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

Ans: Observations 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?

Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Ans: 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?

Ans: Pair-wise scatterplots may be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot

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

Ans: 1. Holiday, 2. Temp, 3. Hum

6. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

7. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

8. What is Pearson’s R?

Ans: Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

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

Ans: Scaling is used to normalize the data within a particular range. 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 modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Normalization/Min-Max Scaling brings all of the data in the range of 0 and 1. Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Ans: 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 dataset 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).

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

Ans: 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