Zewail City of Science and Technology School of Computational Science and AI

Final Course Project: Analyzing Diabetes Dataset

DSAI 307

Statistical Inference

Supervised by

Eng. Rana Abdelfatah & Eng. Riham Hussein

Prepared by

Seif Eldin 202200973, Seif Mohamed 202201554, Ahmed Mostafa 202201114, Ziad Shaaban 202201093

Table of Contents

1. Introduction 4
2. Dataset Overview 4
    2.1. Columns 4
    2.2. Size 4
    2.3. Inconsistencies 4
    2.4. Data Summary: 5
3. Processing and analysis: 5
    3.1. Checking for Nulls 5
    3.2. Checking Duplicates 5
    3.3. Checking for Outliers 6
4. Exploratory Analysis 7
    4.1. The average glucose levels among patients with and without diabetes. 7
    4.2. The average age of patients with and without diabetes. 7
    4.3. The average blood pressure measurements across diabetic and non-diabetic groups. 8
    4.4. The average BMI of diabetic versus non-diabetic patients 8
    4.5. The rate of diabetes among patients in the dataset 9
    4.6. The distribution of BMI values among all patients 9
    The BMI is normally distributed 9
    4.7. The distribution of Diabetes Pedigree Function (DPF) values for diabetic and non-diabetic patients 10
    4.8. The relationship between the number of pregnancies and diabetes occurrence 10
    4.9. The correlation between glucose levels and BMI 11
    4.10. The trend of glucose levels with age among diabetic and non-diabetic patients 11
5. Questions 12
    5.1. Main Questions 12
    5.1.1. Are higher glucose levels associated with a greater likelihood of diabetes? 12
    5.1.2. Are patients with high glucose concentrations also likely to have higher BMI values?
    5.1.3. Are patients with a higher number of pregnancies at greater risk of developing diabetes? 13
    5.1.4. Are older patients more likely to have higher insulin concentrations and blood glucose levels? 13
    5.1.5. Can you identify common “risk profiles” for diabetic patients based on key metrics (glucose, BMI, age, etc.)? 14
    5.2. Other Questions 15
    5.2.1. Is there a relationship between insulin levels and BMI? 15
    5.2.2. Are patients with higher BMI also more likely to have a higher diabetes pedigree function? 15
    5.2.3. How does the age distribution differ between diabetic and non-diabetic patients? 16
    5.2.4. Do pregnant women with a higher BMI tend to have higher glucose levels? 17
    5.2.5. Can age, glucose, and BMI together predict the likelihood of diabetes? 17

6. Hypothesis 18
6.1. Claim: “There is a significant difference in glucose levels between diabetic and non-diabetic patients.” 18
6.2. Claim: "There is a significant difference in Insulin level between diabetic and non-diabetic patients." 18
7. Simulations **19**
7.1. Take 25 Random Samples of Size 15 from the Dataset 19
7.2. Increase the Sample Size to 100 19
7.3. Take 20 Random Samples of Size 10 from the Dataset 19
8. Conclusion **19**
Key Findings and Analysis 19
Final Thoughts 20

1. Introduction

In this project, we aim to analyze and visualize a public diabetes dataset in females link, preprocessing it to account for outliers and other data inconsistencies. Throughout this project, we highlight several factors contributing to diabetes, most probable age groups, and other insights to answer crucial questions, hypotheses, and simulations. Visualizations are provided to help clarify and illustrate the findings under each question.

2. Dataset Overview

2.1. Columns

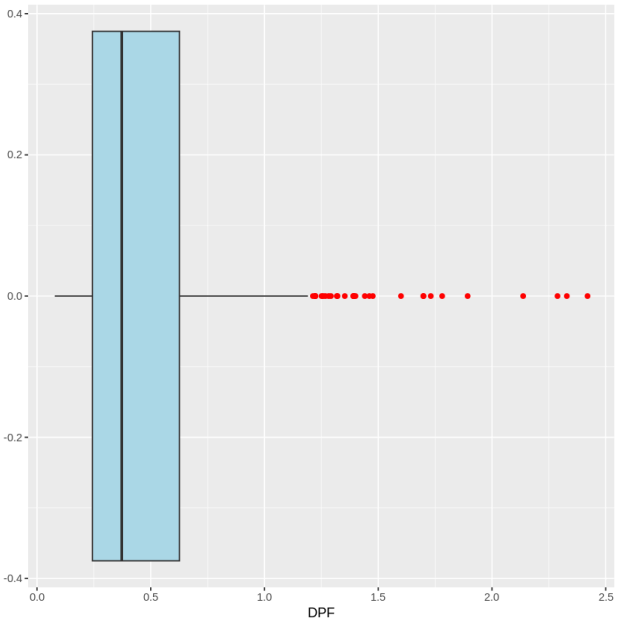
Pregnancies: Number of times the patient has been pregnant.
Glucose: plasma glucose concentration after a 2-hour oral glucose tolerance test.
Blood pressure: Diastolic blood pressure (mm Hg).
SkinThickness: Triceps skinfold thickness (mm).
Insulin: 2-hour serum insulin (mu U/ml).
BMI: Body mass index (weight in kg/(height in m)^2).
DiabetesPedigreeFunction(DPF): A function that represents the patient's diabetes pedigree (i.e., the 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.

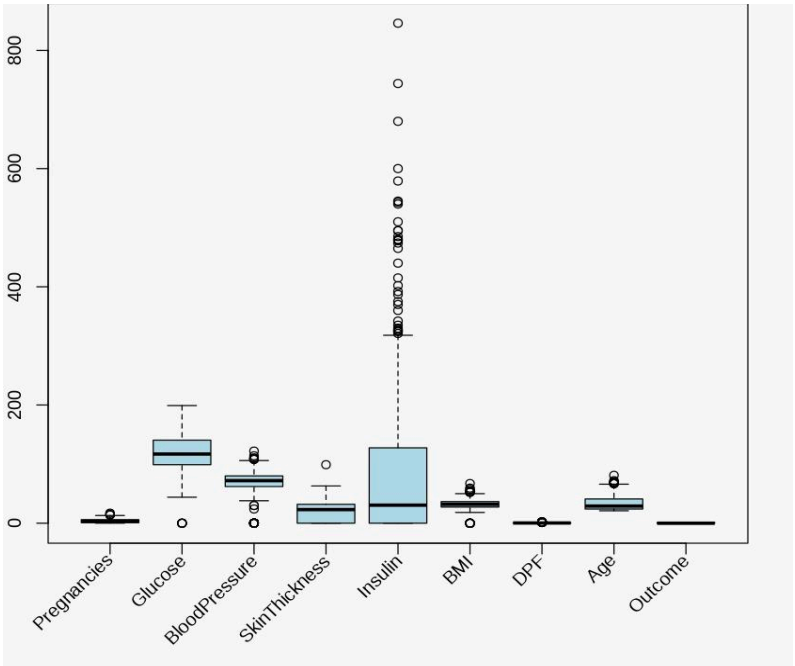
2.2. Size

The dataset has 768 entries spanning 10 columns of real-life female diabetes patients.

2.3. Inconsistencies

The data is free of duplicates and null entries but does contain a significant number of outliers, particularly in the insulin and blood pressure columns.





2.4. Data Summary:

```
Pregnancies        Glucose      BloodPressure    SkinThickness
Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
   Insulin           BMI             DPF              Age
Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
   Outcome
Min.   :0.000
1st Qu.:0.000
Median :0.000
Mean   :0.349
3rd Qu.:1.000
Max.   :1.000
```

3. Processing and analysis:

3.1. Checking for Nulls

We used this method to check how many nulls were in every column

```
colSums(is.na(data))
```

The result was zero in all columns

3.2. Checking Duplicates

We used this method to check if there is any identical rows in the data and delete them

```
duplicate_rows <- nrow(data) - nrow(distinct(data))
if (duplicate_rows > 0) {
  data <- distinct(data)
}
Duplicate_rows
```

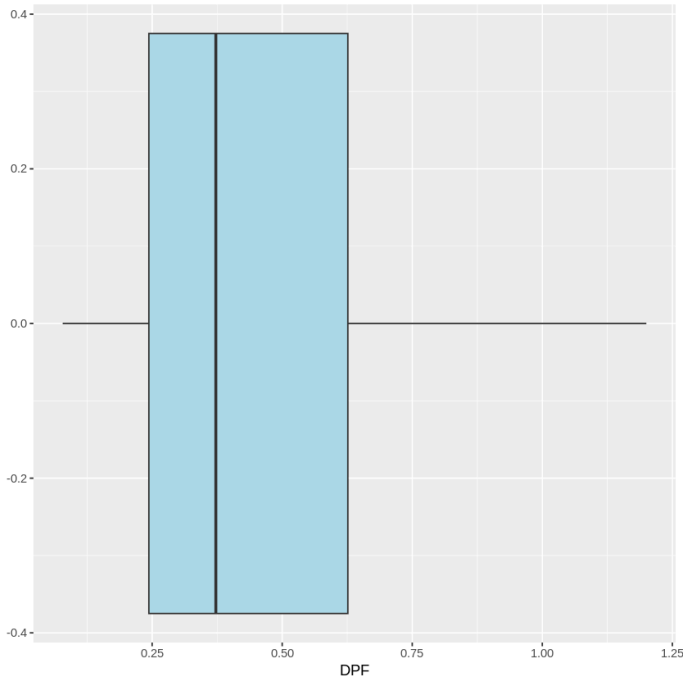
The result was zero duplicates

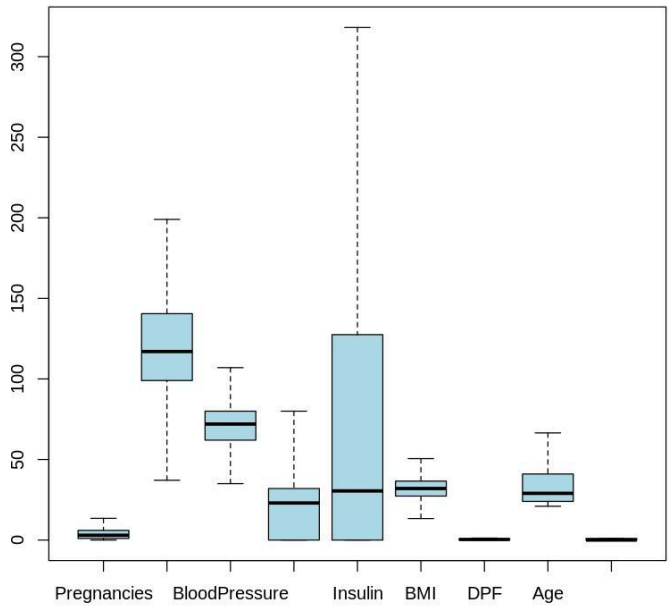
3.3. Checking for Outliers

As shown in the **Inconsistencies section, there were a lot of outliers located, especially in the Insulin and** blood pressure columns, so we used the Q1 and Q3 as a boundary for the outliers to relocate them to the nearest one

```
numerical_cols <- names(data)[sapply(data, is.numeric)]
for (col in numerical_cols) {
  Q1 <- quantile(data[[col]], 0.25, na.rm = TRUE)
  Q3 <- quantile(data[[col]], 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  # Cap outliers to boundaries
  data[[col]] <- ifelse(data[[col]] < lower_bound,
lower_bound,
                        ifelse(data[[col]] > upper_bound,
upper_bound, data[[col]]))
}
```

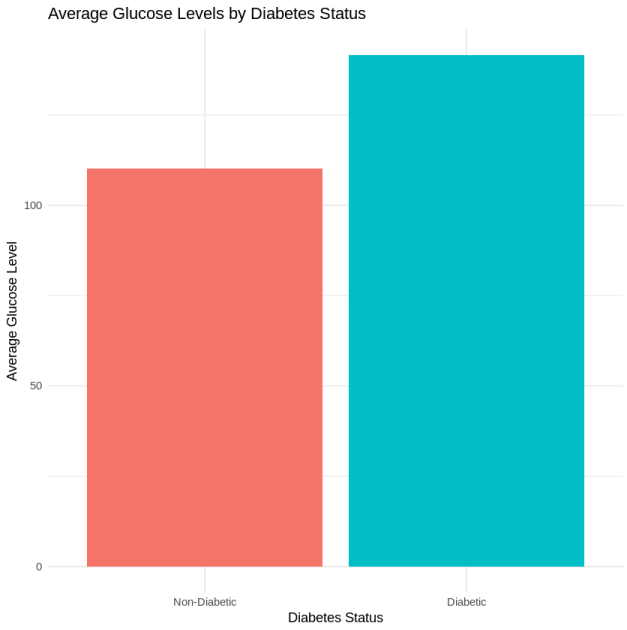
The resulting outcome after running this code was finding a significant amount of outliers and addressing them





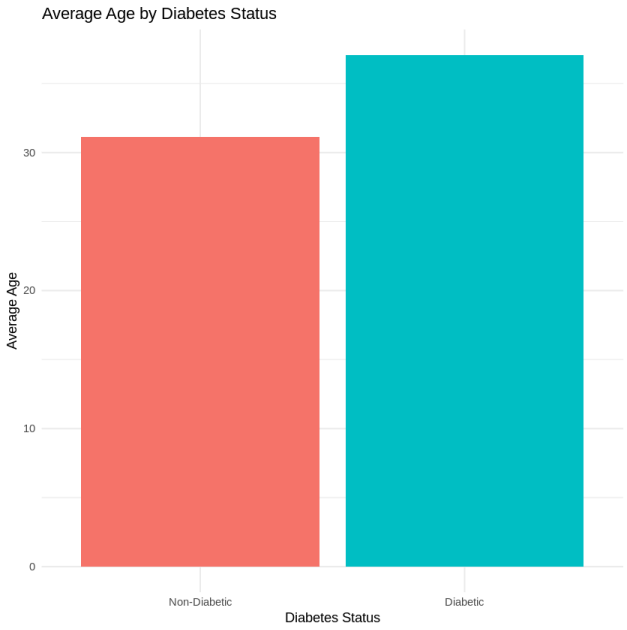
4. Exploratory Analysis

4.1. The average glucose levels among patients with and without diabetes.



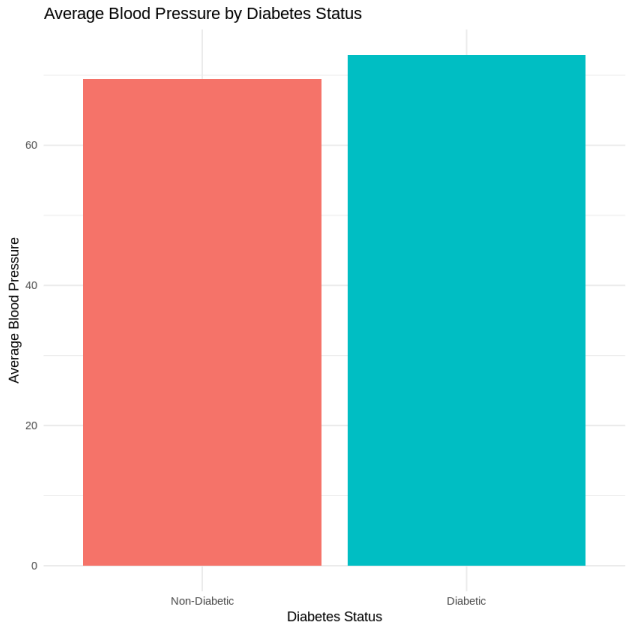
diabetes	Glucose
0	110.2027
1	141.5345

4.2. The average age of patients with and without diabetes.



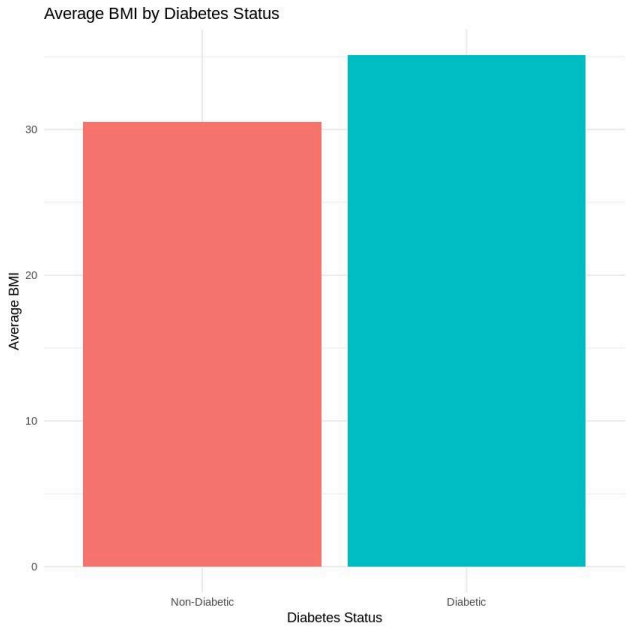
diabetes	Age
0	31.13500
1	37.05224

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



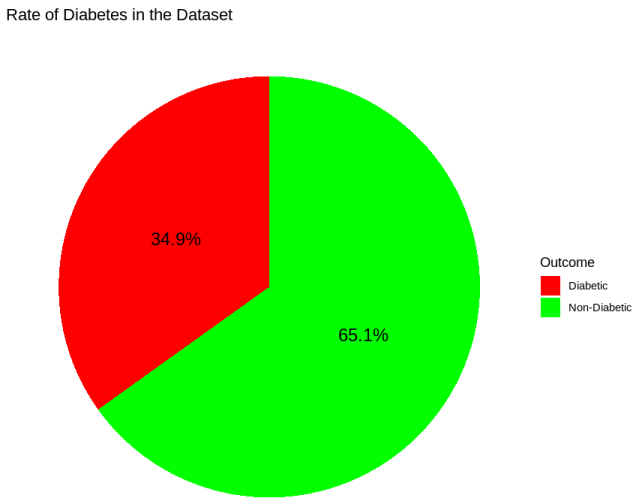
diabetes	Blood Pressure
0	69.5080
1	72.8806

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



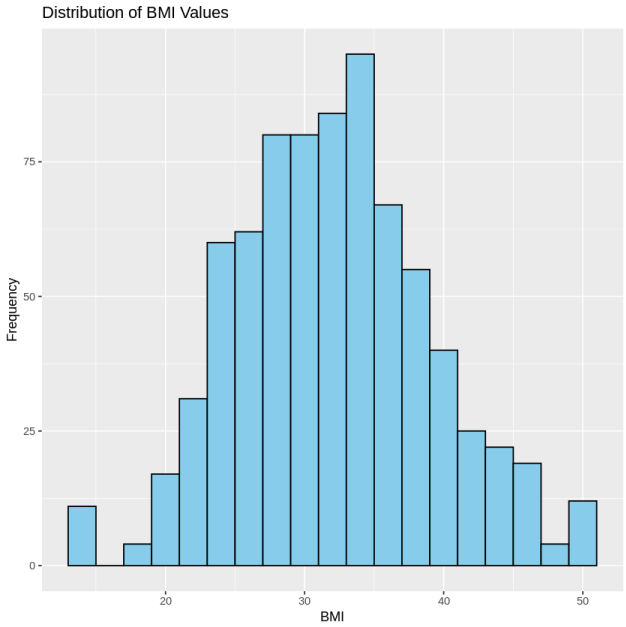
diabetes	BMI
0	30.5275
1	35.1056

4.5. The rate of diabetes among patients in the dataset



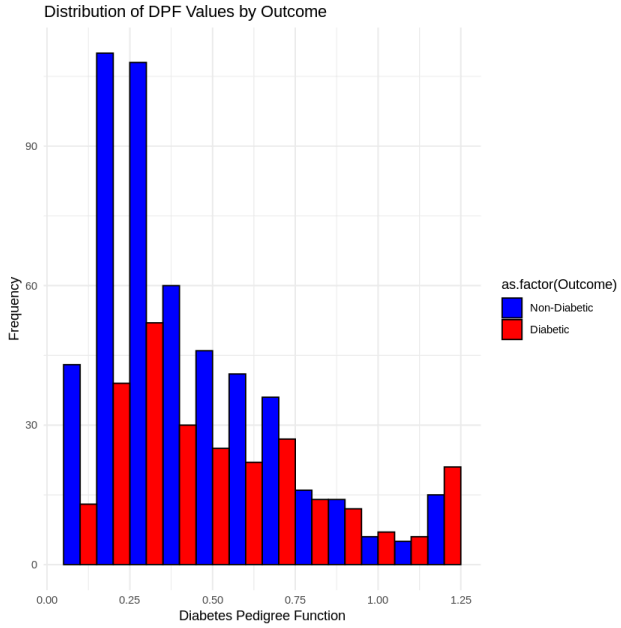
Approximately 34.895% of patients are diabetic.

4.6. The distribution of BMI values among all patients



The BMI is normally distributed

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

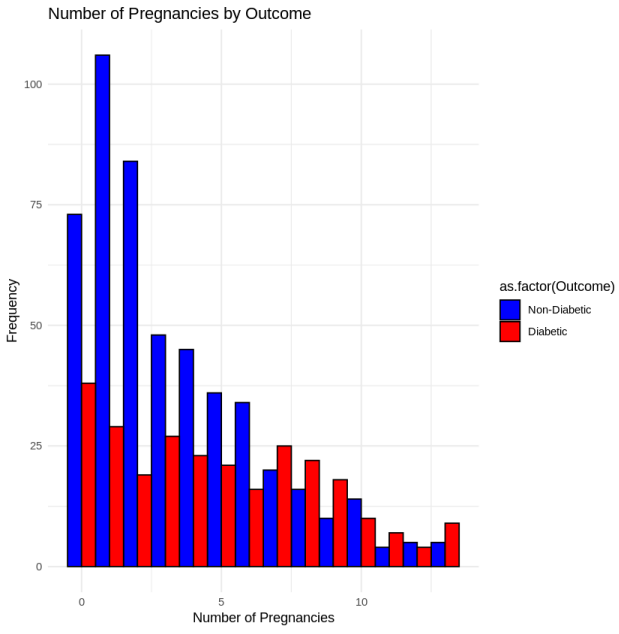


Higher Frequency in Lower DPF Values: Both groups are concentrated in the lower DPF range (e.g., around 0.25), with more non-diabetic individuals in this range.

Higher DPF Values: At higher DPF values (e.g., >0.75), the proportion of diabetic individuals appears relatively higher compared to non-diabetic individuals.

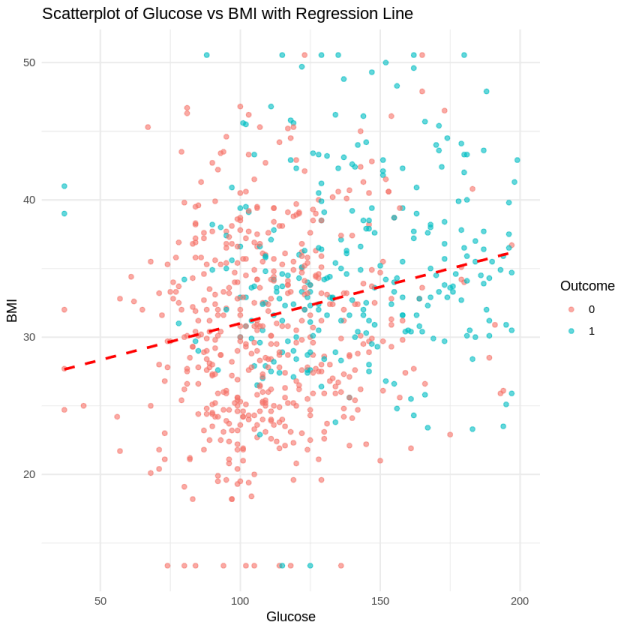
higher DPF value may indicate a greater likelihood of diabetes,

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



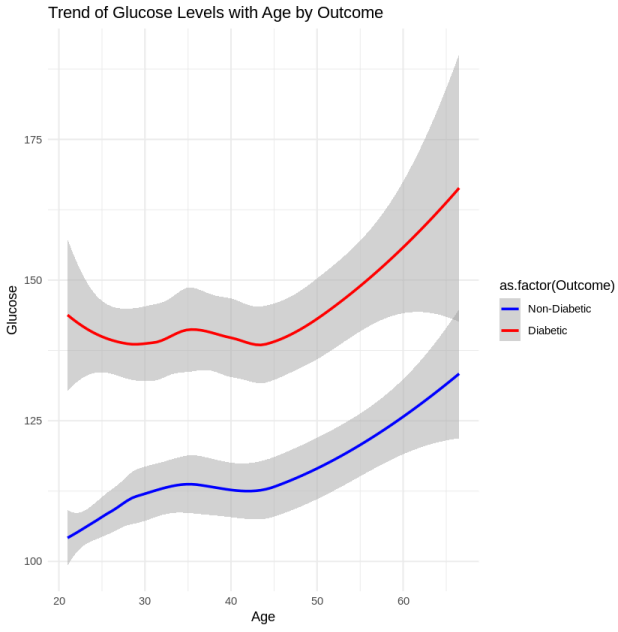
Low Number of Pregnancies: For individuals with fewer pregnancies (e.g., 0–2), the majority are non-diabetic (blue bars dominate), though some diabetic cases (red bars) are present.
Higher Number of Pregnancies (e.g., ≥6): The proportion of diabetic individuals increases as the number of pregnancies rises. In some higher pregnancy categories, diabetic individuals either match or exceed non-diabetic individuals in frequency.
There seems to be a positive association between the number of pregnancies and the likelihood of diabetes

4.9. The correlation between glucose levels and BMI



A slight relation is observed with an almost equal presence in the center region

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



Graphs show a positive, although slightly variable relation between the two factors

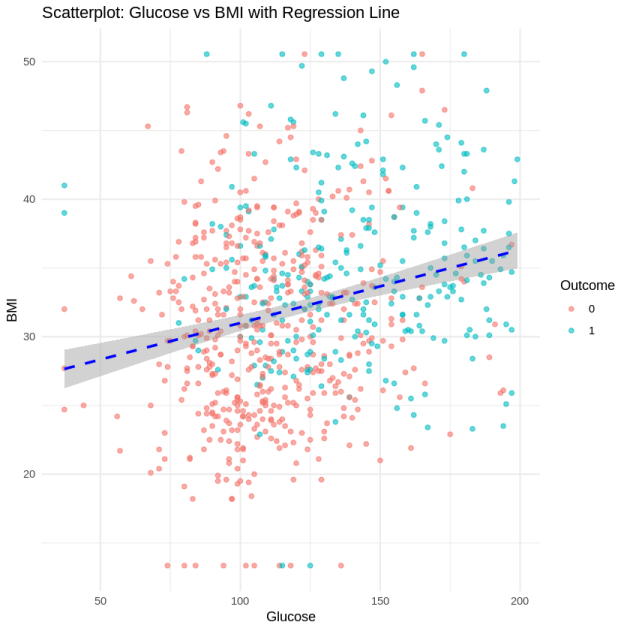
5. Questions

5.1. Main Questions

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

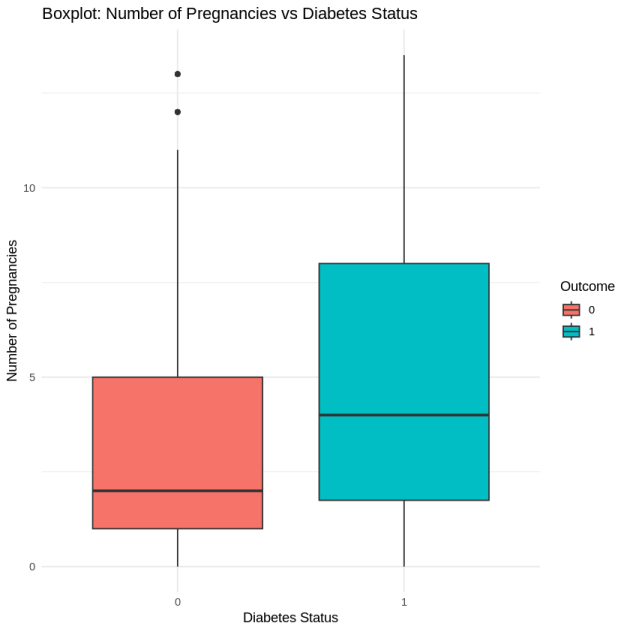
As shown in the graph, higher glucose level do in fact show positive correlation with the likelihood of having diabetes

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



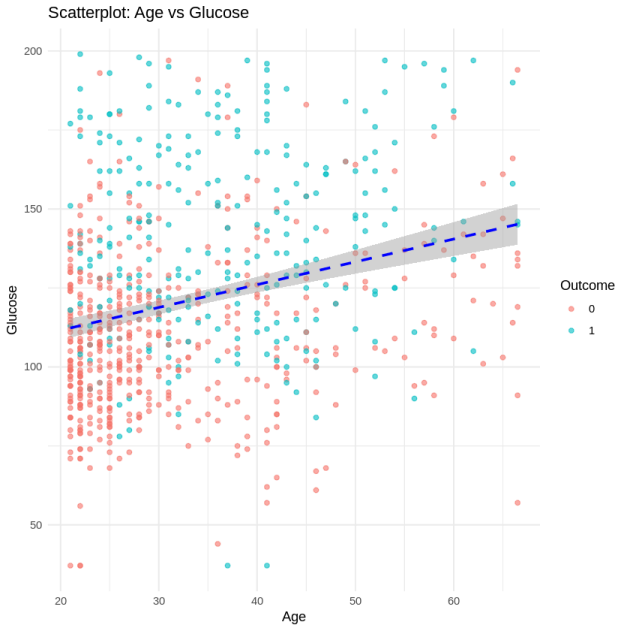
A weak correlation can be observed with some cases expressing a higher chance than others.

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

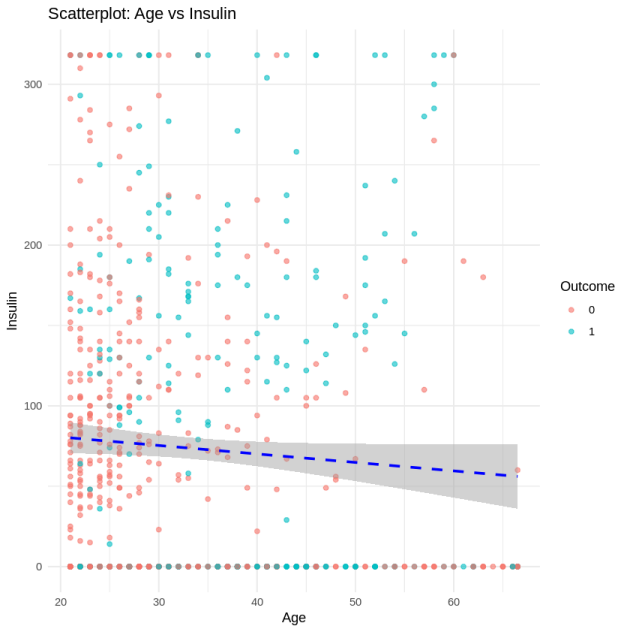


The data indicates that higher pregnancy rates may increase the risk of developing diabetes

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

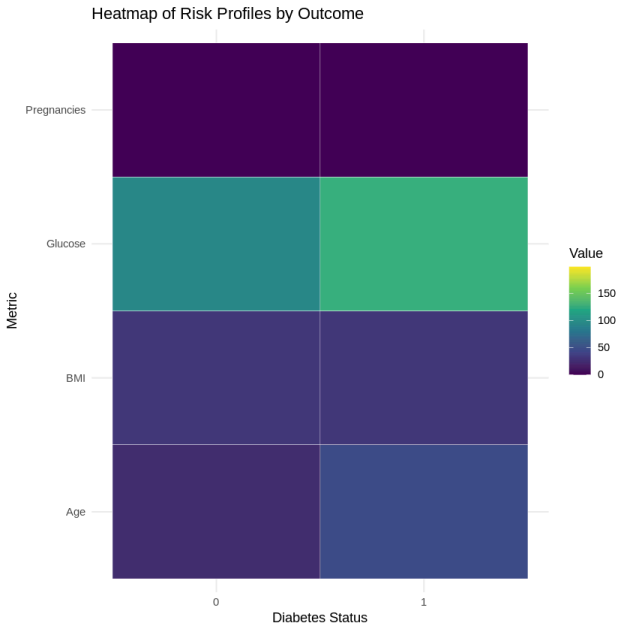


As shown in the graph, there is a higher density of average blood glucose cases around the 20 - 30 ages. While diabetics with high blood glucose levels are equally spread among age groups



As for insulin levels, a similar trend is observed with age groups not showing a strong correlation to insulin levels

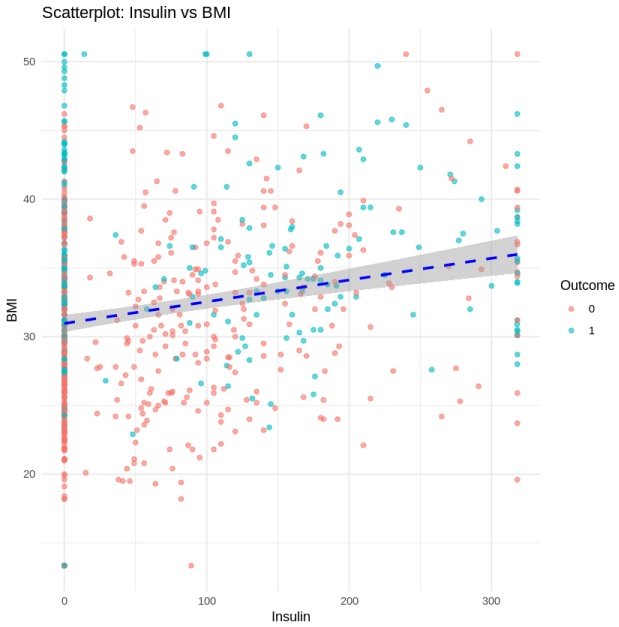
5.1.5. Can you identify common "risk profiles" for diabetic patients based on key metrics (glucose, BMI, age, etc.)?



Glucose levels are shown to be the strongest factor with age and BMI values have a very slight effect and pregnancy being close to neutral

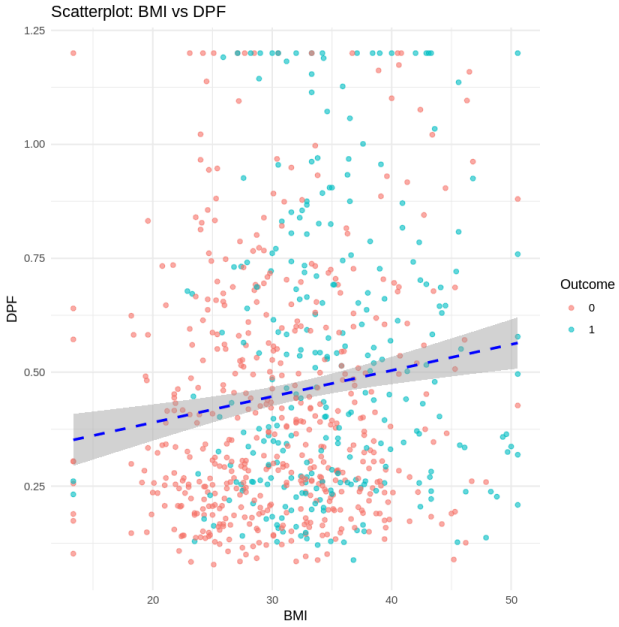
5.2. Other Questions

5.2.1. Is there a relationship between insulin levels and BMI?



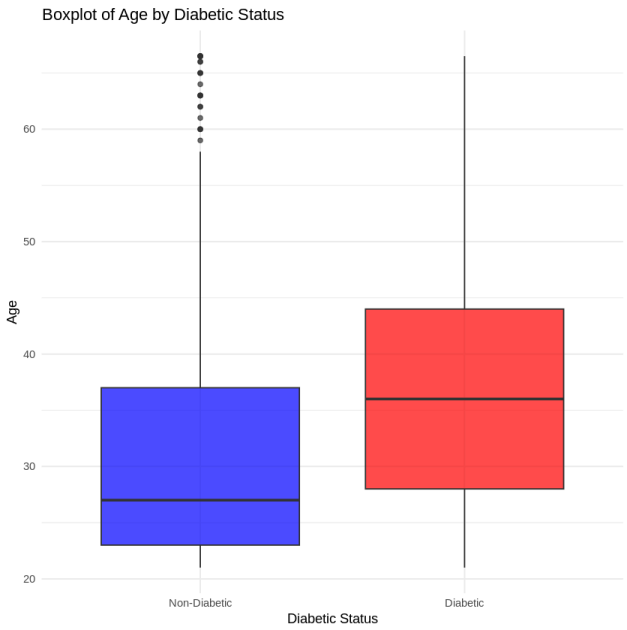
The graph shows a weak relation, which indicates that there is some effect although not a certain key factor

5.2.2. Are patients with higher BMI also more likely to have a higher diabetes pedigree function?

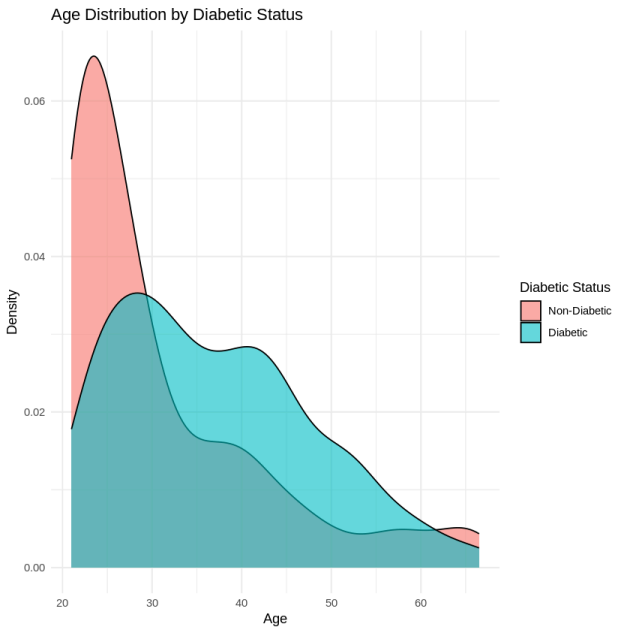


The data points shows an almost equal spread, the regression line indicates a weak positive relation between them

5.2.3. How does the age distribution differ between diabetic and non-diabetic patients?

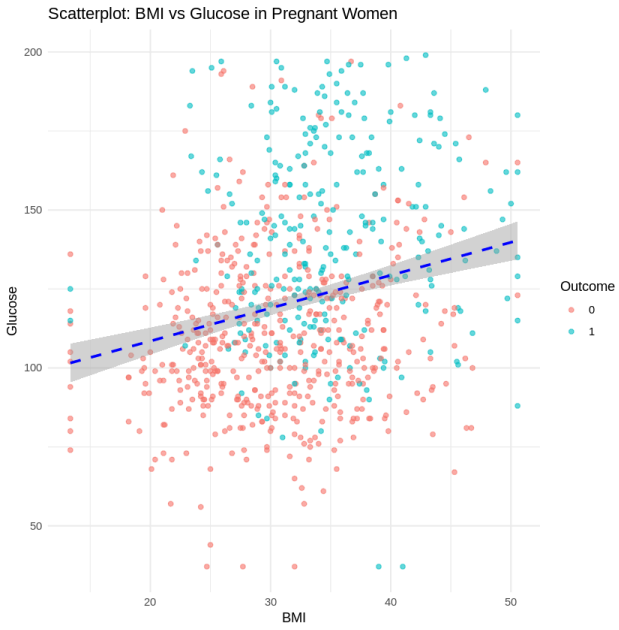


The data indicates that age is certainly a factor though not a strong one



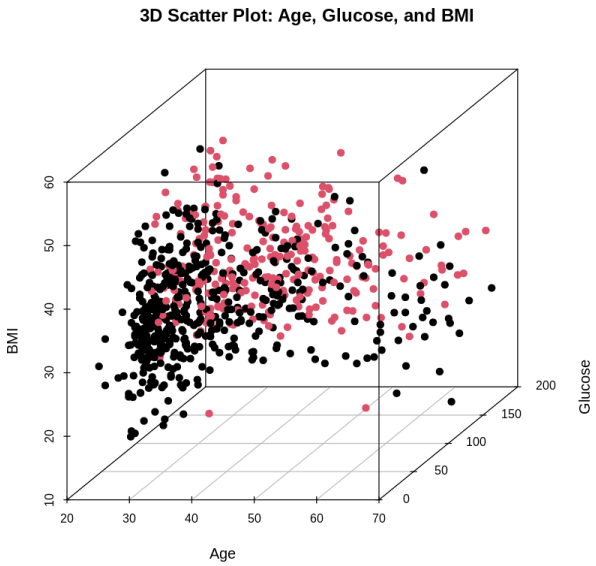
The distribution among ages indicates that as age increases, the diabetic status of patients increases along with it. The younger age groups show a significant peak in non diabetic cases and a smaller one at senior ages

5.2.4. Do pregnant women with a higher BMI tend to have higher glucose levels?



A slightly positive correlation is observed

5.2.5. Can age, glucose, and BMI together predict the likelihood of diabetes?



Predicting diabetes based on 3 key factors alone is a most difficult task, the 3D plot shows clustering among extremes as discussed before which indicates that, should these 3 factors have significant values, they can in fact predict the likelihood of diabetes although not to a high degree of accuracy

6. Hypothesis

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

6.1.1. Test Details

We used a Two-Sample t-test with significance level: 0.05.

6.1.2. Results

```
Two Sample t-test
data:  Glucose by Outcome
t = -15.109,
df = 766,
p-value < 2.2e-16
alternative hypothesis: true difference in means between
group 0 and group 1 is not equal to 0
95 percent confidence interval:
 [-35.40263, -27.26090]
sample estimates:
mean in group 0 = 110.2027
mean in group 1 = 141.5345
P-value: 2.46196e-45
```

The claim is valid since the p_value is smaller than 0.05, we reject the null hypothesis.

6.2. Claim: "There is a significant difference in Insulin level between diabetic and non-diabetic patients."

6.2.1. Test Details

We used a Two-Sample t-test with significance level: 0.05.

6.2.2. Results

```
Two Sample t-test
data:  Insulin by Outcome
t = -3.479,
df = 766,
p-value = 0.0005317
alternative hypothesis: true difference in means between
group 0 and group 1 is not equal to 0
95 percent confidence interval:
 [-38.27697 -10.66251]
sample estimates:
mean in group 0 = 65.11375
mean in group 1 = 89.58349
P-value: 0.0005316746
```

The claim is valid since the p_value is smaller than 0.05, we reject the null hypothesis.

7. Simulations

We explored the behavior of the confidence intervals for the BMI column with different scenarios, for each one we:

1. Calculated the population mean of the BMI column.
2. Generated random samples with different sizes.
3. For each sample: Calculated the sample mean and standard error (SE), calculated the 95% confidence interval using the t-distribution, and checked if the interval contained the true population mean.
4. Calculated the proportion of intervals that included the true mean.

7.1. Take 25 Random Samples of Size 15 from the Dataset

The result was that the proportion of intervals containing the true mean: 0.96 (96%)

7.2. Increase the Sample Size to 100

The result was that the proportion of intervals containing the true mean: 1.0 (100%) and average width of confidence intervals: 2.75.

7.3. Take 20 Random Samples of Size 10 from the Dataset

The result was that the proportion of intervals containing the true mean: 0.95 (95%).

8. Conclusion

The comprehensive analysis of the diabetes dataset has revealed significant insights into the factors associated with diabetes and their respective correlations. The dataset, comprising 768 entries of real-life female patients, was systematically processed and analyzed to ensure data integrity, enabling robust statistical inference and hypothesis testing.

Key Findings and Analysis

1. Glucose Levels and Diabetes Risk:
   The average glucose level among diabetic patients (141.53 mg/dL) is substantially higher than that of non-diabetic patients (110.20 mg/dL). A two-sample t-test confirmed this observation, yielding a t-value of -15.109 and a p-value $< 2.2e-16$. The highly significant result ($p < 0.05$) validates that glucose level is a critical determinant of diabetes risk.

2. BMI and Diabetes:
   Diabetic patients exhibited a higher average BMI (35.10) compared to non-diabetic individuals (30.53). Although a weak correlation was observed between glucose and BMI levels, BMI's influence on diabetes risk is evident.

3. Age as a factor:
   Diabetic patients tend to be older (mean age 37.05 years) compared to non-diabetic patients (mean age 31.14 years). The age distribution suggests that advancing age correlates with an increased risk of diabetes, but this relationship is not as strong as that of glucose levels.

4. Pregnancy and Diabetes Occurrence:
   An increasing number of pregnancies was associated with a higher likelihood of diabetes, indicating a potential relationship that merits further investigation.

5. Insulin levels and Statistical Significance:
   Average insulin levels were significantly different between diabetic (89.58) and non-diabetic groups (65.11), as evidenced by a two-sample t-test (t-value: -3.479, p-value: 0.0005317). This reinforces insulin's role in diabetes pathophysiology.

6. Exploratory Insights:
   Approximately 34.9% of the dataset's patients were diabetic.
   Visualizations of distributions and correlations provided nuanced understanding, including the relationship between BMI, glucose, and age.

7. Simulation Insights:
   Random sampling experiments for the BMI column demonstrated that larger sample sizes yield narrower confidence intervals and greater reliability. For instance, increasing the sample size to 100 resulted in all confidence intervals capturing the true mean (100%), with an average interval width of 2.75.

Final Thoughts

The analysis substantiates that glucose levels are the most decisive factor in determining diabetes risk, followed by BMI and age. While weaker correlations were observed for other factors such as insulin levels and number of pregnancies, their combined consideration can help in profiling at-risk individuals. Predictive modeling leveraging glucose, BMI, and age holds potential, albeit with limitations in accuracy when relying on these factors alone.

The statistical rigor applied throughout, including hypothesis testing and simulations, enhances the validity of these conclusions, providing a solid foundation for further research or practical applications in diabetes risk assessment and management.