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:
   Here are the below points as inference from the categorical variables from the dataset.

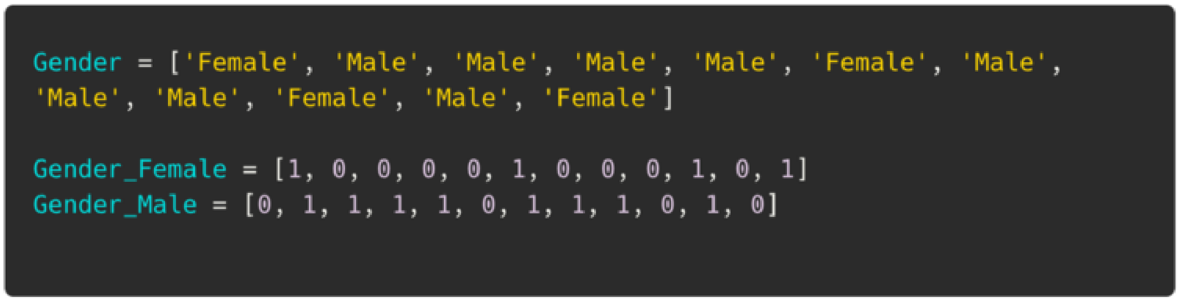
   - **Seasons**: Bike rentals are generally higher in fall and summer, with spring having the lowest rentals.
   - **Month:** Bike rentals peak during the summer months (June to September) and are lowest in winter months (January, February, December).
   - **Weekdays:** Bike rentals are fairly consistent throughout the week with no significant peaks or drops on specific weekdays.
   - **Working_Days:** Bike rentals are slightly higher on non-working days compared to working days.
   - **Weather Situation:** Bike rentals are highest during good weather conditions and drop significantly during bad weather conditions.
   - **Year:** 2019 attracted more numbers of bookings

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

   Answer:
   Categorical variables cannot be used directly in machine learning models. They have to be converted into meaningful numerical representations; this process is called encoding. There are a lot of techniques for encoding categorical variables, but we'll look at the one provided by the Pandas library called get_dummies().

   Issue during encoding

Consider Gender as a feature(Column). Let’s isolate the Gender column from the data set and encode it.



```
Gender = ['Female', 'Male', 'Male', 'Male', 'Male', 'Female', 'Male',
'Male', 'Male', 'Female', 'Male', 'Female']

Gender_Female = [1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1]
Gender_Male = [0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0]
```

If we look closely, Gender_Female and Gender_Male columns are multicollinear. This is because a value of 1 in one column automatically implies 0 in the other. We call this issue a dummy variable trap, which we represent as:
Gender_Female = 1 - Gender_Male

Importance of drop_first(Solution to Multicollinearity)
Multicollinearity is undesirable, and every time we encode variables with pandas.get_dummies(), we’ll encounter this issue. One way to overcome this problem is by dropping one of the generated columns. So, we can drop either Gender_Female or Gender_Malewithout potentially losing any information. Fortunately, pandas.get_dummies() has a parameter called drop_first which, when set to True, does precisely that.
Note: drop_first always gives n-1 dummies out of n categorical level by removing the first one

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

   Answer:

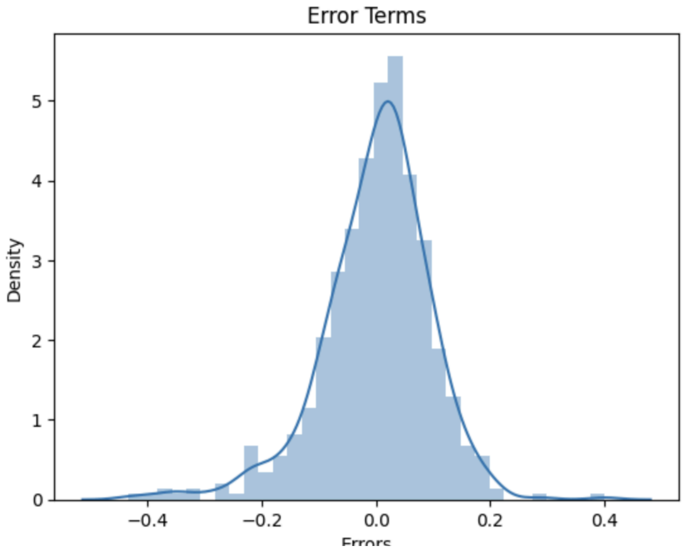
   Variable **temp** has highest correlation with the target variable

