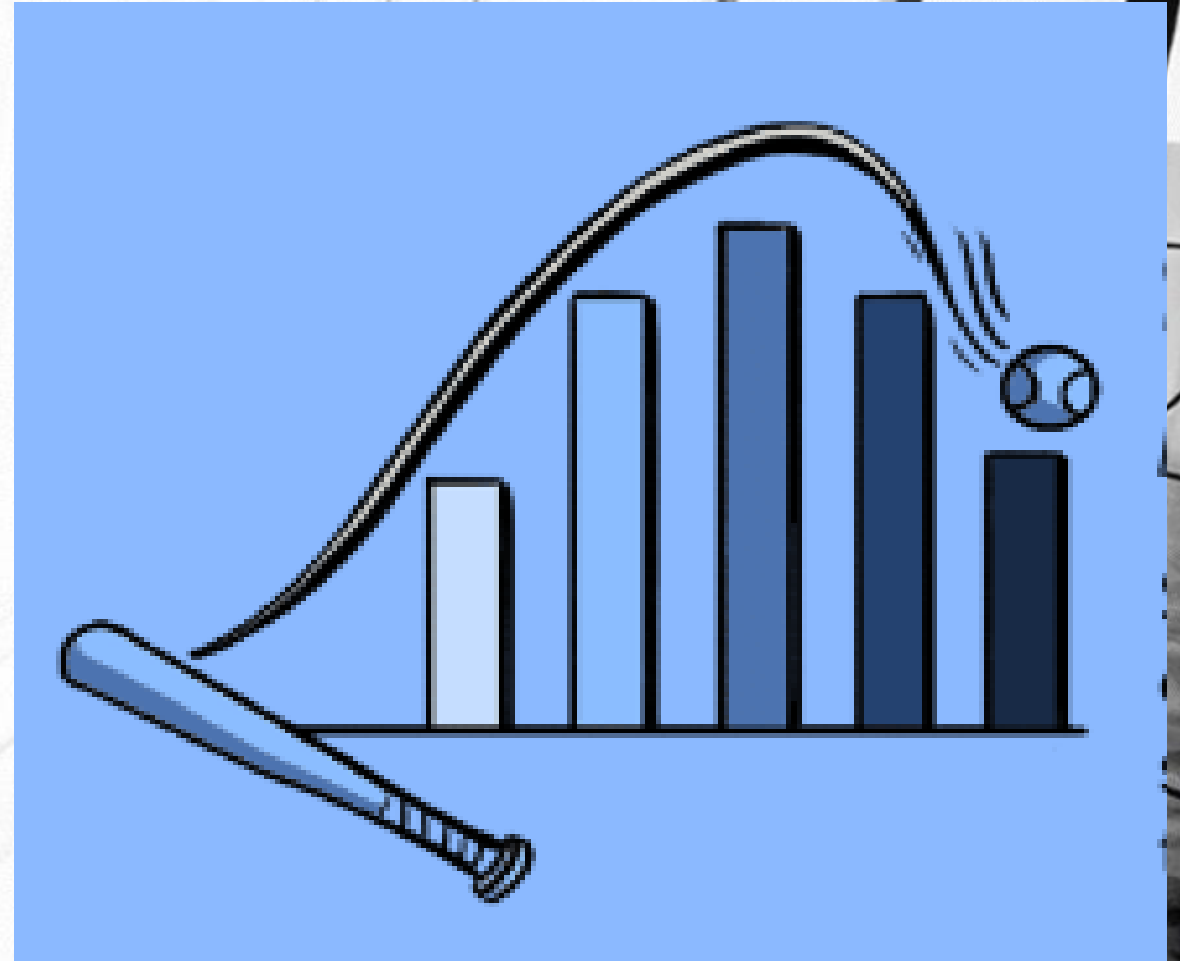


UNIT 3: DESCRIPTIVE STATISTICS

- ❑ Descriptive statistics are **brief informational coefficients that summarize a given data set**, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).



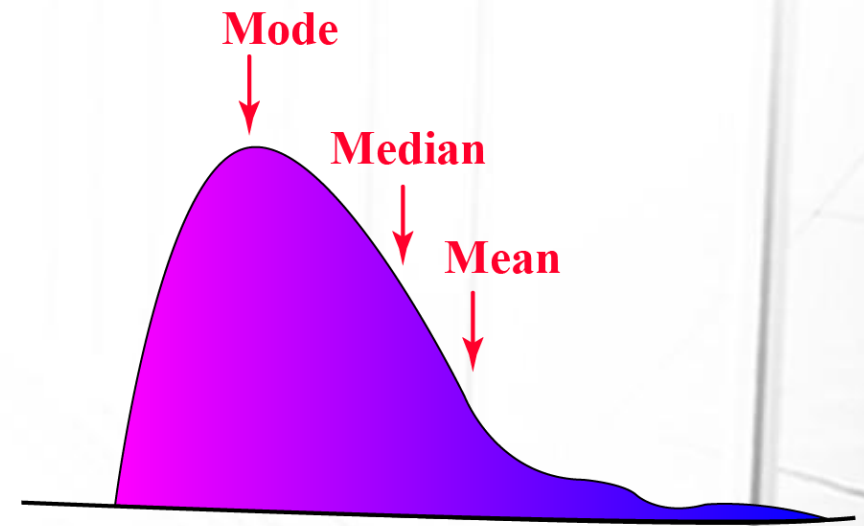
Descriptive statistics in data science refers to techniques used to summarize and describe the main features of a dataset. It provides a clear overview of the data's central tendency, variability, and distribution. Key measures include:

- **Mean:** The average value of the data.
- **Median:** The middle value when the data is ordered.
- **Mode:** The most frequently occurring value.

MEASURES OF CENTRAL TENDENCY

- Measures of central tendency are the values that describe a data set by identifying the central position of the data. There are 3 main measures of central tendency - Mean, Median, Mode
- Mean- Sum of all observations divided by the total number of observations.
- Median- The middle or central value in an ordered set.
- Mode- The most frequently occurring value in a data set.

Measures of Central Tendency





MEASURE OF VARIATION

- **Measure of Variation** **Measure of variation** is the way to extract meaningful information from a set of provided data. Variability provides a lot of information about the data. and some of the information it provides is mentioned below: It shows how far data items lie from each other. It shows the distance from the center of the distribution.

