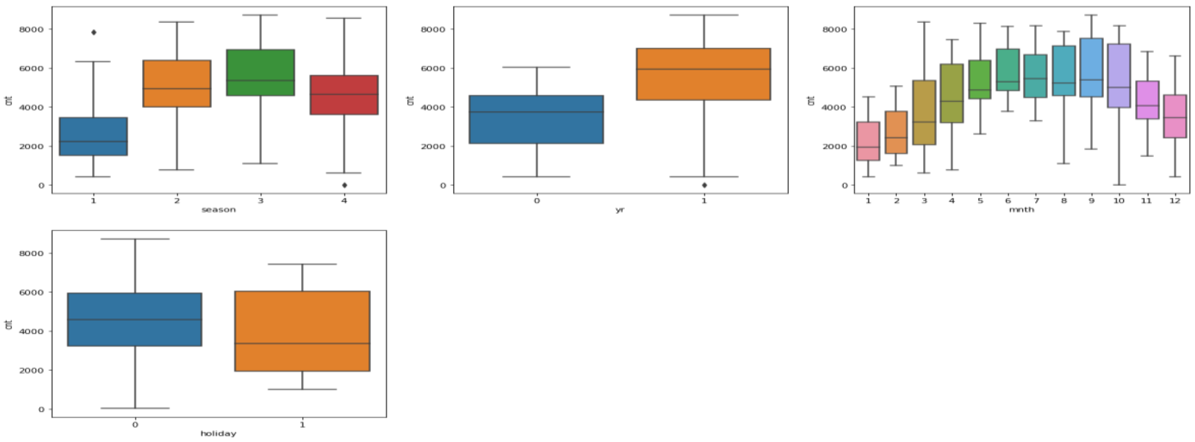
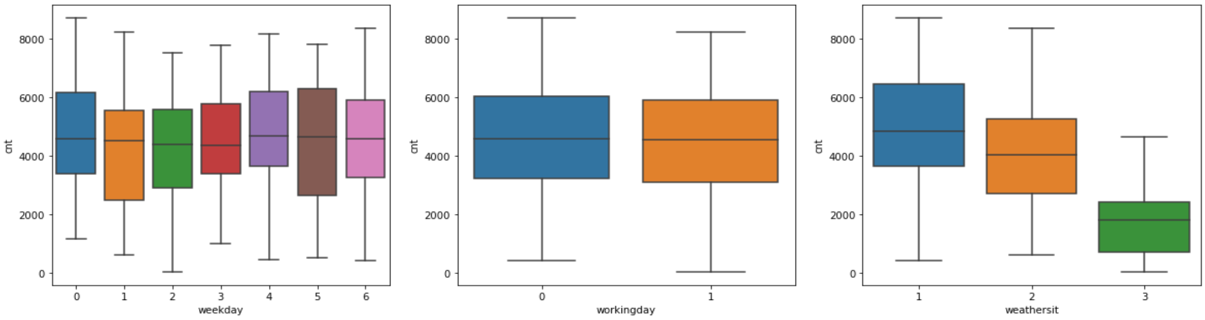
Subjective Questions: Tushar Joshi

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Answer.





Inferences from the boxplots from EDA of categorical variables.

- Yr: The cnt for bike shares incresed in the year 2019
- Holiday:The median values of cnt are almost similar, if it's a working day or weekend/holiday.
- Season:The bike sharing is the least in the spring
- Weekday:The bike sharing is least on Mondays
- Season:The bike sharing values increases in summer months
- Weathersit: Bookings are the highest during clear weather

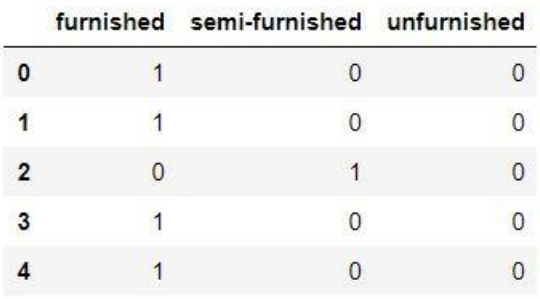
2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

Answer

Dummy Variables are created to show how a categorical columns affects the linear model by converting them to numerical binary variables, to feed to the linear regression model. By using drop_first =True, it drops the first column of the binary columns created, what this does is, it improves the efficiency of the regression model. If the information for n levels can be explained by having n-1 dummy columns.

drop_first=True reduces the extra column created during dummy variable creation. So, it reduces the correlations created among dummy variables.

For example assume the following dataset.

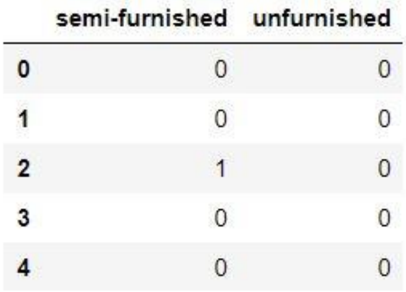


	furnished	semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0

Now, you don't need three columns. You can drop the furnished column, as the type of furnishing can be identified with just the last two columns where —

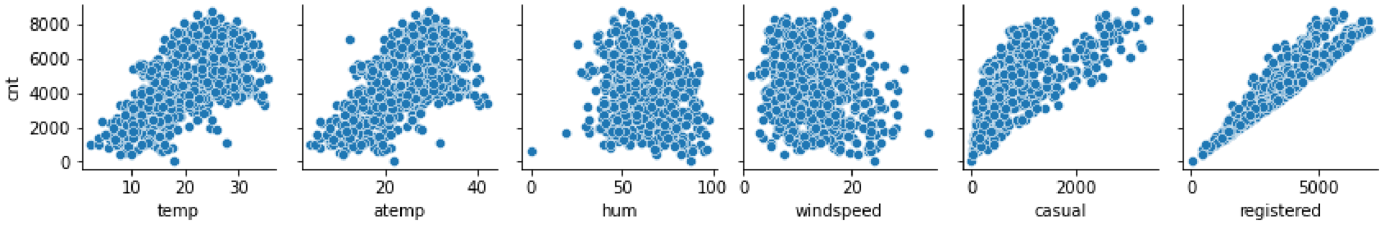
- 00 will correspond to furnished
- 01 will correspond to unfurnished
- 10 will correspond to semi-furnished

After dropping the dataset becomes.



	semi-furnished	unfurnished
0	0	0
1	0	0
2	1	0
3	0	0
4	0	0

3 .Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

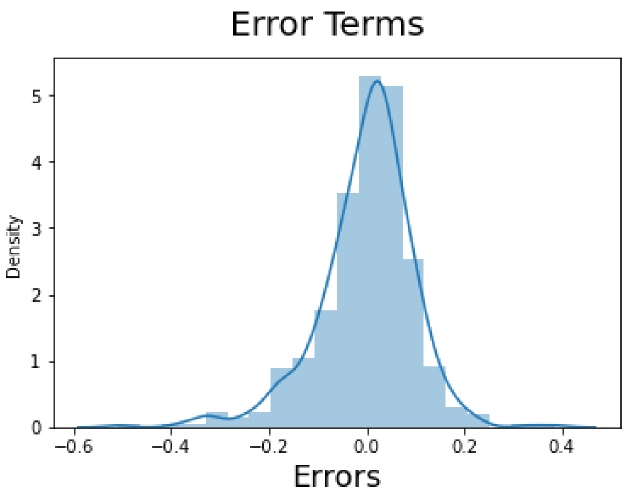


The independent numeric variable ‘registered’ has the highest correlation with the target variable ‘cnt’,before the model building, from the EDA. But after data preparation, and dropping unnecessary features due to multicollinearity the numerical variable ‘atemp’ has the highest correlation with the target variable ‘cnt’.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer)

The assumptions of a linear regression model is validated when the error terms are normally distributed with mean equal to 0.After building the model, we need to verify if the model is not violating this assumption. We just plot a histogram of the error terms to check whether they are normally distributed. And another assumption was that the error terms should be independent of each other. Again for this, we plot the error terms, this time with either of X or y to check for patterns. The histogram is as shown below, which validates the assumption.



