Diabetes Dataset Analysis Report

Nadir NEHILI

August 13, 2025

1 Introduction

This report presents an exploratory analysis of the **Diabetes dataset** from Kaggle. The dataset contains 768 observations and 9 variables, with the goal of investigating factors related to type 2 diabetes occurrence.

2 Dataset Overview

The dataset includes the following columns:

- **Pregnancies** : Number of times pregnant
- **Glucose** : Plasma glucose concentration
- **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** : Diabetes pedigree function
- **Age** : Age in years
- **Outcome** : Target variable (0 = Non-Diabetic, 1 = Diabetic)

3 Missing / Invalid Values Before Cleaning

Some columns contain physically impossible zero values, which should be treated as missing for analysis. The counts of missing or invalid values are presented in Table 1.

Column	Missing / Invalid Values
Pregnancies	0
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0
Outcome	0

Table 1: Number of missing or invalid values per column before cleaning.

4 Data Cleaning Method

To handle invalid values, all zeros in the affected columns (`Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`) were replaced with the **median** of the respective column. This preserves the distribution of the data while correcting unrealistic entries.

After cleaning, the dataset contains no zeros in the previously affected columns, as summarized in Table 2.

Column	Zero Values Remaining
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0

Table 2: Number of zero values remaining after cleaning.

The cleaned dataset has been exported as `data/diabetes_cleaned.csv` for further analysis.

