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:-
season,mnth,weather,holiday,weekday,working day,yr are the categorical variables
1. Season - fall has the maximum count
2.month -September has the max count
3.during Clearsky of weathersit is good
4.Weekdays are more are less same
5.Year 2019 has more count

2. Why is it important to use drop_first=True during dummy variable creation?
Ans:- If we do not use drop_first=True for the dummy variables will be correlated to each other

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans:- error terms should correspond to a normal curve.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans: - 1.Weathersit-Snow
2.Month - Sep
3.Weekday - Sunday

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression model is the one which is used to find the correlation between independent variables and the dependent variable

Once the data cleansing is done split the data into training and test data sets. After checking the collinearity of variables and using the requisite variables to train the model and checking the R-value of the model and the p-values of dependent variables, after dealing/dropping the necessary columns and reiterating the steps (feature elimination), we come to a final model.

According to the conditions of linear regression which states that the error curve must be a normal one, we proceed to testing the model with the test dataset. The conclusion hence drawn on the model would be used to provide valuable insights/predictions on datapoints in the range of the model.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.

3. What is Pearson’s R? (3 marks)

Pearson correlation coefficient or Pearson’s correlation coefficient or Pearson’s r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

In simple words, Pearson’s correlation coefficient calculates the effect of change in one variable when the other variable changes.

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

Scaling is necessary for a model to be functional with the appropriate range of coefficients. For e.g., if there were two independent variables named price and months on which the sale of car depended, the price range would be far too high because there are only 12 months in a year. In that case, scaling the variable price appropriately won't allow decimal errors to happen in the model. There are two types of scaling:

- Normalized scaling: This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a preset range. Typically used in Neural networks broadly.
- Standardized scaling: The example given above is of standardized scaling. Here, the values of variable(s) is/are compressed into a specific range to suit the model.

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

It means that there is a perfect correlation between dependant variable and independent variable

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

This is the tool to test whether the data comes from same statistical distribution .