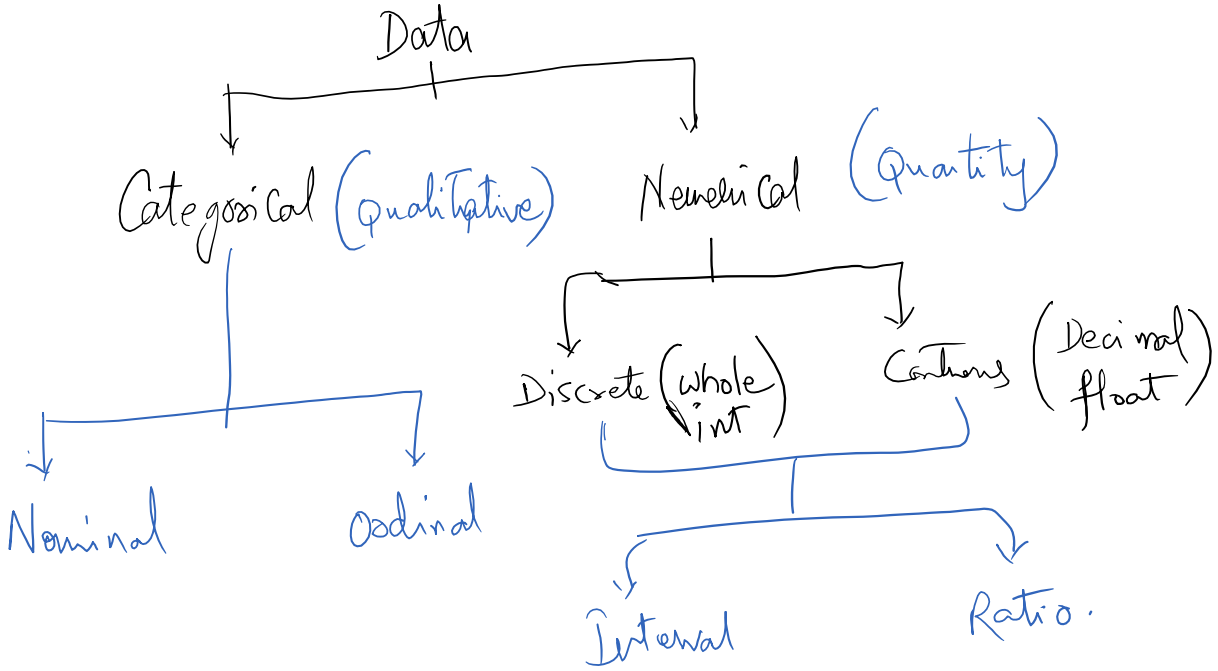
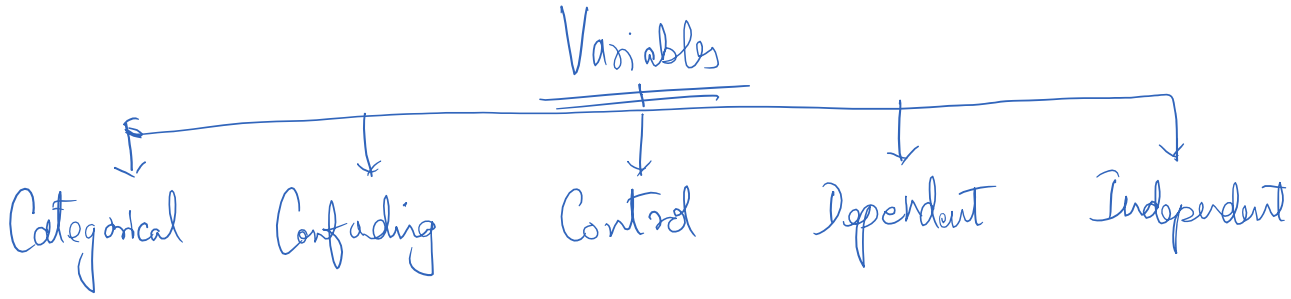
Statistics_Notebook_02_03_Jul_22

