Naga Jithendra Nangineni

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:

Season: Bike usage peaks in summer and fall, with lower counts in winter and spring, indicating a preference for warmer seasons.

Weekday: Counts were higher on Thursday, Friday, Saturday, and Sunday compared to the start of the week.

Holiday: Non-holidays see higher average counts compared to holidays.

Weather Conditions: Bikes were most frequently used in calm and clear weather, less so in misty conditions, even less in light snow, and rarely during heavy rains and thunderstorms.

Month: Usage gradually increased from the beginning of the year to around September/October and then decreased towards the year's end, aligning with seasonal trends. The busiest months were July through October.

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

Answer:

The **drop_first=True** option is crucial to prevent multicollinearity among variables when dealing with categorical data and it reduces the number of predictors, simplifying the model without losing information.

To represent a categorical variable with k levels, we only need k−1 dummy variables. For instance, if a categorical variable has four levels (e.g., seasons = summer, winter, autumn, rainy), we only need three dummy variables to represent it. Without dropping the first category, it results in multicollinearity among the variables.

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

Answer: Based on the pair-plot, Count has the highest correlation with "temp"

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

Answer: Below are the methods I have used to validate the assumptions:

- The residuals should exhibit a normal distribution. Upon plotting the residuals, I observed that they approximate a normal distribution.
- Linear relationship: I confirmed a linear relationship between the independent variables and the dependent variable.
- Homoscedasticity: I ensured that there is no discernible pattern among the residuals.
- Autocorrelation: There was no evidence of autocorrelation among the residuals.
- Multicollinearity: The variables in the final model exhibited low variance inflation factor (VIF) values, meeting acceptable standards for multicollinearity.

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

Answer: Below are the top 3 features significantly contributing

a. temp (Temperature)
b. lightsnow (Weather Situation)
c. yr (Year)

General Subjective Questions

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

Answer:

Linear regression is a predictive modeling technique employed to forecast values of a target variable by examining its known values relative to multiple independent variables. This approach entails constructing a model using a training dataset to establish a linear relationship between the independent variables and the dependent variable. Subsequently, the model is utilized to predict values of the target variable based on new or alternative sets of independent variables. Linear regression is categorized into Simple Linear Regression and Multiple Linear Regression based on the number of independent variables involved.

The primary objective is to establish a mathematical relationship represented by the formula:

Y = mx + c

Where Y = target variable
x = independent variable and
c = constant

The generated model should satisfy some assumptions to be an efficient model:

- ✓ Linearity
- ✓ Absence of multicollinearity
- ✓ Absence of Autocorrelation
- ✓ Normality of residuals
- ✓ Homoscedasticity

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet consists of four datasets that appear nearly identical in terms of basic statistical properties like mean, variance, and correlation. However, each dataset exhibits vastly different distributions and relationships between variables when plotted visually. This demonstrates the limitations of relying solely on numerical summaries and emphasizes the critical importance of graphical exploration in understanding data patterns and relationships. Anscombe's quartet is widely used in statistics education to highlight these principles and to caution against making assumptions without thorough visual examination of data.

3. What is Pearson’s R?

Answer:

Pearson's correlation coefficient, also known as the Coefficient of Correlation, quantifies the strength of the linear relationship between variables. When variables move in the same direction, the correlation coefficient is positive. Conversely, when variables move in opposite directions, with higher values of one associated with lower values of the other, the correlation coefficient is negative. Pearson's correlation coefficient

> **r=1**: Perfect positive correlation.
>
> **r=0**: No correlation between the variables.
>
> **r=-1**: Perfect negative correlation.

A positive r indicates a positive association: as one variable increases, so does the other.

A negative r indicates a negative association: as one variable increases, the other decreases.

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

Answer:

Feature scaling involves adjusting the range of independent variables or features in a dataset to ensure they are treated equally by the model during training. If features are not scaled, their coefficients can vary significantly, making it difficult to determine their true importance in the model. Additionally, variables with different units can skew importance; for instance, a variable in pounds might appear more significant than one in kilograms due to its larger numeric values.

Scaling is typically done in two ways:

a. Normalized Scaling:

   Formula: **x_scaled = (x - min(x)) / (max(x) - min(x))**

b. Standardized Scaling:

   Formula: **x_scaled = (x - mean(x)) / std(x)**

These scaling techniques are essential for preparing data to ensure that models can effectively interpret the relative importance of different features and their contributions to the predictions.

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

Answer:

The Variance Inflation Factor (VIF) is determined by the formula:

VIF = 1 / (1 – R2)

When R2 equals 1, the VIF value becomes infinite. This scenario arises when variables exhibit perfect correlation. A higher VIF indicates stronger correlation between variables, which causes multicollinearity and reduces the model's effectiveness.

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

Answer:

A Quantile-Quantile plot, often abbreviated as Q-Q plot, is a scatter plot used to assess whether two datasets originate from populations with similar distributions. In this plot, quantiles from one dataset are compared against quantiles from another dataset, typically representing a theoretical distribution or a reference dataset. The test dataset is plotted against the actual distribution it is being compared to.

The Q-Q plot displays quantiles of the test dataset against quantiles of the reference distribution. If both datasets share a common distribution, the points should align closely along a diagonal reference line. Deviations from this line indicate differences in the distributions of the two datasets.

The Q-Q plot is crucial for determining if the assumption of a common distribution is valid when comparing two datasets. It provides visual insight into whether location and scale estimators can be pooled across both datasets to derive common estimates. Conversely, discrepancies in the plot reveal differences between the datasets' distributions, aiding in understanding these variations.