Assignment-based Subjective Questions

1> From your analysis of the categorical variables from the data set, what could you infer about their effect on the dependent variable?

Solution:

Observations from (linear regression assignment.ipnyb file) box plots for categorical variables:

- The year box plots indicates that more bikes are rent during 2019.
- The season box plots indicates that more bikes are rent during fall season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The month box plots indicates that more bikes are rent during September month.
- The weekday box plots indicates that more bikes are rent during Saturday.
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

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

Solution:

- To avoid Multicollinearity (if we don't drop, dummy variables will be correlated) and affects the model adversely
- To avoid redundant features

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

Solution:

- By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

- Residual errors follow normal distribution
- Maintains linear relation between defendant variable (Test and predicted

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

Solution:

The Top 3 features contributing significantly towards the demands of share bikes are:

- weathersit_Light_Snow(negative correlation).
- yr_2019(Positive correlation).
- temp(Positive correlation).

General Subjective Questions :

1> Explain the linear regression algorithm in detail

Solution:

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.

Linear regression can be further divided into two types of the algorithm:

A> Simple Linear Regression:
If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

B> Multiple Linear regression:
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Finding the best fit line:

When we are working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

2> Explain the Anscombe's quartet in detail.

Solution:

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

3> What is Pearson's R?

Solution:

Pearson correlation coefficient is a measure of the strength of a linear association between two variables and it's denoted by r.

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

Solution:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1 sklearn Pre-Processing Min-Max Scaling helps to implement normalization in python.

$$\text{Min Max scaling } x = \frac{x-\min(x)}{\text{Max}(x)-\text{Min}(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (u) zero and standard deviation one.

$$\text{Standardization } x = \frac{x-\text{mean}(x)}{\text{Sd}(x)}$$

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

Solution:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the data set which is causing this perfect Multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Solution:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For

example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. 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.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.