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?

Effect of categorical variables:

- Season - The number of bikes rented is more in Fall and least in the Spring season.
- Month - The demand for bikes is highest in months between June to September and there is low demand during December to Febuary.
- Weather - The number of bikes rented is more when the weather is good and the demand decreases on days with bad weather i.e. when it rains or snows.

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

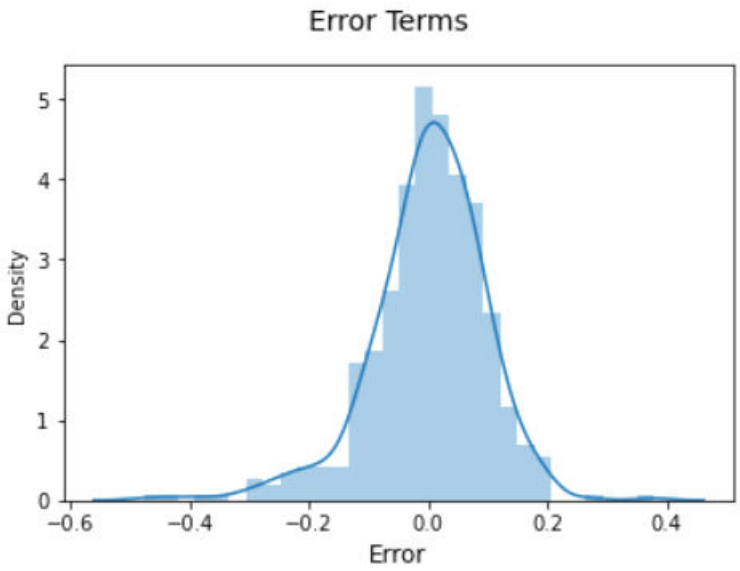
By using the drop_first we are removing an irrelevant column as the information of n columns can be given by n-1 columns.

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

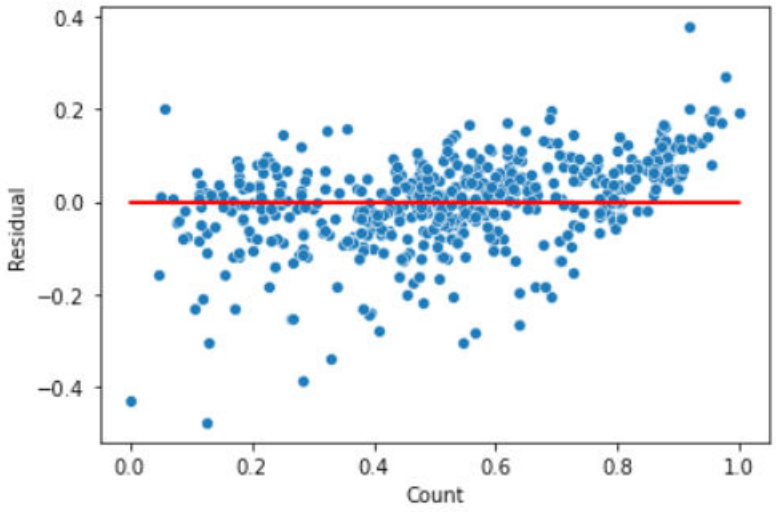
The feature 'temp' i.e. temperature during the day has the highest correlattion with the target variable 'cnt'.

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

- The independant variables showed a linear relationship with the target variable.
- Ther error terms show a normal distribution



- The residual terms show a constant variance and there is no visible pattern.



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

Based on the final model, the features temp (temperature) , yr(year) and weathersit (weather) have a major effect on the demand for bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

The Linear Regression is a method of modeling the best linear relationship between the independent variables and dependent

variables. The simplest form of Linear Regression can be defined by the following equation with one independent and one

dependent variable:

$y = \beta 0 + \beta 1x$, where

x is the independent variable, y is the dependent variable.

Linear regression can be of two types, Simple LR with 1 independent variable and Multiple LR with 2 or more independent variables.

Linear Regression Algorithm:

This algorithm tries to find the best fit line for the available data such that the error in predicted values is minimized.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if

built. They have very different distributions and appear differently when plotted on scatter plots.

