



# STATISTICS FOR DATA SCIENCE

## Power Test & Simple Linear Regression

---

**Dr. Karthiyayini**

Department of Science and Humanities

# STATISTICS FOR DATA SCIENCE

---

## Unit 5 : Power Test & Simple Linear Regression

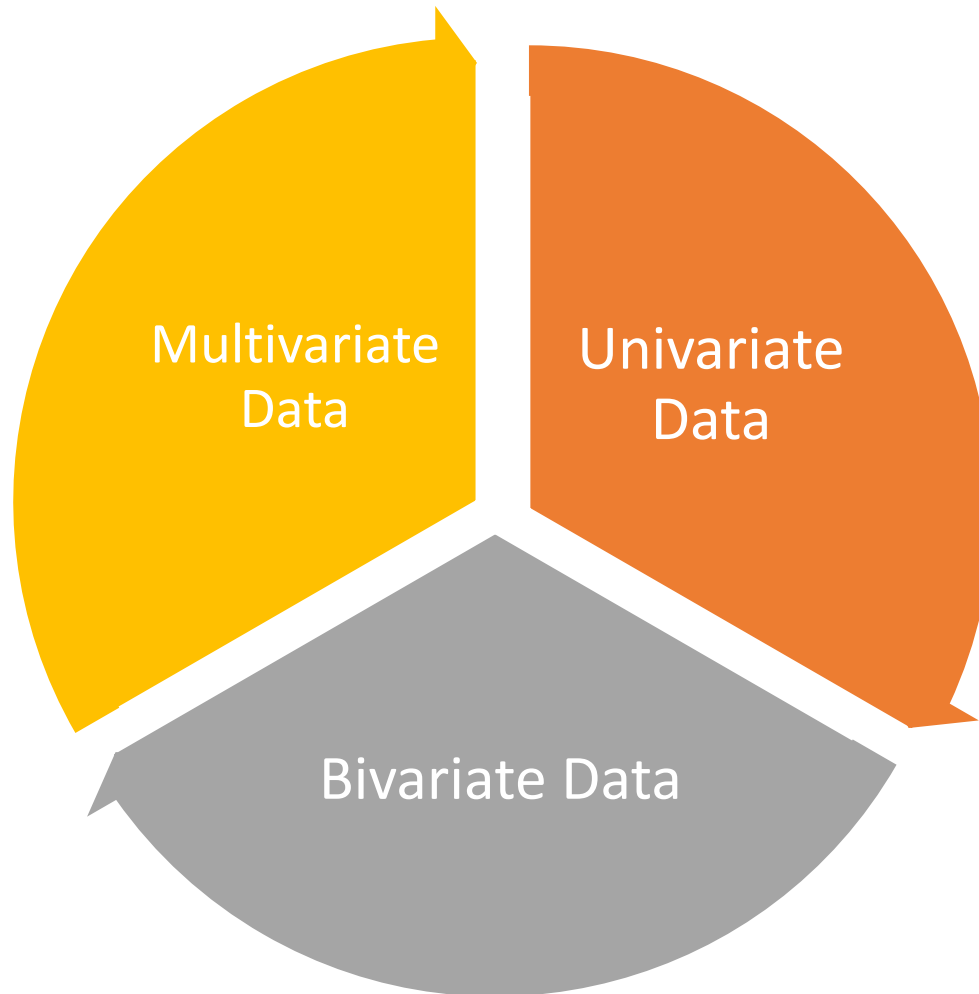
### Session : 4

### Sub Topic : Correlation

**Dr. Karthiyayini**

Department of Science & Humanities

- ❖ Classification of Data
- ❖ What is Correlation ?
- ❖ Pearson's Correlation Coefficient



# STATISTICS FOR DATA SCIENCE

## Example :



SI No.	SRN	10th Marks	12 <sup>th</sup> Marks	PESSAT Ranking	CGPA	Annual Pay Compensation
1.	PESXX001	82%	79%	1228	8.3	8 Lakhs
2.	PESXX002	85%	86%	1119	9.4	10 Lakhs
3.	PESXX003	76%	77%	1302	8.3	7 Lakhs
4.	PESXX004	69%	75%	1356	8.2	6 Lakhs
5.	PESXX005	95%	94%	567	9.8	19 Lakhs
6.	PESXX006	84%	82%	1287	9.1	9 Lakhs
7.	PESXX007	89%	86%	1006	9.4	12 Lakhs
8.	PESXX008	86%	88%	1011	9.3	10 Lakhs
9.	PESXX009	79%	81%	1286	8.7	8 Lakhs
10.	PESXX010	92%	90%	822	9.5	15 Lakhs
11.	PESXX011	90%	91%	799	9.6	16 Lakhs
12.	PESXX012	80%	83%	1021	8.6	8 Lakhs

If an analysis is made by considering values belonging to only one column at a time, then it is called as Univariate Analysis

If an analysis is made by considering values belonging to two columns at a time, then it is called as Bivariate Analysis

If an analysis is made by considering values belonging to three or more columns at a time, then it is called as Multivariate Analysis

## Univariate Data

---

- ❖ This type of data consists of **only one variable**.
- ❖ The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.
- ❖ The analysis of Univariate data can be done using

### 1. Analytical Techniques :

- Central tendency measures (mean, median and mode)
- Dispersion or Spread of data (range, minimum, maximum quartiles, variance and standard deviation)
- Frequency distribution tables

### 2. Visualization techniques :

- Histograms
- Pie Charts
- Frequency Polygon
- Bar Charts.

## Bi - Variate Data

---

- ❖ This type of data involves **two different variables** (one of these variables is independent while the other is dependent)
- ❖ The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
- ❖ Bivariate data analysis involves comparisons, relationships, causes and explanations.
- ❖ The analysis of Bivariate data can be done using
  1. Analytical Technique :
    - Correlation Co-efficient
    - Regression Analysis
  2. Visualization Technique :
    - Scatter Plot

## Multivariate Data

---

- ❖ When the data involves **three or more variables**, it is categorized under multivariate.
- ❖ Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.
- ❖ It is similar to bivariate but contains more than one independent variable.
- ❖ Techniques used for analysis : The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).



## Bi - Variate Analysis

---

- ❖ Bivariate analysis means the analysis of bivariate data; used to find out if there is a relationship between two sets of values.
- ❖ It usually involves the variables  $X$  and  $Y$  and is represented as an ordered pair  $(X, Y)$ .
- ❖ The dependent variables represent the output or outcome whose variation is being studied. [most common symbol for the input is  $y$ ]
- ❖ The independent variables represent inputs or causes, i.e. potential reasons for variation. [most common symbol for the output is  $x$ ]
- ❖ Models test or explain the effects that the independent variables have on the dependent variables. [  $y = f(x)$  ] .

## Alternate terminology for Independent / Dependent variables

### Independent Variable

Input  
variable

Predictor  
variable

Controlled  
variable

Explanatory  
variable

Regressor

Manipulated  
variable

### Dependent Variable

Output/  
Response  
variable

Predicted  
variable

Measured  
variable

Explained  
variable

Regresand

Experimental  
variable

## Identify Dependent and Independent variables

---



- ❖ Relationship between caloric intake and weight.
- ❖ Effect of temperature on pigmentation.
- ❖ Effect of fertilizer on plant growth.

## Solution : Dependent and Independent variables

---

SI No.	Independent Variable (X)	Dependent Variable (Y)
1.	Caloric intake	Weight
2.	Temperature	Pigmentation
3.	Amount of fertilizer used	Growth in height or mass of the plant

