project EDA

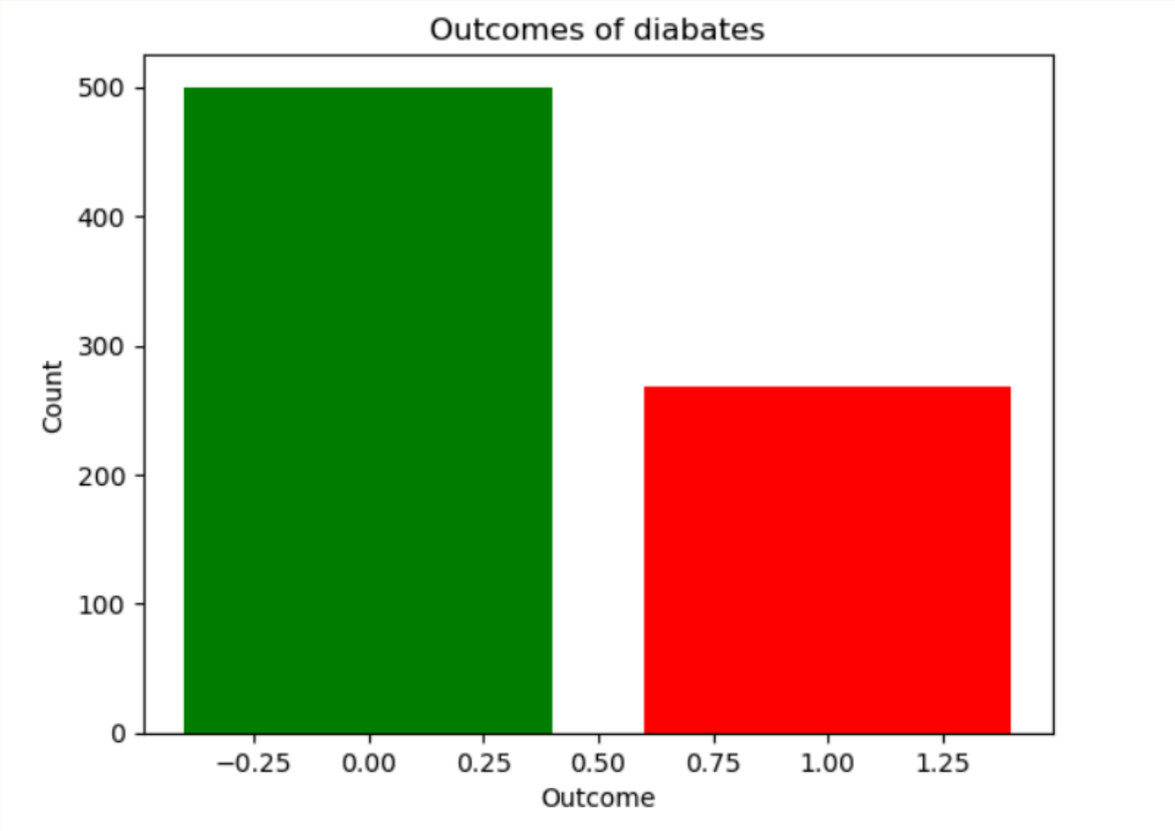
Diabetes Project EDA

Healthcare

Dataset Overview

- The dataset contains key medical measurements related to diabetes such as glucose level, BMI, insulin, and age.
- The target variable Outcome indicates whether a person has diabetes (1) or not (0).

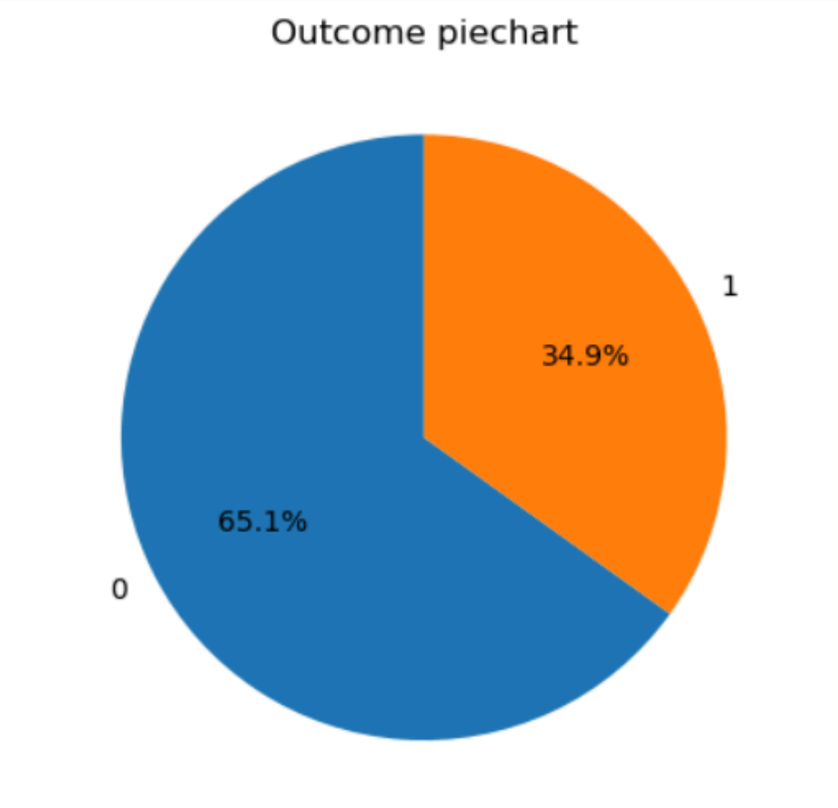
Data info.



