Roadmap

Statistics :→

It is the branch of math that involve Collecting, organizing, Interpreting, presenting the data.

① Descriptive

① It deals with Collection, organization, Analysis, interpretation & presenting the data.

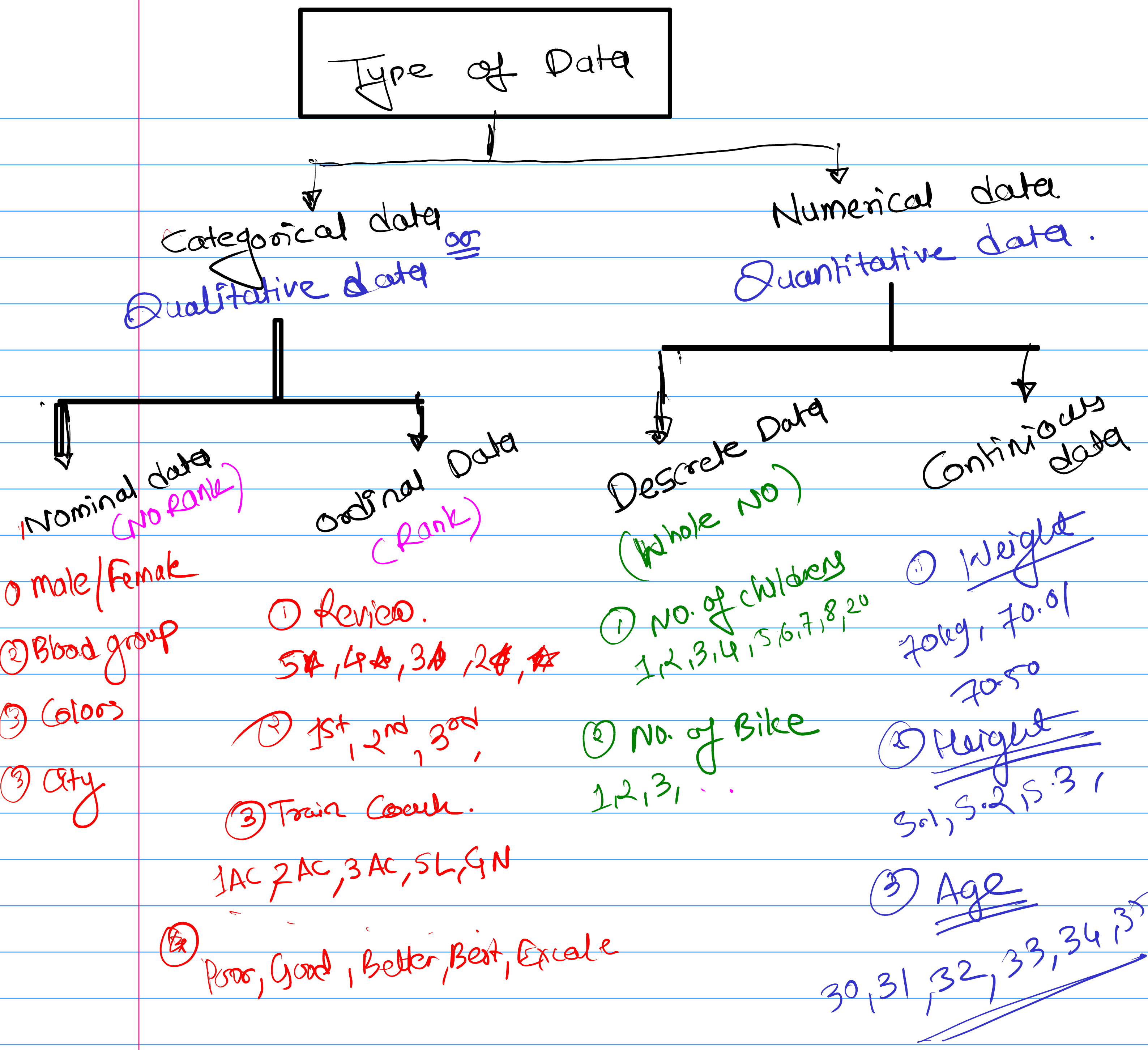
② Inferential

It deals with making a conclusion or a decision, prediction of the population based on Sample.

① Population (N) :→

→ Entire group or Supergroup of data that you are interested in.

② Sample (n) :→ A Sample is a Subset of population data.

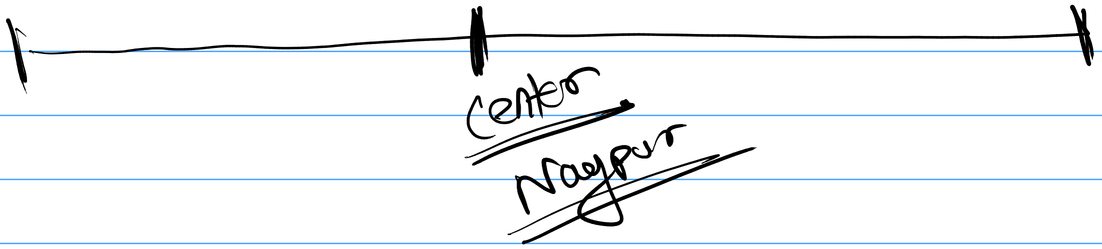


Type of Data
Categorical data or
Qualitative data
Numerical data
Quantitative data.
Nominal data
(No Rank)
Ordinal Data
(Rank)
Descrete Data
(Whole No)
Continious data
① male/Female
② Blood group
③ Colors
③ City
① Review.
5★,4★,3★,2★,★
② 1st, 2nd, 3rd,
③ Train Coach.
1AC, 2AC, 3AC, SL, GN
④ Poor, Good, Better, Best, Excele
① No. of childrens
1,2,3,4,5,6,7,8,20
② No. of Bike
1,2,3, . .
① Weight
70kg, 70.01
70.50
② Height
5.1, 5.2, 5.3,
③ Age
30,31,32,33,34,35

① Measures

① Measure of Central Tendency

Imp



① **Mean** :→ The mean is the sum of all values in the dataset divided by the number of values.

ex = data = [1, 2, 3, 4, 5]

$$\text{mean} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\boxed{\text{mean} = 3}$$

disadvantage :→ It Robhust to outlier.

= It affected by outlier.

ex = [1, 2, 3, 4, 5, 66] → outlier

$$\text{mean} = \frac{1+2+3+4+5+66}{6} = \frac{81}{6} \boxed{= 13.5}$$

② median

:- The median is a middle value in the dataset when the data is sorted.

:→ It dose not affected by outlier.

ex = [1, 2, 3, 4, 5, 666]

1 2 3 4 5 6

$$\text{median} = \frac{n+1}{2} = \frac{6+1}{2} = \boxed{3.5}$$

ex = [10, 20, 30, 40, 50, 600]

$$\text{median} = \frac{n+1}{2} = \frac{6+1}{2} = \boxed{3.5} \rightarrow \frac{3^{rd} + 4^{th}}{2}$$

$$= \frac{30+40}{2} = \boxed{35 = \text{median}}.$$

ex = [20, 40, 60, 80, 100]

$$\frac{5+1}{2} = \frac{6}{2} = \boxed{3}$$

Note :→ mean & median is used to replace null numerical Data.

empty.

mean ✗ median

100 student

Age st

3 Mode :→

The mode is the value that appears most frequently in the dataset.

:→ Generally mode is use for Categorical data.

100 student = Male = 60 =

Female = 40

$\boxed{\text{mode} = \text{Male}}$

4 Weighted Mean :→

It is the sum of product of each value and its weight divided by sum of weight.

AI House price prediction :-]

Algorithm ?

LR = 0.2 = 10L

RF = 0.3 = 11L

Xgboost = 0.5 = 13L

weight value

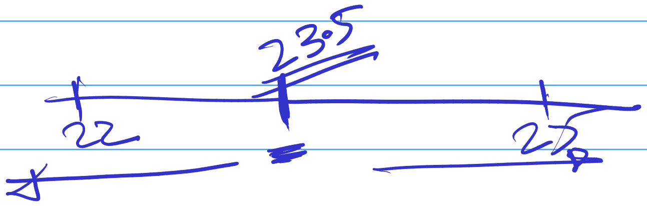
$$\text{W-mean} = \frac{0.2 \times 10L + 0.3 \times 11 + 0.5 \times 13}{0.2 + 0.3 + 0.5}.$$

2. Measure of Dispersion → Spread

= A measure of dispersion is a statistical measure that describe the spread or variability of a dataset.

= It provide information about how data is distributed around the central tendency (mean, median or mode).

Age in Graduation :→ [22, 23, 22, 24, 23, ...
= . . . . 24, 25]



① **Range** :→ Difference between the max & min.

Range = 25 − 22 = 3

:→ It affected by outlier.

$(\sigma^2) \rightarrow$ Sigma

② **Variance** :- → It is the average of squared difference between each data point and the mean.

$$\text{mean} = \frac{3+2+1+5+4}{5} = \frac{15}{5} = 3$$

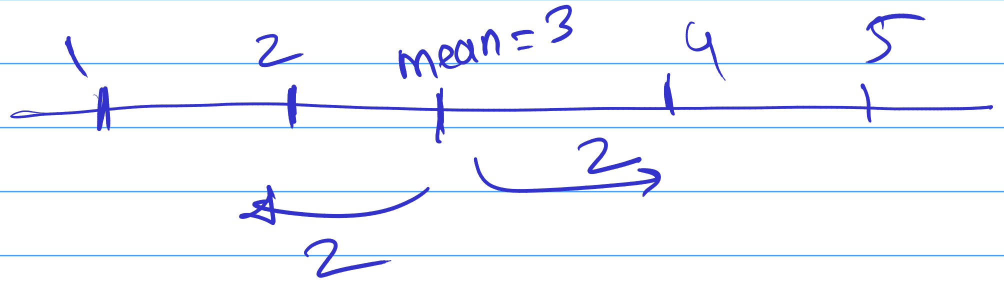
data	(mean - data)	$(\text{mean-data})^2$
3	$3-3=0$	0
2	$3-2=1$	1
1	$3-1=2$	4
5	$3-5=-2$	4
4	$3-4=-1$	1

$$= \frac{0+1+4+4+1}{5}$$

$$= \frac{10}{5} = 2$$

$$\boxed{\text{Variance} = 2}$$

σ^2



km

$\sigma^2 = \text{km}^2$

Age year

$(\text{mean} - x_i)^2 = \text{year}^2$

③ **Standard Deviation** : The square root of the variance.

$$SD = \sqrt{\sigma^2}$$