LINEAR REGRESSION ASSIGNMENT

Subjective Questions

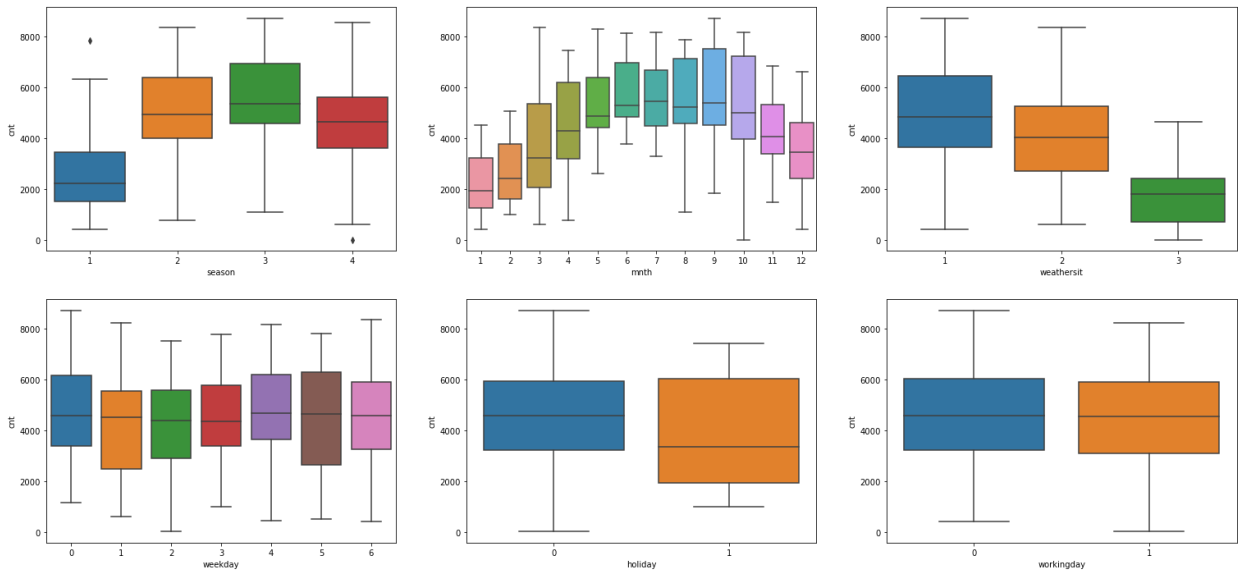
JUNE 9, 2021

Submitted by- Ravi Prakash Gupta

Assignment-based Subjective Questions

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

Ans: From the analysis of the cateforical variables from the dataset, we obtained the following boxplots.



The inference from the above plot is as follows:

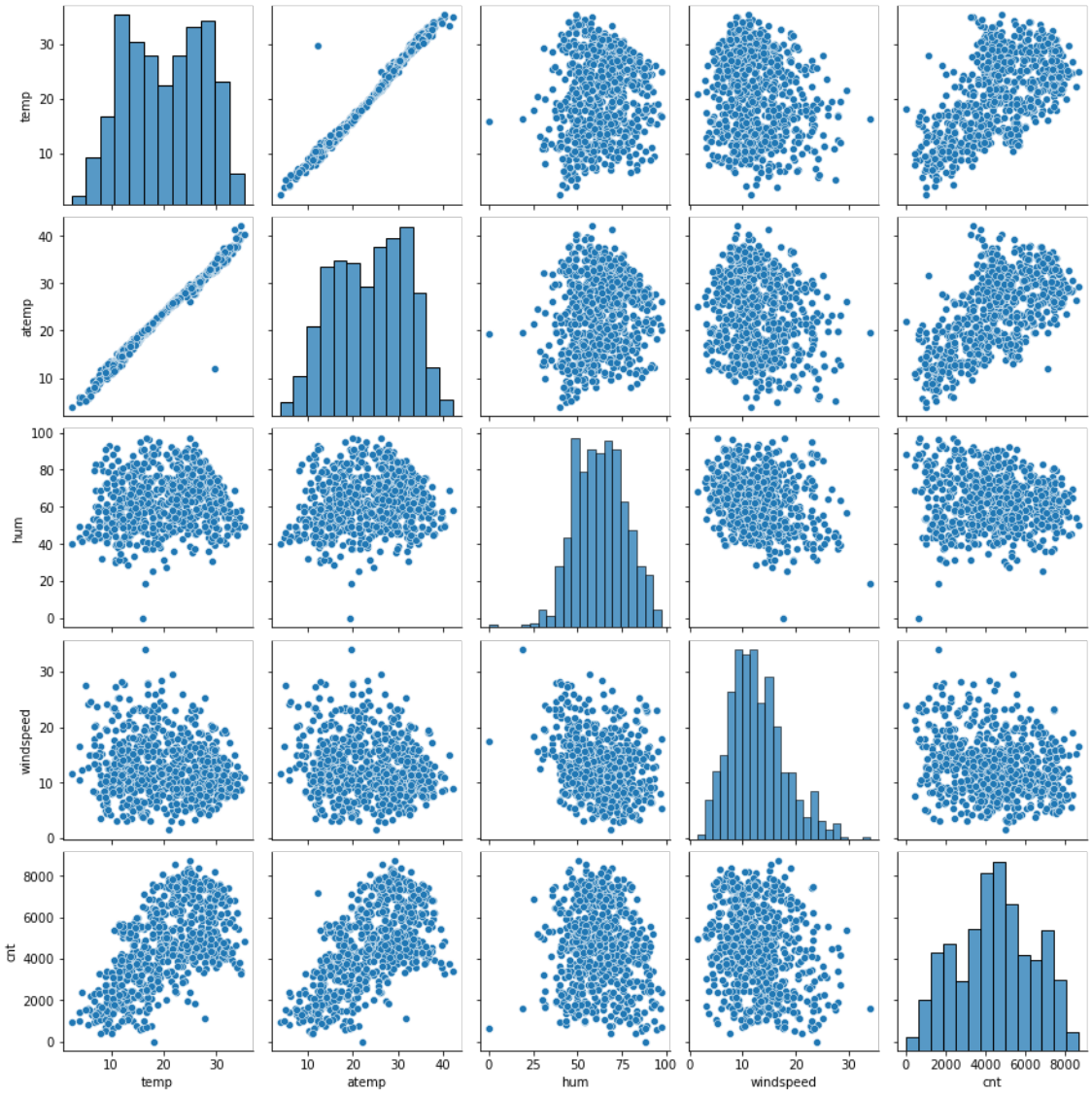
- In the Season boxplot, we can see highest bookings around 5000 in season 3.
- In the Month boxplot we can see the trend following, there is good demand which is over 4000 in May to Oct.
- In the weather boxplot, there is good demand in weather 1 which amounts to around 5000 bookings compared to others.
- In weekday boxplot, we can see that there seems no trend in the weekdays so we can exclude this variable for the prediction.
- In holiday boxplot, most of the bike booking were happening when it is not a holiday indicating holiday cannot be a good predictor for the dependent variable.
- In working day boxplot, we can see maximum bookings happening between 4000 and 6000.

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

Ans: It is important to drop first=True during dummy variable creation as it helps in reducing the extra column that is created while dummy creation. It reduces the correlation created among dummy variable, hence, it is important to remove them.

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

Ans: Plotting the pair-plot we obtain the following graph.

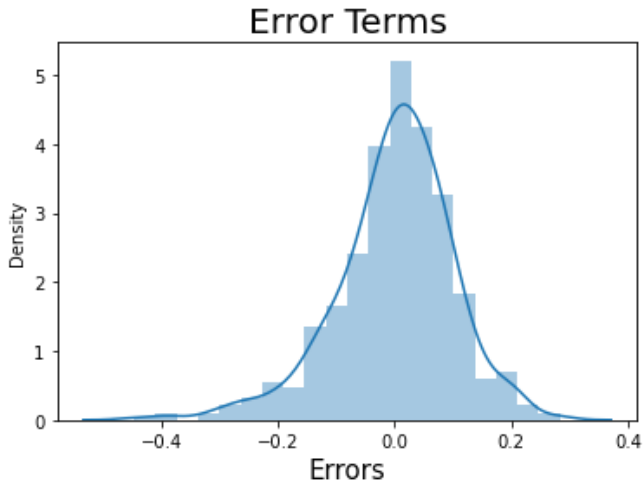


We can clearly see that the target variable i.e. cnt has a highest correlation with temp variables.