Outcome

0 500

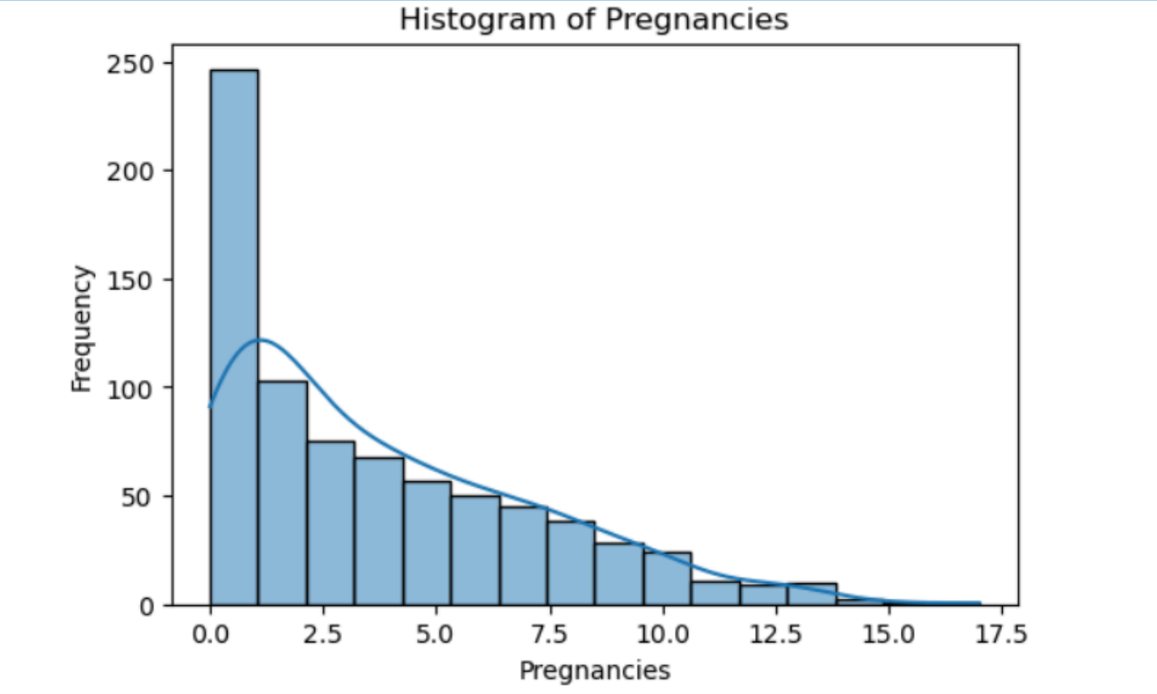
1 268



The dataset is imbalanced: around 65% of the patients are non-diabetic (Outcome=0), while only 35% are diabetic (Outcome=1).

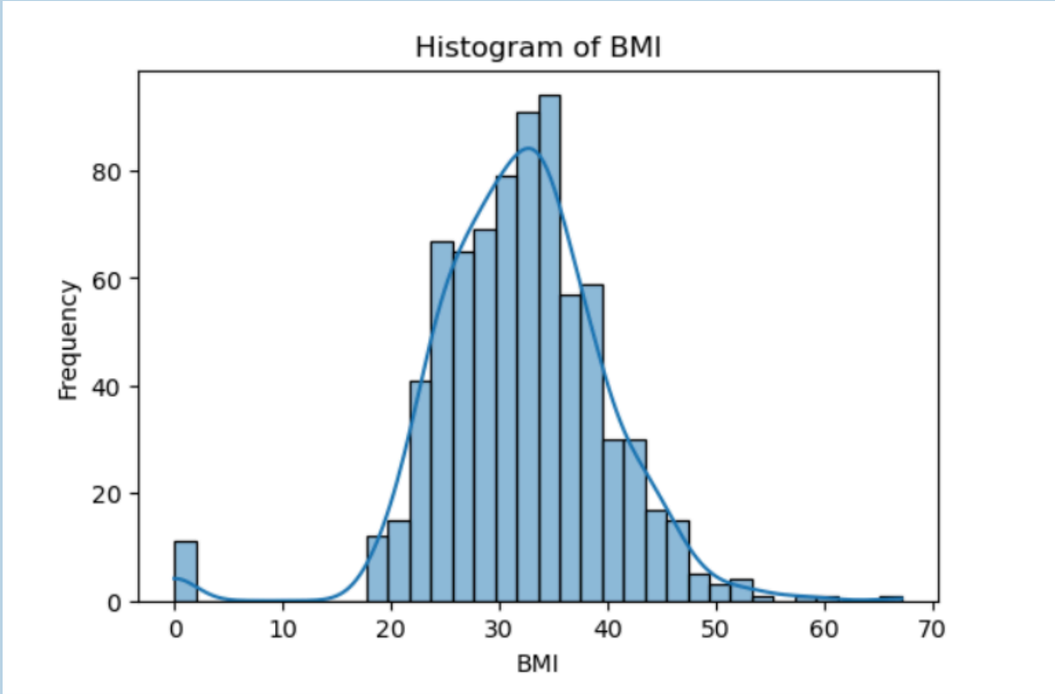
This imbalance indicates that the model may need techniques such as class weighting or resampling to avoid bias toward the majority class.