The distribution is centred around zero and is qualitatively normal. Hence we can say that the model satisfies the assupmtions of linear regression¶

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer

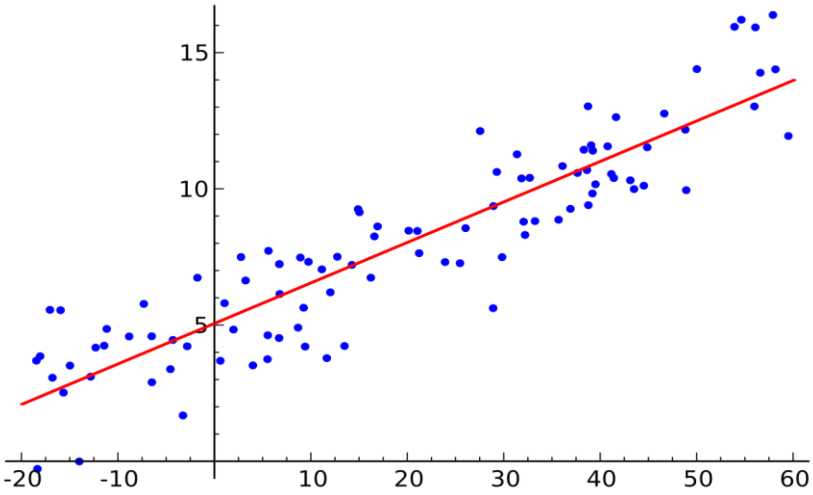
The top three features are
1. ”yr”- coeff=0.2240
2. ‘holiday’, coeff=-0.0897
3. ‘atemp, coeff=0.6176

General Subjective Questions

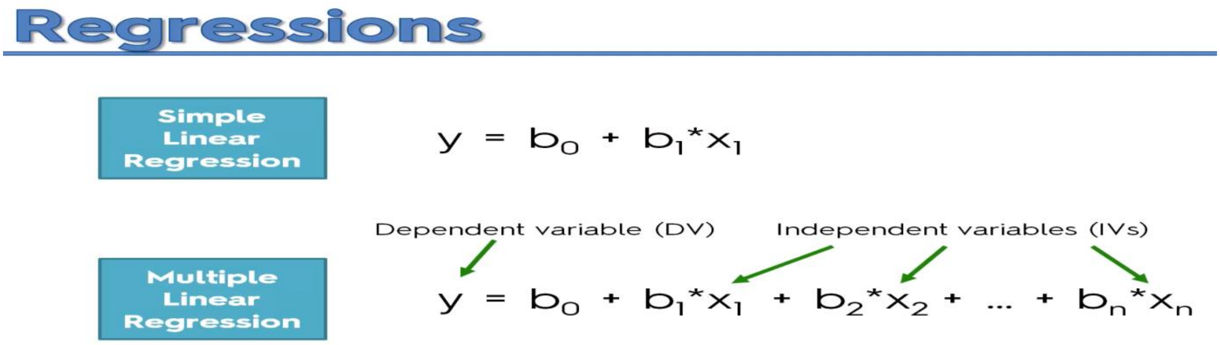
1. Explain the linear regression algorithm in detail. (4 marks)

Answer)

Linear regression is a way to measure the relationship bw predictive variables and target variables, it is a type of machine learning algorithm that uses past data or the labels, which are continuous, hence it is called regression.



Linear regression is one of the basic types of ML algorithms where we train a regression model to forecast the behaviour of data based on predictive variables. There are two types, **simple linear regression and multiple linear regression.** The equations are explained as below.



$$y = b_0 + b_1 * x_1$$

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

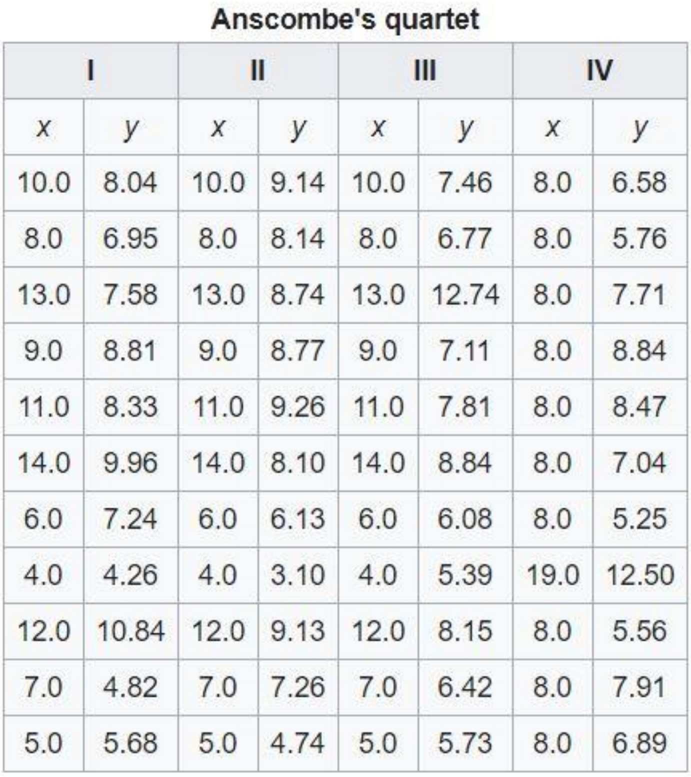
In Python we can the scikit learn library to import the linear regression model and use it directly or we can write our own regression model based on the equations above.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer)

Ascombe's quartet tells us the importance of visualizing the data to make inferences than simply computing summary statistics. The ascombe's quartet contains the values as shown the table.

They were constructed in 1973 by statistician Francis Anscombe to show the importance of graphical visualization of data before analyzing it and the effect of outliers and other influential observations on statistical properties.



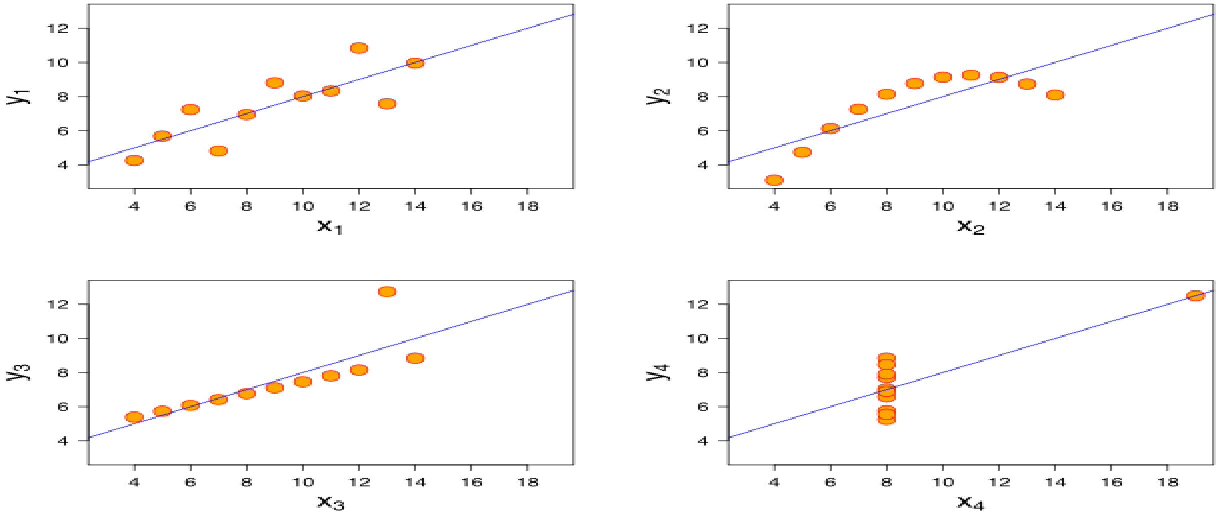
Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All the summary statistics for each data set are

