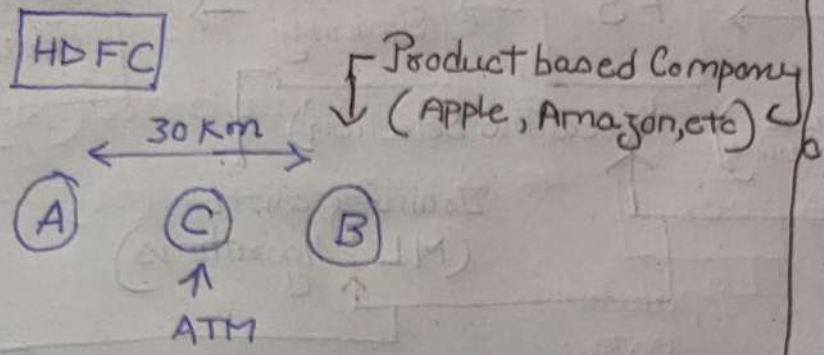
Statistics

→ Use Cases: for banks: examples

①



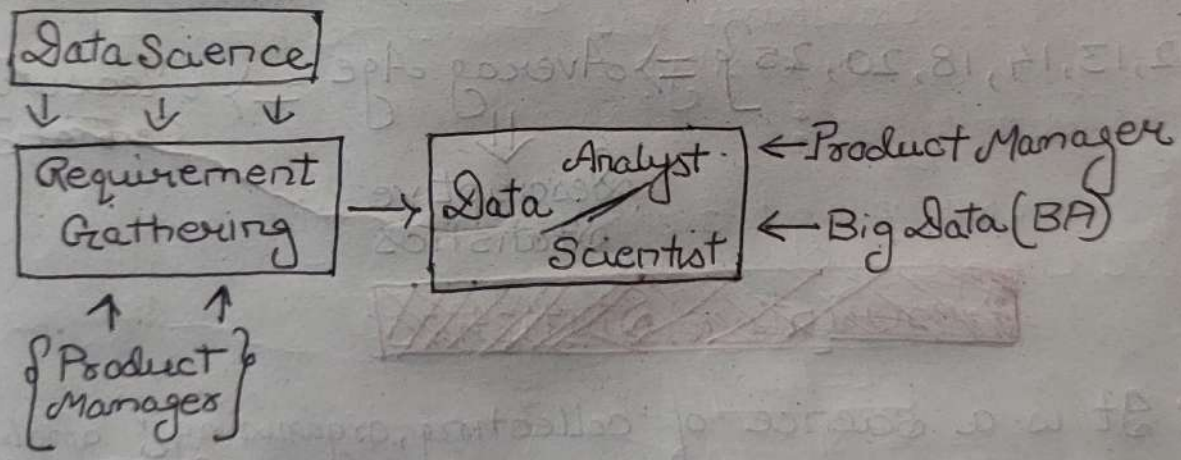
- Product based Company (Apple, Amazon, etc)
  - X-Statisticians - 5 yrs
  - Data Analyst
  - Business Analyst
  - Data Scientist
  - Product Manager

→ HCL, TCS, etc are service based company.

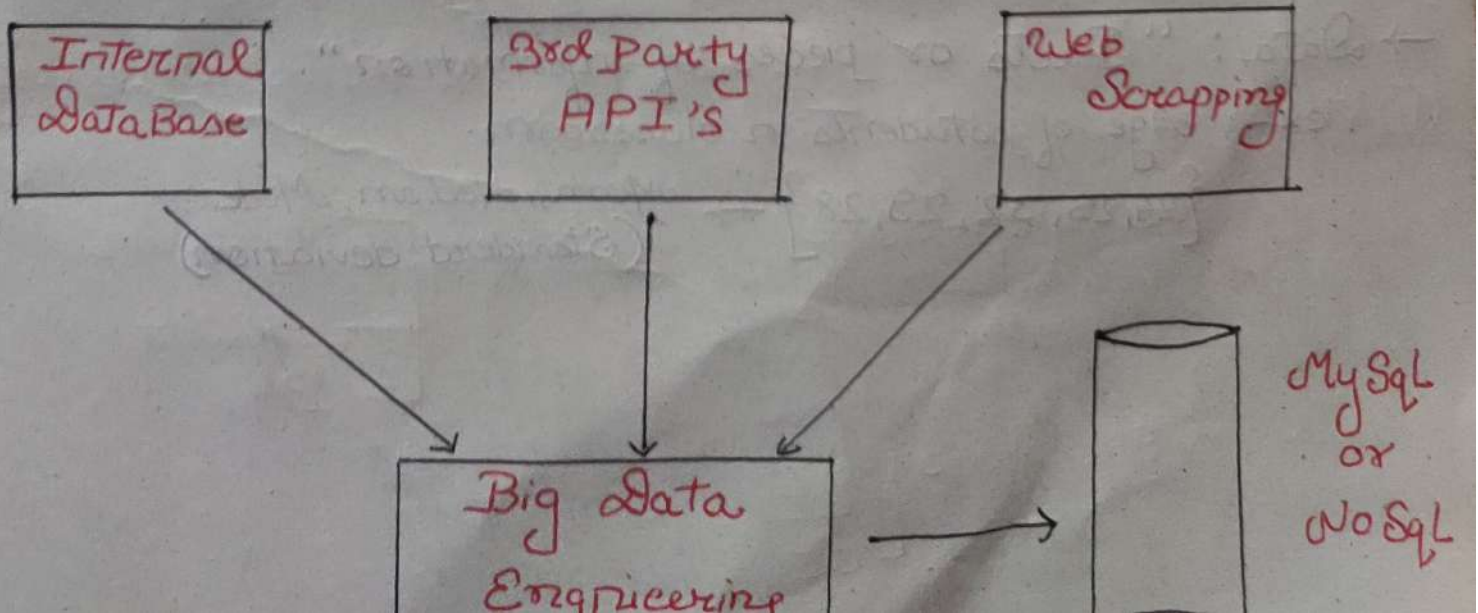
② Identifying the avg. size of the shark in the world.

③ Amazon or Flipkart big-billion day sale, identifying the month of selection.

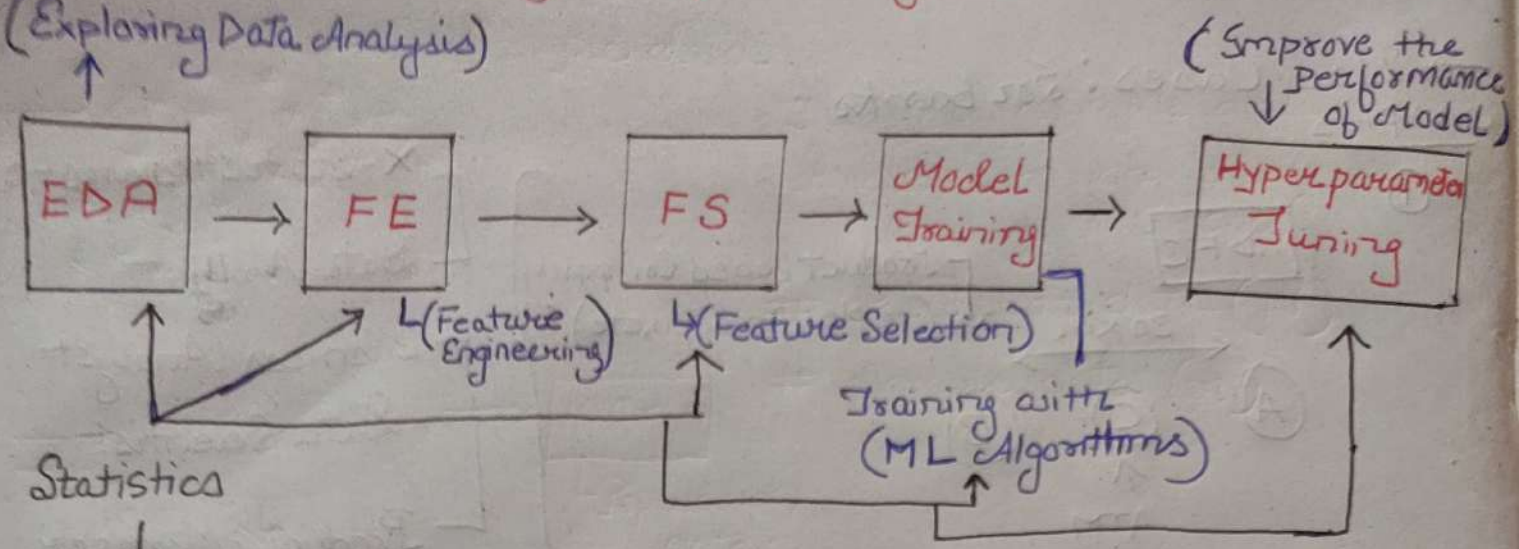
Defination in life cycle of Data Science Project

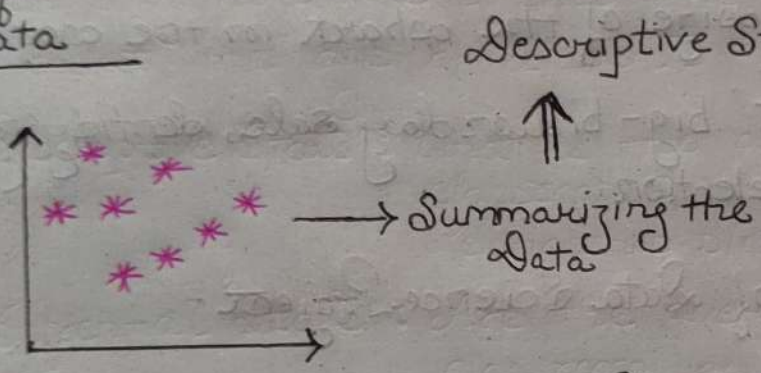


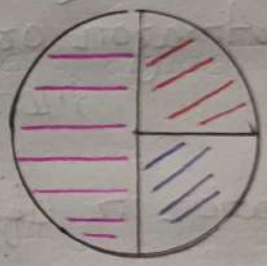
*



Life Cycle of DS Project







Age = {12, 13, 14, 18, 20, 25} ⇒ Averag Age ⇒ Measure o
Central Tend

⇓

Descriptive Statistics

~~Descriptive Statistics~~

* **Statistics:** It is a science of collecting, organising & analysin the data.

