Questions & Answers

Section 1 : 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?**

Different kinds of plots were used to visualize the categorical variables. The following observations were made:

- The demand for shared bikes were higher in 2019 compared to 2018 and after analyzing the weather condition, it was clear that 2018 had worse weather leading to lower demand
- In 2018, most demand was during holidays that fall on weekdays, which could probably mean that people have dedicated more exercising time during their non-working day while in 2019, demand was high both during weekend and public holidays.
- There was a positive trend when it was Summer and fall seasons for both years, however the trend goes downwards in winter.
- The peak month for 2018 was June and demand started dropping drastically from September till December. That was because it was the winter season and the weather had snow, rain and thursderstom making it inconvenient for bike rides.
- The peak month for 2019 was September when the weather was clear even though it was a transition from fall to winter season.
- In 2018, there were no distinctive differences between weekends and workday, this could imply that some people actually use the bike as a means of transport to work.
- It can be deduced that weather conditions have a very big impact on the dependent variable because the demand drops the most in December even though it was a holiday in both 2019 and 2018. It was all due to the weather and winter season.
- Fall and Summer usually come with clear weather, that is why we can notice the demand increasing in both years during those two seasons.

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

When drop_first parameter is set to true, it removes the first column which is created for the first unique value of a column. It is important to do so because it reduces collinearity. Keeping all the dummy will simply introduce redundancy of one level.

Hence, it is better to drop one of the columns and for k levels, k-1 dummies (columns) would suffice.

It is possible to drop the first categorical variable because if the other dummy column is 0, it will simply mean that the first value would have been 1. In conclusion, this approach helps to reduce multicollinearity in the dataset, which is one of the prime Assumptions of Multiple Linear Regression.

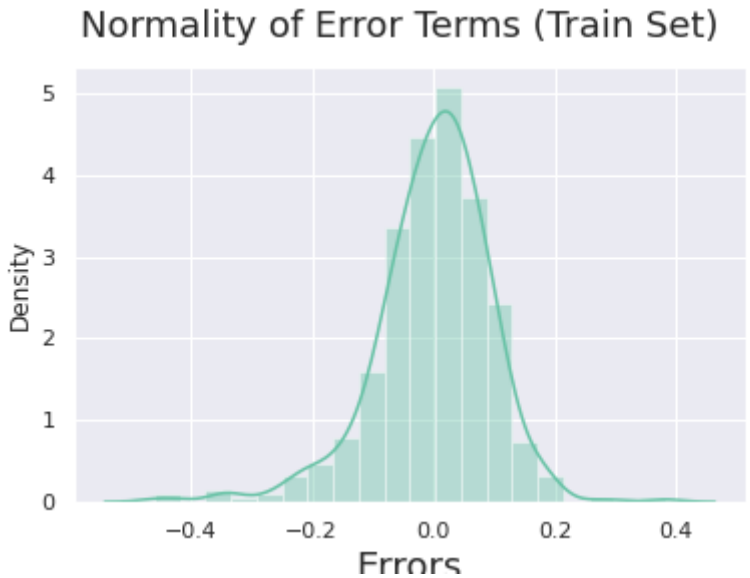
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

The independent variable ***atemp*** has the highest correlation with the target variable. The correlation coefficient of the latter with the ***cnt*** target variable is 0.63.

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

After building the model, the assumptions that can be verified are about the residuals of errors. There are mainly three assumptions and they were all validated by the means of graphical representation together with statistical information.

The first assumption is ***normality***, which states that the error terms should be normally distributed with mean zero.

