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

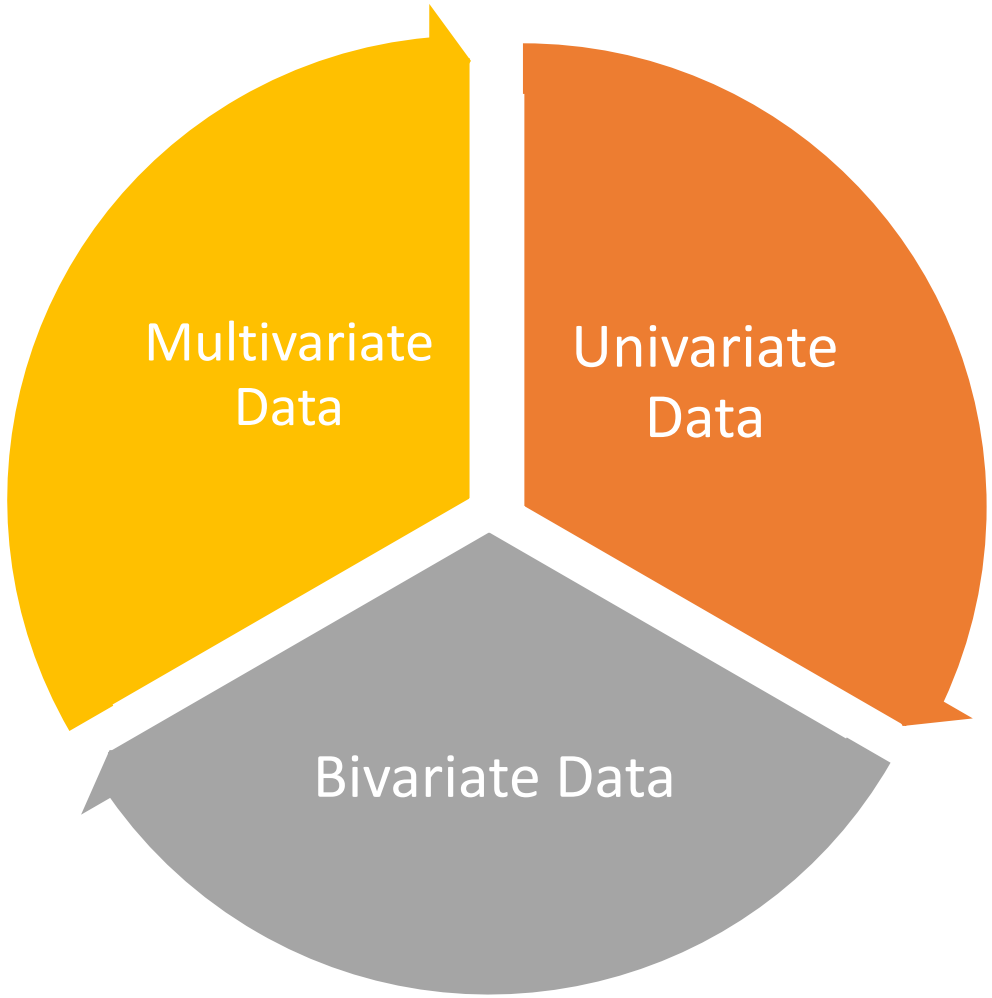
Dr. Karthiyayini
Department of Science & Humanities

STATISTICS FOR DATA SCIENCE

- Classification of Data
- What is Correlation ?
- Pearson’s Correlation Coefficient

STATISTICS FOR DATA SCIENCE

Types of Data



STATISTICS FOR DATA SCIENCE

Types of Data

Sl No.	SRN	10th Marks	12th 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

STATISTICS FOR DATA SCIENCE

Univariate Data

❖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

❖The analysis of Bivariate data can be done using :

1. Analytical Technique :
    - Correlation Co-efficient
    - Regression Analysis
2. Visualization Technique :
    - Scatter Plot

STATISTICS FOR DATA SCIENCE

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 $(X,\ Y)$.

- X represents the independent variable and Y represents the dependent variable.

Alternate terminology for Independent / Dependent variables

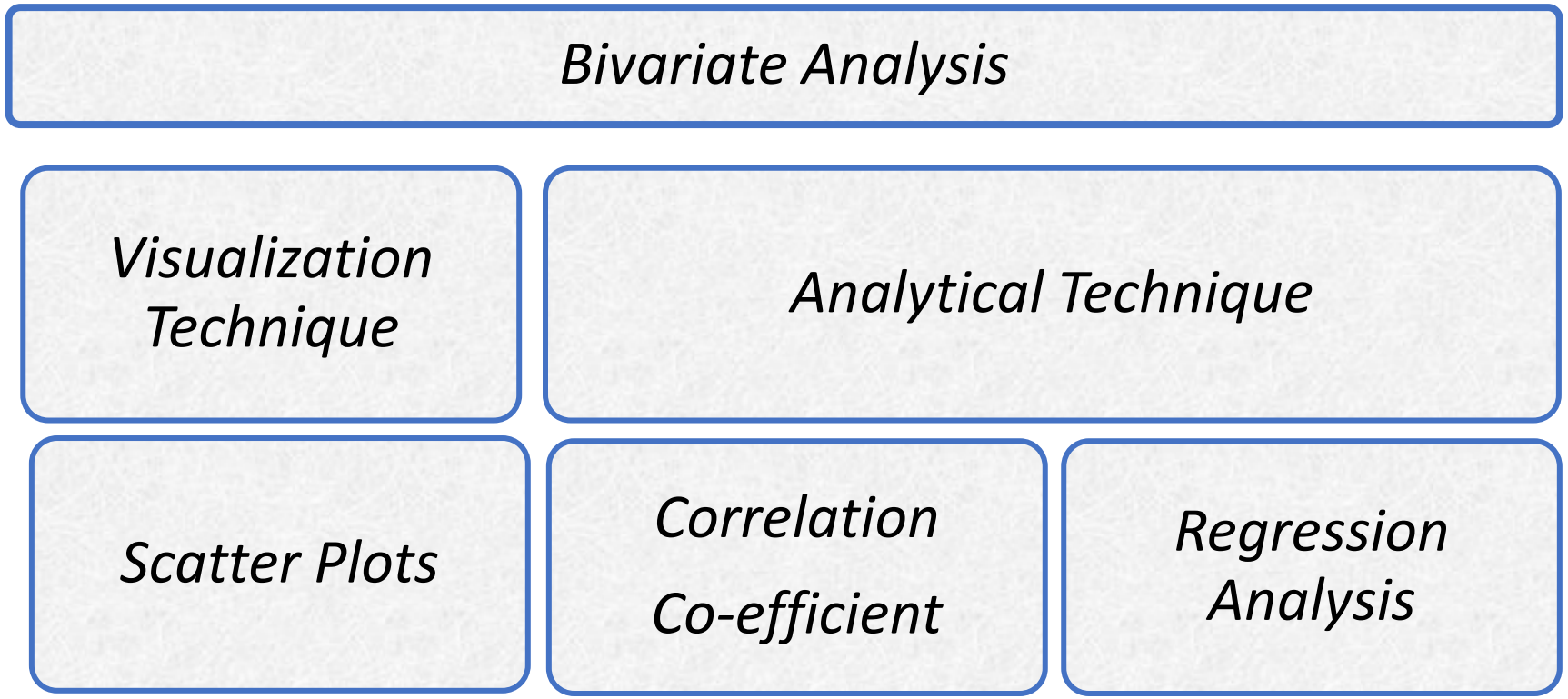
Independent Variable

