Assignment-based Subjective Questions

1. **Effect of Categorical Variables:**

   Typically, categorical variables like `season`, `month`, `weekday`, and `weather condition` affect bike rental demand. For example, the season may indicate that rents are greater in the summer and fall owing to pleasant `weather conditions`. On days with significant rain or snow, weather conditions may signal reduced rental rates.

2. **Importance of drop_first in Dummy Variable Creation:**

   Using drop_first=True reduces multicollinearity in the model by lowering the number of dummy variables. Each category, minus one, is adequate to gather all relevant information. This helps to simplify the model and avoid the dummy variable trap.

3. **Numerical Variable with Highest Correlation:**

   Typically, correlation coefficients or pair-plots are used to determine this. Typically, factors such as temperature have a strong positive association with bike rentals, implying that higher temperatures lead to more rentals.

4. **Validation of Linear Regression Assumptions:**

   Linearity: Check scatter plots of residuals vs. anticipated values to ensure linearity.
   Independence: To determine independence, use the Durbin-Watson statistic to check for residual autocorrelation.
   Homoscedasticity: Identify homoscedasticity by seeing consistent residual variance across forecasts.
   Normality: Use Q-Q plots to ensure residuals are regularly distributed.

5. **Top 3 Features Contributing to Demand:**

   The model's coefficients and p-values suggest that top characteristics may include temperature, year (indicating increase over time), and seasonal or month-specific rental peaks.

General Subjective Questions

1. **Linear Regression Algorithm:**

   Linear regression uses an independent variable (x) to predict the value of a dependent variable, (y). It achieves this by finding the coefficients of the linear equation, incorporating one or more independent variables that best predict the dependent variable. ($y = \beta 0 + \beta 1x1 + ... + \beta nxn$). The least squares approach is used to decrease prediction error.

2. **Anscombe’s Quartet:**

   This collection of four datasets has almost similar simple statistical properties, however they differ dramatically when graphed. Each dataset highlights the importance of showing data before analyzing it, as well as how outliers and other pertinent observations affect statistical aspects.

3. **Pearson’s R:**

   Pearson's correlation coefficient (r) indicates a linear relationship between two variables. It runs from -1 to 1, with 1 indicating perfect positive linear correlation, -1 indicating perfect negative linear correlation, and 0 indicating no linear correlation.

4. **Scaling:**

   - **Purpose**: Scaling is used to normalize or standardize data, making it easier to compare and analyze, particularly in algorithms that rely on input data scale, such as K-means clustering or gradient descent.
   - **Normalized Scaling (Min-Max Scaling)**: Transforms features to a scale between 0 and 1.
   - **Standardized Scaling (Z-score Scaling)**: Features are scaled so that the mean is 0 and the standard deviation is 1.

Infinite VIF:

An infinite VIF (Variance Inflation Factor) shows perfect multicollinearity, which occurs when one independent variable is a perfect linear combination with another. It indicates duplication in data representation.

5. **Q-Q Plot:**

   A Q-Q plot compares a distribution's quantiles to the conventional normal distribution. In linear regression, it aids in verifying the assumption that residuals are normally distributed, which is critical for the reliability of many statistical tests and confidence ranges in the model.