Analyzing Health Patterns Through Data

Presented by

Bosy Ayman 202202076
Jana Ahmed 202201853
Maysam Asser 202200276
Rana Saad 202201853
Riwan Ashraf 202201726

Statistical Inference

27/12/2024

1- Introduction

This dataset contains comprehensive medical diagnostic data aimed at predicting the likelihood of diabetes onset based on various health indicators which are **Pregnancies, Glucose,BloodPressure,SkinThickness,Insulin,BMI, DiabetesPedigreeFunction, Age and outcome** . It includes 768 records of female patients, each described by eight health-related features which are . The Outcome variable specifies whether a patient is diabetic (1) or non-diabetic (0).

The dataset source: Link

2- Data Processing and Analysis Steps

1. Replacing Zeros with Mean Values

zeros in specific columns of the dataset are replaced with the mean value of each respective column. This is done to handle missing things that could affect the analysis.

Steps

- A list of columns (Glucose, BloodPressure, SkinThickness, Insulin, and BMI) is defined.
- For each column:

    - The mean value of the column is calculated, ignoring any missing values (NA).

    -Any occurrence of zero within the column is then replaced by the calculated mean value.

2. Handling Outliers Using the Interquartile Range (IQR) Method

Outliers can significantly distort statistical analysis, so it is important to detect and manage them. The Interquartile Range (IQR) method is applied to identify and cap outliers in the dataset.

Steps

- For each specified column (Glucose, BloodPressure, SkinThickness, Insulin, and BMI):

  -The first quartile (Q1) and third quartile (Q3) are calculated.

  -The IQR is determined by subtracting Q1 from Q3 (IQR = Q3 - Q1).

- Outlier boundaries are then established:

  -The lower bound is defined as Q1 - 1.5 * IQR.

  The upper bound is defined as Q3 + 1.5 * IQR.

- Any values below the lower bound are replaced with the lower bound, and any values above the upper bound are replaced with the upper bound.

4-Challenges, Limitations, and Assumptions

Challenges and Solutions

1-Challenge:

Missing values in columns such as Glucose, BMI, or Blood Pressure.

Solution:

Use means to fill in the missing values(zeros).

2-Challenge:

Outliers, including extremely low values in BMI and Glucose.

3-Solution:

Identify the outliers using the Interquartile Range.

4-Challenge:

Choosing appropriate statistical methods for analysis.

Solution:

Studying, analyzing the problem well to determine the most appropriate method to apply.

5-Challenge:

Deciding on the most effective types of graphs to use.

Solution:

Used the Storytelling book to help us.

Limitations and assumptions

1. Data Quality and Completeness

- **Limitations:** The dataset may have missing values (e.g., zeros in critical columns such as Glucose, Insulin, BMI), which could lead to incomplete or skewed analysis. Although missing values are replaced with the mean of respective columns, this method may not fully address the underlying issue of missingness, potentially leading to biased outcomes.
- **Assumption:** We assume that missing values are missing at random and that replacing zeros with mean values does not significantly alter the underlying distributions.

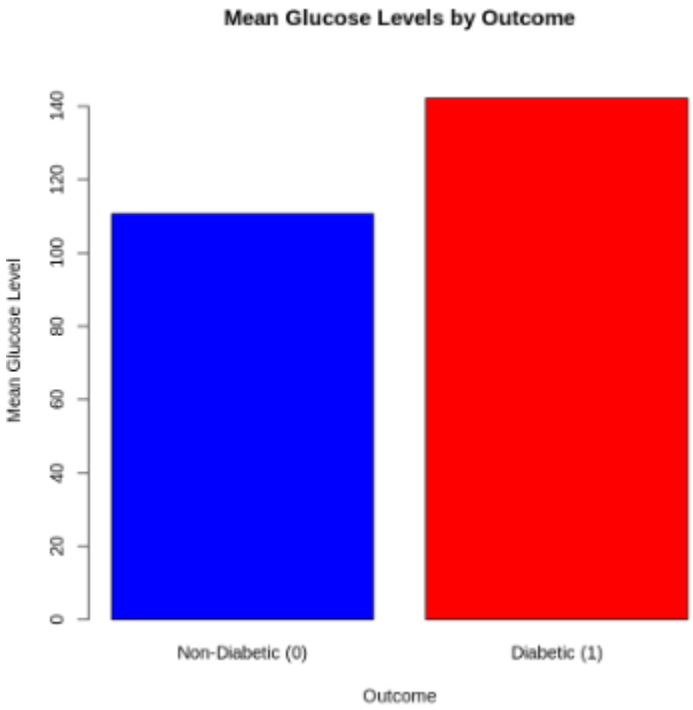
2. Outliers

- **Limitations:** The Interquartile Range (IQR) method is used to handle outliers. However, the IQR approach may not always be effective for extreme outliers or for datasets with skewed distributions.
- **Assumption:** We assume that the outliers identified using the IQR method are indeed anomalies, and that capping them at the calculated bounds helps reduce their influence on the overall analysis.

4-Results and Visualizations

1- Exploratory Analysis

1-The average glucose levels among patients with and without diabetes

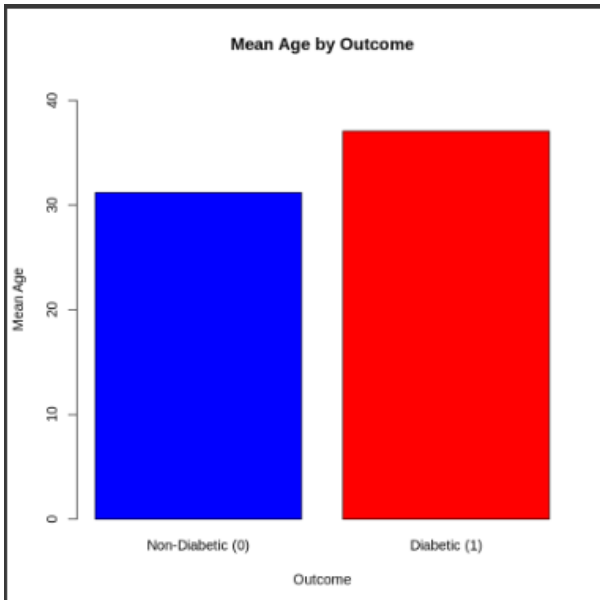


The Bar Chart is used to compare between the glucose levels of the diabetic and non diabetic.

Conclusion:

The bar chart demonstrates that patients with diabetes have significantly higher average glucose levels compared to those without diabetes. The average glucose level for patients without diabetes is approximately 110 mg/dL, while for patients with diabetes, it is significantly higher at approximately 142 mg/dL.

2- The average age of patients with and without diabetes.



The Bar Chart is used to compare between the average age of diabetic and non diabetic

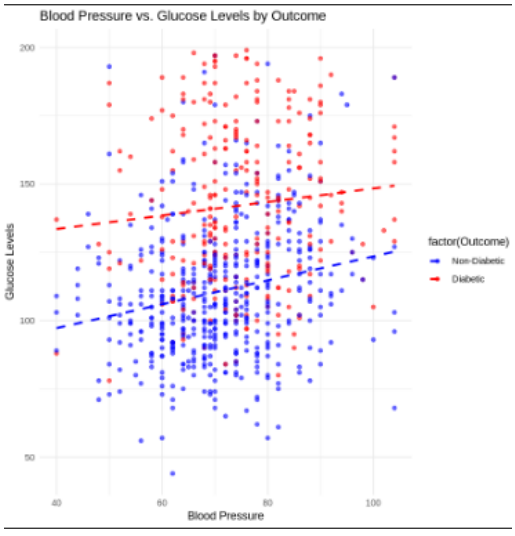
The boxplot is use to indicate the the median and the mean of the again distribution related to the outcome (diabetic or not)

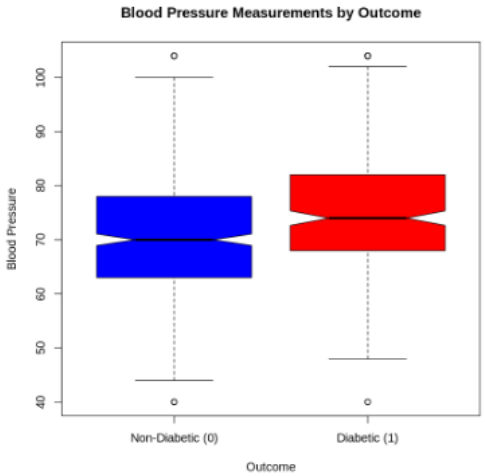
Non-Diabetic : The blue bar indicates the mean age of patients without diabetes. It appears to be around 31 years.

Diabetic : The red bar represents the mean age of patients with diabetes. It seems to be approximately 37 years.

Conclusion: the average age of patients with diabetes is higher that those who are without diabetes.

3- The average blood pressure measurements across diabetic and non-diabetic groups.





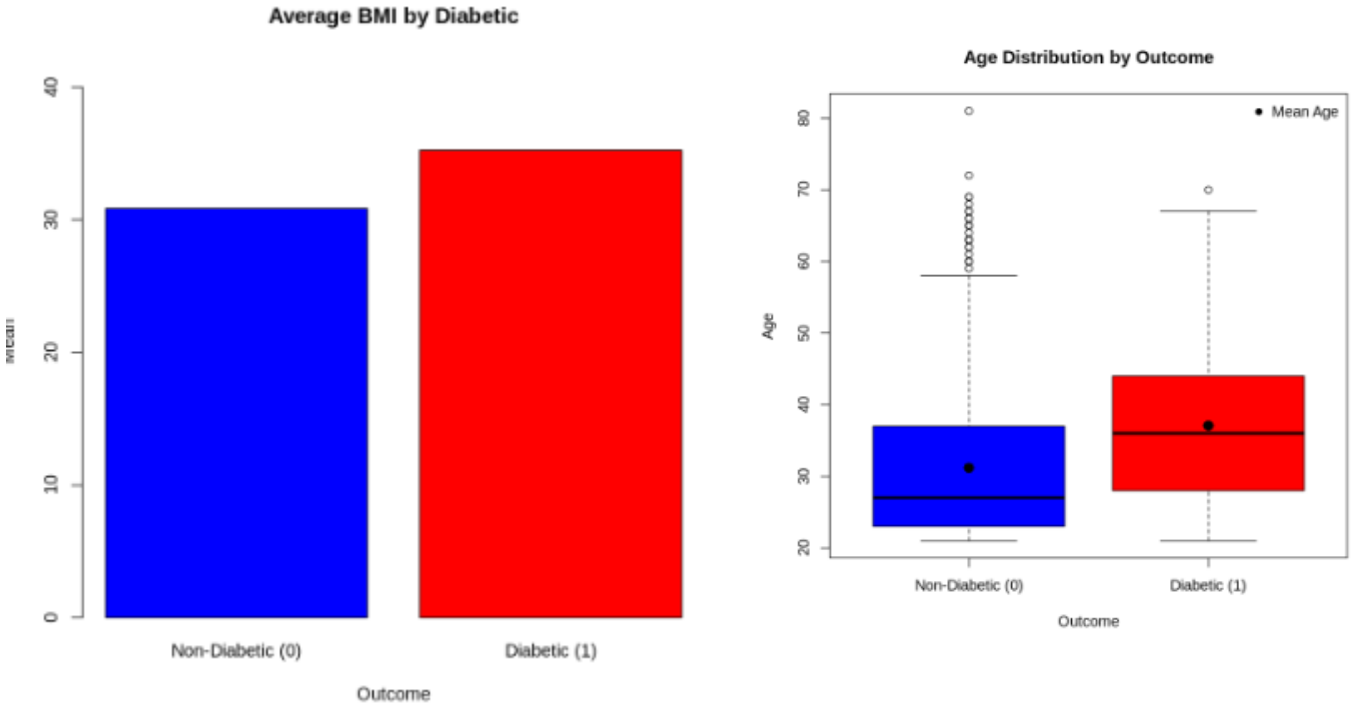
The scatter shows the dispersion of diabetic and non diabetic.

