1. Introduction

This project is based on the diabetes data set available on Kaggle. The objective is to find potential health risks, causes of diabetes, and recommendations. The notions from "Storytelling with Data" are incorporated into analyzing the data, aiming to improve the visualization's efficiency and effectiveness.

1.1 Dataset Overview

- **Source:** Kaggle
- **Key Attributes:** Glucose, BMI, Age, Insulin, SkinThinkness, Blood Pressure, Pregnancies, Diabetes Pedigree Function (DPF), and Outcome (diabetic or non-diabetic).
- **Scope:** Essentially, total patient information is restricted to women patients who are twenty-one years and above, thus the results cannot be generalized.

1.2 Objectives

- Make descriptive analysis to identify patterns and wrinkles in data.
- Respond to significant questions about the risks and factors of developing diabetes.
- Support the claims using a hypothesis testing technique and the confidence interval.
- Give timely reports with good visualization and a perfect story to tell from the analysis.

2. Data Processing and Analysis Steps

2.1 Data Cleaning and Preprocessing

- Inspection: Examined data structure and column definition.
- Missing Values: Imputed for missing data statistically.
- Adding a column [outcome_categorical] to make the outcome categorical
- Split the dataset into two datasets, one for the diabetic and one for non-diabetic

2.2 Analytical Approach

The analysis was divided into the following parts:

1. Exploratory Analysis:
- Compared distribution and mean differences between Diabetic and non-diabetic subjects.
- Conducted correlation and trend analysis on main characteristics.

2. Research Questions:
- Focused specific questions to elicit relations and trends were used.

3. Hypothesis Testing:
- To confirm significance in groups as to glucose levels and other potential variables.

4. Simulation:
- Investigated confidence intervals, to compare levels of accuracy through different sample means.

3. Challenges, Limitations, and Assumptions

3.1 Challenges

Data Imbalance: Some analyses were affected due to a small proportion of diabetic patients in the dataset.

Limited Documentation: Exporting to a new variable from an access table was impossible without metadata; this meant making assumptions about some of the Variables.

3.2 Limitations

Demographic Scope: Limitations of the study relate to the generalizability of results only to the sample population of women.

Static Data: The data collected does not provide information about a patient's status at different points in time other than at the current time.

3.3 Assumptions

Everybody assumed that the data collected was accurate and complete.

Some features such as DPF were used as independent since the presence of relationships between variables could be referred to.

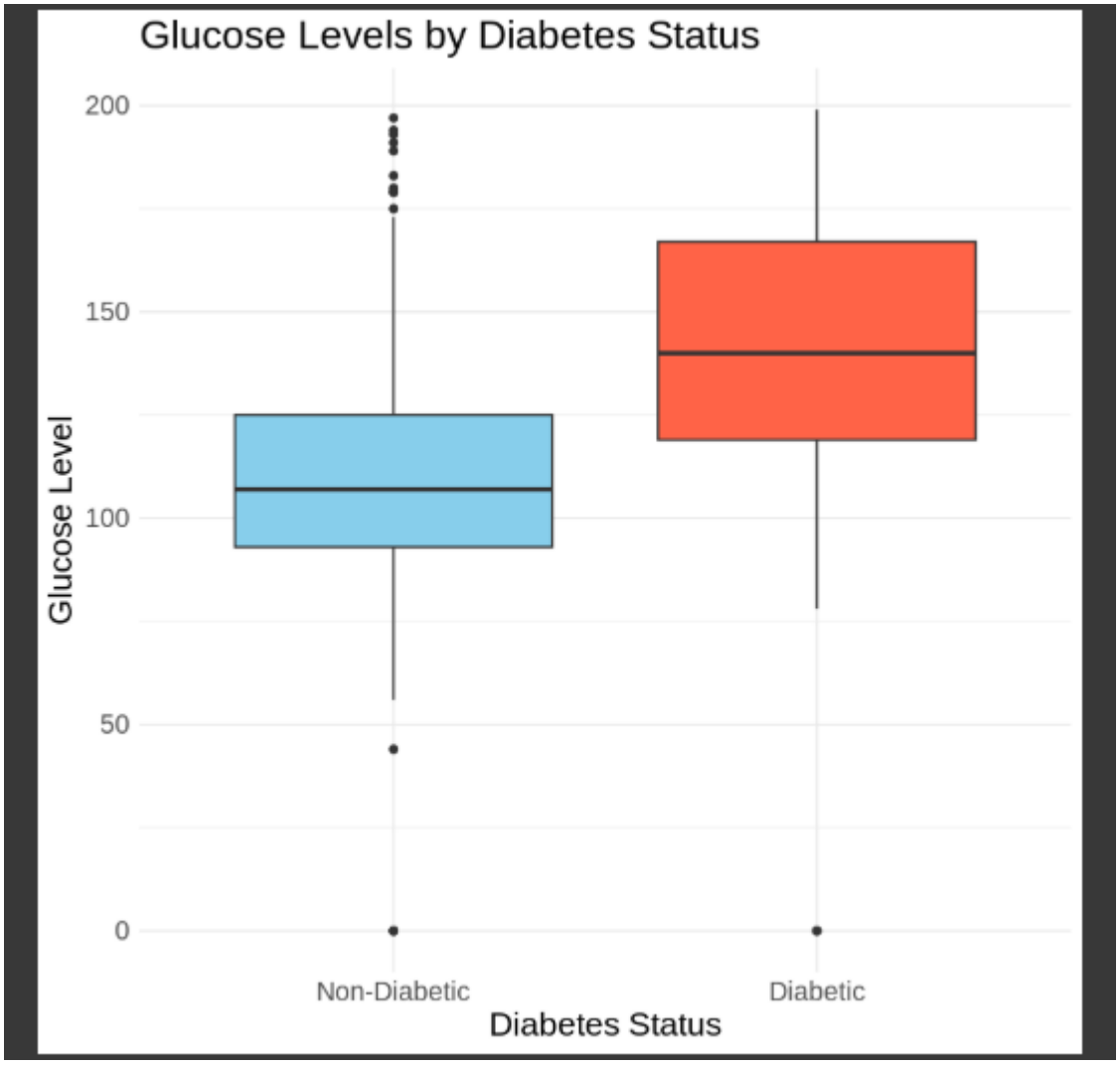
4. Results and Visualizations

4.1 Key Findings and Insights

1. Average Glucose Levels:

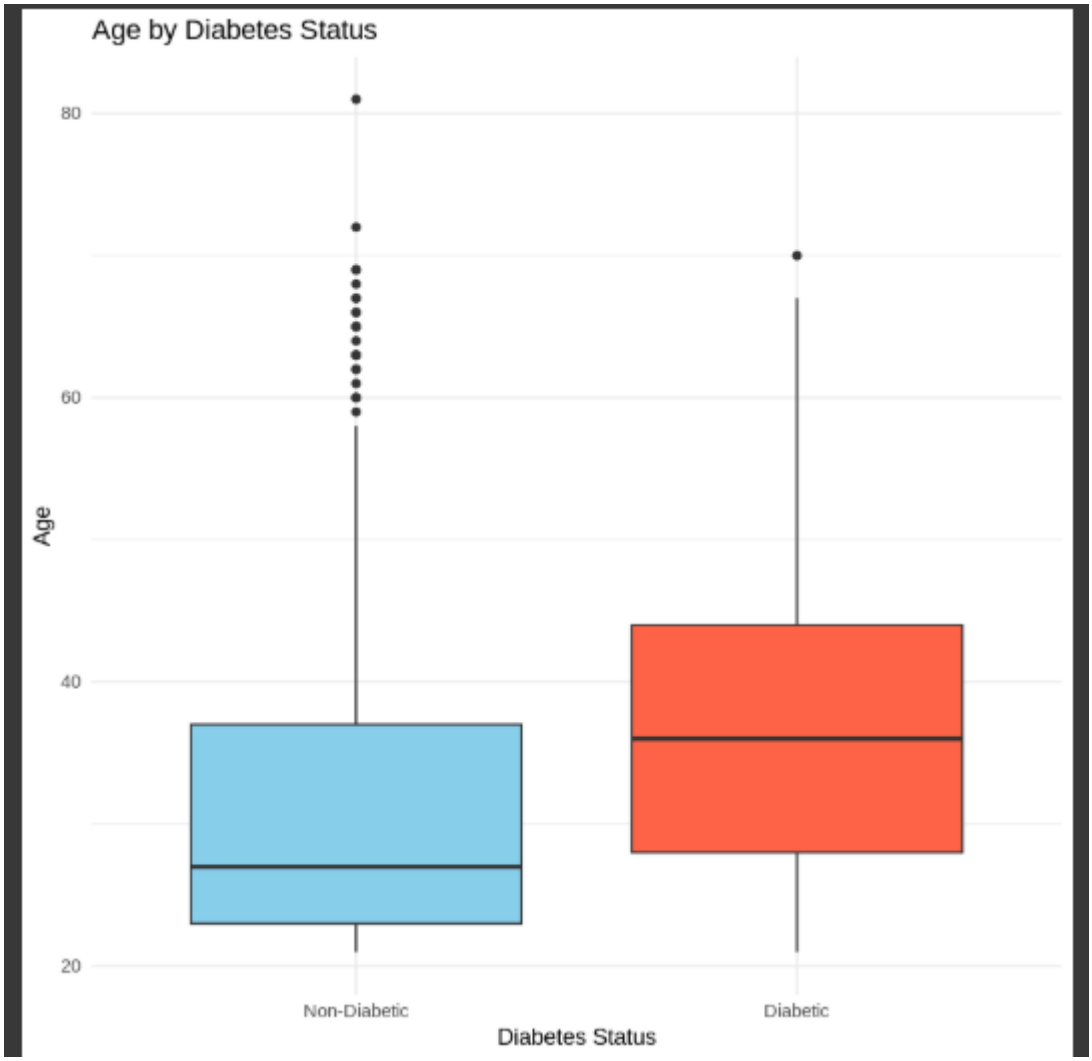
Patients with diabetes had an average glucose concentration of 132.46mg/dL while patients without diabetes had an average of 110.98mg/dL.

Visualization: In each case, boxplots provided a favorable view when comparing the distribution of glucose levels amongst the groups.



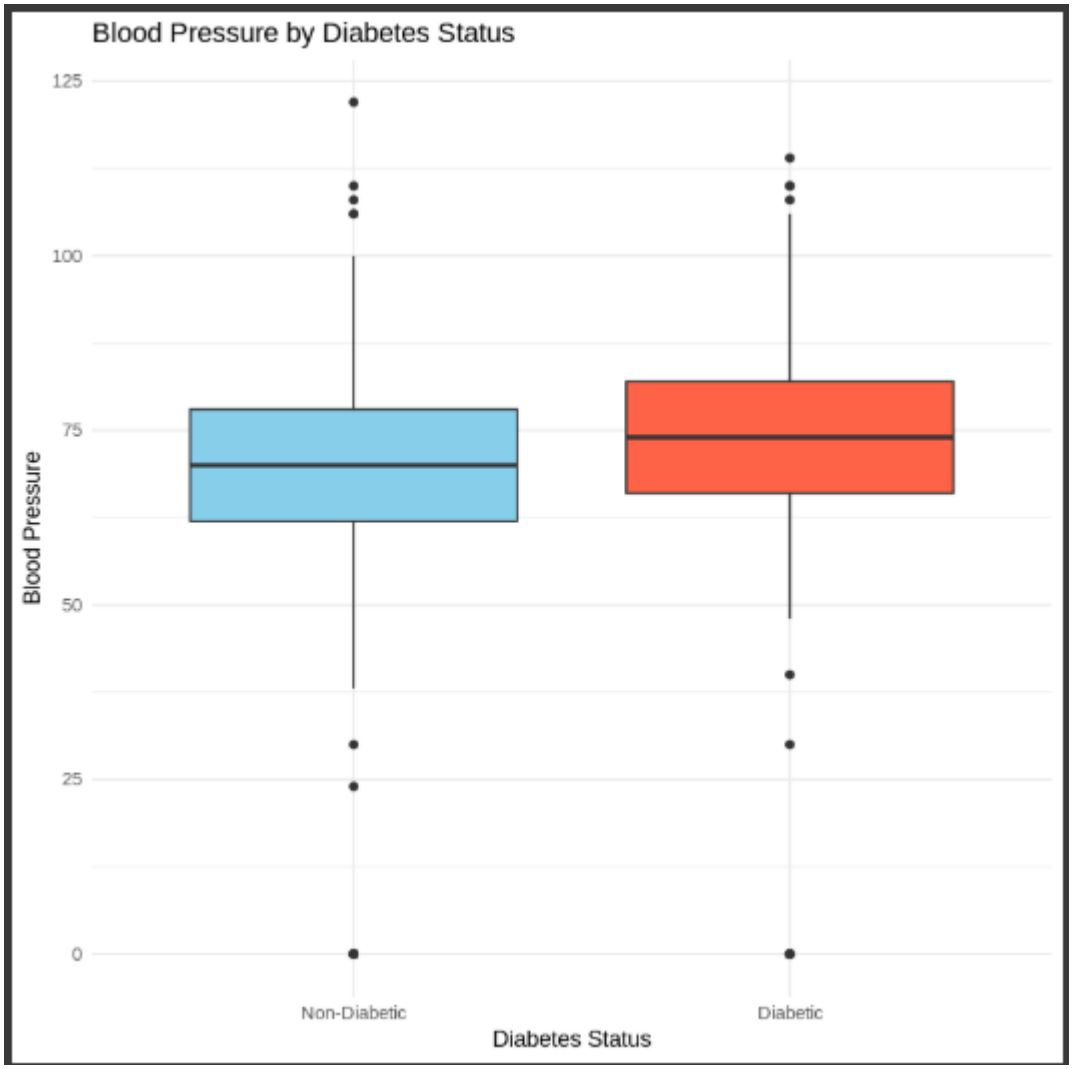
2. Average patients with and without diabetes:

The average age level among diabetic patients are 37 and 31 for non-diabetic patients



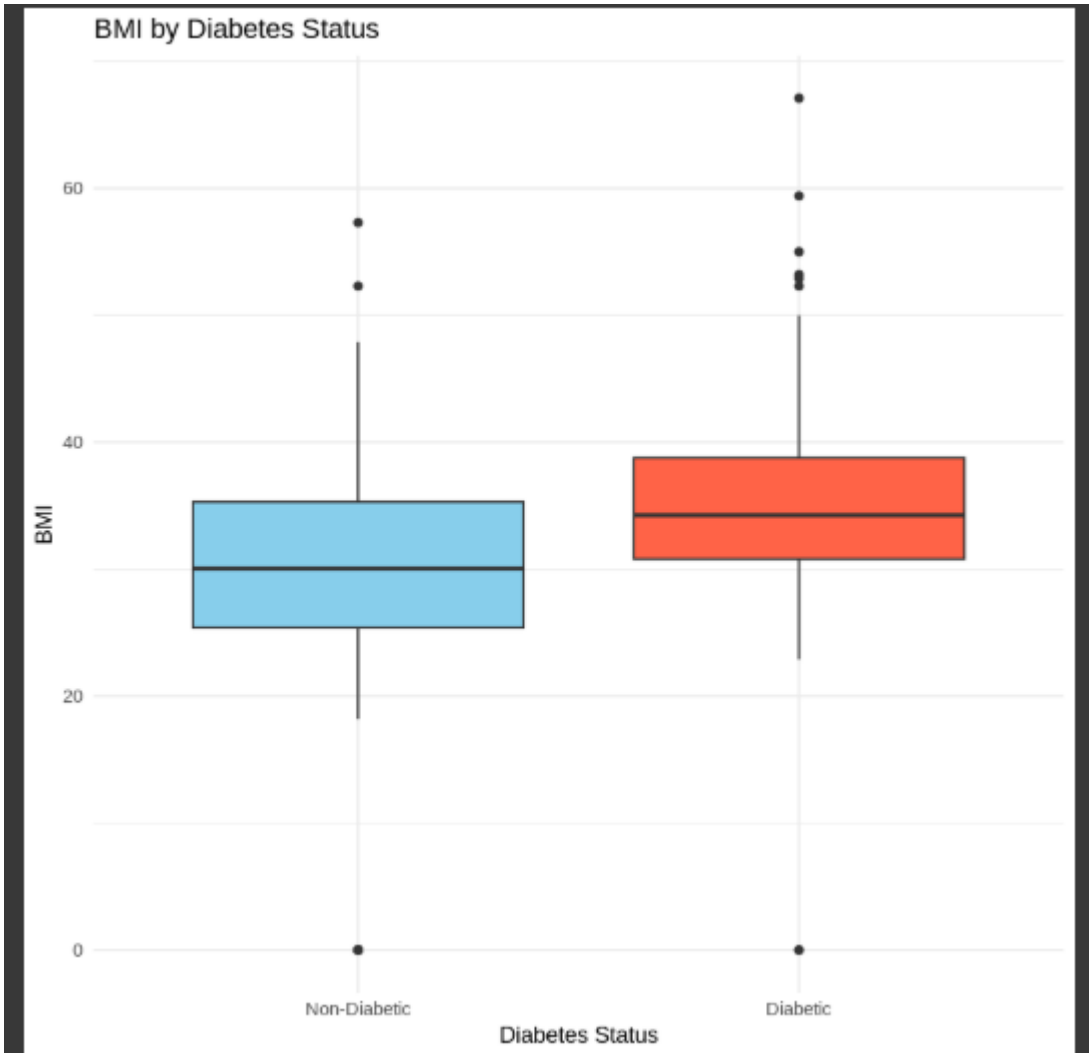
3. The average blood pressure Levels:

The average Blood Pressure levels among diabetic patients are 70.82 and 68.18 for non-diabetic patients



4. BMI Distribution:

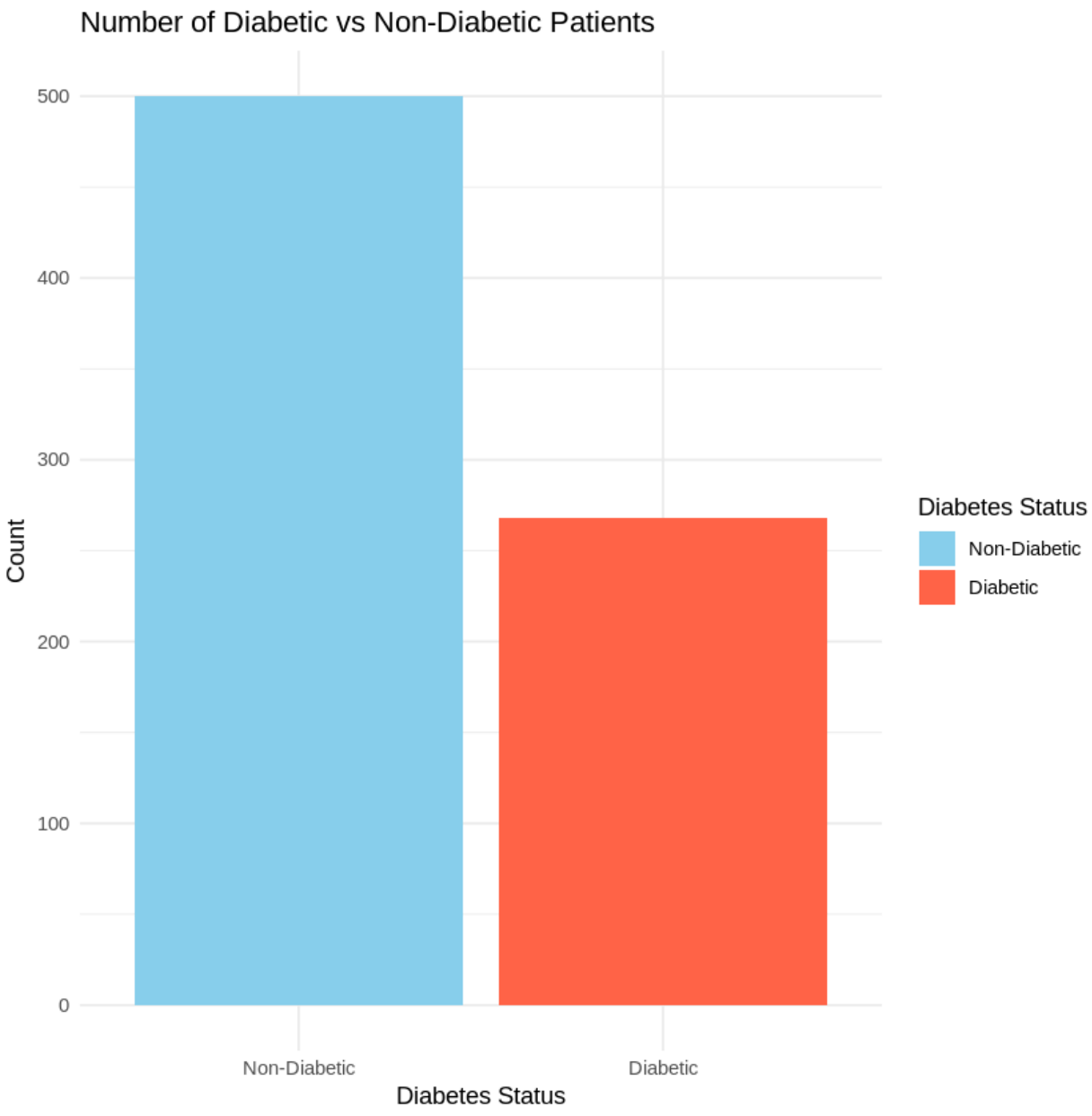
The average BMI Pressure levels among diabetic patients is 35.14 and 30.30 for non-diabetic patients.



5. The rate of diabetes among patients in the dataset:

The rate of diabetes in the dataset is 34.90%

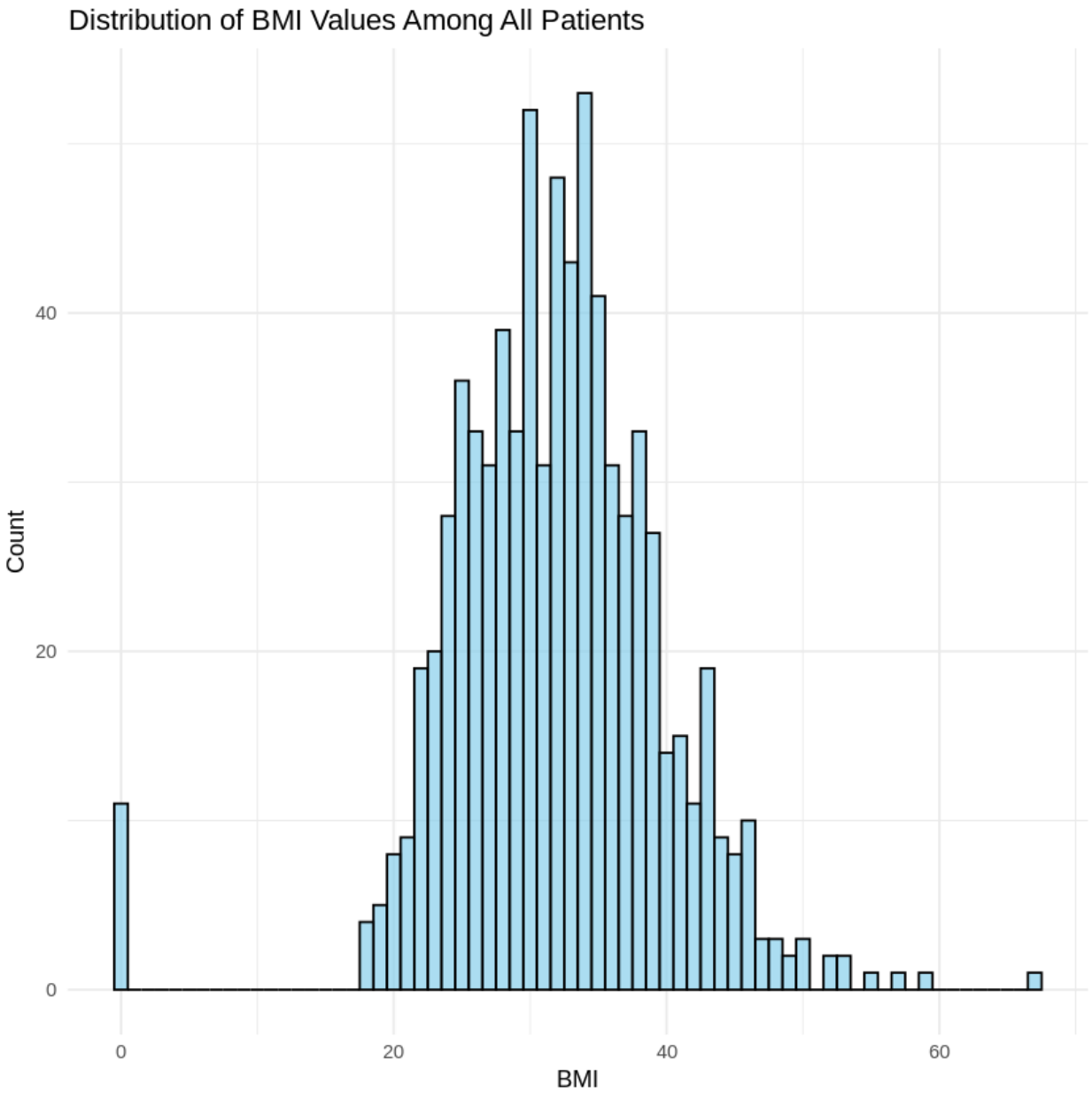
Visualization: Using bar chart between the diabetic and non-diabetic patients among the datasets.



6. The distribution of BMI values among all patients

Higher BMI values were prevalent among diabetic patients (mean: 35.15) compared to the patients with non-diabetes (mean = 30/55).

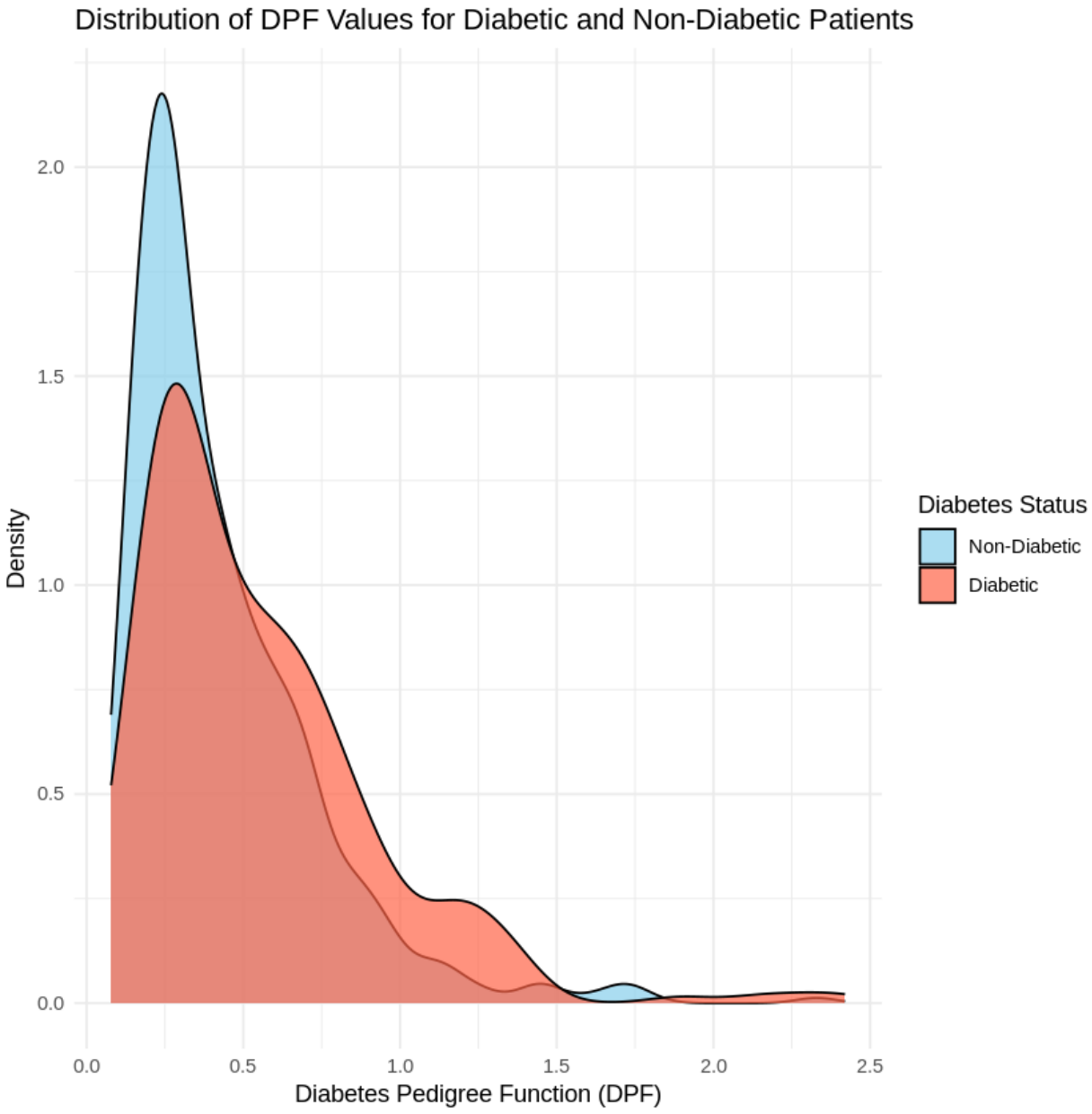
Visualization: Using histograms, the shape of the BMI distribution and the distribution within the population was profiled.



7. The distribution of Diabetes Pedigree Function (DPF) values for diabetic and non-diabetic patients:

The Diabetes Pedigree Function (DPF) values were higher on average for diabetic patients compared to non-diabetic patients. This indicates a stronger genetic predisposition among diabetic individuals.

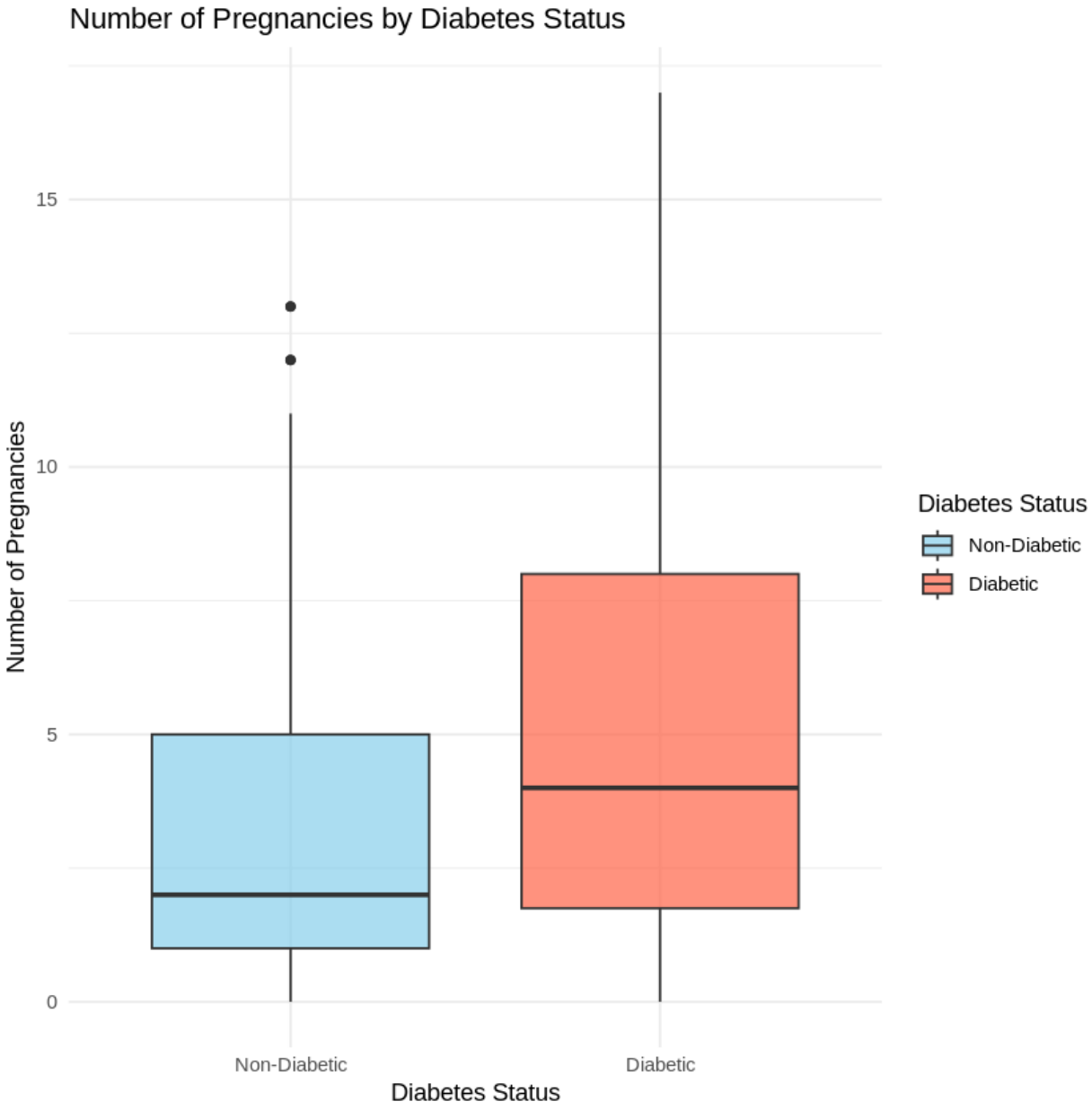
Visualization: The density plot presented above compares the DPF values of both diabetic and non-diabetic patients, and they are similar, but the overall shift of diabetic patients toward the greater DPF values can be observed.



8. The relationship between the number of pregnancies and diabetes occurrence:

Diabetic patients got more pregnancies with an average of 4.87 than the number of pregnancies got by non-diabetic patients an average of 3.30. This may indicate that the more pregnancies a woman has, the higher her likelihood of developing the disease.

Visualization: The down-shown box plot represents the pregnancy counts in cases of diabetes and without diabetes. Diabetic patients seem to have a slightly larger interquartile range and several higher and lower outliers in both groups.



9. The correlation between glucose levels and BMI:

The correlation between glucose levels and BMI is 0.22.

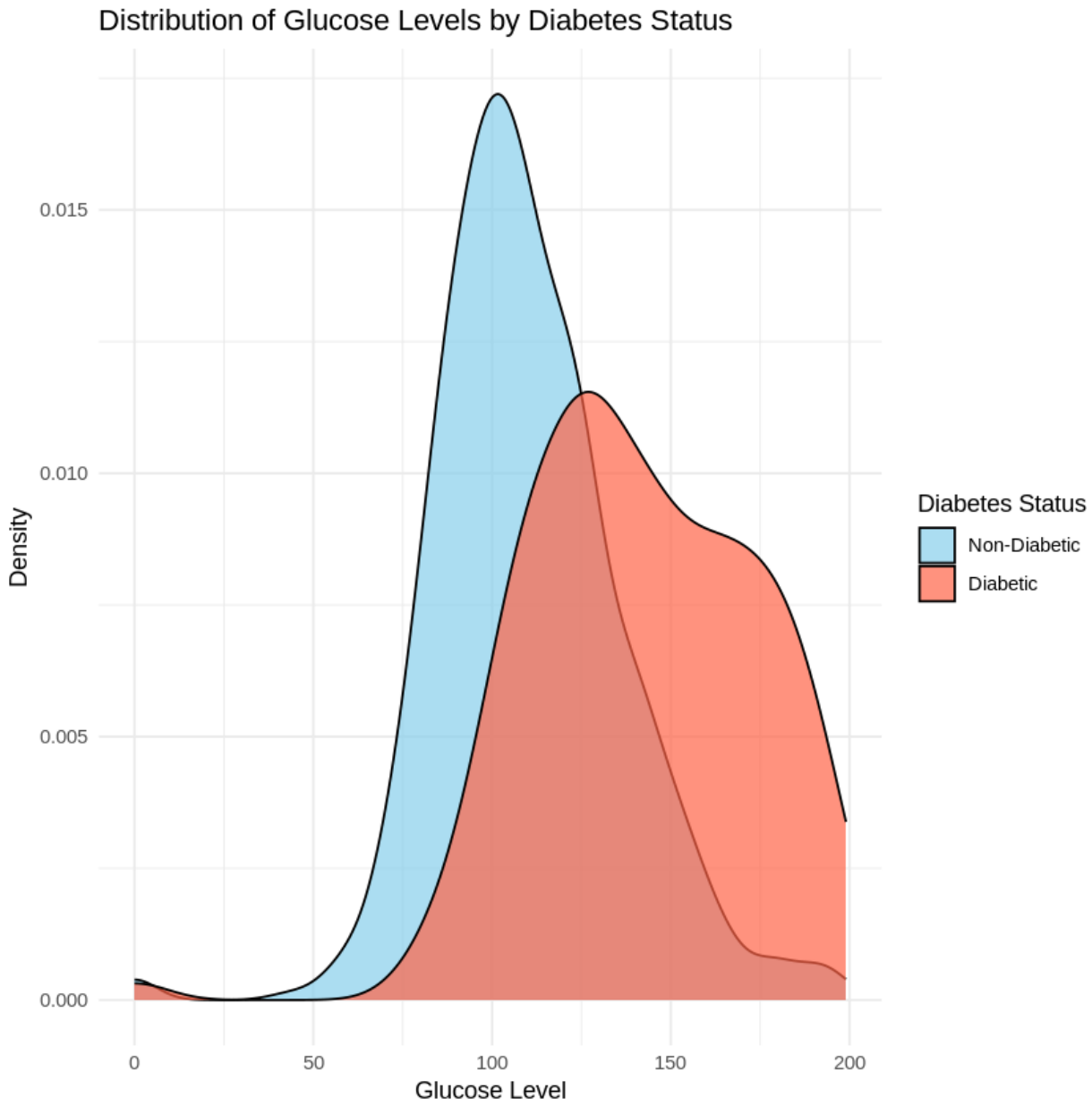
10. The trend of glucose levels with age among diabetic and non-diabetic patients

The visual for both diabetic and non-diabetic patients, and this shows that as age increases so does the glucose level. People with diabetes always have higher blood glucose concentrations than those who do not have diabetes at any age.

4.2 Addressing Research Questions

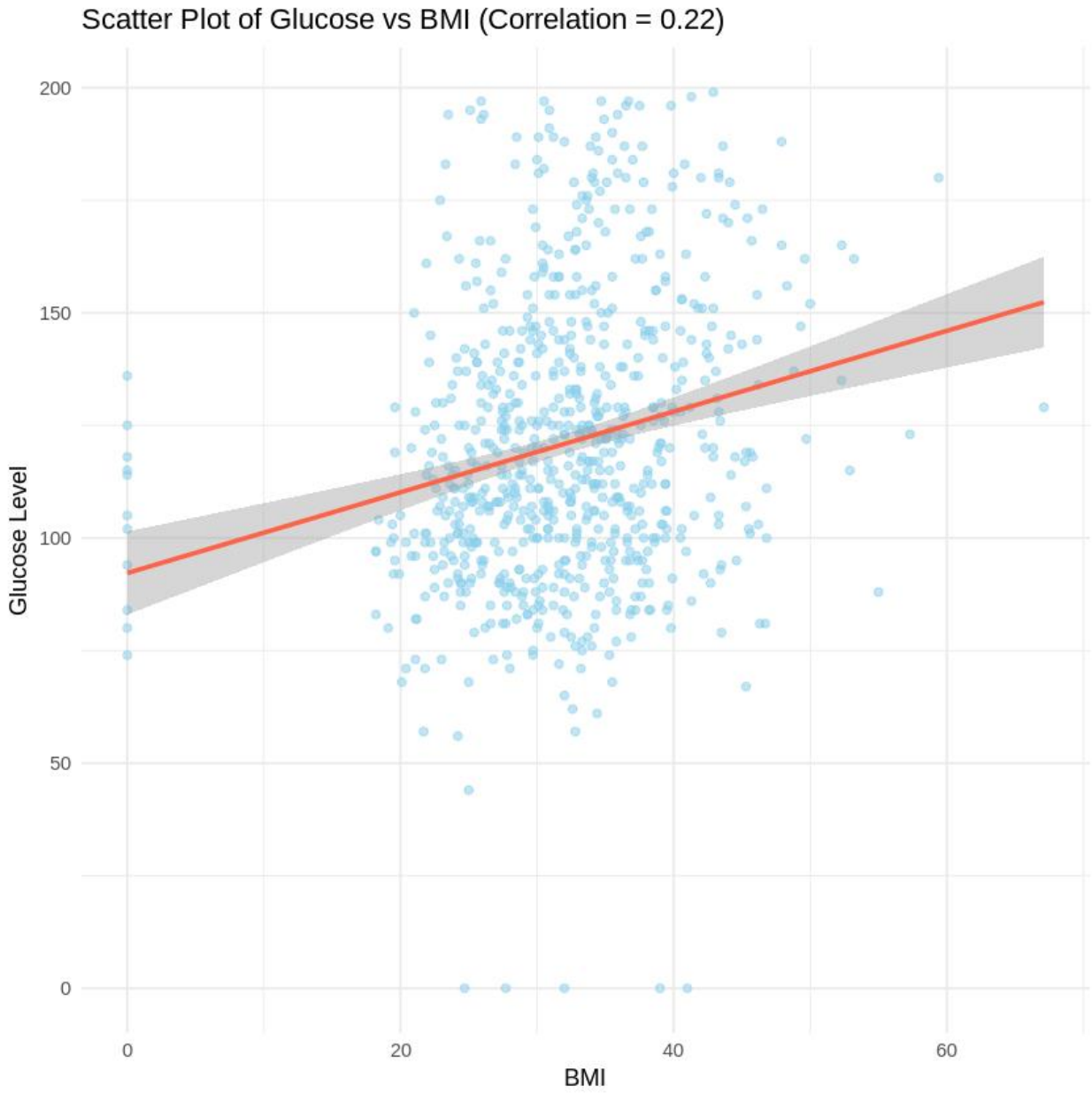
1. Are higher glucose levels associated with a greater likelihood of diabetes?

   It is obvious from the results that higher glucose level increases the probability of diabetes. Diabetic patients on average have an average glucose level of 141.26 while non-diabetic patients have an average glucose level of 109.98. The t-test has further approved this with a p-value below 0.0001 which means that it has strong statistical significance. The density plot also provides more evidence for this trend, as the diabetic patients are shifted toward higher glucose levels.



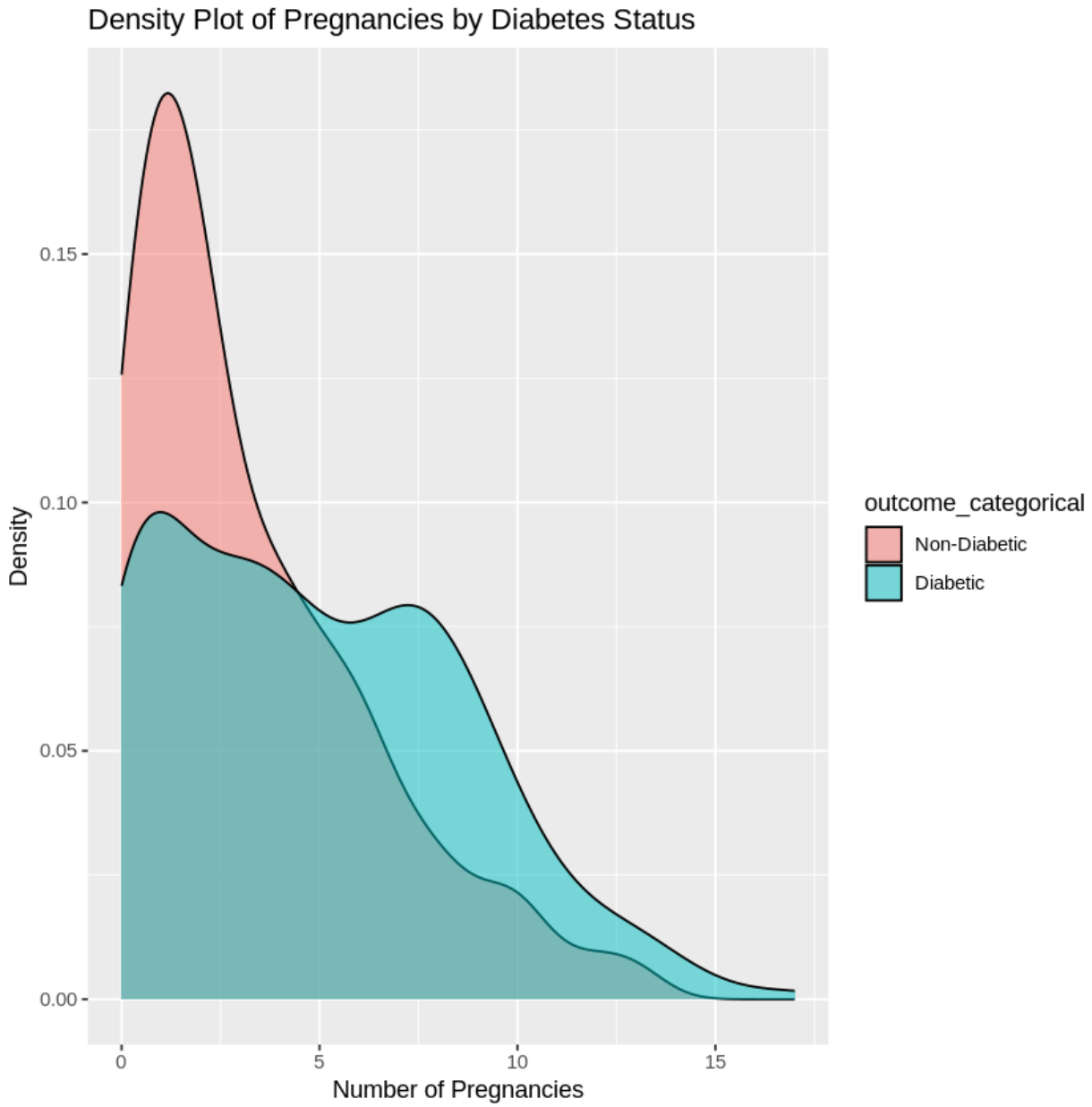
2. Are patients with high glucose concentrations also likely to have higher BMI values?

   The relationship between glucose levels and BMI is positive and very weak with a value of 0.22. This means that a BMI increased is associated with slightly elevated glucose levels, but the intensity is not very high. The scatter plot given below depicts this trend.



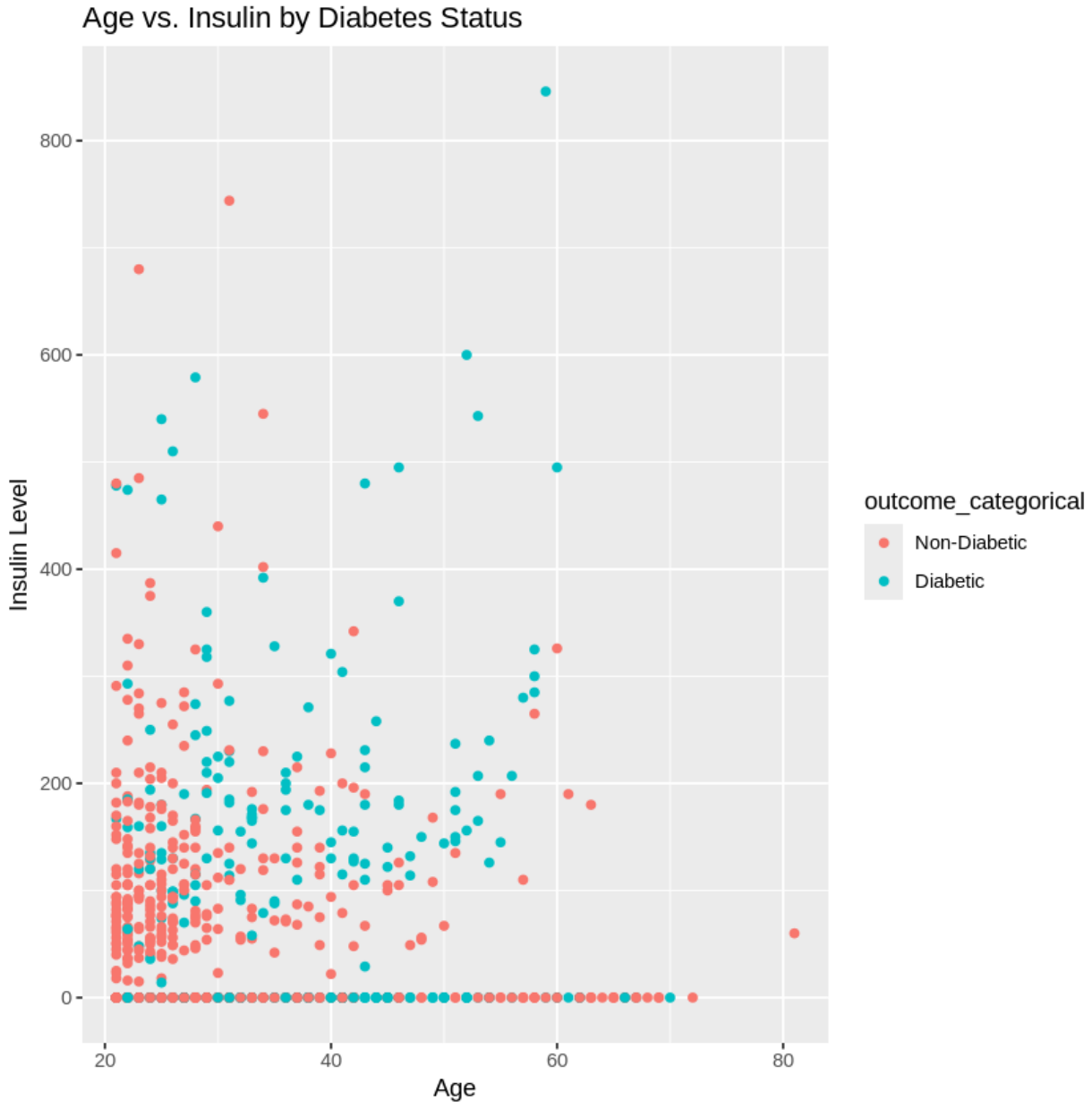
3. Are patients with a higher number of pregnancies at greater risk of developing diabetes?

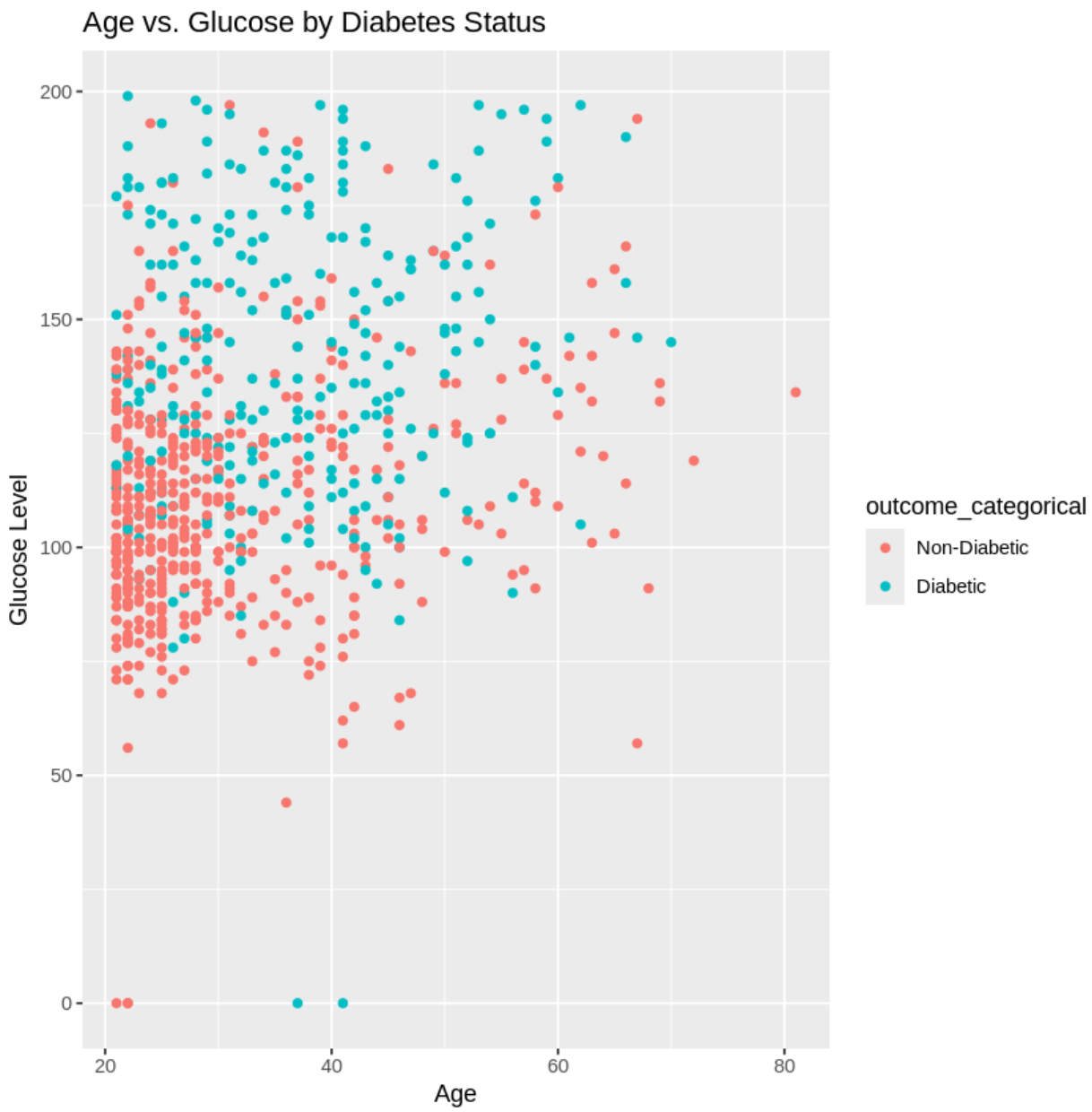
   women with diabetes tend to have a higher average number of pregnancies compared to women without diabetes. This observation suggests a potential association between pregnancy history and the development of diabetes. However, it's important to note that this analysis does not establish a causal relationship.



4. Are older patients more likely to have higher insulin concentrations and blood glucose level?

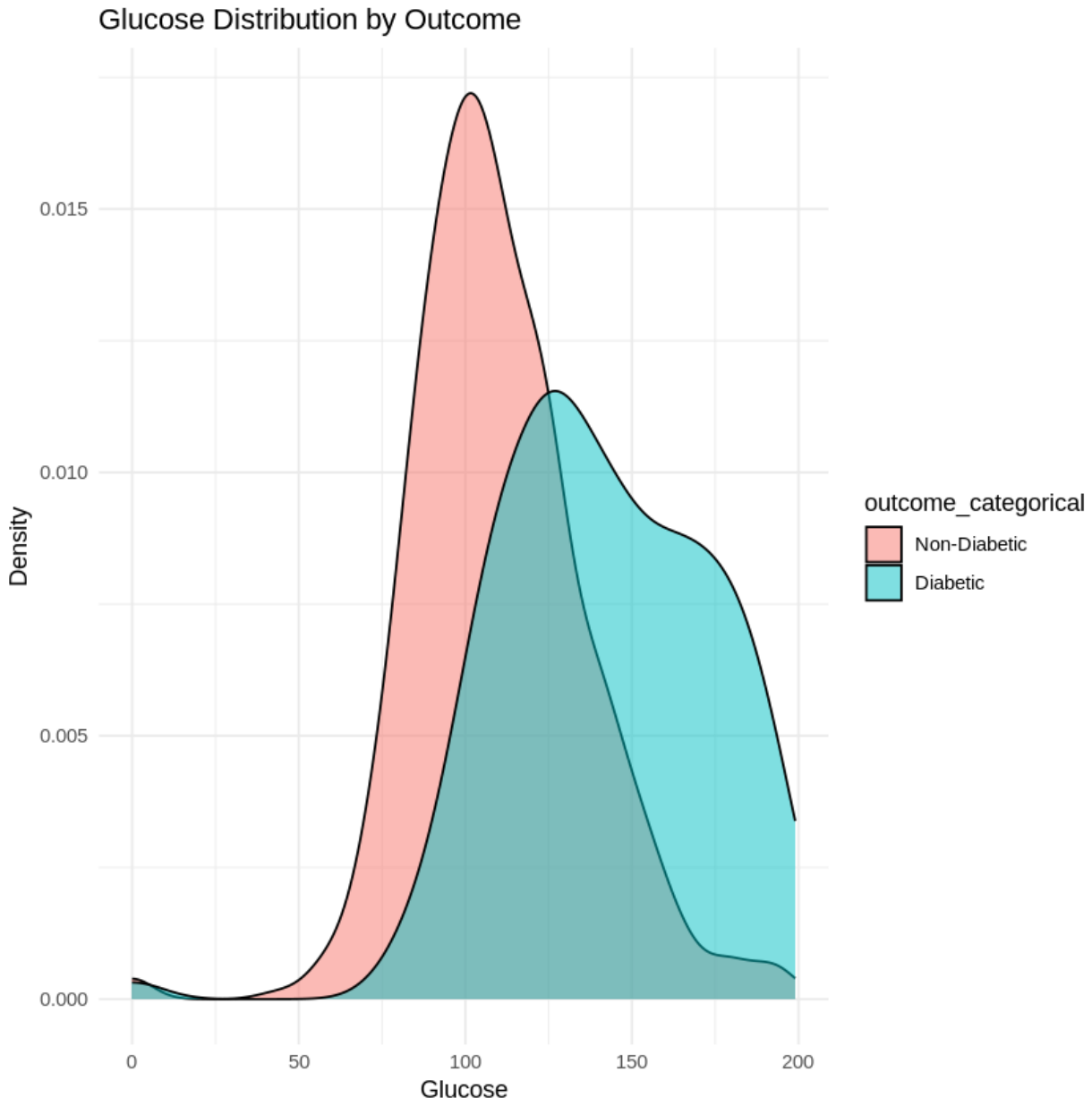
   There is not a strong association between increasing age and increasing insulin levels in this dataset. In contrast, a moderate positive correlation was observed between age and glucose levels, indicating that older individuals tend to have slightly higher glucose levels.
   Correlation between Age and Insulin: -0.04216295
   Correlation between Age and Glucose: 0.2635143

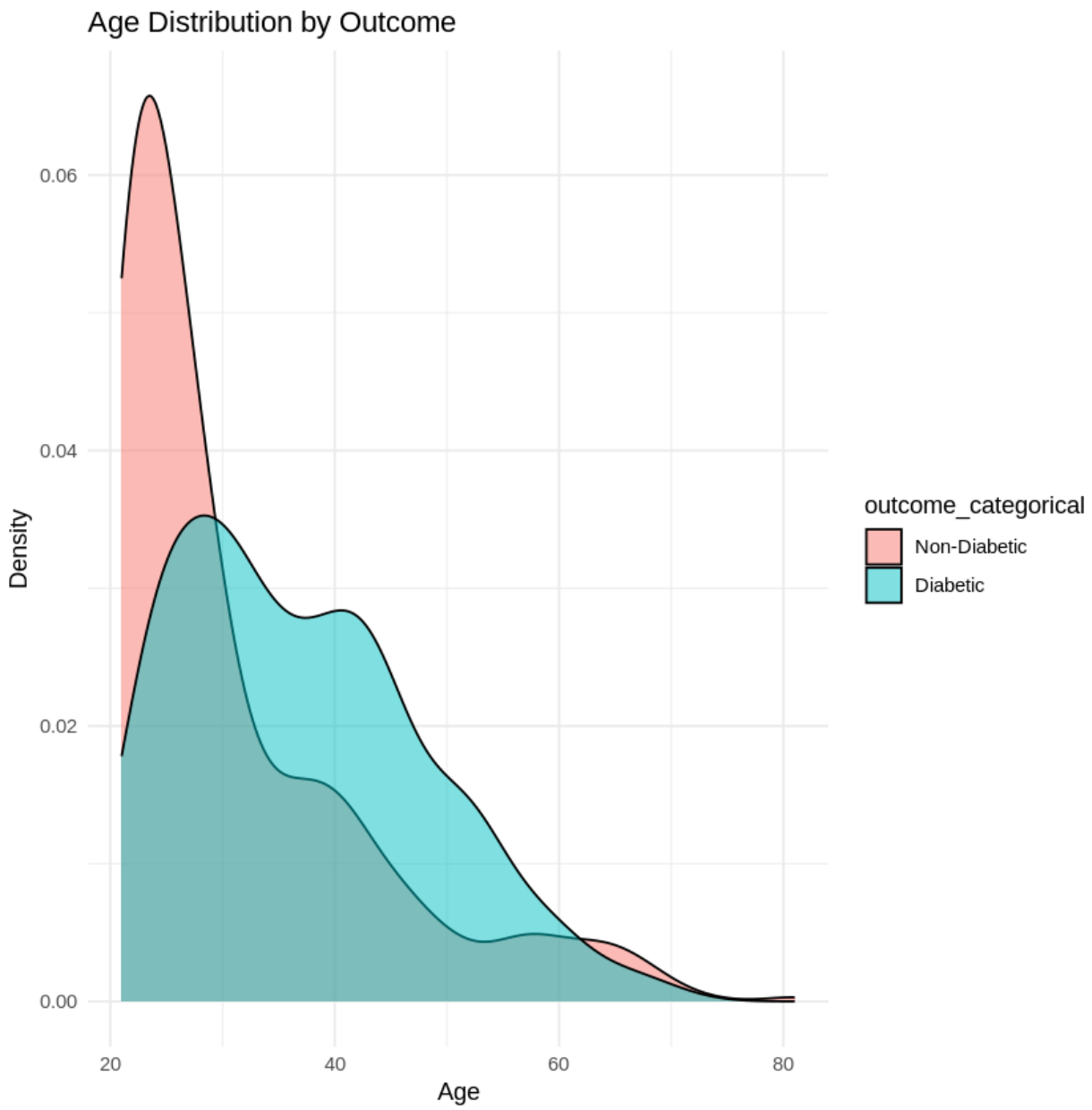



Age vs. Glucose by Diabetes Status
Glucose Level
200
150
100
50
0
20
40
60
80
Age
outcome_categorical
Non-Diabetic
Diabetic

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

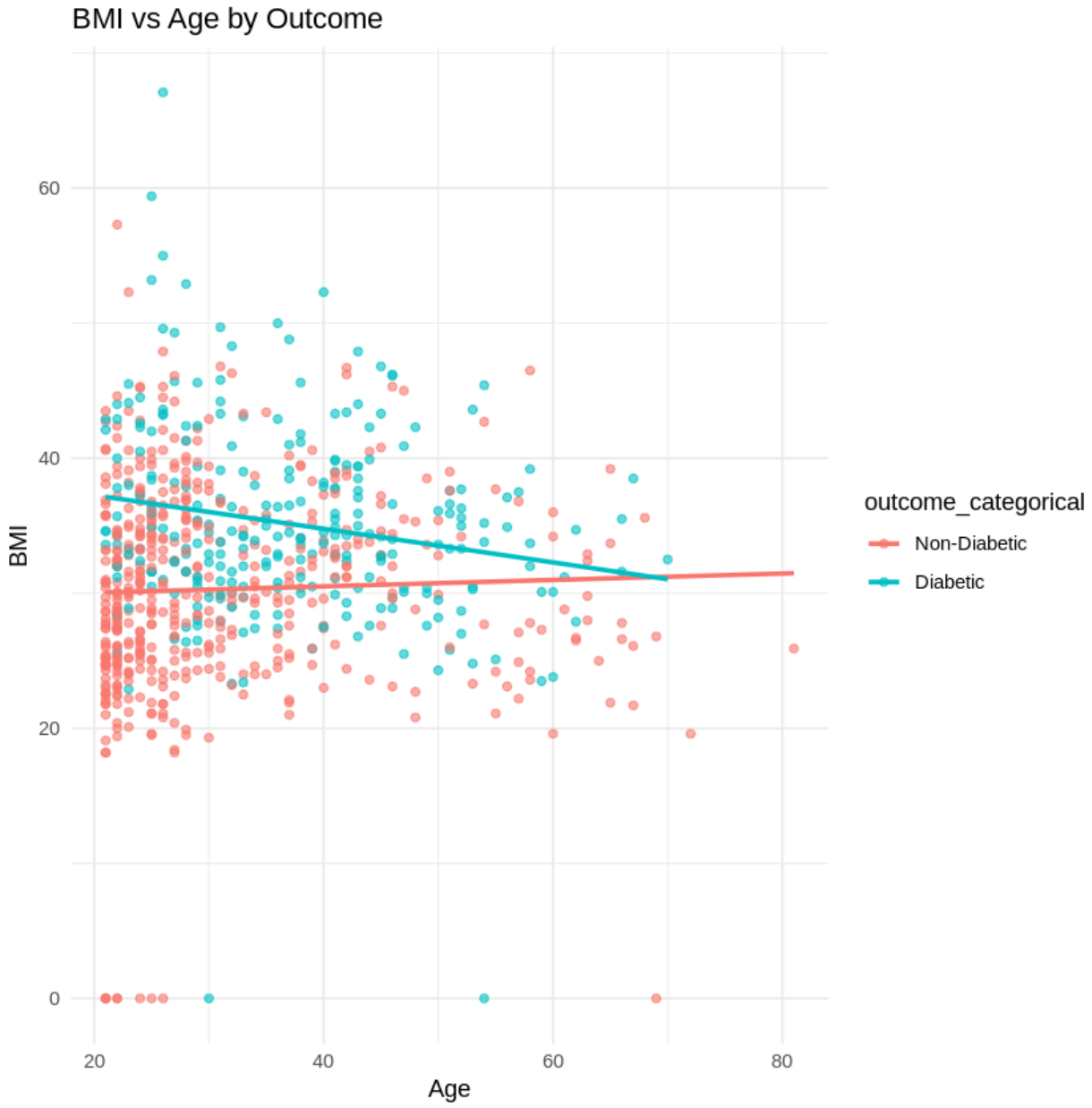
   Diabetic patients tend to have higher glucose levels, BMI, and are generally older compared to non-diabetic patients. The "risk profiles" suggest that individuals with glucose levels above 140 mg/dL, BMI above 35, and age over 40 are at a higher risk of diabetes.




Age Distribution by Outcome
0.06
0.04
0.02
0.00
Density
20
40
60
80
Age
outcome_categorical
Non-Diabetic
Diabetic

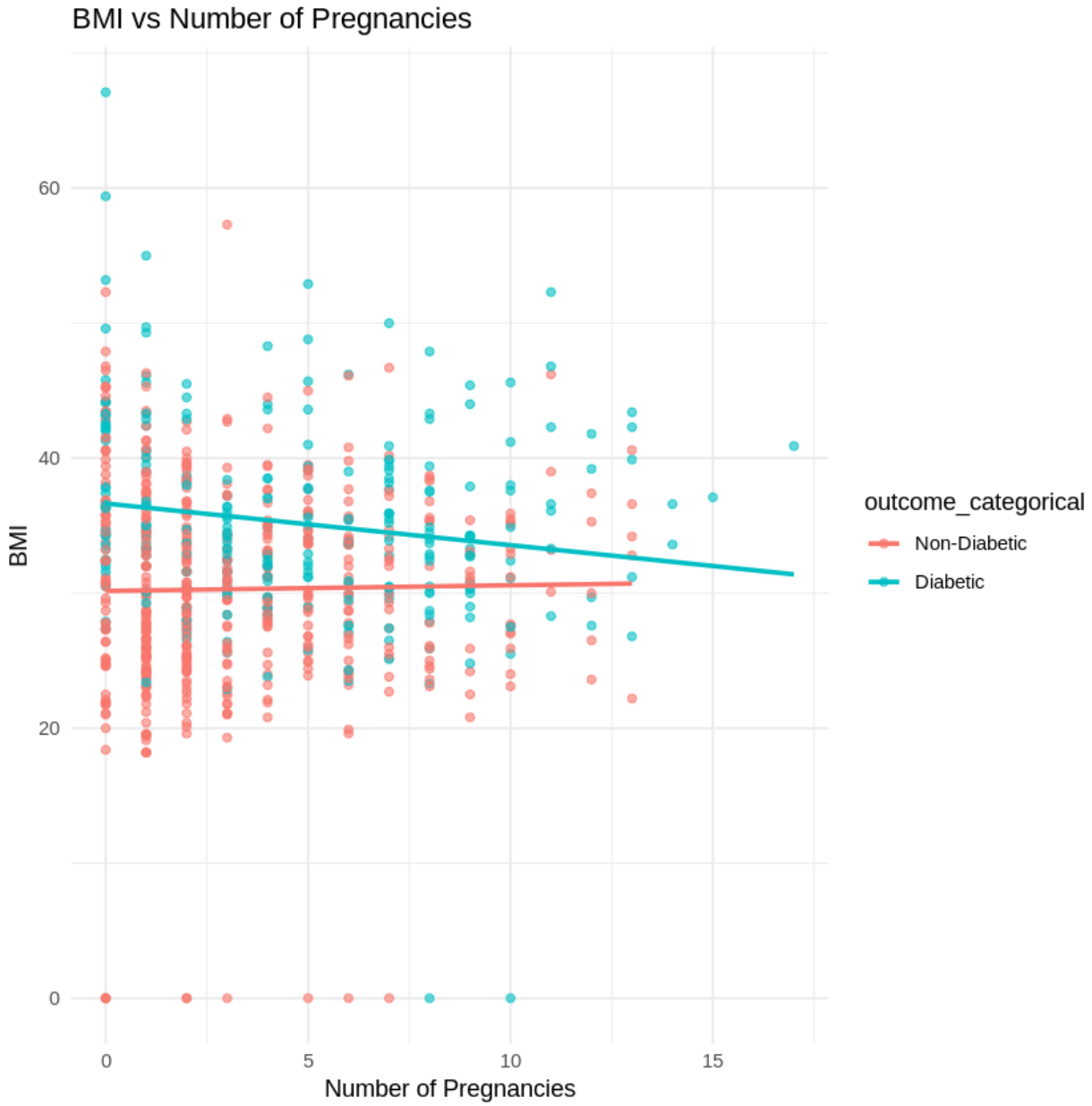
6. Does BMI correlate with age for diabetic and non-diabetic patients?

   For non-diabetic patients, there is a very weak positive correlation between BMI and age (correlation = 0.036), indicating almost no relationship. For diabetic patients, there is a weak negative correlation (correlation = -0.188), suggesting that older diabetic patients tend to have slightly lower BMI. However, in both groups, the relationships are not strong enough to be significant.



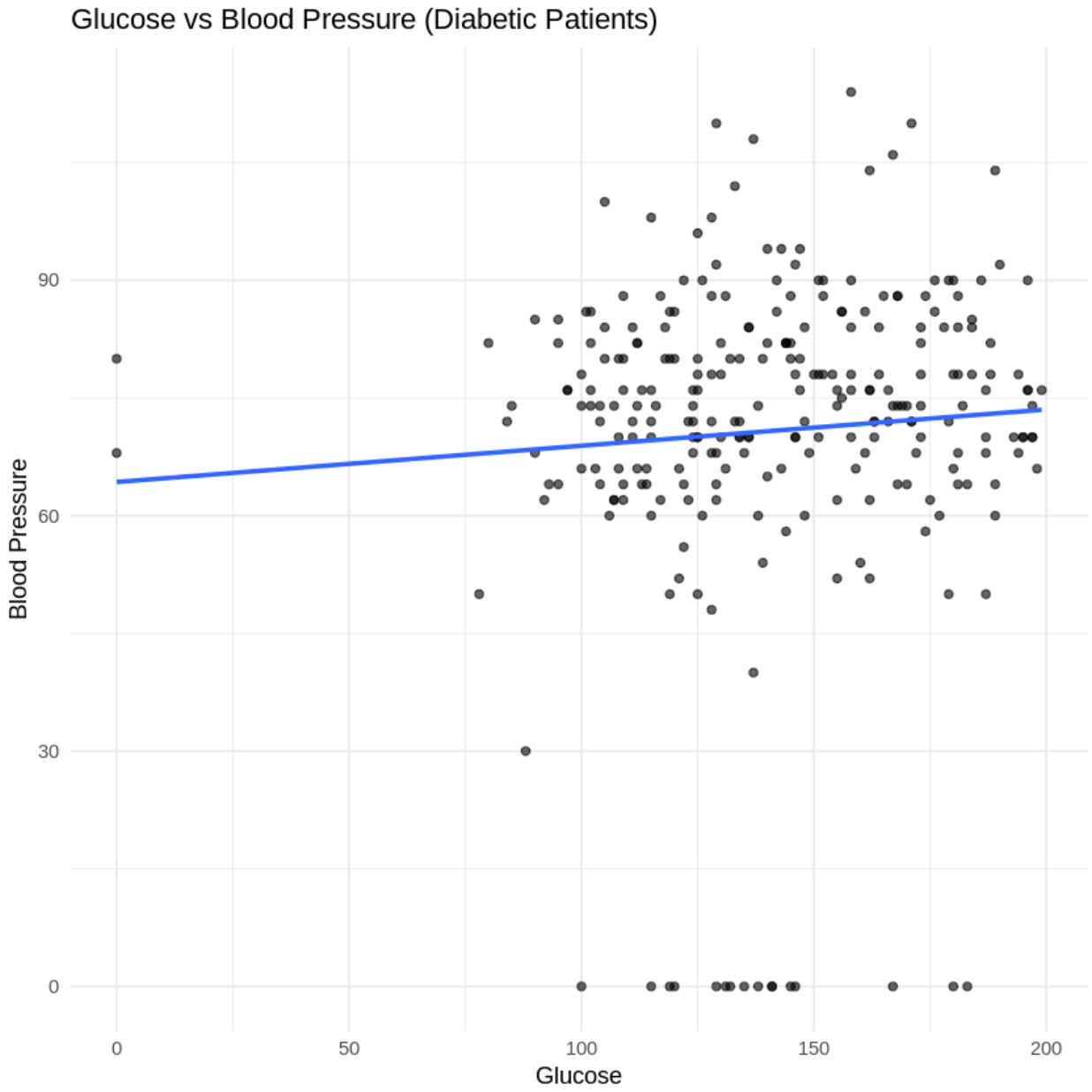
7. Is there a relationship between the number of pregnancies and BMI?

   The scatterplot shows a weak negative trend between BMI and the number of pregnancies for both diabetic and non-diabetic patients. However, the relationship is not strong enough to suggest a significant association between these two variables. Diabetic patients appear to have slightly higher BMI values on average regardless of the number of pregnancies.



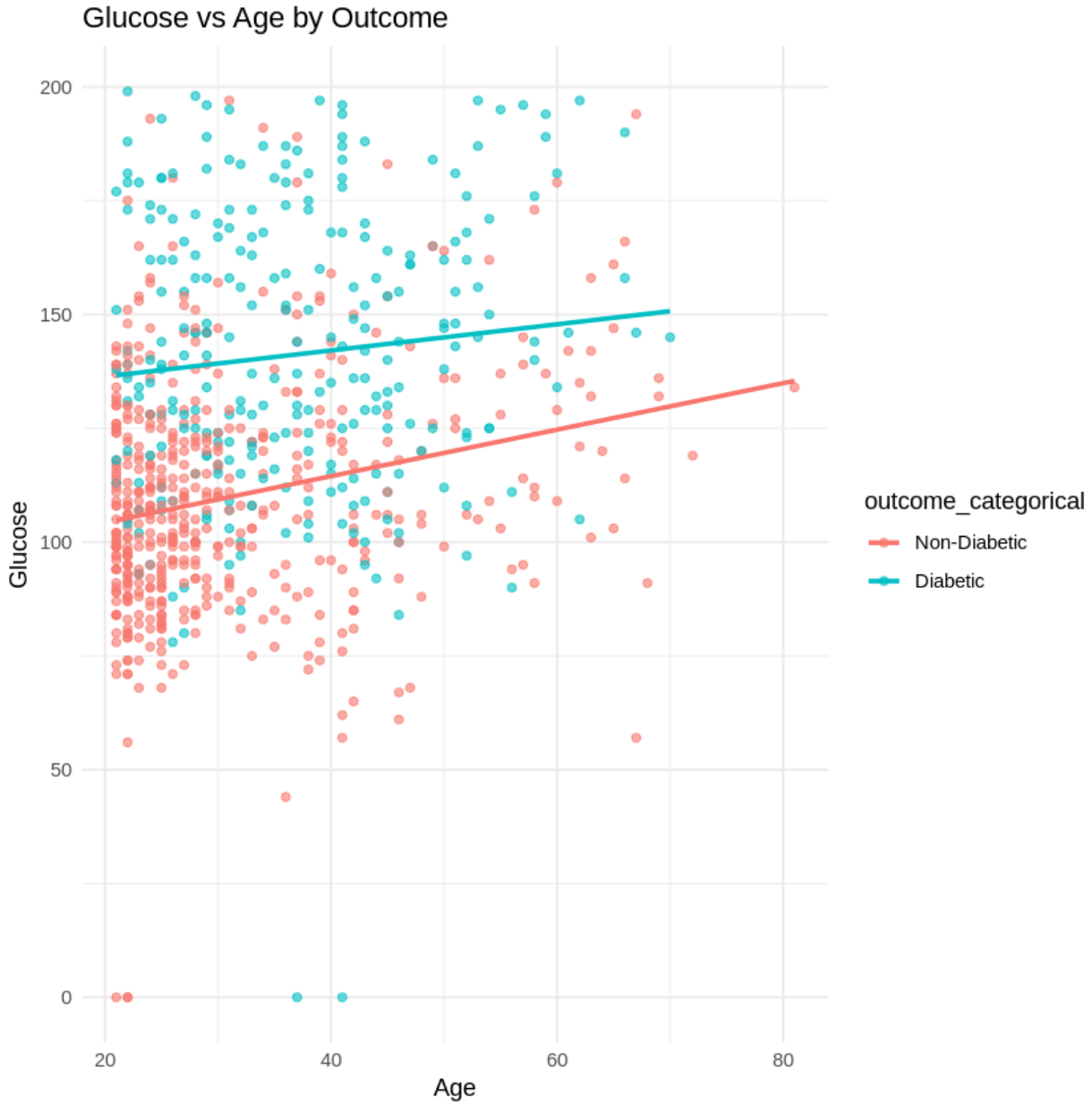
8. Are higher glucose levels associated with higher blood pressure in diabetic patients?

   The scatterplot for diabetic patients shows a weak positive relationship between glucose levels and blood pressure. This suggests that higher glucose levels might be slightly associated with higher blood pressure, but the correlation appears to be weak.



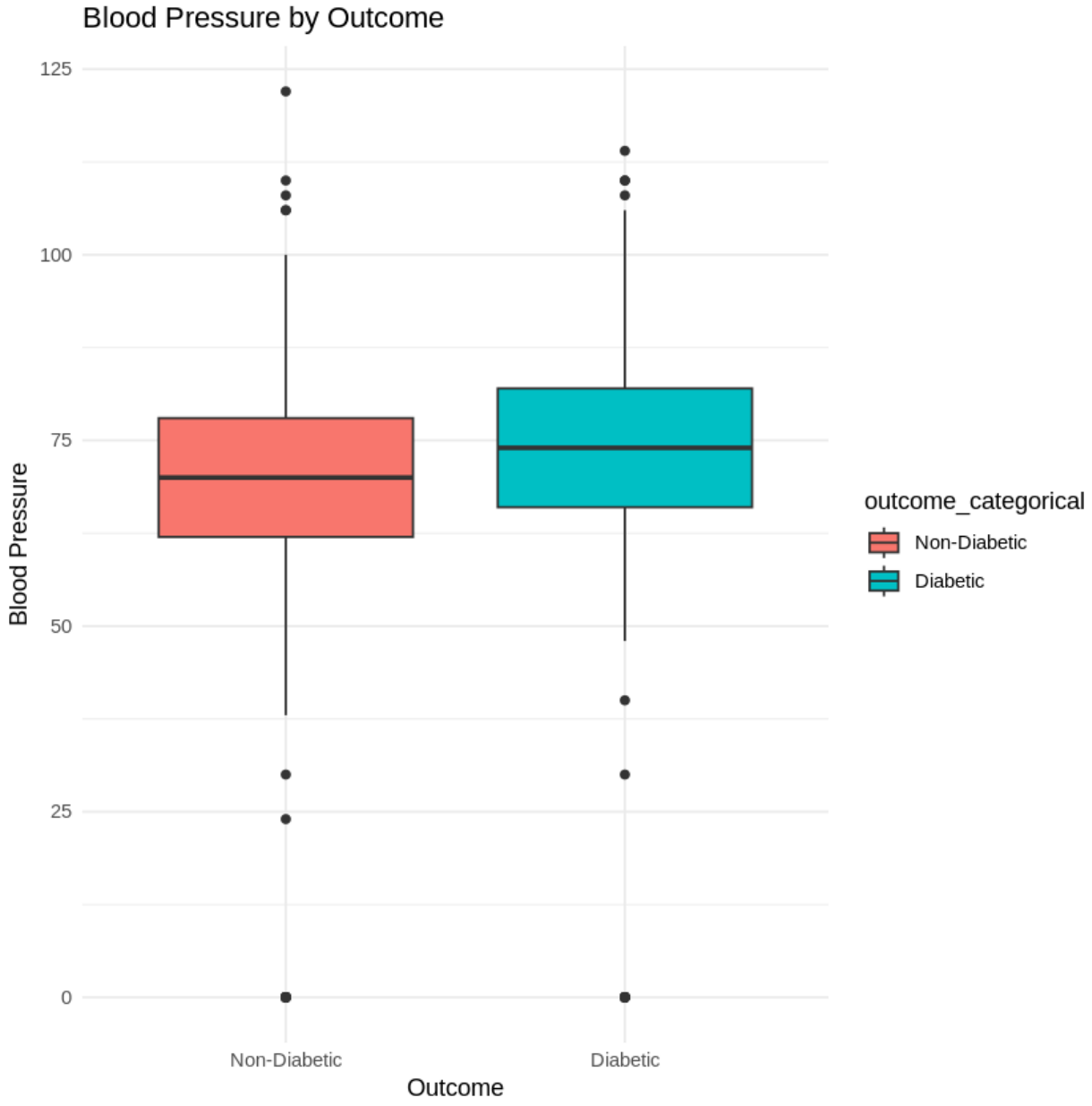
9. Is there a relationship between glucose and age for diabetic and non-diabetic patients?

   The scatterplot shows a weak positive relationship between glucose levels and age for both diabetic and non-diabetic patients. Diabetic patients generally have higher glucose levels than non-diabetic patients across all age groups.



10. Do diabetic patients tend to have higher blood pressure compared to non-diabetic patients?

    The boxplot shows that diabetic patients tend to have slightly higher blood pressure compared to non-diabetic patients. The median blood pressure is higher for diabetics.



4.3 Hypothesis Testing

1. **Validating the Claim: "There is a significant difference in glucose levels between diabetic and non-diabetic patients."**
   Hypothesis:
   Null Hypothesis: There is no significant difference in glucose levels between diabetic and non-diabetic patients.
   Alternative Hypothesis: There is a significant difference in glucose levels between diabetic and non-diabetic patients.

   Test Used: Two-Sample T-test

   Results:

   t = 13.752, degrees of freedom = 461.33

   p-value < 2.2e-16

   95% Confidence Interval: [26.81, 35.75]

   Mean Glucose (Diabetic): 141.26

   Mean Glucose (Non-Diabetic): 109.98

   Conclusion: The p-value is significantly less than 0.05, leading to the rejection of the null hypothesis. This confirms that there is a significant difference in glucose levels between diabetic and non-diabetic patients.

2. **New Claim: "Diabetic patients have a significantly higher BMI compared to non-diabetic patients."**
   Hypothesis:
   Null Hypothesis: Diabetic patients do not have a significantly higher BMI compared to non-diabetic patients.
   Alternative Hypothesis: Diabetic patients have a significantly higher BMI compared to non-diabetic patients.

   Test Used: Two-Sample t-test (one-tailed)

   Results:

   t = 8.6193, degrees of freedom = 573.47

   p-value < 2.2e-16

   95% Confidence Interval: [3.91, Inf]

   Mean BMI (Diabetic): 35.14

   Mean BMI (Non-Diabetic): 30.30

Conclusion: The p-value is significantly less than 0.05, leading to the rejection of the null hypothesis. This confirms that diabetic patients have a significantly higher BMI compared to non-diabetic patients.

4.4 Simulation Task

1. **Confidence Intervals for a Sample Size of 15**
   Method:
   25 random samples, each with a size of 15, were drawn from the dataset's glucose column.
   For each sample, the 95% confidence interval for the mean was calculated.
   The proportion of intervals containing the true population mean was computed.
   Results:
   The proportion of 95% confidence intervals containing the true population mean: **0.92**

2. **Confidence Intervals for a Sample Size of 100**
   Method:
   25 random samples, each with a size of 100, were drawn from the dataset's glucose column.
   For each sample, the 95% confidence interval for the mean was calculated.
   Results:
   Proportion of 95% confidence intervals containing the true population mean: **1**

3. **Confidence Intervals for Sample Size of 10**
   Method:
   20 random samples, each with a size of 10, were drawn from the dataset's glucose column.
   For each sample, the 95% confidence interval for the mean was calculated.
   Results:
   Proportion of 95% confidence intervals containing the true population mean: **1**

5. Conclusion

The exploratory data analysis of the data set has revealed key determinants of diabetes from this analysis of the diabetes set. Some of the most important conclusions are differences within glucose level, BMI, and other diagnostic indicators between the groups of patients with and without diabetes. It is understood that hypothesis testing confirms that higher glucose levels, enlarged BMI, and higher age represent the potential risks for diabetes.

Moreover, it was possible to see the difference in confidence interval simulations of given sample size, which resulted in making the necessity of sufficient data size in clinical and public health research stronger. However, the study has certain drawbacks: first, it focuses on a particular group of the female population of 21 years and older, and, second, the dataset does not allow seeing the dynamics of the situation.

Further research should encompass more diverse samples, longitudinal data and terms of reference as well as entrenched and more sophisticated risk models to mitigate diabetes more effectively. Although some limitations exist, this study contributes a baseline of essential information and practical recommendations for further research and subsequent public health interventions.