AGENDA

01 | Recap

02 | Measures of Spread

03 | Measures of shape and Visuals

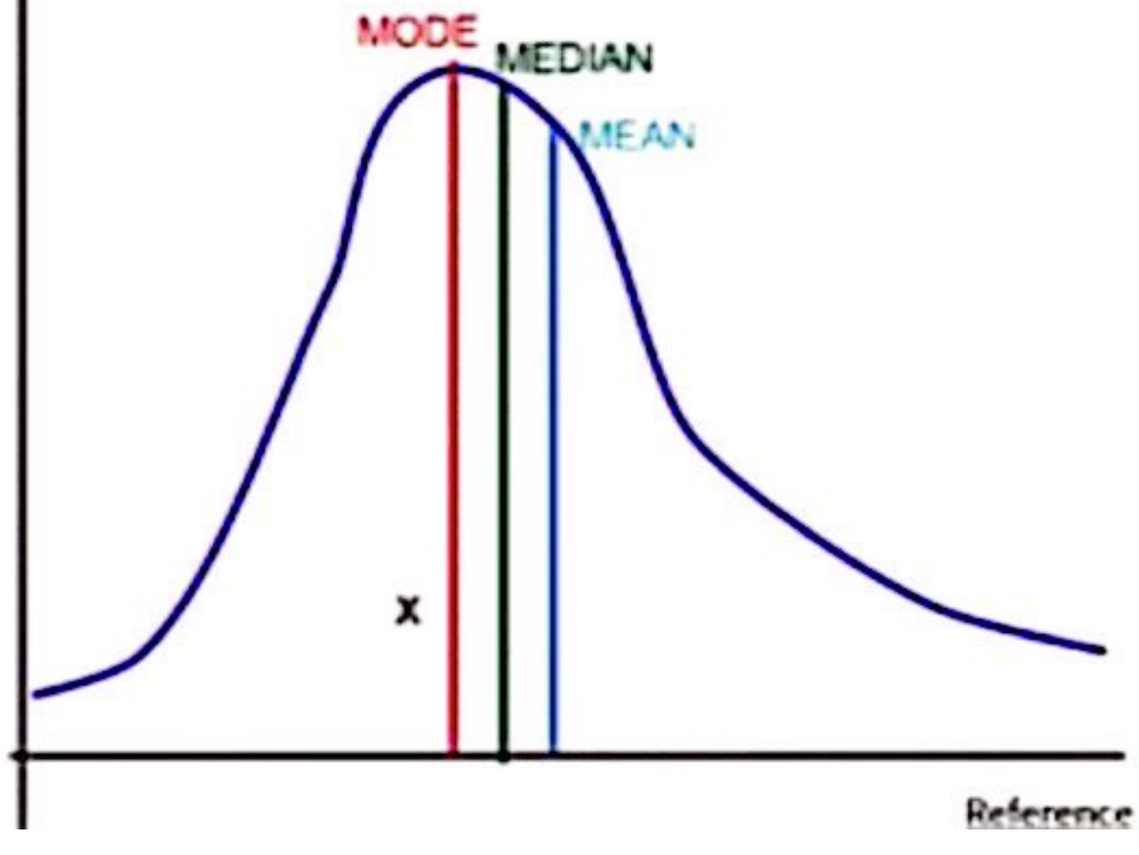
04 | Exploratory Data Analysis

Developer Student Clubs
Al-Azhar University

Measures of location (Central Tendency)

It's the measures that describes the **central** value of a data set, And Its most popular forms are:

1. Mean
2. Median
3. Mode



Developer Student Clubs
Al-Azhar University

Summary statistics types

- A sample summary is a **Statistic**, A population summary is a **Parameter**.