MEASURES OF VARIATION

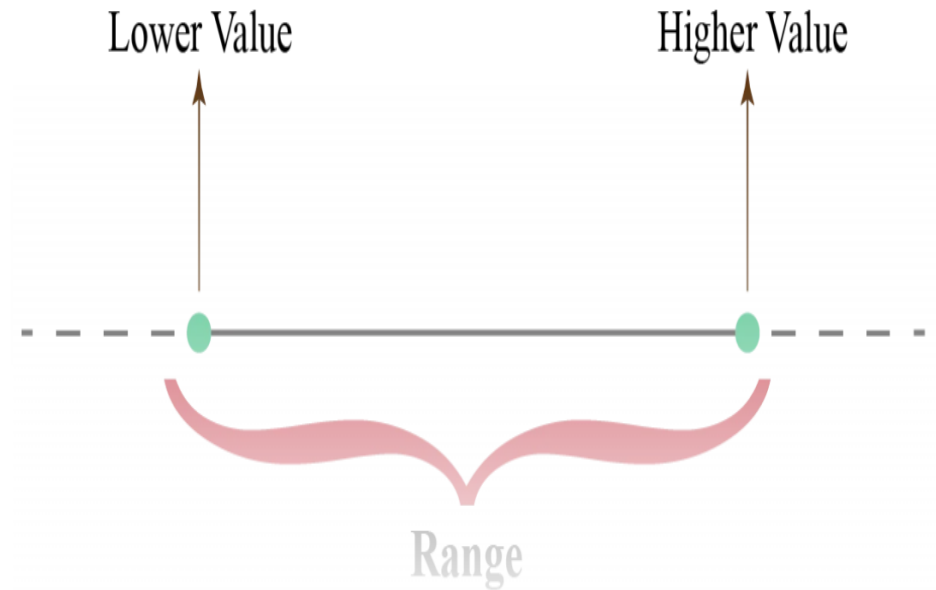
- **Range**
- Variance
- Standard Deviation
- Interquartile Range (IQR)
- Mean Absolute Deviation(MAD)

RANGE

In data science, the **range** is a measure of variation that indicates the spread of data values within a dataset. It is calculated as the difference between the maximum and minimum values:

$$\text{RANGE} = \text{MINIMUM VALUE} - \text{MAXIMUM VALUE}$$

The range provides a simple way to understand the extent of variability in the data, but it is sensitive to outliers, as a single extreme value can significantly affect the range.

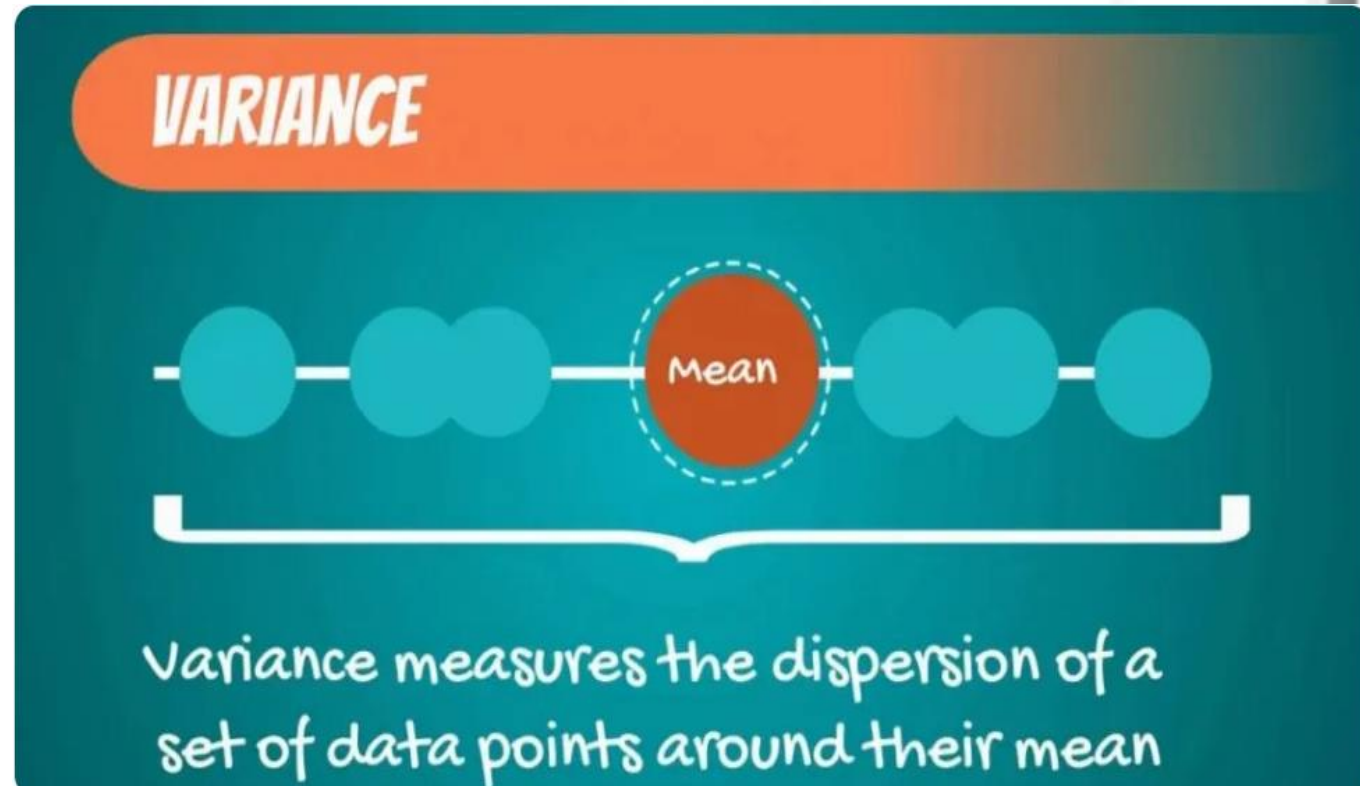


VARIANCE

In data science, **variance** is a measure of variation that quantifies the average squared deviation of each data point from the mean of the dataset. It is calculated using the formula:

$$\text{Variance} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Variance gives an idea of how much the data points differ from the mean, with higher variance indicating more spread. However, it is in squared units of the original data, which can make it less intuitive.



STANDARD DEVIATION

In data science, **standard deviation** is a measure of variation that quantifies the average amount by which data points deviate from the mean of the dataset. It is the square root of the variance and is expressed in the same units as the data, making it more interpretable. The formula for standard deviation is:

$$\text{Standard Deviation} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Where, \bar{x} = Sample Mean

n = Sample size

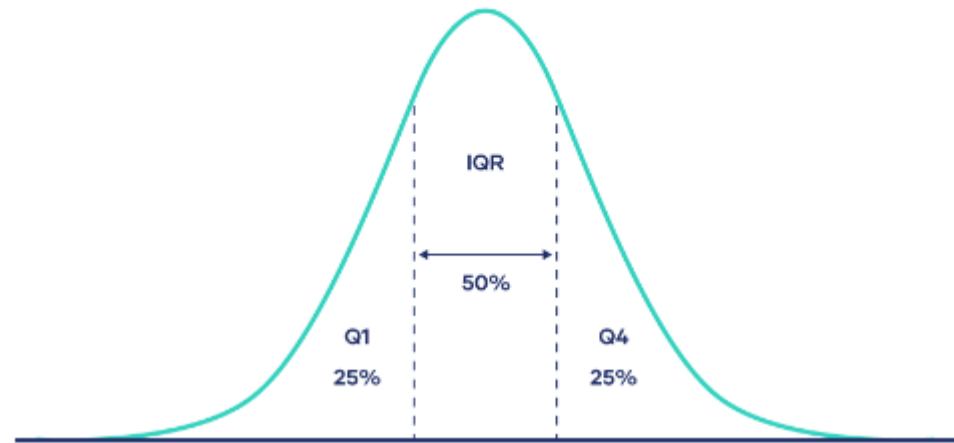
x_i = i th value of the data

INTERQUARTILE RANGE (IQR)

In data science, the **Interquartile Range (IQR)** is a measure of variation that quantifies the range within which the central 50% of data values fall. It is calculated as:

$$\text{IQR} = Q3 - Q1$$

Interquartile range on a normal distribution



MEAN ABSOLUTE DEVIATION(MAD)

In data science, **Mean Absolute Deviation (MAD)** is a measure of variation that quantifies the average absolute deviation of each data point from the mean. It is calculated using:

$$\text{MAD} = \frac{1}{N} \sum_{i=1}^N |x_i - \bar{x}|$$

What is Mean Absolute Deviation (MAD)

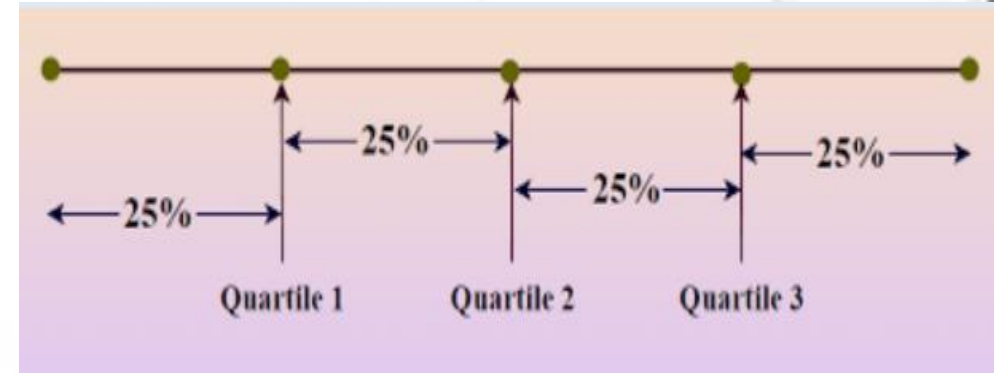


QUARTILES

Quartiles

Quartiles divide a dataset into four equal parts, each containing 25% of the data. They help in understanding the spread and center of the data. The key quartiles are:

- **First Quartile (Q1):** The 25th percentile, or the value below which 25% of the data falls.
- **Second Quartile (Q2):** The 50th percentile, or the median, which divides the data into two equal halves.
- **Third Quartile (Q3):** The 75th percentile, or the value below which 75% of the data falls.



PERCENTILES

Percentiles

Percentiles divide the dataset into 100 equal parts, each representing 1% of the data. They are useful for understanding the relative standing of data points. The key percentiles include:

- 25th Percentile:** Equivalent to the first quartile (Q1).
- 50th Percentile:** Equivalent to the median (Q2).
- 75th Percentile:** Equivalent to the third quartile (Q3).

Percentiles can provide more granularity than quartiles, as they offer a way to understand the distribution of data across a wider range of points. For example, the 90th percentile indicates the value below which 90% of the data falls.



SKEWNESS

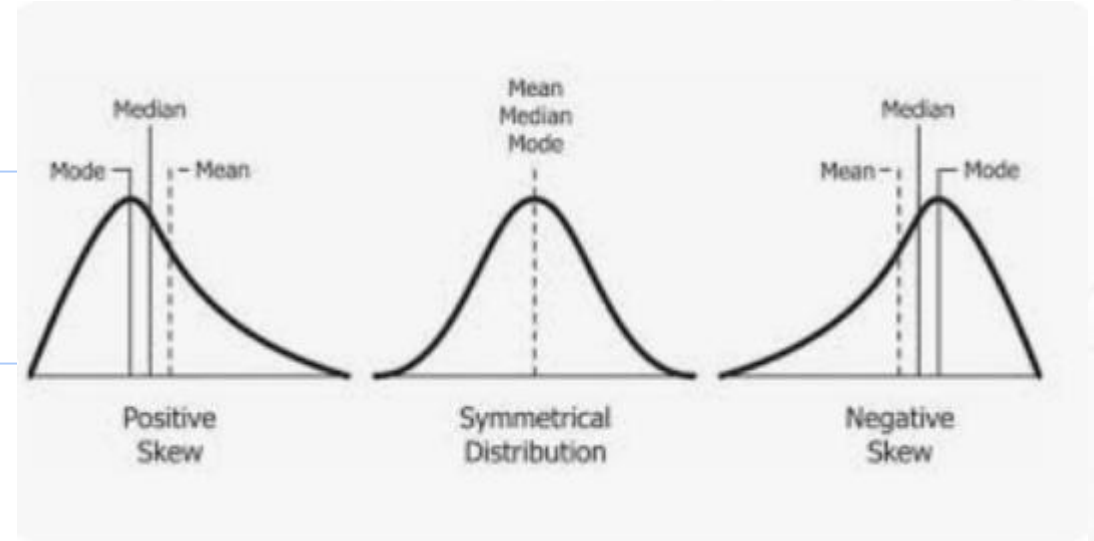
In data science, **skewness** measures the asymmetry of a data distribution around its mean. It indicates whether the data distribution leans to the left or right:

- Positive Skewness:** The distribution's tail extends more to the right, with most data points concentrated on the left. This results in a distribution with a longer right tail.

- Negative Skewness:** The distribution's tail extends more to the left, with most data points concentrated on the right. This results in a distribution with a longer left tail.

- Zero Skewness:** The distribution is symmetrical, resembling a normal distribution.

Skewness helps in understanding the shape of the data distribution and can guide decisions on appropriate statistical methods and transformations.

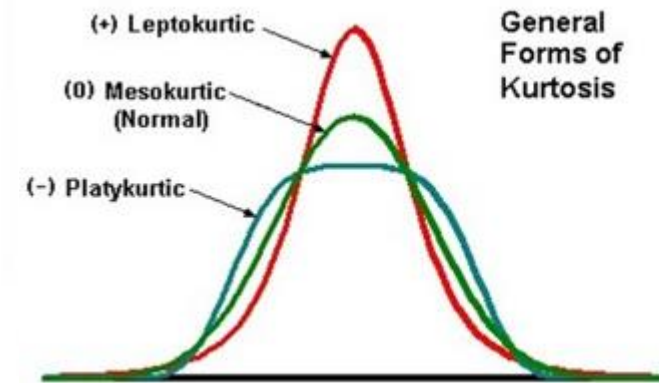


KURTOSIS

In data science, **kurtosis** measures the "tailedness" or the sharpness of a data distribution's peak compared to a normal distribution. It indicates how much data is concentrated in the tails and the peak of the distribution:

- **Leptokurtic:** Positive kurtosis; the distribution has heavier tails and a sharper peak than the normal distribution. It indicates more extreme values or outliers.
- **Platykurtic:** Negative kurtosis; the distribution has lighter tails and a flatter peak compared to the normal distribution. It suggests fewer outliers and a more uniform spread.
- **Mesokurtic:** Kurtosis close to zero (when excess kurtosis is used); the distribution has a shape similar to a normal distribution.

Kurtosis helps in assessing the likelihood of extreme values and understanding the shape of the data distribution.

