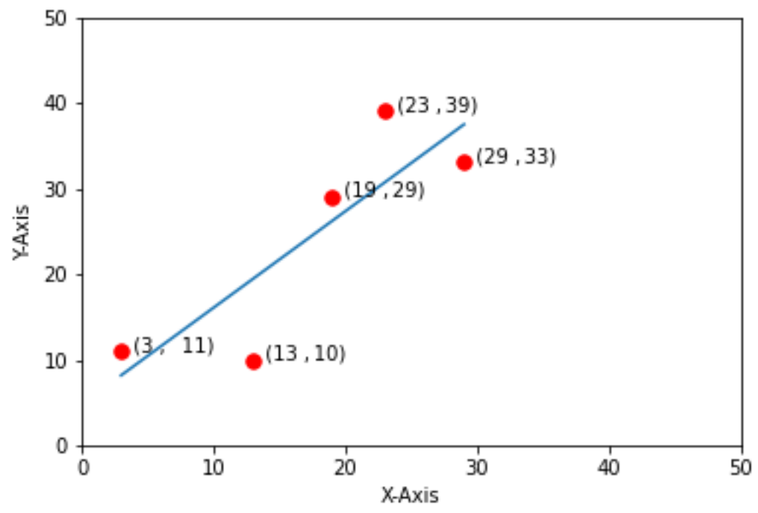
CORRELATION

Correlation is a statistical measure that expresses the extent to which two variables are linearly related

Lets talk about the relationship between data on X-axis and data on Y-axis and we can use st. line to represent the trend.

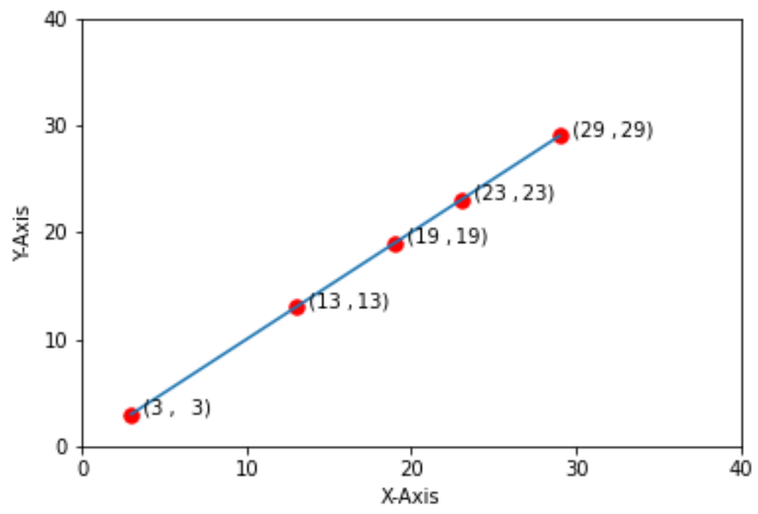


using this line we can predict Y value suppose we got a new value X = 20 then using this line we can predict Y would be somewhere around 27

similarly if we get the Y we can use the line and predict the X

the both cases we made a guess based on the tend

when the data is more closer to the line, we could say there is strong relationship between X and Y



The first plot states the WEAK RELATIONSHIP and the second plot states the STRONG RELATIONSHIP

We can quantify the strength of relationship with correlation value

- Data with weak relationship have small correlation value
- Data with moderate relationship have moderate correlation value
- Data with large relationship have large correlation value

MAX correlation = 1

and this will happens when s st.line with positive slope touches center of every data point like the above plot

this means when we get a point on X and we predicted Y is more accurate

NOTE: Correlation can be 1 irrespective of the slope

NOTE : Correlation can be 1 irrespective of the amount of data we have

Which means if we have only 2 data then we plot them and fit the st.line and then correaltion =1 which make the relationship stronger however we should not have any confidence prediction made using this trend line

reason, we have a very less data

why should we have less confidence in corelation made on less data?

to explain this lets take 2 data point and draw a st.line through it infact we can draw st.line through any two points

Now instead of 2 data point lets say we have 3 and all 3 points lie on same line with correlation=1

however we could be more confidence in pediction made from this line

in general the more data we have the more confidence in prediction made from the line

For a correlation **P-Value** tells us the probability that randomly drawn points will result in similar strong relation or more

The less the **P-Value** the high the confidence in prediction made using the line

- Maximum correlation occurs when st.line passes through center of the point
- Confidence depends on how much data we have

i.e, we must have less confidence on the plot with 2 data point

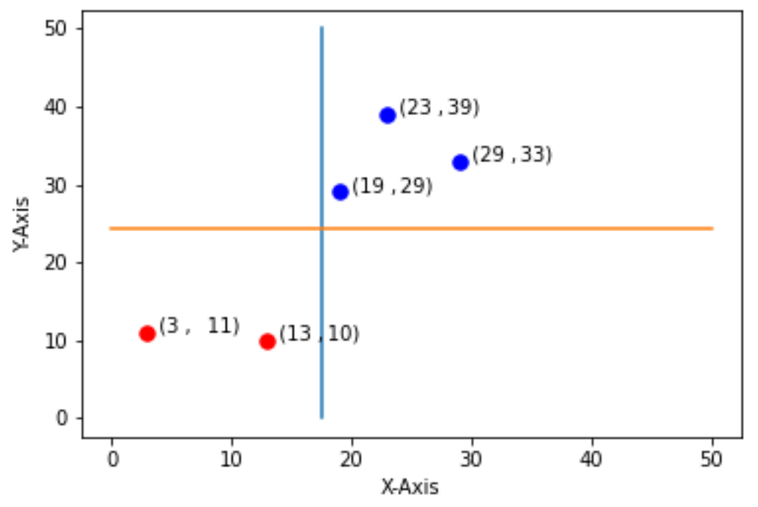
Correlation = -1 for negative slope and confidence depends on high P-Value

some case st.line will not fit to each and every data points then correlation will be near to 0 (0.88 , 0.99,... kind off)

worst cast correlation = 0 where there is no relationship between X and Y

$$Correlation = \frac{Covariance(X, Y)}{\sqrt{variance(x)}\sqrt{variance(y)}}$$

Lets take the same example from covarinace topic



Covariance = 116 ; Variance(x) = 101.8 ; Variance(y)=160.3

putting them into the formula we get

$$Correlation = \frac{116}{\sqrt{101.8}\sqrt{160.3}} = 0.9$$

and in this case P-Value = 0.03