Boom Bikes - Subjective Questions

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

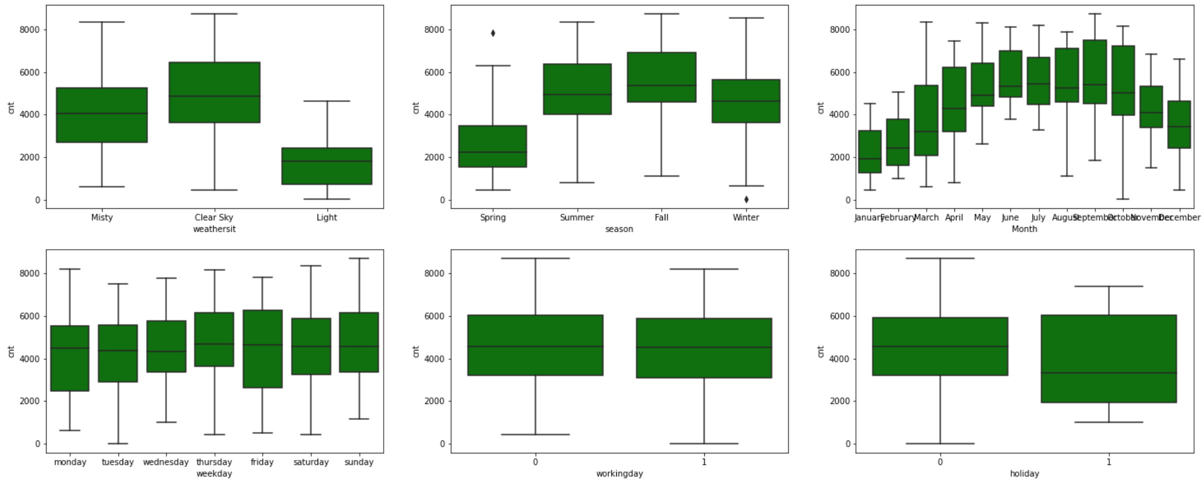
Categorical variables like weekday, working day or holiday – do not have much impact on the target variable - implies whether it is a weekday or a working day or a holiday, the count almost remains the same.

Whereas all other categorical variables are quite significant - like Month, Season, Weathersit and year.

Weather situation like light snow and rain is a significant predictor and it impacts the ride count negatively by a factor of -0.2957

Year is also a very strong predictor - by each passing year the rentals are increasing by 0.2335

Similarly Spring season also negatively impact the ride count by a factor of -0.1122



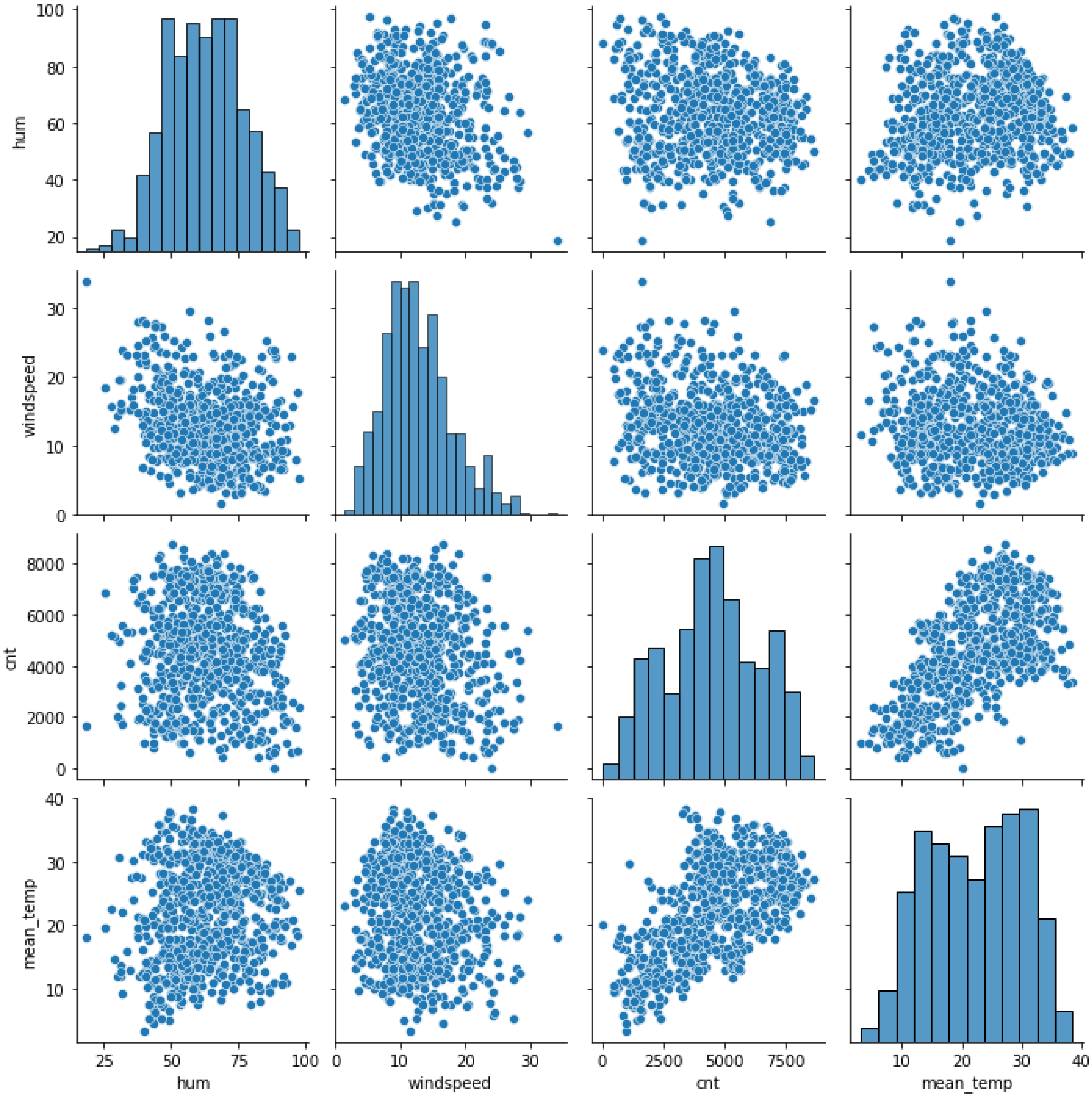
2.Why is it important to use drop_first=True during dummy variable creation?

This is to avoid the dummy variable trap. The N-th dummy variable is obvious since it can be easily calculated by the value of the remaining N-1 variables.

Hence we need only N-1 dummy variables. This is in fact avoid multicollinearity between the dummy variables which is one of the assumption if linear regression.

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

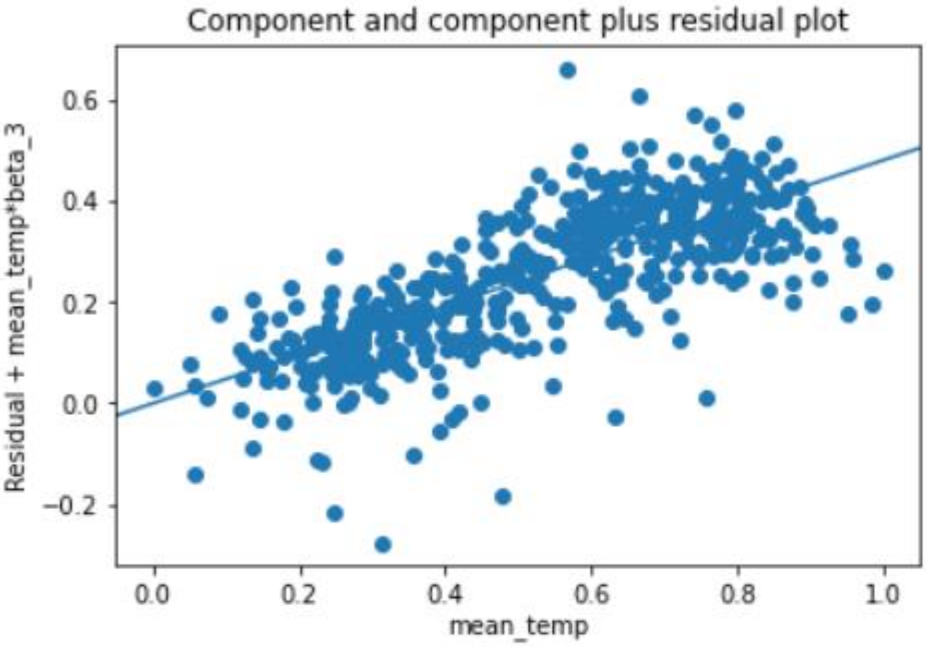
Out of the 3 numeric variables windspeed, hum and mean_temp - mean_temp is having the highest correlation with the target variable with a coefficient of 0.63



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

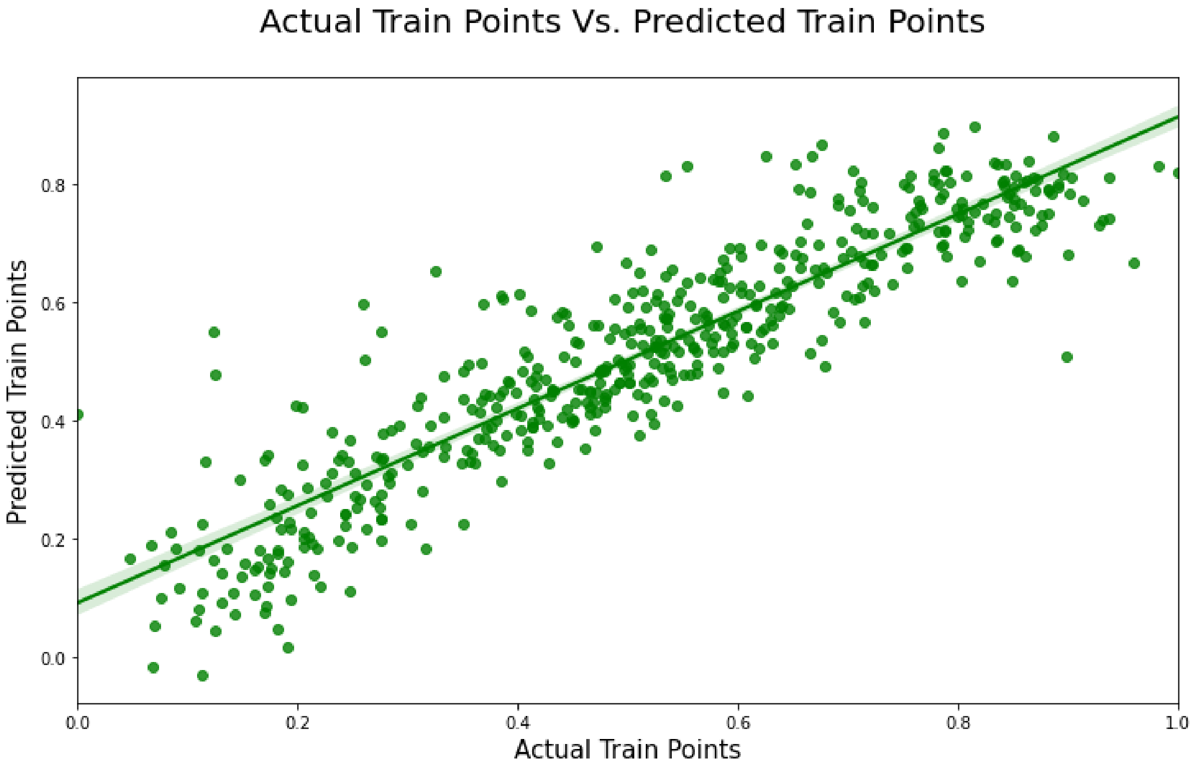
1. **Linearity** - The target variable ‘cnt’ can be explained in terms of the predictors that form of a linear equation. This satisfies the linearity of the model.

cnt = 0.1932 + (0.2335 * Year) - (0.0984 * holiday) + (0.4820 * mean_temp) - (0.0686 * Month_July) + (0.0654 * Month_September) - (0.1122 * season_Spring) + (0.0505 * season_Winter) - (0.2957 * weathersit_Light) - (0.0783 * weathersit_Misty)

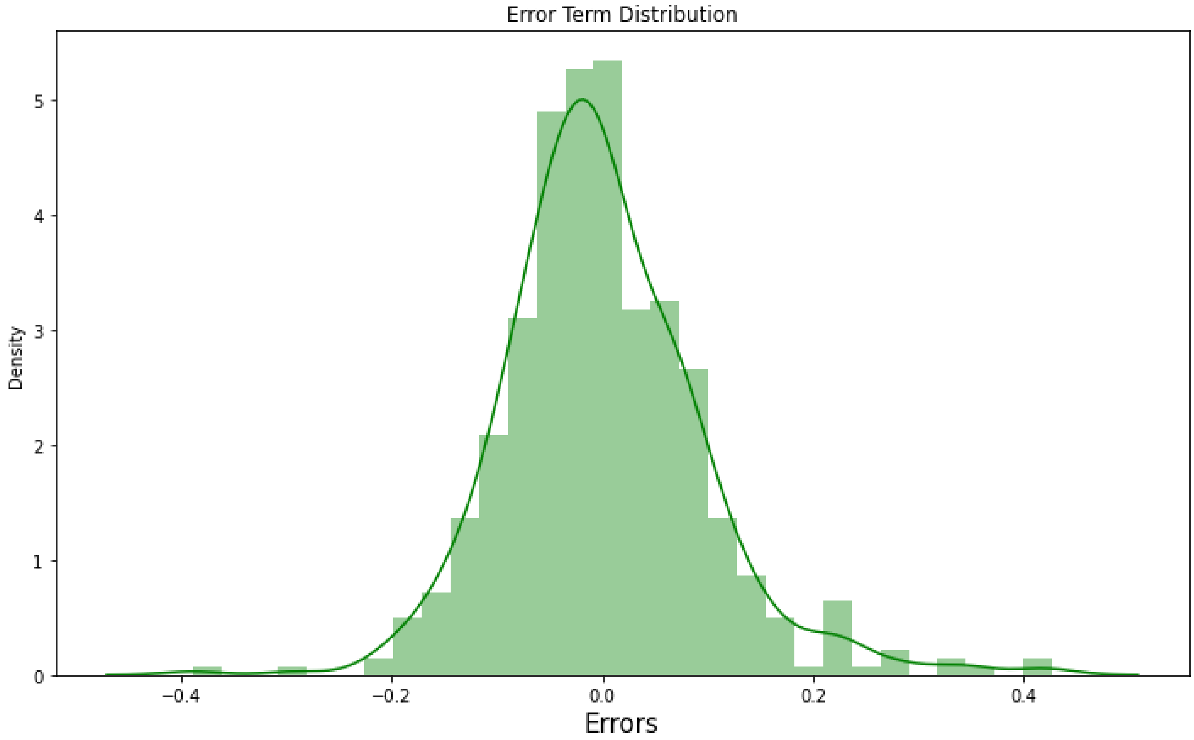


Also the above plots shows the relationship between the model and the predictor mean_temp. As we can see, linearity is well preserved

