Unit 3
Statistical Analysis:Descriptive Statistics

-By Bhagyashree Udagave(Chougule)

Introduction

The first step of any data-related process is the collection of data. Once we have collected the data, what do we do with it?

Data can be sorted, analyzed, and used in various methods and formats, depending on the project's needs. While analyzing a dataset, We use statistical methods to arrive at a conclusion..

Two types of statistical methods are widely used in data analysis: descriptive and inferential.

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Summarize and describe data	Draw conclusions or predictions
Data Sample	Analyzes the entire dataset	Analyzes a sample of the data
Examples	Mean, Median, Range, Variance	Hypothesis testing, Regression
Scope	Focuses on data characteristics	Makes inferences about populations
Goal	Provides insights and simplifies data	Generalizes findings to a larger population
Assumptions	No assumptions about populations	Requires assumptions about populations
Common Use Cases	Data visualization, data exploration	Scientific research, hypothesis testing

Descriptive Statistics

It involves organizing, visualizing, and summarizing raw data to create a coherent picture.

The primary goal of descriptive statistics is to provide a clear and concise overview of the data's main features.

This helps us to identify patterns, trends, and characteristics within the data set without making broader inferences.

Key Aspects of Descriptive Statistics

- **Measures of Central Tendency:** Descriptive statistics include calculating the mean, median, and mode, which offer insights into the center of the data distribution.
- **Measures of Dispersion:** Variance, standard deviation, and range help us understand the spread or variability of the data.
- **Visualizations:** Creating graphs, histograms, bar charts, and pie charts visually represent the data's distribution and characteristics.

Types of Descriptive Statistics

1. **Descriptive Statistics Based on the Central Tendency of Data**

   The central tendency of data is the center of the distribution of data.

   It describes the location of data and concentrates on where the data is located.

   The three most widely used measures of the “center” of the data are **Mean, Median, and Mode.**

Mean

The “Mean” is the average of the data. The average can be identified by summing up all the numbers and then dividing them by the number of observations.

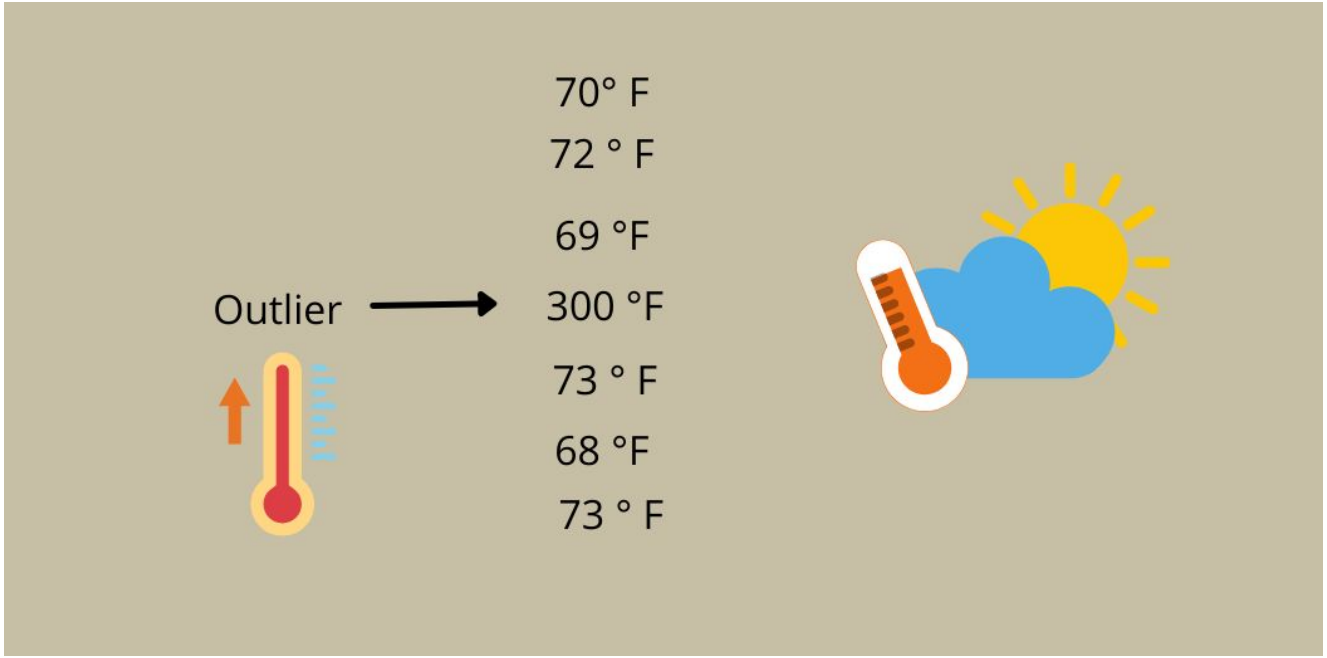
Mean = $X_1 + X_2 + X_3 + \ldots + X_n / n$

Data – 10,20,30,40,50 and Number of observations = 5

Mean = [10+20+30+40+50] / 5

Mean = 30

The central tendency of the data may be influenced by outliers.



E.g. Data – 10,20,30,40,200
Mean = [10+20+30+40+200] / 5
Mean = 60

Solution for the outliers problem: Removing the outliers while taking averages will give us better results.

Median

It is the 50th percentile of the data.

In other words, it is exactly the center point of the data.

The median can be identified by ordering the data, splitting it into two equal parts, and then finding the number in the middle. It is the best way to find the center of the data.

Note that, in this case, the central tendency of the data is not affected by outliers.


MEDIAN

Example:

Odd number of Data – 10,20,30,40,50

Median is 30.

Even the number of data – 10,20,30,40,50,60
Find the middle 2 data and take the mean of those two values.
Here, 30 and 40 are middle values.

30+40 / 2 =35
Median is 35

Mode

The mode of the data is the most frequently occurring data or elements in a dataset.

If an element occurs the highest number of times, it is the mode of that data.

If no number in the data is repeated, then that data has no mode.

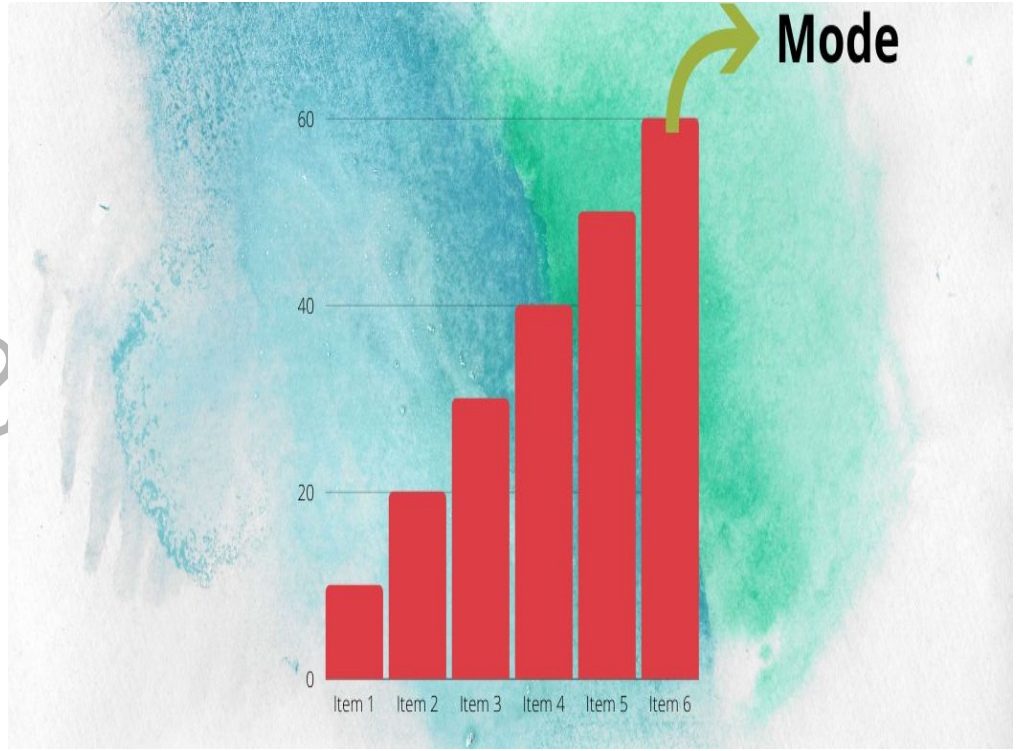
There can be more than one mode in a dataset if two values have the same frequency, which is also the highest frequency.

Outliers don’t influence the data in this case. The mode can be calculated for both quantitative and qualitative data.

Example:

Data – 1,3,4,6,7,3,3,5,10, 3

Mode is 3, because 3 has the highest frequency (4 times)



2. Descriptive Statistics Based on the Dispersion of Data

The dispersion is the “spread of the data”.

It measures how far the data is spread. In most of the dataset, the data values are closely located near the mean.

The values are widely spread out of the mean on some other datasets. These dispersions of data can be measured by the Interquartile Range (IQR), range, standard deviation, and variance of the data.

1. Interquartile Range (IQR)

Quartiles are special percentiles.

1st Quartile Q1 is the same as the 25th percentile.

2nd Quartile Q2 is the same as 50th percentile.

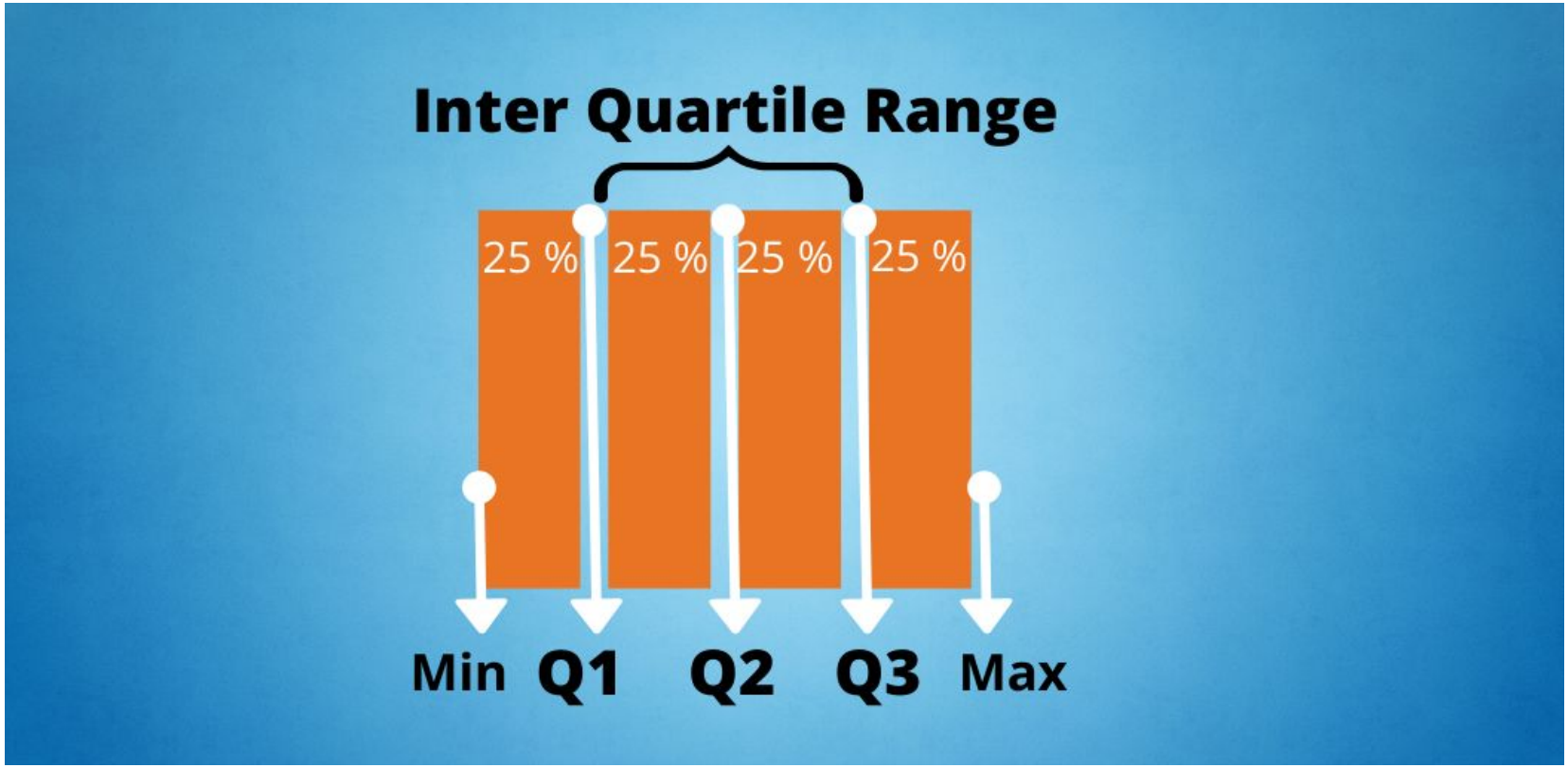
3rd Quartile Q3 is same as 75th percentile

Steps to find quartile and percentile

- The data should sorted and ordered from the smallest to the largest.
- For Quartiles, ordered data is divided into 4 equal parts.
- For Percentiles, ordered data is divided into 100 equal parts.

The Interquartile Range is the difference between the third quartile (Q3) and the first quartile (Q1)

IQR = Q3 – Q1


Inter Quartile Range
25 %
25 %
25 %
25 %
Min
Q1
Q2
Q3
Max

2. Range

The range is the difference between the largest and the smallest value in the data.

