Key Notes

By- Sabyasachi Parida

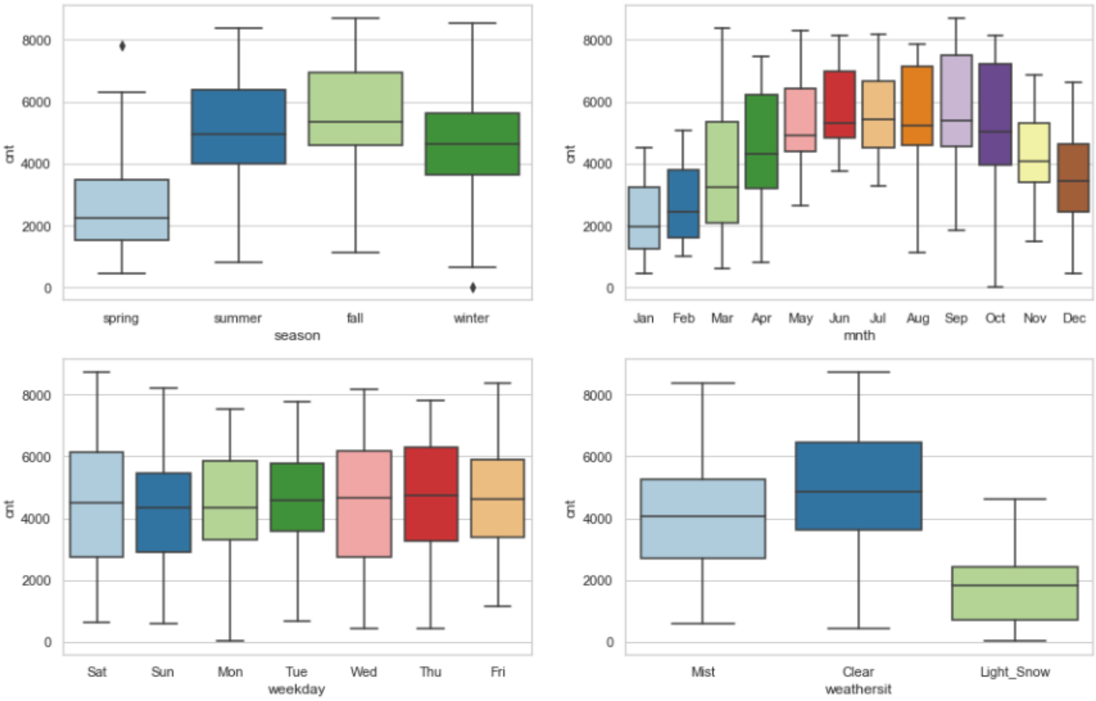
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: In the final model we have the following categorical variables with respective coefficients

CatVals	Coef
summer	781.5361
winter	1222.546
Sep	891.2903
Sat	548.0062
Light_Snow	-2119.89
Mist	-480.7882

From the coefficients we can infer that **summer**, **winter**, **sep** and **sat** have a positive impact on the cnt where as Light_snow and mist has a -ve impact on cnt.

During EDA we had the following observations



Observations

- The Median count is more in fall
- The median count is more in July and September
- Thursday and Saturday have more count
- The count is more in Clear Weathersit

Q2. Why is it important to use drop_first=True during dummy variable creation?

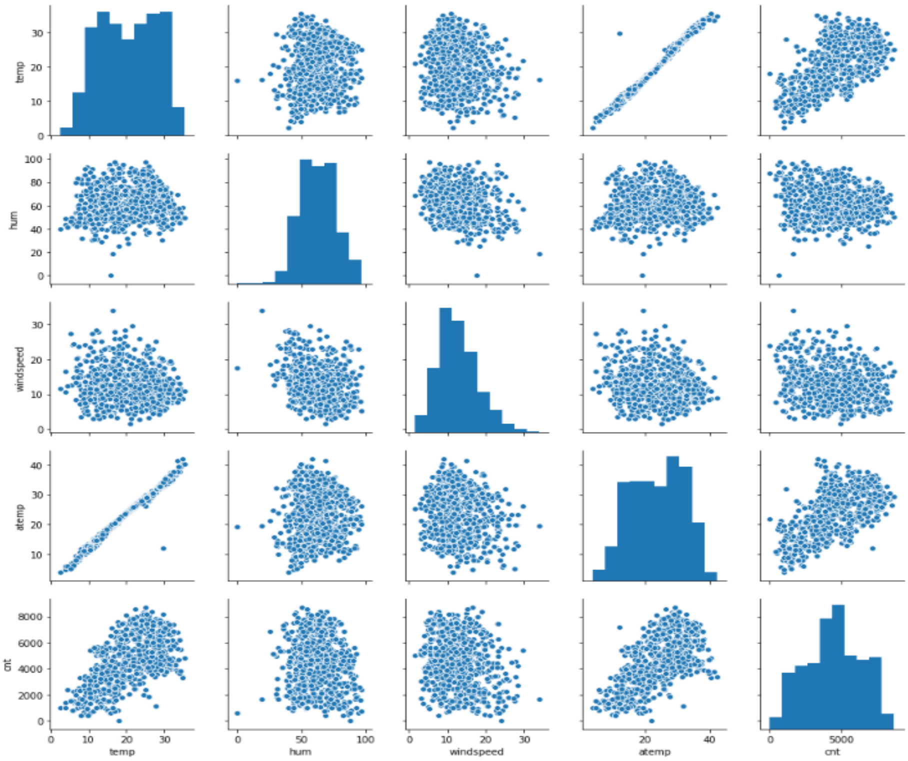
Ans: If we don't use **drop_first=True** we might get a redundant feature**.** In order to understand that better lets take an example. Let's assume we have a feature "Is*male".* If we use getdummies we will get output with 2 columns as follows

iS*male0*	**iS*male1***
0	1
1	0

But we can use only one column to Identify the value we do not need the value of both the columns. Suppose we take Ismale1 in that case if it is **1** person is a male if it is **O** person isn't a male.
So when we have categorical variables with n levels we can use drop_first and create values with n-1 levels which can be used and will prevent us from redundancy.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: Below is the pair plot

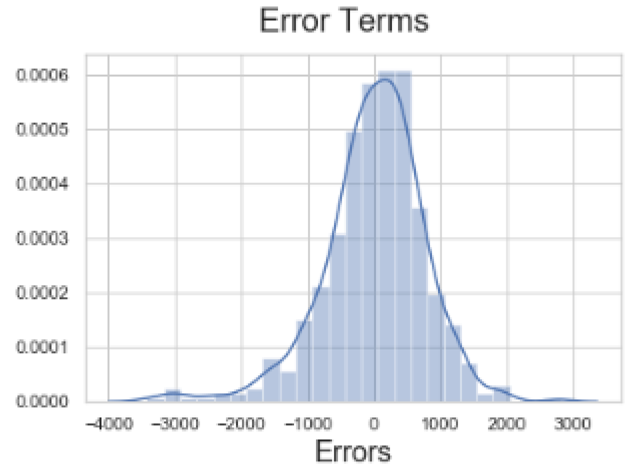


The highest correlation with the Target variable (cnt) is exhibited by **temp** and **atemp.**

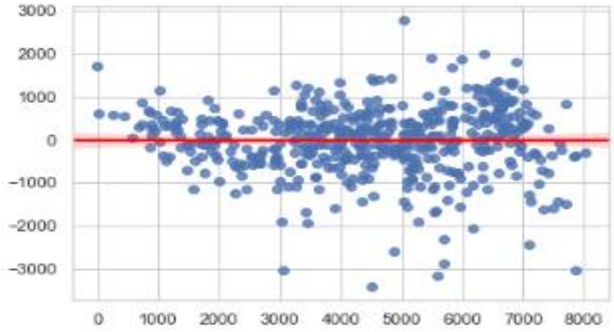
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: After building the model I have verified the below things

- The errors are distributed normally



- The errors are distributed independently and are not dependent on each other.
- The variance of error is uniform.