→ Data: "Facts or pieces of information".

ex-: Age of students in classroom.

{24, 25, 32, 29, 28} → Mean, Median, Mode (Standerd deviation)

Statistics

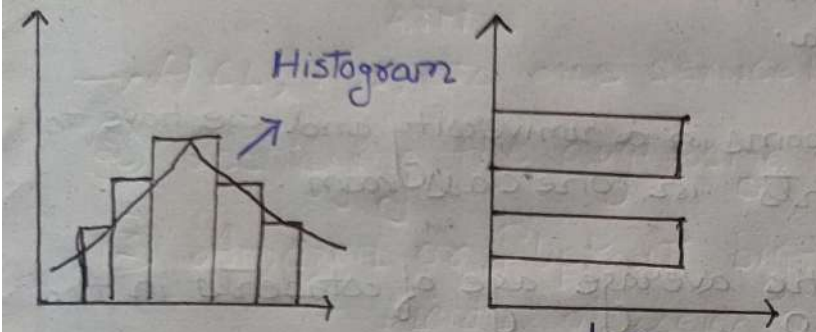
Descriptive Stats [EDA + FE]

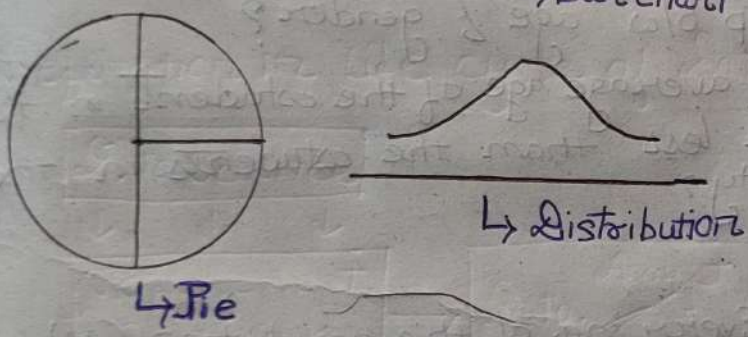
① St consists of organising & summarizing the data

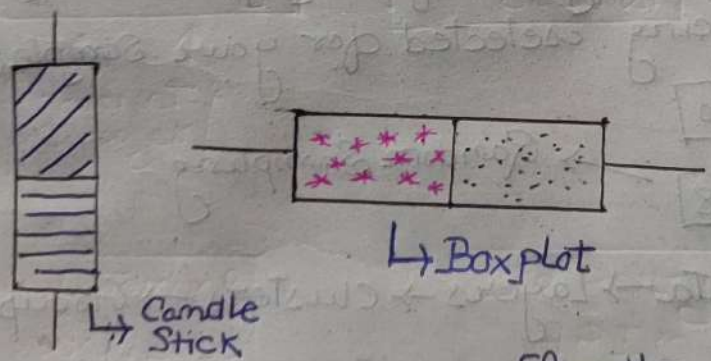
Inferential Stats

St consists of collecting sample data and making conclusions abt Population data using experiments.

→ Hypothesis Testing







University → 500 people

Class A → 60 people

Sample data → Age → Avg. age of entire university

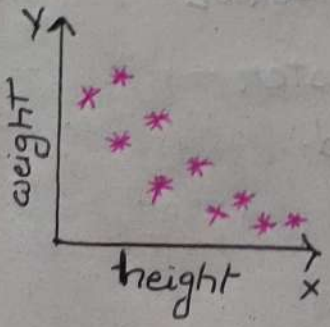
↓

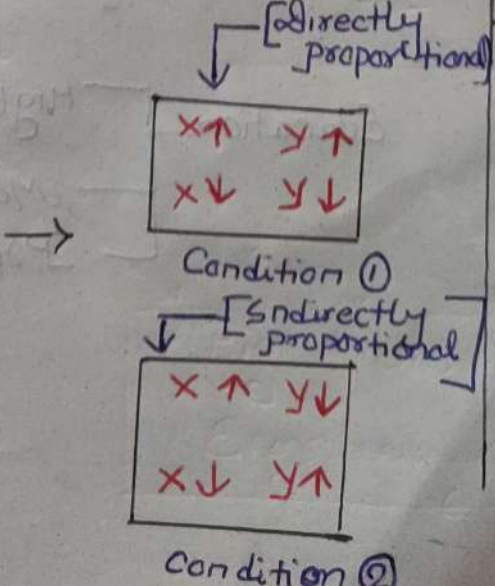
Hypothesis Testing

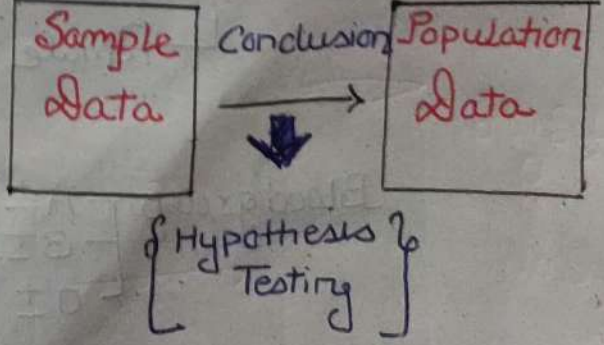
(CI = Convidence Interval)

Types of Testing :

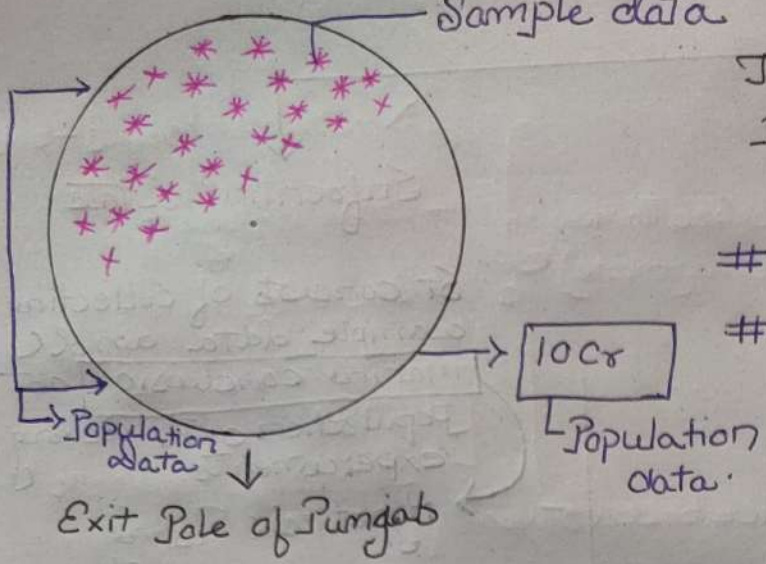
① Z-test
② t-test
③ Chi-squared test
④ f-test







* Sample data Vs Population data:



Through, sample data we can Predict, which party will win or loose in this case.

Population data (N)

Sample data (n)

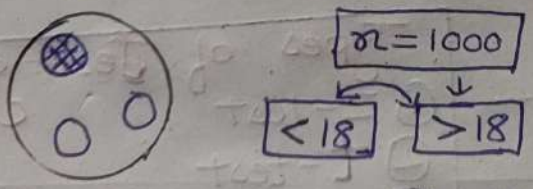
eg: Let say there are 20 classrooms in a university and we have to collect the age of students in one classroom.

→ **Descriptive Stats**: What is the average age of students in the classroom?
Relationship b/w age & gender?

→ **Inferential Stats**: Are the average age of the students in the classroom less than the students in the university?

* Sampling Techniques:

(i) Sample Random Sampling: Every no. of the population (N) has an equal chance of being selected for your sample (n)



Strata → Layers → Clusters → Groups

(ii) Stratified Sampling:

Gender [Male, Female]

Education [High School, Master, Phd]

Blood group [A±, B±, O±]

(iii) Systematic Sampling : → {Airport} nth person
{Credit Card}

ex: Select every nth individual out 5th person of population(N)
or Select every nth individual out of population (N)

(iv) Convenience Sampling : → Only those who are invested in the survey will only participate.

ex : (Data Science Survey → General AI survey)
- filling a survey form at an institution.
- RBI Survey — women → Stratified + Random Sampling.

