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?
   The categorical variables like weekdays, weather condition and season played an important role in the count (target variable).

2. Why is it important to use drop_first=True during dummy variable creation?
   This is important so as to create a set of N-1 dummy variables for N categories. If this argument is not passed in the code, the code created N number of dummy variables. This makes one column redundant.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   The ‘atemp’ variable (feels like temperature) has the maximum correlation with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   By transforming the model on the test set. We matched the R2 score that we got in the training model (71.4%) as well as the test model (70.3%).

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   - Year
   - Windspeed
   - Season
   - Day of the week

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Reading and understanding the data
- Data Prep and EDA
- Creation of dummy variables for categorical columns
- Creating train and test data split
- Rescaling the features of other numerical columns so every column has value between 0 and 1 for the sake of comparison
- Choosing multiple models and choosing the correct one from them
- Residual analysis and predictions on the test set
- Evaluating the usability of the data by calculating R2 for test set and ensuring it is close to that of the training dataset

2. Explain the Anscombe's quartet in detail.

   The Anscombe's quartet is a set of 4 datasets which are identical (mean, variance, std dev) but there are some differences in the dataset that can fool the regression model. It is used to explain the importance of looking at a set of data graphically before analysing.

3. What is Pearson's R?

   Common way of measuring a linear correlation. It's value ranges from -1 to 1. When a variable changes, its correlation ensures that the other variable changes in the same direction.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   Scaling is to ensure that when we are creating a model, the values of different numerical categories fall in the same range so that the regression model can make sense. Normalization is performed by using the Minmax scaler whose value can fall between 0 and 1.
   Standardisation scaling – mean is 0 and Std dev is 1.

5. You might have observed that sometimes the value of VIF is infinite.
   VIF is used to calculate the correlation between a variable and the others variables in the dataset. The higher the correlation, the higher the VIF. If the correlation is perfect, the value for VIF will be infinite.

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

Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If these datasets exhibit normal behaviour, they tend to form a straight line. If the line is curvy at edges, it means the data does not exhibit as normal behaviour as expected.