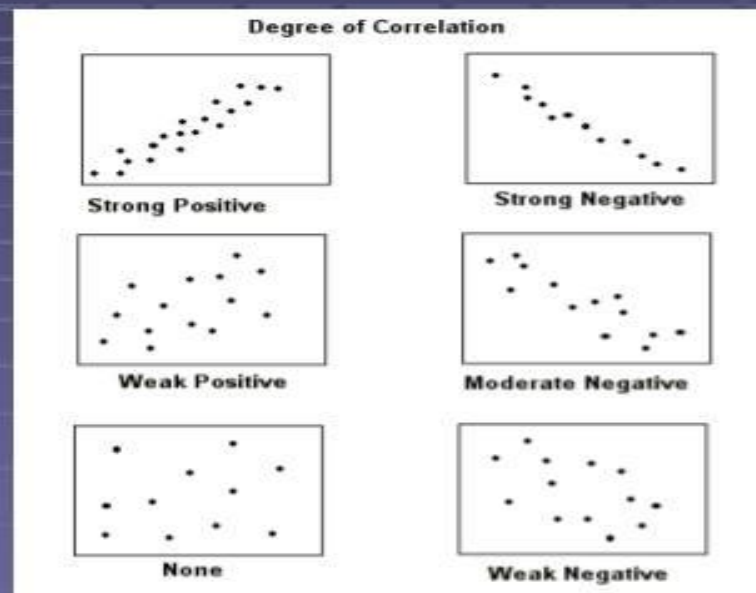




EXPLORATORY DATA ANALYTICS DESCRIPTIVE STATISTICS

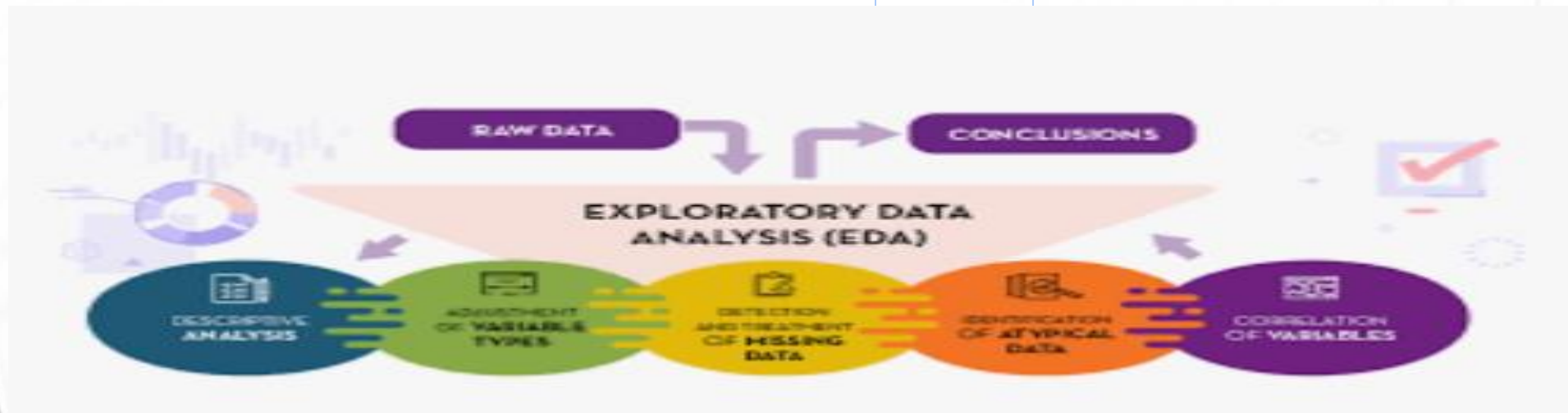
Descriptive Statistics

- Relationship
- Linear Relationship
 - Positive
 - Negative
- Relationship Strength
 - Weak, strong, no relationship
- Correlation Coefficient
 - Between -1 and 1
 - 0 – no relationship
- Regression Analysis
 - Criterion variables (Y)
 - Predictor variables (X)



http://hosting.scanner.ca/etris/remotesensing/1_lectureimages/correlation.gif

Exploratory Data Analysis (EDA) involves using various techniques to summarize and understand the characteristics of a dataset before diving into more complex analyses or modeling. Descriptive statistics are a key component of EDA, providing a foundational understanding of the data. Here are some key descriptive statistics and techniques used in EDA:



❖ Exploratory Data Analysis of Mean

Exploratory Data Analysis (EDA)

- A very specific way to look at data.

	Traditional	Exploratory Data Analysis
Organize	Frequency Distribution	Stem and Leaf Plot
Display	Histogram	Boxplot
Central Tendency	Mean	Median
Variation	Standard Deviation	Interquartile Range

■ Standard Deviation

1. A standard deviation (or σ) is a **measure of how dispersed the data is in relation to the mean**. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.

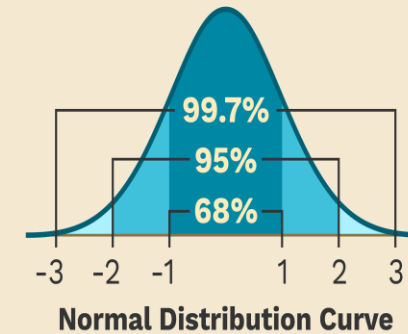
Calculating Standard Deviation

$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

n = The number of data points

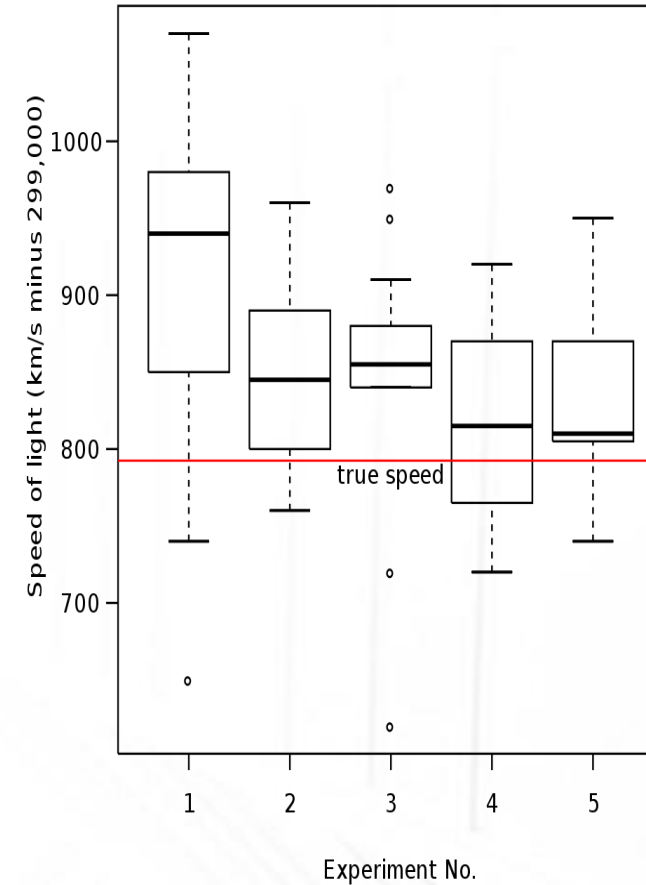
X_i = Each of the values of the data

\bar{X} = The mean of X_i



❑ BOX PLOTS

a **box plot** or **boxplot** is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles. In addition to the box on a box plot, there can be lines extending from the box indicating variability outside the upper and lower quartiles, thus, the plot is also termed as the **box-and-whisker plot** and the **box-and-whisker diagram**.





BOX PLOT :-

BOX PLOT: IT IS A TYPE OF CHART THAT DEPICTS A GROUP OF NUMERICAL DATA THROUGH THEIR QUARTILES. IT IS A SIMPLE WAY TO VISUALIZE THE SHAPE OF OUR DATA. IT MAKES COMPARING CHARACTERISTICS OF DATA BETWEEN CATEGORIES VERY EASY.

COMPONENTS OF A BOX PLOT

1.Box:

1. Represents the interquartile range (IQR) of the data. This is the range between the first quartile (Q1) and the third quartile (Q3), covering the middle 50% of the data.
2. The length of the box indicates the spread of the central 50% of the data.

2.Median Line:

1. A line inside the box that shows the median (the 50th percentile) of the data. It divides the data into two equal halves.

3.Whiskers:

1. Lines extending from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. This range is considered the "inner range" or "non-outlier" range.
2. The whiskers help visualize the spread of the bulk of the data outside the IQR.

4.Outliers:

1. Data points that fall outside the whiskers, i.e., more than 1.5 times the IQR from the quartiles. These points are typically plotted as individual dots or asterisks and indicate potential anomalies or extreme values.

5.Notches (optional):

1. Some box plots include notches around the median line, which represent confidence intervals for the median. Notches can be used to compare medians between multiple groups to assess if they differ significantly.

CODE :

```
import matplotlib.pyplot as plt
```

```
value1 = [82,76,24,40,67,62,75,78,71,32,98,89,78,67,72,82,87,66,56,52]
```

```
value2=[62,5,91,25,36,32,96,95,3,90,95,32,27,55,100,15,71,11,37,21]
```

```
value3=[23,89,12,78,72,89,25,69,68,86,19,49,15,16,16,75,65,31,25,52]
```

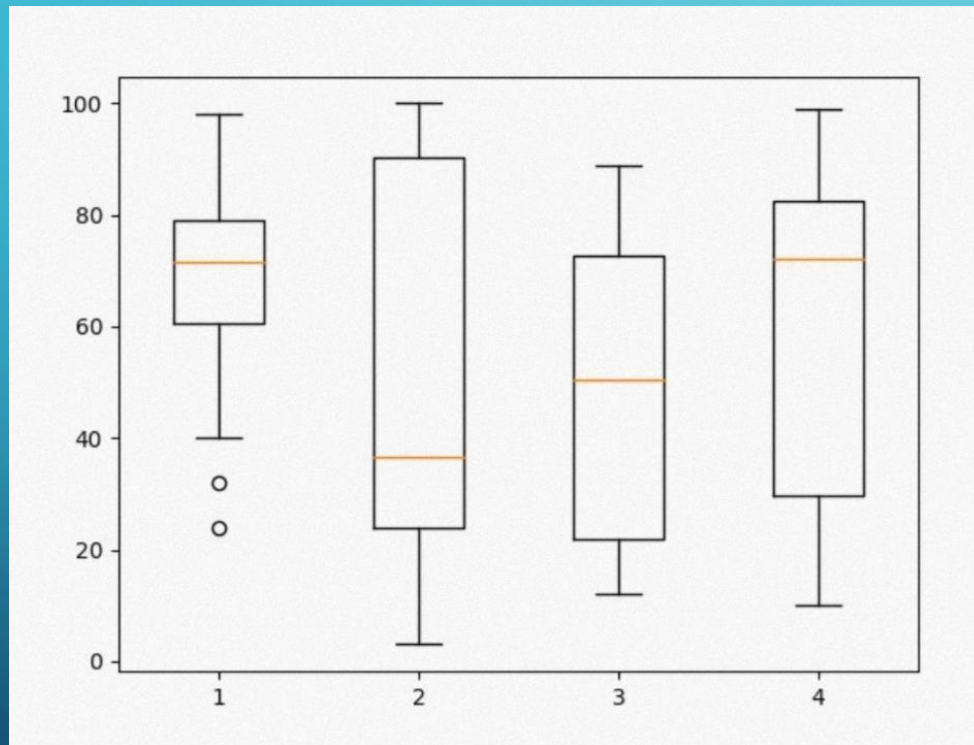
```
value4=[59,73,70,16,81,61,88,98,10,87,29,72,16,23,72,88,78,99,75,30]
```

```
box_plot_data=[value1,value2,value3,value4]
```

```
plt.boxplot(box_plot_data)
```

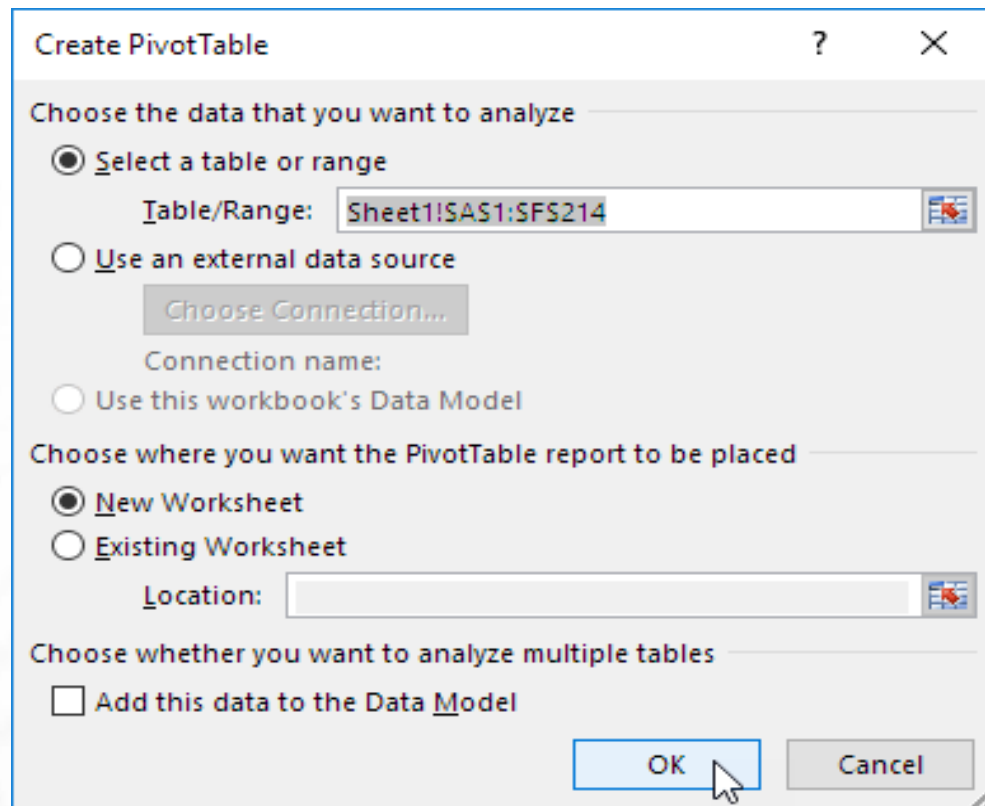
```
plt.show()
```

RESULT OF CODE :



➤ Pivot Table

- Pivot tables are one of Excel's most powerful features. A pivot table allows you to extract the significance from a large, detailed data set.



The screenshot shows the 'Create PivotTable' dialog box with the following settings:

- Choose the data that you want to analyze**
 - ☒ **Select a table or range**
 - Table/Range:
 - ☐ **Use an external data source**
 - Choose Connection...
 - Connection name:
 - ☐ **Use this workbook's Data Model**
- Choose where you want the PivotTable report to be placed**
 - ☒ **New Worksheet**
 - ☐ **Existing Worksheet**
 - Location:
- Choose whether you want to analyze multiple tables**
 - ☐ **Add this data to the Data Model**

Buttons: OK, Cancel

. Insert a Pivot Table

To insert a pivot table, execute the following steps.

