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. From the analysis of the categorical it could be inferred that the rentals are likely to be more during fall and summer season. Moreover, when the weather is clear the demand for the bike is high.

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

Ans. It is imperative to use that because we want to create n-1 level for a categorical variable having n levels to avoid the redundancy.

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

Ans. The temperature has the highest correlation with the target variable.

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

Ans. Assumptions of Linear Regression were validated by confirming that the error terms were normally distributed, there was linear relationship between the dependent variable and independent variable and the VIF and p-value value were within the acceptable limits.

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

Ans. The top three features are temperature, year and weather.

6) Explain the linear regression algorithm in detail.

Ans. Linear Regression is supervised learning type of the ML algorithm, which is used find the linear relationship between the dependent and independent variables. Two of linear regression are simple linear regression and multiple linear regression. In single linear regression there is one dependent and one independent variable, whereas in multiple linear regression there is one dependent and more than one independent variable. Assumptions of the linear regression are: linearity, independence, Homoscedasticity, Normality, No multicollinearity and No endogeneity.

7) Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet consists of the four dataset that have equivalent descriptive statistics but have significantly different distribution and hence appear very different. 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.).

8) What is Pearson's R?

Ans. Pearson's R is the used for identifying the strength of linear relationship between two quantities, and it lies between -1 to 1.

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

Ans. Scaling is a pre-processing where the data of continuous variables are standardized to bring all of them in similar range, so that we can make accurate inference about the coefficients of the independent variables. In normalization all the data is brought between 0 and 1, whereas in standardization data are replaced with their Z-score.

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

Ans. VIF is infinite when there is perfect correlation between the independent variable suggesting a very high multicollinearity. It happens when the R-squared is 1 between the different independent variable.

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

Ans. Quantile-Q**u**antile plot compares two probability distribution using by plotting their quantiles against each other. In linear regression it can used to compare the probability distribution of the y_train and y_pred. If they are similarly distributed then the Q-Q plot would lie on the y=x.