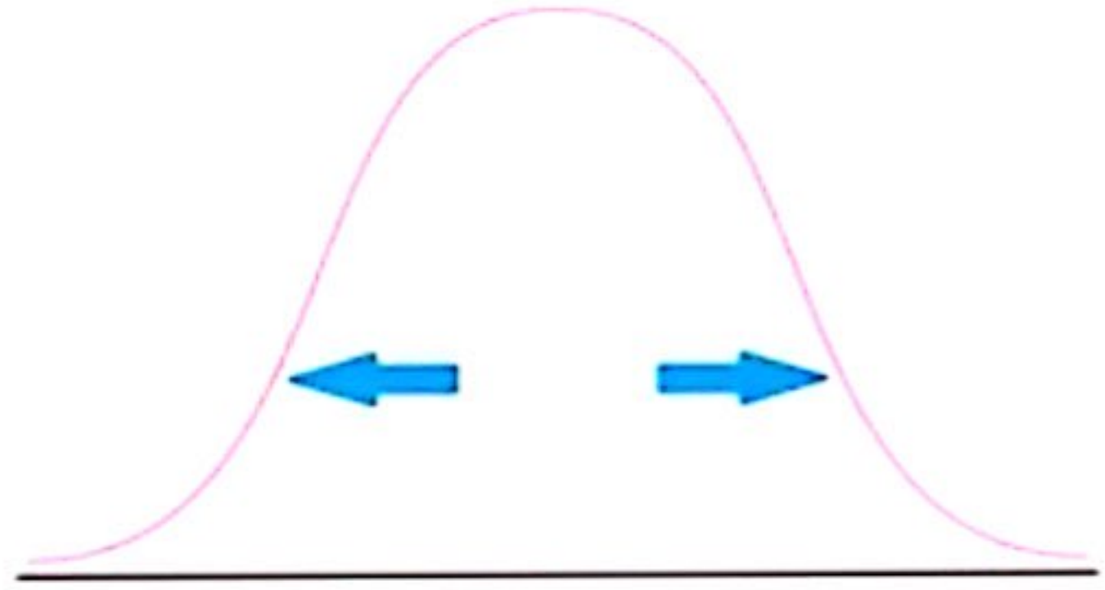
- We can summarize our data with different measures. Each of them adds a certain power to the analysis.

1. Measures of location (Mean, Median, mode)
2. Measures of spread (Min, Max, Variance and Standard Deviation)
3. Measures of shape (Skewness and Kurtosis)

Measures of Spread

Describe how dispersed or varied data is. We can see measures of spread in these forms:

1. Ranges (Maximum - Minimum)
2. IQR
3. Variance
4. Standard Deviations

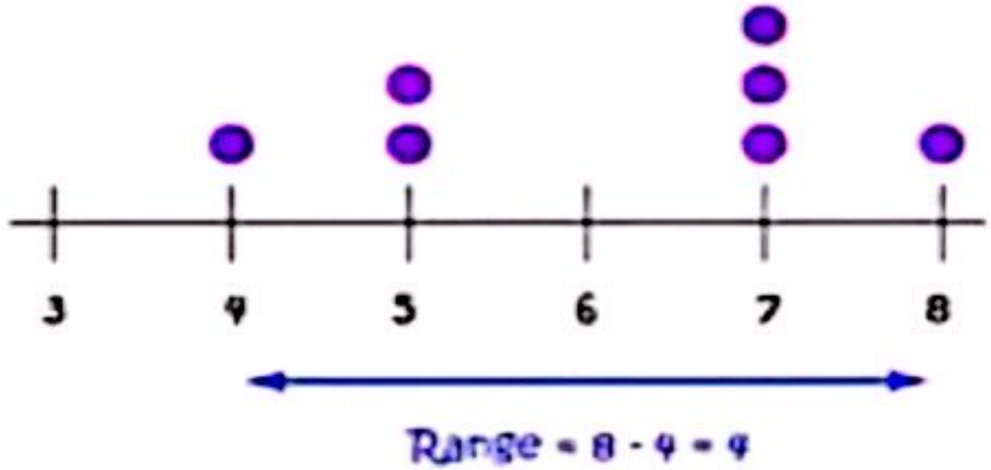


Developer Student Clubs
Al-Azhar University

Ranges

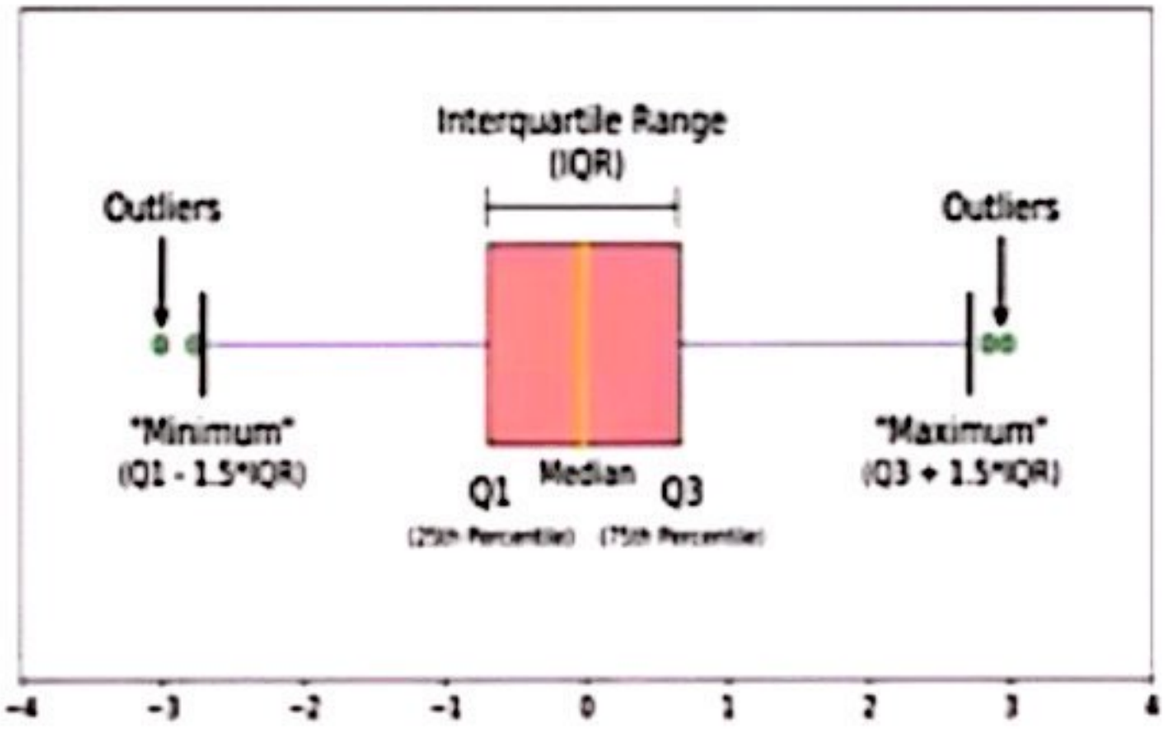
The range is the **simplest** measure of variability to compute, It's simply the **difference between the Maximum Value and the minimum.**

1. It provides a simple view of how varied our data set is.
2. The range can also be used to **estimate** the standard deviation (**The Range Rule**).
3. It's highly sensitive to **outliers**.
4. The range also tells us nothing about the **internal features** of our data set.



The Interquartile range

- We can use the box-plot and the IQR value to represent the variability.
- Remember that **IQR = Q3 – Q1**
- The problem is that IQR doesn't represent the **whole dataset** (Just 50% of it), so it won't be an accurate representation of the dataset's variability.
- It also doesn't tell us nothing about the **internal features** of our data set just like the Range.
- We need to include all of our data sets in the computation … How ?



Measuring variability using all data points

Ranges doesn't give the full picture, so we have to use all of our data points' values to represent the dataset variability. We can think of multiple ways.

1. Find the average distance between any two value.
    - It's inefficient due to the big amount values

2. Find the average between every minimum and maximum value.
    - It's inefficient due to the lack of standardization.

3. Find the average distance between every value and the mean value.
    - And that is the most effective way.

Variance

Variance (σ^2) a measurement of the spread between each number and the average numbers in the dataset.

- To calculate the **population variance** (σ^2), we take the sum of distances between each data point and the population mean divided by the total number of data points.

- In case of **sample variance**, there is a tiny difference. That we divide by the total number of data points in the sample – 1 (n-1). And that is called **Bessel's Correction.**

- This correction is made to correct for the fact that these **sample statistics tend to underestimate** the actual parameters found in the population.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

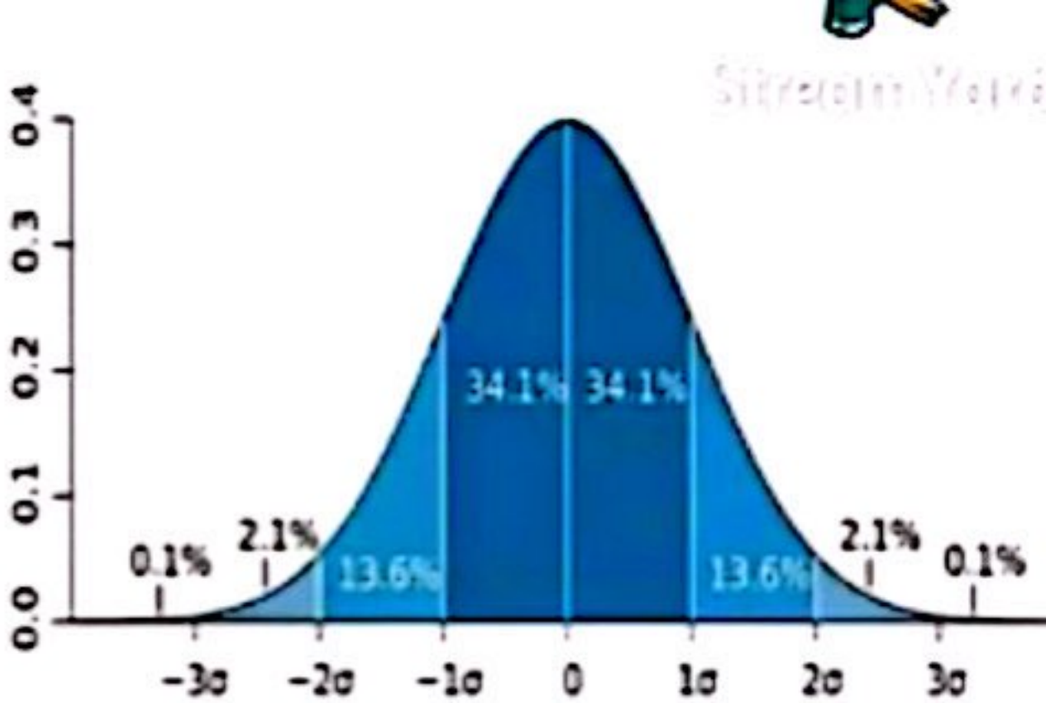
$$\sum_{1}^{2} (x_i - \bar{x})^2$$

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

Standard Deviation

Standard Deviation (σ) is the most common measure of spread for its robustness and unified measurements.

- It's the same calculations of the Variance except we take the **squared root** of the variance output.
- In a **standardized distribution**, we found that:
  - **68%** of data lies in **1σ** from the mean.
  - **95%** lies in **2σ** from the mean.
- So with standard deviation we can know how much data falls in one area. And that helps us making our assumptions and tests.



$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

Statistics Summary

$33

The middle price
(Median)

$28

Avg. price (Mean)

$50

Most Frequent Price
(Mode)

$7

Standard Deviation

$452

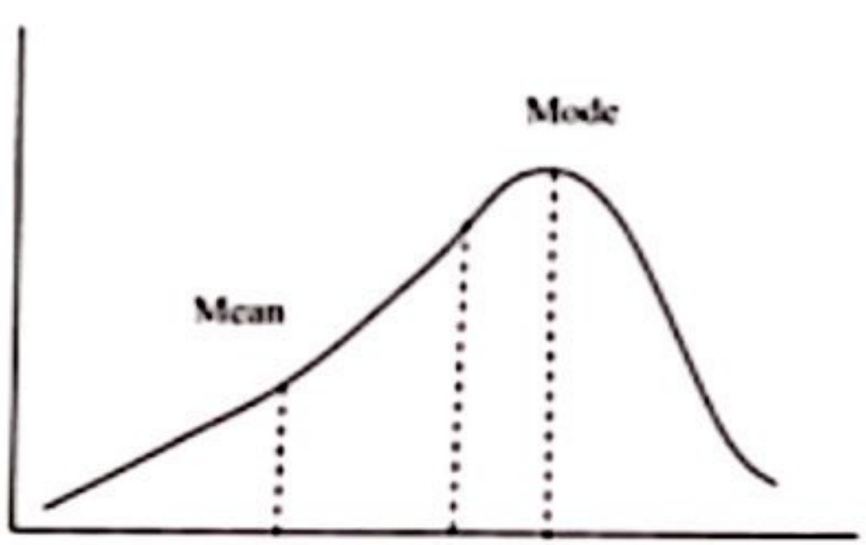
Maximum price

Developer Student Clubs
Al-Azhar University

Measures of Shape

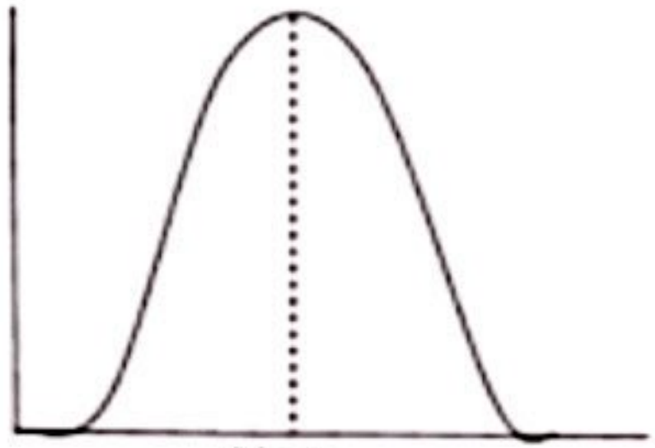
Developer Student Clubs
Al-Azhar University

