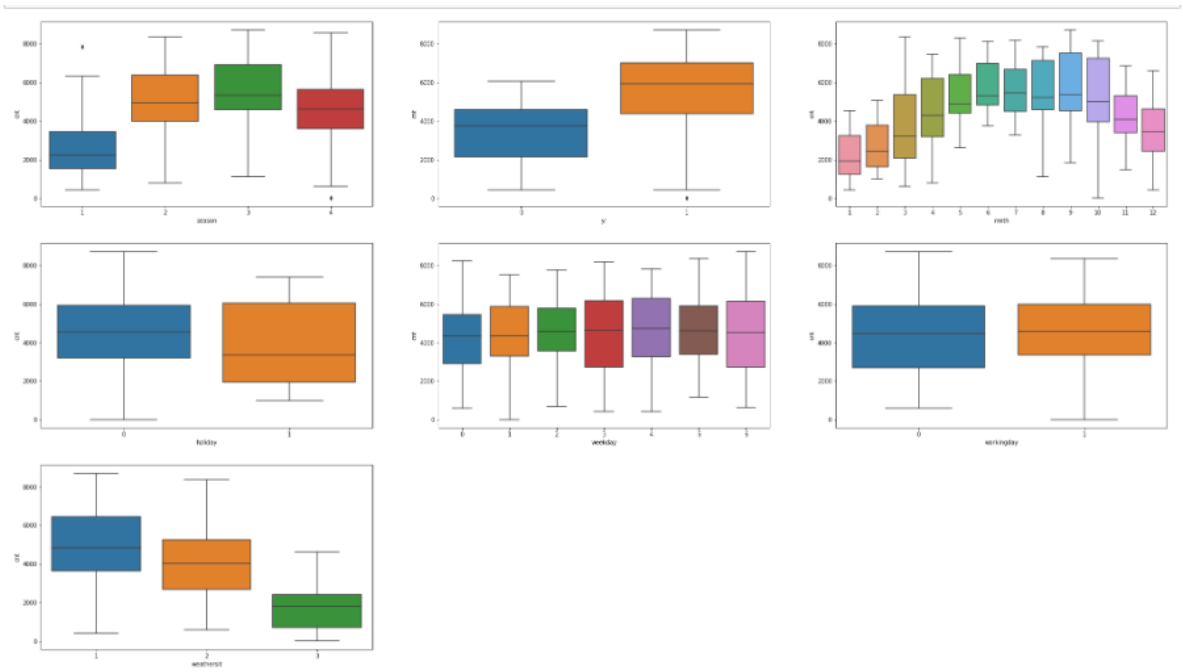
Assignment-based Subjective Questions

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

Answer:



1. Season Vs Cnt (Usage) : From the EDA analysis it can observed that as the season change from Winter to Spring there is a great increase this slightly increases in summer and drops down slightly in Fall. In general, this suggests seasons does have an effect in usage of bikes.
2. Year Vs Cnt (Usage) : As the year moves from 2018 to 2019 this increased usage of the bikes, this can be due to increased popularity.
3. Month Vs Cnt(Usage) : This reflects the information presented through Season vs cnt as this suggest as the month moves from March there is a steady increase until September then it falls down until the next March.
4. Holiday Vs Cnt(Usage) : Median for non-holiday is higher with closer upper and lower quartiles suggesting there might be slight indication that people prefer to use the bikes during non-holiday peiod.
5. Weekday Vs Cnt(Usage) : As the weekday goes into working days Wednesday and Thursday there is a slight increase in median , showing a slight indication that people prefer to use it during mid week more than weekends
6. Workingday Vs Cnt(Usage) : this correlates with holiday and weekday outcome where slight indication that people prefer to use bikes during working days, but not enough to conclude as a critical trend.
7. Weathersit Vs Cnt(Usage) : when the weather is Clear, Few clouds, Partly cloudy, Partly cloudy the usage is at its highest, when the weather worsens to Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds the usage drops same as it worsens to Light Snow, Light Rain +

Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. There is no usage as it turns to Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog. This shows the usage is dependent on weathersit.

