Part 1: Analysis Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   The analysis of categorical variables, such as 'season' and 'weathersit', indicates that these variables have a significant impact on the demand for shared bikes. The data shows seasonal variations, with higher demand during certain seasons like summer and fall, and lower demand during winter. Similarly, the weather situation affects bike demand, with clear weather conditions being associated with higher demand compared to adverse weather conditions like heavy snow or rain.

2. **Why is it important to use `drop_first=True` during dummy variable creation? (2 marks)**

   Using `drop_first=True` during dummy variable creation is important to avoid multicollinearity, which occurs when independent variables are highly correlated. By dropping the first category, we avoid the dummy variable trap, ensuring that the resulting dummy variables are independent and linearly separable. This improves the stability and interpretability of the regression model.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   The pair-plot among the numerical variables indicates that 'registered' has the highest correlation with the target variable 'cnt'. This suggests that the number of registered users is a strong predictor of the overall bike demand.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   After building the model, the assumptions of Linear Regression were validated through residual analysis:

   - **Linearity**: Scatter plots of residuals vs. predicted values were used to check for any patterns. The absence of patterns indicates linearity.
   - **Normality**: A Q-Q plot and a histogram of residuals with a kernel density estimate (KDE) were used to check for the normality of residuals. The residuals followed a normal distribution, indicating normality.
   - **Homoscedasticity**: The scatter plot of residuals vs. predicted values was used to check for constant variance. The residuals were randomly scattered around zero, indicating homoscedasticity.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

   Based on the final model and the coefficients obtained from Lasso Regression, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

  - 'registered': This feature had the highest coefficient, indicating its strong positive impact on bike demand.
  - 'temp': The temperature feature also had a significant positive coefficient, showing that higher temperatures are associated with increased bike demand.
  - 'hum': Humidity had a negative coefficient, indicating that higher humidity levels are associated with decreased bike demand.

Part 2: Theoretical Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression is a supervised learning algorithm used to predict a continuous target variable based on one or more predictor variables. The objective is to determine the best-fitting linear relationship between the independent variables (X) and the dependent variable (y).

The linear regression model can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

where:

- (y) is the dependent variable.
- (\beta_0) is the intercept.
- (\beta_1, \beta_2, \ldots, \beta_n) are the coefficients.
- (x_1, x_2, \ldots, x_n) are the independent variables.
- (\epsilon) is the error term.

The coefficients are estimated using the method of least squares, which minimizes the sum of squared residuals (differences between observed and predicted values). Key assumptions include linearity, independence, homoscedasticity, and normality of residuals.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

   Anscombe's quartet is a set of four different datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but exhibit distinct patterns when graphed. It demonstrates the importance of graphical analysis of data alongside statistical summaries. The quartet emphasizes that relying solely on summary statistics can be misleading and that visualizing data can reveal underlying structures, outliers, and anomalies that are not apparent from the statistics alone.

3. **What is Pearson's R? (3 marks)**

   Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It 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.

Pearson's R is calculated as the covariance of the variables divided by the product of their standard deviations. It provides insight into the strength and direction of the linear relationship between the variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Scaling involves transforming the features of a dataset to a common scale, which is crucial for algorithms sensitive to feature magnitudes, such as linear regression and distance-based algorithms.

- **Normalized Scaling**: This method rescales the features to a range of [0, 1] or [-1, 1] using the formula:

$$X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

  It is useful when the data needs to be bounded within a specific range.
- **Standardized Scaling**: This method transforms the features to have a mean of 0 and a standard deviation of 1 using the formula:

$$X' = \frac{X - \mu}{\sigma}$$

  It is useful for algorithms that assume data is normally distributed and for dealing with features of different units.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

   The Variance Inflation Factor (VIF) quantifies the level of multicollinearity in a set of multiple regression variables. A VIF value becomes infinite when there is perfect multicollinearity, meaning that one predictor variable is a perfect linear combination of other predictor variables. This results in an undefined or infinite VIF, indicating that the predictor is redundant and should be removed from the model to avoid instability and inaccuracies in the regression coefficients.

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 is a graphical tool used to assess if a dataset follows a particular distribution, typically the normal distribution. In a Q-Q plot, the quantiles of the dataset are plotted against the quantiles of the specified theoretical distribution. If the points lie approximately along a straight line, the data is consistent with the specified distribution.

   In linear regression, a Q-Q plot of the residuals is used to check the assumption of normality of the residuals. Normality of residuals is important because it affects the validity of hypothesis tests and confidence intervals for the regression coefficients. A

Q-Q plot helps identify deviations from normality, such as skewness or kurtosis, which may indicate the need for data transformation or alternative modeling approaches.