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

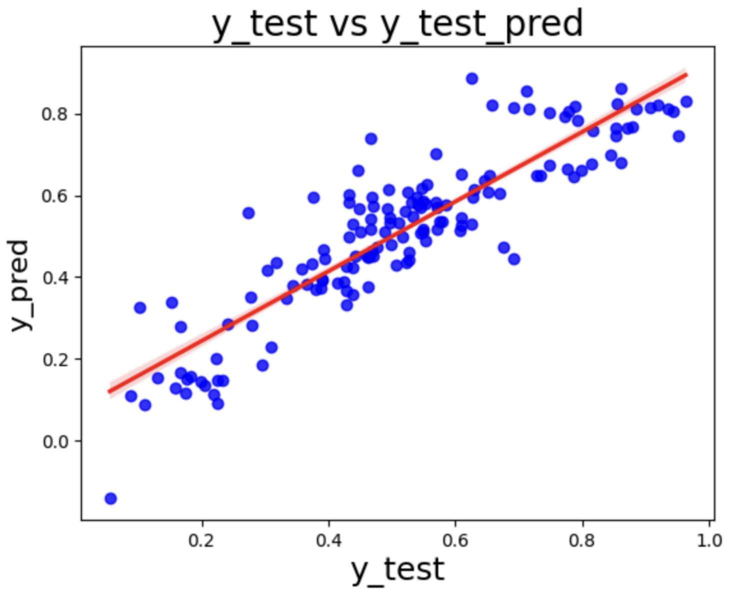
   Answer:

Validated the assumptions of Linear Regression after building a model on the training set.

- Error term normalization:
    - Error terms normally distributed.



- Multicollinearity:
    - Insignificant multicollinearity among the variable
- Linear relationship validation:
    - Linearity is visible among the variables
- Homoscedasticity:
    - There is no pattern in the residual value
- Residual Analysis: We performed R^2 test
    - **Train dataset R^2** : **0.8195564314872248**
    - **Test dataset R^2** : **0.7919791167382211**



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

Answer:

Model:

```
cnt(y)=0.1850+yrX(0.2306)+tempX(0.4885)+season_SpringX(-0.1188)+seas
on_WinterX(0.0582)+mnth_JulX(-0.0715)+mnth_SepX(0.0542)+weathersit_L
ight_Snow_RainX(-0.3016)+weathersit_Mist_CloudyX(-0.0747)
```

By comparing the absolute values of these coefficients, we find:

`temp` (0.4885)(+ve)

`weathersit_Light_Snow_Rain` (0.3016)(-ve)

`yr` (0.2306)(+ve)

These are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

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 and one or more independent variables.
SLR(Simple Linear Regression): When there is only one independent variable, it is called simple linear regression.

The Model

The simple linear regression model can be expressed as:

$y=\beta0+\beta1x+\epsilon$

where:

y is the dependent variable.

x is the independent variable.

$\beta0$ is the y-intercept of the regression line.

$\beta1$ is the slope of the regression line.

ϵ is the error term, representing the difference between the observed and predicted values.

MLR(Multiple Linear Regression): With more than one independent variable, it is called multiple linear regression. The goal is to find the best-fitting linear relationship, which can be used for prediction.
The Model

Multiple linear regression involves more than one independent variable. The model can be expressed as:

$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta pxp+\epsilon$

where:

y is the dependent variable.

x1,x2,…,xp are the independent variables.

β0,β1,…,βp are the coefficients.