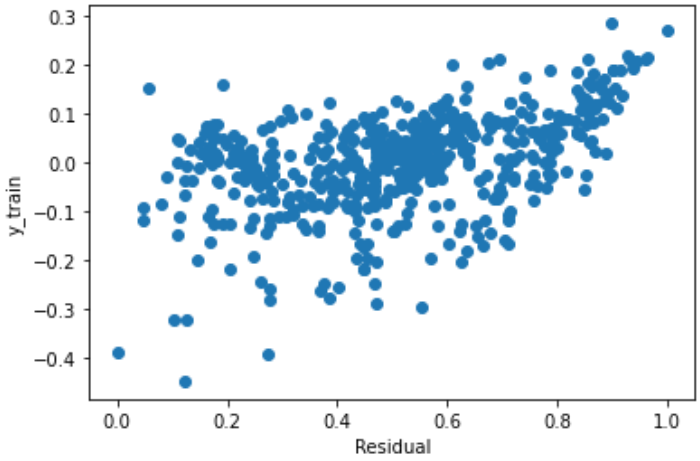
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: I validated the assumptions of Linear Regression after building the model on the training set as the training set satisfied all the following checking as below:

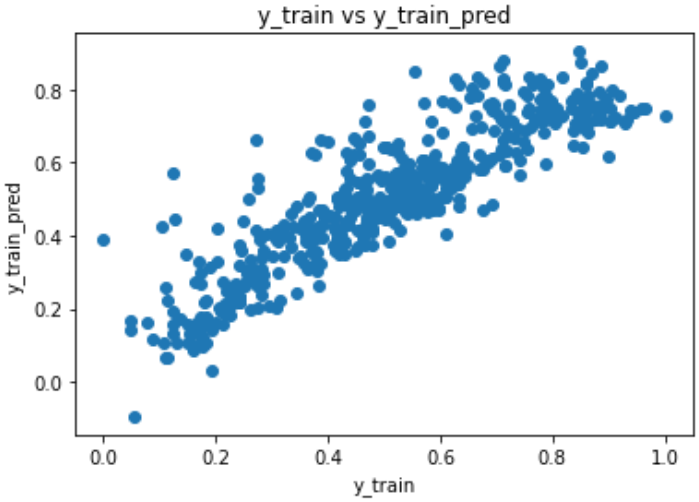
- **Residual Analysis-** Error terms are normally distributed



- **Linear Relationship** exist using pairplot.
- **Residual Pattern Check-** There is no any visible pattern between the residual and y_train



- **Homoscedasticity**- Variance of the residuals (error terms) is constant across the prediction.



- **Multicollinearity**- From the VIF calculation, we find that there is no multicollinearity existing

between the predictor variables, all the values are within the permissible range of below 5.

	Features	VIF
2	temp	3.38
3	windspeed	2.90
0	yr	1.89
4	season_2	1.58
5	season_4	1.33
6	mnth_9	1.19
7	weathersit_3	1.07
1	holiday	1.03

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

Ans: As per the Final Model, the top 3 Predictor Variables that are needed for the prediction purposes are:

- **YR -** Coefficient of yr indicates that a unit increase in yr variable, will increase bike hiring by 0.2234 values.
- **HOLIDAY -** Coefficient of holiday indicates that a unit increase in holiday variable, will decrease the bike hiring by 0.0653 values.
- **TEMP -** Coefficient of temp indicates that a unit increase in temp variable, will increase the bike hiring by 0.5403 values.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is the most basic supervised machine learning algorithm. Supervised algorithm is based on the labeled data that the algorithm uses. It uses the label to predict the outcome. Coming to regression, regression is the method of modeling a target/dependent variable based on independent variable. A regression problem is where the target variable is continuous value, for example salary, weight, etc. The regression model helps in predicting the cause and effect relationship between the variables.

