Linear Regression Subjective QnA

Assignment Based Subjective QnA

1. Effect of Categorical Variables

The categorical variables (season, weathersit, etc.) significantly affect bike demand as they represent variations in weather and time, impacting user behavior. For instance, the 'season' variable captures seasonal trends in bike usage, while 'weathersit' reflects weather conditions that directly influence user decisions to rent bikes. Converting these variables to categorical data helps the model accurately capture these variations.

2. Importance of drop_first=True in Dummy Variable Creation

Using drop_first=True avoids multicollinearity by removing one category from each set of dummy variables, thus preventing the dummy variable trap. This helps ensure that the model does not suffer from redundant information, which can skew results and affect model performance.

3. Highest Correlation with Target Variable

The variable 'registered' shows the highest correlation with the target variable 'cnt'. This indicates that the number of registered users is a strong predictor of the total bike rentals, as registered users contribute significantly to the overall demand.

4. Validating Assumptions of Linear Regression

To validate the assumptions of linear regression after building the model, we performed the following checks:
1. **Linearity**: Checked the scatter plots of residuals vs. fitted values to ensure no patterns exist, indicating linear relationships.
2. **Normality of Residuals**: Used Q-Q plots to check if residuals are normally distributed.
3. **Homoscedasticity**: Verified that residuals have constant variance by inspecting residual plots.
4. **Multicollinearity**: Calculated VIF (Variance Inflation Factor) to detect multicollinearity among independent variables.

5. Top 3 Features Contributing to Bike Demand

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:
1. **Year (yr)**: Indicates the increasing trend in bike demand over the years.
2. **Temperature (temp)**: Higher temperatures generally lead to increased bike rentals.
3. **Winter Season (season_winter)**: Demand for bikes is higher in winter compared to spring (base category).

General Subjective QnA

1. Explain the Linear Regression Algorithm

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The algorithm fits a linear equation (y = b0 + b1*x1 + b2*x2 + ... + bn*xn) to the observed data. The coefficients (b0, b1, ..., bn) are estimated by minimizing the sum of the squared differences between the observed and predicted values (least squares method). This algorithm is useful for predicting continuous outcomes and understanding the influence of multiple factors on a dependent variable.

2. Explain Anscombe's Quartet

Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. The purpose of the quartet is to illustrate the importance of graphing data before analyzing it. Despite having the same mean, variance, correlation, and linear regression line, the datasets exhibit different patterns, outliers, and relationships. This demonstrates that relying solely on statistical measures can be misleading without visual inspection of the data.

3. What is Pearson's R?

Pearson's R, or Pearson correlation coefficient, measures the strength and direction of the linear relationship between two variables. The value of Pearson's R ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. It is commonly used to assess the degree of correlation between variables in statistical analysis.

4. What is Scaling and its Importance?

Scaling is the process of adjusting the range of features in the dataset to a standard scale, typically using normalization or standardization techniques. It is performed to ensure that all features contribute equally to the model, preventing features with larger scales from

dominating. Normalized scaling transforms data to a range between 0 and 1, while standardized scaling centers data around the mean with a standard deviation of 1. Scaling is crucial for algorithms sensitive to feature magnitudes, such as linear regression.

5. Infinite VIF Values

The value of VIF (Variance Inflation Factor) can become infinite when there is perfect multicollinearity, meaning one predictor variable is an exact linear combination of others. This leads to a division by zero in the VIF calculation formula, indicating that the predictor is redundant and can be linearly predicted from other predictors with high precision. Removing such variables is essential to improve model stability and interpretability.

6. What is a Q-Q Plot?

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular distribution, often the normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution. If the points lie approximately along a straight line, the data is likely from the specified distribution. In linear regression, Q-Q plots help validate the normality assumption of residuals, which is critical for accurate hypothesis testing and reliable confidence intervals.