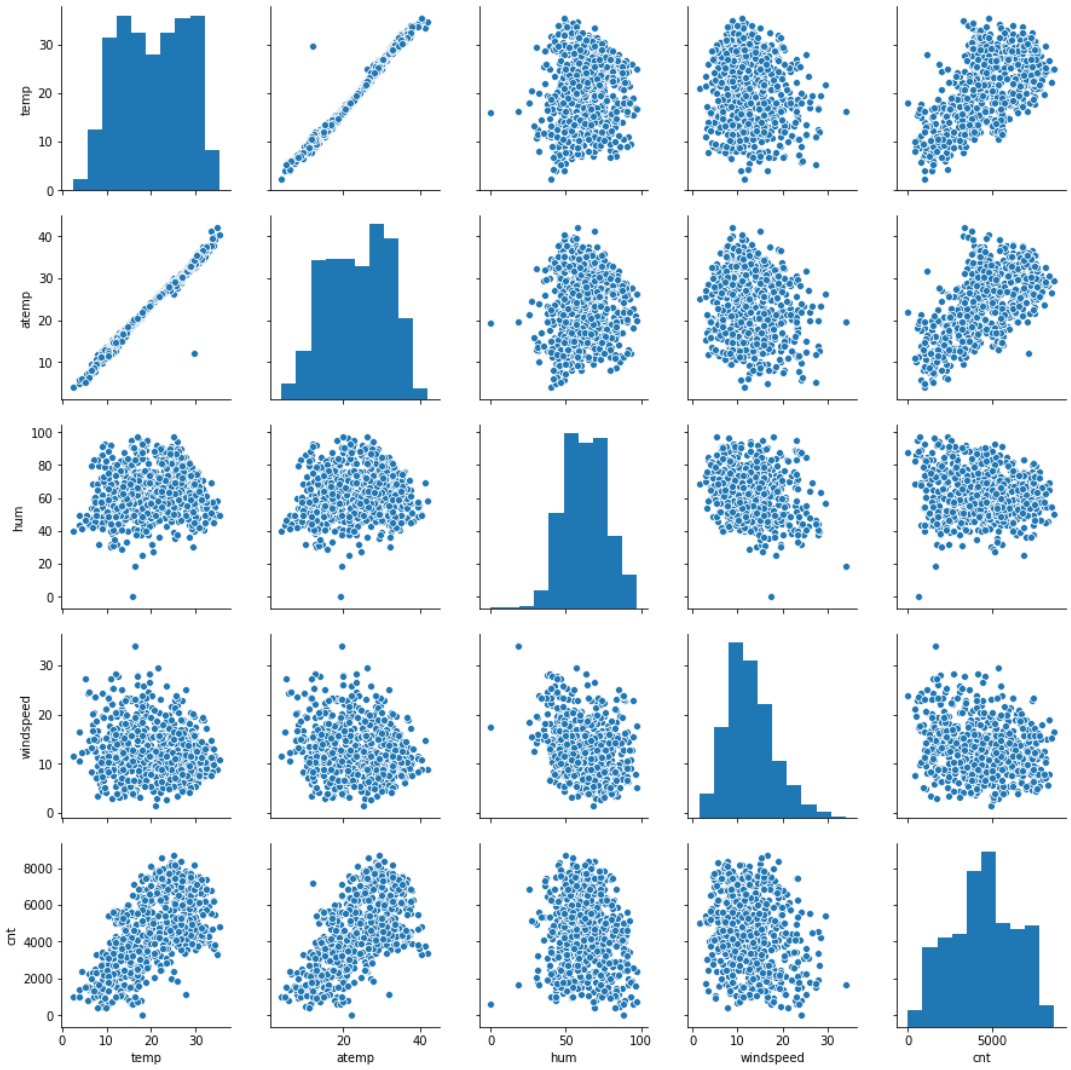
2. Why is it important to use drop_first=True during dummy variable creation?

If first variable not dropped then this can lead to dummy variable trap. This happens when the independent variable are multicollinear. When new dummy variable is created, each categorical variable will be represented using 0 or 1 using the one hot encoding method. 0 represent the absence of it and 1 represent the present of it. A simple explanation is when two or more variables are highly correlated; one variable can be predicted by the outcome of the other variables. For example take Season from this data set, one hot encoding will turn this into

Spring	Summer	Fall	Winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

By removing the Winter we can still predict the outcome, absence of all three spring, summer and fall would suggest that its winter. If not dropped, it will make it harder to predict individual effect on prediction model due to multicollinearity.

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



It can be observed that there is a high correlation between 'temp' and 'atemp' against target variable 'cnt'. Also a linear relationship between 'atemp' and 'temp' can be observed as well. 'Windspeed' and 'hum' does not have any visible relationship with any of the other numerical variables through pariplot.

