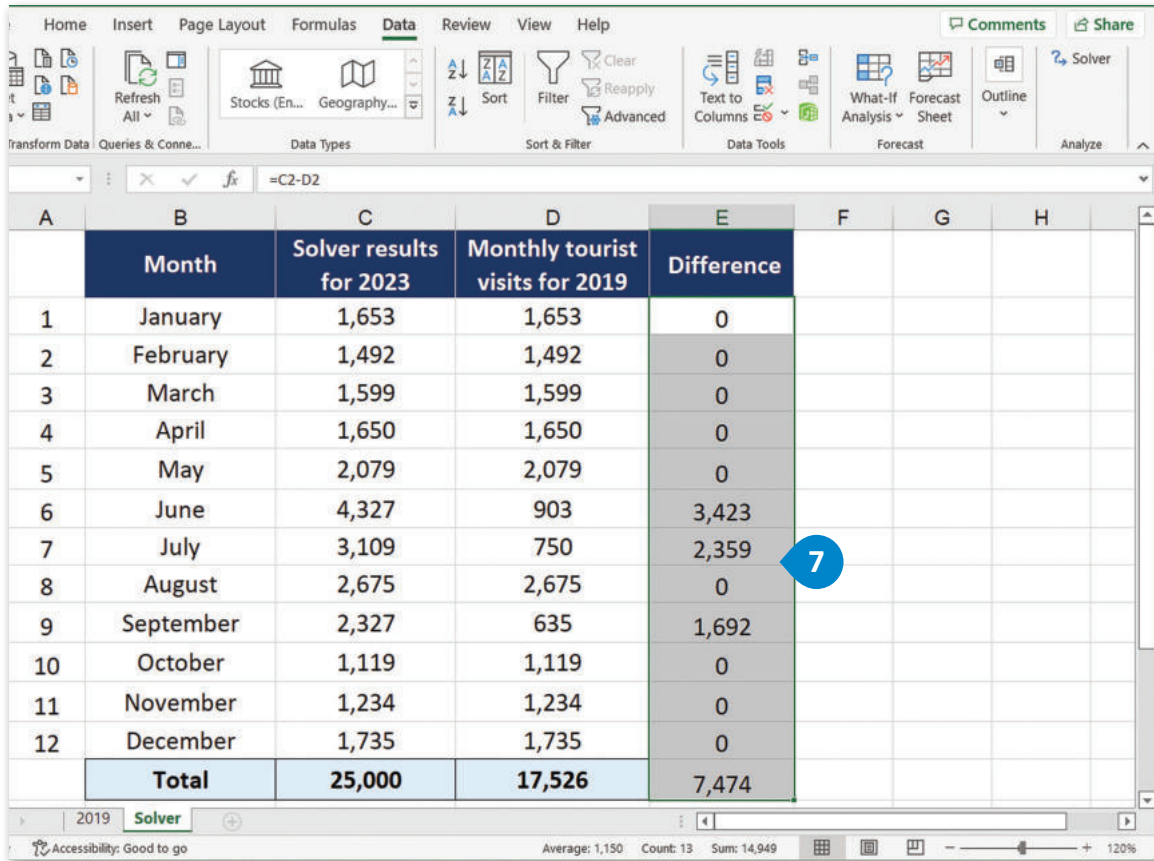


Figure 4.31: Calculate the difference

Now we are ready to take a look to our sheet and assess the results. Taking a closer look at the Solver values for June, July and September, we observe that they are are now extremely high.

The Excel Solver gave us a proposal that says: if we want to reach the goal of 25,000 thousand tourist visits in the year 2023, we have to design our Tourism Campaign in such a way that the number of tourist visits will reach 4,327 thousand in June, 3,109 thousand in July and 2,327 thousand in September. Based on the values of other months, this is an unrealistic objective. When we consider that the highest value in the 2019 data is 2,675 thousand tourist visits, it is clear that whatever tourism campaign we design, we are unlikely reach the target of 4,327 thousand trips in June (160% higher than the number of visits for the best month in 2019) (Figure 4.32).

	Month	Solver results for 2023	Monthly tourist visits for 2019	Difference
1	January	1,653	1,653	0
2	February	1,492	1,492	0
3	March	1,599	1,599	0
4	April	1,650	1,650	0
5	May	2,079	2,079	0
6	June	4,327	903	3,423
7	July	3,109	750	2,359
8	August	2,675	2,675	0
9	September	2,327	635	1,692
10	October	1,119	1,119	0
11	November	1,234	1,234	0
12	December	1,735	1,735	0
	Total	25,000	17,526	7,474

Figure 4.32: Tourist visits values for 2019 and data after Solver

Tourist trip values for June, July and September, generated by the Solver function. These are extremely high values, compared with the values for all the other months.

