Assignment-based Subjective Questions

1. **Inference about Categorical Variables:** From the analysis of categorical variables in the dataset, we can infer their effect on the dependent variable (demand for shared bikes) by examining the coefficients of the corresponding dummy variables. Positive coefficients indicate a positive effect on bike demand, while negative coefficients indicate a negative effect. Large coefficient magnitudes suggest a stronger influence.

2. **Importance of drop_first=True in Dummy Variable Creation**: Using drop_first=True during dummy variable creation is important to avoid multicollinearity issues. When you have a categorical variable with 'n' categories, creating 'n' dummy variables without dropping one (i.e., using drop_first=False) would create perfect multicollinearity because one category can be predicted perfectly from the others. This can cause problems in regression analysis, and it's better to drop one category as a reference level to avoid multicollinearity.

3. **Highest Correlation with the Target Variable**: To determine which numerical variable has the highest correlation with the target variable ('cnt'), you can calculate the correlation coefficients (Pearson correlation) between 'cnt' and each numerical variable. The one with the highest absolute correlation coefficient is the most correlated with 'cnt'.

4. **Validation of Linear Regression Assumptions**: To validate the assumptions of linear regression after building the model on the training set, you can perform the following steps:
   - Residual Analysis: Check the residuals (the differences between actual and predicted values) for normality, linearity, and homoscedasticity using diagnostic plots like Q-Q plots, residual vs. fitted value plots, and residual vs. predictor plots.
   - Multicollinearity: Calculate the Variance Inflation Factor (VIF) for each predictor variable to check for multicollinearity. High VIF values indicate multicollinearity.
   - Independence of Residuals: Check for autocorrelation in residuals using autocorrelation plots or Durbin-Watson statistic.
   - No Heteroscedasticity: Plot residuals vs. predicted values to check for constant variance. You can also perform statistical tests like the Breusch-Pagan test or White test.

5.**Top 3 Features Contributing to Bike Demand**: To identify the top 3 features contributing significantly to explaining the demand for shared bikes, you can examine the coefficients of the predictors in your final linear regression model. The predictors with the largest positive coefficients have the most significant positive effect on bike demand, while predictors with large negative coefficients have the most significant negative effect. You can rank them based on the absolute magnitude of their coefficients.

General Subjective Questions

1. **Explain the linear regression algorithm in detail**:
   - Linear regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (dependent variable) based on one or more predictor variables (independent variables) that are assumed to have a linear relationship with the outcome.
   - The simplest form is Simple Linear Regression, which models the relationship between one independent variable (X) and one dependent variable (Y) using a linear equation: $Y = \beta_0 + \beta_1 X + \varepsilon$.
   - Multiple Linear Regression extends this concept to multiple predictor variables: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta \quad X \quad + \varepsilon$.
   - The goal is to estimate the coefficients (β_0, β_1, β_2, ...) that minimize the sum of squared differences between predicted and actual values (residuals).

- Common methods to estimate these coefficients include Ordinary Least Squares (OLS) and Gradient Descent.
- Linear regression assumes that the relationship between predictors and the response variable is linear, there is no multicollinearity among predictors, and residuals are normally distributed and have constant variance.

2. **Explain the Anscombe's quartet in detail**:
- Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics but appear very different when graphed.
- Each dataset consists of 11 (x, y) data points.
- It was created by statistician Francis Anscombe to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics.
- The four datasets have different patterns: one has a linear relationship, another has a non-linear relationship, another is mostly constant with an outlier, and the last has an outlier that greatly influences the regression line.
- Anscombe's quartet serves as a reminder that graphical exploration of data is crucial to understand underlying patterns, even when summary statistics are similar.

3. **What is Pearson's R:**
- Pearson's correlation coefficient (or Pearson's R) is a statistic that measures the linear relationship between two continuous variables. It quantifies the strength and direction (positive or negative) of the linear association.
- It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.
- Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations.
- Formula: $R = (\Sigma[(X_i - \bar{X})(Y_i - \bar{Y})]) / [n * \sigma(X) * \sigma(Y)]$
- It is widely used in statistics and data analysis to determine how strongly two variables are related linearly.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling**:
- Scaling is the process of transforming data into a standard range or distribution to ensure that all variables have the same scale. It's performed to make sure that the units or magnitude of different variables don't affect their contributions to a model disproportionately.
- Scaling is essential for some machine learning algorithms (e.g., gradient descent) that are sensitive to the scale of input features.
- Two common scaling methods are:
  - Normalized Scaling: This scales the data to a range of [0, 1]. It's useful when you want to preserve the relationships between values and have outliers that you don't want to remove.
  - Standardized Scaling (Z-score scaling): This scales the data to have a mean of 0 and a standard deviation of 1. It centers the data around 0 and is useful when you want to compare variables on a common scale.
- The main difference is the scale they transform data to; normalized scaling uses [0, 1], while standardized scaling uses the z-score scale.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen**:
- Variance Inflation Factor (VIF) measures multicollinearity in regression analysis. A high VIF indicates that an independent variable is highly correlated with other independent variables in the model.
- VIF can become infinite (or extremely high) when there is perfect multicollinearity, which means one independent variable is a perfect linear combination of others.
- This can happen when you include redundant variables in the model, leading to multicollinearity issues. To resolve this, you need to remove one of the correlated variables.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**:

- A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess if a dataset follows a specific theoretical distribution (e.g., normal distribution).
- In a Q-Q plot, the quantiles (ordered data values) of the observed data are plotted against the quantiles of the expected distribution.
- If the data follows the expected distribution, the points in the Q-Q plot should lie along a straight line (the identity line).
- The Q-Q plot is essential in linear regression for checking the assumption of normally distributed residuals. If the residuals follow a normal distribution, the points in the Q-Q plot will be close to the identity line.
- Deviations from the line indicate departures from normality, which can affect the validity of regression results. It helps identify potential problems with the model, such as outliers or heteroscedasticity.