Lesson 1

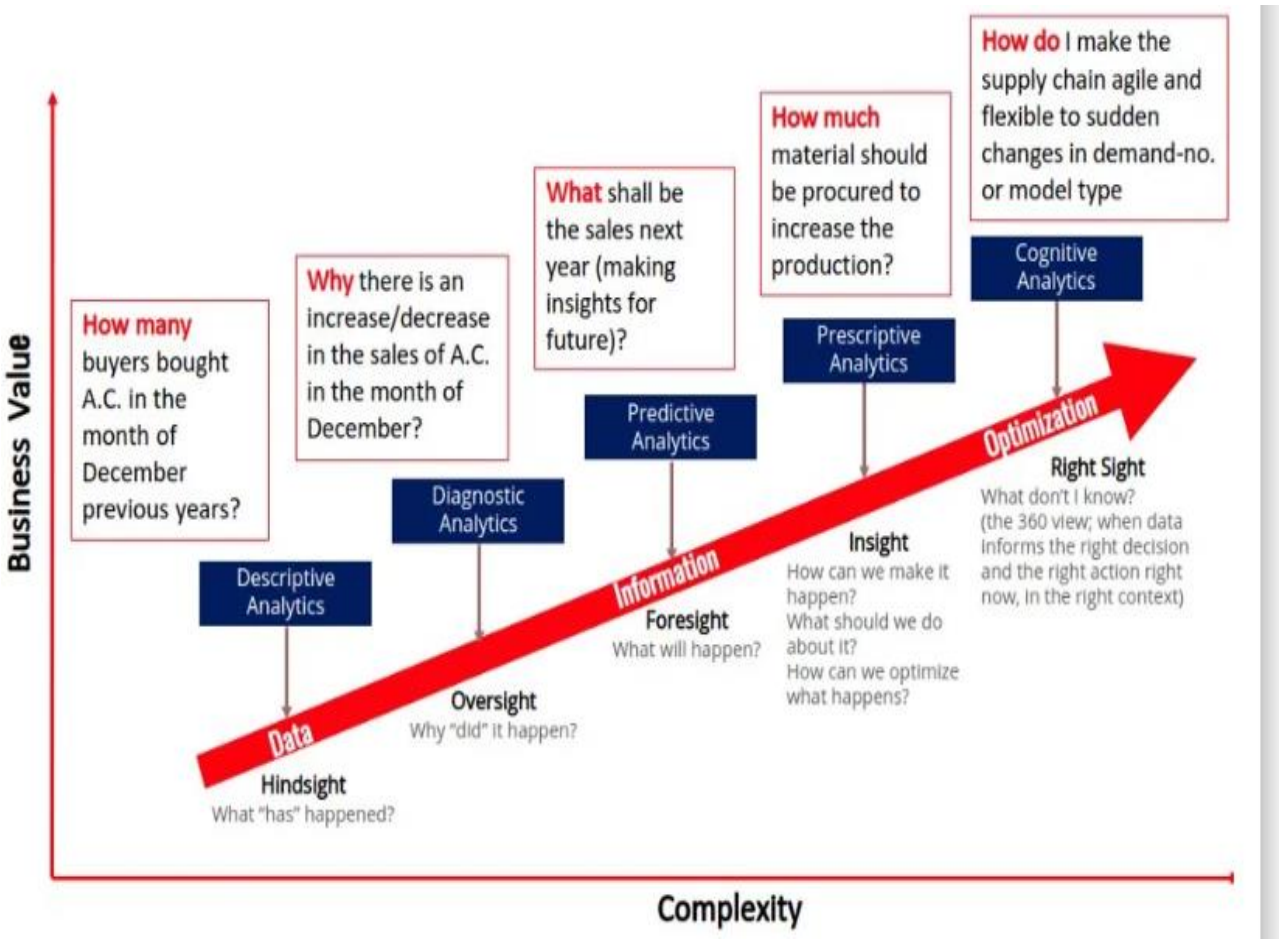
**What is Data Analytics?**

- Data analytics refers *to the process of analyzing raw data to extract actionable insights*. This process transforms incomprehensible data into coherent information that can drive strategic decisions within organizations.
- Data analysts play a crucial role in this process by collecting, organizing, and interpreting data, ultimately providing recommendations based on their findings.

**<u>Types of Data Analysis</u>**

**Data analytics** can be categorized into several types, each serving different purposes:

1. **Descriptive Analytics:** Summarizes historical data to understand what has happened.
2. **Exploratory Data Analysis (EDA):** Investigates data sets to discover patterns or relationships without having specific hypotheses in mind.
3. **Confirmatory Data Analysis (CDA):** Tests hypotheses to confirm or refute existing theories.
4. **Predictive Analytics:** Uses statistical models to forecast future outcomes based on historical data.
5. **Prescriptive Analytics:** Recommends actions based on predictive outcomes, providing the best course of action to achieve desired results.




**Importance of Data Analytics**

Data analytics is essential for modern businesses as it helps in:

1. Informed Decision-Making: By providing insights that guide strategic choices.
2. Operational Optimization: Identifying inefficiencies and areas for improvement.
3. Competitive Advantage: Enabling organizations to understand their market and audience better than their competitors.
4. Enhanced Customer Experience: Tailoring services and products based on customer data.

**Tools and Techniques**

Data analysts utilize various tools and programming languages to perform their analyses, including:

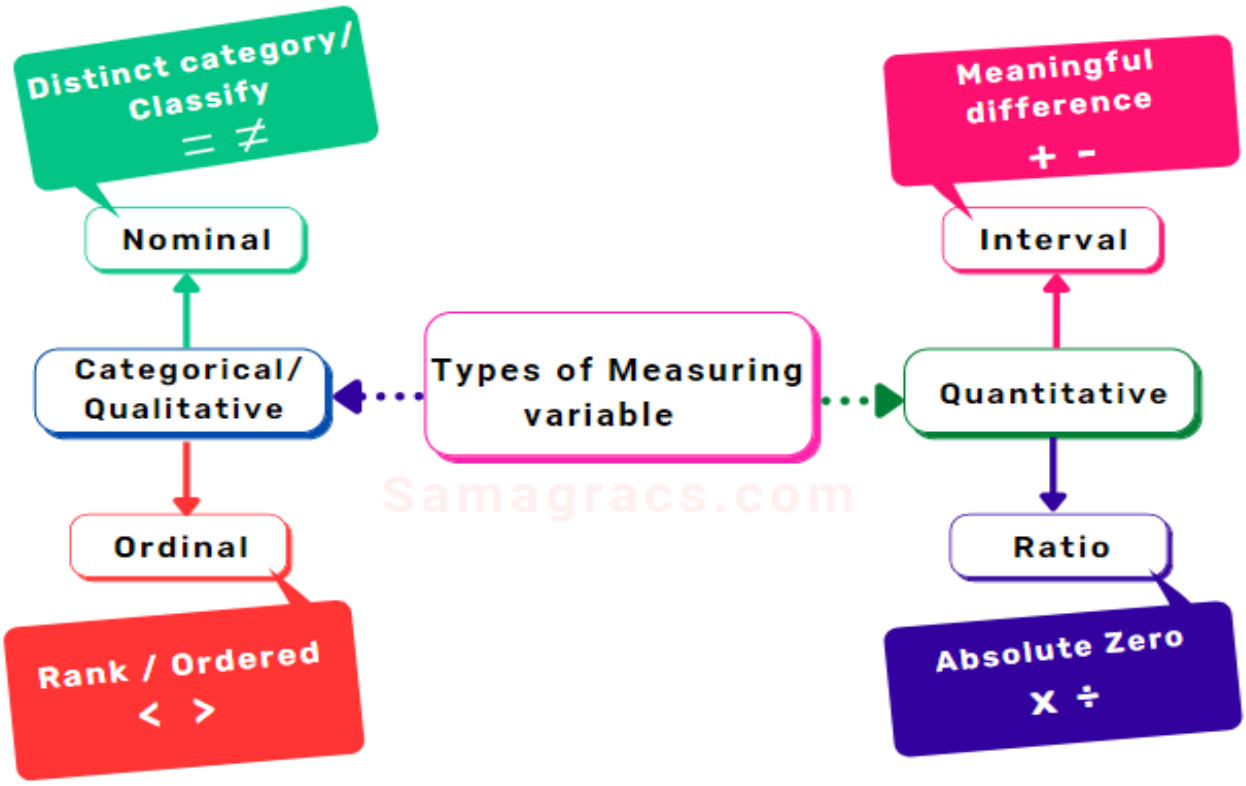
- **Statistical Software**: Such as **R and Python** for data manipulation and statistical analysis.
- **Database Management**: SQL for querying databases and managing data sets.

- **Visualization Tools:** Tools like Tableau or Power BI to present data insights visually.

**Skills Required**

- To be effective in data analytics, individuals typically need:
- Proficiency in programming languages (e.g., SQL, R, Python).
- Strong analytical and statistical skills.
- Ability to communicate findings clearly to stakeholders.
- Familiarity with data visualization tools and techniques

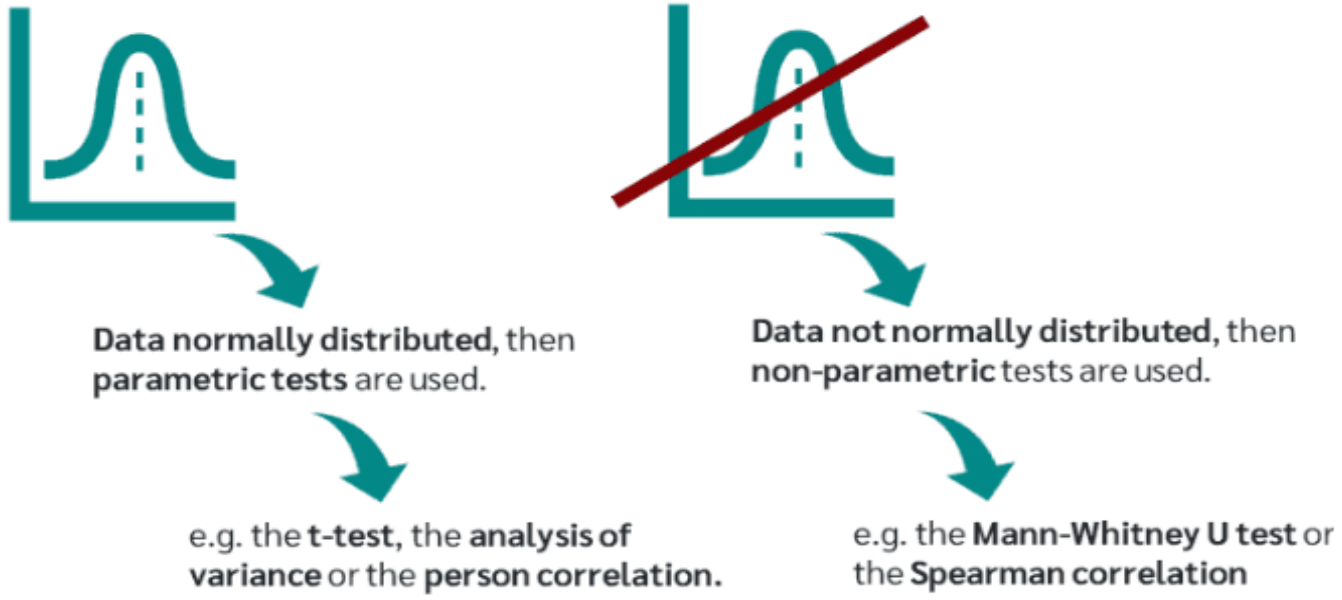
**scales of measurements**



| Scales | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| **Description** | Label | Label, order | Label, order, equal distance units | Label, order, equal distance units and absolute zero |
| **Nature** | Qualitative | Qualitative | Quantitative | Quantitative |
| **Data** | Discrete Data | Discrete Data | Continuous Data | Continuous Data |
| **Test** | Non- Parametric Test | Non- Parametric Test | Parametric Test | Parametric Test |
| **Example** | Gender can be male or female.<br><br>Eye colour can be black, blue, green. | Height can be short, middle and tall.<br><br>Customer satisfaction can be unhappy, neutral, happy. | What is temperature in your city? It can be.<br>• below 0°C.<br>• between 0°C - 20°C.<br>• between 20°C - 40°C and<br>• above 40°C. | What is your weight in kilograms?<br>• Less than 50 KG.<br>• 51- 100 KG.<br>• 101- 150 KG.<br>• More than 150 KG. |

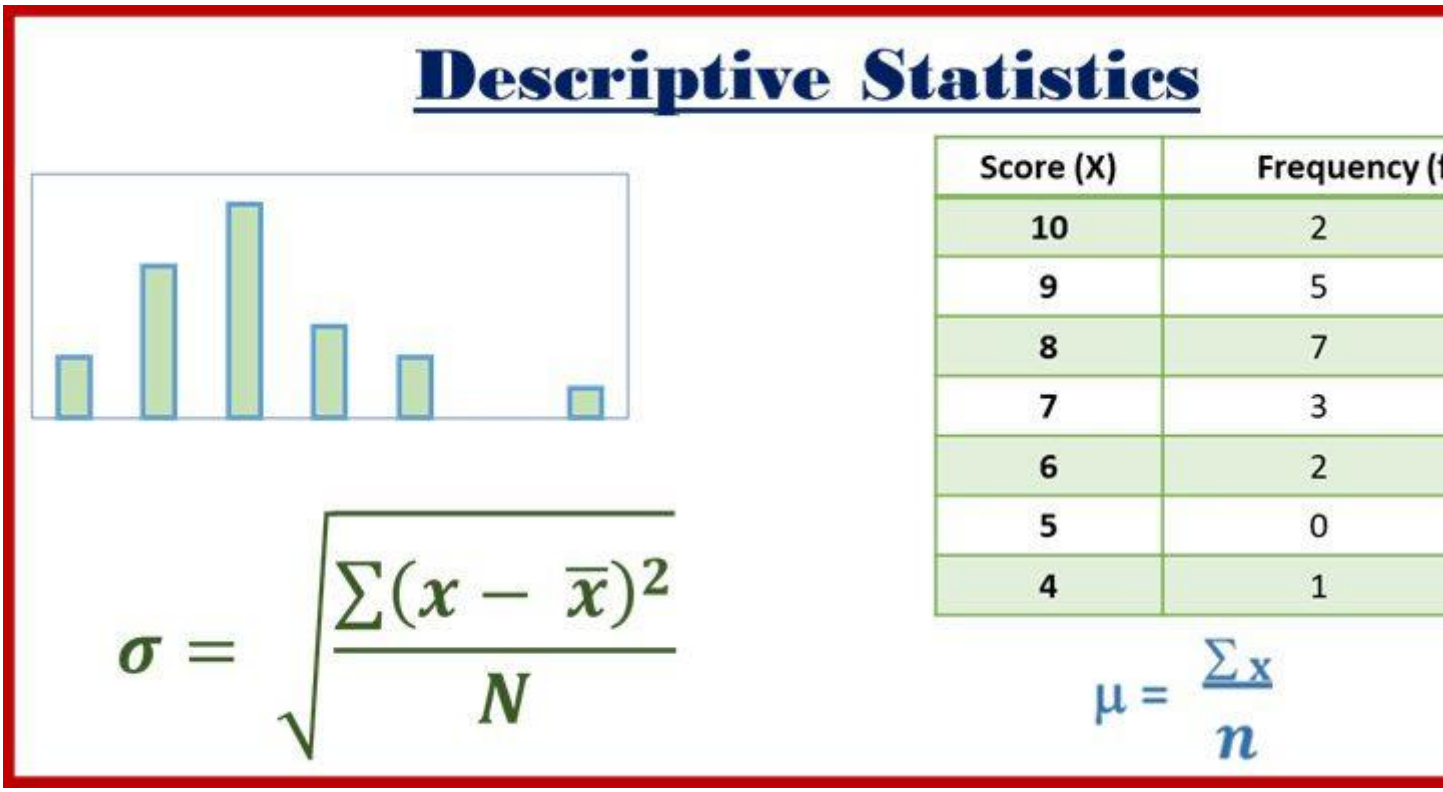


## Simplified:



| | Parametric tests | Nonparametric tests |
|---|---|---|
| **One sample** | Simple t-Test | Wilcoxon test for one sample |
| **Two dependent samples** | Paired Sample t-Test | Wilcoxon-Test |
| **Two independent samples** | Unpaired Sample t-Test | Mann-Whitney U-Test |
| **More than two independent samples** | One-way ANOVA | Kruskal-Wallis-Test |
| **More than two dependent samples** | Repeated Measures ANOVA | Friedman-Test |
| **Correlation between two variables** | Pearson Correlation | Spearman-Correlation |

**Descriptive statistics summarize and organize characteristics of a data set/entire population.**



- In quantitative research, after collecting data, the first step of statistical analysis is to describe the characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).
- The next step is inferential statistics, which helps you decide whether your data confirms or refutes your hypothesis and whether it is generalizable to a larger population.

---

Types of descriptive statistics
There are 3 main types of descriptive statistics:

1. **Central tendency**: describes averages of the data points.
2. **Variability** : describes – variation between the values.
3. **Distribution** : describes – the frequency of each value.

---

**Central tendency**

**What are the measures of central tendency?**

- It is a measure to describe a single value middle or centre value (of the whole data set).
- It is also called a measure of centre or central location.
- Each of these measures describes a different indication of central value in the distribution.
- There are three main measures of central tendency:
    -
    - Mean
    - Mode and
    - Median

---

Mean (Arithmetic)
The mean is the sum of each value divided by the number of observations. This is also known as the arithmetic average.

- The mean (or average) is the most popular and well-known measure of central tendency.
- It can be used with both discrete and continuous data
- As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.
- If continuous data values are $x_1$, x2, ……..……,$x_n$ and n is the number of data

$$\text{Mean} = \frac{x1 + x2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + xn}{n}$$

$$\mu = \frac{\sum x}{n}$$

---

Mode

- The mode is the most commonly occurring value in a distribution.
- The mode is the most frequent value in our data set.
- It is the highest bar in a histogram.

---

Median

The median is the *middle value* in distribution when the values are arranged in ascending or descending order.

The median divides the distribution in half (there are 50% of observations on either side of the median value).

In a distribution with an odd number of observations, the median value is the middle value.

- If no of the data is even, the median is the average of the middle two values.
- The median is less affected by outliers and skewed data.

---

Application of Mean, median and mode with data type

- The best measure of central tendency with respect to the type of variable

| Variable | Best measure of central tendency |
|---|---|
| Nominal | Mode |
| Ordinal | Median |
| Interval/Ratio (not skewed) | Mean |
| Interval/Ratio (skewed) | Median |

---

How do outliers influence the measures of central tendency?

- Outliers are extreme data value(s) that are notably different from the rest of the data.
- Outliers can alter the results of Mean, Mode, and Median analysis.
- The mean is more sensitive to the existence of outliers than the median or mode.

---

**Variability or dispersion**

- The goal for variability is to obtain a measure of how spread out the scores are in distribution.
- A measure of variability usually accompanies a measure of central tendency as basic descriptive statistics for a set of scores.
- Variability serves both as a descriptive measure and as an important component of most inferential statistics.
- As a descriptive statistic, variability measures the degree to which the scores are spread out or clustered together in a distribution.
- In the context of inferential statistics, variability provides a measure of how accurately an individual score or sample represents the entire population.
- When the population variability is small, all of the scores are clustered close together and any individual score or sample will necessarily provide a good representation of the entire set.
- When variability is large and scores are widely spread, it is easy for one or two extreme scores to give a distorted picture of the general population.

---

Why Understanding Variability is Important

- A low dispersion indicates that the data points tend to be clustered tightly around the centre.
- High dispersion signifies that data points are far away.

---

Measuring Variability

- Variability is determined by measuring *distance*. It can be measured by calculating –

1. Range
2. Interquartile range
3. Standard deviation or variance.

---

a) Range

- The range is the simplest measure of variability.
- Take the smallest number and subtract it from the largest number to calculate the range. This shows the spread of our data.
- The range is sensitive to outliers or values that are significantly higher or lower than the rest of the data set, and should not be used when outliers are present.
- The **range** is the total distance covered by the distribution, from the highest value to the lowest value.

---

b) Interquartile range

- The IQR, or the middle fifty, is the range for the middle fifty percent of the data. The IQR only considers middle values, so it is not affected by the outliers.
- The **interquartile range** is the distance covered by the middle 50% of the distribution (the difference between Q1 and Q3).

**IQR = Q3 -Q1**

- Steps to calculate IQR :-
  -

1. List the data in numerical order.
2. Find out the range and median.
3. Consider data points above the Median in Q3 Zone
4. Consider data points below the median in Q1 Zone
5. Find the median of the data in Zone Q1.
6. Find the median of the data in Zone Q3
7. Find the interquartile range using the formula IQR = Q3 – Q1

---

b) Standard deviation

- **Standard deviation** measures the standard distance between a data value and the mean.
- Follow the following steps for calculating the standard deviation :

1.

1. Find out the mean of all values
2. Subtract the mean from each data point to get the distance from the mean.
3. Square each distance.
4. Add up all of the squared distance.
5. For Population: Divide the sum of the squared distances by N (N- number of data points in a Population)
6. For sample: Divide the sum of the squared distances by $n - 1$ (n- number of data points in a sample)
7. Do the square root of the above value to get the Standard deviation.

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{N}}$$

---

Properties of the Standard Deviation

- If a constant is added to every score in a distribution, the standard deviation will not be changed.