- The rsquared value is about 0.84
- The features have VIF less than 5 which tells that they don’t exhibit multicollinearity.
- They also have low P value which tells that the features are significant

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

CatVals	Coef
temp	4962.0848
2019	1994.2695
winter	1222.546
Sep	891.2903
summer	781.5361
Sat	548.0062
workingday	465.5211
Mist	-480.7882
hum	-1401.8639
windspeed	-1617.3921
Light_Snow	-2119.8904

Features which contribute positively are as follows

- Temp
- 2019 (yr)

- Winter

Features which affects significantly (positive or negative both)

- Temp
- Light_snow(weathersit)
- 2019(yr)
- Windspeed
- Hum

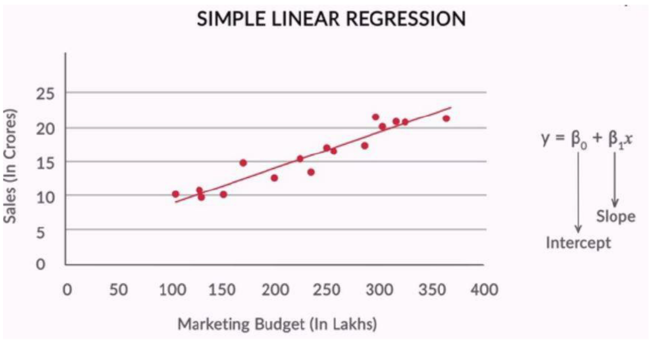
General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

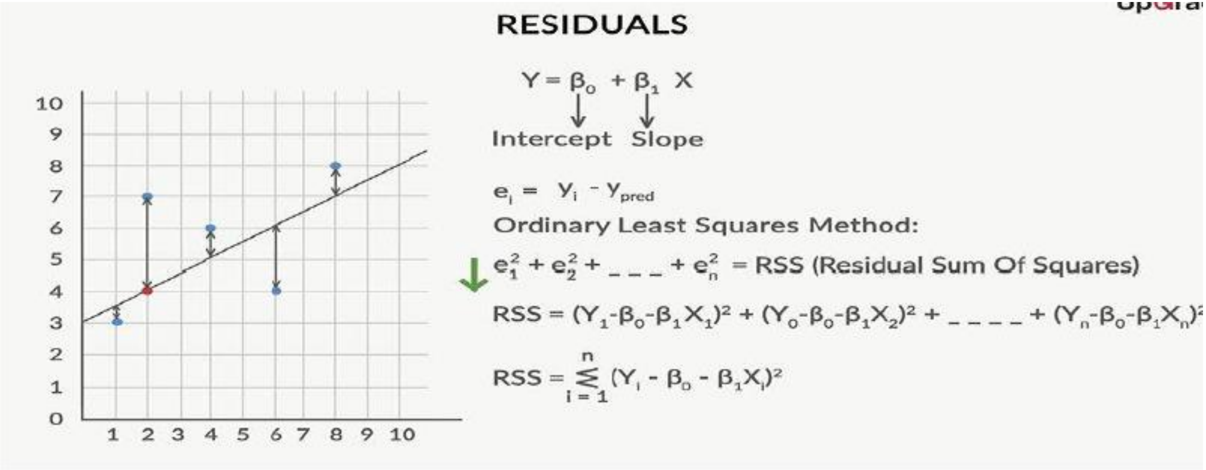
Ans: A simple linear regression model attempts to explain the relationship between a dependent variable and an independent one using a straight line.
The independent variable is also known as predictor variable and the dependent variable are also known as output variables. There are 2 types

- Simple linear regression
- Multiple linear regression

In simple linear regression the relationship is explained between a dependent and one independent variable using a straight line which is plotted on the scatterplot of these 2 pts.
Example:-



The best fit line is obtained by minimising the RSS. So, a best fit line is a line that fits the given scatterplot in the best way.



