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? (3 marks)
Ans:

1. Bike demand is less in Season 1(Spring) when compared with other seasons
2. Bike demand is more in month of June compared with other months
3. Bike demand is less in weekends than weekdays
4. Bike demand is increased in year 2019 compared with 2018

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans:

1. Number of columns in the data set will increase and its redundant data.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans:

1. aTemp and temp

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans:

1. By Residual analysis on the train data

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Ans:

1. Temp
2. weathersit_Clear_FewClouds
3. yr

General Subjective Questions:

1. Explain the linear regression algorithm in detail. (4 marks)
Ans: 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.

2. Explain the Anscombe's quartet in detail. (3 marks)
Ans: 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.

3. What is Pearson's R? (3 marks)
Ans: It measures of linear correlation between two sets of data

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Ans:
Scaling is a personal choice about making the numbers feel right, e.g. between zero and one, or one and a hundred. For example converting data given in millimeters to meters because it's more convenient, or imperial to metric.

Normalizing can either mean applying a transformation so that you transformed data is roughly normally distributed, but it can also simply mean putting different variables on a common scale. Standardizing, which means subtracting the mean and dividing by the standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans: An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans:
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight