DIABETES ANALYSIS REPORT

Introduction

People around the world suffer from diabetes, a state of chronic health that has negative effects on them. Knowing factors that contribute to diabetes can help in early detection, prevention, and control of the disease. This report reflects such insight based on the dataset that contains several health-related measures such as glucose, BMI, and pregnancy, among many others.

This analysis aims to identify patterns and relationships from the dataset that show variation between individuals suffering from diabetes and not suffering from diabetes. These findings will result in a better strategy for healthcare providing .

Dataset Overview

The dataset provides detailed measurements for medical diagnosis aimed directly at predicting the on-set of diabetes based on some health parameters. It involves 768 records from female patients, each characterized by 8 health-related attributes. The Outcome variable indicates whether or not the patient is diabetic (1 or 0).

Dataset Columns:

- **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.

Source and Relevance

The data from Kaggle, it is one of most widely used datasets in diabetes research. Because it captured all attributes relating to diabetes, the relevance lies in exploring contributing factors.

Problem Statement

Diabetes is affected not only by lifestyles, environment, and genetics, but by other factors. However, it is more important to identify the most predictive factors and relationships among them. So, the main goal of this project is:

Analysis of the relationship between important health metrics and the occurrence of diabetes.

Identifying trends and patterns between diabetic and non-diabetic individuals.

By these objectives, the analysis provides a better understanding of the problem of diabetes.

Data Preprocessing

Check:

Missing Values,Duplicate Rows which not found

Unlogical Values: replace zero values where the zeros were non-logical (e.g., zero blood pressure).

To address invalid zero values:

- **Mean Replacement**: Applied to columns with low skewness (Glucose, BloodPressure, BMI).
- **Median Replacement**: Applied to columns with high skewness (SkinThickness, Insulin).

The summary statistics shows information regarding distribution and central tendencies of the data.

The histograms were plotted for important columns on the following: Glucose, Blood Pressure, Skin Thickness, Insulin, BMI to visualize the distributions.

Some outliers have been identified by thresholds associated with their logical ranges.

Challenges

Challenges, Limitations, and Assumptions

Challenges Faced

- Handling Missing Data: We conducted checks for missing values across variables like blood pressure and insulin levels. No missing data was found, except for some illogical values (e.g., zeros), which were handled by we replace zero values with either the mean or the median of column, based on the skewness of the data distribution.

Approach:

1. **Mean**:
   - Used for columns with low skewness distribution (Glucose, BloodPressure, and BMI)
2. **Median**:
   - Used for columns with highly skewed distribution to reduce the impact of outliers (SkinThickness and Insulin)

- Outliers and Skewed Data: Extreme values were identified in variables like glucose and insulin. We adjusted these extreme values (such as outliers) through assumptions, ensuring that the analysis remained accurate.

- Multicollinearity:Some predictors, such as blood pressure, glucose, and insulin, were highly related, making it hard to see their separate effects.

Limitations in Data or Methodology

- Data Representation:The dataset might not show all types of people, so results might not work for everyone.

- Causality vs. Correlation: We can see relationships, but we can’t say for sure if one thing causes another.

- Sample Size: The number of data points may not be enough to trust the results completely.

Assumptions

- Data Accuracy: We believed the data was correct and came from good sources.

- Stability of Variables: We thought the relationships between variables didn’t change during the study.

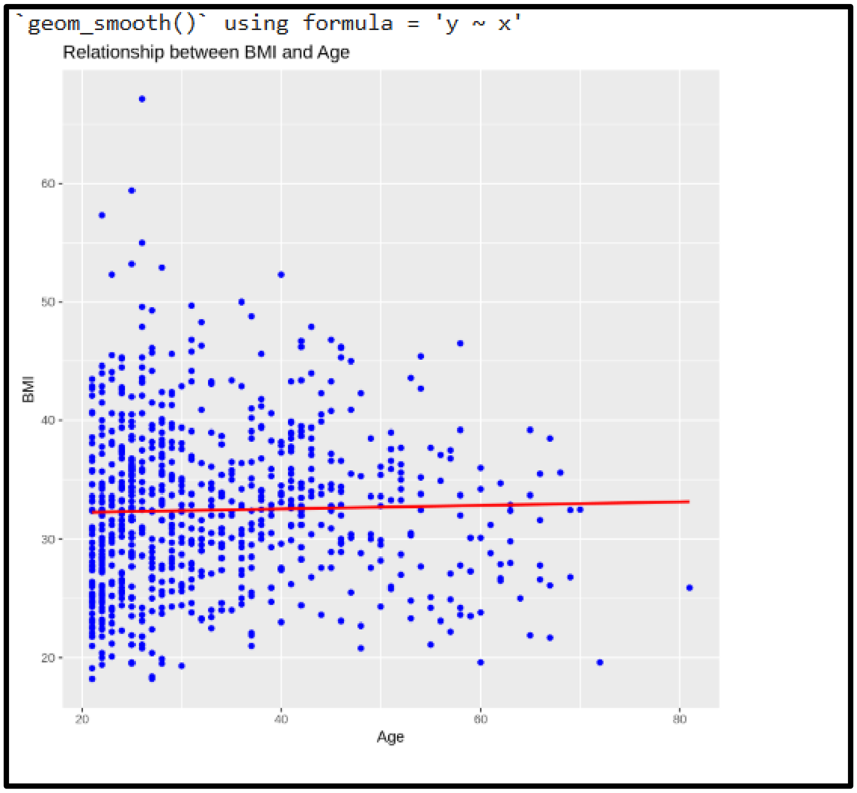
- External Factors: We assumed there were no big outside factors changing the results.

-Normality of Variables: We assumed some variables followed a normal distribution for some hypothesis tests and confidence interval calculation

Results and Visualizations

1. **BMI and Age:** In this analysis, the relationship between Body Mass Index (BMI) and age was examined. shows a weak relationship where there is a slight increase in BMI with age. The red regression line indicates a minor increase in BMI as age increases.

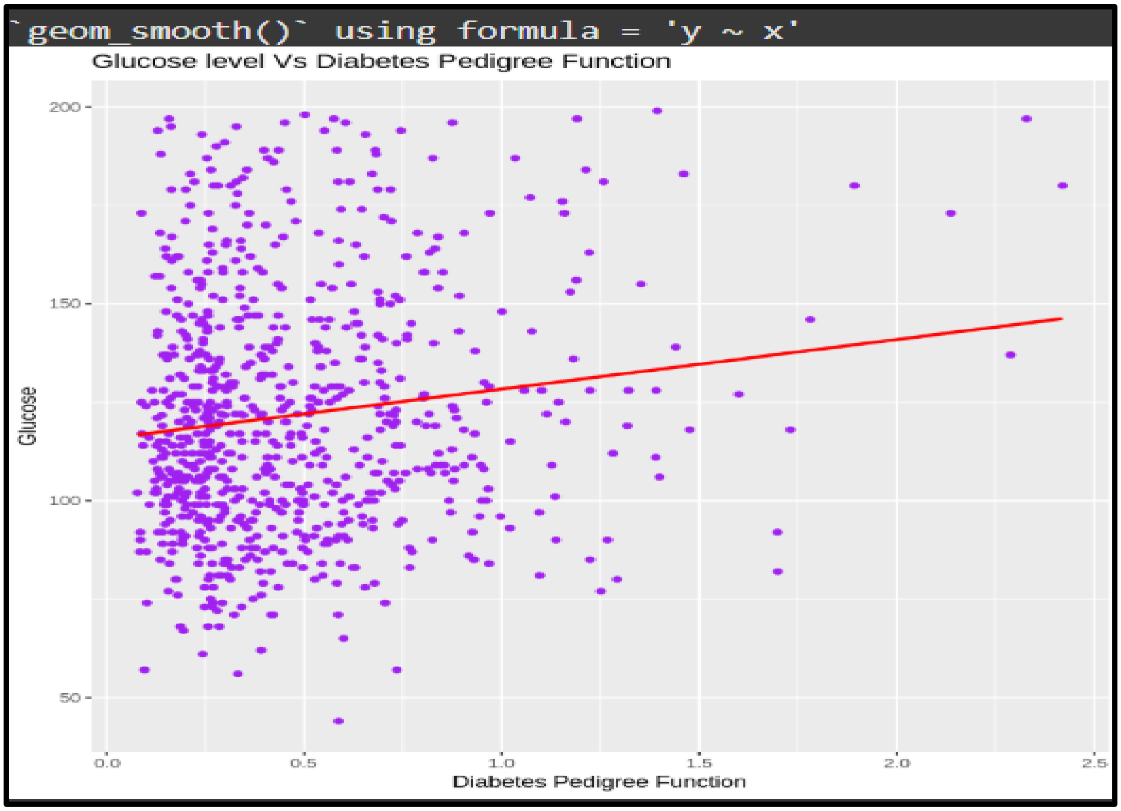
   ○ **Interpretation:** *There is a weak relationship between age and BMI where increasing age may cause an increase in BMI, but this relationship is not strong enough to confirm that age is the main cause of the increase in BMI, as the red regression line indicates that the increase in BMI is slight with increasing age.*



```
`geom_smooth()` using formula = 'y ~ x'
```

2. Effect of Family History (Diabetes Pedigree Function) on Glucose Levels: This analysis explored the effect of family history (diabetes pedigree function) on glucose levels. While there is a slight increase in glucose levels, the wide spread around the regression line suggests that other factors are influencing glucose levels as well.

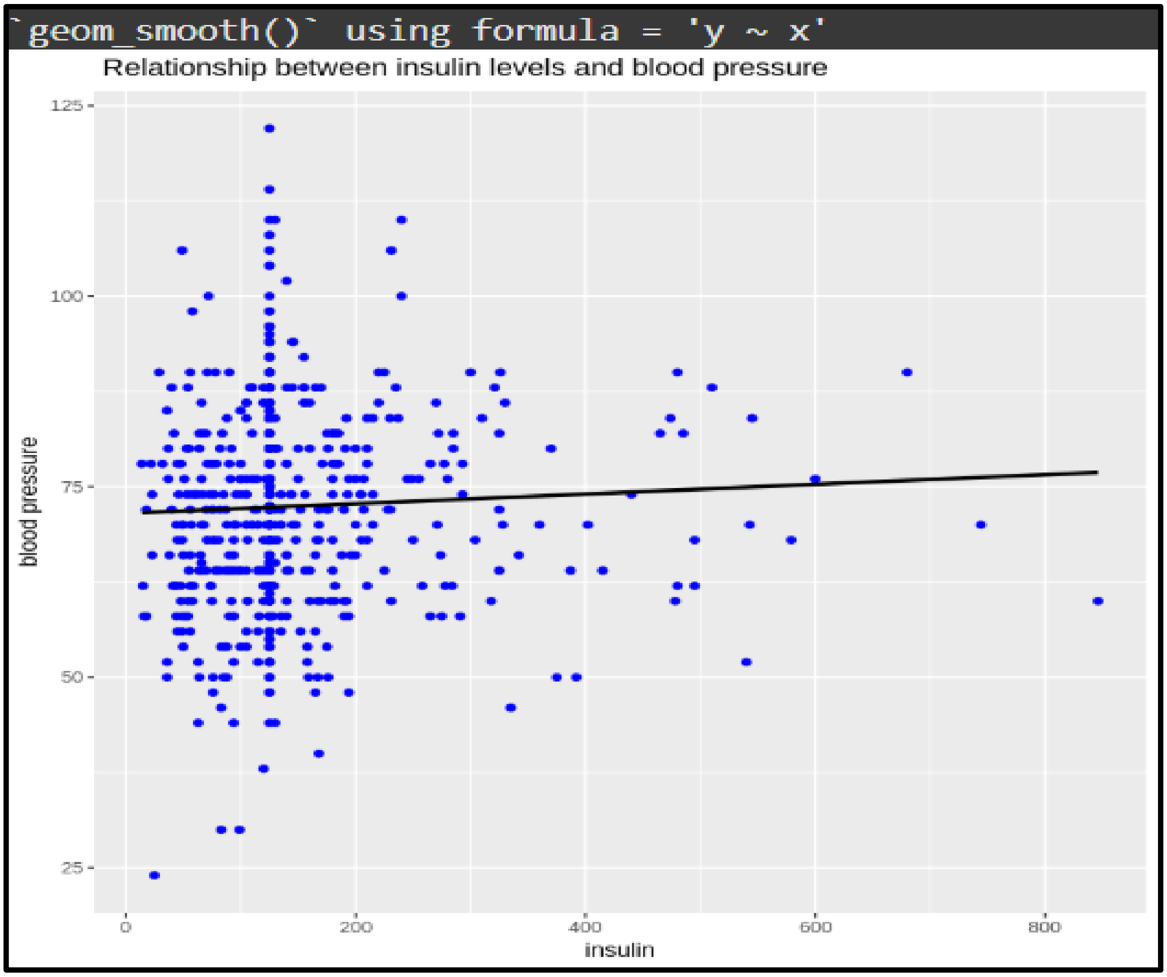
- **Interpretation:** *Family history can be a slight factor in glucose levels, and although there is an increase in glucose level, there is a wide spread around the red regression line, and this indicates that there are other factors influencing besides family history, so this relationship is relatively small*.



```
`geom_smooth()` using formula = 'y ~ x'
```

3.**Relationship between Insulin Levels and Blood Pressure:** This analysis showed a slight positive relationship between insulin levels and blood pressure. As insulin levels increase, blood pressure tends to increase slightly.