Non-Diabetic (0): The blue bar indicates the mean blood pressure of patients without diabetes. It appears to be around 70.8 mmHg.

Diabetic (1): The red bar represents the mean blood pressure of patients with diabetes. It seems to be approximately 74 mmHg.

Conclusion: The visualization indicates that the average blood pressure is significantly higher for patients with diabetes compared to those without diabetes.

4- The average BMI of diabetic versus non-diabetic patients.



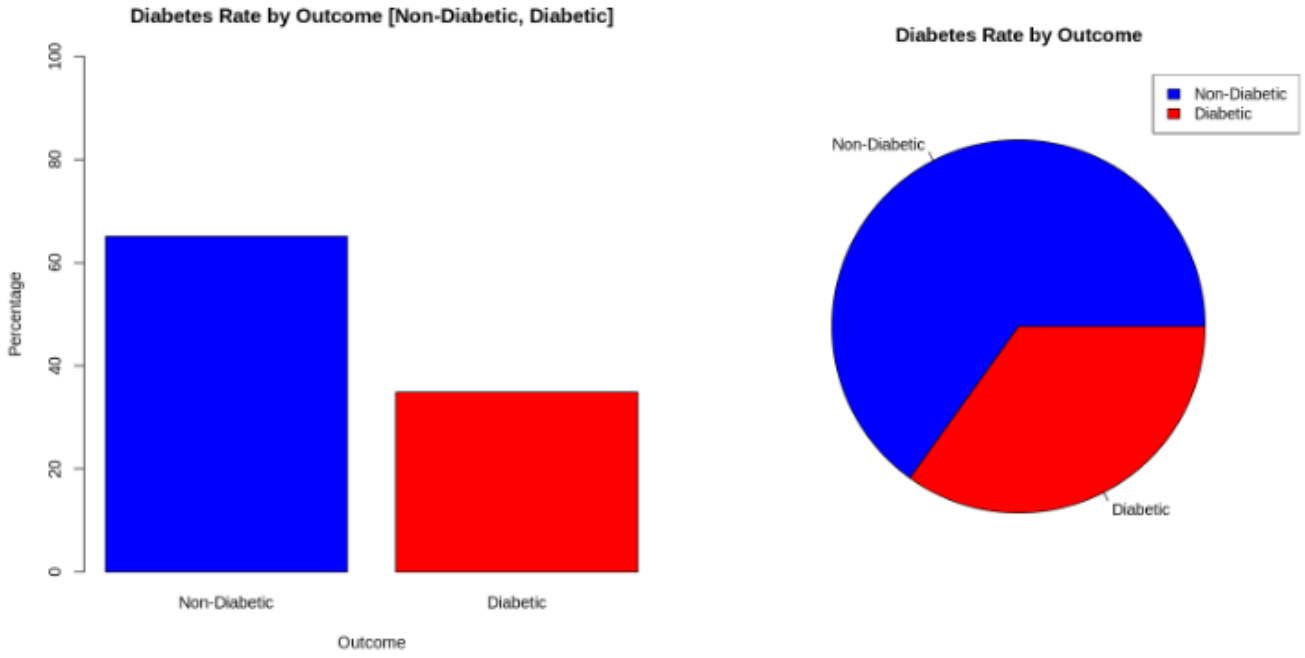
The Bar Chart is used to compare between the BMI of diabetic and non diabetic

Non-Diabetic (0): The blue bar indicates the mean BMI of patients without diabetes. It appears to be around 30 kg/m2.

Diabetic (1): The red bar represents the mean BMI of patients with diabetes. It seems to be approximately 38 kg/m2. Conclusion:

The visualization shows that the average BMI is significantly higher for patients with diabetes compared to those without diabetes.

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



Pieplot shows the percentage of diabetic and nondiabetic.

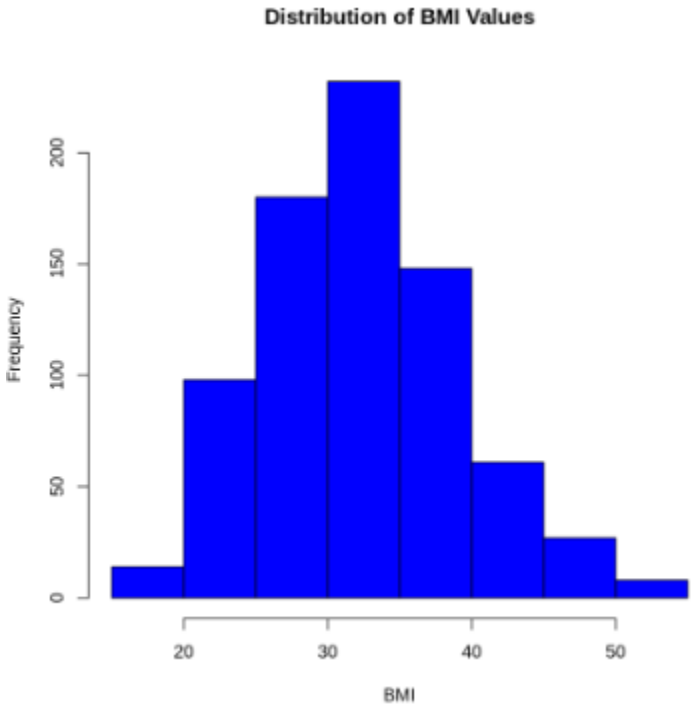
Non-Diabetic: It is around 65%.

Diabetic: it is around 35%.

Conclusion:

The visualization shows that approximately 35% of the patients in the dataset have diabetes, while the remaining 65% are non-diabetic.

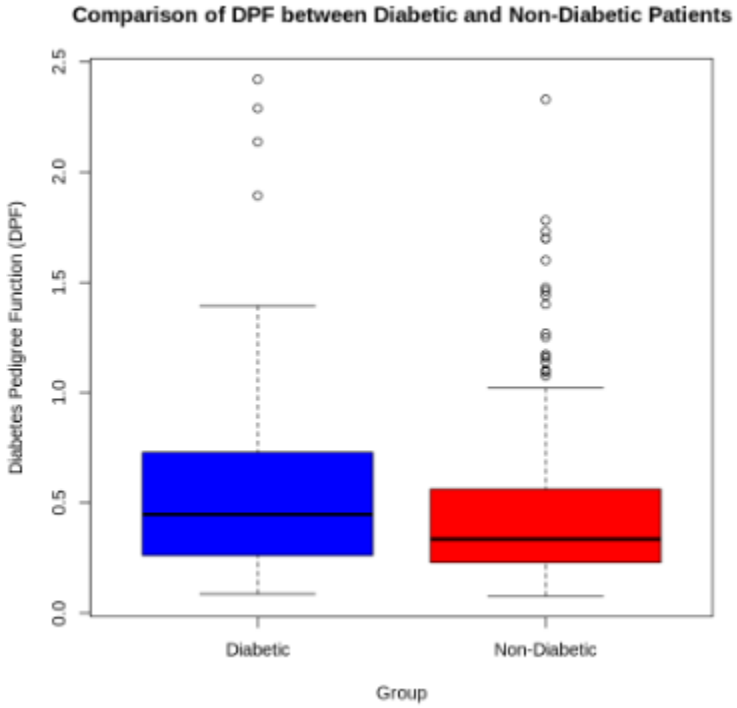
6- The distribution of BMI values among all patients.



The histogram is used to show the distribution of the BMI.

Conclusion : The distribution of BMI values among the patients is right-skewed, with a mean BMI of 32.45 and a median of 32. This indicates that a larger proportion of patients have lower BMI values, while a smaller proportion have higher values. The standard deviation of 6.88 suggests a moderate level of variability in BMI within the sample.

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

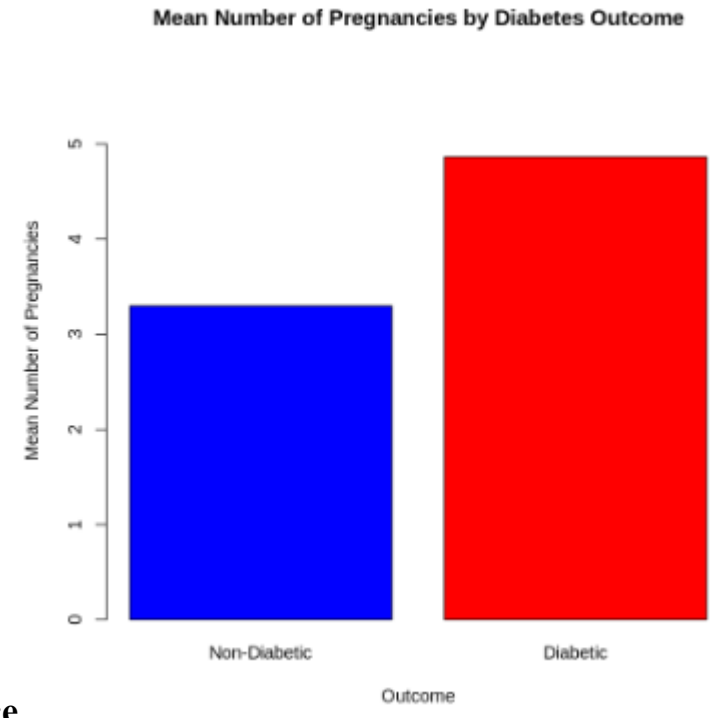


The box plot is used to compare the DPF existence for the diabetic and non diabetic .

Conclusion:

Diabetic patients exhibit a higher median DPF compared to non-diabetic patients, indicating a stronger family history of diabetes in this group.

8- The relationship between the number of pregnancies and diabetes



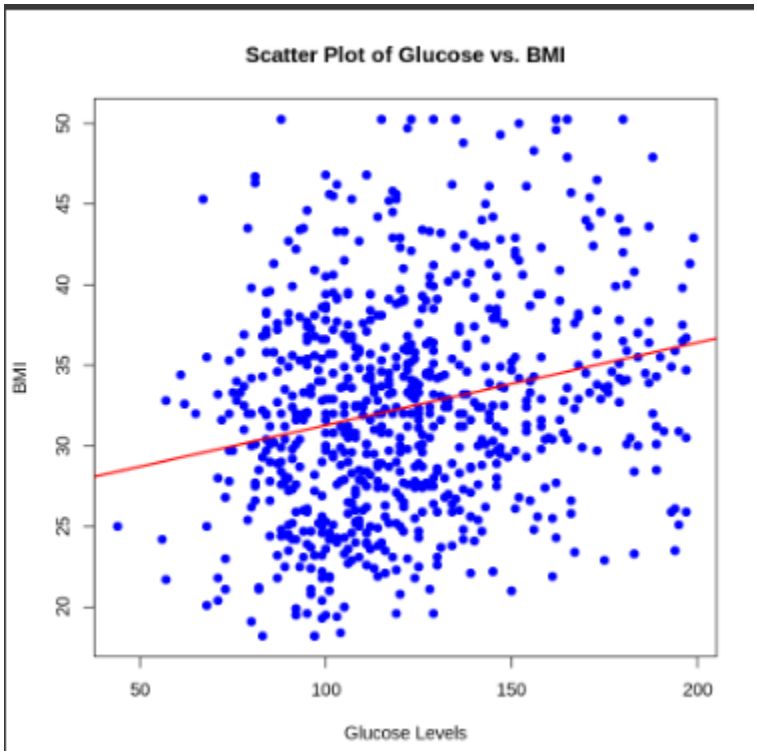
occurrence

The barplot is used to compare the mean of pregnancies that have diabete or not.

Conclusion:

The bar plot reveals a clear association between the number of pregnancies and the risk of developing diabetes. On average, individuals with diabetes have a higher mean number of pregnancies compared to those without diabetes.

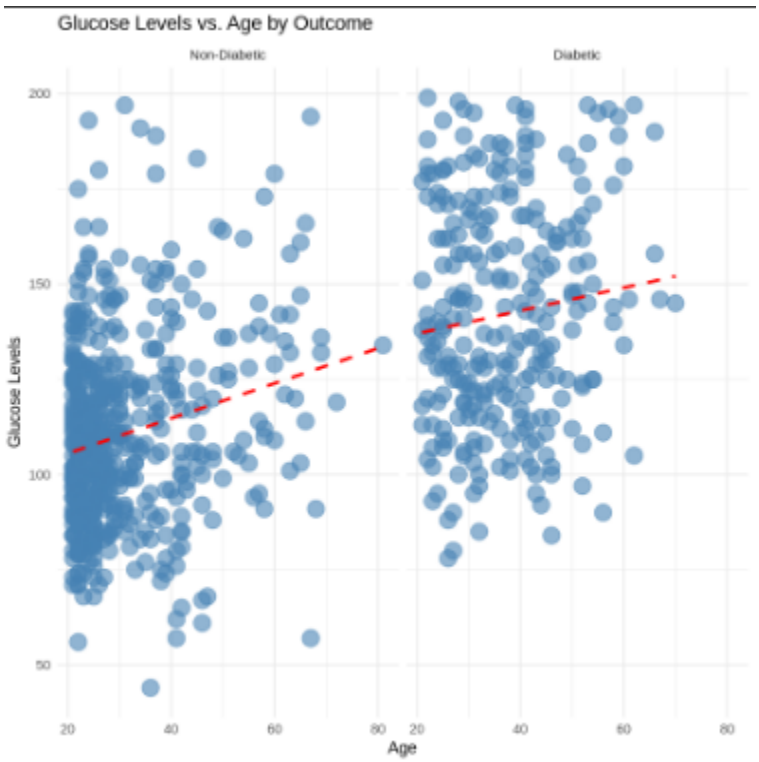
9- The correlation between glucose levels and BMI.



The scatter plot indicates the correlation between the glucose and the BMI.

Conclusion: The correlation coefficient of 0.23 indicate a weak positive correlation between Glucose Levels and BMI.So the relation is not very strong

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



The scatter plot is used to show the disparse of the data(diabetic or not) in relation to age,

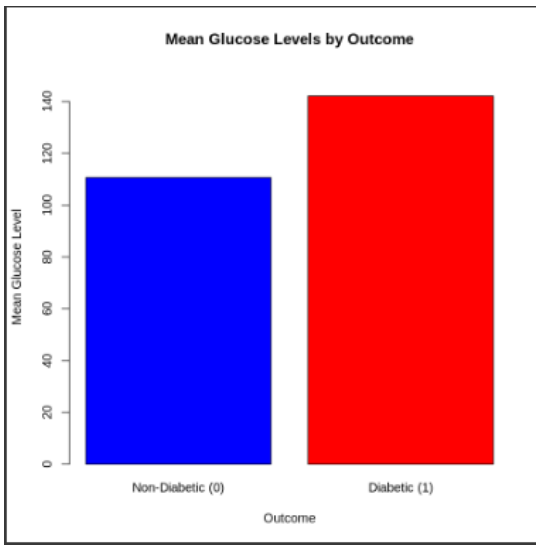
Conclusion:

The scatter plot reveals that glucose levels tend to increase with age in both diabetic and non diabetic individuals. However, the rate of increase shows that it is higher in diabetic patients.

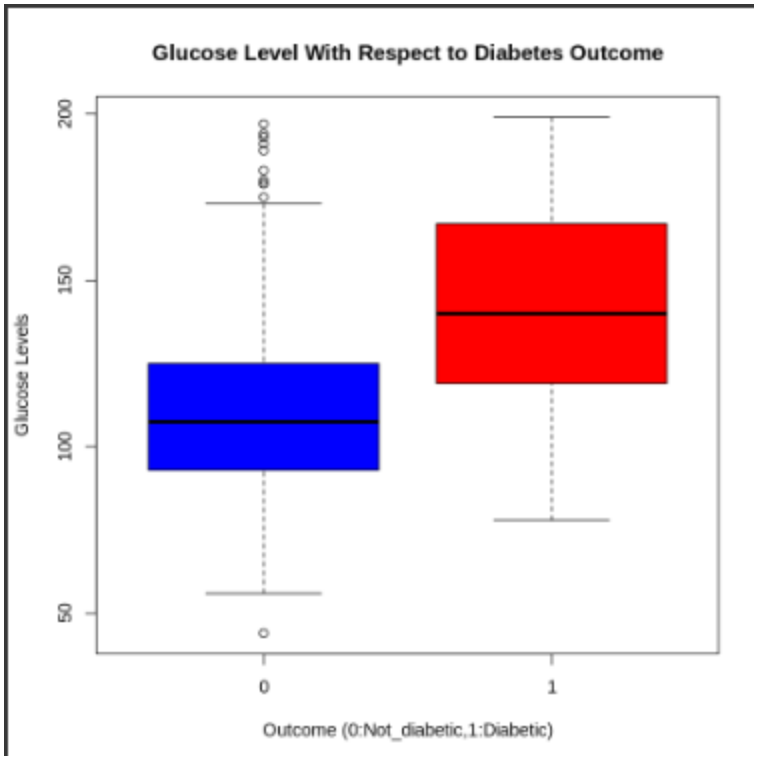
Part 2

Section 2.1

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

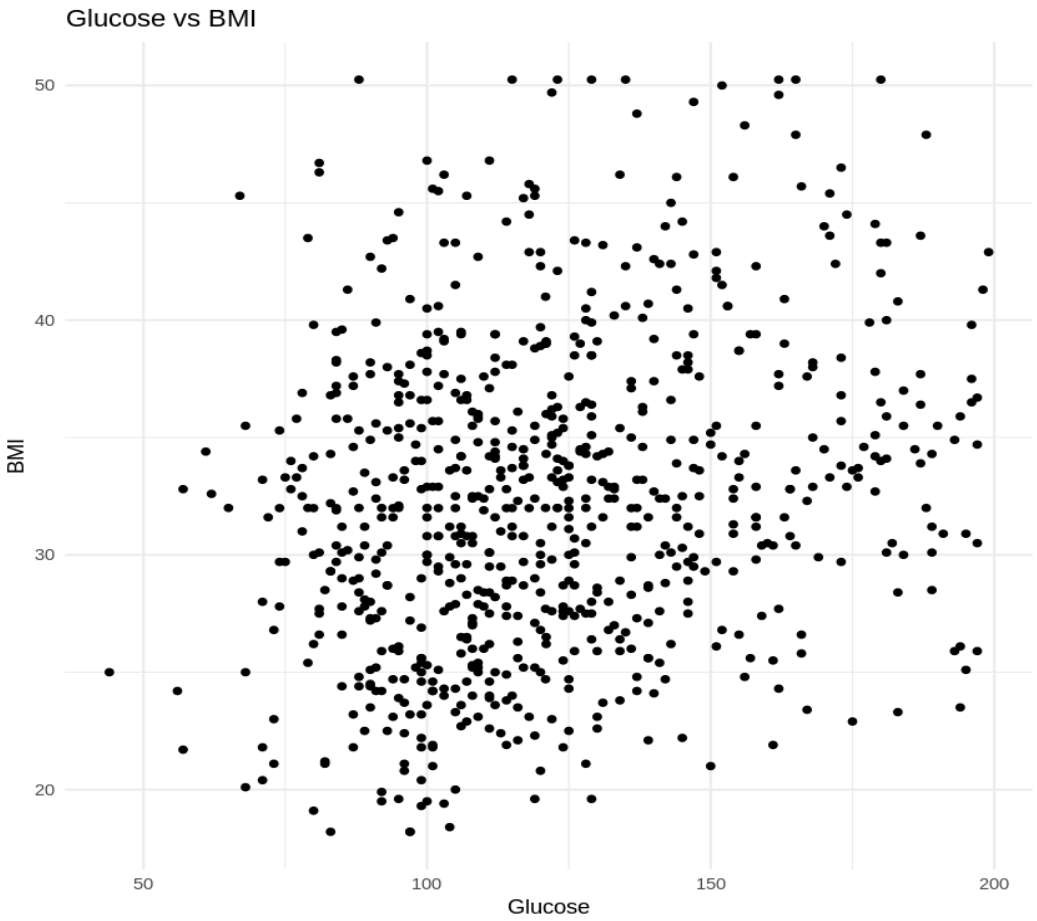


The boxplot will be suitable as it will show mean ,median ,interquartile also the outliers fot glucose level with respect to diabetes



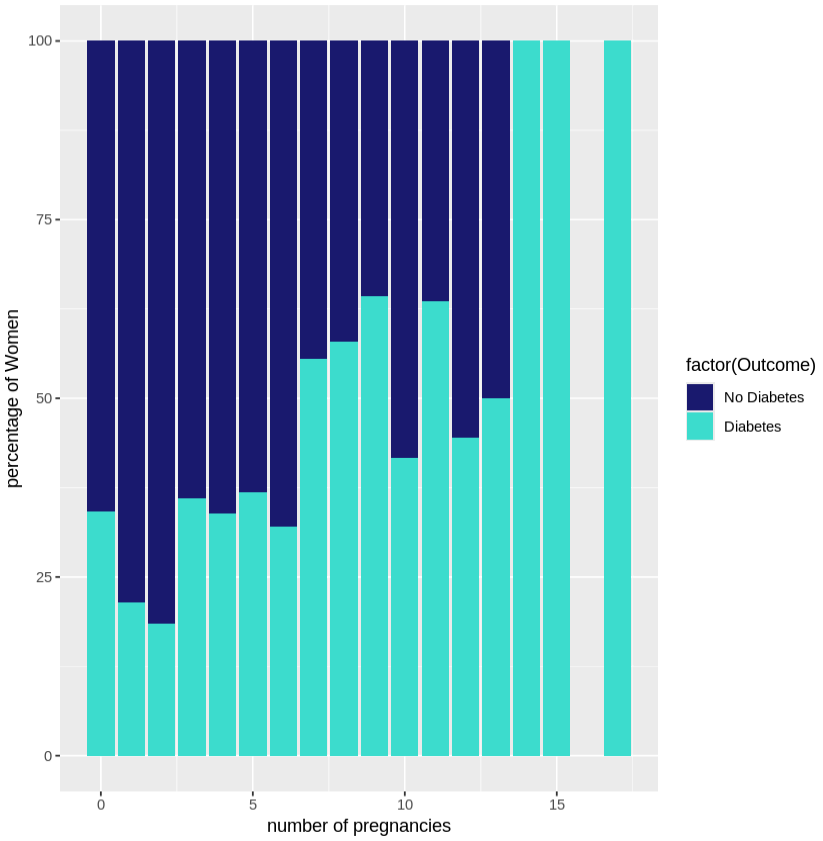
Conclusion: **Yes** higher glucose levels are strongly associated with a greater likelihood of diabetes as shown in the visualizations and statistic

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



CONCLUSION: NO, The scatter plot shows no clear trend between high glucose levels and high BMI

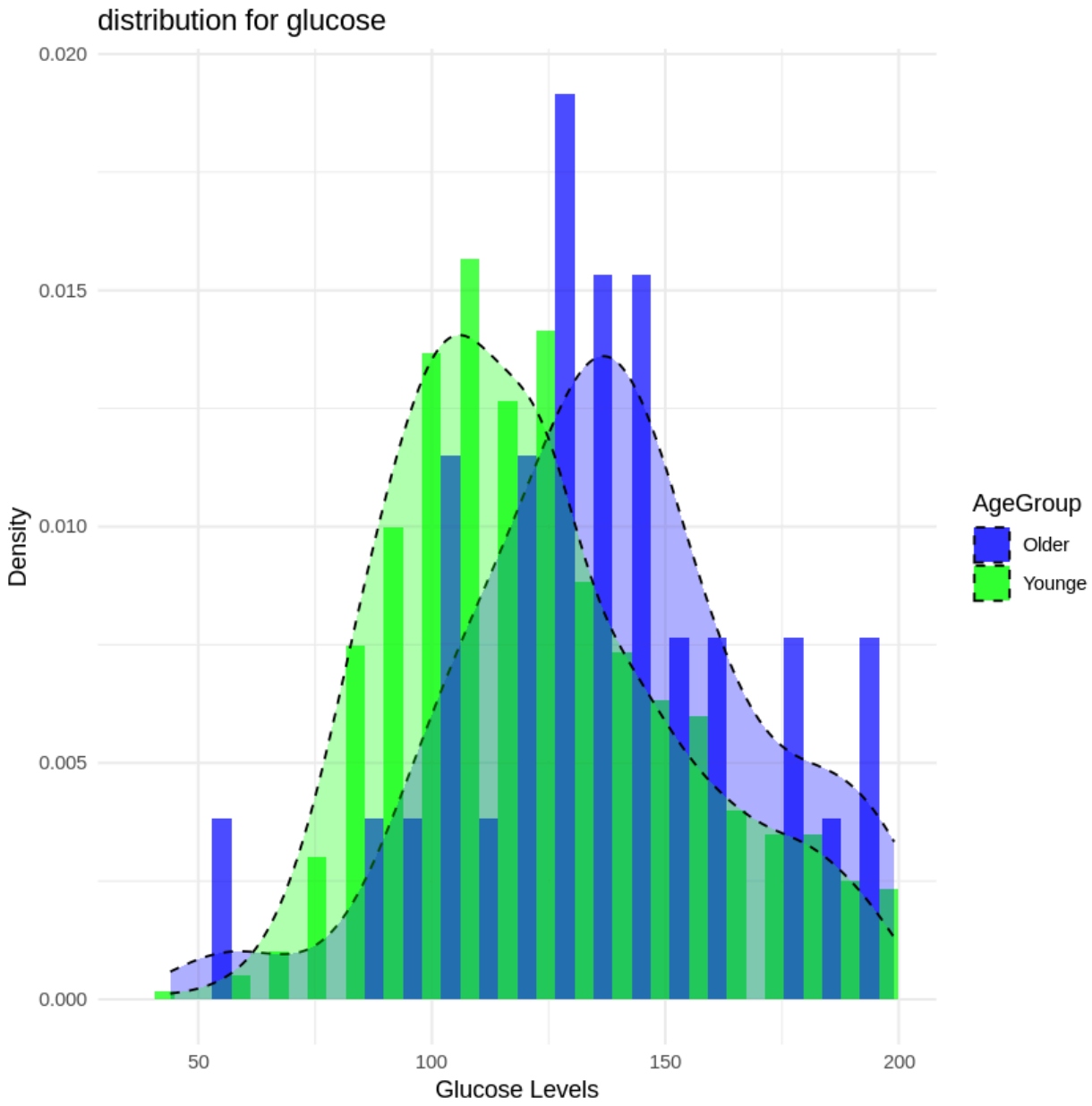
3-Are patients with a higher number of pregnancies at greater risk of developing diabetes?

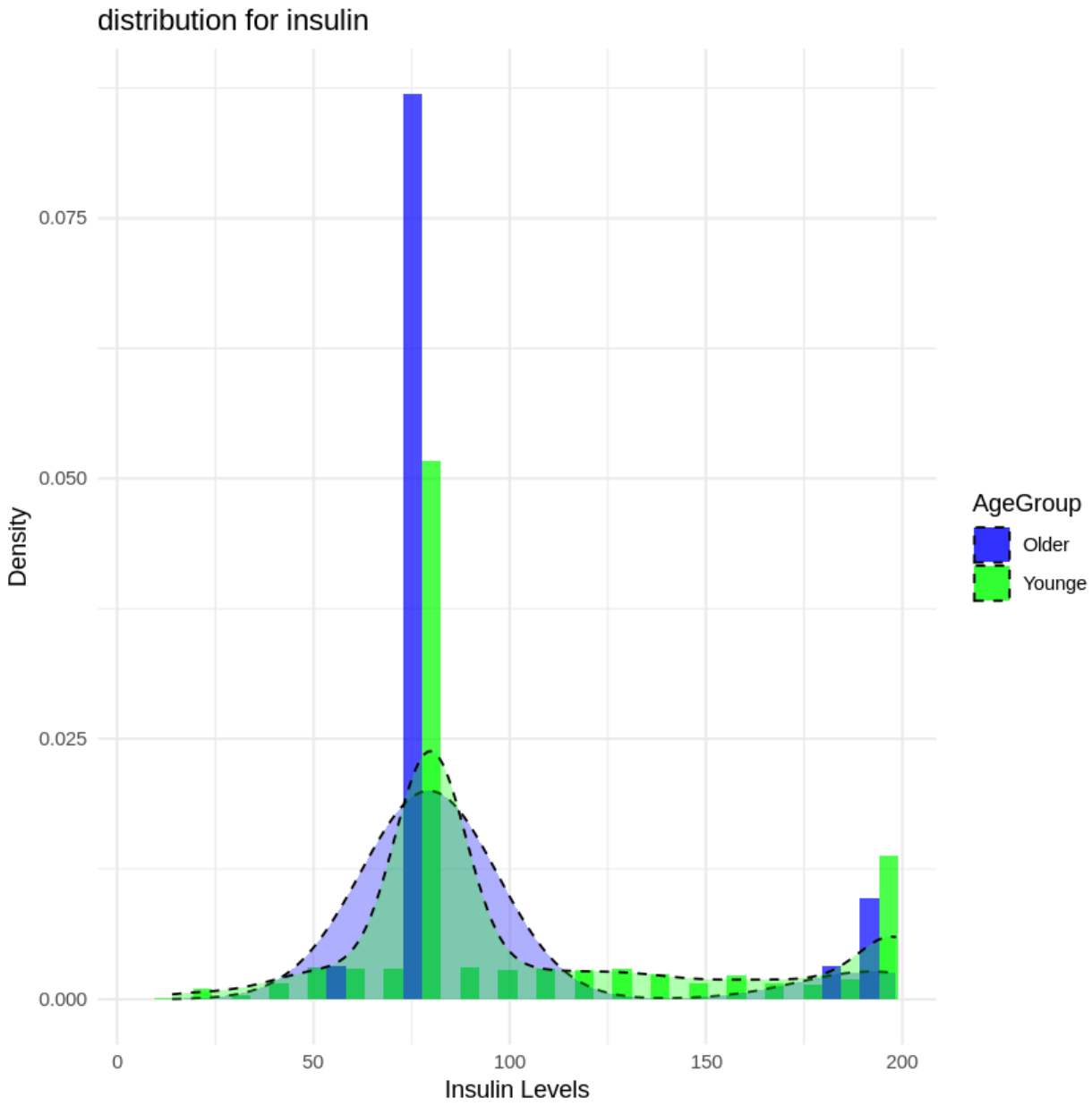


stacked bar chart because The chart allows for a clear comparison between women with diabetes (light blue) and those without diabetes (dark blue) within each pregnancy group.

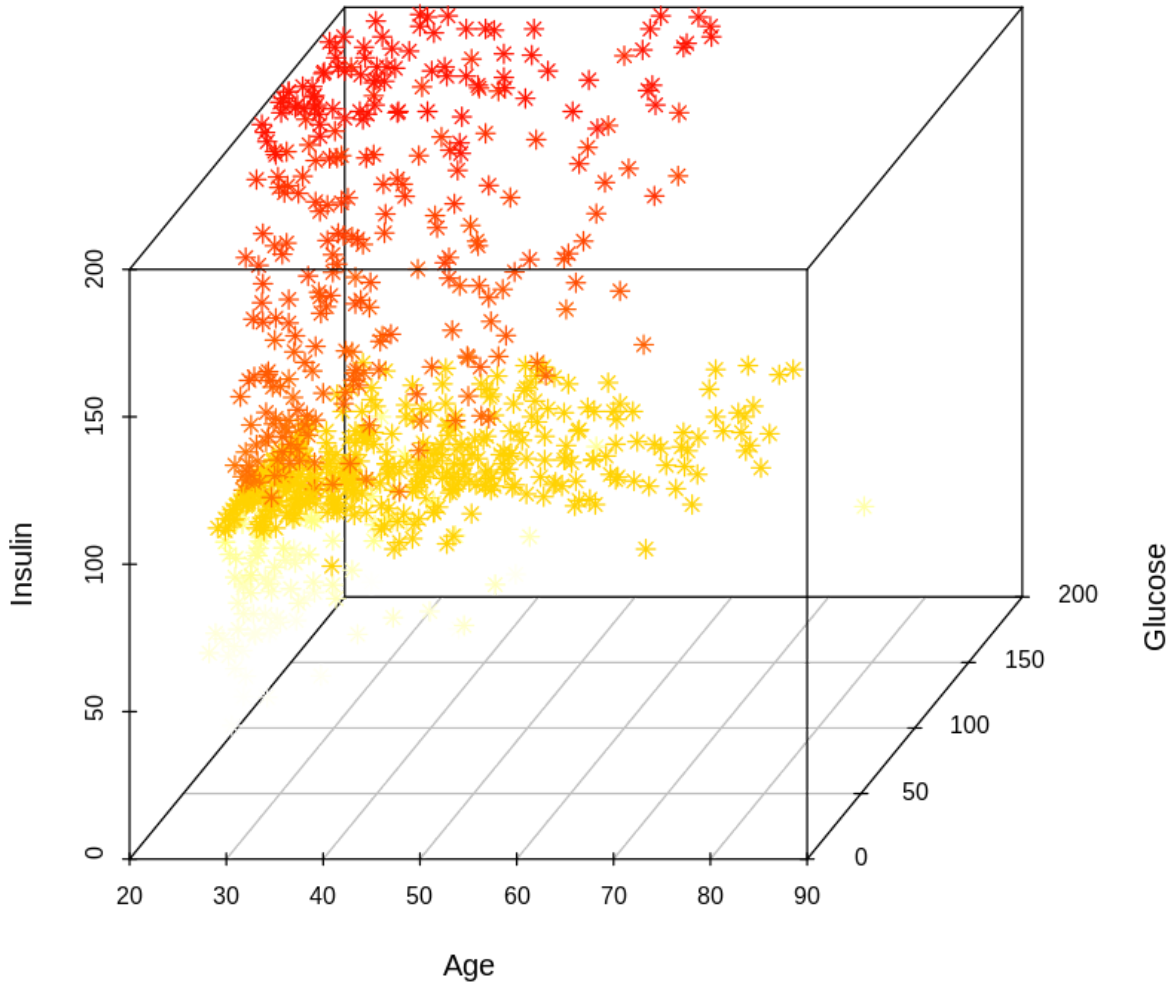
CONCLUSION: yes As the number of pregnancies increases, the proportion of women with diabetes increases, while the proportion of woman with no diabetes decreases so women with a higher number of pregnancies have a greater risk of developing diabetes

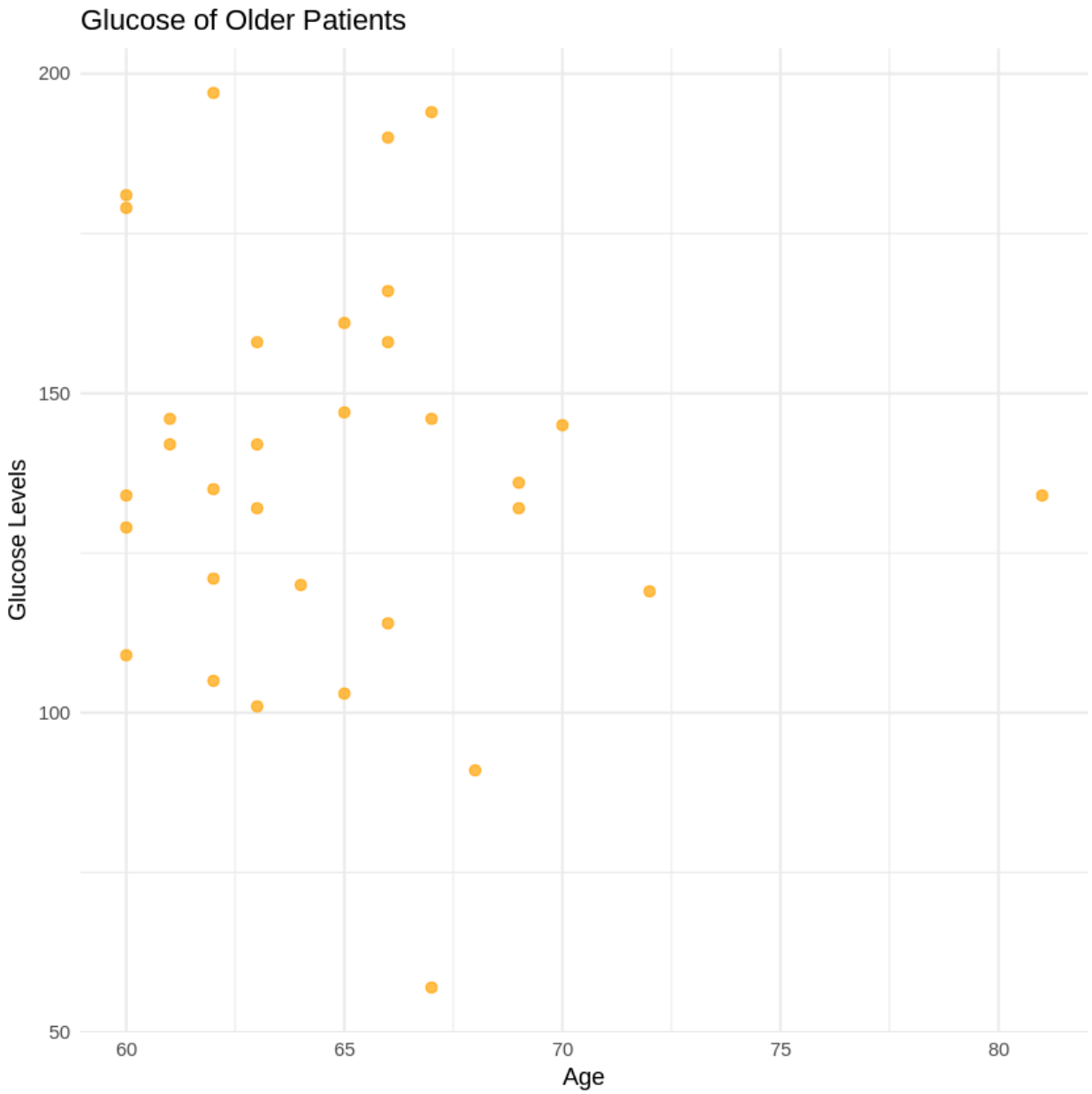
4-Are older patients more likely to have higher insulin concentrations and blood glucose levels?

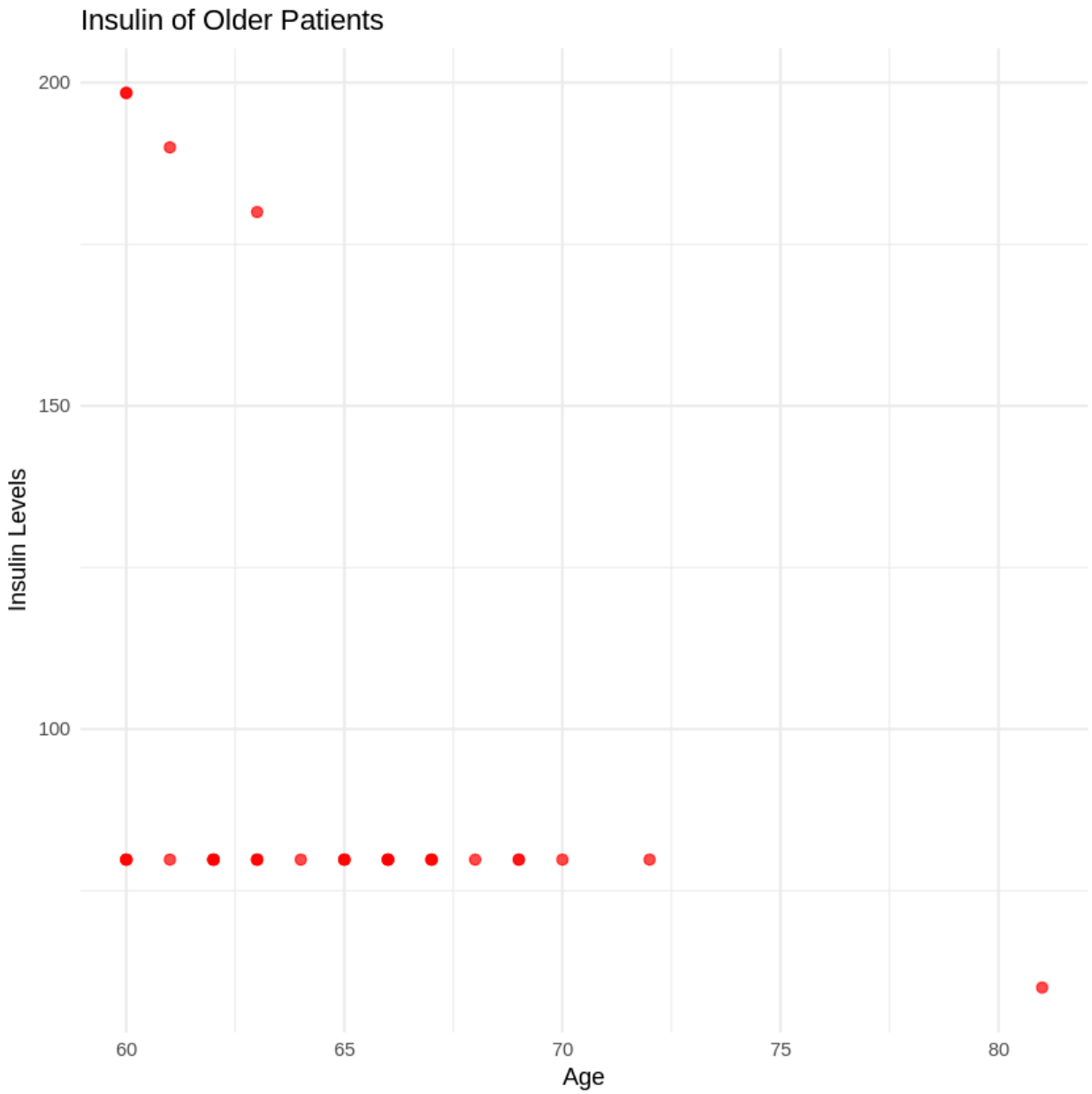

distribution for glucose
0.020
0.015
0.010
0.005
0.000
Density
50
100
150
200
Glucose Levels
AgeGroup
Older
Younge

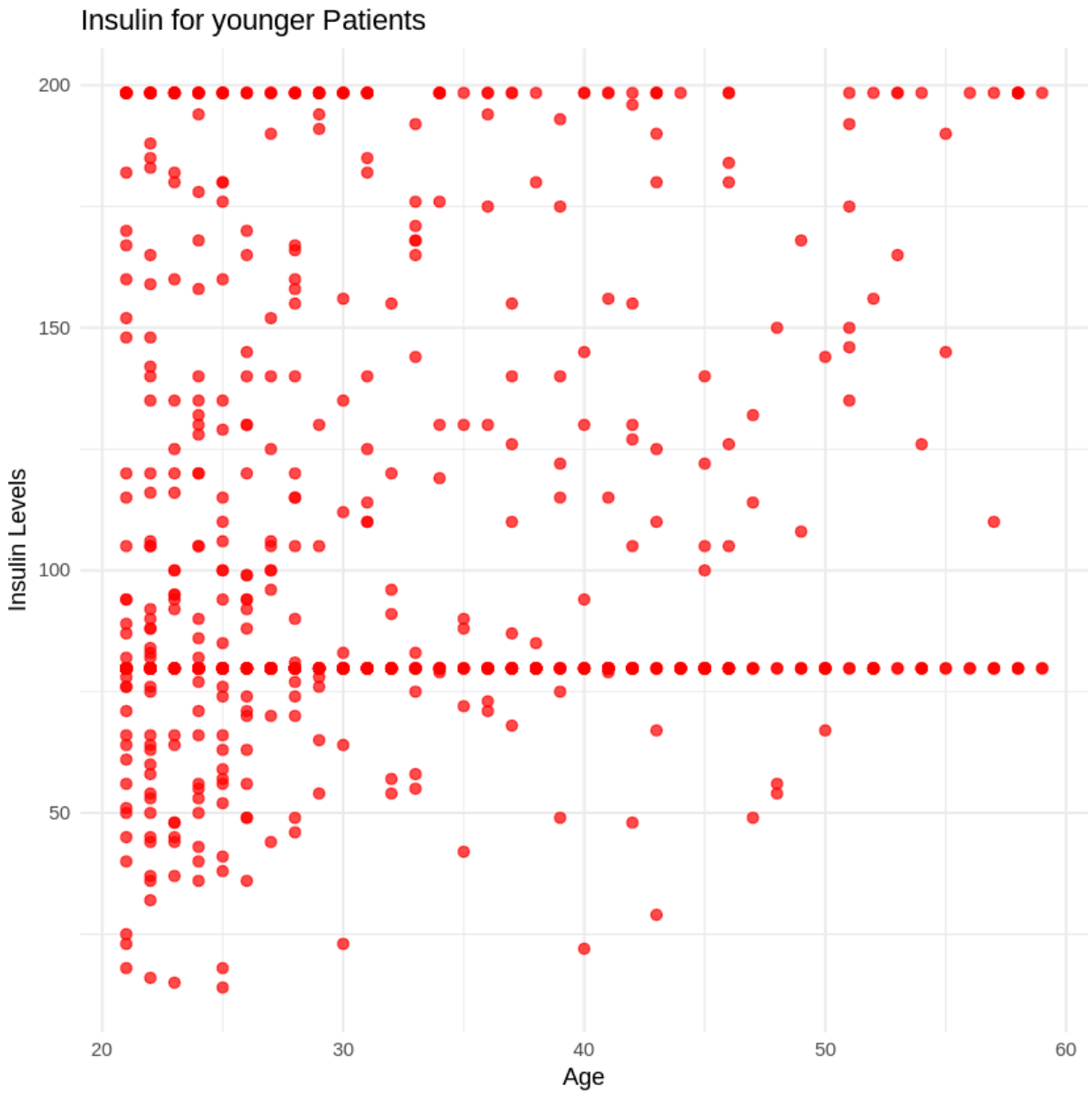

distribution for insulin
0.075
0.050
0.025
0.000
Density
0
50
100
150
200
Insulin Levels
AgeGroup
Older
Younge

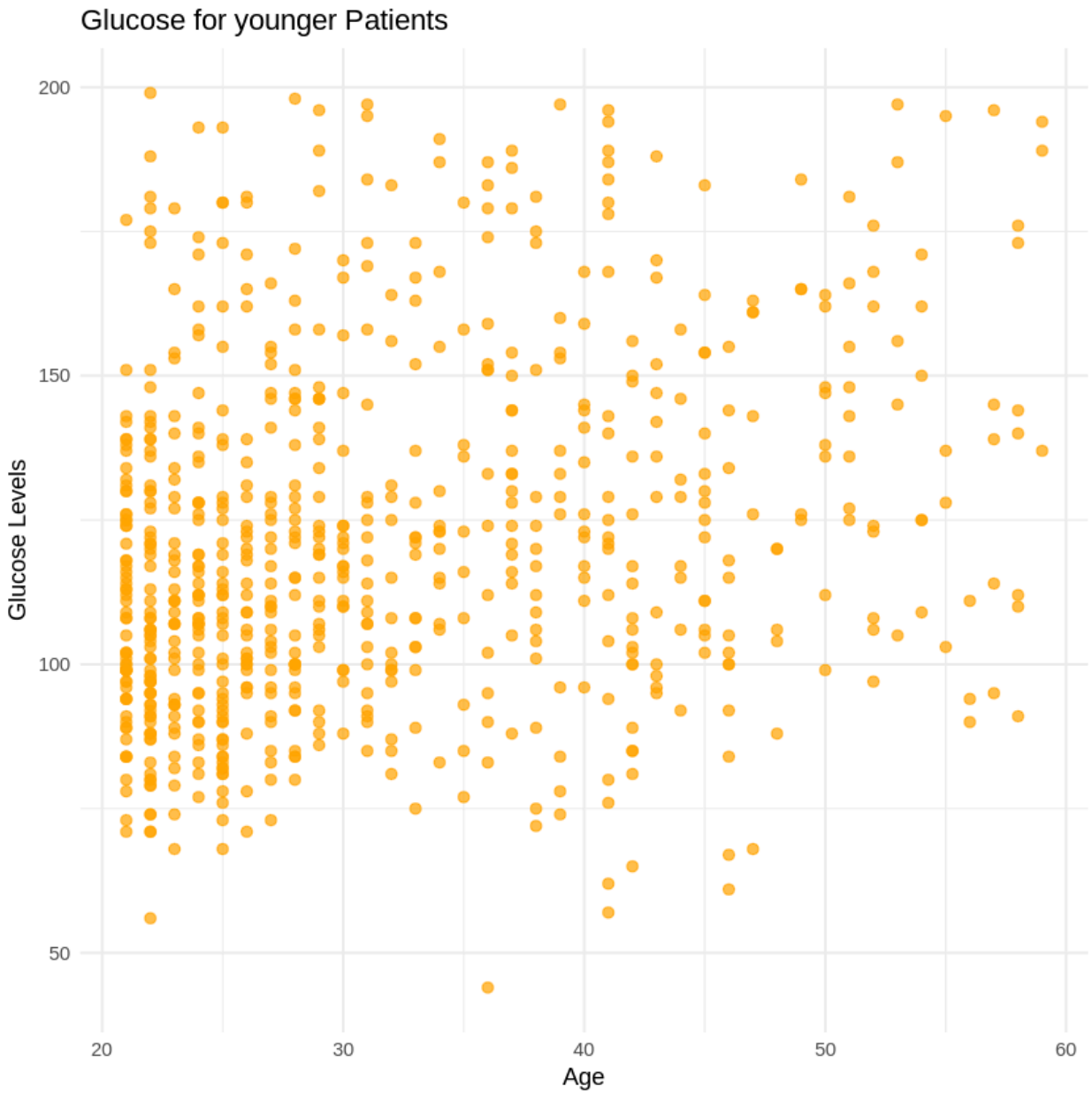
Patients that are more likely to have higher insulin and glucose in blood




Glucose of Older Patients
200
150
100
50
Glucose Levels
60
65
70
75
80
Age

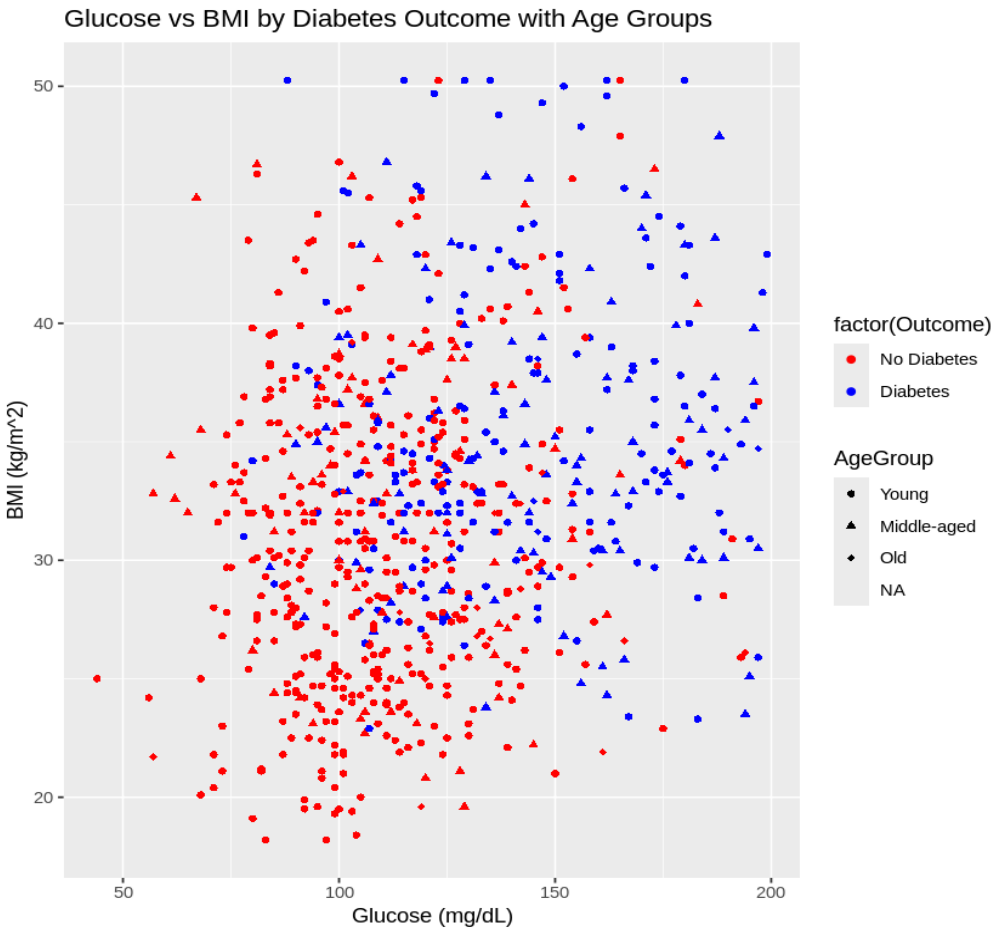

Insulin of Older Patients
200
150
100
Insulin Levels
60
65
70
75
80
Age


Insulin for younger Patients
Insulin Levels
200
150
100
50
20
30
40
50
60
Age



Conclusion: yes Older patients tend to have slightly higher glucose levels than younger ones, but they have lower insulin levels.

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



CONCLUSION: Glucose Levels as a Key Indicator: glucose levels are strongly associated with diabetes.

BMI Influence: the visualization indicates that higher BMI might be a contributing risk factor for diabetes.

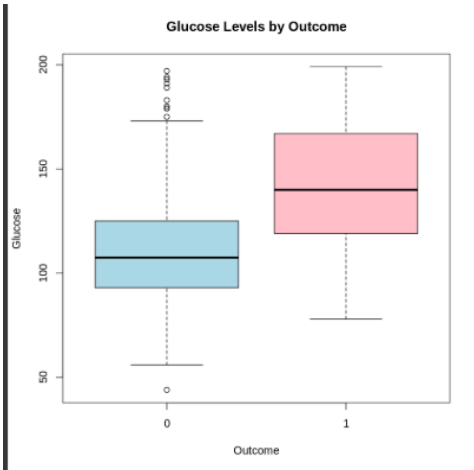
Age Group Impact:

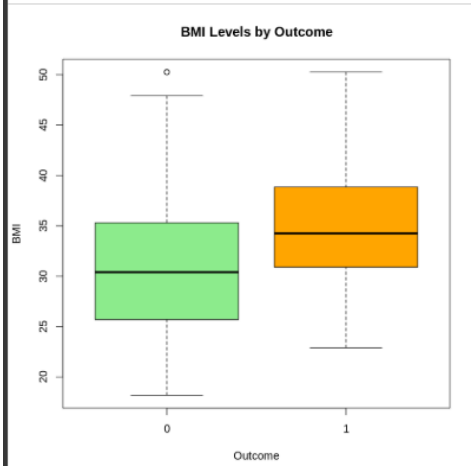
Middle-aged and older individuals are more likely to have diabetes

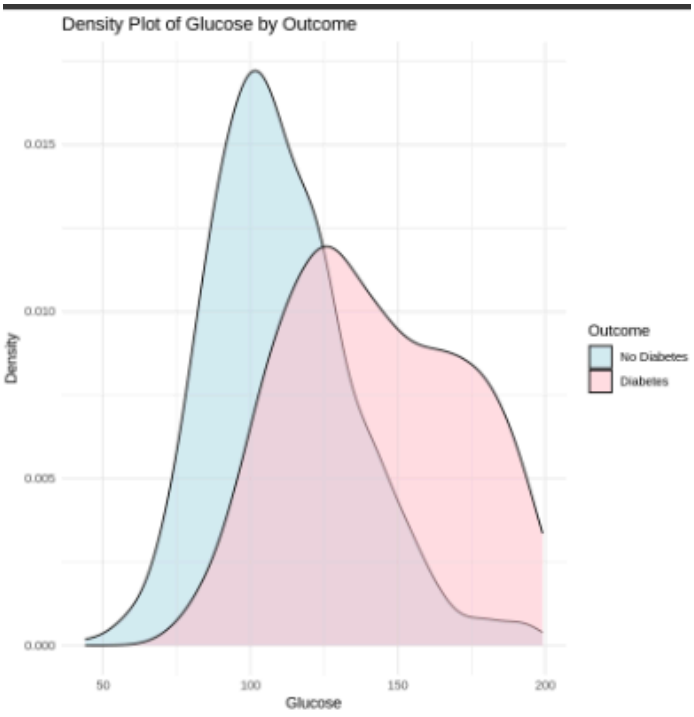
Section 2.2

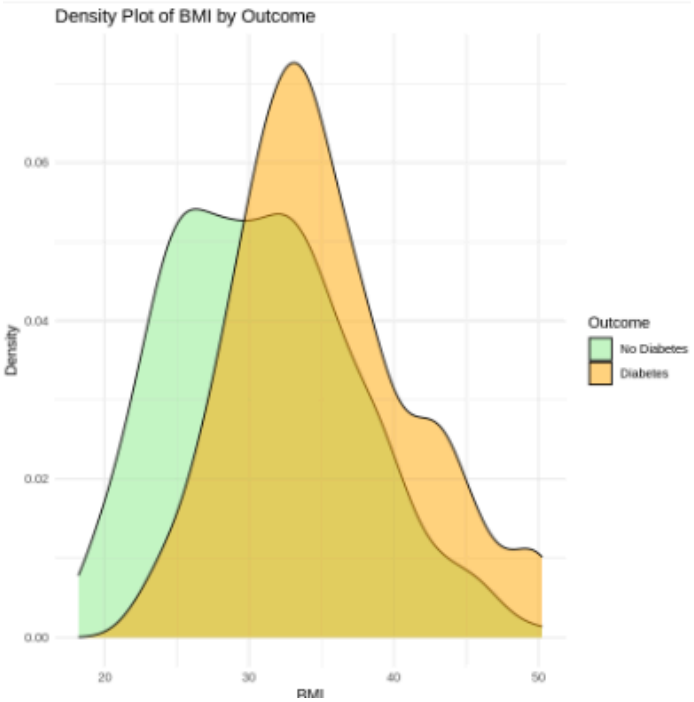
1- Do high glucose levels with high BMI indicate a higher probability of being diabetic?

conclusion: yes as shown in the visualizations people who has high glucose levels and high BMI are more likely to have diabetes

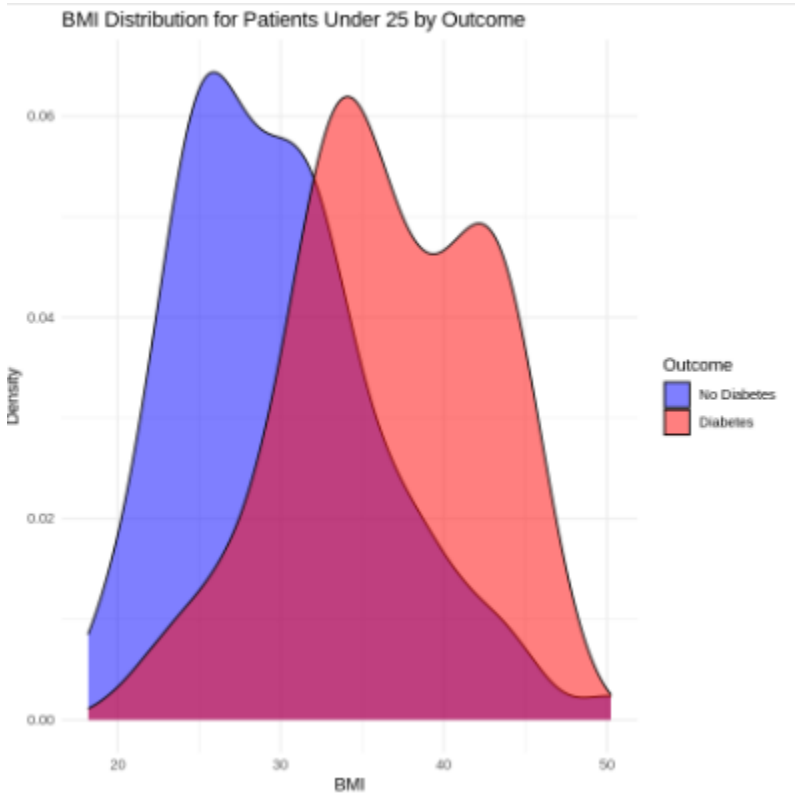

Glucose Levels by Outcome
Glucose
Outcome
0
1


BMI Levels by Outcome
BMI
Outcome
0
1


Density Plot of Glucose by Outcome
Density
Glucose
Outcome
No Diabetes
Diabetes


Density Plot of BMI by Outcome
Density
BMI
Outcome
No Diabetes
Diabetes

2- Do young samples with high BMI more likely to be diagnosed as diabetic?

- 

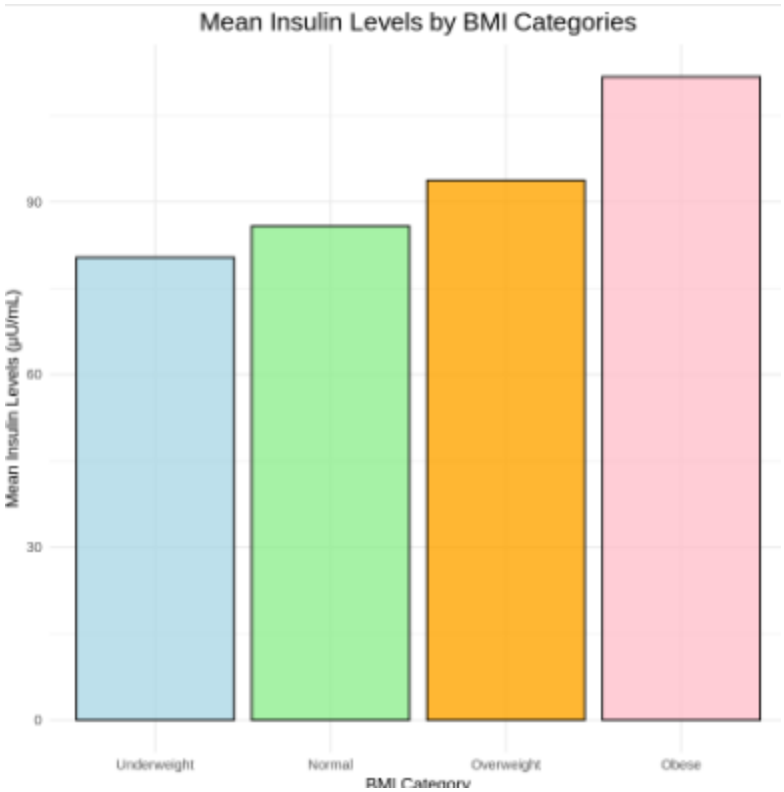
From the graph we notice:

- non diabetes people : The BMI distribution peaks at around 25 (lower BMI values)
- Diabetes people: The BMI distribution peaks at around 34

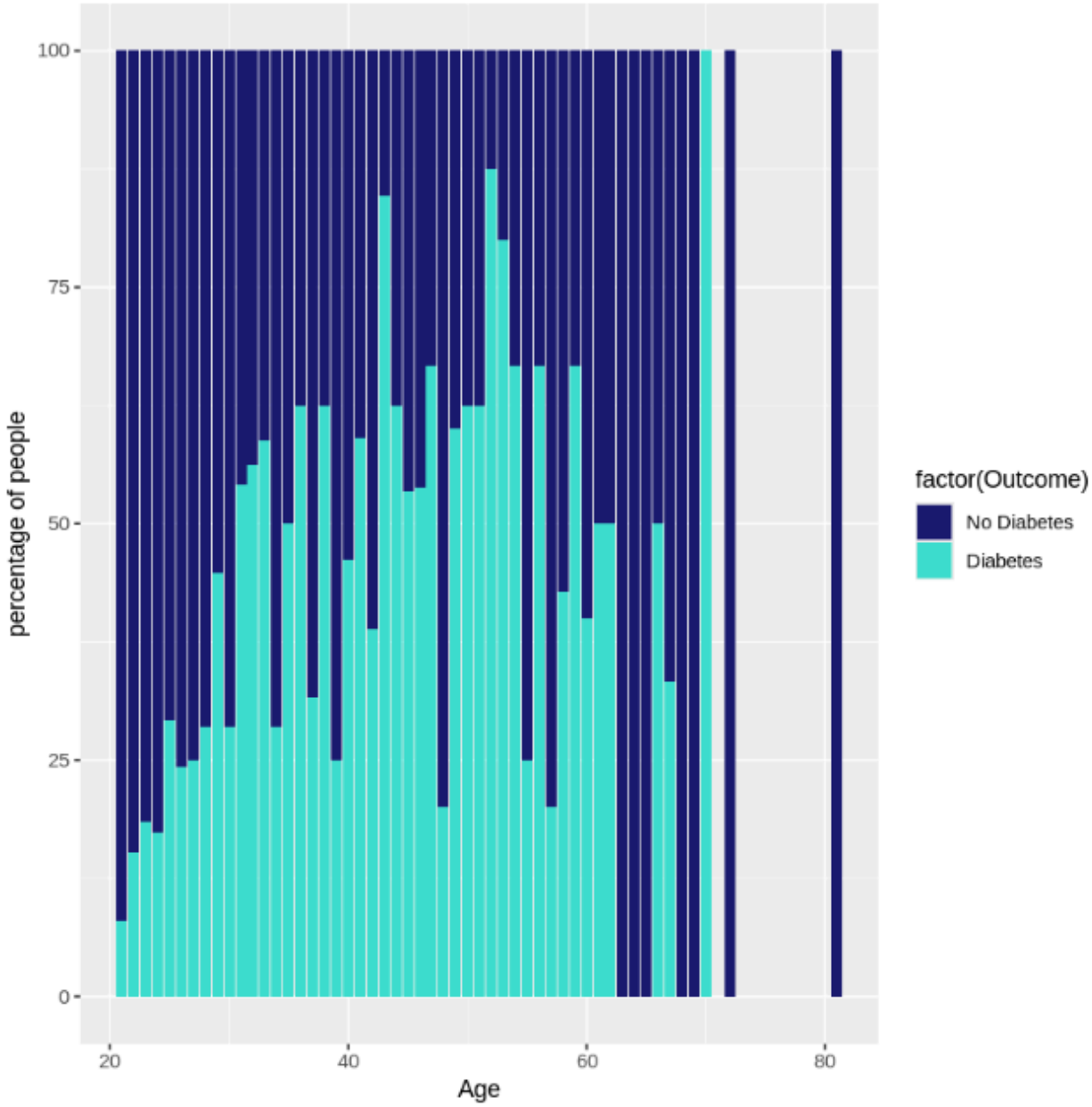
conclusion: yes, higher BMI is associated with a higher likelihood of diabetes for young people

3- Do insulin levels vary by BMI categories?

Yes ,as shown in the graph there is an direct relationship BMI categories and mean insulin levels as people in higher BMI categories has higher mean insulin levels compared to those in lower BMI categories

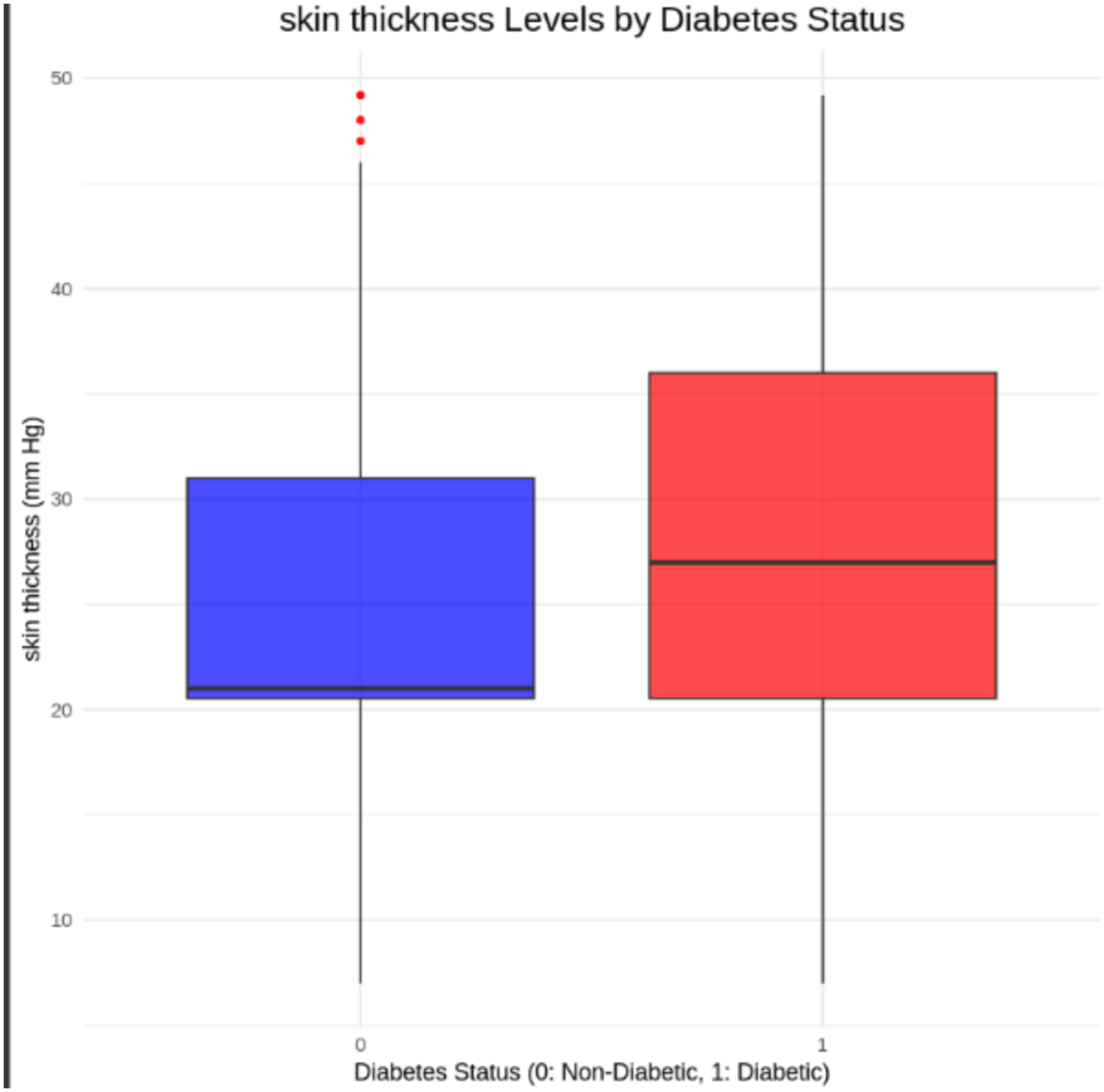

Mean Insulin Levels by BMI Categories
Mean Insulin Levels (µU/mL)
90
60
30
0
Underweight
Normal
Overweight
Obese
BMI Category

4- Does age affect diabetes?



yes , age is an important factor that affect diabetes
From the graph we notice older group seems to have higher percentage of people with diabetes than the younger group

5- does skin thickness affect diabetes?



yes from the graph we can notice that high skin thickness is more likely related with diabetes

Part 3

Section 3.1

Claim :**There is a significant difference in glucose levels between diabetic and non-diabetic patients.**

- **We will use t-test of two samples**
- **Reasons:**
  - Comparison is between Diabetic and Non Diabetic which are Two independent groups
  - Population standard deviation(σ) is unknown so the sample standard deviation(S) is used
  - n≥30 (Central Limit Theorem).

- **Stating the hypotheses**

H0: μ1 = μ2 vs. Ha: μ1 ≠ μ2

Where:

μ1 is the mean of glucose in diabetic

μ2 is the mean of glucose in non diabetic

- **We conducted T-test using the following rule:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The result was that t-statistic (t): 14.86244

- **Then calculated df get the p value**

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Degrees of Freedom = 469.498

- **We calculate the p-value :**

  p-value = 0 because its value is extremely small as the difference between the means of the diabetic and non-diabetic groups is very large

- **Since p_value is less than alpha (0.05)**

  Therefore our conclusion is that we **reject the null hypothesis (H0)**

  And accept H alternative

  Therefore

"There is a significant difference in glucose levels between diabetic and non-diabetic patients."

Section 3.2

We Claimed that : **non-diabetic people have a higher insulin level than diabetic people.**

- **We will use t-test of two samples**
- **Reasons:**
  - Comparison is between Diabetic and Non Diabetic which are Two independent groups
  - Population standard deviation(σ) is unknown so the sample standard deviation(S) is used
  - n≥30 (Central Limit Theorem).

- **Stating the hypotheses**

H0: μ1 ≥ μ2 vs. Ha: μ1 < μ2

Where:

μ1 is the mean of insulin to people who have diabetes

μ2 is the mean of insulin to people who don't have diabetes

- **We conducted T-test using the following rule:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The result was that t-statistic (t): 5.664438

- **Then calculated df get the p value**

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Degrees of Freedom = 484.6704

- **We calculate the p-value :**

p-value = 1

- **Since p_value is greater than alpha (0.05)**

Therefore our conclusion is that we **Fail to reject the null hypothesis (H0)**

So

We reject H alternative hypothesis ,therefore :

“ non-diabetic people have not a higher insulin level than diabetic people”

Part 4

1. Sampling from the Dataset

We randomize samples from the Glucose column of the dataset to explore how the sample size and confidence interval affect the coverage of the true population mean.

Steps

- For each sample size, we generated 25 samples. The sample sizes used are 15, 100, 10, and 10.
- These samples are drawn without replacement, and the size of each sample is fixed.

2. Calculating Confidence Intervals

Once the samples are drawn, we calculate the 95% confidence interval for the mean of each sample. This process involves using two different methods depending on the sample size:

- For sample sizes less than 30, the **t-distribution** is used.

- For sample sizes greater than 30, the **z-distribution** is used.

The formula for the confidence interval is:

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

Where:

CI: Confidence interval

x̄:Sample mean

Z:Confidence level value

S:sample standard deviation

n:Sample size

3. Assessing the Coverage Proportion

We evaluate how many of the confidence intervals contain the true population mean. This is done by checking if the true population mean lies between the lower and upper bounds of each confidence interval.

- The proportion of confidence intervals that contain the true population mean is calculated for each sample size.

4. Calculating the Average Width of Confidence Intervals

For each sample size, we also compute the average width of the confidence intervals. A narrower confidence interval indicates more precision, while a wider interval indicates less precision.

Results

Does the width of the confidence intervals increase or decrease?

- The width of the confidence intervals decreases when the sample size increase

Does increasing the sample size result in more or fewer intervals containing the true population mean?

- Increasing the sample size results in more intervals containing the true population mean

A data.frame: 4 × 4

Number_of_samples	sample_size	Coverage_proportion	Average_width_of_interval
<dbl>	<dbl>	<dbl>	<dbl>
25	15	0.92	29.17569
25	100	0.96	10.13981
25	10	0.84	35.44327
20	10	0.90	36.69492