Saturday, July 2, 2022 5:58 PM

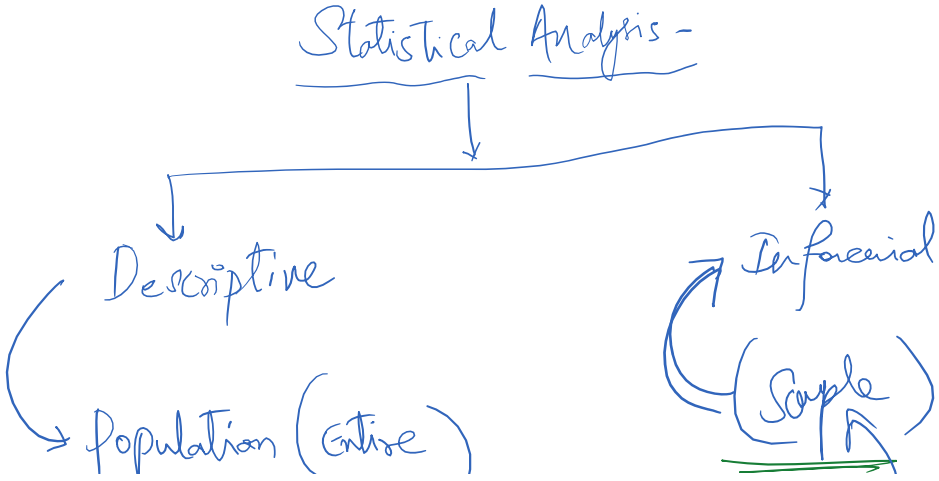
Data → Set of Recorded facts → Information : Processed data with Meaning



Variables : Represents an unknown value / value that changes



Statistical Analysis -



→ Population (Entire data)

Sample
From Population

Descriptive stats

- Central Tendency
  - Mean (Avg)
  - Median (middle)
  - Mode (most freq)
- Dispersion
  - Range (Max - min)
  - IQR
  - Variance
  - Std deviation
  - Skewness
  - Kurtosis
- Position
  - Percentile
  - Quartile
  - Std Scores

IQR : Inter Quartile Range (4)

$IQR = Q_3 - Q_1$

Variance : Measurement of how far a number is from mean.

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

x → data point
$\bar{x}$ → mean
n → Total data pts.

$S^2 = \frac{\sum(x_i - \bar{x})^2}{n}$

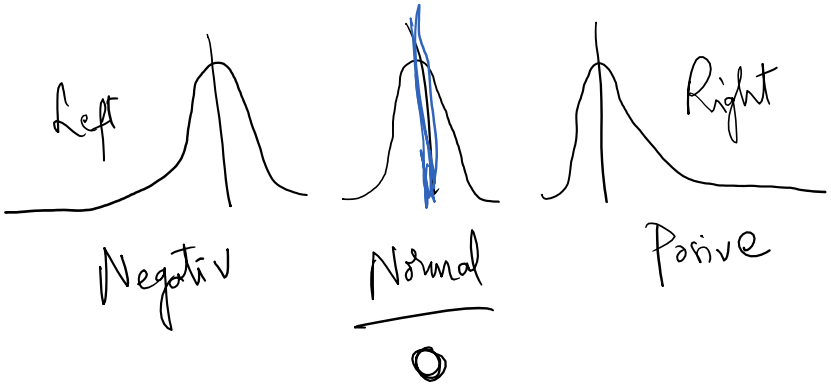
$\bar{x}$ → mean

n → Total data pts.

