

Unit 04

Data and Analysis



After completing this lesson, you will be able to:

- relate the role and importance of model building with their real-world applications.
- understand and explain experimental design in data science.
- analyze pre-existing datasets to create summary statistics and data visuals (such as bar charts, pie charts, line graphs, etc.)



UNIT INTRODUCTION

Data analysis is a process, where data is cleaned, explored, and summarized, providing insights and information for the construction of statistical models to make predictions or draw inferences about real-world phenomena. The procedure helps to reduce the risks inherent in decision-making by providing useful insights and statistics. It is often presented in the form of charts, images, tables, and graphs.

A simple example of data analysis is whenever we decide in our daily lives by evaluating what has happened in the past or what will happen in the future if we make the decision in the same way? Basically, this is the process of analyzing the past or future and making a decision based on that analysis.

4.1 Statistical Modeling

Statistical modeling is the use of mathematical models and statistical assumptions to generate sample data and make predictions about the real world. A statistical model is a collection of probability distributions on a set of all possible outcomes of an experiment. Data in a model has a relationship between two or more variables. For example, in a science experiment you collect some data about the height and weight of some specific group of students. You represent the students' weight by variable x and students' height by variable y . After plotting the values, you find the relationship $y = 5x + 2$. It shows that variable y is a dependent variable because it is dependent on variable x . Figure 4.1 shows this relationship graphically.

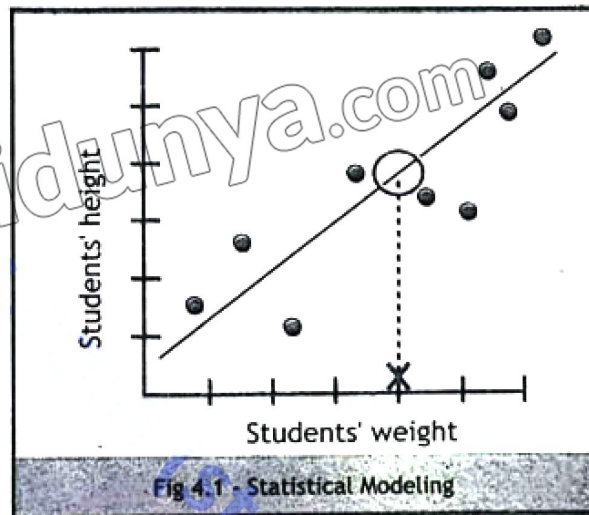


Fig 4.1 - Statistical Modeling

Statistical modeling is described as the mathematical relationship between random and non-random variables

4.1.1 Use cases for statistical modeling

A use case is a methodology used in system analysis to identify, clarify, and organize a system. A data science use case is a concrete real-world task to be solved using the available data. A use case is a description of how the users who use the system will perform tasks or achieve the goal. In the framework of a particular company, many variables are analyzed using data science techniques in the context of the company's specific industry. Data science use cases can be a problem to be resolved, a hypothesis to be checked, or a question to be answered. Essentially, doing data science means solving real-world use cases.

Although the content of each use case will likely be very different, there are some common things to always keep in mind:

- Data science use case planning is outlining a clear goal and expected outcomes, understanding the scope of work, assessing available resources, providing required data, evaluating risks, and defining KPI (key performance indicators) as a measure of success.
- The most common approaches to solve data science use cases are forecasting, classification, pattern and anomaly detection, recommendations, and image recognition.
- Some data science use cases represent typical tasks across different fields, and you can rely on similar approaches to solve them, such as customer churn rate prediction, customer segmentation, fraud detection, recommendation systems, and price optimization.

4.1.2 How to Solve a Data Science Case Study?

The answer to this question varies from case to case. It depends on the company's business strategy, and on the core of the case study itself. However, the following points can be helpful to outline a general roadmap to follow with any data science use case:

1. **Formulating the right question:** This first step is reviewing available literature.
2. **Data collection:** At this step, we conduct a data inventory and gathering process.
3. **Data wrangling:** The data is cleaned, preprocessed, transformed, manipulated, and mapped from its raw form into a more suitable format.
4. **Data analysis and modeling:** The cleaned data is analyzed from the standpoint of its current status, then fit into a selected predictive statistical model.
5. **Result communication:** It includes sharing the most important findings and ultimate conclusions with management, shareholders, and any other involved parties.

Some examples from the real life as a case study are as follows:

- **Record of Production Goods and Services:** Production analysis may help in improving the quality.
 - **Stock Market Data Analysis:** Investors and consumers can get help from predictive analysis of existing data of stock market.
 - **Weather Forecasting:** It could be helpful for aviation, agriculture, mountaineering, and businesses.
 - **Medical Records:** Data obtained and analyzed at healthcare organizations could be helpful for medical research and timely diagnosis of diseases.
 - **Sales Tracking:** A good data analysis helps to plan next strategy model to prevent loss and get profit.
 - **Population Record:** The analysis of population data is helpful for government and other organizations for better planning and distribution of resources.
-

- **Educational Data:** It helps in better planning of educational resources for students and teachers.
- **Natural Disaster Prediction:** It helps the people to take preventive measures against earthquakes, floods, landslides, snowfall, cyclones, and other natural disasters.
- **Pandemic Analysis:** It helps in controlling the pandemic and taking preventive measures timely.

A sample case study about weather forecasting:

Weather conditions play significant role in our daily life, from dressing to travelling, planning activities and events etc. Unfavourable weather conditions can cause damage to life and properties. The following are steps that how predictive analysis about weather forecast can be performed and would be helpful in our real life.



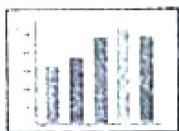
1. **Formulating the right question:** The existing data about the comparable weather situations will be analyzed and reviewed. You can observe from your past experiences that which weather conditions are likely to occur before a rainy day, a sunny day or otherwise.

2. **Data collection:** In this step you can collect data about temperature, amount of rainfall, wind speed and direction, and barometric pressure. You can collect this data by yourself for learning purposes, by using the thermometer, rain gauge, compass, and barometer. For accurate data collection an existing data from meteorological department is acquired.



3. **Data wrangling:** At this stage the collected data will be rearranged, processed, and grouped for the next stage. The data will be cleaned by removing duplicate values, impute missing values etc.

4. **Data analysis and modeling:** The cleaned data will be fitted to some statistical model for analysis.



5. **Result communication:** The accuracy of our results obtained from the statistical model will determine its reliability. Such models are trained on sample data and can be used to predict the weather condition. For example, when we give some new data about temperature, amount of rainfall, wind speed and direction, atmospheric pressure to this model, it will give us the prediction that how likely be the weather considering this new data.

4.1.3 Statistical modeling techniques

Data gathering is the foundation of statistical modeling. The data may come from the cloud, spreadsheets, databases, or other sources. There are two categories of statistical modeling methods used in data analysis: Supervised learning and unsupervised learning.

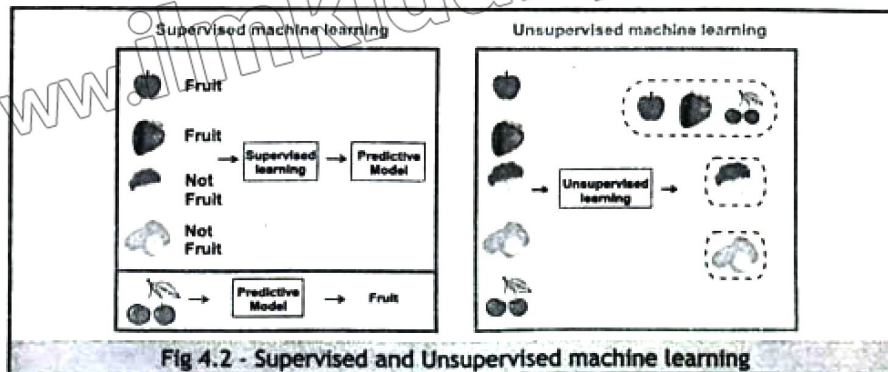


Fig 4.2 - Supervised and Unsupervised machine learning

Supervised learning

In the supervised learning model, the algorithm uses a labeled dataset for learning, with an answer key the algorithm uses to determine accuracy as it trains on the data. For example, if you give a dataset of fruit and vegetables with labels "Fruit" and "Not Fruit". The model will learn from the input data and labels that how does a fruit or vegetable should look like. Then if you give it some fruit or vegetable without label, it will recognize it and assign it the correct label. Supervised learning techniques in statistical modeling include regression model and classification model.

