Assignment-based Subjective Questions

1. From the analysis of categorical variables such as `season`, `weathersit`, `yr`, `mnth`, `holiday`, and `weekday`, it can be inferred that:
   - **Season**: The demand for shared bikes varies with the season, with higher demand typically observed in warmer seasons like summer and fall compared to colder seasons like winter.
   - **Weathersit**: Weather conditions significantly impact bike demand. Clear or partially cloudy days see higher bike usage, while adverse weather conditions like heavy rain or snow result in lower demand.
   - **Year (yr)**: The year variable indicates a trend in bike usage over time. An increase in bike demand is often seen in the second year (2019) compared to the first year (2018), reflecting growing popularity.
   - **Month (mnth)**: Monthly trends can show variations in bike demand, with certain months like July and August having higher demand due to favorable weather.
   - **Holiday**: Demand may drop on holidays since fewer people commute to work.
   - **Weekday**: The day of the week can affect demand, with weekdays generally showing higher demand due to commuting patterns compared to weekends.

2. Using `drop_first=True` is important because it helps to avoid the dummy variable trap. The dummy variable trap occurs when the model includes all the dummy variables of a categorical feature, leading to multicollinearity (i.e., perfect correlation) among the variables. By dropping the first dummy variable, we avoid this issue and ensure that the model matrix is of full rank, making the model estimation more stable.

3. From the pair-plot among numerical variables, the variable `temp` (temperature) often shows the highest correlation with the target variable `cnt` (count of total bike rentals). Warmer temperatures generally lead to higher bike usage.

4. After building the model on the training set, the assumptions of linear regression were validated through:
   - **Residuals vs. Fitted Values Plot**: This checks for homoscedasticity, ensuring residuals have constant variance.

- **Distribution of Residuals**: A histogram or density plot of residuals checks for normality.
- **Q-Q Plot**: This plot checks if residuals follow a normal distribution.
- **Residuals Over Time**: This ensures residuals are independent over time and not autocorrelated.
- **Variance Inflation Factor (VIF)**: This checks for multicollinearity among the predictors.

5. Based on the final model, the top 3 features significantly contributing to the demand for shared bikes typically include:
   - **temp**: Temperature, as warmer weather encourages biking.
   - **yr**: Year, indicating an upward trend in bike usage over time.
   - **season**: Season, with higher demand in warmer seasons.

General Subjective Questions

1. Linear regression is a supervised machine learning technique used to model the relationship between a dependent variable (Y) and one or more independent variables (X). It aims to find a linear equation that best fits the data points. The algorithm follows these steps:

- **Data Preparation:** Gather data with a dependent variable (what you want to predict) and independent variables (features influencing the prediction).
- **Model Representation:** The model is typically represented by a linear equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$, where:
  - Y is the predicted value of the dependent variable.
  - β_0 is the intercept (the Y-axis value when all X's are zero).
  - β_i (i = 1 to n) are the coefficients (slopes) for each independent variable X_i.
- **Loss Function (Cost Function):** This function measures the difference between predicted (Y) and actual values. The most common loss function is the Least Squares method, which minimizes the sum of squared residuals (differences between predicted and actual Y).

- **Optimization:** The algorithm iteratively adjusts the coefficients (β) to minimize the loss function. This can be achieved using various optimization algorithms like gradient descent.
- **Evaluation:** Metrics like R-squared (goodness of fit) and p-value (statistical significance) are used to assess the model's performance.

2. Anscombe’s quartet consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but very different distributions and graphical representations. The quartet demonstrates the importance of visualizing data before drawing conclusions based solely on statistical measures. The four datasets illustrate how datasets with the same statistical properties can have different trends, outliers, and patterns, highlighting the necessity of graphical analysis in addition to numerical analysis.

3. Pearson’s R, also known as the Pearson correlation coefficient, measures the linear correlation between two variables. Its value ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

The formula for Pearson’s R is:

$r=\sum(xi-\bar{x})2\sum(yi-\bar{y})2\ \sum(xi-\bar{x})(yi-\bar{y})$

where xi and yi are the individual sample points and $\bar{x}$ and $\bar{y}$ are the means of the variables.

4. Scaling refers to transforming data to fit within a specific range, typically to improve the performance of certain machine learning algorithms. Scaling ensures that features contribute equally to the result, preventing features with larger ranges from dominating the model.

- **Normalized Scaling**: Transforms data to fit within a range, typically [0, 1].
- **Standardized Scaling**: Centers data around the mean with a standard deviation of 1.

The main difference is that normalization adjusts the range of the data, while standardization adjusts the scale based on the distribution of the data.

5. The Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity, meaning one predictor variable is a perfect linear combination of other predictor variables. In such cases, the regression model cannot compute the coefficient estimates uniquely, resulting in an undefined VIF - often represented as infinite.

6. A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a dataset follows a particular distribution, typically the normal distribution. It plots the quantiles of the dataset against the quantiles of the specified distribution. If the points lie approximately along a straight line, the dataset is likely to follow the specified distribution. In linear regression, Q-Q plots are used to check if the residuals are normally distributed, validating one of the assumptions of the regression model.