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- Following categorical variables impact the dependent variable

   I. **Year**

   The number of customers increases every year. The number of customers next year will even be higher

   II. **Season**

   The number of customers is higher in Summer, Fall and Winter. The number of customers drop in Spring season.

   III. **Months**

   The average number of customers is higher in April, May, June, July, Aug, Sep and Oct. Whereas it is less in Dec, Jan, Feb and March.

   IV. **Weather Situation**

   - Weather Situation 1 i.e. Clear, Few clouds, Partly cloudy, Partly cloudy whether attracts more customers.
   - Weather situation 2 i.e. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist causes less customers to avail bike sharing services.
   - Weather situation 3 i.e. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds causes even fewer customers to avail bike sharing services.

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

   Ans – When drop_first is set to True, n-1 dummy variables are created for categorical variables. The n-1 dummy variables are sufficient to analyze data. Thus, by avoiding creation of unnecessary columns, the size of data can be reduced which helps make analysis easier.

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

   Ans – "temp" and "atemp" numerical variables have 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)

   Ans – After plotting distribution plot of the residuals, it was visibly clear that the error terms were normally distributed.

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 – The top 3 features contributing towards the demand are

   1) Fall season is favorable factor. Number of customers increase in Fall season.

2) Weather situation 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) is quite unfavorable situation. Number of customers decrease significantly during this type of weather situation.
3) Summer season is the next favorable factor. Number of customers increase in summer season.

General Subjective Questions

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

   Ans – Linear regression algorithm is a supervised machine learning algorithm that tries to explain or predict the values/behavior of a target variable based on the data provided for the independent variables. Regression algorithm is used for predicting continuous data. As the name suggests, linear regression is used to predict the target variable when it is expected that there is a linear relationship between target variable and predictor variables.

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

   Ans – Anscombe's quartet is a group of 4 data sets that appear similar when viewed in tabular format, however, appear completely different when put to graphical format. By using graphical demonstration, Anscombe established the importance of using graphs while analyzing data.

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

   Ans – Pearson's R is a correlation coefficient that helps explain the correlation between two sets of data. It has a range of -1 to 1, where higher absolute value indicates higher correlation between data and a lower value i.e. values nearer to 0 indicate lower correlation.

   A positive Pearson's R value indicates that both the data sets either increase or decrease together while a negative value indicates that one data set increases with decrease in the value of other data set. The higher the value, the more the correlation between data sets.

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

   Ans – Scaling is a technique through which data can be standardized and scaled within a set range. Thus scaling will fit data of varying magnitude into a fixed scale, hence helps the machine learning algorithm to give proper weightage to all variables.

   Normalized scaling is a process where the variables are scaled in the range of 0 and 1.

   Standardized scaling is a process where the variables are scaled in a way that each data has it's mean as 0 and a standard deviation of 1.

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

   Ans – VIF is a measure of multicollinearity of an independent variable with respect to other variables.

   The formula of VIF is

   $$1/(1-R^2)$$

   Where R^2 is the R-squared value of a model built for a predictor using other predictor variables.

When R-squared value is 1 or very near to 1, then as per the formula, VIF tends to become infinite, which means that the predictor variable is highly multicollinear. In other words, the predictor variable can be explained very well using other predictor variables and hence is redundant.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Ans – Q-Q plot refers to Quantile-Quantile plots, where Quantiles of two data sets are plotted to find out if the quantiles belong to the same data distribution. If the quantiles follow a linear relationship, then they belong to same data distribution.