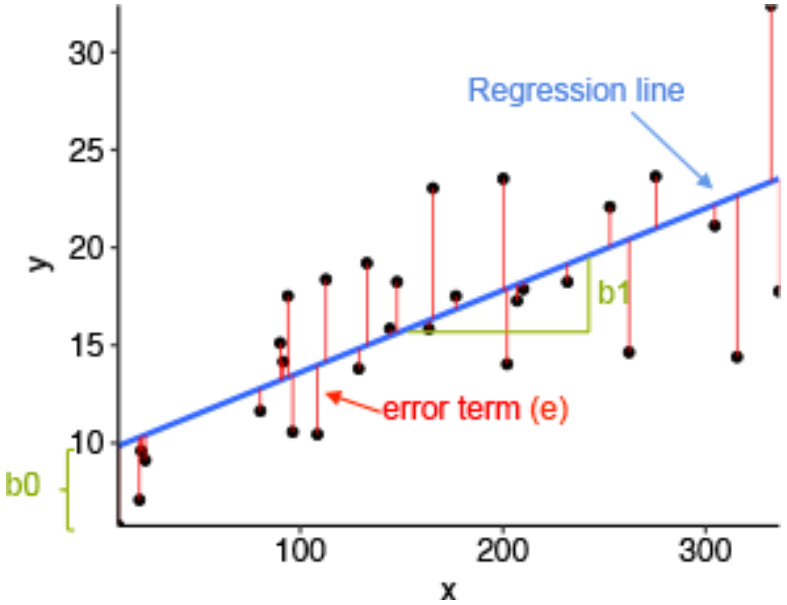
Simple linear regression is the type of regression analysis where there is only one independent variable and it has a linear relationship with the dependent variable.

The following are the basic assumptions of a linear regression model –

- A linear relationship exists between the independent variable (X) and dependent variable (Y).
- Little or no multicollinearity between the different features.
- Residuals should be normally distributed.
- Homoscedasticity of the errors.

The mathematical equation of linear regression can be written as **y = b0 + b1*x + e**, where –

- y is the predicted (dependent variable)
- x is the independent variable
- b0 is the intercept of the regression line i.e. the predicted value when x = 0
- b1 is the slope of the regression line
- e is the error term i.e. the part of y that can be explained by the regression model



The graph of linear regression above shows –

- the best-fit regression line is in blue
- the intercept (b0) and the slope (b1) are shown in green
- the error terms (e) are represented by vertical red lines

After the linear regression model is made, there are a few metrics we use to calculate error in the model.

- R-Square –

$$R^2 = (TSS - RSS)/TSS$$

TSS (Total sum of squares) – It tells how much variation there is in the dependent variable.

RSS (Residual Sum of Squares) – It is the sum of the squared differences between the actual y and the predicted y.

(TSS – RSS) – These measures the amount of variability in the response that is explained by performing the regression.

Properties of R^2:

1. R^2 will always range between 0 to 1.
2. If R^2 is 0, it means that there is no correlation between the dependent and the independent variable.
3. If R^2 is 1, it means the dependent variable can be predicted from the independent variable without any error.
4. R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable.

- Root Mean Square Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

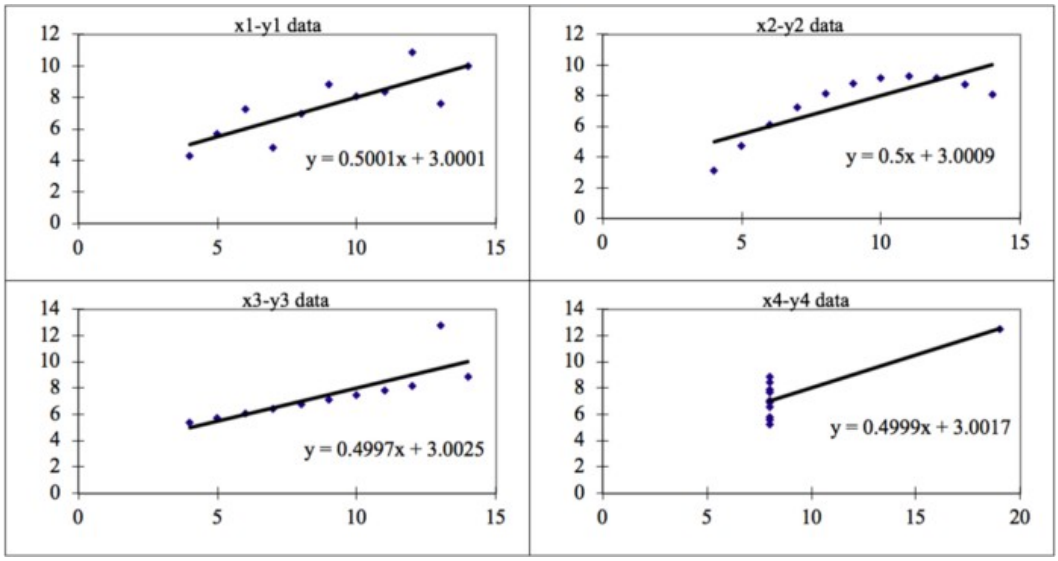
2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four datasets that have the same descriptive statistics and linear regression fit. The datasets are however very different from each other. These datasets have very similar statistical properties but are very different from each other when graphed.