### *Bivariate Analysis*

*Visualization  
Technique*

*Analytical Technique*

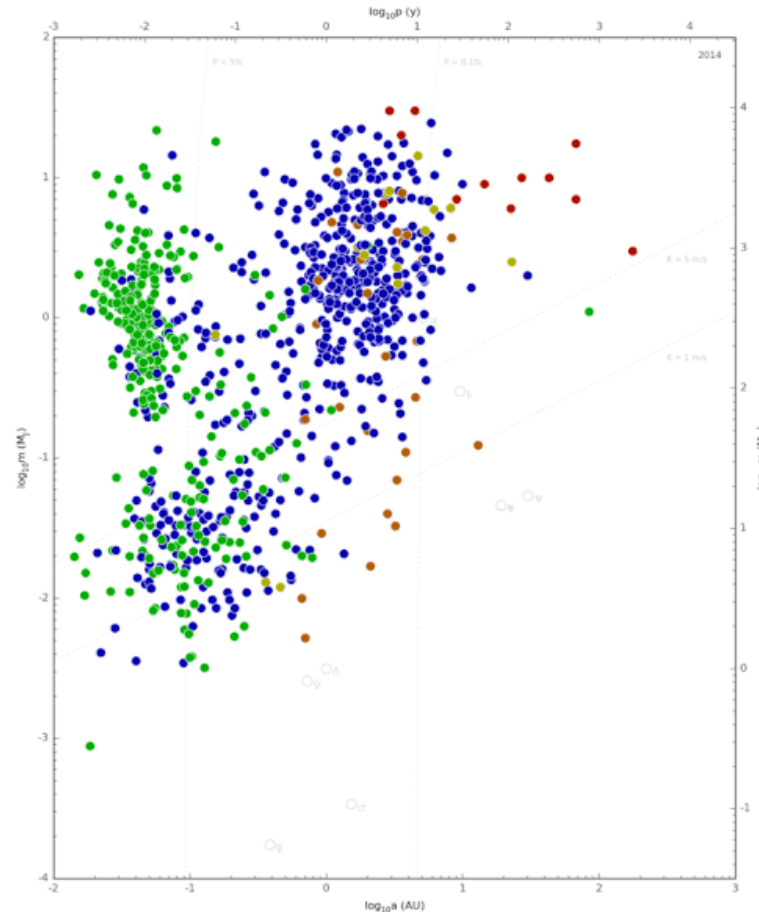
*Scatter Plots*

*Correlation  
Co-efficient*

*Regression  
Analysis*

## Scatter Plots

- ❖ The Scatter Plot is a mathematical diagram that plots pairs of data on an X-Y graph in order to reveal the relationship between the data sets.
- ❖ [Scatter plots](#) give you a visual idea of the pattern that your variables follow.
- ❖ Scatterplots can show you visually the strength of the relationship between the variables, the direction of the relationship between the variables and whether any outliers exist.



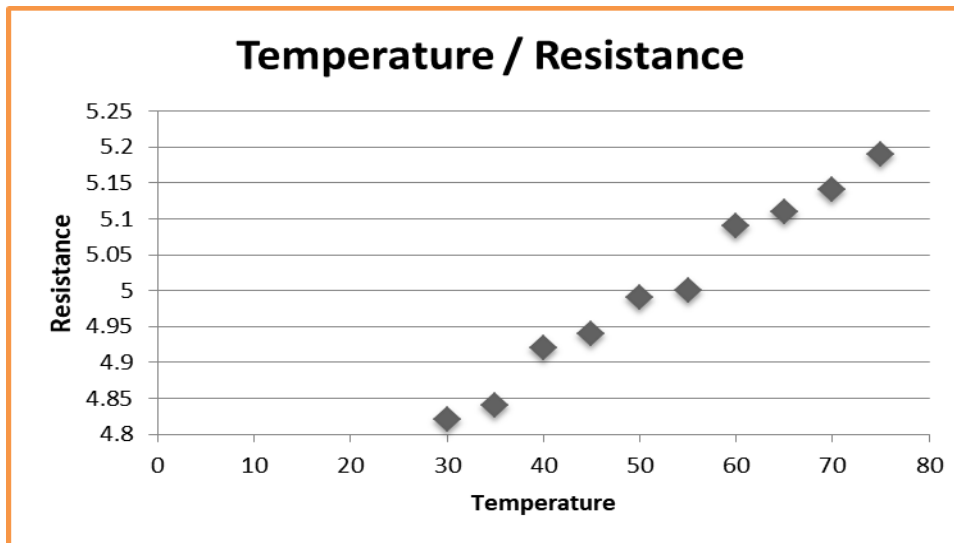
## Example for Physics Lab :

❖ Variation of resistance with change in temperature of a Semiconductor/ Conductor.



