Assignment-based Subjective Questions and Answers

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

Answer- The demand of bike is less in the month of spring when compared with other seasons and decreases in the winter season. The demand bike increased in the year 2019 when compared with year 2018.

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

Answer - 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?

Answer- The numerical variable 'registered' has the highest correlation with the target variable 'cnt', if we consider all the features.

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

Answer- The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

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

Answer- temperature, year and holiday

General Subjective Questions and Answers

1. Explain the linear regression algorithm in detail.

Answer - 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.

2. Explain the Anscombe's quartet in detail.

Answer - 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.

3. What is Pearson's R?

Answer - In statistics, the Pearson correlation coefficient also known as Pearson's *r*, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

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

   Answer - It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

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

Answer - If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.