* Variable : It is a property that can take any values.
ex- age = 14 ; Variables ⇒ Ages = [25, 26, 32, 34] Collection of ages.

→ Types of Variable :

① Quantitive Variable → measured numerically {Mathematical operation}
ex: Age, weight, height, rainfall, temp, distance etc.

② Qualitive Variable → Catagorical Variables {based on some characteristics they are grouped together}
ex- Gender, type, class, Species. etc

Quantitave Variable

Discrete Variable
eg: Whole no. — fixed
- no. of bank account
  └ (it can't be in decimal)
- no. of children
- Pincode — fixed

Continous Variable
eg. Decimal values
height, weight, age, rainfall, speed

Assignment —— :

(i) What kind of variable is Marital Status? — Categorical
(ii) " " " " " River length? — Continous
(iii) " " " " " Movie duration? — Continous
(iv) " " " " " Pincode ? — Discrete
(v) " " " " " IQ ? — Discrete

Objectives

1. Histograms
2. Measure of Central Tendency
3. Measure of Dispersion
4. Percentiles & Quartiles
5. 5 Number Summary (Box Plot)

→ Histogram :

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 37, 40, 41, 42, 51, 60, 65, 68, 78, 90, 95, 100}

Step-1 : SORT the numbers

Step-2 : Bins → No. of groups

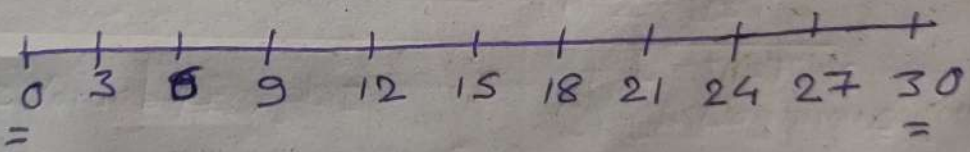
Step-3 : Bin Size – Size of bins

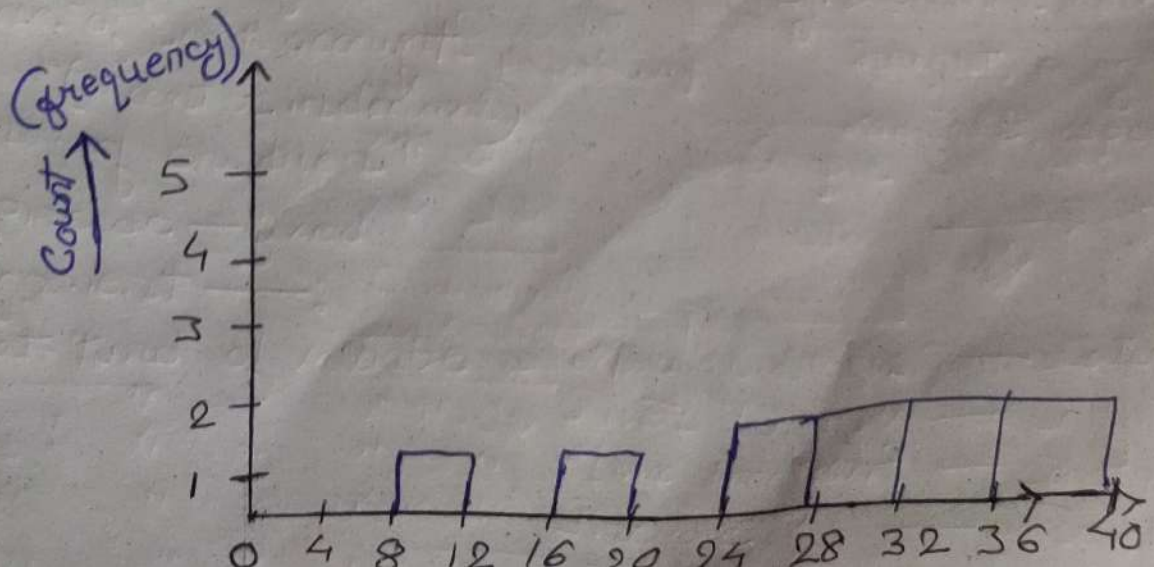
[10, 20, 25, 30, 35, 40]

bins = 10

min = 10
max = 40

$\frac{40-10}{10} = \frac{30}{10} = 3$



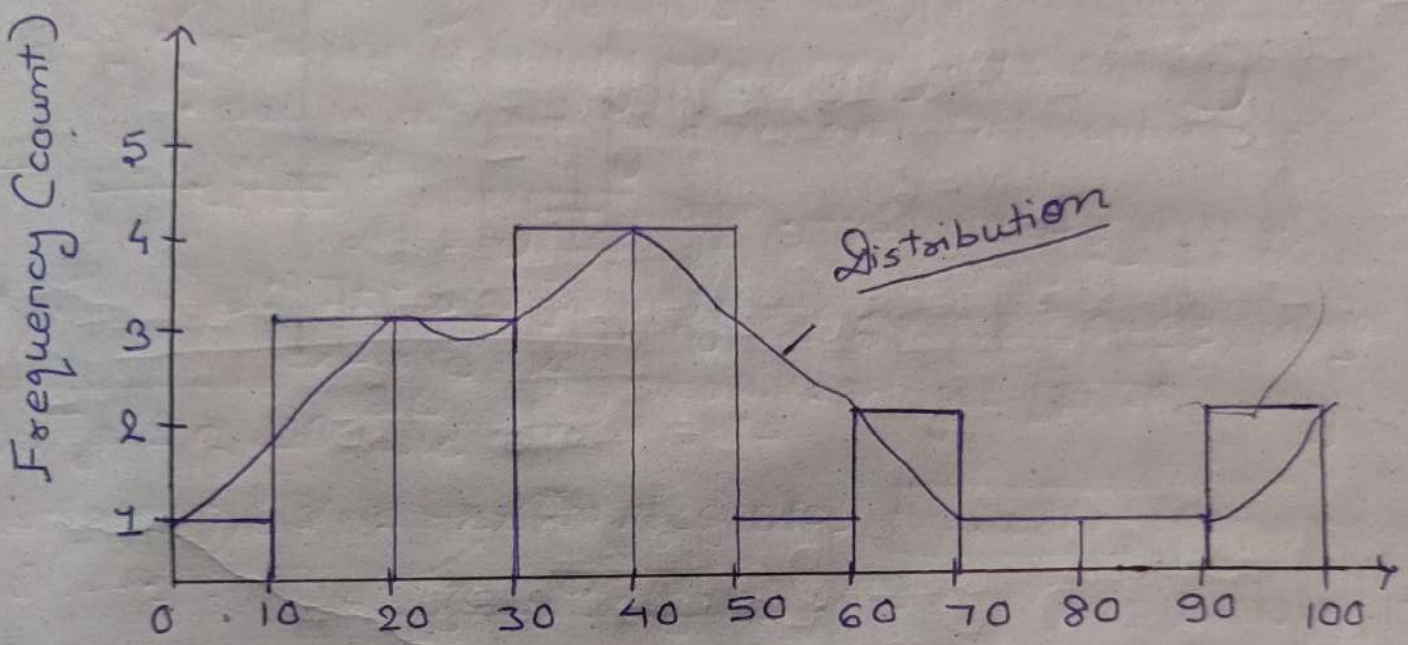


Example: 1

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 98, 95, 100}

binsize = 10

{binsize = $\frac{100}{10}$ = 10 (Max / Min)}



Example - 2

Weight: {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

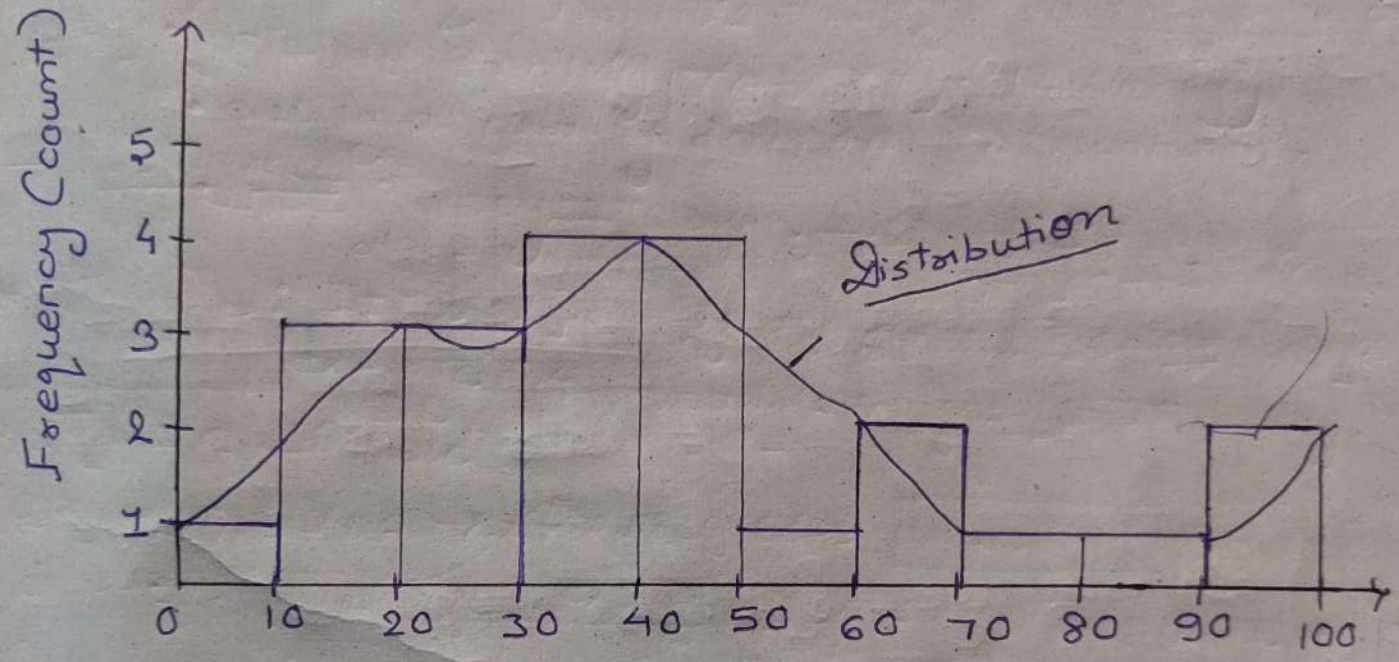
bins = 10 , bin size = $\frac{95-30}{10}$ = 6.5