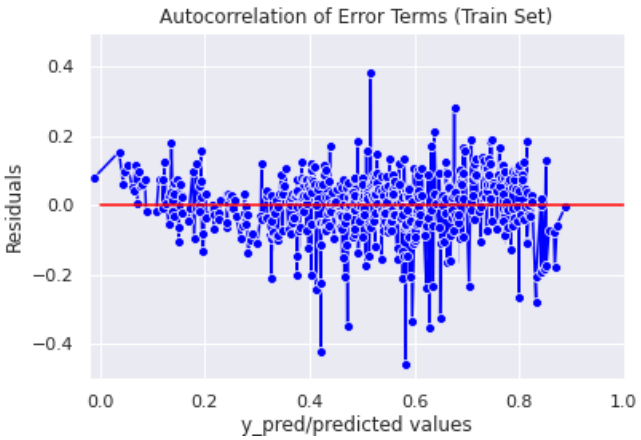


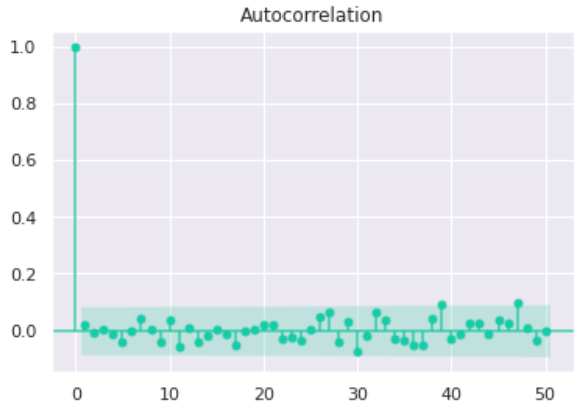
The plot represents the error distribution in the training set and the distribution plot shows a perfect normal distribution that has its mean at 0.

Moreover, we used the Anderson-Darling test. This test rejects the hypothesis of normality when the p-value is less than or equal to 0.05. The computed value we obtained is 5.718, which is clearly higher than 0.05.

Based on the plot and statistics, we concluded that the normality assumption has been met.

The second assumption is ***no autocorrelation***, which means that the error terms should be independent of each other.

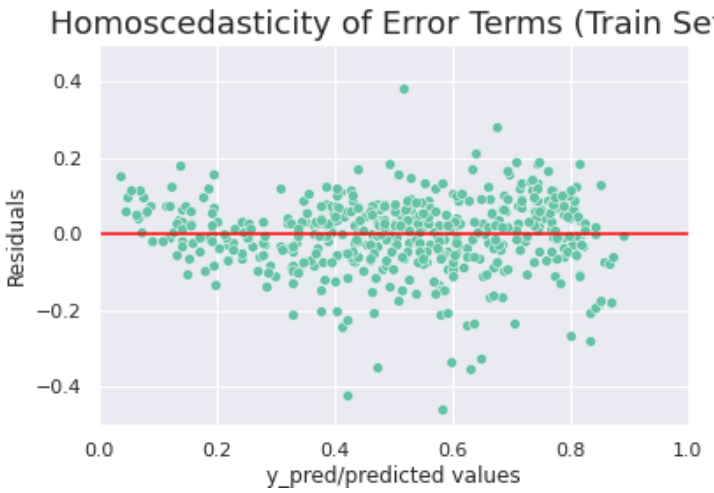




The first plot represents the autocorrelation of the error terms of the whole train set while the second plot presents partial autocorrelation using statsmodel. We can observe that there are no distinctive patterns from the first plot. Additionally, there are barely any spikes outside the red confidence interval region from the second plot.

The Durbin Watson statistic was also used to validate the autocorrelation assumption. The latter says that if the value is close to 2, it implies that there is little to no correlation. The value we obtained is 1.95, which is extremely close to 2, hence we can conclude that the assumption that there should be no autocorrelation between residuals of error has been satisfied

The third assumption is ***homoscedasticity***, which means that the error terms should have a constant variance.



From the above plot, we observe minor changes only as the error changes but nonetheless, the variance does not follow any pattern. The F-statistics and P-value was computed using goldfeld-quandt statistics and their values are: F-statistic: 1.0090037830352643 and p-value: 0.473111786766807.

As such, we can say that the third assumption is also met.

Given that this is a multiple linear regression, there is an additional assumption called ***No Multicollinearity*** that assumes that the independent variables are not highly correlated with each other. This assumption was verified using the Variance Inflation Factor (VIF).