Histogram of Pregnancies



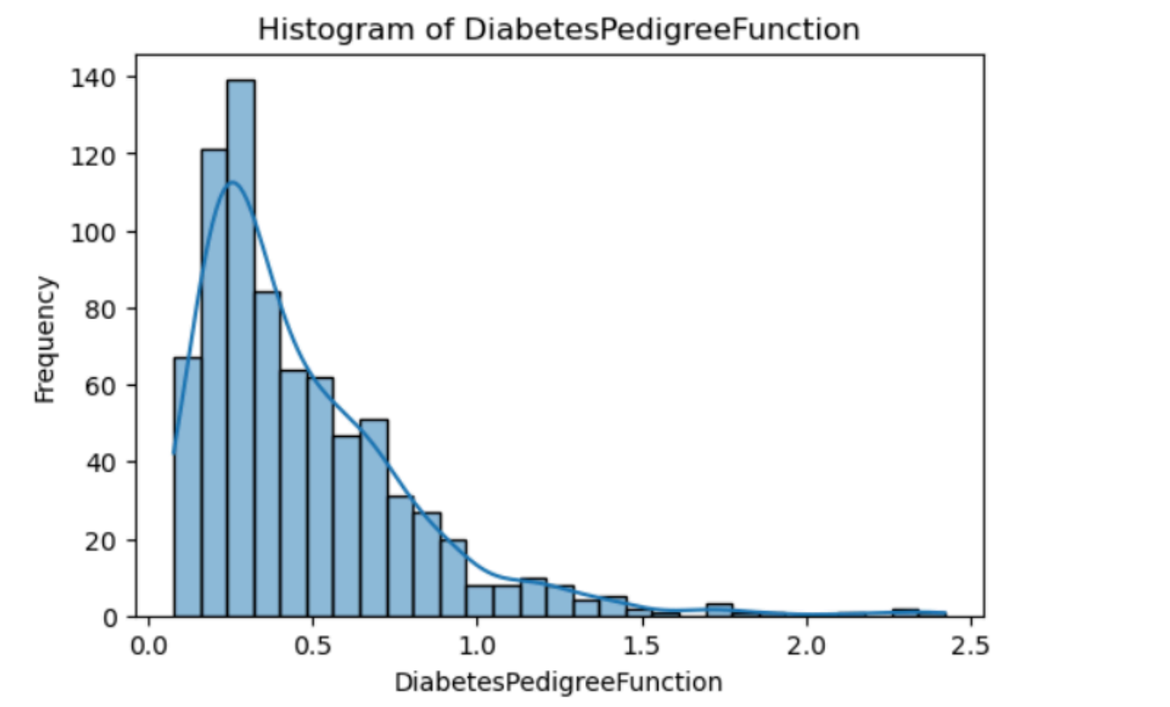
Pregnancies show a strong right-skew, with most women having 0–2 pregnancies. Higher pregnancy counts are rare but are clinically important, as increased pregnancies are linked to higher diabetes risk.

Histogram of BMI



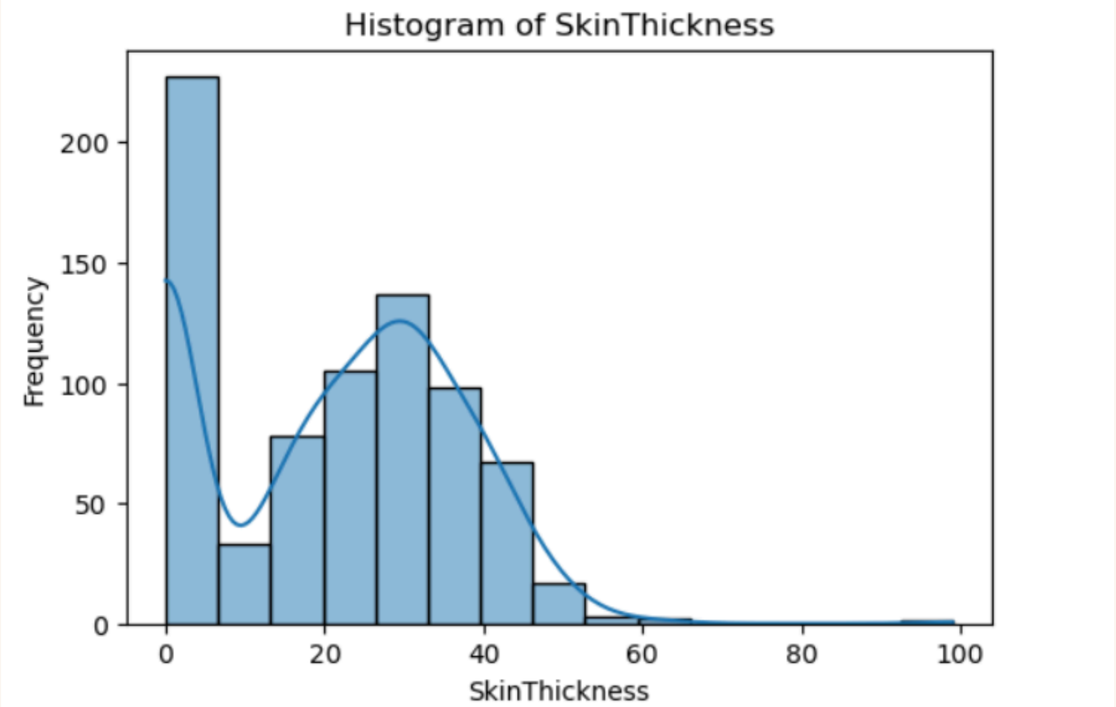
Most participants have BMI values between 25 and 35, indicating that overweight individuals are highly represented in the dataset. The right skew suggests the presence of individuals with obesity, which is clinically relevant since higher BMI is associated with increased diabetes risk

