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”

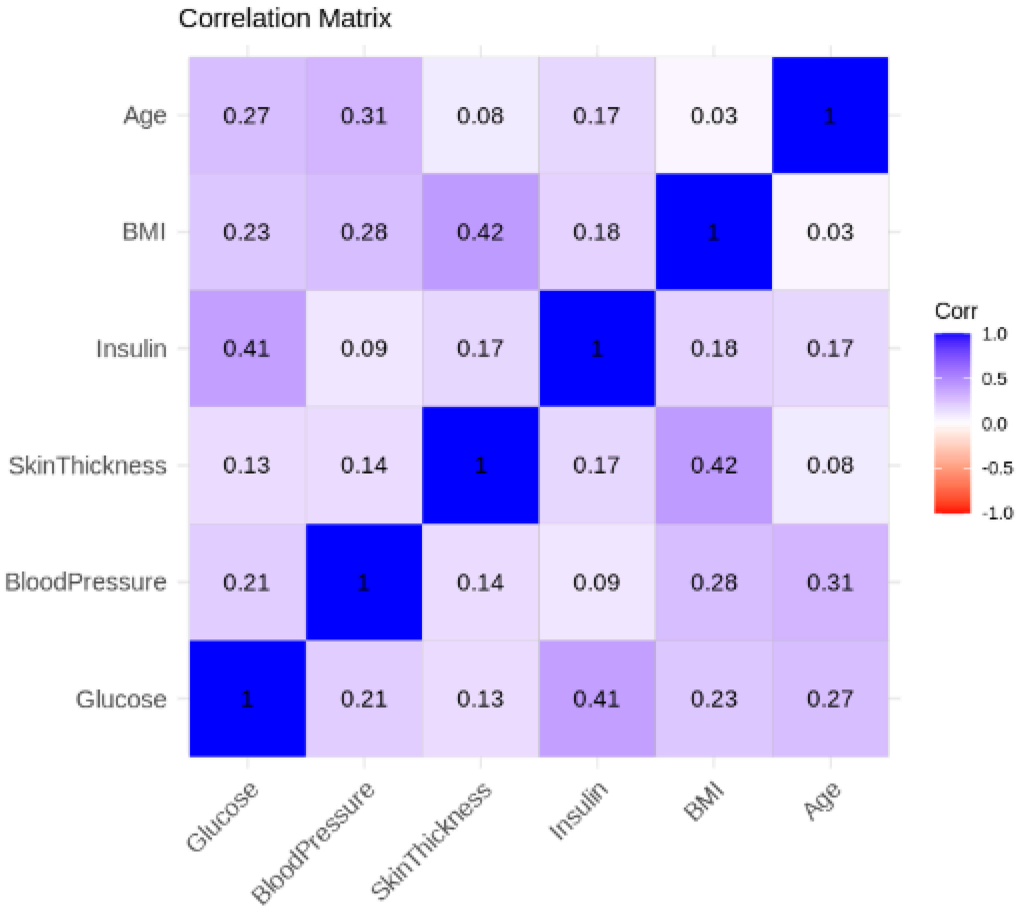
In this Chapter, we will explore the key characteristics of the diabetes dataset and explore the relationships between features. First,To deal with missing values and guarantee statistical validity, the dataset was cleaned and preprocessed. Bar charts, histograms, and scatter plots were used to depict key data in order to identify trends and distinctions between people with and without diabetes.

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

Step 2 : for outliers for 3 columns filled using cap , for 2 column using the mean

Step 3 : checking and saving the new preprocessed dataset

Step 4: correlation matrix for numerical columns to see the relation between data features



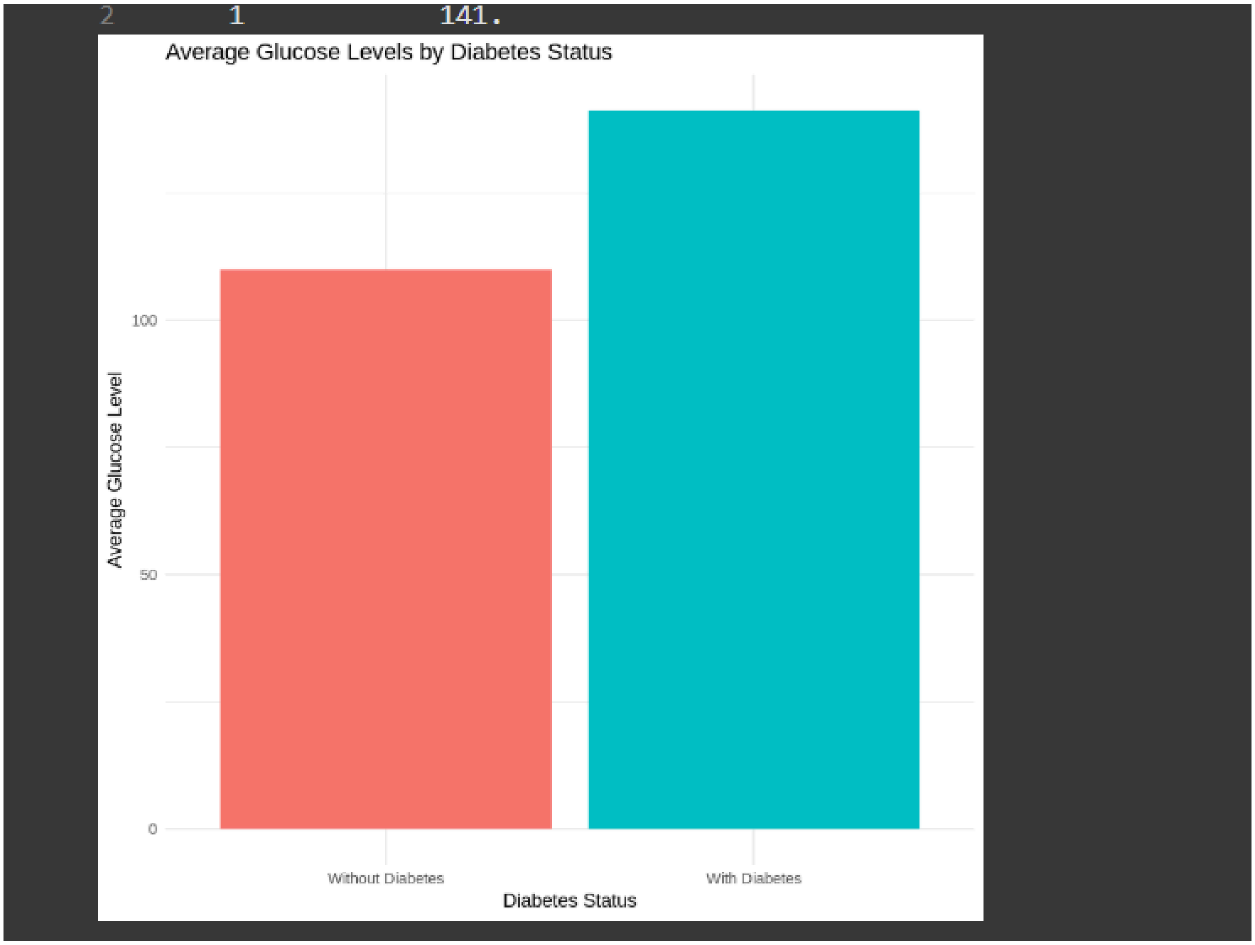
	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Age
Age	0.27	0.31	0.08	0.17	0.03	1
BMI	0.23	0.28	0.42	0.18	1	0.03
Insulin	0.41	0.09	0.17	1	0.18	0.17
SkinThickness	0.13	0.14	1	0.17	0.42	0.08
BloodPressure	0.21	1	0.14	0.09	0.28	0.31
Glucose	1	0.21	0.13	0.41	0.23	0.27

Results and Visualizations

- 3.1 Are higher glucose levels associated with a greater likelihood of diabetes?

  Diabetes is substantially linked to elevated glucose levels. People with diabetes have higher average blood glucose levels than people without the disease.

The average blood glucose levels of patients with diabetes (1) and those without diabetes (0) are depicted in this bar chart. It draws attention to the connection between blood glucose levels and the risk of developing diabetes.

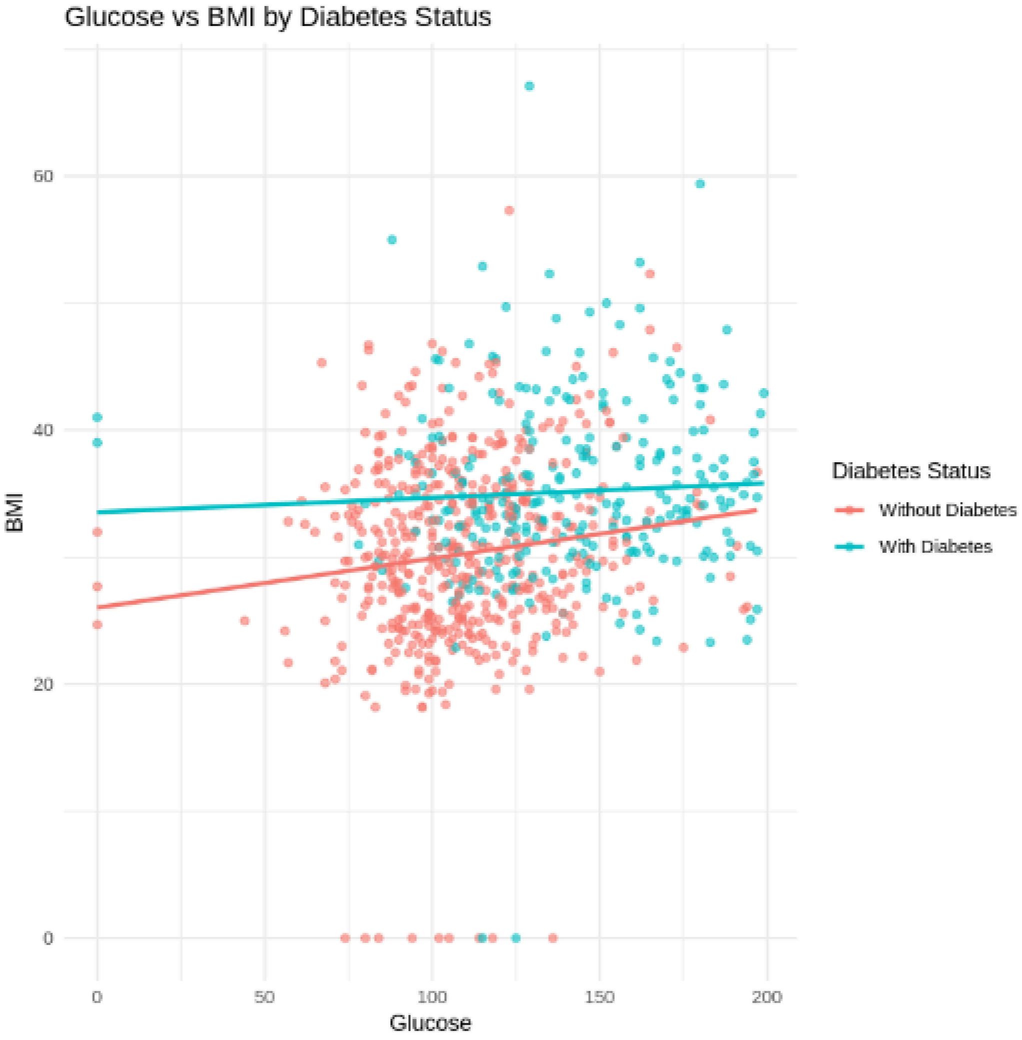


```
2          1                141.
```

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

 Among diabetic patients, there is a discernible relationship between elevated blood glucose levels and higher BMI values.

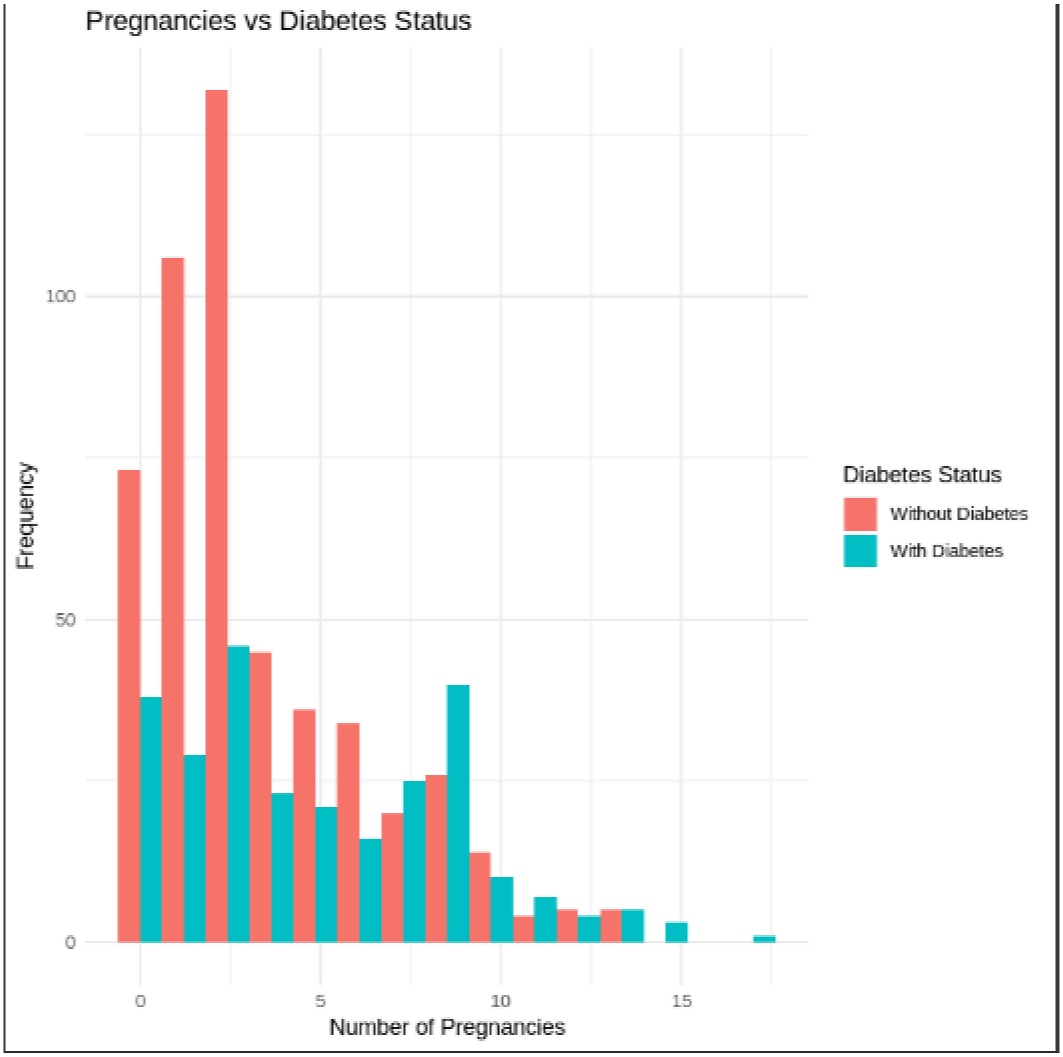
The association between BMI and glucose levels in patients with and without diabetes is investigated in this scatter plot. It aids in determining whether elevated blood glucose levels are linked to elevated body mass index.



- 3.3 Are patients with a higher number of pregnancies at greater risk of developing diabetes?

  The distribution suggests that patients with a larger number of pregnancies are more likely to have diabetes.

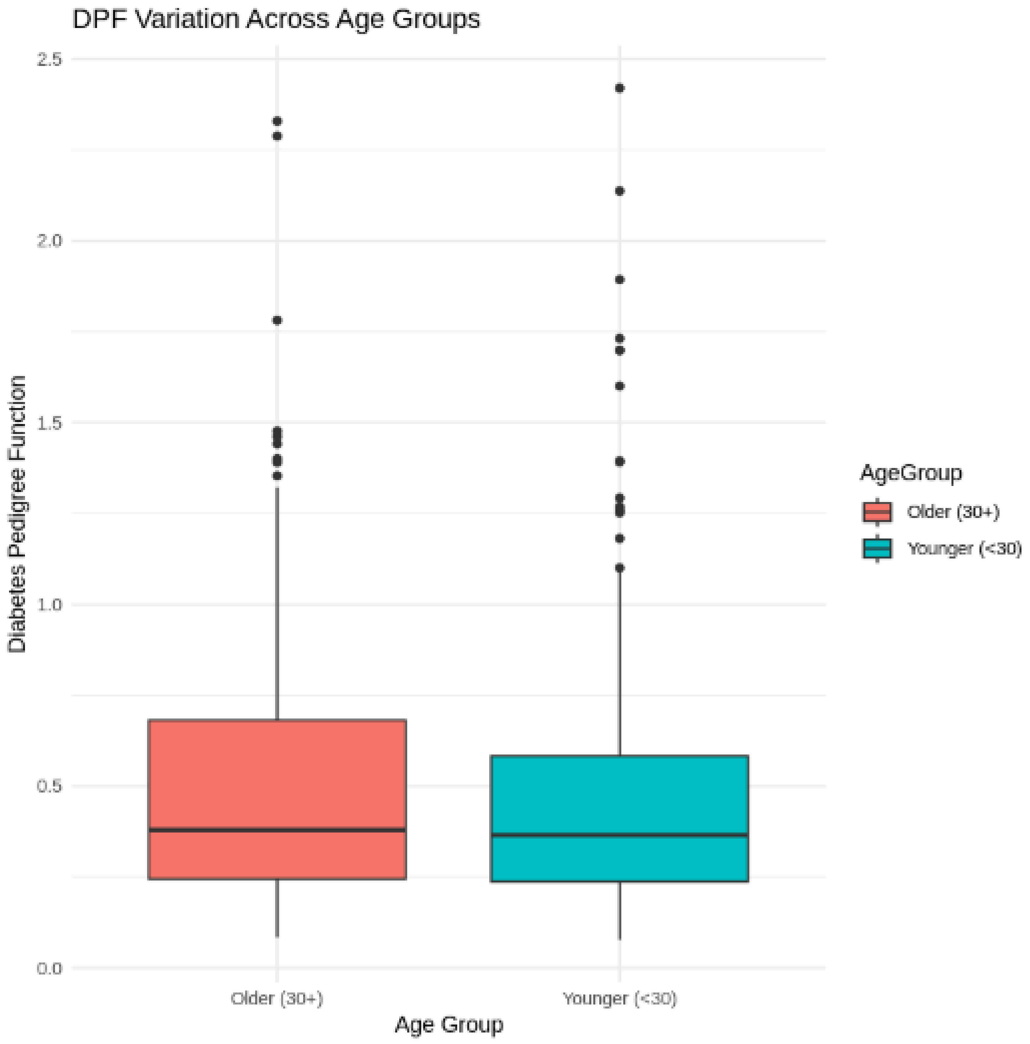
  The distribution of pregnancies amongst patients with and without diabetes is depicted in this histogram. It looks into whether those who had more pregnancies are more likely to get diabetes.



- 3.4 Are older patients more likely to have higher insulin concentrations and blood glucose levels?

  The visualizations show that older individuals typically have higher blood glucose levels and insulin concentrations.

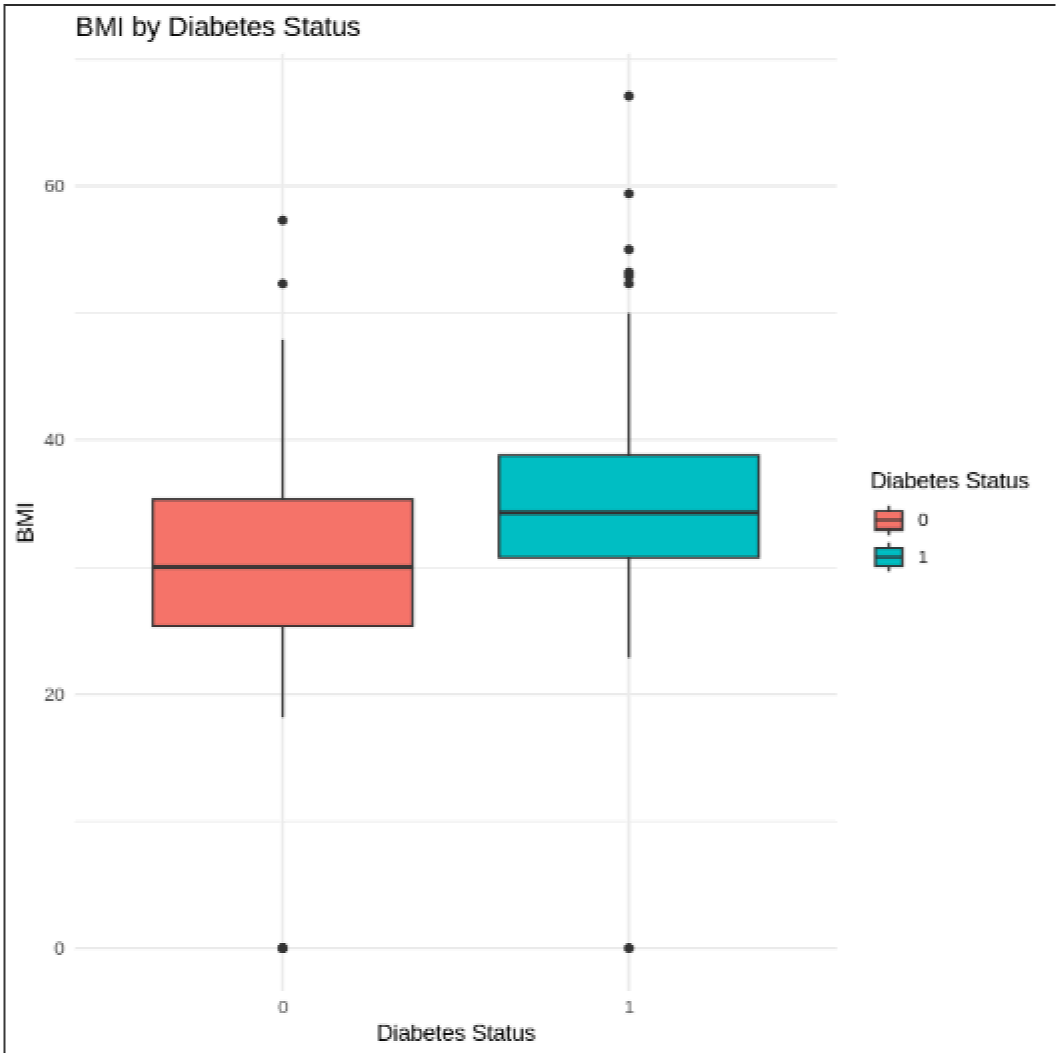
  The difference in Diabetes Pedigree Function (DPF) values between younger (less than 30 years old) and older (30+ years old) age groups is depicted in this box plot. It investigates if age affects DPF.



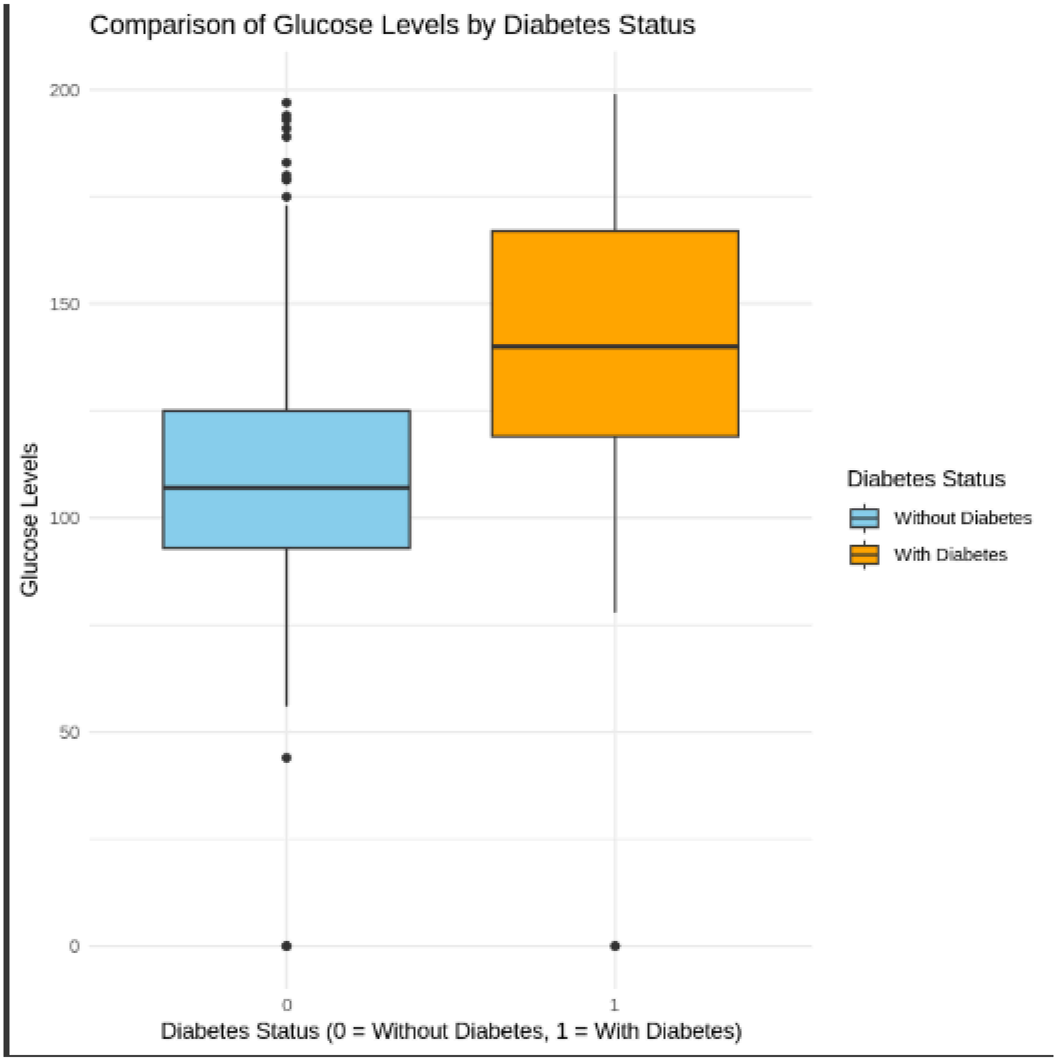
- 3.5 Can you identify common “risk profiles” for diabetic patients based on key metrics (glucose, BMI, age, etc.)?

  Patients with diabetes typically have higher BMIs, higher blood sugar levels, and are frequently older. Together, these traits create a typical risk profile for diabetes.

  The BMI distributions of persons with diabetes (1) and those without the disease (0) are contrasted in this box plot. It determines variations in BMI according to the presence of diabetes.



The glucose levels of patients with diabetes (1) and those without the disease (0) are contrasted in this box plot. It draws attention to variations in glucose levels.



Challenges, Limitations, and Assumptions

- Challenges: Possible biases in demographic representation and a small dataset size.and handling missing data
- 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

Significant relationships and trends between diabetes and health measures are highlighted in this study. According to the analysis, age, elevated BMI, and elevated glucose levels are all reliable markers of diabetes. To validate these results, further research could build on this analysis with bigger, more varied datasets.

Potential improvement: apply that in several dataset to get high level data with Advanced EDA techniques and apply Classification models