Assignment-based Subjective Questions

1. Effect of Categorical Variables on the Dependent Variable:

From the analysis of categorical variables, we might infer that factors such as `season`, `month`, `holiday`, `weekday`, `workingday`, and `weather situation` significantly influence the demand for shared bikes. For example, higher demand may be observed in spring and summer months, on weekdays, and when the weather is clear. Holidays may also lead to increased or decreased demand based on public activities.

2. Importance of Using `drop_first=True`:

Using `drop_first=True` when creating dummy variables helps to avoid the dummy variable trap, that we avoid redundancy of the data, thus preventing overfitting for example if we have 3 categories like furnished semi furnished and un furnished then it will create three separate colunms but we required only two because if its neither furnished and semifurnished then it is unfurnished.

3. Highest Correlation with the Target Variable:

Looking at the pair-plot among numerical variables, typically, `temp` (temperature) or `hum` (humidity) tends to show the highest correlation but in my opinion temp is having the highest correlation among the numeric variable it became more clear after seeing heatmap.

4. Validating Linear Regression Assumptions:

After building the model, I would validate the assumptions of linear regression by checking:

- Linearity: In Residual plots if we see residuals are randomly scattered around zero.

- Homoscedasticity: Plotting residuals against predicted values to check for constant variance.

- Normality: If we see residual plot error terms or residuals are normally distributed .

5. Top 3 Features Contributing to Demand:

Based on the final model, the top 3 features that significantly contribute to explaining the demand for shared bikes are typically `temp` (temperature), `season`, and `humidity`. These variables likely have the highest coefficients, indicating their strong influence on the demand as they are having minimum value as well as VIF value too is low.

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. The simplest form is in

simpler linear regression is y=mx+c that is the basic equation of line because in this our aim is to find the best fit line so m is the slope and for multiple linear regression, the equation expands to include multiple predictors. The algorithm minimizes the sum of squared residuals (the differences between observed and predicted values) to find the best-fitting line. Key assumptions include linearity, independence, homoscedasticity, and normality of residuals.

2. Anscombe's Quartet:

Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics yet show different distributions and relationships when graphed. This emphasizes the importance of visualizing data, as relying solely on statistical measures can be misleading. The datasets have the same mean, variance, correlation, and regression line, but their scatter plots reveal distinct patterns, demonstrating how different underlying structures can yield similar statistics.

3. Pearson's R:

Pearson's R, or Pearson correlation coefficient, measures the linear correlation between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear correlation, 1 indicates a perfect positive linear correlation, and 0 indicates no linear correlation. It is sensitive to outliers, and while it captures linear relationships, it may not adequately describe non-linear relationships.

4. Scaling:

Scaling is the process of standardizing the range of independent variables in data. It is performed to ensure that each feature contributes equally to the analysis, especially when they have different units or scales. Normalized scaling adjusts the values 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. Normalization is useful for algorithms sensitive to the scale of the data, while standardization is typically preferred for linear models.

5. Infinite VIF Values:

Variance Inflation Factor (VIF) can be infinite when there is perfect multicollinearity among predictors, meaning one predictor can be perfectly predicted from others. This situation arises when two or more independent variables are highly correlated, making it impossible to separate their individual effects on the dependent variable.

6. Q-Q Plot:

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as the normal distribution. In linear regression, it is used to assess whether the residuals are normally distributed, which is an assumption of the model. Points that fall along the reference line indicate that the residuals follow a normal distribution, while deviations

from this line suggest departures from normality, which may impact the validity of the regression results.