3. Standard Deviation

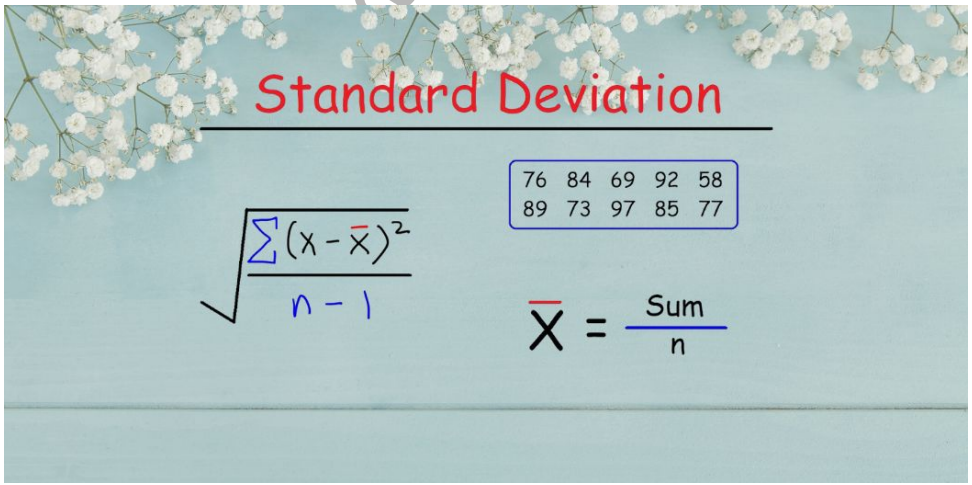
The most common measure of spread is the standard deviation. The Standard deviation measures how far the data deviates from the mean value.

- Sample Standard Deviation – “s”
- Population Standard Deviation – “σ”

Steps to find the Standard Deviation

If x is a number, then the difference “x – mean” is its deviation. The deviations are used to calculate the standard deviation.

