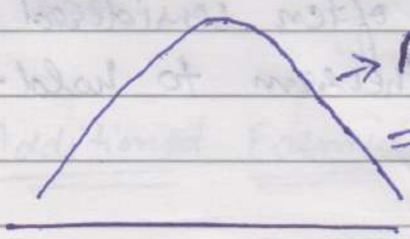
DAY 4 - STATISTICS.

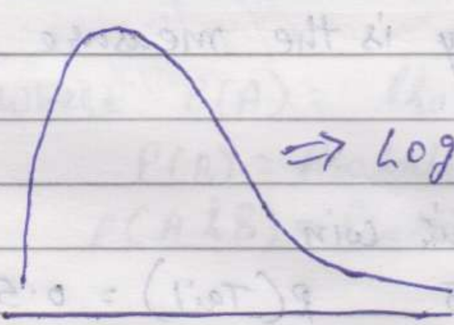
1. Central Limit Theorem
2. Probability
3. Permutation And Combination.
4. Covariance, Pearson Correlation, Spearman Rank Correlation.
5. Bernoulli's Distribution.
6. Binomial Distribution.
7. Power Law [Pareto Distribution]

* Central Limit Theorem:-



→ Population data [N]

⇒ Gaussian/Normal distribution.



⇒ Log Normal distribution.

⇐ Left Skewed.

Sample data $\boxed{n}$

$$\{x_1, x_2, x_3 \cdots\cdots x_n\} \rightarrow \bar{x}_1$$
$$\{x_1, x_2, x_3 \cdots\cdots x_n\} \rightarrow \bar{x}_2$$
$$\{x_1, x_2, x_3 \cdots\cdots x_n\} \rightarrow \bar{x}_3$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$\{x_1, x_2, x_3 \cdots\cdots x_n\} \rightarrow \bar{x}_n$$

The Central Limit Theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution. Sample size equal to or greater then 30 are often considered sufficient for the central limit theorem to hold.

★ **Probability** =
Probability is the measure of the likelihood of an event.

Example:- Tossing a fair coin
$P(\text{Head}) = 0.5 \qquad P(\text{Tail}) = 0.5$

Example 2 :- Rolling a Dice
$P(1) = \frac{1}{6} \qquad P(2) = \frac{1}{6} \qquad P(3) = \frac{1}{6}$

① ⇒ Mutual Exclusive Event :-
Two Events are mutually exclusive if they cannot occur at the same time.
$$P(A \text{ or } B) = P(A) + P(B)$$
Eg:- Tossing a Coin   Eg :- Rolling a dice.

(2) => Non - Mutual Exclusive Events:-

Two Events can occur at the same time.

Eg:- Picking randomly a card from a deck of cards, two events "Heart" and "King" can be selected.

Eg:- Bag of Marbles:- 10 Red, 6 Green, 3 (R & G)

Red Marble Green Marble Red and Green Marbel

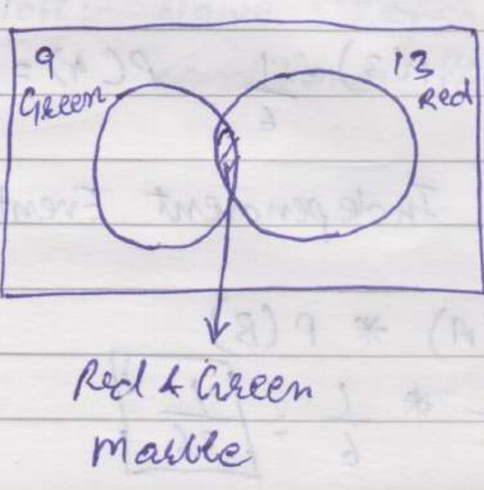
There is a probability that we choose red and Green marble.

* Additional Formula for Non Mutual Exclusive Event

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

Where $P(A)$ = Probability of Red marble picked up
$P(B)$ = Probability of Green marble picked up.
$P(A \& B)$ = Probability of Red and Green marble picked up.