Example: 1

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 98, 95, 100}

bin size = 10

{bin size = $\frac{100}{10}$ = 10 ; 100 (Max), 10 (Min)}



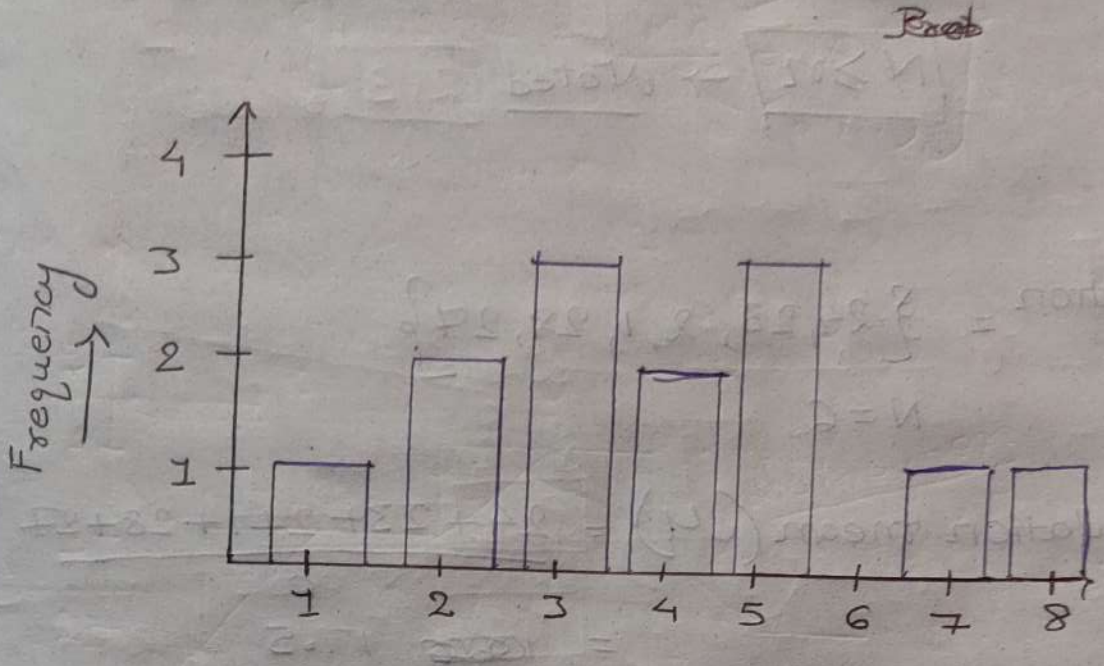
Example - 2

Weight : {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

bins = 10 , bin size = $\frac{95-30}{10}$ = 6.5 =

② Discrete Continous

No. of bank accounts = {2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5}



PdF: Probability density function → Continous
Pmf: Probability mass function → Discrete

* Measure of Central Tendency:

(i) Mean
(ii) Median
(iii) Mode

A measure of central tendency is a single value that attempts to describe a set of data identifying the central position:

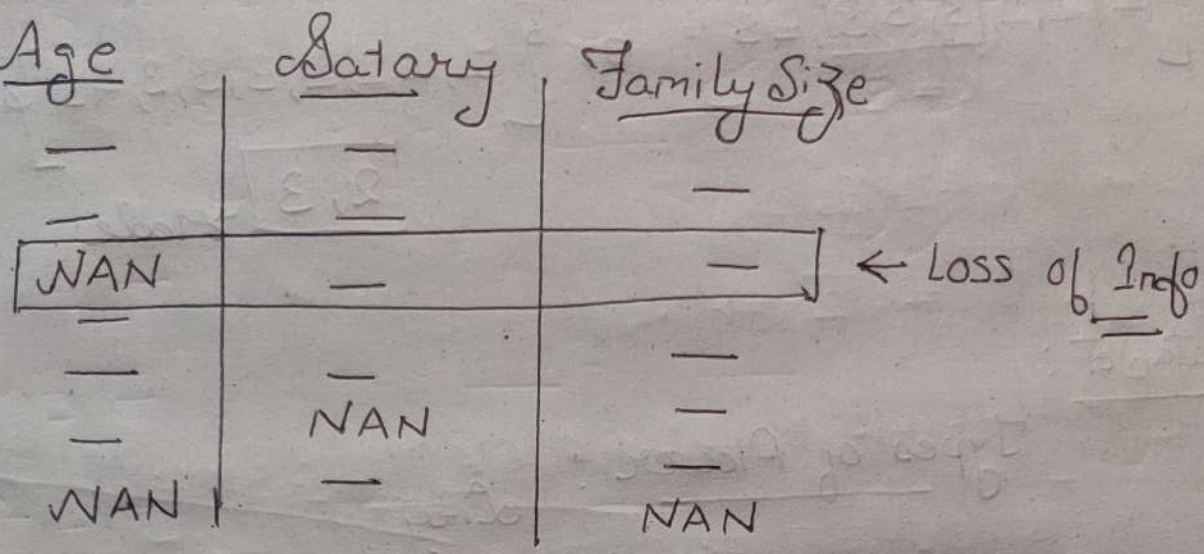
(1) Mean: $x = \{1, 2, 3, 4, 5\}$

$$\text{Avg / Mean} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

→ Population (N)

$$\text{Population mean } (\mu) = \boxed{\sum_{i=1}^{N} \frac{x_i}{N}}$$

* Practical Application (Feature Engineering)



② Median :

→ Steps to find the median :

① Sort the numbers.
② Find the central number

① if the no. of elements are even we find the average of central elements

② if the no. of elements are odd we find the central elements.

~~{1, 2, 3, 4, 5, 6, 7, 8, 100, 120}~~

{1, 2, 3, 4, 5, 6, 7, 8, 100, 120}

③ Mode : Most frequent occuring elements.

ex:

{1, 2, 2, 3, 3, 3, 45} or {1, 2, 2, 2, 3, 3, 3, 45}

2, 3 → Mode

Dataset:

Types of Flower:

Lily
Sunflower
Rose — Mode
NAN ← Rose
Rose
Sunflower
Rose
NAN ← Rose

Replace NAN with Rose

* Measure of Dispersion:

(i) Variance (σ^2) → Spread of Data
(ii) Standerd deviation (σ)

Variance → Population Variance (σ^2)
→ Sample Variance (s^2)

(i) Population Variance :-

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

(ii) Sample Variance :

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

ex-1 {1,2,3,4,5}

$$\text{Mean}(\mu) = \frac{1+2+3+4+5}{5} = 3$$

$$\sigma^2 = \frac{[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2]}{5}$$

$$\sigma^2 = \frac{4+1+0+1+4}{5} \Rightarrow \frac{10}{5} = 2$$

Ex-2 {1, 2, 3, 4, 5, 6, 80}

$$\text{Mean}(\mu) = \frac{1+2+3+4+5+6+80}{7}$$

$$= \frac{101}{7} \Rightarrow 14.4$$

$$\sigma^2 = \frac{[(1-14.4)^2 + (2-14.4)^2 + (3-14.4)^2 + (4-14.4)^2 + (5-14.4)^2 + (6-14.4)^2 + (80-14.4)^2]}{7}$$

$$\sigma^2 = 719.10$$

Note → As variance increases spread of data coould be higher.

