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

- The demand for bikes increasing from 2018 to 2019.
- The season affecting bike demand, with demand increasing from spring to fall, and then decreasing in winter.
- Holidays have less bike demand compared to regular days.
- Working days have more bike demand compared to non-working days.
- Weather situation affects bike demand, with clear days having the highest demand and light rain or snow days having the lowest demand.

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

The drop_first=True option is used during dummy variable creation to avoid correlating variables as there might opposite variables for the same use cases. So, we drop the first dummy variable to avoid this issue.

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

Based on the heatmap that I've created I can find that casual and registered and temp and atemp are highly correlational to the output.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

We can validated the assumptions of Linear Regression by checking for linearity (through scatter plots),

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

Based on final model the top 3 features contributing significantly towards explaining the demand of the shared bikes are 'yr' (year),temp, atemp,workingday).

General Subjective Questions:

Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task, which is a kind of predictive modeling technique. The variable to be predicted is the dependent variable, and the one used to predict the value of the dependent variable is the

independent variable. In simple linear regression, there is one independent variable. In multiple linear regression, there are two or more independent variables.

Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties (mean, variance, correlation, and linear regression line), yet appear very different when graphed.

What is Pearson's R?

Pearson's R, also called the Pearson correlation coefficient, measures the direction of association between two continuous variables. The value ranges between -1 and 1.

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

It basically helps to normalize the data within a particular range. Normalized scaling scales all values in a fixed range between 0 and 1. Standardized scaling transforms the data between -1 to 1

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

The Variance Inflation Factor (VIF) is a measure of multicollinearity in a set of multiple regression variables.

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

A Q-Q (Quantile-Quantile) plot is a plot of the quantiles of two distributions against each other. In the context of a linear regression, a Q-Q plot of the residuals is used to check the assumption that the error terms are normally distributed.