~~$\therefore P(A \text{ or } B) = \frac{10}{19} + \frac{6}{19} - \frac{3}{19} = \frac{13}{19}$~~



$$P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$$
$$= \frac{13}{19} + \frac{9}{19} - \frac{3}{19}$$
$$= \frac{19}{19} = 1$$

Another Example → What is the probability of choosing Heart ♡ or Queen.

$$P(\heartsuit \text{ or Queen}) = P(\heartsuit) + P(\text{Queen}) - P(\heartsuit \text{ or Queen})$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \boxed{\frac{16}{52}}$$

* Multiplication Rule for Non Mutual Exclusive Event.

* Dependents Evens:- Two Events are dependent if they affect one another.

White Marble

Bag of marble $\left\{ \begin{matrix} o & o & o & * \\ o & o & o & \end{matrix} \right\}$

Yellow marble

$$P(\text{white marble}) = \frac{4}{7} \longrightarrow P(\text{yellow marble}) = \frac{3}{6}$$

White 7 marble.

* Independent Events.-

Question:- What is the probability of rolling a "5" and then a "3" with a normal 6 sided dice?

Ans:- $P(1) = \frac{1}{6}$ $P(2) = \frac{1}{6}$ $P(3) = \frac{1}{6}$ $P(4) = \frac{1}{6}$

Multiplication Rule For Independent Events.

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

$$= \frac{1}{6} * \frac{1}{6} = \boxed{\frac{1}{36}}$$

* Example of Dependent Events:-

Bag of marble $\left\{ \begin{array}{c} \overbrace{0\,0\,0\,0}^{\text{White marbles}} \\ \underbrace{0\,0\,0}_{\text{Yellow marbles}} \end{array} \right\}$

Q:- What is the probability of drawing a "White" and then drawing a "yellow" marble from the bag?

Ans

$\boxed{\begin{array}{c} 0\,0\,0\,0 \\ 0\,0\,0 \end{array}} \rightarrow P\left(\text{White marble}\right) = \dfrac{4}{7}$

Now If we remove 1 White marble then.

$\boxed{\begin{array}{c} 0\,0\,0 \\ 0\,0\,0 \end{array}} \rightarrow P\left(\dfrac{\text{Yelow marble}}{\text{White marble}}\right) = \dfrac{3}{6} \rightarrow \boxed{\text{Conditional Probability}}$

$$\boxed{P(\text{White and Yellow}) = P\left(\text{White marble}\right) * P\left(\dfrac{\text{Yellow}}{\text{White marble}}\right)}$$

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

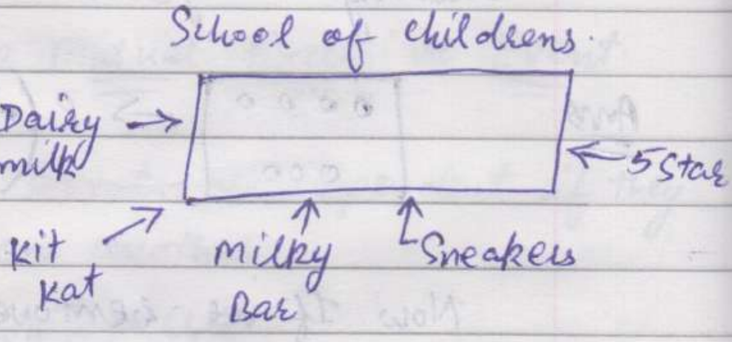
Note:- Naive Baye's ML Algorithm is derived from conditional Probability.

★ $\boxed{\text{Permutation :-}}$

A permutation is an arrangement of objects in a definite order.

For Eg:- The permutation of set $A = \{1, 6\}$ is 2 Such as $\{1, 6\}$, $\{6, 1\}$. As you can see, there are no other ways to arrange the elements of set A.

ANOTHER EXAMPLE.



$\underline{5} * \underline{4} * \underline{3}$

= 60 ways

⇓

Permutation.

means there will be 60 possible arrangements like.

{DM KK MB} { } { }
{KK DM MB} { } { }
{KK MB DM} { } { }
{ } { } { } and so on.

$$\boxed{{}^{n}P_{r} = \frac{n!}{(n-r)}} = \frac{5!}{(5-3)!}$$

$$= \frac{5 \times 4 \times 3 \times \cancel{2} \times \cancel{1}}{\cancel{2} \times \cancel{1}} = 60$$

where n = Total no. of objects
r = No. of Selections.