- 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**

Bivariate Analysis

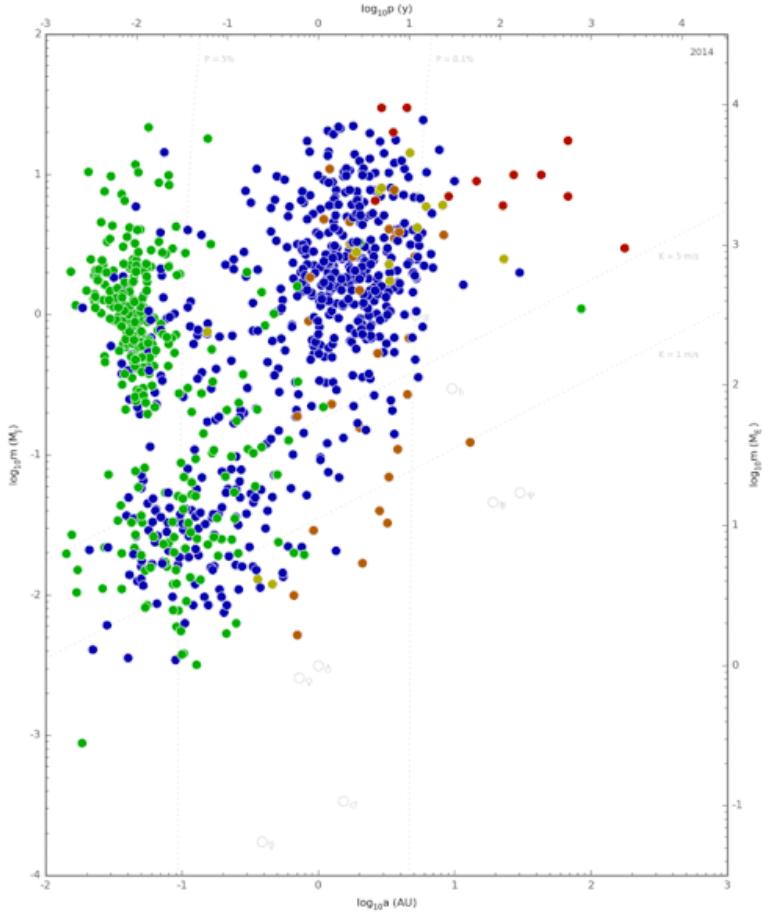


Scatter Plots

❖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.



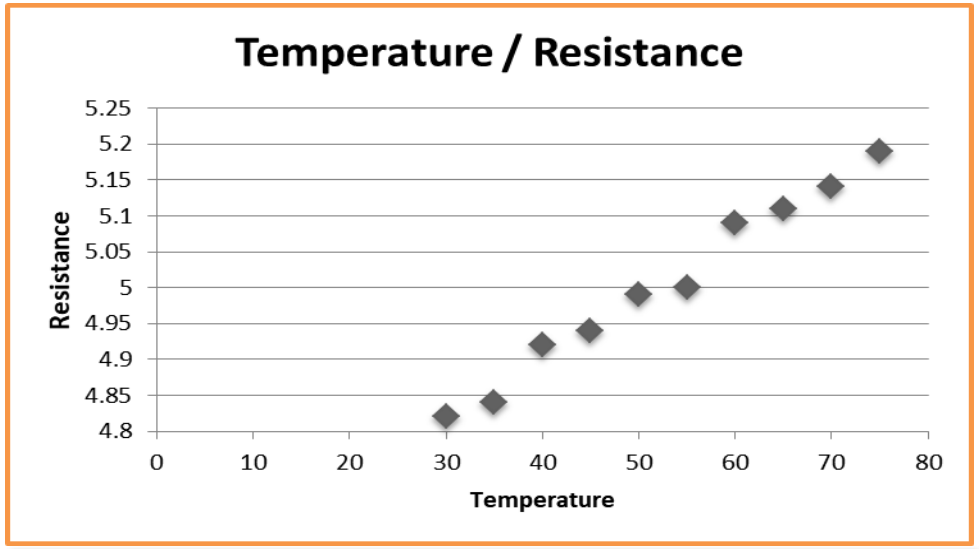
Source : commons.Wikimedia.org

STATISTICS FOR DATA SCIENCE

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



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

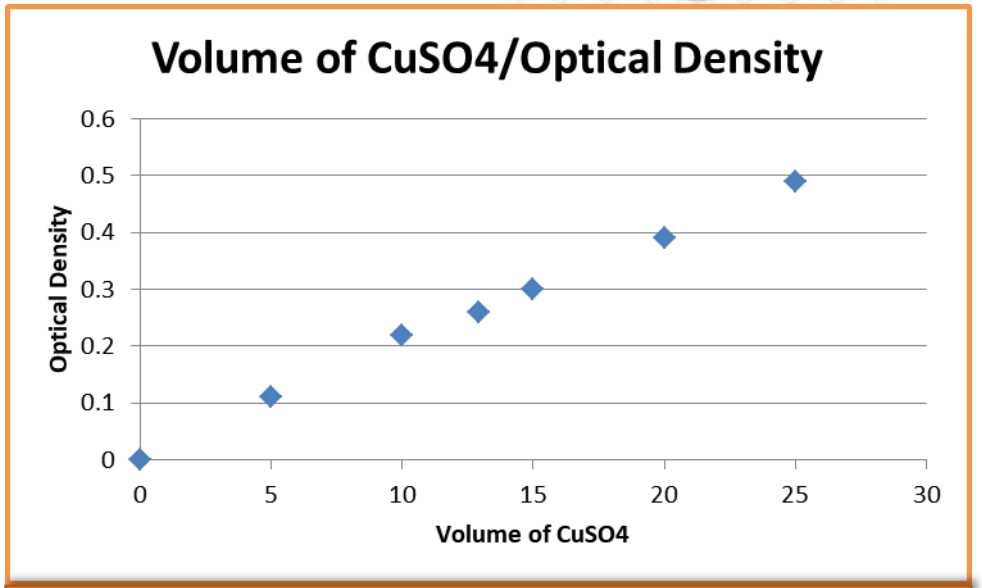
Source :semiconductor.org

STATISTICS FOR DATA SCIENCE

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



The optical density increases with an increase in the volue of CuSO4.

Source :iconscout.com

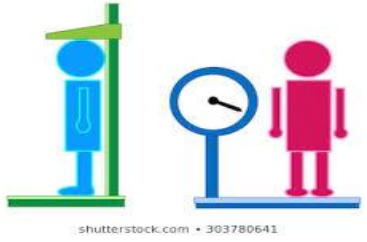
STATISTICS FOR DATA SCIENCE

Correlation

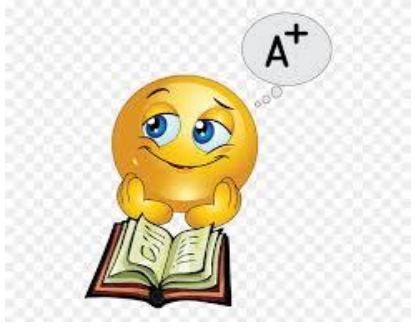
- Does height have an impact on the performance of a player in a Basket ball match?

- Is there a relationship between internet bandwidth and time taken for data transfer?

- Are Height and Weight of an individual related?



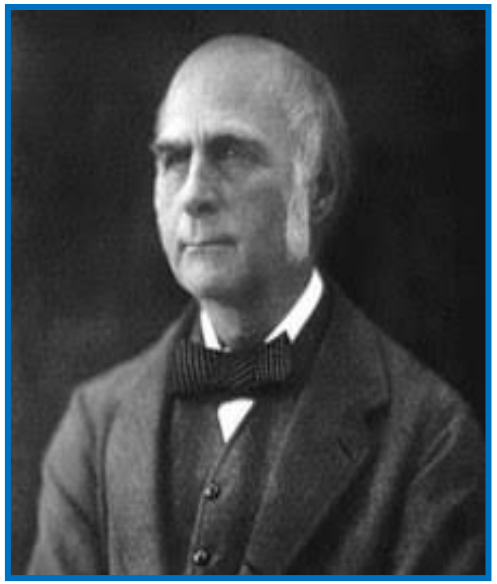
- Does no. of hours effort have an impact on CGPA scored?

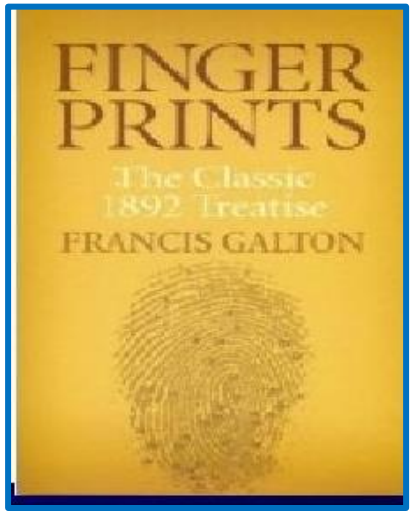


Sources :shutterstock.com, hdsmileys.com, internet sources

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, amazon.in

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.

Sources :aliexpress.com, freepik.com

Case Study : Galtons

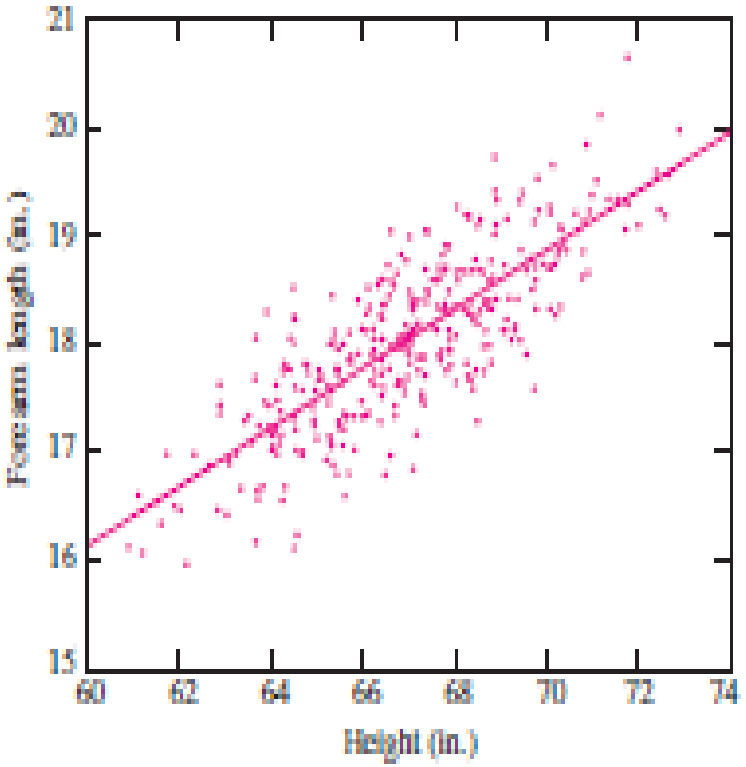
- The data set that he considered consisted of the *heights and forearm lengths of 348 adult men*.

  (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 ith man* be = x_i

- Let the *length of the forearm of the ith man* be = y_i

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



Source : Statistics for Engineers and Scientists, William Navidi

Correlation Coefficient

- 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

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

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

The Correlation Coefficient is the average of the product of 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 Correlation Coefficient?

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

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

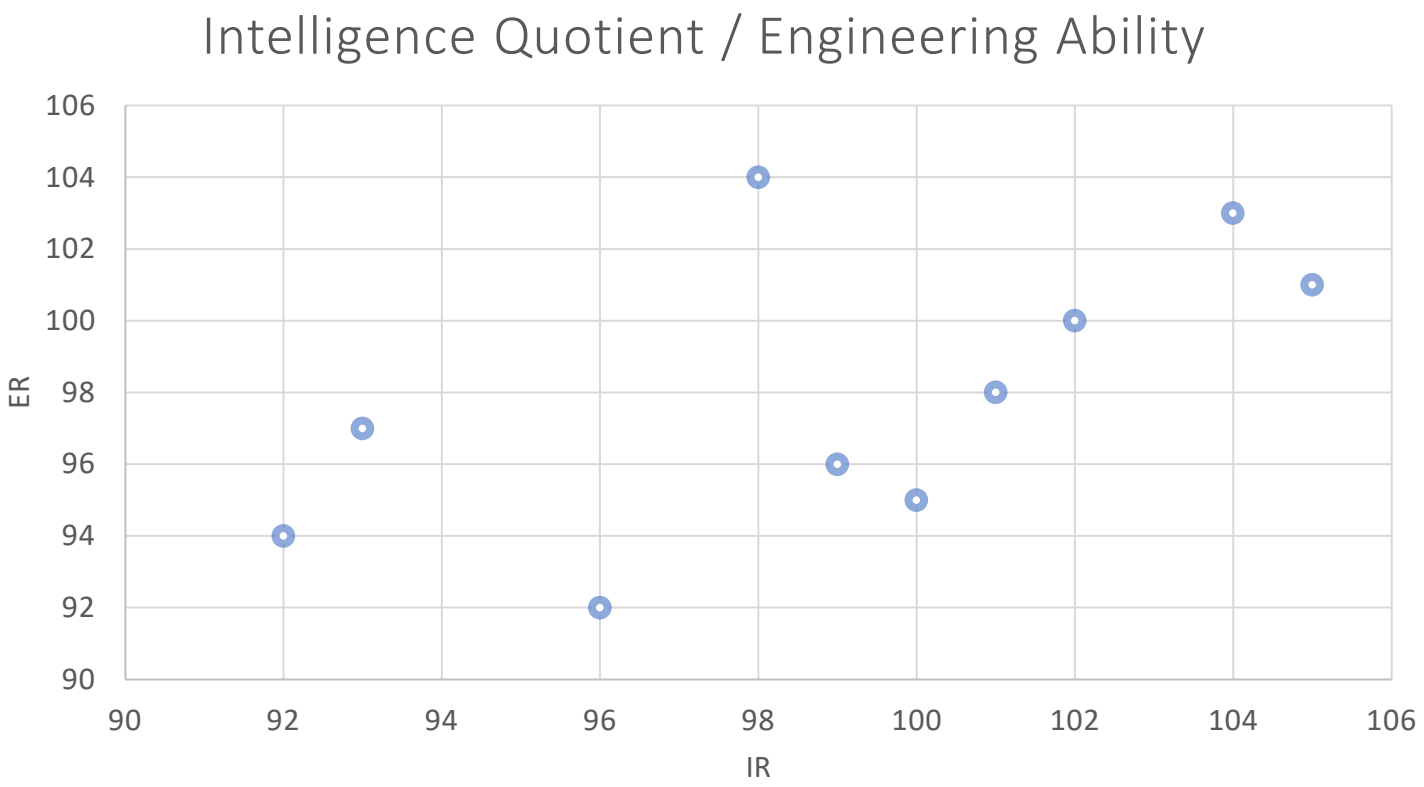
$$r = \frac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$$

$$= \frac{92}{\sqrt{170}\sqrt{140}}$$

$$= 0.59$$

STATISTICS FOR DATA SCIENCE

Scatter Plot



Correlation Co-efficient, $r = 0.59$

PES UNIVERSITY ONLINE

THANK YOU

Dr. Karthiyayini
Department of Science & Humanities

Karthiyayini.roy@pes.edu

+91 80 6618 6651