Assignment Questions:

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer1: An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer2: It helps in reducing the extra column created during dummy variable creation.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: yr

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

Assumption1: There should not be multicollinearity between independent variables, VIF is checked to validate this assumption, Variables with high VIF were removed one by one.

Assumption2: Error should be normally distributed.

Histogram was plotted to check the distribution of errors and error were distributed normally

Assumption3: There should not be autocorrelation between error values.

Dubin Watson coefficient value is around 2, which validates that there is no autocorrelation between error values.

Q5: 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 most significant variables

yr: A coefficient value of '0.2462' indicated that a unit increase in yr variable, increases the bike hire numbers by 0.2462 units.

windspeed :A coefficient value of '-0.1648 ' indicated that a unit increase in windspeed variable, decreases the bike hire numbers by 0.1648 units.

mnth_2 :A coefficient value of ‘0.0528 ’ indicated that a unit increase in mnth_2 variable, increase the bike hire numbers by 0.0528 units.

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 (also called the response or target variable) and one or more independent variables (also called the predictors or features) that may influence it. The goal of linear regression is to find a linear equation that can best explain the relationship between the variables.

The general form of a linear regression equation for a single independent variable is:

$y = \beta 0 + \beta 1x1 + \varepsilon$

Where y is the dependent variable, x1 is the independent variable, β0 and β1 are the coefficients of the intercept and slope, respectively, and ε is the error term representing the random variability in the data that is not explained by the model.

Question2: What is Pearson’s R

Answer:

Pearson's correlation coefficient (often denoted by r) is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is named after Karl Pearson, a British mathematician who first introduced the concept in 1895.

The value of Pearson's correlation coefficient ranges from -1 to +1, where -1 indicates a perfect negative correlation (when one variable increases, the other variable decreases), 0 indicates no correlation, and +1 indicates a perfect positive correlation (when one variable increases, the other variable also increases). The sign of r indicates the direction of the relationship, while the magnitude of r indicates the strength of the relationship.

Question 3: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

scaling refers to the process of transforming variables in a dataset to a common scale, typically to make the variables comparable and easier to interpret. It is an important pre-processing step in many data analysis and machine learning tasks.

Scaling is performed for several reasons, including:

To ensure that variables have a similar range of values. This is important because many algorithms are sensitive to the scale of the variables, and variables with larger values may dominate the analysis.

To make variables dimensionless. This is important in situations where the units of measurement for different variables are not comparable.

To improve the performance of certain algorithms. For example, gradient descent optimization algorithms in machine learning tend to converge faster when variables are scaled.

There are two common types of scaling: normalized scaling and standardized scaling.

Question4: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: Variable is perfectly correlated to the other features that's why the VIF is infinite.

Question5: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer : A Q-Q plot (quantile-quantile plot) is a graphical method for comparing the distribution of a sample to a theoretical distribution, typically a normal distribution. The Q-Q plot plots the quantiles of the sample against the quantiles of the theoretical distribution, and if the sample is normally distributed, the points should fall along a straight line.

If the Q-Q plot suggests that the residuals are not normally distributed,