Temperature	Resistance
55	5
45	4.94
35	4.84
65	5.11
75	5.19
70	5.14
60	5.09
50	4.99
40	4.92
30	4.82

Source :semiconductor.org



*The resistance decreases with increase in temperature in a Semiconductor whereas in a Conductor, the resistance increases with an increase in the temperature.*

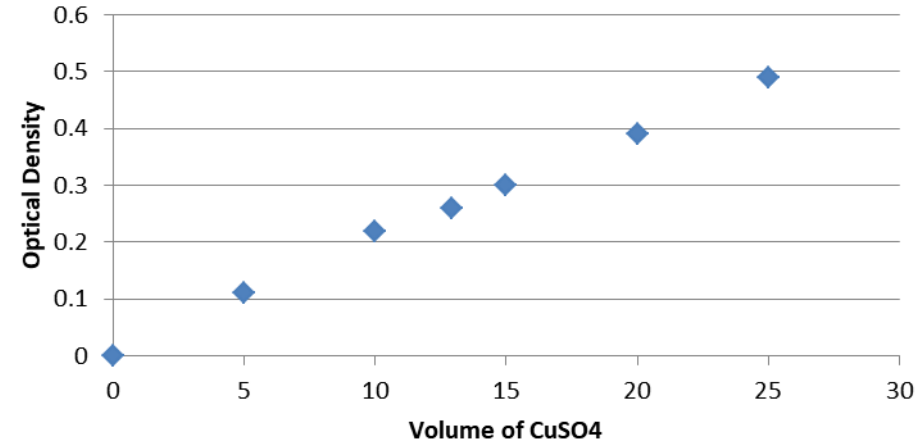
## Example from Chemistry Lab :

❖ Estimation of copper in copper sulphate solution by means of Calorimetry



Copper Sulphate	Optical Density
0	0
5	0.11
10	0.22
15	0.3
20	0.39
25	0.49
12.94	0.26

**Volume of CuSO<sub>4</sub>/Optical Density**



*The optical density increases with an increase in the value of CuSO<sub>4</sub>.*



## Remark :

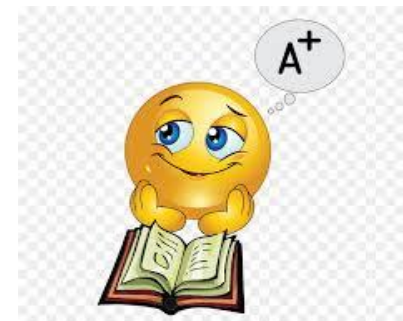
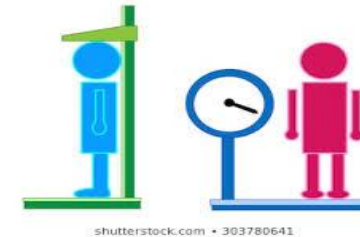
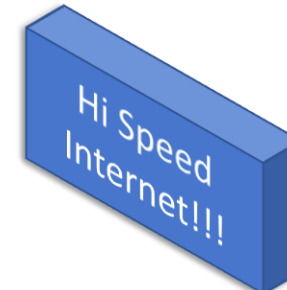
---

- ❖ Though Scatter plots display the strength and direction of the relationship between 2 variables, the measure of the strength of a relationship cannot be obtained from a Scatter plot.
- ❖ Hence we need to study Correlation.
- ❖ The Correlation coefficient enables us to obtain the measure of the strength of the relationship between the 2 variables.

# STATISTICS FOR DATA SCIENCE

## Correlation

- ❖ Does height have an impact on the performance of a player in a Basket ball match?
- ❖ Is there a relationship between internet bandwidth and time taken for data transfer?
- ❖ Are Height and Weight of an individual related?
- ❖ Does no. of hours effort have an impact on CGPA scored?

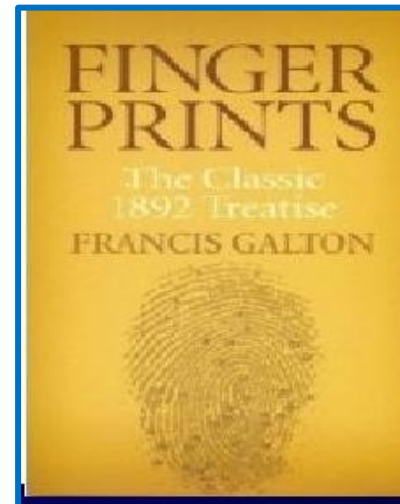
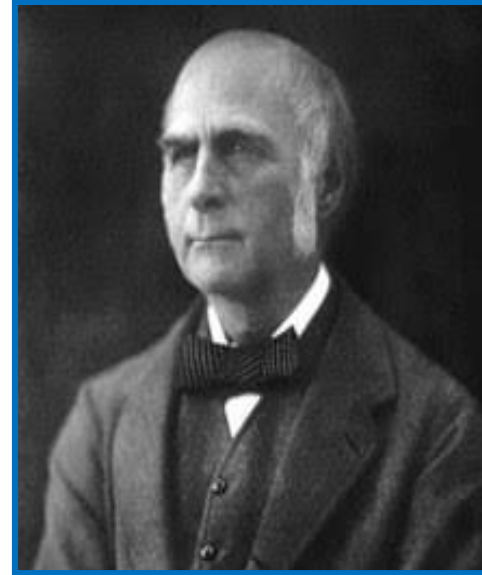