Regression model: If the result, label, or outcome of a model is continuous value then it will be called regression. The most common regression model is a linear model. A linear regression model is a mathematical equation that allows us to predict a response for a given predictor value. The variable which is used for prediction is known as independent variable (x) and the variable to be predicted based on the value of variable x is known as dependent variable (y). In simple linear regression the equation of the line is $y = mx + b$, where m is the slope and b is the y intercept. If the values of independent variable is plotted along the x-axis and the values of dependent variable are plotted along the y-axis and a line is drawn then the intercept and slope will be as follows:

Intercept: It is the point where the graph line meets the y-axis. It is also called y-intercept.

Slope: It is calculated by the term, rise/run where rise is the distance from the x-axis to the y-intercept, and the run is the distance from y-axis to the y-intercept.

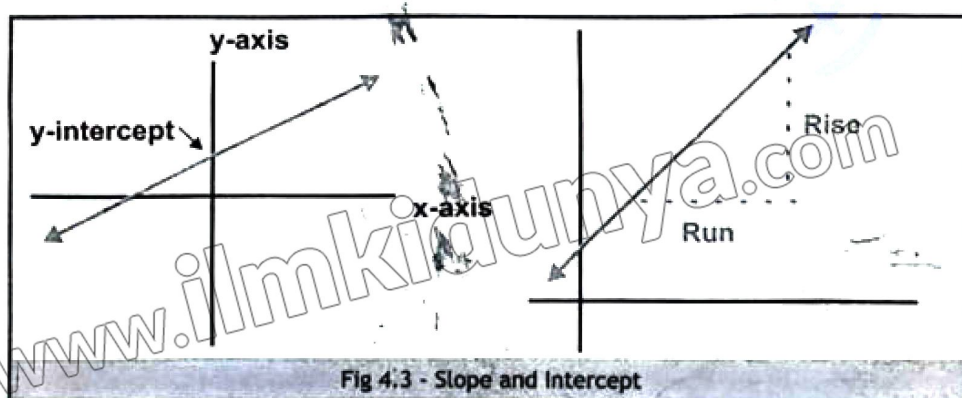


Fig 4.3 - Slope and Intercept

Classification model: If the result, label, or outcome of a model is a discrete value then it will be called classification. For example, if we predict that will a certain employ get raise in salary or not then it will be classification example. But if we make a prediction that how much salary raise an employee will get, based on some statistical data, it will be regression model.

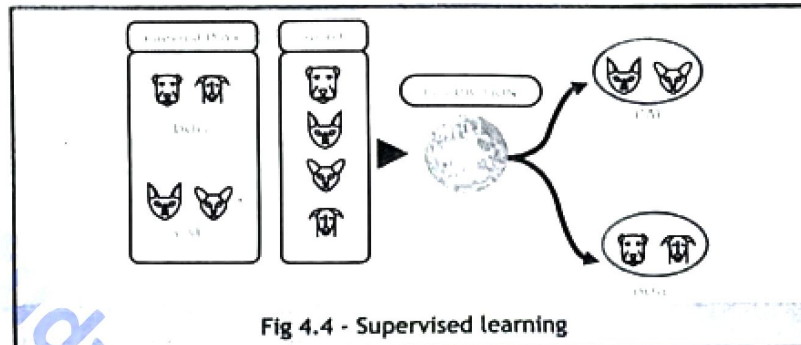


Fig 4.4 - Supervised learning

Unsupervised learning

In the unsupervised learning model, the algorithm is given unlabeled data and attempts to extract features and determine patterns independently. Clustering algorithms and association rules are examples of unsupervised learning. K-mean clustering algorithms and reinforcement learning algorithms are used in unsupervised learning. The term clustering means the group of data items with maximum similarities. For example, to reduce the churn rate of customer a telecom company observes the customers' usage and divide them into three clusters, the customers with long call duration, the customers with heavy internet usage, and the customers with short calls and average internet usage. Now the company will give economical call rates to the cluster of having long calls, and attractive internet packages to the customer with heavy internet usage. The term association means how likely is to do second action if first action is done. For example, in a supermarket customer 1 buys bread, milk, tea and tissues. Customer 2 buys bread, milk, coffee, and eggs. Now if the customer 3 buys bread, then there are more chances that he will also buy milk. In this example an association between bread and milk has been found.

K-means clustering:

The algorithm combines a specified number of data points into specific groupings based on similarities. An example based on the image could be a machine learning algorithm that instead of classifying by known labels separates data into groups where members are similar as possible as shown in Figure 4.5.

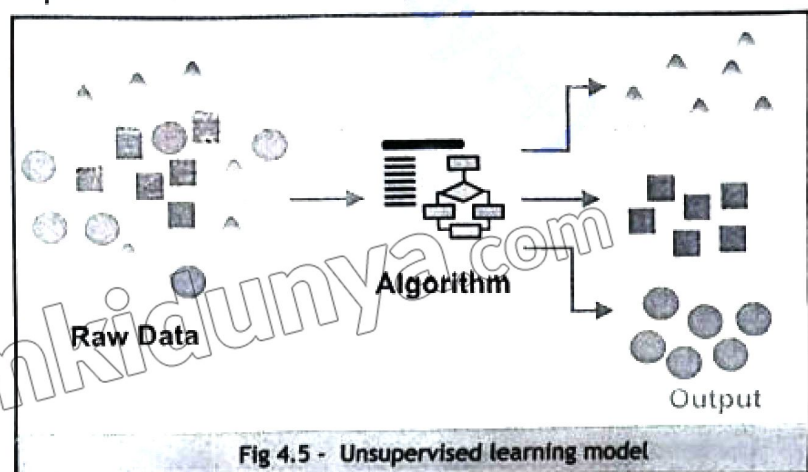


Fig 4.5 - Unsupervised learning model

4.1.4 Build a statistical model using Python

You can build statistical models to perform predictive analysis. Several tools are available for this task such as MS-Excel, Weka, R Studio and Python. Many datasets are available on the internet. www.kaggle.com/ and <https://github.com/> are the most popular platform to download various datasets. However, to create your experimental statistical model you don't need to download any dataset.

You can create your own dataset by

generating some random numbers by using python code. We will plot these numbers on a linear graph by using correlation coefficients $y = mx + c$ ($y = 3x + 4$, where x will also be generated randomly) using slope and intercept. You could write python code in your favorite IDE if you already had installed any, else you can open browser on the computer and type the URL <https://colab.google/> and press enter key. The online compiler for Python will open, click on "Open Colab" button on the top right corner of the browser window (See Figure 4.6)

