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: a. Yr and target variable has a positive correlation.
b. Season and target variable has positive correlation.
c. Weathersit and target variable has negative correlation.

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

Ans: drop_first=True will help reduce the creating extra column during dummy variable creation.

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

Ans: temp and atemp has the highest correlation with larget variable.

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

Ans: Creating a scatter plot between x and y after the model is being trained. There should be linearlity between the variables, mean should be centered at zero.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?
Ans: Season, yr and temp are the 3 features significantly contributing towards explaining the demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression performs the tasks to predict a target variable(dependent variable) based on given independent variables. It finds out a linear relationship between a dependent and other given independent variables.
Equation for best fitted line for linear regression is y=b0 + b1x

2. Explain the Anscombe's quartet in detail.

Ans: Anscombes quartet can be defined as a group of 4 datasets which are nearly identical in simple descriptive statistics. This tells us about the importance of visualizing data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in data like outliers etc.

3. What is Pearson’s R?

Ans: Pearson’s R is a correlation coefficient. It is the most common way of measuring a linear correlation. It lies between -1 and 1 which measures the strength and direction of relationship between two variables.

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

Ans: Scaling is a technique to bring all the variables to measure at same scale.

Scaling is performed for easy interpretation of coefficients, if we have variables at different scales then interpretation of coefficients after training the model becomes very difficult. Another important reason that scaling is performed is that the computing time of mathematical optimization functions like gradient descent will be much faster therefore reducing the cost of the function.

Normalized scaling is min-max scaling where the data is compressed between 0 and 1.

Standardization converts data so that mean is centered at 0 and standard deviation is 1.

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

Ans: if there’s a perfect correlation between all the variables then VIF is infinite.

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

Ans: Q-Q plots are plots of two quantiles against each other. It is used to find out if two sets of data come from same distribution. If the distributions are linearly related then Q-Q plots will approximately lie on a line.