Tourist visits values for June, July and September 2019.

So, given the fact that our initial Solver results are not completely satisfactory, we will now execute Solver with constraints in order to obtain more realistic objectives. For example, we can set Solver to change all monthly values and set a constraint for the values for June, July and September, in order not to get unrealistic results. This could be achieved by using the average of our data values. More specifically, we will allow Solver to increase the values for all months, but include constraints that specify that the values for June, July and September must be above the average value for all months in 2019.



So, to continue working in the same Excel file, we will create another copy of the sheet "2019", called "Solver (constraints)" and we will run the Solver function again, this time with constraints.

Calculate the Average

The average (also known as the arithmetic mean) is calculated by adding a group of numbers together and then dividing by the count of those numbers. In our case, we will add all the monthly tourist visits values and we will divide this sum by 12.

Arithmetic mean

In descriptive statistics, the average value is calculated by adding the scores together and then dividing the total by the number of scores.

To calculate the average:

- > In the Excel file, create a new sheet and give it the name **Solver (constraints)**. **1**
- > In cell **B15**, type **average**. **2**
- > In cell **C15**, type the average formula **=average(C2:C13)**. **3**
- > Press **Enter** and the average will appear in cell **C15**. **4**

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D
1		Month	Tourist visits	
2	1	January 2019	1,653	
3	2	February 2019	1,492	
4	3	March 2019	1,599	
5	4	April 2019	1,650	
6	5	May 2019	2,079	
7	6	June 2019	903	
8	7	July 2019	750	
9	8	August 2019	2,675	
10	9	September 2019	635	
11	10	October 2019	1,119	
12	11	November 2019	1,234	
13	12	December 2019	1,735	
14		Total	17,526	
15		Average	=average(C2:C13)	

The formula bar shows the formula `=average(C2:C13)` being entered into cell C15. Blue callouts 1, 2, and 3 point to the sheet name, the word 'Average', and the formula respectively.

The screenshot shows the same Excel spreadsheet as the previous one, but with the average calculated. The formula bar now shows `=AVERAGE(C2:C13)`.

	B	C	D
	Month	Tourist visits	
	January 2019	1,653	
	February 2019	1,492	
	March 2019	1,599	
	April 2019	1,650	
	May 2019	2,079	
	June 2019	903	
	July 2019	750	
	August 2019	2,675	
	September 2019	635	
	October 2019	1,119	
	November 2019	1,234	
	December 2019	1,735	
	Total	17,526	
	Average	1,460	

Blue callout 4 points to the calculated average value of 1,460 in cell C15.

Figure 4-33: Calculate the average

Solver with Constraints

Now that we've calculated the average for the tourist visits values for the year 2019, we will execute Solver with constraints. This time, in the Solver function parameters, the objective cell will be the total tourist visits and the variable cells will be the tourist visits for all months. Specific constraints will be added, setting the estimated values for Solver function for June, July and September to be greater than or equal to 1,460 thousand the average for all months (average). The value of the objective cell will again be set to 25,000 thousand.

Constraint

Specification of what may be contained in a data or metadata set in terms of the content or, for data only, in terms of the set of key combinations to which specific attributes (defined by the data structure) may be attached.

To use Solver with constraints:

- > In the **Data** tab, **1** click the **Solver** button. **2**
- > In the **Set Objective** field choose cell **C14**. **3**
- > Select **Value of: 25000**. **4**
- > In the **By Changing Variable Cells** field, choose the cells **C2:C13**. **5**
- > Click **Add** to add a constraint. **6**

The screenshot shows the Excel interface with the Solver Parameters dialog box open. The spreadsheet data is as follows:

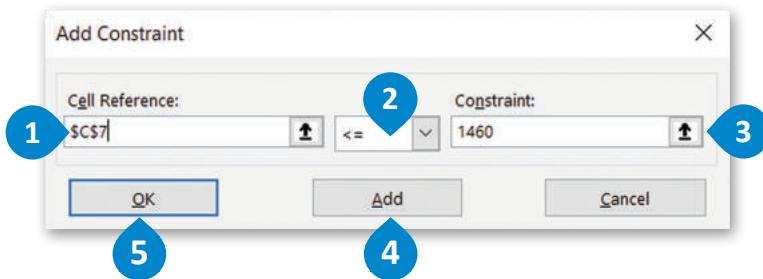
	A	B	C	D	E	F	G	H	I
1		Month	Tourist visits						
2	1	January 2019	1,653						
3	2	February 2019	1,492						
4	3	March 2019	1,599						
5	4	April 2019	1,650						
6	5	May 2019	2,079						
7	6	June 2019	903						
8	7	July 2019	750						
9	8	August 2019	2,675						
10	9	September 2019	635						
11	10	October 2019	1,119						
12	11	November 2019	1,234						
13	12	December 2019	1,735						
14		Total	17,526						
15		Average	1,460						

The Solver Parameters dialog box is configured as follows:

- Set Objective:** \$C\$14 (Cell C14)
- To:** Max Min Value Of: 25000
- By Changing Variable Cells:** \$C\$2:\$C\$13 (Cells C2:C13)
- Subject to the Constraints:** (Empty list, with 'Add' button highlighted)
- Make Unconstrained Variables Non-Negative
- Select a Solving Method:** GRG Nonlinear
- Solving Method:** Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

To set the constraints:

- > In the **Cell Reference** box, select the cell **C7**. 1
- > Choose the symbol **<=**. 2
- > Write **1460** in the **Constraint** box. 3
- > Click **Add**. 4
- > Set the same constraint for cells **C8** and **C10**, and click **OK**. 5

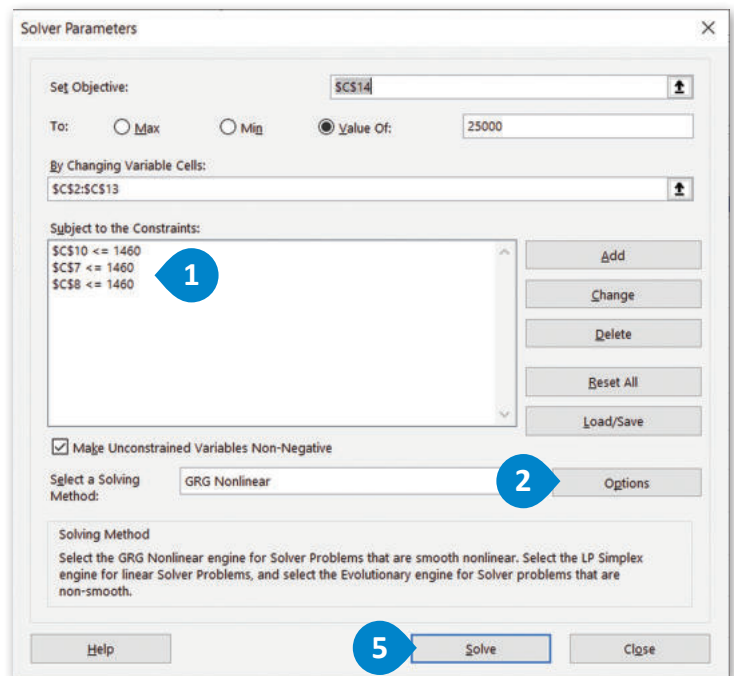


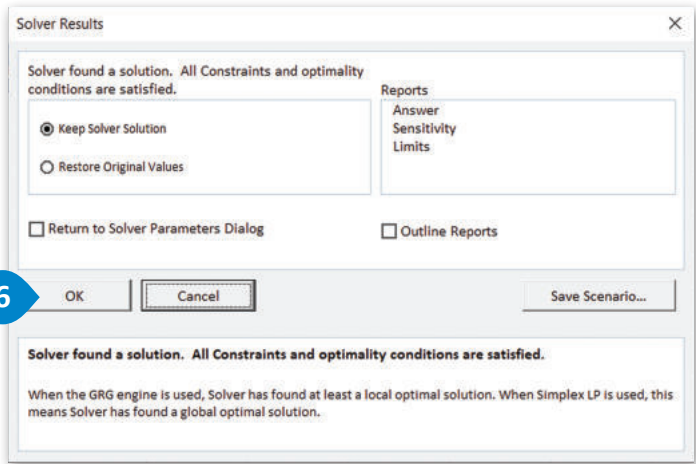
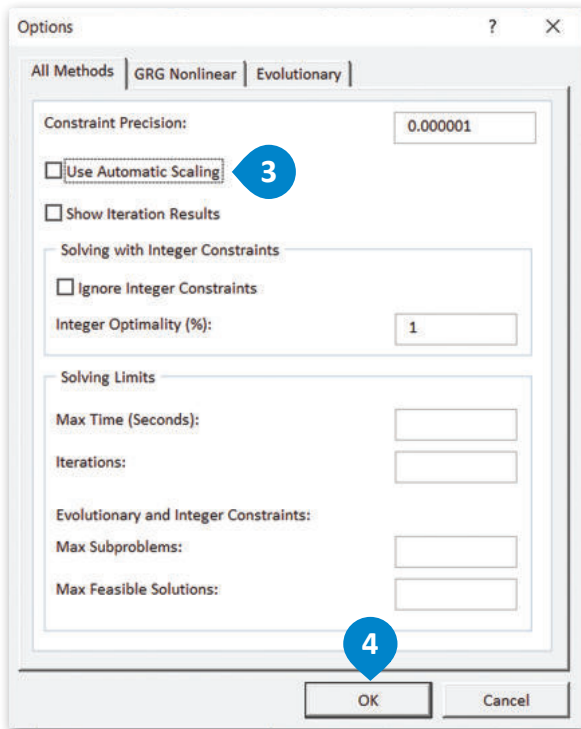
Month	Tourist visits
January 2019	1,653
February 2019	1,492
March 2019	1,599
April 2019	1,650
May 2019	2,079
June 2019	903
July 2019	750
August 2019	2,675
September 2019	635
October 2019	1,119
November 2019	1,234
December 2019	1,735
Total	17,526
Average	1,460

Figure 4.35: Set the constraints

To set the Solver parameters:

- > Make sure that the list of constraints is shown correctly in the **Subject to the Constraints** box. 1
- > Click **Options**. 2
- > In the **Options** window, uncheck the **Use Automatic Scaling** option. 3
- > Click **OK**. 4
- > In the **Solver Parameters** window, click **Solve**. 5
- > In the **Solver Results** window, click **OK**. 6
- > Changes will appear in the selected cells. 7





AutoSave Off

Tourist visits for the year 2019

Binary Academy BA

File Home Insert Page Layout Formulas Data Review View Help

Get Data Refresh All Stocks (En... Geography... Sort Filter Clear Reapply Advanced Text to Columns What-If Analysis Forecast Sheet Outline Solver

	A	B	C	D	E	F	G	H	I
1		Month	Tourist visits						
2	1	January 2019	2,156						
3	2	February 2019	1,901						
4	3	March 2019	2,069						
5	4	April 2019	2,151						
6	5	May 2019	2,874						
7	6	June 2019	1,610						
8	7	July 2019	1,563						
9	8	August 2019	3,989						
10	9	September 2019	1,534						
11	10	October 2019	1,349						
12	11	November 2019	1,514						
13	12	December 2019	2,289						
14		Total	25,000						
15		Average	2,083						

Changes will appear in the cells C2:C13.

The values in the cells C7, C8 and C10 are above the average (1,460).

Assess the Solver with Constraints Results

Now that we've executed the Solver function with constraints, we can again create a five column table (Identification number, Month, 2019 monthly tourist visits, Solver results 2023 and Difference) in order to easily compare the before and after for the Solver process. By taking a look at the results, we can observe that, this time, Excel Solver gave us a proposal that says: if we want to reach the goal of the 25,000 thousand tourist visits in the year 2023, we have to design our tourism campaign in such a way that the number of tourist visits for all the months of the year will be increased, meaning that we will have a more holistic tourism campaign, targeting the whole year, and not only the months of June, July and September, where we observed the problematic numbers in the first place. Specifically for the months of June, July and September, the Solver results suggest that our tourism campaign should focus on increasing the number of visits, but not to unrealistic levels.

Month	Solver (constraints) results for 2023	Monthly tourist visits for 2019	Difference
January	2,156	1,653	503
February	1,901	1,492	409
March	2,069	1,599	470
April	2,151	1,650	501
May	2,874	2,079	795
June	1,610	903	707
July	1,563	750	813
August	3,989	2,675	1,314
September	1,534	635	899
October	1,349	1,119	230
November	1,514	1,234	280
December	2,289	1,735	554
Total	25,000	17,526	7,474

Figure 4.37: Tourist trips values for 2019 and data after Solver with constraints

Tourist visits values for June, July and September, generated by the Solver with constraints function. The values are realistic and they will be useful for informing future decisions.

Tourist visits values for June, July and September 2019.

In conclusion, the results of the Solver with constraints show that a holistic tourism campaign should be designed in order to increase tourist visits in every month of the year, with an increase ranging from approximately 500 thousand to 1000 thousand tourist visits per month. Taking into account these suggestions, a tourism agent could decide, for example, to boost advertising for tourism in the Kingdom of Saudi Arabia during the whole year with special focus on the months of June, July and September, when special offers on air tickets could be offered, or cruises or festivals could be organized to attract more tourists.

Exercises

1

Read the sentences and tick ✓ True or False.	True	False
1. Solver is an Excel tool that helps us with optimization modeling.	<input type="radio"/>	<input type="radio"/>
2. The design of a tourism campaign is considered an optimization problem.	<input type="radio"/>	<input type="radio"/>
3. It is not essential to formulate the problem under study in advance.	<input type="radio"/>	<input type="radio"/>
4. The objective cell is always set to a specific value.	<input type="radio"/>	<input type="radio"/>
5. The Excel Solver function is rarely executed with constraints.	<input type="radio"/>	<input type="radio"/>
6. Assessing Solver results is always part of the optimization process.	<input type="radio"/>	<input type="radio"/>
7. It is important to compare past data values with predicted values in order to come to better conclusions.	<input type="radio"/>	<input type="radio"/>
8. The results of the Excel Solver should never exceed the average of the selected values.	<input type="radio"/>	<input type="radio"/>
9. The variable cells are chosen based on the phenomenon under study.	<input type="radio"/>	<input type="radio"/>
10. The objective cell and the variable cells should not be related to each other.	<input type="radio"/>	<input type="radio"/>

Project

1

Let's suppose that you are working as a travel agent and you want to design the tourism campaign for the year 2024 for the Kingdom of Saudi Arabia, in order to increase the number of tourist visits to 50,000 thousand.

2

More specifically, you should:

- download tourist visits data for the year 2018 (<https://data.gov.sa>).
- create a forecast of this data for the year 2024.
- assess the forecast results in order to formulate an optimization problem.
- use Excel Solver to get information on how you will design your tourism campaign.
- make suggestions for a tourism campaign based on the Excel Solver results.

3

Create a PowerPoint presentation using suitable Excel charts, illustrating your forecast and Excel Solver results. Explain your charts and present your suggestions for the tourism campaign.



Wrap up

Now you have learned:




- > what predictive modeling is.
- > the difference between parametric and non-parametric modeling.
- > the different types of predictive models.
- > the process of creating a predictive model.
- > the benefits and challenges of predictive modeling.
- > the applications of predictive modeling.
- > what forecasting is.
- > the different types of forecast charts.
- > how forecasting is applied to specific data.
- > what a confidence interval is.
- > what an optimization problem is.
- > how to perform optimizations using Excel Solver.
- > how to assess Excel Solver results.

KEY TERMS

Classification Model	Forecast	Optimization Problem
Clustered Column Chart	Forecast Model	Outlier Detection Model
Clustering Model	General Linear Model	Parametric
Column Chart	Gradient Boosted Model	Predictive Data Modeling
Confidence Interval	Line Chart	Prophet Model
Constraints	Linear Regression	Quality Enhancement
Data Collection	Lower Confidence Bound	Risk Assessment
Data Cleaning	Model Formulation	Stacked Column Chart
Data Transformation	Neural Network	Time Series Model
Decision Tree	Non-Parametric	Upper Confidence Bound
Excel Solver	Objective Cell	Variable Cells

Python programming prerequisite

Programming is one of the most important skills that should be acquired by students enrolled in Computer Science and Engineering pathway, as it is a requirement for a number of curricula in this pathway, including the Engineering and Data Science curricula. To facilitate the student's acquisition of Python programming basics, the following content is designed to be accessible by scanning the QR code for each topic. The student is advised to make a time plan to complete the reading of these units with the help of the suggested durations in the table below. The student can also put a tick mark (✓) to mark the completion of each unit.

Unit	Suggested duration	QR Code	Did you complete the unit?
1. Introduction to Python	One day		
2. Input-Output and Mathematical Operations	One day		
3. Conditional Statements	Two days		



Unit	Suggested duration	QR Code	Did you complete the unit?
4. Loops and Functions	Two days		
5. Lists, Tuples and Python Libraries	One week		
6. Dictionary, Nested Lists and Data Files	One week		
7. Advanced Data Structures and Recursion	Two weeks		
8. Introduction to Object Oriented Programming	Two weeks		