Diabetes Data Analysis Report

December 25, 2024

Team Member	ID
Hothifa Hamdan	202201792
Mazen Khaled	202201534
Toqa Elmasry	202201868
Layla Mohammad	202201906
Nehal Mohamed	202202020

Contents

1 Introduction

Diabetes is a chronic metabolic condition characterized by elevated levels of blood glucose, posing significant risks for cardiovascular diseases, nerve damage, and other severe complications. According to global health organizations, the incidence of diabetes has risen sharply, demanding effective strategies for both prevention and management.

In this project, our primary objective is to investigate a specific diabetes dataset to uncover patterns, risk factors, and statistical relationships that may inform healthcare practices. The dataset contains features such as **Glucose**, **BMI**, **Insulin**, **Blood Pressure**, and other clinical/demographic measurements alongside a binary *Outcome* variable indicating diabetes status.

1.1 Problem Statement and Objectives

The principal goals of this project include:

- **Identifying Key Risk Factors:** Assess which clinical and demographic variables (e.g., glucose, BMI, age) are most strongly associated with diabetes.
- **Comparative Analysis:** Determine significant differences in key metrics between diabetic and non-diabetic groups (e.g., glucose, BMI, blood pressure).
- **Hypothesis Testing:** Formally test statistical claims regarding differences in mean values or proportions related to diabetes.
- **Confidence Interval Simulation:** Examine coverage and behavior of confidence intervals across various sample sizes.

2 Data Processing and Analysis Steps

2.1 Data Collection and Cleaning

The dataset was obtained from a medical repository containing anonymized patient records. Key steps in cleaning included:

- **Zero to NA Conversion:** Certain columns (e.g., `Glucose`, `BloodPressure`, `BMI`, `SkinThickness`, `Insulin`) had biologically implausible values of 0, which were replaced with `NA`.

- **Missing Data Imputation:** We used median imputation to fill in missing values, thus reducing bias without overly distorting the data.
- **Verification:** Summary statistics (`str`, `summary`) were checked to ensure the distribution remained realistic post-imputation.

2.2 Exploratory and Descriptive Analyses

- **Descriptive Statistics:** We calculated mean, median, and standard deviation for each feature to understand typical ranges and variances.
- **Visualization:** Used box plots, histograms, and density plots to detect outliers, identify skew, and compare distributions between diabetic (Outcome = 1) and non-diabetic (Outcome = 0) groups.

3 Theoretical Framework and Hypothesis Testing

This section outlines the statistical methods used in the analysis and provides the mathematical background behind each test.

3.1 Hypothesis Testing Concepts

Hypothesis testing generally involves:

- Stating a **null hypothesis** (H_0), which typically represents the status quo or no difference.
- Stating an **alternative hypothesis** (H_a), which posits a deviation from the null (e.g., a difference in means).
- Computing a **test statistic** and **p-value** to determine whether to reject H_0 at a chosen significance level α (commonly $\alpha = 0.05$).

3.2 Two-Sample t-Test

To compare mean glucose levels between diabetic and non-diabetic patients, we employ an independent two-sample t-test. Suppose we have two groups:

- Group 1: Non-diabetic patients, sample size n_1, mean $\overline{X}_1$, variance s_1^2.
- Group 2: Diabetic patients, sample size n_2, mean $\overline{X}_2$, variance s_2^2.

The null and alternative hypotheses can be:

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_a : \mu_1 \neq \mu_2.$$

If normality assumptions and equal variance are (approximately) satisfied, the t-statistic is given by:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad \text{(Pooled standard deviation)}.$$

A non-pooled version (Welch's t-test) is used if variances are suspected to differ.

3.3 Two-Proportion z-Test

To compare the proportion of obese patients (BMI > 30) in diabetic versus non-diabetic groups, we use a two-proportion z-test:

$$H_0 : p_1 = p_2,$$
$$H_a : p_1 > p_2 \quad (\text{or } p_1 \neq p_2 \text{ depending on the research question}),$$

where p_1 is the true proportion of obese individuals in group 1 (diabetic), and p_2 is the true proportion in group 2 (non-diabetic). The test statistic is:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}},$$

where

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2},$$

x_1 is the number of obese individuals in group 1, and x_2 is the number of obese individuals in group 2.

3.4 Confidence Intervals and Coverage

A 95% confidence interval for a mean μ (assuming normality or large sample sizes) typically takes the form:

$$\overline{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}},$$

or, if σ is unknown,

$$\overline{X} \pm t_{\alpha/2,\, df} \times \frac{s}{\sqrt{n}},$$

where $t_{\alpha/2,\, df}$ is the critical value from the Student's t-distribution for the given degrees of freedom. In our simulations:

- We select different sample sizes (e.g., 10, 15, 100).
- For each sample, we compute the sample mean $\overline{X}$ and the interval bounds.
- We check whether the interval captures the **true population mean** from the entire dataset.

4 Challenges, Limitations, and Assumptions

4.1 Challenges

- **Handling Missing Data:** Some values needed careful imputation to avoid skewing the results.
- **Sample Diversity:** The dataset may not fully capture all demographic groups, limiting generalizability.

4.2 Limitations

- **Self-Reporting Bias:** Certain metrics (e.g., number of pregnancies) might rely on patient self-reporting and may be imprecise.

- **Confounding Variables:** Important factors (e.g., diet, genetics) might be omitted, affecting interpretability.

4.3 Assumptions

- The dataset is representative of broader populations at risk for diabetes.
- Median imputation for missing values retains overall distribution properties without major distortion.

5 Results and Visualizations

This section provides the core findings from our exploratory analysis and the relevant plots/graphs generated. We highlight several key investigations:

5.1 Glucose Levels by Diabetes Status

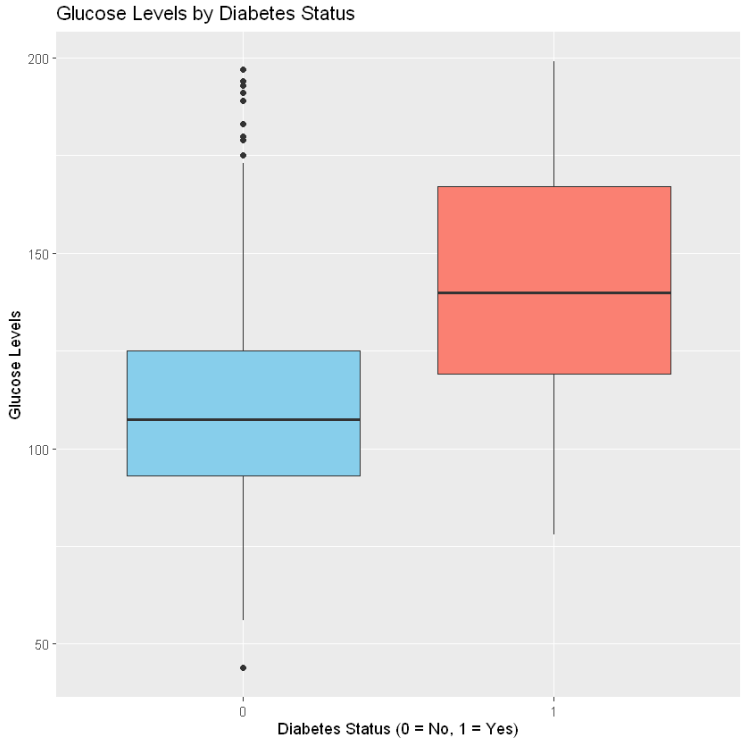

Figure 1: Box Plot of Glucose Levels by Diabetes Status (0 = No, 1 = Yes).

Interpretation: Diabetic individuals (Outcome = 1) show a visibly higher median glucose level, indicating a strong relationship between hyperglycemia and diabetes.

5.2 Age by Diabetes Status

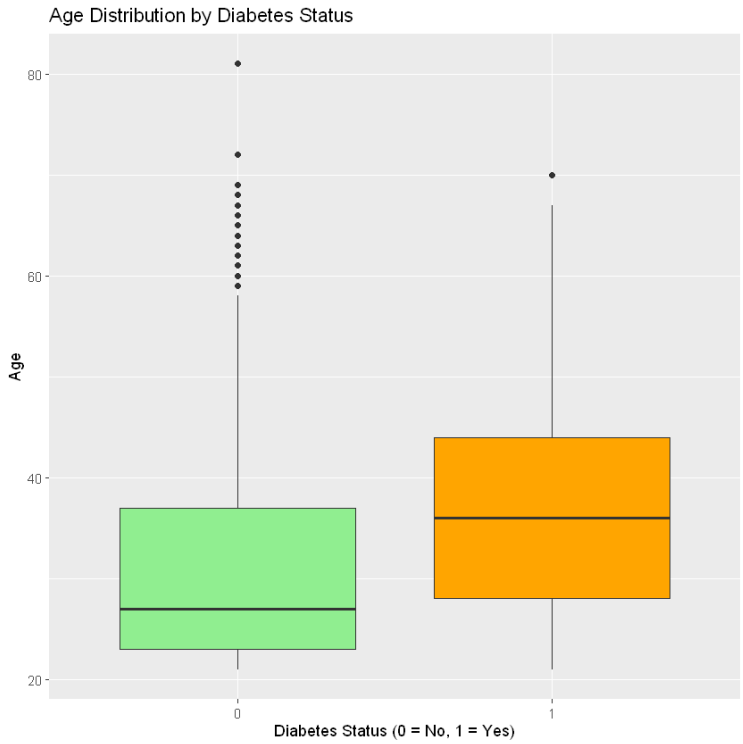

Figure 2: Box Plot of Age by Diabetes Status (0 = No, 1 = Yes).

Interpretation: The distribution of age suggests diabetic patients may be slightly older on average, though overlap exists.

5.3 Blood Pressure by Diabetes Status

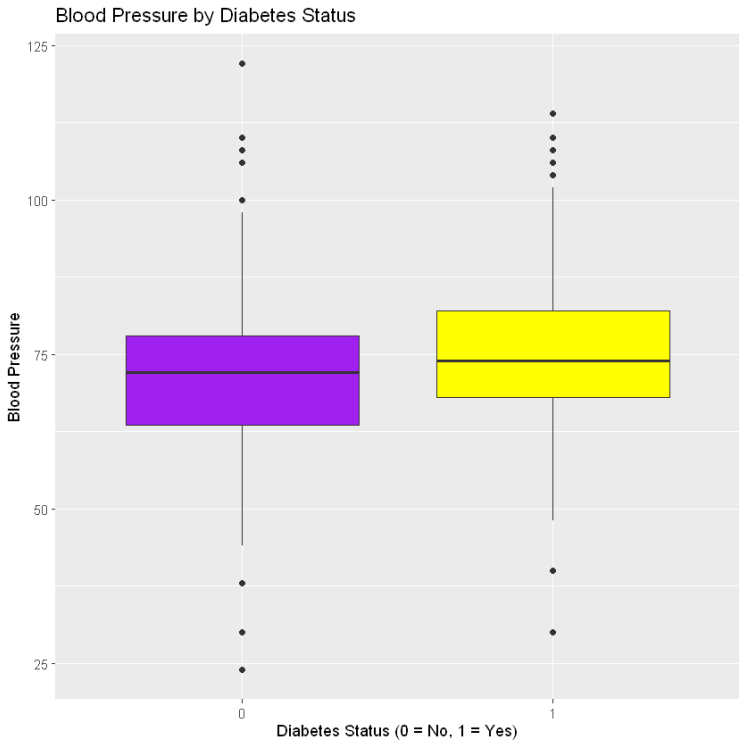

Figure 3: Box Plot of Blood Pressure by Diabetes Status (0 = No, 1 = Yes).

Interpretation: Average blood pressure appears somewhat higher in diabetic patients, though not as pronounced as glucose differences.

