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?
   Ans: Categorical variables such as 'Season' , 'Weathersit' and 'Mnth' are highly correlated to dependent variable 'cnt'.

2. Why is it important to use drop_first=True during dummy variable creation?
   Ans: Generally, we create n number of dummy variables for n unique values. We use 'drop_first' to delete the first dummy variable because we can still interrupt the values by deleting one of the dummy variables.
   It is used to keep the dataset as small as possible.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans: Perform Residual analysis of the error terms

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: Top 3 features are yr(Year), holiday and temp(temperature)

General Subjective Questions

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

Ans: It is defined as relationship between dependent variable (Y) and one or more independent variable(X) to form a straight line.

$Y = mX + c$

Y is the target variable

m is the coefficient

X is the dependent variable

C is the intercept

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

Ans: Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

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

Ans: The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

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

Ans: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Mostly data set contains features highly varying in the magnitudes, units and ranges. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization Scaling: It brings all of the data in the range of 0 and 1.

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Ans: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.

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 is a graphical plotting of the quantiles of two distributions with respect to each other. In other words, we can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line. We also call it the 45-degree line in statistics. It entails that each of our distributions has the same quantiles.