1. Click any single cell inside the data set.
2. On the Insert tab, in the Tables group, click PivotTable.
3. Click ok.

PIVOT TABLES :

- **Pivot Tables:** A pivot table is a table of statistics that summarizes the data of a more extensive table (such as from a database, spreadsheet, or business intelligence program). This summary might include sums, averages, or other statistics, which the pivot table groups together in a meaningful way.

COMPONENTS OF PIVOT TABLES

Rows and Columns:

- Rows:** These represent the categories or dimensions along the rows of the table. Each row represents a unique category or grouping of data.
- Columns:** These represent the categories or dimensions along the columns. Each column represents another level of grouping or comparison.

•Values:

- Aggregations:** The cells in the pivot table show aggregated values based on the row and column dimensions. Common aggregations include sum, average, count, maximum, and minimum.

•Filters:

- Filters allow you to include or exclude specific data from the pivot table, enabling you to focus on subsets of the data.

CODE :-

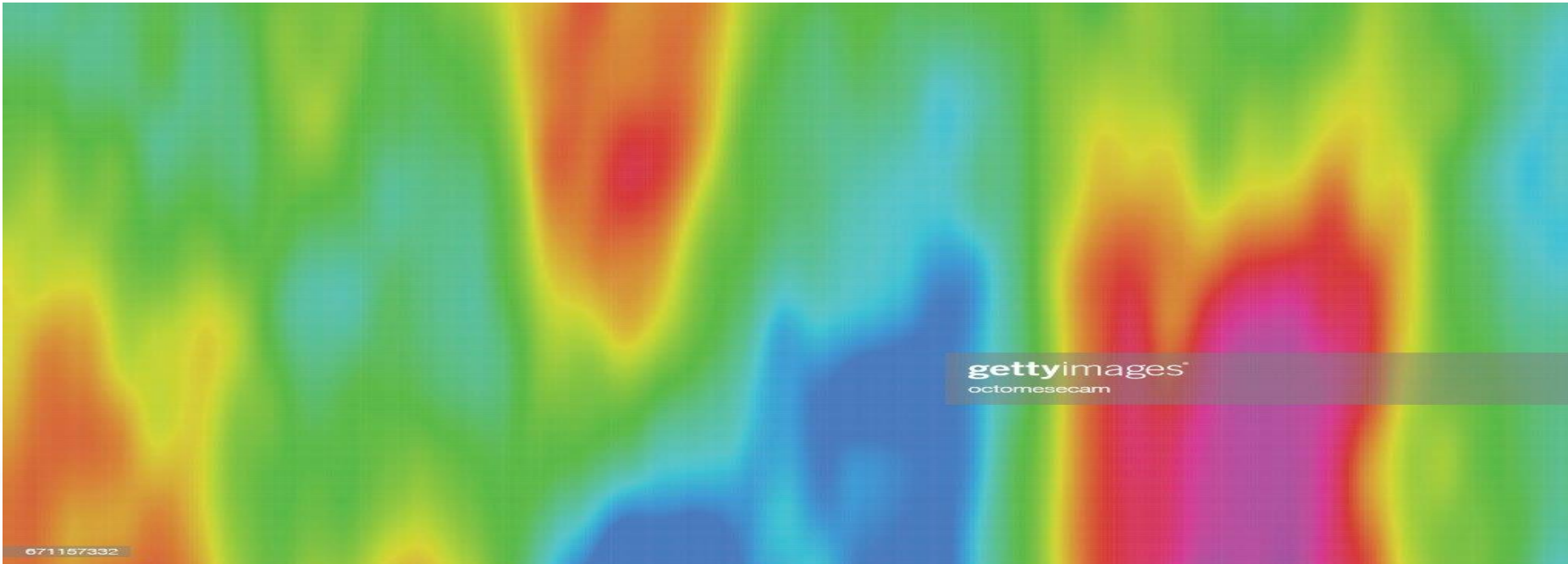
- `import pandas as pd`
- `data = {'person': ['A', 'B', 'C', 'D', 'E', 'A', 'B', 'C', 'D', 'E', 'A', 'B', 'C', 'D', 'E', 'A', 'B', 'C', 'D', 'E'],
'sales': [1000, 300, 400, 500, 800, 1000, 500, 700, 50, 60, 1000, 900, 750, 200, 300,
1000, 900, 250, 750, 50], 'quarter': [1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4],
'country': ['US', 'Japan', 'Brazil', 'UK', 'US', 'Brazil', 'Japan', 'Brazil', 'US', 'US', 'US', 'Japan',
'Brazil', 'UK', 'Brazil', 'Japan', 'Japan', 'Brazil', 'UK', 'US']}`
- `df = pd.DataFrame(data)`
- `pivot = df.pivot_table(index=['person'], values=['sales'], aggfunc='sum')`
- `print(pivot)`

RESULT OF CODE :-

person	sales
A	4000
B	2600
C	2100
D	1500
E	1210

❏ HEAT MAP

- A heat map (or heatmap) is a **data visualization technique that shows magnitude of a phenomenon as color in two dimensions**. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.



HEAT MAPS :-

- A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colours

COMPONENTS OF HEAT MAP

- Matrix/Grid:**

- Cells:** The fundamental unit of a heat map. Each cell represents the value at the intersection of two variables or dimensions.

- Rows and Columns:** Define the matrix structure. Typically, one dimension is represented by rows and the other by columns.

- Color Scale:**

- Gradient:** Colors in a heat map represent the magnitude of the values. Commonly, a gradient from one color to another (e.g., from blue to red) is used to show low to high values.

- Color Bar:** A legend that shows the mapping between colors and their corresponding values. This helps in interpreting the color scale used in the heat map.

- Annotations:**

- Labels:** Text labels for rows and columns that help identify the categories or variables being represented.

- Value Annotations (optional):** Actual values displayed within the cells or alongside the color gradient, which can be useful for precise reading.

- Title and Axis Labels:**

- Title:** Descriptive text that provides context about what the heat map represents.

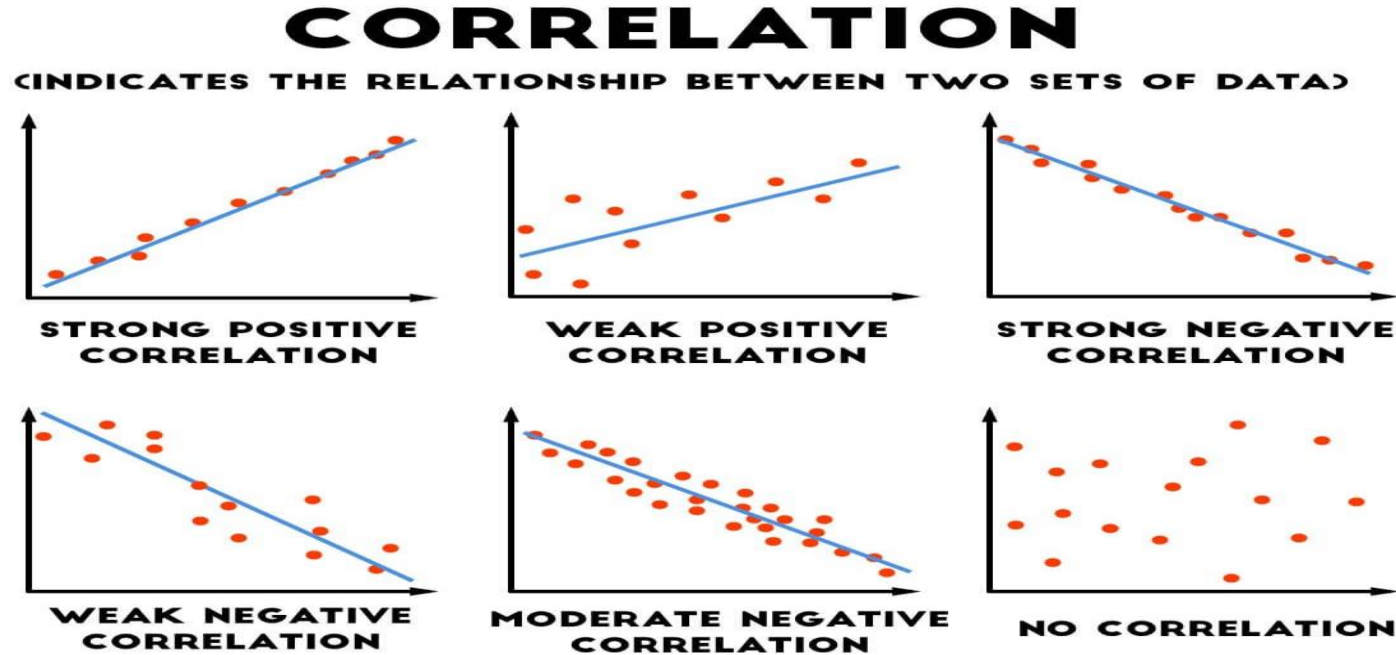
- Axis Labels:** Labels for the rows and columns that indicate the variables or categories being compared.

