Muhammad Sikandar Dar, IMCB, G-11/1, Islamabad. 0333-5813882

UNIT NO. 04
DATA AND ANALYSIS

Q.No.01: List out the parameters and statistics from given statements:
Ans:

a) Average length of height of Giraffe.
Ans: This represents a parameter. A parameter is a value that describes a characteristic of an entire population. In this case, it refers to the average height of all giraffes**.**

b) Average weight of Watermelon.
Ans: This is also a parameter. It describes the average weight of all watermelons in the population.

c) There are 430 Doctors in Hospital.
Answer: This is a statistic. A statistic is a value that describes a characteristic of a sample, in this case, the number of doctors in a specific hospital.

d) Average age of students of 6th class in a school is 12 years.
Ans: This is a statistic. It describes the average age of students in a specific class, which is a sample of the entire population of students.

e) The number of a basket ball team player having height above 6 feet.
Answer: This is a statistic. It refers to the number of players in a specific team, which is a sample of all possible basketball players.

Q.No.02: If you want to make report regarding the products exported from Pakistan in last five years, how libraries can help you to collect data? Write steps.
Ans: Libraries can provide access to a variety of resources for data collection:

Access to Research Papers: Libraries house academic journals and research papers that contain detailed studies and analyses on export trends.

Business Directories and Annual Reports: Libraries maintain collections of business directories and annual reports that can provide historical data on export statistics.

Government Publications: Libraries often have archives of government publications and reports that include statistical data on exports.

Digital Repositories: Many libraries offer access to digital repositories and databases that can be searched for relevant data on exports.

Interlibrary Loan Services: If a particular resource is not available, libraries can use interlibrary loan services to obtain materials from other institutions.

Q.No.03: Make a Pie chart of vegetable prices in the market. Consider five to ten vegetables.

Ans: Collect Data: Gather the prices of five to ten different vegetables from the market.

Organize Data: Create a list or table that includes the names of the vegetables and their corresponding prices.

Choose a Tool: Use a data visualization tool such as Matplotlib in Python.

Create the Chart:

Import the required libraries (e.g., matplotlib.pyplot).

Define the data for the vegetables and their prices.

Use the pie() function to create the pie chart.

Customize the chart with labels, title, and colors.

Display the Chart: Use the show() function to display the pie chart.

Example Python Code:

```
import matplotlib.pyplot as plt
# Data
vegetables= ['Tomato', 'Potato', 'Carrot', 'Onion', 'Pepper']
prices= [30, 20, 15, 25, 10]
# Creating pie chart
plt.pie (prices, labels=vegetables, autopct='%1.1f%%')
# Adding title
plt.title('Vegetable Prices in the Market')
# Display the chart
plt.show()
```

Q.No.04: Enlist steps to represent the monthly temperature of a Pakistani city in 2023 from January till December using line chart.

Ans: Collect Temperature Data: Gather the average monthly temperature data for the city from January to December.

Organize Data: Create a list or table that includes the months and their corresponding average temperatures.

Choose a Tool: Use a data visualization tool such as Matplotlib in Python.

Create the Line Graph:

Import the required libraries (e.g., matplotlib.pyplot).

Define the data for the months and their temperatures.

Use the plot () function to create the line graph.

Customize the graph with labels, title, and grid lines.

Display the Graph: Use the show () function to display the line graph.

Example Python Code:

```
import matplotlib.pyplot as plt
# Data
months = ['Jan','Feb','Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec']

temperatures = [10, 12, 15, 20, 25, 30, 35, 34, 28, 22, 15, 11]
# Creating line graph
plt.plot (months, temperatures, marker='o')
# Adding title and labels
plt.title('Monthly Temperatures in 2023')
plt.xlabel('Months')
plt.ylabel('Temperature (°C)')
# Display the graph
plt.show()
```

Q.No.05: Define statistical modeling.

Ans: Statistical modeling is the process of applying statistical techniques to analyze data. It involves creating a model that represents the relationships between two or more variables. Statistical modeling is used to understand these relationships, draw meaningful conclusions and make predications about real world situations. For example, statistical modeling can be applied to sales data from the past few years to predict sales for the upcoming year.

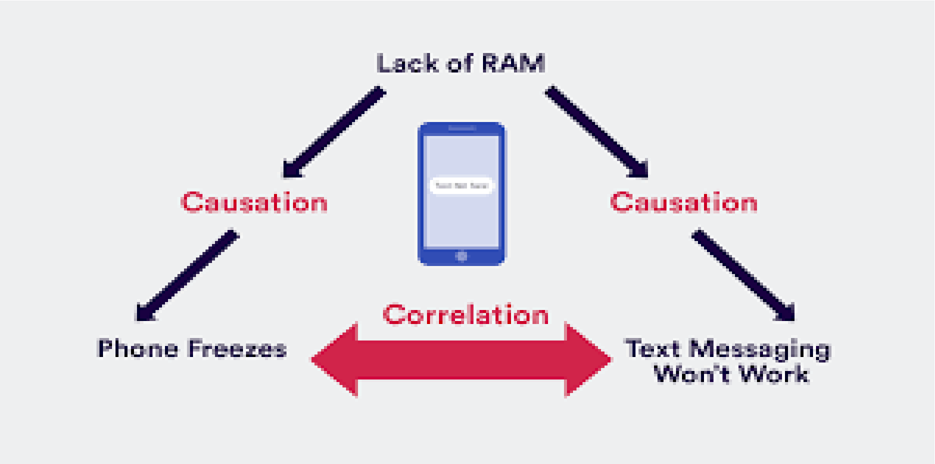
Q.No.06: List the names of statistical tools used to perform predictive analysis?

Ans: Following are some tools which are used for predictive analysis:

MS Excel, Weka, R Studio, and Python.

Q.No.07: Define Correlation and Causation.

Ans: While causation and correlation can exist simultaneously, correlation does not imply causation. Causation means one thing causes another—in other words, action A causes outcome B. On the other hand, correlation is simply a relationship where action A relates to action B—but one event doesn’t necessarily cause the other event to happen.

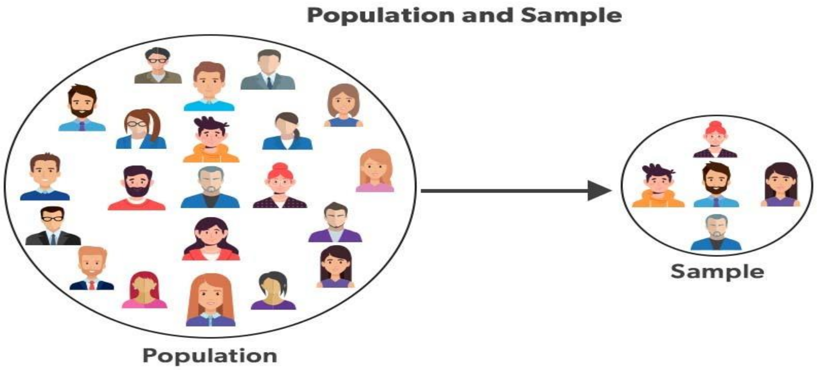


Q.No.08: Differentiate between Population and Sample.

Ans: A population is the entire group that you want to draw conclusions about.

A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

In research, a population doesn't always refer to people. It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms, etc.



Q.No.09: Differentiate between Parameters and Statistics.

Ans: The difference between parameter and statistics is as follows:

Parameter	Statistics
It refers to a numerical value that describe the characteristics of population	It refers to a numerical value that describes a characteristics of sample.
It is used to describe the whole population accurately.	It is used to estimate the population parameter based on sample data.

It is typically unknown as it is not possible to measure the entire population.	It is known when the sample data is available.
Examples: is the average income of all citizens in Pakistan; Mean weight of Persian cats; Proportion of all people who prefer soft drinks.	Examples: is the average income of government employees in Pakistan; Mean weight of 100 Persian cats; Proportion of random sample of 500 persons who prefer soft drinks.

Q.No.10: List down the primary data collection methods.

Ans: Following are most common primary data collection methods:

Interviews, Observations, Surveys and Questionnaires, Focus groups and Oral histories.

Q.No.11: List down the secondary data collection methods.

Ans: Following are most common secondary data collection methods:

Internet, Government Archives and Libraries.

Q.No.12: Which tools are used for data visualization?

Ans: Following tools are used for data visualization:

Charts, graphs, boxplots etc.

Q.No.13: Define data, analytics, insights, data visualization, data trends, data pattern and outliers.

Ans: Data: data is a collection of facts, representing measurements or descriptions of a specific situation. These facts can be in the form of numbers, symbols or words and are typically stored digitally.

Analytics: Analytics is the use of tools and processes to combine and examine sets of data to identify patterns, relationships and trends. The goal of analytics is to answer specific questions, discover new insights, and help organizations make better, data-driven decisions.

Insights: Data insights refers to the deep understanding an individual or organization gains from analyzing information on a particular issue. This deep understanding helps organizations make better decisions than by relying on gut instinct.

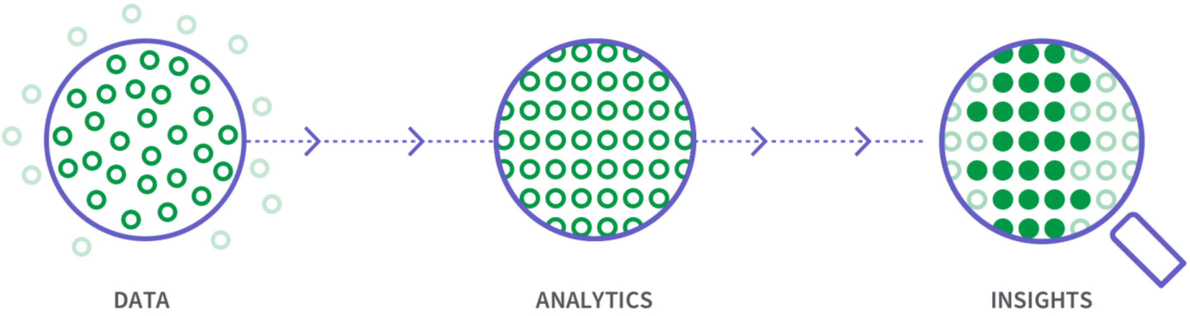
The differences become clear when we crystalize the definitions:

Data = a collection of facts.

Analytics = organizing and examining data.

Insights = discovering patterns in data.

There's also a linear aspect to these terms that differentiates them. Data is collected and organized, then analysis is performed, and insights are generated as follows:



Data Visualization: Data visualization translates complex data sets into visual formats that are easier for the human brain to comprehend. This can include a variety of visual tools such as:

- **Charts**: Bar charts, line charts, pie charts, etc.
- **Graphs**: Scatter plots, histograms, etc.
- **Maps**: Geographic maps, heat maps, etc.
- **Dashboards**: Interactive platforms that combine multiple visualizations.

The primary goal of data visualization is to make data more accessible and easier to interpret, allowing users to identify patterns, trends, and outliers quickly. This is particularly important in the context of big data.

Data Trends: In data analysis, a trend refers to the underlying direction or tendency of the data over a specific time period. It represents a consistent pattern of increasing, decreasing, or stable behavior observed in the data. Trends can be crucial in uncovering meaningful insights, predicting future outcomes, and making informed decisions. For example, in financial markets, analyzing trends helps identify potential investment opportunities or market risks. Similarly, in climate science, understanding temperature trends over time provides insights into global warming patterns.

Data Patterns: Data patterns refer to the recurring structures or behaviors found in the data. They can exhibit either regular or irregular characteristics. Regular patterns repeat in a predictable manner, such as daily or weekly seasonal variations in sales data. Irregular patterns, on the other hand, represent random or unpredictable fluctuations. These patterns can reveal hidden relationships, detect anomalies, and guide strategic planning. For instance, detecting sudden spikes in website traffic patterns can help businesses troubleshoot technical issues.

Additional Information

Pattern — repeats in a recognizable way
Trend — the general direction in which something is developing
Outlier — unlikely or rare event
Anomaly — a result that can't be explained
Distribution — how your data is spread out
Relationship — how two or more variables interact
Comparison — examining data sets side-by-side

Q.No.14: What is the purpose of following libraries i.e Pandas, Matplotlib, Numpy and Sklearn.

Ans: Pandas: "Pandas" is a contraction of the words "Panel" and "Data," but it is also a contraction of the term "Python Data Analysis."

Panel Data is a form of multidimensional data that logs the behaviors of multiple subjects over multiple time periods or points in time.

Python Data Analysis is basically any form of analysis that's being streamlined by Python-based tools.

So, the name says a lot about Pandas' function, which is to make quick work of messy data, clarifying and organizing it for relevance, and deleting NULL values as needed.

Python Pandas library provides two primary data structures, DataFrame and Series. These streamline the processes of tabular data management for both textual and numerical data, including:

data loading
data tabulating
data cleaning
data filling
NULL data deletion
data normalization
data inspection
statistical data analysis
data saving, and more.

Matplotlib: Matplotlib is a powerful plotting library in Python used for creating static, animated, and interactive visualizations. Matplotlib's primary purpose is to provide users with the tools and functionality to represent data graphically, making it easier to analyze and understand. It was originally developed by John D. Hunter in 2003 and is now maintained by a large community of developers.

Pandas allows for efficient and flexible numerical data and textual data handling and, when you combine Pandas module with other, complementary Python modules, it streamlines all aspects of data cleaning, manipulation, and analysis.

Scikit-learn: is an open-source Python library that implements a range of machine learning, pre-processing, cross-validation, and visualization algorithms using a unified interface. It is an open-source machine-learning library that provides a plethora of tools for various machine-learning tasks such as Classification, Regression, Clustering, and many more.

