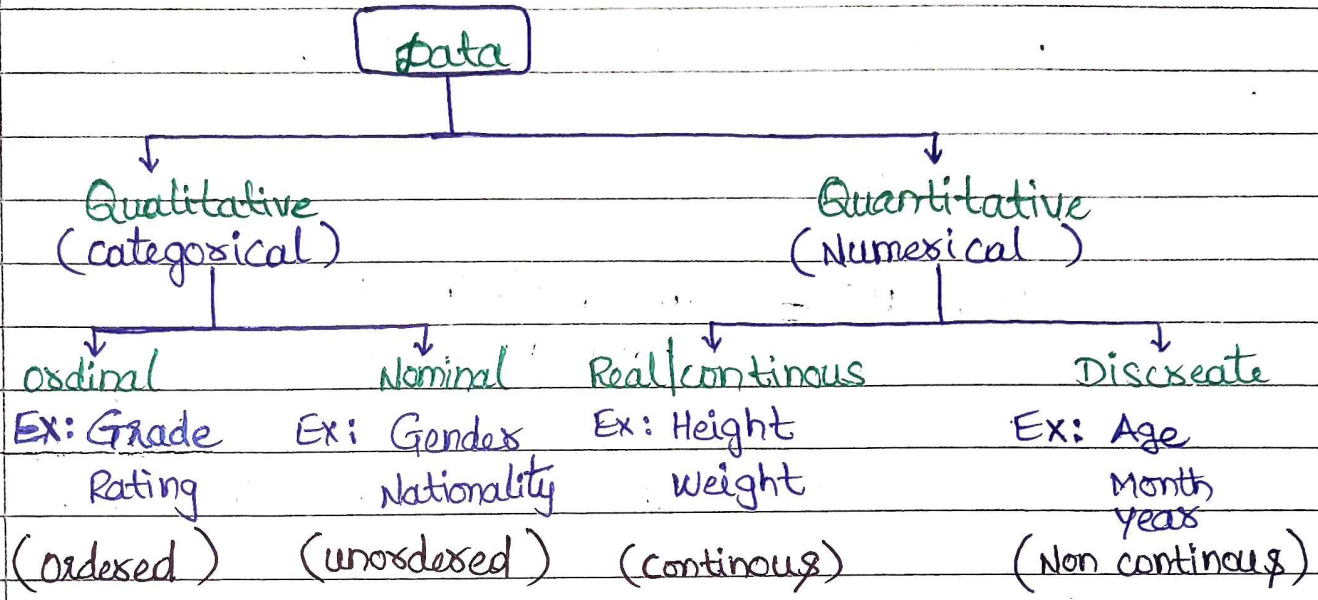
Descriptive statistics

__/__/____

- Describe and summarizing the Data.



* Descriptive statistics allows us to do non-visual data Analysis technique for data Analysis.

Case study:

let us assume Vijetha Super Market hires you to Analyze their Data?

(Why? → so that they can improve their Business.

* No matter in what format the data comes.

For Ex:- .csv, .xlsx, mySQL, postgre SQL, Mango DB, Oracle, .json, .xml, .txt, .jpg, .png, Google sheets, AWS, GCP, Azure etc...)

* you should be able to LOAD the data in your favorate tool.
* you should be able to perform TRANSFORMATION of data so that it is ready for Analysis:.
* perform Data Analysis.

__/__/____

Measuse of central Tendency:

* These axe the statics that helps us measuse/ computes the central point in the data.

① Mean (Avesage):

$$\text{Mean } (\mu) = \frac{\text{sum of all the values}}{\text{total no.of values}}$$

② Median: It is the central value in your given dataset.

Ex: 10,20,30,40,50 → mean $= \frac{10+20+30+40+50}{5} = 30$

→ Median ⇒ 10 20 (30) 40 50

computing the median:

① Sort the data

② Check the count (data) $= n$ values

-②ⓐ 'n' is odd:

$$\text{median} = \left(\frac{n+1}{2}\right)^{th} \text{ term}$$

-②ⓑ 'n' is Even:

$$\text{median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

Mean & Median allows us to find the central Tendency of the data.

Ex: 10 20 40 50 300

__/__/____

$$\text{mean} = \frac{10+20+40+50+300}{5} = 84$$

median ⇒ Sort data ⇒ 10 20 (40) 50 300

$$\text{median} = \frac{n+1}{2} = 40$$

Note : Mean is badly affected by outliers, whereas Median is some what it manage.

Outliers : Any extremely large (OR) extremely small values when compared to rest of the data.

(3) Mode : Applied on Categorical (OR) Discreate Numerical columns.

↳ Most freequent value in the data.

- unimodel : Ex: 2 1 2 3 1 2 3 2 2

mode = 2

- Bimodel : Ex: 2 3 2 3 1 2 3 2 3

mode = 2, 3.

__/__/____

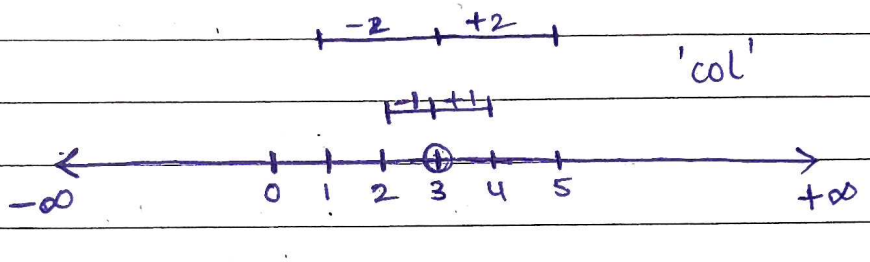
Measure of spread / Dispersion :

① **Range :** Range = maximum - minimum.

② **variance :** It helps us measure the spread of the given data Across Mean.

Ex: col

col
5
4
3
2
1



* variance measures how the data is spread across mean.

- In simple terms ↴

variance measures how far away are the given data points from centre of your data.

		x_i - mean ↖ distance of x_i from mean
x_1	5	+2
x_2	4	+1
x_3	3	0
x_4	2	-1
x_5	1	-2

$\frac{\sum_{i=1}^{n} (x_i - \text{mean})}{n}$ → problem with this formula these are distances with +ve and -ve magnitude & they are cancelling each other.

Solution :-

$\frac{\sum}{n}$ abs $(x_i - \text{mean})$

- This was the not solution

$$\sigma^2 = \frac{\sum}{n} (x_i - \text{mean})^2$$

$$= \frac{2^2 + 1^2 + (-2)^2 + (-1)^2 + 0^2}{5}$$

$$= \sqrt{2}$$

$$\text{std} = \sqrt{\text{variance}}$$

__/__/____

x_i - mean → How far away is x_i from mean

In order to aggregate all distances ↓

$$\text{variance} = \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \text{mean})^2}{n}$$

← TO Nagate the Effect of negative value.

→ standard deviation = σ = $\sqrt{\text{variance}}$

percentiles:

use cases: ① It helps us know the position / Rank.

② It helps you Break the data into Equal parts.

Example:

* you scored 95 percentage:

It tells how much you scored. It won't help you compare your performance with others.

* you scored 95 percentile:

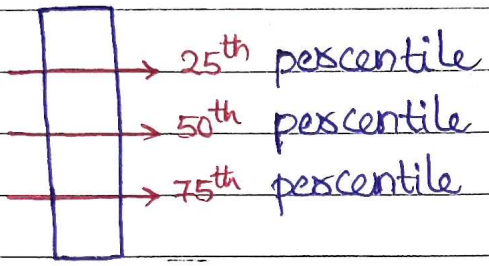
It tells out of all the participants you scored Better than 95% of participants.

Rank / position in the data (i.e; out of all the quiz participents)

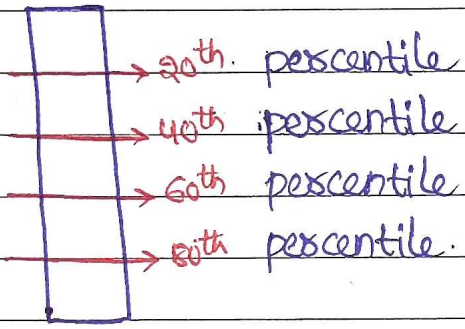
__/__/____

- Some popular percentiles are :

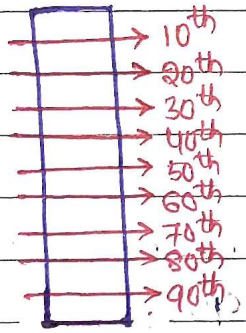
(A) Quantile (Divides the data into 4 equal parts)



(B) Quintile (Divides the data into 5 equal parts)



(C) Decile (Divides the data into 10 equal parts)



__/__/____

③ IQR (Inter Quartile Range) → max - min

↳ 25th, 50th, 75th percentiles.

IQR = 75th percentile − 25th percentile

④ Mean Absolute Deviation (spread across mean)

$x_1 - \mu_x$

$x_2 - \mu_x$

$x_3 - \mu_x$

⋮

$x_n - \mu_x$

↓

Mean Aggregation

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

Mean Square Deviation

(or)

Variance (σ^2)

$$\frac{\sum_{i=1}^{n} abs(x_i - \mu_x)}{n}$$

Mean Absolute Deviation.

Note: For Example your manager asked you to calculate 'MAD' of the data.
you should always get a clarity, MAD means do you mean Median Absolute Deviation (or) Mean Absolute Deviation.