CODE :-

- `From pandas import DataFrame`
- `import matplotlib.pyplot as plt`
- `data=[{2,3,4,1},{6,3,5,2},{6,3,5,4},{3,7,5,4},{2,8,1,5}]`
- `Index= ['I1', 'I2','I3','I4','I5']`
- `Cols = ['C1', 'C2', 'C3','C4']`
- `df = DataFrame(data, index=Index, columns=Cols)`
- `plt.pcolor(df)`
- `plt.show()`



❖ CORRELATION STATISTICS



In **statistics**, **correlation** or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. Although in the broadest sense, "**correlation**" may indicate any type of association, in **statistics** it normally refers to the degree to which a pair of variables are linearly related.

CORRELATION :-

- A **correlation coefficient** is a number between -1 and 1 that tells you the strength and direction of a relationship between variables.
- Correlation coefficients quantify the association between variables or features of a dataset. These statistics are of high importance for science and technology, and Python has great tools that you can use to calculate them. SciPy, NumPy, and Pandas correlation methods are fast, comprehensive, and well-documented

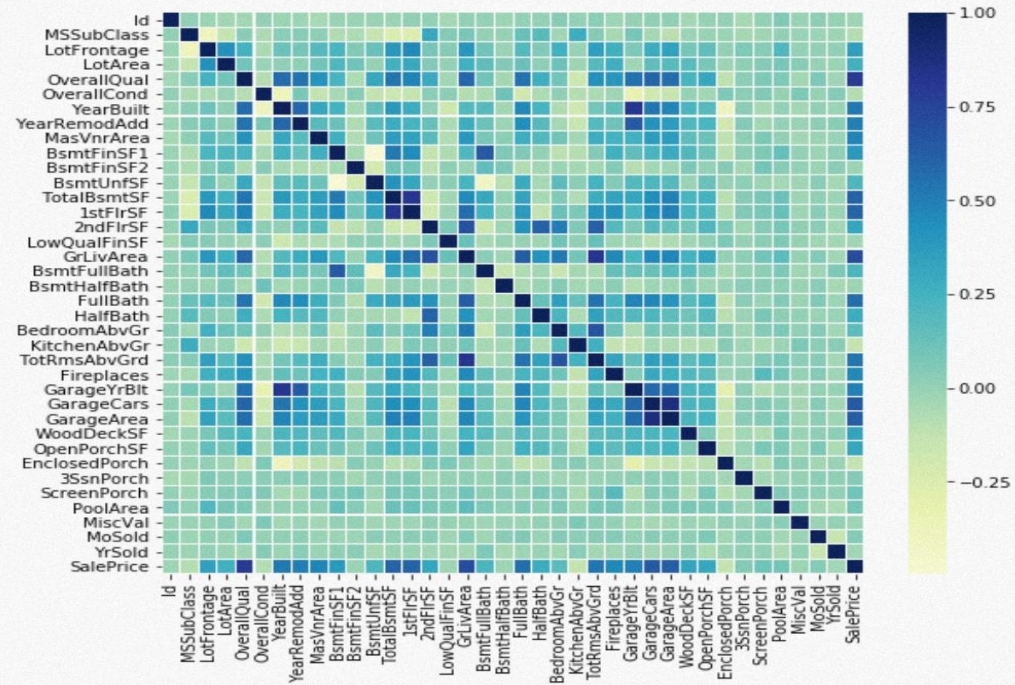
CODE :-

```
corrmat = data.corr()

ax = plt.subplots(figsize =(9, 8))
sns.heatmap(corrmat, ax = ax, cmap ="YlGnBu", linewidths = 0.1)
```

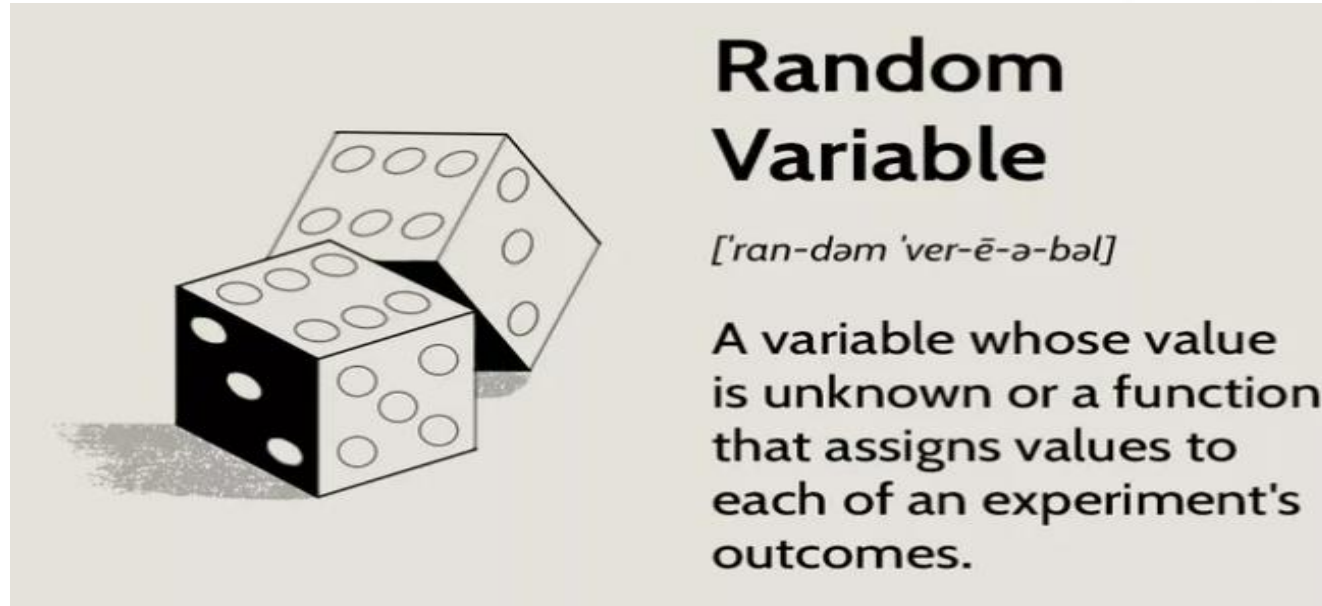

RESULT OF CODE :-

Output:



❑ Random Variable

- A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's



- The use of random variables is most common in probability and statistics, where they are used to quantify outcomes
- Risk analysts use random variables to estimate the probability of an adverse event occurring.

