# Diabetes Risk Analysis

Venelina Dimitrova

04.06.2026

## 1. Introduction

Diabetes is a chronic metabolic disease that affects millions of people worldwide and is increasingly becoming a major public health concern. Characterized by elevated blood glucose levels over prolonged periods, diabetes can lead to severe complications such as heart disease, kidney failure, blindness, and limb amputation if left unmanaged. Understanding the factors that contribute to the development of diabetes is therefore essential not only for early diagnosis but also for prevention and intervention strategies.

For this project, I chose to work with the Pima Indians Diabetes dataset, a well-known and widely used dataset in medical and data science research. This dataset contains clinical and personal information from female patients of Pima Indian heritage, including features such as the number of pregnancies, blood glucose levels, BMI, age, insulin levels, and family history of diabetes (pedigree function).

What makes this dataset particularly interesting is its real-world medical relevance and the opportunity to explore how common health metrics relate to diabetes risk. All individuals in the dataset are women aged 21 or older, which allows for focused analysis on female-specific health trends — such as pregnancy history — in relation to diabetes risk. Additionally, the dataset includes both numerical and categorical data, making it ideal for a comprehensive statistical and visual analysis.

**Project Objectives:** - To clean and prepare the data, identifying and addressing missing or inconsistent values. - To engineer meaningful features that help capture medical risk groups, such as age ranges, BMI categories, and glucose level thresholds. - To perform grouped analysis and statistical testing to determine which factors are most associated with diabetes. - To visualize important patterns and relationships through clear and informative plots. - To identify high-risk subpopulations based on combinations of medical indicators.

Ultimately, to draw clinically relevant insights that could support early screening and targeted intervention in real-world healthcare contexts. This project serves as both a technical exercise in data analysis with R and a meaningful contribution toward understanding how simple health metrics can inform diabetes risk at a population level.

## 2. Dataset Overview

The dataset used in this analysis contains 2,768 observations and 10 variables related to individual health indicators and diabetes outcomes. Each row represents data from a single female patient, and each column captures a specific medical measurement or demographic detail.

The structure of the dataset includes:

**Pregnancies:** Number of times the patient has been pregnant.

**Glucose:** Plasma glucose concentration (mg/dL) during an oral glucose tolerance test.

**BloodPressure:** Diastolic blood pressure (mm Hg).

**SkinThickness:** Triceps skinfold thickness (mm).

**Insulin:** 2-hour serum insulin level (mu U/ml).

**BMI:** Body mass index (kg/m$^2$).

**DiabetesPedigreeFunction:** A function that scores likelihood of diabetes based on family history.

**Age:** Age in years.

**Outcome:** Target variable — 1 indicates diabetes, 0 indicates no diabetes.

**Id:** Unique identifier.

```
data <- read_csv("Healthcare-Diabetes.csv")
```

```
## Rows: 2768 Columns: 10
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (10): Id, Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin,
B...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
str(data)
```

```
## spc_tbl_ [2,768 × 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                      : num [1:2768] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Pregnancies             : num [1:2768] 6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : num [1:2768] 148 85 183 89 137 116 78 115 197
125 ...
##  $ BloodPressure           : num [1:2768] 72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : num [1:2768] 35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : num [1:2768] 0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num [1:2768] 33.6 26.6 23.3 28.1 43.1 25.6 31
35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num [1:2768] 0.627 0.351 0.672 0.167 2.288
...
##  $ Age                     : num [1:2768] 50 31 32 21 33 30 26 29 53 54
...
##  $ Outcome                 : num [1:2768] 1 0 1 0 1 0 1 0 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   Pregnancies = col_double(),
##   ..   Glucose = col_double(),
##   ..   BloodPressure = col_double(),
##   ..   SkinThickness = col_double(),
##   ..   Insulin = col_double(),
##   ..   BMI = col_double(),
##   ..   DiabetesPedigreeFunction = col_double(),
##   ..   Age = col_double(),
##   ..   Outcome = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
dim(data)
```

```
## [1] 2768   10
```

```
head(data)
```

```
## # A tibble: 6 × 10
##      Id Pregnancies Glucose BloodPressure SkinThickness Insulin   BMI
##   <dbl>       <dbl>   <dbl>         <dbl>         <dbl>   <dbl> <dbl>
## 1     1           6     148            72            35       0  33.6
## 2     2           1      85            66            29       0  26.6
## 3     3           8     183            64             0       0  23.3
## 4     4           1      89            66            23      94  28.1
## 5     5           0     137            40            35     168  43.1
## 6     6           5     116            74             0       0  25.6
## # ℹ 3 more variables: DiabetesPedigreeFunction <dbl>, Age <dbl>, Outcome
<dbl>
```

```
summary(data)
```

```
##        Id          Pregnancies        Glucose      BloodPressure
##  Min.   :   1.0   Min.   : 0.000   Min.   :  0.0   Min.   :  0.00
##  1st Qu.: 692.8   1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00
##  Median :1384.5   Median : 3.000   Median :117.0   Median : 72.00
##  Mean   :1384.5   Mean   : 3.743   Mean   :121.1   Mean   : 69.13
##  3rd Qu.:2076.2   3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00
##  Max.   :2768.0   Max.   :17.000   Max.   :199.0   Max.   :122.00
##  SkinThickness       Insulin            BMI          
DiabetesPedigreeFunction
##  Min.   :  0.00   Min.   :  0.00   Min.   : 0.00   Min.   :0.0780  
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:27.30   1st Qu.:0.2440  
##  Median : 23.00   Median : 37.00   Median :32.20   Median :0.3750  
##  Mean   : 20.82   Mean   : 80.13   Mean   :32.14   Mean   :0.4712  
##  3rd Qu.: 32.00   3rd Qu.:130.00   3rd Qu.:36.62   3rd Qu.:0.6240  
##  Max.   :110.00   Max.   :846.00   Max.   :80.60   Max.   :2.4200  
##       Age           Outcome      
##  Min.   :21.00   Min.   :0.0000  
##  1st Qu.:24.00   1st Qu.:0.0000  
##  Median :29.00   Median :0.0000  
##  Mean   :33.13   Mean   :0.3439  
##  3rd Qu.:40.00   3rd Qu.:1.0000  
##  Max.   :81.00   Max.   :1.0000
```

```
mean(data$Age)
```

```
## [1] 33.13223
```

```
sd(data$BMI)
```

```
## [1] 8.076127
```

```
max(data$Age)
```

```
## [1] 81
```

```
table(data$Outcome)
```

```
## 
##    0    1 
## 1816  952
```

**Key Observations:** - The dataset consists of only female patients, which is especially relevant for pregnancy-related risk factors. - The average age is approximately 33.13 years. - The standard deviation of BMI is about 8.08, showing variation in body weight. - The maximum age observed is 81 years. - Outcome distribution shows the following: - 1816 non-diabetic individuals (Outcome = 0) - 952 diabetic individuals (Outcome = 1)

These initial insights provide the foundation for deeper exploration into how health indicators are associated with diabetes risk.

# 3. Data Cleaning

In this step, we identified and addressed missing values in the dataset.

Although the dataset does not contain explicit NA values, several numeric variables used the value 0 as a placeholder for missing or biologically implausible measurements. This is common in medical datasets where, for example, a glucose or blood pressure reading of zero is not realistic.

We considered the following variables as potentially affected:

**Glucose**

**BloodPressure**

**SkinThickness**

**Insulin**

**BMI**

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

```
##                       Id              Pregnancies                  Glucose
##                        0                        0                        0
##            BloodPressure            SkinThickness                  Insulin
##                        0                        0                        0
##                      BMI DiabetesPedigreeFunction                      Age
##                        0                        0                        0
##                  Outcome
##                        0
```

```
cols_with_missing <- c("Glucose", "BloodPressure", "SkinThickness",
"Insulin", "BMI")
dat <- data %>%
  mutate(across(all_of(cols_with_missing), ~na_if(., 0)))
colSums(is.na(data))
```

```
##                       Id              Pregnancies                  Glucose
##                        0                        0                        0
##            BloodPressure            SkinThickness                  Insulin
##                        0                        0                        0
##                      BMI DiabetesPedigreeFunction                      Age
##                        0                        0                        0
##                  Outcome
##                        0
```

These columns were processed by converting all zero values to NA, allowing for accurate statistical summaries and visualizations.This step ensures the integrity of the data for all downstream analyses, especially when calculating averages, standard deviations, or performing statistical tests.

#4. Feature Engineering: Creating Categorical Risk Groups

To support more insightful analysis, we created several categorical variables from continuous medical measurements. This step allows for meaningful subgroup comparisons and risk profiling. Below are the newly created variables:

BMI_Category: Body Mass Index (BMI) was categorized based on WHO standards:

Underweight: BMI < 18.5

Normal: 18.5 ≤ BMI < 25

Overweight: 25 ≤ BMI < 30

Obese: BMI ≥ 30

Pregnancy_Group: Since all individuals in the dataset are female, pregnancy count is a key variable:

No pregnancy

1–3 pregnancies

4 or more pregnancies

Age_Group: Age was grouped to analyze trends across life stages:

Young: < 30 years

Middle-aged: 30–50 years

Older: > 50 years

Glucose_level_Group: Based on medical diagnostic thresholds:

Normal: Glucose < 140

Prediabetes: 140 ≤ Glucose < 200

Diabetes: Glucose ≥ 200

Low_risk_Group: A binary variable indicating individuals with a low genetic predisposition to diabetes:

Yes: Diabetes Pedigree Function < 0.2

No: ≥ 0.2

```r
data <- data %>%
  mutate(BMI_Category = case_when(
    BMI < 18.5 ~ "Underweight",
    BMI >= 18.5 & BMI < 25 ~ "Normal",
    BMI >= 25 & BMI < 30 ~ "Overweight",
    BMI >= 30 ~ "Obese"
  ))

data <- data %>%
  mutate(Pregnancy_Group = case_when(
    Pregnancies == 0 ~ "No pregnancy",
    Pregnancies >= 1 & Pregnancies <= 3 ~ "1-3 pregnancies",
    Pregnancies > 3 ~ "4 or more pregnancies"
  ))

data <- data %>%
  mutate(Age_Group = case_when(
    Age < 30 ~ "Young",
    Age >= 30 & Age <= 50 ~ "Middle-aged",
    Age > 50 ~ "Older"
  ))

data <- data %>%
  mutate(Glucose_level_Group = case_when(
    Glucose < 140 ~ "Normal",
    Glucose >= 140 & Glucose < 200 ~ "Prediabetes",
    Glucose >= 200 ~ "Diabetes"
  ))

data <- data %>%
  mutate(Low_risk_Group = ifelse(DiabetesPedigreeFunction < 0.2, "Yes",
"No"))
```

These engineered features provide interpretable dimensions for evaluating diabetes prevalence and were used throughout the grouped analysis, hypothesis testing, and visualization stages of the project.

#5.Analysis by Risk Subgroups

To understand how diabetes risk varies across key health-related categories, we analyzed and visualized grouped summaries based on Body Mass Index (BMI) and pregnancy history.

5.1 Diabetes Risk by BMI Category We grouped patients by BMI category and calculated:

The total number of individuals in each category.

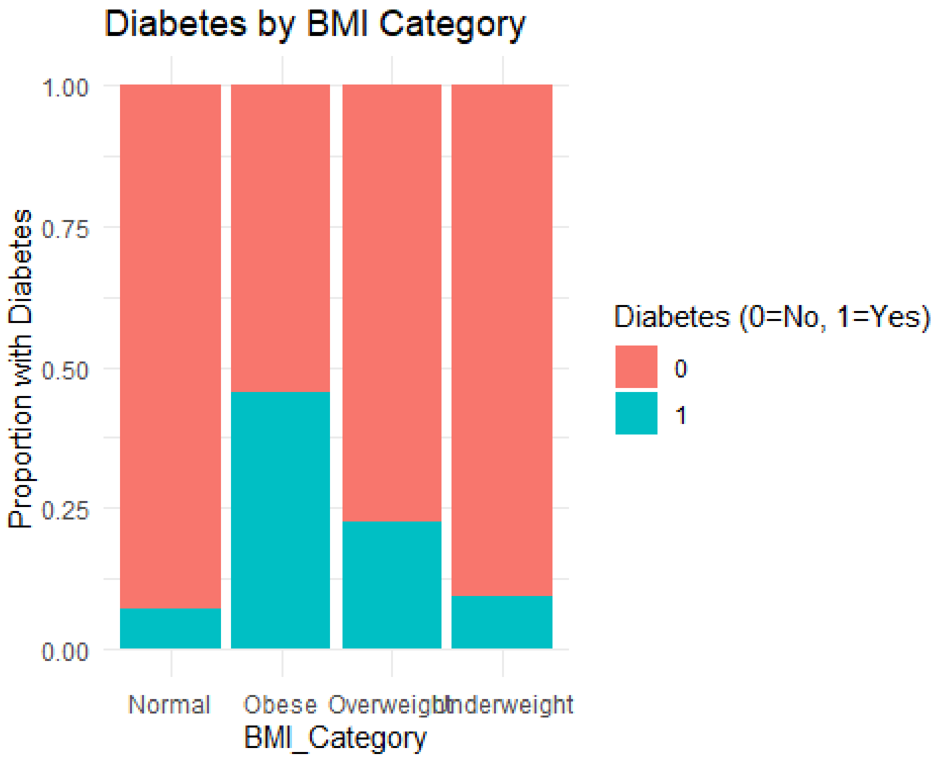
The average glucose level.

The rate of diabetes within each group.

```
bmi_summary <- data %>%
  group_by(BMI_Category) %>%
  summarise(
    Count = n(),
    Diabetes_Rate = mean(Outcome, na.rm = TRUE),
    Avg_Glucose = mean(Glucose, na.rm = TRUE)
  )
print(bmi_summary)

## # A tibble: 4 × 4
##   BMI_Category Count Diabetes_Rate Avg_Glucose
##   <chr>        <int>         <dbl>       <dbl>
## 1 Normal         357          0.0700        108.
## 2 Obese         1704          0.454         126.
## 3 Overweight     654          0.226         118.
## 4 Underweight     53          0.0943         99.7
```

The bar chart below shows the average glucose level per BMI category. We observe a rising trend, with obese individuals showing the highest average glucose, indicating a greater risk of developing diabetes.

```
ggplot(bmi_summary) +
  aes(x = BMI_Category, y = Avg_Glucose, fill = BMI_Category) +
  geom_col() +
  labs(y = "Average Glucose", title = "Average Glucose by BMI Category") +
  theme_minimal()
```



5.2 Diabetes Risk by Pregnancy History Next, we grouped patients based on the number of pregnancies and calculated:

The number of individuals per group.

The rate of diabetes.

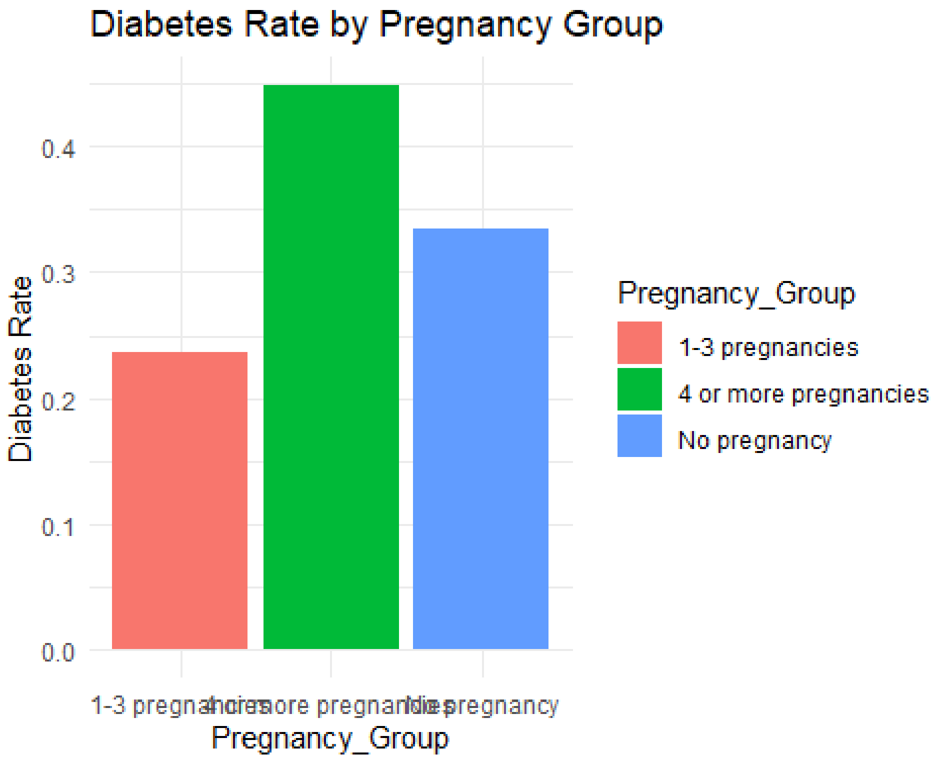
The average insulin level.

```
pregnancy_summary <- data %>%
  group_by(Pregnancy_Group) %>%
  summarise(
    Count = n(),
    Diabetes_Rate = mean(Outcome, na.rm = TRUE),
    Avg_Insulin = mean(Insulin, na.rm = TRUE)
  )
print(pregnancy_summary)
```

```
## # A tibble: 3 × 4
##   Pregnancy_Group       Count Diabetes_Rate Avg_Insulin
##   <chr>                 <int>         <dbl>       <dbl>
## 1 1-3 pregnancies        1148         0.237        92.1
## 2 4 or more pregnancies  1208         0.449        67.7
## 3 No pregnancy            412         0.335        83.1
```

The bar chart below presents the diabetes rate by pregnancy group. Notably, women with four or more pregnancies show a considerably higher rate of diabetes, suggesting that high parity may be a relevant factor in diabetes risk assessment.

```
ggplot(pregnancy_summary) +
  aes(x = Pregnancy_Group, y = Diabetes_Rate, fill = Pregnancy_Group) +
  geom_col() +
  labs(y = "Diabetes Rate", title = "Diabetes Rate by Pregnancy Group") +
  theme_minimal()
```

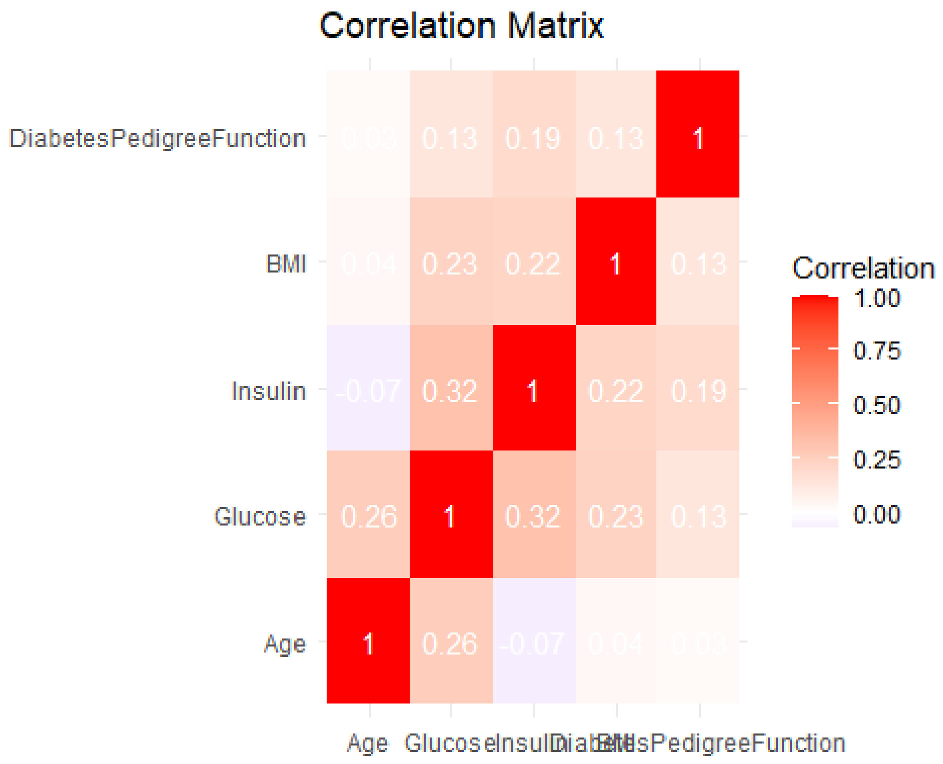


#6.Data Visualization and Relationship Analysis

To better understand the relationships between key variables and their role in diabetes risk, we used a series of data visualizations. These plots reveal patterns, trends, and associations that might not be obvious from raw summaries alone.

6.1 Correlation Heatmap - We first computed the pairwise correlation between major numeric health indicators, including Age, Glucose, Insulin, BMI, and Diabetes Pedigree Function. The heatmap below displays these correlations with both color intensity and value labels.

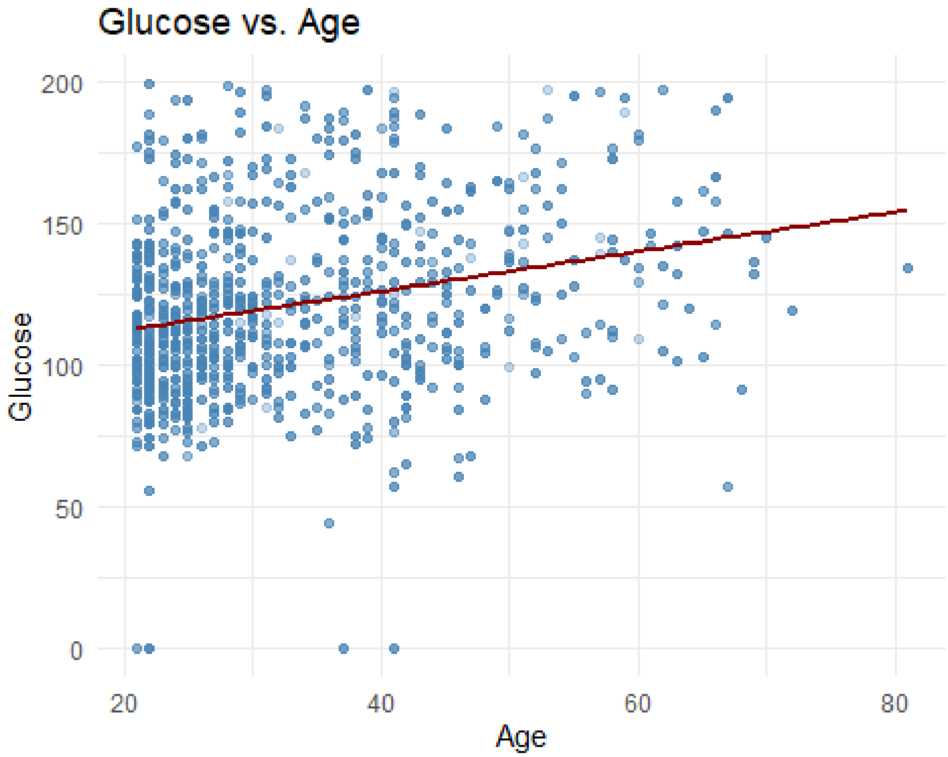
```
install.packages("reshape2")

## Warning: package 'reshape2' is in use and will not be installed

library(reshape2)

cor_vars <- data %>%
  select(Age, Glucose, Insulin, BMI, DiabetesPedigreeFunction)

cor_matrix <- round(cor(cor_vars, use = "complete.obs"), 2)
cor_melt <- melt(cor_matrix)

ggplot(cor_melt, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = value), color = "white", size = 4) +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint =
0) +
  labs(title = "Correlation Matrix", x = NULL, y = NULL, fill =
"Correlation") +
  theme_minimal()
```



Insight: We observe moderate positive correlations between glucose and insulin, as well as between age and glucose, suggesting potential age-related trends in glucose regulation.

6.2 Glucose vs. Age (Regression Plot) To visualize the relationship between age and glucose levels, we created a scatter plot with a linear regression line.

```
ggplot(data) +
  aes(x = Age, y = Glucose) +
  geom_point(alpha = 0.3, color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "darkred") +
  labs(title = "Glucose vs. Age", x = "Age", y = "Glucose") +
  theme_minimal()

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

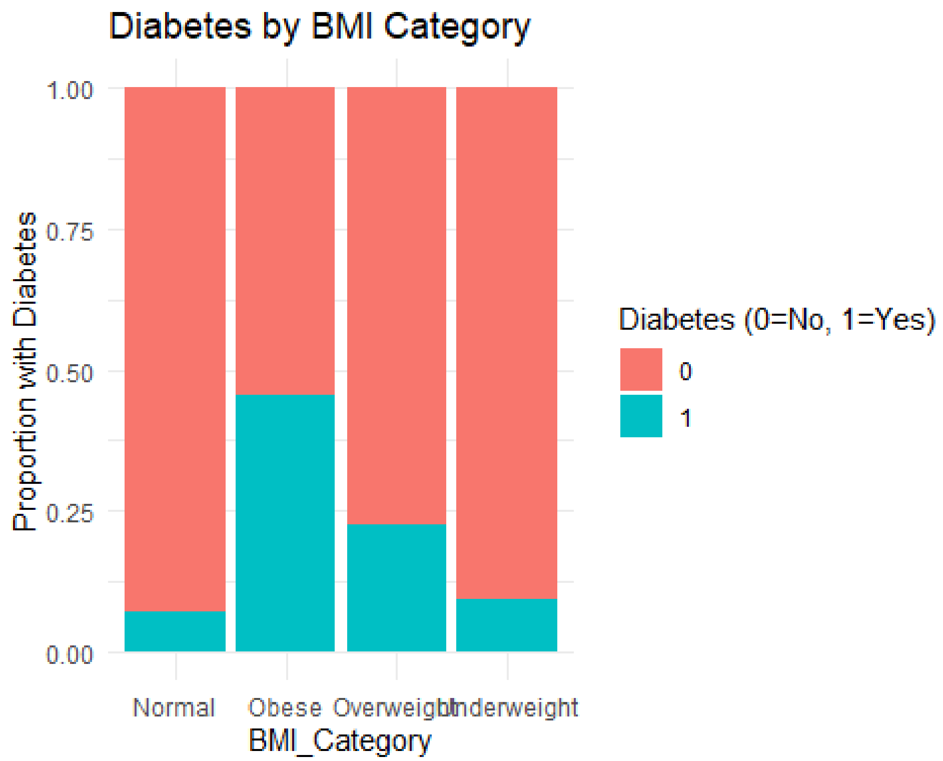


Insight: There is a visible upward trend, indicating that glucose levels tend to increase with age. This is consistent with the known biological decline in insulin sensitivity over time.

6.3 Proportion of Diabetes by BMI Category The stacked bar chart below shows the proportion of diabetic and non-diabetic individuals in each BMI category. The bars are normalized to allow comparison of proportions.

```
ggplot(data) +
  aes(x = BMI_Category, fill = factor(Outcome)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion with Diabetes", fill = "Diabetes (0=No, 1=Yes)", title
```

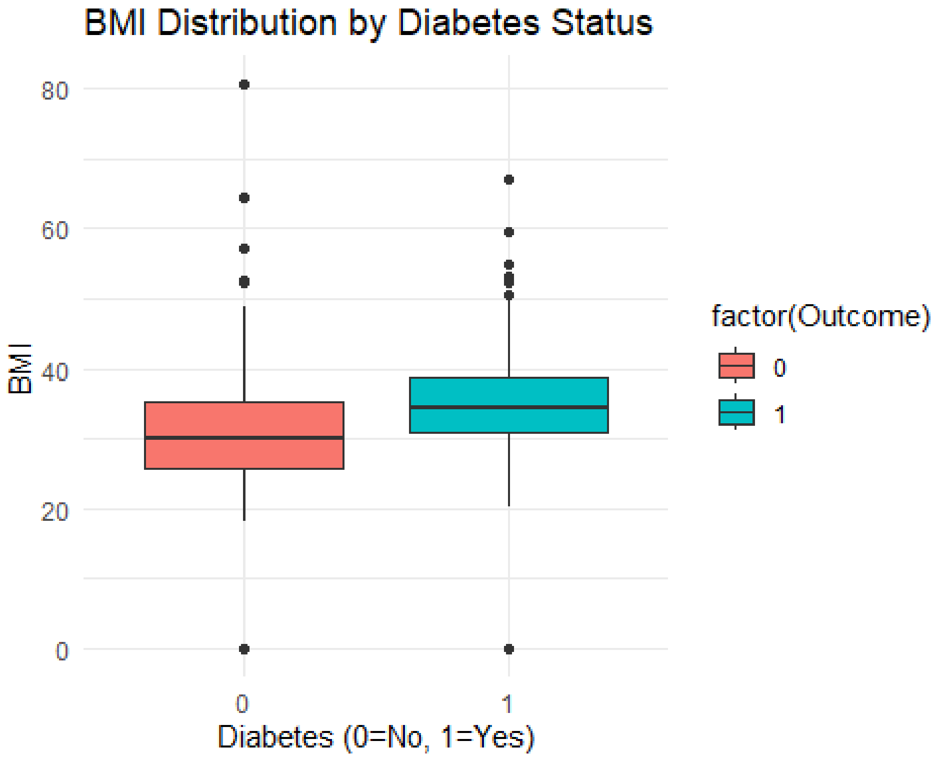
```r
= "Diabetes by BMI Category") +
  theme_minimal()
```



Insight: The proportion of individuals with diabetes increases sharply in the Obese group, further highlighting BMI as a strong predictive indicator.

6.4 BMI Distribution by Diabetes Status Boxplots allow us to explore the distribution of BMI values for diabetic and non-diabetic individuals:

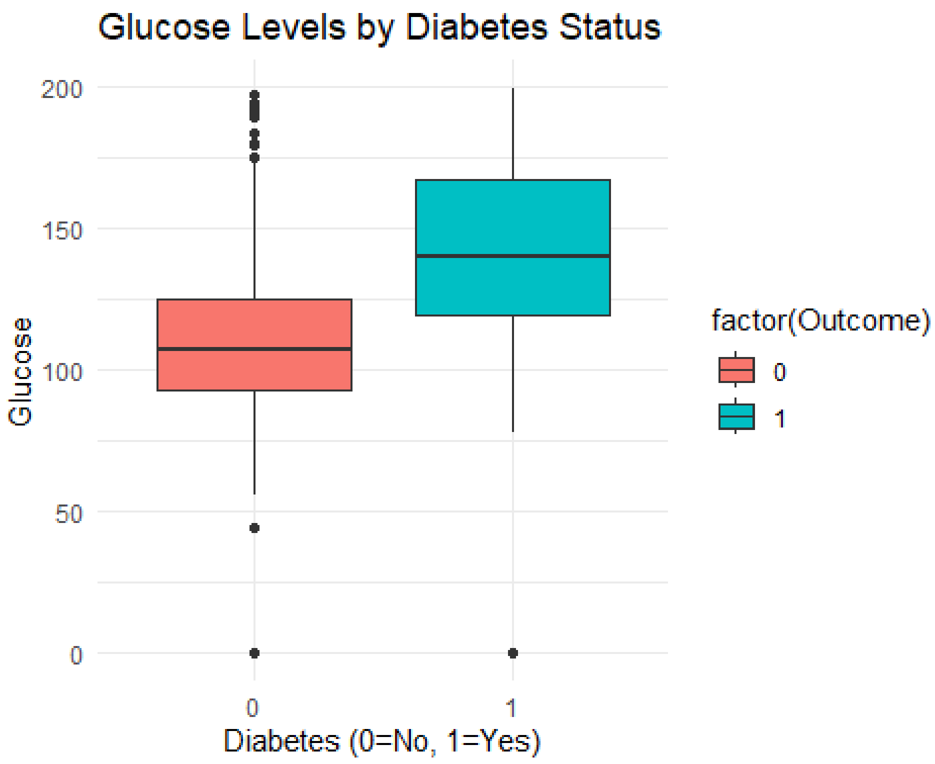
```r
ggplot(data) +
  aes(x = factor(Outcome), y = BMI, fill = factor(Outcome)) +
  geom_boxplot() +
  labs(x = "Diabetes (0=No, 1=Yes)", y = "BMI", title = "BMI Distribution by
Diabetes Status") +
  theme_minimal()
```



Insight: The diabetic group shows a higher median BMI and more outliers with extreme values, reinforcing the pattern observed in the previous plot.

6.5 Glucose Distribution by Diabetes Status Similarly, we examined glucose levels by diabetes status using boxplots:

```
ggplot(data) +
  aes(x = factor(Outcome), y = Glucose, fill = factor(Outcome)) +
  geom_boxplot() +
  labs(x = "Diabetes (0=No, 1=Yes)", y = "Glucose", title = "Glucose Levels
by Diabetes Status") +
  theme_minimal()
```



Insight: The glucose levels of diabetic individuals are significantly higher and more variable, indicating clear separation between the two groups.

#7. Conclusion This analysis demonstrates clear relationships between various clinical indicators and the likelihood of diabetes among female patients. Key risk factors identified include elevated glucose levels, obesity, increased age, and higher pregnancy counts. Statistical testing and visual exploration further confirmed that these variables are strongly associated with diabetes outcomes.

However, the dataset is limited in scope, as it includes only female participants and omits potentially significant lifestyle or demographic features. For example, factors such as physical activity level, dietary habits, family history beyond the pedigree score, socioeconomic status, and ethnicity could provide additional predictive power.

Moreover, incorporating male patients or gender-diverse populations would allow for more inclusive and generalized insights. Future studies that include a broader population and more lifestyle-related variables would likely improve the precision and applicability of diabetes risk prediction models.

In conclusion, this project successfully highlights patterns and relationships using available data, but it also emphasizes the importance of richer, multidimensional datasets for advancing preventive healthcare and personalized treatment planning.