All these questions  
can be answered  
by the Correlation  
coefficient

## Brief history of Correlation

- ❖ Sir Francis Galton, (16 February 1822 – 17 January 1911).
- ❖ He was an English Victorian era statistician and a Fellow of the Royal Society.
- ❖ Galton produced over 340 papers and books.
- ❖ In 1892, he published the book “Finger Prints” and proposed the use of fingerprints as a means of personal identification.

Sources : [en.wikipedia.org](https://en.wikipedia.org), [amazon.in](https://amazon.in)



# STATISTICS FOR DATA SCIENCE

## *Galton's case study on relationship between heights & forearm length*

---



- ❖ He also created the statistical concept of [correlation](#) and widely promoted [regression toward the mean](#). He was the first to apply statistical methods to the study of human differences and [inheritance of intelligence](#), and introduced the use of [questionnaires](#) and [surveys](#) for collecting data on human communities.
- ❖ His primary example was the relationship between height and forearm length.

# STATISTICS FOR DATA SCIENCE

## *Galton's case study!!*



Is there any relation between  
the *height of an individual* and  
the *length of his forearm*???



*Sir Francis Galton introduced the concept of 'Correlation' in 1888 with a paper discussing how to measure the relationship between two variables.*

## Case Study : Galtons

---

- ❖ The data set that he considered consisted of the *heights and forearm lengths of 348 adult men*.

(He measured the distance from the elbow to the tip of the middle finger which is called as a cubit)

- ❖ Let the *height of the  $i$ th man* be  $= x_i$

- ❖ Let the *length of the forearm of the  $i$ th man* be  $= y_i$

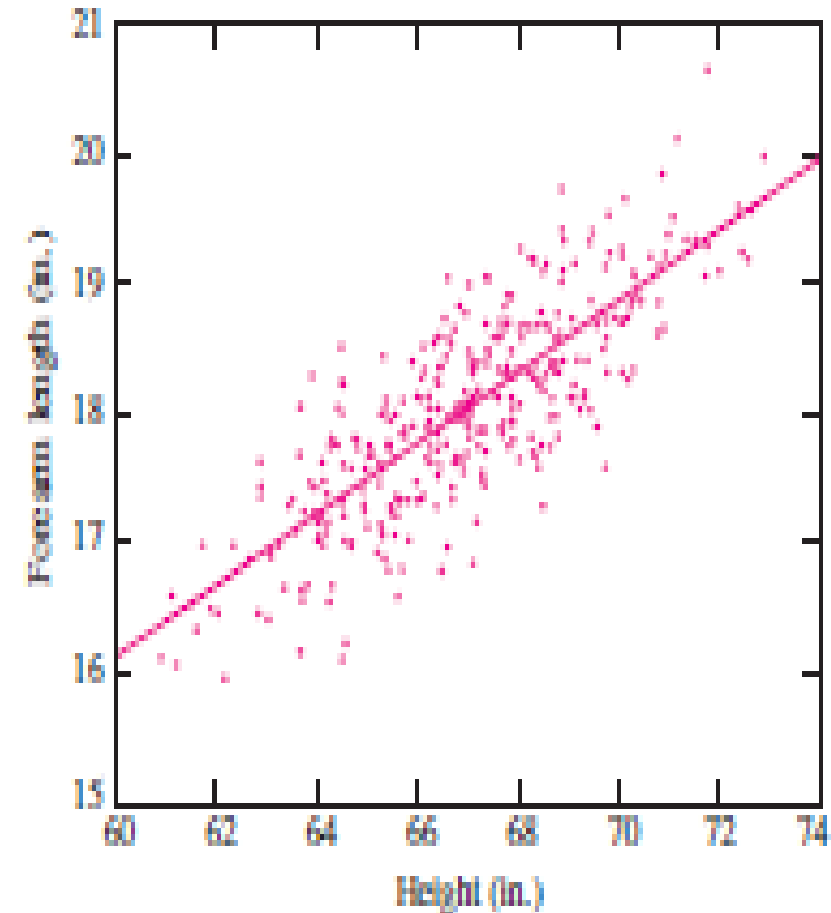
- ❖ Then Galton's data consists of *348 ordered pairs  $(x_i, y_i)$*