- If you visualize the scores in a frequency distribution histogram, then adding a constant will move each score so that the entire distribution is shifted to a new location.
- The centre of the distribution (the mean) changes, but the standard deviation remains the same.
- If each score is multiplied by a constant, the standard deviation will be multiplied by the same constant.
- Multiplying by a constant will multiply the distance between scores, and because the standard deviation is a measure of distance, it will also be multiplied.

---

Descriptive Statistics- Mean and Standard Deviation

- If you are given numerical values for the mean and the standard deviation, you should be able to construct a visual image (or a sketch) of the distribution of scores.
- As a general rule, about 70% of the values will be within one standard deviation of the mean, and about 95% of the values will be within a distance of two standard deviations of the mean.

---

Difference between Central Tendency and Variability

- Central tendency describes the central point of the distribution, and variability describes how the data values are scattered around that central point.
- Together, central tendency and variability are the two primary values that are used to describe a distribution of a population or a sample.

---

**Frequency distributions**

A frequency distribution is an organized tabulation of the number of individuals located in each category on the scale of measurement.

The following set of N = 20 scores was obtained from a 10-point statistics quiz. We will organize these scores by constructing a frequency distribution table. Scores:

8, 9, 8, 7, 10, 9, 6, 4, 9, 8, 7, 8, 10, 9, 8, 6, 9, 7, 8, 8

---

Frequency Table:

| Score (X) | Frequency (f) |
|---|---|
| 10 | 2 |
| 9 | 5 |
| 8 | 7 |
| 7 | 3 |
| 6 | 2 |
| 5 | 0 |
| 4 | 1 |

1.

    1.

        1. Score (X) is in the first column
        2. The frequency associated with each score is recorded in the second column
        3. *X* values in the above frequency distribution table represent the scale of measurement, *not* the actual set of scores. Example: The x column lists the value 10 only one time, but the frequency column indicates that there are actually two values of $X = 10$
        4. The highest score is $X = 10$, and the lowest score is $X = 4$
        5. No one had a score of X = 5
        6. Plot the score vs frequency on a Bar graph for a visual understanding of the frequency distribution:

---

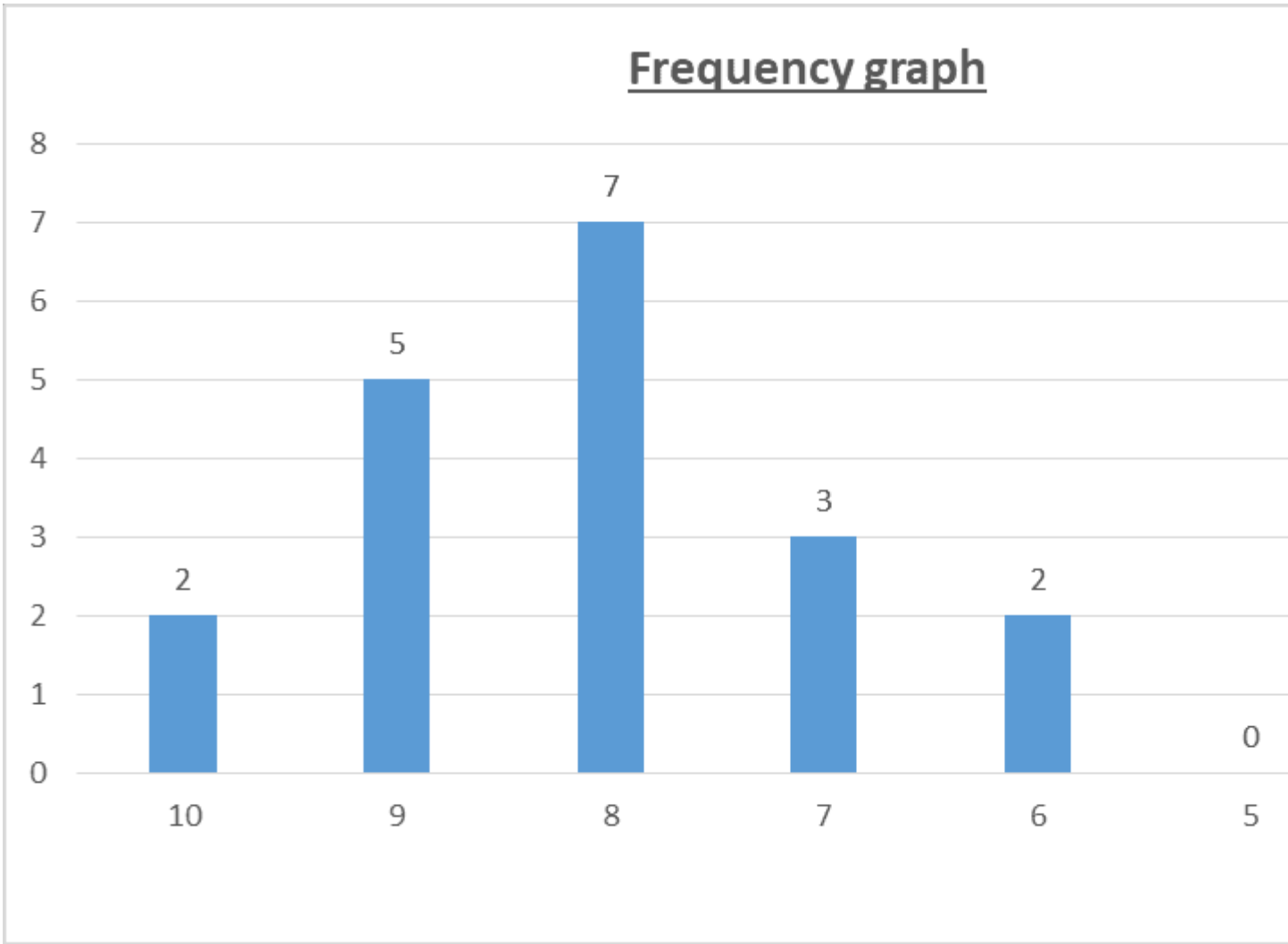
Frequency Bar Graph

1.