- The average *x* = 9
- The average *y* = 7.50 for each dataset
- The variance for *x* = 11 and
- The variance for *y* = 4.12
- The correlation between *x* and *y* = 0.816 for each dataset
- A linear regression (line of best fit) is *y = 0.5x + 3*

But the plots of these four data sets on an x/y coordinate plane, we get the following results:



The quartet is still used to show the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistical properties for describing realistic datasets

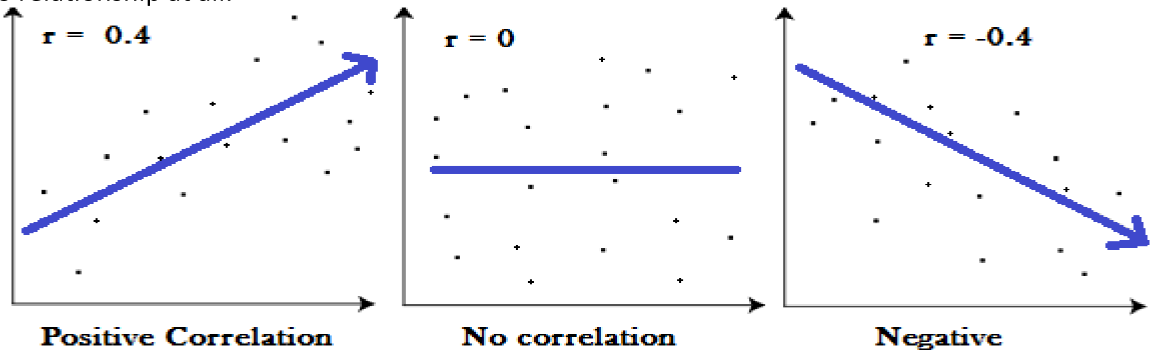
3. What is Pearson's R? (3 marks)

Answer)

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear relation between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, it indicates how far away all these data points are to this line of best fit or how well the data points fit this new model.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1= strong positive relationship.
- -1 =strong negative relationship.
- 0 = no relationship at all.



3. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   Answer)

In Simple Linear Regression, scaling doesn't impact your model. So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very inefficient at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling

1. Min-Max scaling (Normlisation)

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

In Normalisation all the numerical features are adjusted between 0-1, thus taking care of the outliers

2. Standardisation (mean-0, sigma-1)

$$X' = \frac{X - \mu}{\sigma}$$

.

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

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answer)

VIF - the variance inflation factor - The VIF shows the variance of the coefficient
Estimate which is being shown by collinearity. (VIF) =1/(1-R_1^2). If there is perfect correlation, then
VIF = infinity.Where R-1 is the R-square value of that independent variable which we want to
show how well this independent variable is explained well by other independent variables-

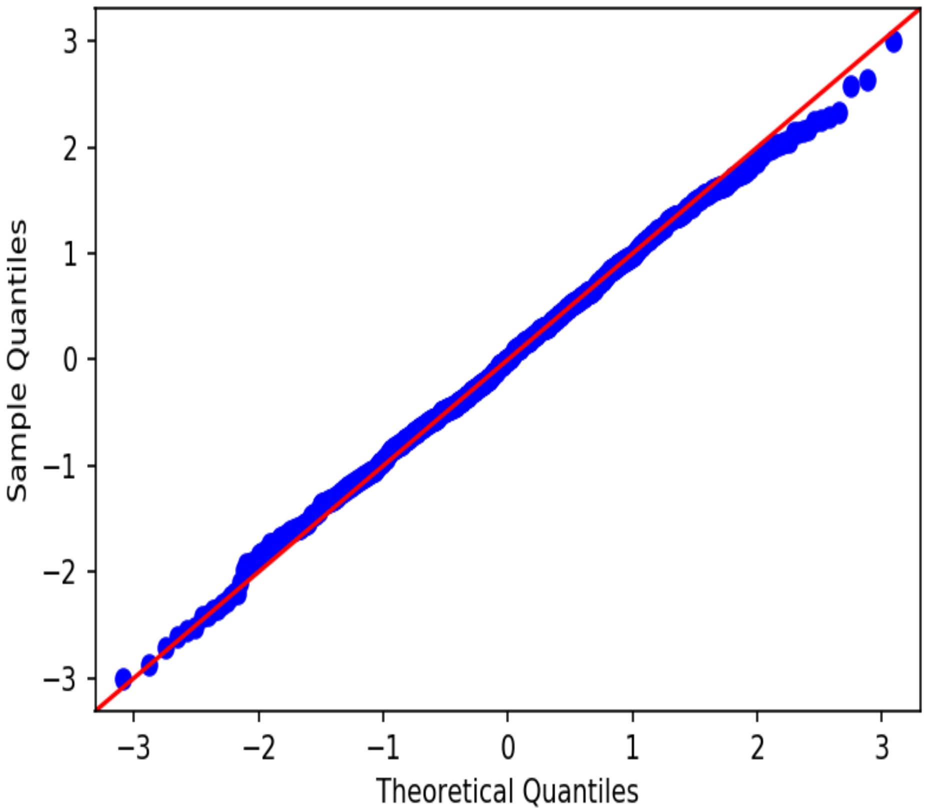
If that independent variable can be explained perfectly by other independent variables, then it will
have perfect correlation and it’s R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives
VIF = 1/0 which results in “infinity”

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

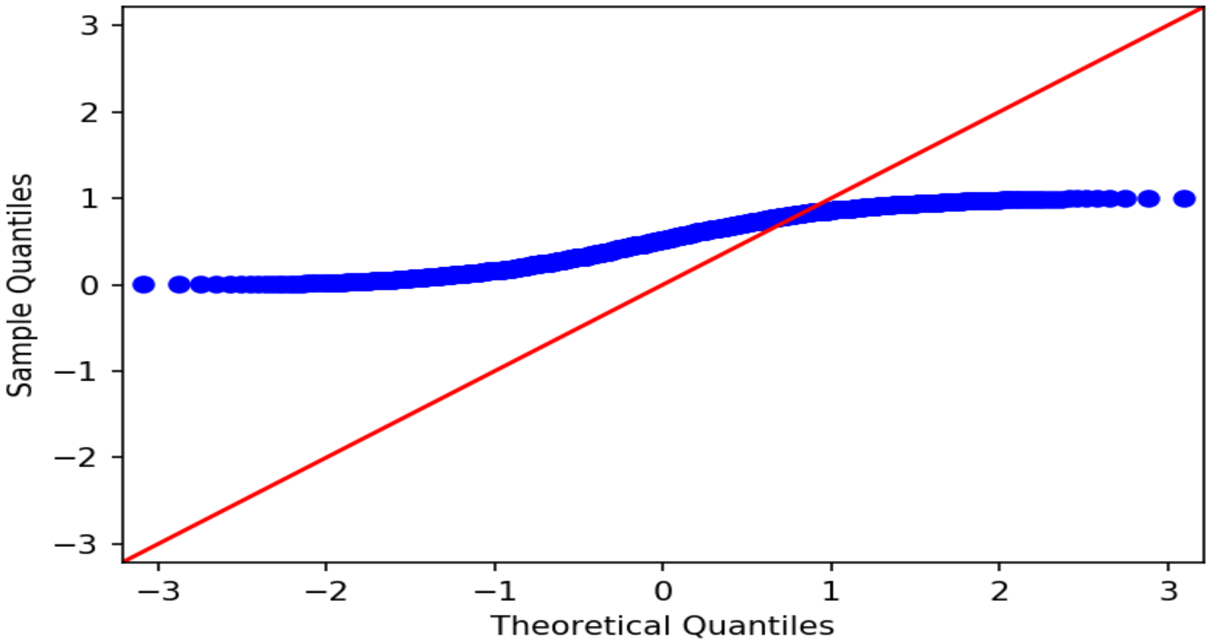
(Q-Q) plot stands for Quantile-Quantile Plot, it is a graphical tool to assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This 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.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis



Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis


Sample Quantiles
Theoretical Quantiles