	Features	VIF
0	temp	4.47
11	workday	3.17
2	Winter	2.59
5	2019	2.06
1	Spring	2.03
9	Nov	1.77
7	Jan	1.65
4	Misty & Cloudy	1.52
6	Dec	1.47
8	Jul	1.35
10	Sep	1.20
3	Light Snow or Rain	1.07

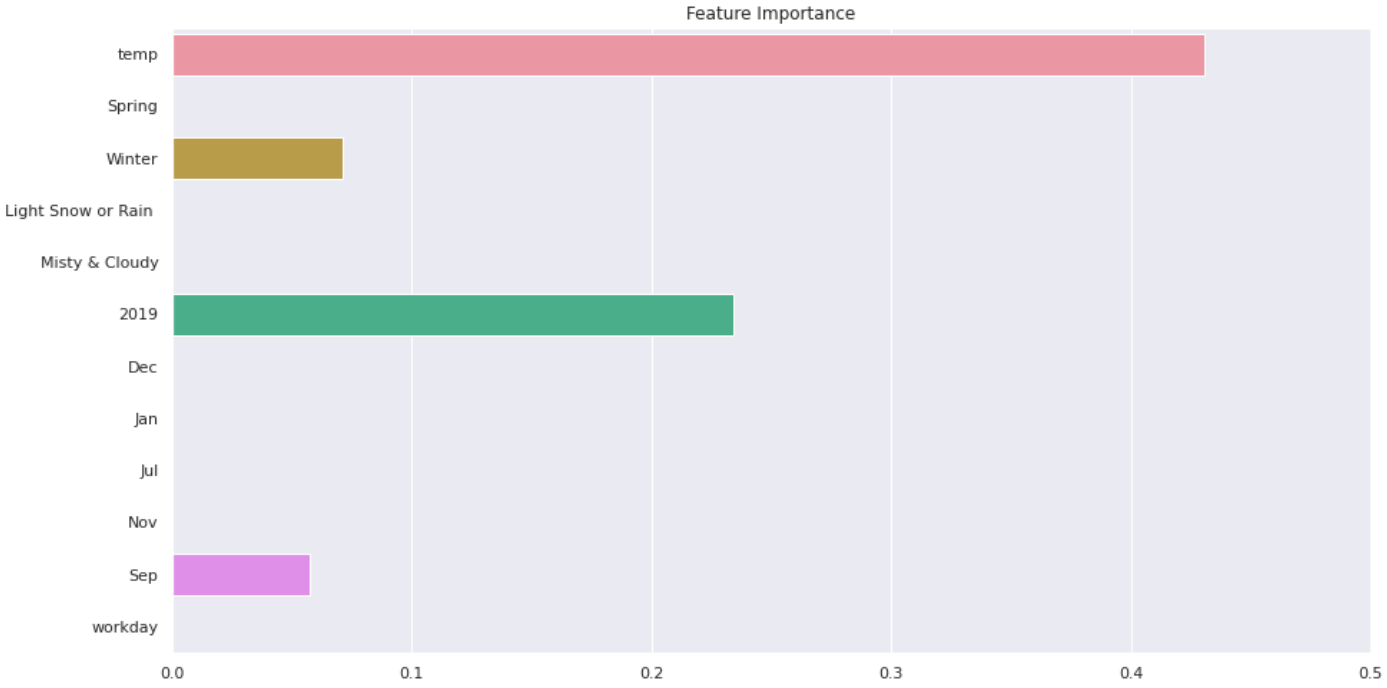
It can be observed that all the feature's VIF are within the acceptable range of less than 5. Hence we can conclude that the no Multicollinearity assumption has been satisfied.

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

The top three features that contributes significantly toward the demand of shared bikes are :

1. temp : with a coefficient of 0.431
2. 2019(yr) : with a coefficient of 0.234
3. Winter : with a coefficient of 0.071

The importance of each feature is measured by their coefficients. The higher the coefficient the more important the feature is. The feature importance of the final model is as illustrated:



Section 2: General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear regression is a statistical method used for predictive analysis. It is a model that assumes a linear relationship between the input variables (x) and the single dependent variable (y). In other words, a linear combination of the independent variables can be used to compute the target variable (y).

Linear regression is a supervised machine learning algorithm. It is used to predict a continuous value like Sales and the predicted output has a constant slope. There are two main types of linear regression:

1. **Simple Linear Regression** : It is used to determine the relationship between two quantitative variables (x and y).

   It uses the traditional slope-intercept form. It is represented as follows:

   $$y = \beta 0 + \beta 1X + \varepsilon$$

   - x represents the independent variable
   - y represents the predicted variable
   - β0 represents the intercept
   - β1 represents the regression coefficient
   - ε represents the residual error

   It identifies the line of the best-fit by searching for the regression coefficient that minimizes the residual error.