★ $\boxed{\text{Combination:-}}$ {Repeatation will not occur}

In combination all 60 possible arrangements are not possible only unique groups are possible.

$$\boxed{{}^{n}C_{r} = \frac{n!}{r!(n-r)!}} = \frac{5!}{3!(5-3)!}$$

$$= \frac{5 \times 4 \times 3 \times 2 \times 1}{3!\,(2!)} = \frac{5 \times \overset{2}{\cancel{4}} \times \cancel{3 \times 2} \times 1}{\cancel{3 \times 2} \times \cancel{2}}$$

$$= \boxed{10}$$

Where n = Total no. of objects.
r = No. of selections.

★ $\boxed{\text{Covariance}}$:- {Feature Selection}

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

Age ↑	weight ↑
Age ↓	Weight ↓

⇑

Quantify the relationship of x & y using mathematical question.

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

This formula is derived from variance.

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \quad \therefore S^2 = \frac{\sum(x_i - \bar{x}) * (x_i - \bar{x})}{n-1}$$

$\therefore$ We can say that

$$Cov(x, x) = Var(x)$$

Now According to our Age and Weight data.

$$Cov(x, y) = \frac{\sum (x_i - \bar{x}) * (y_1 - \bar{y})}{n-1}$$

~~Here $x_1 - \bar{x}$~~

		$x_i - \bar{x}$	$y_1 - \bar{y}$
	~~27~~	$12 - 15 = -3$	$40 - 51 = -11$
	~~28~~	$13 - 15 = -2$	$45 - 51 = -6$
$\bar{x} = 15$	~~30~~	$15 - 15 = 0$	$48 - 51 = -3$
$\bar{y} = 51$	~~32~~	$17 - 15 = 2$	$60 - 51 = 9$
$n = 5$	~~33~~	$18 - 15 = 3$	$62 - 51 = 11$

$$\text{Now } \sum (x_i - \bar{x}) * (y_1 - \bar{y}) = (-3 \times -11) + (-2 \times -6) + (0 \times -3) + (2 \times 9) + (3 \times 11) = 96$$

$$\therefore Cov(x, y) = \frac{96}{n-1} = \frac{96}{4} = \boxed{24}$$

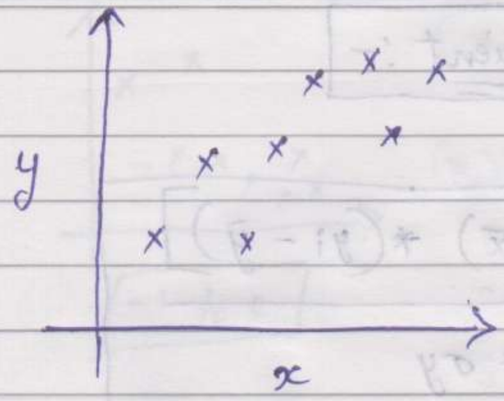
24 is +ve covariance.

If +ve covariance $\boxed{\begin{matrix} x\uparrow & y\uparrow \\ x\downarrow & y\downarrow \end{matrix}}$ x increase y increase.

If -ve covariance $\boxed{\begin{matrix} x\uparrow & y\downarrow \\ x\downarrow & y\uparrow \end{matrix}}$

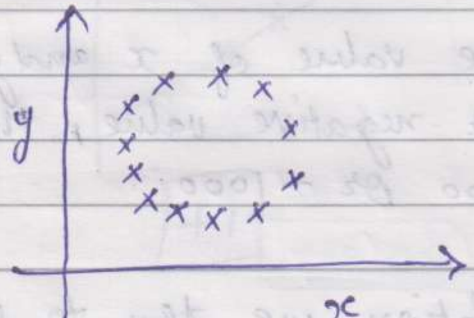
x increase then y decrease.

If covariance is 0 $\boxed{\text{No relation with } x \text{ \& } y}$

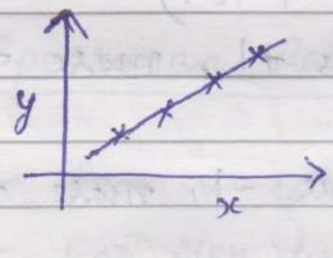


=> +ve covariance.