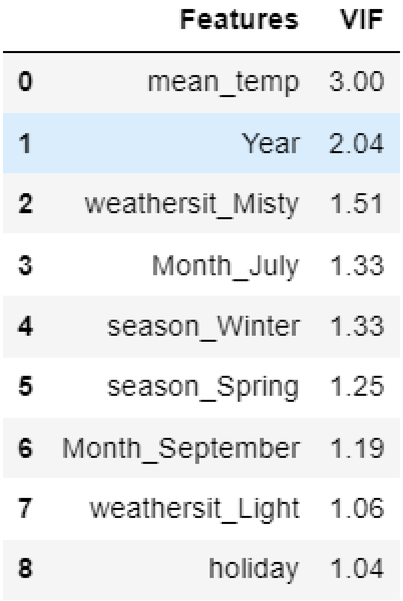
2. **No heteroscedasticity** : Plot y_train and y_train_prediction in a scatter plot - this is in fact the residual plot, here the scatter plot shows a uniform distribution without much variance. Hence, we can ignore the heteroscedasticity issue.



3. **Normality of error terms**: Potted histogram of residual term show a distribution which resembles a bell curve, hence, it satisfies the normality of error terms.

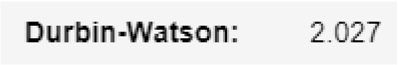


4. **No Multicollinearity** - This condition was met by the final model where all the VIF values are well within range.



	Features	VIF
0	mean_temp	3.00
1	Year	2.04
2	weathersit_Misty	1.51
3	Month_July	1.33
4	season_Winter	1.33
5	season_Spring	1.25
6	Month_September	1.19
7	weathersit_Light	1.06
8	holiday	1.04

5. **No autocorrelation**: Here the final model got a Durbin Watson score of 2.027 which shows almost zero auto-correlation among variables. The Durbin Watson statistic is a test for autocorrelation in a data set. A value of 2.0 means there is no autocorrelation detected in the sample. Values from zero to 2.0 indicate positive autocorrelation and values from 2.0 to 4.0 indicate negative autocorrelation.



```
Durbin-Watson:     2.027
```

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

As per the final model, the top 3 predictors that influences bike booking are:

1. Temperature (mean_temp) - Coefficient of ‘0.4820’ indicated that temperature has significant impact on bike rentals
2. Light Rain & Snow (weathersit =3) - Coefficient of ‘0.2957’ indicated that the light snow and rain negatively impact the rentals.
3. Year (yr) - Coefficient of ‘0.2335’ indicated that a year wise the rental counts are increasing

General Subjective Questions

1.Explain the linear regression algorithm in detail

Linear regression is a supervised learning algorithm. The algorithm tries to find out the best fitting line or hyperplane for a given set of data points which predict the target variable with some level of accuracy.

Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages –

1. analysing the correlation and directionality of the data
2. estimating the model, i.e., fitting the line
3. evaluating the validity and usefulness of the model

The first step enables the researcher to formulate the model, i.e. that variable X has a causal influence on variable Y and that their relationship is linear.

The second step of regression analysis is to fit the regression line. Mathematically least square estimation is used to minimize the unexplained residual.

The distance between the regression line and the data point represents the unexplained variation, which is also called the residual. The method of least squares is used to minimize the residual.

The key measure to the validity of the estimated linear line is R^2. R^2 = explained variance/ total variance.

The last step for the linear regression analysis is the test of significance. Linear regression uses two tests to test whether the found model and the estimated coefficients can be found in the general population the sample was drawn from. Firstly, the F-test tests the overall model. The null hypothesis is that the independent variables have no influence on the dependent variable. In other words, the F-tests of the linear regression tests whether the $R^2=0$. Secondly, multiple t-tests analyse the significance of each coefficient and the intercept. The t-test has the null hypothesis that the coefficient/intercept is zero.

We also see the bias and variance trade off to assess the accuracy of the model.