2. **Multiple Linear Regression**: It is used to determine the relationship between more than one independent variable (x) and only one dependent variable (y).

   The mathematical representation is:

   $$y = \beta 0 + \beta 1X1 + \ldots + \beta nXn + \varepsilon$$

   - y represents the predicted variable
   - β0 represents the y-intercept
   - β1 represents the regression coefficient of the first input variable X1
   - βnXn represents the coefficient of the last independent variable.
   - ε represents the residual error

Multiple linear regression computes three statistics in order to find the best-fit line for each input variable. They are : The P-value, the t-statistic of the model and the coefficients with the least error

For both types, the regression coefficient indicates the effect that the independent variable(s) has on the predicted value when the latter is increasing. The model error indicates the amount of variable in the model estimate of y.

The most common technique used to train linear regression is Ordinary Least Squares (OLS).

The relationship between the dependent and independent variables can be both positively or negatively linear in nature.

Assumptions of linear regression

Linearity: There is a linear relation between dependent and independent variables. That is the best fit line is a straight line.

Homoscedasticity: The error term is the same for all the values of independent variables and there should be no clear pattern distribution of data in the scatter plot.

Normality: The error terms should follow the normal distribution pattern, i.e, error terms should be normally distributed with a mean of 0.

Independence: The error terms should not be dependent on each other. That is there should be no hidden relationships between them, no autocorrelation. Otherwise, the accuracy of the model will lessen.

Small or no multicollinearity: This applies to Multiple linear regression only. Two or more independent variables should have very small to no correlation with one another. Otherwise this will impact the model in finding the relationship between the predictors and the target variable.

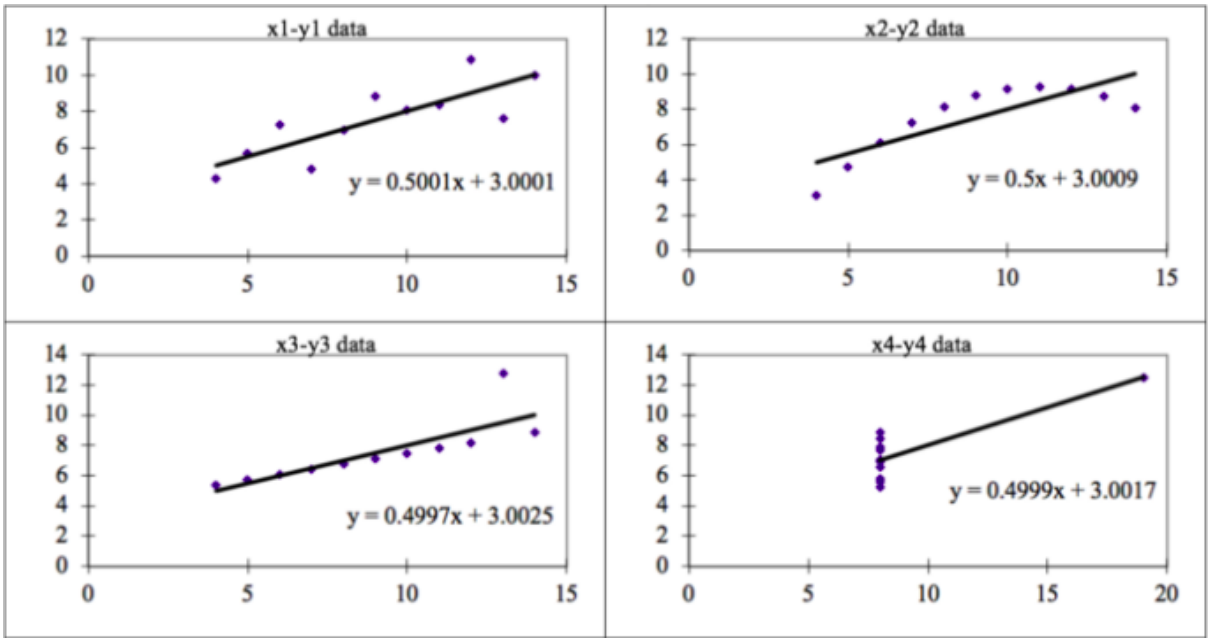
2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet consists of a group of four data sets that are nearly identical in simple descriptive statistics, however they have different distributions and can only be noticed when graphed. Anscombe's quartet illustrates the importance of data visualization before building a model. By plotting the data, anomalies in the latter can be detected.

