Assignment-based Subjective

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

ANS: Analysis of categorical variables indicates higher bike rental rates during summer and fall, particularly in September and October. Rentals are also elevated on Saturdays, Wednesdays, Thursdays, and holidays, with 2019 demonstrating overall increased demand.

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

ANS : Setting `drop_first=True` eliminates redundant columns created during dummy variable encoding, preventing multicollinearity.

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

ANS : Temperature is the variable most strongly correlated with the target variable.

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

ANS: The model's assumptions were validated by assessing multicollinearity through VIF, residual normality, and the linearity of the relationship between the dependent variable and independent features.

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

ANS: Temperature, year, and holiday status are the three most significant predictors of shared bike demand.

General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

ANS : Linear regression is a supervised machine learning algorithm used to predict a numerical value (dependent variable) based on one or more input features (independent variables).

It models the relationship between these variables as a linear equation.

There are two primary types:

- Simple linear regression: Uses a single independent variable to predict the target variable.
- Multiple linear regression: Employs multiple independent variables for prediction.

The line representing this relationship is called the regression line. A positive linear relationship exists when the dependent variable increases as the independent variable increases. Conversely, a negative linear relationship occurs when the dependent variable decreases as the independent variable increases.

Question 2. **Explain the Anscombe's quartet in detail.**

ANS: Anscombe's quartet is a set of four datasets with nearly identical summary statistics but vastly different visual representations. Each dataset contains eleven data points. This example emphasizes the critical role of data visualization in exploratory data analysis. By revealing underlying patterns, trends, and outliers that numerical summaries alone cannot capture, graphical exploration prevents misleading interpretations and ensures a comprehensive understanding of the data.

Question 3 . **What is Pearson's R?**

ANS: Pearson's correlation coefficient measures the strength and direction of a linear relationship between two numerical variables.Its value ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 suggests no linear relationship.

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

ANS : Scaling is a crucial preprocessing step in machine learning that involves transforming numerical features into a standardized range. This is essential when features have vastly different scales or units, as it prevents features with larger magnitudes from dominating the model. By standardizing features, we ensure that all features contribute equally to the model's learning process.

Normalization and **standardization** are two common scaling techniques. Normalization rescales features to a specific range, typically between 0 and 1, while standardization transforms features to have a mean of 0 and a standard deviation of 1. The choice between the two depends on the specific algorithm and data distribution.

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

ANS : A perfect correlation between two independent variables results in an infinite Variance Inflation Factor (VIF) due to an R-squared value of 1 in the auxiliary regression. This indicates severe multicollinearity, where one variable provides no additional information beyond what is already captured by the other. To address this issue and build a reliable regression model, it's essential to remove one of the highly correlated variables.

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

ANS: A quantile-quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a specific distribution, such as normal, uniform, or exponential. It compares the quantiles of the dataset to the quantiles of the theoretical distribution. If the data points on the Q-Q plot closely follow a straight line, it suggests that the data is likely from the specified distribution. Q-Q plots are particularly useful for checking the normality assumption of residuals in regression analysis.