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?

Ans: From the analysis of the Categorical variables we found:

- Fall and summer seasons had the greatest number of bikes hired, followed by winter and the least by spring
- 2019 had the higher number of bikes hired leading 2018
- July, September has the highest number of bikes hiring least being during the fall season which is January and February
- The amount of bikes hired during holidays were slightly lower than working days
- The day of the week did not have much significance deciding the amount of bikes hired
- Clear weather was a contributing factor on the amount of bikes hired

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

Ans: Drop_first =True is important during dummy variable because it helps in reducing Multicollinearity. Multicollinearity occurs when two or more predictor variables are correlated with each other, meaning they are measuring the same thing, with Drop_first = True, The first dummy variable column is dropped, reducing the issue of multicollinearity.

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

Ans: Temperature and Feels like temperature(atemp) have the highest correlation with the Target Variable

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

Ans: Validation of the assumptions of Linear Regression was done as follows:

- Normally distributed: Error terms were normally distributed

- Multicollinearity : There were insignificant multicollinearity between the variables
- Linear model: There is a Linear Relation among the variables
- Homoscedasticity : There is no Visible pattern in residual values
- Independence of Residuals

5 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 towards explaining the demand of the shared bikes are:

1. Year - The Coefficient value of Year indicates, that for a unit increase in Year the bike hiring number increases by coef val 0.2258
2. Temperature - The Coefficient value of temperature indicates, that for a unit increase in temperature the bike hiring number increases by coef val 0.6783
3. Light Snow/ Rain – The Coefficient value of light snow/rain indicates, that for a unit increase in light snow/rain the bike hiring number decreases by coef val 0.1408

General Subjective Questions

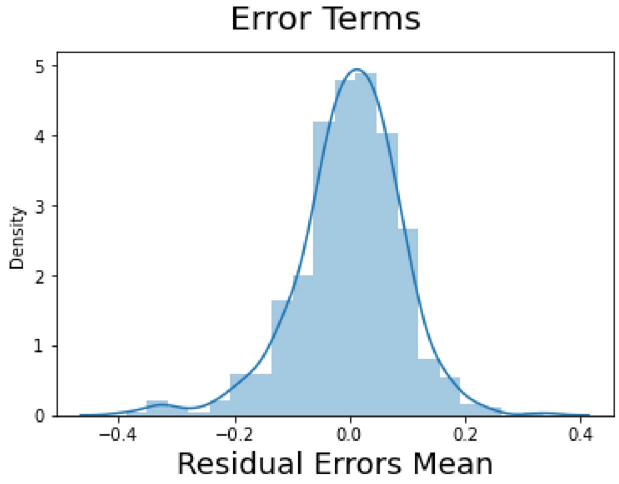
1 Explain the linear regression algorithm in detail.

Ans: Linear regression is a statistical method which is used to determine the linear relationship between and independent variable X and the dependent variable Y. It is used to predict the value of Y(dependent variable) using the value of X(independent variable)

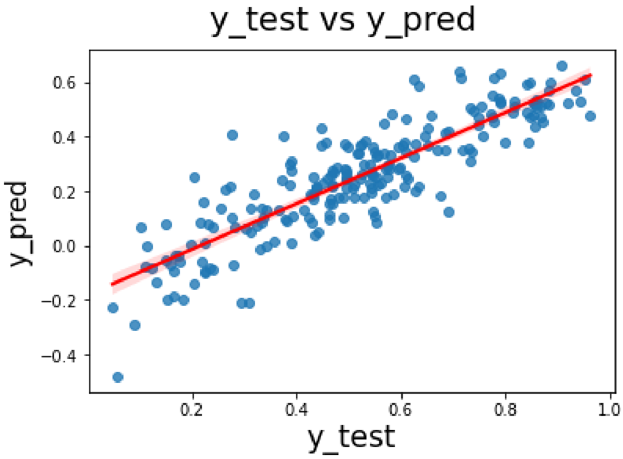
Linear Regression algorithm attempts to find the best fitting line for the variable data points, It does this by minimizing the sum of the squared errors(SSE) between the predicted and the actual values also known as Residual Sum of Square (RSS). The BFL also known as Best fit line is determined by the linear equation y = c + mx where Y is the dependent variable , c is the intercept, m is the slope , x is the independent variable

The basic steps involved in the Linear Regression Algorithm are as follows:

1. Analyze the data: Analyze and gather the data points that is used in the regression analysis
2. Calculate the mean values: Calculate the mean values of x bar and y bar
3. Calculate the coefficients : Calculate the slope (m) and the intercept (c) of the best fit line. Which is done by minimizing the Residual sum of Squares (RSS)
4. Residual analysis : It tells us about the errors in the dataset between actual data points – predicted data points, plotting a graph of this will show that the dataset is normally distributed meaning the mean of the errors is 0



5. Making Predictions using the model : We make predictions on the test dataset by transforming it onto the trained dataset, afterwards calculating the R^2 score of the test data. The R^2 of the Trained data and the Test data will be somewhat similar, Hence by plotting the graph of the test dataset vs the predicted dataset we get a linear model like below



2 Explain the Anscombe’s quartet in detail.

Ans: Ansombe’s quartet is a set of four datasets, each comprising of eleven(x,y) points. The datasets were created by statistician Francis Anscombe in 1993 to demonstrate two things, The importance of graphing data before analysing it and the effects of the outliers on statistical properties (Mean, Variance, Correlation). This is the importance of visualizing the data before analysing it statistically.