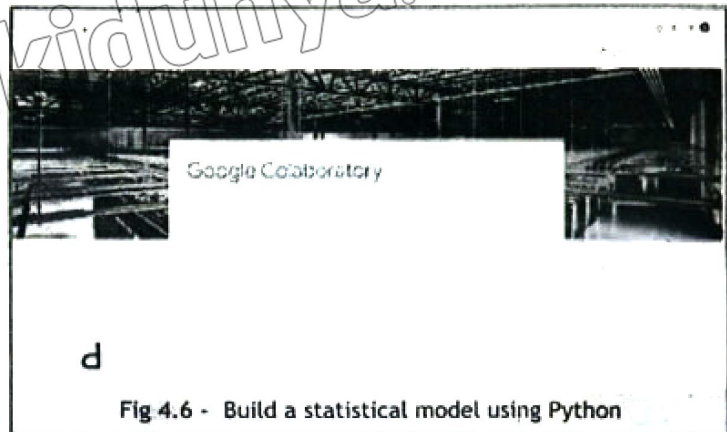


Fig 4.6 - Build a statistical model using Python

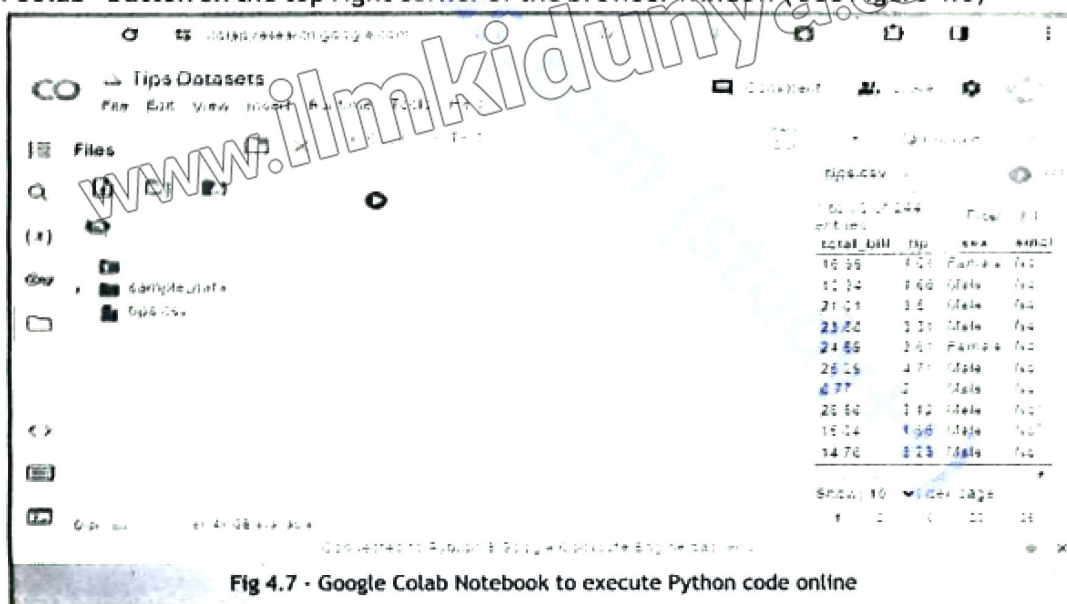


Fig 4.7 - Google Colab Notebook to execute Python code online



Teacher's Guide

To explore more tools to build statistical model using python following are some useful links to download Weka and R Studio:

<https://www.weka.io>

<https://www.rstudio.com/tags/rstudio-ide>

Type the following code in the browser window:

```
1. # Import necessary libraries
2. import numpy as np
3. from sklearn.linear_model import LinearRegression
4. import matplotlib.pyplot as plt
5. # Generate a small dataset
6. np.random.seed(42)
7. X = 2 * np.random.rand(100, 1)
8. y = 4 + 3 * X + np.random.randn(100, 1)
9. # Create and fit the linear regression model
10. model = LinearRegression()
11. model.fit(X, y)
12. # Make predictions on new data points
13. X_new = np.array([[0], [2]])
14. y_pred = model.predict(X_new)
15. # Plot the original data and the linear regression line
16. plt.scatter(X, y, label='Original data')
17. plt.plot(X_new, y_pred, 'r-', label='Linear regression line')
18. plt.xlabel('X')
19. plt.ylabel('y')
20. plt.legend()
21. plt.show()
```

Click the Run button or press Ctrl+E keys, the linear regression graph will be fitted on newly generated data set. See Figure 4.8.

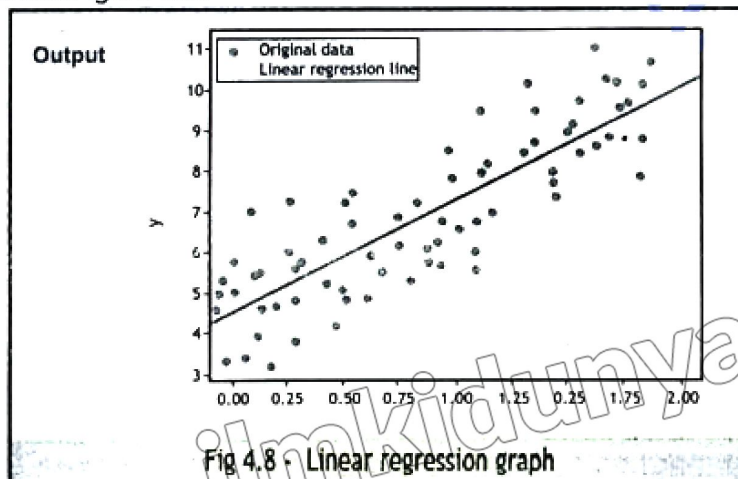


Fig 4.8 - Linear regression graph

Experimental design is about precise planning and design to ensure that you have the appropriate data and design for your analysis or studies so that erroneous conclusions can be prevented.

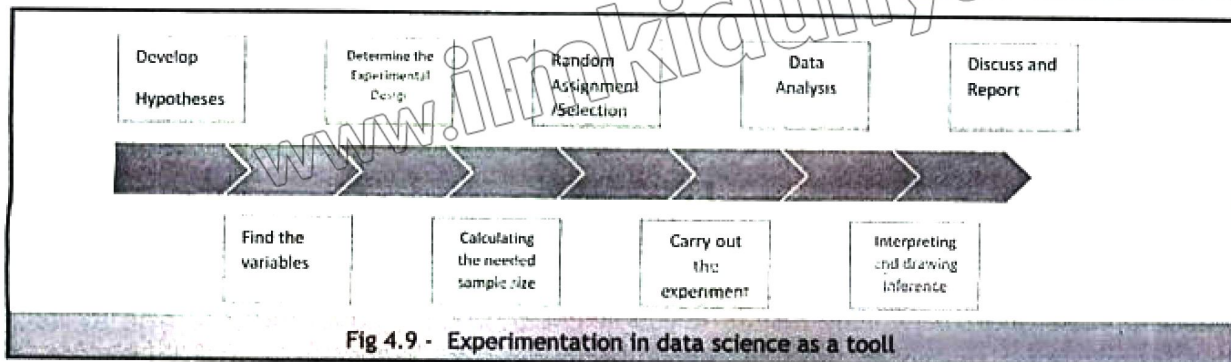
4.2.1 Experimentation in data science as a tool

Experimental design is a tool or technique used to organize, conduct, and clarify the results of an experiment efficiently. It leads to the important factors that must be investigated. An experiment will produce precise and trustworthy results if you select the right sample size and plan it carefully. Experimental design is used in many disciplines like engineering, psychology, agriculture, and medicine.

Experimental Design Flow

In data science, the experimental design flow often adheres to a standardized procedure to guarantee careful preparation, execution, and evaluation of experiments. Here is a general description of the data science experimental design flow. It will indicate what the research question is. Clearly define the research topic or problem statement that your experiment will attempt to solve. This will serve as the design's compass throughout.

- **Develop Hypotheses:** The first step is developing hypotheses which will clarify the differences and correlations between various variables involved in the experiment.
- **Find the variables:** In this step you must decide which variables will be fixed and which will be changed. The fixed variables will be called independent variables. The values that will change according to fixed variable will be called dependent variables. For example, if the salary of the person will be higher for the employees with more work experience, then years of experience will be independent variable and salary will be dependent variable. You should also consider other potential confounding factors that might need to be taken into consideration in this case the skill type of an employee can also affect the salary. For example, employees having C++ programming skills might be paid higher than employees having Java programming skills.
- **Determine the Experimental Design:** Choose an appropriate experimental design according to the nature of research and available resources. Factorial designs, randomized block designs, totally randomized designs, and more are examples of common designs. Each design has its own benefits and factors to consider. (The details of these experimental designs is beyond the scope of this book.)
- **Calculating the Needed Sample Size:** Calculate the necessary sample size to generate appropriate statistical results and identify their significant effects.
- **Random Assignment and Selection:** To produce more accurate results choose random samples. It ensures that groups are identical in there is no bias. It will ensure that any change in dependent variable is purely based on the change in independent variable.
- **Carry out the experiment:** After taking careful sample size follow the instructions and gather the data. While collecting data about the samples, you should strictly follow the defined



- **Data Analysis:** Use the collected data for appropriate statistical analysis to evaluate the hypotheses and develop conclusions. Depending on the experimental design and research issue, this may require different procedures including hypothesis testing, regression analysis, ANOVA, and other statistical methods. ANOVA (analysis of variance) is a statistical test for detecting differences in the group means, when there is one dependent variable and one or more independent variable.
- **Interpreting and Drawing Conclusions:** After data analysis you will reach to the conclusion. Observe the results carefully and consider any practical factors which might affect the results.
- **Discuss and Report:** Give a clear and concise presentation of the design of the experiment, methodology, findings, and conclusions. To enable reproducibility and transparency, the experiment should be thoroughly documented. In this way if the same experiment is repeated by different group of data scientists, it will produce more or less the same results.

