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)

   The following inferences are drawn about the effect on categorical variables on target variable,

   i. Demand for the bike is higher during Fall season, followed by Summer season.
   ii. The above point correlates with the calendar months Aug - Oct.
   iii. Monday seems to have higher bike demand.
   iv. Bike demand is higher during Clear weather.
   v. There is an increase in demand in 2019 when compared to 2018.
   vi. During holidays, the demand falls compared to working days.

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

   By default **drop_first** takes **False**, in order to reduce the number of dummy columns it is assigned **True.**
   For example, there are four categories (spring, winter, summer, fall) defined within a feature (season). When the season is column is mapped into four dummy variables, when one variable is not mapped to spring, winter, summer, then it is obvious to be fall. On assigning **drop_first = True,** it will reduce this column 'fall', keeping remaining three columns (spring, winter, summer).

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

   'temp' variable has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   i. Multi-collinearity: After selecting the features using RFE and Manual method, VIF Is checked for them. And it is observed that none of them has a VIF greater than 5, hence there is no significant multi-collinearity among the independent variables used for predictions.
   ii. Normality of Error Terms: Plotted frequency distribution error terms, and it is visible that the residuals are normally distributed around mean = 0.
   iii. Homoscedasticity of Error Terms: On plotting the variance of the error terms, it is visible that variance is almost constant.
   iv. Independent Error terms: By plotting the residuals, it is observed that there is no pattern or trend among themselves. This shows they are independent of each other.
   v. Linear relationship: From the final regression equation, it is observed that target variable is linearly related to independent variable.

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

   Top 3 features contributing significantly towards the demand of shared bikes are 'light_snow', 'spring' and 'windspeed'.

General Subjective Questions

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

   Linear regression is one of the supervised learning algorithms which is used to predict a continuous variable. Through this algorithm, a linear relationship is established between the predictor variables and the target variable. For example: y = m0 + m1x1 + m2x2 + m3x3; where m1, m2, m3 are the coefficients of predictor variables x1, x2, x3 respectively and m0 is the intercept or constant. The coefficients provides an idea about the weights of the independent variable in predicting the target variable.

   Linear Regression has following assumptions in order to establish the relationship between predictor and target variables:

   1. Linearity: Target variables are linearly related to predictor variables.
   2. Independence of residual terms: There should not be any dependency between residual terms. If the residual is dependent on other residual, it gives rise to autocorrelation.
   3. Normality of error terms: The residuals are normally distributed with mean around zero.
   4. Homoscedasticity: The variance of error terms is constant.
   5. No Multicollinearity: There should not be high correlation between independent variables.

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

   Anscombe's quartet is a group of four data-sets that are nearly identical in descriptive summary statistics like mean, variance, standard deviation, but they look completely different from one another when it is plotted. It helps us to understand the importance of data visualization.

   Due to same descriptive property of the date sets, they easily fool the linear regression algorithm if built. These data sets were constructed to illustrate the importance of graphical representation before analyzing the statistical summary of the datasets

3. What is Pearson's R? (3 marks)

   Pearson's R is the correlation coefficient which is used mostly to measure the linear relationship between two variables. It ranges between – 1 and +1. +1 shows both the variables are perfectly positively correlated whereas -1 shows that both the variables are perfectly negatively correlated. A value of 0 shows there is no relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Scaling is a technique to standardize the independent variables present in the data within a fixed range for a comparable scale. It is performed to make it easy for the Machine Learning algorithm to learn from the data and understand the problem because it may be impacted by the magnitude of

the independent variables. There are two types of scaling techniques – MinMax Scaling and Standardized Scaling.

MinMax Scaling:

Through this technique, the data points are scaled to a range between 0 to 1. Here each data point is normalized by subtracting the minimum value and dividing by the difference between max and min. Hence, it is affected by the outliers preset in the dataset.

Standardized Scaling:

Each of the data points are normalized by subtracting the mean and dividing the difference with standard deviation. This scaling is used when the distribution in normal. This distribution is not affected by outliers present in the dataset.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

   VIF stands for Variance Inflation Factor, which shows the correlation effect between the variables. An infinite value of VIF indicates there is perfect correlation between the variables. Infinite VIF depicts that R-square is 1, which means there is over fitting of the model.
   Since VIF = 1/(1 – R_Square).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   A Q-Q plot provides a powerful visual assessment of the data points whether drawn from populations which follows a common probability distribution. It compares the quantiles for two distribution.
   It is used in Linear Regression to verify the normality of error terms. When the error terms are not normally distributed then the graph deviates from the straight line. Normality of the error terms is a major assumption of linear regression. This is also helpful in studying the train and test dataset, by plot Q-Q plot, we can confirm that both the datasets follows the same probability distribution.