Answers for Assignment-based Subjective Questions

1. From the analysis of categorical variables, we can infer their effect on the dependent variable as follows:

    - **Season**: Different seasons may have varying effects on bike rentals. For example, the demand for bikes might increase during warmer seasons like summer and decrease during colder seasons like winter.
    - **Month**: Certain months may exhibit seasonal trends or events that influence bike rentals. For instance, bike rentals might be higher during vacation months or summer holidays.
    - **Weekday**: Weekdays versus weekends may show distinct patterns in bike rentals. Typically, weekdays might see more bike rentals for commuting purposes, while weekends might have higher leisure-based rentals.
    - **Weather Situation**: Weather conditions like clear skies, cloudy weather, or precipitation can significantly impact bike rentals. For example, unfavorable weather conditions such as rain or snow might deter people from renting bikes, leading to a decrease in demand.

2. It is important to use `drop_first=True` during dummy variable creation to avoid multicollinearity issues and the dummy variable trap. By dropping the first dummy variable, we ensure that the information contained in the other dummy variables is sufficient to represent all categories of the original categorical variable. This prevents perfect multicollinearity and makes the regression model more interpretable and stable.

3. From the pair-plot among the numerical variables, the variable with the highest correlation with the target variable (dependent variable) is identified as the one that exhibits the strongest linear relationship. This can be determined by visually inspecting the pair-plot or by calculating the correlation coefficients between each numerical variable and the target variable. The numerical variable with the highest correlation coefficient is considered to have the highest correlation with the target variable.

4. To validate the assumptions of Linear Regression after building the model on the training set, several diagnostic tests and checks can be performed:

    - **Residual Analysis**: Check the distribution of residuals to ensure they are normally distributed and centered around zero. Additionally, plot residuals against predicted values to detect any patterns or heteroscedasticity.
    - **Collinearity Check**: Examine the variance inflation factor (VIF) to identify multicollinearity issues among predictor variables.
    - **Homoscedasticity**: Assess whether the residuals exhibit constant variance across different levels of the predictors.
    - **Linearity**: Validate the assumption of linearity by plotting observed values against predicted values and ensuring that they form a linear pattern.
    - **Independence of Errors**: Check for autocorrelation by analyzing the Durbin-Watson statistic or plotting residuals against time or other relevant variables.

5. Based on the final model, the top three features contributing significantly towards explaining the demand for shared bikes are determined by examining the coefficients or importance scores assigned to each feature. The features with the highest coefficients or importance scores indicate their relative influence on the target variable.

In this case, the top three features contributing significantly to explaining bike demand would be those with the highest coefficients in the regression model or the highest feature importance scores in the chosen machine learning algorithm.

Answers for General Subjective Questions

1. **Linear Regression Algorithm:**

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. Here's a detailed explanation:

- **Assumptions:** Linear regression assumes that there is a linear relationship between the independent variables and the dependent variable, the errors are normally distributed, the errors have constant variance (homoscedasticity), and the errors are independent of each other.
- **Model Representation:** The linear regression model can be represented as:

```
Y = β0 + β1*X1 + β2*X2 + ... + βn*Xn + ε
```

  where Y is the dependent variable, X1, X2, ..., Xn are the independent variables, β0 is the intercept term, β1, β2, ..., βn are the coefficients corresponding to each independent variable, and ε is the error term.

- **Objective:** The objective of linear regression is to find the coefficients (β0, β1, β2, ..., βn) that minimize the sum of squared differences between the actual values and the predicted values.
- **Cost Function:** The cost function, also known as the loss function, is typically the mean squared error (MSE) or the sum of squared errors (SSE), which measures the average squared difference between the actual and predicted values.
- **Optimization:** The coefficients are estimated using optimization algorithms such as gradient descent or analytical methods like the ordinary least squares (OLS) method.
- **Evaluation:** Linear regression models are evaluated using metrics like R-squared (coefficient of determination), adjusted R-squared, mean squared error (MSE), root mean squared error (RMSE), etc., to assess how well the model fits the data.
- **Interpretation:** After fitting the model, the coefficients provide insights into the relationship between the independent and dependent variables. For example, the coefficient β1 represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant.

2. **Anscombe's Quartet:**

Anscombe's quartet consists of four datasets that have nearly identical statistical properties, including mean, variance, correlation coefficient, and linear regression line. However, upon visual inspection, these datasets are significantly different from each other. This quartet highlights the importance of visualizing data before drawing

conclusions solely based on summary statistics. It was created by the statistician Francis Anscombe in 1973 to emphasize the limitations of relying solely on numerical summaries.

3. **Pearson's R:**

   Pearson's cor1relation coefficient (Pearson's R) is a measure of the linear relationship between two variables. It ranges from -1 to 1, where:

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

   Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It is sensitive to outliers and assumes that the relationship between variables is linear.

4. **Scaling:**

   Scaling is the process of transforming data such that it fits within a specific scale, usually to make the data comparable or to improve the performance and convergence of certain machine learning algorithms.

   - **Why Scaling is Performed:** Scaling is performed to ensure that the features contribute equally to the analysis. It helps in cases where the features have different units or different ranges of values, preventing one feature from dominating others.
   - **Difference Between Normalized Scaling and Standardized Scaling:**
     - **Normalized Scaling:** Normalization scales the values of features to a fixed range, often between 0 and 1. It is done by subtracting the minimum value from each observation and dividing by the range (maximum value minus minimum value).
     - **Standardized Scaling:** Standardization (also known as z-score normalization) transforms the data such that it has a mean of 0 and a standard deviation of 1. It is done by subtracting the mean from each observation and dividing by the standard deviation.

5. **Infinite Value of VIF:**

   The Variance Inflation Factor (VIF) measures the extent of multicollinearity in a regression analysis. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other. If there is perfect multicollinearity, where one independent variable can be perfectly predicted by another, it leads to an infinite value of VIF. This situation typically arises when one independent variable is a linear combination of other independent variables.

6. **Q-Q Plot:**

   A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a given set of data follows a particular distribution, such as the normal distribution. It compares the quantiles of the observed data to the quantiles of a theoretical distribution. In linear regression, Q-Q plots are often used to check the assumption of normality of residuals. If the residuals follow a normal distribution, the points on the Q-Q plot will fall approximately along a

straight line. If there are deviations from the straight line, it suggests departures from normality, which may indicate that the linear regression model is not appropriate for the data. Q-Q plots are important for validating the assumptions of linear regression and ensuring the reliability of the model's results.