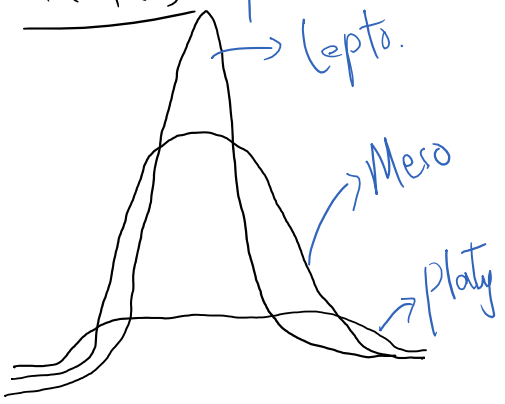
Std dev → Root of variance

$\sigma = \sqrt{\text{Variance}}$

Skewness : Symmetry



Kurtosis : Peakedness



Percentile : Divide data in to 100 parts equally → each part is 1 percentile

Quartile : —— 4 parts. —— Quartile

Standard Score : How many standard deviation, a value is away from Mean

→ Z Score

$$Z = \frac{x - \bar{x}}{\sigma}$$

x → Data point

$\bar{x}$ → Mean

σ → Std deviation

Age	Salary
23	25k
24	26k
32	50k
46	75k

→ Digfrent Scale ↔

Age(z)	Sal(z)
0.3	0.8
0.4	0.85
0.6	0.93
0.8	0.96

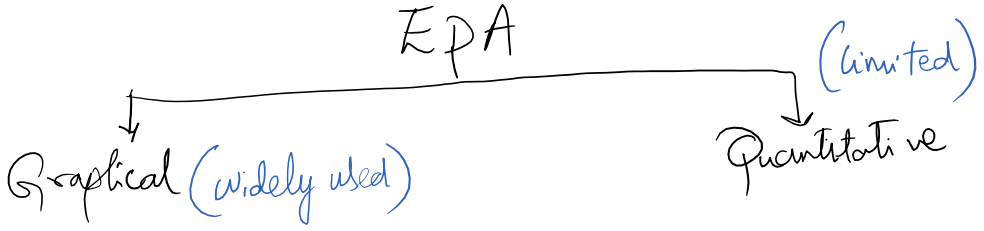
EDA → Exploratory data Analysis.

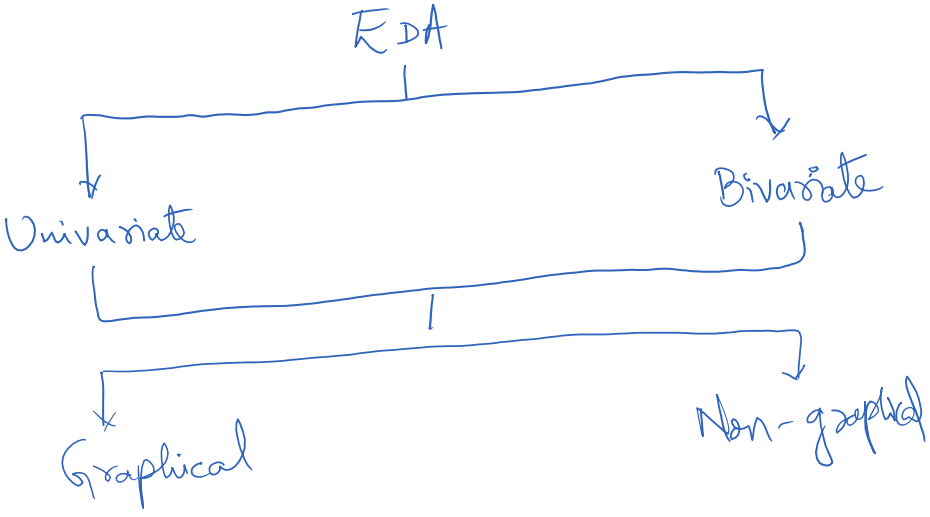
→ Insight To the data

→ Extract important Variables

→ Detect outliers & Anamolies.

→ Understand data better





Univariate → Non-graphical EDA

⇒ Measure of Centre (Mean, Median, Mode)

⇒ Measure of Dispersion (Range, IQR, Variance, Stddev)

⇒ Measure of position (percentile, Quartile, Std Score)

⇒ Measure of Shape (Skewness (−, 0, +), Kurtosis)

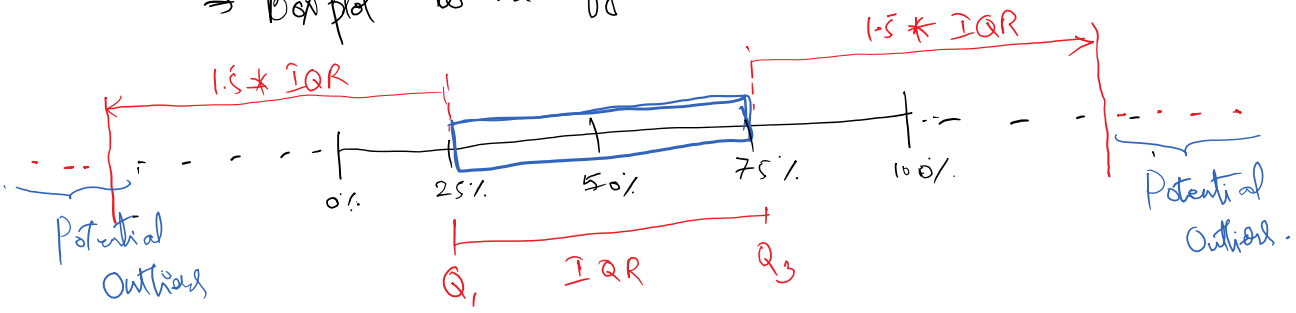
⇒ Test for Outliers

Outlier : Data that deviates highly from main stream data.