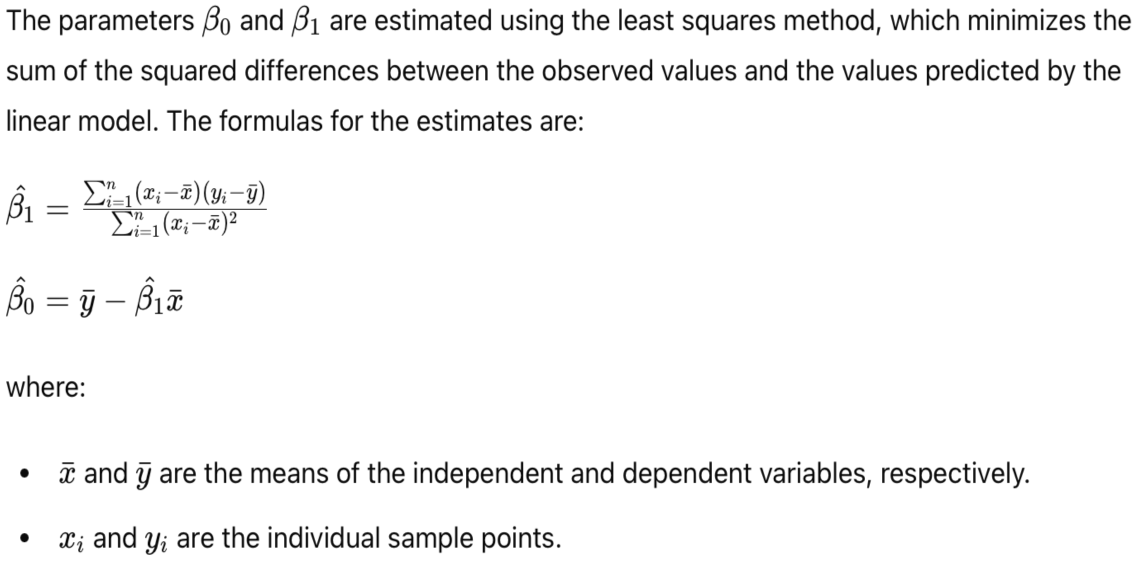
ϵ is the error term.

Note: Linear relationships can be **positive** or **negative.**

This algorithm follows the following steps to get the right model for the provided datasets.

- **Assumptions**
  - **Linearity**: The relationship between the independent and dependent variables is linear.
  - **Independence**: Observations are independent of each other.
  - **Homoscedasticity**: The variance of the residuals (errors) is constant across all levels of the independent variable.
  - **Normality**: The residuals (errors) of the model are normally distributed.

- **Estimation of parameters**
  - ***For SLR***



The parameters β_0 and β_1 are estimated using the least squares method, which minimizes the sum of the squared differences between the observed values and the values predicted by the linear model. The formulas for the estimates are:

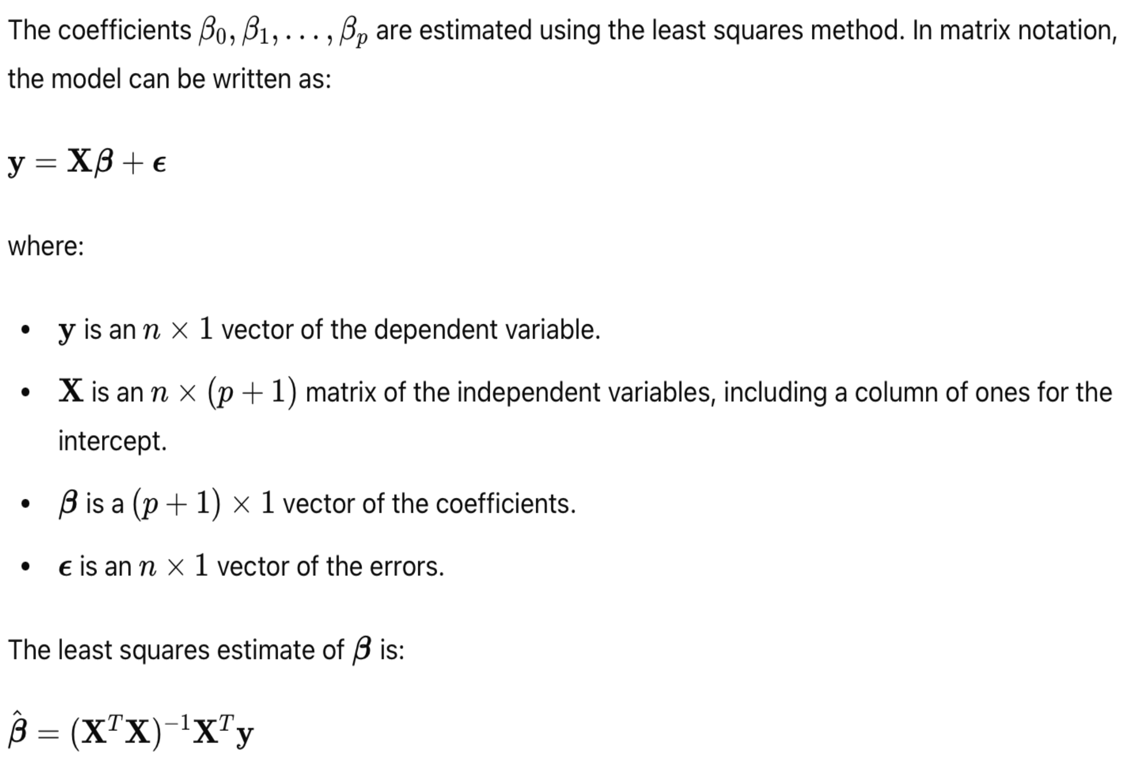
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

