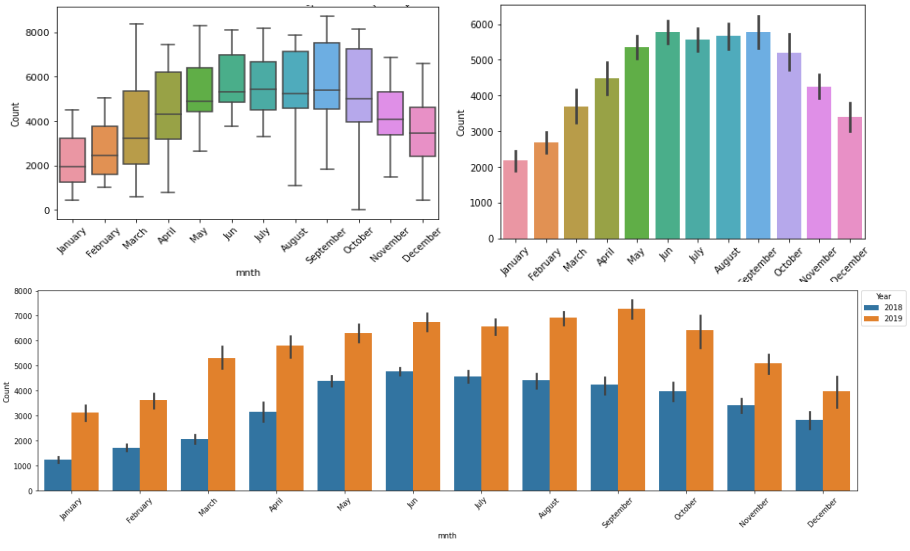
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

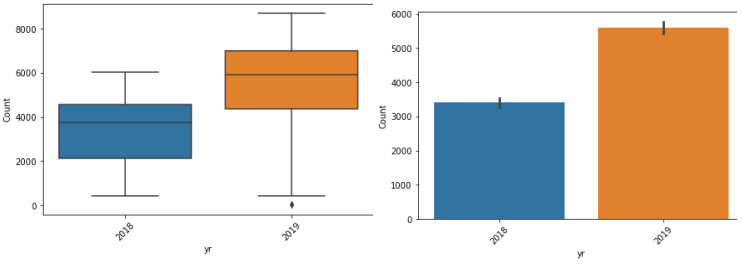
Ans: mnth: a. There is an increase in user counts during may - September: it can be due to the holiday season

b. The pattern still holds in both years though the number of users in 2019 increased from 2018. - Popularity increase might have caused more users to join in 2019.

c. March and September seem to have received a higher percentage of growth compared to other months in the number of users.

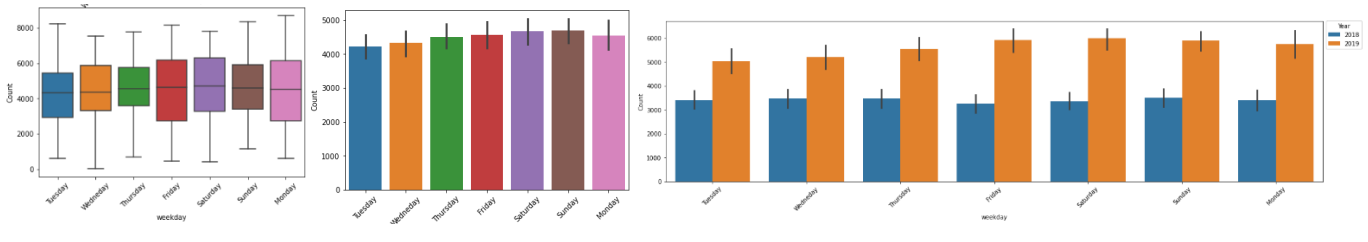


yr: a. There is an increase in usage from 2018 to 2019: as popularity increases it is set to increase the usage

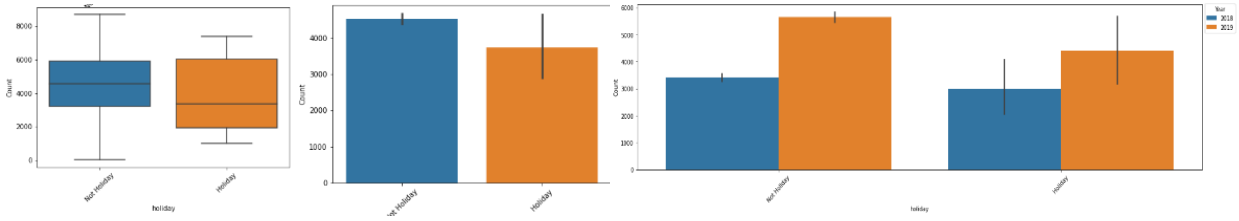


weekday: a. There is not much variation as per weekdays: usage is constant as users can be using the bike constantly

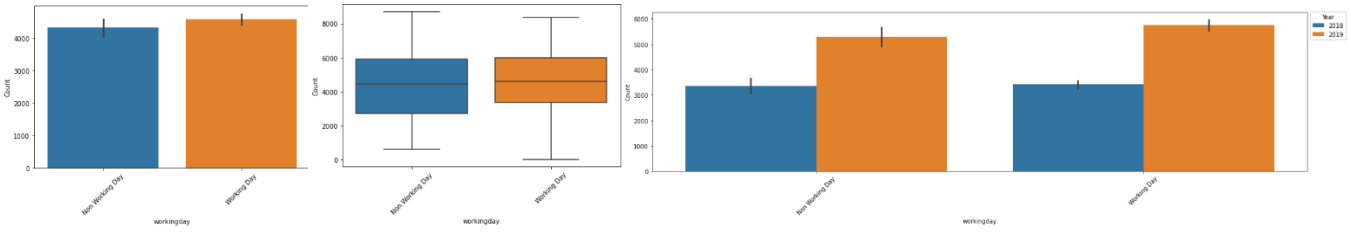
b. The pattern still holds in both years, though the number of users in 2019 increased from 2018. - popularity increase might have caused more users to join in 2019.



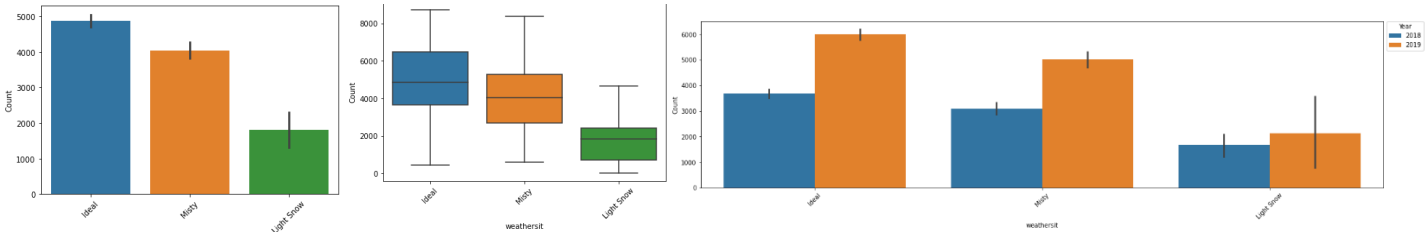
holiday: a. Usage is more during non-holidays: usage seems to be for users commuting to work/studies
b. The pattern still holds in both years, though number of users in 2019 increased from 2018. - Popularity increase might have caused more users to join in 2019.



workingday: a. Working days see more usage marginally: it seems commuting is more used for this bike rentals
b. It is not much difference in usage count: Showing the usage is almost constant
c. The pattern still holds in both years, though number of users in 2019 increased from 2018. – popularity increase might have caused more users to join in 2019.



weathersit: a. deal conditions and misty has the most usage: Bike usage is much better during clear or near clear conditions rather than during snowy weather
b. The pattern still holds in both years, though number of users in 2019 increased from 2018. – popularity increase might have caused more users to join in 2019.
c. The percentage change in users increased in 2019 is more for Ideal and Misty situation as compared to lighly snow - The difference in percentage seems to be due to more understanding of the use case with experience.



Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans: It is important to use drop_first=True during dummy variable creation as it reduces the complexity of the data and helps to the algorithm to converge faster. It's just for convenience.

Example: Let's assume we have a dataframe with 3 unique levels, say Level_1, Level_2 and Level_3.

If we keep drop_first=False, then all the three levels will be included as dummy variables.

On the other hand if we will have drop_frist=True, it will only have 2 levels, which can anyway justify the 3rd level, say Level_3 got dropped, then if either Level_1 or Level_2 are 1, it will say we are talking about those levels, in case both are 0 then it is definitely is Level_3, thus redundancy can be avoided.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

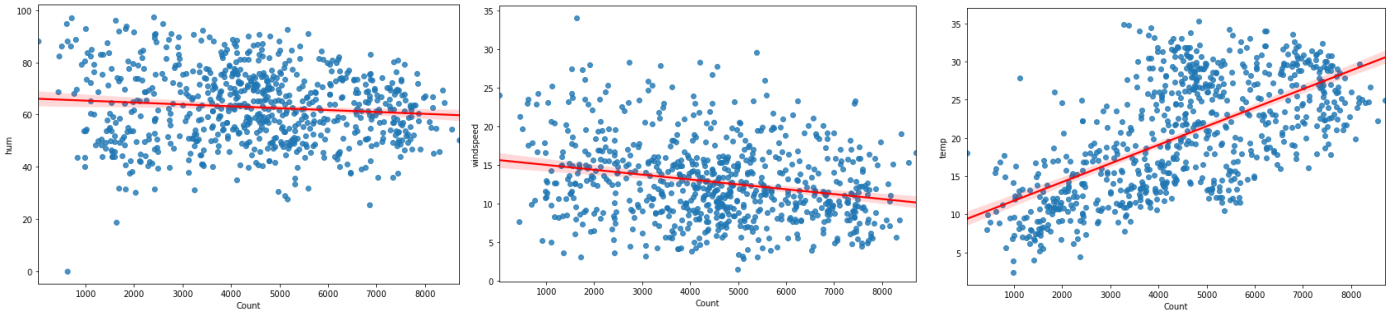
Ans: temp and atemp have the highest correlation with the target variable (0.63).

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

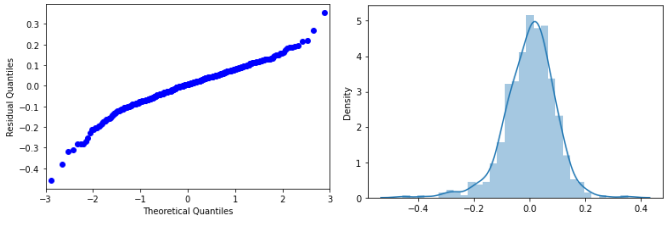
Ans: Assumptions are:

1. Linearity

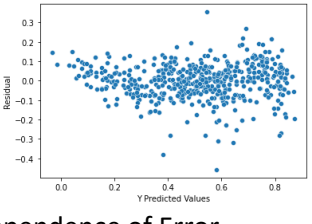
   Give the linearity of the data a confirmation - we can use linear regression on it **cnt** increase as:
   - **temp** increases
   - **hum** decreases
   - **windspeed** decreases



2. Normality of error
   1. qqplot roughly follows a straight line - proving normality
   2. We can see mean of residual is -6.62^-16, i.e almost zero and SD is close to 1



3. Homoscedasticity
   1. By plotting a scatter plot between predicted values on x-axis and residual values on y-axis
   2. The data is following homoscedasticity



4. Independence of Error
   1. By plotting a scatter plot between predicted values on x-axis and residual values on y-axis
   2. The Graph above proves the independence of error too.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: The top 3 features contributing significantly are: 1. Temp(0.4026), 2. Light Snow(-0.2949), 3. Yr(0.2348).

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Linear regression algorithm is a kind of supervised learning algorithm, which tries to find the best fit line between the independent and dependent variables, i.e., the relationship between them.

Linear Regression is of 2 types:

1. Simple Linear Regression

   1 independent variable and the model has to find it's relation with the dependent variable.

   Equation:

$$y = b_0 + b_1 x$$

   y: dependent variable

   x: independent variable

   b_0: intercept

   b_1: slope/coefficient

2. Multiple Linear Regression

   More than 1 independent variable and the model has to find it's relation with the dependent variable.

   Equation:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_i x_i$$

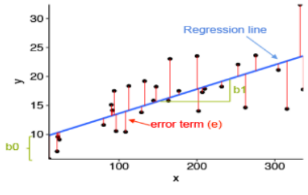
   y: dependent variable

   x_i: i^{th} independent variable

   b_0: intercept

   b_i: i^{th} slope/coefficient of the independent variable x_i

Linear regression model aims at finding the best fit linear line as well as optimal values of the intercept and coefficients such that the error(difference between the actual and the predicted value) is minimized.



The graph above shows a good insight into linear regression,

x: plotted on x-axis is the independent variable.

y: plotted on y-axis is the dependent variable.

Black dots: represents the actual data points.

b_0: is the intercept, which is 10 here

b_1: is the slope of the x variable

blue line: the best fit line predicted by our model(regression line).

Vertical distance between each point and the regression line: error or residual

Sum of all these differences known as sum of residuals/Errors.

We have to minimize this sum of residuals, to do that we find the squares and sum them and minimizes it.

In other words, minimize the rss, residual sum of squares:

$$rss = \sum(Actual\ Value - Predicted\ Value)^2$$

There are a few basic assumptions for linear regression:

1. Linearity: It states that the dependent variable should be linearly dependent on the dependent variables.
   This can be checked by plotting a scatter plot between the dependent and independent variables.
2. Normality of error: This states that the error follows a normal distribution.
   Can be tested using a qqplot and ploting a histogram of the residuals.
3. Homoscedasity: This states that the variance of the error terms should be constant for all the values of x.
   Can be tested by plotting the residual vs fitted values, and see that there should be no cone shape and be spread across constantly.
4. Independence of erros: This states that the error values should be independent of the independent variables.
   This is easily tested by the variance information we receive above.

These assumptions being satisfied signifies that out model is working on the dataset and the liner regression was apt for this dataset.

A few evaluation metrics useful for linear regression are:

1. R squared or Coefficient of Determination: Value lies between 0 and 1, value closer to 1 signifies better model.

$$R^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

2. Adjusted R squared: Value lies between 0 and 1, a value closer to 1 signifies better model, this is usually smaller than R squared values. This is a much better metric when it comes to multiple independent variable as it has a a penalty associated with adding extra variable, thus making the metric much more reliable.

$$R^2{}_{adjusted} = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

3. Mean Squared Error(MSE): It takes in the mean of sum of squared difference of the actual and predicted values.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

