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)

   Some of the categorical variables are significant .
   Some months have high demand of Bikes such as September.
   Some Season has high demand for Bikes such as Winter.
   Bad Weather which is Weather Type as 3 has negative effect on demand of Bikes .

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

   drop_first=True reduces the number of independent variables which needed to be dealt with while building the Linear Regression Model . drop_first=True would not change the resulting linear model . First column can still be explained by other Columns. So, it is good to dop the first column while creating dummy variables .

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

   tmp and atemp have the highest correlation with the target variable cnt.

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

   We validate the assumptions of Linear Regression by drawing the histogram of error terms.
   And we observe that histogram follows a normal distribution also known as Gaussian distribution .

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

   The top 3 features contributing significantly towards explaining the demand of the shared bikes are temperature, year, and weather .

General Subjective Questions

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

   Linear regression algorithm is basically focused on developing a best possible linear equation which can explain the relationship between the dependent variables and independent variables .
   In this algorithm , we study the relationship between dependent variable and independent variables and select the most important variables which influence the target variable positively or negatively .

   Best fit line is drawn based on where the sum of the squared differences is least .

First , we perform following steps in Sequence.

1) Data cleaning steps such as removing invalid values or filling invalid values with appropriate values , removing unwanted columns , removing rows if applicable .
2) After performing cleanup , we introduce dummy variables as applicable .
3) Now, we divide the dataset into training and test data set .
4) We need to scale some values as some column values may be either too small or too big.
5) After scaling , we built the linear model using training dataset . We use the recursive feature elimination.
6) Check the model stats.
7) Remove insignificant variables with p values less than 0.05.
8) Check VIF for multicollinearity.
9) Remove variables with VIF less than 7. Selection of cut-off VIF can vary .
10) Check whether error terms between y_train and y_pred is normal distribution .
11) Execute Linear model with Test Data Set

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

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

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

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

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