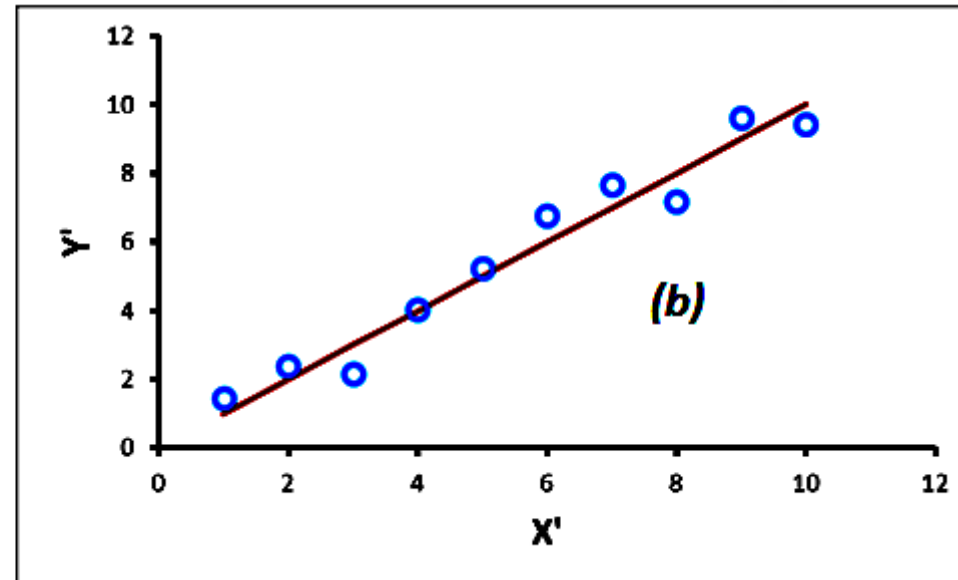
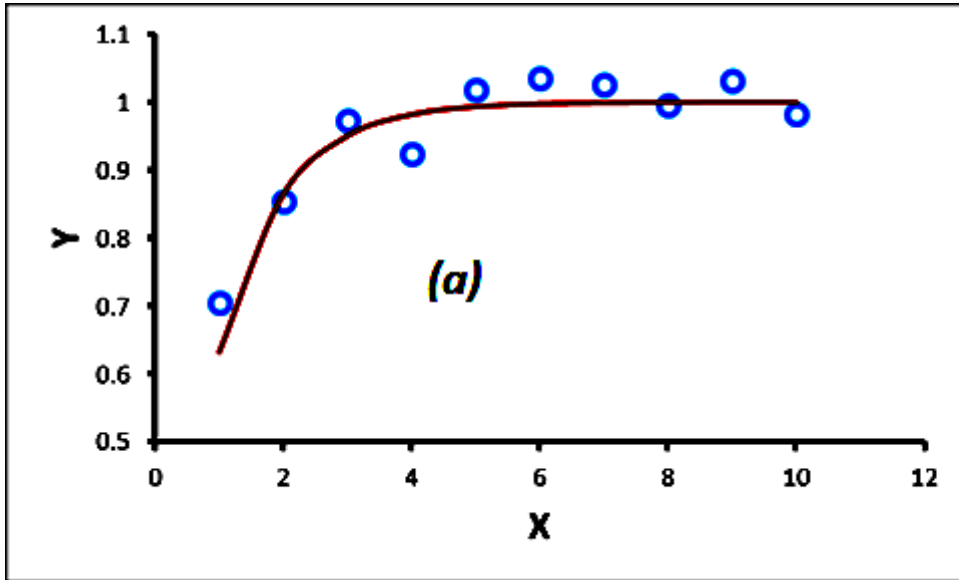
➤ Variance

- Variance is a measure of how data points differ from the mean.
According to Layman, a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value.
- Variance means to find the expected difference of deviation from actual value. Therefore, variance depends on the standard deviation of the given data set.
- The more the value of variance, the data is more scattered from its mean and if the value of variance is low or minimum, then it is less scattered from mean. Therefore, it is called a measure of spread of data from mean.

✓ COVARIANCE

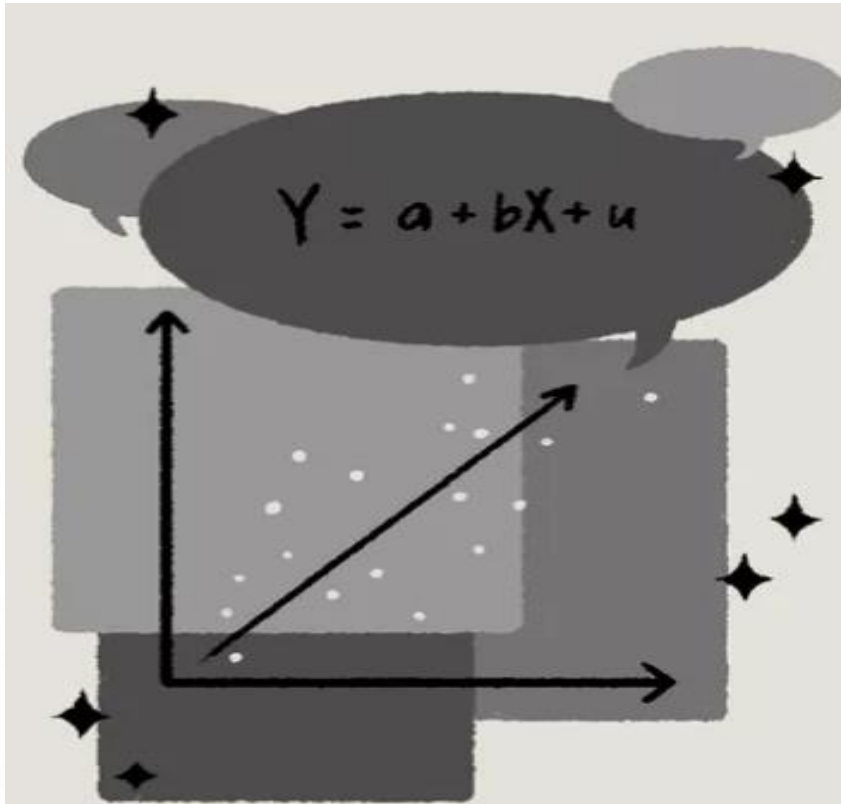
- **Covariance** is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.
- Covariance can have both positive and negative values. Based on this, it has two types:
 1. positive covariance
 2. Negative covariance

■ Correlation Linear Transformations of Random Variable



A linear rescaling is a transformation of the form $g(u) = a + bu$. A linear rescaling of a random variable does not change the basic shape of its distribution, just the range of possible values. A linear rescaling transforms the mean in the same way the individual values are transformed.

➤ REGRESSION



Regression

[ri-'gre-shən]

A statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

THANK YOU