Each of the Datasets consists of eleven points with an x-value, a y- value and a dataset number. The x- values range between 4 to 19 and the y- values range from 7 to 29. The datasets are identical in many aspects:

- The mean of the x- values is the same in all four datasets (9.0)
- The mean of the y- value is also the same in all four datasets (7.50)
- The variance of the x- values is the same in all four datasets(11.0)
- The variance of the y-values is the same in all four datasets(4.125)

3 What is Pearson’s R?

Ans: Pearson’s R, also known as Pearson’s correlation coefficient, is the measure of linear correlation between two variables. It ranges from -1 to 1, where -1 being the least correlated or indicating a perfect negative linear correlation, 0 indicated no linear correlation between the variables and +1 being the most correlated or indicating a perfect positive correlation

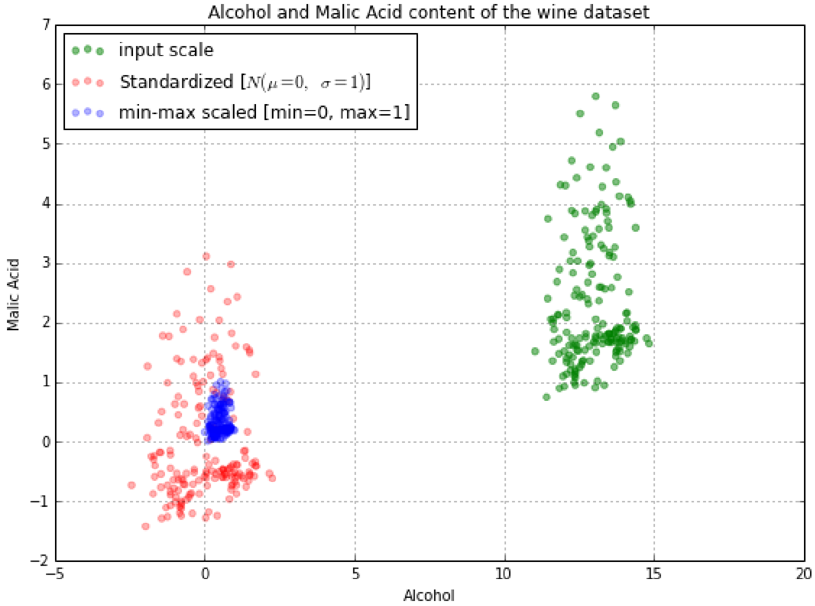
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

Pearson’s R can be calculated by dividing the covariance between the two variables by the product of the two variables standard deviations. This formula is used to determine the strength of the linear relationship between two variables. Pearson’s R is also used to compare the correlations between multiple variables and often used to assess the strength of relationships between variables in Research studies

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

Ans: Scaling is the process of transforming data from one range to another. The reason it is performed is to ensure that the data points in the dataset have the same weight when making calculations and comparisons

The difference between Normalized scaling and Standardized scaling is that normalized Scaling also known as min- max scaling, rescales the data from a range of 0 to 1 , whereas the standardized scaling rescales the data so that it has a mean of 0 and a standard deviation of 1



Above Graph explains the Scaling process with normalization and Standardization

The Blue points in the graph shows the normalization scale which is graphed between 0 and 1. Whereas the red points show the standardization scale being between negative and positive values with the mean being 0.

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

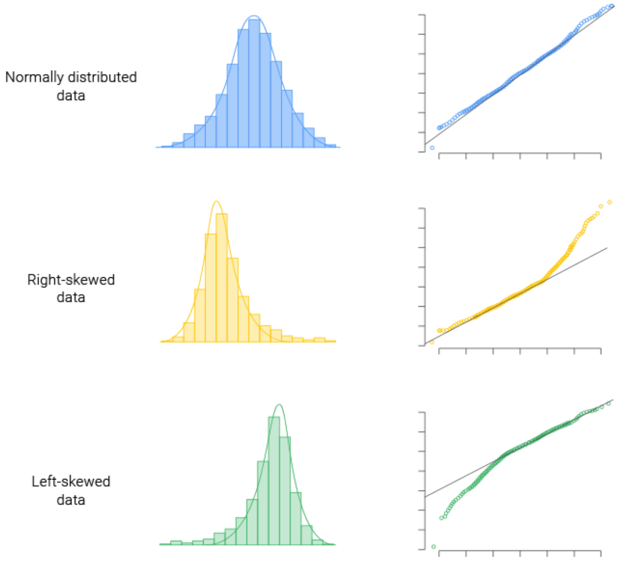
Ans: The value of VIF can sometimes be infinite when two or more predictor variables are perfectly correlated with one another. This means that one of the variables can be expressed as a linear combination of the other variables, Hence it does not have no uniqure effect on the response variable and should be removed from the model to avoid multicollinearity. During perfect correlation we get R2 = 1, which lead to 1/ (1-R2) infinity.

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

A Q-Q plot also known as quantile- quantile plot is a graphical tool which is used to compare two probability distributions by plotting their quantiles against each other. It determines whether the residuals of the model is normally distributed, If they are not normally distributed, it indicates the model is not a good fit for the data

The Q-Q plot is also used to identify potential outliers, compare distributions and identify influential observations, which are observations that have an abnormal amount of effect on regression line

In linear regression a Q-Q plot is used to assess the assumptions of linearity and normality that are necessary for the model to be correct, If the data points are in a straight line in a Q-Q plot , it indicates the data is following a linear pattern, If they are on a 45-degree angle it indicates the data is normally distributed. An example of Q-Q plot is given below


Normally distributed data
Right-skewed data
Left-skewed data