Skewness

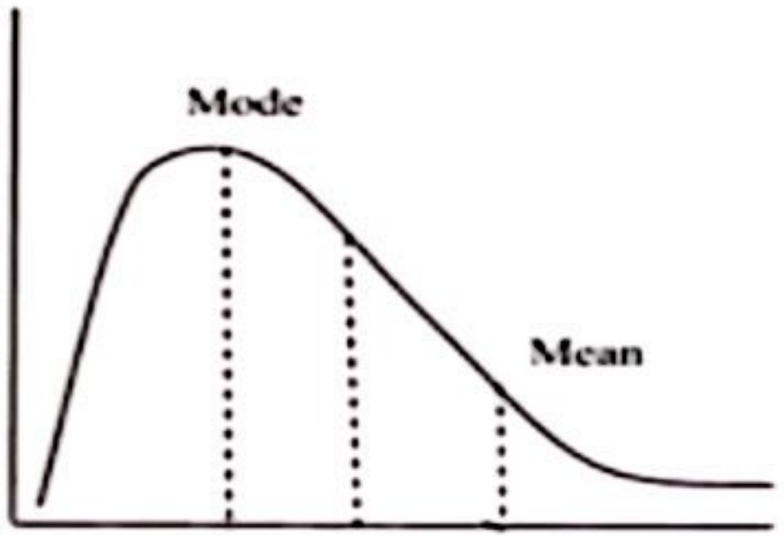


mean<median<mode

Negative direction



Symmetrical data
mean=median=mode



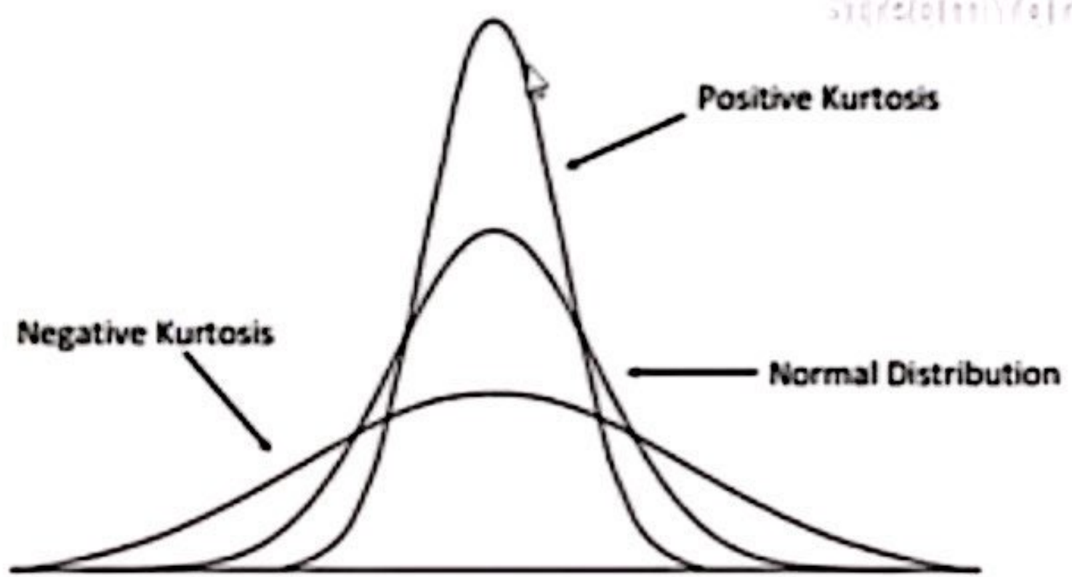
mode<median<mean

Positive direction

Kurtosis

Kurtosis is a statistical measure used to describe the degree to which scores cluster in the tails or the peak of a frequency distribution.

- **Low kurtosis** is an indicator that data has light tails or **lack of outliers**. It's called **(PlatyKurtic)**



- Datasets with **high kurtosis** tend to have heavy tails, and it indicates to **outliers.** It's called **(LeptoKurtic)**

- It's closely related to skewness as both represents measures of distribution shapes.

