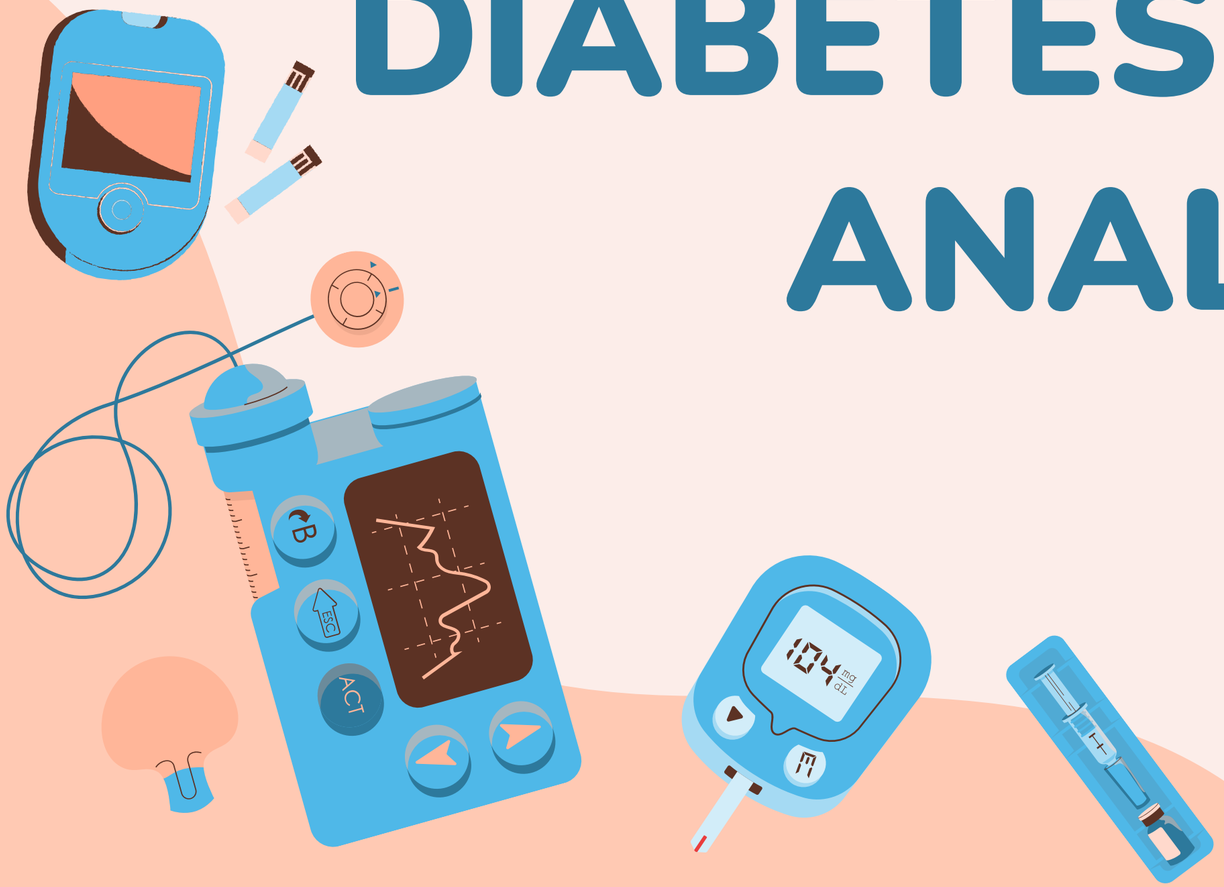
DIABETES DATASET ANALYSIS



TODAY'S AGENDA

1. PRESENT KEY FINDINGS FROM OUR ANALYSIS ON DIABETES PREDICTION

2-EXPLAIN THE PREPROCESSING AND ANALYSIS TECHNIQUES USED

3. HIGHLIGHT INSIGHTS THROUGH VISUALIZATIONS

4. DISCUSS THE IMPACT OF STATISTICAL METHODS AND MODELS

5. OUTLINE RECOMMENDATIONS AND CONCLUSIONS FOR IMPROVING OUTCOMES

ABOUT THIS PROJECT

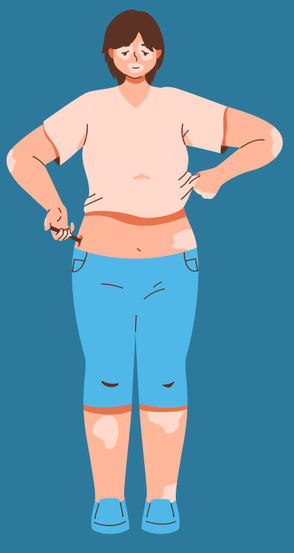
This project investigates diabetes onset using statistical and data-driven techniques, focusing on:

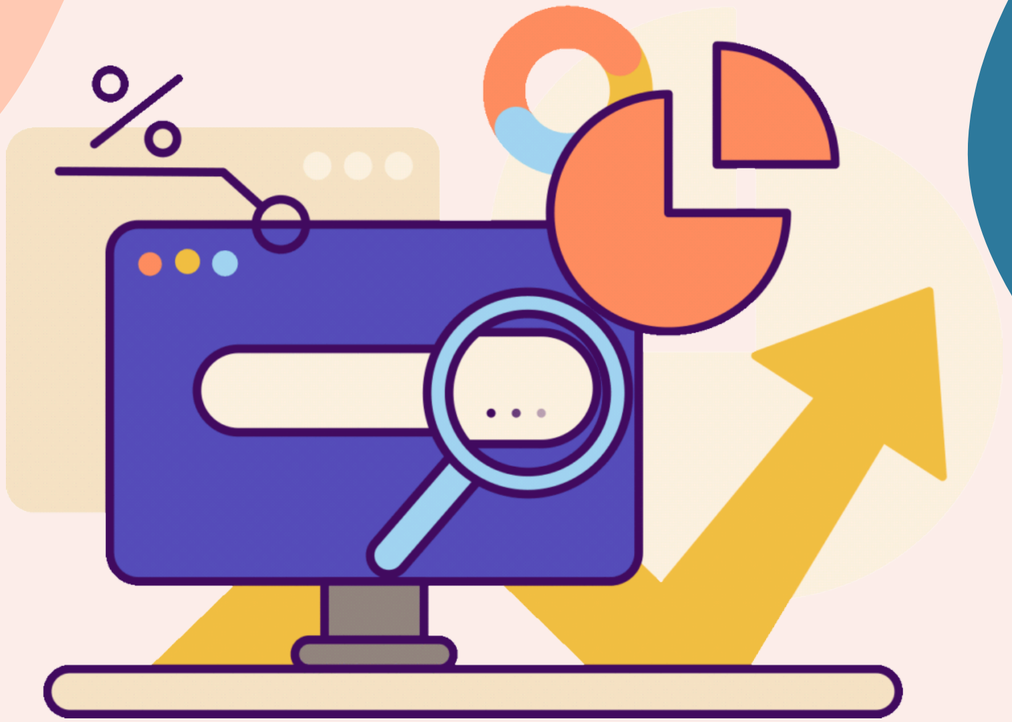
- Exploratory Data Analysis (EDA): To uncover trends and relationships in key health metrics like glucose, BMI, and age.
- Statistical Testing: To validate hypotheses about factors contributing to diabetes risk.
- Visualization Techniques: To present insights effectively and support data-driven conclusions.
- Predictive Modeling: To identify individuals at risk of diabetes and propose interventions.

Objective: Provide actionable insights to healthcare professionals for better diabetes prevention and management.

ABOUT DATASET

- Purpose: Analyze health metrics to understand factors contributing to diabetes onset and predict outcomes.
- Attributes:
  - Medical measures: Glucose levels, BMI, Blood Pressure, Insulin.
  - Demographics: Age, Pregnancies.
  - Family history: Diabetes Pedigree Function (DPF).
  - Outcome: Diabetes status (1 = Diabetic, 0 = Non-Diabetic).
  - Target Population: Primarily female patients, aged 21+, of Pima Indian heritage.





STEP-BY-STEP OVERVIEW OF DIABETES ANALYSIS

→

1.

Data Preparation and preprocessing

2.

Exploratory Analysis

3.

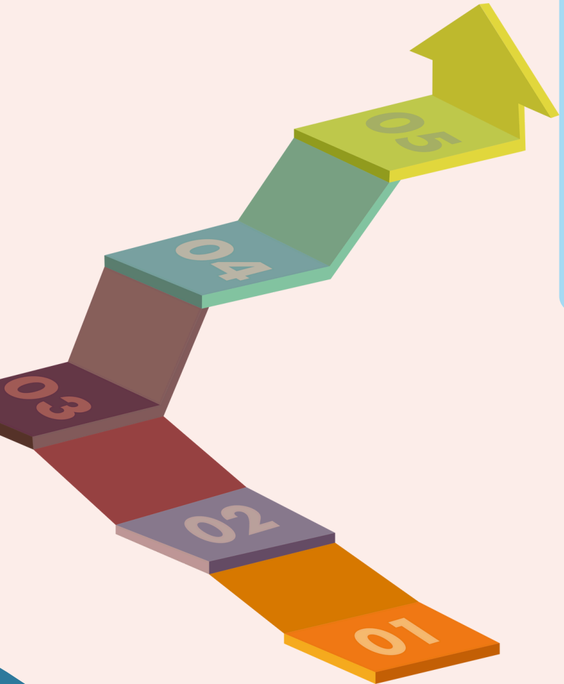
Exploring Key Relationships

4.

Hypothesis Testing

5.

Simulation Task



1-DATA PREPARATION AND PREPROCESSING

Data Loading and Inspection

Imported the dataset and reviewed structure, statistics, and anomalies.

Categorical Encoding

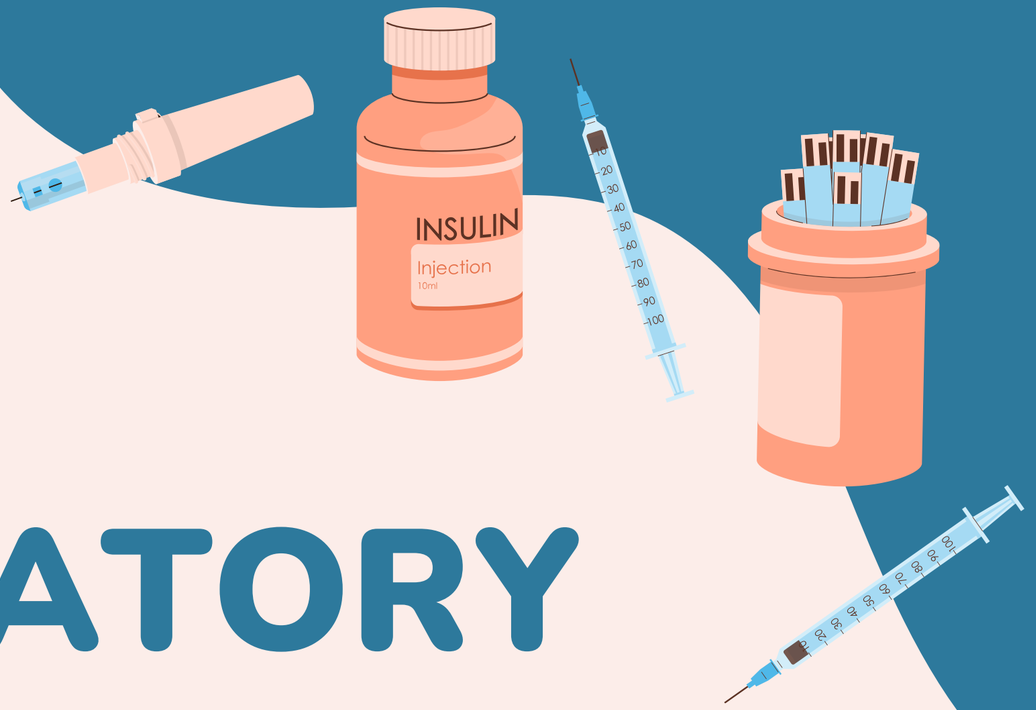
Converted Outcome into labeled categories (Diabetic, Non-Diabetic) for clarity.

Data Splitting

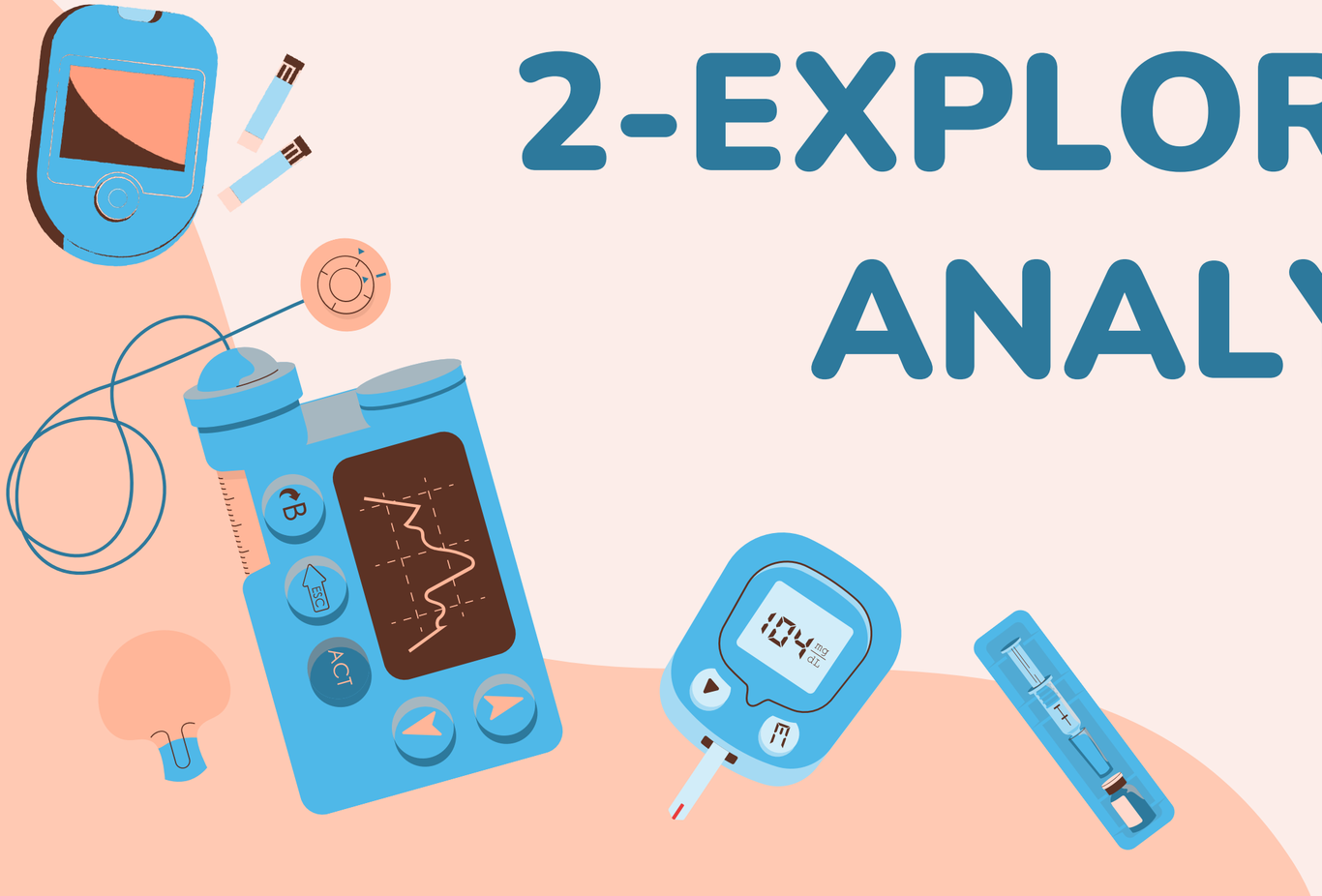
Divided data into diabetic and non-diabetic groups for comparative analysis.

Missing Values and Visualization Prep

Identified missing values and prepared data for visualizations by adding categorical columns.



2-EXPLORATORY ANALYSIS

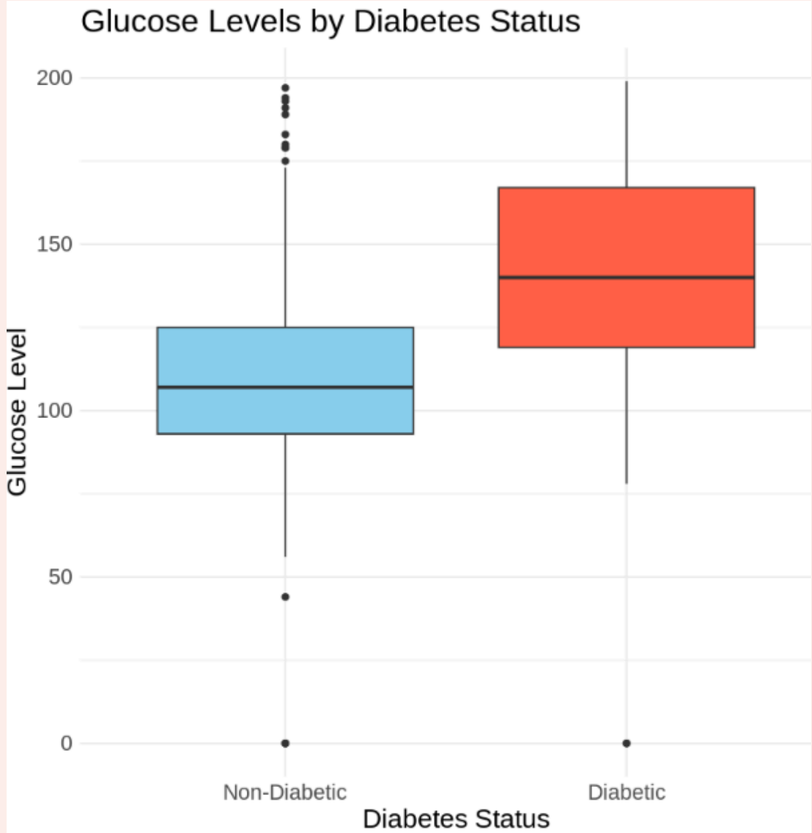


OBJECTIVE :

- Compare glucose levels between diabetic and non-diabetic patients.

KEY FINDINGS :

- Diabetic: 141.26 mg/dL, Non-Diabetic: 109.98 mg/dL.
- Diabetic patients consistently show higher glucose levels.

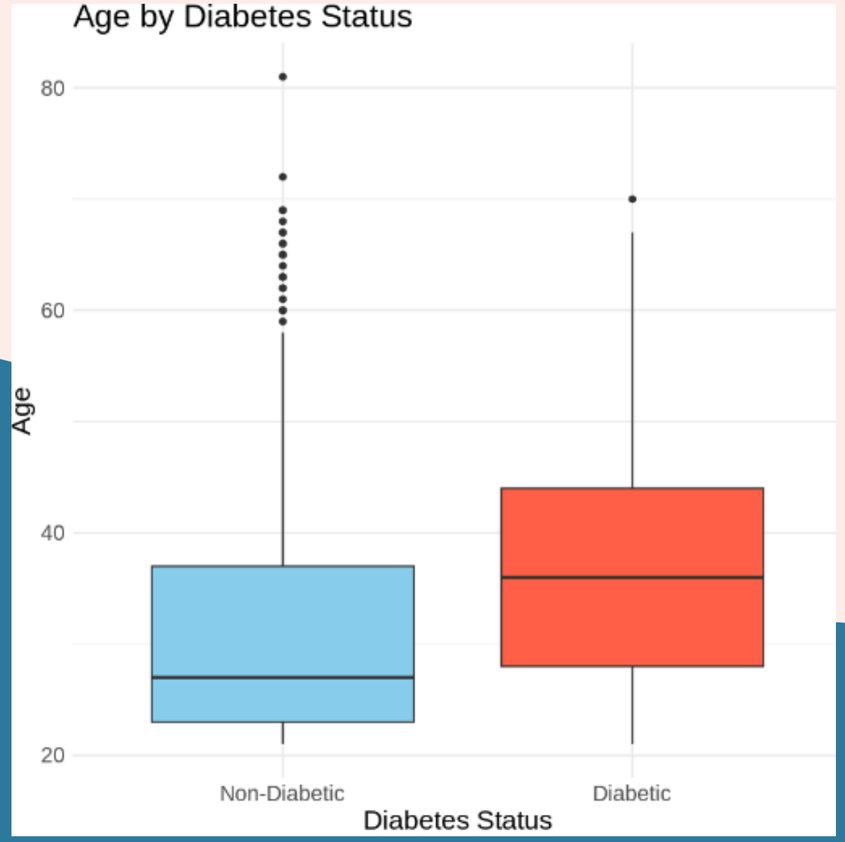


OBJECTIVE :

- Compare the average age of diabetic and non-diabetic patients.

KEY FINDINGS :

- Average age of diabetic patients: 37 years.
- Average age of non-diabetic patients: 31 years.

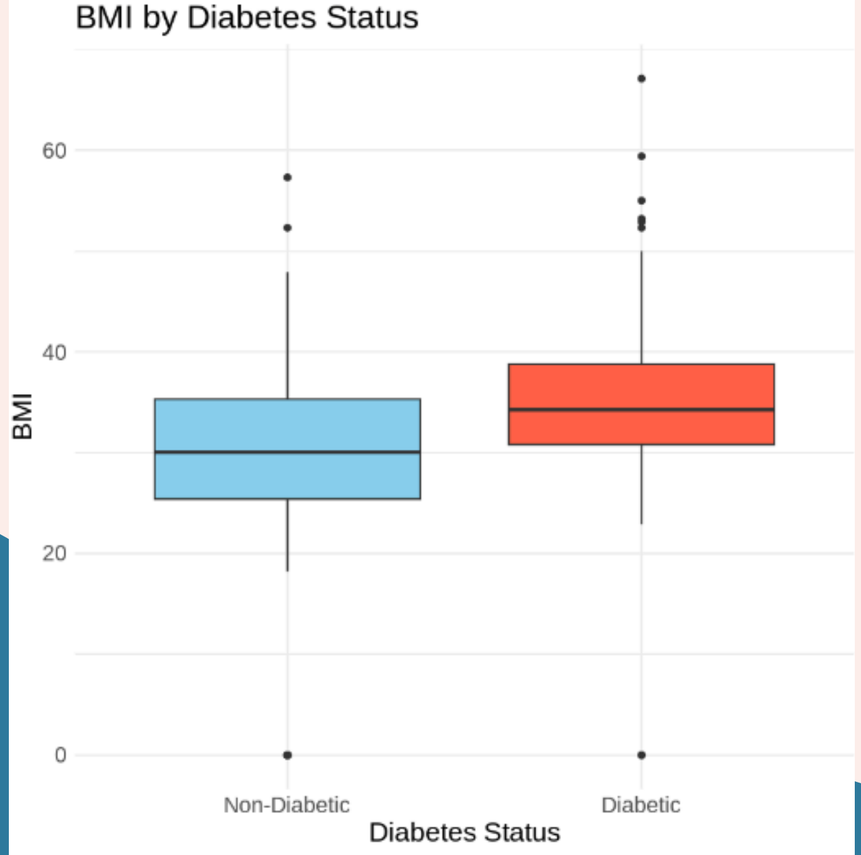


OBJECTIVE :

- Compare the average BMI (Body Mass Index) of diabetic and non-diabetic patients.

KEY FINDINGS :

- Average BMI of diabetic patients: 35.14.
- Average BMI of non-diabetic patients: 30.30.
- Diabetic patients tend to have a higher BMI compared to non-diabetic patients.

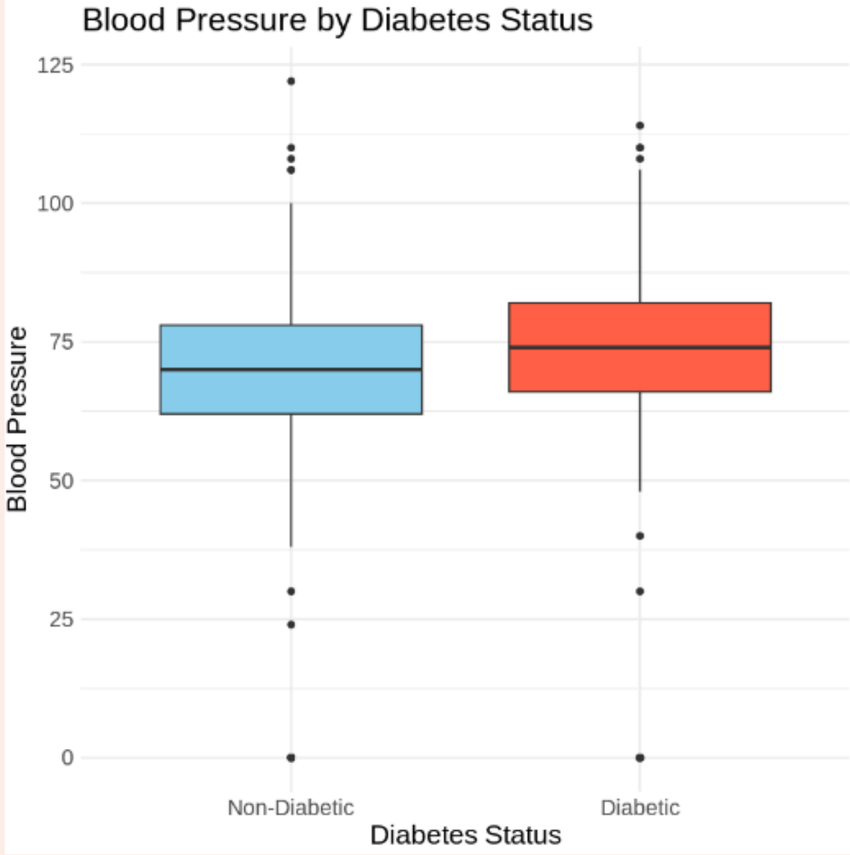


OBJECTIVE :

- Compare blood pressure measurements between diabetic and non-diabetic groups.

KEY FINDINGS :

- Average blood pressure for diabetic patients: 70.82 mmHg.
- Average blood pressure for non-diabetic patients: 68.18 mmHg.
- Diabetic patients show slightly higher average blood pressure, but the difference is minimal.

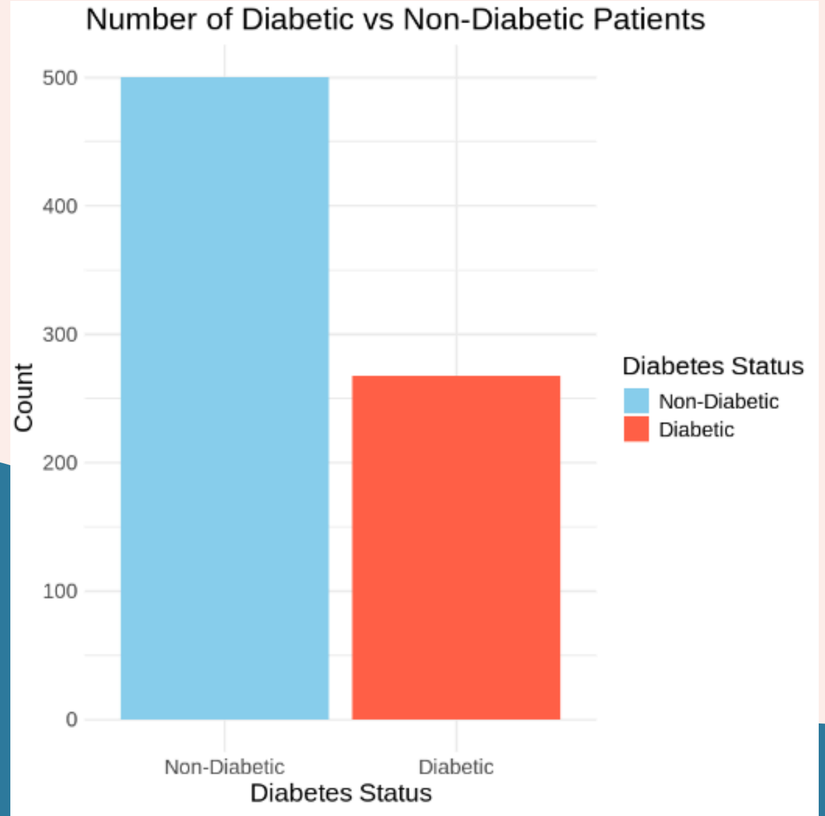


OBJECTIVE :

- Determine the prevalence of diabetes among patients in the dataset.

KEY FINDINGS :

- Rate of Diabetes: 34.90% of the patients in the dataset are diabetic.
- The majority (approximately 65%) of patients are non-diabetic.

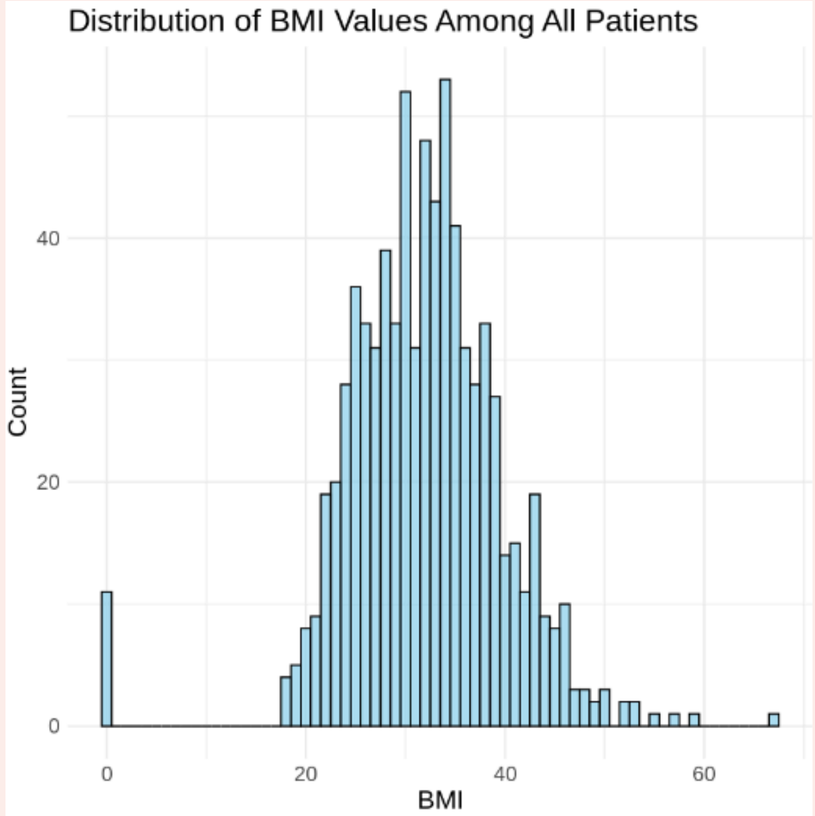


OBJECTIVE :

- Visualize the distribution of BMI values among all patients.

KEY FINDINGS :

- The distribution is approximately bell-shaped, indicating a normal distribution.
- Most patients have a BMI in the range of 25 to 35, with the peak around 30, which is in the overweight category.
- There are a few outliers on both ends, with extremely low or high BMI.

