Name: Akash Patel
Roll no: 281029
Batch: A2

Assignment 3

Statement:

Q. Visualize the data using R/Python by plotting the graphs for assignment no. 1 and 2. Consider a suitable dataset. Use the following types of plots:
a) Scatter Plot
b) Bar Plot
c) Box Plot
d) Pie Chart
e) Line Chart

Objective:

1. This assignment focuses on exploring and preparing a dataset through statistical and visualization techniques.
2. Develop an understanding of computing and interpreting summary statistics for various features.
3. Leverage visualizations to examine data distributions and identify patterns.
4. Execute essential data cleaning, integration, and transformation procedures.
5. Construct a classification model for predictive analysis.

Resources Used:

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

Introduction to Data Analysis and Classification:

1. Data analysis involves summarizing, visualizing, and refining datasets for effective modeling.
2. Classification models predict categorical outcomes using input features.
3. The dataset includes maternal health indicators such as blood pressure, glucose levels, heart rate, and corresponding risk labels.

Methodology:

1. Summary Statistics Computation:
   - Calculate essential statistical measures, including minimum, maximum, mean, range, standard deviation, variance, and percentiles for each feature.
2. Feature Distribution Visualization:
   - Utilize histograms and other visualization techniques to analyze numerical feature distributions.

3. Data Cleaning and Preprocessing:
    - Detect and manage missing values while addressing inconsistencies.
    - Normalize or scale numerical attributes when required.
4. Data Integration and Transformation:
    - Merge datasets where necessary and encode categorical variables for improved processing.
    - Apply feature engineering to enhance data representation.
5. Model Development (Classification):
    - Select an appropriate classification algorithm, such as Logistic Regression, Decision Tree, Random Forest, or SVM.
    - Train and assess the model using relevant performance metrics.

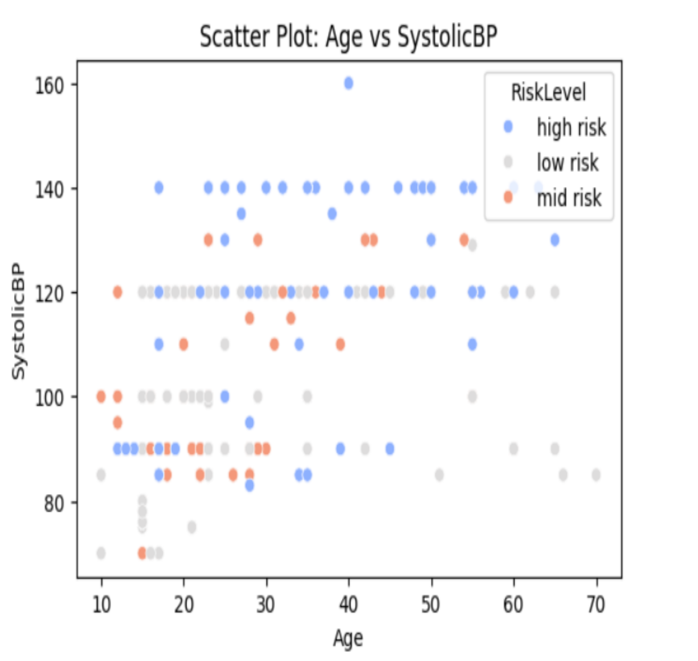
Advantages:

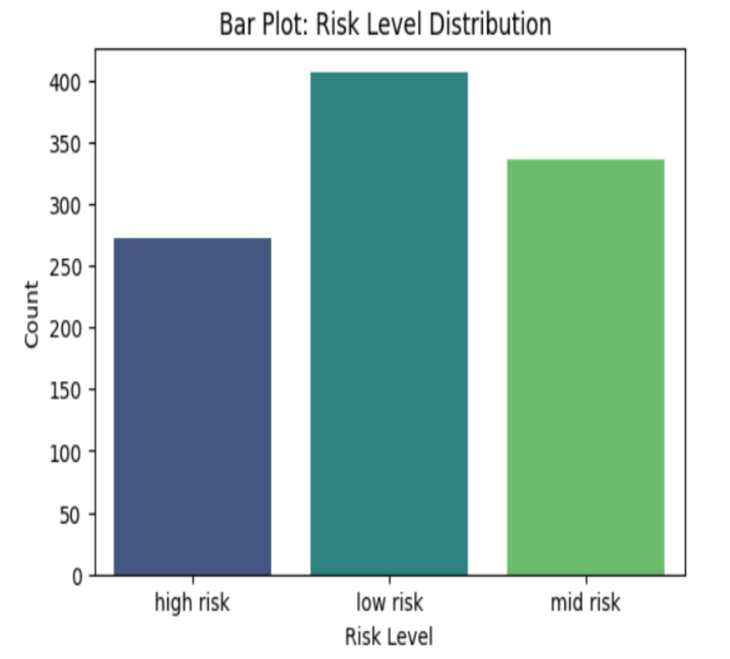
1. Strengthens comprehension of data attributes and distributions.
2. Enhances predictive model accuracy through effective preprocessing.
3. Supports data-driven decision-making, particularly in healthcare applications.

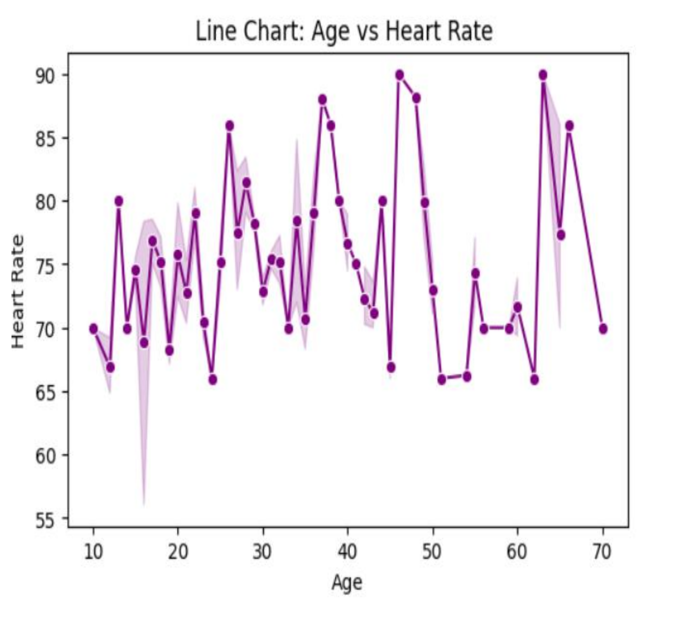
Disadvantages:

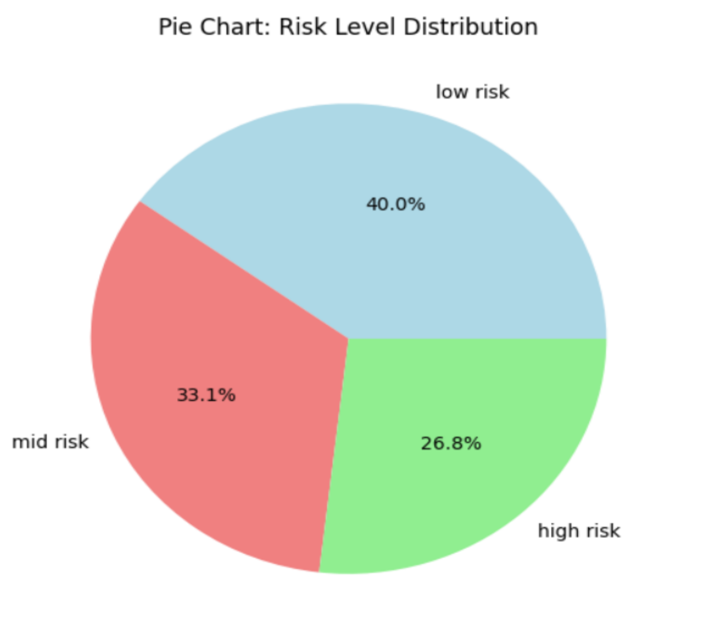
1. Requires meticulous handling of missing and inconsistent data to prevent biases.
2. Model performance is influenced by dataset quality and preprocessing strategies.

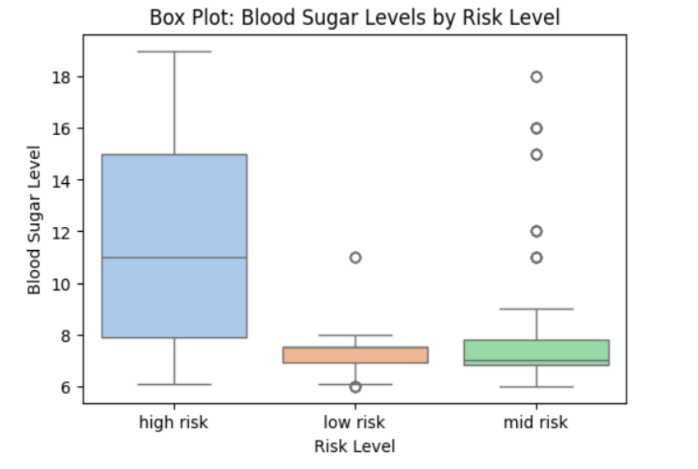
Graphs:











Conclusion:

This assignment provided practical experience in dataset analysis by computing summary statistics and visualizing feature distributions. We implemented key preprocessing steps, including data cleaning, integration, and transformation. Finally, we developed and assessed a classification model to predict maternal health risks. These processes are essential for making informed, data-driven decisions in healthcare analytics.