Sample Standard Deviation, s = Square root of sample variance

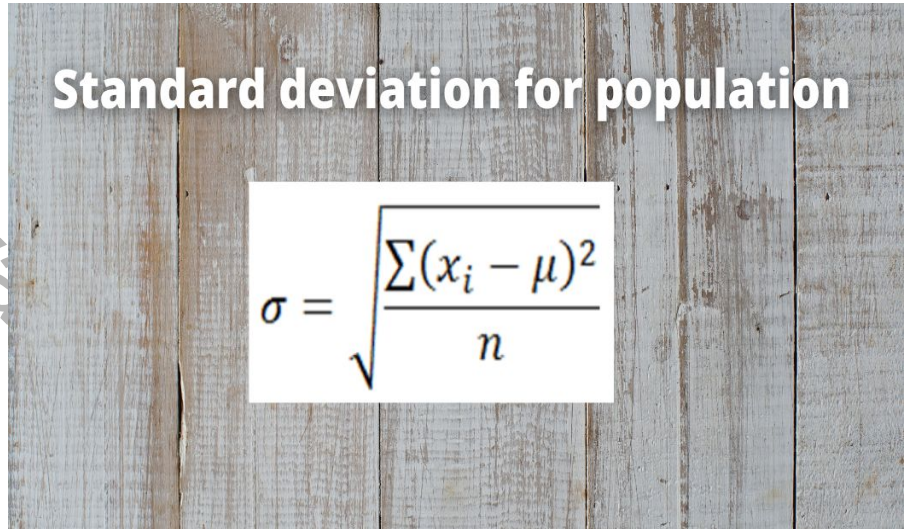


Population Standard Deviation,

σ = Square root of population variance

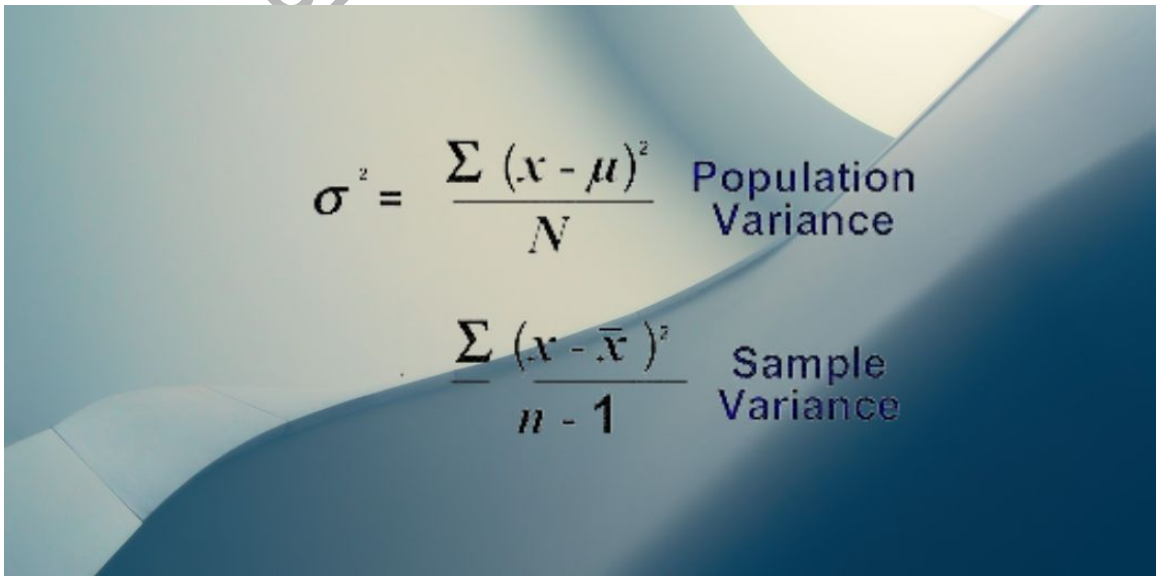
The standard deviation is always positive or zero.

It will be large when the data values are spread out from the mean.



Variance

The variance is a measure of variability. It is the average squared deviation from the mean.



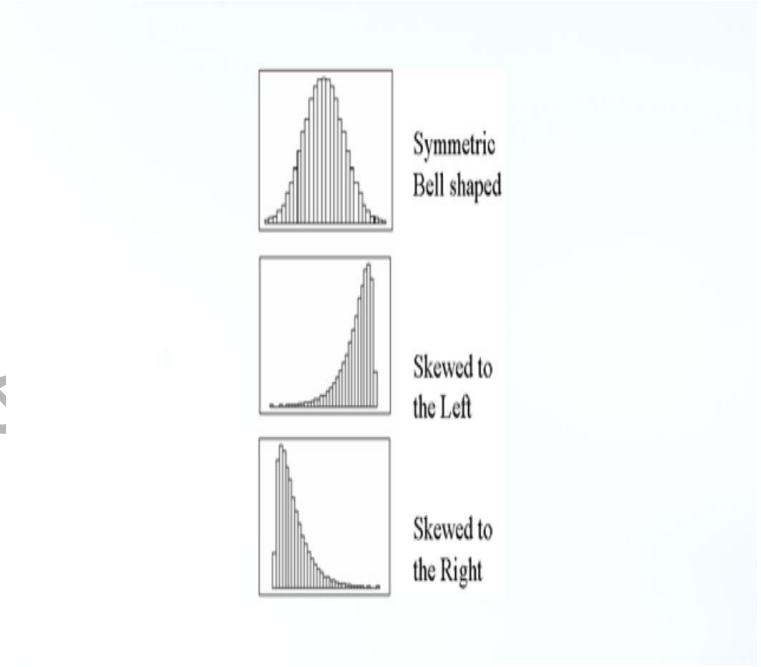
Correlation

Refer PDF given

Descriptive Statistics Based on the Shape of the Data

The shape of the data can be measured by three methodologies: symmetric, skewness, kurtosis. The shape of the data is important because deciding the probability of data is based on its shape.

1. **Symmetric**
   In the symmetric shape of the graph, the data is distributed the same on both sides.
   In symmetric data, the mean and median are located close together.
   The curve formed by this symmetric graph is called a normal curve.

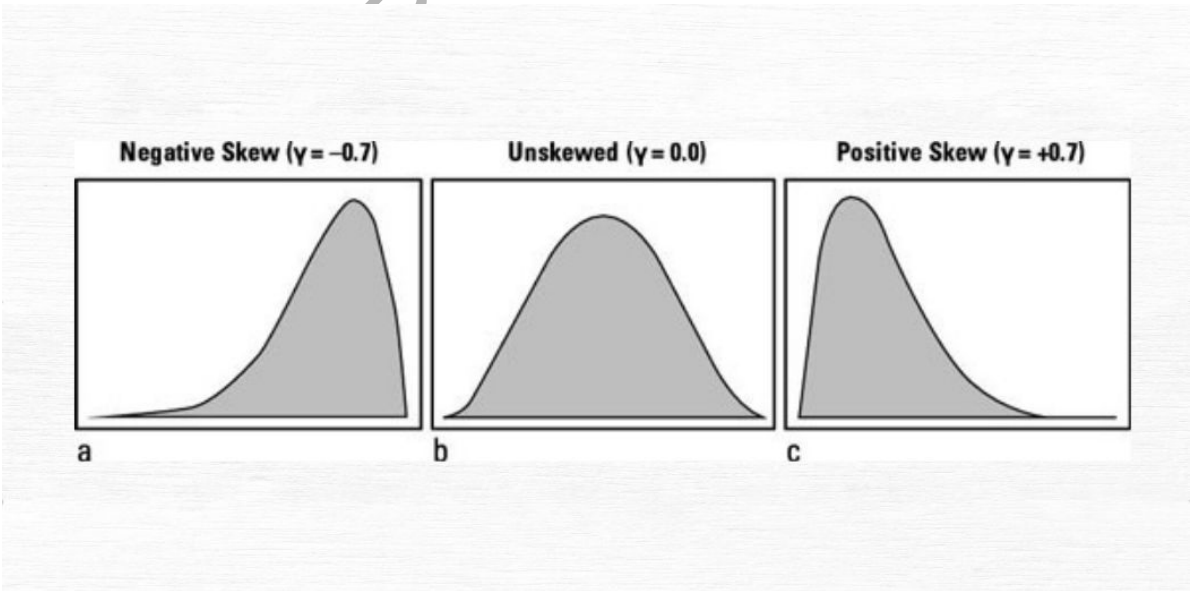


2. Skewness

Skewness is the measure of the asymmetry of the distribution of data. The data is not symmetrical (i.e.) it is skewed towards one side. Skewness is classified into two types: positive skew and negative skew.

- **Positively skewed:** In a Positively skewed distribution, the data values are clustered around the left side of the distribution, and the right side is longer. The mean and median will be greater than the mode in the positive skew.

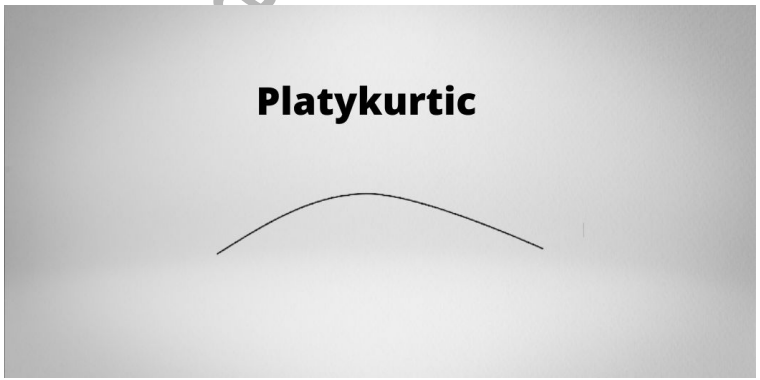
- **Negatively skewed:** In a Negatively skewed distribution, the data values are clustered around the right side of the distribution, and the left side is longer. The mean and median will be less than the mode.



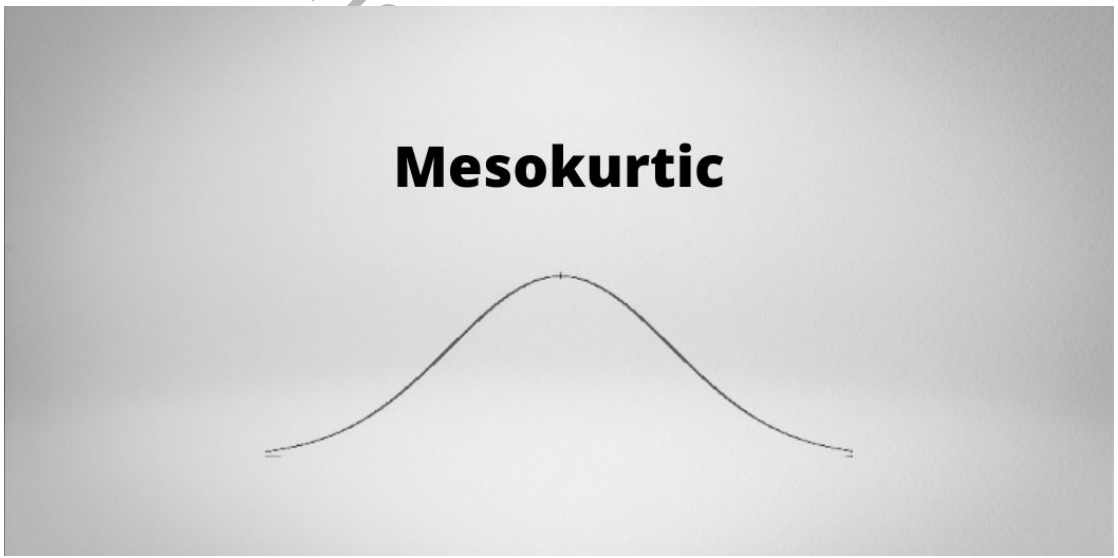
3. Kurtosis

Kurtosis is the measure of describing the distribution of data. This data is distributed in three different ways: platykurtic, mesokurtic, and leptokurtic.

- Platykurtic: The platykurtic shows a distribution with flat tails. Here, the data is distributed fairly. The flat tails indicated the small outliers in the distribution.

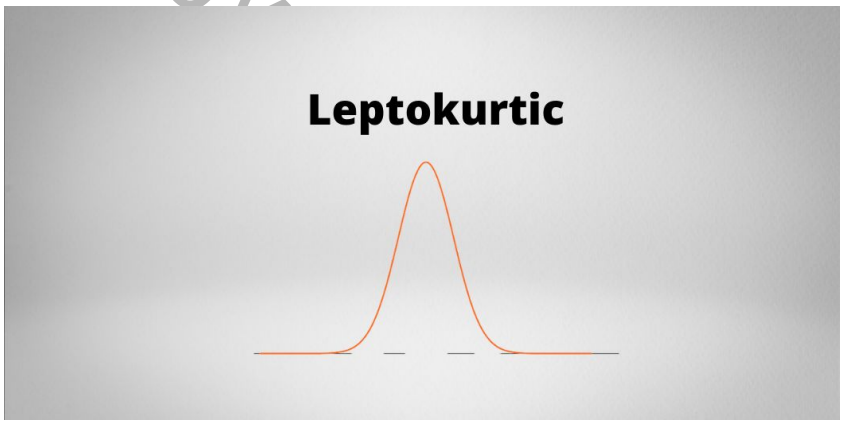


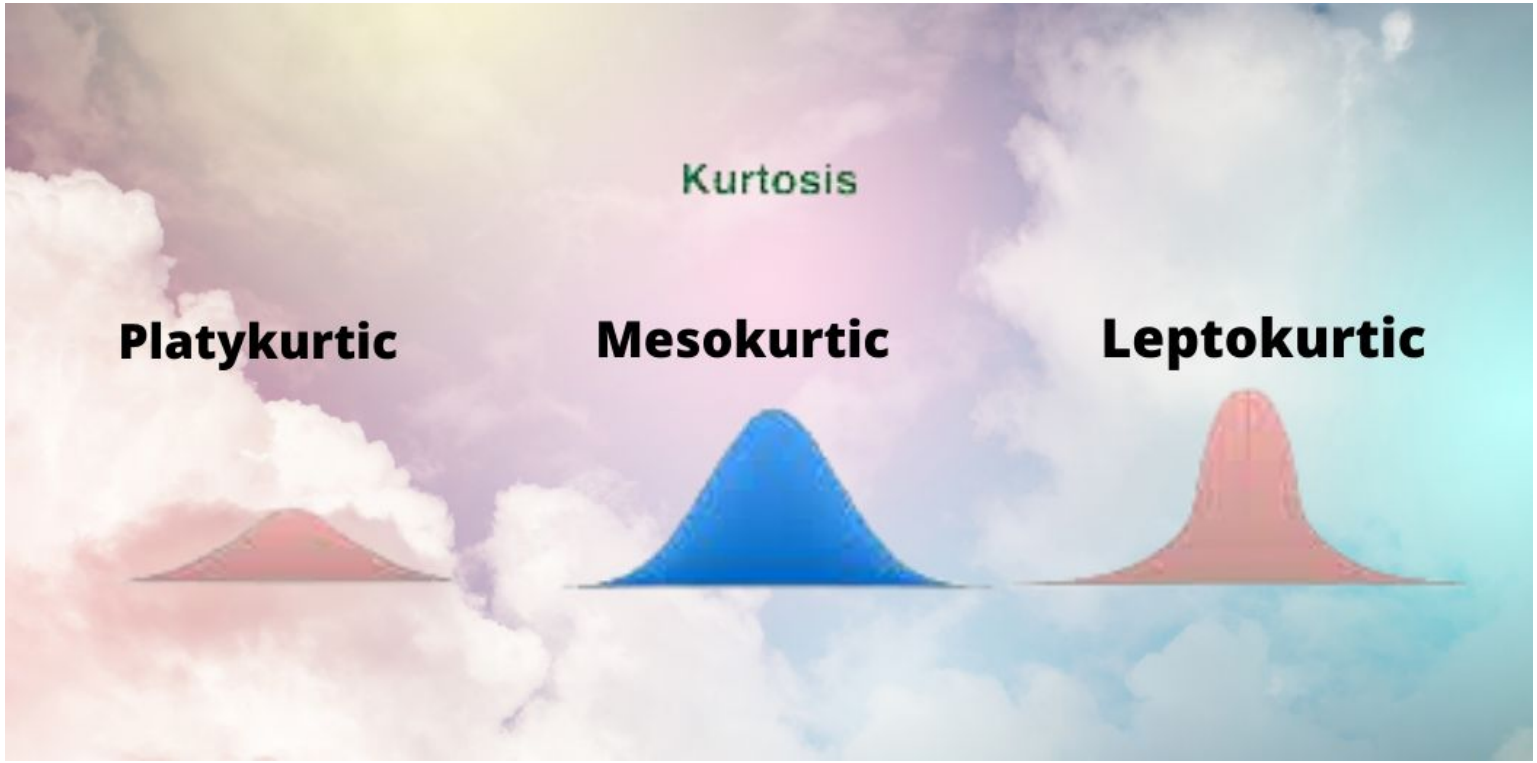
- Mesokurtic: In Mesokurtic, the data is widely distributed. It is normally distributed, and it also matches normal distribution.



