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?

The categorical variables are: season, weathersit, month, weekday

Holiday and workingday need to be mentioned here although they are mapped to numeric categories 0 and 1 in the dataset representing No/Yes, False/True etc.

The observations for the categorical variables from the dataset are as follows:

- Demand of rental bike is highest at the fall and lowest at the spring. The decrease in the rental count might have other factors involved such as holiday season, bad weather etc. which should be analysed in the bivariate/multivariate analysis.
- The demand is maximum when weather is clear or partly cloudy and it decreases when there is mist, snow, rain, thunderstorm, scattered clouds i.e. when the weather turns bad.
- Demand varies a lot throughout the year. Count is comparatively lower during the beginning of every year gradually increasing reaching its peak at September. Holidays/weather condition might play a part.
- Demand is more on Weekends and lower on Mondays but the mean/average remains comparatively uniform throughout the week.
- Average Count of rental bikes does not show significant change for working days
- Demand of rental bikes decreases considerably on holidays.

Except Working days, all other categorical variables show influence on the dependent variables and should be considered as factors of the Regression Model.

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

   The drop_first = True is used when we can drop the first column of the dummy representation of a categorical variable with multiple levels and still can uniquely identify all the levels. Since we do not get any additional information from this column, it is safe to drop it thereby reducing the count of columns and making the dataset simpler.

   Sometimes the if we create n dummy variables for n discrete categorical variables, these variable columns create correlation among themselves which increases the multicollinearity of the model. This is also known as Dummy variable Trap. To avoid this we prefer dropping the first column.

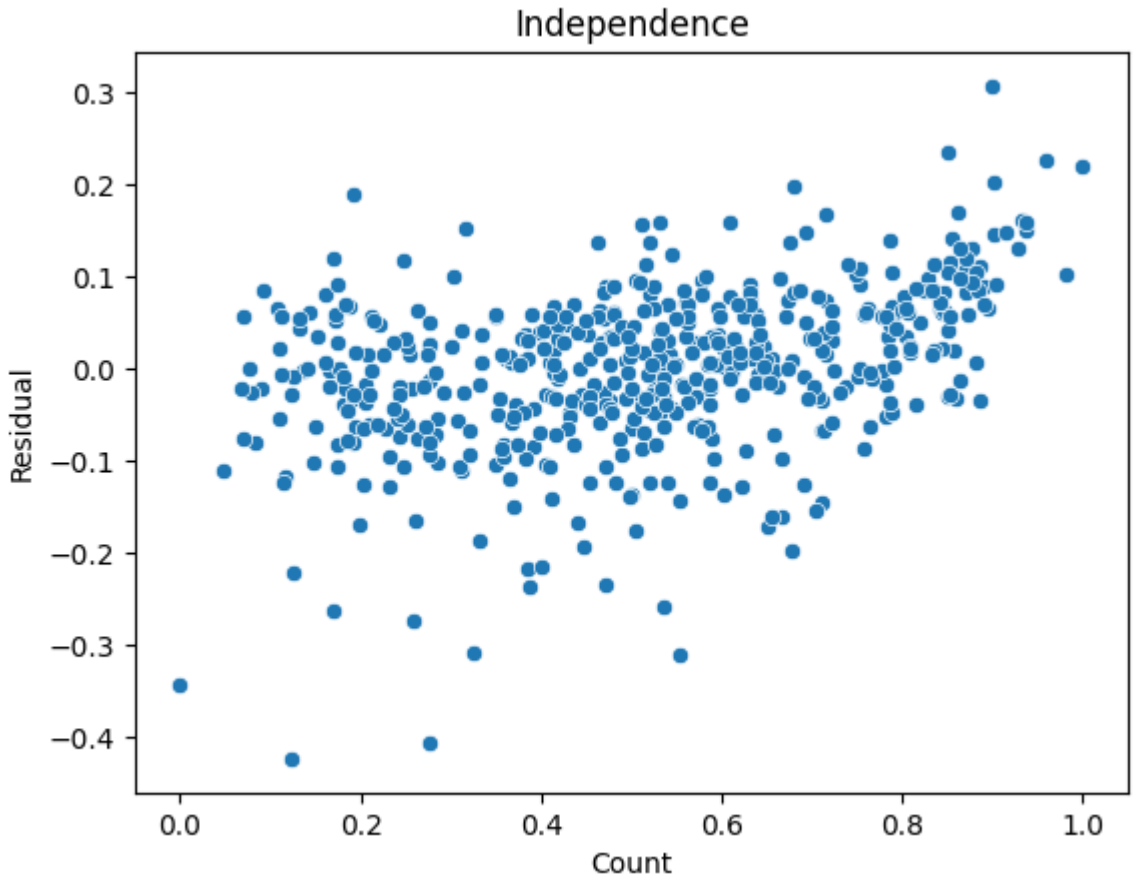
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temp and Atemp has the highest correlation with the target variable.

