kSchool of Computational Sciences and Artificial Intelligence
DSAI 307: Statistical Inference
Project

Final Course Project
Analyzing Diabetes Dataset

Team Members:

Doha Hemdan 202200701

Galal Qasas 202201379

Ibrahem Ali 202201987

Salma Sherif 202200622

Ziad Moutaz 202201252

1. Introduction

- Diabetes dataset, from Kaggle is used for this project. The set indeed contains different health indications, including all those factors linked with the risk for diabetes, like: blood pressure and glucose concentration.

- The dataset is relevant because it may lead to the identification of at-risk individuals based on health metrics, thus providing better prevention techniques and increasing public health awareness.

- The project aims at analyzing the dataset in order to establish the key factors that contribute to the development of diabetes. The analysis will try to establish health indicators that best correlate with diabetes, besides establishing patterns and trends and succinctly.

- Here's what the columns represent:
    - **Pregnancies:** Number of times the patient has been pregnant.
    - **Glucose:** Plasma glucose concentration after a 2-hour oral glucose tolerance test.
    - **BloodPressure:** Diastolic blood pressure (mm Hg).
    - **SkinThickness:** Triceps skinfold thickness (mm).
    - **Insulin:** 2-hour serum insulin (mu U/ml).
    - **BMI:** Body mass index (weight in kg/(height in m)^2).
    - **DiabetesPedigreeFunction:** A function that represents the patient’s diabetes pedigree (i.e., likelihood of diabetes based on family history).
    - **Age:** Age of the patient (years).
    - **Outcome:** Binary outcome (0 or 1) where 1 indicates the presence of diabetes and 0 indicates the absence.

2. Data Processing and Analysis Steps

Description of the methods used for data cleaning and preprocessing.

we checked the values in each column. We found some columns have 0 values which is not biologically possible

- **Glucose:** 5
- **BloodPressure:** 35
- **SkinThickness:** 227
- **Insulin:** 374
- **BMI:** 11

  Replacing these biologically implausible zero values with NA allows a more accurate representation of the missing information, and helps us use packages to impute the missing values.

Imputing Missing Values using K-Nearest Neighbors (KNN):

- After identifying missing values, we used the knnImputation() function from the **DMwR2** package to impute them.

  Imputation is often preferred over removing rows with missing data, especially when the data is small. KNN imputation is simple and does the required job.

Removing Outliers:

We defined a function to identify and remove outliers. It gets the first and third quartiles, and subtracts them to get the interquartile range (IQR), then defines the lower bound as Q_1 - 1.5 * IQR and the upper bound as Q_3+ 1.5 * IQR.

Outliers influence our statistical analysis leading to skewness and bias in our results. Removing them guarantees that our analysis is robust and reliable.

- Overview of the analytical approach

  The analysis started by answering the main proposed questions by getting information for all the features in the dataset, for example getting the “Mean of Glucose” for both diabetic and non-diabetic

patients, and the distribution of the ages, blood pressure, and BMI of those patients.

Then moving to the relationship between some features like “Pregnancy” and being diabetic or not which shows a piece of exciting information by increasing the pregnancy counts the percentage of being diabetic or not is getting closer, showing also the relation of BMI and Glucose which show nearly no correlation and

Finally, the trend of glucose level over the age for both diabetic and non-diabetic patients shows an interesting thing that the control for glucose level is getting harder by increasing the age and causes weird variances.

3. Challenges, Limitations, and Assumptions

3.1. Challenges

Missing Data (Null Values):

Challenge: One of the major problems we had to face was missing values in columns, which can tamper with the integrity and accuracy of results.

Solution: We employed K-Nearest Neighbors (KNN) for imputation since it does a pretty efficient filling in of missing values in the dataset using data point similarity.

Outlier in the Data:

Challenge: Most columns have outliers, which will distort the statistical analysis.

Solution: Outliers were identified and removed by the Interquartile Range in order not to affect the analysis.

Weak Correlation:

Challenge: A significant challenge is the lack of strong correlations between most column outputs.

Solution: Advanced techniques like regression, clustering, and feature engineering enable us to analyze the data effectively and uncover hidden patterns to enhance insights.

3.2. Limitations

Data Scope & Generalizability:

This data demographic - women patients only and that does not represent the general population, and will not lead to generalization.

Limited variables:

The dataset lacks features relevant for insight into the development of diabetes, such as family medical history related to diabetes, hypertension, and other diseases.

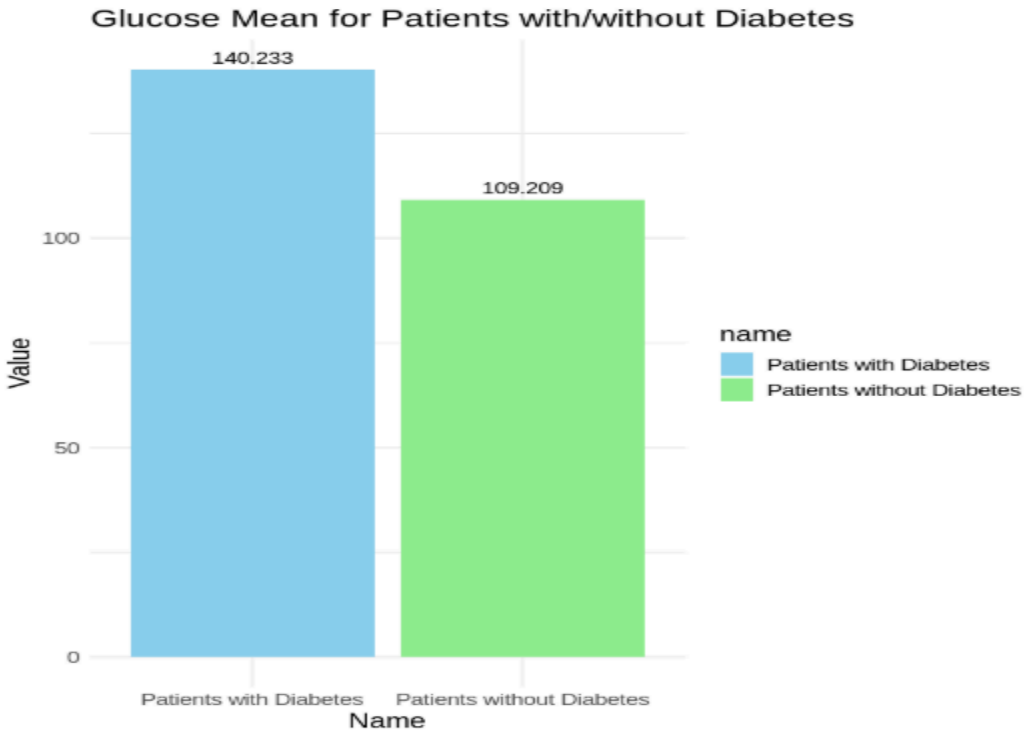
3.3. Assumptions

- Random Sampling: Data points were randomly sampled with replacement.

- KNN Imputation: Missing data is assumed to be missing at random, allowing KNN to be effective.

4. Results and Visualizations

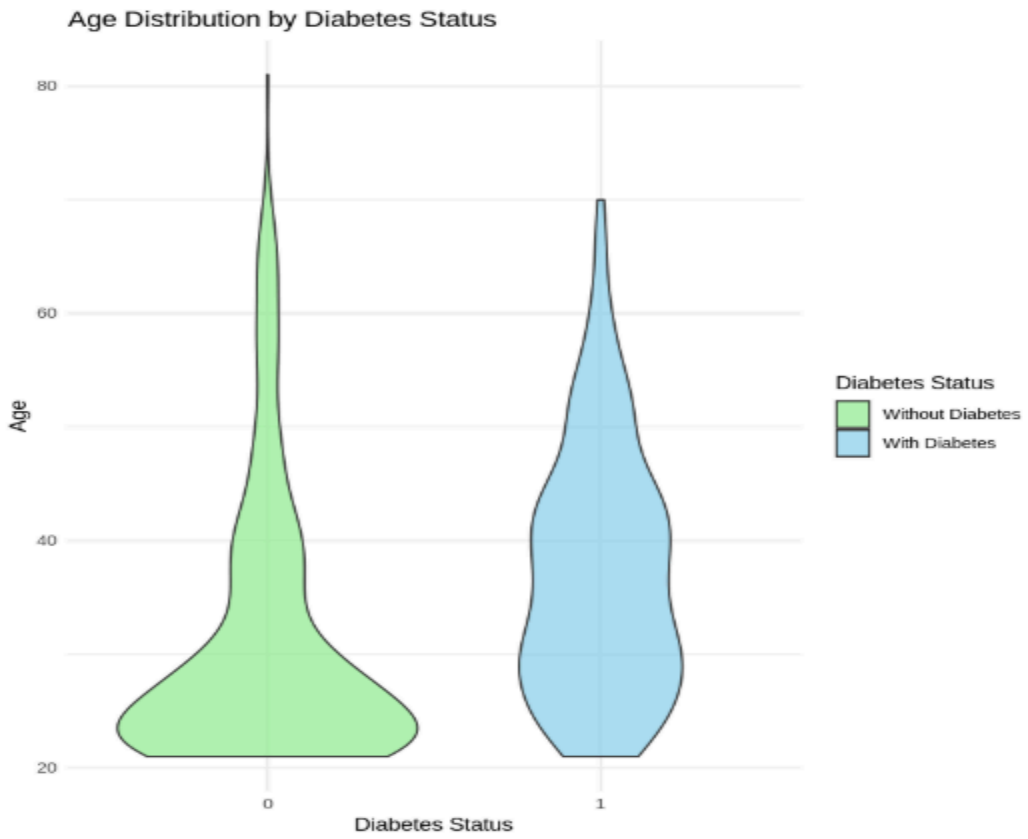
- **Exploratory Analysis:**



Glucose Mean for Patients with Glucose and without Glucose

Interpretation:

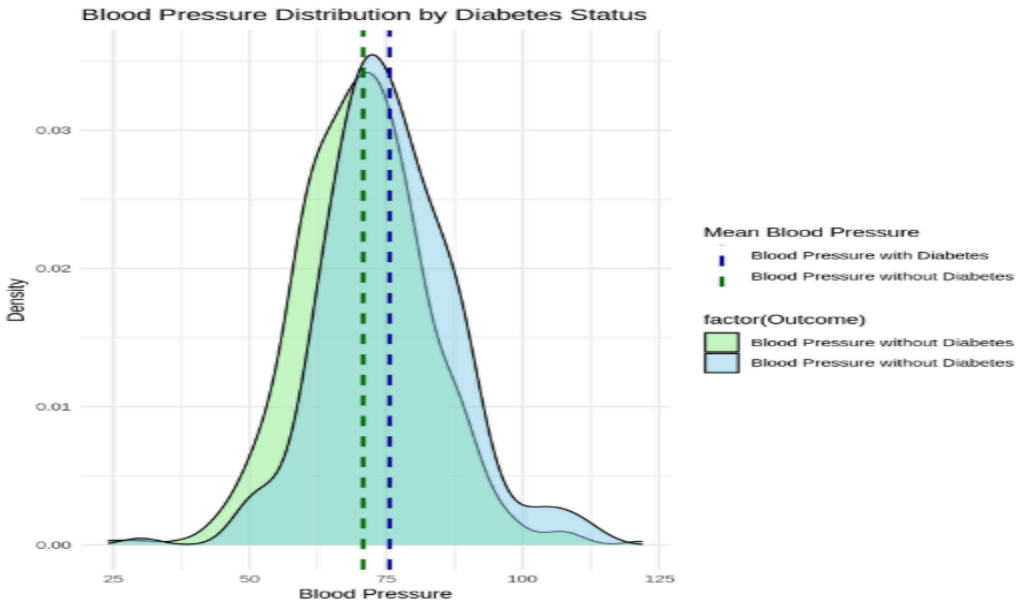
The average glucose level in people with diabetes is higher than the average glucose level in people without diabetes with a difference of 30 that



Age Distribution by Diabetes Status

Interpretation:

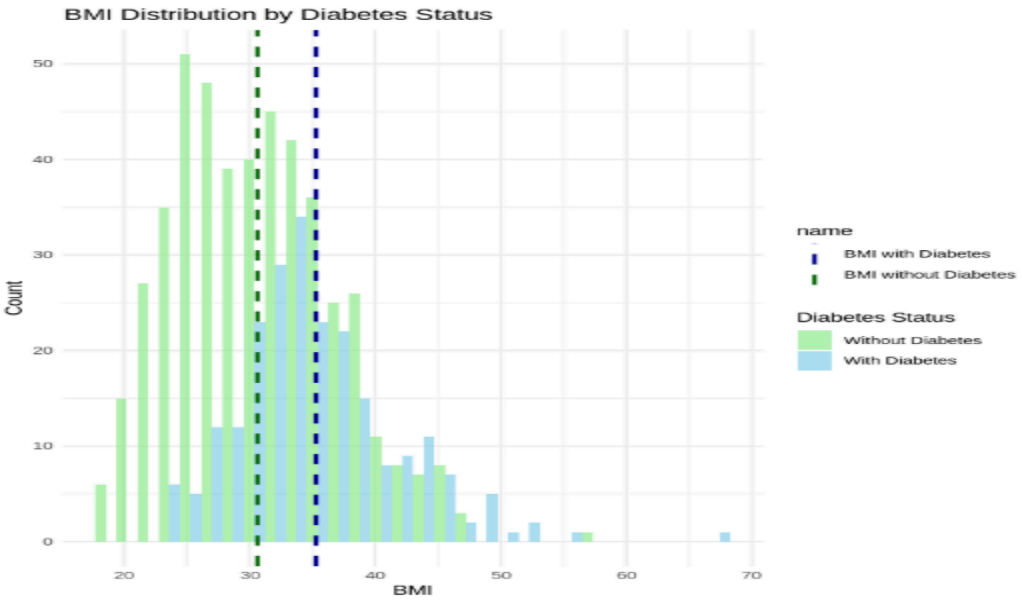
The violin graph shows that people ranging from 20 to 30 years mostly do not have diabetes; however, this percentage is lower in ages above 30



Blood Pressure Distribution by Diabetes Status

Interpretation:

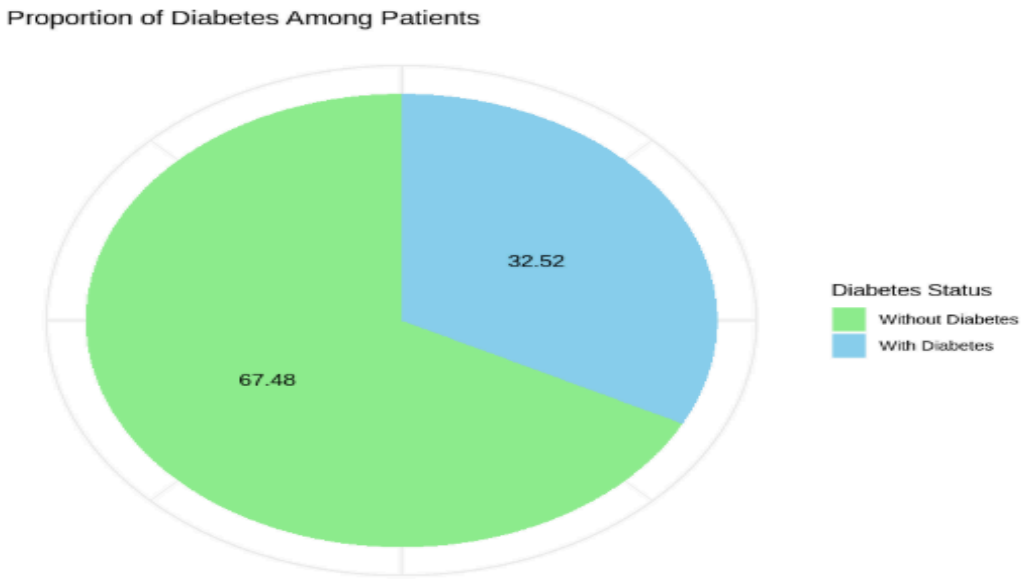
The Density plot shows the distribution of the blood pressure without diabetes and the blood pressure with diabetes showing that they are nearly close in distribution. This means that the mean and variance are very close in the two types with diabetes and without. It also highlights the mean in the plot



BMI Distribution by Diabetes Status

Interpretation:

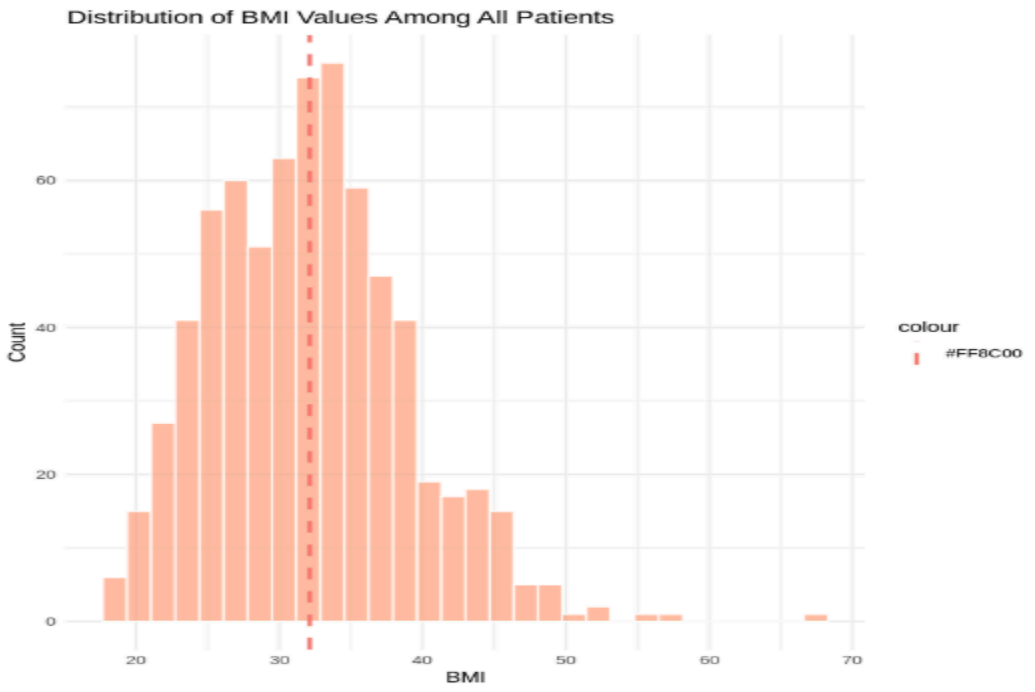
The histogram shows that the number of patients with a low BMI is higher among those without diabetes, while the number of patients with a high BMI is greater among those with diabetes. Interestingly, there are no high BMI values for individuals without diabetes.



Rate of Diabetes Among Patients

Interpretation:

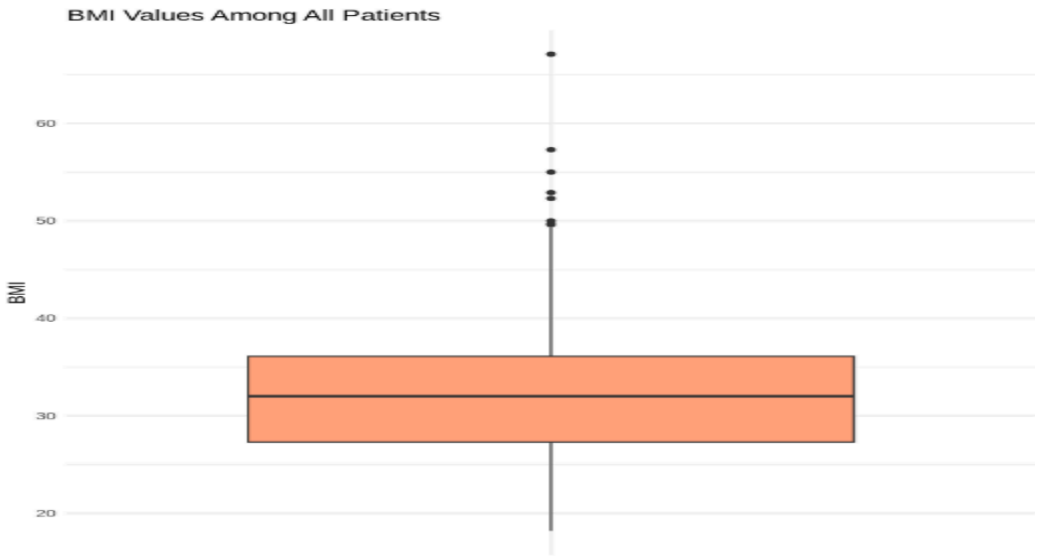
This indicates that approximately one-third of the patients in the dataset are affected by diabetes, while the majority (two-thirds) do not have the condition. The rate of diabetes among patients is therefore 32.52%.



Distribution of BMI Values Among all Patients

Interpretation:

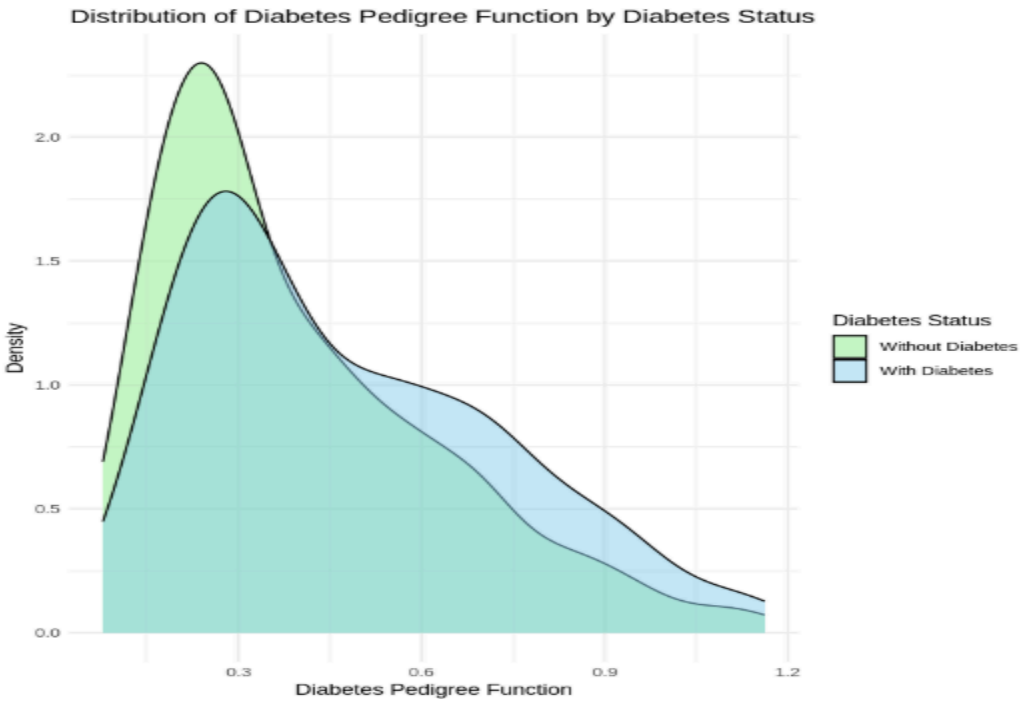
The Distribution of BMI shows that it mostly ranges from 25 to 40 with a mean of nearly 32



BMI Values among all Patients

Interpretation:

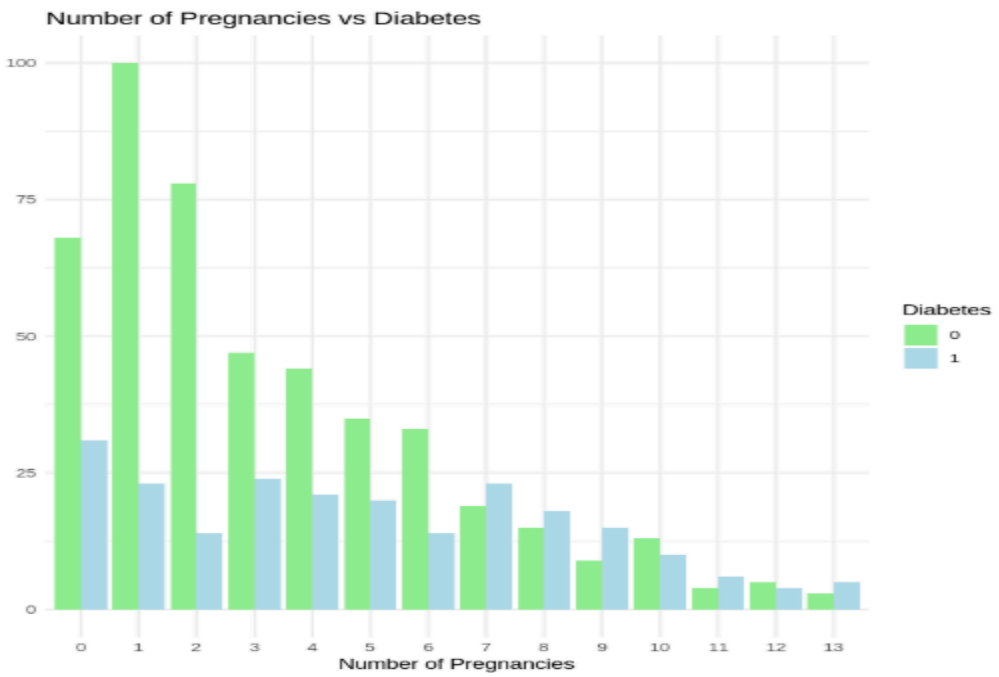
Boxplot shows that the min value of the patient's BMI in the data is nearly 27 and the max is 36 except for some outliers with BMI more than that.



Distribution of Diabetes Pedigree Function by Diabetes Status

Interpretation:

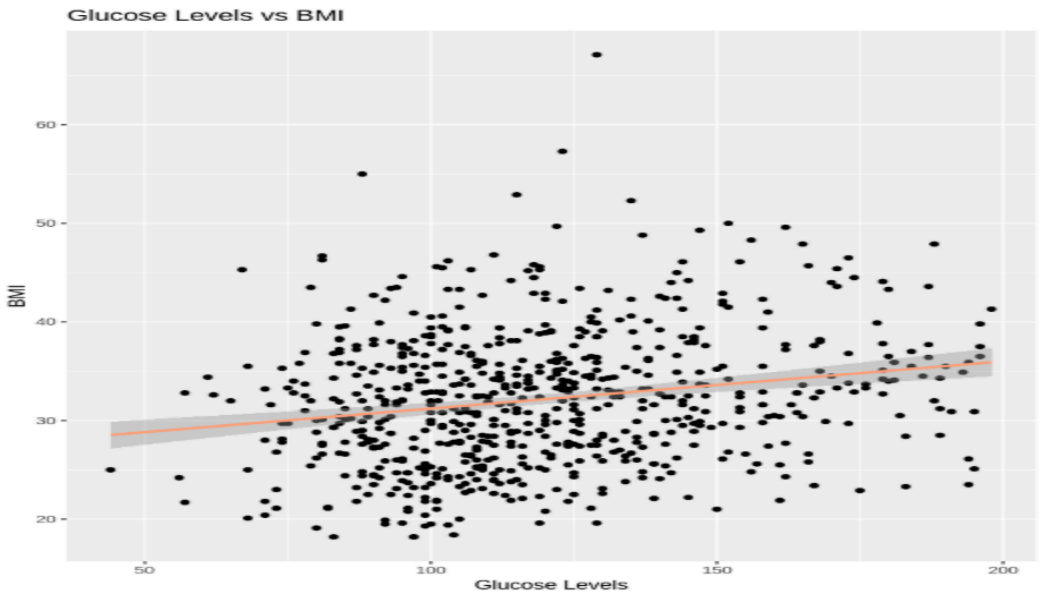
The density plot indicates that patients without diabetes have a higher density when the diabetes degree function is low. However, as the diabetes degree function increases, the density of patients without diabetes decreases, while the density of patients with diabetes increases.



Number of Pregnancies vs Diabetes

Interpretation:

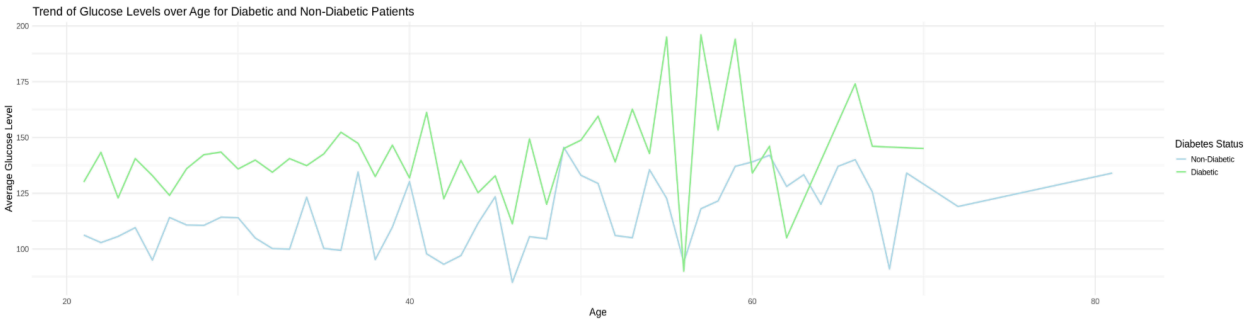
The plot shows that as the number of pregnancy increase the difference proportion between patients with diabetes and without decrease for example 1 pregnancy show difference between of nearly 75% between patients with diabetes and without ;however, at 6 pregnancy it shows nearly difference of 10%.



Glucose Level vs BMI

Interpretation:

There is a weak positive correlation between glucose levels and BMI, as indicated by the slightly upward-sloping trend line.



Trend of Glucose Levels over Age for Diabetic and Non-Diabetic Patients

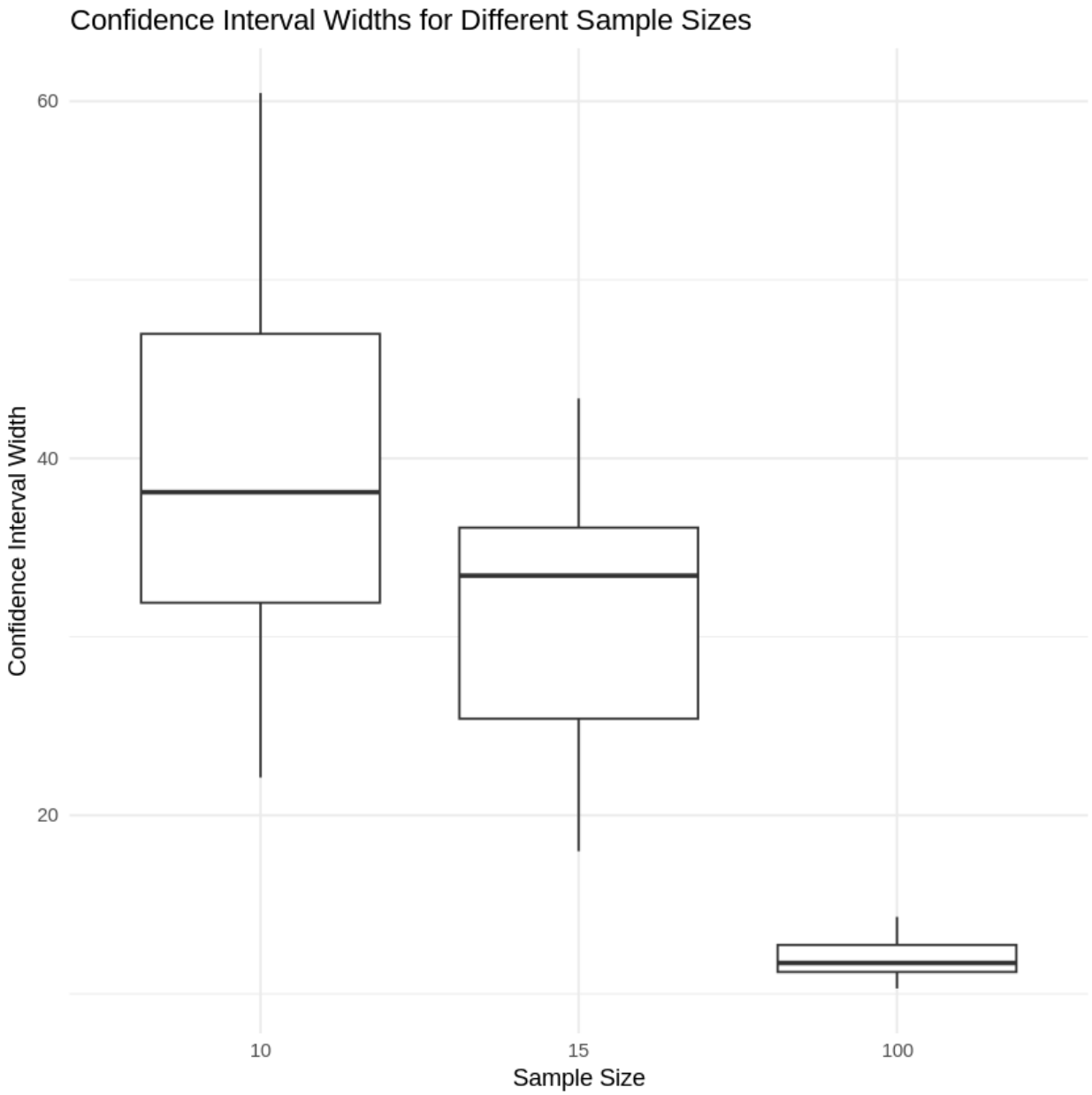
Interpretation:

The trend in the line plot shows that Diabetic patients consistently show higher glucose levels compared to non-diabetic patients and that glucose control becomes more challenging for diabetic patients in their late

50s and early 60s as there is high variances as shown by the chart

- **Hypothesis Testing:**
  - The claim was to see if there is any significant difference in glucose values between the diabetic and nondiabetic patients.
    - We used the t test with significance value 0.05 to test the hypothesis.
    - We measured the P-Value and we found that it is less than the significance value, which leads us to reject the null hypothesis.
    - Statistical Summary:
      - Standard Error: 2.18569
      - CI for the mean: [26.72725, 35.32242]
      - Mean Estimation of diabetic patients: 140.2334
      - Mean Estimation of Non-diabetic patients: 109.2086
    - Conclusion:
      - Since the p-value is less than the significance value, there is a significant difference between glucose levels between diabetic and non-diabetic patients.
  - The claim was to see if there is any significant difference in blood pressure values between the diabetic and nondiabetic patients, having diabetic people less than non diabetic people.
    - As the above claim we used the t test with significance value 0.05 to test our hypothesis.
    - We measured the P-Value and we found that it is more than the significance value, which leads us to accept the null hypothesis.
    - Statistical Summary:
      - Standard Error: 0.9653083
      - CI for the mean: [-∞, 6.447348]
      - Mean Estimation of diabetic patients: 75.62776
      - Mean Estimation of Non-diabetic patients: 70.77151
    - Conclusion:
      - Since the p-value is more than the significance value, there is no evidence to conclude that there is a significant difference having diabetic patients less than non-diabetic patients.

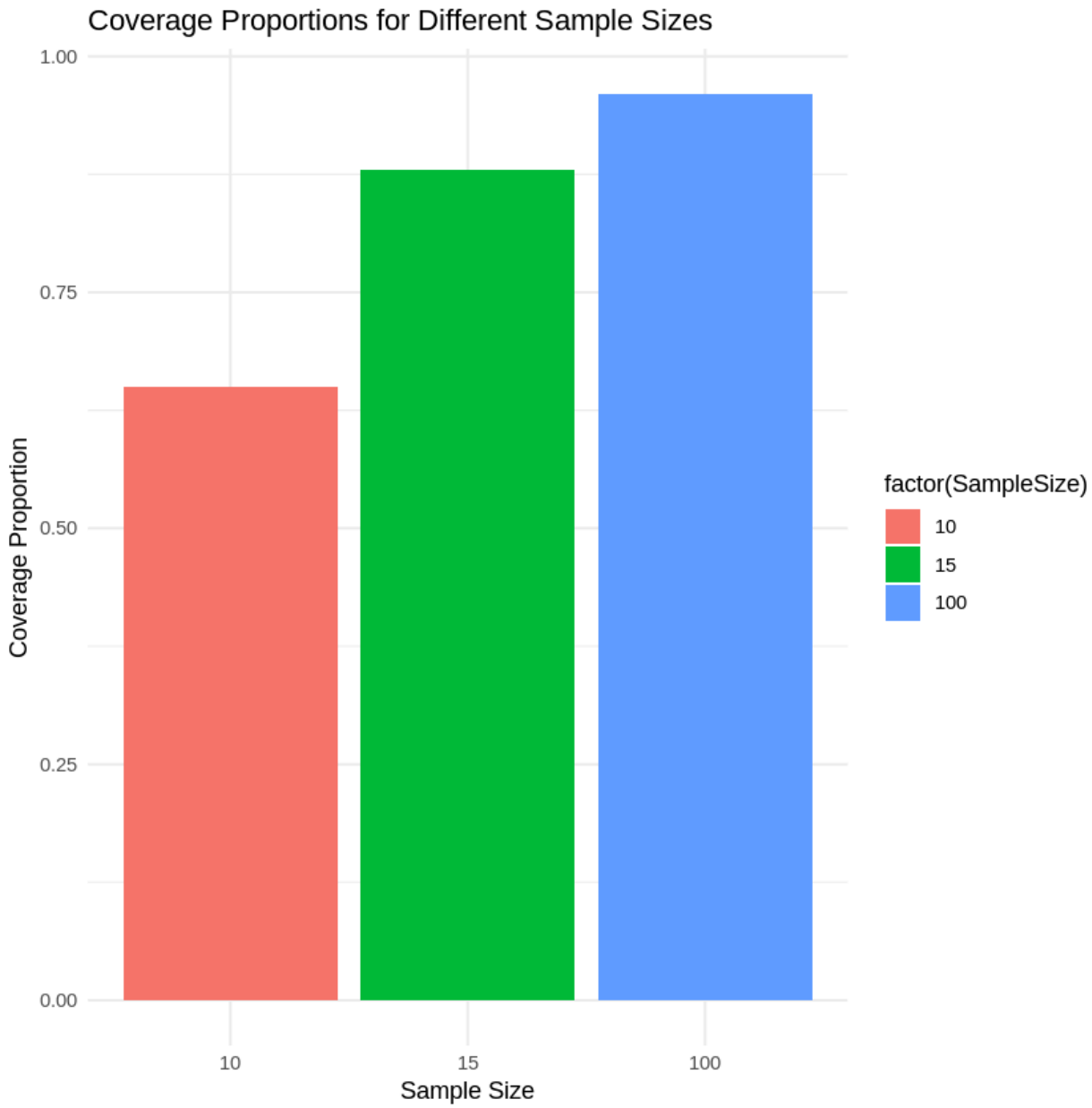
- **Simulation Task:**



Sample Size vs Confidence Interval Width

Interpretation:

From the box plot, it was observed that, with an increase in sample sizes, the width of the CI tends to decrease. It also highlights that larger sample sizes provide a better estimation to the population mean.



Sample Size vs Coverage Proportion

Interpretation:

From the bar graph it was observed that when the sample size

increases the coverage proportions tend to increase to reach the expected 95% coverage. It also highlights that for smaller sample sizes the coverage proportions are less than 95%.

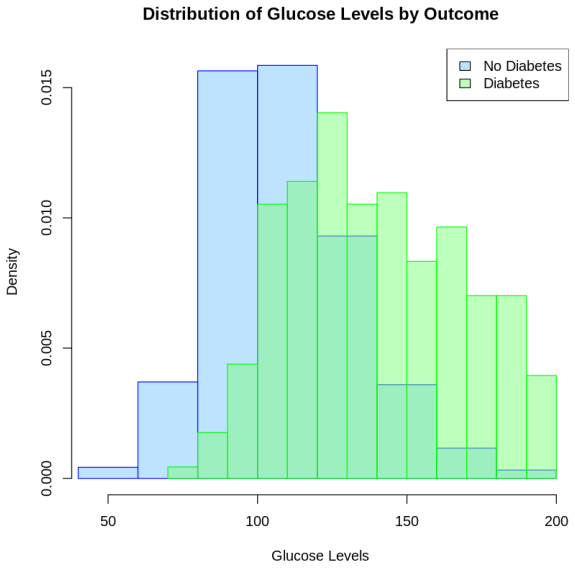
5. Answering Questions

5.1. Using the appropriate statistics and plots to answer the following questions:

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

In Figure 1:

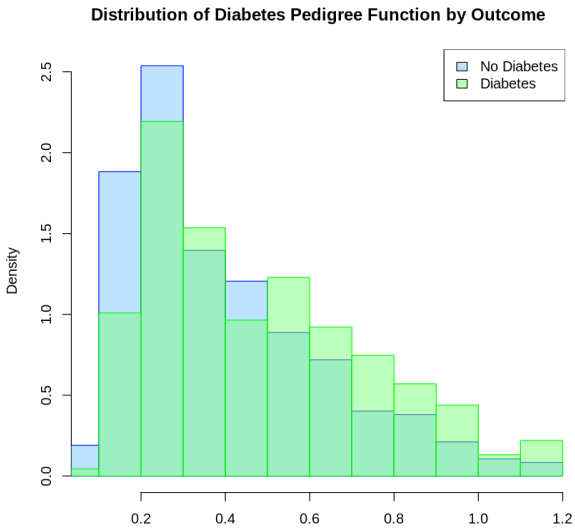
- Individuals without diabetes generally have glucose levels concentrated in the lower range, peaking between 80 and 110. (Normal)
- Individuals with diabetes exhibit a broader distribution, with glucose levels skewed towards higher values, often exceeding 120.



Conclusion: Higher glucose levels are more strongly associated with diabetes This supports the hypothesis that elevated glucose levels increase the likelihood of diabetes.

In Figure2:

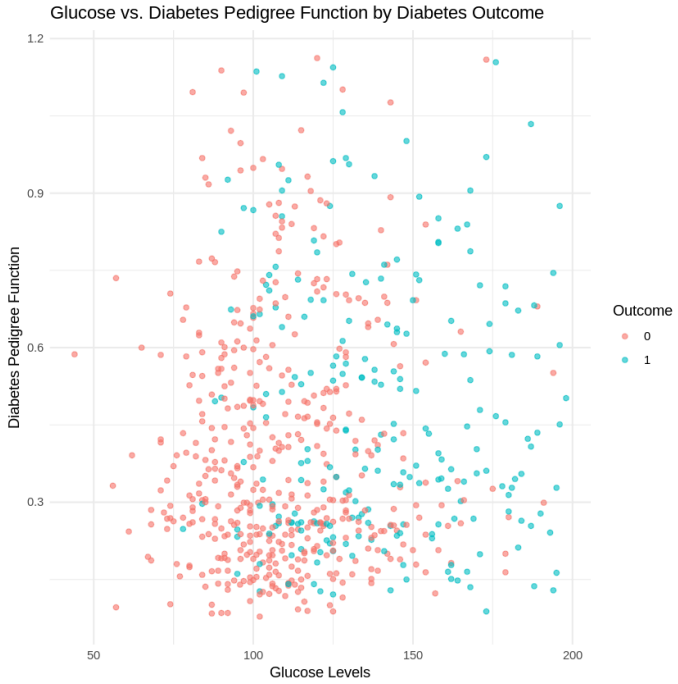
- For individuals without diabetes, the distribution of values is concentrated towards the lower range, with most values below 0.5.
- Individuals with diabetes exhibit a broader spread, with higher



frequencies at values above 0.5, and a tail extending to higher values.

Conclusion: Higher Diabetes Pedigree Function values are associated with a greater likelihood of diabetes.
The histograms suggest that individuals with a stronger family history of diabetes or higher genetic predisposition are more likely to have diabetes.

In this figure, There appears to be a positive association between higher glucose levels and the likelihood of having diabetes. Specifically, individuals with higher glucose levels above 120 are more likely to have diabetes
Conversely, lower glucose levels below 100 are more associated with having diabetes.



Perform t-test to compare Glucose levels between Outcome groups:

- **t-value:** -14.195, indicating a large difference between group means.
- **p-value:** < 2.2e-16, which is highly significant, rejecting the null hypothesis.

Conclusion: There is a significant difference in glucose levels between the two groups, with higher glucose levels in the diabetes group.

Perform t-test to compare DiabetesPedigreeFunction values between Outcome groups

- **t-value:** -4.0092, indicating a noticeable difference between group means.
- **p-value:** 7.23e-05, which is highly significant, rejecting the null hypothesis.

Conclusion: There is a significant difference in Diabetes Pedigree Function values between the two groups, with higher values in the diabetes group.

Preform Pearson's product-moment correlation to get the correlation Between Glucose level and DiabetesPedigreeFunction

Correlation Coefficient (cor = 0.0612052):

- The correlation coefficient between Glucose and Diabetes Pedigree Function is 0.061, indicating a very weak positive correlation.

P-value (p-value = 0.1054):

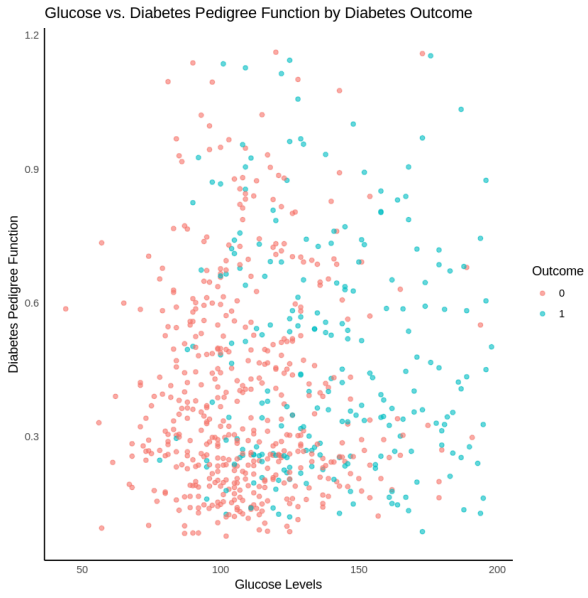
- The p-value is greater than 0.05, indicating that the result is not statistically significant.
- This means we fail to reject the null hypothesis, and there is no strong evidence to suggest a meaningful correlation between Glucose and Diabetes Pedigree Function.

t-Statistic (t = 1.6212) and Degrees of Freedom (df = 699):

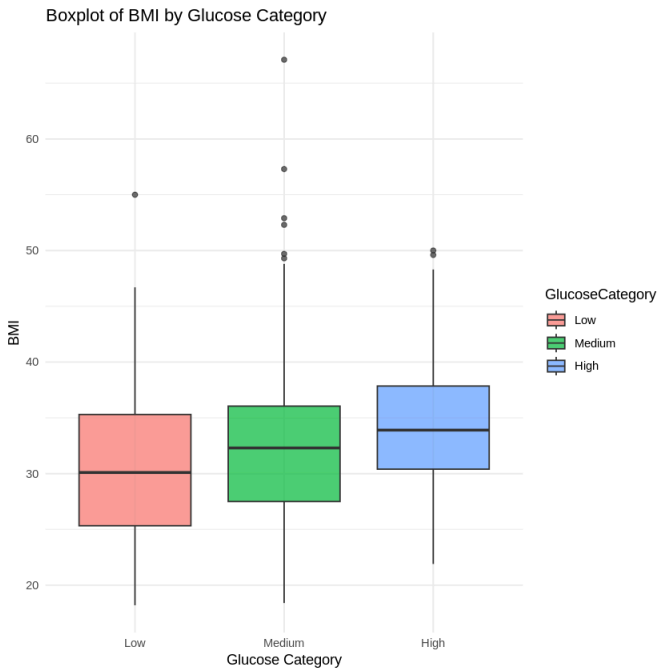
- The t-statistic quantifies the strength of evidence against the null hypothesis. Here, the value is small (1.62), suggesting weak evidence.

5.1.2. Are patients with high glucose concentrations also likely to have higher BMI 5.2.2 values?

In the figure, It appears that there is no strong or clear trend indicating that higher glucose levels are consistently associated with higher BMI values.



In the figure, It appears that there is a slight association between higher glucose levels with higher BMI values.



Preform Pearson's product-moment correlation to get the correlation Between Glucose level and BMI

Correlation Coefficient (cor = 0.2059253):

- This indicates a weak positive correlation, meaning that as Glucose levels increase, BMI tends to increase slightly, but the relationship is not very strong.

P-value (p-value = 3.766e-08):

- This result is statistically significant, so we can confidently reject the null hypothesis that there is no correlation.

t-Statistic (t = 5.5636) and Degrees of Freedom (df = 699):

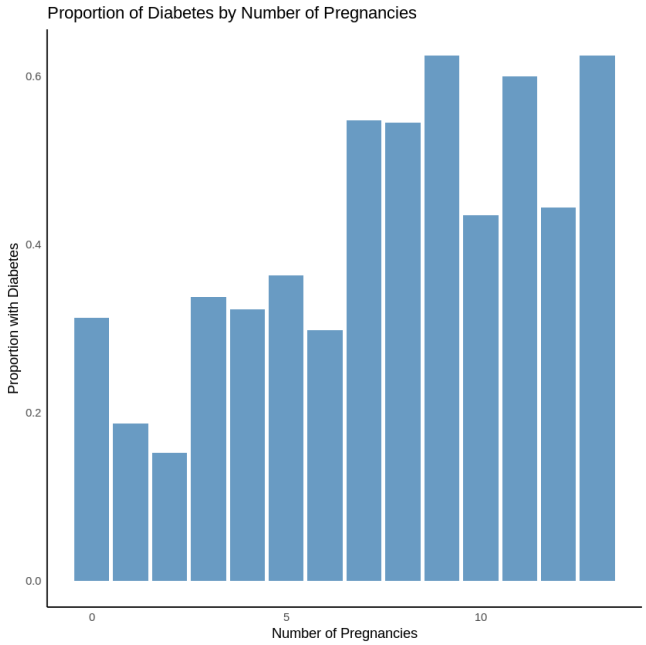
- The t-statistic quantifies the strength of evidence against the null hypothesis, and here it is quite large (5.56).

Conclusion:

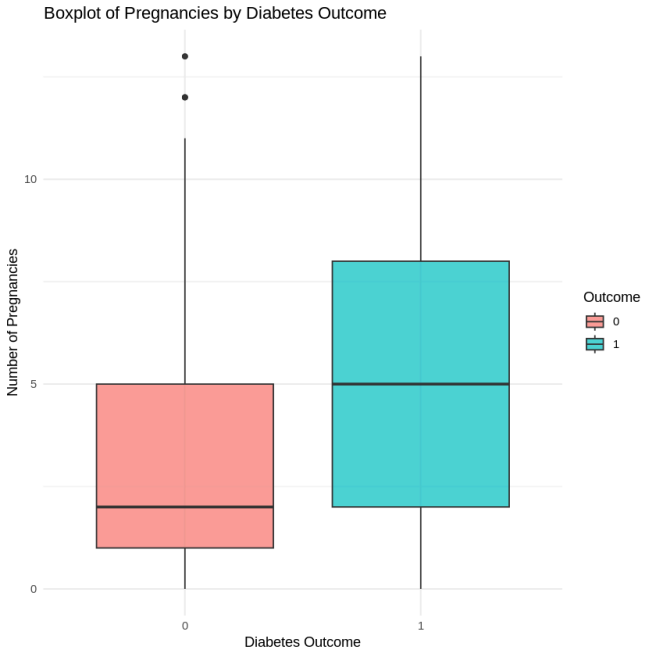
There is a **weak positive correlation** between Glucose levels and BMI.
Higher Glucose levels are associated with slightly higher BMI, but the relationship is **not strong enough to make reliable predictions.**

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

In this figure, The proportion of patients with diabetes tends to increase as the number of pregnancies increases. Patients with higher numbers of pregnancies 10 or more appear to have a higher proportion of diabetes.



It appears that there is a slight association between higher pregnancy numbers with having diabetes.



Preform Pearson's product-moment correlation to get the correlation Between Glucose level and DiabetesPedigreeFunction

Correlation Coefficient (cor = 0.2287):

- The correlation coefficient indicates a weak positive relationship between the number of pregnancies and the Outcome variable.
- A value of 0.2287 means that as the number of pregnancies increases, the likelihood of the Outcome being 1 (or positive) slightly increases, but the relationship is not strong.

Statistical Significance (p-value = 9.081e-10):

- The extremely small p-value (< 0.05) indicates that the correlation is statistically significant, meaning the observed relationship is unlikely to be due to chance.

Test Statistic (t = 6.2102, df = 699):

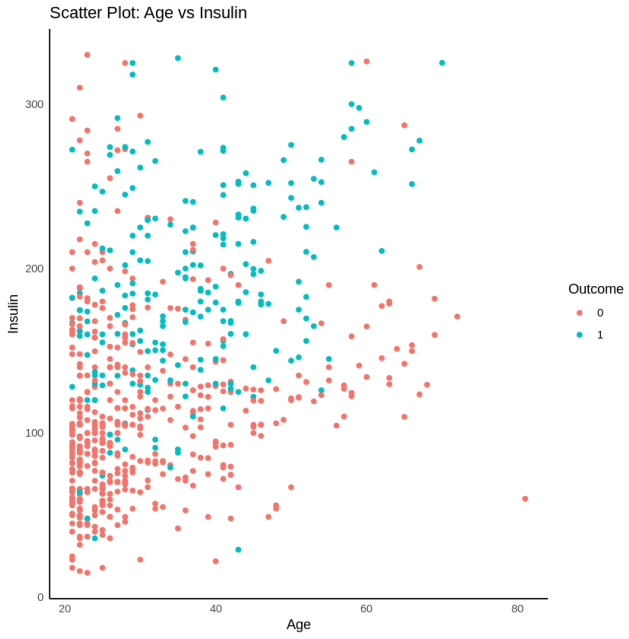
- The high t-value further supports the conclusion that the relationship between pregnancies and Outcome is statistically significant.

Conclusion:

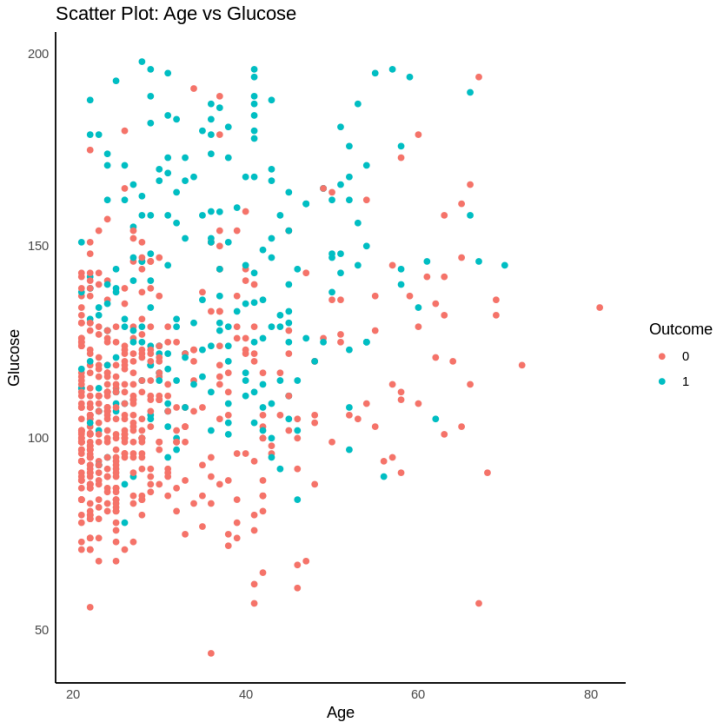
- There is a statistically significant, weak positive correlation between the number of pregnancies and the Outcome variable. While there is a relationship, it is not strong, suggesting that pregnancies have only a **limited impact on predicting the Outcome.**

5.1.4. Are older patients more likely to have higher insulin concentrations and blood glucose levels?

In this figure,There is a very slight positive correlation between age and insulin levels, suggesting that older patients may be somewhat more likely to have higher insulin concentrations.



In this figure, There is even smaller positive correlation between age and blood glucose levels, suggesting that older patients may be somewhat more likely to have higher blood glucose levels.



Perform linear Regression for Insulin and age [1]

The linear regression model is:

$$\textit{Insulin} = 75.2560 + 1.9204 \times \textit{Age}$$

- There is a positive and statistically significant relationship between age and insulin levels. For each additional year of age, insulin levels are expected to increase by 1.92 units.
- The R-squared value of 11.5% suggests that age alone does not explain a large proportion of the variation in insulin levels, indicating that other factors likely contribute significantly to insulin levels.

Perform linear Regression for Glucose and age [1]

The equation for the linear regression model is:

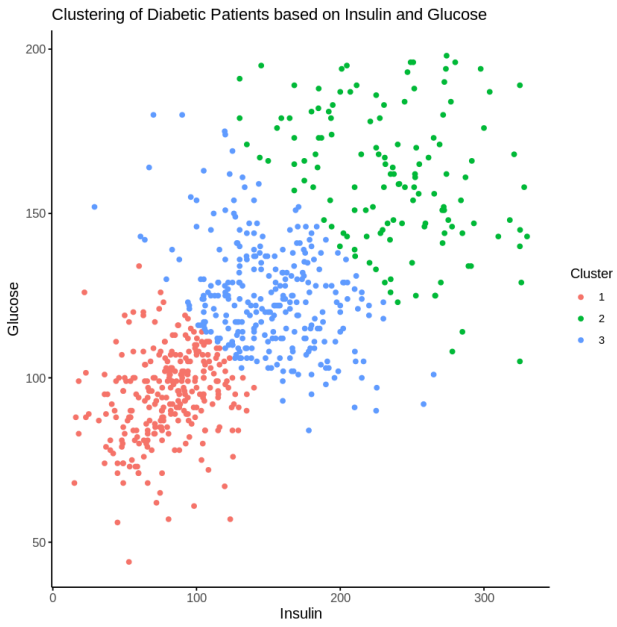
$$\textit{Glucose} = 95.77412 + 0.71061 \times \textit{Age}$$

- There is a significant positive relationship between age and glucose levels, - as indicated by the p-value (< 0.05).
- Weak strength of the relationship: The R-squared value of 0.08208 means that only 8.21% of the variance in glucose levels is explained by age.

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

The clustering analysis based on insulin and glucose levels reveals three distinct groups of diabetic patients [2]:

Cluster 1 (Red): This group is characterized by relatively low insulin and glucose levels. These patients likely represent individuals with milder insulin resistance or better controlled diabetes.



Cluster 2 (Green): This group has high insulin and glucose levels. These individuals may have more severe insulin resistance or poorly controlled diabetes.

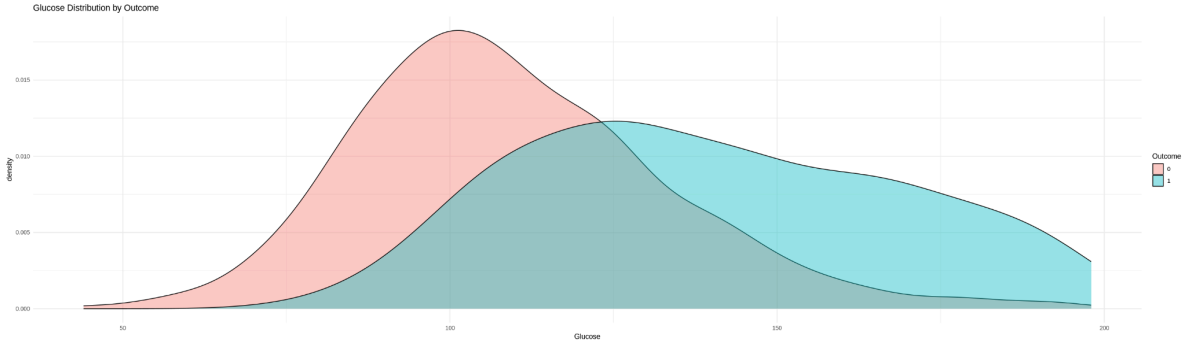
Cluster 3 (Blue): This group falls between clusters 1 and 2, with moderate insulin and glucose levels. These patients may represent those with intermediate-risk or diabetes management outcomes.

Conclusion:

- Older patients are more likely to have higher insulin concentrations and glucose levels, based on the positive correlations identified in the regression models.
- The relationships are weak, meaning age alone does not account for much of the variability in insulin or glucose levels. Other factors play a larger role.

5.2. Come up with 5 more bivariate/multivariate analysis questions and similarly answer each with appropriate visuals and commentary.

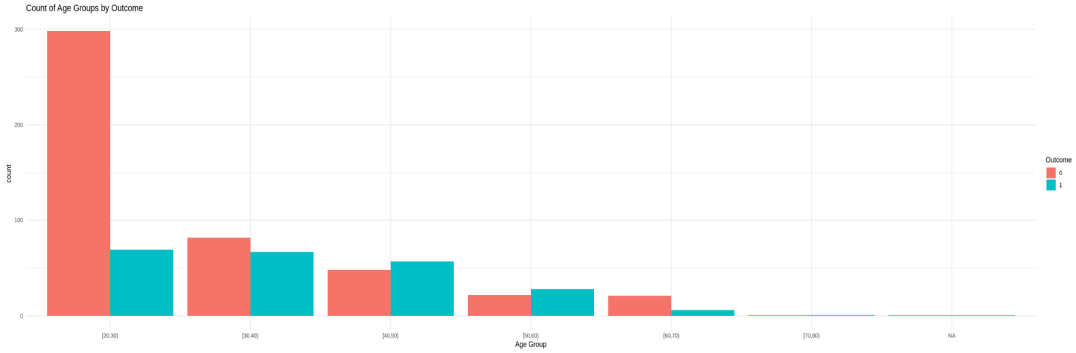
5.2.1 Glucose Distribution by Outcome



The red curve (people with diabetes) skews higher. This indicates glucose levels among diagnosed individuals are noticeably higher.

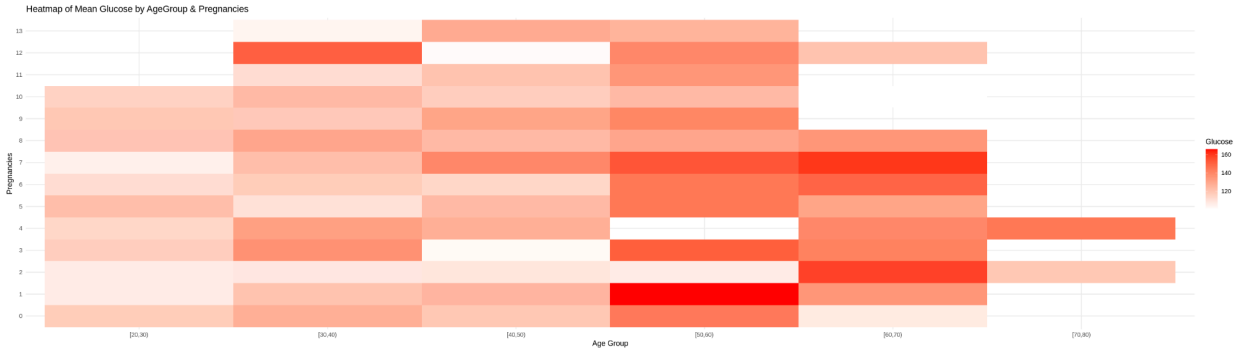
We can see that the red curve is dense at lower Glucose levels, while the blue curve is dense at higher Glucose levels and has a wider spread.

5.2.2 Count of Age Groups by Outcome



This bar chart shows the count of individuals in different age groups, categorized by Outcome. The 20-30 age group has the most significant difference.

5.2.3 Mean Glucose by AgeGroup & Pregnancies

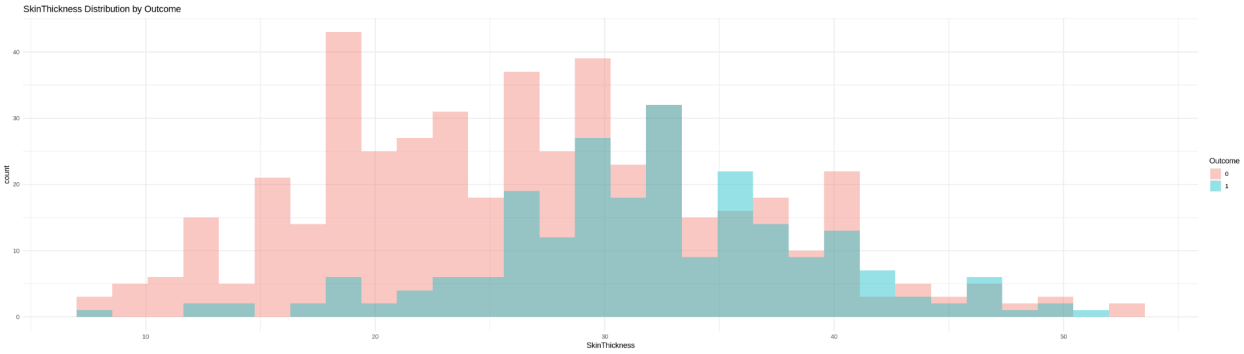


This heatmap visualizes the average Glucose levels across different age groups and pregnancy counts.

We notice that:

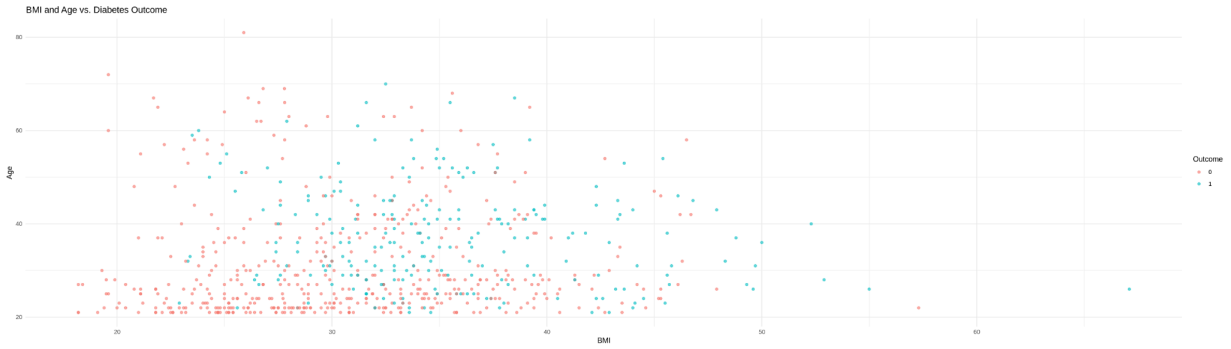
- The highest mean Glucose levels are observed in the 50-60 age group with 1 pregnancy and the 30-40 age group with 12 pregnancies.
- Generally, Glucose levels appear to increase with age for a given number of pregnancies.

5.2.4 SkinThickness Distribution by Outcome



This histogram shows the distribution of skin thickness for people with diabetes and people without. The graph shows that people with diabetes tend to have higher skin thickness values.

5.2.5 BMI and Age vs. Diabetes Outcome

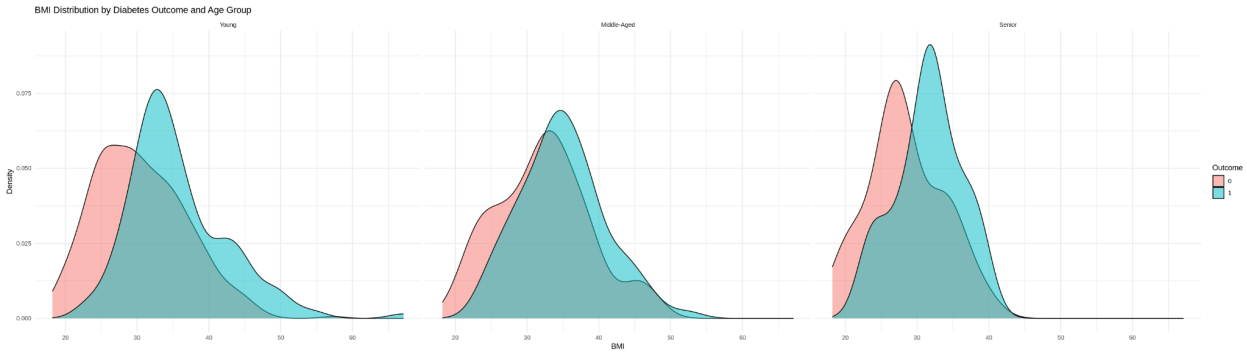


Individuals with diabetes generally have higher BMI and are older than those without. This suggests higher BMI and older age may be associated with increased diabetes risk.

5.2.6 Glucose vs. Insulin

This scatter plot shows the relationship between Glucose and Insulin, with trend lines. People with diabetes show a smoother positive relationship between Glucose and Insulin compared to people without. This suggests that for a given increase in Glucose, people without diabetes tend to have a larger increase in Insulin.

5.2.7 BMI Distribution by Diabetes Outcome and Age Group



The BMI distribution differs between outcome groups. Individuals with diabetes tend to have higher BMI in these age groups. In the "Senior" group, the difference is less pronounced.

Conclusion

- Summarize the findings and insights derived from the analysis.

  Univariate Analysis:

  - The average of Glucose for Diabetic patients is larger than the average for non-diabetic patients; however, it is just slightly higher at which the diabetic shows an average of 140 and the non-diabetic shows an average of 110.
  - The distribution of age shows that as the age increases the diabetic patients increase.
  - Distribution of Blood Pressure for both diabetic and non-diabetic patients is close to each other and the average for both is very close as the diabetic has average blood pressure of 75 and non-diabetic has average blood pressure of 70.
  - BMI for diabetic patients is higher than the non-diabetic patients however the diabetic shows higher variance due to some outliers
  - Only one-third of the data patients are affected by diabetes.
  - The average BMI for all patients is nearly 32 which we can say it is less than BMI for diabetic patients which was 35 and bigger than BMI for non-diabetic patients which was 30

  Bivariate Analysis:

    - As the count of pregnancy increase the percentage of becoming diabetic or remain non-diabetic is nearly being the same
    - There is a weak correlation between glucose level and BMI
    - Glucose control becomes more challenging for diabetic patients in their late 50s and early 60s as there is high variances as shown by the chart
- Suggest future directions or potential improvements.

Future Improvements

We can predict the risk of diabetes before it happens using techniques like regression or clustering. Regression helps us estimate the chances of diabetes based on health data, while clustering groups of people with similar health patterns to find those at risk early.

References

[1] STATS4STEM, "Chapter 11: Inference for Linear Regression," STATS4STEM, (n.d.). [Online]. Available: https://www.stats4stem.org/inference-linear-regression. [Accessed: 24-Dec-2024].
[2] Pina, A., Macedo, M. P., & Henriques, R. (2019). Clustering clinical data in R. Methods in Molecular Biology, 309–343. https://doi.org/10.1007/978-1-4939-9744-2_14