Assignment-based Subjective Questions

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

   With each category assigned a coefficient representing its impact on the dependent variable compared to a reference category. Understanding these coefficients helps infer how different categories of a variable affect the predicted outcome in the model.

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

   In dummy variable creation for linear regression we avoid multicollinearity by omitting one categorical level to improve model interpretability and stability. So this ensures each dummy variable represents a distinct category's impact on the dependent variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Highest correlation with the target variable is with the cnt column.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   After building a Linear Regression model on the training set, assumptions are validated by checking residuals for linearity, normality, and constant variance; additionally, multicollinearity is assessed using VIF scores or correlation matrices.

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

   The top 3 features that affect the demand of shared bikes are Holiday, weekday and Temp.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression models the relationship between dependent variable y and independent variable X by fitting a linear equation. Assumptions which include linearity, independence of residuals, homoscedasticity, and normality of residuals. Coefficient provides insights into variable impacts, making it interpretable and widely applicable despite limitations like sensitivity to outliers and the assumption of linear relationships.

2. Explain the Anscombe’s quartet in detail.

   Anscombe's quartet comprises four datasets designed to have nearly identical summary statistics (mean, variance, correlation, regression line), yet visually and analytically they vary significantly: one is linear, another non-linear, the third with an outlier affecting the regression, and the fourth with an extreme outlier. This highlights the importance of graphical exploration in data analysis to detect nuances and anomalies that summary statistics alone might obscure, emphasizing the limitations of purely numerical summaries in understanding complex data relationships and patterns.

3. What is Pearson’s R?

   Pearson's R is a statistic that measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1, where r=1 indicate a perfect positive linear relationship, r = −1 indicates a perfect negative linear relationship, and r = 0 indicates non linear relationship. It is widely used in statistics to assess correlation between variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   Scaling in linear regression involves transforming variables to a consistent range to prevent large scale variables from dominating the model. Normalization scales variables to a range maintaining relative relationship among variables. Standardization adjusts variables to have a mean of zero and a standard deviation of one, facilitating comparison of variable importance through standardized coefficients. These techniques aid in improving the stability and interpretability of the regression model.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

   In linear regression, perfect multicollinearity occurs when one or more independent variables are perfectly predictable from a linear combination of other variables. This leads to a situation where the correlation matrix of the variables is singular or near-singular, resulting in infinite values in the VIF calculation. Infinite VIF values indicate that the affected variables are redundant in the model, as their effects cannot be uniquely separated from each other. Addressing perfect multicollinearity involves identifying and possibly removing one of the correlated variables to stabilize the regression model's estimates.

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 compares the quantiles of observed data against those of a theoretical distribution (e.g., normal). In linear regression, it is used to assess whether residuals (differences between observed and predicted values) are normally distributed. A straight line on the Q-Q plot suggests residuals follow the assumed distribution, validating the regression model's statistical assumptions.