variable - column / attribute

x	y
1	1
2	4
3	9
4	16
5	25

y depends upon x

$y = x^2$ → formula/ model

Covariance

Correlation

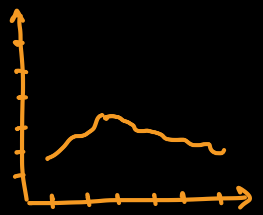
Measures of Correlation

relationship between variables

Terminologies

temperatures = [28, 29, 30, 28, 27..]

marks = [80, 85, 90, 70, 60..]



temperatures

marks

- **Univariate**
  - This type of data consists of only one variable
  - It does not deal with causes or relationships → can not be used to find the model
  - the main purpose of the analysis is to describe the data and find patterns that exist within it

↳ data visualization

↳ trend

- **Bivariate** → two variables
  - This type of data involves two different variables
  - The analysis of this type of data deals with causes and relationships
  - the analysis is done to find out the relationship among the two variables

Experience	1	2	2.5	4.5	5	6
salary	10	15	18	40	50	55

# study hrs	2	5	3	4	6
marks	80	85	70	90	95

- **Multivariate** → ≥ three variables
  - When the data involves three or more variables
  - It is similar to bivariate but contains more than one ~~dependent~~ independent variables

(x)	Experience	1	3	2.5	4.5	5	6	8.5
(y)	salary	10	15	18	40	50	55	?

model / formula ⇒ $\overset{\overset{\text{dependepent}}{\downarrow}}{y} = f(\underset{\underset{\text{independent}}{}}{\underset{\downarrow}{x}})$

variable

→ dependent :

- → also known as output variable (y)
- → value of this variable needs to be predicted
- → in any ML problem there will be one & only one dependent variable

→ independent :

- → also known as input variable (x)
- → value of this variable will be used to predict the dependent var
- → in any ML problem there may be one or more independent variables

house price prediction

		X1		X2	(y)	X3
No	owner name	#rooms	color	sq. area	price	location
1	abcd	2	white	1000	1cr.	swargate
2	pqrs	3	gray	1500	1 cr.	Hinjawadi
3	xyz	4.5	red	2500	2cr.	Sinhgad Road
4	acd	2.5	pink	1200	2.5 cr.	Kalyani Nagar
5	yz6	1	purple	500	1.5 cr.	Boat club

dependent variable : price

independent variable(s) : #rooms, sq.area, location

Introduction

- In practice, we come across a large number of problems involving the use of two or more number of variables → multivariable
- If two quantities vary in such a way that, movements (changes) in one are accompanied by movements in the other, these quantities are correlated
- E.g. price of commodity and its demand, experience of an employee and his salary etc.
- The degree of relationship is measured between the variables under consideration is measured through the correlation analysis
- The problem of analyzing relationship between variables can be divided into following steps
  - Determine if the relationship exists between the variables
  - If it exists, then measure it using measure of correlation
  - Testing whether it is significant
  - Establish the cause and effect relation if any → causal relationship

one variable causes change in another variable

Significance of study of correlation

- Most of the variables show some kind of relationship. With the help of correlation analysis, we can measure its degree of relationship in one figure → strong, weak, no
  - E.g. relationship between price and supply
- Once we know that two variables are closely related, we can estimate the value of one variable given the value of another

  ↳ measure ⇒ $y = f(x)$ = model
- Correlation analysis contributes to the understanding of economic behavior, aid in locating the critically important variables on which other depends
- In business, the correlation analysis enables the execution to estimate costs, sales, prices and other variables on the basis of some other series with which these my be functionally related

Correlation and Causation

Correlation analysis does NOT provide causal relationship

Regression analysis

- Correlation analysis helps us in determining the degree of relationship between two or more variables
- It does not tell us anything about the cause and effect relationship
- Even a high degree of correlation does not necessarily mean that the relationship of cause and effect exists between the variables
- The explanation of significant degree of correlation may be one or both of the following
  - The correlation may be due to pure chance, especially in a small sample
  - Both the correlated variables may be influenced by one or more other variables
  - Both the correlated variables are mutually influencing each other

-ve correlation

No Correlation

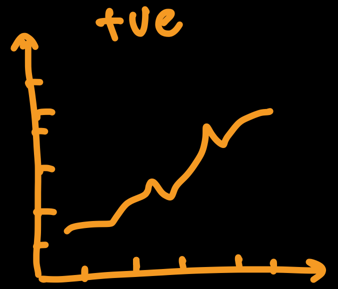
+ve correlation

Types of Correlation

Positive and Negative Correlation

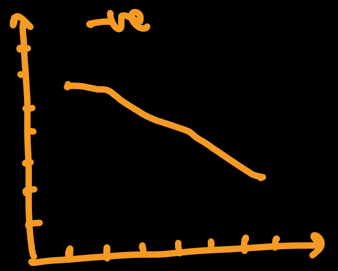
- Whether the correlation is positive or negative would depend upon the direction of the change of the variables
- If both he variables are varying in the same direction i.e., if as one variable is increasing and the other one is also increasing or vice-a-versa, is said to be positive correlation
- On the other hand, if the variables are varying in opposite directions, the correlation is said to be negative correlation
- E.g.
  - Positive

X	10	12	15	18	20	24	30
Y	15	20	25	30	35	40	45



  - Negative

X	10	12	15	18	20	24	30
Y	45	40	35	30	25	20	15



Simple, Partial and Multiple Correlation

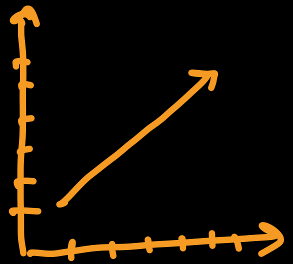
in ML, only simple correlation can be used

- The distinction between simple, partial and multiple correlation is based upon the number of variables under study → dependent variable
- When only two variables are studied, it is a problem of simple correction → only one dependent variable
- When three or more variable are studied, it is a problem of either multiple or partial correlation
- In multiple correlation, three or more are studied simultaneously
  - E.g. studying relationship between yield of rice per acre and amount of rainfall and amount of fertilizers used
- On the other hand, in partial correlation, we recognize more than two variables, but consider only two variables keeping other constant

Linear and Non-Linear Correlation

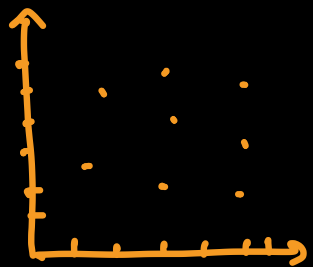
- The distinction between line and non-linear correlation is based upon the constancy of the ratio of change between the variables
- Linear Correlation
    - If the change in one variable tends to bear constant ratio to the amount of change in the other variable, then the correlation is said to be linear
    - If the variables are plotted on graph, all potted points would fall on a straight line

X	10	20	30	40	50
Y	70	140	210	280	350



- Non-Linear Correlation
    - Correlation is called as non-linear of curvilinear if the amount of change in one variable does not bear a constant ratio to the amount of change in other
    - If the variables are plotted on graph, the points will never fall on a straight line

X	10	20	30	40	50
Y	10	40	80	30	49



(Graph) Graphical

Algebric (formula)

- covariance
- correlation

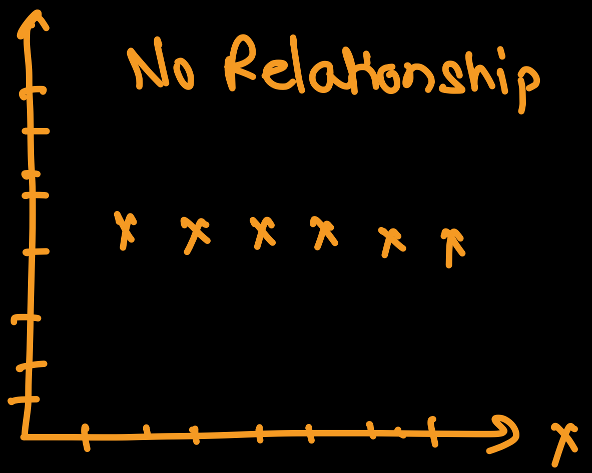
Methods of Studying Correlation

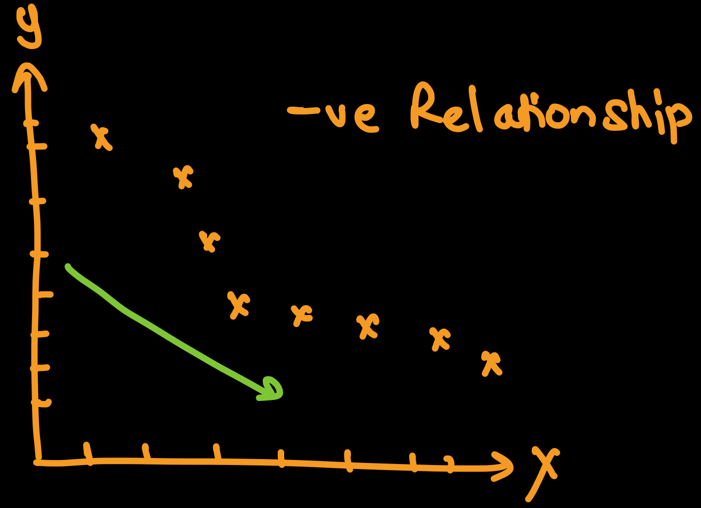
Chart : does not use any graph paper & does NOT strictly follow the scale