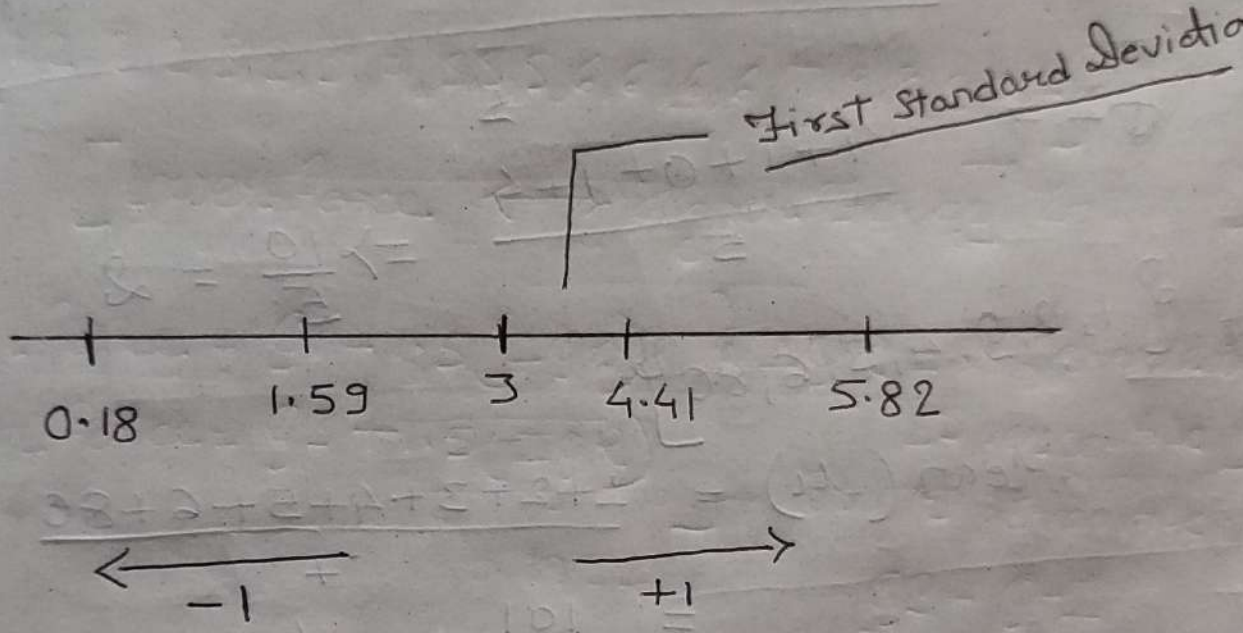
(ii) Standard Deviation ($\sqrt{\sigma^2}$)

ex:- $\{1, 2, 3, 4, 5\}$

$\mu = 3$

$\sigma^2 = 2$

$\sigma = \sqrt{2} \Rightarrow 1.41$



* Percentile & ~~Quaters~~ Quartiles :-

Percentage = $\{1, 2, 3, 4, 5, 6, 7, 8\}$

Percentage of even no. = $\frac{\text{No. of even no.}}{\text{Total no. of numbers}}$

$= \frac{4}{8} = 0.5 = 50\%$

Percentiles : GATE, CAT, SAT, GRE, JEE, NEET are getting in percentiles

Percentile is a value below which a certain percentage of observation lies.

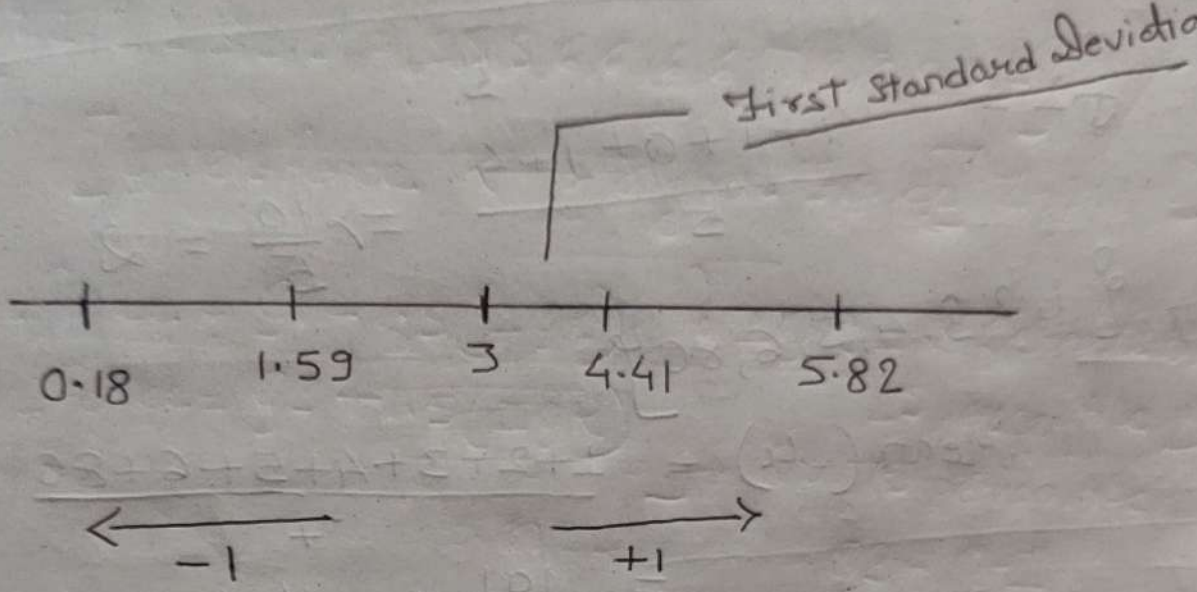
(ii) Standard Deviation (SD)

ex:- $\{1,2,3,4,5\}$

$\mu = 3$

$\sigma^2 = 2$

$\sigma = \sqrt{2} \Rightarrow 1.41$



* Percentile & ~~Quaters~~ Quartiles :-

Percentage = $\{1,2,3,4,5,6,7,8\}$

Percentage of even no. = $\frac{\text{No. of even no.}}{\text{Total no. of numbers}}$

$= \frac{4}{8} = 0.5 = 50\%$

Percentiles : GATE, CAT, SAT, GRE, JEE, NEET are getting in percentiles

Percentile is a value below which a certain percentage of observation lies.

99 percentile = St means the person has got better marks, than 99% of the entire students.

Dataset : 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

Q. Find the percentile rank of 10

Ans:

$$\text{Percentile Rank of } x = \frac{\text{No. of value below } x}{n}$$

$$= \frac{16}{20} = 80 \text{ percentile}$$

Q - What is the value that consists at 25 percentile.

$$\text{Value} = \frac{\text{Percentile}}{100} * n \quad \begin{cases} \text{for even } (n+1) \\ \text{for odd } (n) \end{cases}$$

$$= \frac{25}{100} \times 20 = 5^{th} \text{ index}$$

Hence, 5th index is '5' in the dataset

* 5 number Summary :

(i) Minimum
(ii) First Quartile
(iii) Median
(iv) Third Quartile
(v) Maximum

Dataset

{1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

[Lower Fence ⟷ Higher Fence]

Lower Fence = $Q1 - 1.5(IQR)$
Higher Fence = $Q3 + 1.5(IQR)$

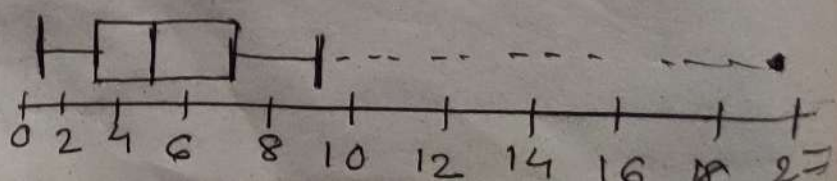
IQR = Difference b/w 25 percentile & 75 percentile.
$Q3 - Q1$

$Q1 = \frac{25}{100} \times 21 = 5.25$ Index = 5

$Q3 = \frac{75}{100} \times 21 = 15.75$ index = $\frac{8+7}{2} = 7.5$

Lower Fence = $3 - (1.5)(4.5) = -3.65$
Higher Fence = $7.5 + (1.5)(4.5) = 14.25$

(i) Minimum = 1
(ii) $Q1 = 3$
(iii) Median = 5
(iv) $Q3 = 7.5$
(v) Maximum = 9



② Standard Normal Distribution :

$x \approx$ Gaussian Distribution (μ, σ)

[Converting x to y] ↓

$y \approx SND\ (\mu = 0, \sigma = 1)$

ex: $x = \{1, 2, 3, 4, 5\}$

↓

$\mu = 3 \quad \sigma = 1.41$ (Applying Z-score)

③ Z-Score :

$$\left.\begin{array}{l} \dfrac{x_i - \mu}{\sigma} \\ \dfrac{\sigma}{\sqrt{n}} \end{array}\right\} \rightarrow \text{Standard Error}$$

$$= \frac{x_i - \mu}{\sigma} \quad ; \text{if } n = 1$$

$x = \{1, 2, 3, 4, 5\}$

$\mu = 3, \sigma = 1.414$

$= \frac{1-3}{1.414} = -1.414$ $\quad \{ y = \{ -1.414, -0.707, 0, 0.707, 1.414 \}$

