Correlation – relation between 2 variables(only).

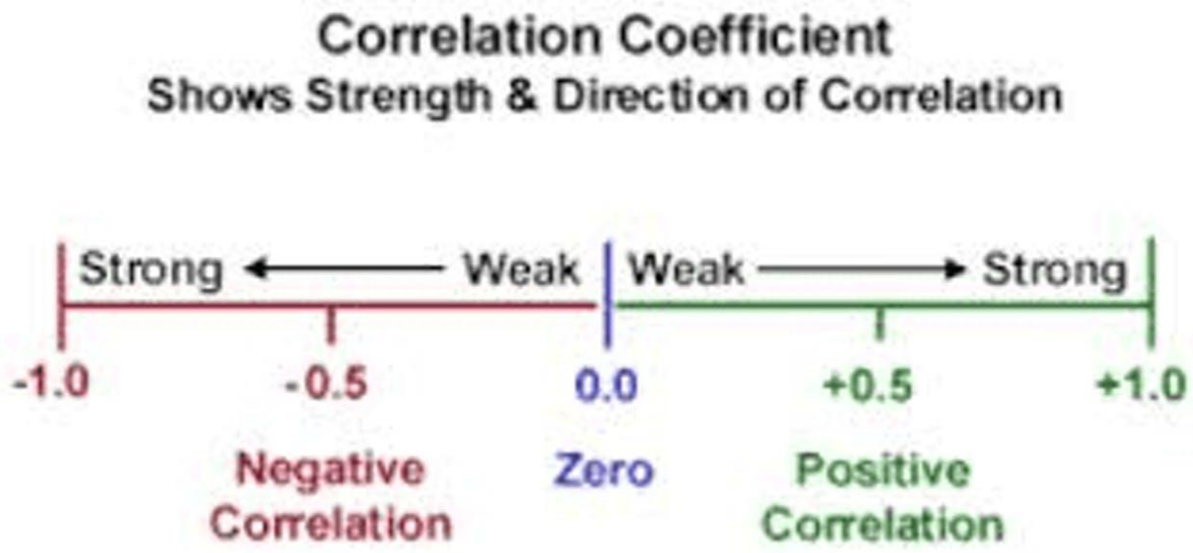
- It is determined using the correlation coefficient ‘r’ or ‘R’.(Pearson Correlation coefficient)
- R= Covariance/product of std deviations

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

- Covariance – If we are changing one variable, how much variation is there in the other variable.
- R always lies between -1 to +1.



- Positive Correlation: r : 0 to +1

The graph is directed upwards i.e it is directly proportional.

e.g – The number of marks obtained in the exam is directly proportional to the number of hours of studying. So, they are positively correlated.

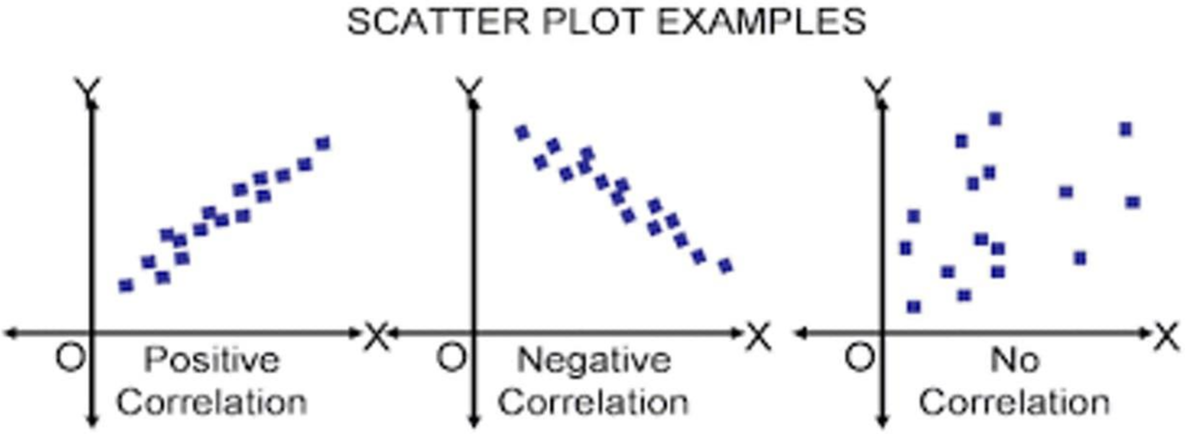
- Negative Correlation: r : -1 to 0

The graph is downwards and is inversely proportional.

e.g – Price of pen is inversely proportional to its sale.

- Zero Correlation :r=0

The graph is scattered and there is no relation between the 2 variables.



Coefficient of determination(R^2) –

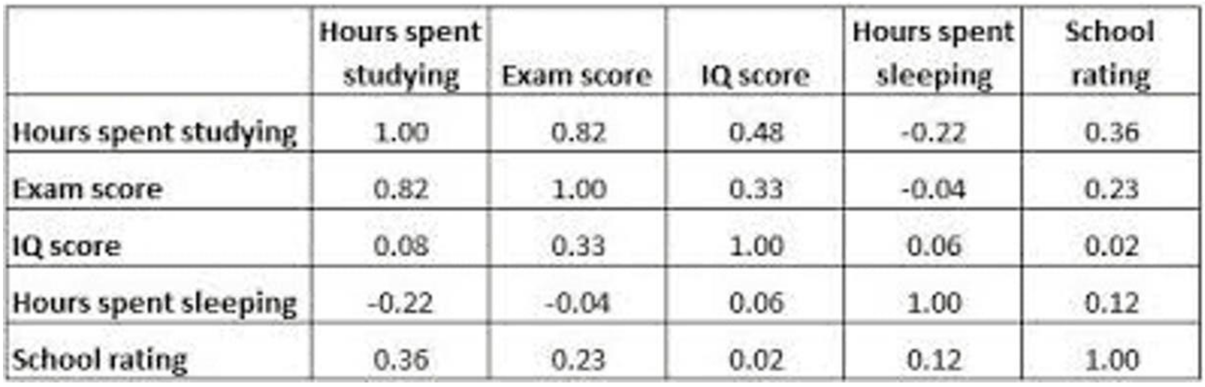
R^2 = Explained Variance/Total variance

e.g – If the correlation coefficient between attrition and age is -0.635.

Then, R^2 = 0.403. so, 40.3% of variance in attrition can be explained by variance in age and vice versa.

Correlation matrix – Finds correlation in multiple variable combinations.

- Same variables in rows and columns
- Diagonal entries are 1
- Only entries below diagonal are filled



	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00

Correlation Classification –

1. **Product Moment :**

- Both variables are continuous**.**
- **E.g –** Suppose a company is sending its employees to different places. And it is noting the **attitude** towards the place and **duration** of residence. It observes that it is positively correlated. If a person is sent to good places like Canada, he/she has a good attitude(high value) and tries to increase its duration of residence. But, if they are sent to some dangerous places like Nigeria, the attitude is bad(low value) and hence tries to lower the duration of residence and return back as soon as possible.
- **Decomposition of Total variation –**

When it is computed for a population rather than sample, the product-moment correlation is referred to as “r”.

Data we have is a sample of 12 employees- so, we have Sample mean, population mean and sample standard deviation.

1.H0: r = 0 Þ There is no significant correlation between attitude towards the place and duration of residence.

H1: r ≠ 0.

2. Two- tail, t-test is used as population std deviation is not known.

3.α =0.05 , d.f = rows-columns = 12 – 2=10

4.calculated t-value :

$$t = \sqrt{\frac{n-2}{1-r^2}} \cdot r$$

Suppose t calcualted comes out to be = 8.144.

5. t table val = 2.228

6. So, as t-calculated > t table : H0 is rejected. So, H1 is accepted. This implies that There is correlation between attitude towards the place and duration of residence.

2.Partial correlation :

- One variable is continous and one categorical.
- E,g – sales vs product category, age vs attrition rate.

3.Non-metric correlation :

- Both variable are categorical.
- E.g – gender and attrition rate.

Python Implementation –

- `from scipy.stats import pearsonr`

```
stats,p = pearsonr(dataset.A,dataset.B)
```

stats -> prints the r value from which we can decide whether it is positive, negative or neutral correlated.

p-> prints the p-value. So, if

p<0.05: reject H0

p>=0.05 : accept H0.

Here, H0: There is no significant relation between A and B.

- Correlation Matrix-

```
dataset.corr()
```

- If any categorical data entry is there, convert it into numeric-

  e.g – Attrition column conataining ‘Yes’ or ‘No’

```
dataset["Attrition"] = dataset["Attrition"].astype('category')
dataset["Attrition"] = dataset["Attrition"].cat.codes
```

This will convert attrition “No” to 0 and “Yes” to 1.

Then we can proceed with pearsonr.