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: The demand for bikes is less in the month of spring when compared with other seasons. The demand for bikes increased in the year 2019 when compared with the year 2018.

2. Why is it important to use drop_first=True during dummy variable creation?
   Ans: The drop_first=True is important 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: 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?
   Ans: The easy way is to draw the distribution of residuals against levels of the dependent variable.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: As per the Final Model, the top 3 predictor that influences the bike booking are:

   Temperature (temp) - A coefficient value of '0.5636' means that there is an increase in the bike hire numbers by 0.5636 units.

   Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3070', means that there is a decrease in the bike hire numbers by 0.3070 units.

   Year (yr) - A coefficient value of '0.2308' which means there is an increase in the bike hire numbers by 0.2308 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task and models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

2. Explain the Anscombe's quartet in detail.
   Ans: Anscombe's Quartet is a group of four data sets that provide a useful caution against blindly applying statistical methods to data. Each data set consists of ten x- and y-values such that the mean and variance of x and y, the correlation coefficient, regression line, and error of fit using the line are the same.

3. What is Pearson's R?
   Ans: The Pearson's R, is a statistical calculation of the strength of two variables' relationships. In other words, it's a measurement of how dependent two variables are on one another.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans: Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.
   In regression, it is often recommended to scale the features so that the predictors have a mean of 0. This makes it easier to interpret the intercept term as the expected value of Y when the predictor values are set to their means.
   The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

5. 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.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Ans: The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot, if the two data sets come from a common distribution, the points will fall on that reference line. This particular type of Q-Q plot is called a normal quantile-quantile (QQ) plot.