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:

Categorical require a special attention in regression analysis, It identifies important covariates related to how likely an individual move from one response category to another. Bike_Sharing_company(season, weathersit, yr, mnth, holiday, workingday)

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

ANS:

To avoid dummy variable trap, we need to drop the first variable from each category. Dummy variable trap where there are attributes that are highly correlated(Multicollinear).

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

ANS:

cnt

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

We can identify the assumption is met or not is by a creating scatter plot between x and y. Describe the linear relationship between the dependent variable and independent variables from a dataframe.

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

Bike demands – It depends on the different variables like year, holiday, temp, windspeed, sep, Light_snowrain, Misty, spring, summer and winter.

General Subjective Questions:

1. Explain the linear regression algorithm in detail. ?

Linear Regression is a supervised machine learning method that is used by the train using AutoML and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variables.

2. Explain the Anscombe’s quartet in detail ?

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.

3. What is Pearson’s R?

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?

Feature scaling is the process of normalizing the range of independent variables or features in a dataset. In linear regression, feature scaling is important because it can affect the performance of the model.

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

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.