The latest version of Scikit-learn is 1.1 and it requires Python 3.8 or newer.

Scikit-learn requires:

- NumPy
- SciPy as its dependencies.

Before installing scikit-learn, ensure that you have NumPy and SciPy installed.

Numpy: NumPy stands for **Numerical Python,** is an open-source Python library that provides support for large, multi-dimensional arrays and matrices.

It also have a collection of high-level mathematical functions to operate on arrays. It was created by Travis Oliphant in 2005.

Q.No.15: Why and where Experimental design is used?

Ans: Experimental design is a tool or technique used to organize , conduct and clarify the results of an experiment efficiently. Experimental design is used in many disciplines like engineering, psychology, agriculture and medicine etc.

Q.No.16: What is Use Case?

Ans: A Use Case is a technique in system analysis that describes how a system will be used to achieve a specific goal or solve a real world problem. It is a list of actions or events steps typically defining the interaction between a role(UML as actor) and a system to achieve a goal.

Q.No.17: What is the first step in solving a data science case study?

Ans: The first step is formulating the right question that involves reviewing the available literature in order to understand the business problem and translate it into a clear question.

Q.No.18: What is the use of Data Wrangling?

Ans: Data wrangling is the process of cleaning, structuring, and transforming raw data into a usable format for analysis. Also known as data munging, it involves tasks such as handling missing or inconsistent data, formatting data types, and merging different datasets to prepare the data for further exploration and modeling in data analysis or machine learning projects.

Q.No.19: What is Machine learning?

Ans: Machine Learning, a subset of AI, crafts algorithms and statistical models to empower computers to learn from data and make decisions autonomously without direct programming. It encompasses employing mathematical and statistical methodologies to train models on data and subsequently utilizing these models for prediction or decision-making tasks.

Q.No.20: Differentiate between supervised and unsupervised learning.

Ans: Supervised learning:

1. You train the model with a set of input data and a corresponding set of paired labeled output data.
2. Predict an output based on known inputs.
3. Minimize the error between predicted outputs and true labels.

Unsupervised Learning:

1. You train the model to discover hidden patterns in unlabeled data.
2. Identify valuable relationship information between input data points. This can then be applied to new input to draw similar insights.
3. Find patterns, similarities, or anomalies within the data.

Q.No.21: What is linear regression?

Ans: Linear regression refers to supervised learning models that, based on one or more inputs, predict a value from a continuous scale. An example of linear regression is predicting a house price. You could predict a house's price based on its location, age, and number of rooms, after you train a model on a set of historical sales training data with those variables.

Q.No.22: When to use: supervised vs. unsupervised learning?

Ans: You can use supervised learning techniques to solve problems with known outcomes and that have labeled data available. Examples include email spam classification, image recognition, and stock price predictions based on known historical data.

You can use unsupervised learning for scenarios where the data is unlabeled and the objective is to discover patterns, group similar instances, or detect anomalies. You can also use it for exploratory tasks where labeled data is absent. Examples include organizing large data archives, building recommendation systems, and grouping customers based on their purchasing behaviors.

Q.No.23: What does the equation y=mx+b represent in linear regression.
Ans: In linear regression, **y=mx+b** represents the equation of a line where **y** is the dependent variable, **x** is the independent variable, m is the slope of the line and b is the **y** intercept.

Q.No.24: What are the roles of the intercept and slope in a linear regressions model?
Ans: The intercept (y-intercept) is the point where the line crosses the y-axis and the slope represents the rate of change in the dependent variable with respect to the independent variable.

Q.No.25: What is a classification model in statistical modelling?
Ans: A classification model is used when the result is a discrete value such as predicting whether an employee will receive a salary raise or not. It categorizes data into predefined categories.

Q.No.26: How does classification differ from regression in statistical modelling?
Ans: Regression models are used to predict continuous values such as predicting a person’s salary or age. The output is a number that can take any value within a range. Classification models are used to predict discrete values such as yes/no or true/false. For example, classifying an email as spam or not spam. The output is a category or class label not a number.

Q.No.27: What does clustering mean in the context of unsupervised learning?
Ans: Clustering involves grouping data items based on their similarities. For example, customer may be clustered into groups based on their usage patterns such as long call duration or heavy internet usage.

Q.No.28: What is the purpose of association rules in unsupervised learning?
Ans: Association is a method that identifies the relationships or associations among a set of items within large datasets. It identifies the combinations of items that often occur together. For example, an association exists among bread, milk and butter. If a customer buys bread, he may also buy butter and milk. Many retailers and e-commerce platforms often use this method to recommend products to their customers to increase sales.

Q.No.29: What is the K-means clustering algorithm?
Ans: K-means clustering is a popular clustering algorithm. It combines a specified number of data points into specific groups based on similarities.

Q.No.30: What is the use of surveys and questionnaires?
Ans: Surveys and questionnaires are used to gather information from a large group of people quickly and efficiently. They can be conducted face-to-face, by post or over 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.

Q.No.31: What role do libraries play in secondary data collection?
Ans: The libraries have a large collection of important and authentic information on different topics. They also have business directories, annual reports and other similar documents that help businesses in their research.

Q.No.32: How does Airbnb use A/B testing?
Ans: Airbnb uses statistical experimentation such as A/B testing to improve its platform and user experience. For example, when testing a new feature such as a new booking process or search method, Airbnb creates two different versions. They randomly show each version to different users. They collect and analyze data on how users interact with each versions to determine which one performs better.

Q.No.33: What is Google Colab and how is it used in statistical modelling with Python?
Ans: Google Colab is an online platform that provides a Python environment for coding and executing Python scripts. It is used for building and testing statistical models by allowing users to write and run Python code in a browser.

Q.No.34: How to predict house prices by using linear regression using python code.
Ans: Following is code for linear regression:
Before diving into code you must have a dataset in CSV format to create model then another CSV file which is used to predict prices on the basis of area…

```
from sklearn import linear_model
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
# import libraries for linear regression
df=pd.read_csv('prediction.csv')
# # TO load CSV file from your project (Dataset for model)
# #print(df)
# # To check file is properly loaded or not?
plt.xlabel("Area (square fee)")
plt.ylabel("Price (US $)")
plt.scatter(df.area,df.price,color="red")
#plt.plot(df.area,reg.predict(df[['area']]),color='blue')
reg=linear_model.LinearRegression()
reg.fit(df[['area']],df.price)
# #plt.show()
print(reg.predict([[5000]]))
#print(predictedvalue)
#print(reg.coef_)    get slope value
#print(reg.intercept_) get y-intercept value
# # Another file in which there is only areas
da=pd.read_csv('areas.csv')
#print(da)
p=reg.predict(da[['area']])
da['prices']=p
da.to_csv('areas.csv',index=False)
```

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

Q.No.01: Simulate on paper, an experimental design for awareness of food security(Narrative Visualization).

