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?

   - The fall season has a high number of bike rentals.
   - June, and Aug, Oct is having a high number of bike rentals.
   - People prefer bike rental when weather situations is good.
   - People don't prefer bike rental on weekends.
   - People don't prefer renting bike holidays.
   - Year 2019 has higher Bike Rental than the year 2019.

2. Why is it important to use **drop_first=True** during dummy variable creation?
   - 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?
   - The ‘temp’ and ‘atemp’ variables have the highest correlation when compared to the rest with target variable as ‘cnt’.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   - The assumptions of Linear Regression have been validated by the application of Residual Analysis.
   - Plotted the histogram of the error terms and discovered that "Error Distribution" Is Normally Distributed Across 0, which implies that our model has handled the assumption of Error Normal Distribution appropriately.
   - Assuming Terms for Errors Being Independent: we observe that the residual and the expected value seldom ever correlate.
   - Homoscedasticity: from both ends of the fitted line, we can observe that variance is comparable.

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

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   - One kind of supervised machine learning is linear regression. Linear regressions come in two varieties.
     One predictor is used in a simple linear regression.
     For example, y=b0+b1x, where x is the predictor and b0 are the interceptors and b1 are the coefficients or slop.
   - Multiple Linear Regression: When there are more than one predictor.
     ex: y= b0+b1x1+b2x2 +...+bnxn, where b0 is the interceptor and b1, b2,b3,...bn are the coefficients/slopes of the predictors x1, x2, x3,...xn.
   - A continuous value is the target variable in linear regression. In other words, the goal of linear regression is to find a fitted line—or, in the case of multiple linear regression, the fitted plane—so that the total error between the target value and the projected value is as little as possible.

2. Explain the Anscombe's quartet in detail.
   - Anscombe's Quartet, comprising four datasets with nearly identical summary statistics, underscores the limitations of relying solely on numerical metrics.
   - This article explores the quartet's datasets, emphasizing the importance of visualizing data for a comprehensive understanding.

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. With formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   - It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
   - Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the

variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Normalization/Min-Max Scaling
    - It brings all of the data in the range of 0 and 1.
    - sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

- Standardization Scaling
    - Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
    - One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.
    - sklearn.preprocessing.scale helps to implement standardization in python.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? The value of VIF is calculated by the below formula:

Variance Inflation Factor

VIF

$$VIF_j = \frac{1}{1\text{-}RSquare_{x_j}}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
- This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.