These four plots can be defined as follows:

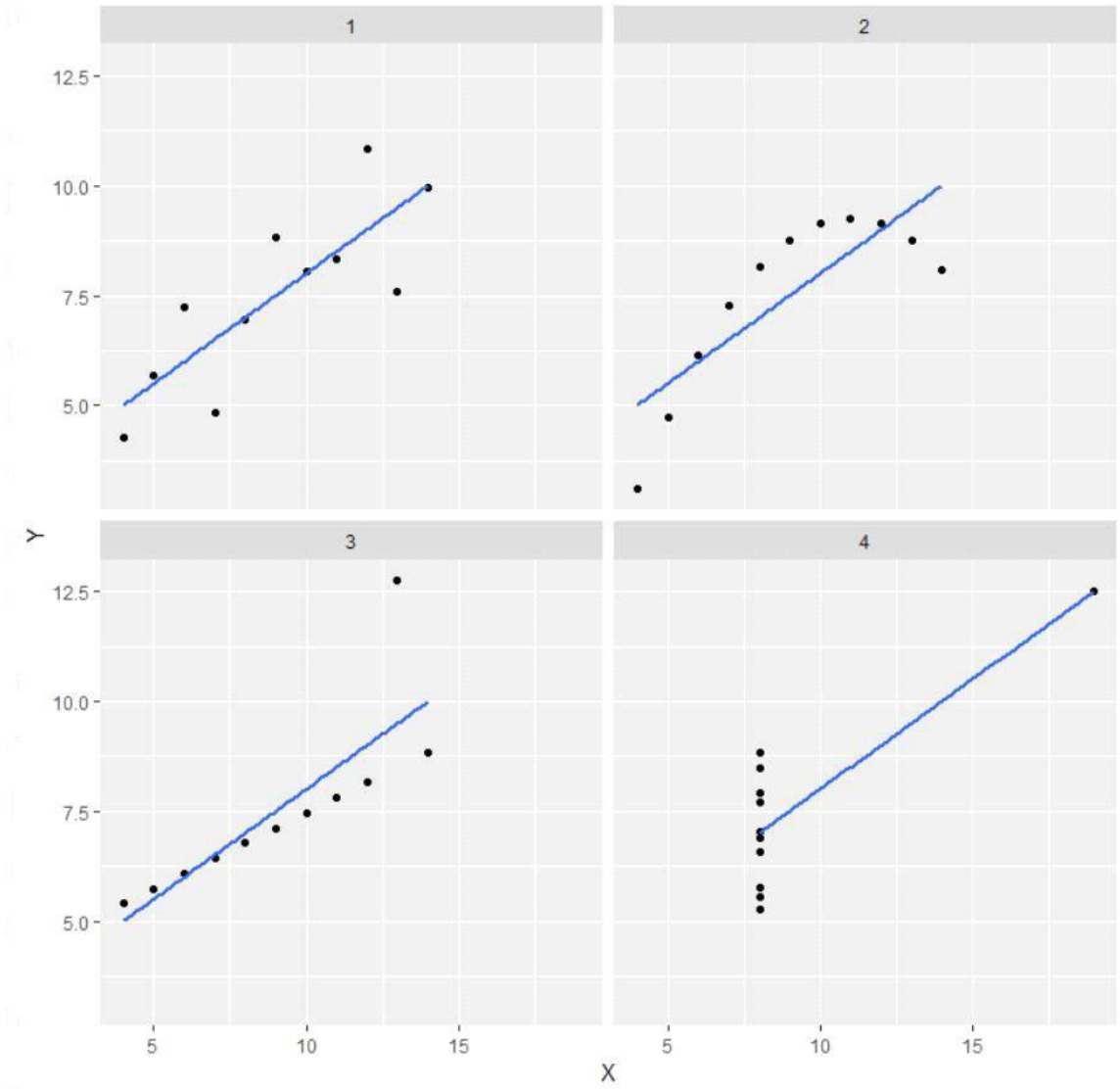
The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model



The four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

Pearson's correlation coefficient or Pearson's R defines the strength of linear relationship between the concerned variables. It varies between -1 to 1, where 1 means a perfectly linear relationship with positive slope and -1 is a perfectly linear relationship with negative slope.

A value of 0 means there is no linear relationship between the variables.

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

The values of variable have different scales depending on the the feature they represent.

Feature scaling is done to normalize the values to a common scale ensuring the features used in the model have values within similar range.

This is done as if the feature scale among variables differs a lot , the weightage given to a variable may be biased depending on its scale. Hence distance based algorithms and gradient descent based algorithms need the features to be scaled.

In normalized scaling the values are rescaled to be in the range 0 and 1.

In standardized scaling the values are rescaled in a way that the mean is 0 and standar deviation is 1.

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

VIF is infinite when there is perfect correlation between the variables that are being checked for collinearity.

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

Q-Q plots also called as quantile-quantile plots are used to plot the quantiles of two different distributions.

Using Q-Q plots we can check if the distribution is of any specific type like normal or unifrom distribution.

The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common

distribution, the points will fall on that reference line.