where:

- $\bar{x}$ and $\bar{y}$ are the means of the independent and dependent variables, respectively.
- x_i and y_i are the individual sample points.

  - ***For MLR***



The coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are estimated using the least squares method. In matrix notation, the model can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

- $\mathbf{y}$ is an $n \times 1$ vector of the dependent variable.
- $\mathbf{X}$ is an $n \times (p+1)$ matrix of the independent variables, including a column of ones for the intercept.
- $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of the coefficients.
- $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of the errors.

The least squares estimate of $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- **Prediction**
  The predicted value (y^) for a given set of independent variables is:
  y^=β0^+β1^x1+β2^x2+⋯+βp^xp
  Note: For SLR this will be ‘β0^+β1^x’

- **Model Evaluation**
  In this step we evaluate the model with below methods.
  - R^2 Validation
    - R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
  - R^2 Adjusted Validation
    - Adjusted R-squared adjusts the R2R2 value for the number of predictors in the model. It is useful when comparing models with different numbers of predictors.
  - Residual Analysis
    - Residuals (errors) are analyzed to check the assumptions of the regression model (linearity, independence, homoscedasticity, and normality). Plotting residuals and conducting statistical tests help in this evaluation

Linear regression is a powerful tool for understanding and predicting relationships between variables. By making and checking assumptions, estimating parameters, and evaluating the model, we can make informed predictions and gain insights from data.

2. Explain the Anscombe’s quartet in detail.

Answer:

Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Four datasets consider Anscombe's quartet.

If the mean for x and y for all four datasets.

the standard deviations for x and y for all four datasets.

the correlations with their corresponding pair of each datasets.

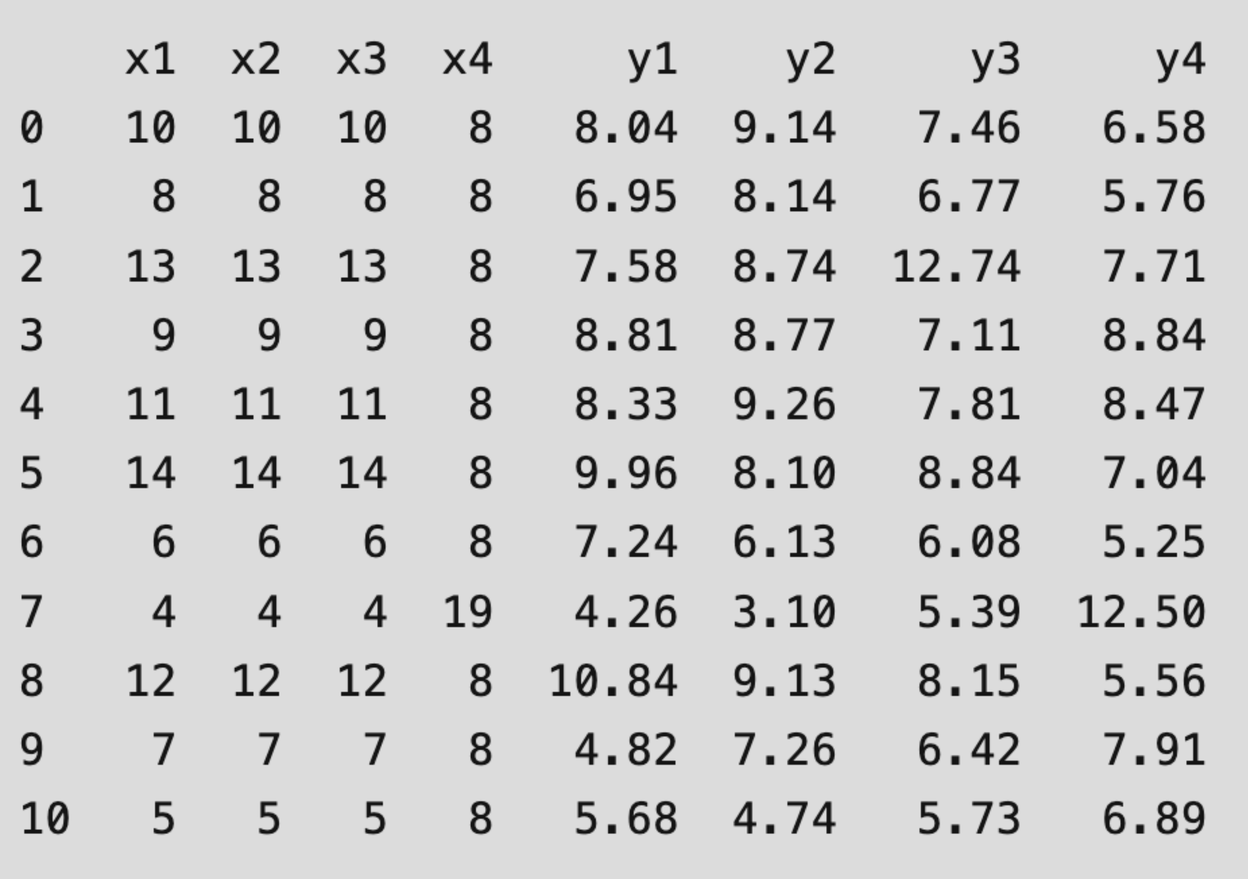
the slope and intercept for each datasets.

