Answers to Subjective Questions

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?

The categorical variables in the dataset include `season`, `weathersit`, `mnth`, and `weekday`. Upon converting these variables to categorical strings and analyzing their impact on the dependent variable `cnt`:

- Season: Different seasons significantly affect bike demand. For example, summer and fall might see higher bike demand due to favorable weather conditions.

- Weathersit: The weather situation also impacts bike demand. Clear weather conditions are likely to see higher demand, while heavy snow/rain conditions see lower demand.

- Month (mnth): The month of the year affects bike rentals, with certain months (like summer months) having higher demands.

- Weekday: The day of the week can influence bike demand, with weekends potentially showing different patterns compared to weekdays.

2. Why is it important to use `drop\_first=True` during dummy variable creation?

Using `drop\_first=True` during dummy variable creation helps to avoid multicollinearity by dropping the first category. This ensures that the dummy variables are independent of each other, which is crucial for maintaining the interpretability and stability of the regression coefficients.

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

From the correlation matrix and the pair-plot, the variable `atemp` (feeling temperature) has the highest correlation with the target variable `cnt`.

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

The assumptions of Linear Regression were validated through:

- Residual Analysis: Checking the residuals to ensure they are randomly distributed around zero, indicating no obvious patterns.

- Linearity: Ensuring a linear relationship between the independent and dependent variables by visualizing the scatter plots.

- Homoscedasticity: Checking that the residuals have constant variance across all levels of the independent variables.

- Normality: Ensuring the residuals are normally distributed using Q-Q plots.

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 contributing significantly towards explaining the demand for shared bikes are:

- `temp` (actual temperature)

- `yr` (year, indicating the trend over time)

- `atemp` (feeling temperature)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. The algorithm fits a linear equation to the observed data:

\[

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 goal is to minimize the sum of squared residuals (difference between observed and predicted values). This is typically done using the Ordinary Least Squares (OLS) method.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It demonstrates the importance of visualizing data before analyzing it:

- Each dataset has the same mean, variance, correlation, and linear regression line.

- Despite these similarities, their scatter plots are markedly different.

- The quartet highlights how relying solely on statistical measures can be misleading without graphical analysis.

3. What is Pearson’s R?

Pearson’s R, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1:

- \( r = 1 \) indicates a perfect positive linear relationship.

- \( r = -1 \) indicates a perfect negative linear relationship.

- \( r = 0 \) indicates no linear relationship.

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

- Scaling: The process of adjusting the range of features in a dataset so that they are on a comparable scale.

- Purpose: To ensure that features contribute equally to the analysis, especially important in algorithms that rely on distance metrics (e.g., k-nearest neighbors, SVM).

- Normalized Scaling: Rescales the data to a range of [0, 1].

- Standardized Scaling: Transforms the data to have a mean of 0 and a standard deviation of 1.

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

An infinite VIF occurs when there is perfect multicollinearity among the independent variables, meaning one variable is an exact linear combination of others. This makes the regression coefficients indeterminate.

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 to assess if a dataset follows a particular distribution. In linear regression, it's used to check the normality of residuals:

- Use: Plot the quantiles of the residuals against the quantiles of a normal distribution.

- Importance: Deviations from a straight line in a Q-Q plot indicate departures from normality, which can affect the validity of the regression model.