__/__/____

Measure of Relationship (Bi-variate Analysis)

① Co-variance:

$$\text{Variance}, \sigma^2(x,x) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(x_i - \mu_x)}{n}$$

$$\text{Co-variance}(x,y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_x)}{n}$$

Question: How do we Interpret from Cov(x,y)?
(or)
What does cov(x,y) implies?

Answer: Cov(x,y) helps us understand the Relationship between the two columns.

- What kind of a Relationship?

$x \propto y$ → Whenever cov(x,y) is +ve

$x \propto \frac{1}{y}$ → Whenever cov(x,y) is -ve

- Range of covariance (x,y) ↓

$$-\infty \leq \text{cov}(x,y) \leq +\infty$$

- What is the meaning of $x \propto y$ (or) $x \propto \frac{1}{y}$

$x \propto y$ → Means if 'x' increases, 'y' also increases
i.e; There is a directly proportional Relationship between x & y.

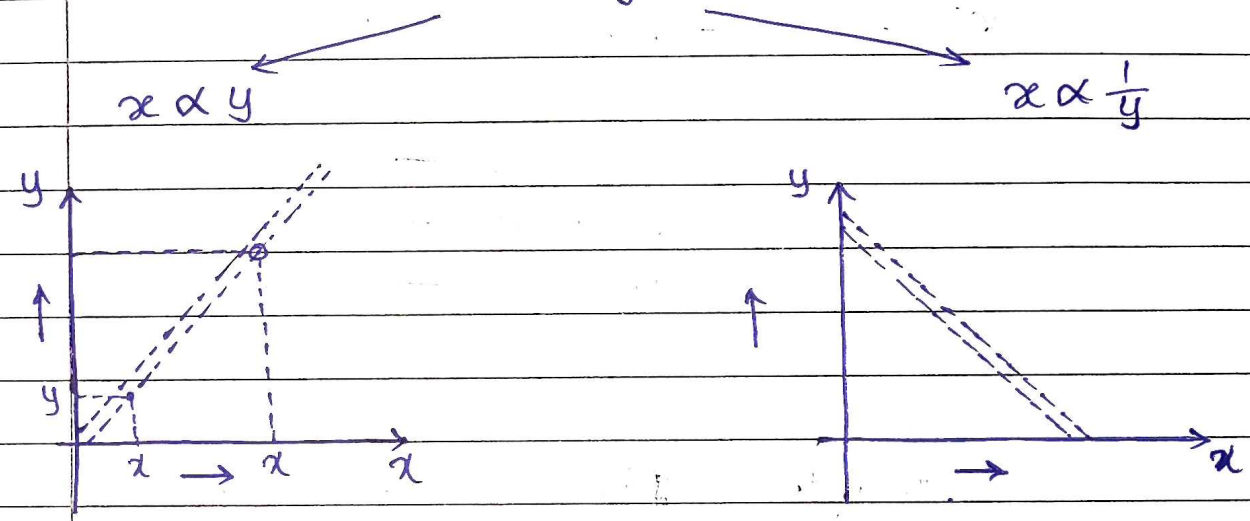
$0 \leq \text{cov}(x,y) \leq +\infty$ → cov(x,y) is +ve

__/__/____

$x \propto \frac{1}{y}$ → Means it 'x' increases ; 'y' decreases
i.e; There is an inversely proportional relationship between x & y.

$-\infty < Cov(x,y) < 0$ → Cov(x,y) is Negative.

Bivariant
(Num Vs Num)
↓
Viz. Analysis



② pearson correlation coefficient ($\rho_{x,y}$) :

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma(x) * \sigma(y)}$$

$$-1 \leq \rho_{x,y} \leq +1$$

Note : { Earlier Cov(x,y) was able to determine it $x \propto y$ (or) $x \propto \frac{1}{y}$ (direction of Relationship)

__/__/____

$S_{x,y}$ can determine even the strength of a Relationship along with it $x \propto y$ (or) $x \propto \frac{1}{y}$

- **Direction :** if $S_{x,y} < 0 \Rightarrow x \propto \frac{1}{y}$
  elif $S_{x,y} > 0 \Rightarrow x \propto y$
  else $S_{x,y} == 0 \Rightarrow$ No Relation

- **Strength of Relationship :**

$0.5 \leq S_{x,y} \leq 1 \rightarrow$ Strong $\leftarrow -1 \leq S_{x,y} \leq -0.5$

$-0.5 \leq S_{x,y} \leq 0.5 \rightarrow$ weak

$S_{x,y} = 0 \rightarrow$ No Relationship

0	→ No Relation
0 to 0.5	→ weak
0.5 to 1	→ Strong

Note : pearson correlation coefficient can only determine Strength & Direction of a Linear Relationship.