# Analyzing Diabetes Dataset

**Submitted by**

Adham Ashraf 202200953

Mohamed Hassan 202201257

Mohamed Mohamed 202201507

Seif Mahdy 202200990

Yousef Moustafa 202201699

**Instructor**

**Name:** Dr.Rasha Mandouh

School of Computational Sciences and Artificial Intelligence

**University of Science and Technology**

**Zewail City of Science, Technology, and Innovation**

**Table of Contents:**

1. Introduction 3
2. Data Processing and Analysis Steps 6
2.1 Description of the methods used for data collection, cleaning, and preprocessing 6
2.2 Exploring Dataset through Visuals 6
2.3 Confidence Interval analysis 6
3. Challenges, Limitations, and Assumptions 6
3.1 Discussion of challenges faced and solutions implemented 6
3.2 Explanation of limitations in data or methodology and assumptions made 6
4. Conclusion 6
4.1 Summarize the findings and insights derived from the analysis 6
4.2 Suggest future directions or potential improvements 6

# 1. Introduction

This project analyzes a publicly available diabetes dataset from Kaggle, with the goal of identifying health trends and some factors contributing to the development of diabetes. Diabetes is a chronic condition with very high public health importance due to its continuously increasing prevalence in most parts of the world. Its risk factors, when identified early, will lead to better prevention and management strategies that reduce the burden of healthcare.

The data set is a collection of medical diagnostic measurements that was obtained to forecast the appearance of diabetes from several health factors. This dataset has 768 instances of female patients, each described by 8 health-related variables. The Outcome variable identifies the patient as having diabetes (1) or non-diabetic (0). This data acts as a solid foundation for the implementation of statistical methods, development of visual insights, and claims verification by hypothesis testing-not excluding other uses that could be applied within the course project.

## Dataset Overview:

- **Source:** Kaggle
- **Features:** Includes variables such as **glucose levels**, **blood pressure**, **BMI**, **age**, **number of pregnancies**, and **diabetes pedigree function**.
- **Relevance:** Understanding correlations and patterns in this dataset can provide insights into diabetes risk factors.

# 2. Data Processing and Analysis Steps

## 2.1 Description of the methods used for data cleaning and preprocessing

- Checked for missing values and anomalies in all numerical features  no missing values were identified, so imputers were not required.
- We checked for outliers in each numerical feature and addressed them using the **Z-score method**.
- We have Encoded the Categorical Column **“Outcome” ( 0 : Non Diabetic , 1 : Diabetic)**

## 2.2 Exploring Datasets Through Visuals

- By displaying the feature's distribution histogram, we presented the following figure:

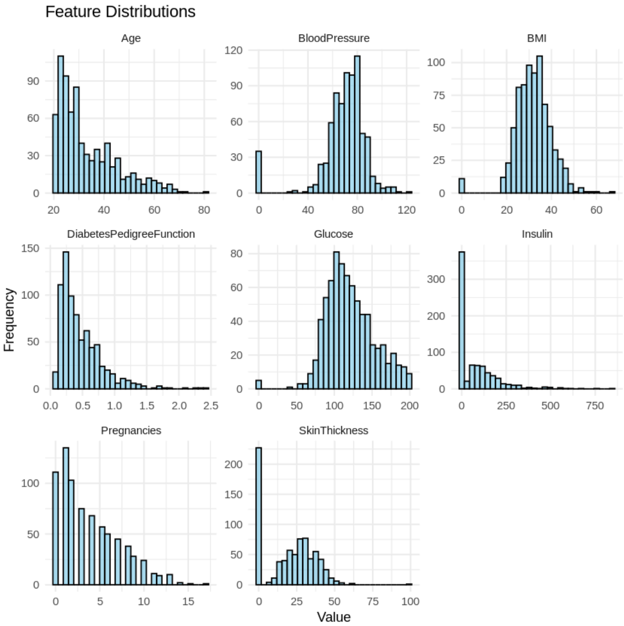


**Figure [1] : Distribution of Diabetes Dataset Features**

**Age**: The distribution is fairly uniform with a slight bias toward younger ages.

**BloodPressure**: It is highly right-skewed with a long tail to the right.

**BMI**: The distribution is roughly bell-shaped with a slight skew toward higher values.

**DiabetesPedigreeFunction**: Distribution is skewed towards lower values with a tail of higher values.

**Glucose**: Distribution is skewed toward the lower values with a tail towards the higher values.

**Insulin**: Distribution is heavily skewed towards lower values with a long tail towards higher values.

**Pregnancies**: This dataset is positively skewed; there is a long tail toward the higher values.

**SkinThickness**: The data follows a positive-skewed distribution, with a long tail in the higher values.

**To address these variations, we scaled down the numerical features.**

- By displaying the BMI distribution histogram, we presented the following figure:

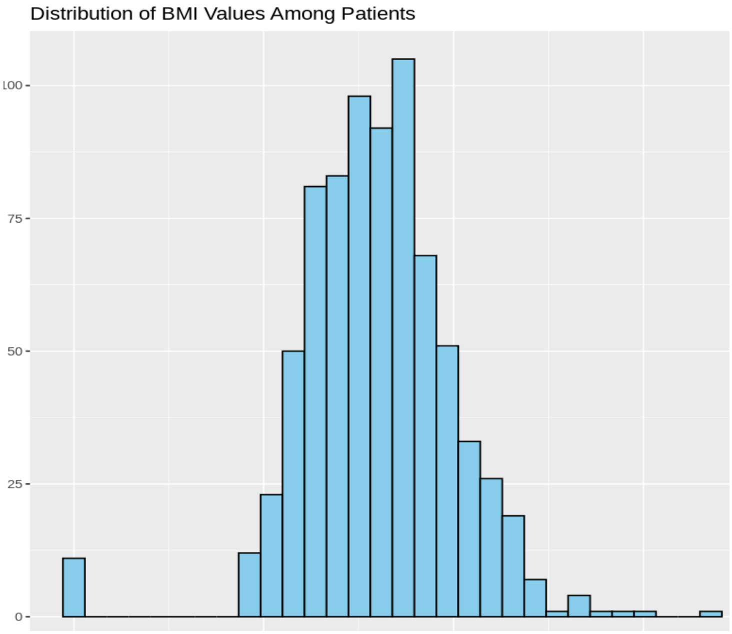
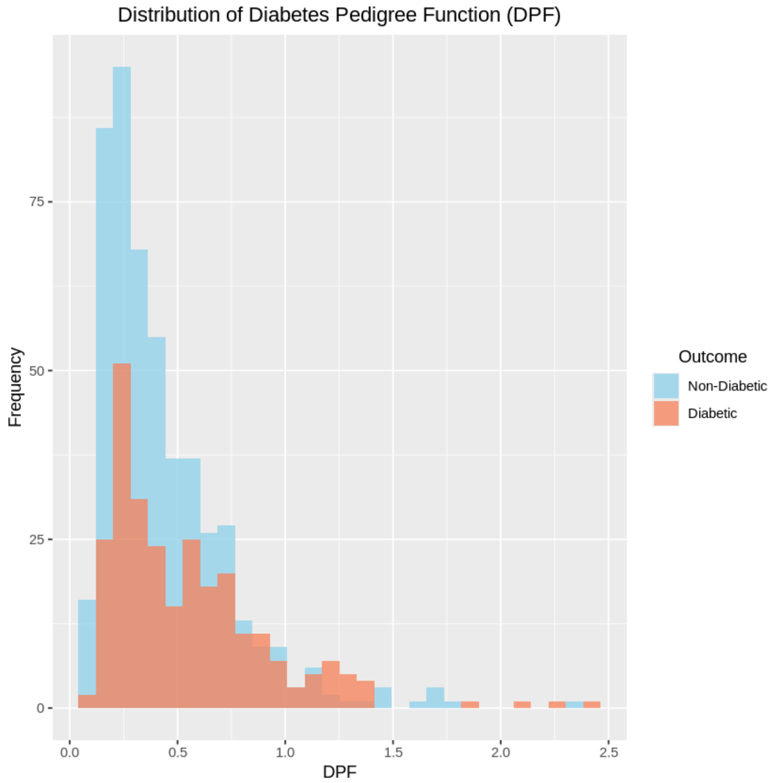


**Figure [2] : Distribution of BMI Values among Patients**

The Following figure shows the distribution of BMI, which is **normally** distributed. From this, it is clear that most patients have their values of BMI concentrated around the mean, while a few patients have values considerably lower or higher. The normal distribution suggests that the pattern of the BMI of the patient population is predictable, hence valuable for statistical analysis and modeling.

- By displaying the Diabetes Pedigree Function distribution histogram among diabetic and non-diabetic patients, we present the following figure:



-

**Figure [3] : Distribution of DBF Function among Patients**

The following figure  is the distribution of the Diabetes Pedigree Function, DPF, for both diabetic and non-diabetic patients. **The DPF provides the possibility of diabetes for the people based on the genetic disposition given by the pedigree chart**. From the following figure, it is clearly observed that in both the groups, distribution for DPF is highly asymmetric to lower values. More than a major group, irrespective of status whether diabetic or not, had less than a DPF of 0.5. However, when the DPF increases at each stage, the diabetic subjects showed a higher proportion as compared to the nondiabetic subjects. **This indicates that the higher the DPF value, the higher the risk of being a diabetic**. The following figure infers the influence of genetic predisposition on the susceptibility to diabetes since with the increase in the value of DPF, higher is the chance that the individual belongs to a diabetic group.

- By displaying the Relationship between Number of pregnancies and diabetes occurrence , we present the following figure:

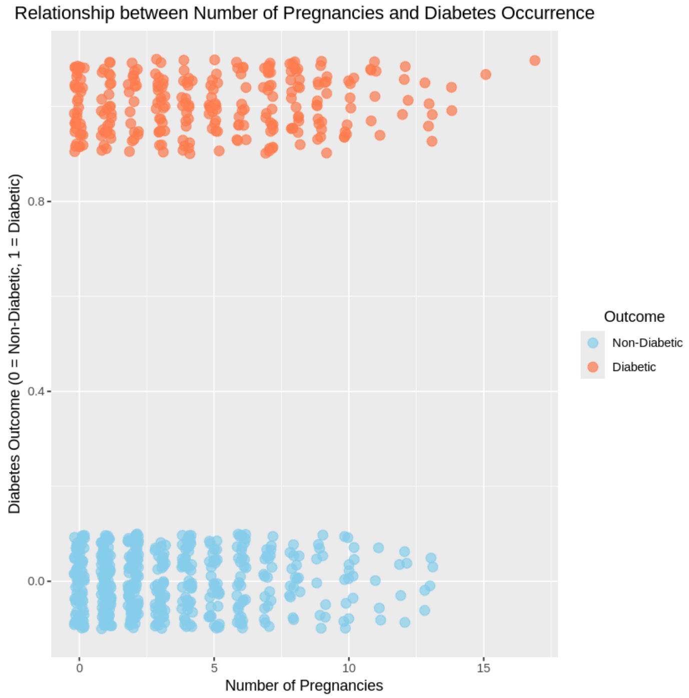


**Figure [4] : Relationship between the number of pregnancies and diabetes occurrence**

The following Figure explores the relationship between the number of pregnancies and the incidence of diabetes. It identifies two clear groups-one for non-diabetics and the other for diabetics. Although there is some overlap to indicate that multiple pregnancies do not always result in diabetes, a slight upward trend is seen in the diabetic group as the number of pregnancies increases. This is a trend that may indicate that the more pregnancies one has, the greater one's risk of diabetes.

- By displaying the Trend of glucose levels with age among diabetic and non-diabetic patients , we present the following figure:

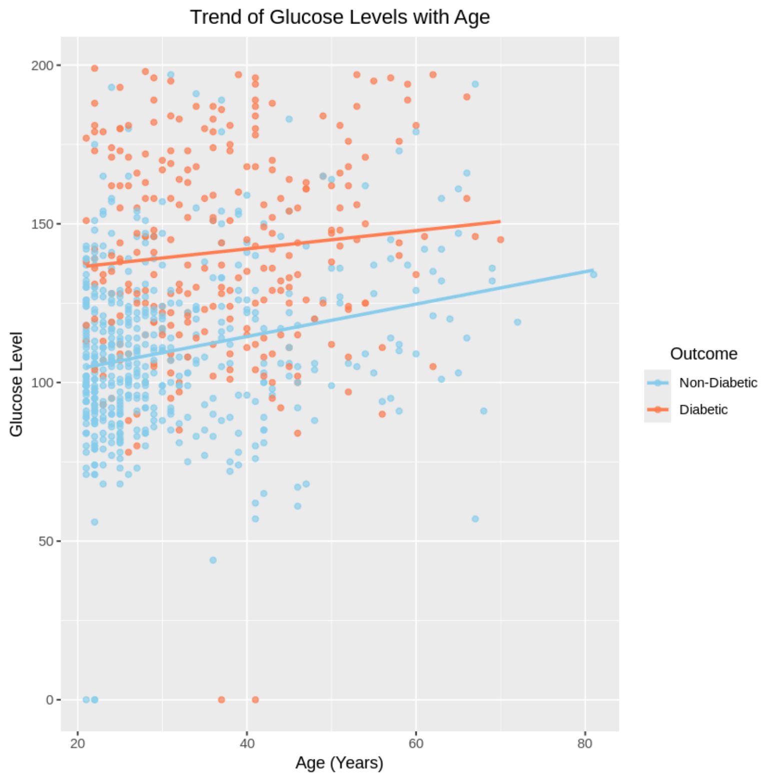


**Figure [5] : Trend of glucose levels with age among diabetic and non-diabetic patients**

The following figure examines the impact of age on glucose levels for diabetic and non-diabetic individuals. There are two trend lines: a very gradual upward slope for non-diabetics and one for diabetics that goes up much more steeply, reflecting that for diabetic individuals, glucose levels tend to rise more steeply with age. Whereas these lines show the trend for the majority, the scattering of points indicates that age is not the sole factor in determining glucose levels , there is an intersection between the two groups: high levels of glucose do not always indicate diabetes, and vice versa.

-

- By displaying the Trend of glucose levels with age among diabetic and non-diabetic patients , we present the following figure:

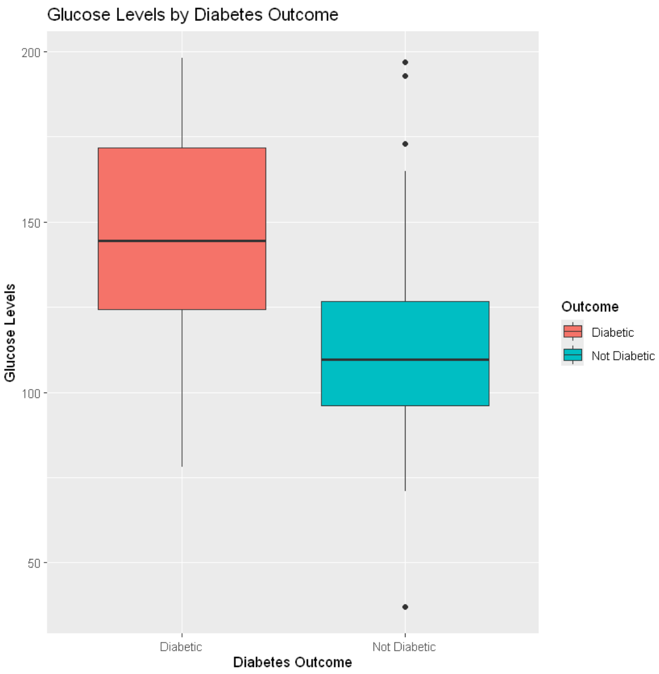


**Figure [6] : Box Plot of glucose levels with age among diabetic and non-diabetic patients**

The boxplot clearly shows that glucose levels tend to be higher in the diabetic group compared to the non-diabetic group. The diabetic group also has a wider spread of glucose levels and some high outliers. The interquartile range (IQR) is notably higher for diabetic patients, suggesting that higher glucose levels are associated with diabetes.

- By displaying the Distribution of Glucose Levels Among diabetic and non-diabetic patients, we present the following figure:

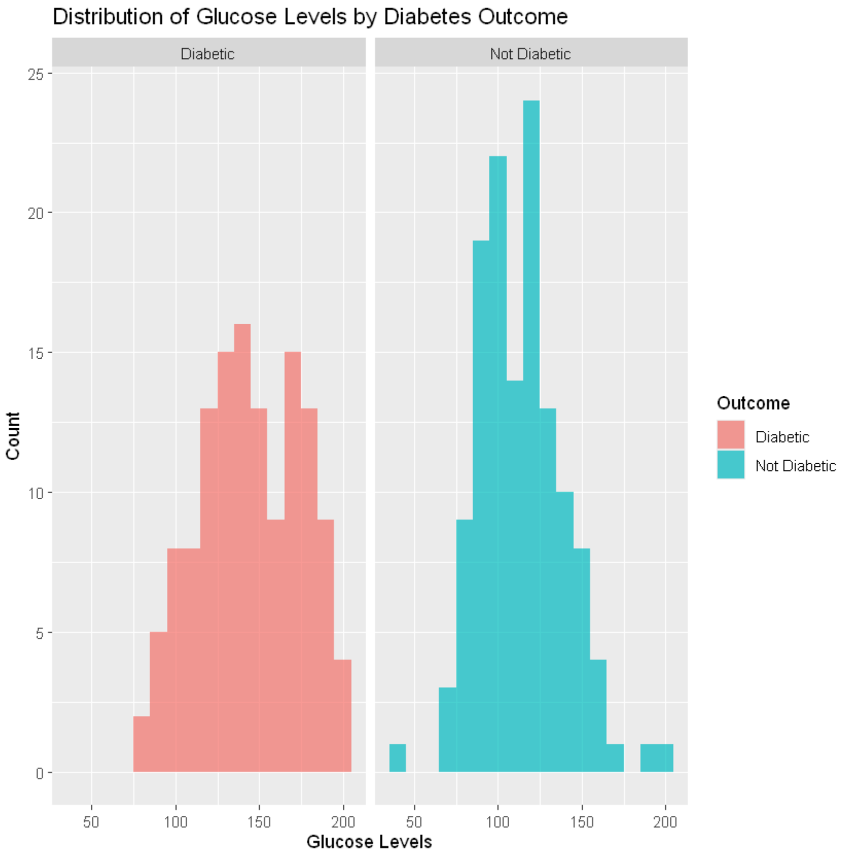


**Figure [7] : Two Histograms for diabetic patients according to their Glucose Levels.**

The following figure has two histograms, one for diabetic patients and one for non-diabetic patients, illustrating the distribution of their glucose levels. The x-axis represents glucose levels, and the y-axis represents the count of individuals within each glucose level range. The diabetic group's histogram is skewed to the right, which indicates a higher number of individuals with high glucose levels, while the non-diabetic group's histogram is more symmetrical.

- By displaying the scatterplot of Glucose Levels vs BMI Regression line , we present the following figure:

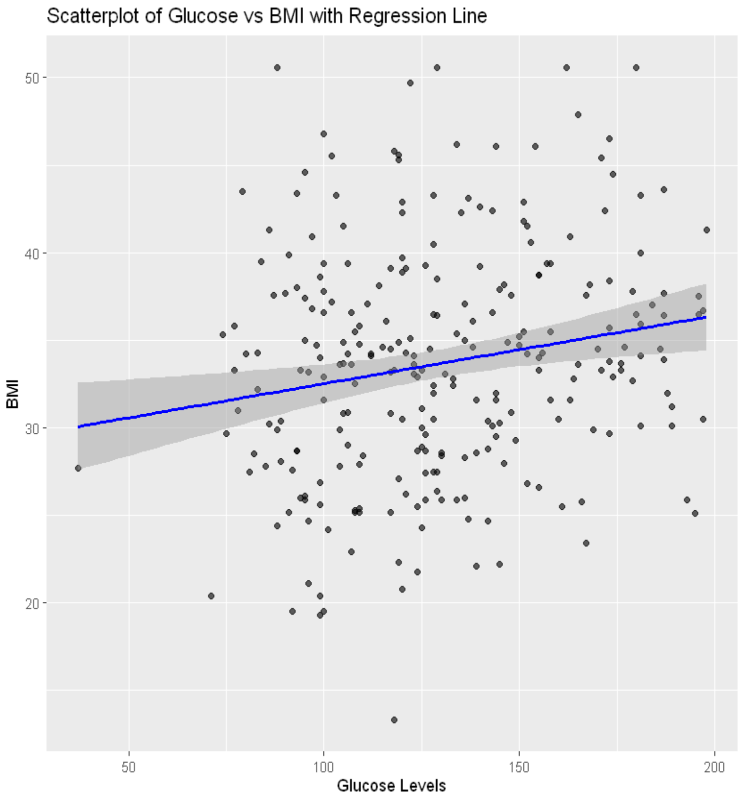


**Figure [8] : Scatterplot for glucose levels and Body Mass Index (BMI)**

This figure shows that the scatterplot suggests a weak positive correlation between glucose levels and BMI. While there's a slight tendency for BMI to increase with higher glucose levels, the relationship is not strong, and other factors likely play a more significant role in determining BMI. The wide spread of data points and the relatively flat slope of the regression line emphasize the weakness of this association.

By displaying the relation between pregnancies and diabetes outcome, we present the following figure:

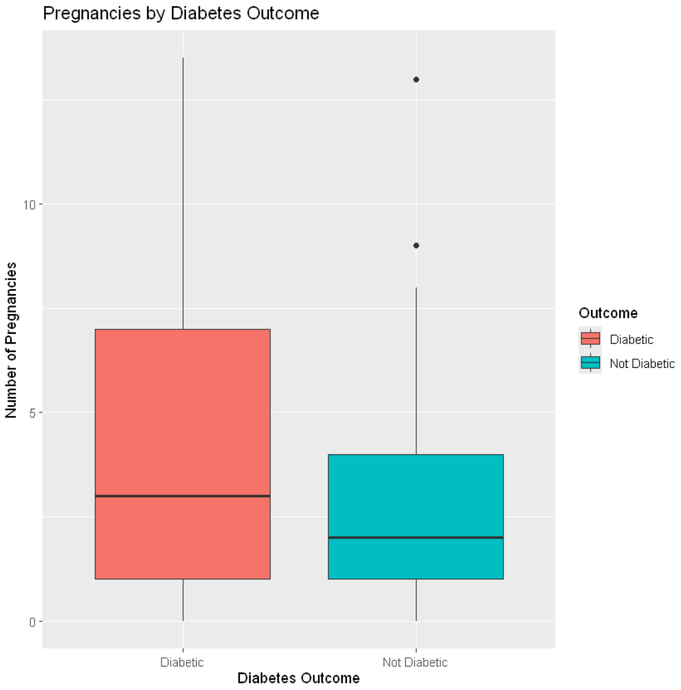


**Figure [9] : Boxplot for Number of pregnancies and Diabetes Outcome**

The box plot compares the distribution of the number of pregnancies between diabetic and non-diabetic individuals. The x-axis represents the diabetes outcome, and the y-axis indicates the number of pregnancies. The box plot shows a higher median number of pregnancies compared to the non-diabetic group. The plot suggests that there is an association between diabetes and a higher number of pregnancies.

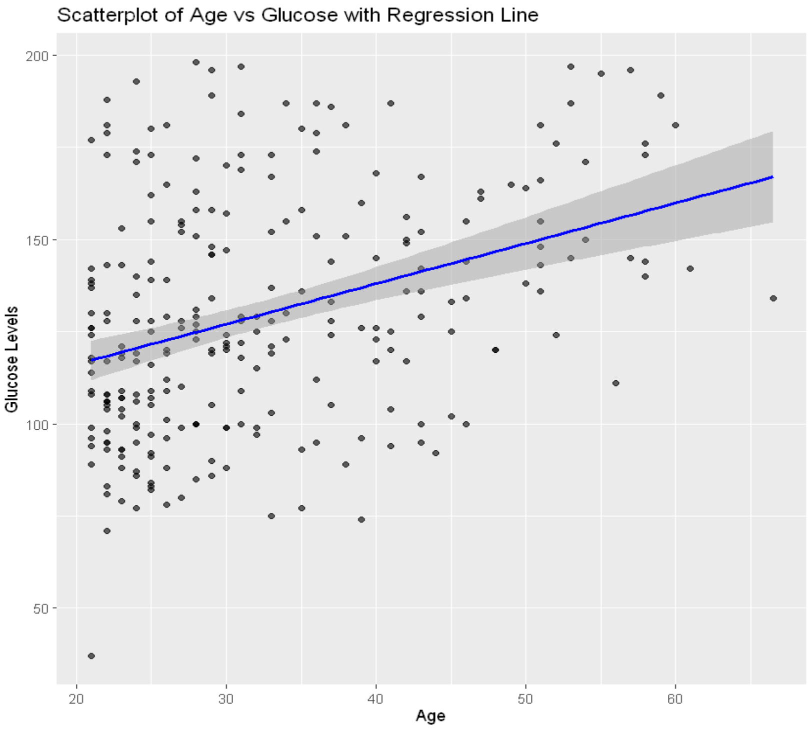


**Figure [10] : Scatterplot for the relationship between age and glucose levels**

The scatterplot suggests a positive correlation between age and glucose levels. While there is still some variability, the upward trend of the regression line is more pronounced than in the BMI vs. Glucose plot, implying a stronger relationship. This observation aligns with general medical knowledge that glucose tolerance can decrease with age.

- By displaying the Blood Pressure among diabetic and non-diabetic patients, we presented the following figure:

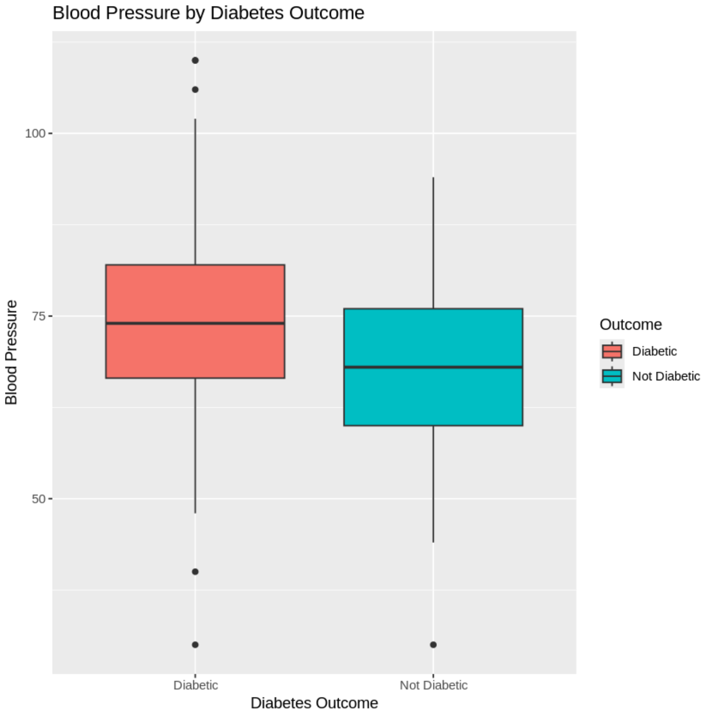


**Figure [11.1] : Box Plot for Blood Pressure among diabetic and non-diabetic patients**

The Following  Figure describes the relationship between diabetes and blood pressure. It can be observed that generally, persons with diabetes have higher median blood pressures than persons who do not have diabetes. Also, the box for diabetics is slightly wider, showing greater variability in blood pressures measured for the group with diabetes. There are only a few outliers in the two groups, which include persons with far higher or even lower blood pressures compared to the usual.

- By displaying the Blood Pressure among diabetic and non-diabetic patients , we presented the following figure:

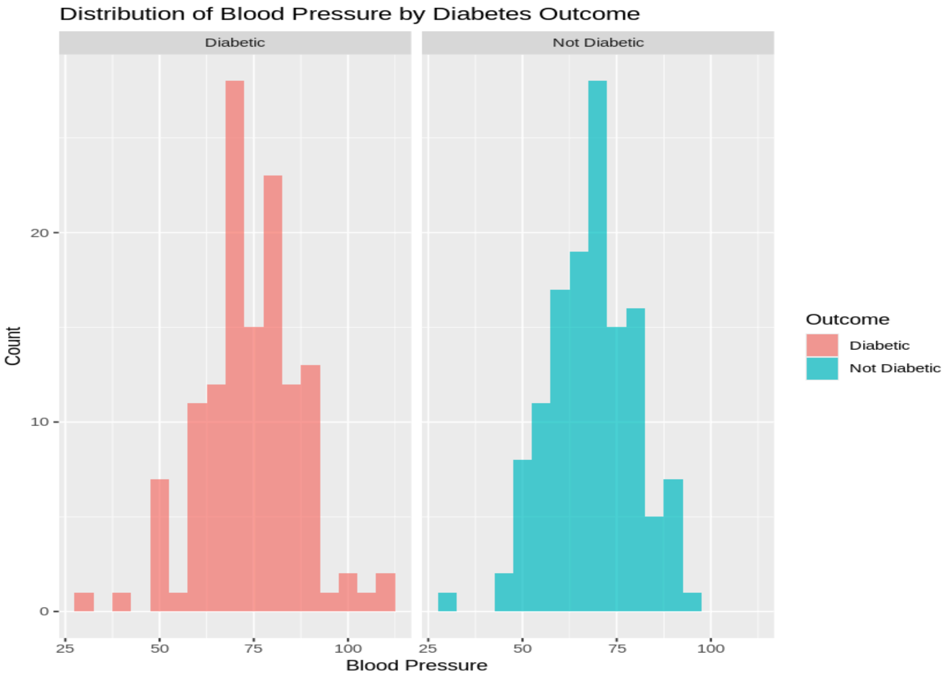


**Figure [11.2] : Distribution for Blood Pressure among diabetic and non-diabetic patients**

This histogram shows the distribution of blood pressure for both diabetes and non-diabetes patients. Both distributions are somewhat bell-shaped but clearly shifted to the right. Thus, individuals with diabetes tend to have higher blood pressure than those without diabetes. There is considerable overlap in the two distributions so high blood pressure is possible even for an individual without diabetes. This underlines that high blood pressure is not restricted to people with diabetes and gives a strong basis for a comprehensive approach to blood pressure management in all.

- By displaying the relationship between Insulin and BMI , we presented the following figure:

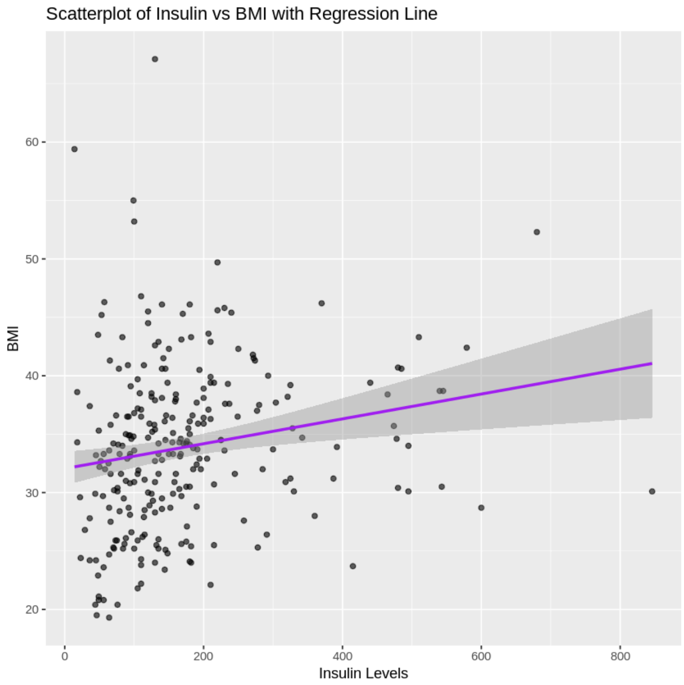


**Figure [12] : Relationship Between Insulin and BMI with Regression Line**

This scatterplot displays the relation between insulin levels and Body Mass Index. This graph presents a great deal of data points that are the results of individual observations of corresponding insulin levels and BMI values. A fitted regression line has been conducted on this data, suggesting a positive trend between insulin levels and BMI. With the increase in insulin levels, the general trend is that the BMI also goes up. However, the relation is not linearly perfect, and the scatter of the data indicates that the determining variables for BMI are other than insulin levels. The grey shaded area around the regression line likely depicts the confidence interval and gives the range of the predicted values of BMI over given insulin levels.

- By displaying the relationship between Age and Glucose Level , we presented the following figure:

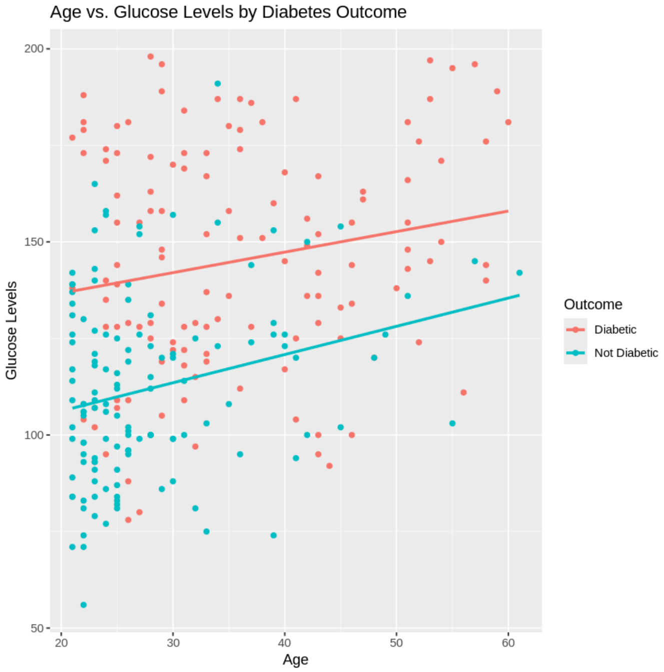


**Figure [13.1] : Relationship Between Age and Glucose Level among diabetic and non-diabetic patients**

This scatterplot explores the relationship between age and glucose levels for both people with and without diabetes. It reflects two trend lines-one for non-diabetics with a gentle upward slope and another for diabetics with a steeper upward slope, indicating that glucose levels tend to rise more significantly with age in people with diabetes. These lines, while giving the general trend of data, the scatter gives an indication that age isn't the only influencing variable on glucose levels. It also overlaps between groups, bringing into light that high glucose level does not always point toward diabetes and vice versa. This chart shows the relation between age and glucose levels, with special consideration to those affected by diabetes, while individual variability should be considered.

- By displaying the relationship between Age and Diabetes Outcome , we presented the following figure:

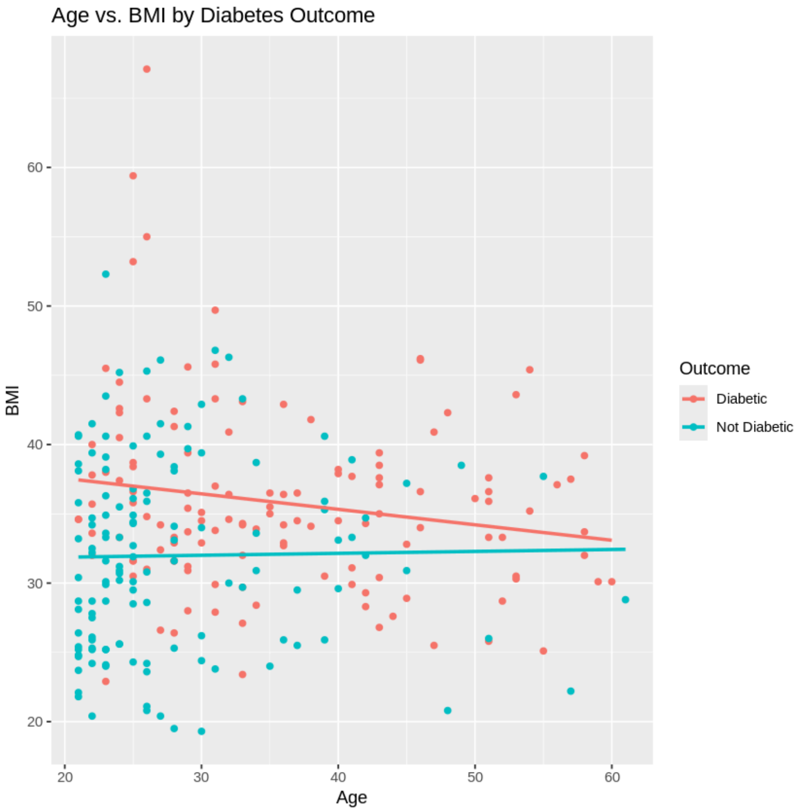


**Figure [13.2] : Relationship Between Age and Diabetes Outcome among diabetic and non-diabetic patients**

This chart compares age versus BMI for individuals with and without diabetes. This chart reflects two trend lines: a very slight downward slope for the non-diabetic group and a slightly greater downward slope in the diabetic group, thus suggesting perhaps a minimal decline in BMI with age increase in both groups. However, the data points are all over the lines. Therefore, age seems not to be one of the major factors determining BMI. Other factors probably play a much more important role in the variations of BMI observed in this dataset.

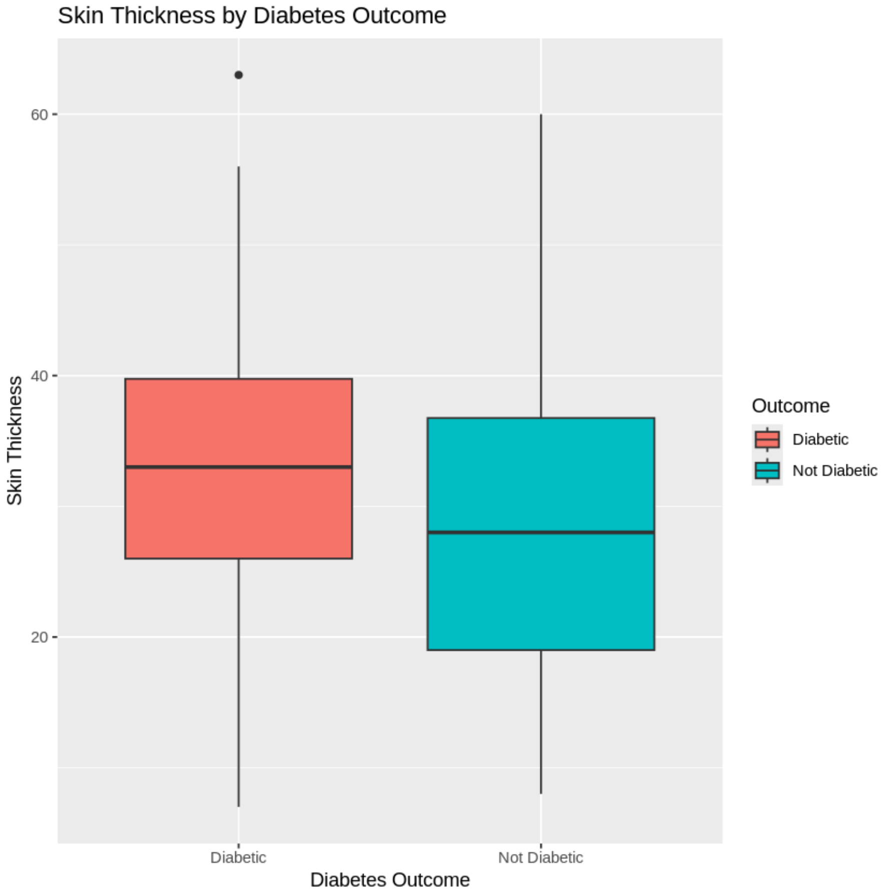


**Figure [14.1] : Skin thickness between individuals with and without diabetic outcome**

This boxplot suggests that skin thickness is not substantially different between individuals with and without a diabetic outcome. The medians and IQRs are similar, indicating comparable distributions. The presence of a few outliers in both groups is the main point of variation, but it doesn't represent a systematic difference between the groups.

- By displaying the distribution  Skin Thickness among diabetic and non-diabetic patients , presented the following figure:

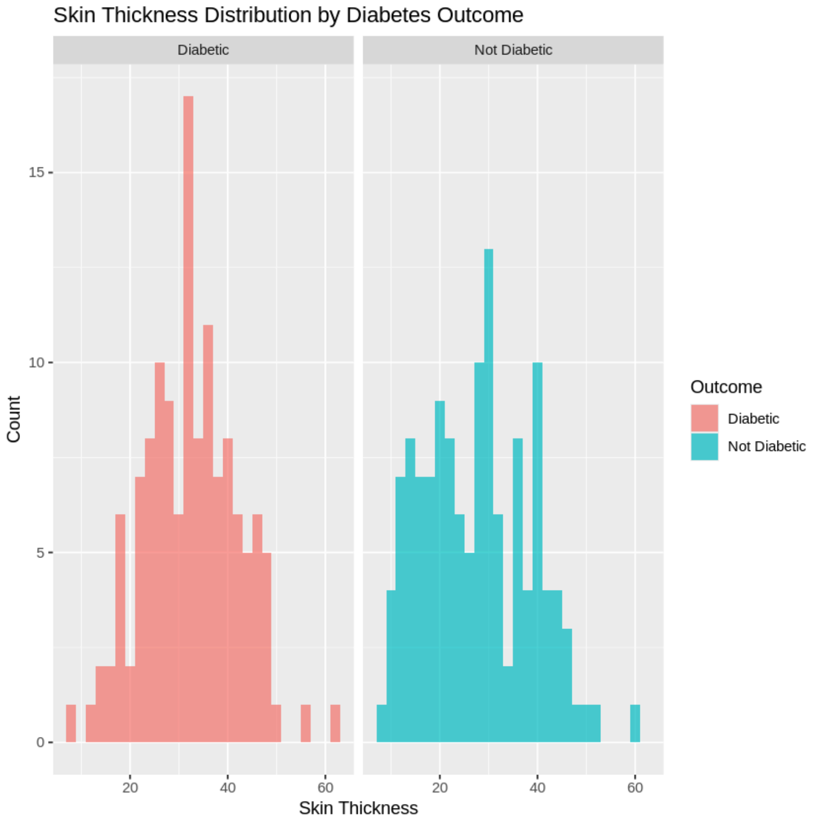


**Figure [14.2] : Distribution for Skin Thickness among diabetic and non-diabetic patients**

This chart presents the distribution of skin thickness measurements for individuals with and without diabetes. Both distributions are right-skewed, with more individuals having lower skin thickness values. However, the distribution for individuals with diabetes appears shifted slightly to the right compared to those without diabetes, suggesting a higher average skin thickness in the diabetic group. Despite this difference, there is overlap between the distributions, indicating that higher skin thickness can be observed in both individuals with and without diabetes.

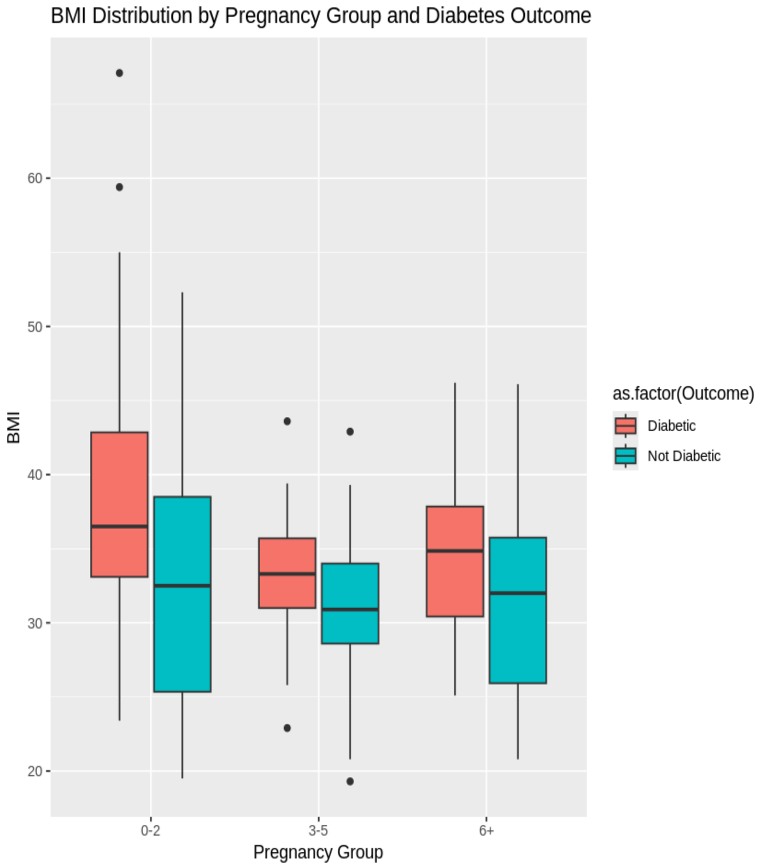


**Figure [15.1] : Distribution of BMI across different pregnancy groups, divided by diabetes outcome**

This plot shows that BMI is consistently higher in individuals with a diabetic outcome, regardless of the number of pregnancies. The number of pregnancies doesn't seem to have a strong independent effect on BMI, especially within the diabetic group. While there is a slight suggestion of decreasing BMI with increasing pregnancies in the non-diabetic group, the difference is small. The primary takeaway is the consistent difference in BMI based on diabetes outcome across all pregnancy groups.

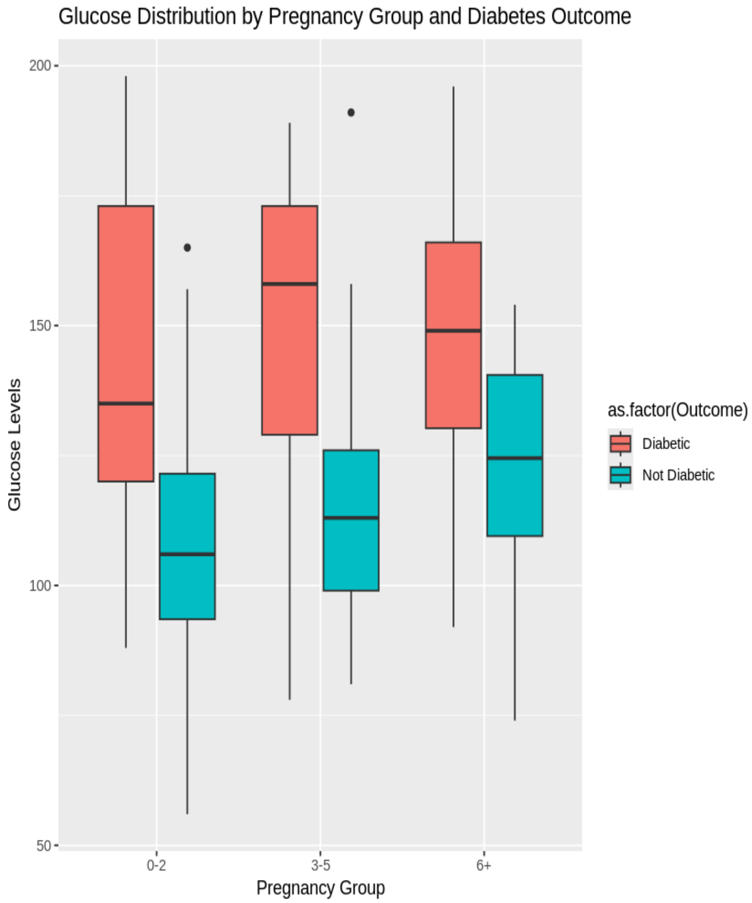


**Figure [15.2] : Distribution of glucose levels across different pregnancy groups, further broken down by diabetes outcome**

This plot indicates that diabetes outcome is the primary factor influencing glucose levels, with diabetic individuals having substantially higher glucose levels than non-diabetic individuals, regardless of the number of pregnancies. The number of pregnancies has a much weaker influence on glucose levels. There might be a slight increase in glucose with more pregnancies in the non-diabetic group, but the effect is small compared to the difference between diabetic and non-diabetic individuals

### 2.3 Confidence Interval Analysis for the Glucose Column

We analyzed the glucose column in the dataset due to its correlation with the presence of diabetes. We applied a 95% confidence interval using a different number of samples and sample size.

To calculate the mean of our samples we used this formula:

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}$$

To calculate the standard deviation of our sample we used this formula:

$$\sigma = \sqrt{\frac{\sum_{i=0}^{n} (x_i - \bar{x})^2}{n - 1}}$$

To calculate the confidence interval we  used this formula:

$$CI = \bar{x} \pm t_{\alpha,df} \times \frac{\sigma}{\sqrt{n}}$$

The interpretation of the mathematical symbols is as following:

- $\bar{x}$: Sample mean
- $x_i$: Sample element
- $n$: Sample size
- $\sigma$: Standard deviation
- $CI$: Confidence interval
- $\alpha$: Alpha of confidence
- $df$: Degree of freedom
- $t$: T-Score at given alpha and degree of freedom

In the 25 samples with size 15, we got a mean width of confidence interval of 26.78511 and 84% of samples' confidence intervals (21 out of 25 samples) contained the true mean. For sample size 100 (25 samples), the mean width of confidence interval was 27.27171, and 100% of intervals (25 out of 25) contained the true mean. For sample size 10 (20 samples), the mean width of confidence interval was 25.67561, and 90% of intervals (18 out of 20) contained the true mean. Observations show that larger sample sizes improve the likelihood of capturing the true mean, while the interval width does not consistently decrease with increasing sample size.

## 3. Challenges, Limitations, and Assumptions

### 3.1 Discussion of challenges faced and solutions implemented

The distribution of the Diabetes pedigree function (DPF) is right-skewed so we could use the natural logarithm as a solution for the data.

### 3.2 Explanation of limitations in data or methodology and assumptions made

We have 2 hypotheses:

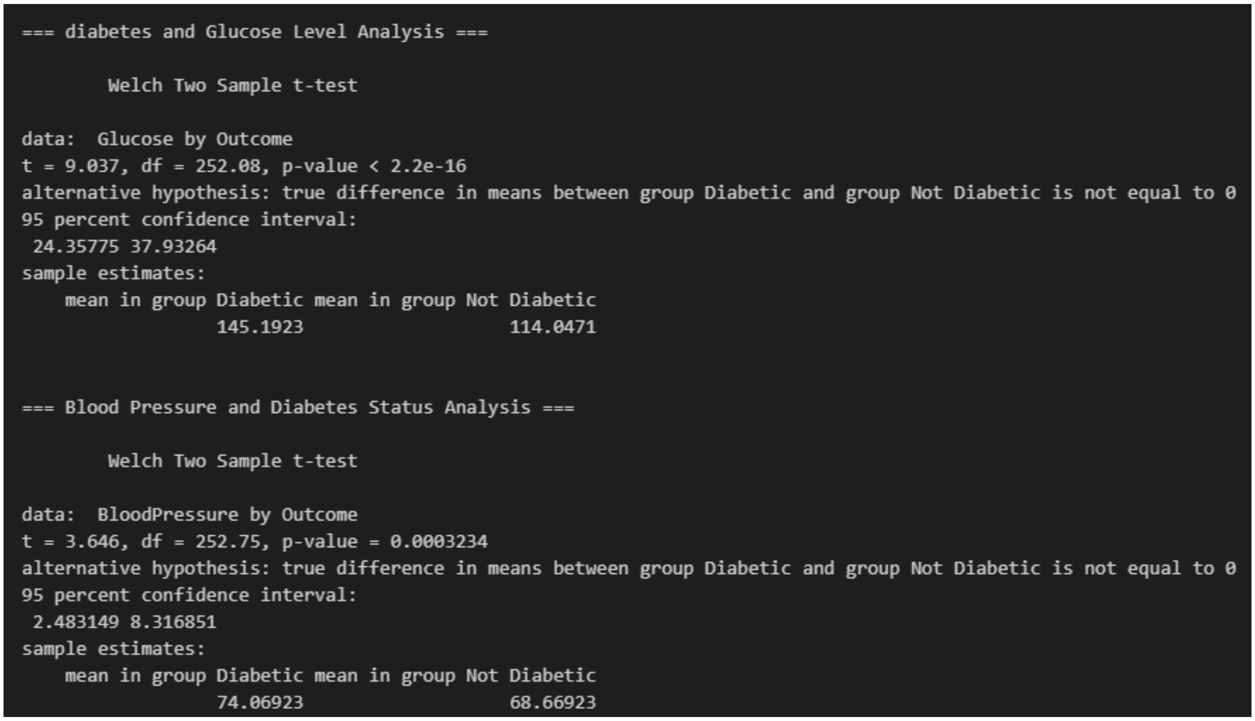
1. **Claim + Hypothesis:** There is a significant difference in glucose levels between diabetic and non-diabetic patients.
    1.1. **Justification:** The t-test is appropriate for comparing the means of two groups, which is the case here (diabetic vs. non-diabetic patients). The t-test assumes that the data is normally distributed, which is a reasonable assumption given the large sample size.
    **1.2. Hypothesis:**
        1.2.1. H0: $\mu_{\text{diabetic}} = \mu_{\text{non-diabetic}}$ → There is no significant difference in glucose levels between diabetic and non-diabetic patients.
        1.2.2. H1: $\mu_{\text{diabetic}} \neq \mu_{\text{non-diabetic}}$ → There is a significant difference in glucose levels between diabetic and non-diabetic patients.
2. **Claim + Hypothesis:** patients with high blood pressure are diabetic.
    2.1. **Justification:** The t-test is appropriate for comparing the means of two groups, which is the case here (diabetic vs. non-diabetic patients). The t-test assumes that the data is normally distributed, which is a reasonable assumption given the large sample size.

## 2.2. Hypothesis:

2.2.1. H0: $\mu_{\text{diabetic}} = \mu_{\text{non-diabetic}}$ → There is no difference in blood pressure between diabetic and non-diabetic patients.

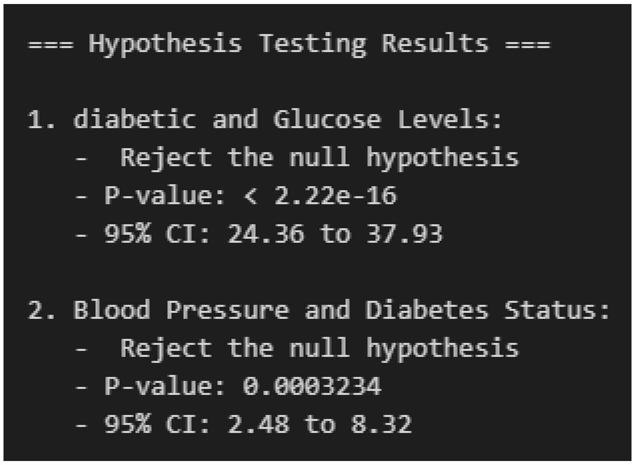
2.2.2. H1: $\mu_{\text{diabetic}} \neq \mu_{\text{non-diabetic}}$ → Patients with high blood pressure are diabetic.

3.



```
=== diabetes and Glucose Level Analysis ===

        Welch Two Sample t-test

data:  Glucose by Outcome
t = 9.037, df = 252.08, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Diabetic and group Not Diabetic is not equal to 0
95 percent confidence interval:
 24.35775 37.93264
sample estimates:
    mean in group Diabetic mean in group Not Diabetic
                  145.1923                   114.0471


=== Blood Pressure and Diabetes Status Analysis ===

        Welch Two Sample t-test

data:  BloodPressure by Outcome
t = 3.646, df = 252.75, p-value = 0.0003234
alternative hypothesis: true difference in means between group Diabetic and group Not Diabetic is not equal to 0
95 percent confidence interval:
 2.483149 8.316851
sample estimates:
    mean in group Diabetic mean in group Not Diabetic
                  74.06923                   68.66923
```

4.



```
=== Hypothesis Testing Results ===

1. diabetic and Glucose Levels:
   -  Reject the null hypothesis
   - P-value: < 2.22e-16
   - 95% CI: 24.36 to 37.93

2. Blood Pressure and Diabetes Status:
   -  Reject the null hypothesis
   - P-value: 0.0003234
   - 95% CI: 2.48 to 8.32
```

# 4. Conclusion

## 4.1 Summarize the findings and insights derived from the analysis

- Glucose Level is the most important factor, high glucose level is highly related to diabetic patients. It is observed from the distribution and box plots that the glucose level of the diabetic patient is much higher compared to the nondiabetic patients.
- **Age vs Glucose**: Glucose level increases with increasing age, especially for diabetic patients. So it can be concluded that age is also a factor contributing to glucose intolerance and development of diabetes.
- In this connection, it is to be pointed out that the diabetic patients have a higher BMI. Although glucose and BMI are weakly positively correlated, yet the latter was found to be a strong marker of diabetic tendency in the population under study.
- Diabetes Pedigree Function was clearly correlated with diabetes risk. The higher the value of DPF, the greater the chance of being diabetic which indicates the role of family history in diabetes susceptibility.
- **Pregnancy versus Diabetes:** There is a trend, which can be analyzed between the number of pregnancies and diabetes. Though this is not the major predictor, it does indicate that women who have had multiple pregnancies have a greater chance of having diabetes.
- **Blood Pressure:** The diabetic subjects have higher blood pressure on average compared to the non-diabetic subjects. This indicates that management of blood pressure should be considered in diabetic populations.
- **Insulin and BMI:** There was a positive trend seen between insulin levels and the BMI. This may be interpreted to mean that individuals with higher BMI need higher insulin, which is also helpful in metabolic responses.
- **Skin Thickness:** Although there was some variation, this variable did not seem to significantly discriminate between diabetic versus nondiabetic subjects in this particular study.
- **Confidence Interval Analysis:** The greater the sample sizes, the greater the percentage of the confidence intervals which contained the true population mean, whereas the width of the

interval did not always reduce with the increase in sample sizes, suggesting sample variability could be a big player in the width of the confidence interval.

### 4.2 Suggest future directions or potential improvements

- **Hypothesis Testing**: It conducts the formal hypothesis testing using t-tests or ANOVA in order to establish the feature-wise differences among diabetic and nondiabetic subjects observed in this study. These will provide statistical evidence for features with a significant difference.
- **Machine Learning Models:** Construction of machine learning models to predict diabetes based on the features present. A number of algorithms will be explored, including logistic regression, support vector machines, and neural networks. Performance metrics such as precision, recall, F1-score, and AUC will be used to evaluate these models.
- **Longitudinal Analysis:** If the data is longitudinal, perform time-series analysis to track changes in these factors over time for diabetic and non-diabetic subjects. This will be very informative about the development of the condition and those factors that influence this.