Ans: Research Question:

The primary research question for this experimental design is, "What are the key factors affecting food security awareness among different demographics, and how can narrative visualization be used to enhance this awareness?"

Develop Hypotheses:

1. Income level significantly affects food security awareness.

2. Education level significantly impacts food security awareness.

3. Geographic location influences food security awareness.

Identify Variables:

- **Independent Variables:** Income level (low, medium, high), education level (primary, secondary, tertiary), and geographic location (urban, rural).
- **Dependent Variable:** Level of food security awareness, measured through a survey.

Determine Experimental Design:

A factorial design will be used to study the interaction between different independent variables (income, education, geographic location) and their impact on food security awareness. This design allows the examination of main effects and interactions between factors.

Sample Size Calculation:

Use a power analysis to determine the necessary sample size to detect significant effects. This involves specifying the expected effect size, desired power (typically 0.80), and significance level (typically 0.05). For simplicity, let's assume a sample size of 300 participants is required, evenly distributed across different demographics.

Random Assignment:

Randomly assign participants to groups based on income level, education level, and geographic location to eliminate bias and ensure a representative sample.

Q.No.02: Sketch primary data collection methods in context of disease outbreak, like seasonal flu.

Ans: The primary data collection methods in context of disease outbreak like flu can be as follows:

1. **Surveys and Questionaries:**

   The purpose of survey and questionnaire is to collect information on symptoms, treatment and health behaviors from individuals. For example, the surveys can be distributed to individuals in affected areas to gather data on flu symptoms, vaccination history and healthcare utilization.

2. **Interviews:**

   The purpose of interviews is to obtain detailed information from patients, healthcare providers and public health officials. For example, the interviews can be conducted with patients to understand their flu symptoms, timeline and contact with others.

3. **Observations:**

   The purpose of observation is to monitor and record real-time data on flu cases and response efforts. For example, the researcher can observe and document the number of patients arriving at clinics with flu-like symptoms.

4. **Medical Records Review:**

   The purpose of reviewing medical record is to analyze patient records for trends and patterns related to the flu outbreak. For example, the hospital records can be reviewed to identify the number of flu cases, severity and outcomes.

5. **Field Reports:**

   The purpose of checking the field reports is to gather data from field teams on the ground in affected areas. For example, field reports can be collected from public health workers about the spread of flu, vaccination rates and community response.

6. **Focus Groups:**

   The purpose of focus group is to gather qualitative data from groups of individuals about their experiences and perceptions of the flu outbreak. For example, focus groups can be held with community members to discuss their concerns and experiences with the flu.

Q.No.03: Argue about the use of statistical modeling techniques. Highlight all techniques discussed in the Book.

Ans:Introduction:

Statistical modeling techniques are fundamental tools in data science used to analyze data, make predictions, and derive insights. These techniques transform raw data into meaningful information, guiding decision-making across various domains.

Key Statistical Modeling Techniques:

1. **Regression Models:**

- **Linear Regression:** A technique used to predict a continuous outcome based on one or more predictor variables. The relationship is modeled through the equation y=mx+by = mx + by= mx+b, where yyy is the dependent variable, xxx is the independent variable, mmm is the slope, and bbb is the y-intercept. Linear regression is essential for understanding trends, forecasting, and making quantitative predictions.
- **Multiple Regression:** An extension of linear regression that involves multiple predictor variables. It is useful for modeling complex relationships and predicting outcomes based on several factors simultaneously.

.

2. **Classification Models:**

- **Logistic Regression:** Used to predict binary outcomes (e.g., yes/no, success/failure). It models the probability of a certain class or event. Logistic regression is widely used in fields like medical diagnosis, fraud detection, and marketing.
- **Decision Trees:** A non-parametric model used for classification and regression tasks. It splits data into subsets based on the value of input features, creating a tree-like structure. Decision trees are interpretable and can handle both numerical and categorical data.
- **Support Vector Machines (SVM):** A powerful classification technique that finds the optimal hyperplane separating different classes in the feature space. SVM is effective in high- dimensional spaces and is used in image recognition, text classification, and bioinformatics.

3. **Clustering Algorithms:**
    - **K-Means Clustering:** An unsupervised learning algorithm that partitions data into K clusters based on similarity. It minimizes the variance within each cluster. K-

means is useful for market segmentation, customer analysis, and pattern recognition.

- **Hierarchical Clustering**: Builds a hierarchy of clusters using either an agglomerative (bottom- up) or divisive (top-down) approach. It provides a tree-like structure called a dendrogram. Hierarchical clustering is beneficial for exploratory data analysis and identifying natural groupings in data.

4. Association Rules:
   - **Apriori Algorithm:** Used to find frequent item sets and generate association rules in transactional datasets. It helps in market basket analysis, discovering patterns of co-occurrence among items. For example, it can identify that customers who buy bread often buy milk as well.

Q.No.04: Compare linear regression and classification. Emphasize on their respective roles in statistical modeling.

Ans: Linear regression: Data analysts use **regression models** to examine relationships between variables. Regression models are often used by organizations to determine which independent variables hold the most influence over dependent variables—information that can be leveraged to make essential business decisions.

“The most traditional regression models that have been used for a long time are logistic regression, linear regression, and polynomial regression,” Mello says. “These are the most common.”

Other examples of regression models can include stepwise regression, ridge regression, lasso regression, and elastic net regression.

Classification is a process in which an algorithm is used to analyze an existing data set of known points. The understanding achieved through that analysis is then leveraged as a means of appropriately classifying the data. Classification is a form of machine learning that can be particularly helpful in analyzing very large, complex sets of data to help make more accurate predictions.

“Classification models are a form of supervised machine learning which is often used when the analyst needs to understand how they got to a certain point,” Mello says. “They give you more than just an output; [they give you] more information that you can use to explain the results of the prediction to your boss or stakeholder.”

Some of the most common classification models include decision trees, random forests, nearest neighbor, and Naive Bayes.

There are also the neural networking models that are more used in AI. "These are very powerful models, and they can make accurate predictions very well," Mello says, "but you typically cannot explain what is happening behind the scenes."

Q.No.05: Define either supervised learning and unsupervised learning. Give reasons for your preference to the other.

Ans: Supervised Learning:

Definition: Supervised learning involves training a model on labeled data, where the outcome is known. The model learns to map input features to the correct output based on the provided labels. Common techniques include regression (linear and logistic) and classification (decision trees, support vector machines).

Advantages of Supervised Learning:

1. **Predictive Accuracy:**
    - **Precision:** Supervised learning models are trained on labeled data, allowing them to make precise predictions. For example, a supervised learning model can predict house prices based on features like size, location, and number of bedrooms with high accuracy.
    - **Error Minimization:** The model adjusts its parameters during training to minimize prediction errors, leading to better performance.

