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?

- During spring season the total count is clearly lower than other season.
- The Overall total of bike rental is increasing, as the year increases.
- When the weather is Clear, few clouds or Partly cloudy, the total count is clearly higher than rest of the weather condition

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

When N category present in categorical variable, N-1 dummy variables are enough to identify all the categories. Because when N-1 dummy variables hold 0 as value, then it will be identified as dropped category. It will reduce the redundant independent variables.

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

Temperature variable temp has the highest correlation with target variable. In the temp vs cnt plot, cnt clearly increases as the temp variable increase.

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

By creating the histogram of error terms, I have verified the error terms(Residual) is normally distributed with the mean of 0.

By creating scatter plot for error terms to verify the error terms doesn't follow any pattern and distributed randomly around zero.

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

Temperature(temp), Year(yr), Light (weathersit category "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds") are the 3 features which contributes significantly to the target variable explanation.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm is a supervised machine learning algorithm which aims to predict the target variable based on the independent variable by finding out the linear relationship between them. In this, model finds the best fit line which represents the independent and target variable linear relation, to predict the target variable based on the independent variables.

2. Explain the Anscombe's quartet in detail.

The Anscombe's quartet comprises of 4 datasets with 11 points which have nearly identical descriptive statistics, but when graphed it all had completely different distribution and will appear differently. It is created to counter the impression of statistics are exact and graph are rough.

- In the first dataset, when we look at the scatter plot we will see that there seems to be a linear relationship between x and y.
- In the second dataset, when we plot the scatterplot, it will look like there is a non-linear relationship between x and y.
- In the third one, we can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one, when we graphed it, we can find one high-leverage point is enough to produce a high correlation coefficient.

The main aim of this quartet is to illustrate the importance of looking at the set of data graphically before proceeding with the Analysis.

3. What is Pearson's R?

It is a measure of linear correlation between two datasets. It is the ratio between the covariance of two variables and the product of their standard deviations. Its value always between -1 to 1. 1 indicates perfect positive correlation and -1 indicates perfect negative correlation and 0 represents no correlations between the 2 datasets. It is not possible to practically obtain 1 or -1 as value between 2 datasets. So close to 1 represents strong positive correlation and closure to -1 indicates strong negative correlation and close to 0 represents weak correlation.

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

Scaling is the technique to standardize the variables in a fixed range. It is performed for easy interpretation and faster convergence for gradient descent methods. When the variables are on different scale model will try to weigh greater value as higher and smaller value as lower regardless of the unit of the values.

In standardized scaling the variables are scaled in such a way that their mean is 0 and standard deviation is 1.

$X = (x-mean(x))/(std(x))$

In Normalised scaling, the variables are scaled in a way that all the values lie between 0 and 1 using the maximum and minimum values in the data.

X = (x-min(x))/(max(x)-min(x))

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

When a variable is perfectly correlated with another variable, we will get VIF as Infinite. Because when there is a perfect correlation R value will become 1 and VIF which is a measure of 1/(1-R^2) will become infinite.

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

Q-Q plot are the plots of two quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.