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?

   Answer:
   Tt's apparent that several categorical variables have been considered in the model to predict bike demand. The coefficients associated with these categorical variables help in understanding their impact on the dependent variable (bike demand).
   Here are the inferences about the effect of categorical variables on the dependent variable based on the coefficients:
   - **Working Day:** The coefficient of 0.0473 suggests that being a working day has a positive impact on bike demand. It implies that on days categorized as working days, there tends to be a slight increase in bike demand compared to non-working days.
   - **Months (Mon, Dec, Jan, July, Nov, Sep):** Several months have negative coefficients (-0.1163, -0.0974, -0.1163, -0.0164, -0.0883, 0.0427). This indicates that during these specific months, there might be a decrease in bike demand compared to the omitted reference month (probably due to seasonal variations or other factors associated with those months).
   - **Seasons (Spring, Winter):** Both spring and winter have negative coefficients (-0.2501, -0.0472). It implies that during these seasons, there tends to be a decrease in bike demand compared to the omitted reference season, indicating a potential impact of weather or other seasonal factors on demand.

   The negative coefficients for certain months and seasons suggest a potential seasonal effect on bike demand, where certain times of the year might see reduced demand compared to others.
   Additionally, the positive coefficient for working days indicates that these days generally witness slightly higher bike demand, possibly due to commuting or work-related travel.
   These inferences provide insights into how specific categorical variables (such as days, months, and seasons) influence the bike demand captured by the model.

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

   Answer
   Using **drop_first=True** during dummy variable creation in regression helps to prevent multicollinearity issues. It ensures that one category is omitted as a reference, avoiding perfect correlation among dummy variables, which could affect model stability and interpretation. This approach maintains independence among variables and enhances the reliability of the regression analysis.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   (1 mark)
   Answer
   Temp 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?
   Answer

Validated the assumptions of linear regression after building a model on the training set involves several checks to ensure the model's reliability. Checked the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable

5. Based on the final model, which are the top 3 features contributing significantly towards
   explaining the demand of the shared bikes?
   1. Spring
   2. Year (yr)
   3. Windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer
   Linear regression is a supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables by assuming a linear relationship. It predicts the dependent variable based on the coefficients of the independent variables and an intercept term. The goal is to minimize the difference between observed and predicted values. It follows certain assumptions like linearity, independence, constant variance of residuals, and normality of residuals. Variants include simple linear regression (one independent variable), multiple linear regression (multiple independent variables), and polynomial regression (nonlinear relationships). The model is trained using methods like Ordinary Least Squares (OLS) or Gradient Descent, and it's used for prediction and inference tasks in various fields.

2. Explain the Anscombe’s quartet in detail.

Answer

Anscombe's quartet consists of four datasets that share identical statistical properties (mean, variance, correlation) but reveal dramatically different patterns when graphed. It emphasizes the critical need for data visualization alongside summary statistics, cautioning against drawing conclusions based solely on numbers. This quartet highlights the limitations of relying solely on statistical measures without considering the actual graphical patterns within the data.

3. What is Pearson’s R?

Answer

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 positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 implies no linear relationship. It's sensitive to outliers and only measures linear associations between variables. It's widely used for correlation analysis in various fields to understand the connections between variables.

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

Answer

Scaling is about adjusting values in a dataset to make them comparable. It's done to ensure all variables have similar importance in analysis. Normalized scaling brings values between 0 and 1, keeping original relationships. Standardized scaling gives a mean of 0 and a standard deviation of 1, preserving data shape. Normalized focuses on a specific range, while standardized maintains data distribution. Each method suits different needs in analysis or machine learning.

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

Answer

Infinite VIF (Variance Inflation Factor) occurs due to perfect multicollinearity, where one variable is a precise linear combination of others. It indicates extremely high multicollinearity, making coefficient estimation impossible and models unreliable. To address this, removing redundant variables, combining them, or using regularization techniques like Ridge Regression can help mitigate multicollinearity issues. Resolving infinite VIF is vital for stable and accurate regression models.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer

A Q-Q (Quantile-Quantile) plot compares the distribution of a dataset to a theoretical distribution, like the normal distribution. In linear regression, it's crucial for checking if the

residuals (prediction errors) follow a normal distribution. A Q-Q plot helps assess the assumption of normally distributed residuals, ensuring the model's validity and reliability. If the points on the plot align closely with a straight line, it indicates good conformity to the assumed distribution, validating the regression model's assumptions.