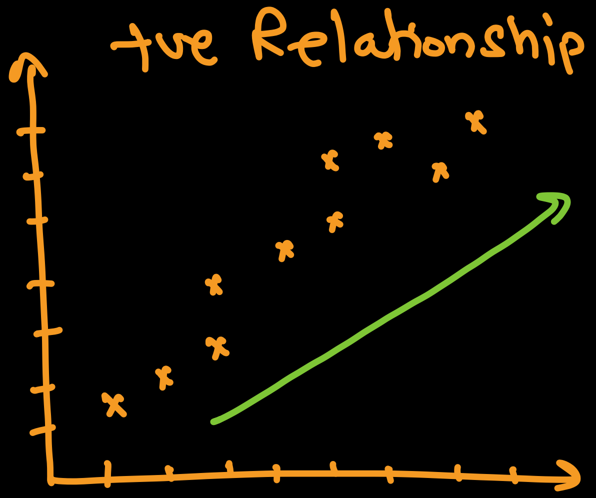
Graph : uses graph papers & uses a well defined scale

Scatter Diagram → Graphical method →

- The simplest way of asserting whether two variables are related or not is to prepare a dot chart called as scatter diagram
- When this method is used, the dataset is plotted on the graph (for every value we simply put a dot)
- By looking at the scatter of the of the various points, we can form an idea as to whether the variables are related or not







Merit and Limitations of Scatter Method

- Merits
  - It is a simple and non-mathematical method of studying the correlation between the variables
  - It can be easily understood and a rough idea can very quickly can be formed about the relationship
  - It is not influenced by the size of extreme items whereas most mathematical methods are influenced by them
  - Making a scatter diagram is usually the first step in investigating the relationship
  - If he variables are related, we can see what kind of line or estimating equation describes the relationship
- Limitations
  - By applying this method, we can get an idea about the direction of correlation, but we can not establish the exact degree of correlation between two variables

Graphic Method

- When this method is used, the individual values of the two variables are plotted on the graph paper
- We thus obtain two curves, one for X variable and another for Y variable
- By examining the direction and closeness of the two curves so drawn, we can infer whether or not the variables are related
- If both the curves drawn on the graph are moving in the same direction (either upward or downward), correlation is said to be positive
- On the other hand, if the curves are moving in the opposite directions, correlation is said to be negative

Covariance

→ algebric

- A measure of the relationship between two random variables
- The metric evaluates how much – to what extent – the variables change together
- A positive covariance would indicate a positive linear relationship between the variables
- A negative covariance would indicate the opposite

$$cov(x, y) = \frac{\sum(x - \bar{x})(y - \bar{y})}{N}$$

- Where
  - X_i – the values of the X-variable
  - Y_i – the values of the Y-variable
  - $\overline{x}$ – the mean (average) of the X-variable
  - $\overline{y}$ – the mean (average) of the Y-variable
  - n – the number of the data points in sample
  - N – the number of the data points in Population

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
1	3	-2	-4	8
2	5	-1	-2	2
3	7	0	0	0
4	9	1	2	2
5	11	2	4	8
				20

$\Sigma(x-\bar{x})(y-\bar{y})$

$$\boxed{y = 2x+1}$$

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{y} = \frac{3+5+7+9+11}{5} = 7$$

$$\mathrm{cov}(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{N}$$

$$= \frac{20}{5} = \underline{\underline{4}}$$

$$\left[-\infty \quad 0 \quad +\infty \right]$$

No strong / weak Relationship

cov(x,y) is 4, means x & y are +vely correlated

Karl Pearson Coefficient of Correlation

- Of the several mathematical methods of measuring correlation, the Karl Pearson's method popularly known as Pearson's coefficient of correlation, is most widely used in practice
- It is denoted as r

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{N\sigma_1\sigma_2}$$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{N\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	3	-2	-4	8	4	16
2	5	-1	-2	2	1	4
3	7	0	0	0	0	0
4	9	1	2	2	1	4
5	11	2	4	8	4	16
				20	10	40

$$\boxed{y = 2x+1}$$

$\bar{x} = 3 \quad \bar{y} = 7 \quad \text{cov}(x, y) = +4$

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}} \quad , \quad \sigma_1 = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.41$$

$$\sigma_2 = \sqrt{\frac{40}{5}} = \sqrt{8} = 2.8$$

$$\text{cor}(x, y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{N \times \sigma_1 \times \sigma_2}$$

$$= \frac{20}{5 \times 1.41 \times 2.8}$$

$$= \frac{20}{19.74} \approx \underline{\underline{1.0}}$$

Direct Method of finding Correlation Coefficient

- Correlation Coefficient can be calculated without taking deviations of the items either from the actual mean or assumed mean

$$r = \frac{N\sum XY - \sum X\sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\,\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Assumptions of Pearsonian Coefficient

