Descriptive Statistics

Agenda

- What is Statistics ?
- Population Vs. Sample
- Descriptive Statistics
- Univariate Analysis
- Measures of Center (Central Tendency)
- Five Number Summary
- Measure of Spread (Dispersion)
- Shape of Data
- Normal Distribution
- Outliers & Boxplot
- Covariance & Correlation

What is Statistics ?

What is Statistics ?

- *The science of collecting, analyzing, presenting, and interpreting data.*
- *To get information from Data.*

Population Vs. Sample

Source : datatab.net

Population Vs. Sample

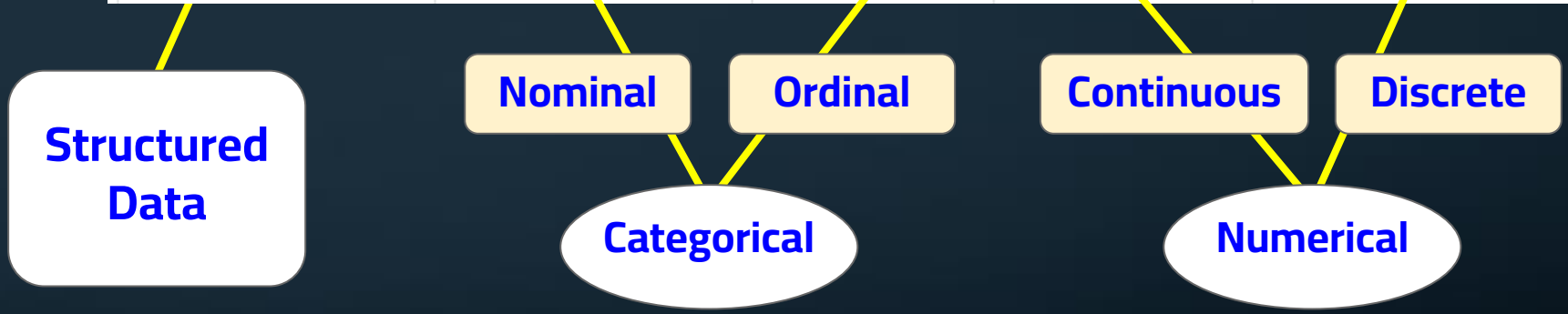
- ***Population*** *is the entire group that you want to draw conclusions about.*
- ***Sample*** *is the specific representative group that you will collect data from.*

Descriptive Vs. Inferential

- ***Descriptive statistics*** *are used to describe the characteristics or features of a dataset (also known as 'summary statistics')*
- ***Inferential statistics*** *focus on making generalizations about a larger population based on a representative sample of that population.*

Descriptive Statistics

Employee	Position	Level	Salary	# supervision
Ahmad	HR Specialist	A	6500	2
Mohamed	Accountant	B	7500	1
Sayed	Lawyer	C	8450	0
Alaa	Engineer	B	9150	3
Adel	Trainer	C	4757	0
Osama	Sales	A	7546	2



Employee	Position	Level	Salary	# supervision
Ahmad	HR Specialist	A	6500	2
Mohamed	Accountant	B	7500	1
Sayed	Lawyer	C	8450	0
Alaa	Engineer	B	9150	3
Adel	Trainer	C	4757	0
Osama	Sales	A	7546	2

Analysis

Bivariate

Univariate

Univariate Analysis

Measures of Center (Central Tendency)

Why Center ?

- *To give one representative number about some feature.*

Mean

The average value

How to find the Mean:

1. Add up all the numbers.
2. Divide the sum by the number of values.

E.g. The mean of 3,2,10,5 is

$$\frac{3+2+10+5}{4} = \frac{20}{4} = 5$$

Median

The middle number

How to find the Median:

1. Put the numbers from smallest to largest.
2. The number in the middle is the median. If there are two middle numbers, add them and divide by two.

Mode

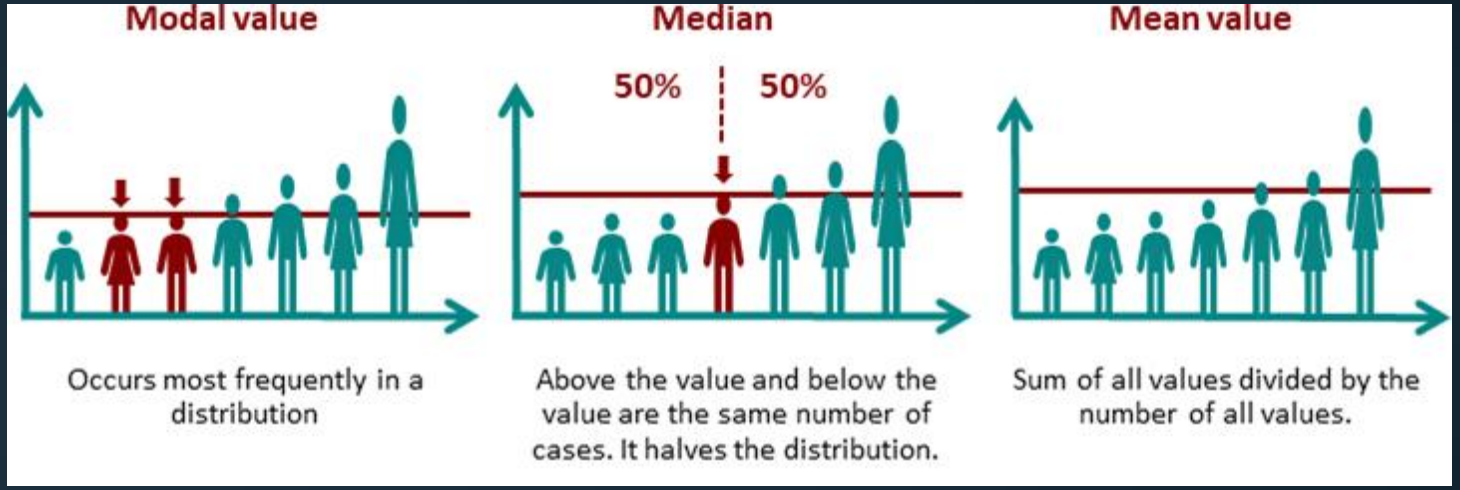
The most frequent number

Special Cases:

- No Mode if all the numbers occur the same amount of times.
- More than one Mode if more than one number is the most frequent.

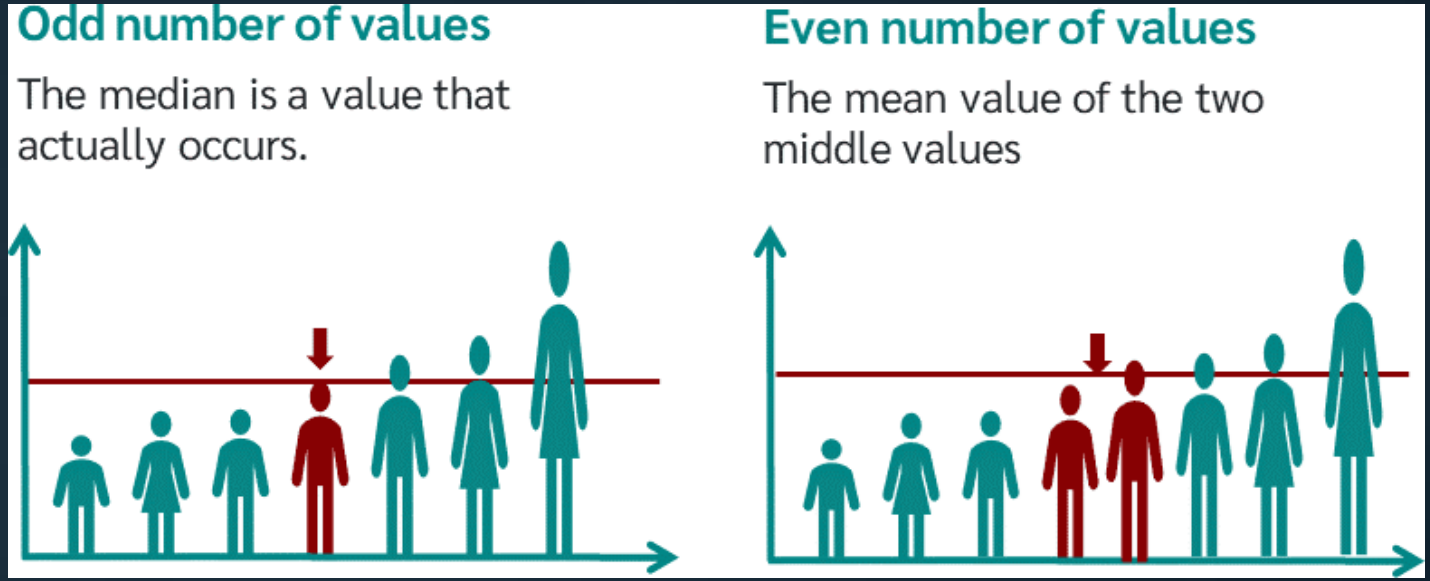
Source : onlinemathlearning.com

Example



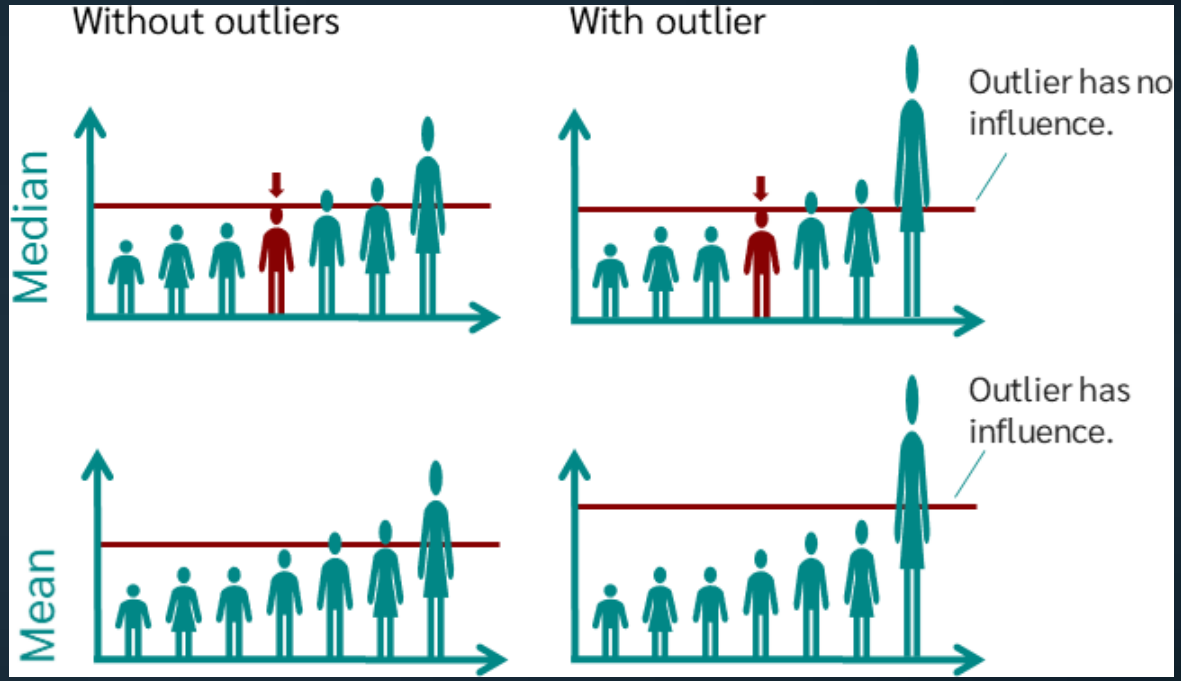
Source : datatab.net

Median for Odd & Even Numbers



Source : datatab.net

Mean Vs. Median



Source : datatab.net

Example

Employee	Salary	Country
Ahmad	6500	Egypt
Mohamed	7500	Iraq
Sayed	8450	UAE
Alaa	9150	Egypt
Adel	8450	Libya
Osama	7500	Egypt