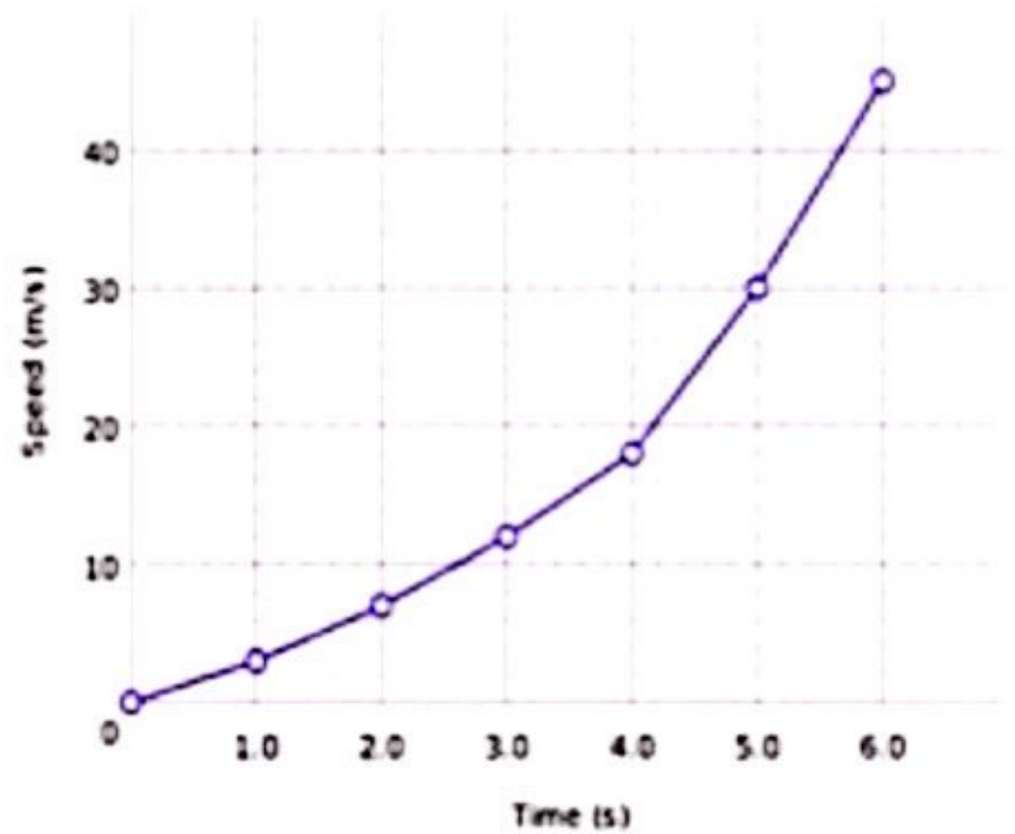
- **Read More**

The power of Visualization

Developer Student Clubs
Al-Azhar University

Line Plot

- A **line plot** is a **graph** that shows the frequency of data occurring along a number **line (Usually a timeline)**.
- Line graphs are used to track changes **over short and long periods of time.**
- Line graphs can also be used to compare changes over the same period of time for **more than one group**.
- It's very easy to interpret and use.

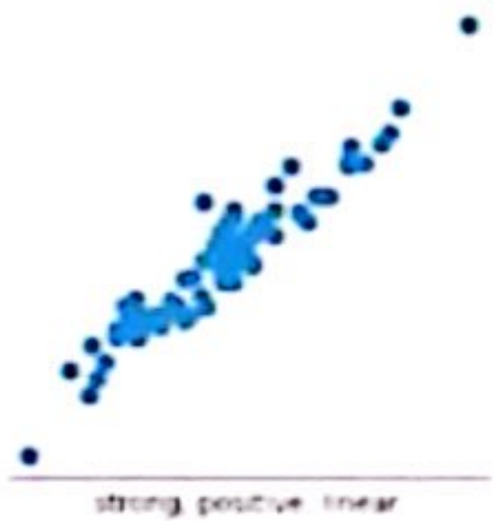


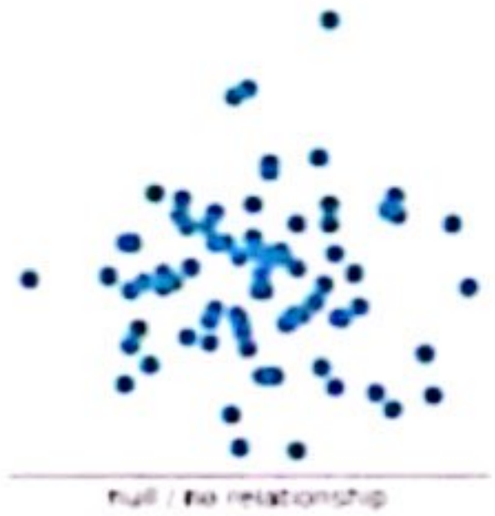
Scatter Plot

- Scatter Plot is a type of plot or mathematical diagram using **Cartesian coordinates** to display values for typically **two variables** for a set of data.

- Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show **how much one variable is affected by another. (Correlation test)**

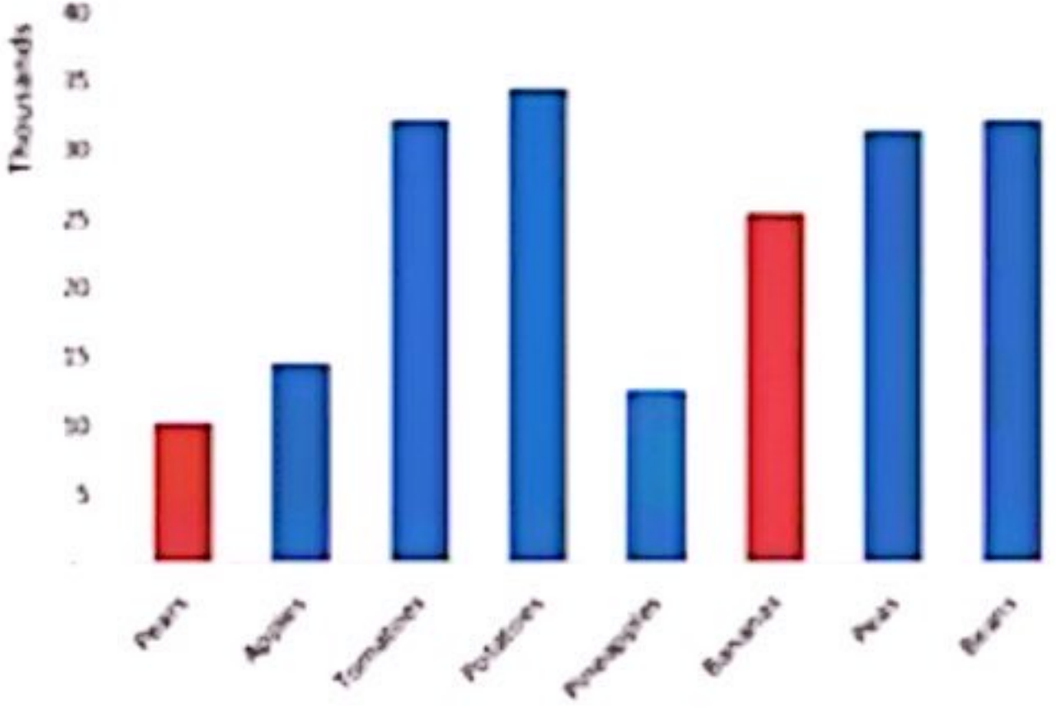
- There are many types of correlation relationships that scatter plots can reveal.





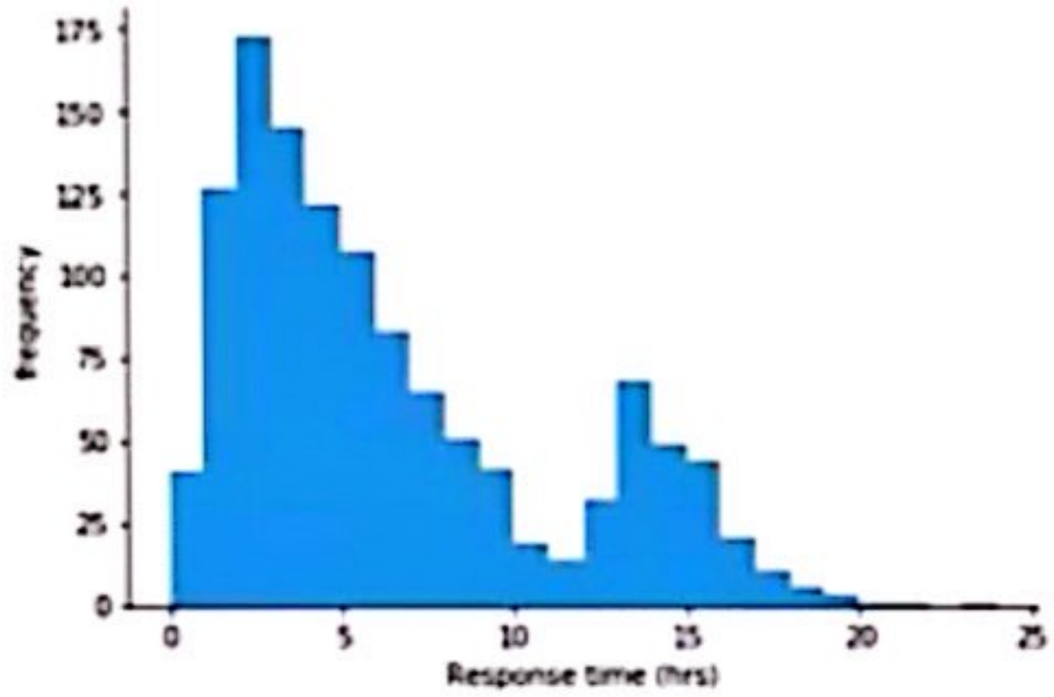
Bar Chart

- A bar chart or bar graph is a chart or graph that presents **categorical data** with rectangular bars.
- Bar graphs are used to compare things between different groups or to track changes over time.
- When trying to measure change over time, **bar graphs** are best when the changes are **larger**. But line plots are more suitable for visualizing changes over time.
- It can be graphed horizontally or vertically.



Histogram

- Histograms are the most frequently used chart in frequency distribution.
- Frequency distribution shows how often each different value in a set of data occurs.
- Taller bars show that more data falls in that range.
- A histogram displays the shape and spread of continuous sample data.



Further Reading