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 number of customers for Boom Bikes is recorded high in the season of Summer & Fall where the weather situations are Clear_FewClouds and Mist_Cloudy.
   So, the citizens of US are more likely to utilize the BoomBikes service during the Fall season and Summer season on all the days of the week (whether it is a workingday/non-working day) except during holidays.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   The n-1 Dummy variables are created for n unique values of a each categorical variable when we used the pandas function pd.get_dummy(cat_var). But we should use another parameter drop_first=True to reduce one extra column as well as reduce the correlation between the encoded dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Both the numeric variables 'temp' & 'atemp' has highest correlation of 0.63 with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Hypothesis Testing is used to validate the assumptions of BoomBikes Linear regression model.

   H0:B1=B2=...=Bn=0

   H1: at least one Bi!=0

   So all the coefficient values B1,B2,----Bn are not equal to zero from the model.Thus the null hypothesis is rejected.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of shared bikes?**
   As per the final model, the top 3 features that contribute significantly are:
   Temperature (temp)
   Light Rain & Snow (weathersit =3)
   Year (yr)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear Regression algorithm is a supervised learning technique that performs a regression task to predict the target value which is of continuous value based on the independent variable values of historical data.

   It finds the linear relationship between the independent variable and target variable and plot a regression line with a minimum residual sum of squares from each data point. This regression line is the line of best fit for the model.

   The mathematical equation behind the Linear Regression model is

   $y = \theta1X+ \theta2$

   y - target variable

   X – independent variable(capitalized because there can be any number of independent variables)

   where θ1 is the slope/coefficient value of X which represents the increase of coefficients times of target variable when there is a unit increase in the independent variable (X)

   θ2 is the constant/intercept value that added while building the model using Ordinary Least Squares algorithm from stats library to not pass the regression line through origin.

   Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.

3. **What is Pearson's R?**

   Pearson's R is a statistical calculation of the strength of two variables' relationships.

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

   Scaling is a technique used to transform the data to fit into a specific scale(0-1 / 1-10 etc,)

   Most of the time, the collected data set contains features highly varying in magnitudes, units, and ranges. If scaling is not done then the algorithm only takes magnitude into account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalized scaling brings all of the data in the range of 0 and 1.

Standardized scaling 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?**
   If there is multicollinearity between independent variables then the VIF value becomes infinity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Q – Q plot known as Quantile – Quantile plot, is used to check the normality of data(reference line and value distribution). All the values lie on the reference line then the distribution is normal.