the R-square for each datasets.

Looks similar, however in graphical representations it looks different.

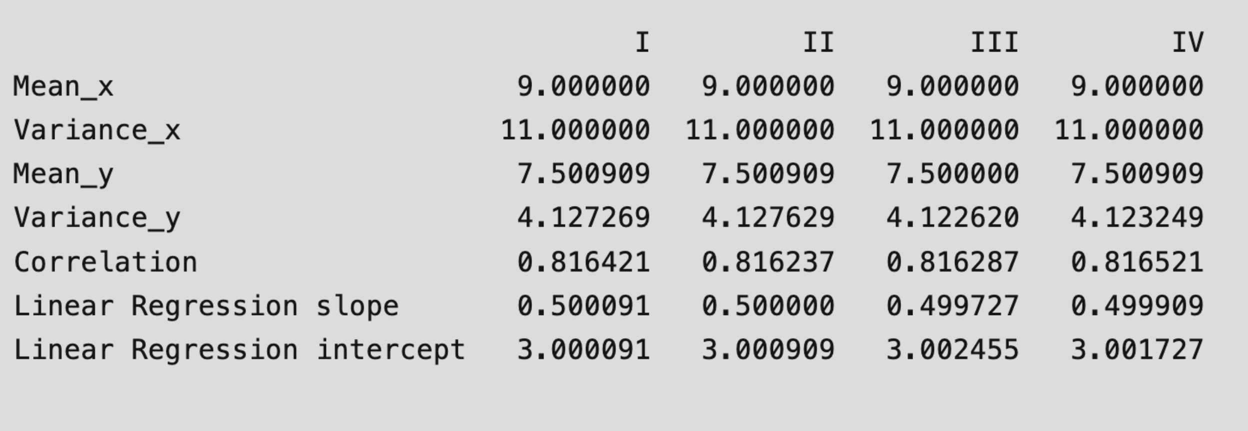
Example:

Datasets:



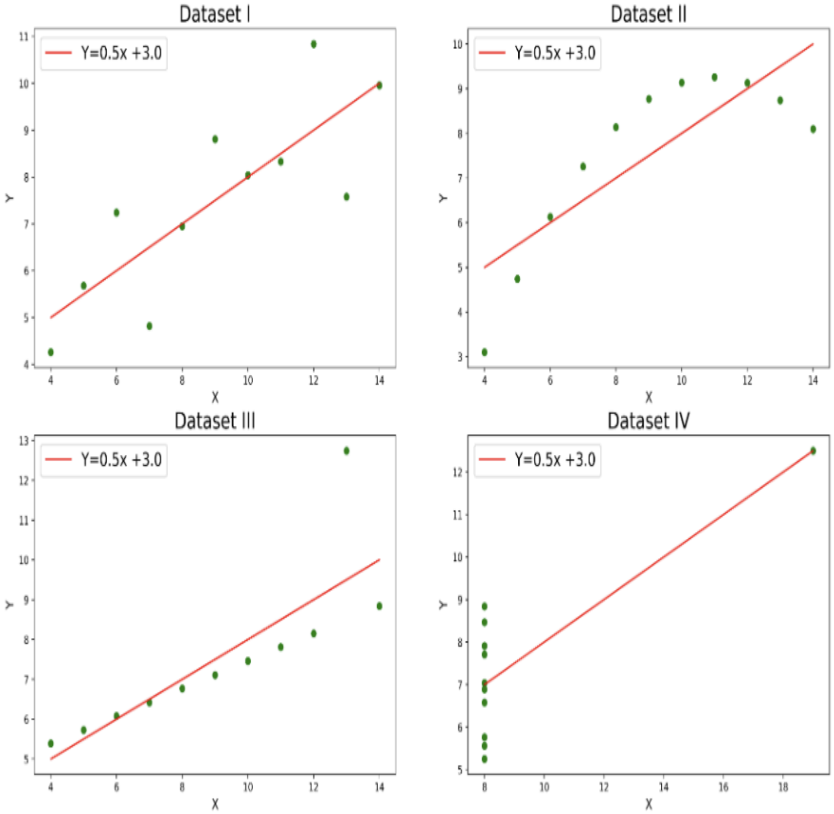
```
    x1  x2  x3  x4     y1    y2     y3     y4
0   10  10  10   8   8.04  9.14   7.46   6.58
1    8   8   8   8   6.95  8.14   6.77   5.76
2   13  13  13   8   7.58  8.74  12.74   7.71
3    9   9   9   8   8.81  8.77   7.11   8.84
4   11  11  11   8   8.33  9.26   7.81   8.47
5   14  14  14   8   9.96  8.10   8.84   7.04
6    6   6   6   8   7.24  6.13   6.08   5.25
7    4   4   4  19   4.26  3.10   5.39  12.50
8   12  12  12   8  10.84  9.13   8.15   5.56
9    7   7   7   8   4.82  7.26   6.42   7.91
10   5   5   5   8   5.68  4.74   5.73   6.89
```

Summary:



```
                                     I          II         III          IV
Mean_x                        9.000000    9.000000    9.000000    9.000000
Variance_x                   11.000000   11.000000   11.000000   11.000000
Mean_y                        7.500909    7.500909    7.500000    7.500909
Variance_y                    4.127269    4.127629    4.122620    4.123249
Correlation                   0.816421    0.816237    0.816287    0.816521
Linear Regression slope       0.500091    0.500000    0.499727    0.499909
Linear Regression intercept   3.000091    3.000909    3.002455    3.001727
```

Graphical Representation:



Anscombe's quartet Plot

3. What is Pearson's R?

Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. In the context of regression analysis, it quantifies the strength and direction of the linear relationship between the independent variable (predictor) and the dependent variable (response). Here's a breakdown of its characteristics:

Range: Pearson's R ranges from -1 to 1.

- R=1 indicates a perfect positive linear relationship.
- R=-1 indicates a perfect negative linear relationship.
- R=0 indicates no linear relationship.

In regression analysis, Pearson's R helps in understanding how well the independent variable predicts the dependent variable. However, it's important to note that Pearson’s R only measures linear relationships and does not capture non-linear relationships between variables.

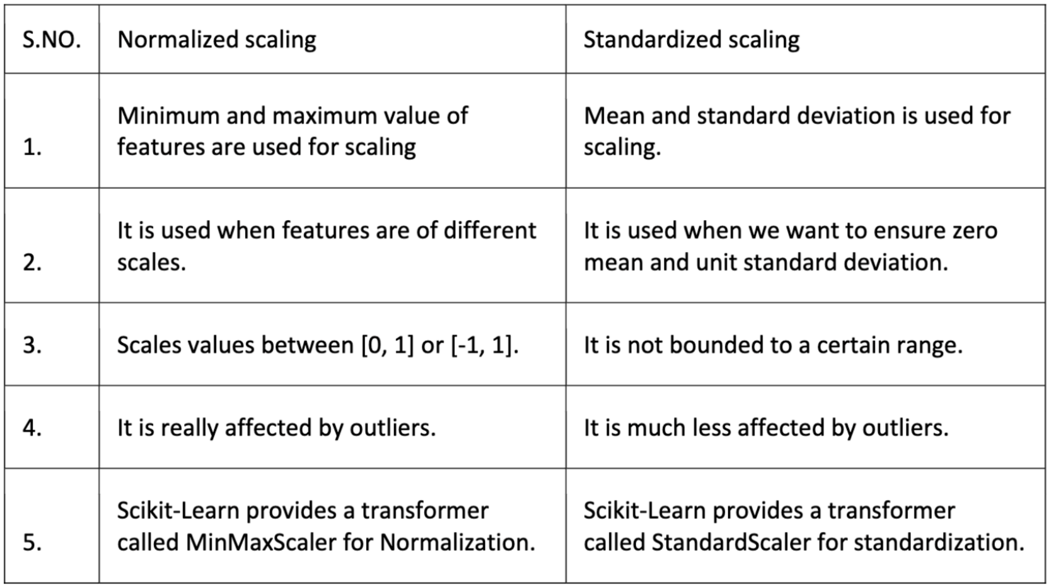
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:
Scaling is a data preprocessing technique used to adjust the range and distribution of features in a dataset. It is often performed before applying machine learning algorithms to ensure that the features are on a similar scale, which can improve the performance and convergence speed of the models.

Why Scaling is Performed

1. **Algorithm Performance**: Many machine learning algorithms, such as gradient descent-based methods, are sensitive to the scale of the features. Features with larger ranges can dominate the learning process, leading to suboptimal models.
2. **Convergence Speed**: Scaling can help in faster convergence of optimization algorithms used in training machine learning models.
3. **Model Interpretation**: Scaled features can make it easier to interpret the coefficients of linear models.
4. **Distance Metrics**: Algorithms that rely on distance metrics, such as k-nearest neighbors (KNN) or support vector machines (SVM), perform better when the features are scaled.

Min-Max VS Standardized



S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

To Summarize:

Normalized Scaling (Min-Max Scaling): Transforms data to a specific range (e.g., [0, 1]). Sensitive to outliers.
Standardized Scaling (Z-score Normalization): Transforms data to have a mean of 0 and a standard deviation of 1. Less sensitive to outliers and suitable for normally distributed data.

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

Answer:
In Multiple Linear Regression (MLR), the Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases due to collinearity among the predictors. When VIF is infinite, it indicates perfect multicollinearity, meaning that one predictor variable is a perfect linear combination of one or more other predictor variables.
Let’s discuss with the formula:
VIF=1/1-R^2
If VIF is infinite that means R^2 is 1 and that explains everything.

Here are some reasons why VIF might be infinite:
Perfect Multicollinearity: This occurs when there is an exact linear relationship between two or more predictor variables. For example, if you have two variables X1 and X2 such that X2=aX1+b (where a and b are constants), the VIF for X1 and X2 will be infinite.

Dummy Variable Trap: When using dummy variables for categorical data, including a full set of dummy variables without excluding one category can lead to perfect multicollinearity. For example, if you have a categorical variable with three categories and you create three dummy variables without dropping one, the sum of the dummy variables will always equal 1, leading to perfect multicollinearity.

Redundant Variables: Including redundant predictors that are linear combinations of other predictors can cause infinite VIF values. For example, if you include both total score and individual component scores of a test in the model, this redundancy can cause perfect multicollinearity.

6. 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 tool used to assess whether a set of data follows a particular distribution, typically a normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution. If the data follow the specified distribution, the points will lie approximately on a straight line.

Interpreting a Q-Q Plot

- **Straight Line**: If the points form a straight line, the data follow the theoretical distribution.
- **S-Shaped Curve**: An S-shaped curve indicates heavier tails (leptokurtic distribution).

- **Convex or Concave Curve**: A convex or concave curve suggests lighter tails (platykurtic distribution).

Example of a Q-Q Plot in Linear Regression

- **Fit the Model**: Fit a linear regression model to the data.
- **Obtain Residuals**: Calculate the residuals (observed values - predicted values).
- **Generate Q-Q Plot**: Create a Q-Q plot of the residuals against a normal distribution.