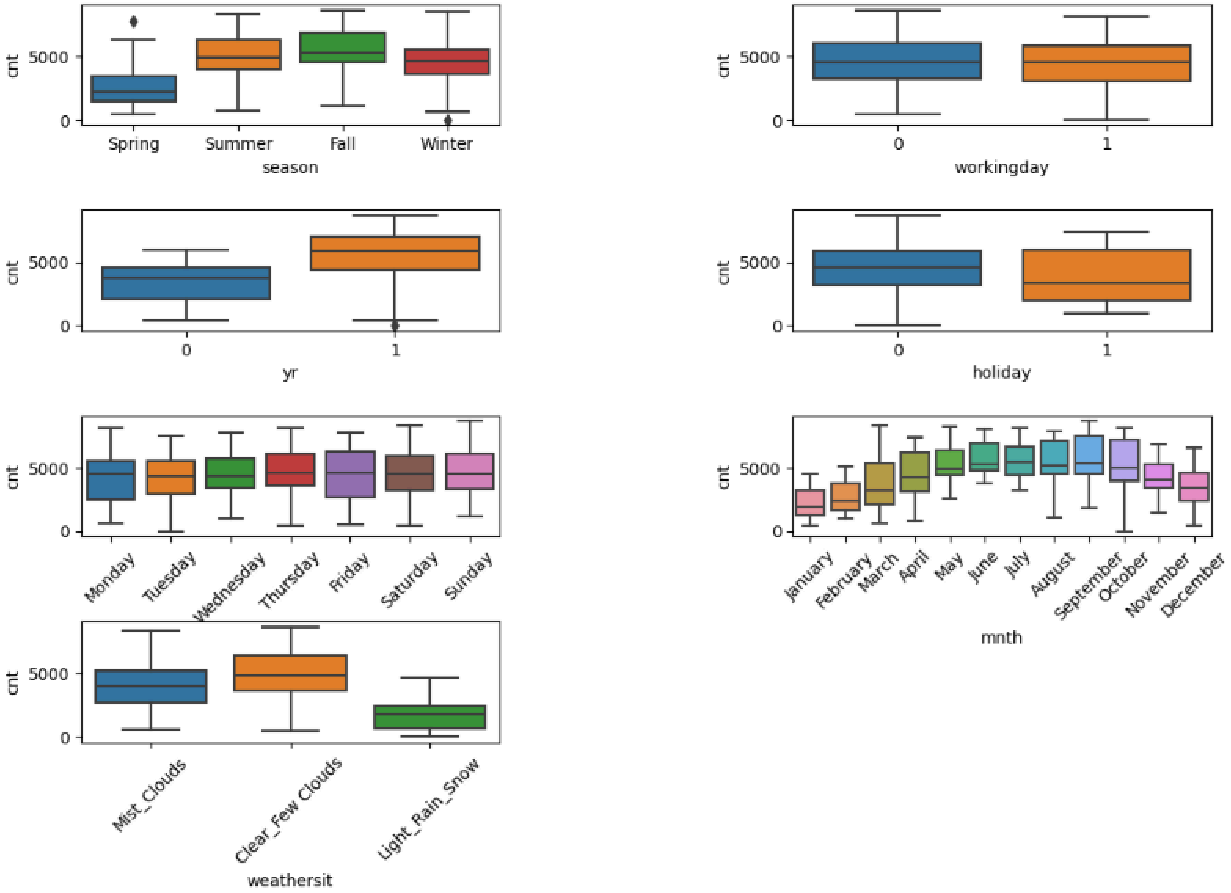
Bike Sharing LR - Assignment

Shayak Majumder

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? (3 marks)*

A: The dependent variable, *cnt* (the number of bikes rented), is impacted by several categorical variables.



- The **fall season has seen the most bike rentals**, followed by Summer, as seen in the ‘mnth’ plot as well.
- **Not much difference** between working days or not, very similar median as well.

- The year **2019** has seen more number of rentals, so there has been a remarkable YoY growth.
- More rentals on holidays, but the median is higher for non-holidays in the ‘holiday’ plot
- **Mondays and Fridays have more number of rentals**, although the median is very similar for all 7 days.
- **Clear weather (with few clouds/partly cloudy) has seen more rentals**, which isn't surprising at all.

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

A: We need to use *drop_first=True* to avoid facing the Dummy Variable Trap.

When we create dummy variables, let’s say for *n* number of categories, *n* number of dummy variables will be created. This is problematic because the sum of all these variables will always equal 1, which in turn leads to a perfect multicollinearity (one dummy variable is perfectly predicted by other dummy variables).

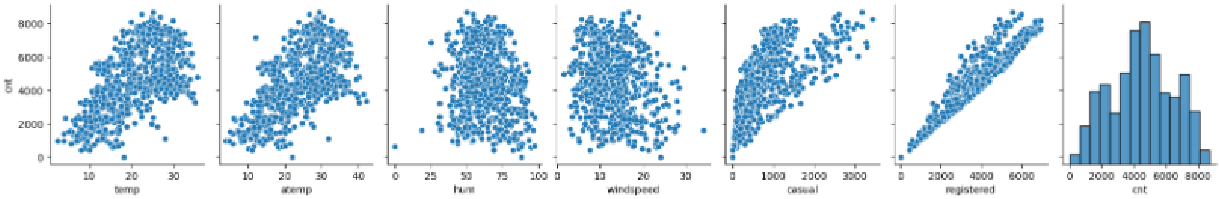
To avoid this, *drop_first=True* helps drop one dummy variable for each categorical variable, creating *n-1* dummy variables. This dropped category becomes the baseline for other categories.

For example, if we have four *Weather* conditions: *Clear, Cloudy, Light Rains*, & *Heavy Rains*, *drop_first=True* will create 3 dummies: *Weather_Cloudy, Weather_Light_Rains*, & *Weather_Heavy_Rains*.

So, in the above case, *Weather_Clear* becomes the baseline, which means if all dummies are 0, it indicates clear weather.

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

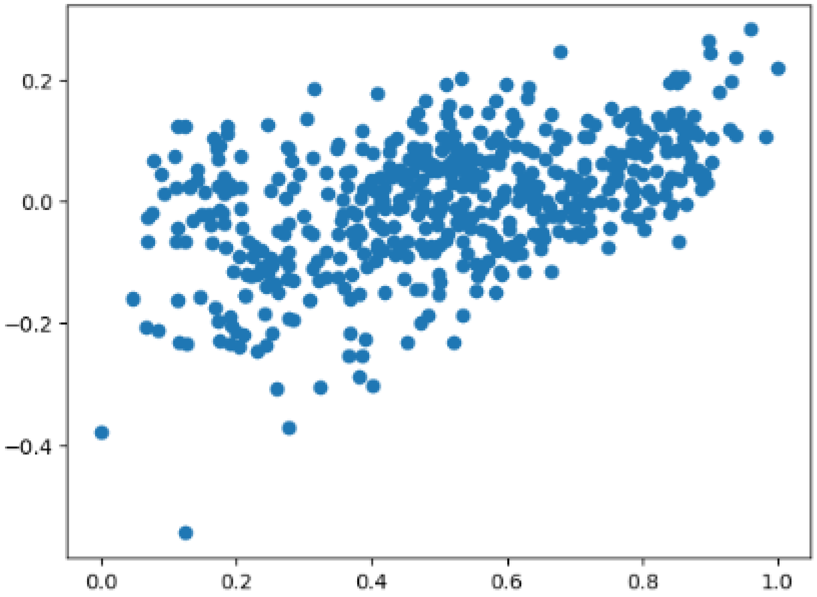
A: For our target variable, *cnt*, the numerical variable that has the highest correlation is *registered*, then followed by *casual*. This makes sense, as *registered* + *casual* = *cnt*.



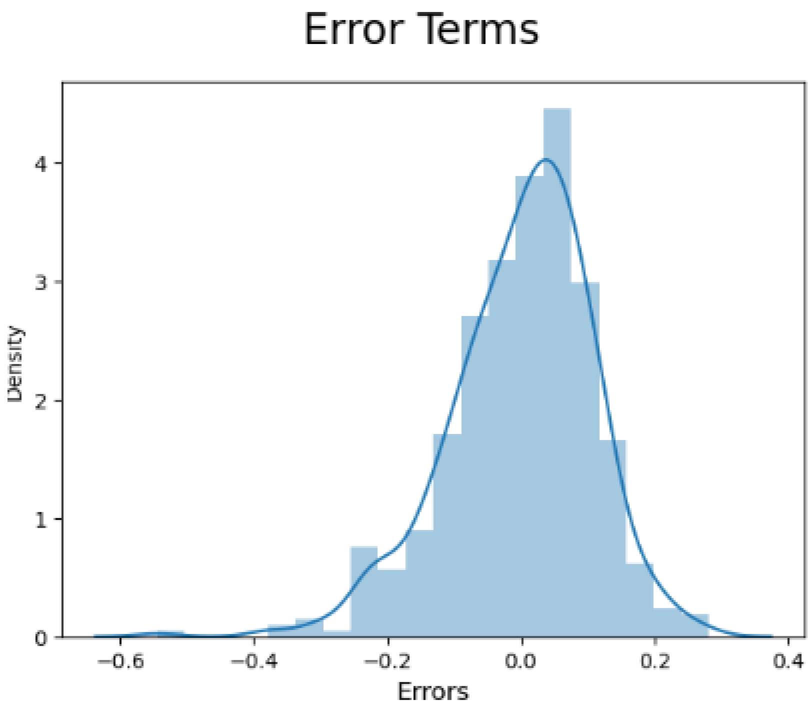
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A: To validate the assumptions of Linear Regression after building my model, I checked a few things:

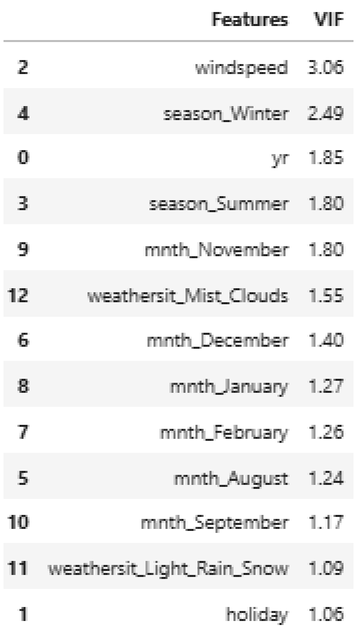
1. I created a scatter plot *(see below)* of residuals, or errors, vs predicted values. No notable pattern was seen, which meant that all the points were scattered randomly, validating the assumption.



2. The scatter plot also showed that homoscedasticity is maintained - residuals were constant through all levels.
3. I created a histogram *(see below)* of residuals. It showed a nice bell-shaped distribution, which meant residuals were normally distributed.

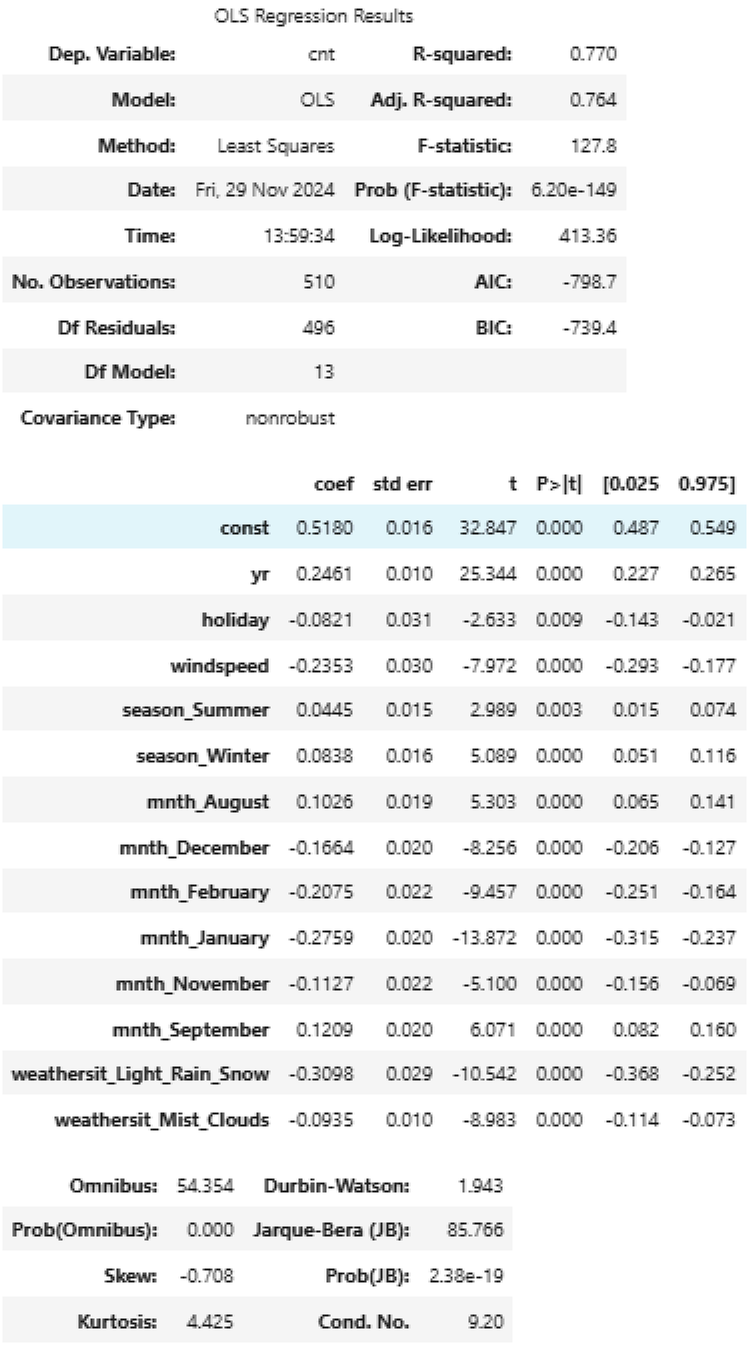


4. Lastly, I calculated the Variance Inflation Factors (VIFs) of all features *(see below)*, and none of them were found to be more than 5.0.



	Features	VIF
2	windspeed	3.06
4	season_Winter	2.49
0	yr	1.85
3	season_Summer	1.80
9	mnth_November	1.80
12	weathersit_Mist_Clouds	1.55
6	mnth_December	1.40
8	mnth_January	1.27
7	mnth_February	1.26
5	mnth_August	1.24
10	mnth_September	1.17
11	weathersit_Light_Rain_Snow	1.09
1	holiday	1.06

5. Of course, the training model summary *(see below)* showed consistently good p-values of well under 0.05. Also, the Durbin-Watson statistic stood at 1.943, which is within the safe zone of 1.5 and 2.5.



OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.770
Model:	OLS	**Adj. R-squared:**	0.764
Method:	Least Squares	**F-statistic:**	127.8
Date:	Fri, 29 Nov 2024	**Prob (F-statistic):**	6.20e-149
Time:	13:59:34	**Log-Likelihood:**	413.36
No. Observations:	510	**AIC:**	-798.7
Df Residuals:	496	**BIC:**	-739.4
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5180	0.016	32.847	0.000	0.487	0.549
yr	0.2461	0.010	25.344	0.000	0.227	0.265
holiday	-0.0821	0.031	-2.633	0.009	-0.143	-0.021
windspeed	-0.2353	0.030	-7.972	0.000	-0.293	-0.177
season_Summer	0.0445	0.015	2.989	0.003	0.015	0.074
season_Winter	0.0838	0.016	5.089	0.000	0.051	0.116
mnth_August	0.1026	0.019	5.303	0.000	0.065	0.141
mnth_December	-0.1664	0.020	-8.256	0.000	-0.206	-0.127
mnth_February	-0.2075	0.022	-9.457	0.000	-0.251	-0.164
mnth_January	-0.2759	0.020	-13.872	0.000	-0.315	-0.237
mnth_November	-0.1127	0.022	-5.100	0.000	-0.156	-0.069
mnth_September	0.1209	0.020	6.071	0.000	0.082	0.160
weathersit_Light_Rain_Snow	-0.3098	0.029	-10.542	0.000	-0.368	-0.252
weathersit_Mist_Clouds	-0.0935	0.010	-8.983	0.000	-0.114	-0.073

Omnibus:	54.354	**Durbin-Watson:**	1.943
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	85.766
Skew:	-0.708	**Prob(JB):**	2.38e-19
Kurtosis:	4.425	**Cond. No.**	9.20

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

A: Based on the final model, this was my formula for the best-fitted line:

cnt = (0.2461 x *yr*) - (0.0821 x *holiday*) - (0.2353 x *windspeed*) + (0.0445 x *season_Summer*) + (0.0838 x *season_Winter*) + (0.1026 x *mnth_August*) - (0.1664 x *mnth_December*) - (0.2075 x *mnth_February*) - (0.2759 x *mnth_January*) - (0.1127 x *mnth_November*) + (0.1209 x *mnth_September*) - (0.3098 x *weathersit_Light_Rain_Snow*) - (0.0935 x *weathersit_Mist_Clouds*) + 0.5180

Now, based on the coefficients, my model's biggest 3 features are:

1. Year or *yr*, with a coefficient of +0.2461. This indicates that demand for bike rentals increased over time.
2. *weathersit_Light_Rain_Snow,* with a coefficient of -0.3098. This indicates that light rain/snow significantly reduces bike rentals.
3. Lastly, *mnth_January,* with a coefficient of -0.2759. This indicates that bike rentals dipped considerably in January, owing to the cold weather.

General Subjective Questions

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

A: There are 4 main steps in the linear regression algorithm.

1. First, we need to **formulate the hypothesis**. We will need to assume a linear relationship between a dependent variable, which can be y, and independent variables, which can be x1, x2, x3, and so on up to xn. The coefficients of each feature is represented by β. The linear relationship is represented by y^ = β0 + β1x1 + β2x2 + B3x3 + ….. + βnxn. β0 is intercept.
2. Then, we need to **define the cost function**. We can use Mean Squared Error (MSE) to measure the error between actual values, or y, and predicted values, or y^, minimizing average squared differences.
3. Now, we will need to **optimize coefficients** by applying gradient descent to update the coefficients one by one by reducing cost function.
4. Lastly, we will have to **train and evaluate** the model. We can evaluate performance by using several metrics, including R-squared. We need to validate assumptions and make sure there is no multicollinearity.

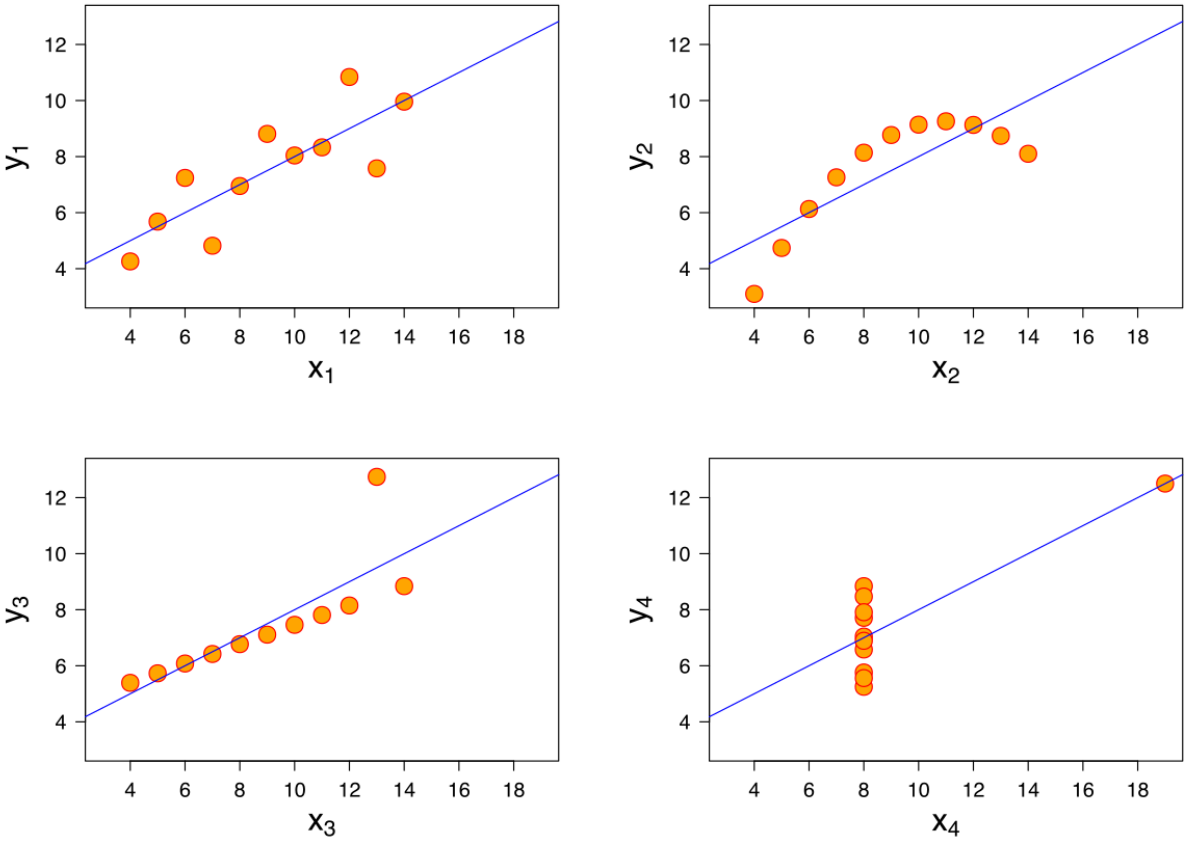
***2.** Explain the Anscombe's quartet in detail. (3 marks)*

A: The brainchild of British statistician Francis Anscombe, the namesake Quartet is a set of four datasets which have identical summary statistics, but look completely different. This means that statistics such as mean, variance, or correlation for four datasets could be similar, but the distribution of data points or the visual appearance of the four datasets will be completely different.

The purpose of Anscombe's Quartet is to remind statisticians and analysts that trusting or depending only on numbers and statistics shouldn't be the end-all solution - we must visualize the data to understand the data distribution and relationship.

The Quartet shows us the importance of elements like linearity, non-linearity, and outliers, which stresses how important exploratory data analysis (EDA) is.

Given below is a visual representation of Anscombe's Quartet:



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

The brainchild of yet another British mathematician, Karl Pearson, the Pearson Correlation Coefficient (popularly called Pearson's R and denoted by r) helps quantify the strength and direction of the linear relationship between two continuous variables.

Pearson's R always ranges from -1 to 1.

If r = -1 or is between -1 and 0, there is a negative correlation, which means if one variable increases, the other shall decrease and vice versa, as we saw in the case of heavy precipitation weather and bike rentals.

If r is between 0 and +1, there is a positive correlation, which means if one variable increases, the other shall also increase and vice versa. A simple example would be the more hours you put in for your workout, the more calories you lose (unless there's a freak anomaly).

If r is equal to 0, then there is simply no linear relationship between the two variables.

Of course, Pearson's R assumes that data is normally distributed and is linear.

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

A: Scaling is the process of transforming variables or features in a dataset to a common scale.

During machine learning, scaling is performed primarily to give all data an equal weightage. A dataset can contain two variables such as age and income. Now, while income might be a numerical of 5 or 6 or even more figures, age will hardly be more than 3 digits. So, to make sure that the model's performance is stable and valid, we must scale all data to a certain measure.

In normalized scaling, or Min-Max scaling, data is adjusted to a specific range, mostly 0 and 1. Standarized scaling, on the other hand, transforms the data so that it has a mean of 0 and an STD of 1. This means that while the former is sensitive to outliers, the latter is not as sensitive to them and might not be suitable for produce data within a fixed range.

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

A: As we know, Variance Inflation Factor, or VIF, helps us understand how much a variable is inflating the variance of the estimated regression coefficient due to correlations with other features.

So, VIF can become infinite when there is perfect multicollinearity between two or more independent variables in a dataset.

An infinite VIF means that at least one or more predictor variables must be removed, or combined with others so that redundancy can be reduced and model can be made more stable and reliable.

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

A: A Quantile-Quantile plot, or a Q-Q plot can help us understand if a dataset follows a normal distribution or not. It does so by comparing the quantiles of a dataset with the quantiles of a chosen theoretical distribution.

As per a Q-Q plot, if the data distribution is normal, the points will form a straight line.

It is generally used to check the assumption of normality of residuals in linear regression. If Q-Q plot generates a heavy-tailed curve, then that means there are outliers or extreme values present in the dataset. On the other hand, if it forms an ‘S’ shape, it means the distribution is skewed negatively or positively.

So, Q-Q plot is a handy tool to understand the normality of residuals, and in turn, the reliability of the model.