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)
   - It is interpreted similar as a continuous variable. As seen from the model analysis, categorical variables like weather, year, month, season has strong correlation with the target variable

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   - It is to avoid curse of dimensionality or multicollinearity issue

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

   - temp, atemp

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

   - using the residual analysis , checking for residual normality and patterns in error terms

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

   - year, atemp

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   - Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting

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

- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

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

   - It measures the linear association (correlation) between two variables.

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

   - scaling is a way of standardizing the data, it is performed to bring all continuous variables onto a single unit/scale. Normalized scaling uses min_max scaler, while Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation.

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

- 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)

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.