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: All Categorical data has impact on target variable and multicollinearity is possible in categorical variables as well. For model perspective these categorical values need to be encoded to proper numeric values if they are text categorical values.

2. Why is it important to use **drop_first=True** during dummy variable creation?
   Answer: it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Answer: 2019 has more correlation.

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

   Normality of Error
   Error values (ε) are normally distributed for any given value of X
   Homoscedasticity
   The probability distribution of the errors has constant variance
   Independence of Errors
   Error values are statistically independent

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Answer: atemp (feeling temperature), year (2019), season (fall) are significantly more correlated with target variable.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Answer:
   Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of data based on some variables.

Linear regression is used to predict a quantitative response Y from the predictor variable X

y=a+bx

Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots

3. What is Pearson's R?

Answer:

The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. It attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.

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

Answer:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude

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

Answer:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

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

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.