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

Assumption 1: **Linearity**: Relationship between the dependant variable and independent variable to be linear

Pair plot shows the linearity between dependant and independent variable model to be linear. Then it was confirmed using r2, which was 0.822517 in this model. R2 explains the difference between variance explained by the model divided by the total variance, the value was closer to 1 hence there a clear indication of linearity.

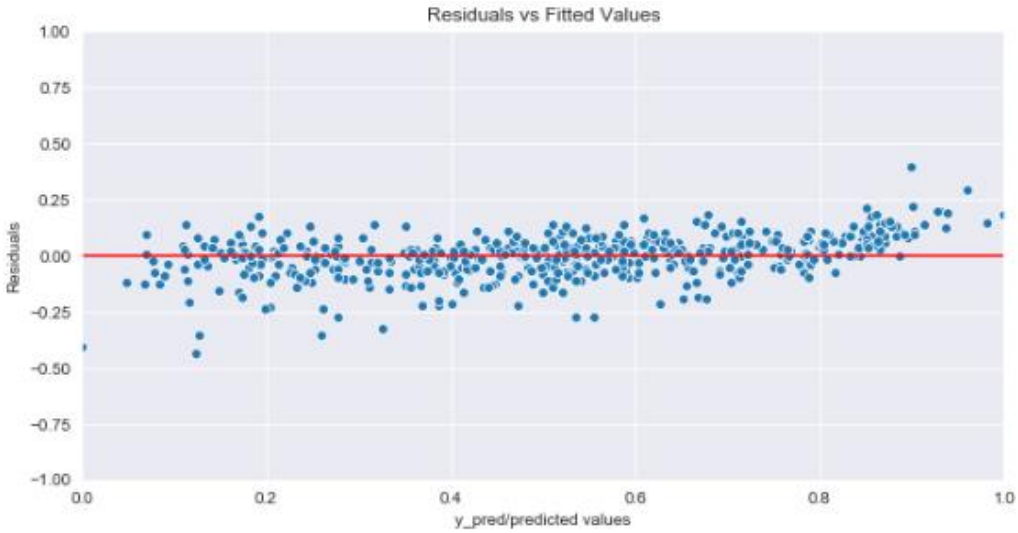
Assumption 2: **Mean of Residuals**: Residual is the difference between the actual value and predicted value, the assumption is the value should be 0. In the designed model value is close to 0, hence proving the model.

```
Mean Residual for the Training Set :
-3.4792430374650126e-16
R-squared for the Training Set :

: 0.8225173906649428
```

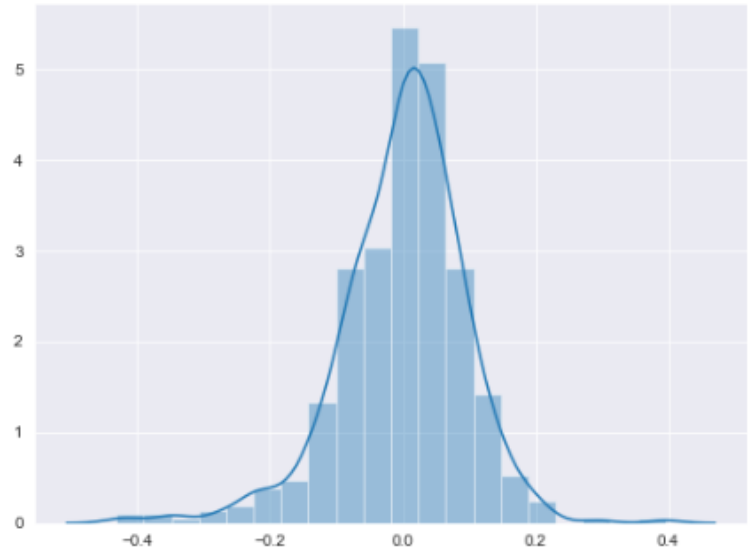
Assumption 3: **Homoscedasticity**: Homoscedasticity means the residuals have a equal or almost equal variance across the regression line.

By plotting, the residual it is observed as shown in the image below the data is equally distributed across both side of the 0 point.