Principles of Experimental Design

The fundamentals of this design determine whether an experiment is successful. These principles of experimental design consist of:

Principle of Randomization: According to this principle you divide the people into various treatment groups randomly. The randomization technique ensures that there is no bias or preplanned grouping for example some groups might have people with one capability and others might have some other strengths or weaknesses, such grouping cannot provide accurate results. Therefore, groups should be decided randomly. For example, if we conduct an experiment about the effectiveness of a new medicine for asthma. There are 100 patients. The principle of randomization means we must randomly make the two groups of patients. 50 patients will be in each group. Some people in the first group might have severe asthmatic problem while the other might have less chronic condition.

Principle of local Control: A control group is a group that has not received any treatment; hence a control group is used to compare the treatment group's findings to its outcomes. This approach guarantees that the treatment is the reason that any other factors are to blame for the outcomes. In our example we will add the new medicine to one group and the other group will be treated with their regular medication. The group which will not be given the new medicine is called

control group.

Principle of Blocking: Grouping participants according to a particular trait that could have an impact on the results is known as blocking. This principle makes sure that the results aren't affected by group differences. For example, the patients belong to two different genders, male and female. So we can make two blocks, one for women and other for men. Suppose there are 56 women and 44 men, we will make two groups of women 28 each and two groups of men 22 each. One group of 28 women and 22 men will be treated with new medicine while the other group of 28 women and 22 men will be treated with their regular medicines.

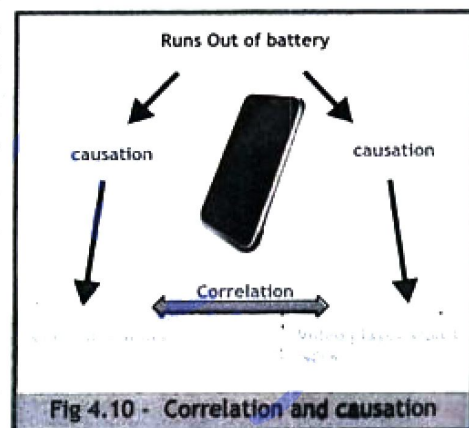
Principle of Replication: If you repeat the experiment again and again, it is called replication. The replication ensures that results are reliable. This principle ensures that experimental outcomes are not coincident, erroneous or a result of some randomness. In above example if we get the results that new medicine is more effective than the previously used medicines, then different data scientists should repeat this test for new medicine, before finally declaring it effective. Because there might be some coincidence, or patients taking the new medicine might have some psychological impact of taking new medicine. Therefore, for principle of replication ensures that these experiments should be repeated with different group of patients, may be from some other demographic background. If the medicine gives the same results repeatedly then we can declare that the new medicine is more effective.

4.2.2 Correlation and causation

Correlation describes an association between different types of variables. It is a statistical measure of the degree to which changes to value of one variable predicts the change in the value of other variable. It is a statistical indicator of the relationship between variables. These variables change together which means they co-vary. But this covariation is not necessarily due to a direct or indirect causal link.

Causation means that changes in one variable brings about changes in the other. There is a **cause-and-effect** relationship between variables. The two variables are correlated with each other and there is also a causal link between them. A correlation does not imply causation, but causation always implies correlation.

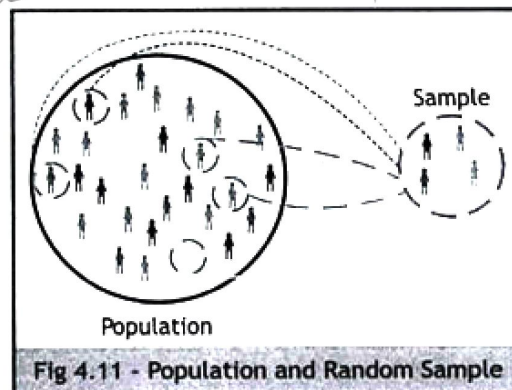
For example, a smart phone user complains, "Whenever I try to send text message, my phone lags.." A quick look at the smartphone confirmed that there were five game apps open at the same time plus Facebook and YouTube. The act of trying to send a text message wasn't causing the lagging, the lack of RAM was causing the phone to lag. But the user immediately connected it with the last action she was doing before the lags. She was implying a causation where there was only a correlation.



4.2.3 Population and Random Sample

Population and random sample are the terms used in statistics. Population is a complete collection or set of similar items or events, which are of interest or under consideration regarding some problem or experiment. Therefore, a population is a complete set of people with a specialized set of characteristics.

Whereas, a random sample is a subset of population. each member of the population has an equal chance of being selected as a sample. A population and random sample is shown in Figure 4.11.



4.2.4 Parameter and Statistic

Parameter is a number or quantity in statistical population, that describes an aspect of entire population. While a statistic is a number or quantity that describes characteristics or an aspect of a sample. Parameters are the key things we want to learn about, they are usually unknown. Sample's statistics give us estimates for parameters. There will always be some uncertainty about how accurate estimates are. More certainty gives us more useful knowledge. For every parameter we want to learn about we can get a sample and calculate a sample statistic, which gives us an estimate of the parameter. Some Important Examples are:

Table 4.1 - Parameter and Statistic

Parameter	Statistics
Average income for Pakistan.	Average income for government employees in Pakistan.
Mean weight of Persian cats.	Mean weight of 100 Persian cats.
Proportion of all people who prefer soft drinks.	Proportion of random sample of 500 persons who prefer soft drinks.
Standard deviation of all the transaction times in the state bank.	Standard deviation of a random sample of 1000 transaction times in the state bank.

Tip

A parameter is a fixed, unknown numerical value, while the statistic is a known number and a variable which depends on the portion of the population.

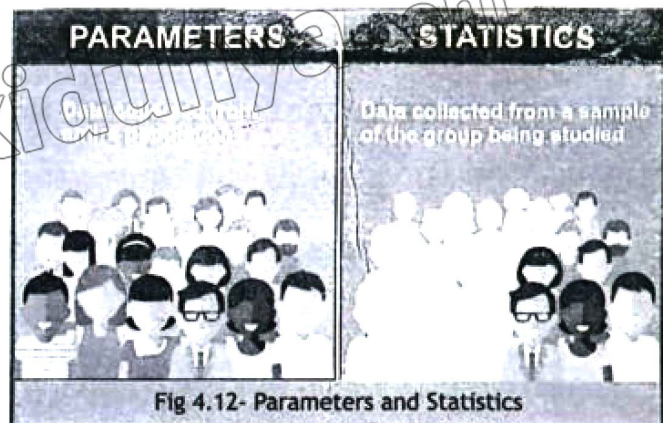
Figure 4.12 shows the difference between parameters and statistics.

Mean, median and mode are different types of averages used to represent typical values in a population. Therefore you can say that the mean value of a population is parameter while the mean value of a sample is statistics. Some examples of parameters and statistics are as follows:

The typical age of people in a country

The typical profits of a company

The typical range of an electric car



4.2.5 Data collection methods

Data collection is the process of collecting, measuring and analyzing different types of information using a set of standard validated techniques. The main objective of data collection is to gather information-rich and reliable data, and analyze them to make critical business decisions. Once the data is collected, it goes through a rigorous process of data cleaning and data processing to make this data truly useful for businesses. There are two main methods of data collection in research based on the information that is required, namely:

- Primary Data Collection
- Secondary Data Collection

Primary Data Collection Methods

Primary data refers to data collected from firsthand experience directly from the main source. It refers to data that has never been used in the past. Here are some of the most common primary data collection methods:

Interviews

Interviews are a direct method of data collection. It is simply a process in which the interviewer asks questions, and the interviewee responds to them. It provides a high degree of flexibility because questions can be adjusted and changed anytime according to the situation.