    1.

        1. The frequency of each object is calculated
        2. And a Bar Graph is plotted – object vs Frequency
        3. On the x-axis: Objects

4. On the y-axis: Frequencies



```
e/docker-python
# For example, here's several helpful packages to load in

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
import matplotlib.pyplot as plt
# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will
list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# Any results you write to the current directory are saved as output.
```

## 2.Load the file

In [2]:
```
income_df=pd.read_csv("/kaggle/input/incomeexpenditure-dataset/Inc_Exp_Data.csv")
```

In [3]:
```
income_df.head()
```

Out[3]:

| | Mthly_HH_Income | Mthly_HH_Expense | No_of_Fly_Members | Emi_or_Rent_Amt | Annual_HH_Income |
|---|---|---|---|---|---|
| 0 | 5000 | 8000 | 3 | 2000 | 64200 |
| 1 | 6000 | 7000 | 2 | 3000 | 79920 |
| 2 | 10000 | 4500 | 2 | 0 | 112800 |
| 3 | 10000 | 2000 | 1 | 0 | 97200 |
| 4 | 12500 | 12000 | 2 | 3000 | 147000 |

## 3.Analyze the data

In [4]:
```
income_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 7 columns):
Mthly_HH_Income             50 non-null int64
Mthly_HH_Expense            50 non-null int64
No_of_Fly_Members           50 non-null int64
Emi_or_Rent_Amt             50 non-null int64
Annual_HH_Income            50 non-null int64
Highest_Qualified_Member    50 non-null object
No_of_Earning_Members       50 non-null int64
dtypes: int64(6), object(1)
memory usage: 2.9+ KB
```

In [5]:
```
income_df.shape
```

Out[5]:
```
(50, 7)
```

In [6]:
```
income_df.describe().T
```

Out[6]:

| | count | mean | std | min | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|
| Mthly_HH_Income | 50.0 | 41558.00 | 26097.908979 | 5000.0 | 23550.0 | 35000.0 | 50375.0 |
| Mthly_HH_Expense | 50.0 | 18818.00 | 12090.216824 | 2000.0 | 10000.0 | 15500.0 | 25000.0 |
| No_of_Fly_Members | 50.0 | 4.06 | 1.517382 | 1.0 | 3.0 | 4.0 | 5.0 |
| Emi_or_Rent_Amt | 50.0 | 3060.00 | 6241.434948 | 0.0 | 0.0 | 0.0 | 3500.0 |
| Annual_HH_Income | 50.0 | 490019.04 | 320135.792123 | 64200.0 | 258750.0 | 447420.0 | 594720.0 |
| No_of_Earning_Members | 50.0 | 1.46 | 0.734291 | 1.0 | 1.0 | 1.0 | 2.0 |

In [7]:
```
income_df.isna().any()
```

Out[7]:
```
Mthly_HH_Income             False
Mthly_HH_Expense            False
No_of_Fly_Members           False
Emi_or_Rent_Amt             False
Annual_HH_Income            False
Highest_Qualified_Member    False
No_of_Earning_Members       False
dtype: bool
```

No null values in the dataset

## 4.What is the Mean Expense of a Household?

In [8]:
```
income_df["Mthly_HH_Expense"].mean()
```

Out[8]:
```
18818.0
```

## 5.What is the Median Household Expense?

In [9]:
```
income_df["Mthly_HH_Expense"].median()
```

Out[9]:
```
15500.0
```

## 6.What is the Monthly Expense for most of the Households?

In [10]:
```
mth_exp_tmp = pd.crosstab(index=income_df["Mthly_HH_Expense"], columns="co
unt")
mth_exp_tmp.reset_index(inplace=True)
mth_exp_tmp[mth_exp_tmp['count'] == income_df.Mthly_HH_Expense.value_count
s().max()]
```

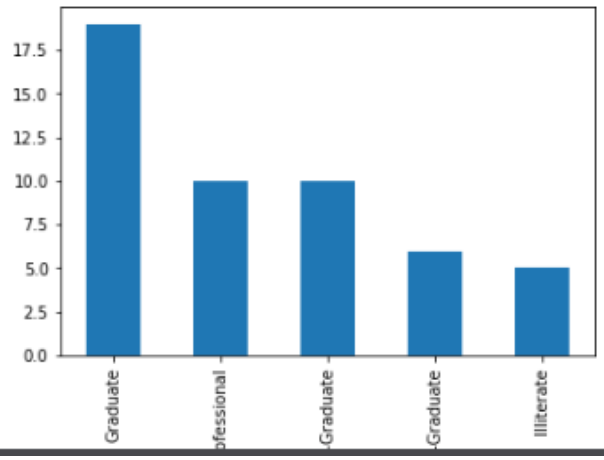
Out[10]:

| col_0 | Mthly_HH_Expense | count |
|---|---|---|
| 18 | 25000 | 8 |

## 7.Plot the Histogram to count the Highest qualified member

In [11]:
```
income_df["Highest_Qualified_Member"].value_counts().plot(kind="bar")
```

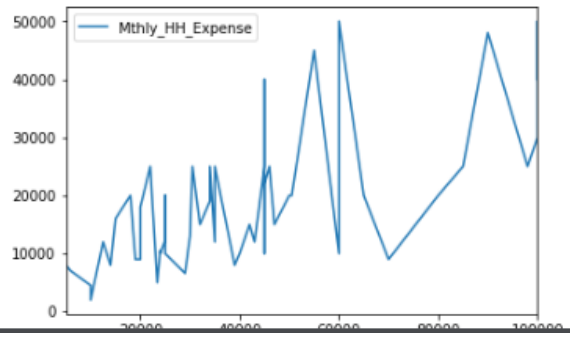
Out[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fec6df5dc88>
```



8.Calculate IQR(difference between 75% and 25% quartile)

In [12]:
```
income_df.plot(x="Mthly_HH_Income", y="Mthly_HH_Expense")
IQR=income_df["Mthly_HH_Expense"].quantile(0.75)-income_df["Mthly_HH_Expense"].quantile(0.25)
IQR
```

Out[12]:
```
15000.0
```

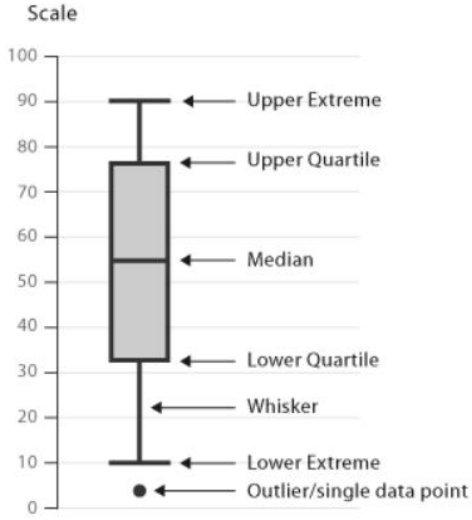



**Data visualisation examples**

The best visual representation of a data set is determined by the relationship data scientists want to convey between data points. Do they want to present the distribution with outliers? Do they want to compare multiple variables or analyse a single variable over time? Are they presenting trends in your data set? Here are some of the key examples of data visualisation.

- A **bar chart** is used to compare two or more values in a category and how multiple pieces of data relate to each other.
- A **line chart** is used to visually represent trends, patterns and fluctuations in the data set. Line charts are commonly used to forecast information.
- A **scatter plot** is used to show the relationship between data points in a compact visual form.
- A **pie chart** is used to compare the parts of a whole.
- A **funnel chart** is used to represent how data moves through different steps or stages in a process.
- A **histogram** is used to represent data over a certain time period or interval.

- **Box and Whisker** Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartiles.




**Box plots are useful as they show the average score of a data set**

  - The median is the average value from a set of data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value, and half are less.

**Box plots are useful as they show the skewness of a data set**

  - The box plot shape will show if a statistical data set is normally distributed or skewed.

  - **Skewness** is the degree of asymmetry observed in a probability distribution.
  - Skewness tells you where the outliers occur, although it doesn't tell you how many outliers occur.

- Distributions can be positive and right-skewed, or negative and left-skewed. A normal distribution exhibits zero skewness.
  - Example: Skewness is often found in stock market returns or the distribution of average individual income.

**What Does Skewness Tell Investors?**

  - **Investors commonly use standard deviation to predict future returns**, but the standard deviation assumes a normal distribution. As few return distributions appear normal, skewness is a better measure to base performance predictions.

**1. Types of Skewness**

  - **Positive Skewness:**