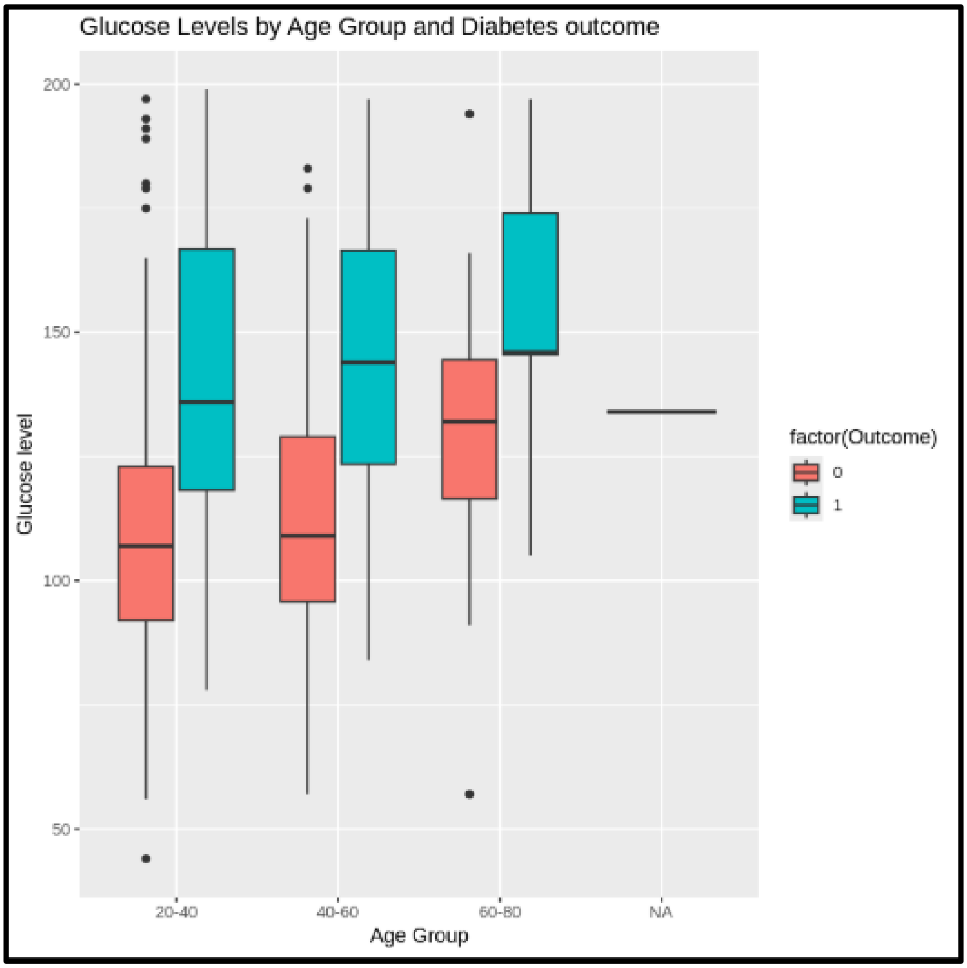
- **Interpretation:** *There is a slight positive relationship between the level of insulin and blood pressure, with the blood pressure tends to increase slightly.*



```
`geom_smooth()` using formula = 'y ~ x'
```

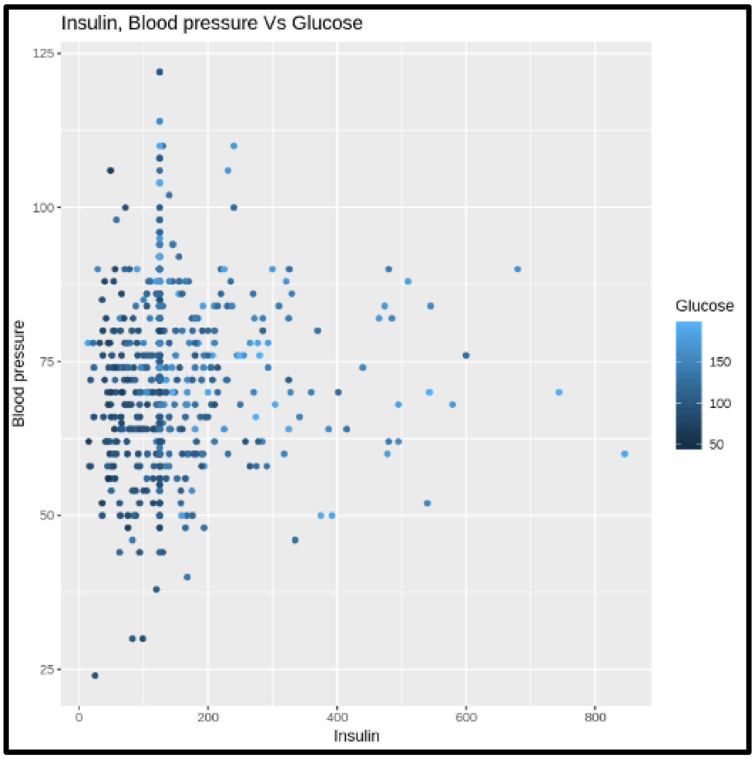
4.Glucose Levels by Age Group and Diabetes Outcome: The prevalence of high glucose levels increases with age, particularly among individuals with diabetes. Therefore, age and diabetes are key factors in determining blood glucose levels.

