

## **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?**

=> R2 values for predictions test data i.e., 0.815 is almost same as R2 values of train data i.e., 0.818.

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

=> In the absence of drop\_first = True, n dummy variables will be generated, and since these predictors (n dummy variables) are correlated with one another, this phenomenon is referred to as multicollinearity and it results in the Dummy Variable Trap.

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

=> In pair plot, temp and atemp has highest correlation with the target variable.

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

=> We can validate the assumptions by plotting a scatter plot between features (season, mnth, weekday, weathersit) and target (target=casual + registered).

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

=> These are top 3 factors which affect demands of bikes: -

A) holiday

B) temp

C) hum

## **General Subjective Questions: -**

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

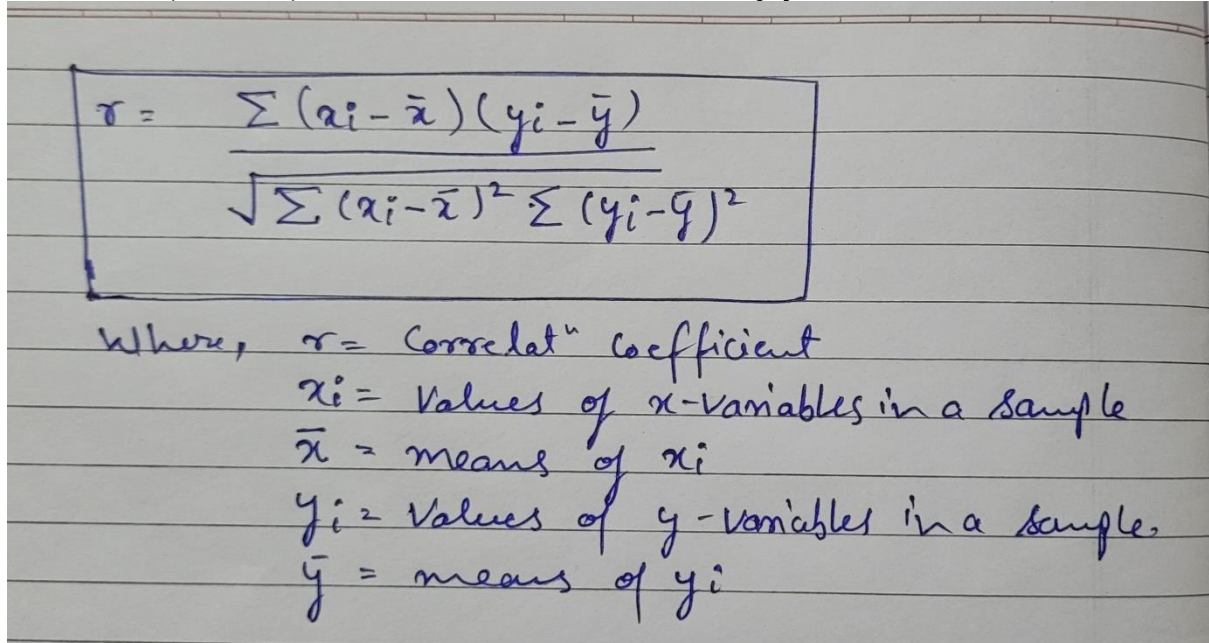
=> The Train Using AutoML tool employs the supervised machine learning technique of linear regression to identify the linear equation that most accurately captures the relationship between the explanatory variables and the dependent variable. This is accomplished by utilising least squares to fit a line to the data.

**2.Explain the Anscombe's quartet in detail.**

=> Anscombe's Quartet is a set of four data sets that, while they appear to be almost equal in simple descriptive statistics, have certain anomalies that, if a regression model is developed, would deceive it. When displayed on scatter plots, they have significantly different distributions and show up differently.

### 3. What is Pearson's R?

=> A measure of linear correlation between two sets of data is the Pearson correlation coefficient, often known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or simply the correlation coefficient.



The image shows a handwritten formula for Pearson's correlation coefficient  $r$  enclosed in a rectangular box. The formula is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Below the box, the components are defined:

- Where,  $r$  = Correlation coefficient
- $x_i$  = values of  $x$ -variables in a sample
- $\bar{x}$  = means of  $x_i$
- $y_i$  = values of  $y$ -variables in a sample
- $\bar{y}$  = means of  $y_i$

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

=> Scaling is a data pre-processing procedure used to normalise data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations.

=> If scaling is not done, the algorithm will only consider magnitude and not units, which will result in inaccurate models. We must scale all the variables to the same degree of magnitude in order to resolve this problem.

=> In Normalization It brings all of the data in the range of 0 and 1. While in Standardization. It replaces values according to their Z scores. It transforms all of the data into a normal distribution with a mean () of zero and a standard deviation () of one.

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

=> VIF = 1.0 if all independent variables are orthogonal to one another. VIF = infinity if there is perfect correlation. A high VIF score denotes a strong connection between the variables.

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

=> Q-Q plots are also known as Quantile-Quantile plots. They plot the quantiles of a sample distribution versus quantiles of a theoretical distribution, as the name implies. By doing this, we may establish whether a dataset adheres to a specific kind of probability distribution, such as a normal, uniform, or exponential one.