## Case Study : Galtons

Interpretation from the Scatter Plot :

- ❖ The points tend to slope upward and to the right which indicates that taller men tend to have longer forearms.
- ❖ In this the height and forearm lengths are said to be positively correlated.
- ❖ The line superimposed on the scatter plot is called as the Least Squares lines which will be discussed in one of the later sessions.
- ❖ The slope of the line is approximately constant throughout. This indicates that all the points are clustered around a straight line.

*Note : The degree to which the points cluster around the Least Squares line reflects the strength of the linear relationship between the two variables.*



## Correlation

- ❖ Correlation is a statistical measure of the strength of relation between two variables.
- ❖ Correlation also refers to the extent to which two variables have a linear relationship with each other (or related to each other).
- ❖ Correlations are useful because they can indicate a predictive relationship that can be exploited in practice.
- ❖ For Example :  
Meditation improves concentration levels.



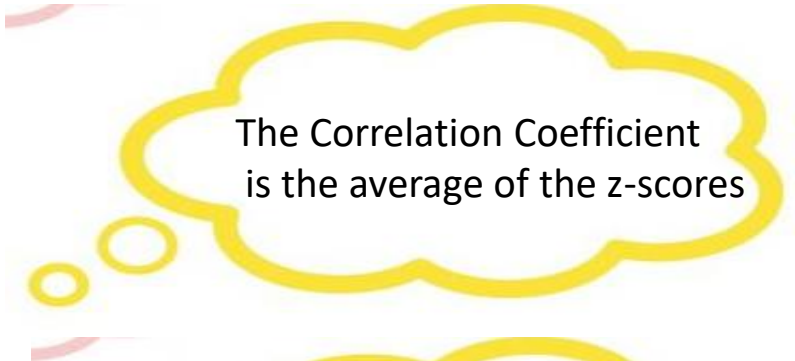


## Correlation Coefficient

- ❖ The Correlation Co-efficient is a numerical measure of the strength and direction of the linear relationship between two variables.
- ❖ Let  $(x_i, y_i)$  = ordered pairs that represent points on a scatter plot.
- ❖  $\bar{x}$  = mean of the 'x' values
- ❖  $\bar{y}$  = mean of the 'y' values
- ❖  $S_x$  = standard deviation of 'x' values
- ❖  $S_y$  = standard deviation of 'y' values
- ❖ Correlation Co-efficient is given by

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{S_x} \right) \left( \frac{y_i - \bar{y}}{S_y} \right)$$

$$\Rightarrow r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$



The Correlation Coefficient  
is the average of the z-scores



Pearson's Correlation Coefficient

### Example Problem :

The Psychological tests of intelligence and of engineering ability were applied to 10 students. Here is a record of ungrouped data showing intelligence ratio (I.R) and engineering ratio (E.R) calculate the co-efficient of correlation?

Student	A	B	C	D	E	F	G	H	I	J
I.R	105	104	102	101	100	99	98	96	93	92
E.R	101	103	100	98	95	96	104	92	97	94

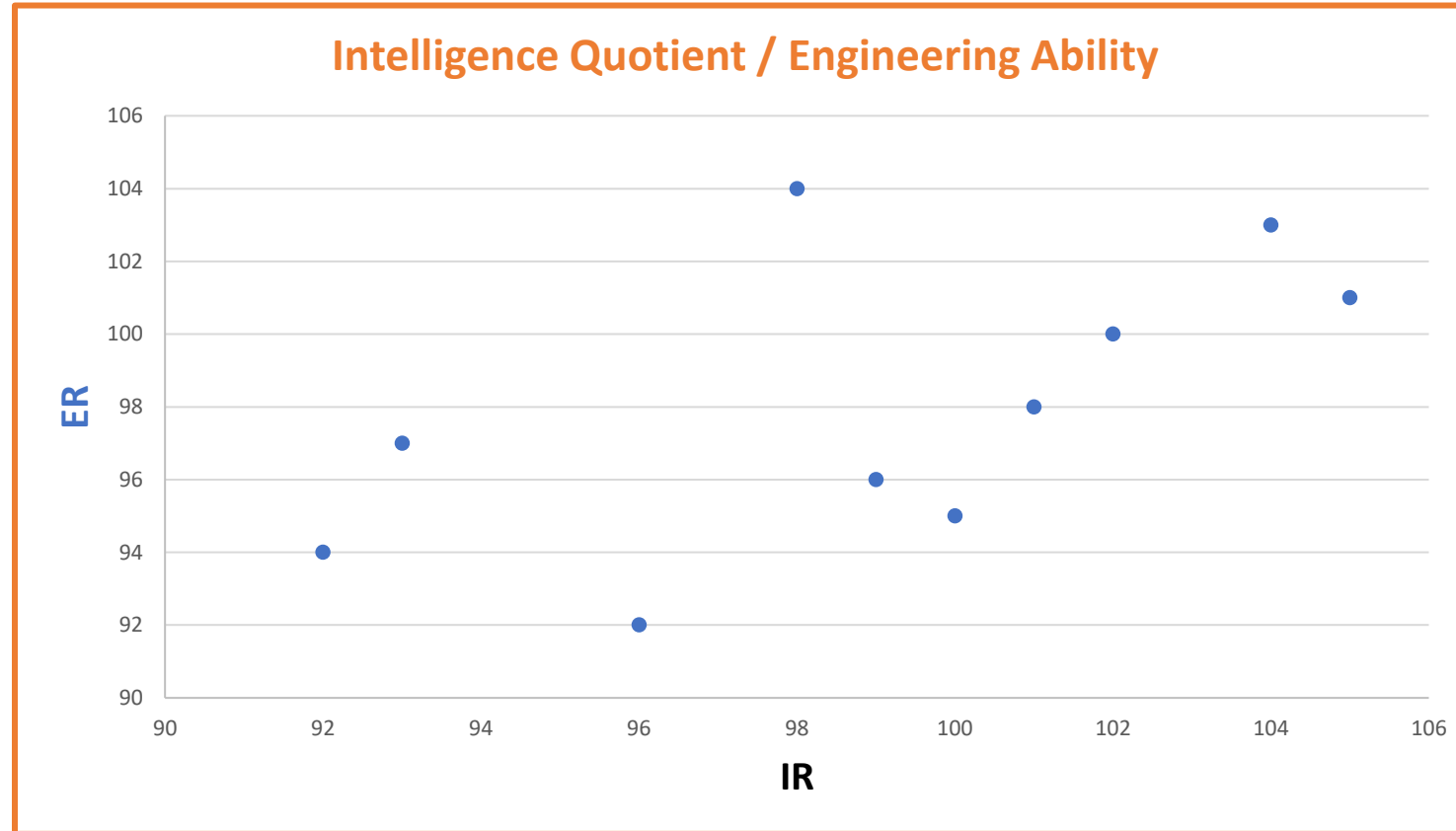
# STATISTICS FOR DATA SCIENCE

## Solution :

Students	IR ( $x$ )	ER ( $y$ )	$X = x - \bar{x}$	$Y = y - \bar{y}$	$X^2$	$Y^2$	$XY$
1	105	101	6	6	36	9	18
2	104	103	5	5	25	25	25
3	102	100	3	3	9	4	6
4	101	98	2	2	4	0	0
5	100	95	1	1	1	9	-3
6	99	96	0	0	0	4	0
7	98	104	-1	-1	1	36	6
8	96	92	-3	-3	9	36	18
9	93	97	-6	-6	36	1	6
10	92	94	-7	-7	49	16	28
	990	980	0	0	170	140	92

## Example

$$\begin{aligned} r &= \frac{\sum XY}{\sqrt{\sum X^2} \sqrt{\sum Y^2}} \\ &= \frac{92}{\sqrt{170 \times 140}} \\ &= 0.59 \end{aligned}$$



Interpretation : There is a moderate correlation between Intelligence Quotient and Engineering Ability.



# THANK YOU

---

**Dr. Karthiyayini**

Department of Science & Humanities

**Karthiyayini.roy@pes.edu**

+91 80 6618 6651