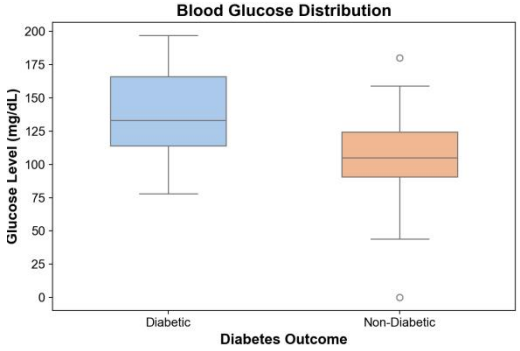
Individual Contribution			
CWID	Name	Contribution (description)	Percent Contribution
A20563452	Changming Yao	Find datasets and write code	33.3%
A20563416	Daiyang Chen	Modify code and write reports	33.3%
A20563428	Haozhe Ye	Modify code and write reports	33.3%

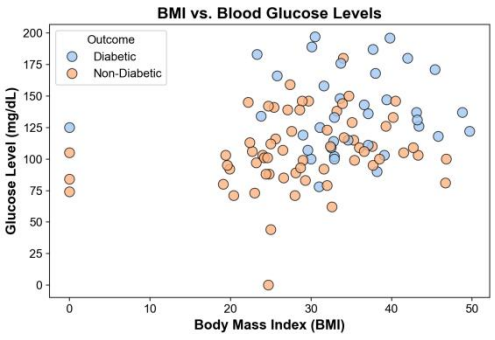
Report on the Diabetes Dataset.

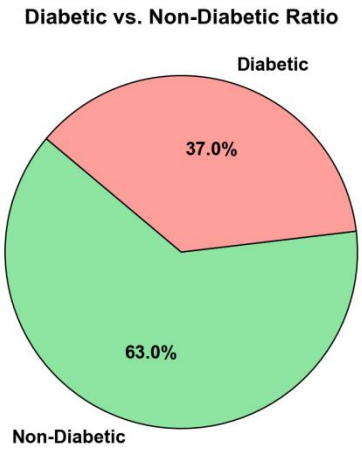
1.Dataset Description.

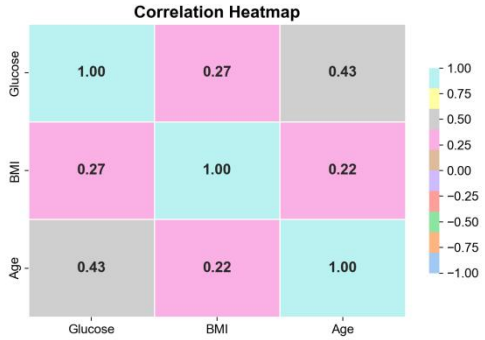
This dataset records health information of diabetic patients, containing 768 entries. Each entry represents a patient's physical examination information, including the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, age, and other features. The last column "Outcome" indicates whether the patient has diabetes (1 represents having diabetes, 0 represents not having diabetes).

2.Display Visual Charts.









3.Methods for Creating Charts.

First, the data preprocessing selects the top 100 rows and key feature columns, while converting categorical labels into text labels. By using a consistent font and color scheme, the chart style remains uniform. The code sequentially creates four types of charts: a box plot showing the distribution differences in glucose levels between diabetic and non-diabetic groups, a scatter plot visually displaying the relationship between BMI and glucose levels with color indicating disease status, a pie chart illustrating the proportion of diabetic cases, and finally, a heatmap revealing the correlations between glucose, BMI, and age.

4.Programming Libraries Used.

Pandas: Used for data reading, filtering, and preprocessing, such as loading CSV files, selecting specific columns, and converting categorical labels.

Matplotlib: Provides basic plotting functions, responsible for creating charts, setting titles, labels, and overall layout.

Seaborn: An advanced visualization library based on Matplotlib, simplifying data visualization and offering aesthetically pleasing default themes, mainly used for drawing box plots, scatter plots, and heatmaps.

Numpy: Underlying library potentially used by Seaborn and Pandas for data calculations and matrix operations.

5.Analysis Results and Insights for the Selected Dataset.

It reveals important relationships between variables in the dataset, particularly the impact of BMI and glucose levels on diabetes. Diabetic patients in the dataset tend to have higher BMI and glucose levels, which can help us consider BMI and glucose

as key features when building predictive models. Additionally, the charts also display the diabetes prevalence and correlations, providing a basis for further analysis of potential risk factors for diabetes, with a particular emphasis on the influence of BMI on diabetes.