The dataset to depict Anscombe's quartet is shown below, the statistical information of the four dataset has been computed (last section in the figure) and they are pretty much the same as illustrated:

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

The difference in the four dataset can only be spotted by using graphs:



The four datasets can be described as:

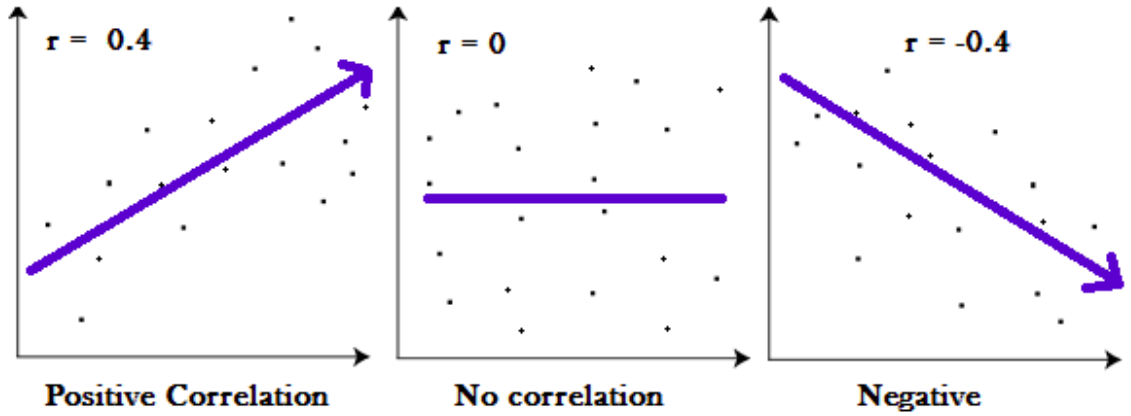
- Dataset 1: It has a set of (x,y) points that shows a linear relationship with some variance
- Dataset 2: The data appears to be non-linear in nature and this is shown by the curve shape.
- Dataset 3: The data appears to have a linear relationship, however there are outliers in the dataset which cannot be handled by a linear regression model
- Dataset 4: The dataset contains one outlier even though the remaining values of x are constant.

To conclude, the four datasets have almost the same statistical information in terms of variance and mean for each x and y point. The different sample distributions only become visible when the data is plotted. Hence, it is important to first visualize the dataset before interpreting or building a model with it.

3. **What is Pearson's R?**

Pearson's R is the most common correlation coefficient used in linear regression. Correlation coefficients measure the strength of the relationship between two variables, denoted by r. It draws the best fit line through the two data variables and it indicates the distance between the data point and the best fit line.

The Pearson's R values can range from +1 to -1. When there is no association between the two variables, the value is zero. On the other hand, a positive relationship is represented by a correlation coefficient of greater than 0, it simply means that as the value of one variable increases so does the value of the other variable. A negative relationship takes a value of less than 0, this implies that as the value of one variable increases, the value of the other variable has a negative decrease. It is shown in the diagram:



A correlation coefficient of +1 or -1 implies that none of the data points has any variation away from the line, i.e, all the points are on the best fit line.

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

Scaling is very important in the data preparation phase. It is applied on continuous predictor variables in order to normalize the data within a specific range. Scaling helps to accelerate the algorithm computation since it brings all the continuous variables to the same magnitude..

The need for scaling arises because most of the time the data varies in terms of range, units and magnitude. If scaling is not performed, the model might unnecessarily assign

higher importance to higher value, thus invalidating the model accuracy. Scaling only affects the variable coefficients.

There are two types of scaling that can be performed :

1. **Normalization** : It is also known as min-max scaling and it rescales the features' range to a scale of [0,1]. The formula is given below, where max(x) is the maximum value of the feature while min(x) is its minimum value:

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

2. **Standardization**: It is also known as z-score scaling. This technique centers the values around the mean with a unit of standard deviation. This implies that the mean of the attribute becomes 0 and the resultant distribution has a unit standard deviation. The formula is given below, where, $\bar{x}$ is the feature's vector average while σ is the feature's vector standard deviation:

$$x' = \frac{x - \bar{x}}{\sigma}$$

The table below summarizes the differences between normalization and standardization.

