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: From the analysis of categorical variables, it can be inferred that certain categories may have a significant effect on the dependent variable. The effect can be seen through higher or lower average values of the dependent variable across different categories.

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

Answer: Setting drop_first=True during dummy variable creation is important to avoid multicollinearity in the regression model. When creating dummy variables from categorical variables, we convert each category into a binary (0 or 1) representation. If we include all the dummy variables, it can lead to perfect multicollinearity, where one dummy variable can be predicted perfectly from the others. By dropping the first dummy variable, we maintain linear independence among the variables and prevent multicollinearity issues.

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

Answer: By examining the pair-plot among the numerical variables, the variable that has the highest correlation with the target variable are “temp”, “atemp”, “casual” and “registered”.

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

Answer: After building the Linear Regression model on the training set, I validated its assumptions through various techniques:

- Residual Analysis: Check if the residuals are normally distributed around zero with constant variance, using Q-Q plots and residual vs. fitted plots.

- Linearity Check: Verify if the relationship between the dependent variable and each independent variable is approximately linear using scatter plots or partial regression plots.

- Multicollinearity Check: Ensure that independent variables are not highly correlated with each other, using VIF (Variance Inflation Factor) calculation.

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

Answer: Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes can be identified from the regression coefficients. These features are “yr”, “weathersit” and “windspeed”.

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to the observed data. The algorithm aims to find the best-fitting straight line that minimizes the sum of squared errors between the predicted and actual values.

The equation of a simple linear regression is represented as:

$y = mx + b$

Where:
- y is the dependent variable (target)
- x is the independent variable (feature)
- m is the slope of the line
- b is the y-intercept

The algorithm estimates the values of 'm' and 'b' based on the given data points and their corresponding target values using techniques like the least squares method. Once the coefficients are determined, the linear equation can be used to predict the target variable for new data points.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but significantly differ when visualized. These datasets were introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization in understanding and analyzing data.

Each dataset consists of eleven (x, y) points, and when graphed, all four datasets appear to follow linear relationships. However, upon closer inspection, they reveal distinct patterns such as outliers, non-linear relationships, and the effect of influential data points.

The quartet highlights that relying solely on summary statistics like mean, variance, and correlation can be misleading, as the underlying patterns might not be adequately captured. Visualizing the data can provide deeper insights and reveal important features that simple statistics may overlook.

3. What is Pearson's R?

Answer: Pearson's correlation coefficient (often denoted as 'r') is a statistical measure that quantifies the linear relationship between two continuous variables. It is used to assess the strength and direction of the association between the variables. The value of 'r' ranges between -1 and +1.

- If 'r' is close to +1, it indicates a strong positive correlation, implying that when one variable increases, the other tends to increase as well.
- If 'r' is close to -1, it signifies a strong negative correlation, suggesting that when one variable increases, the other tends to decrease.
- If 'r' is close to 0, there is little or no linear correlation between the variables.

Pearson's correlation is widely used in various fields, including data analysis, machine learning, and scientific research, to understand the relationships between different variables in a dataset.

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

Answer: Scaling is the process of transforming the values of different variables to a common scale, which is typically done to facilitate comparisons or improve the performance of certain machine learning algorithms. Scaling is essential when variables have different ranges, as it helps bring them to a similar magnitude.

Normalized Scaling:
- In normalized scaling (also known as Min-Max scaling), the values of the variable are transformed to a range between 0 and 1.
- The formula for normalized scaling is: `x_scaled = (x - min(x)) / (max(x) - min(x))`
- This scaling method preserves the relative relationships between the data points but may be sensitive to outliers.

Standardized Scaling:
- In standardized scaling (also known as z-score scaling), the values of the variable are transformed to have a mean of 0 and a standard deviation of 1.
- The formula for standardized scaling is: `x_scaled = (x - mean(x)) / standard_deviation(x)`
- Standardized scaling is useful when the data has outliers, as it is not affected by extreme values.

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

Answer: VIF (Variance Inflation Factor) is a measure used to detect multicollinearity in a multiple regression model. It quantifies how much the variance of an estimated regression coefficient is inflated due to multicollinearity.

VIF can become infinite in situations of perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables in a regression model can be perfectly predicted from a linear combination of other independent variables. In such cases, the VIF for the affected variable becomes infinite because its variance cannot be accurately estimated, leading to computational issues.

To handle multicollinearity, it's essential to identify and remove correlated predictors from the model or consider techniques like regularization to mitigate the impact of multicollinearity.

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 if a dataset follows a specific theoretical distribution (e.g., normal distribution). It compares the quantiles of the data against the quantiles of the chosen theoretical distribution.

To construct a Q-Q plot, the data is sorted in ascending order, and each data point's quantile is determined. Then, the quantiles are plotted against the corresponding quantiles from the theoretical distribution.

The use and importance of Q-Q plots in linear regression are as follows:
- Normality Check: In linear regression, it is often assumed that the residuals (the differences between observed and predicted values) follow a normal distribution. Q-Q plots help verify this assumption by visually inspecting if the residuals align with the straight line on the plot.
- Identifying Departures from Normality: Q-Q plots can reveal departures from normality. If the points deviate significantly from the straight line, it indicates non-normality in the residuals, which may affect the reliability of the regression model.

- Outlier Detection: Q-Q plots can also help identify outliers. Outliers can cause distortion in the linear regression model and should be carefully examined and potentially treated or removed.

In summary, Q-Q plots are a valuable diagnostic tool in linear regression, enabling researchers to assess the assumptions of the model and make appropriate adjustments if necessary.