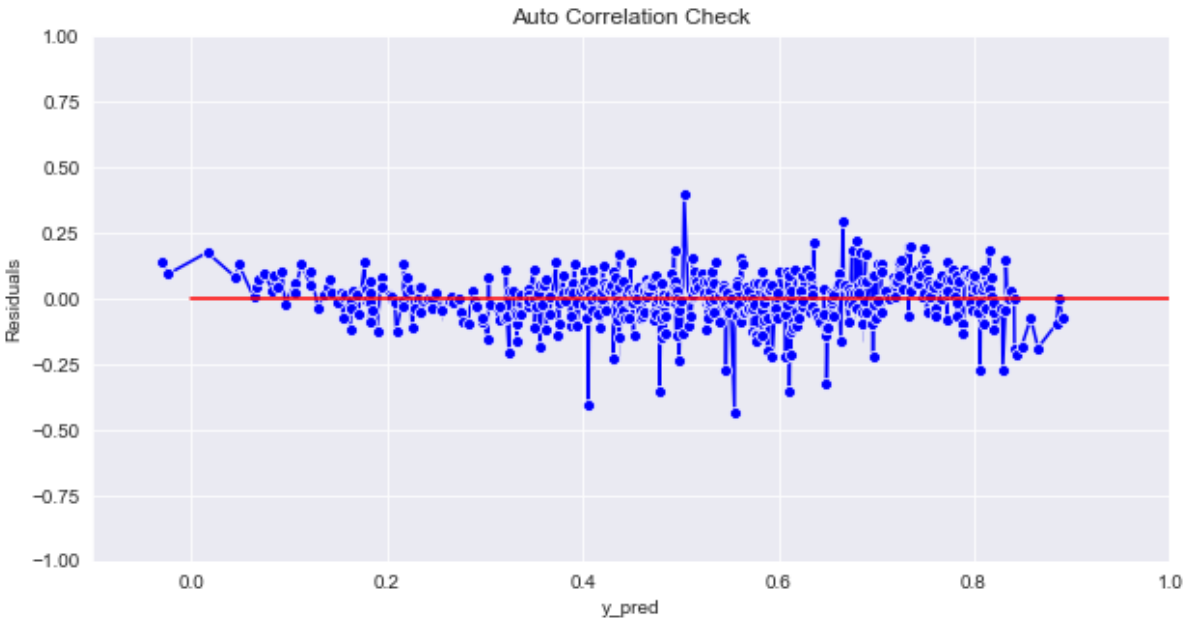


Assumption 4 : **Normality of error terms/residuals** : Residual terms are normally distributed and meets the central limit theorem, it will also be centred around 0.

This can be observed in the image below from the model and the curve is narrower.



Assumption 5 : **No Residual Autocorrelation** : When there is unexplained pattern in the Y variable this can show up in Y variable. Which means the current value is dependent on the previous one. From the image below it can be seen that it did not form any pattern.



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

Top three features from the model would be as follows:

- Temp : Temperature
- Light Snow : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- Yr : year

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1938	0.022	8.750	0.000	0.150	0.237
yr	0.2330	0.009	27.300	0.000	0.216	0.250
holiday	-0.1011	0.027	-3.740	0.000	-0.154	-0.048
temp	0.4712	0.031	15.063	0.000	0.410	0.533
Spring	-0.1107	0.016	-7.051	0.000	-0.142	-0.080
Winter	0.0558	0.013	4.393	0.000	0.031	0.081
July	-0.0688	0.018	-3.824	0.000	-0.104	-0.033
September	0.0658	0.016	4.042	0.000	0.034	0.098
Light Snow	-0.3002	0.026	-11.768	0.000	-0.350	-0.250
Mist	-0.0796	0.009	-8.752	0.000	-0.097	-0.062

Omnibus:	66.905	Durbin-Watson:	2.010
Prob(Omnibus):	0.000	Jarque-Bera (JB):	173.340
Skew:	-0.661	Prob(JB):	2.29e-38
Kurtosis:	5.532	Cond. No.	13.2

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is an analysis of fitting a straight line to the data, this is a popular known technique to model and understand real world problems to predict them. It falls under the supervised learning methods, where previous data used to build the model and commonly used as a predicative analysis model. In laymen's terms it can explained as a relationship between a dependent variable (y) and an independent variable (x) using a straight line.

The best fit of the line derived by minimising the residual sum of squares; this is sum of squares of the residual from each data position in the regression plot. Residual is the predicted value of dependant variable subtracted by the actual value of the dependant variable.

Equation for linear regression taken from straight-line equation, which is

$$y = mx + c$$

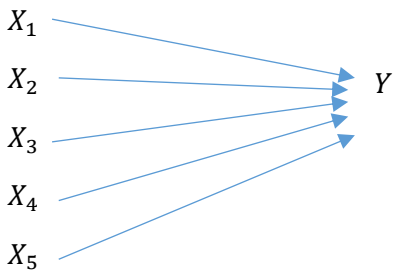
Rewritten to the following for univariate linear regression

$$Y = \beta_0 + \beta_1 X$$

Where β_0 is the intercept and β_1 is the slope.

As shown above univariate linear only involves one variables, but when it comes to multivariate linear regression there are more than one predictor variables so the equation can be re-written as following

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \ldots \ldots \ldots . + \beta_n X_n$$



This shows the outcome of the multivariate linear regression to be the combination of all variables (X).

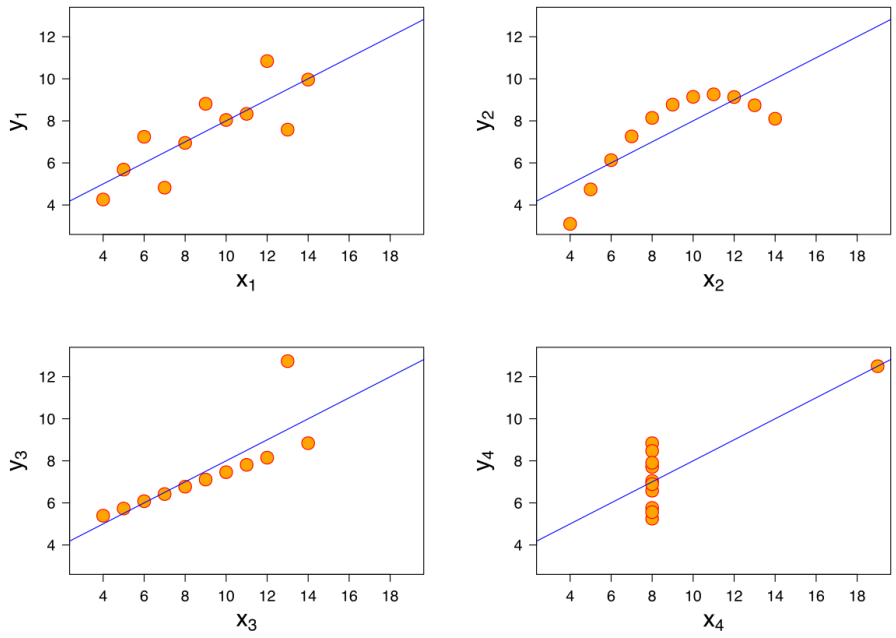
Validity of the model assessed by checking if the assumptions have been which as

- Linearity
- Mean of Residuals
- Homoscedasticity
- Normality of error terms/residuals
- No Residual Autocorrelation

2. Explain the Anscombe's quartet in detail.

Francis Anscombe illustrated that four sets of data, which got the same statistical observation, which provides same statistical observations can provide the similar information such as variance, mean on all four data sets. However, when plotted they show completely different pattern to the data, which shows the important of plotting. See the example below where all statistical information matches while

1. Graph 1 : shows Linear regression model fits pretty well with the data
2. Graph 2 : this could not fir the model and shows data is non linear
3. Graph 3 : There are outliers on the data which cannot be handled by the linear model that well
4. Graph 4 : shows the data not distributed and outliers cannot be handled by the model that well.



3. What is Pearson's R?

Pearson's correlation co-efficient is method of measuring linear correlation. This will summarize the behaviour of the data by measuring the strength and direction of the relationship between two variables; it is a number between -1 to 1. Below table show the relationship can be read

Pearson's R	Direction	Strength
Greater than 0.5	Positive	Strong
Between 0.3 to 0.5	Positive	Moderate
Between 0 to 0.3	Positive	Weak
0	None	None
Between 0 to -0.3	Negative	Weak
Between -0.3 to -0.5	Negative	Moderate
Less than -0.5	Negative	Strong

It is the ratio where the two variables vary together divided by two variables vary separately. One of the example used in this project is the pairplot from sns library. The equation for pearsons r shown below

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

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

Scaling is when normalising a feature and putting it into a same range as the other feature. If the scaling is not performed this could have an effect on the co-efficient of the model and also makes it hard to interpret the data.