- **Leptokurtic:**

  In leptokurtic, the data is very closely distributed. The height of the peak is greater than the width of the peak.




Kurtosis
Platykurtic
Mesokurtic
Leptokurtic

Data Visualization Principles

Incredibly, around 2.5 quintillion bytes of new data are generated every day, and this number is increasing day by day.

One effective method for you to understand complex data in an easier manner is through data visualization.

Data visualization:

Data visualization involves making sense of rows and columns of data by presenting it in an easily understandable format.

Data may, therefore, be represented by pictures, charts, or graphs, to make it easy-to-understand or to identify new patterns.

By using data visualization, you may also be able to recognize relationships and patterns between certain parameters and discover emerging trends as well.

Key principles of effective data visualization:

1. **Determine the best visual**

   To begin with, it is imperative to understand the volume of data in hand. You may then identify the aspects that you wish to visualize along with the information that you wish to convey. After this, you may select the best-suited and simplest visual format for your target audience.

2. **Balance the design**

This refers to equally distributed visual elements across the plot such as texture, color, shape, and negative space. You may select the visual to be symmetrical, asymmetrical or radial, and figure out the right balance of elements that work best for visualizing your data.

3. Focus on the key areas

Ensure that the key areas are well- highlighted. You may place key data points towards the top-left corner as a user's attention is generally drawn towards that quadrant.

4. Keep it simple

Ensure that your visuals are simple and easy-to-understand. Adding unwanted information may make it confusing, which defeats the purpose of data visualization. Keep in mind that the ultimate goal of data visualization is simplicity. You may build additional visuals if you wish to convey a multi-faceted story.

5. Incorporate interactivity

You may deploy data visualization tools that infuse interactivity into your graphs or charts. However, ensure that this does not confuse the target audience since the main purpose is to clarify doubts or queries.

6. Use patterns

You may display similar types of information like one with the help of patterns. You may establish a pattern by using similar chart types, colors, or any other element.

7. Compare aspects

You may display a side-by-side comparison of aspects to make understanding of data easier. You may also align data either horizontally or vertically so that it can be compared accurately.

Data Visualization Techniques

Data visualization provides an important suite of tools for identifying a qualitative understanding.

This can be helpful when we try to explore the dataset and extract some information to know about a dataset and can help with identifying patterns, corrupt data, outliers, and much more.

Data visualization is defined as a **graphical representation** that contains the **information** and the **data**.

By using visual elements like **charts, graphs, and maps**, data visualization techniques provide an accessible way to see and **understand trends, outliers, and patterns in data.**

The basic uses of the Data Visualization technique are as follows:

- It is a powerful technique to explore the data with **presentable** and **interpretable** results.
- In the **data mining process**, it acts as a primary step in the pre-processing portion.
- It supports the **data cleaning process** by finding incorrect data and corrupted or missing values.
- It also helps to **construct and select variables**, which means we have to determine which variable to include and discard in the analysis.
- In the process of **Data Reduction**, it also plays a crucial role while combining the categories.

Univariate Analysis Techniques for Data Visualization

1. Distribution Plot
2. Box and Whisker Plot
3. Violin Plot

Bivariate Analysis Techniques for Data Visualization

1. Line Plot
2. Bar Plot
3. Scatter Plot

Refer link for explanation:

https://www.analyticsvidhya.com/blog/2021/06/must-known-data-visualization-techniques-for-data-science/