Submitted By – K Sumanth

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?

   'yr', 'temp', 'atemp', 'casual', and 'registered' display significant correlations, and there are inter-variable correlations like (season, month), (weathersit, hum), and (temp, atemp). These strong correlations may harm model performance. Hence, it's crucial to manage multicollinearity by judiciously eliminating certain features.

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

   Employing drop_first=True is essential as it reduces the generation of unnecessary columns when creating dummy variables.

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

   'temp' and 'atemp' show significant correlations with the target variable 'cnt.' Although 'casual' and 'registered' are highly correlated, they are not directly considered as 'cnt' is a derived metric from these two variables.

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

   Check for multicollinearity by calculating Variance Inflation Factors (VIF) among independent variables. Examine outliers using scatter plots, leverage plots, or studentized residuals, as they can significantly influence regression outcomes. Validate the model on a separate test set to assess performance on new data, comparing predictions to actual values and evaluating overall metrics. Analyze residuals to identify patterns or trends, revealing potential violations of assumptions like linearity or homoscedasticity.

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

   ‘mnth_9’, ‘weathersit_3’ and ‘mnth_8’

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression, a supervised machine learning algorithm, models the connection between a dependent variable and independent variables by fitting a linear equation to observed data. It minimizes the sum of squared differences between predicted and actual outcomes, presenting a straight line equation for making predictions with new input values. Assumptions include linearity, homoscedasticity, and normality of residuals. It finds application in predicting house prices, stock values, and scenarios assuming a linear relationship between variables.

2. Explain the Anscombe’s quartet in detail.

   Anscombe's Quartet is akin to a set of four data enigmas. Despite nearly identical numerical statistics, plotting the data reveals distinct narratives. With 11 (x, y) pairs for each set, it emphasizes the significance of visualizing data, highlighting that numerical values alone may not unveil the complete story. Anscombe's Quartet underscores the importance of incorporating graphs to comprehend data, as even seemingly similar mathematical figures can convey vastly different insights. It teaches us that a comprehensive understanding requires both numerical analysis and graphical interpretation.

3. What is Pearson’s R?

   Pearson's correlation coefficient, known as "Pearson's R," measures the strength and direction of a linear relationship between two continuous variables. With values between -1 and 1, it signifies perfect positive (1) or negative (-1) correlation, while 0 implies no linear association. Widely utilized, it detects linear relationships but isn't sensitive to non-linear patterns.

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

   In data preprocessing, scaling transforms features to a consistent scale, ensuring equal contribution to models. Normalized scaling places data between 0 and 1, while standardized scaling (z-score normalization) centers data at 0 with a standard deviation of 1. Both enhance model performance and interpretability.

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

   An infinite Variance Inflation Factor (VIF) often arises due to perfect multicollinearity among predictor variables, where one variable can be precisely predicted from others. This condition leads to division by zero in the VIF calculation, resulting in an infinite value. Addressing this issue involves identifying and managing highly correlated variables through removal or transformation during the modeling process.

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

   A Quantile-Quantile plot, or Q-Q plot, visually juxtaposes sample quantiles with those of a theoretical distribution. In linear regression, it aids in evaluating the normality of residuals. Alignment of points along a straight line indicates that residuals adhere to a normal distribution, confirming a fundamental assumption of linear regression.