RESIDUALS

$Y = \beta_o + \beta_1 X$

Intercept Slope

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

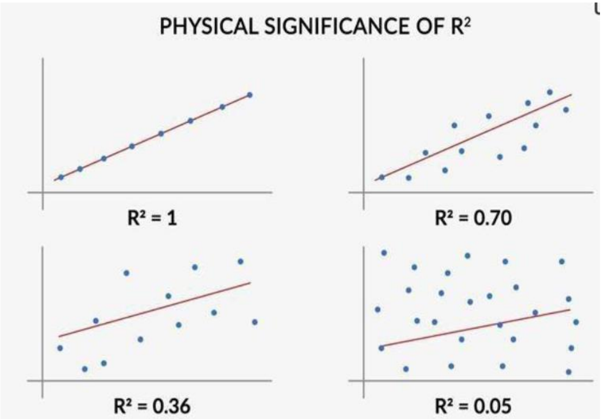
$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

$RSS = (Y_1 - \beta_o - \beta_1 X_1)^2 + (Y_o - \beta_o - \beta_1 X_2)^2 + \_\_\_\_ + (Y_n - \beta_o - \beta_1 X_n)^2$

$RSS = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_i)^2$

Rss for any regression line is the sum of the squared differences between the actual and predicted value(residuals).
Tss is Total sum of squares

The accuracy of a model is given by R2 statistics given by **R² = 1 - (RSS / TSS)**



Higher the R^2 better the model .

Gradient descent is an optimization algorithm that optimizes the objective function to reach the optimal solution.

While building the model we take the following assumptions

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance

In case of multiple regression, the resultant is a hyperplane.

Steps that we follow to while doing regression: -

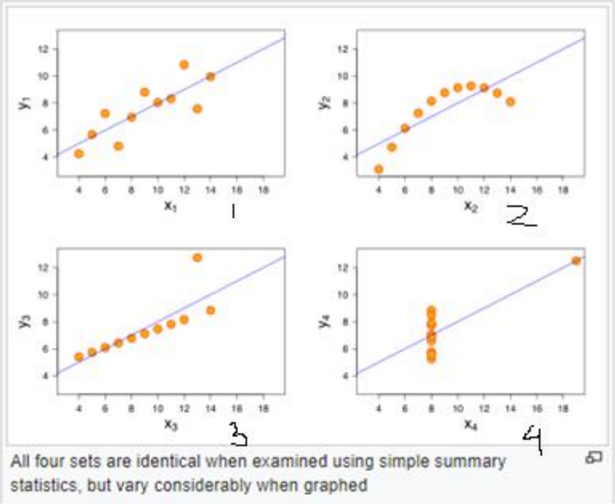
1. Reading & Understanding Data
   To understand the type of variables, number of nulls in rows and columns etc .
2. Data Preparation for EDA
   Prepare the data so that it can be visualized
3. Data Visualization & Exploratory Data Analysis (EDA)
   To find and visualise the outliers and get a sense of data i.e. what trend it follows
4. Data Preparation
   Add Dummy variables and perform scaling.
5. Data Modelling
   Build the regression model using Sklearn and stats model.
6. Residual Analysis.
   Perform the residual analysis, observe the errors and validate the assumptions
7. Predictions & Evaluation
   Predict the target variable using the trained model.
   If the result has r-squared value closer to the training r squared and all the features exhibit low VIF then we consider it as a good model.

Properties of regression line :-

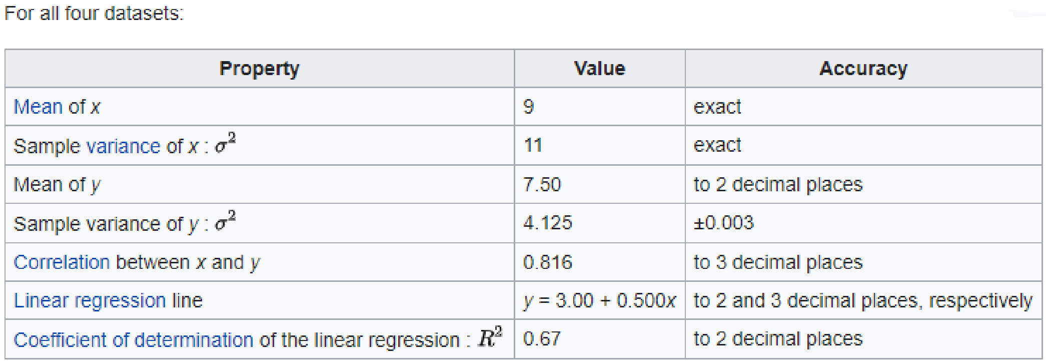
- Regression passes through the mean of independent variable (x) as well as mean of the dependent variable (y).
- Regression line minimizes the sum of “Square of Residuals”.
- B_i explains the change in Y with a change in x by one unit. In other words, if we increase the value of ‘x’ it will result in a change in value of Y.