5 Exploratory Data Analysis

5.1 Boxplots by Outcome

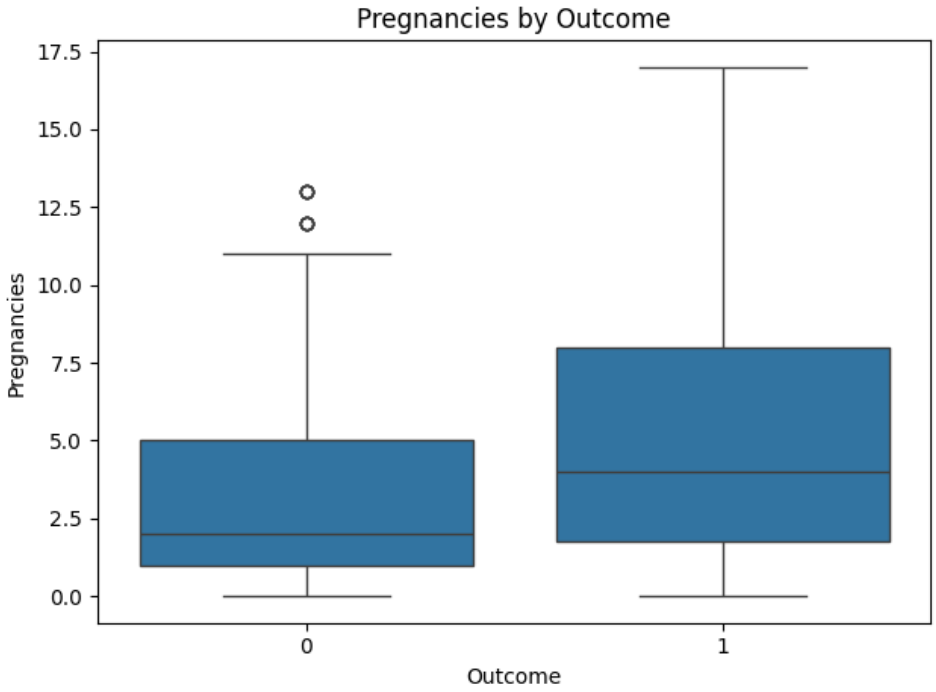

Figure 1: Pregnancies by Outcome

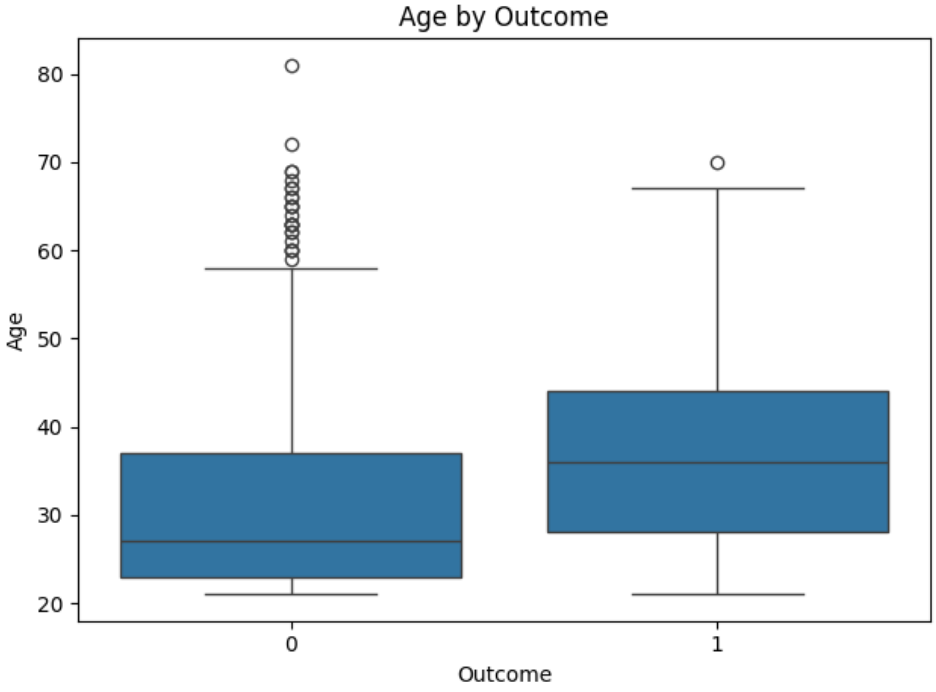

Figure 2: Age by Outcome

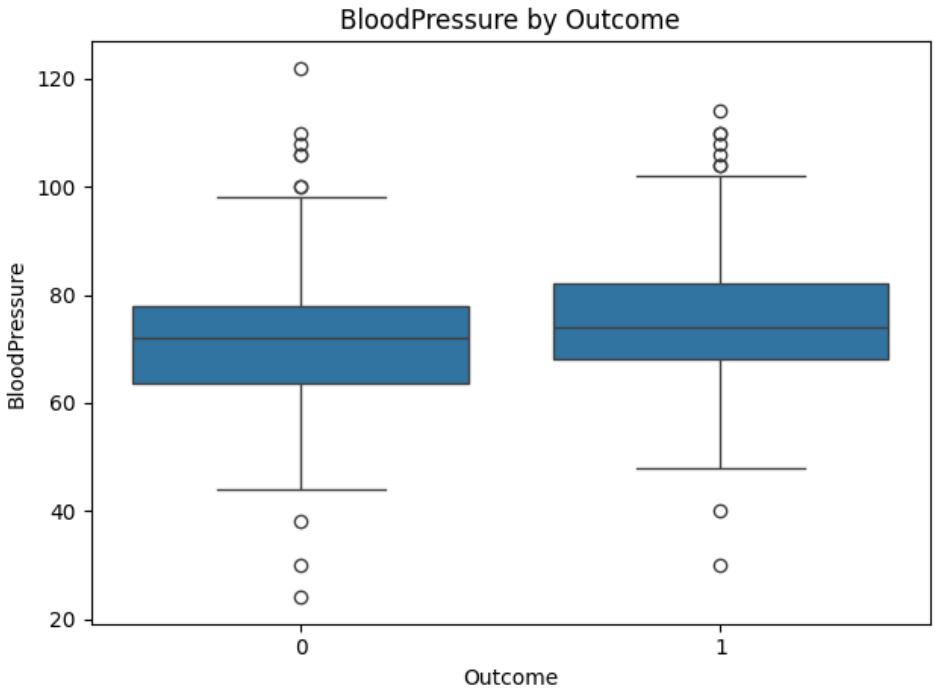

Figure 3: Blood Pressure by Outcome

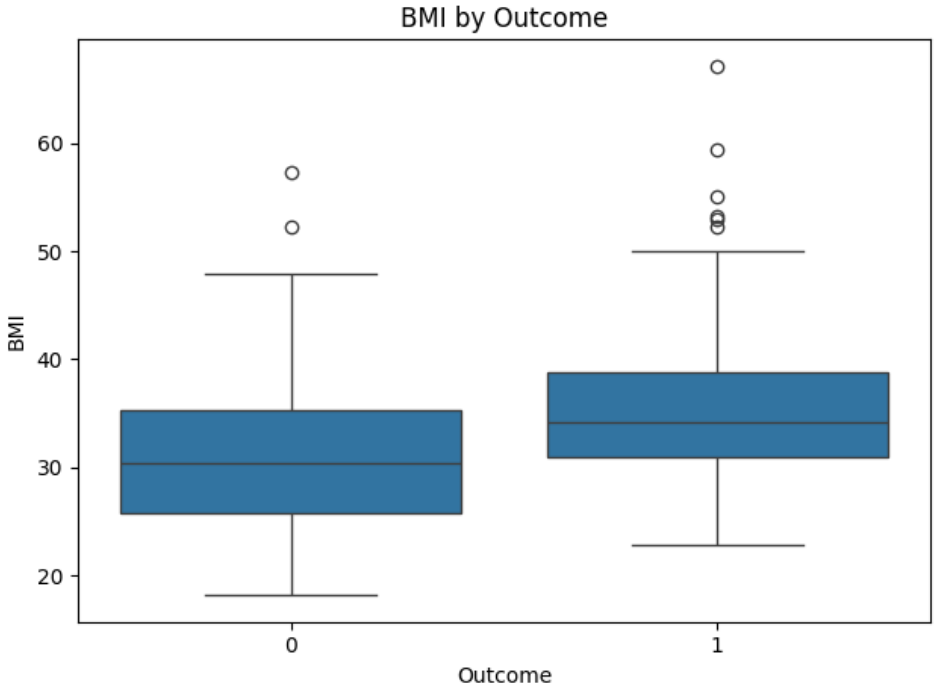

Figure 4: BMI by Outcome

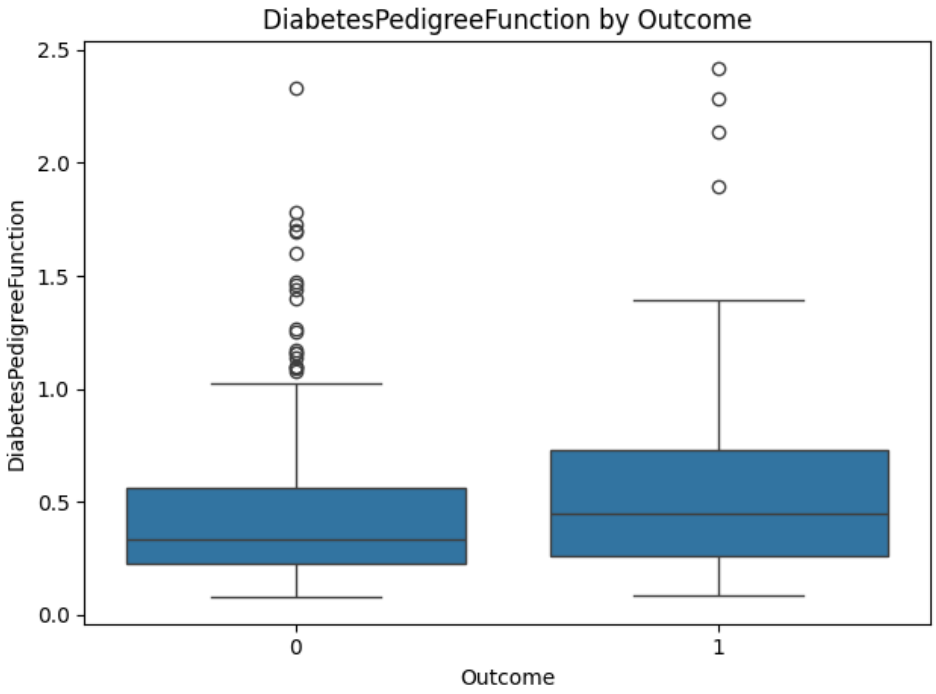

Figure 5: Diabetes Pedigree Function by Outcome

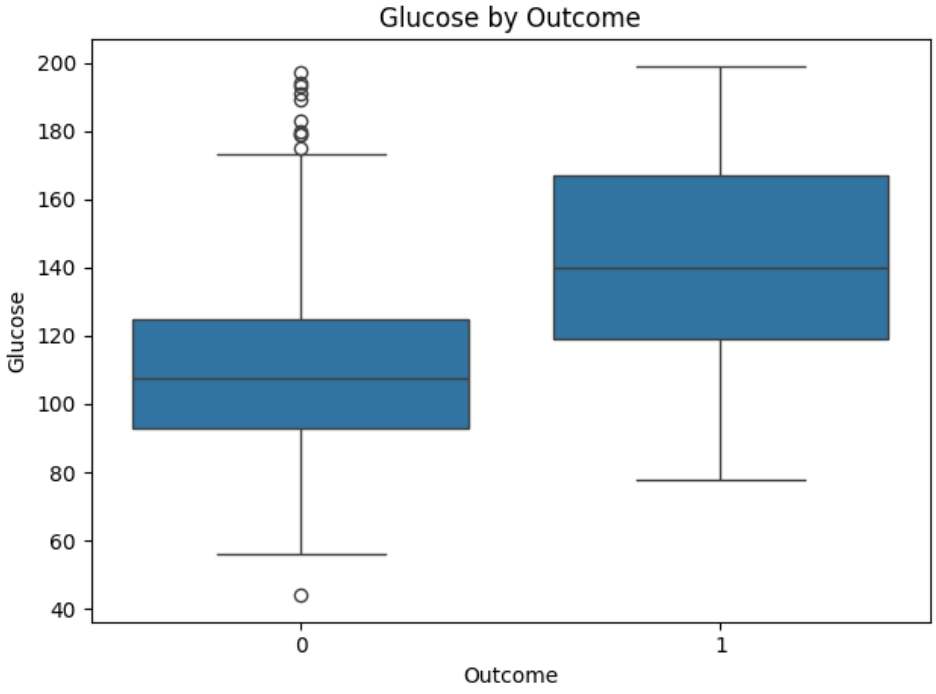

Figure 6: Glucose by Outcome

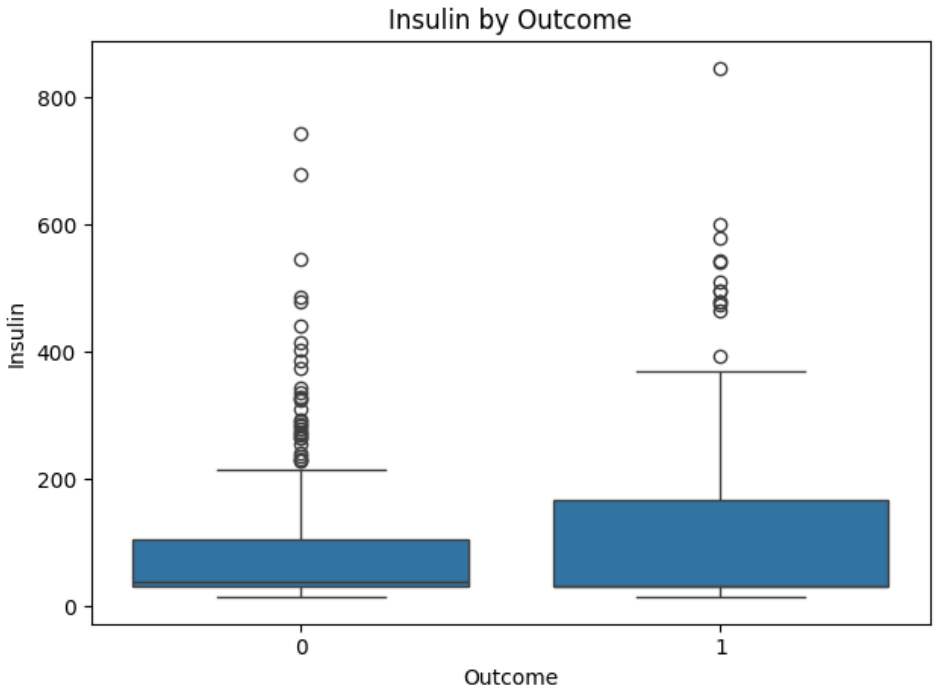

Figure 7: Insulin by Outcome

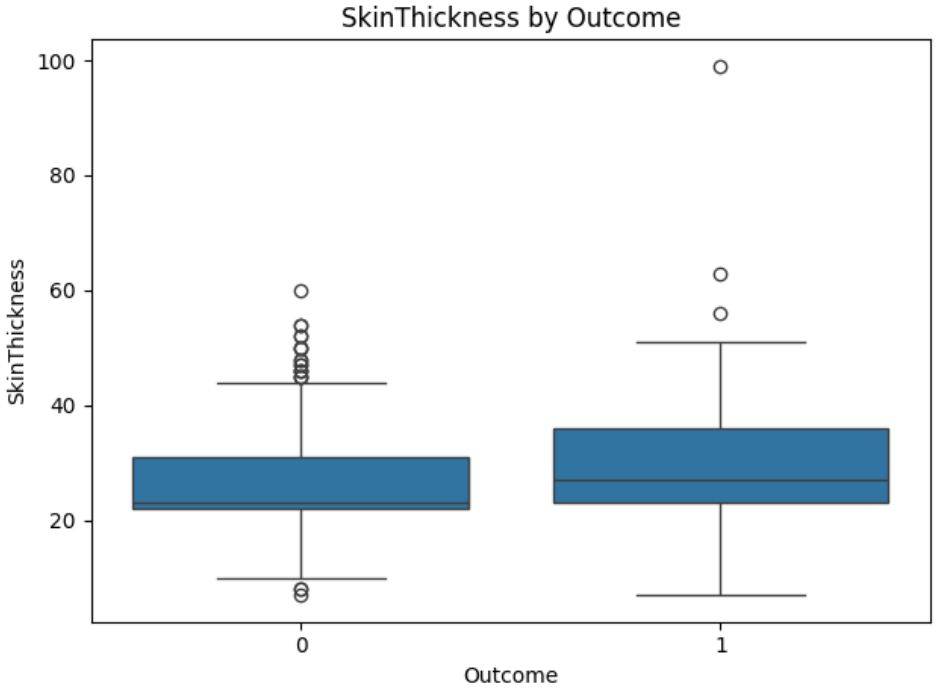

Figure 8: Skin Thickness by Outcome

5.2 Correlation Heatmap

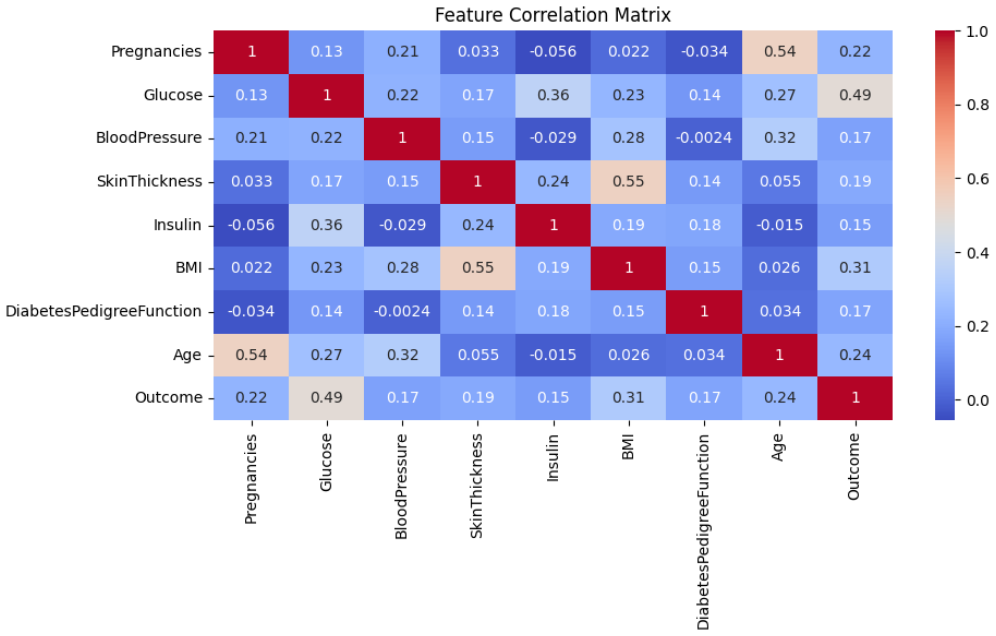

Figure 9: Correlation matrix of features

5.3 Detailed Analysis of Plots

Based on the provided box plots and correlation heatmap, here is a detailed analysis of the factors associated with diabetes.

5.3.1 Box Plots Analysis

Glucose: This variable shows the clearest distinction between the two groups. The median and the entire interquartile range (IQR) for diabetic individuals (Outcome 1) are significantly higher than for non-diabetics (Outcome 0). This strongly suggests that high glucose levels are a key indicator of diabetes.

Age: The median age of the diabetic group (1) is higher. This group also shows a wider spread of ages, indicating that diabetes affects an older population on average.

BMI: The diabetic group (1) has a higher median BMI and a generally higher distribution of values than the non-diabetic group (0). This points to a strong link between a higher BMI and the presence of diabetes.

DiabetesPedigreeFunction: The median and distribution of this function are higher for diabetic individuals (1). This suggests that a family history of diabetes, as measured by this function, is a significant risk factor.

Pregnancies: The median and spread of pregnancies are higher in the diabetic group (1). This could indicate that the number of pregnancies is a factor associated with diabetes.

Insulin: The medians of both groups are very similar. However, the distribution is wider for diabetics (1), and both groups show a large number of outliers, especially the non-diabetic group (0). Insulin alone does not seem to be as clear an indicator as glucose.

BloodPressure: The medians of both groups are very close, and the distributions overlap considerably. This suggests that blood pressure, as a single variable, is not a very distinctive factor for predicting diabetes in this dataset.

SkinThickness: The median for diabetics (1) is slightly higher, but the difference is not as pronounced as for glucose or BMI. The distributions of the two groups are quite similar.

5.3.2 Correlation Heatmap Analysis

By examining the "Outcome" row/column, we can see the correlation of each variable with the presence of diabetes:

Strong Correlation: **Glucose**: The highest correlation (0.49). This confirms what was observed in the box plot: glucose levels are the factor most strongly linked to diabetes.

BMI: Moderate correlation (0.31).

Age: Moderate correlation (0.24).

Weak to Moderate Correlation: **Pregnancies**: Correlation of 0.22.

SkinThickness: Correlation of 0.19.

DiabetesPedigreeFunction: Correlation of 0.17.

Weak Correlation: **BloodPressure**: The correlation is very weak (0.17), which confirms the observation from the box plot.

Insulin: The correlation is also very weak (0.15).

6 Conclusion

The initial analysis identified missing or invalid values in several columns, which were cleaned by replacing zeros with the median. The EDA plots provide an overview of feature distributions, correlations, and differences between diabetic and non-diabetic individuals.

Based on this exploratory data analysis, we can conclude that Glucose levels are the most significant factor associated with diabetes, followed by BMI and Age. Other variables such as **Pregnancies** and the **DiabetesPedigreeFunction** also show a notable association. **Blood Pressure** and Insulin appear to be the least correlated with the outcome in this dataset. The cleaned dataset is now ready for more advanced analysis, including predictive modeling, using these key variables as primary predictors.