**Definition:** The tail on the right side of the distribution is longer or fatter, indicating that there are outliers with higher values.

- **Example:** Consider a tech startup stock that has shown modest returns most of the time but has had a few years with exceptionally high returns due to a breakthrough product. This stock may have a positive skewness, suggesting that while most returns are moderate, there is potential for substantial upside.

### Negative Skewness:

**Definition:** The tail on the left side is longer or fatter, indicating that there are outliers with lower values.

- **Example:** A distressed asset, such as a company facing bankruptcy, might show negative skewness. Most of the time, the returns are slightly negative, but there are occasional very large losses. This suggests a higher risk of significant downturns.

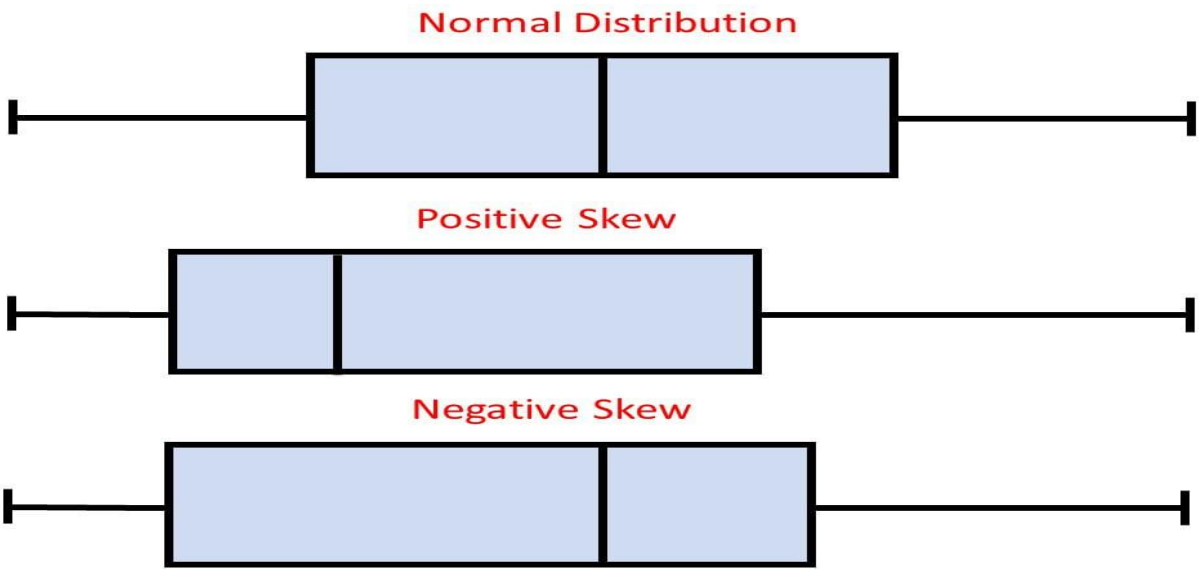
**2. Risk Assessment**

### Positive Skewness:

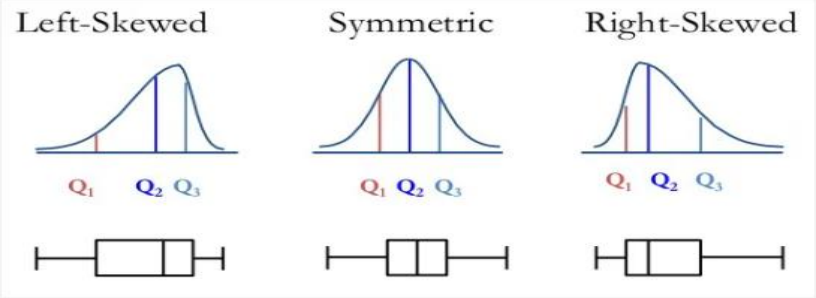
- **Interpretation:** Investments with positive skewness can be appealing as they offer the chance for high returns with lower probabilities of large losses.
- **Example:** A venture capital fund investing in startups may exhibit positive skewness. While most startups may fail (small losses), a few might achieve extreme success, leading to large gains for the fund.

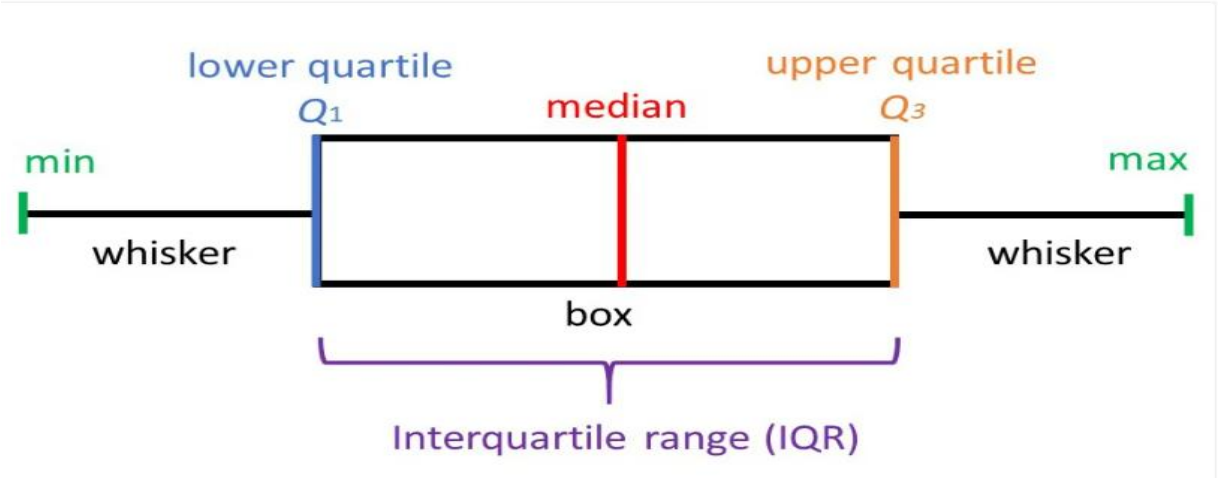
### Negative Skewness:

- **Interpretation:** Investments with negative skewness are riskier because they indicate a higher likelihood of substantial losses.
- **Example:** A high-yield bond fund may show negative skewness if it frequently experiences defaults. While investors may receive regular interest payments, the risk of a few large defaults can lead to significant losses.




- When the **median** is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution **is positively skewed (skewed right).**
- When the **median** is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is **negatively skewed (skewed left).**



Formula for Pearson's Skewness

$$\frac{Sk_1 = \frac{X - Mo}{s}}{Sk_2 = \frac{3(X - Md)}{s}}$$

**where:**

$Sk_1$ = Pearson's first coefficient of skewness and $Sk_2$ the second

$s$ = The standard deviation for the sample

$\bar{X}$ = Is the mean value

$Mo$ = The modal (mode) value

$Md$ = Is the median value

Pearson's first coefficient of skewness is used if the data exhibit a strong mode. Pearson's second coefficient may be preferable if the data have a weak mode or multiple modes, as it does not rely on mode as a measure of central tendency.

**Here are the types of observations one can make from viewing a Box Plot:**

- What the key values are, such as: the average, median, 25th percentile, etc.
- If there are any outliers and what their values are.
- If the data is symmetrical or not.
- How tightly is the data grouped.
- If the data is skewed and if so, in what direction.