Q2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe’s quartet has 4 data sets that have nearly identical descriptive statistics , but they have a very different distribution and appear very different when graphed .
Below is the representation



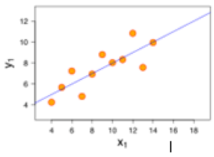
All four sets are identical when examined using simple summary statistics, but vary considerably when graphed



For all four datasets:

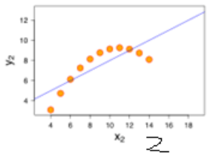
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : σ^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : σ^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

The 1. scatter plot



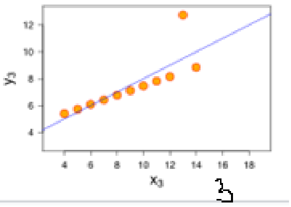
appears to be a simple linear relationship, corresponding to two variables correlated where **y** could be modelled as gaussian with mean linearly dependent on X.

The 2. Scatter plot



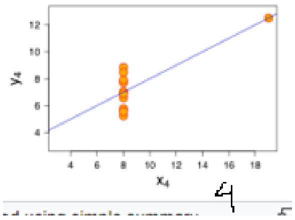
is not distributed normally, while a relationship between the two variables is obvious, it is not linear.

The 3· Scatter plot



the distribution is linear, but should have a different regression line. And it dot offset by the presence of outlier

The 4. Scatter plot



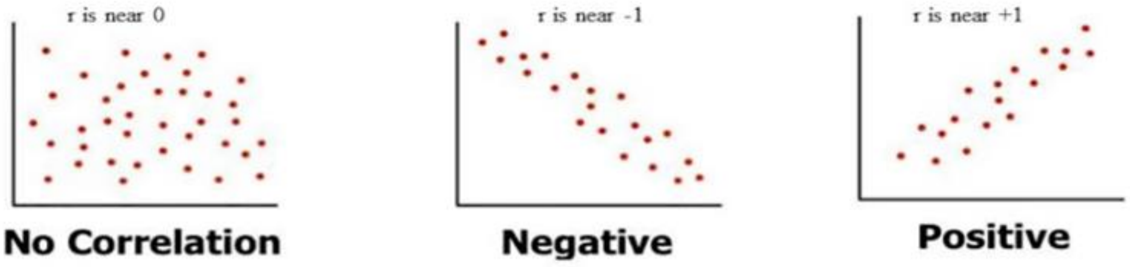
Is an example when one high leverage point is enugh to produce high correlation coefficient even though the other data points do not indicate any relationship between the variables.

The quartet illustrates the importance of looking at the set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistics for describing the realistic dataset.

Q3. What is Pearson’s R?

Ans: Mathematically correlation between 2 variables indicate how closely their relationship follows a straight line. **Pearson correlation coefficient** is also referred to as **Pearson's *r*.** which is given by the below formula

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



The value of which ranges between -1 to 1

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

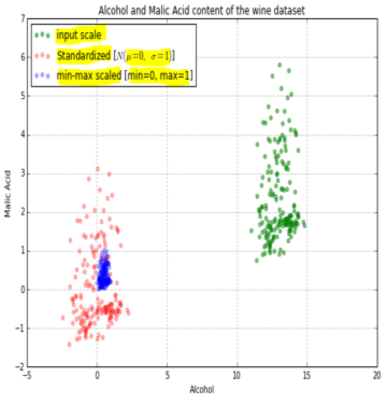
Ans: Scaling or Feature scaling is a method used to normalize/standardise the range of independent variables or feature of data. We do scaling for below reasons

- Ease of interpretation
- Faster convergence of gradient descent model.

