مدينة زويل

ZEWAIL CITY

Course:Statistical Inference

Course Code: DSAI 307

University of Science and Technology at ZewailCity

Academic Year : Fall 2024-2025

Name	ID
Reem Ehab Helmy	202201373
Tasneem Ashraf El Sadaany	202201573
Hanin Khaled	202201043
Abdulrahman Omar	202202254

Analysis of Diabetes Dataset

Introduction

In order to investigate several health metrics and their correlations with diabetes status, This report presents an analysis of a publicly available diabetes dataset. Finding important trends and connections between important characteristics like age, BMI, glucose levels, and other variables is the goal.to investigate potential relationships between these factors and diabetes.

The primary goal of this research is to explore the dataset to identify trends and validate specific claims using statistical inference. Our analysis has the following objectives: first, to conduct exploratory data analysis to understand the key characteristics of the dataset. Second, to perform hypothesis tests on claims about relationships between different variables. Third, to apply confidence interval analysis to determine the coverage of confidence intervals. Finally, getting a conclusion of results to understand the factors of diabetes.

Exploring Data Analysis “EDA”

In this Chapter, we will explore the key characteristics of the diabetes dataset and explore the relationships between features. First,To deal with missing values and guarantee statistical validity, the dataset was cleaned and preprocessed. Bar charts, histograms, and scatter plots were used to depict key data in order to identify trends and distinctions .

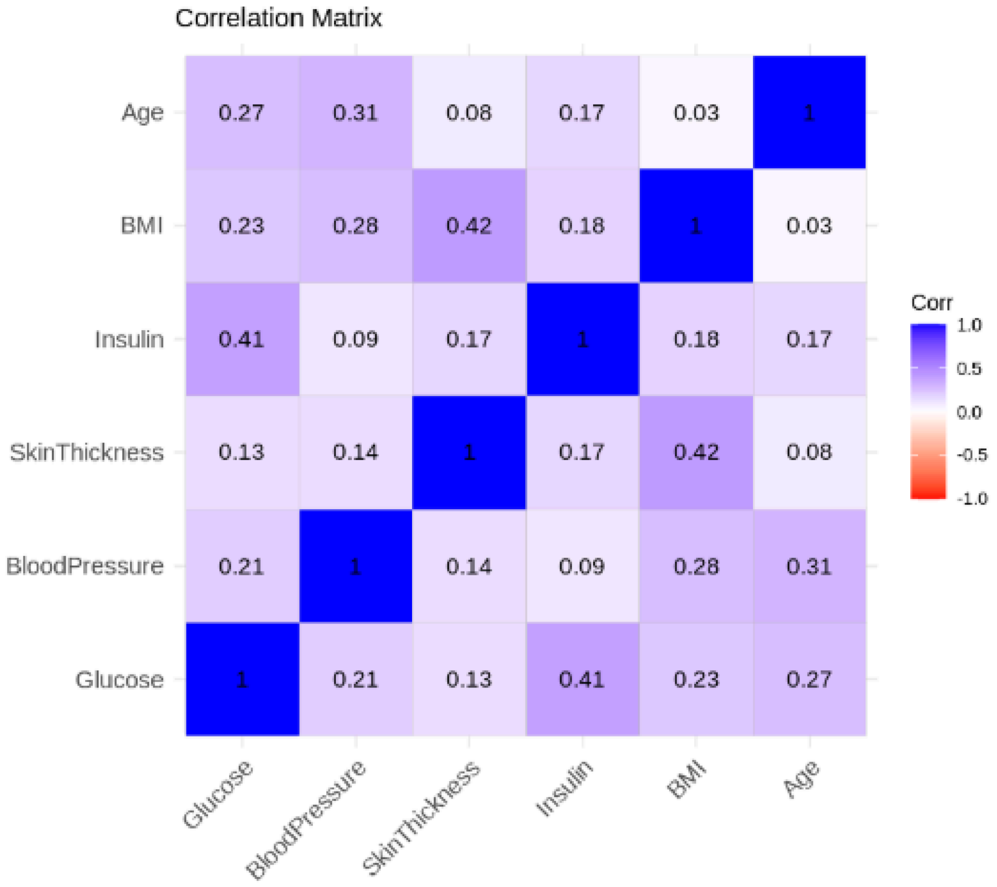
Step 1 : check for missing values and replace it if there is missing value

Step 2 : checking for outliers

Step 3 : for outliers for columns “ glucose , insuline , BMI “ we used cap , for columns “ bloodpressure , skinthikness ” filled with the mean

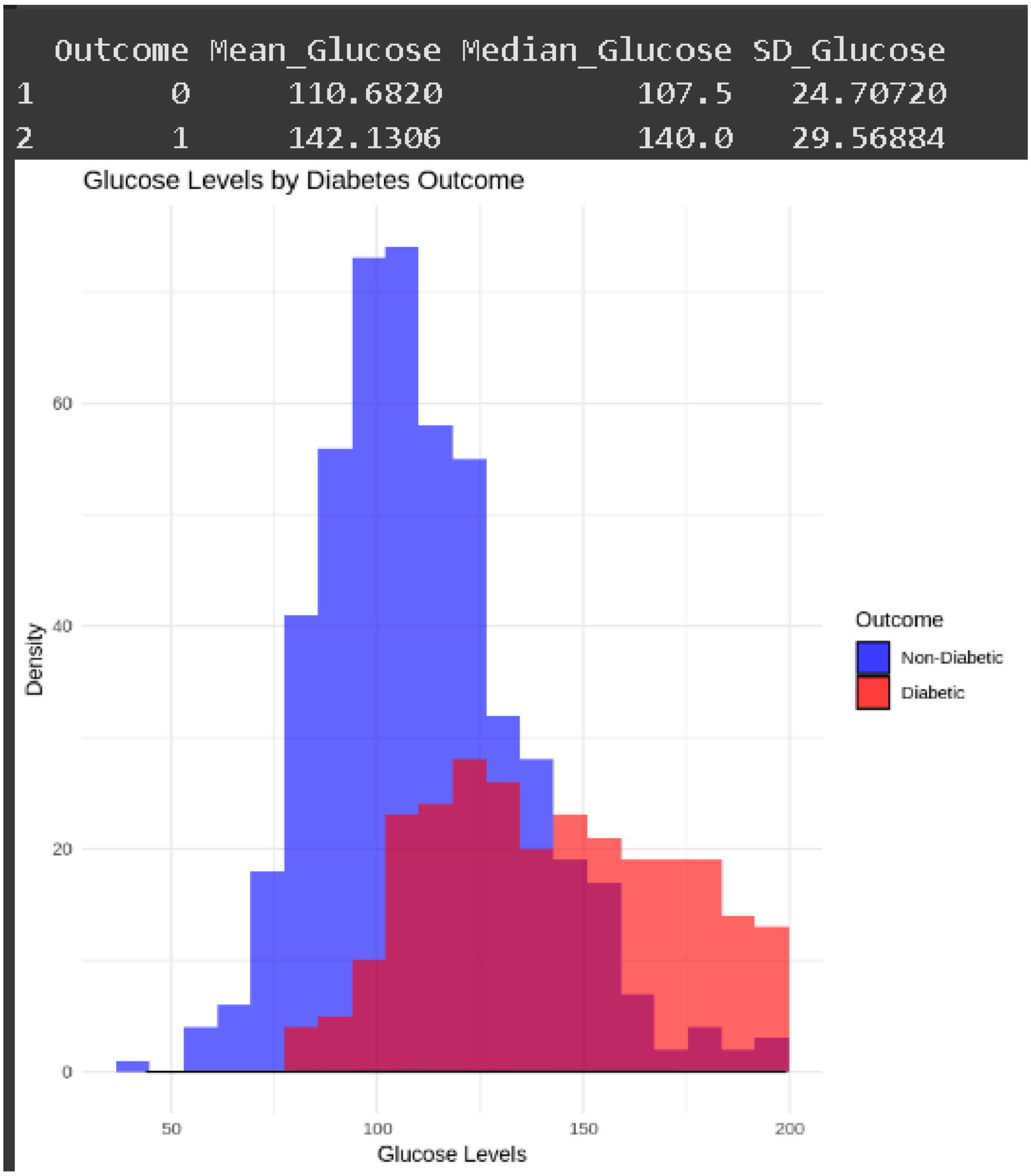
Step 4 : checking and saving the new preprocessed dataset

Step 5 : correlation matrix for numerical columns to see the relation between data features



Results and Visualizations

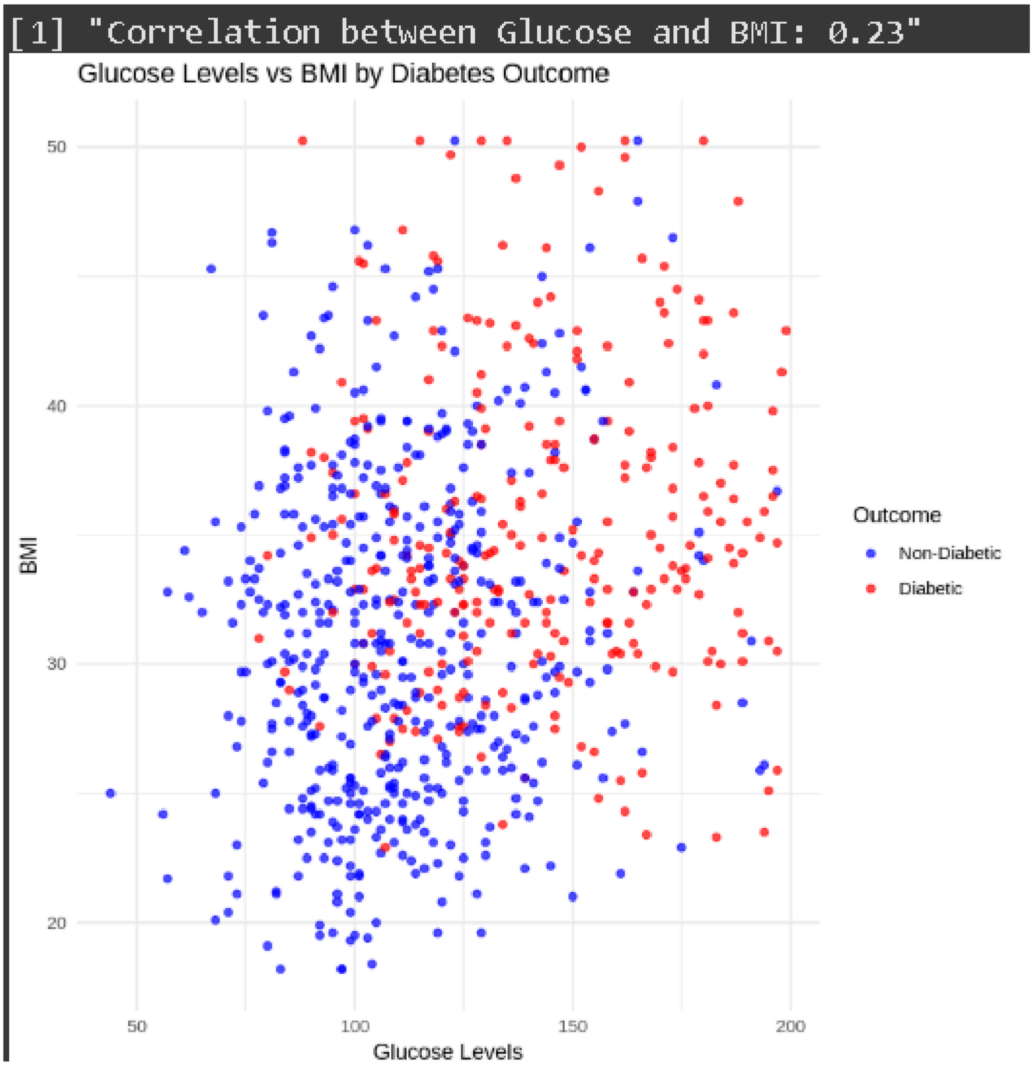
- 2.1 Are higher glucose levels associated with a greater likelihood of diabetes?
- Histogram with density for glucose levels by outcome
- plot clearly show that diabetic patients have significantly higher glucose levels than non-diabetic patients.



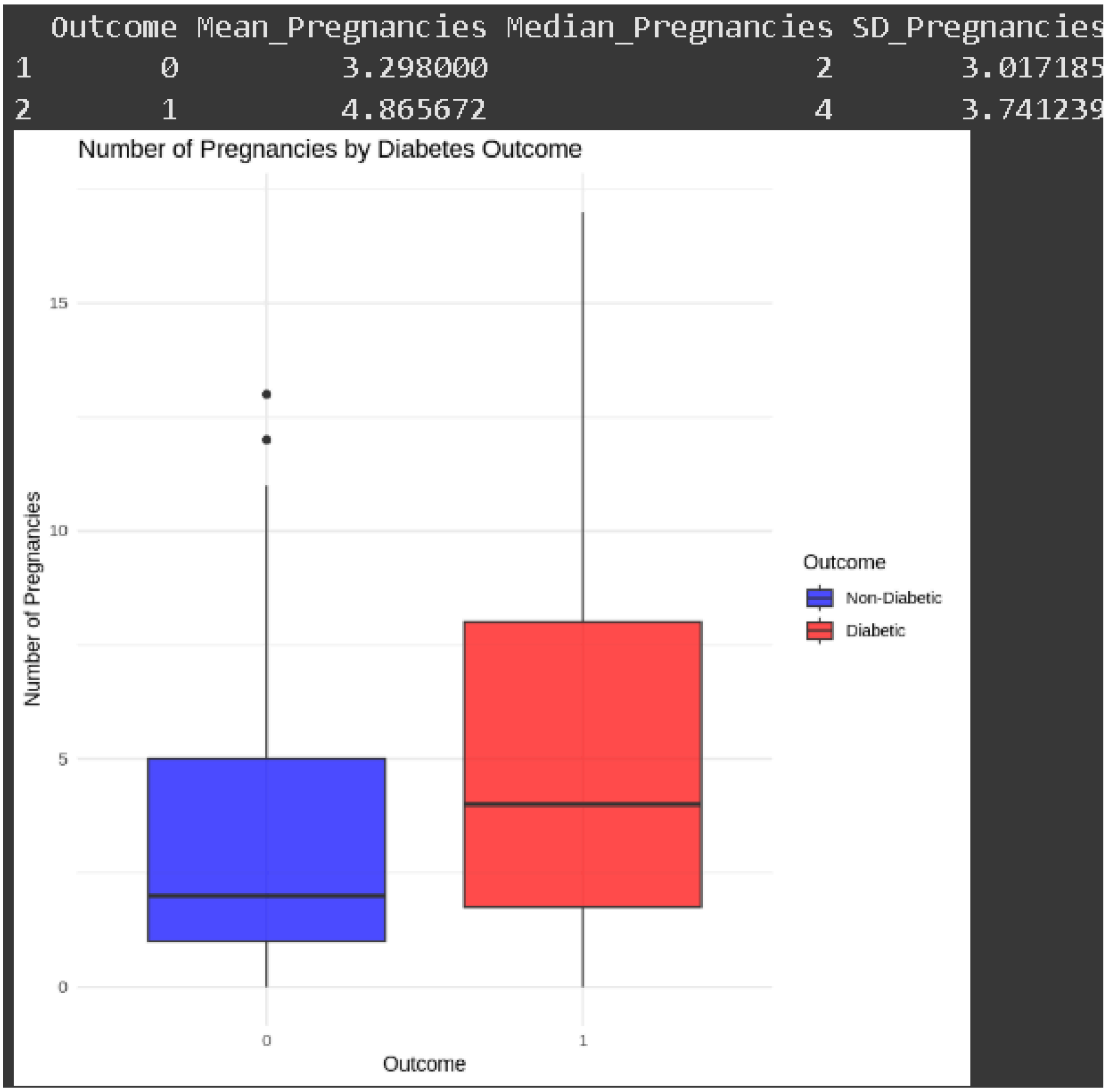
```
  Outcome Mean_Glucose Median_Glucose SD_Glucose
1       0     110.6820          107.5   24.70720
2       1     142.1306          140.0   29.56884
```

- 2.2 Are patients with high glucose concentrations also likely to have higher BMI values?

- Scatter plot of Glucose vs BMI
- giving the 0.23 correlation Patients with higher glucose levels are more likely to have higher BMI values this is more common among diabetic patients.

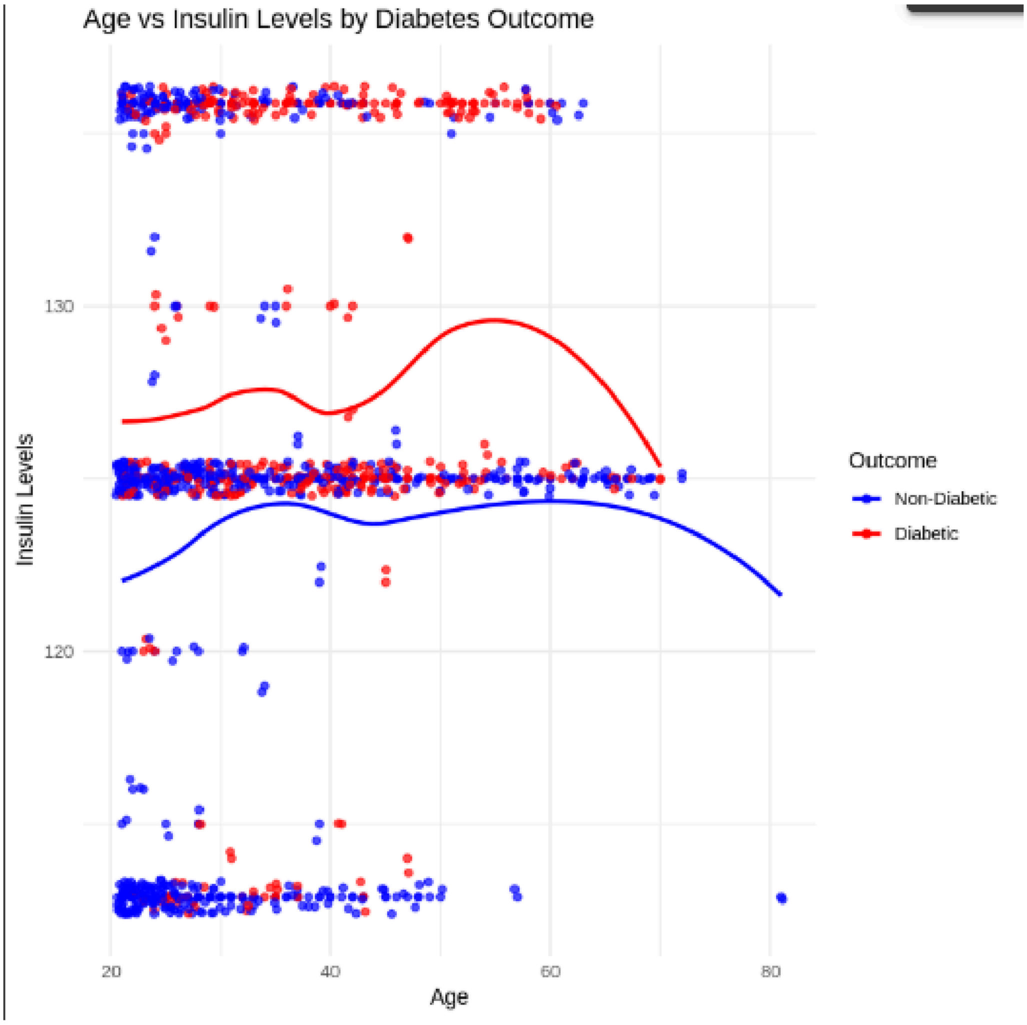


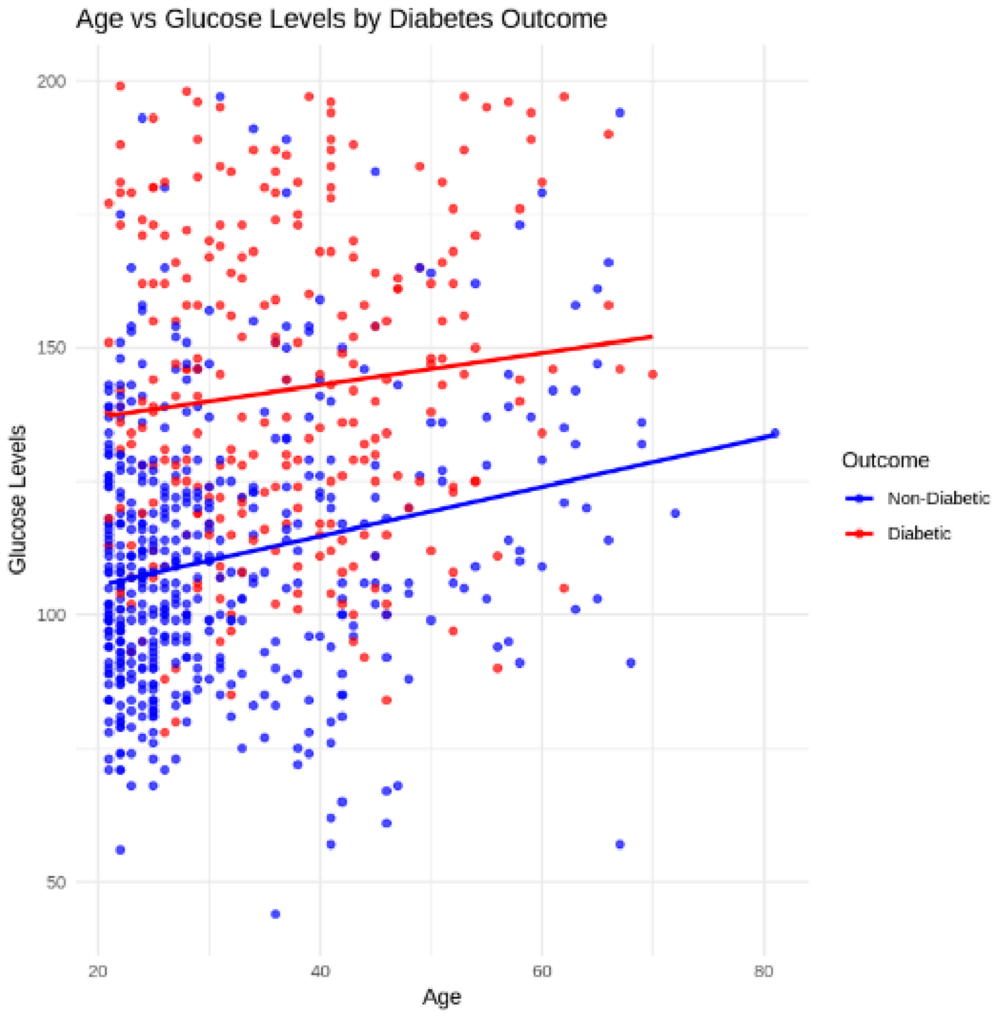
```
[1] "Correlation between Glucose and BMI: 0.23"
```

- 2.3 Are patients with a higher number of pregnancies at greater risk of developing diabetes?
- Boxplot for number of pregnancies by diabetes outcome
- mean non diabetic : 3.30 , diabetic : 4.87 Patients with diabetes tend to have a higher number of pregnancies giving a potential risk factor.
- 

```
  Outcome Mean_Pregnancies Median_Pregnancies SD_Pregnancies
1       0         3.298000                  2       3.017185
2       1         4.865672                  4       3.741239
```

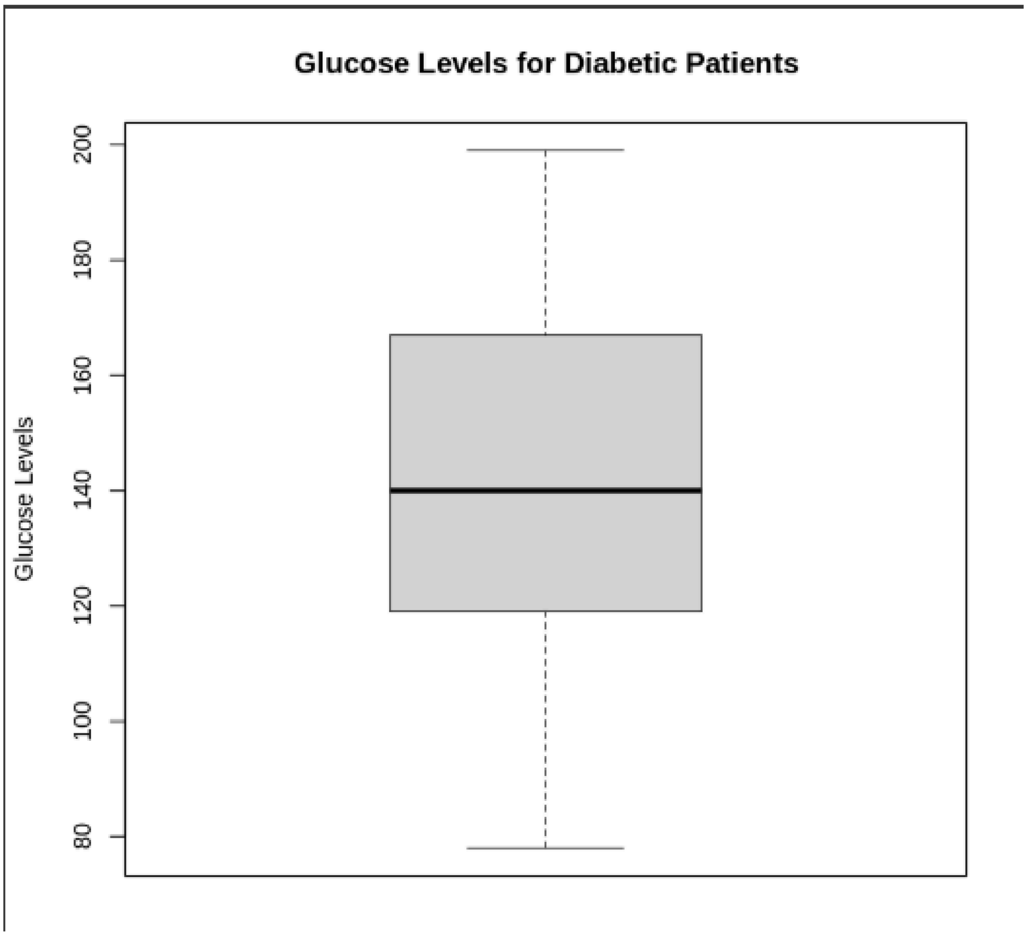
- 2.4 Are older patients more likely to have higher insulin concentrations and blood glucose levels?
- Boxplot for number of pregnancies by diabetes outcome
- mean non diabetic : 3.30 , diabetic : 4.87 Patients with diabetes tend to have a higher number of pregnancies giving a potential risk factor.

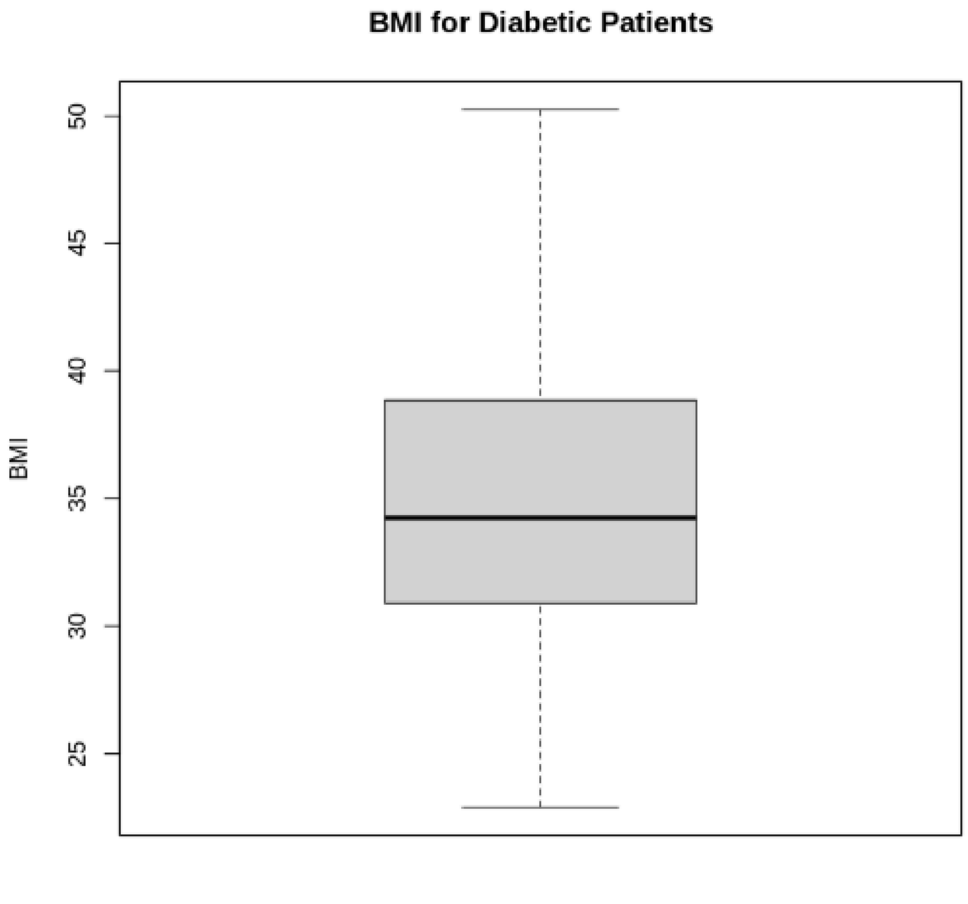


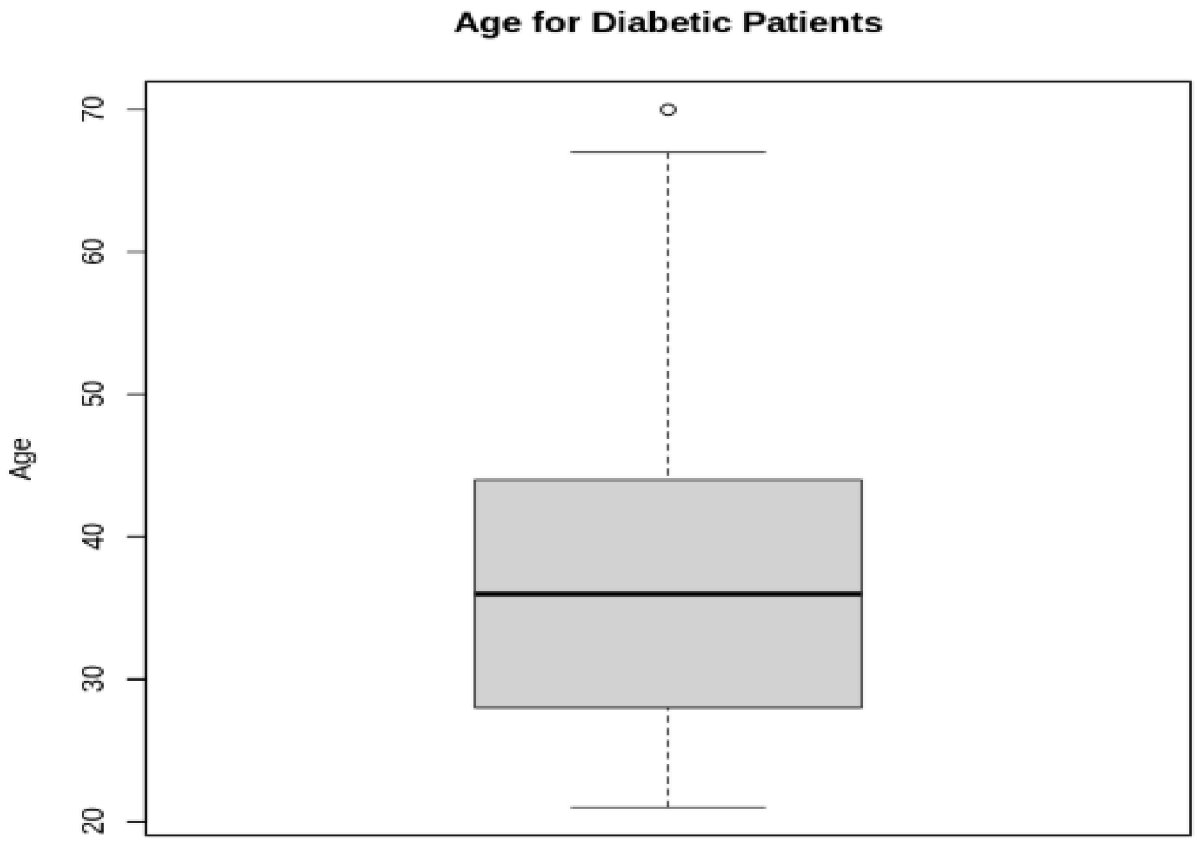


- 2.5 Can you identify common “risk profiles” for diabetic patients based on key metrics (glucose, BMI, age, etc.)?
- Box plot for each key metrics
- based on the calculations :
- Average Glucose Level: 142.13
- Average BMI: 35.24
- Average Age: 37.07 , diabetic patients have high glucose level and high bmi no specific pattern for the age but we can say that the are middle-aged

follow the second cell plot for more better visualization


Glucose Levels for Diabetic Patients
Glucose Levels
80
100
120
140
160
180
200


BMI for Diabetic Patients
BMI
50
45
40
35
30
25



High-Level Interpretations ;

Glucose Levels Are Key : High glucose is the most reliable indicator of diabetes. Monitoring and controlling glucose levels should be a top priority for prevention and treatment.

BMI and Obesity Matter: Higher BMI is strongly linked to diabetes, showing the critical role of managing weight in reducing diabetes risk.

Age Plays a Role: Older people are more likely to develop diabetes, but lifestyle factors like weight and glucose levels are even more influential.

Insulin Levels Are Complex: Insulin variability shows its importance but also highlights data issues. Glucose and BMI remain more reliable predictors for now.

What This Means :

The results confirm that focusing on weight management, glucose monitoring, and regular screenings can help prevent diabetes or catch it early. Addressing data gaps and including additional factors like diet and exercise could make future analyses even more effective.

Challenges, Limitations, and Assumptions

- **Challenges:** Possible biases in demographic representation and a small dataset size.and handling outliers
- **Limitations**: The dataset may not fully generalize across varied populations because it exclusively focuses on particular traits and specific demographics.
- Normal distributions for statistical tests and the accuracy of the recorded data are assumed in this research.

Conclusion

The analysis of the diabetes dataset provided valuable insights into the factors associated with diabetes. Key findings include:

Glucose Levels: Diabetic patients have consistently higher glucose levels compared to non-diabetic patients, making this the strongest predictor of diabetes.

BMI: Higher BMI is strongly linked to diabetes, reinforcing the connection between obesity and diabetes risk.

Age: Older patients are more likely to have diabetes, though the relationship is less pronounced compared to glucose and BMI.

Insulin Levels: The variability in insulin levels suggests potential as a predictor, but it requires careful preprocessing due to missing or extreme values.

Throughout the analysis, we identified and addressed issues like outliers, which helped improve the quality of the insights.

Potential improvements :

Improve Data Quality: Ensure more accurate data collection and address missing values or unrealistic entries more effectively.

Expand Features: Include additional factors like physical activity, diet, and family history of diabetes to enrich the analysis.

Build Predictive Models: Use machine learning algorithms to predict diabetes more accurately and explore which features contribute most to the predictions.