Assignment-based Subjective Questions

1. Effect of Categorical Variables: From the analysis of categorical variables, the inference on their effect on the dependent variable can be drawn based on factors like the distribution of categories, frequency counts, and possibly statistical tests like chi-square test for independence or ANOVA. For example, if certain categories within a variable are heavily skewed towards one outcome of the dependent variable, it suggests a potential influence.

2. Importance of drop_first=True: When creating dummy variables from categorical variables, it's important to avoid multicollinearity, where one predictor variable can be linearly predicted from the others with a high degree of accuracy. By setting drop_first=True, we drop one of the dummy variables created for each categorical variable, thus avoiding perfect multicollinearity and preventing the "dummy variable trap" in regression analysis.

3. Highest Correlation with Target Variable: By examining the pair-plot among numerical variables, the variable showing the highest correlation with the target variable can be identified visually. This is usually the variable with the strongest linear relationship with the target, indicated by a clear, upward or downward trend in the scatter plot.

4. Validation of Linear Regression Assumptions: After building the linear regression model on the training set, assumptions such as linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of residuals are validated. This can be done through techniques like residual analysis, scatter plots of residuals against predicted values, Q-Q plots, and statistical tests like Shapiro-Wilk test for normality.

5. Top 3 Features Contributing Significantly: Based on the final model, the top three features contributing significantly towards explaining the demand of shared bikes can be determined by examining the coefficients of the regression model. Features with larger absolute coefficients indicate stronger influence on the dependent variable.

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. It assumes a linear relationship between the variables and estimates the coefficients of the linear equation using methods like ordinary least squares (OLS) to minimize the sum of squared errors.

2. Anscombe's Quartet: Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and relationships when graphed. It

demonstrates the importance of visualizing data and the limitations of relying solely on summary statistics.

3. Pearson's R: Pearson's correlation coefficient, denoted as r, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.

4. Scaling: Scaling is the process of transforming data so that it falls within a specific range. It is performed to standardize the range of independent variables or features of the data, which helps in improving the performance and convergence of machine learning algorithms. Normalized scaling scales the data to a range between 0 and 1, while standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.

5. Infinite VIF: The value of Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among predictor variables. This happens when one or more predictor variables can be perfectly predicted from the others, leading to an inflated estimate of the variance of the regression coefficients.

6. Q-Q Plot: A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a given data follows a particular distribution, usually the normal distribution. It compares the quantiles of the observed data with the quantiles of a theoretical distribution. In linear regression, Q-Q plots are used to check the assumption of normality of residuals, where deviations from a straight line indicate departures from normality.