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?**
   Categorical variables such as `season`, `yr`, `mnth`, `weekday`, `workingday`, `holiday`, and `weathersit` have varying effects on bike demand:
   - **Season:** Demand varies significantly across seasons, with higher demand in summer and fall compared to spring and winter.
   - **Year (`yr`):** The demand for shared bikes increases year over year, indicating growing popularity.
   - **Weather Situation (`weathersit`):** Clear weather conditions lead to higher bike demand, while adverse weather (like heavy rain or snow) reduces demand.
   - **Holiday and Working Day:** Bike usage tends to differ on holidays and working days, reflecting changes in commuting patterns.

2. **Why is it important to use `drop_first=True` during dummy variable creation?**
   Using `drop_first=True` avoids the dummy variable trap, which occurs when dummy variables created from categorical variables are highly correlated (multicollinear). This option drops the first category, providing k-1 dummies for k categories, ensuring that the model has a full-rank design matrix and avoiding multicollinearity.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   The variable `temp` (temperature) shows the highest correlation with the target variable `cnt` (total bike rentals), indicating that bike rentals increase with rising temperatures.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   To validate the assumptions of linear regression:
   - **Linearity:** Checked the scatter plots between predictors and the target variable.
   - **Normality:** Examined the residuals' distribution using histograms and Q-Q plots.
   - **Homoscedasticity:** Plotted residuals versus fitted values to ensure constant variance.
   - **Multicollinearity:** Calculated Variance Inflation Factor (VIF) to ensure no predictor had high multicollinearity.
   - **Independence:** Verified residual independence using the Durbin-Watson test.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?**
   The top 3 significant features are:
   - Temperature (`temp`)
   - Year (`yr`)
   - Weather Situation (`weathersit`): particularly clear weather

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Linear regression is a statistical method to model the relationship between a dependent variable y and one or more independent variables X. The goal is to find the best-fitting linear equation $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$, where:
   - $\beta_0$is the intercept,
   - $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients representing the relationship between each independent variable and the dependent variable,
   - ϵ is the error term.

   The algorithm minimizes the sum of squared errors (differences between observed and predicted values) to find the optimal coefficients. It assumes linearity, independence, homoscedasticity, normality, and no multicollinearity among predictors.

2. **Explain the Anscombe's quartet in detail.**
   Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. It demonstrates the importance of graphing data before analyzing it:
   - **Identical Statistics:** Mean, variance, correlation, and regression line are similar across all datasets.
   - **Different Graphical Representations:** Each dataset, when plotted, shows different patterns, outliers, and relationships, highlighting the necessity of visualizing data to uncover underlying structures.

3. **What is Pearson's R?**
   Pearson's correlation coefficient(r) measures the linear relationship between two variables, ranging from -1 to 1:
   - r = 1: Perfect positive linear relationship.
   - r = -1: Perfect negative linear relationship.
   - r = 0: No linear relationship.

   It quantifies the strength and direction of the linear relationship.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Scaling transforms data to fit within a specific range or distribution:
   - **Importance:** Ensures all features contribute equally to the model, particularly for algorithms sensitive to feature magnitudes (e.g., linear regression, SVM).
   - **Normalized Scaling:** Rescales features to a [0, 1] range, often using min-max scaling.
   - **Standardized Scaling:** Transforms features to have zero mean and unit variance, using z-score normalization.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   VIF (Variance Inflation Factor) measures multicollinearity among predictors. An infinite VIF occurs when a predictor is a perfect linear combination of other predictors, indicating perfect multicollinearity. This situation arises from redundant or highly correlated features.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   A Q-Q (Quantile-Quantile) plot compares the quantiles of a dataset's distribution with a theoretical distribution (usually normal). It assesses if data follows a specific distribution:
   - **Use in Linear Regression:** Validates the normality assumption of residuals. If residuals lie along the 45-degree line in the Q-Q plot, they are normally distributed.
   - **Importance:** Ensures reliable hypothesis testing and confidence intervals for model parameters.