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)**
   - Holiday: It has a significant negative coefficient (-737.4982, p-value < 0.001), suggesting that holidays tend to decrease bike rentals.
   - Weekday: The coefficient (18.5348) is positive, but not statistically significant (p-value = 0.298), indicating weekdays might not strongly influence bike rentals.
   - Workingday: It shows a negative coefficient (-168.9381, p-value = 0.033), indicating fewer rentals on working days.
   - Seasons (season_2, season_3, season_4): All have positive coefficients (significant with p-values < 0.001), indicating higher rentals in seasons 2, 3, and 4 compared to season 1.
   - Year (yr_1): Positive coefficient (1954.2960, p-value < 0.001) indicates an increasing trend in bike rentals over the years.
   - Weather conditions (weathersit_2, weathersit_3): Both have negative coefficients (weathersit_2: -489.2353, weathersit_3: -1831.4602, both p-values < 0.001), indicating fewer rentals during poor weather conditions.
   -

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

   Using drop_first=True during dummy variable creation is crucial to avoid the dummy variable trap, which can occur due to multicollinearity. When all dummy variables for a categorical feature are included, they can be perfectly collinear, leading to redundancy. Dropping the first category removes this redundancy, ensuring the model coefficients are interpretable and the matrix operations are stable.

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

   Based on the regression summary, none of the numerical variables (like temp, atemp, hum, windspeed) have notably high correlations explicitly listed. However from the EDA it is assumed that the numerical variable temp (temperature) has the highest correlation with the target variable cnt (total bike rentals). This indicates that higher temperatures are associated with increased bike rentals.

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

   The assumptions of Linear Regression were validated by:

   - **Residual Plot: Examined for homoscedasticity and linearity around zero.**

- **Q-Q Plot: Checked for normality of residuals against a theoretical normal distribution.**
- **Histogram of Residuals: Assessed the overall distribution of residuals for normality and spread.**

These analyses collectively verify the assumptions of Linear Regression, including normality, homoscedasticity, and the absence of significant patterns in residuals.

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 contributing significantly towards the demand for shared bikes are:

   - temp (temperature): Warmer temperatures lead to increased bike rentals.
   - yr (year): Indicates an overall increase in bike demand over time.
   - hum (humidity): Higher humidity levels tend to decrease bike rentals.

General Subjective Questions

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

   Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The objective is to find the best-fitting straight line (linear relationship) described by the equation y=β0+β1x1+β2x2+…+βnxn+ϵy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilony=β0+β1x1+β2x2+…+βnxn+ϵ, where yyy is the dependent variable, β0\beta_0β0 is the intercept, β1,β2,…,βn\beta_1, \beta_2, \ldots, \beta_nβ1,β2,…,βn are the coefficients for the independent variables, and ϵ\epsilonϵ is the error term. The coefficients are estimated using methods like Ordinary Least Squares (OLS), which minimizes the sum of squared residuals (the differences between observed and predicted values).

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

   Anscombe's quartet consists of four datasets that have nearly identical simple statistical properties (mean, variance, correlation, etc.) but differ greatly when graphed. This underscores the importance of graphing data to uncover patterns that summary statistics might miss. Despite having similar statistical properties, each dataset in Anscombe's quartet reveals different relationships and insights when visualized, demonstrating how reliance on statistical summaries alone can be misleading.

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

   Pearson's R is a measure of the linear correlation between two variables, ranging from -1 to 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. It quantifies the degree to which two variables are linearly related, providing insight into the strength and direction of their relationship.

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

   Scaling adjusts the range of feature values to a common scale, ensuring that all features contribute equally to the model. This is crucial for algorithms that rely on distance measurements, such as linear regression. Normalized scaling (min-max scaling) rescales features to a range [0, 1], whereas standardized scaling (z-score normalization) centers the data around zero with a standard deviation of one. Scaling enhances model performance and convergence speed.

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

   An infinite VIF value occurs when there is perfect multicollinearity, meaning one independent variable is a perfect linear combination of others. This makes it impossible to estimate the regression coefficients uniquely, as the matrix inversion required in the calculation of VIF is not possible. This indicates redundancy among the variables, leading to instability in the regression model.

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

   A Q-Q (quantile-quantile) plot is a graphical tool to assess if a dataset follows a particular distribution, usually normal. In linear regression, a Q-Q plot of residuals helps verify the assumption that the residuals are normally distributed. Deviations from the diagonal line in a Q-Q plot indicate departures from normality, which can affect the validity of regression inferences.