Normalized scaling is when the value is mapped between 0 and 1. It is also knows as min-max scaling. The equation for the min-max scaling shown below

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardized scaling is when the data is not enforced to a specified range instead this is transformed to have a mean 0 then also a standard deviation of 1 this can be represented using the equation below.

$$z = \frac{x - \mu}{\sigma}$$

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

This happens when there is a perfect correlation between the two, which is equal to 1. Since the equation for VIF is as shown below

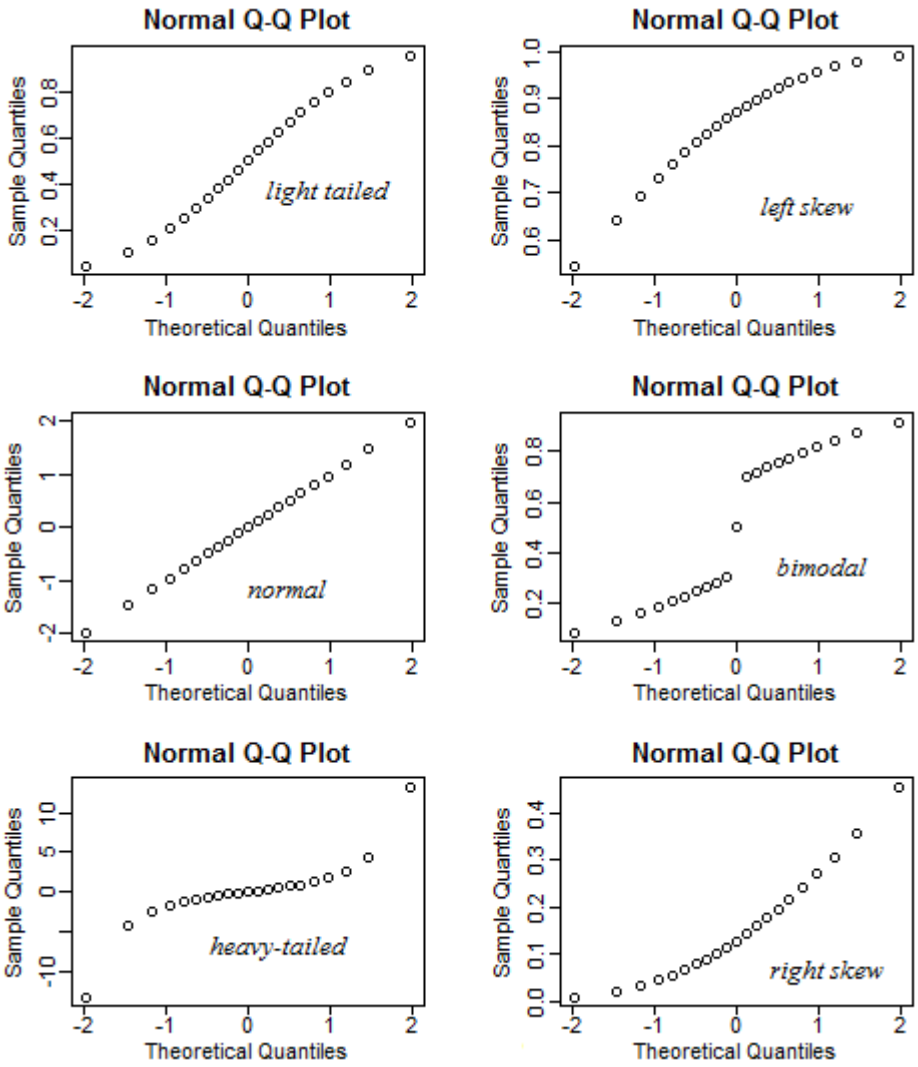
$$VIF = \frac{1}{1 + R^2}$$

When the r2 is equal to 1, VIF value will become infinite. This can also happen if one of the dummy variable is not removed from the table this will introduce a collinearity in the model. To overcome this the columns must be removed to optimize the model.

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

Q-Q plot used to graph the compare the theoretical against the sample distribution. It will plot quantiles of the data versus the quantiles of the distribution hence known as Q-Q plot. From this, we can determine if the data follows any certain pattern such as normal, uniform or if it is exponential.

Each of the behaviour can be explained to determine the model. For example if the Q-Q plot is Light tailed it shows that there are more data located at the extremes of the distribution and less in the centre. Left skew suggest that there is longer tail to left so best bit fit line falls below (vice-versa for the right skew)



References:

https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot