Name: Pathan Firdos Maheraj
Roll no: 281073
Batch: A3

Assignment 2

Statement:

Q. 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) Illustrate the feature distributions using histograms.
c) Data cleaning, Data integration, Data transformation, Data model building (e.g., Classification).

Objective:

1. This assignment aims to analyze and preprocess the dataset using various statistical and visualization techniques.
2. Learn how to compute and interpret summary statistics for different features.
3. Visualize feature distributions to understand data distribution.
4. Perform essential data cleaning, integration, and transformation steps.
5. Build a classification model for predictive analysis.

Resources used:

1. Software used: Google Colab
2. Libraries used: Pandas, Scikit-learn, Matplotlib, Seaborn

Introduction to Data Analysis and Classification:

1. Data analysis involves summarizing, visualizing, and preparing data for modeling.
2. Classification models predict categorical outcomes based on input features.
3. The dataset contains various maternal health attributes such as blood pressure, glucose levels, heart rate, and risk labels.

Methodology:

1. **Computing Summary Statistics:**
   - Calculate minimum, maximum, mean, range, standard deviation, variance, and percentiles for each feature.
2. **Feature Distribution Visualization:**
   - Use histograms to display the distribution of numerical features.
3. **Data Cleaning and Preprocessing:**
   - Handle missing values and remove inconsistencies.
   - Normalize or scale numerical features if required.
4. **Data Integration and Transformation:**
   - Merge datasets if needed and encode categorical variables.
   - Apply feature engineering techniques.
5. **Model Building (Classification):**
   - Choose a suitable classification algorithm (e.g., Logistic Regression, Decision Tree, Random Forest, or SVM).
   - Train and evaluate the model on the dataset.

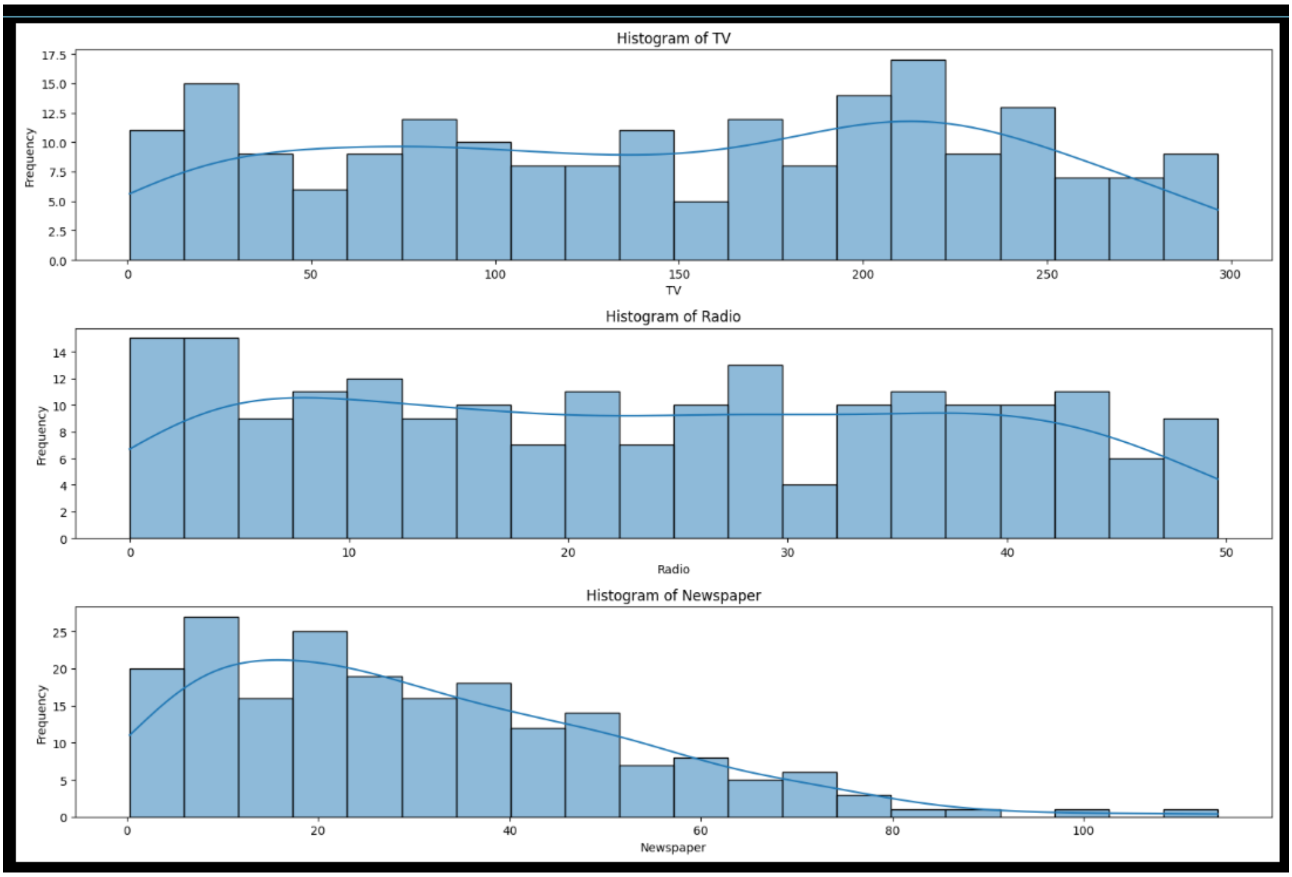
Advantages:

1. Helps in understanding data characteristics and distributions.
2. Improves predictive model accuracy through data preprocessing.
3. Enables better decision-making in healthcare applications.

Disadvantages:

1. Requires proper handling of missing and inconsistent data.
2. Model performance may vary based on dataset quality and preprocessing techniques.

Results:



Conclusion:

In this assignment, we analyzed the dataset by computing summary statistics and visualizing feature distributions. We performed essential data preprocessing steps such as cleaning, integration, and transformation. Finally, we implemented a classification model to predict maternal health risks. These steps are crucial for effective data-driven decision-making in healthcare analytics.