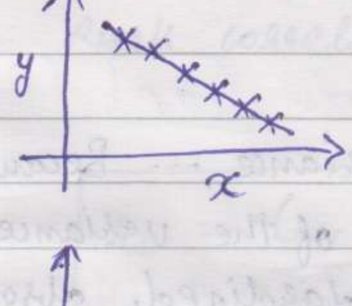
=> -ve covariance.



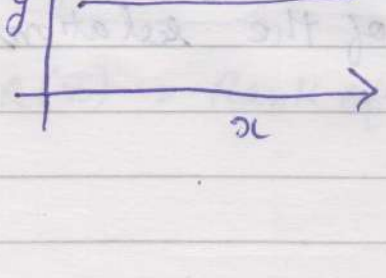
=> 0 covariance, No relation with x & y



=> +ve covariance.



=> -ve covariance.



=> 0 covariance, No relation with x & y.

Pearson Correlation Coefficient :-

rho $$\boxed{\rho(x,y) = \frac{\Sigma[(x_i - \bar{x}) * (y_i - \bar{y})]}{\sigma x * \sigma y}}$$

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

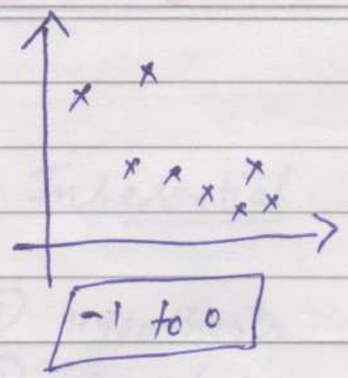
Difference between Pearson Correlation and Covariance.

In Covariance(x,y), the value of x and y can be any positive value or negative value, there is no limit, it can be + 1000 or - 1000

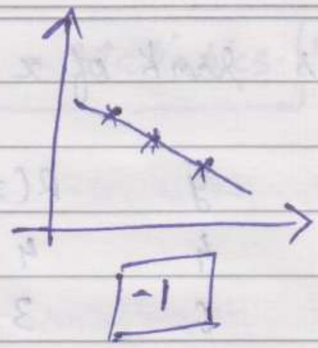
By using Pearson Correlation, we try to restrict value of x and y between (-1 to 1)

Note :- - more the value towards +1, more +ve correlated it is.
- more the value towards -1, more -ve correlated it is.

Correlation is better than covariance --- Because correlation removes the effect of the variance of the variables, it provides a standardized, absolute measure of the strength of the relationship, bounded by -1.0 and 1.0.

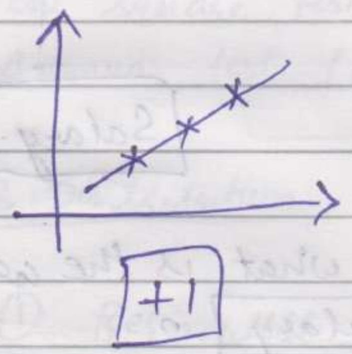


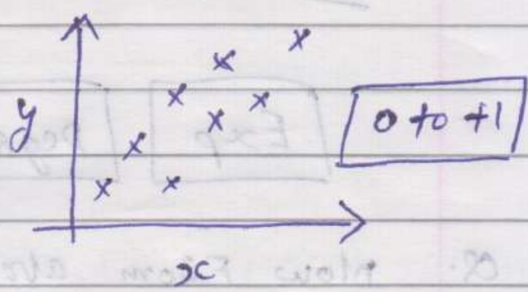
Here we are getting $\boxed{-1 \text{ to } 0}$, because all are not falling in the same line.



Since all the points are in same line rho(ρ) is -1

Similarly





★ **Spearman Rank Correlation:-**

- Pearson Correlation is only good for Linear Data.
- For Non linear data we have to use Spearman Rank correlation.

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

$R(x)$ = Rank of x

What is rank of x? $R(x)$

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

$R(x)$ - 1,2,3,4 are the ranks assigned to x, and all the ranks are assigned in the ascending order.

Why this correlation will be used?

Exp   Degree   City   Salary.

Q. Now From above example what is the correlation between Exp and Salary?

Ans. It is +ve correlated.

Q. What is the correlation between City and Salary?

Ans. +ve correlated as the city is good like banglore or girgaon the salary is much better.

Q. What is the correlation between Exp, Degree?

Ans. No relation.

Q. Correlation between Degree and Salary?

Ans. +ve correlated.