Observations



In this method, researchers observe a situation around them and record the findings. It can be used to evaluate the behavior of different people in controlled (everyone knows they are being observed) and uncontrolled (no one knows they are being observed) situations. This method is highly effective because it is straightforward and not directly dependent on other participants. For example,

a person looks at random people that walk their pets on a busy street, and then uses this data to decide whether or not to open a pet food store in that area.

Surveys and Questionnaires

Surveys and questionnaires provide a broad perspective from large groups of people. They can be conducted face-to-face, mailed, or even posted on the Internet to get respondents from anywhere in the world. The answers can be yes or no, true or false, multiple choice, and even open-ended questions.



Focus Groups

A focus group is similar to an interview, but it is conducted with a group of people who all have something in common. The data collected is similar to in-person interviews, but they offer a better understanding of why a certain group of people thinks in a particular way. However, some drawbacks of this method are lack of privacy and domination of the interview by one or two participants. Focus groups can also be time-consuming and challenging, but they help reveal some of the best information for complex situations.

Oral Histories

Oral histories also involve asking questions like interviews and focus groups. However, it is defined more precisely, and the data collected is linked to a single phenomenon. It involves collecting the opinions and personal experiences of people in a particular event that they were involved in. For example, it can help in studying the effect of a new product in a particular community.



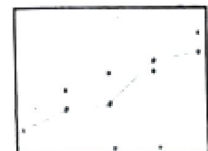
Secondary Data Collection Methods

Secondary data refers to data that has already been collected by someone else. It is much more inexpensive and easier to collect than primary data. While primary data collection provides more authentic and original data, there are numerous instances where secondary data collection provides great value to organizations.

Here are some of the most common secondary data collection methods:

Internet

The use of the Internet has become one of the most popular secondary data collection methods in recent times. There is a large pool of free and paid research resources that can be easily accessed on the Internet. While this method is a fast and easy way of data collection, you should only source from authentic sites while collecting information.



Government Archives

There is lots of data available from government archives that you can make use of. The most important advantage is that the data in government archives are authentic and verifiable. The challenge, however, is that data is not always readily available due to several factors. For example, criminal records can come

under classified information and are difficult for anyone to have access to them.

Libraries

Most researchers donate several copies of their academic research to libraries. You can collect important and authentic information based on different research contexts. Libraries also serve as a storehouse for business directories, annual reports and other similar documents that help businesses in their research.



4.2.6 Real world experimentation examples

Facebook A/B testing: A/B testing, also known as split testing, is a method used by companies like Facebook to compare two versions of something to figure out which one performs better. In the context of Facebook or other online platforms, this often involves testing different versions of a webpage, app feature, or advertisement to see which one leads to better results.

AirBnB uses data science to help renters set their prices: Airbnb uses statistical experimentation, such as A/B testing, to enhance its platform and user experience. For instance, when testing a new feature like a redesigned booking flow or a different search algorithm, Airbnb would create two versions (A and B) and randomly assign users to each group. By collecting and analyzing data on user interactions, conversion rates, or other relevant metrics, Airbnb can determine which version performs better.

This data-driven approach allows Airbnb to make informed decisions about implementing changes that positively impact user satisfaction, booking rates, or other key performance indicators. Through ongoing experimentation, Airbnb continuously refines its platform, ensuring that updates are based on empirical evidence rather than assumptions, ultimately contributing to a more effective and user-friendly service.

YouTube and statistical experimentations: YouTube utilizes statistical experimentation to improve user engagement, content discovery, and overall user experience. A/B testing is one of the key methodologies they use. For example, when introducing a new recommendation algorithm or modifying the layout of the platform, YouTube may create two versions (A and B) and randomly assign users to each group. YouTube collects data on user interactions, such as click-through rates, watch time, and user feedback, to evaluate the performance of each version. By comparing metrics between the control group (A) and the experimental group (B), YouTube can make data-driven decisions about the effectiveness of the changes. This iterative process allows them to fine-tune algorithms, recommend more relevant content to users, and optimize features to enhance overall satisfaction.

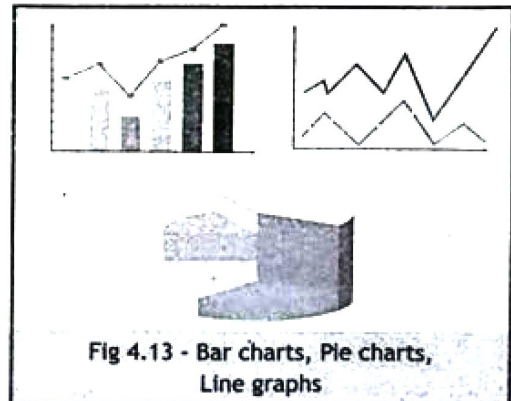
4.3 Analyze pre-existing datasets to create summary statistics and data visuals. (bar charts, pie charts, line graphs, etc.)

To analyze a dataset a series of steps is involved. By following all those steps, you can extract meaningful insight and information from the data. These steps include:

• Data exploration

• Data cleaning

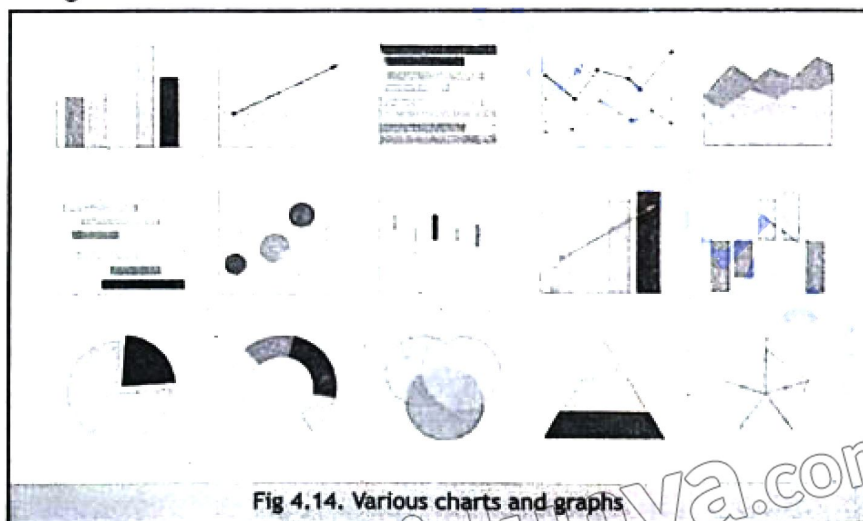
You eliminate unwanted data to get more meaningful insight. Summary statistics are calculated to understand the tendency (how closely similar is data) and dispersion (how scatter is the data). After computing, mean, median, mode you will calculate count and frequently etc. After all these steps of data exploration, data cleaning, calculation of summary statistics, data will be ready for visualization. Various diagrams/tools are used for data visualization like charts, graphs, boxplots etc. various graphs are shown in Figure 4.13.



4.3.1 Data products (charts, graphs, statistics)

A data product is a tool that uses data to help businesses to improve their decisions and business processes. Data products, in turn, leverage the results of data analysis to automate processes, provide real-time recommendations, or deliver data-driven services to users or organizations.

In simple words, data analysis is the process of deriving insights and knowledge from data, while data products are the applications or tools that use those insights to deliver value to users. Together, they form the core of data science, enabling organizations to harness the power of data for decision-making and innovation.



4.3.2 Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see

and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

4.3.3 Data analysis through Python

In today's world, a lot of data is being generated daily. And sometimes to analyze this data for certain trends, patterns may become difficult if the data is in its raw format. To overcome this data visualization comes into play. Data visualization provides a good, organized pictorial representation of the data which makes it easier to understand, observe, and analyze data. Python provides various libraries that come with different features for visualizing data. All these can support various types of graphs.

Dataset of Tips in a restaurant (Tips.csv)

Tips database is the record of the tip given by the customers in a restaurant for two and a half months in the early 1990s. It contains 6 columns such as total bill, tip, male/female, smoker, day, time, size.

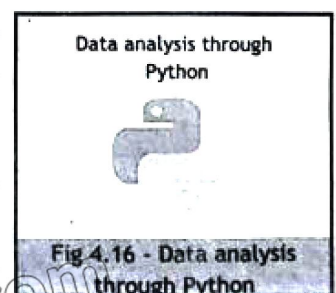
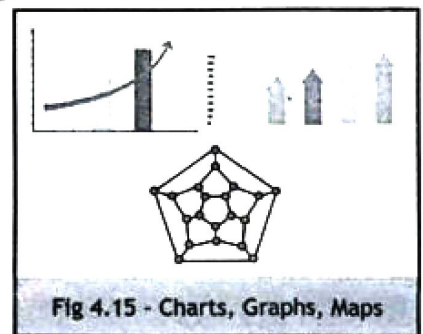
You can download the tips database from the Internet (<https://www.mongodb.com> or <https://www.kaggle.com/datasets/jsphyg/tipping>) To analyze dataset or database in Python you need to install certain libraries. Pandas is the most important library to install for data analysis. To install pandas library the following is the command:

```
pip install pandas
```

To run the Python code you can use google colab as mention earlier in this unit. However, this time you need to upload the dataset "tips.csv" to your google drive, so that Python code in colab notebook can access it directly. You can see the dataset in the left pane of the browser in Figure 4.17.

Example Python Code is as following:

```
import pandas as pd
# reading the database
data = pd.read_csv("tips.csv")
# printing the top 10 rows
display(data.head(10))
```



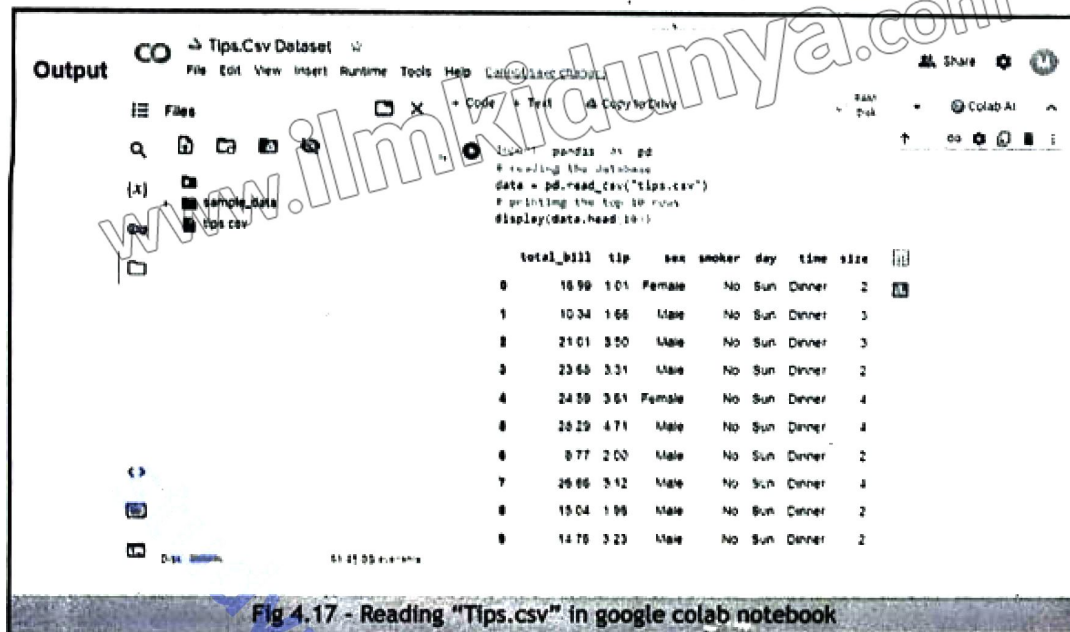


Fig 4.17 - Reading "Tips.csv" in google colab notebook

Matplotlib: Matplotlib is an easy-to-use, low-level data visualization library that is built on NumPy arrays. It consists of various plots like scatter plot, line plot, histogram, etc. Matplotlib provides a lot of flexibility. To install this type the below command in the terminal.

Scatter Plot: Scatter plots are used to observe relationships between variables and uses dots to represent the relationship between them. The scatter() method in the matplotlib library is used to draw a scatter plot. In the browser window click on + Code button to type new piece of code given below:

1. import pandas as pd
2. import matplotlib.pyplot as plt
3. # reading the database
4. data = pd.read_csv("tips.csv")
5. # Scatter plot with day against tip
6. plt.scatter(data['day'], data['tip'])
7. # Adding Title to the Plot
8. plt.title("Scatter Plot")
9. # Setting the X and Y labels
10. plt.xlabel('Day')
11. plt.ylabel('Tip')
12. plt.show()

Output

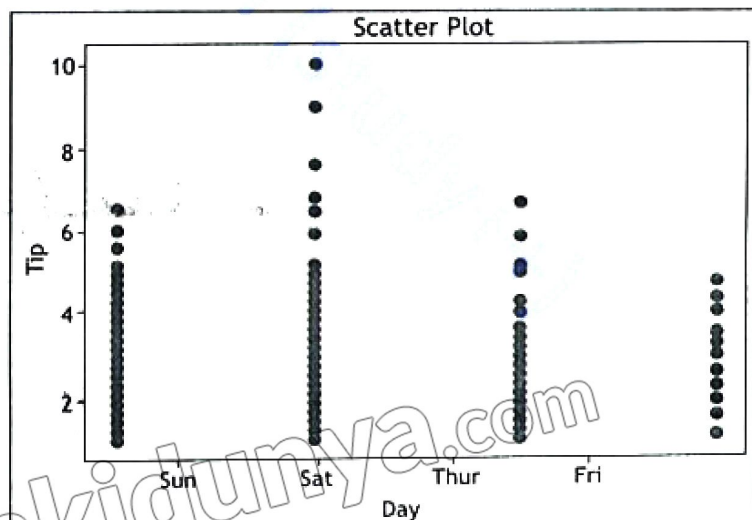


Fig 4.18 - Python Code

Line Chart: Line Chart is used to represent a relationship between two data X and Y on a different axis. It is plotted using the plot() function. You can use the following code to make line chart.

```
1. import pandas as pd
2. import matplotlib.pyplot as plt
3. # reading the database
4. data = pd.read_csv("tips.csv")
5. # Scatter plot with day against tip
6. plt.plot(data['tip'])
7. plt.plot(data['size'])
8. # Adding Title to the Plot
9. plt.title("Scatter Plot")
10. # Setting the X and Y labels
11. plt.xlabel('Day')
12. plt.ylabel('Tip')
13. plt.show()
```

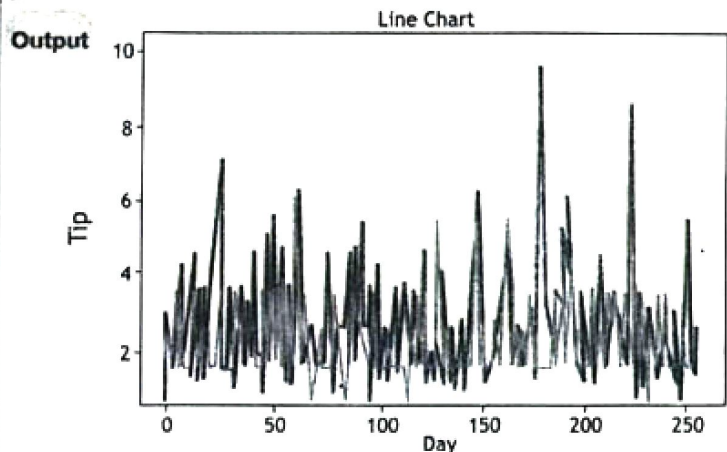


Fig 4.19 - Python Code

Bar Chart: A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. It can be created using the bar() method.

```
1. import pandas as pd
2. import matplotlib.pyplot as plt
3. # reading the database
4. data = pd.read_csv("tips.csv")
5. # Bar chart with day against tip
6. plt.bar(data['day'], data['tip'])
7. plt.title("Bar Chart")
8. # Setting the X and Y labels
9. plt.xlabel('Day')
10. plt.ylabel('Tip')
11. # Adding the legends
12. plt.show()
```

Output

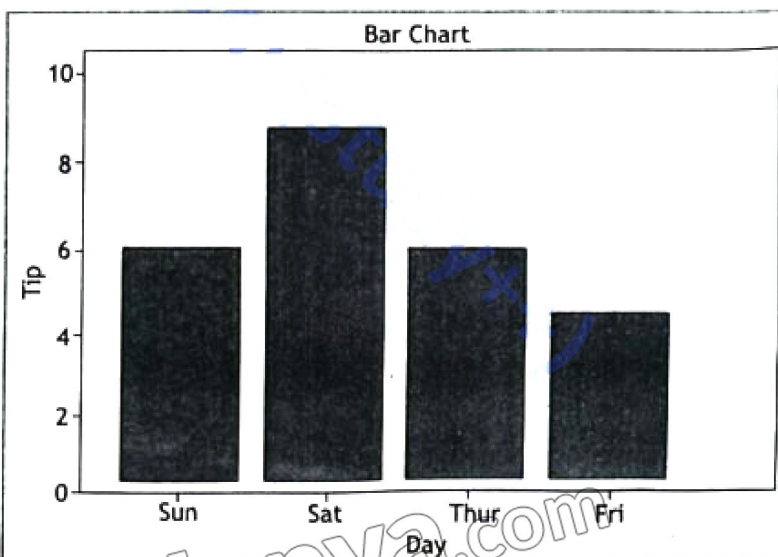
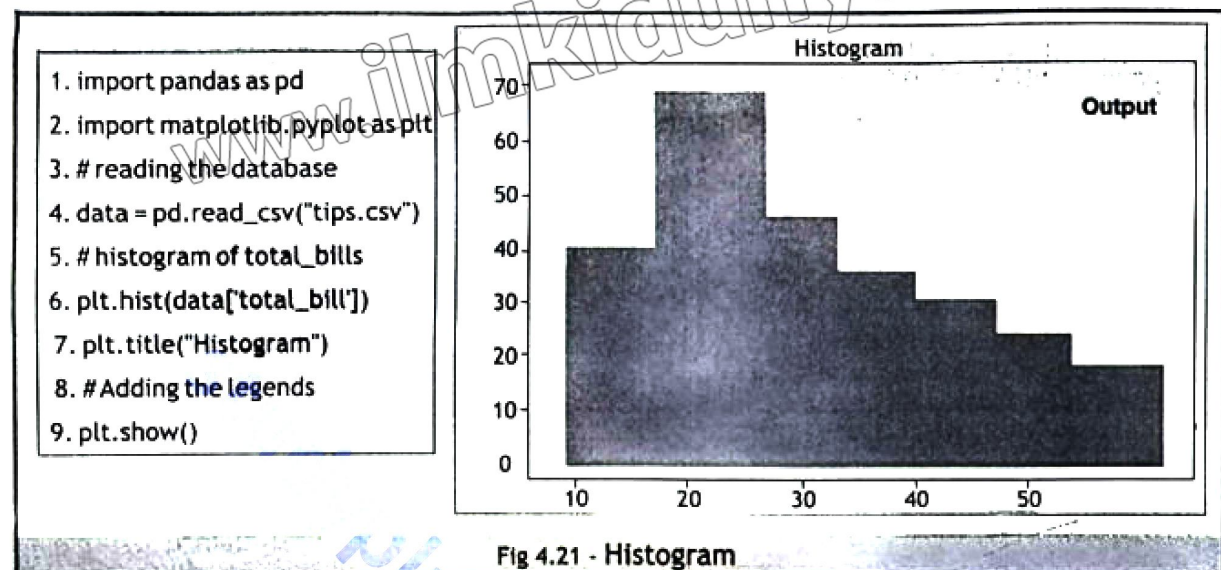


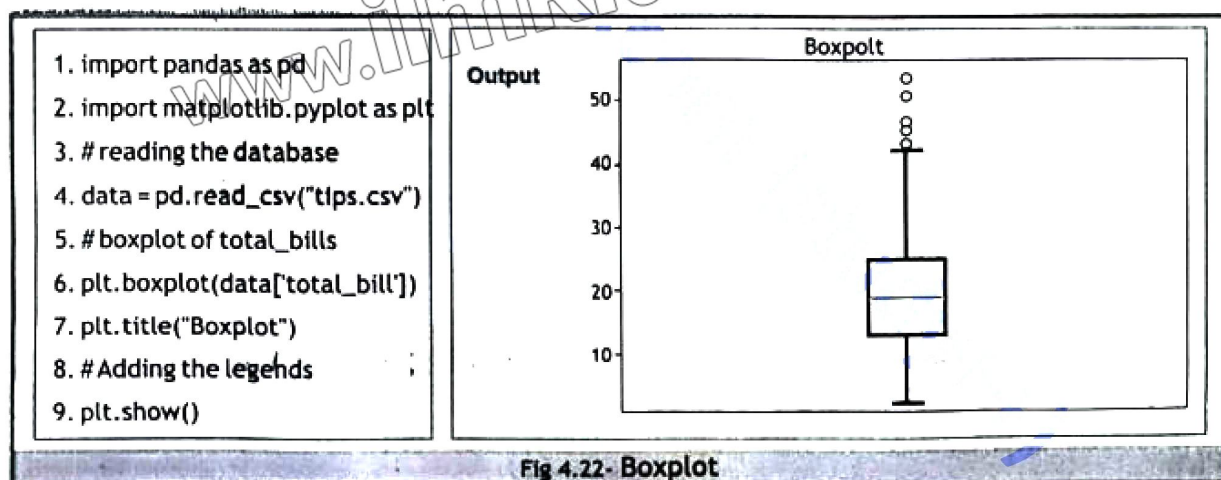
Fig 4.20 - Bar Chart

Histogram: A histogram is basically used to represent data in the form of some groups. It is a type of bar plot where the X-axis represents the bin ranges while the Y-axis gives information about frequency. The hist() function is used to compute and create a histogram. In histogram, if we pass

categorical data then it will automatically compute the frequency of that data i.e. how often each value occurred.



Boxplot: A boxplot, also known as a box-and-whisker plot, is used to represent distribution of a dataset. It provides a summary of key statistical measures, allowing you to visualize the central tendency, spread, and identify potential outliers in the data. The boxplot() function is used to compute a boxplot



Pie Chart: A pie chart is a circular statistical chart, which is divided into sectors to illustrate numerical proportion. Given a set of categories or groups with their corresponding values you can make use of the pie function from matplotlib to create a pie chart in Python. By default, the area of the slices will be calculated as each value divided by the sum of values. With pyplot, you can use the pie() function to draw pie charts.


```

1. import pandas as pd
2. import matplotlib.pyplot as plt
3. # reading the database
4. data = pd.read_csv('tips.csv')
5. # Pie chart of tips on various days
6. plt.pie(data['day'].value_counts(),
    autopct="%.1f%%", radius=1,
    labels=['Sat', 'Sun', 'Thur', 'Fri'])
7. # Adding the legends
8. plt.title("Pie Chart")
9. plt.show()

```

Output

Pie Chart

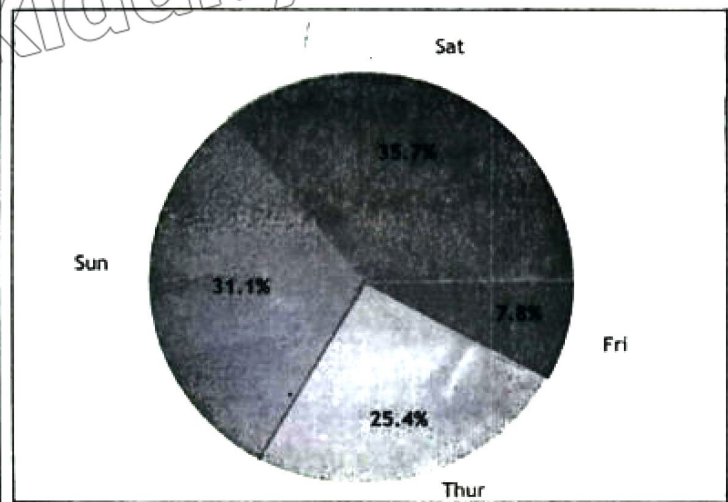


Fig 4.23- Pie Chart

Note: you can display more than one graph in a single instance of code, according to the need of data representation. Here, only one graph in each piece of code is shown for simplicity and better understanding of the students.

Summary

- **Data analysis:** is the primary step, where data is cleaned, explored, and summarized, providing insights and information for the construction of statistical models to make predictions or draw inferences about real-world phenomena.
- **Statistical modeling:** is the use of mathematical models and statistical assumptions to generate sample data and make predictions about the real world.
- **Statistical model:** is a collection of probability distributions on a set of all possible outcomes of an experiment.
- **Use case:** is a methodology used in system analysis to identify, clarify, and organize a system. A use case is a description of how the users will perform tasks or achieve the goal.
- **Supervised learning model:** uses a labeled dataset for learning, with an answer key to determine accuracy as it trains on the data.
- **Regression model:** is predictive model designed to analyze the relationship between independent and dependent variables.
- **Linear regression model:** is a mathematical equation that allows us to predict a response for a given predictor value.
- **Classification model:** analyzes and classifies a large and complex set of data points. Common models include decision trees, Naive Bayes, the nearest neighbor, random forests, and neural networking models.
- **Unsupervised learning:** is given unlabeled data and attempts to extract features and determine patterns independently.
- **K-means clustering:** combines a specified number of data points into specific groupings based on similarities.
- **Reinforcement learning:** involves training the algorithm to iterate over many attempts using deep learning.
- **Experimental design:** is about precise planning and design to ensure that you have the appropriate data and design for your analysis or studies so that erroneous conclusions can be prevented.
- **Randomization:** is the technique of dividing people into various treatment groups at random. This approach guarantees that prejudice or innate variations between the groups did not affect the outcomes.
- **Control group:** is a group that has not received any treatment, hence a control group is used to compare the treatment group's findings to its outcomes. This approach guarantees that the treatment is the reason that any other factors are to blame for the outcomes.

- **Blocking:** is grouping the participants according to a particular trait that could have an impact on the results. This principle makes sure that the results aren't affected by group differences.
- **Correlation:** describes an association between types of variables.
- **Causation** means that changes in one variable bring about changes in the other variable.
- **Population** is a complete set of people with a specialized set of characteristics.
- **Sample:** is a subset of the population.
- **Parameter:** is a number that describes something (usually mean) about the whole population
- **Statistics:** are a number that describes something (usually mean) about the sample or describe properties of sample.
- **Mean, MEDIAN AND Mode** are different types of averages used to represent typical values in a population.
- **Data collection:** is the process of collecting, measuring, and analyzing different types of information using a set of standard validated techniques.
- **Primary data:** refers to data collected from firsthand experience directly from the main source. It refers to data that has never been used in the past. Here are some of the most common primary data collection methods:
 - **Interviews:** are a direct method of data collection.
 - **Observations:** is the method, in which researchers observe a situation around them and record the findings.
 - **Surveys and Questionnaires:** provide a broad perspective from large groups of people.
 - **Focus Groups:** is similar to an interview, but it is conducted with a group of people who all have something in common.
 - **Oral Histories:** involve asking questions like interviews and focus groups. However, it is defined more precisely, and the data collected is linked to a single phenomenon.
- **Secondary data:** refers to data that has already been collected by someone else. It is much more inexpensive and easier to collect than primary data.:
- **Internet:** is one of the most popular secondary data collection methods in recent times.
- **Government Archives:** contains a lot of authentic and verifiable data.
- **Libraries:** are important and authentic sources of information based on different research contexts.
- **Data visualization:** is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an

accessible way to see and understand trends, outliers, and patterns in data.

- **Tips database:** is the record of the tip given by the customers in a restaurant for two and a half months in the early 1990s. It contains 6 columns such as total bill, tip, sex, smoker, day, time, size.
- **Matplotlib:** is an easy-to-use, low-level data visualization library that is built on NumPy arrays. It consists of various plots like scatter plot, line plot, histogram, etc. Matplotlib provides a lot of flexibility.
- **Scatter Plot:** Scatter plots are used to observe relationships between variables and uses dots to represent the relationship between them.
- **Line Chart:** is used to represent a relationship between two data X and Y on a different axis.
- **Bar Chart** is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent.
- **Histogram:** is used to represent data in the form of groups. It is a type of bar plot where the X-axis represents the binary ranges while the Y-axis gives information about frequency.
- **Boxplot:** is used to represent distribution of a dataset. It provides a summary of key statistical measures, allowing you to visualize the central tendency, spread, and identify potential outliers in the data.

Exercise



Select the best answer for the following Multiple-Choice Questions (MCQs).

1. Supervised learning is one of the main category of _____.
a) Statistical modeling b) Data waggling
c) Data collection d) Evaluation modeling
2. For accurate data collection, an existing data from _____ is acquired.
a) Sampling department b) Modeling techniques
c) Meteorological department d) Educational department
3. Natural disaster prediction helps the people to take preventive measures against _____.
a) Pandemic b) Cyclones
c) Data analysis d) Temperature
4. A linear regression model is a mathematical equation that allows us to predict a response for given _____.
a) Predictor value b) Variable
c) Intercepts d) Model
5. A point where graph line meet the y-axis is called _____.
a) Y-axis b) Y-intercept
c) Y-coordinates d) Y-dimensions
6. Clustering algorithms are examples of _____.
a) Classical learning b) Professional learning
c) Unsupervised learning d) Cognitive learning
7. We can create our own database by generating some random number by using _____.
a) Python code b) Data code
c) System code d) Pseudo code
8. During summer people of hot regions consume more cold items. This information is called _____.
a) Parameter b) Population
c) Statistics d) Random sample
9. If there are some flaws in data collection method, after analyzing such data the result would be more _____.
a) Accurate b) Huge
c) No result can be drawn c) Erroneous
10. Experimental design is a tool used to _____ the result effectively.
a) Add b) Evaluate
c) Organize d) Test



Give short answers to the following Short Response Questions (SRQs).

1. List out the parameters and statistics from given statements:
 - a) Average length of height of a giraffe.
 - b) Average weight of watermelon.
 - c) There are 430 doctors in a hospital.
 - d) Average age of students of 6th class in a school is 12 years.
 - e) The number of a basketball team players having height above 6 feet.
2. If you want to make a report regarding the products exported from Pakistan in last five years, how libraries can help you to collect data? Write steps.
3. Make a pie chart of vegetable prices in the market. Consider five to ten vegetables.
4. Enlist steps to represent the monthly temperatures of a Pakistani city in 2023 from January till December using a line graph.



Give long answers to the following Extended Response Questions (ERQs).

1. Simulate on paper, an experimental design for awareness of food security (Narrative Visualization).
2. Sketch primary data collection methods in context of disease outbreak, like seasonal flu.
3. Argue about the use of statistical modeling techniques. Highlight all techniques discussed in this unit.
4. Compare linear regression and classification. Emphasize on their respective roles in statistical modeling.
5. Defend either of supervised learning and unsupervised learning. Give reasons for your preference to the other.
6. Write a Python code to generate a dataset with two variables where $y = x^2 + 2x$. Fit scatter plot and box plot on this data.
7. Relate some real world examples (other than Airbnb, Facebook and YouTube) where data science was used to improve marketing strategies and enhance the business.



Activity 1

Practice all the Python code given in this Unit.

Objective: To help students develop the skill of data analysis by using Python.

Outcome: This activity will help students understand the basic syntax of Python for data analysis.



Activity 2

Data collection by interview method:

Objective: To help students understand the process of data collection through survey and questionnaires.

- Divide the students into small groups(not more than 3 students in each group).
- Ask the students to collect data by interviewing their fellow students, about their academic performance, study schedule, outdoor games etc.
- Ask them to organize data and plot them by using their favorite tool.
- Supervise them to draw conclusions about correlation among various factors and academic performance.
- Various groups will compare their results/conclusions and find the similarities and differences.

Outcome: This activity will help students understand positive or negative correlation among academic and not academic activities. They will also develop the skill of conducting interview as an interviewee as well as interviewer.



Activity 3

Build more statistical models on new datasets.

Objective: To help students develop the skill of data analysis by using Python and other tools.

- Ask the students to search and download some dataset of their choice from the internet.
- Perform the steps of data cleaning and feature selection.
- Visualize the data and create various graphs according to the nature of dataset. Students can use other tools like MS-Excel, Weka, Pytorch, R studio etc.
- Supervise and help them finding easy and simple tutorials on internet, students can later share small video tutorial through Whatsapp.

Outcome: This activity will help students learn new tools at their own.