Assignment 2

Problem statement

Perform the following operations using R/Python on the data sets
a) Compute and display summary statistics for each feature available in the dataset. (e.g. minimum value, maximum value, mean, range, standard deviation, variance and percentiles
b) Data Visualization-Create a histogram for each feature in the dataset to illustrate the feature distributions.
c) Data cleaning, Data integration, Data transformation, Data model building (e.g. Classification).

S/W Packages and H/W apparatus used:

OS: **Ubuntu/Windows, Google Colab**
Packages: **Numpy, Pandas, Matplotlib and Seaborn**

Theory :

1. Methodology :

- **Compute and Display Summary Statistics:**
  - **Python (using pandas):**
    - Use **describe()** function to compute summary statistics.
  - **R:**
    - Use **summary()** function to compute summary statistics.
- **Data Visualization - Histogram Creation:**
  - **Python (using matplotlib or seaborn):**
    - Use **hist()** function to create histograms for each feature.
  - **R:**
    - Use **hist()** function to create histograms for each feature.
- **Data Cleaning, Integration, Transformation, Model Building:**
  - **Data Cleaning:**

        - Identify and handle missing values using techniques such as imputation or deletion.
    - **Data Integration:**
        - Merge or join multiple datasets based on common variables.
    - **Data Transformation:**
        - Normalize or scale features, encode categorical variables, and handle outliers.
    - **Model Building:**
        - Split data into training and testing sets.
        - Choose an appropriate machine learning algorithm (e.g., classification algorithm).
        - Train the model on the training data and evaluate its performance on the testing data.

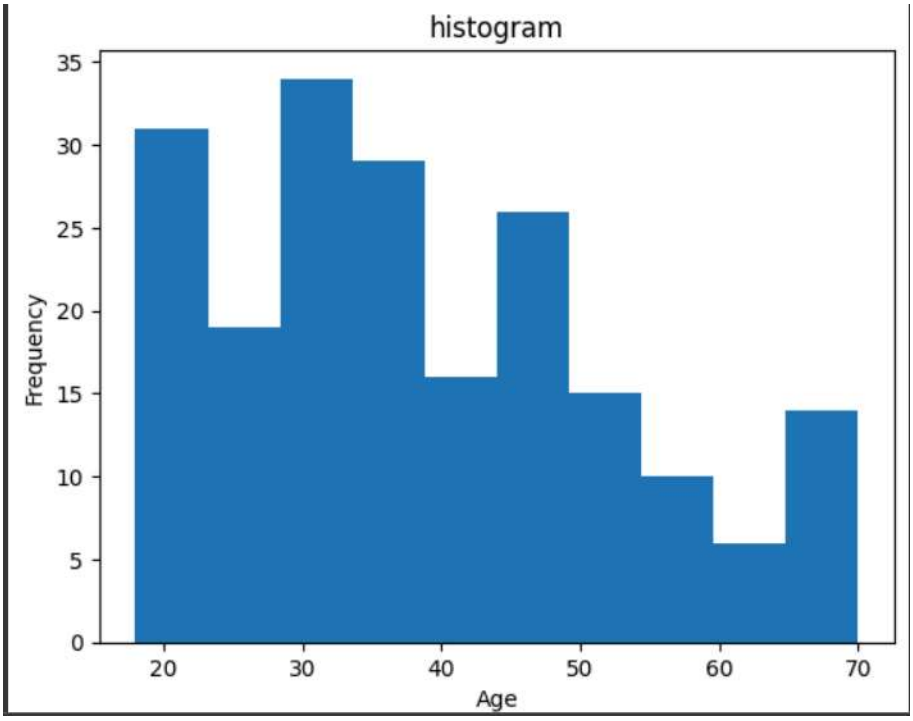
2. Advantages and Disadvantages & Limitations/Example:

1. **Advantages:**
    - **Summary Statistics:**
        - Provides a quick overview of the dataset's characteristics.
        - Helps in identifying outliers and understanding the distribution of features.
    - **Data Visualization:**
        - Enables intuitive understanding of feature distributions.
        - Facilitates identification of patterns and trends in the data.
    - **Data Cleaning, Integration, Transformation, Model Building:**
        - Enhances data quality and prepares it for analysis.
        - Facilitates the development of predictive models for classification tasks.
2. **Disadvantages & Limitations/Example:**
    - **Summary Statistics:**
        - May not capture all nuances of the data distribution, especially in complex datasets.
        - Outliers can skew summary statistics, affecting their interpretability.
    - **Data Visualization:**

- Histograms may not provide sufficient detail for understanding complex relationships.
- Interpretation of histograms can be subjective and influenced by binning choices.

- **Data Cleaning, Integration, Transformation, Model Building:**
  - Data cleaning and transformation can be time-consuming, especially for large datasets.
  - Model performance heavily depends on data quality, feature selection, and algorithm choice.

Diagram :



Conclusion :

The methodology involves using R and Python for data analysis, including computing summary statistics, creating histograms, and cleaning and transforming data for classification modeling.

Though versatile, these methods might oversimplify data and lack detailed insights. However, they can still provide valuable insights for decision-making when used carefully.