Histogram of Diabetes Pedigree Function



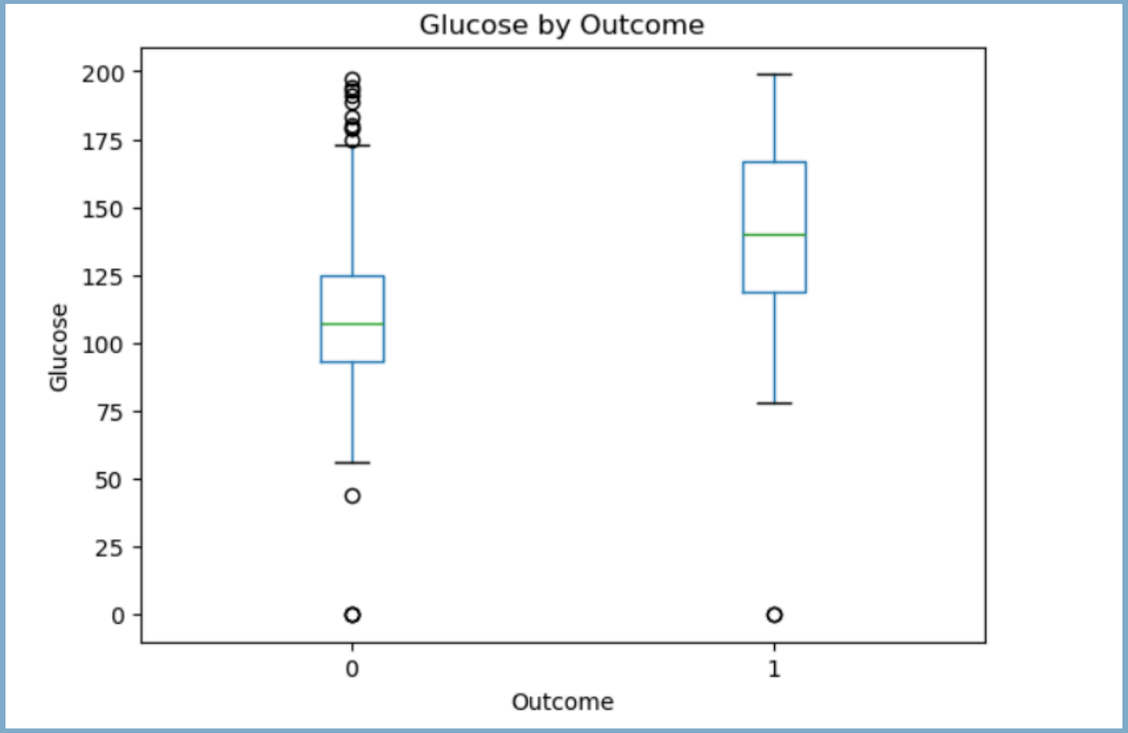
Most individuals have low DPF values, but higher values—although rare—indicate a strong family history of diabetes and therefore higher risk.

Histogram of Skin Thickness



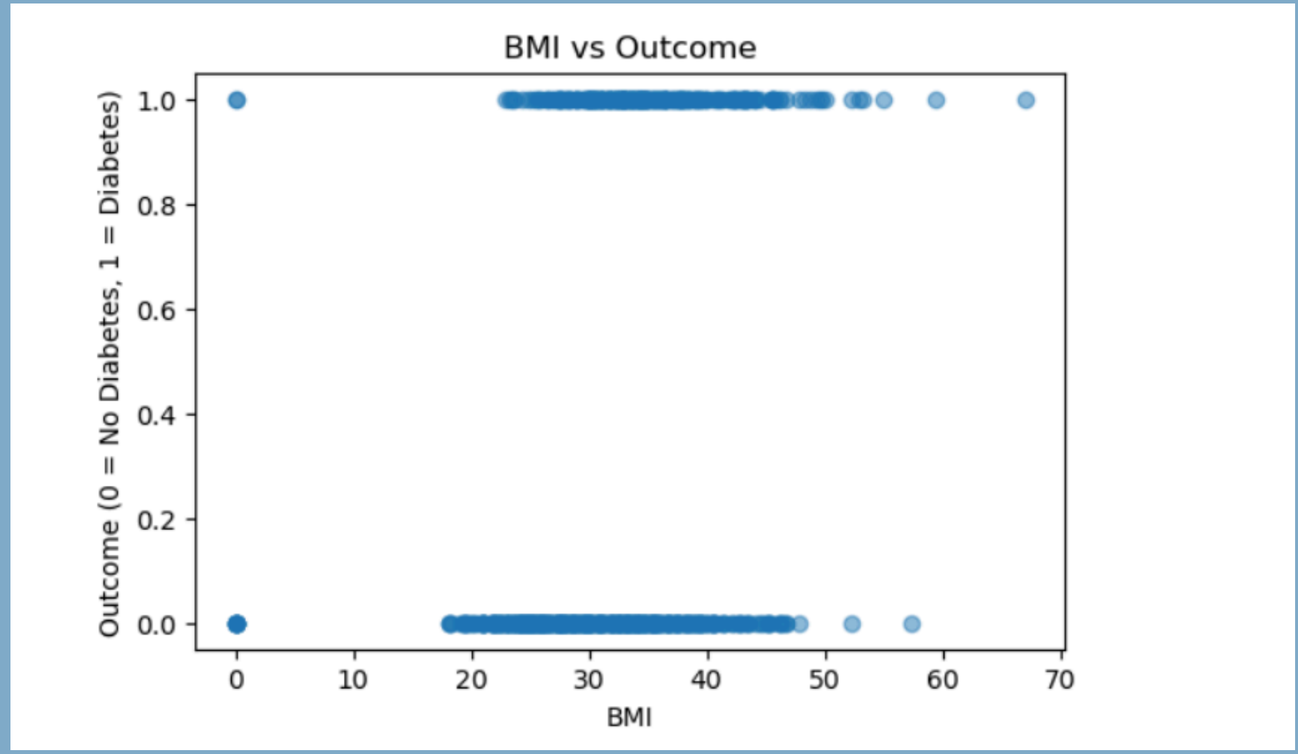
Skin Thickness contains many zero values, suggesting missing or inaccurate measurements. The distribution is highly skewed, making it less reliable without additional cleaning

