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?

   Answer: From the analysis of the categorical variables from the dataset. People mostly used the Bike Rental are more in season summer and fall, in the months September and October uses more ,Saturday, Wednesday and Thursday in Weekday, in clear Weather use more,2019 year more bike rents compare to 2018 year, in the holiday most of people use Bikes.

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

   Answer: drop_first=True helps in reducing the extra column created during the dummy variable creation, avoids the redundancy.

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

   Answer: Temperature variable has high correlation with target variable.

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

   Answer: Assumptions of Linear Regression by checking the VIF, P-Value and error distribution.

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

   Answer: The Top features contributing significantly towards the demand of the shared bikes: year, working day and wind speed.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer: Linear Regression is a popular supervised learning algorithm used in Machine Learning
   And Statistics for predicting a continuous output variable Y based on one or more input variable X. In
   Simple terms, it models the relationships between the input variables and the output variable by finding the best linear fit that describes the data.

2. Explain the Anscombe's quartet in detail.

   Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. What is Pearson's R?
   Answer: The Person correlation measures the strength oh the linear relationships between two variables. It has a value between -1 to 1,with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and +1 meaning a total positive correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Answer: Feature Scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If features scaling is not done, then a machine learning algorithm tends to units of the values.
   - Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithm that do not assume any distribution of the data like K- Nearest Neighbours and neural Networks.
   - Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So even if you have outliers in your data, they will not be affected by standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer: The Value of VIF is Infinite when there is a perfect correlation between the two independent variables. The R-Squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggest that is there is a problem of multi-collinearity and one of these variable need to be dropped in order to define a working model for regression.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
   Answer: Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quintiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.
   - Do two data sets come from populations with a common distribution?
   - Do two data sets have common locations and scale?
   - Do two data sets have similar distributional shapes?
   - Do two data sets have similar tail behaviour?