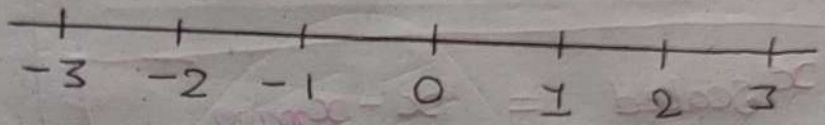
$= \frac{2-3}{1.414} = -0.707$

$= \frac{3-3}{1.414} = 0$

$= \frac{4-3}{1.414} = 0.707$

$= \frac{5-3}{} = 1.414$

$y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

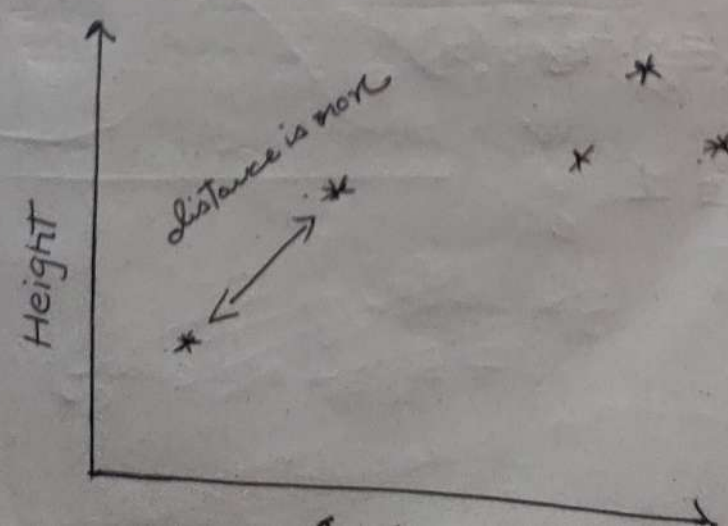


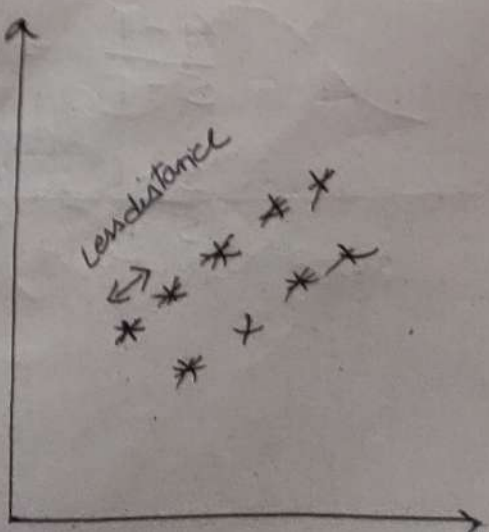
→ Why to Convert ?

Age (years)	Weight (kg)	Height (cm)
24	72	150
26	78	160
32	84	165
33	92	178
34	87	150
28	83	180
29	80	175

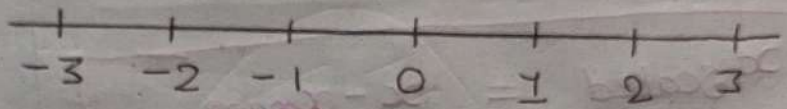
Machine Learning {Maths equation ↓ applied}

Algorithm → Mathematical Model





$y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

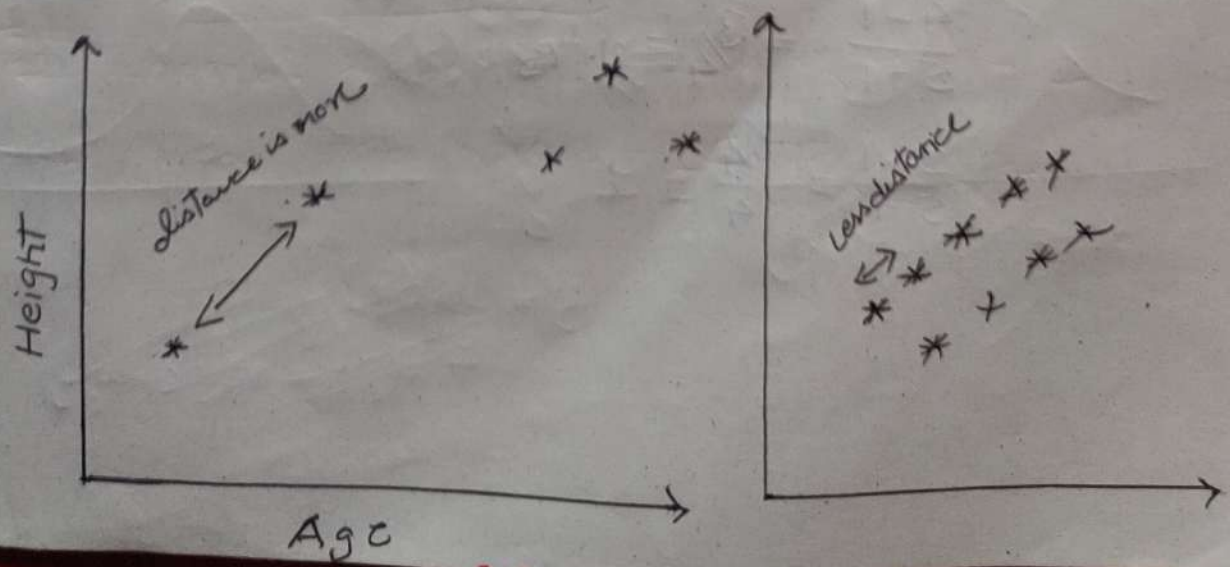


→ Why to Convert ?

Age (years)	Weight (kg)	Height (cm)
24	72	150
26	78	160
32	84	165
33	92	178
34	87	150
28	83	180
29	80	175

Machine Learning {Maths equation applied}

Algorithm → Mathematical Model



Z-score rule helps to put data in a same scale, this is called standardization; where μ (mean) = 0 & σ (standard deviation) = 1.

* Feature Scaling :-

(1)* Normalization => We try to normalise data from lower scale to higher scale.

(i) Min Max Scalar : [0-1] [Value varies b/w 0-1]

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$x = \{1, 2, 3, 4, 5\}$

$y = \{0, 0.25, 0.5, 0.75, 1\}$

Applying formula :

$$\frac{1-1}{5-1} = 0$$

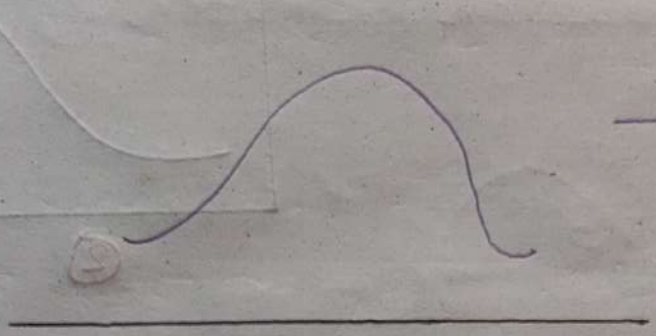
$$\frac{2-1}{5-1} = \frac{1}{4} = 0.25$$

$$\frac{3-1}{5-1} = \frac{1}{2} = 0.50$$

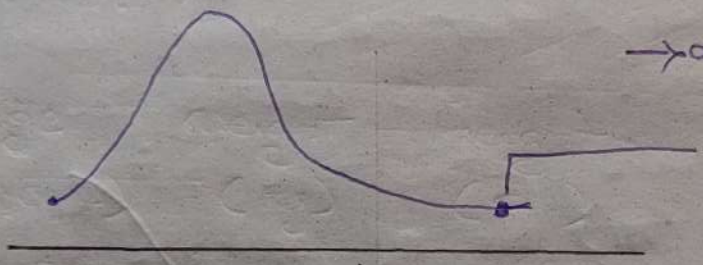
$$\frac{4-1}{5-1} = \frac{3}{4} = 0.75$$

$$\frac{5-1}{5-1} = \frac{4}{4} = 1$$

* Log Normal Distribution :



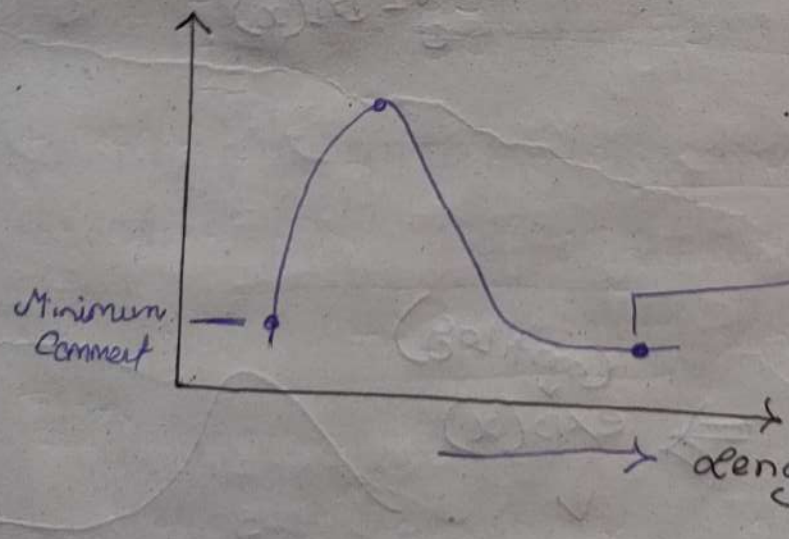
→ Normal / Gaussian distribution



→ Log Normal Distribution

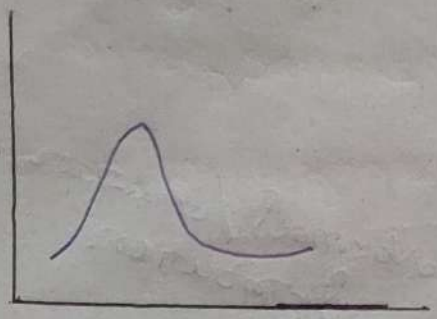
Maximum (Richer)

example → Wealth Distribution

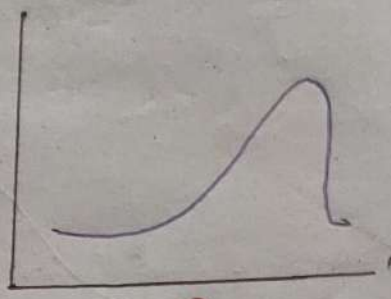


Assignment

Q: From ascending order give the relation of mean, median, mode from given graph

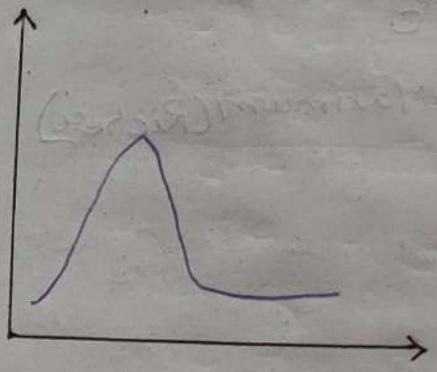


(1)

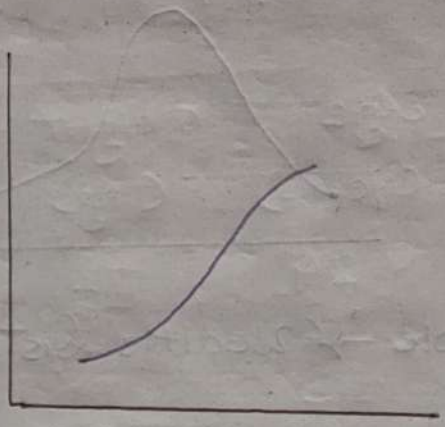


(2)

*

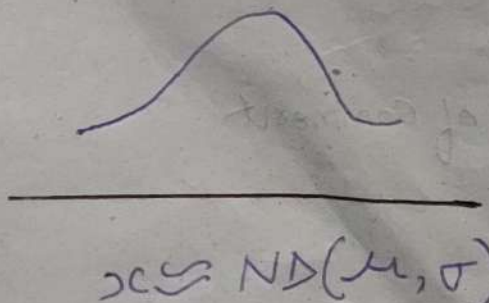


$x \approx \log ND$



$y = \ln(x)$

*



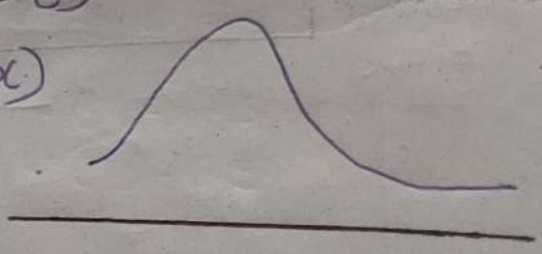
$x \sim ND(\mu, \sigma)$

(antilog)

$\Rightarrow \exp(x)$

$\downarrow$

y



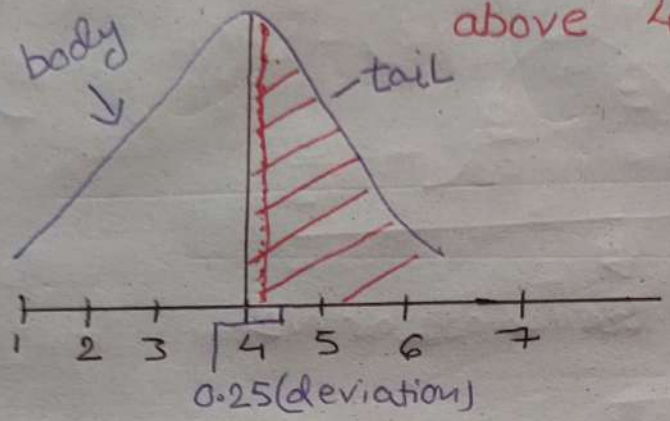
$y = \ln(x)$

$\rightarrow \log_e$ (natural log)

* Inversing LND to ND: by applying: $\log_e(x)$

* " ND to LND: by " : ~~exp(xx)~~ $\exp(x)$

Q $x = \{1,2,3,4,5,6,7\}$; if $\mu = 4$ & $\sigma = 1$, then what is the percentage of scores that falls above 4.25?



$$\text{Z score} = \frac{4.25 - 4}{1} = 0.25$$

Ans 0.25 deviation from mean.

* Z-table (Area under the table): There are two types of z-table: (i) Negative z-table (2) Positive z-table or left-z-table & right z-table respectively.

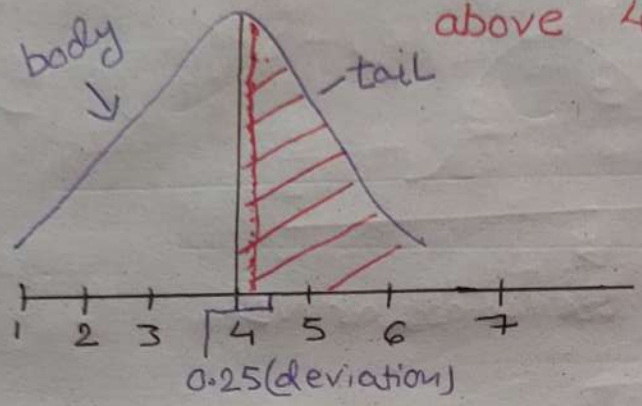
→ According to z-table area under z-score 0.2 would be 59% then area of the tail will be $1 - 59\% = \boxed{0.41\%}$

Ans

* Inversing LND to ND: by applying: $\log_e(x)$

* " ND to LND: by " : ~~exp(x)~~ $\exp(x)$

Q. $x = \{1,2,3,4,5,6,7\}$; if $\mu = 4$ & $\sigma = 1$, then what is the percentage of scores that falls above 4.25?



$$\text{Z score} = \frac{4.25 - 4}{1} = 0.25$$

Ans 0.25 deviation from mean.

* Z-table (Area under the table): There are two types of z-table: (i) Negative z-table (2) Positive z-table or left z-table & right z-table respectively.

→ According to z-table area under z-score 0.2 would be 59% then area of the tail will be $1 - 59\% = \boxed{0.41\%}$

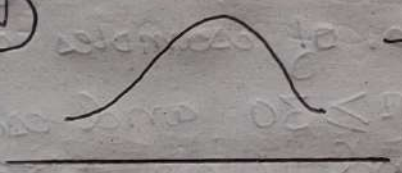
Ans

* Agenda :

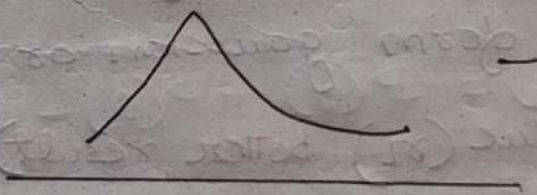
1. Central Limit Theorem
2. Probability
3. Permutation & Combination
4. Covariance, Pearson Corelation, Spearman Rank Corelation
5. Bernollis Distribution
6. Binomial Distribution
7. Power Law {Perito Distribution}

1. Central Limit Theorem:

Population Data (N)
(all graph are N)



— Gaussian / Normal Distribution



— LND

— Left Skewed

$N \geq 30$

Sample Data (n) ← extracting mean from sample data

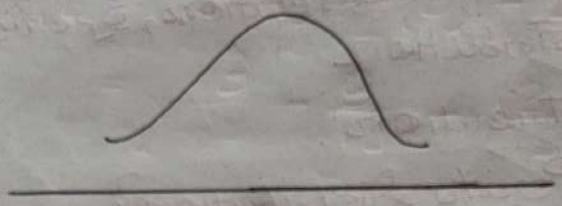
$S_1 - \{x_1, x_2, x_3 \ldots x_n\} \rightarrow \bar{x}_2$
$S_2 - \{x_1, x_2, x \ldots x_n\} \rightarrow \bar{x}_3$
$S_3 - \{x_1, \ldots x_n\} \rightarrow \bar{x}_4$
$S_m - \{x_1, \ldots x_n\} \rightarrow \bar{x}_n$

} sample mean $(\bar{x}_m)$ of s_m

n → Size of sample

m → Number of samples

this, will form Gaussian Distribution



↓

Gaussian or ND

(i)

(ii)

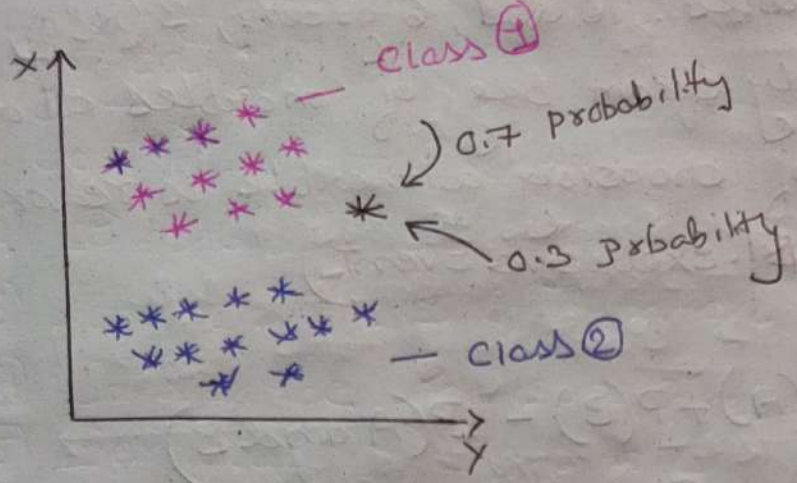
Central theorm says that whether your population data(N) may be gaussian normal distributed or is not normal gaussian distributed or any but if we take sample data (m) no. of samples and if size of the sample is $n \geq 30$ and some no. of samples and if we take mean of all the samples and put it in histogram this will form gaussian or ND.

→ (Larger the value (n) better result)

② Probability : Probability is a measure of the likelihood of an event.

example : (i) Tossing a fair coin $P(H) = 0.5$
$P(T) = 0.5$

(ii) Rolling a dice = $P(1) = 1/6$
$P(2) = 1/6$
$P(3) = 1/6$
$P(4) = 1/6$



(i) → Mutual Exclusive Event : Two events are mutual exclusive, if they cannot occur at the same time.

Ex : ① Tossing a coin ② Rolling a dice

(ii) → Non-mutual exclusive event : Two events can occur at the same time.

Ex : (i) Picking a card from a deck of cards, two events "heart" and "king" can be selected
(ii) Bag of marble.

Q- What is the probability of a coin landing on heads or tails?

Ans :

$$P(A \text{ or } B) = P(A) + P(B)$$
$$= \frac{1}{2} + \frac{1}{2} = 1$$

Q - What is the probability of getting 1, or 6 or 3 while rolling a dice?

Ans :

$$P(1 \text{ or } 6 \text{ or } 3) = P(1) + P(6) + P(3)$$
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \text{ Ans}$$

Q- What is the probability of rolling a "5" & then a "3" with a normal 6 size dice?

Ans:

$$P(1) = 1/6 \;,\; P(2) = 1/6$$

[Multiplication Rule of independent event]

$$P(A) \& (B) = P(A) * P(B)$$

$$= 1/6 * 1/6 = 1/36$$

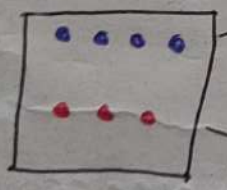
$P(A \text{ or } B) \Rightarrow$ → Mutual Exclusive (ME)
→ Non-Mutual Exclusive (NME)

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \rightarrow NME$$

$$P(A \text{ or } B) = P(A) + P(B) \rightarrow ME$$

→ Dependent Event : used in Naive Bayes in ML.

ex : Bag of marbles.



Blue marbles (B)
Possibility = ?
Red marbles (R)

$P(B) = \frac{4}{7}$ ↓ after picking one blue marble

$P(R/B) = \frac{3}{6} = 1/2$

$P(B \text{ and } R) = P(B) * P(R/B)$

$= \frac{4}{7} * \frac{3}{6} = 2/7$ Ans

→ Independent Event :

$$P(A \text{ and } B) = P(A) * P(B)$$

Ex: Tossing a coin, rolling a dice

* Permutation :

Q:- School of children

{Dairy milk, Kitkat, Milkybar, Sneaker, 5-star}

] Picking random chocolate
Possibility of total ways will be

$5 * 4 * 3$

= 60 ways → (Permutation)

→ All the possible arrangement.

$^{n}P_{r}$ where; n = Total no. of objects
r = no. of selections

↳ $\frac{n!}{(n-r)!}$

then, going to upper school based ques :-

$$\frac{5!}{(5-3)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1}$$

$$= 60$$

* Combination :

Repeation will not occur, only unique combination occur.

$^{n}C_{R}$

$$\frac{n!}{r!(n-r)!}$$

going to upper ques = $\frac{5 \times 4 \times 3 \times 2 \times 1}{3! \times 2}$

$$= 10$$

* Covariance :

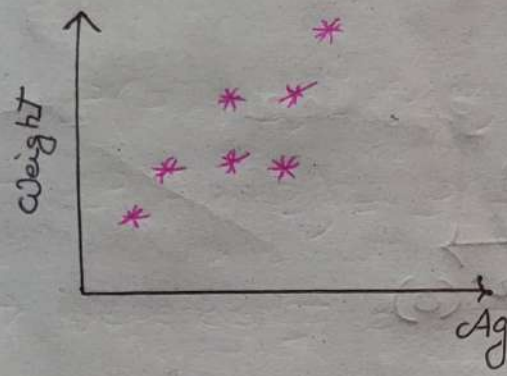
{Feature Selection}

ex:-

Age	Weight
12	40
13	45
15	48
17	60
18	62

Age ↑	Weight ↑
Age ↓	Weight ↓

↑↑ Quantity the relationship x & y using



$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

it, is similar to variance's formula.

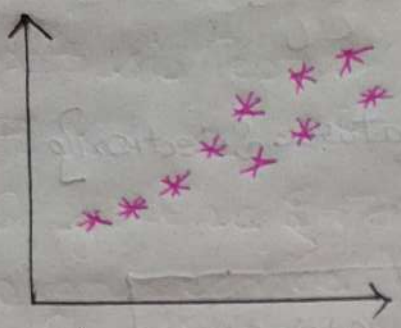
{ $Cov(x,x) = Var(x)$ } → important for interview

if Cov is (+) then, | x↑ y↑ / x↓ y↓ | = equally proportional

if cov is (-) then, | x↑ y↓ / x↓ y↑ | = inversely proportional

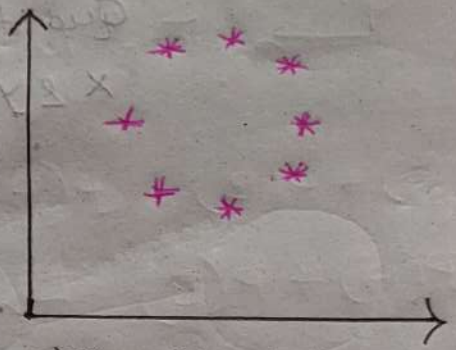
if cov is (0) then, | No relation b/w x & y | = no relation

Graph → Covariance through graph:



(i) (+) covariance
(0 to 1)

(ii) (-) Covariance
(-1 to 0)



(iii) no relation (0)

* Pearson Corelation Co-efficient : (ρ) row

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

Restricting the range of covariance b/w (-1 to 1) - defining the scale.

— more the value to +1 → more (+) corelated.

— more " " " -1 → " (-) " .

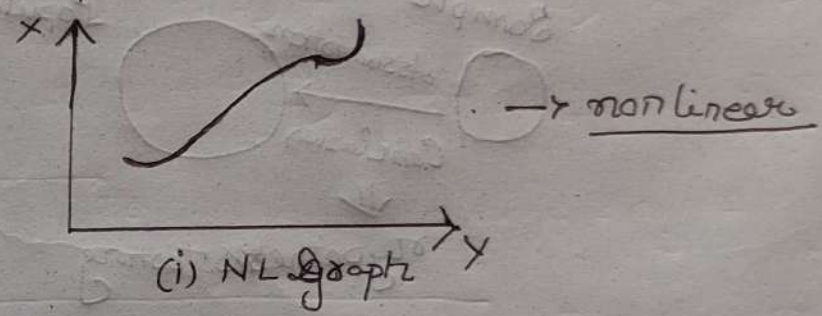
→ It only hold for linear data.

* Spearman Rank Corelation : It only hold for non-linear datas
It assigns ranks.

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

ex-(i) Rank follow in acending order :

x	y	R(x)	R(y)
10	4	4	1
8	6	3	2
7	8	2	3
6	10	1	4



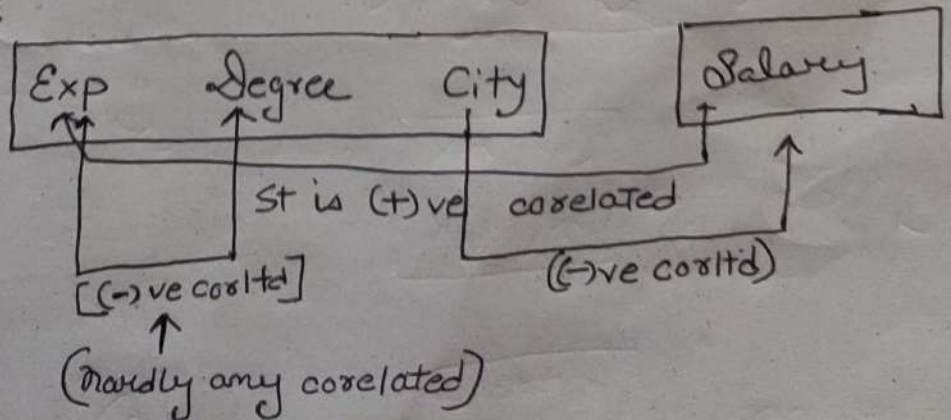
(i) NL graph

Q Why this corelation is used ?

Ans

→ It is important to predict the output

exmp :



We are not Loosing data by this.