Submitted By – Varun Laroiya

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?

   Features such as 'yr', 'temp', 'atemp', 'casual', and 'registered' exhibit strong correlations. Additionally, there are inter-variable correlations between (season and month), (weathersit and hum), as well as (temp and atemp). These high correlations among variables can adversely affect model performance. Therefore, it is essential to address multicollinearity by selectively dropping some features.

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

   Using **drop_first=True** is crucial because it helps minimize the creation of redundant columns when generating 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' exhibit strong correlations with the target variable 'cnt.' While 'casual' and 'registered' are the most highly correlated variables, they are not being considered directly 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?

   Multicollinearity: Assess the presence of multicollinearity among independent variables. Calculate variance inflation factors (VIF) to identify variables that may be highly correlated

   Outlier Analysis: Identify and examine any outliers in the data. Outliers can significantly impact regression results. Scatter plots, leverage plots, or studentized residuals can be helpful in detecting outliers

   Validation on Test Set: Apply the model to a separate test set to evaluate its performance on new, unseen data. Compare the model's predictions to the actual values and assess its overall performance metrics

   Residual Analysis: Evaluate the residuals (the differences between predicted and actual values). Check for patterns or trends in the residuals, which may indicate violations of assumptions such as 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?

   ‘yr’, ‘temp’ & ‘windspeed’

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression is a supervised machine learning algorithm that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The algorithm aims to minimize the sum of squared differences between predicted and actual outcomes. The resulting equation represents a straight line, allowing predictions based on new input values. It assumes a linear relationship between variables and relies on assumptions like homoscedasticity and normality of residuals. Linear Regression is widely used for tasks such as predicting house prices, stock values, or any scenario where a linear relationship is assumed.
2. Explain the Anscombe's quartet in detail.

   Anscombe's Quartet is like a set of four secret codes for data. Even though the numbers (statistics) for each set look almost the same, when you draw pictures of the data, they actually tell different stories. Imagine you have 11 pairs of numbers (x, y) for each set. This shows how important it is to not just use numbers but also draw graphs to understand data. Even if the math numbers seem similar, the pictures can show very different things. Anscombe's Quartet teaches us that just looking at numbers might not give the full story – we need to use graphs too!

3. What is Pearson's R?

   Pearson's correlation coefficient, denoted as "Pearson's R," quantifies the strength and direction of a linear relationship between two continuous variables. Ranging from -1 to 1, it indicates perfect positive (1) or negative (-1) correlation, with 0 suggesting no linear correlation. Widely used, it is sensitive to linear associations but not non-linear patterns.

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

   Scaling in data preprocessing involves transforming features to a common scale. It's done to ensure variables contribute equally to models. Normalized scaling scales data between 0 and 1, while standardized scaling (z-score normalization) centers data around 0 with a standard deviation of 1. Both aid in improved model performance and interpretability

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

   The occurrence of an infinite Variance Inflation Factor (VIF) typically results from perfect multicollinearity among the predictor variables. Perfect multicollinearity happens when one variable in a regression model can be exactly predicted from others. In such cases, the mathematical calculation of VIF involves dividing by zero, leading to an infinite VIF. To address this issue, it is necessary to identify and handle highly correlated variables by removing or transforming them in the modeling process.

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

   A Q-Q plot (Quantile-Quantile plot) visually compares the quantiles of a sample with the quantiles of a theoretical distribution. In linear regression, it helps assess the normality of residuals. If the points align with a straight line, the residuals follow a normal distribution, validating a key assumption of linear regression.