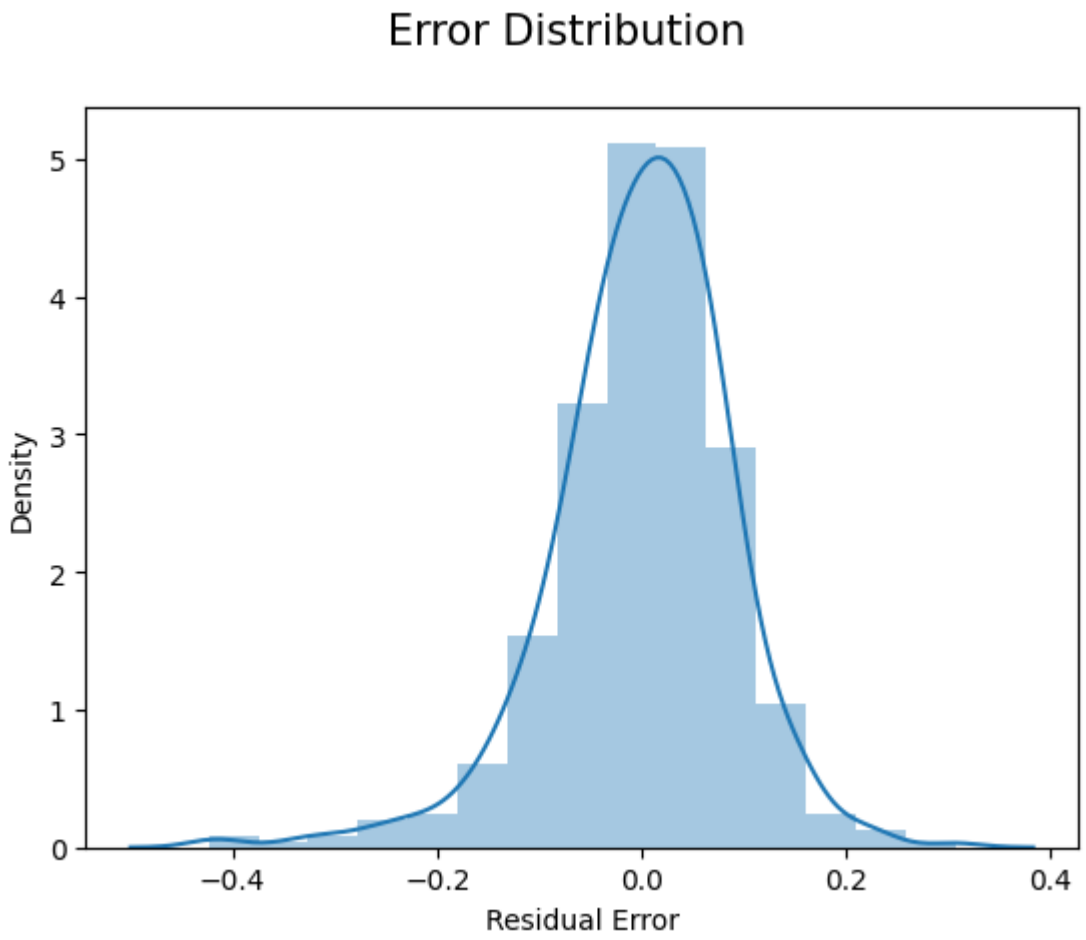
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Linearity: The predictor variables had linear relationship with the target variables. The variables such as day of the month, which did not show linear relationship were dropped.

