Assignment-based Subjective Questions

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

Answer :

a) Most Bikes were rented on Friday.

b) Most Bikes were rented on working days and not on holidays or non working days.

c) Most bikes were rented when the weather was Clear and not cloudy or rainy.

d) Most Bikes were rented in the month of August.

e) Most bikes were rented in the Fall Season.

f) The Saturday and Sunday being non working days the count of bikes rented on these days is almost negligible for these days compared to other days. But surprisingly there is an interesting observation, that whenever Wednesday is holiday or non working day the more number of bikes are being rented compared to when Wednesday is a working day.

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

Answer:
It helps to avoid multicollinearity by removing one category, which serves as a reference. It helps in preventing dummy variable trap.

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

Answer :
The columns temp and atemp are highly correlated.

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

Answer:

a. **Residual Analysis**: Checked for constant variance (homoscedasticity) by plotting residuals against predicted values. Also, plot residuals against each predictor to identify potential patterns.
b. **Linearity**: Used scatter plots of predicted values vs. actual values for each predictor to verify linear relationships.
c. **Multicollinearity**: Examined correlation matrix or variance inflation factors (VIF) to detect high correlation among predictors.

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

Answer:

Base on the coefficients of the independent variables the three most important variables are year, season_winter and atemp.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer:

Linear Regression is a supervised machine learning algorithm used for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the features and the target. The algorithm estimates coefficients that minimize the sum of squared differences between predicted and actual target values. It involves two main steps: 1) fitting a linear equation to the data using least squares, and 2) making predictions by substituting feature values into the equation. The model's accuracy and quality of fit are evaluated using metrics like Mean Squared Error or R-squared.

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

Answer:

Anscombe's quartet consists of four datasets that have nearly identical summary statistics (mean, variance, correlation, and linear regression coefficients), yet exhibit vastly different scatter plots and regression lines. It highlights the importance of visualizing data and not relying solely on summary statistics. The quartet emphasizes that diverse data distributions can yield similar numerical properties, underscoring the need for graphical analysis and challenging assumptions in statistical analysis. Anscombe's quartet serves as a cautionary example when interpreting results and illustrates the limitations of relying solely on quantitative measures.

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

Answer:

Pearson's correlation coefficient (Pearson's R) is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear correlation, 1 indicates a perfect positive linear correlation, and 0 indicates no linear correlation.

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

Answer:

Scaling is the process of transforming the features of a dataset to a common scale to ensure that they have similar magnitudes. Scaling is performed to prevent certain features from dominating others and to improve the performance and convergence of machine learning algorithms.

Normalized Scaling: Normalization scales features to a range between 0 and 1, using the formula (x - min) / (max - min). It maintains the distribution's shape but brings all features to a similar scale.

Standardized Scaling (Z-score normalization): Standardization scales features to have a mean of 0 and a standard deviation of 1, using the formula (x - mean) / standard deviation. It centers the data around 0 and preserves the shape of the distribution while giving equal weight to all features.

The key difference is that normalized scaling adjusts values to a specific range, while standardized scaling transforms values to have a standard deviation of 1, making it suitable for algorithms sensitive to feature scales.

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

Answer:

The occurrence of infinite values in the Variance Inflation Factor (VIF) is due to perfect multicollinearity in the data. Perfect multicollinearity arises when two or more predictor variables are perfectly linearly related, meaning one variable can be exactly predicted from a combination of others. This leads to a situation where the determinant of the correlation matrix is zero, causing the VIF calculation to involve division by zero and resulting in infinite VIF values.

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 is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, typically a normal distribution.

In linear regression, a Q-Q plot is important for checking the assumption of normality of residuals. If the residuals follow a normal distribution, the points on the Q-Q plot should approximately lie on a straight line. Deviations from a straight line can indicate departures from normality, which may affect the validity of statistical inference and hypothesis tests.