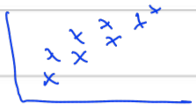
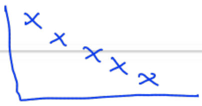
V) Covarience & Correlation :-

Suppose we have 2 features or 2 Random Variables (X, Y) :-

X	Y
2	3
4	5
6	7
8	9

So, can we find a relationship b/w X and Y. (i.e like if X ↑es then Y ↑es etc.)

Plotting : $\boxed{X\uparrow, Y\uparrow \quad X\downarrow, Y\downarrow}$ ≡ 

Plotting : $\boxed{X\downarrow, Y\uparrow \quad X\uparrow, Y\downarrow}$ ≡ 

5.1) Covarience (X, Y) or Cov (X, Y) :-

now, covarience w.r.t random variables X and Y is given by formulae :-

$$\boxed{Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}}$$

Note ⟹ What the difference b/w Varience and Covarience

Varience $(X^2) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$ & Covarience $(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$

(signifies spread)

So; Varience is nothing but Covarience of random variable X with itself i.e Cov(X, X)

i.e

$$Cov(X,X) = \frac{\sum_{i=1}^{n}(X_i-X)(X_i-X)}{n-1} = \frac{\sum_{i=1}^{n}(X_i-X)^2}{n-1} = Var(X^2)$$

∘) For $(X\uparrow, Y\uparrow / X\downarrow, Y\downarrow)$ Relationship : Covarience = +ve

X	Y
2	3
4	5
6	7
$\bar{X}=4$	$\bar{Y}=5$

$$Cov(X,Y) = \frac{(2-4)(3-5)+(4-4)(5-5)+(6-4)(7-5)}{3-1}$$

$$= \frac{4+0+4}{2} = \boxed{4}\ (+ve)$$

$\therefore$, X and Y are having a +ve covarience

Q7 Example where Covarience can be used?

Sol^n -> Suppose we have a dataset where Price and size of different houses is given, and In such case we have to predict the price of the house
So, Covarience can be used to find the relationship b/w the size and prices of the house.

Advantages and Disadvantages of Covarience:-

- **Advantage:-** Using covarience we can find a relation b/w X and Y. It can be +ve or -ve

- **Disadvantage:-** Covarience doesnot have a specific limit value

  ex: Suppose we have X, Y, Z
  if $Cov(X,Y) = +100$ & $Cov(X,Z) = +500$
  it does not mean X is heavily corelated to Z as compared to Y coz there is no specific limit to Cov value to compare high or low

So;
in order to overcome this advantage of covarience we specifically use another corelation technique c/a "Pearson Corelation Coefficient"

5.2) Pearson Correlation Coefficient(ρ)[range: -1 to +1]

as its value is restricted b/w -1 and +1 ∴, now we can differentiate which variables are highly correlated and which are not.

$$\boxed{\text{Correlation}(X,Y)(\rho_{X,Y}) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = [-1 \text{ to } 1]}$$

- more the value is towards +1, the more +vely correlated X & Y are.
- more the value of ρ_{XY} is towards -1, the more -vely correlated X & Y are

Q) Where can this correlation concept is applied in data science project?

Sol^n Suppose a dataset with 1000 features,
now, for an ML project it is not feasible to take all of the features

Ex: dataset features: Size of house, No of Rooms, Location, No. of PPL staying, Price

$\underbrace{\text{Size of house, No of Rooms, Location, No. of PPL staying}}_{\text{Independent features}}$ $\underbrace{\text{Price}}_{\text{O/P or dependent feature}}$

now using Feature Selection :-
if Cov (No. of ppl, Price) ≈ or near 0 then it can be removed from our ML model.

5.3) Spearman Rank Correlation :- (r_s)

Sometime it is consider to be better than Pearson Correlation.

here instead of considering the value of X, & Y we consider the rank of both X and Y i.e

$$\boxed{r_s = \frac{Cov[R(X), R(Y)]}{\sigma_{R(X)} \sigma_{R(Y)}}}$$

ex: ij

X	Y	R(X)	R(Y)
1	2	5	5
3	4	4	4
5	6	3	3
7	8	2	1
0	7	6	2
8	1	1	6