Boxplot – Glucose by Outcome



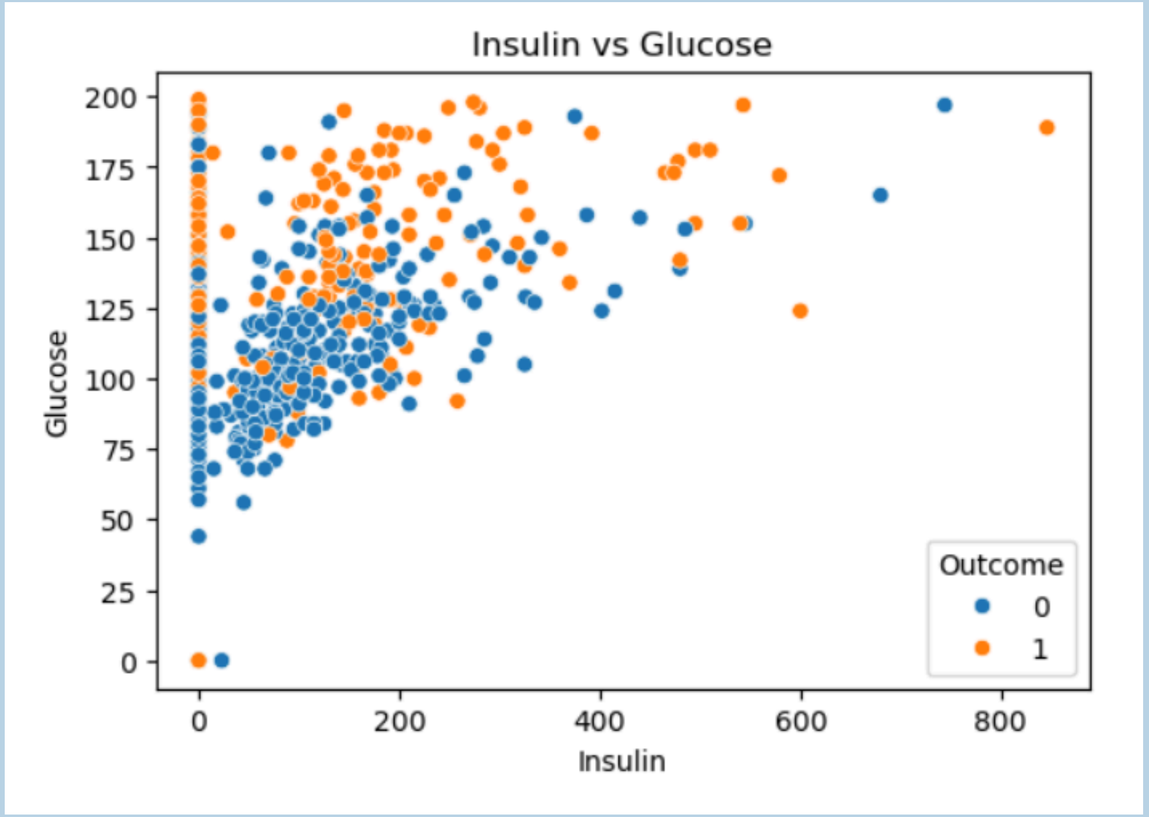
Glucose is the strongest differentiator between diabetic and non-diabetic individuals. The median glucose in diabetic patients is significantly higher, confirming its importance as a key predictive feature.

Scatter Plots_BMI vs Outcome



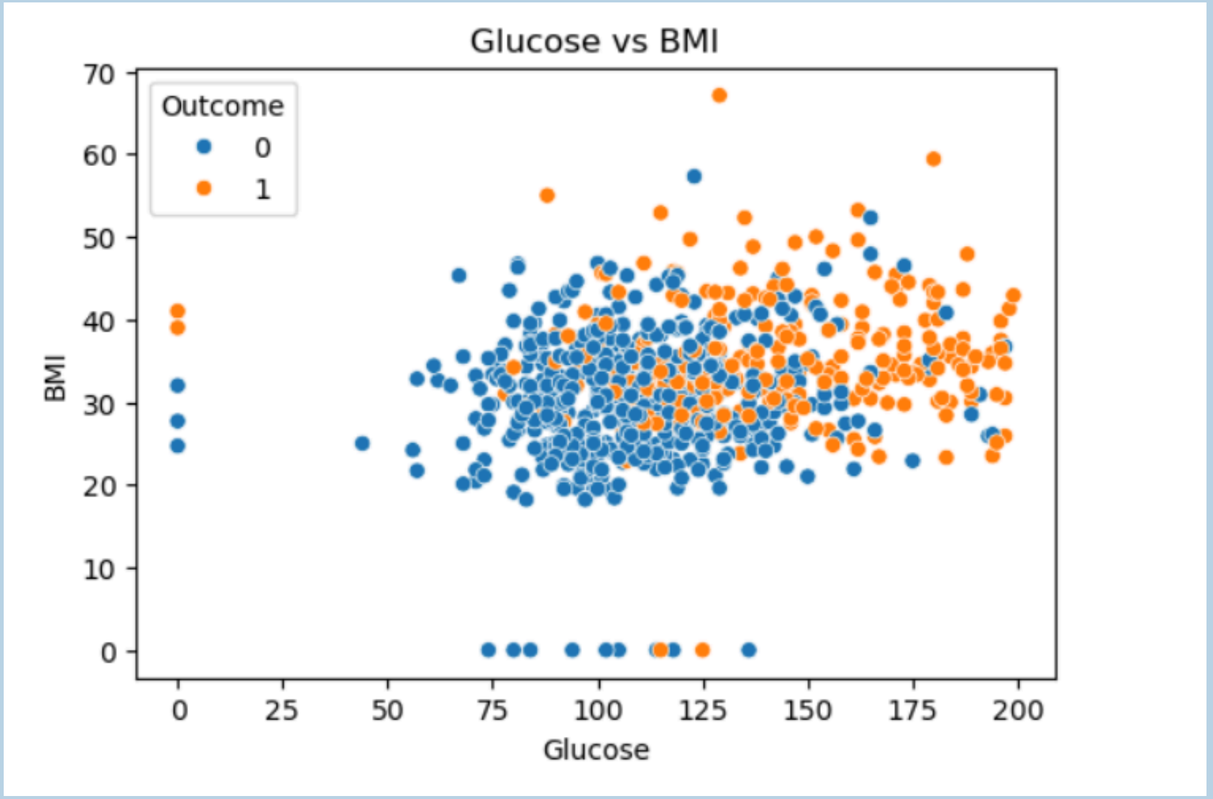
Diabetic individuals appear more frequently in higher BMI ranges (above 30), supporting the known medical relationship between obesity and diabetes risk

