Assignment-based Subjective Question

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

   Categorical variable played a major role as it was identified that the categories themselves did not indicate much meaning but when they replaced by the meaningful numerical dummy variable, they made more sense to how the dependent variable was impacted by them

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

   In order to remove the redundant column where the rest of the columns will be giving the required information. When we have a K categorical variable we would need K-1 dummy columns to represent those.

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

   temp (Temperature)

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

   By plotting a normal distribution curve for the residuals values of the training set. The values were clustered around 0 indicating that most of the residuals were 0 and the model was a good fit on the training set.

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

   Season, Weather situation, Holiday.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   1. Linear regression in a method of predicting the value of the dependent variable (variable of interest) using the independent variables (impacting variables).
   2. Linear regression models are easy to develop and gives meaningful insights on the variables (independent variables) impacting the prediction (dependent variable).
   3. Linear regression is basically explained by straight line formula y = mx + c
   where m is the slope and is the intercept wrt to ML m is the regression co-efficient and y is the independent variable and x is the dependent variable.
   4. Linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.

2. Explain the Anscombe's quartet in detail.

   Anscombe's quartet gives the necessary information about the data at hand on which we can build linear regression model or not. This would give us the visual representation of correlation between the dependent and independent variables. So, we can conclude which variable would help us in building the model.

3. What is Pearson's R?

   Pearson's R the correlation co-efficient between the dependent variable dataset and independent variable dataset. It explains how the strong the 2 variables of interest are correlated.
   The variables can be related positively or negatively.

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

   Scaling is a method of bringing down the independent variables which are having high values with respect to each other so that all the independent variables are of similar scale.
   Scaling is done in order generalize the dependent variable and independent variable such that the distance between them is reduced, with this the accuracy increases.
   Normalized scaling or Normalization gets the overall values between the range of 0 to 1 where the max value will be represented by 1 and min value will represented by 0, Accordingly the intermediate values by the decimals between 0 to 1. Formula for Normalization X_new = (X - X_min)/(X_max - X_min).
   Standardized scaling is a transformation of the variable value to relatively lower value by subtracting the X from mean and dividing by standard deviation.

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

   VIF would be infinite when there are variable which having perfect correlation with each other.

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

General Subjective Questions

Q-Q plot is quantile analysis between 2 datasets, it helps in determining if the 2 datasets have same distribution. To understand the skewness. If the residuals have the normal distribution or not.