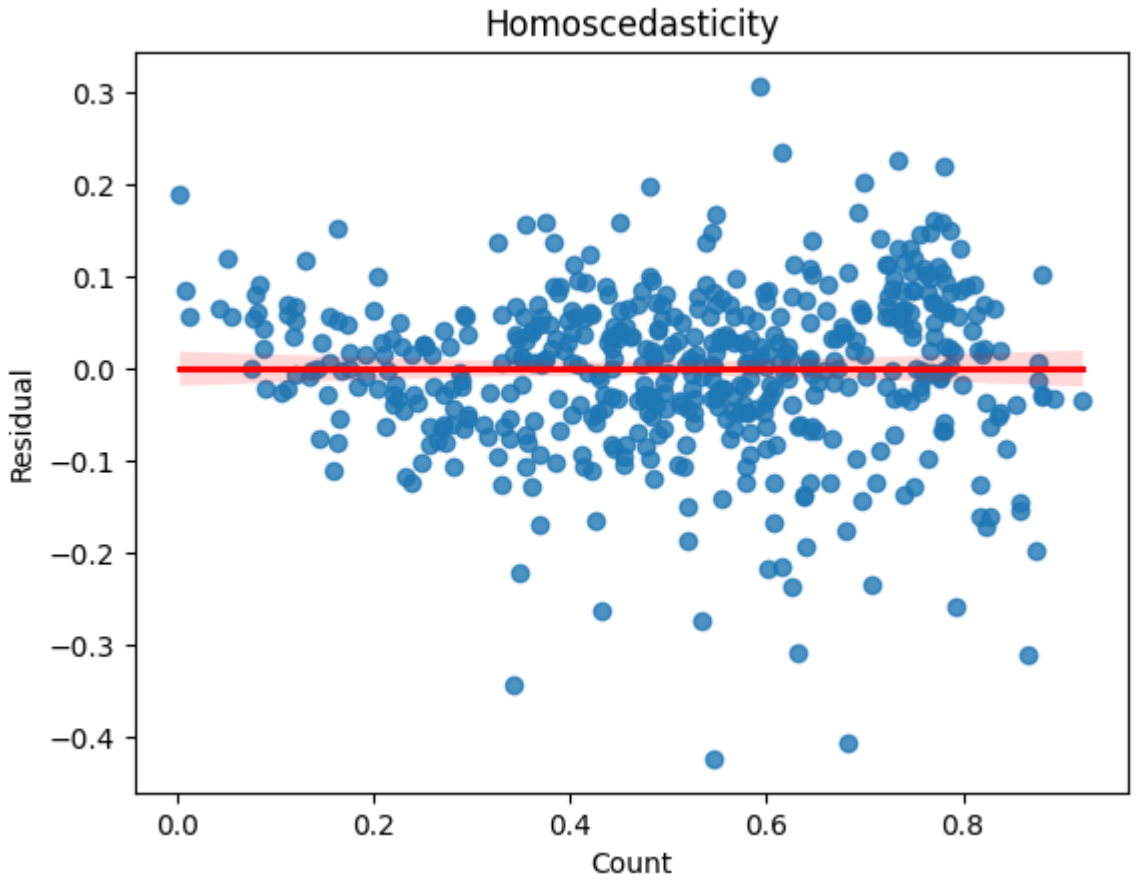
- Independence: On plotting the target variable (Train data) with the residual, we found the error terms are randomly scattered without any visible pattern.



- There is little or no multicollinearity in the data. All VIF values are <5
- On plotting the error distribution for train data, we found that it is distributed normally with the average as zero.



- Homoscedasticity - residual vs predicted data do not have distinctive pattern. The scatter points are evenly distributed in both sides of the line. We can confirm Homoscedasticity is maintained.



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

According to our final Model, the top 3 features contributing towards increase/decrease of demand of the shared bikes are as follows –

Temperature: Increase in 1 unit of temperature will increase the demand by 0.55 units.

Weather: The demand of bikes decrease by 0.25 units for every unit change in the weather towards Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.

Year: The demand of bikes seems to increase by 23 times every year.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is the Machine Learning algorithm where output variable and input variables are linearly related. The output variable is predicted to be continuous.

The algorithm for Linear Regression can be listed as follows:

I. Create a training data set and a test data set (usually in a ratio of volume as 70:30)
II. Create a model using all available variables
III. Continuously check the VIF and the summary of the model
IV. Identify the variables which are insignificant (P-value>=0.05)
V. Remove these variables one at a time
VI. In case we have multiple variables with high VIF, we will remove the one with higher P-value
VII. Continue these steps until we reach a point where we are left with only significant variables with acceptable VIF (<=5)
VIII. Predict the output variables from the training dataset
IX. Identify the best fit line by minimizing the residuals/errors between the actual data and the predicted data.
X. Use the best fit line to predict data for test data set
XI. Check the strength of the model using Coefficient of Determination or Root mean square error calculations

There are certain assumptions behind this algorithm-

I. The dependent and independent variables have linear relationship
II. The residual error from the model evaluation is normally distributed with mean at 0
III. The independent variables do not show strong correlations
IV. The variance of the residuals should be consistent

2. Explain the Anscombe's quartet in detail

Anscombe's quartet is a set of four data set created by Statistician Francis Anscombe. These data sets have identical statistical summary but they are visually different. It is often used to define the importance of visual check of dataset instead of just relying on their statistical summary.

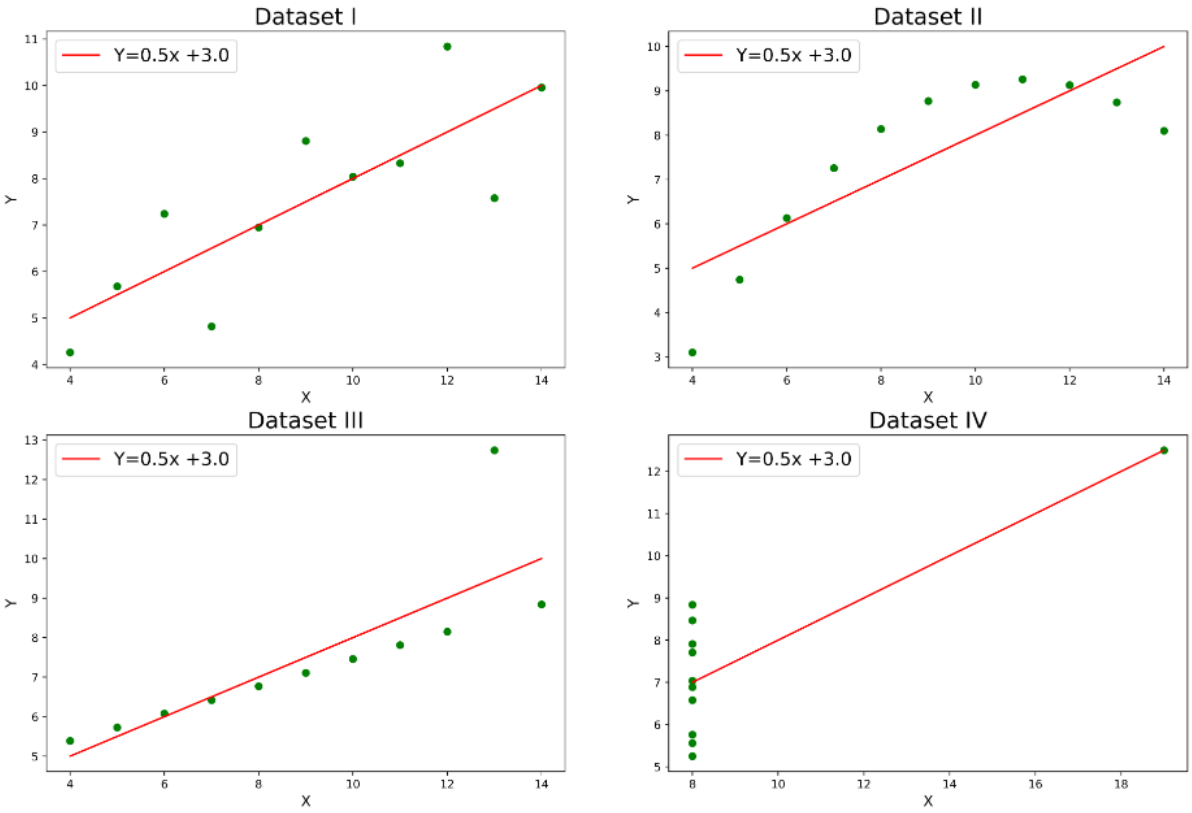
Below are the four datasets of Anscombe's quartet-

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

If we look into the summary we can see the mean, variance, correlation etc for all four of these data sets are exactly same.

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

But if we plot the scatterplot of the data points we can see the difference in the patterns.



3. What is Pearson's R?

Pearson correlation coefficient (R) is used to check how strongly two variables are correlated to each other. The range of R varies between -1 to 1.

-1: The variables have very strong negative correlation

0: The variables are not correlated to each other

1: The variables have very strong positive correlation

The statistical formula to find R is –

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

r = Pearson correlation coefficient
x = Values in the first set of data
y = Values in the second set of data
n = Total number of values.

i = number of sample

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

   The process of bringing all data of a data set in the same range, magnitude or unit before we perform Regression analysis on the data is called scaling.

   If we do not scale data of different magnitudes or units, the regression algorithm will only take the magnitude of the data into account ignoring its units. This will reduce the accuracy of the Regression Model.

   E.g.: We have the record of temperature (as columns) from different regions of India. Some of the variables (temperature) is logged as Celsius and some are logged in Fahrenheit. Obviously, we need to scale all the temperature data into one single unit to ensure we are considering the same meaning for all.

Points	Normalized Scaling	Standardized Scaling
Range	Between 0 to 1	Between -1 to 1
Data Distribution	Data is not distributed normally	Data follows Gaussian Distribution
Outliners	Can affect the scaling	Does not affect scaling

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

   For VIF to be infinity, R Squared should be 1, which happens when the variables have perfectly strong correlations (R=-1/+1) with each other.

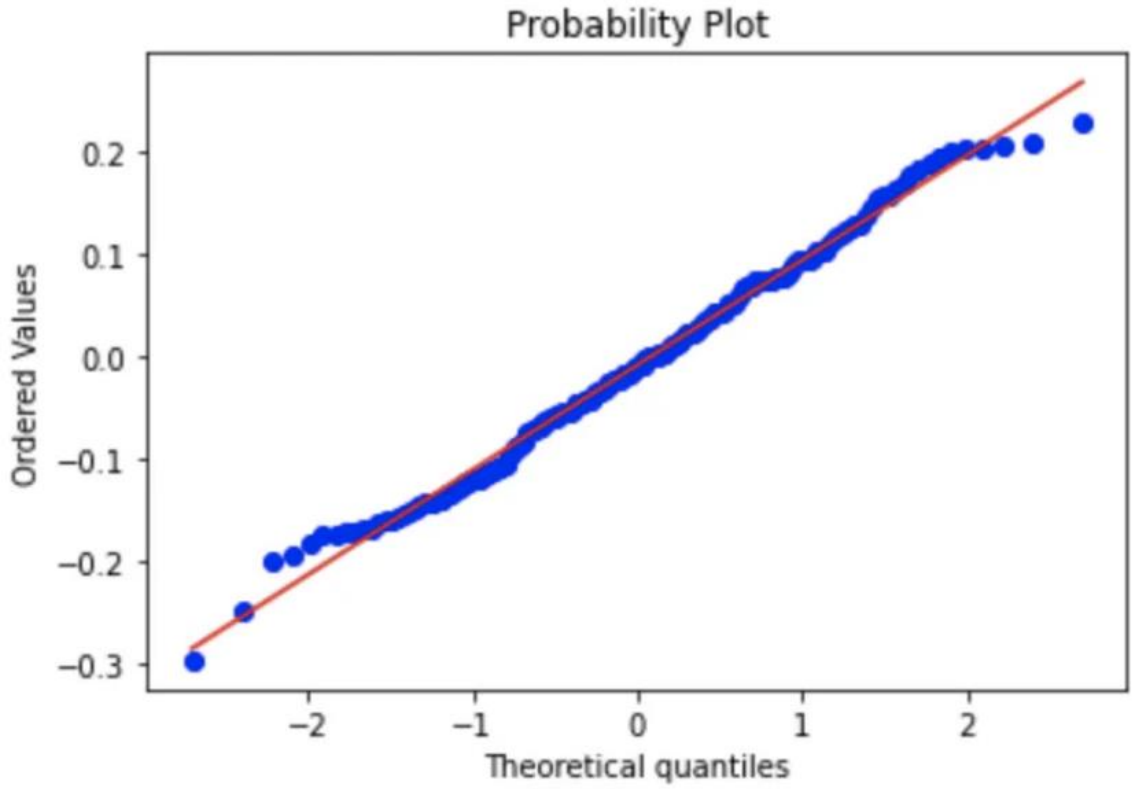
   In such scenarios we have to identify the variable which is causing the perfect correlation and we need to drop it from the dataset.

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

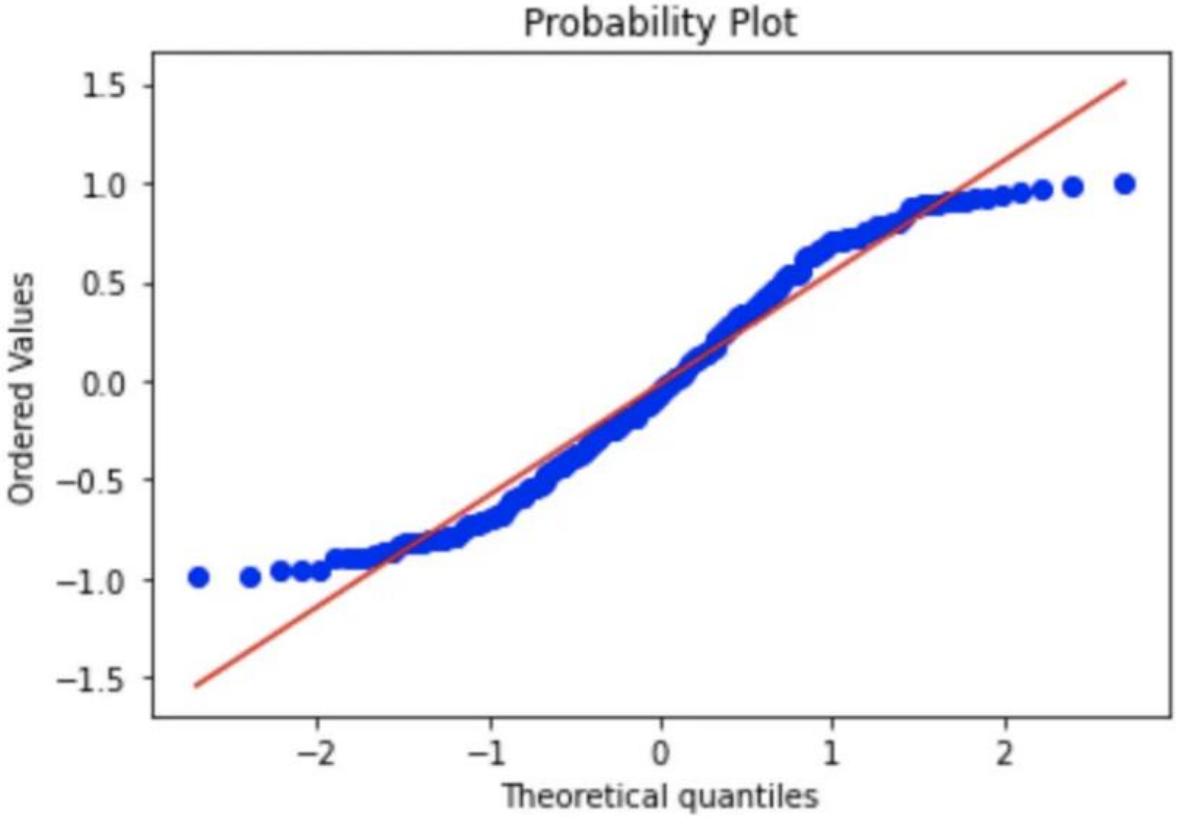
   Q-Q plot is a graphical method which helps us to understand whether two dataset follows same probability distribution or belongs to the same population or not.

If two quantiles are sampled from the same distribution, they should roughly fall in a straight line. If they don't then we can conclude that the residuals are not normally distributed.

In the below graphical illustration, we can see most of the data points lie on a straight line, this indicates that the data point is normally distributed.



In the below diagram we can see the data points do not lie on a straight line, indicating that they are not normally distributed.


Probability Plot
Ordered Values
Theoretical quantiles
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1
2