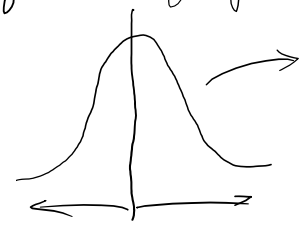
Outlier : Data that deviates highly from main stream data.

Data > / < 1.5 * IQR → Generally considered outlier

→ Boxplot to identify outliers.



Skewness : Measure of lack of Symetry



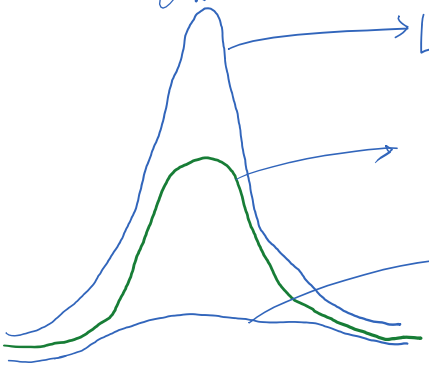
Symetric → Skewness is 0

Asymetric distribution

Positive Skewed / Right skewed

Negative Skewed / Left skewed

Kurtosis : Measure of Peakedness (Compared to Normal dist)



Lepto kurtic (positive)

Meso kurtic (Normal)

platy kurtic (Negative)

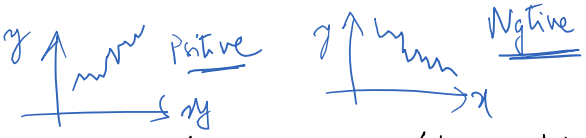
Univariate Graphic EDA : Shape & Distribution

→ Bar graph

→ Histograms
→ pie chart
→ Box plot → Quartiles → Outliers
→ Scatter plot

Bivariate EDA : plot the graphical Relation btw Variables.

Bivariate EDA → Non graphic



→ Covariance
→ Correlation
} Help to identify the linear Relation btw Variables.

Covariance

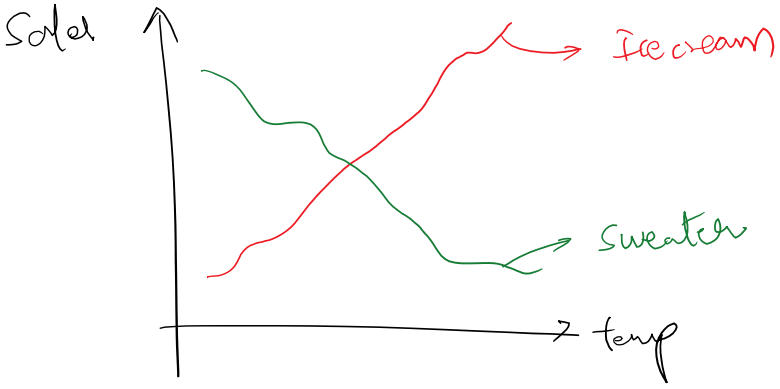
→ Var's are +/− related

→ Only direction

Correlation (Association)

Degree of the relation of Variables

Direction + Strenght



Correlation → Pearson, Spearman, Kendall

	Var1	Var2	Var3
Var1	1	0.3	0.8
Var2	0.3	1	0.4

Var 2	0.3	1	-0.4
Var 3	0.8	-0.4	1

Covariation : Change in one variable is similar to another which has no relation with the variable

Bivariate EDA : Graphical

→ Scatter plot (Outliers, Positive/Neg, Density, Trend,)

→ Heat Map (Highlight the important & Not important one)

Guide for graphic EDA

Variable	GEDA
Categorical	Bar, pie
Univariate + Continous	Line plot, Histogram
Bivariate + Continous	Scatterplot

objective	Grap EDA
Idea of Distribution	Histogram
Outliers	Box, Scatter, Histogram
Relation btn variables	Scatter, Cov & Corr

Cross tab

State	Gender	People	Total

State	Gender	People	Total
KA	M	2	7
	F	5	
TN	M	3	9
	F	6	
KL	M	1	3
	F	2	
AP	M	4	12
	F	8	