These four data sets are as follows –

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



- Top Left – This first plot shows a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.
- Top Right – The second plot was not distributed normally, and the relationship between the two variables is not linear.
- Bottom Left – In the third plot, the relationship is linear between the two variables except for the one variable which seems to be an outlier. The calculated regression is offset by the one outlier that exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Bottom Right – In the fourth plot, one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship

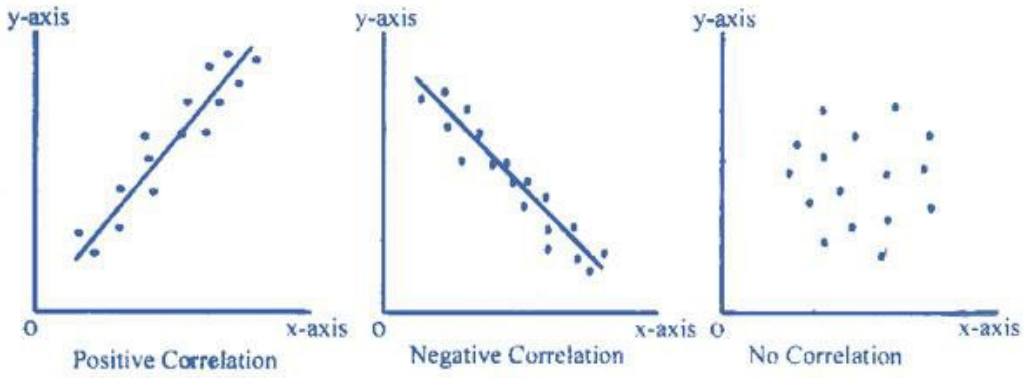
between the variables. It shows the outliers involved in the dataset which cannot be handled by the model.

This Anscombe's quartet is used to describe the importance of visualizing the dataset and how the regression algorithm can be fooled by the same. It shows the importance of visualizing the data before applying various algorithms. It suggests that the data features must be plotted to see the distribution of the sample to understand the various anomalies present like outliers, etc.

3. What is Pearson's R?

Pearson's R or Pearson's correlation coefficient measures the statistical relationship, or association, between two continuous variables. It is the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Pearson's R measures the strength of the Linear Relationship between two variables. Pearson's R will always lie between -1 and 1.



- Positive correlation means that both the variables increase together. The value of the coefficient is 1 for a perfect positive correlation.
- Negative correlation means that both the variables decrease together. The value of the coefficient is -1 for a perfect negative correlation.
- No correlation means that there is no relationship between the variables. The value of the coefficient is 0 for no correlation.

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

Ans: Feature Scaling is one of the important pre-processing that is required for standardizing/normalization of the input data. Scaling can make a difference between a weak model and a better one. It is performed to handle varying magnitudes or values or units.

Most of the times, the collected data contains highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units, hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

The following are the techniques to perform Feature Scaling –

- Normalization –

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

$$X = (X – Xmin)/(Xmax – Xmin)$$

where –

Xmax – maximum value of the feature

Xmin – minimum value of the feature

- Standardization –

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X = (X - \mu)/ \sigma$$

where –

μ - the mean of the feature

σ – the standard deviation of the feature

Normalization vs. Standardization –

- Normalization is mostly used when the distribution of the data does not follow a Gaussian distribution. This can be useful for algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization is used in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$$X\_1=C+ \alpha\_2 X\_2+\alpha\_3 X\_3+\cdots$$

$$VIF\_1 =1/(1-R\_12)$$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$$X_2=C+ \alpha_1 X_1+\alpha_3 X_3+\cdots$$

$$VIF_2=1/(1-R_{22})$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

Ans: Q-Q plots (Quantile-Quantile plots) are plots of two quantiles against each other. Q-Q plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or uniform distribution. It also helps us determine if two data sets come from populations with a common distribution.

Q-Q plots helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

- It can be used with sample sizes also.
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets -

- **Similar distribution** – If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis.
- **Y-values < X-values** – If y-quantiles are lower than the x-quantiles.