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.

   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. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

2. Explain the Anscombe's quartet in detail.

   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.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R?

   The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

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

   The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.
   Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
   Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

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

   Multicollinearity occurs when there are two or more independent variables in a multiple regression model, which have a high correlation among themselves. When some features are highly correlated, we might have difficulty in distinguishing between their individual effects on the dependent variable. Multicollinearity can be detected using various techniques, one such technique being the Variance Inflation Factor (VIF)

General Subjective Questions

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

   The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. If two populations are of the same distribution

   If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.

   Skewness of distribution