Mean or Median

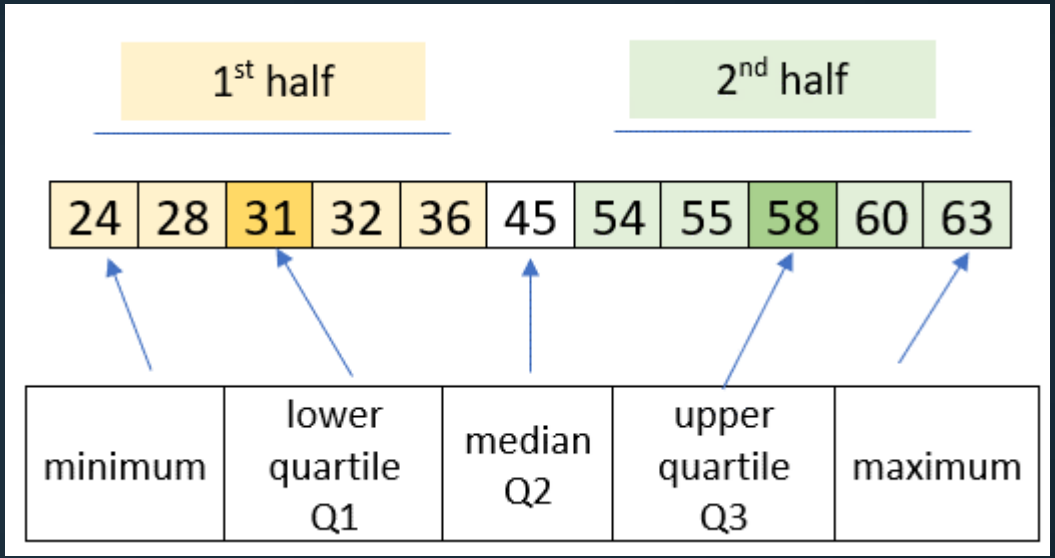
Mode

Five Number Summary

Tell me More!

1. Minimum: The smallest number in the dataset.
2. Q1 : The value such that 25% of the data fall below.
3. Q2 : The value such that 50% of the data fall below.
4. Q3 : The value such that 75% of the data fall below.
5. Maximum: The largest value in the dataset

Five Number Summary



Source : math-salamanders

Measure of Spread (Dispersion)

Spread of Data

- The mean can be affected by extreme values.
- Dispersion, or spread of data, is measured in terms of how far the data differs from the center.

Salary	Salary
6500	650
7500	7500
8450	8450
9150	9150
8450	8450
7500	13350

7925 ← Mean → 7925

1270

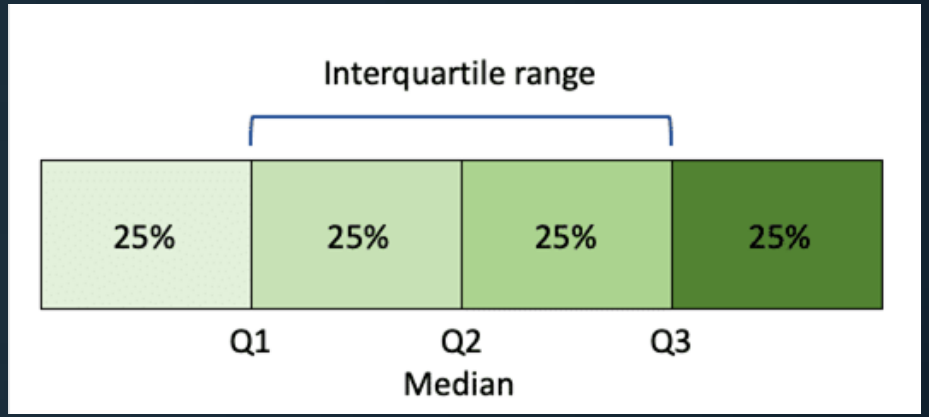
2650 ← Range → 0

Spread of Data

1. Range : max - min
2. Interquartile Range :
   IQR = Q3 - Q1

3. Variance

4. Standard Deviation



Source : scribbr.co.uk

Spread of Data

- ***Variance*** is the average squared difference of each observation from the mean.
- ***Standard deviation*** is the square root of the variance

Variance	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard deviation	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Source : datatab.net

Sample/Population Variance & Standard Deviation

Sample Variance:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{(N-1)}$$

Sample Standard Deviation:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(N-1)}}$$

Population Variance:

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

Population Standard Deviation:

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

Source : Screened-Instructor

Spread of Data

Salary	Mean	Diff. from Mean	Diff. from Mean ^ 2	Diff. from Mean ^ 2 / n-1
6500	7925	-1425	2030625	888750
7500		-425	180625	
8450		525	275625	
9150		1225	1500625	
8450		525	275625	
7500		-425	180625	

Spread of Data

Salary	Salary
6500	650
7500	7500
8450	8450
9150	9150
8450	8450
7500	13350

Mean	7925	7925
Variance	888750	16917750
Std. Deviation	943	4113
Range	2650	12700
IQR	950	1238