	NORMALIZATION	STANDARDIZATION
1.	Known as min-max normalization	Known as z-score normalization
2.	Scales value between [0,1] or [-1, 1]	Not bounded to a specific range
3.	Outliers have high impact	Outliers has lower impact
4.	Uses minimum and maximum value of features	Uses mean and standard deviation
5.	Used when features varies in scale	Used when zero mean and unit deviation wants to be ensured.
6.	Useful when we have no knowledge of the distribution	Useful when feature has Gausssian distribution
7.	Transformer called MinMaxScaler is provided by Scikit-Learn	Transformer called StandardScaler is provided by Scikit-Learn
8.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube	It translates the data to the mean vector of original data to the origin and squishes or expands.

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

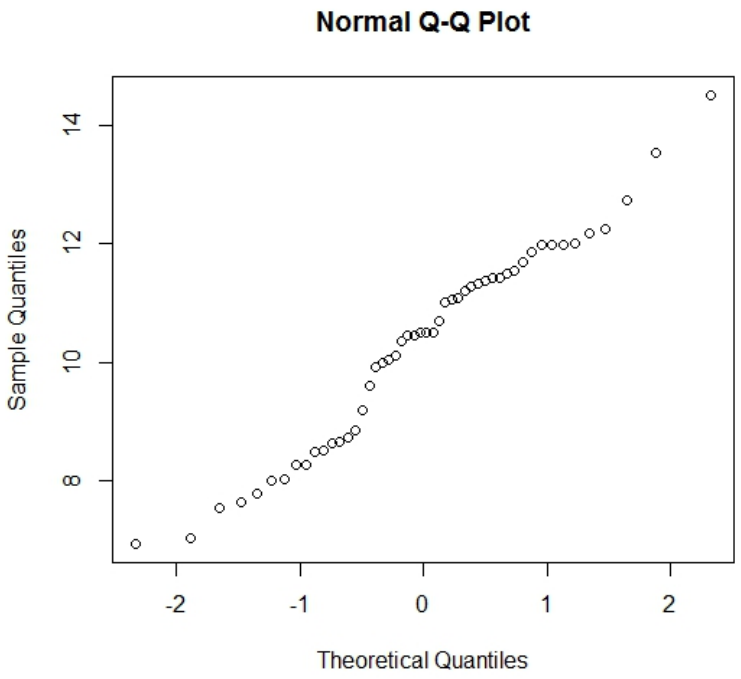
The Variance Inflation Factor measures the level of collinearity between independent variables in a multiple linear regression. It is computed by dividing the ratio of the variance of all the model's beta by the variance of one beta if it were to fit alone.

VIF is infinite when there is a perfect correlation between two independent variables. In such a situation, R-squared (R2) is equal to 1 which evaluated to $\frac{1}{(1-R2)}$, hence resulting in infinity. A VIF of infinity signifies that the variable expresses itself with the corresponding variable by linear combination.

To address this issue, one of the predictor variables causing this perfect multicollinearity has to be dropped from the dataset.

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

The quantile-quantile plot is a graphical technique that helps in determining the type of distribution of a random variable. A quantile is a fraction and some values can fall below that quantile. A Q-Q plot is basically a scatter plot that plots two sets of quantiles against one another. The data points should form a nearly straight line when both sets of quantiles come from the same distribution. An example of when both quantiles sets come from normal distribution:



A Q-Q plot is used to identify the type of distribution for any random variable be it normal distribution, uniform distribution, exponential distribution or pareto distribution and so on. The first data set is plotted along the x-axis while the second dataset is plotted along the y-axis. Generally, it is about normal distribution where the normal distribution is the base distribution and its quantiles are along the x-axis named as Theoretical Quantiles. While the sample quantiles are along the y-axis known as Sample Quantiles. This can be seen from the diagram above. Another use of Q-Q plot is to estimate parameters in a location-scale distribution family with the means of graphical representation.

Q-Q plot is important in linear regression because it helps to determine:

- If the two datasets come from population having the same distribution
- If the two datasets have common location and scale
- If error terms follow a normal distribution. This being an assumption of linear regression and can be verified using Q-Q plots.
- The distribution skewness