Preference for Supervised Learning:

While unsupervised learning has its strengths, such as discovering hidden patterns and groupings in unlabeled data, supervised learning is preferred for several reasons:

1. Predictive Power:

Supervised learning models can achieve high predictive accuracy, making them suitable for tasks requiring precise outcomes. For example, in predicting loan defaults, supervised learning models can provide accurate risk assessments based on historical data.

Unsupervised Learning:

Definition: Unsupervised learning involves training a model on unlabeled data, where the outcome is unknown. The model identifies patterns and structures within the data independent Common techniques include clustering (K-means, hierarchical) and association (Apriori algorithm).

Advantages of Unsupervised Learning:

1. **Discovering Hidden Patterns:**
    - **Clustering:** Unsupervised learning can uncover natural groupings within data, such as customer segments with similar purchasing behaviors. This is valuable for exploratory data analysis and market segmentation.
    - **Dimensionality Reduction**: Techniques like Principal Component Analysis (PCA) reduce data dimensionality, making it easier to visualize and interpret high-dimensional datasets.

Both supervised and unsupervised learning are essential in machine learning, each serving unique purposes. However, supervised learning is often preferred for its predictive accuracy, interpretability and wide applicability. The availability of labeled data allows for precise predictions, thorough validation, and actionable insights, making supervised learning a reliable choice for many real-world applications. While unsupervised learning excels in discovering hidden patterns and adapting to new data, the controlled and interpretable nature of supervised learning provides a significant advantage in making data-driven decisions.

Q.No.06: 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.

Ans:

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

# Generate dataset
x = np.linspace (-10, 10, 100)
y=x**2 + 2*x# Generate 100 points between -10 and 10 y = x**2 +2*x #Calculate y based on the equation y = x^2 + 2x
# Create a DataFrame
data = pd.DataFrame({'x': x, 'y': y})
# Scatter plot
plt.figure(figsize=(10, 5))
plt.scatter (data['x'], data['y'], color='blue')
plt.title ('Scatter Plot of y = x^2 + 2x')
plt.xlabel('x')
plt.ylabel('y')
plt.grid(True)
plt.show()
# Box plot
plt.figure (figsize=(10, 5))
plt.boxplot (data['y'], vert=False)
plt.title('Box Plot of y')
plt.xlabel('Value of y')
plt.grid(True)
plt.show()
```

Q.No.07: Relate some real world examples (other than Airbnb, Facebook and YouTube) where data science used to improve marketing strategies and enhance the business.

Ans:

Q.No.08: Explain random and relevant functions of NumPy in python.

Ans: Following are some important functions used in unit 4:

numpy.random.rand(): is a function in the NumPy library used to generate random numbers. It returns an array of specified shape with random values uniformly distributed between 0 and 1.

Example:

```
# Get 1-dimensional array of random values
import numpy as np
arr = np.random.rand(6)
print("After getting a 1D array of random values:\n",arr)
# Output:
# After getting a 1D array of random values:
#  [0.54019408 0.50597572 0.18926709 0.27325964 0.6813522  0.63123887]
```

Seed function: function is used to seed the random number generator in NumPy. Seeding is important for reproducibility. By setting the seed, you ensure that the sequence of random numbers generated by NumPy is the same every time you run your code with that seed. In the below example, **numpy.random.seed(5)** sets the seed for NumPy's random number generator to the integer value 5. Setting the seed is crucial for reproducibility. If you run this code multiple times, you will get the same sequence of random numbers each time because the seed is fixed. **numpy.random.random(5)** generates an array of 5 random numbers from a uniform distribution between 0 (inclusive) and 1 (exclusive). The result is assigned to the variable arr.

Example:
```
# Import numpy
import numpy as np
# Generate random numbers
np.random.seed(5)
arr = np.random.random(5)
print("The random numbers are:\n",arr)
```

Numpy.random.randint(): is a function in the NumPy library used to generate random integers. It allows you to generate random integers between specified low (inclusive) and high (exclusive) values. The generated integers can be used for various purposes such as simulations, random sampling, and generating random data for testing and analysis.

Example:
```
# Get the random integers of array
import numpy as np
arr = np.random.randint(low=1, high=6, size=4)
print("After generating random integers of the array:\n",arr)
```

numpy.random.randn(): is a numpy library function that returns an array of random samples from the standard normal distribution. It allows specified dimensions as an argument and returns an array of specified dimensions. If you don't provide any argument, it will return the float value. The **np.random.randn()**function returns all the samples in float form, which are from the univariate "normal" (Gaussian) distribution of mean 0 and variance 1.

The dimensions of the returned array must be non-negative. If you provide a negative argument, then it will return an error.

The function is used to generate a single random value from a standard normal distribution. Since no size argument is provided, the result is a scalar (a single float). You might get different random numbers when you run the same code multiple times.

Example:
```
# Import numpy module
import numpy as np
# Use random.randn() function
arr = np.random.randn(3)
```

```
print("Random normal distribution:\n",arr)
```

Standard normal distribution: The standard normal distribution is one of the forms of the normal distribution. It occurs when a normal random variable has a mean equal to zero and a standard deviation equal to one. In other words, a normal distribution with a mean 0 and standard deviation of 1 is called the standard normal distribution. Also, the standard normal distribution is centered at zero, and the standard deviation gives the degree to which a given measurement deviates from the mean.

The random variable of a standard normal distribution is known as the **standard score or a z-score**. It is possible to transform every normal random variable X into a z score using the following formula:

$$\mathbf{z = (X - \mu) / \sigma}$$

where X is a normal random variable, μ is the mean of X, and σ is the standard deviation of X. You can also find the normal distribution formula here. In probability theory, the normal or Gaussian distribution is a very common continuous probability distribution.

Probability distribution: A probability distribution is an idealized frequency distribution.

A frequency distribution describes a specific sample or dataset. It’s the number of times each possible value of a variable occurs in the dataset.

The number of times a value occurs in a sample is determined by its **probability** of occurrence. Probability is a number between 0 and 1 that says how likely something is to occur:

- 0 means it’s impossible.
- 1 means it’s certain.

The higher the probability of a value, the higher its frequency in a sample.

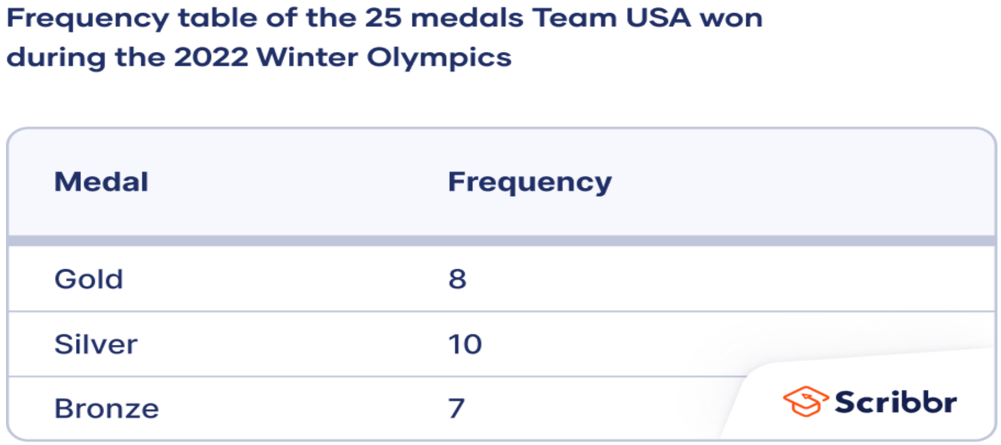
More specifically, the probability of a value is its relative frequency in an infinitely large sample.

Infinitely large samples are impossible in real life, so probability distributions are theoretical. They’re idealized versions of frequency distributions that aim to describe the population the sample was drawn from.

Probability distributions are used to describe the populations of real-life variables, like coin tosses or the weight of chicken eggs. They’re also used in hypothesis testing to determine *p* values.

Frequency distribution: A frequency distribution describes the number of observations for each possible value of a variable. Frequency distributions are depicted using graphs and frequency tables.

Example: Frequency distributionIn the 2022 Winter Olympics, Team USA won 25 medals. This frequency table gives the medals' values (gold, silver, and bronze) and frequencies:



Frequency table of the 25 medals Team USA won during the 2022 Winter Olympics

Medal	Frequency
Gold	8
Silver	10
Bronze	7

Scribbr

Q.No.09: Discuss Trends, Patterns and Outliers.

Ans: Line graphs are an important tool in visualizing data. They are widely used across different fields, ranging from science to finance. Line graphs represent data over time, displaying trends, patterns, and outliers. Interpreting line graphs is crucial to understanding data and making informed decisions. In this section, we will discuss the process of interpreting line graphs and how to identify trends, patterns, and outliers.

1. Understanding Trends: Trends are the general direction in which data is moving. In a line graph, trends are identified by the slope of the line. If the line is going up, then the trend is positive, while if the line is going down, then the trend is negative. If the line is horizontal, then there is no trend. For example, if we plot the sales of a product over time, we may observe *an increasing trend*. This indicates that sales are growing over time.

2. Identifying Patterns: Patterns are the recurring characteristics in data. In line graphs, patterns can be identified by the shape of the line. For example, if the line has a series of peaks and valleys, then there may be *a repeating pattern*. Patterns can also be identified by the distance between the peaks and valleys. If the distance between the peaks is the same, then there is a regular pattern. For instance, if we plot the temperature of a city over the course of a year, we may observe *a repeating pattern* with higher temperatures in *the summer and lower temperatures* in the winter.

3. Detecting Outliers: Outliers are data points that are significantly different from the rest of the data. They can be identified in line graphs by points that are far away from the general trend or pattern. Outliers can occur due to errors in data collection or real-world events that affect the data. For example, if we plot the stock price of a company over time, we may observe a sudden drop in the price due to *a major event* like a recession.

Interpreting line graphs is an essential skill in data analysis. Understanding trends, patterns, and outliers can provide valuable insights into the data and help make informed decisions. Whether you are analyzing financial data or studying *scientific experiments*, line graphs can provide a clear picture of the data over time.

Q.No. 10: Explain Boxplot with the help of sample data.

A box plot gives a five-number summary of a set of data which is-

- Minimum – It is the minimum value in the dataset excluding the outliers.
- First Quartile (Q1) – 25% of the data lies below the First (lower) Quartile.
- Median (Q2) – It is the mid-point of the dataset. Half of the values lie below it and half above.
- Third Quartile (Q3) – 75% of the data lies below the Third (Upper) Quartile.
- Maximum – It is the maximum value in the dataset excluding the outliers.

The area inside the box (50% of the data) is known as the Inter Quartile Range. The IQR is calculated as –

IQR = Q3-Q1

Outlies are the data points below and above the lower and upper limit. The lower and upper limit is calculated as –

Lower Limit = Q1 - 1.5*IQR

Upper Limit = Q3 + 1.5*IQR

The values below and above these limits are considered outliers and the minimum and maximum values are calculated from the points which lie under the lower and upper limit.

How to create a box plots?

Let us take a sample data to understand how to create a box plot.

Here are the runs scored by a cricket team in a league of 12 matches – *100, 120, 110, 150, 110, 140, 130, 170, 120, 220, 140, 110.*

To draw a box plot for the given data first we need to arrange the data in ascending order and then find the minimum, first quartile, median, third quartile and the maximum.

Ascending Order

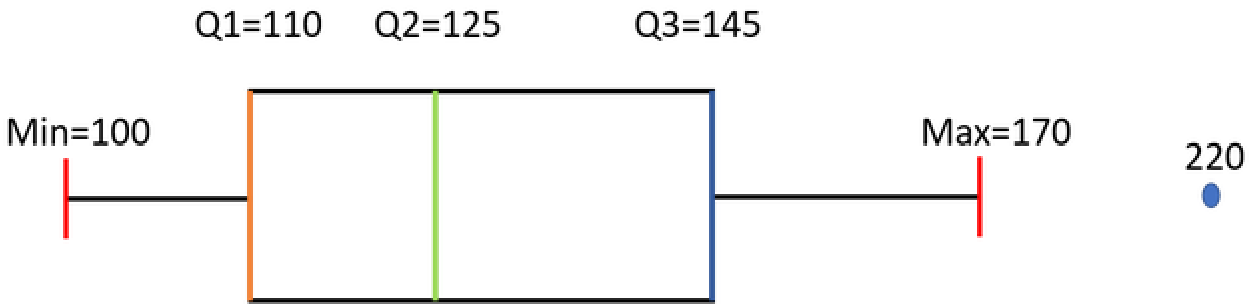
100, 110, 110, 110, 120, 120, 130, 140, 140, 150, 170, 220

Median (Q2) = (120+130)/2 = 125; Since there were even values

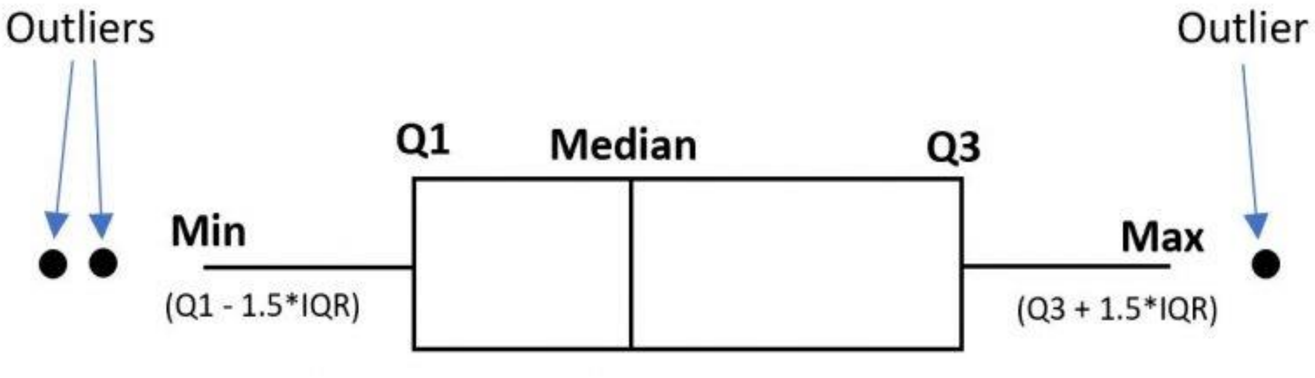
To find the First Quartile we take the first six values and find their median.
Q1 = (110+110)/2 = 110
For the Third Quartile, we take the next six and find their median.
Q3 = (140+150)/2 = 145
Note: If the total number of values is odd then we exclude the Median while calculating Q1 and Q3. Here since there were two central values we included them. Now, we need to calculate the Inter Quartile Range.
IQR = Q3-Q1 = 145-110 = 35
We can now calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any.
Lower Limit = Q1-1.5*IQR = 110-1.5*35 = 57.5
Upper Limit = Q3+1.5*IQR = 145+1.5*35 = 197.5

So, the minimum and maximum between the range [57.5,197.5] for our given data are –
Minimum = 100
Maximum = 170

The outliers which are outside this range are –
Outliers = 220
Now we have all the information, so we can draw the box plot which is as below-



We can see from the diagram that the Median is not exactly at the center of the box and one whisker is longer than the other. We also have one Outlier.



- **Lower outlier threshold:** A data point is a lower outlier if it is less than Q1 - 1.5 * IQR
- **Upper outlier threshold:** A data point is an upper outlier if it is greater than Q3 + 1.5 * IQR

Q.No. 11: Explain Histogram with the help of sample data.

Ans: A histogram is one of the most commonly used graphs to show the frequency distribution. As we know that the frequency distribution defines how often each different value occurs in the data set. The histogram looks more similar to the bar graph, but there is a difference between them.

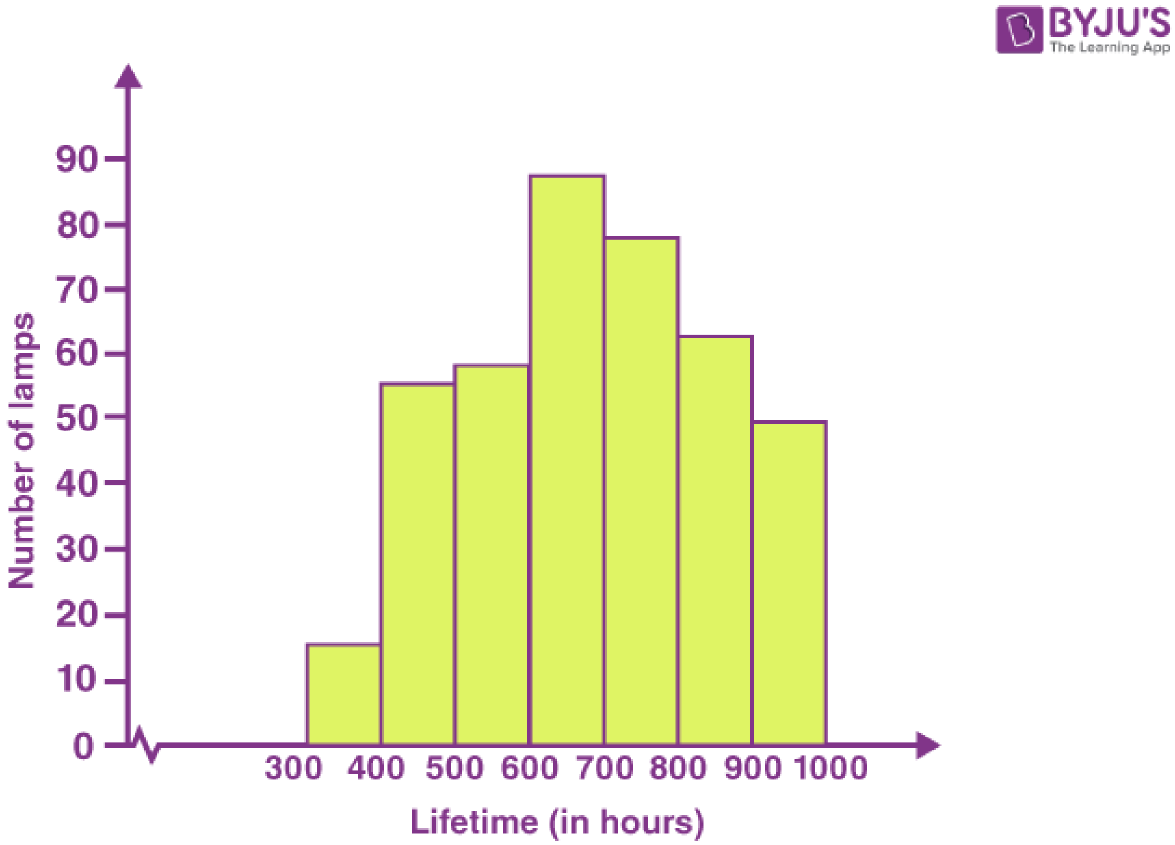
The following table gives the lifetime of 400 neon lamps. Draw the histogram for the below data.

Lifetime (in hours)	Number of lamps
300 – 400	14
400 – 500	56
500 – 600	60
600 – 700	86
700 – 800	74

800 – 900	62
900 – 1000	48

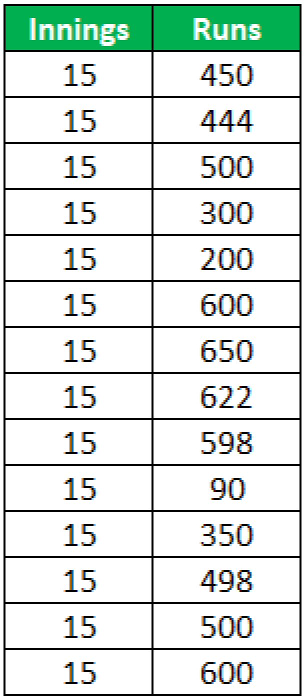
Solution:

The histogram for the given data is:



Another Example:

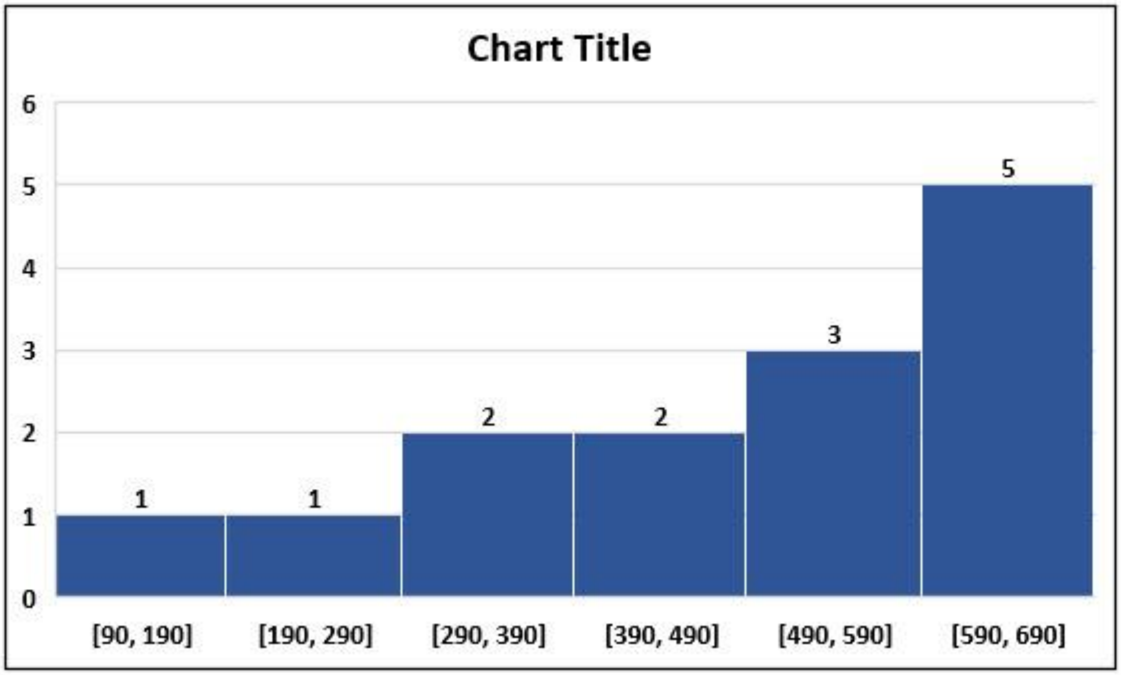
Shastri, the coach of an Indian cricket team, is analyzing batters' average scores and wants to finalize the chosen batters for the upcoming world cup. However, he first wants to create a benchmark to shortlist the batters. He has received a list of below batters in their last 15 innings; however, he wants to know the odd one from this list. Use the histogram to find out and comment on the distribution.



Innings	Runs
15	450
15	444
15	500
15	300
15	200
15	600
15	650
15	622
15	598
15	90
15	350
15	498
15	500
15	600

Solution:

We have created a histogram using 6 bins with 6 different frequencies, as seen in the chart below. The Y-axis shows the number of batters falling in that particular category. In addition, on the X-axis, we have a range of runs. For example, the 1st bin range is 90 to 190. We can note that the count is 1 for that category from the table, as seen in the below graph.



We can see that the above table shows a left-skewed distribution. That is because many data values occur on the right side and a smaller number of data on the left side.

90 runs in 15 innings appear to be the odd one out and appear to be of a bowler and hence need to be removed.

Q.No. 12: What do you know about central tendency?

Ans: List of numbers: 4, 10, 7, 15, 2. Calculate the median.

Solution: Let us arrange the numbers in ascending order.

In ascending order, the numbers are: 2,4,7,10,15

There are a total of 5 numbers. Median is (n+1)/2th value. Thus, the Median is (5+1)/2th value.

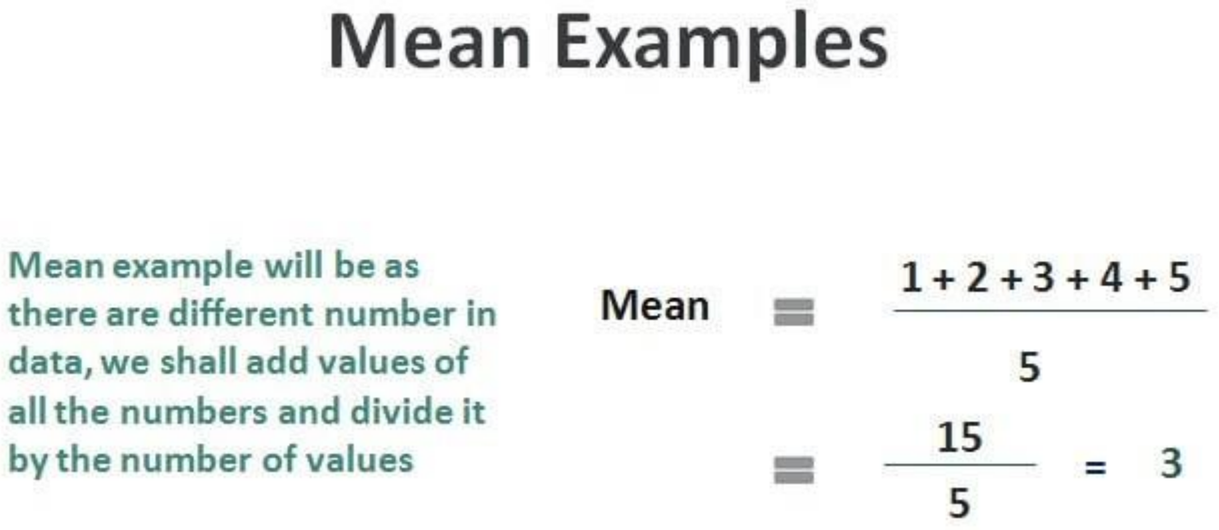
Median = 3rd value.

The 3rd value in list 2, 4, 7, 10, 15 is 7.

Thus, the median is 7.

How to find Mode in statistics?

Hence to find the Mode in statistics, one must determine that item in a set has occurred the highest number of times in the data set. For example, if the students are taking admission to history in large numbers concerning other subjects, then history is the Mode of all the subjects.



Mean Examples

Mean example will be as there are different number in data, we shall add values of all the numbers and divide it by the number of values

$$\text{Mean} = \frac{1+2+3+4+5}{5}$$

$$= \frac{15}{5} = 3$$

Q.No. 13: Explain Pandas library with the help of Data frame and Series.

Ans: Pandas is a Python library used for working with data sets.
It has functions for analyzing, cleaning, exploring, and manipulating data.
Pandas allows us to analyze big data and make conclusions based on statistical theories.
Pandas can clean messy data sets, and make them readable and relevant.
Relevant data is very important in data science.

What is a Series?

A Pandas Series is like a column in a table.
It is a one-dimensional array holding data of any type.

Example 1:

```
import pandas as pd
a = [1, 7, 2]
myvar = pd.Series(a)
print(myvar)
```

Example: 2

```
import pandas as pd
calories = {"day1": 420, "day2": 380, "day3": 390}
myvar = pd.Series(calories)
print(myvar)
```

Data Frames

Data sets in Pandas are usually multi-dimensional tables, called Data Frames.
Series is like a column, a Data Frame is the whole table.

Example:1

```
import pandas as pd
data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}
myvar = pd.DataFrame(data)
print(myvar)
```