2. Explain the Anscombe’s quartet in detail

Anscombe's quartet shows how descriptive statistics could be misleading at times if we do not try to find out the hidden patterns visually.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

3. What is Pearson’s R?

The Pearson correlation coefficient, also called Pearson’s R, is a statistical calculation of the strength of two variables’ relationships. In other words, it is a measurement of how dependent two variables are on one another.

The equation for Pearson's R is given by

P(x,y)= covariance(X)*Covariance (y)/(std. dev(x)* std.dev(y))

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

Feature scaling is one of the most important data pre-processing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled. Feature scaling helps machine learning, and deep learning algorithms train and converge faster.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

This scales the range to [0, 1] or sometimes [-1, 1]. Geometrically speaking, transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Normalization is useful when there are no outliers as it cannot cope up with them.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std Dev

Standardization can be helpful in cases where the data follows a Gaussian distribution.

Standardization does not get affected by outliers because there is no predefined range of transformed features. It is not bounded to a certain range.

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

An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model. This is a strong indication of multicollinearity in the model.
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.

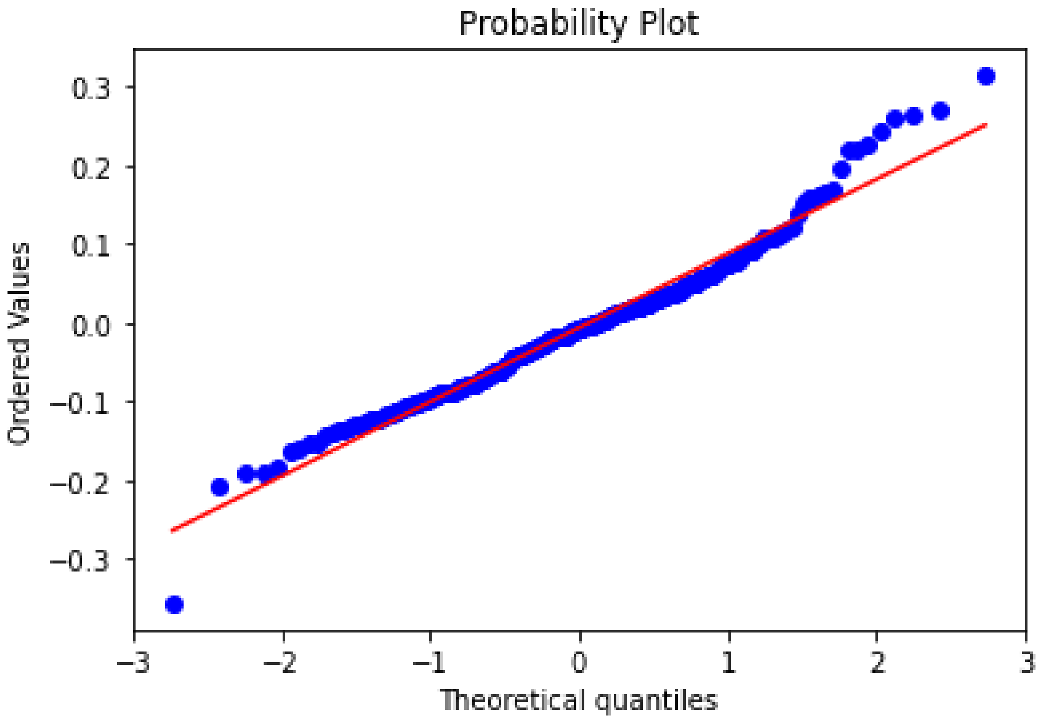
6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q(quantile-quantile) plots play a very vital role to graphically analyse and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x, means it forms a 45 degree angle line.

The most fundamental question answered by Q-Q plot is: Is this curve Normally Distributed?

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. 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.

Also in simple forms we can see whether the error terms exhibits a normal distribution in a linear regression model by plotting the y values against y_pred values and see whether it forms a 45 angled straight line and can confirm the normality of error terms.



By Gloriya Thomas