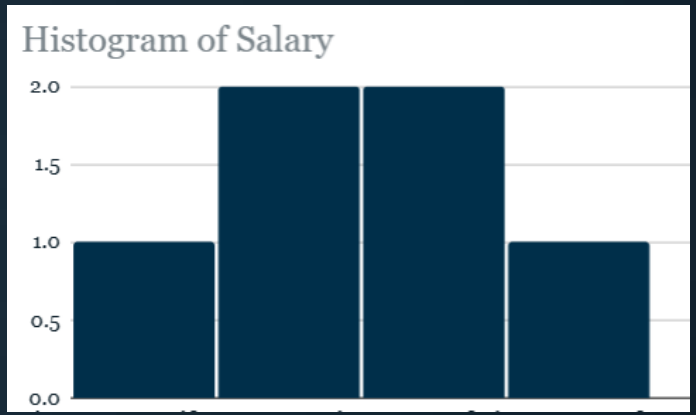
Shape of Data

Histogram

Salary
6500
7500
8450
9150
8450
7500

Frequency Table

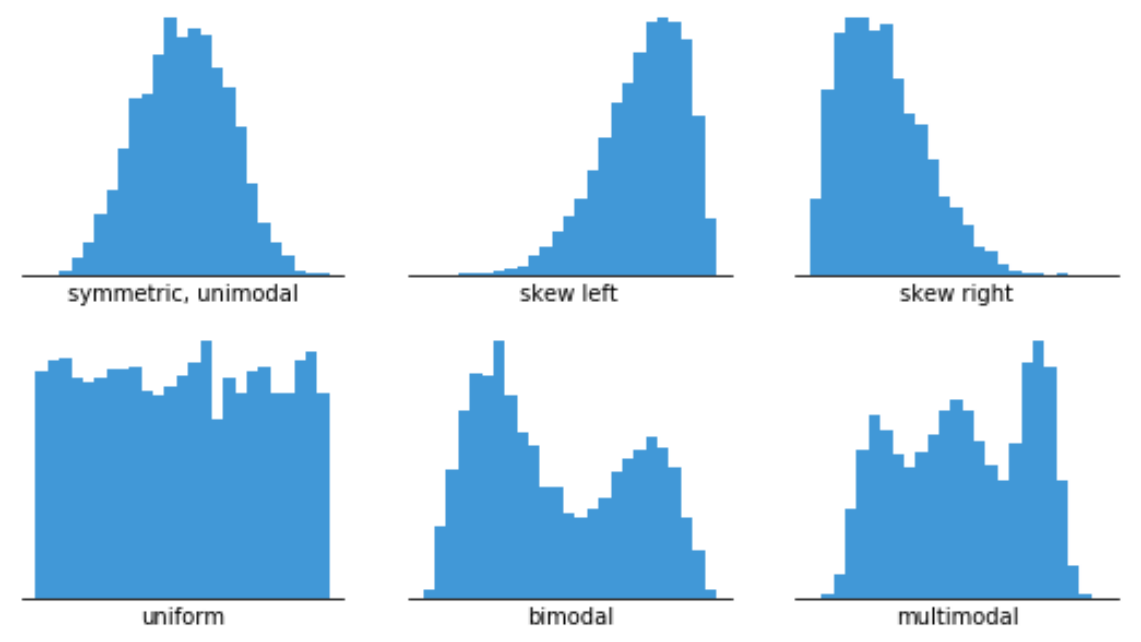
Groups	Frequency
6000-7000	1
7000-8000	2
8000-9000	2
9000-10000	1



Histogram

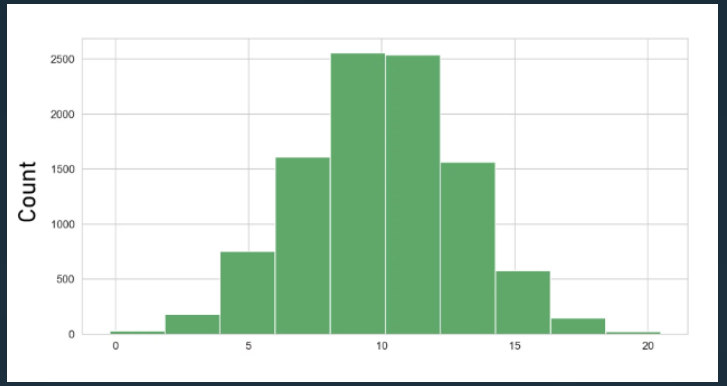
- Each bar typically covers a range of numeric values called a bin or class
- A bar’s height indicates the frequency of data points
- Histograms are good for showing general distributional features of dataset variables.
- You can see roughly where the peaks of the distribution are, whether the distribution is skewed or symmetric, and if there are any outliers.

Histogram

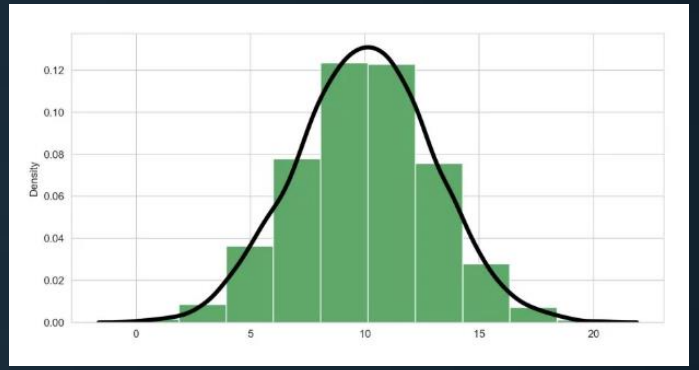


Source : chartio.com

Density Plot

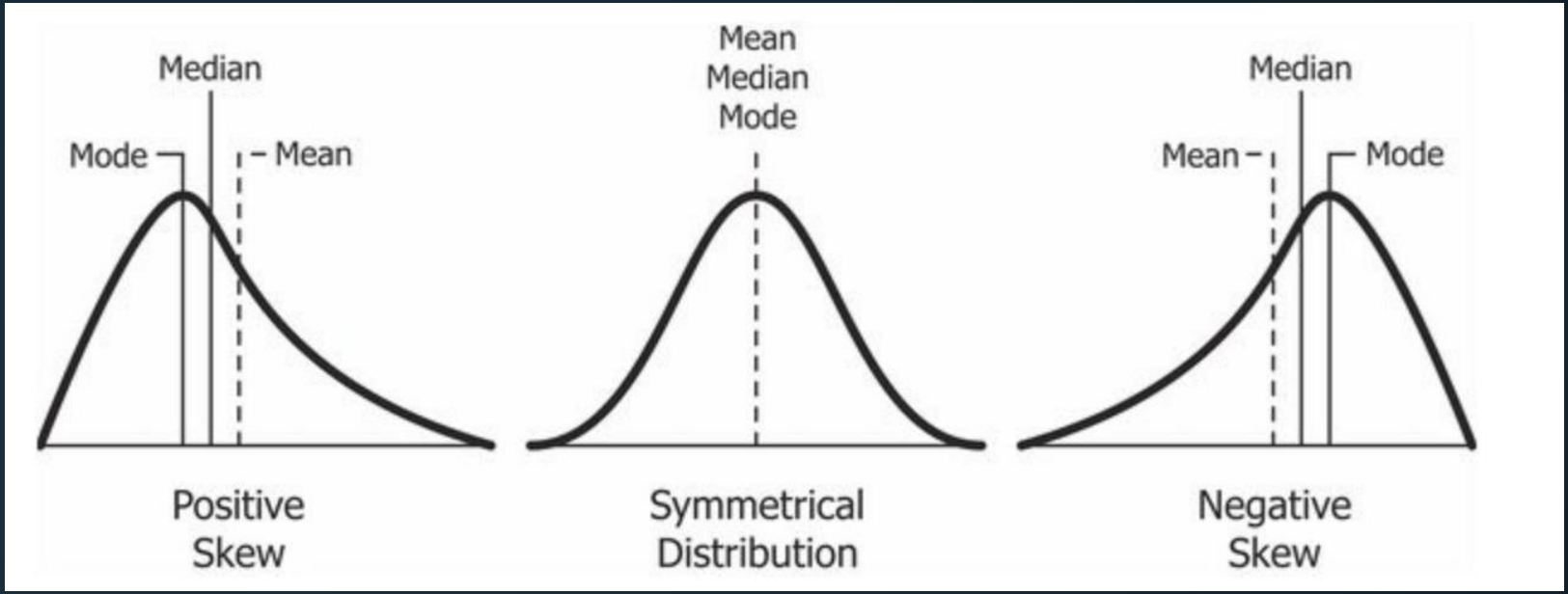


continuous and
smoothed version



Source : askpython

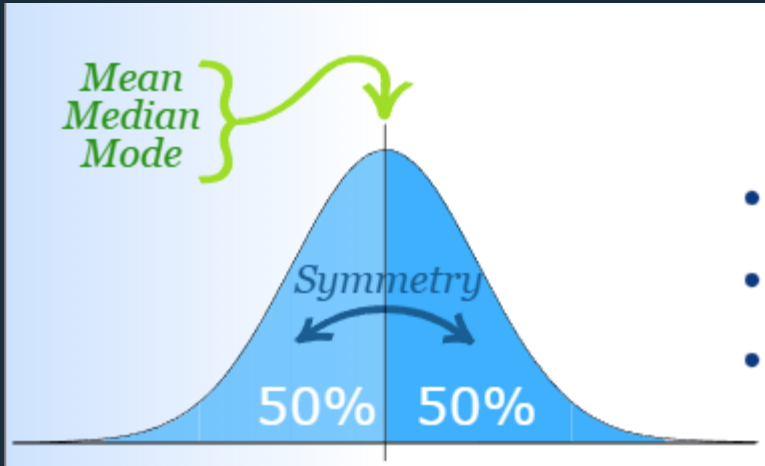
Skewness



Source : analyticsvidhya

Normal Distribution

Normal Distribution

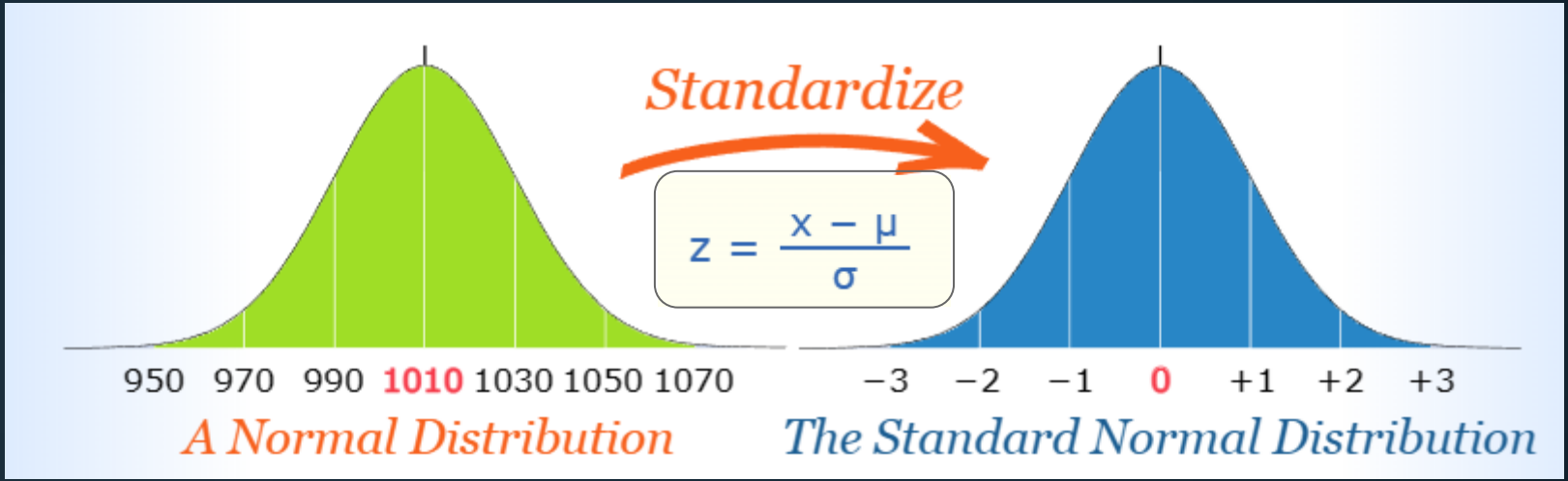


The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

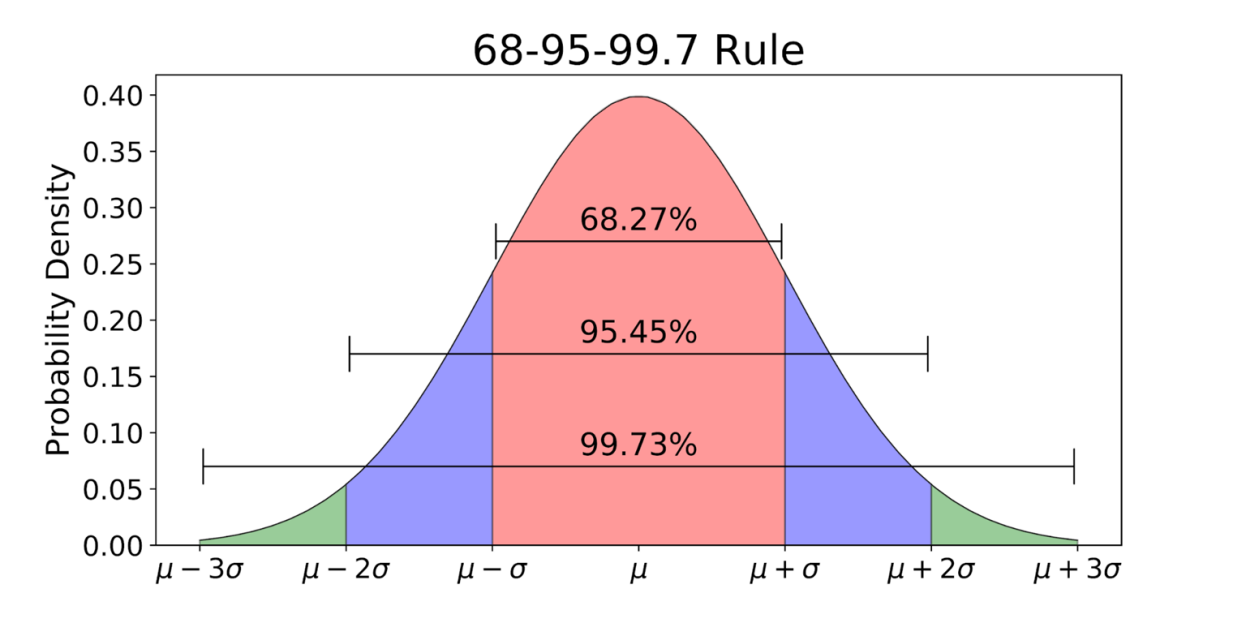
Source : mathsisfun

Standard Normal Distribution



Source : mathsisfun

Empirical Rule

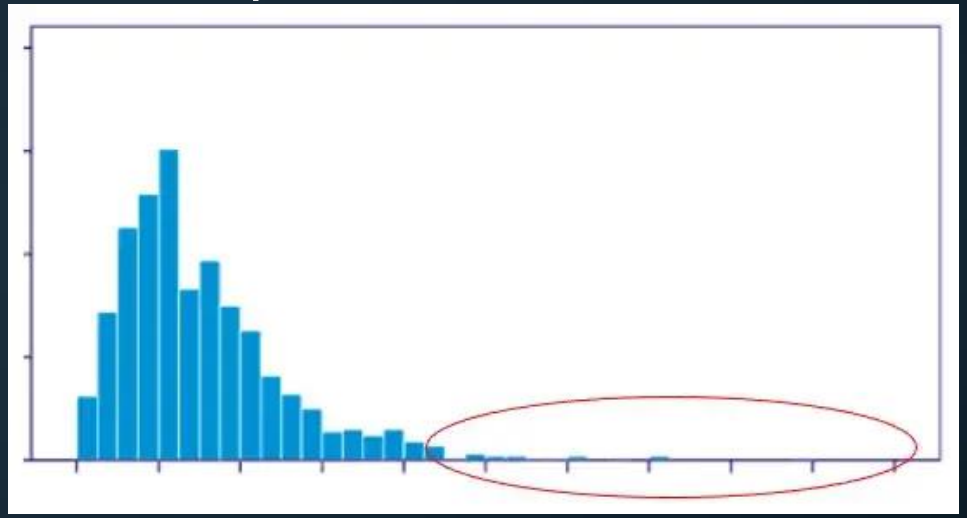


Source : algaestudy.

Outliers

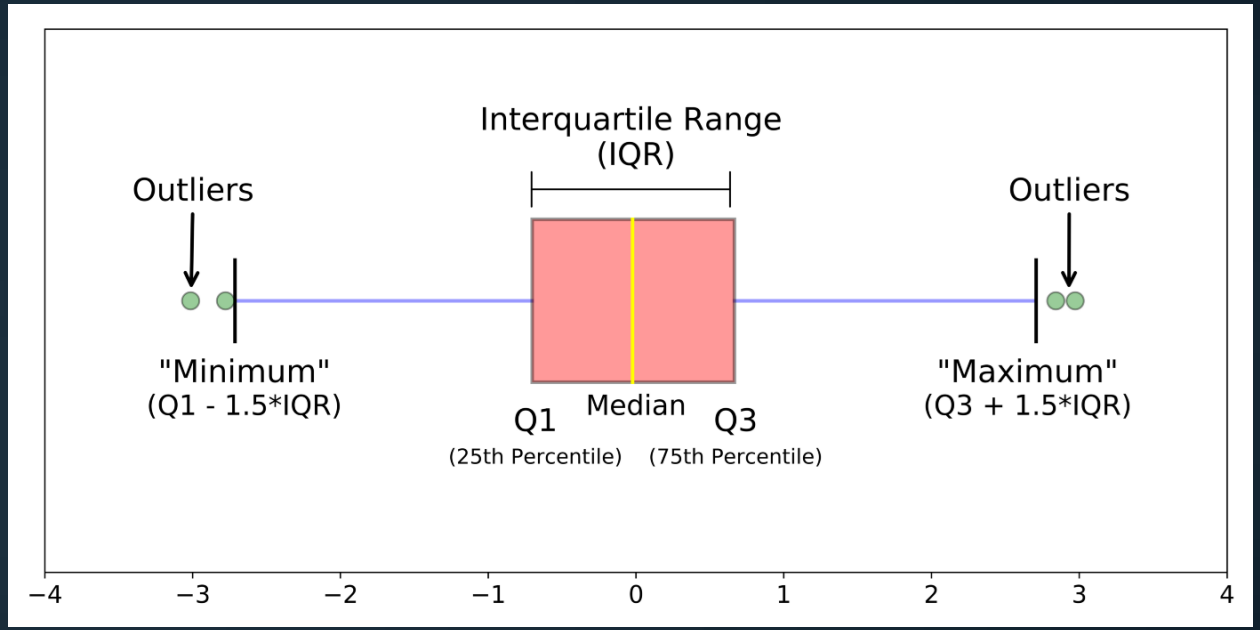
Outliers

- At least note they exist and the impact on summary statistics.
- If typo - remove or fix