4. Root Mean Squared Error(RMSE): It is the root of MSE value. RMSE penalizes large errors, which MSE fails to.

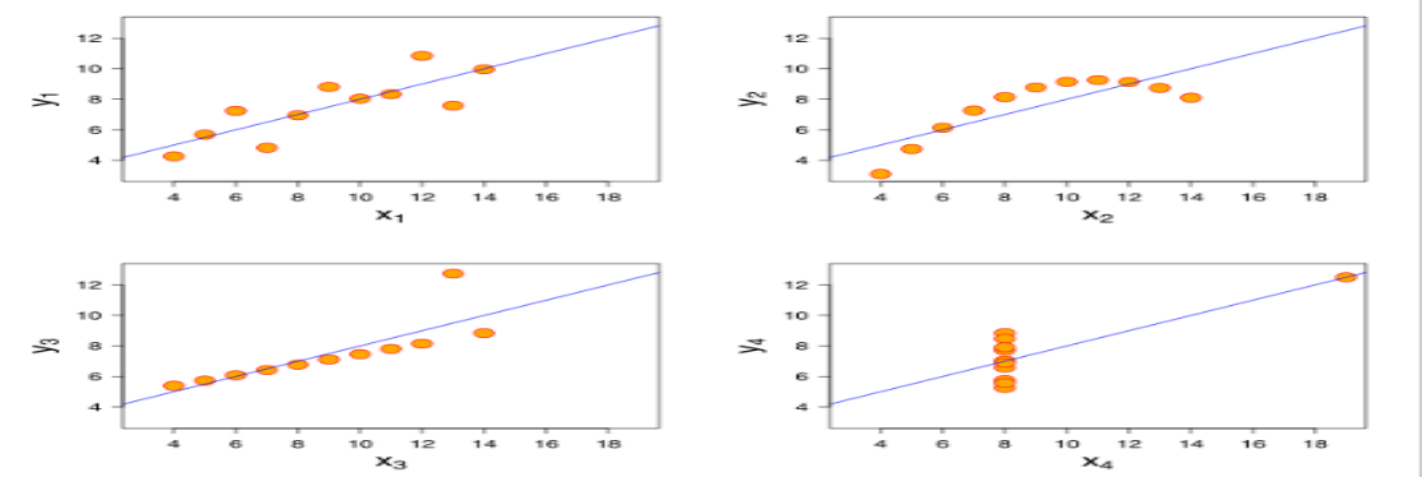
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Q2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet are a set of 4 nearly identical simple descriptive stats, but have a very different distribution which can be seens once graphed. It is an important display to understand the importance of graphs and the effect of outliers and other influential observations on statistical properties.

Anscombe's quartet Data Example							
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

Statistics for data	
Properties	**Values**
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125 ± 0.003
Correlation between x and y	0.816
Linear Regression Line	y = 3.00 + 0.500x
Coefficient of determination of linear Regression	0.67



As we can clearly see, even though statistically all 4 datasets are similar, the graphs show the true picture of what is actually happening behind the scenes with the datasets.

x1, y1 – shows an approx. simple linear relationship.

x2, y2 – shows a relationship that is clearly not linear, but y can be estimated using x through some non-linear function.

x3, y3 – shows a linear relationship, but clearly, the regression line can be better to accommodate the outlier.

x4, y4 – clearly shows how one outlier can cause a high correlation value even though all other points seem to be not correlated at all.

Q3. What is Pearson's R?

Ans: Pearson's R is one of the most common method used to calculate the linear correlation.

It is a value between -1 and 1, denoting the strength and the direction of the relationship of the two variables.

It is denoted by r,

r > 0 – positive correlation – both moves in the same direction.

r < 0 – negative correlation – both the variables move in the opposite direction

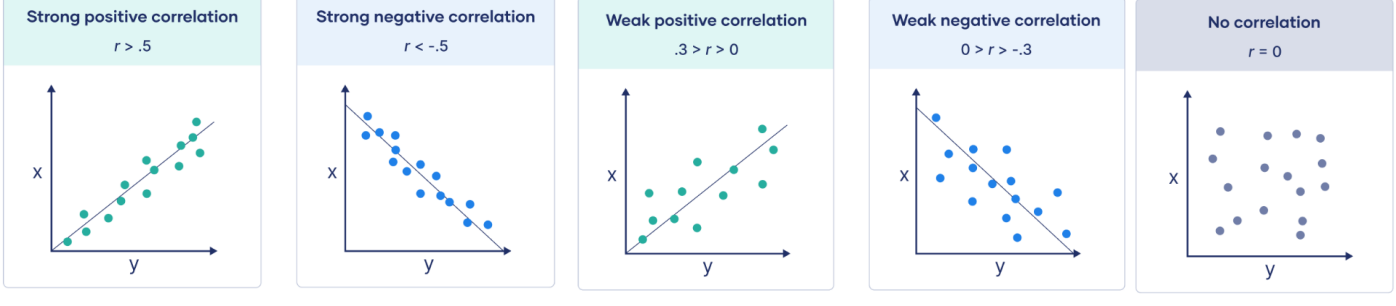
r = 0 – no correlation – no relationship between the variables.

It is what is called a descriptive statistic, one that summarizes the relationship between the variables(both direction and strength).

It is also an inferential statistic, one that can allow us to test the significance of a hypothesis.

There are certain thumb rules to this value r,

r Value	Strength	Direction
r > 0.5	Strong	Positive
0.3 < r < 0.5	Moderate	Positive
0 < r < 0.3	Weak	Positive
0	None	None
-0.3 < r < 0	Weak	Negative
-0.5 < r < -0.3	Moderate	Negative
r < -0.5	Strong	Negative



Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is a step while pre-processing data, it is applied on independent variables and help us to normalize the data in a particular range.

It helps in speeding up an algorithm, for example speeding up gradient decent optimization.

It allows us to bring all variables to a common range and allows the algorithm to not only use the magnitude but also the units, thus reducing the risk of incorrect modeling. It only affects the coefficients, all other statistically important parameters are untouched. So we are able to get similar results but much faster.

There are two types of popular scaling techniques:

a. Normalized Scaling:

This type of scaling brings the data in range of 0 and 1 both included. In python this can be done using Min-Max Scaling from sklearn.preprocessing.MinMaxScale.

This is internally done using the following equation:

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

b. Standardized Scaling:

This type of scaling bring out all the values of the variable to their z score, i.e., brings all of them to their standardized normal distribution form with mean(μ) = 0 and standard deviation(σ) to 1.

In python sklearn.preprocessing.scale allows us to use this scaling.

This scaling is done using the following equation:

Standardization: $x = \frac{x - mean(x)}{sd(x)}$

One major disadvantage of normalization is that it loses information such as outliers in the data.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: There are some times when VIF values is infinity, before reasoning on it, lets first look at what VIF is mathematically

$$VIF_i = \frac{1}{1 - R_i^2}$$

So for VIF_i to be infinite R_i^2 must be 1, that denotes that there is a perfect correlation between two independent variables in question, thus will denote that the feature in question is perfectly explained by other variables, which will be not ideal for our model as it will introduce redundancy, we need to keep multicollinearity low in our model and this value causes an issue.

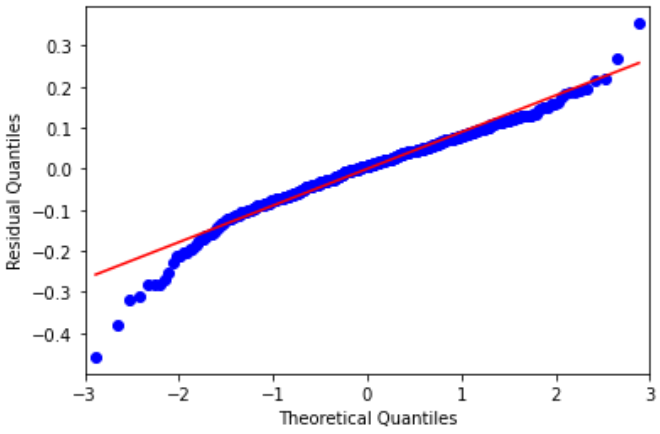
This infinite basically shows that this variable, can be perfectly represented by a linear combination of the other Variables. This is can be fixed by removing/dropping one of the variable from the dataset causing this perfect multicollinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: This Q-Q plot is used to check if two populations come from dataset that has a common distribution.

This plot specifically work with quantile information and compare the 0.3 quantile of one dataset to the same of the other. Also a reference line is also plotted along with the points.

Here is an example from the assignment,



If the two sets come from population that have same distribution, the points will fall approximately on this reference line.
The greater the deviation, more the proof that both the set come from population that have different distribution.

In linear regression, it can help us with verifying that the train and test dataset have same distribution,
also it allows for testing of one of the assumptions we make before starting with linear regression on a dataset, the normality of error, we use the residual vs our fitted values to see if the error follows a normal distribution throughout.