- **Interpretation:** Glucose levels tend to increase with age, especially among diabetic individuals.



5.Insulin, Blood Pressure, and Glucose Levels: This analysis examined the relationship between insulin levels, blood pressure, and glucose levels. Higher glucose levels were found to be associated with greater variation between insulin and blood pressure levels.

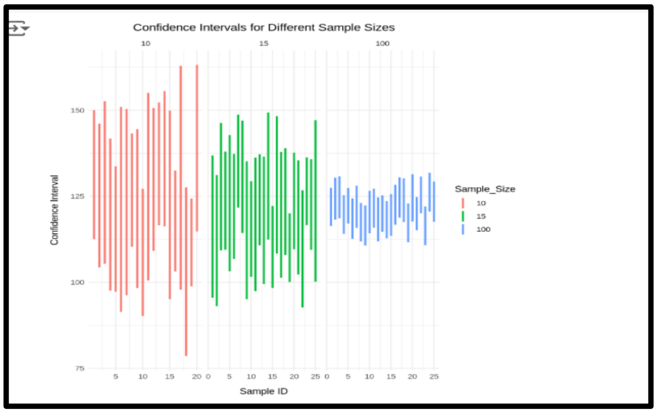
- **Interpretation:** High levels of glucose appear to be associated with a greater disparity between insulin and blood pressure, so glucose plays an important role in the diversity in blood pressure and insulin levels.



Other Visualizations:

confidence intervals:

illustrates confidence intervals for three different sample sizes: 10 (red), 15 (green), and 100 (blue). As the sample size increases, the confidence intervals become narrower, indicating more precise estimates of the mean.



∘

Hypothesis Testing

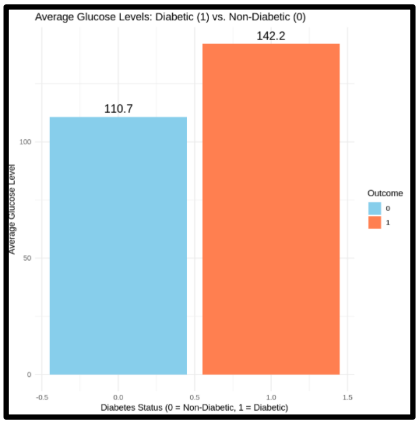
:

Claim: "There is a significant difference in glucose levels between diabetic and non-diabetic patients."
Null Hypothesis (H0) : mean glucose level of diabetic == mean glucose level of non-diabetic

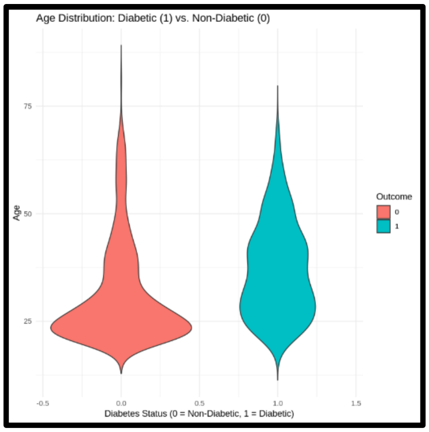
Alternatiive Hypothesis (Ha) : mean glucose level of diabetic != mean glucose level of non-diabetic
Result:Decision: Reject H0: There is a significant difference in glucose levels between diabetic and non-diabetic patients.
Claim 2 My own claim is : "There is a significant difference in BMI levels between diabetic and non-diabetic patients."Null Hypothesis (H0) : mean BMI level of diabetic == mean BMI of non-diabetic
Alternatiive Hypothesis (Ha) : mean BMI level of diabetic != mean BMI of non-diabetic
Result:Decision (BMI): Reject H0: There is a significant difference in BMI levels

Exploratory Analysis:

1. The average glucose levels among patients with and without diabetes.
Plot:

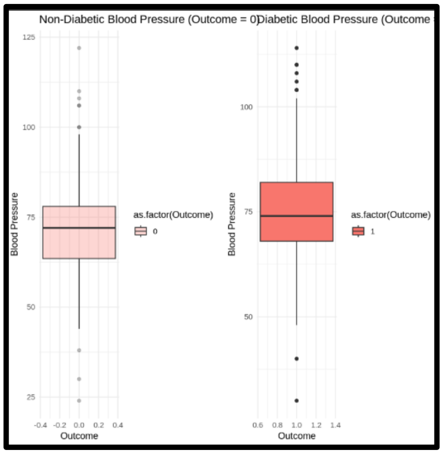


2. The average age of patients with and without diabetes.
Plot:



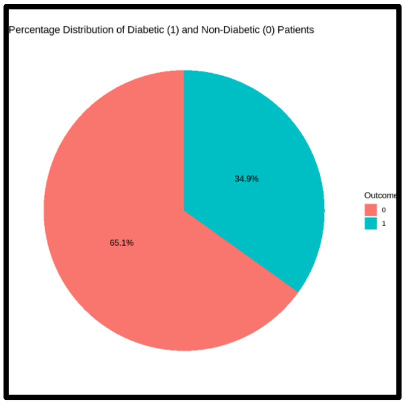
3. The average blood pressure measurements across diabetic and non-diabetic groups.

Plot:



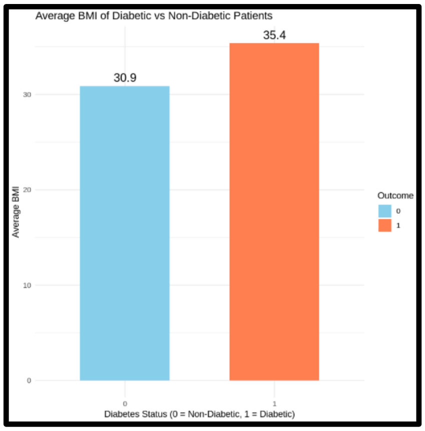
4. The average BMI of diabetic versus non-diabetic patients.

Plot:



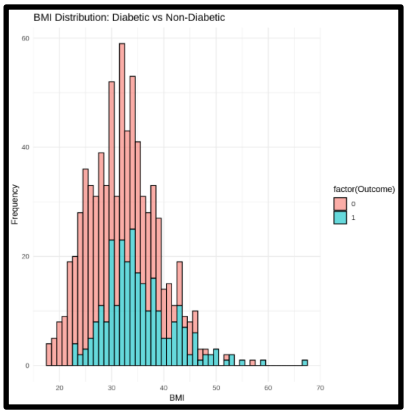
5. The rate of diabetes among patients in the dataset.

Plot:



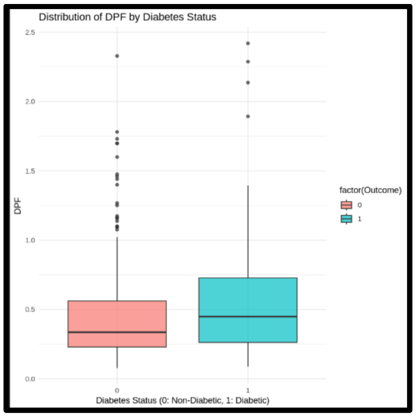
6. The distribution of BMI values among all patients.

Plot:



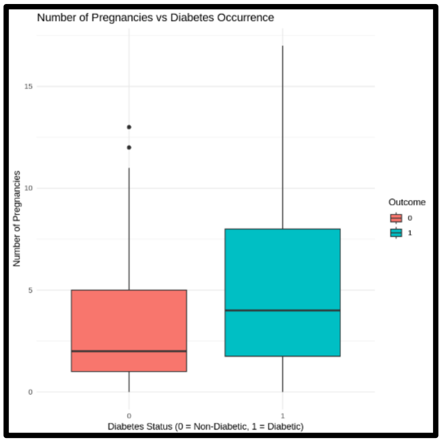
7. The distribution of Diabetes Pedigree Function (DPF) values for diabetic and non-diabetic patients.

Plot:



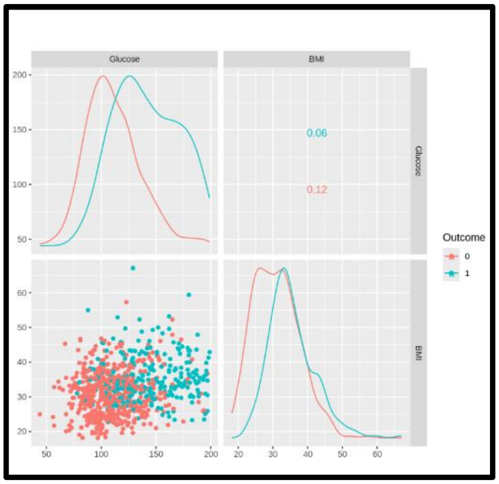
8. The relationship between the number of pregnancies and diabetes occurrence.

Plot:



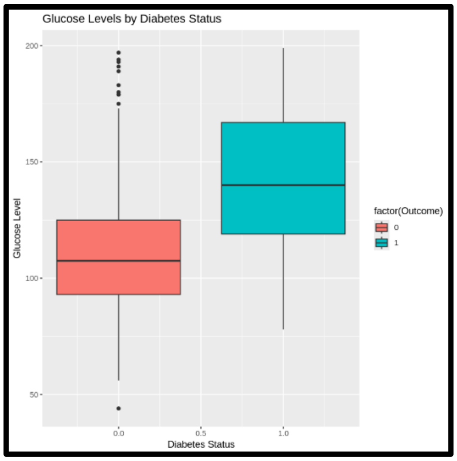
9. The correlation between glucose levels and BMI.

Plot:



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

Plot:

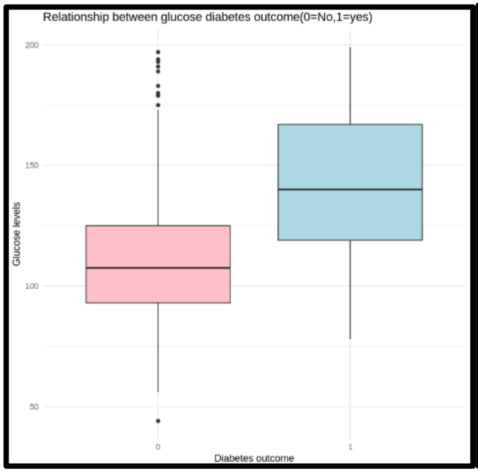


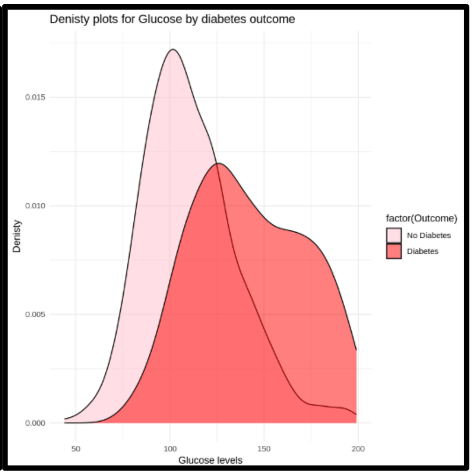
Part2:Answering Questios:

2.1 Use the appropriate statistics and plots to answer the following questions:

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

Plot:

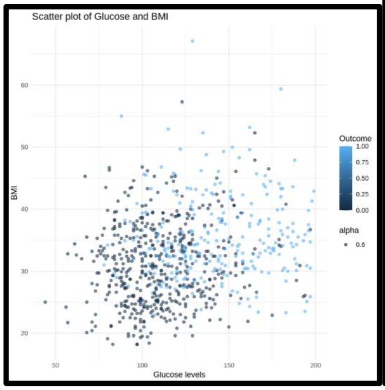


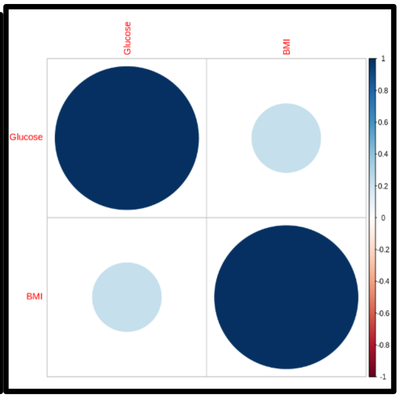


- **Interpretation:** The results from boxplot :
- showing that diabetics had a significantly higher glucose level than non-diabetic patients, as well as the IQR of the diabetic group was clearly high, indicating that most glucose measurements were very high in diabetics.
- The results from Denisty plot:
- The density curve of diabetic patients is concentrated at a higher glucose level than those without diabetes and the overlap between the two distributions is strong evidence that glucose level is an important factor in distinguishing diabetics from non-diabetics.
- The correlation between glucose and outcome is about 0.4665814, which indicates a relationship between high blood glucose level and the presence of diabetes

2. **2. Are patients with high glucose concentrations also likely to have higher BMIvalues?**

   Plots:

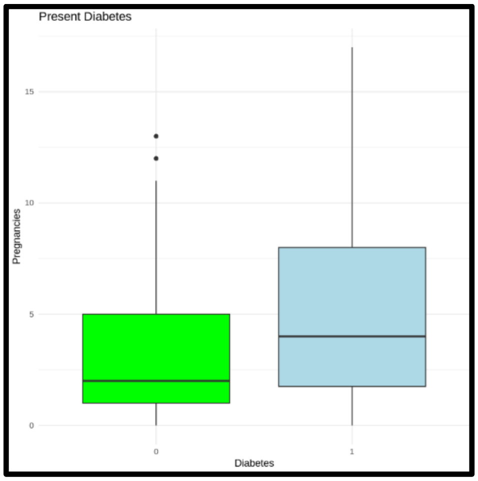




- **Interpretation**: Depending on Pearson's product-moment correlation there is a weak positive correlation with glucose levels and BMI
-
- Based on the scatter plot with the presence of the constant variable which is the level of glucose on the x-axis and the dependent variable for the independent variable on the y-axis there is a positive correlation between weight and glucose level as with increasing the level of glucose in the blood may increase weight.
-
- Based on correlation matrix, the presence of blue circles indicates a positive correlation between weight and BMI and Glucose levels.
-
- Although there is a relation, it is weak, so it must be analyzed based on other factors such as age or genetics in order to clarify the relationship better

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

Plots:

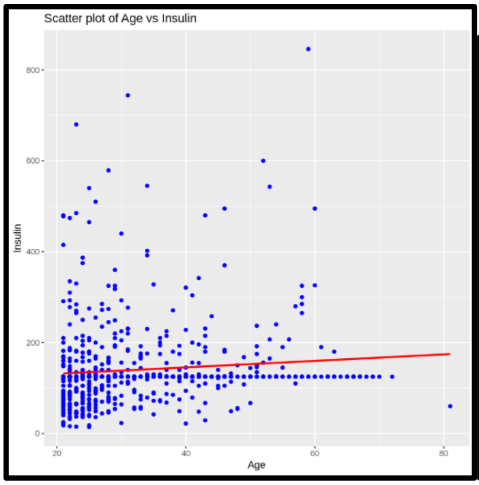


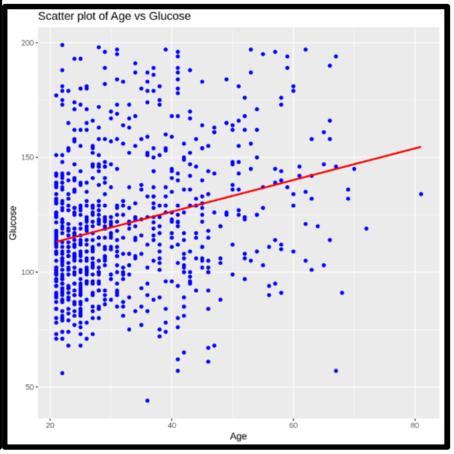
- **Interpretation:**Box plot shows the relationship between the number of pregnancies and the presence of diabetes and the median number of pregnancies in patients with diabetes is higher compared to patients without diabetes

-
- As well as IQR for people with diabetes is wider than those without diabetes, which indicates a greater disparity with increasing the number of pregnancies
-
- So the results show that patients with diabetes tend to have a higher number of pregnancies and there is also a greater disparity in the number of pregnancies among patients with diabetes.

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

Plots:





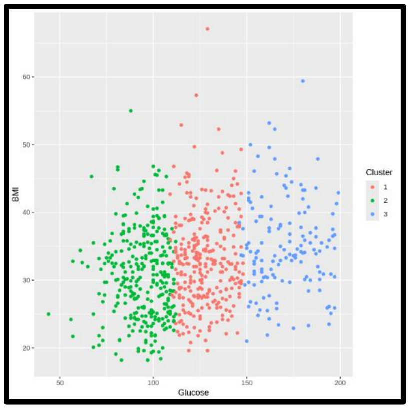
Interpretation:A weak and negative relationship appears between insula levels and age, as the correlation coefficient is -0.04216295, which indicates that there is negative correlation between age and insulin levels.

While a positive relationship appears between age and glucose level, as the correlation coefficient is 0.2635143, and this indicates a weak positive relationship between them, which indicates the possibility of increasing blood glucose levels with age.

So older patients may have a higher glucose level, but age is not a major factor in blood insulin toxicity.

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

Plots:



Interpretation: Based on cluster analysis using the k-means algorithm to divide patients into three groups based on glucose and BMI level

The first group contains patients with low glucose levels as well as low BMI, and these patients have a risk of developing diabetes.

The second group contains patients with a medium glucose level and high BMI, and those have a moderate risk of diabetes.

The third group has a high glucose level as well as BMI, which is evidence that the profile of these patients is high in relation to the risk of diabetes.

Conclusion

- **Summary of Findings:**
  The analysis reveals key insights into the relationships between BMI, glucose levels, age, insulin, blood pressure, and diabetes outcomes. It confirms that diabetes significantly affects glucose levels and that other health factors like BMI, insulin, and age play important roles in understanding the disease. Statistical tests and visualizations further emphasize the importance of sample size in determining the reliability of results.