Scatter Plot (Insulin vs Glucose)

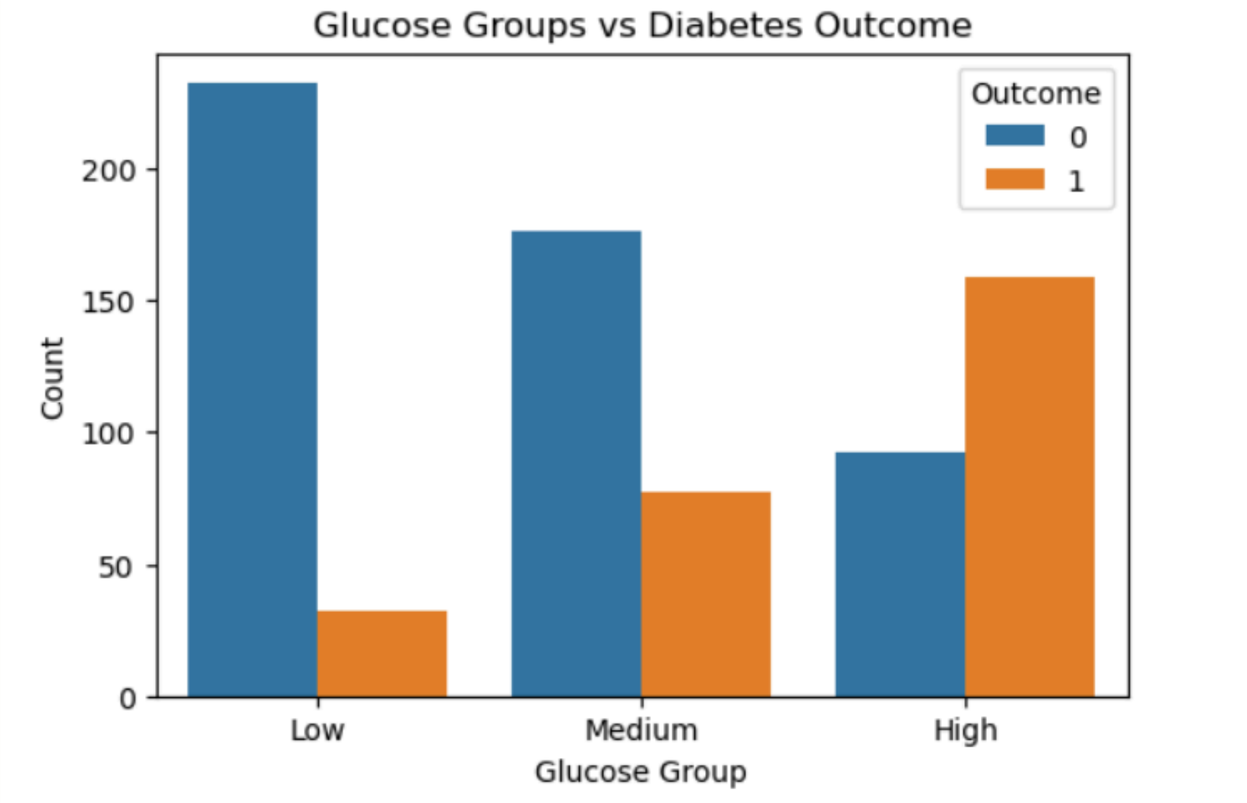


Glucose increases with insulin, and diabetic individuals cluster more in higher-glucose ranges. Several zero insulin values indicate potential data-entry issues or missing-information patterns.

Glucose vs BMI



Diabetic individuals tend to appear in regions with higher glucose and slightly higher BMI values, indicating that both features contribute significantly to diabetes prediction



"This chart shows the relationship between Glucose levels and Diabetes Outcome.

