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?

Answer : From my analysis I have noticed most of the bookings are on Thursday, Friday and Sunday and from may, June, July, Aug and Sep months. Bookings are same on all days.

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

Answer : It's important to use, it helps in reducing the extra column created during dummy variable creation.

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

Answer : temp

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

Answer : Linear relationship validation and Multicollinearity check

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

Answer : temp, winter and sep

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer : Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

2. Explain the Anscombe's quartet in detail.

Answer : Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R?

Answer : The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and

1 that measures the strength and direction of the relationship between two variables.

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

Answer : Scaling is a technique to standardize the independent features present in the data in a fixed range.

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

Answer : if there is a perfect correlation, then VIF = infinity

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

Answer : 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.