- There is a linear relationship between the variables, i.e., when the two variables are plotted on a scatter diagram, a straight line will be formed by the points so plotted
- The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc, are affected by such forces that a normal distribution is formed
- There is a cause and effect relationship between the forces affecting the distribution of the items in the two series
- If such a relationship is not formed between the variables, i.e, if the variables are independent, there cannot be any correlation
  - For example, there is no relationship between income and height because the forces that affect these variables are not common

Merits and Limitations of Pearsonian Coefficient

- Merits
  - Amongst the mathematical methods used for measuring the degree of relationship, Karl Pearson's method is the most popular one
  - The correlation coefficient summarizes in one figure not only the degree of correlation but also the direction, i.e., whether correlation is positive or negative
- Limitations
  - The correlation coefficient always assumes linear relationship regardless of the fact whether that assumption is correct or not
  - Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted
  - The value of the coefficient is unduly affected by the extreme items
  - As compared with other methods, this method takes more time to compute the value of correlation coefficient

Interpreting Coefficient of Correlation

→ strong
→ weak
→ No

- The coefficient of correlation measures the degree of relationship between two sets of figures
- As the reliability of estimates depends upon the closeness of the relationship, it is imperative that utmost care must be taken while interpreting the value of coefficient of correlation, otherwise, fallacious conclusions can be drawn
- following general rules are given which would help in interpreting the value of r
  - When r = + 1, it means that there is perfect positive relationship between the variables
  - When r =-1, it means that there is perfect negative relationship between the variables
  - When r = 0, it means that there is no relationship between the variables, i.e., the variables are uncorrelated.
- The closer r is to +1 or -1, the closer the relationship between the variables and the closer r is to 0, the less close the relationship

Correlation coefficient

r = +1 = perfect +ve

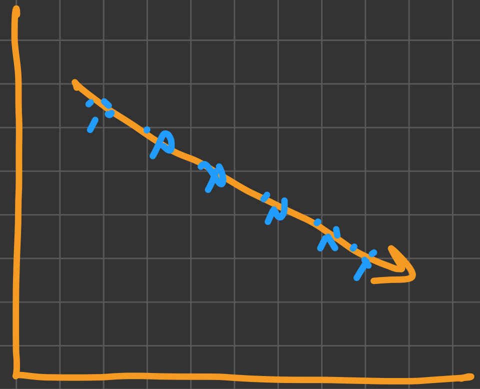
r = +0'3 = weak +ve

r = -0'3 = weak -ve

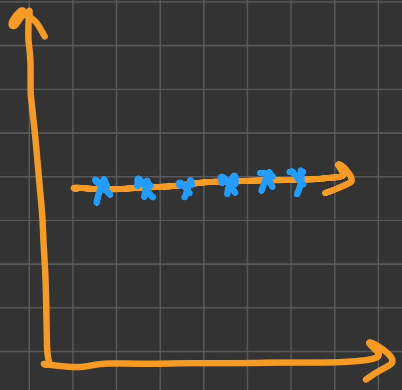
r = -0'8 = strong -ve

-1 ← stronger weaker · 0 weaker stronger → +1

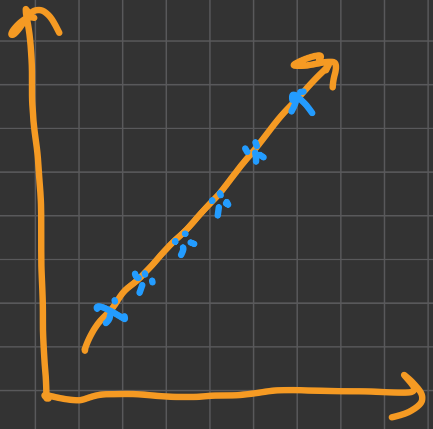
perfect -ve



No relation



perfect +ve



Properties of Correlation Coefficient

- The coefficient of correlation lies between -1 and +1, symbolically -1 <= r <= +1
- The coefficient of correlation is independent of the change of scale and origin of the variables
- It is the geometric mean of two regression coefficient * *
- The degree of relationship between two variables is symmetric
  - $r_{xy} = r_{yx}$

$$cov(x,y) = cov(y,x)$$

$$cor(x,y) = cor(y,x)$$