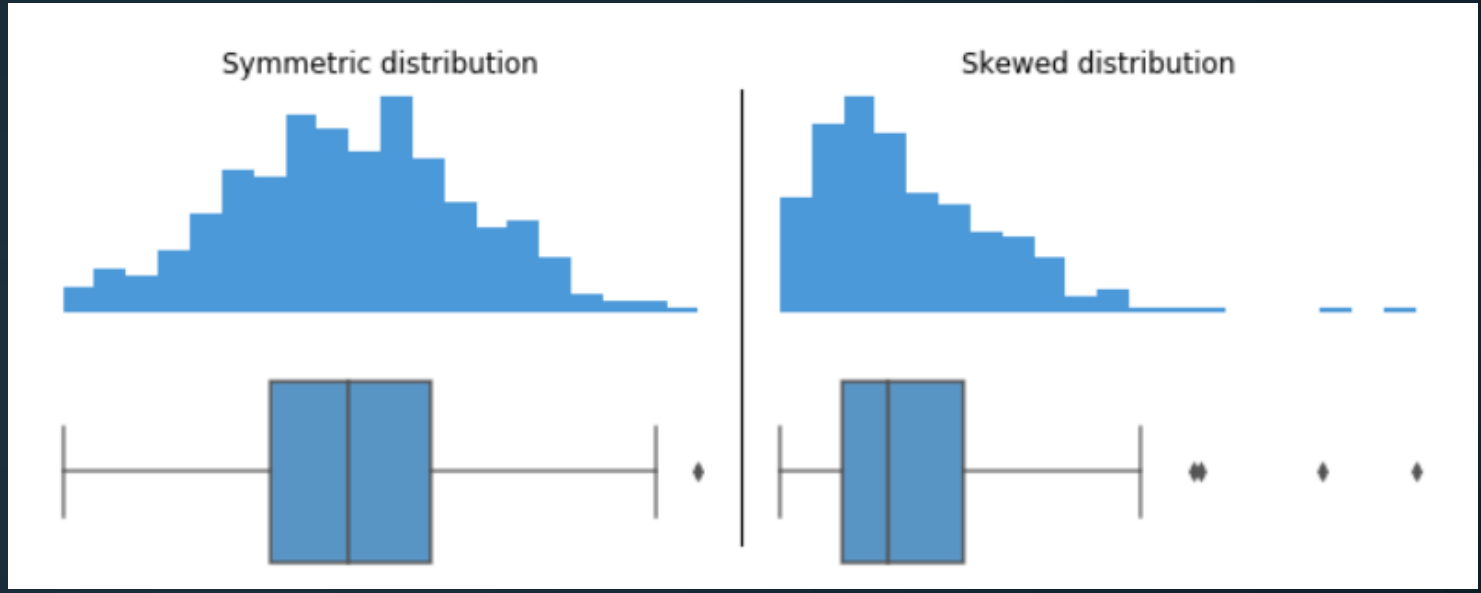
Source : medium

Box Plot



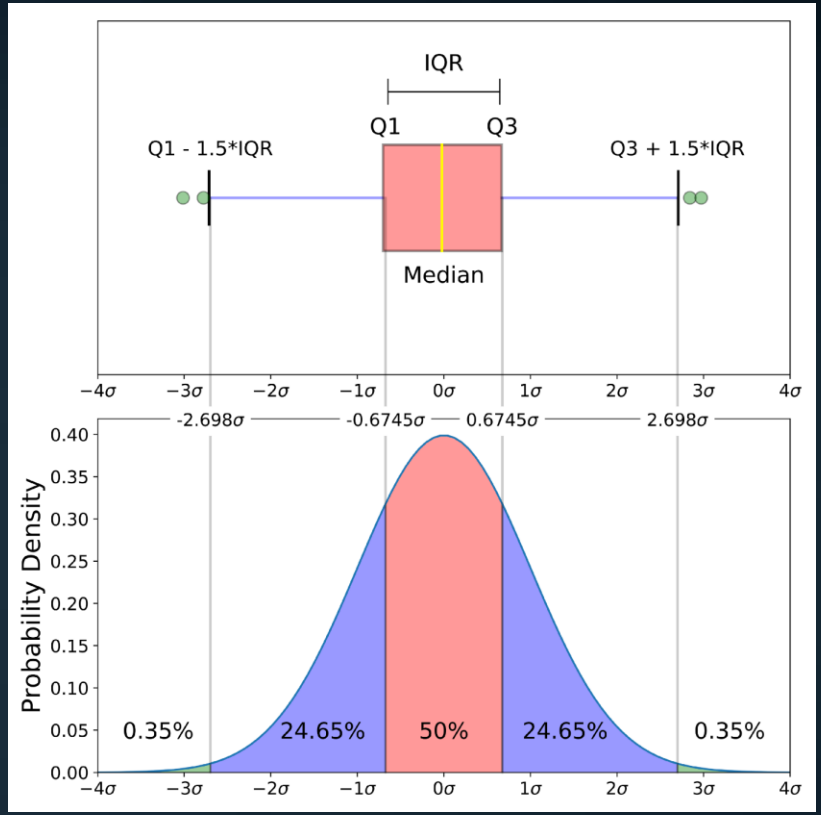
Source : kdnuggets

Box Plot & Histogram



Source : chartio.com

Box Plot



Source : kdnuggets

From Univariate To Bivariate

Measures of Center

- Mean
- Median
- Mode

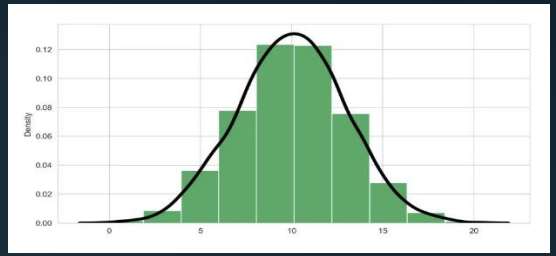
Measures of Spread

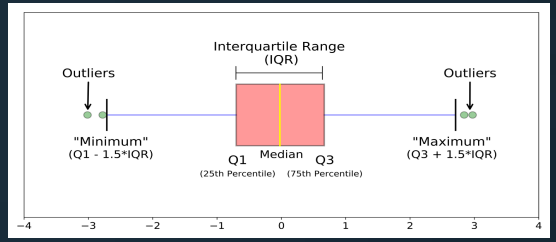
- Range
- IQR
- Var.
- Std.

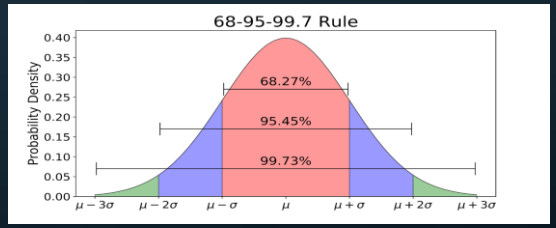
Five Number Summary

- Min
- Q1
- Q2
- Q3
- Max

Shape of Data



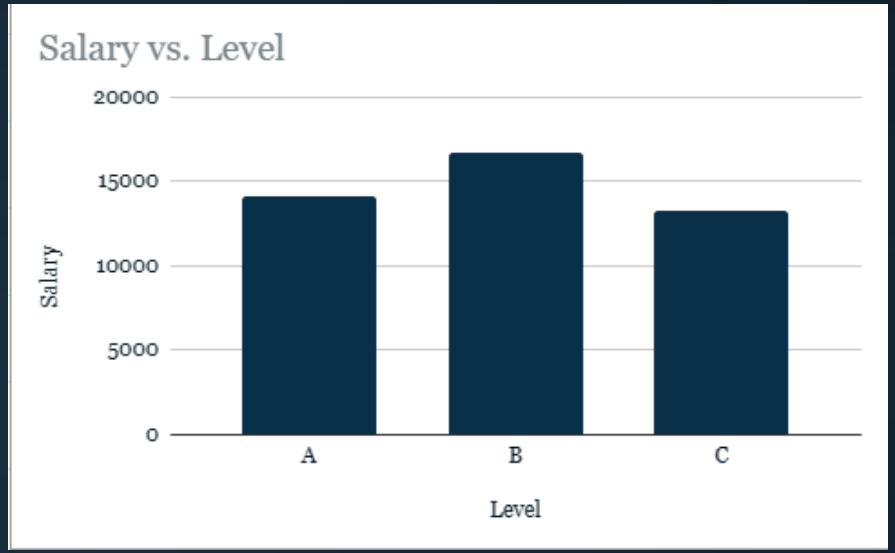




Salary
6500
7500
8450
9150
8450
7500

Bivariate Analysis

Level	Salary
A	6500
B	7500
C	8450
B	9150
C	4757
A	7546



Bar Plot

Bivariate Analysis

Salary	# Experience
6500	1
7500	1.5
8450	2
9150	2.7
8450	2.2
10000	3
6500	2
6500	1.8
8450	2.4
6500	1.2
5000	1
13500	3.5
14000	3.4
8450	3
9150	3.1

Scatter Plot

Bivariate Analysis

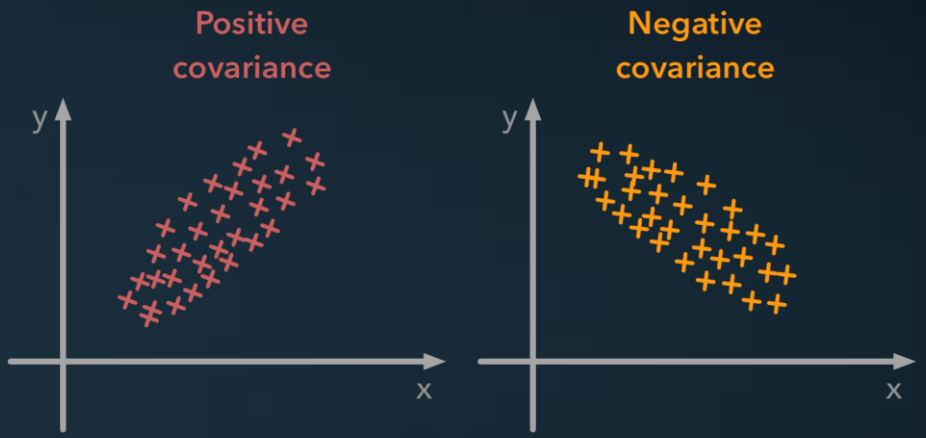
- What is the relation between Salary and No. of years of Experience ?
- ***Covariance***
  - It tells us if the paired values tend to rise together, or if one tends to rise as the other falls.

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

Source : mathsisfun

Covariance

- It is used for the linear relationship between variables.
- It can take any value between -∞ and +∞



x and y values tend to rise together

x rises as y falls (and vice versa)

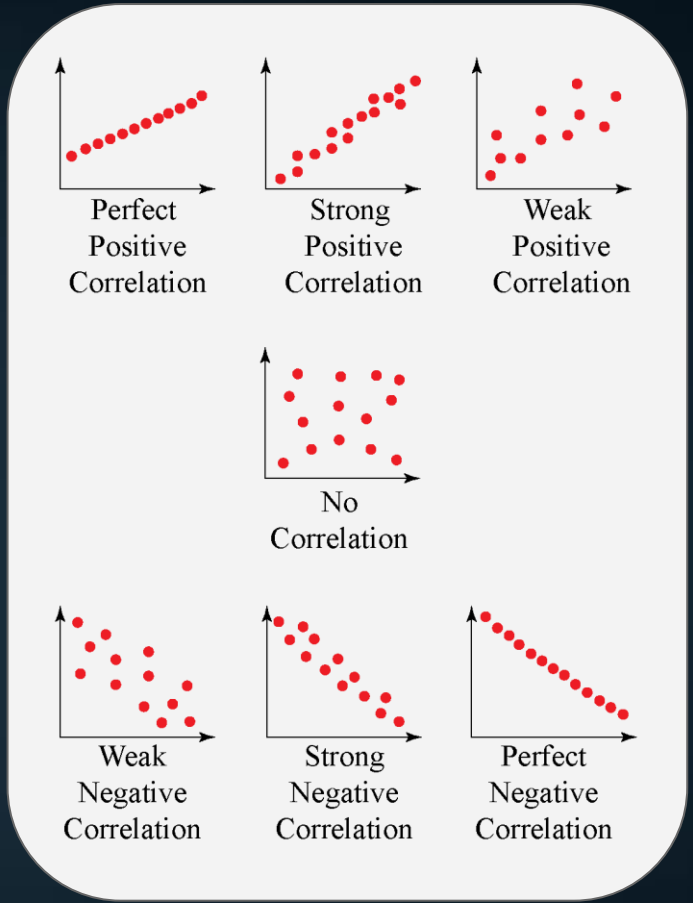
Source : medium

Correlation

- how strong the relationship is.
- a dimensionless metric and its value ranges from -1 to +1.
- The closer it is to +1 or -1, the more closely the two variables are related.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Source : analyticsvidhya



Source : cuemath

Go for Practice

Contact me:

https://www.linkedin.com/in/ahmadmmostafa