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?**

   Answer:
   Season, weather, and workdays impact bike rentals. More rentals in spring/summer than winter, good weather leads to more rentals, and weekdays see higher rentals than weekends.

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

   Answer:
   Using drop_first=True in dummy variable encoding helps avoid multicollinearity by dropping the first category. This maintains the independence of dummy variables and preserves the integrity of the linear regression model.

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

   Answer:
   Among the numerical variables, temperature typically shows the highest correlation with the target variable. As temperature increases, the number of bike rentals tends to increase, indicating a strong positive correlation.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   Answer:
   The assumptions of linear regression were validated by:
   - Checking for linearity through scatter plots or Pair Plot.
   - Assessing multicollinearity through Variance Inflation Factor (VIF) values.
   - Verifying Residuals and Checking Error Terms

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?**

   Answer:
   The top 3 features contributing significantly towards explaining bike sharing demand are:
   - Temperature (temp): Higher temperatures generally lead to more bike rentals.
   - Season: Bike rentals depend upon season, in winter we have less rental bikes.
   - Weather situation (weathersit): Favorable weather conditions (e.g., clear skies) result in increased bike usage.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Answer:
   Linear regression is a statistical method that models the relationship between a dependent variable and one or more independent variables. The algorithm aims to find the best-fitting line that minimises the sum of squared differences between the observed and predicted values.

   The basic formula is $Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \varepsilon$,
   where Y is the dependent variable, β0 is the intercept, β1 to βn are the coefficients for the independent variables X1 to Xn, and ε is the error term.

2. **Explain the Anscombe's quartet in detail.**

   Answer:
   Anscombe's quartet consists of four datasets with nearly identical simple descriptive statistics (mean, variance, correlation) that reveal different distributions and relationships when graphed. It illustrates the importance of visualizing data before analyzing it and shows that relying solely on summary statistics can be misleading. The quartet emphasises the necessity of plotting data to understand its underlying structure and detect anomalies.

3. **What is Pearson's R?**

   Answer:
   Pearson's R, or Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. Its value ranges from -1 to 1, where 1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation. Pearson's R is used to determine how strongly the variables are related.

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

   Answer:
   Scaling adjusts the range of features to a standard range, typically 0 to 1 or a normal distribution with a mean of 0 and a standard deviation of 1. It is performed to ensure that no single feature dominates the model due to its scale and to improve the convergence of gradient-based optimization algorithms. Normalized scaling rescales the data to a range of [0, 1] or [-1, 1], while standardized scaling transforms the data to have a mean

of 0 and a standard deviation of 1.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

   Answer:
   VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity, meaning one predictor variable can be perfectly predicted from the others. This occurs when there is an exact linear relationship between variables, causing the regression model to fail due to singularity or redundancy among the predictors.

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

   Answer:
   A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a dataset follows a particular distribution, typically a normal distribution. It compares the quantiles of the data against the quantiles of a theoretical distribution. In linear regression, Q-Q plots are used to check the normality of residuals. If the residuals are normally distributed, the points will lie approximately on a straight line. This validation is crucial as the normality of residuals is an assumption for many inferential statistics in linear regression.