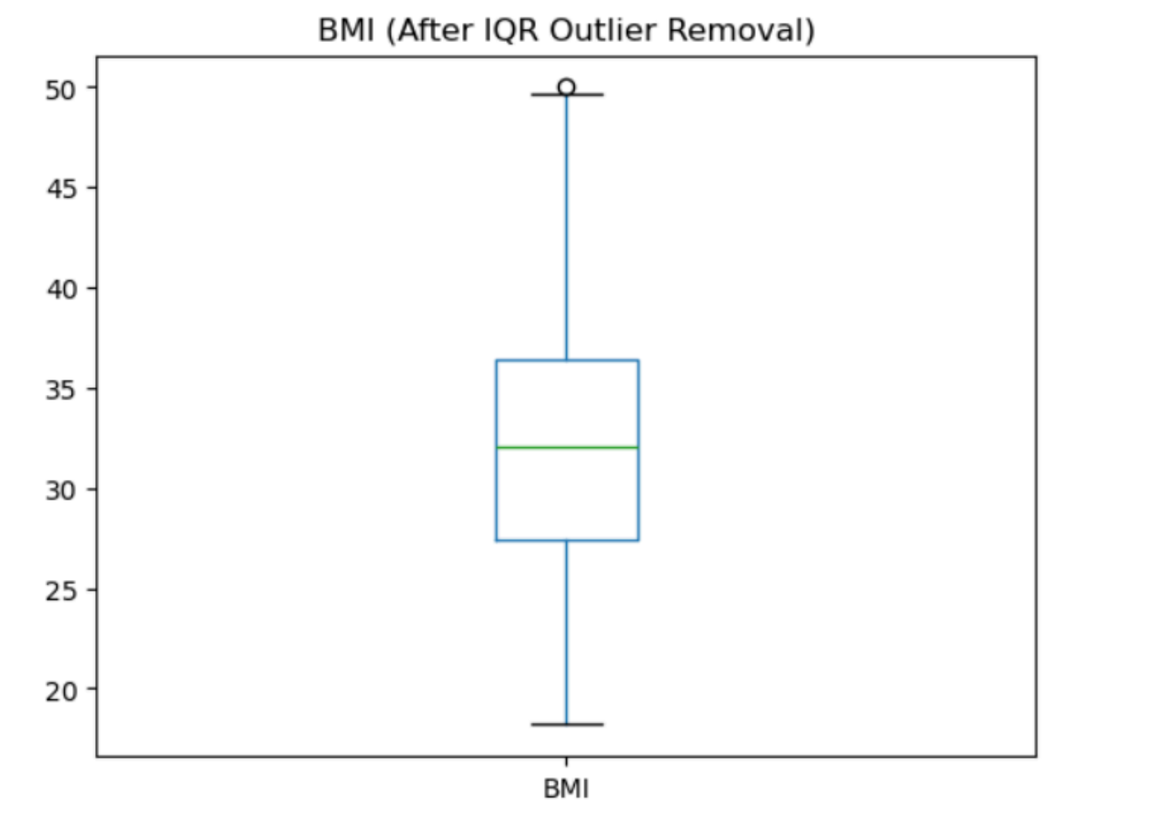
We can see that:

- Most non-diabetic cases (Outcome=0) are in the Low and Medium glucose groups.
- The High glucose group has the highest number of diabetic cases (Outcome=1).
- This indicates that higher glucose levels are strongly associated with diabetes."

Data Cleaning Summary

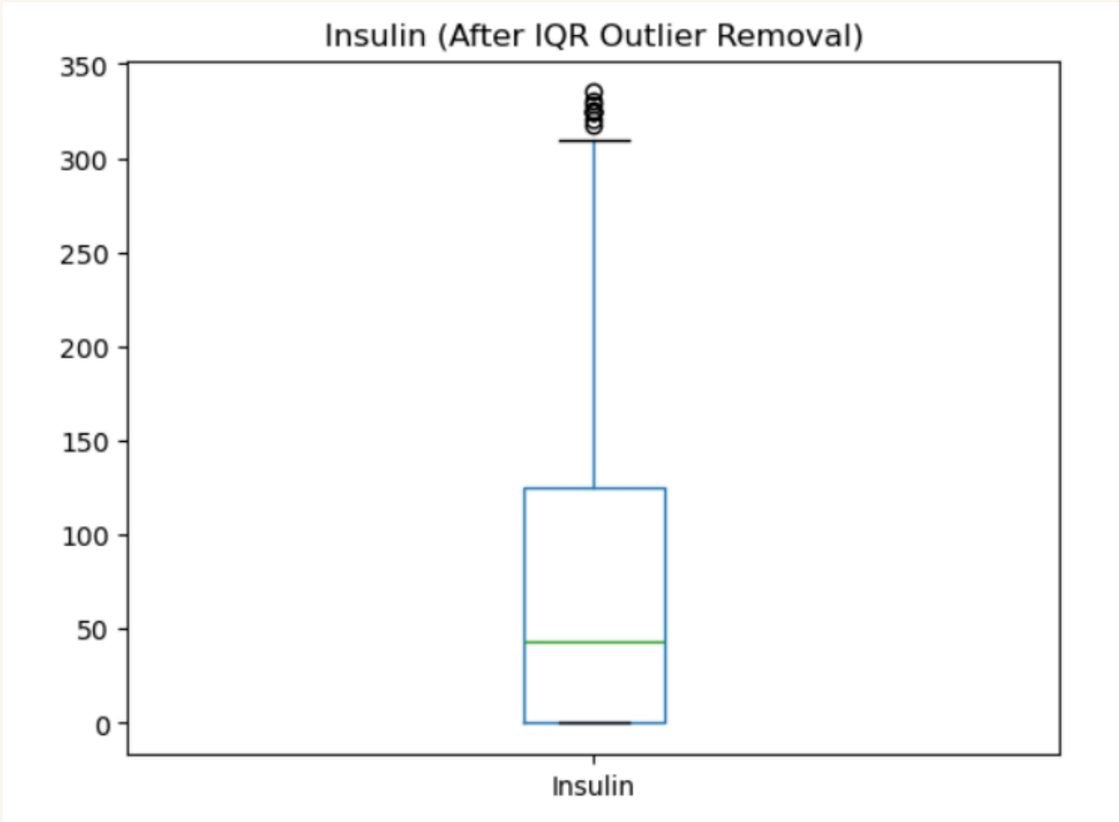
- No missing values were found in the dataset.
- No duplicate rows were detected.
- Therefore, no imputation or duplicate removal was required.
- The dataset was already clean and ready for further analysis.

BMI (After IQR Outlier Removal)



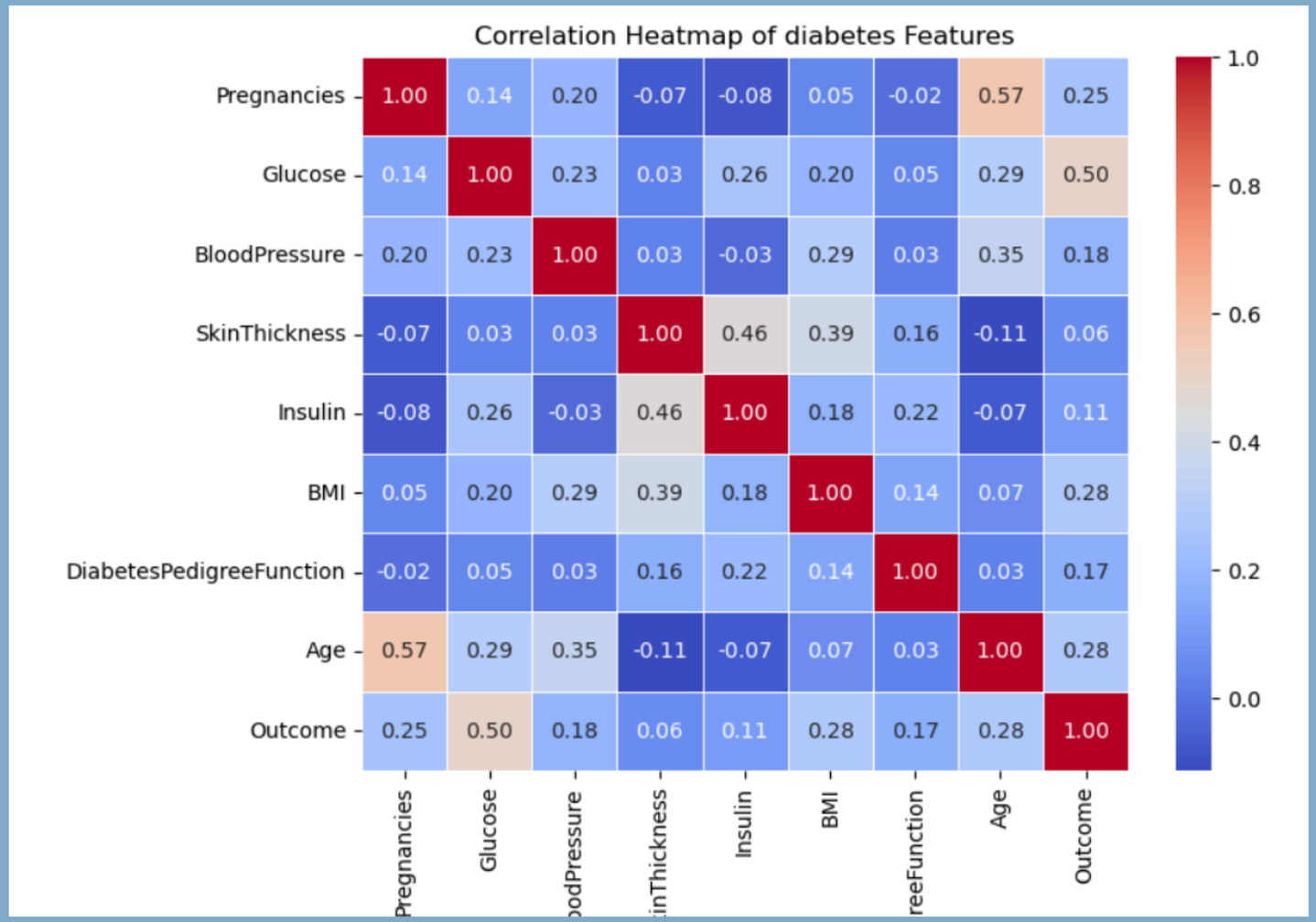
After removing outliers using IQR, BMI values are more consistent and realistic. The cleaned distribution improves model reliability and reduces noise.

Insulin



“After applying the IQR method, most extreme insulin values were removed. The cleaned distribution shows that insulin levels remain highly skewed, but outliers have been significantly reduced. This helps improve model stability and reduces noise in the data.”

Correlation Heatmap



Glucose shows the strongest correlation with diabetes outcome ($r \approx 0.50$), followed by BMI and Age. These features are therefore the most impactful in predicting diabetes, while BloodPressure and SkinThickness show weak impact.

Feature Importance Summary

Most important features influencing diabetes:

1. Glucose (strongest)
2. BMI
3. Age
4. Pregnancies
5. DPF

Less important:

BloodPressure, SkinThickness, Insulin (due to missing values)

- Features with importance greater than 0.05 were selected as key predictors.
- These variables have the strongest influence on the model and provide the most meaningful contribution for diabetes classification."

Final Insights

- Glucose is the primary indicator of diabetes.
- High BMI strongly increases diabetes likelihood.
- Age and Pregnancies contribute moderately.
- Features with many missing values (Insulin, SkinThickness) require careful treatment.
- Dataset benefits significantly from outlier removal and proper cleaning.

project EDA

Thank you very much!