5.4 BMI by Diabetes Status

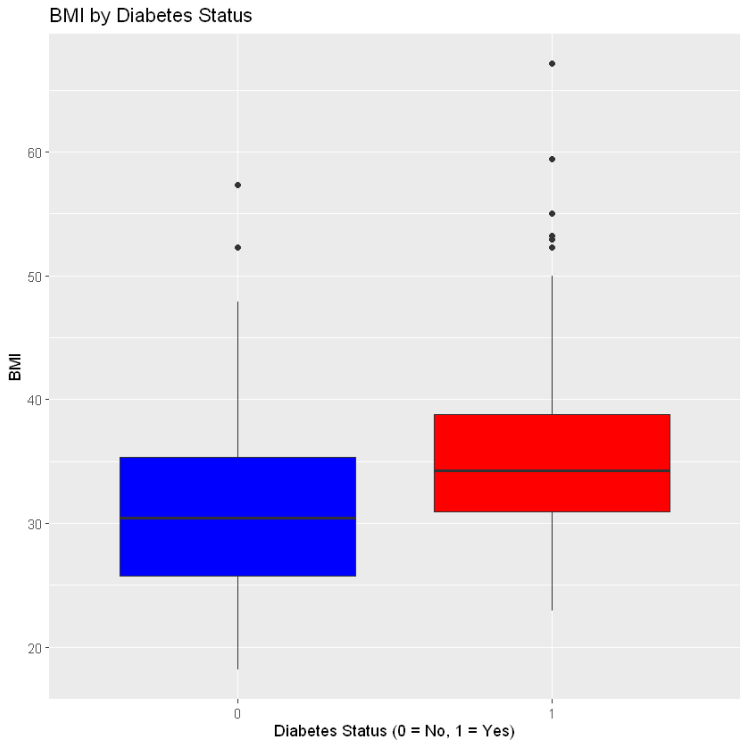

Figure 4: Box Plot of BMI by Diabetes Status (0 = No, 1 = Yes).

Interpretation: BMI is typically higher for diabetic patients, consistent with known links between obesity and insulin resistance.

5.5 Distribution of BMI (All Patients)

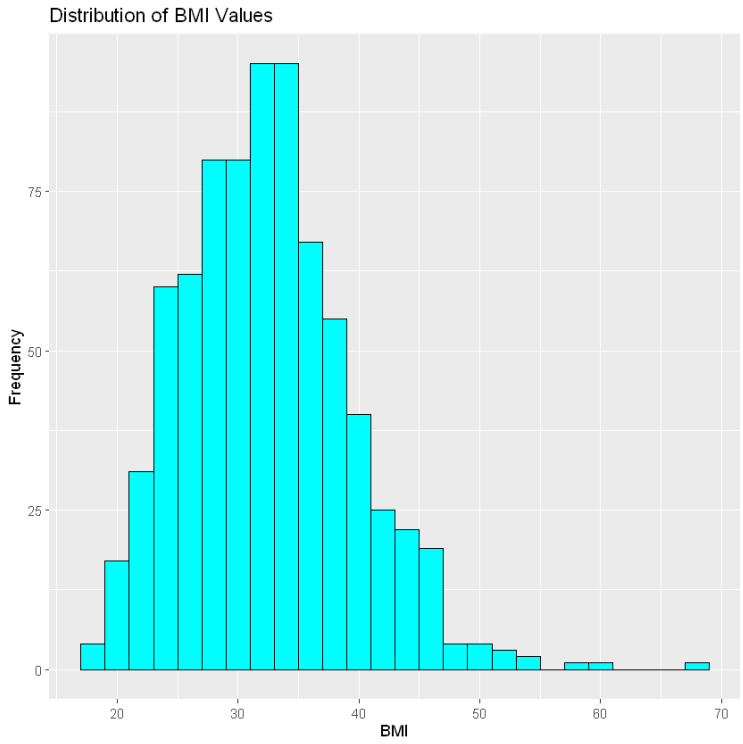

Figure 5: Histogram of BMI Values Among All Patients.

Interpretation: The distribution skews right, indicating many patients fall into overweight or obese ranges.

5.6 Relationship Between Pregnancies and Diabetes

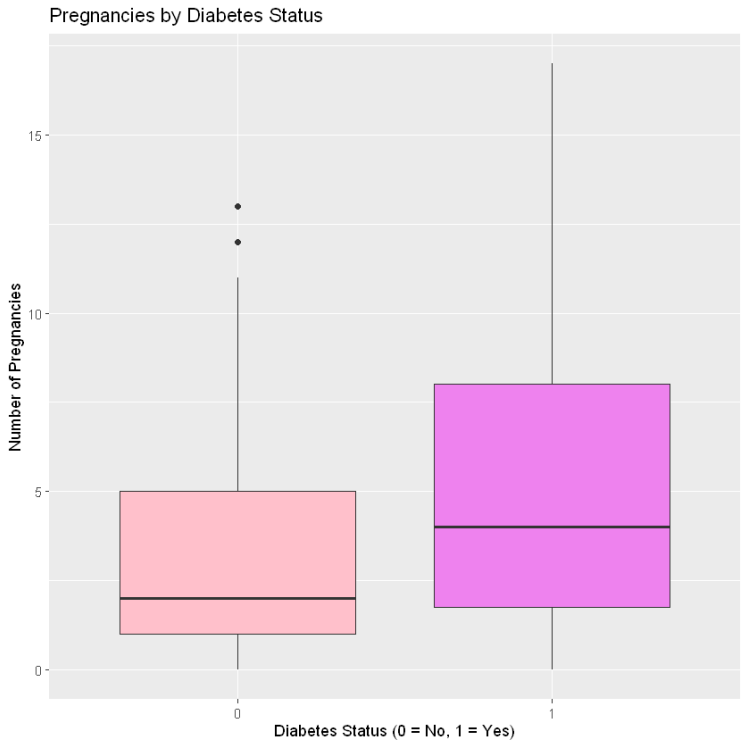

Figure 6: Box Plot of Number of Pregnancies by Diabetes Status.

Interpretation: On average, diabetic patients have had slightly more pregnancies, possibly hinting at gestational diabetes links or post-pregnancy metabolic changes.

5.7 Correlation Between Glucose and BMI

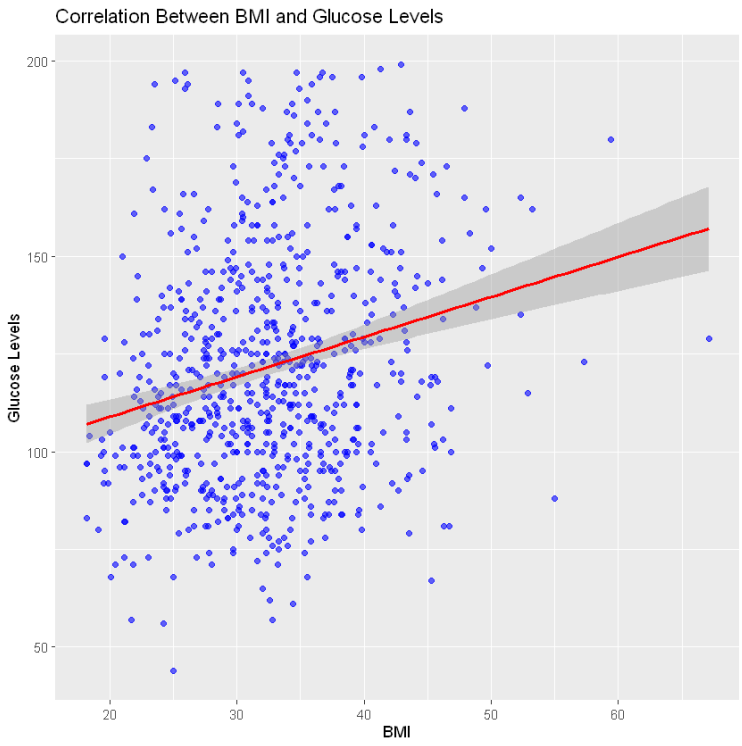

Figure 7: Scatter Plot of Glucose vs. BMI.

Interpretation: A modest positive correlation indicates that higher BMI often pairs with higher glucose levels, consistent with clinical knowledge of insulin resistance.

5.8 Age vs. Glucose Trend by Diabetes Status

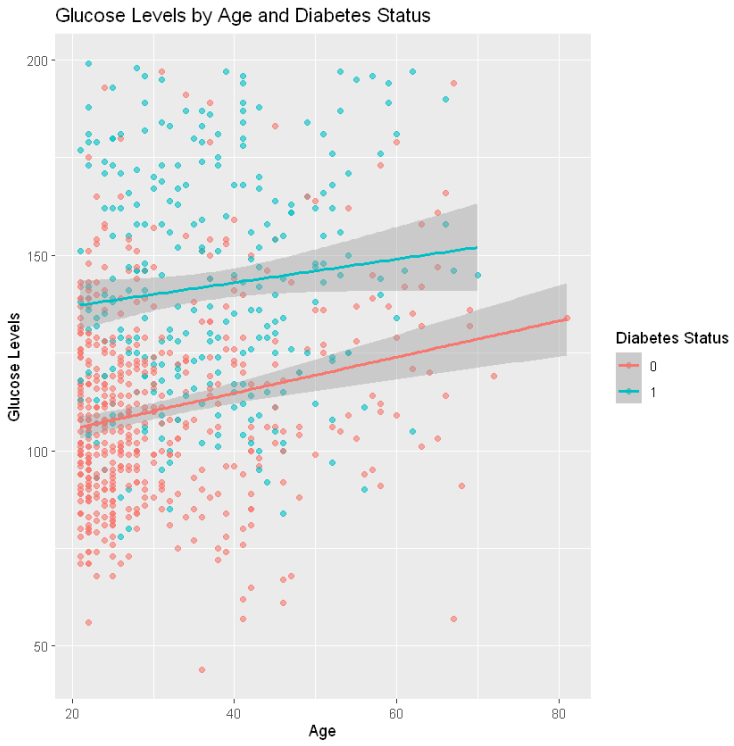

Figure 8: Scatter Plot of Age vs. Glucose, Color-Coded by Outcome (0 or 1).

Interpretation: A slight upward trend is seen, particularly among diabetic patients, suggesting that glucose can climb gradually with age.

6 Conclusion

6.1 Summary of Key Findings

- **Glucose Differences:** Diabetic patients exhibit significantly higher glucose levels than non-diabetic patients (verified by the t-test described in Section 3.2).
- **BMI and Obesity:** Significantly more diabetic patients are obese (BMI ¿ 30) than non-diabetic patients, confirmed by the two-proportion z-test (Section 3.3).
- **Age Factor:** Older patients show slightly higher glucose and insulin levels, though the effect size is less pronounced compared to BMI or glucose differences.
- **Confidence Intervals:** Simulations with varied sample sizes (Section 3.4) indicate that larger samples yield narrower 95% CIs and higher coverage of the true population mean.

6.2 Future Directions

- **Extended Variables:** Including dietary habits, activity levels, and family history could clarify additional risk factors.
- **Longitudinal Data:** Tracking how BMI, glucose, and insulin change over multiple years to better understand disease progression.
- **Predictive Modeling:** Developing logistic regression or other machine learning classifiers to predict diabetes risk more accurately.

Final Conclusion (Technical Overview)

Overall, the analyses confirm that hyperglycemia (high glucose levels) and elevated BMI are strong predictors and correlates of diabetes within this dataset. Hypothesis tests validate these findings statistically, and confidence interval simulations demonstrate how sample size impacts the reliability of our estimates. Despite certain limitations in the dataset (missing covariates, potential biases), the evidence robustly supports the need for focused interventions on weight management and regular glucose monitoring to mitigate diabetes risk.