Scaling just affects the coefficients and none of the other parameters such as t-statistics, f-statistics, P-value, R squared value etc.

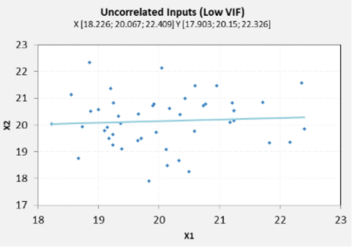
Standardisation brings all the data into a standard normal distribution with **Mean=0** and **standard deviation =1**
Formula = > X = (X-Mean(X))/sd(X)

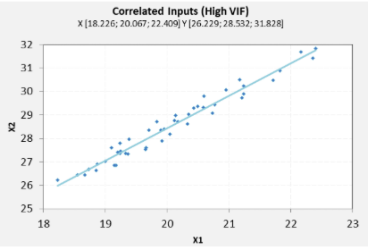
Min- max scaling or normalized scaling brings all the data in the **range of 0-1**
Formula = > X = (X-min (X))/(Max(X)-Min(X))
Below is the pictorial representation

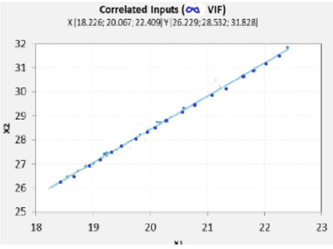


Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: VIF is denoted by the formulae => VIFi= 1/(1-Ri^2)
Hence, if there is a perfect correlation, then VIF= infinity. Which tells that there exists another X in the features which moves exactly as that of one of the other feature and this exhibit perfect multicollinearity.







Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

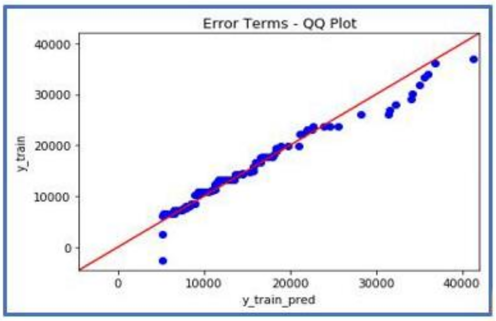
Ans: Q-Q plot or quantile-Quantile plot helps us assess if the set of data came from same distribution. This helps in a scenario of linear regression when we have train and test data set which we received separately , then we can go for qq plot to confirm if both the datasets are from population with same distribution or not.

It helps us in the following ways

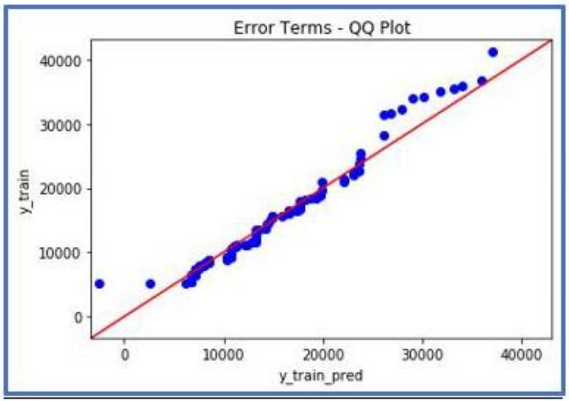
- Helps us identify if both data sets come from common distribution or not.
- Helps us identify if both the satasets have common location and scale
- Helps us check f both datasets have similar distribution shapes or not.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

- Similar distribution : if all points of quantiles lie on or close to straight line at an angle of 45 degrees from x axis.
- Y-values<X values: if y quantiles are lower than X quantiles



-
- If X-values < Y-values : if X-quantiles are lower than y-quantiles



-
- Different distribution : If all points lie away from the straight line which is drawn at 45 degrees from X axis