## 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 categorical variables like weathersit, workingday, season etc have a reasonable effect on the dependent variable.

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

- The drop_first=True option makes sure there is always n-1 dummy variables created for a given categorical variable.

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

- The registered 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)

- The distribution plot (sns.displot) helped validate the assumptions after the building the model. The graph was normalized.

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 are the temperature, weather and the registered users over the years.

## General Subjective Questions

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

- Linear Regression is a machine learning algorithm that is used in supervised learning. It performs the task of predicting a dependent or target variable based on one or more independent variables.

The Linear regression is called simple linear regression if there is a single input variable. It is called multiple linear regression of there are more than one input variables.

The algorithm gives a sloped straight line which describes the relationship between the variables.

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

- Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four dataset and each dataset consists of eleven (x, y) points. The basic thing to analyze about these datasets is that they all share the same descriptive statistics (mean, variance, standard deviation etc.) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

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

- Pearson's R is a measure of linear correlation between two sets of data. It indicates how far away all these data points are to this line of best fit.

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

- Scaling is bringing all variables to a common scale. In regression, it is often recommended to center the variables so that the predictors have mean 0. This makes it easier to interpret the intercept term as the expected value of Yi when the predictor values are set to their means. Otherwise, the intercept is interpreted as the expected value of Yi when the predictors are set to 0, which may not be a realistic or interpretable situation.

**Normalized** scaling rescales the values between 0 and 1. It is also known as min-max scaling.

**Standardized** scaling rescales the data to have a mean of 0 and a standard deviation of 1. Hence it gives a normalized graph.

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

- If VIF is infinite, it indicates that there is a perfect correlation between the variables. When there is perfect correlation, R2 (r-squared) is 1. Hence VIF= 1/(1-R2) will lead to infinity.

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

- The Q-Q plot (Quantile-Quantile plot) is a graphical technique for determining if two data sets come from a populations with common distribution.
  The advantages of the q-q plot are:
  1. The sample sizes do not need to be equal.
  2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot