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?

   i. Sol: From analysis on dataset of multiple linear regression of bike sharing assignment in fall label of season variable has most no of rented bikes, year 2019 has most number of rented bikes, Month of september has most no of bikes, In clear weather has most no of rented bikes, also in the working day also have most no of rented bikes.

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

   Sol: we need drop the first label of dummy variable becoz by from the other labels we can clearly identify the another one and also we can minimize our no of variables much less variable its is simple model which is easy to analysis.

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

   Sol:Temperture and atemp variable have the highest co-relation with dependent variable.

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

   Sol: Mostly from error terms which are normally distributed there is no pattern in error term of scatter plot, and independent variables are having linear relationship with dependent varibales.

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

   Sol:Three features which are significant towards the most rented bikes are: 1.temperture
   2.Year
   3.Light snow

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Sol: Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

   Equation: y=mx+b

2. Explain the Anscombe's quartet in detail.

   Sol: 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?

   Sol: This is used for finding the co-relation between the variables:
   Pearson correlation: Pearson correlation evaluates the linear relationship between two continuous variables.

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

   Sol: Scaling is used to normalize the data, whereas standard scaling is used to convert the data into z-score according to mean and standard deviation and minmaxscaling is used for the converting the data between 0 and 1.

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

   Sol: When there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 5 this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

   Sol: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us to know if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
   ithelps to determine if two data sets come from populations with a common distribution.