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?

-From my analysis of the categorical variables from the dataset we could infer that most effect is from year then from the month of june there category affects the most to dependent variable.

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

-drop_first=True removes the first dummy variable from the dataset so we have one column less because for this particular column all the other dummy variables will have 0 values.

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

-registered column has the highest corelation with the target varaible i.e., cnt.Correlation value=0.95.

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

-By plotting the histigram of error tems if its normally distributed then its validated.

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

-Based on final model for me the top 3 features contributing significantly towards explaining the demand of the shared bikes are

-year

-casual

-june

General Subjective Questions

1. Explain the linear regression algorithm in detail.

-Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. In the simplest form, linear regression assumes a linear relationship between the dependent variable Y and the independent variable X

Linear regression assumes several key assumptions:

-Linearity: The relationship between the independent and dependent variables is linear.

-Independence: The residuals (errors) are independent of each other.

-Homoscedasticity: The variance of the residuals is constant across all levels of the independent variable(s).

-Normality of Residuals: The residuals follow a normal distribution.

-No Multicollinearity: The independent variables are not highly correlated.

Linear regression is widely used in various fields for prediction and understanding the relationships between variables. Extensions such as multiple linear regression allow for modeling with more than one independent variable. Regularization techniques, such as Ridge and Lasso regression, can be applied to prevent overfitting in the presence of multicollinearity.

2. Explain the Anscombe's quartet in detail.

-Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear quite distinct when graphed. This quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphing data before analyzing it and to illustrate the impact of outliers on statistical properties.

Despite having the same means, variances, and correlation coefficients, these datasets have notably different patterns when graphed. This highlights the importance of visual exploration and the limitations of relying solely on summary statistics.

Anscombe's quartet is often used to illustrate the concept that summary statistics alone may not be sufficient to understand the nature of data. Graphical representations, such as scatter plots and regression lines, can provide valuable insights into the underlying relationships within a dataset.

3. What is Pearson's R?

-Pearson's correlation coefficient, often denoted as r or Pearson's

r, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of a linear association between the variables. The value of r ranges from -1 to 1, where:

r=1:Perfect positive linear correlation

r=-1:Perfect negative linear correlation r=0:No linear correlation

Key points about Pearson's correlation coefficient:

Direction: The sign of rindicates the direction of the linear relationship. A positive

r implies a positive correlation (both variables increase or decrease together), while a negative r implies a negative correlation (one variable increases as the other decreases).

Strength: The absolute value of r indicates the strength of the linear relationship. The closer |r| is to 1, the stronger the linear correlation.

Assumption: Pearson's r assumes that the relationship between the variables is linear. It may not accurately reflect non-linear relationships.

Outliers: Pearson's r is sensitive to outliers. Outliers can strongly influence the correlation coefficient.

Unitless: r is a unitless measure, meaning it is not affected by the scale of measurement of the variables.

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

and standardized scaling?

-Scaling is the process of transforming variables to a standardized range or distribution. It involves adjusting the values of different variables so that they are on a comparable scale. The primary goal of scaling is to bring all variables to a common scale, preventing some variables from dominating others in analyses. Scaling is particularly important in machine learning and statistical modeling techniques that rely on distance measures or gradients, as it helps algorithms converge faster and makes the interpretation of coefficients or feature importance more meaningful.

Why Scaling is Performed:

Algorithm Sensitivity: Many machine learning algorithms are sensitive to the scale of input features. Algorithms like k-nearest neighbors, support vector machines, and neural networks often perform better when features are on a similar scale.

Gradient Descent: Gradient-based optimization algorithms, such as those used in linear regression or neural networks, converge faster when variables are scaled. This is because the step sizes during optimization are influenced by the scale of the variables.

Distance Measures: Algorithms that rely on distance measures (e.g., k-means clustering) can be influenced by the scale of variables. Scaling ensures that all variables contribute equally to distance calculations.

Differences:

Normalized scaling transforms data to a specific range (usually 0 to 1), while standardized scaling transforms data to have a mean of 0 and a standard deviation of 1.

Normalized scaling is less affected by outliers, making it suitable for datasets with extreme values. Standardized scaling can be influenced by outliers, but it often handles them better than raw data.

The choice between normalized and standardized scaling depends on the nature of the data and the requirements of the specific algorithm being used.

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

-The occurrence of infinite values in the Variance Inflation Factor (VIF) is typically a result of perfect multicollinearity among the independent variables in a regression model. Perfect multicollinearity means that one or more independent variables can be exactly predicted by a linear combination of the others.

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) is a graphical tool used to assess whether a given sample or dataset follows a specific theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of the expected distribution. The Q-Q plot is particularly useful for identifying departures from normality.