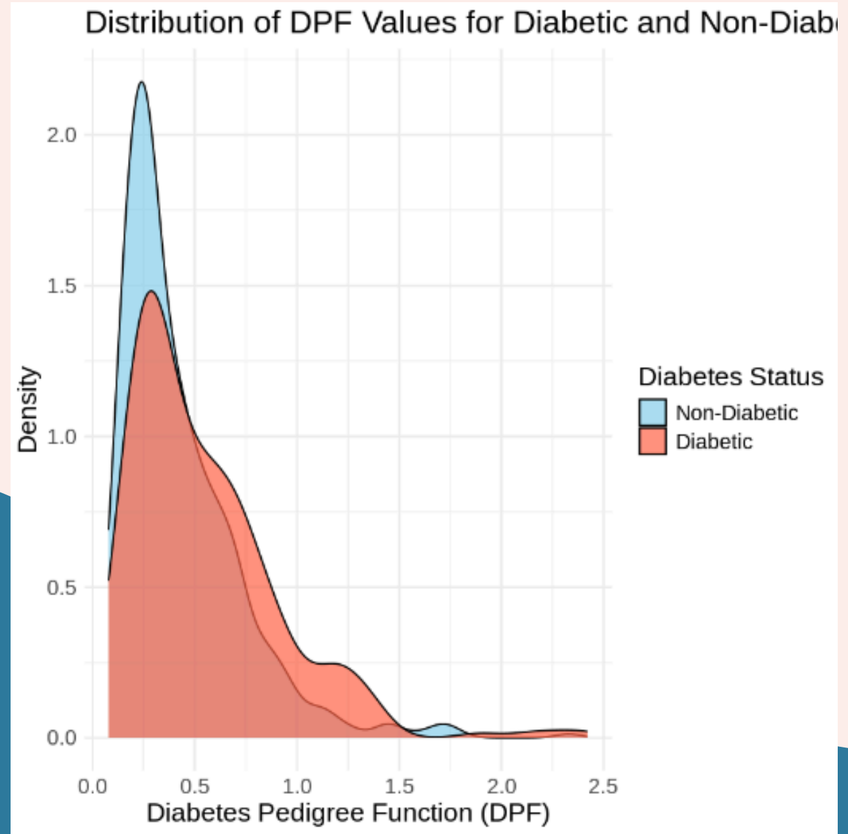


OBJECTIVE :

- Compare the distribution of Diabetes Pedigree Function (DPF) values between diabetic and non-diabetic patients.

KEY FINDINGS :

- Non-diabetic patients (blue) generally have lower DPF values, with a peak density around 0.2–0.3.
- Diabetic patients (red) tend to have a wider range of DPF values, extending beyond 1.0, though the peak density is similar to non-diabetic patients.

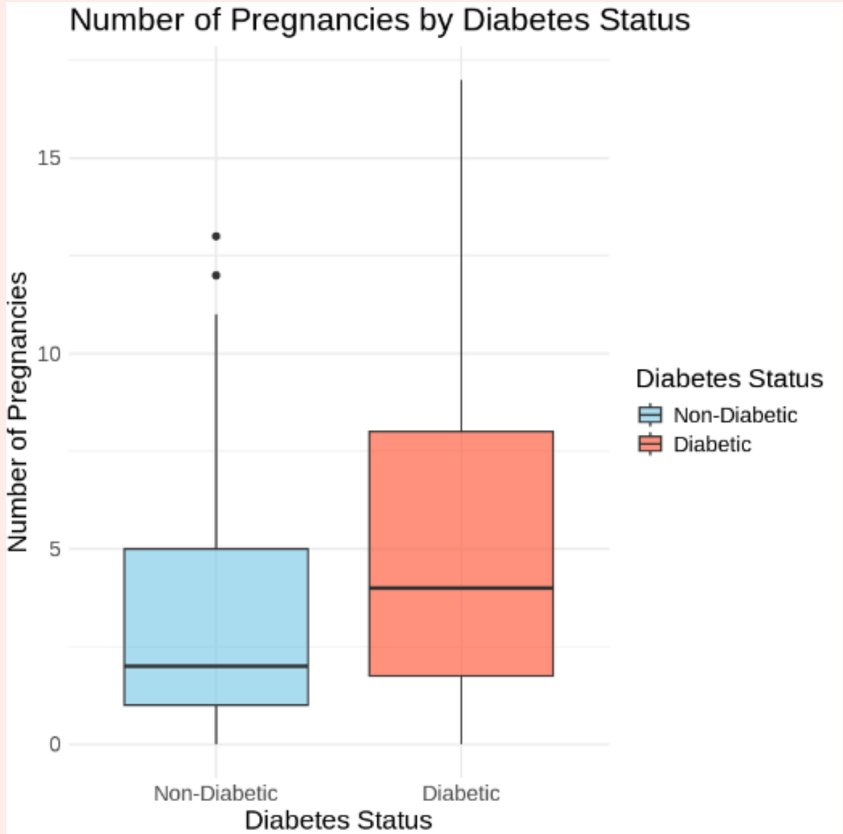


OBJECTIVE :

- Analyze the link between the number of pregnancies and diabetes status.

KEY FINDINGS :

- Average number of pregnancies for diabetic patients: 4.87.
- Average number of pregnancies for non-diabetic patients: 3.30.
- Diabetic patients tend to have more pregnancies on average compared to non-diabetic patients.

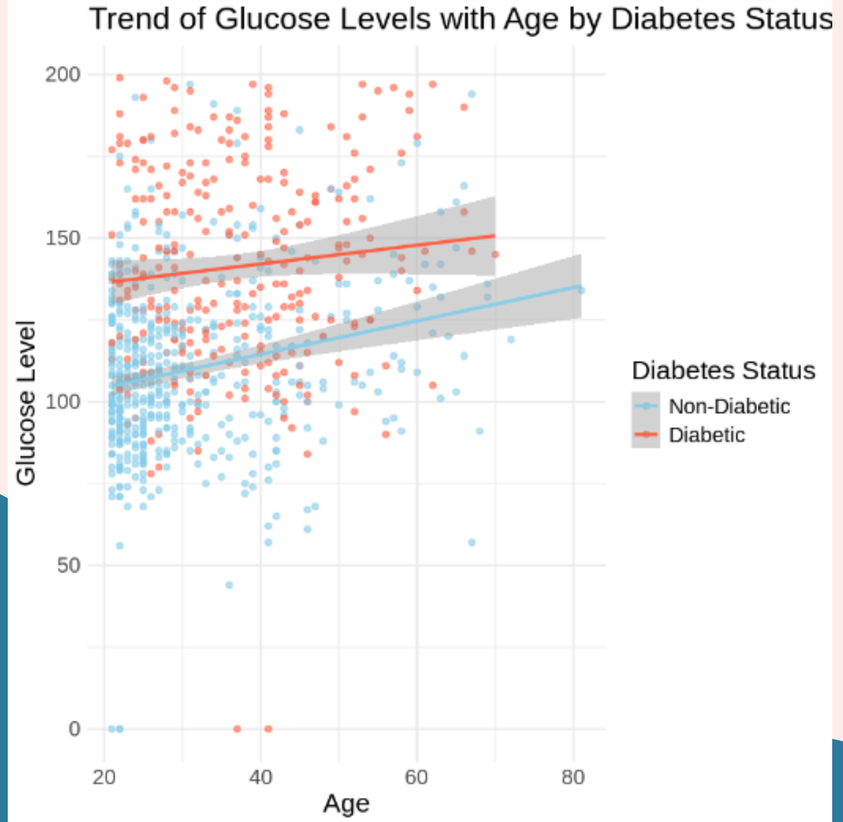


OBJECTIVE :

- Analyze how glucose levels vary with age among diabetic and non-diabetic patients.

KEY FINDINGS :

- Diabetic patients show a slight increase in glucose levels with age.
- Non-diabetic patients have stable glucose levels across all ages.
- Diabetic patients consistently have higher glucose levels than non-diabetic patients.

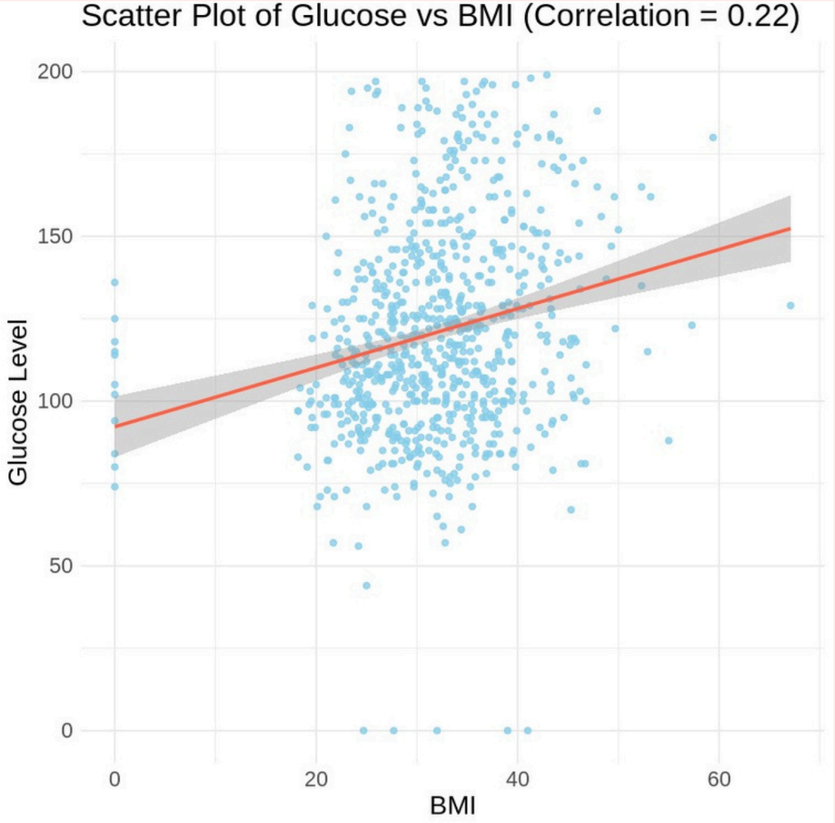


OBJECTIVE :

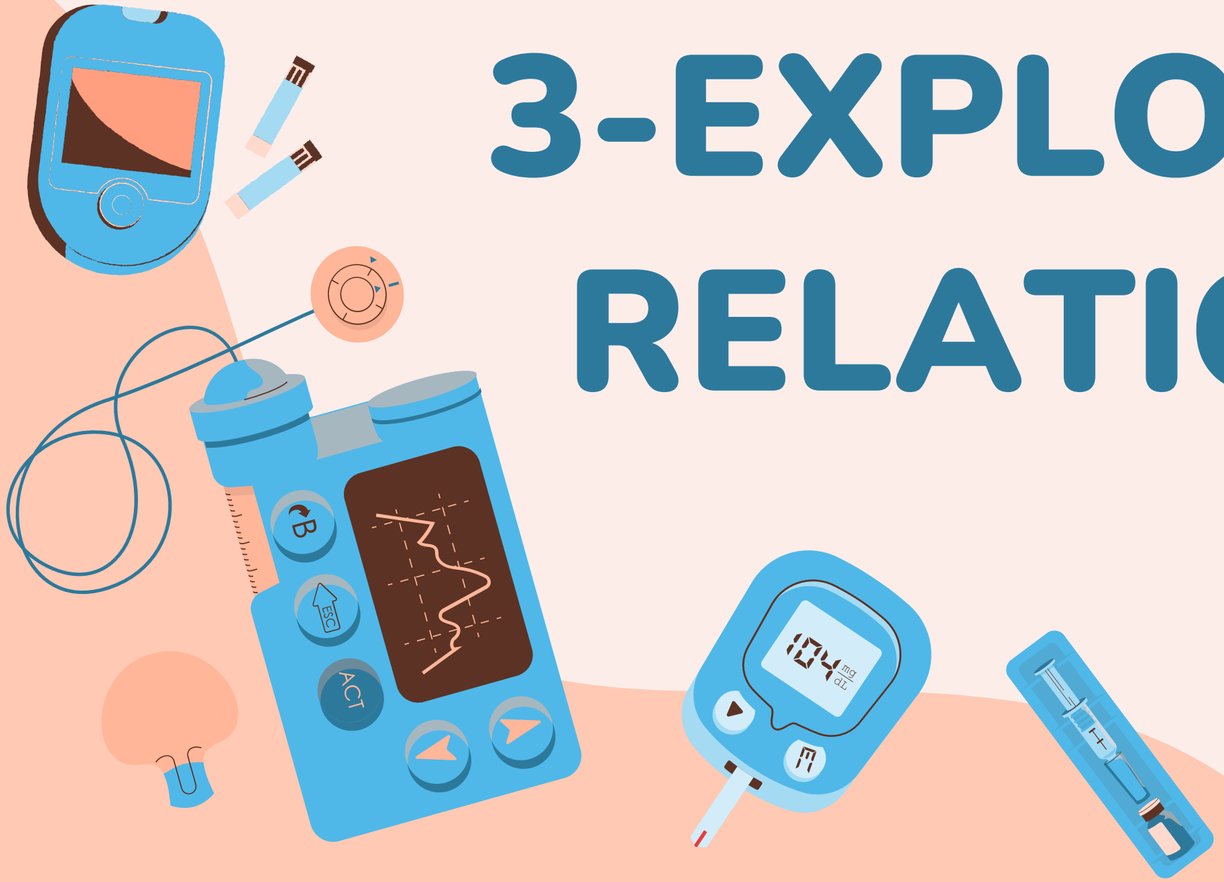
- Explore the relationship between BMI and glucose levels to identify patterns.

KEY FINDINGS :

- Correlation Coefficient: 0.22.
- Indicates a weak positive correlation: higher BMI slightly correlates with higher glucose levels.
- The scatter plot visualizes this relationship, showing a positive trend with significant variability.

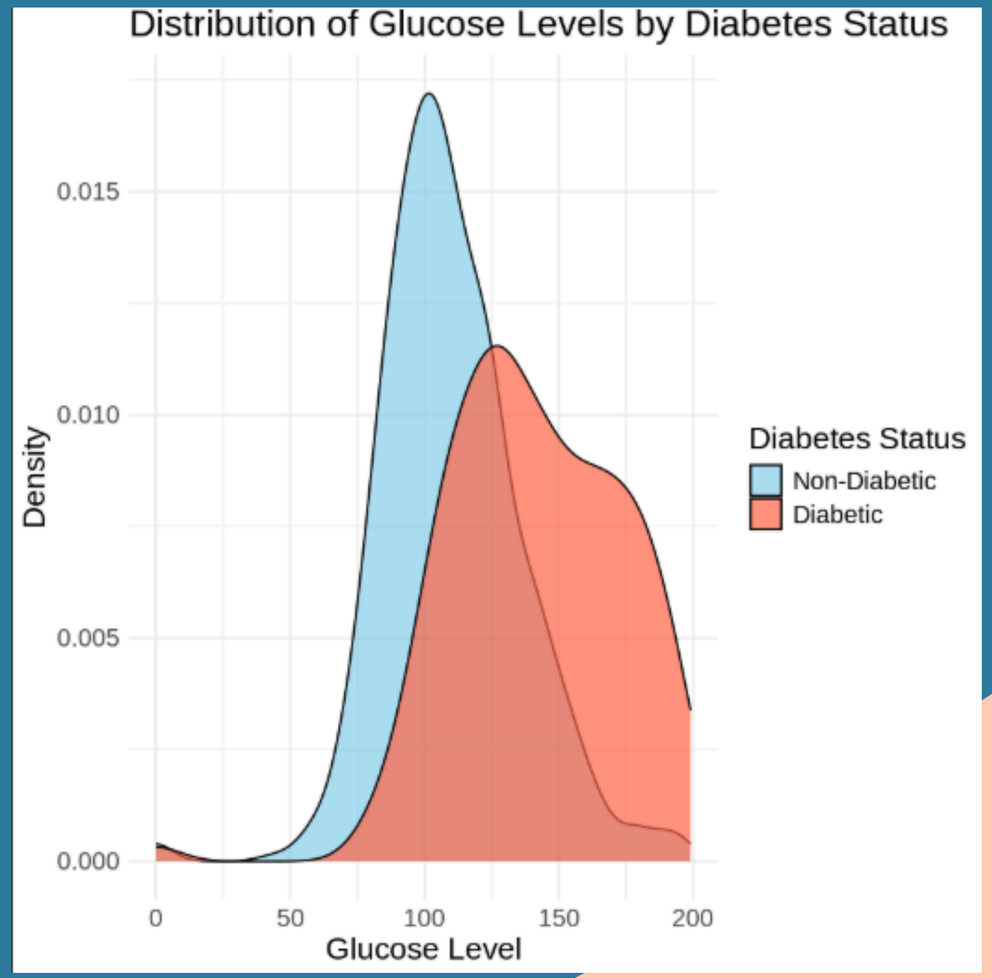


3-EXPLORING KEY RELATIONSHIPS



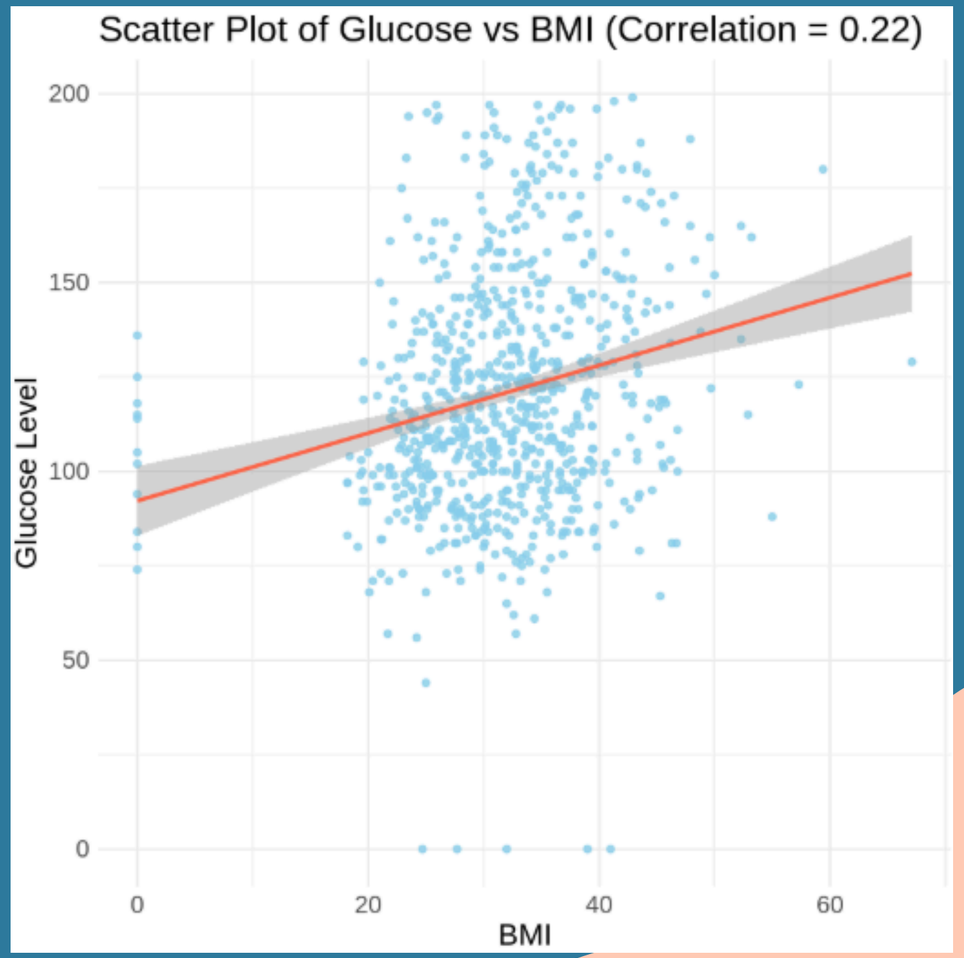
Are higher glucose levels associated with a greater likelihood of diabetes?

- Compare average glucose levels between diabetic and non-diabetic patients using a t-test.
- Diabetic patients have significantly higher glucose levels (average 141.26 vs. 109.98, p-value < 0.0001).



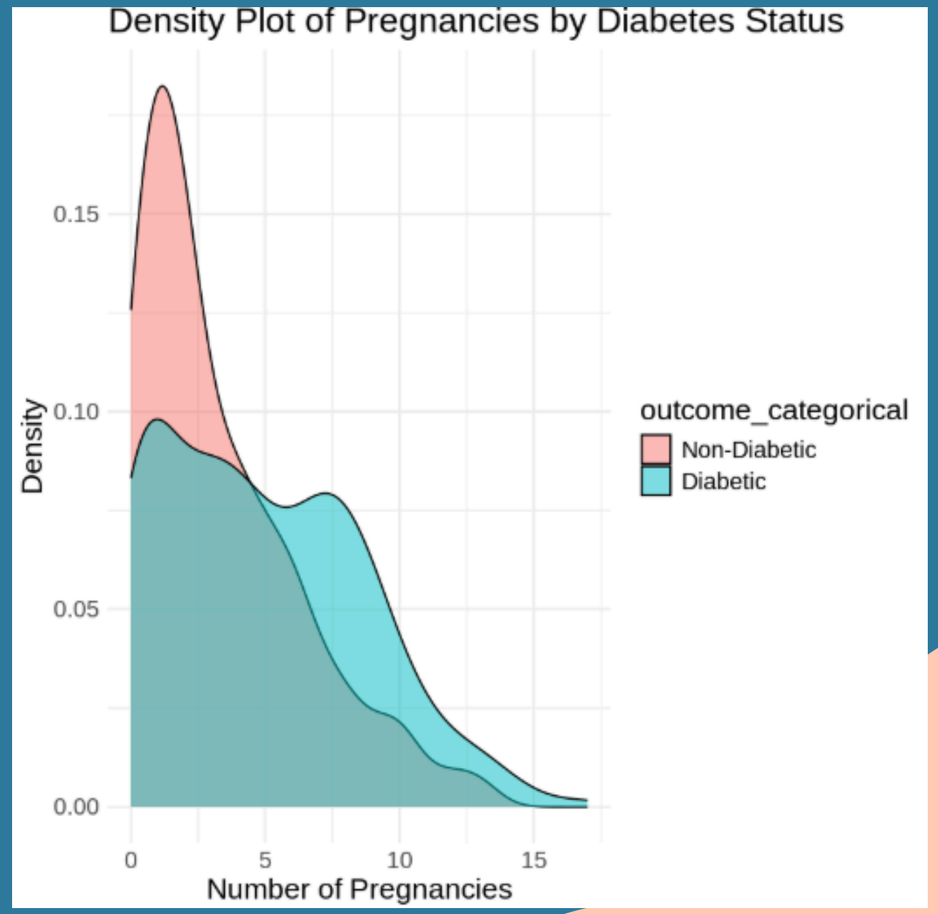
Are patients with high glucose concentrations also likely to have higher BMI values?

- Assess the correlation between glucose levels and BMI.
- Weak positive correlation (r = 0.22) indicating a slight association.



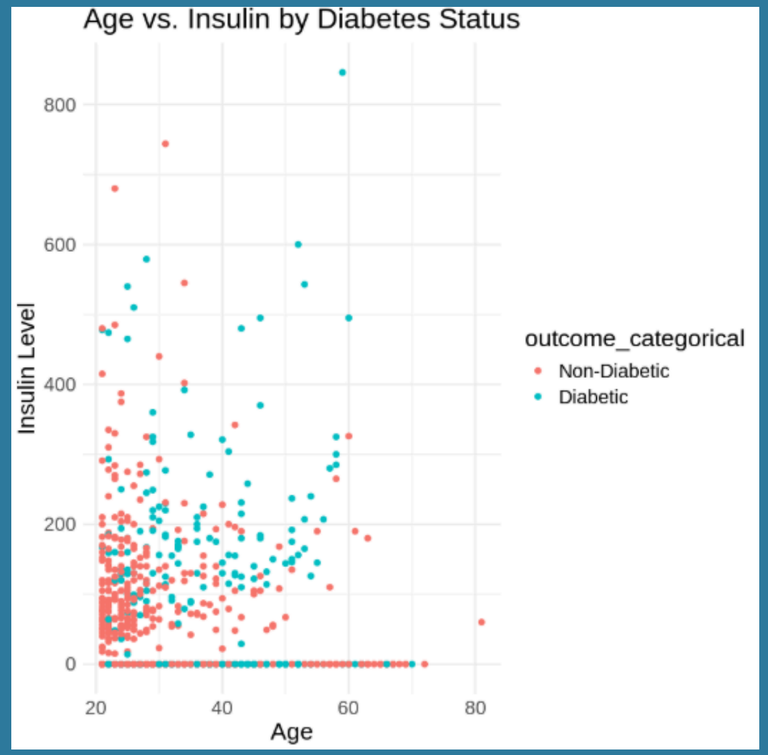
Are patients with a higher number of pregnancies at greater risk of developing diabetes?

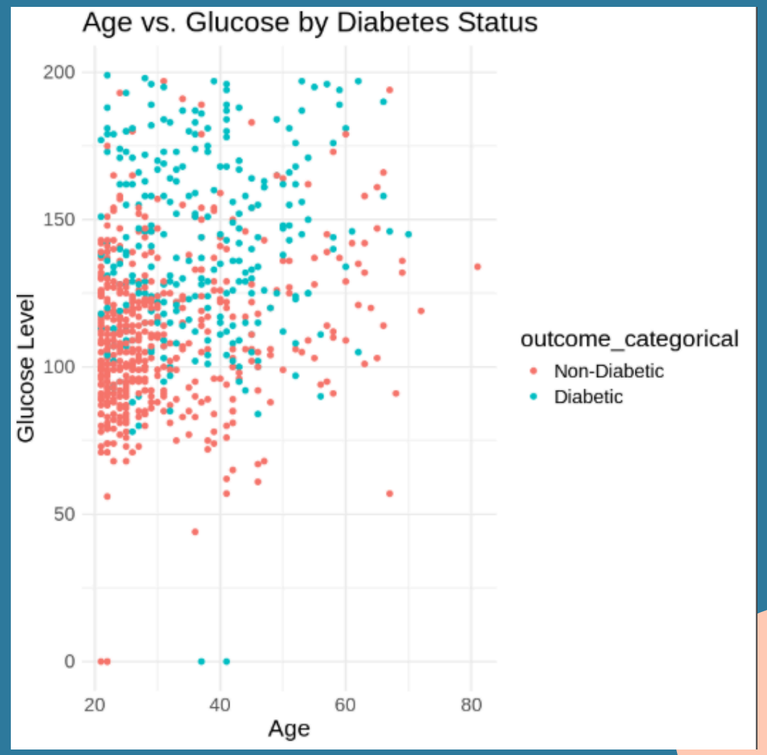
- Compare pregnancy rates between diabetic and non-diabetic patients using a t-test.
- Diabetic women have a slightly higher average number of pregnancies.



Are older patients more likely to have higher insulin concentrations and blood glucose levels?

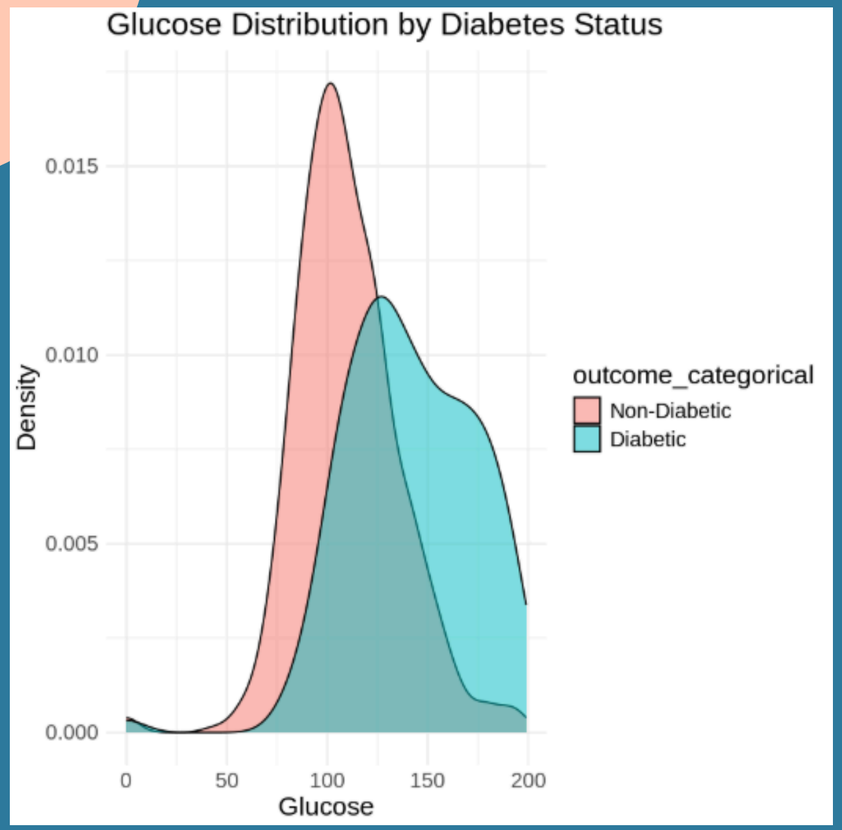
- Analyze the relationship between age, insulin, and glucose.
- Moderate positive correlation between age and glucose levels, but no significant relationship with insulin.

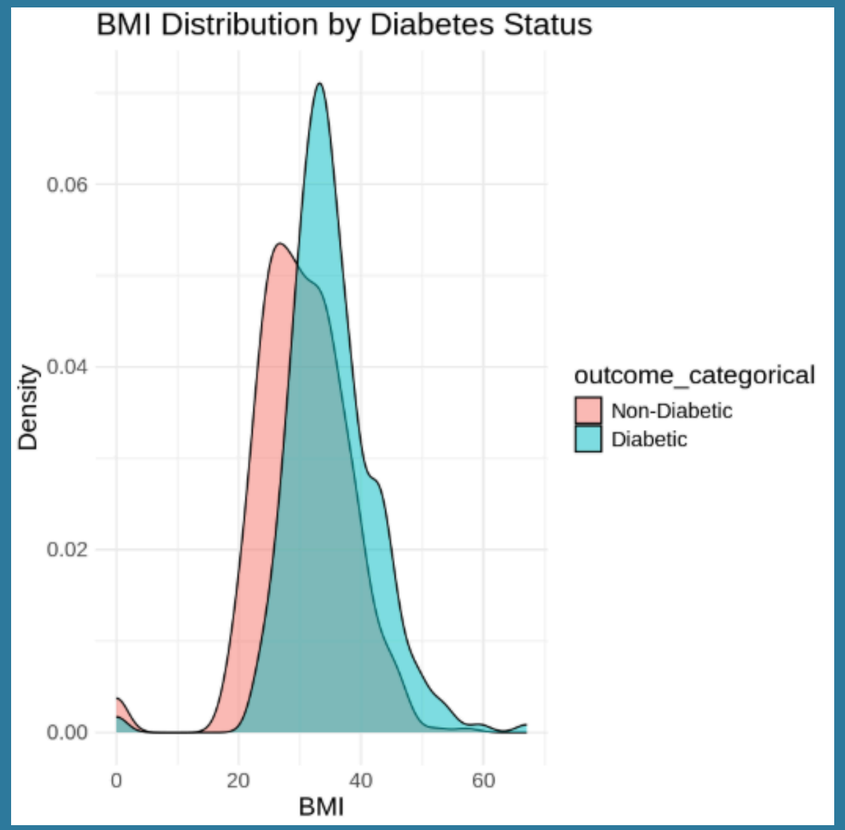


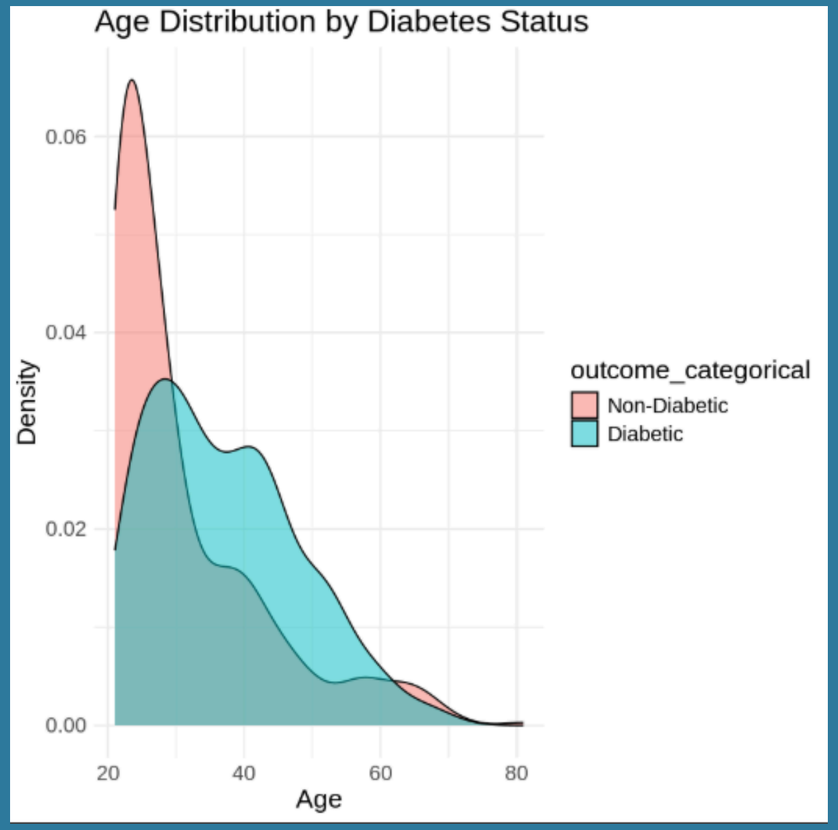


Can common “risk profiles” for diabetic patients be identified?

- Compare key metrics (glucose, BMI, and age) between diabetic and non-diabetic groups.
- Diabetic patients tend to have: Glucose > 140 mg/dL - BMI > 35 - Age > 40.

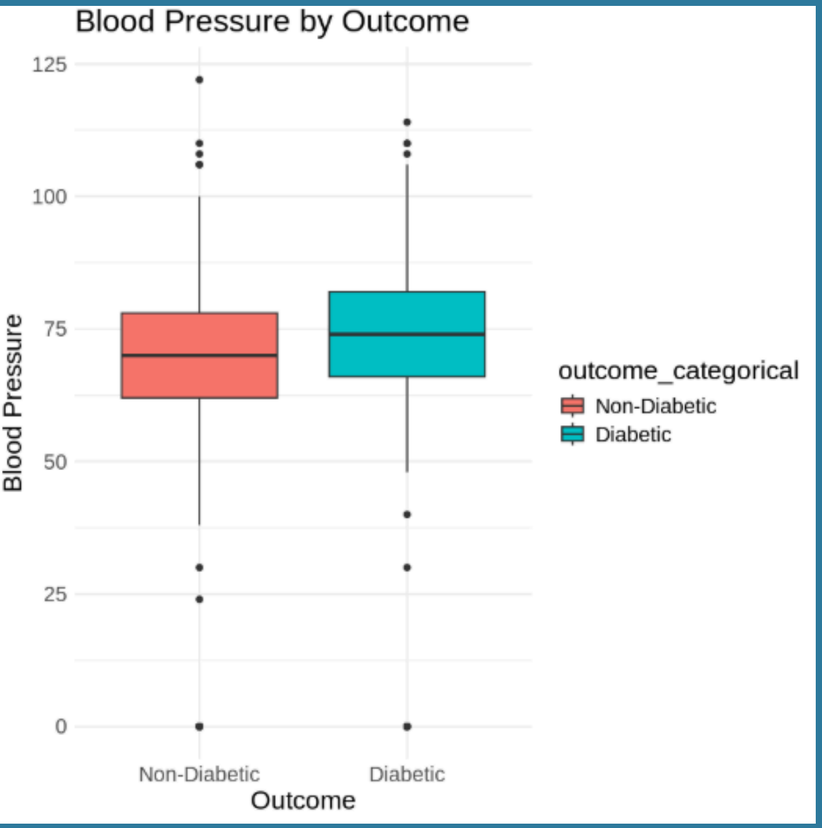




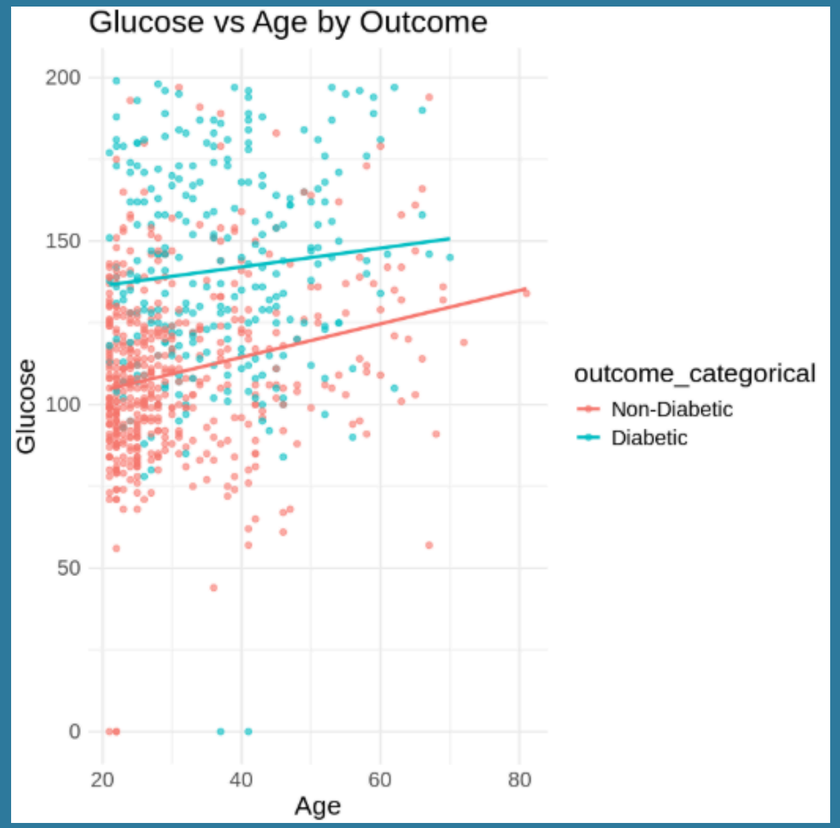


OTHER QUESTIONS

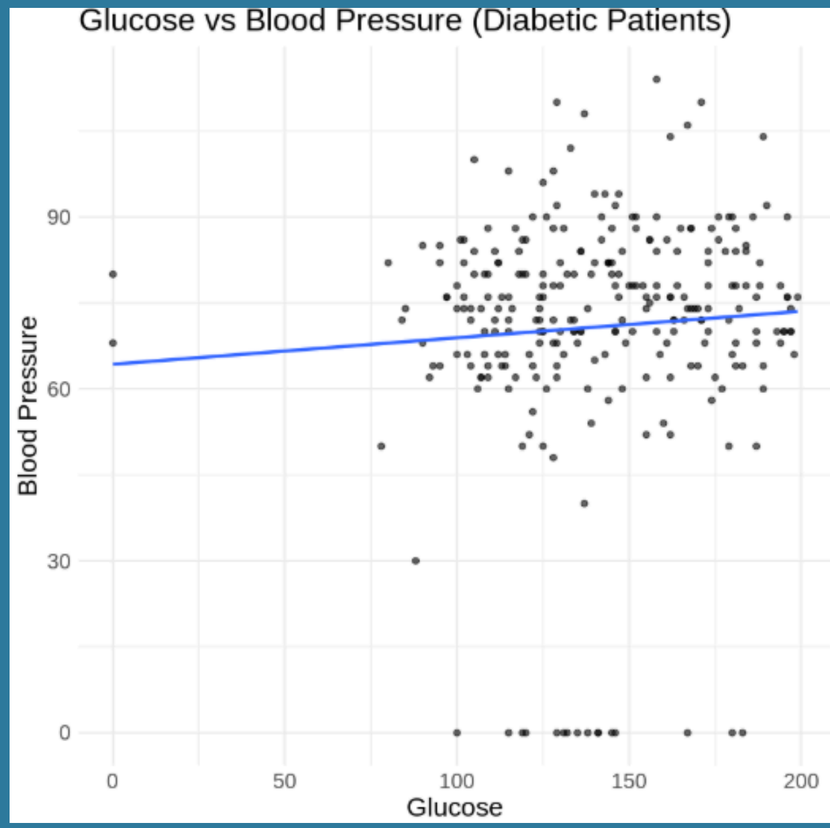
Blood Pressure by Outcome: Diabetics have slightly higher blood pressure.



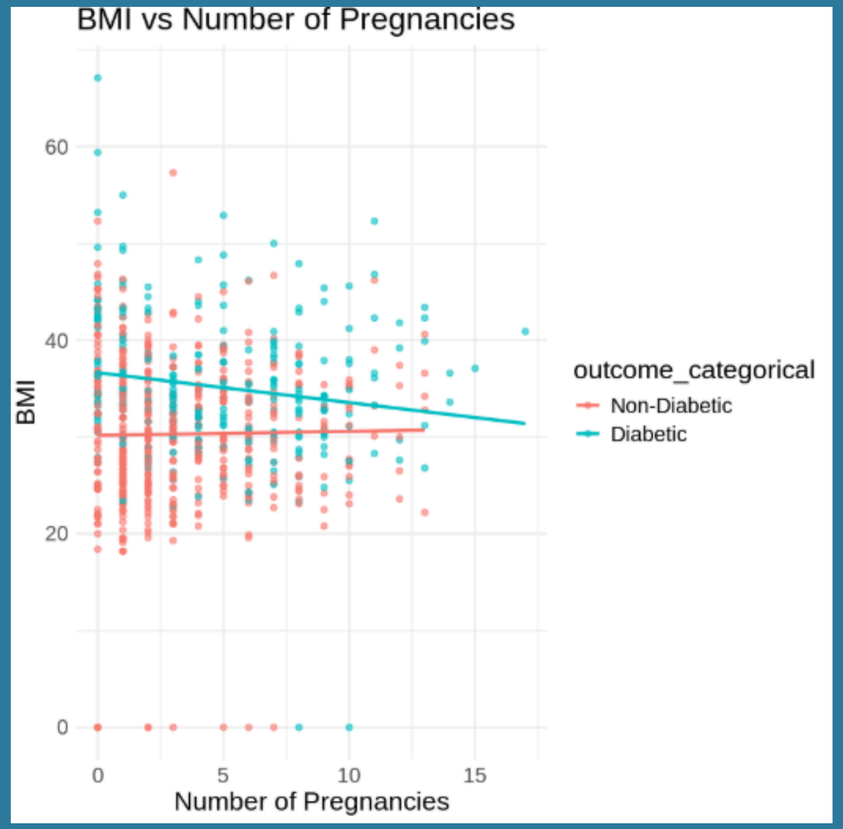
Glucose vs Age: Glucose rises with age, higher in diabetics.



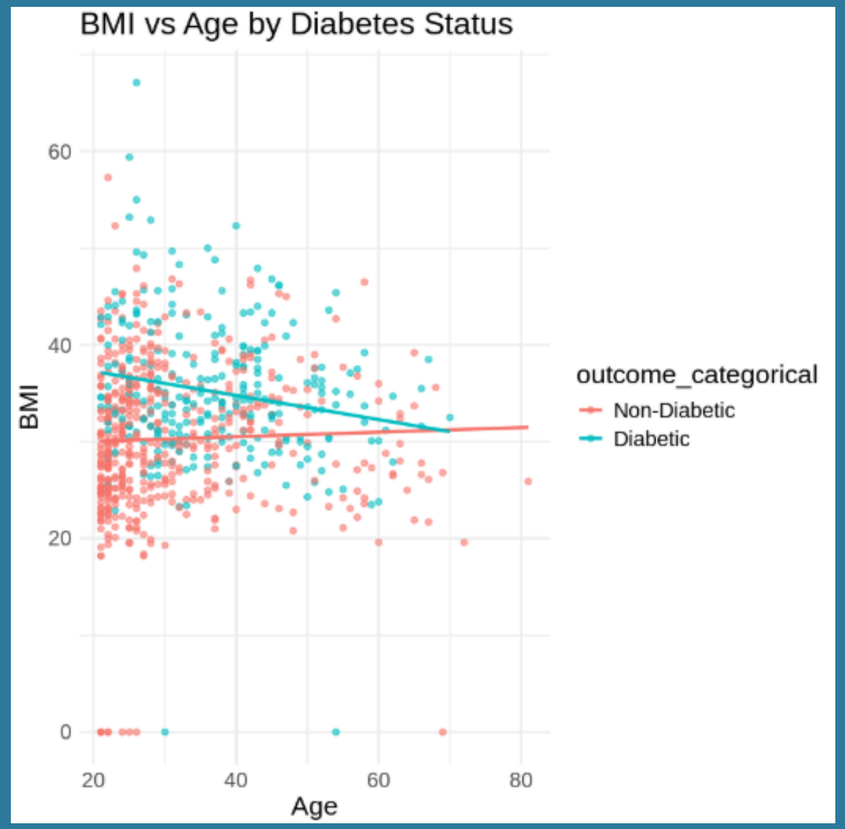
Glucose vs Blood Pressure: Weak positive correlation in diabetics.



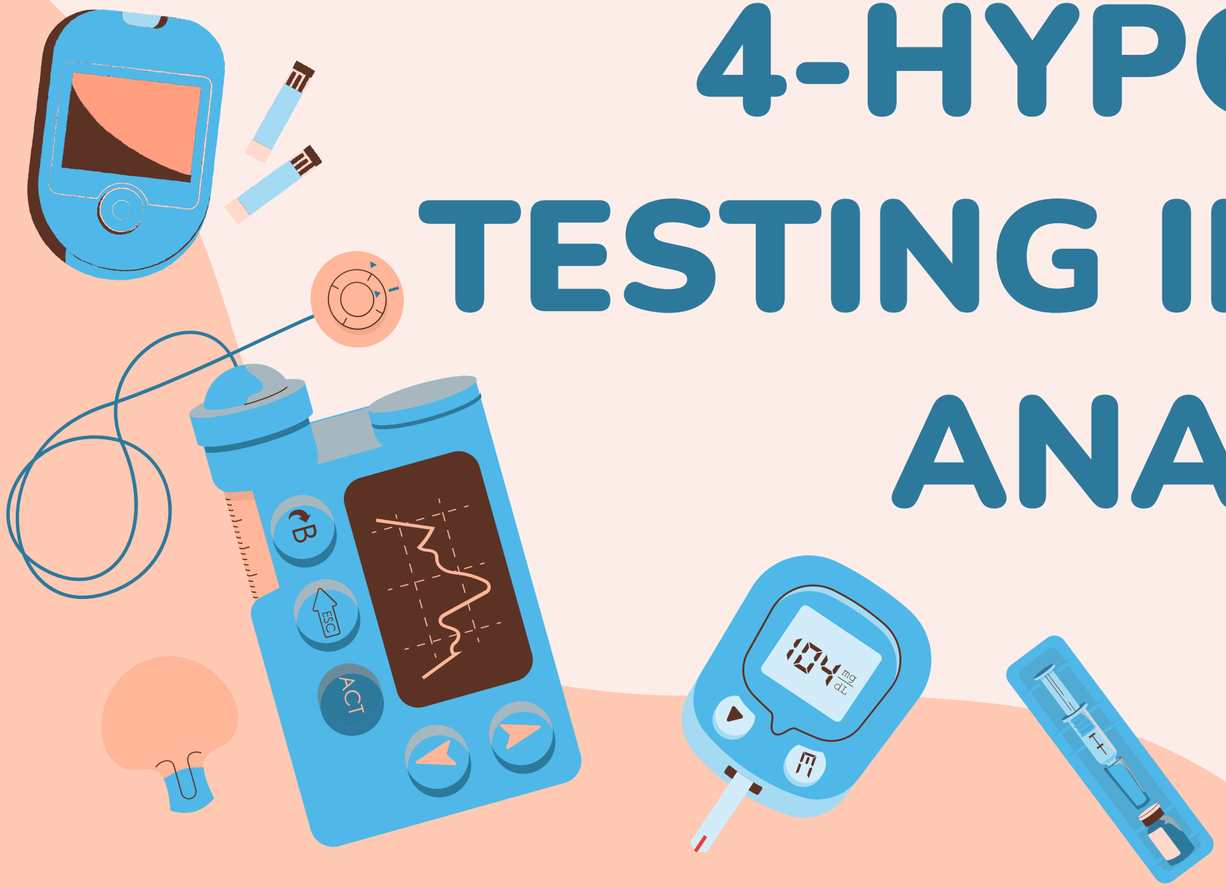
BMI vs Number of Pregnancies: Slight negative trend; diabetics have slightly higher BMI regardless of pregnancies.



BMI vs Age by Diabetes Status: Weak negative correlation between BMI and age, especially for diabetics.



4-HYPOTHESIS TESTING IN DIABETES ANALYSIS



4-HYPOTHESIS TESTING IN DIABETES ANALYSIS

WHAT IS HYPOTHESIS TESTING?

- A statistical method to assess the validity of claims using data.
- Helps determine whether observed differences or patterns are statistically significant.

GOAL :

- Evaluate specific claims related to diabetes and patient characteristics.

CLAIMS BEING TESTED :

- Claim 1: There is a significant difference in glucose levels between diabetic and non-diabetic patients
- Claim 2: Diabetic patients have a significantly higher BMI compared to non-diabetic patients.

(CLAIM 1: DIFFERENCE IN GLUCOSE LEVELS)

CLAIM :

- There is a significant difference in glucose levels between diabetic and non-diabetic patients.

STEPS :

- Statistical Test: Perform a two-sample t-test (Welch t-test) to compare means of glucose levels.
- Threshold: Compare the calculated p-value to the significance level ($\alpha = 0.05$).

RESULT :

- p-value: < 0.05
- The null hypothesis is rejected.

CONCLUSION :

- There is a statistically significant difference in glucose levels between diabetic and non-diabetic groups.

(CLAIM 2: BMI DIFFERENCES)

CLAIM :

- Diabetic patients have a significantly higher BMI compared to non-diabetic patients.

STEPS :

- 1.Statistical Test: Perform a one-sided t-test to evaluate whether diabetic patients have a higher BMI.Threshold: Compare the calculated p-value to the significance level ($\alpha = 0.05$).

- 2.Threshold: Compare the calculated p-value to the significance level ($\alpha = 0.05$).

RESULT :

- p-value: < 0.05
- The null hypothesis is rejected.

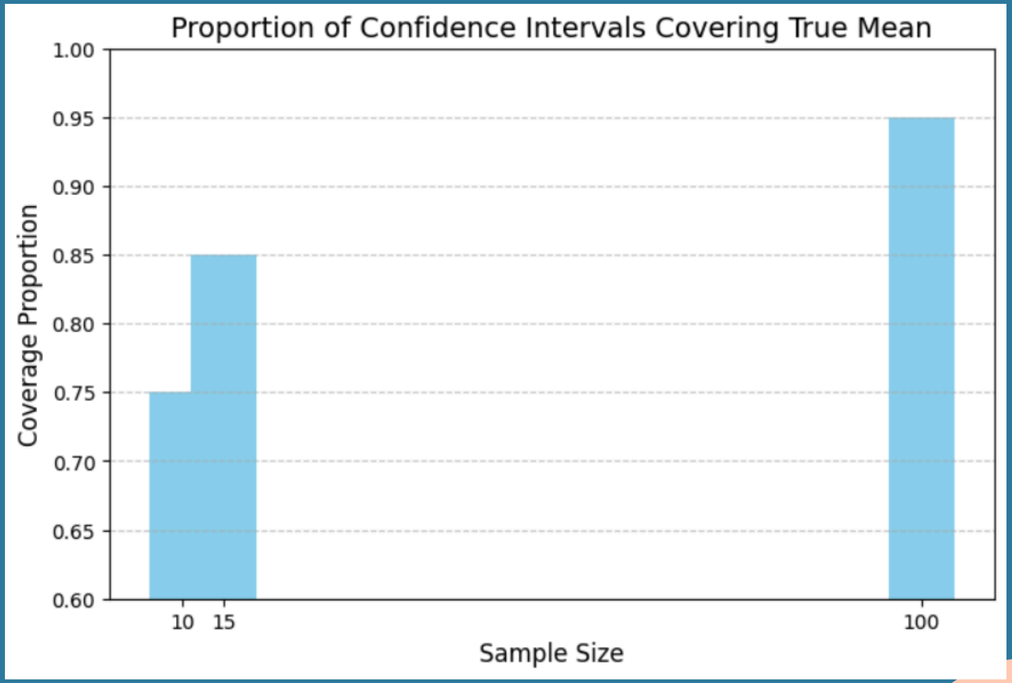
CONCLUSION :

- Diabetic patients have a statistically significant higher BMI compared to non-diabetic patients.

5-SIMULATION TASK

- **Objective:**
  - To evaluate the behavior of confidence intervals (CIs) for the population mean of glucose levels with varying sample sizes.

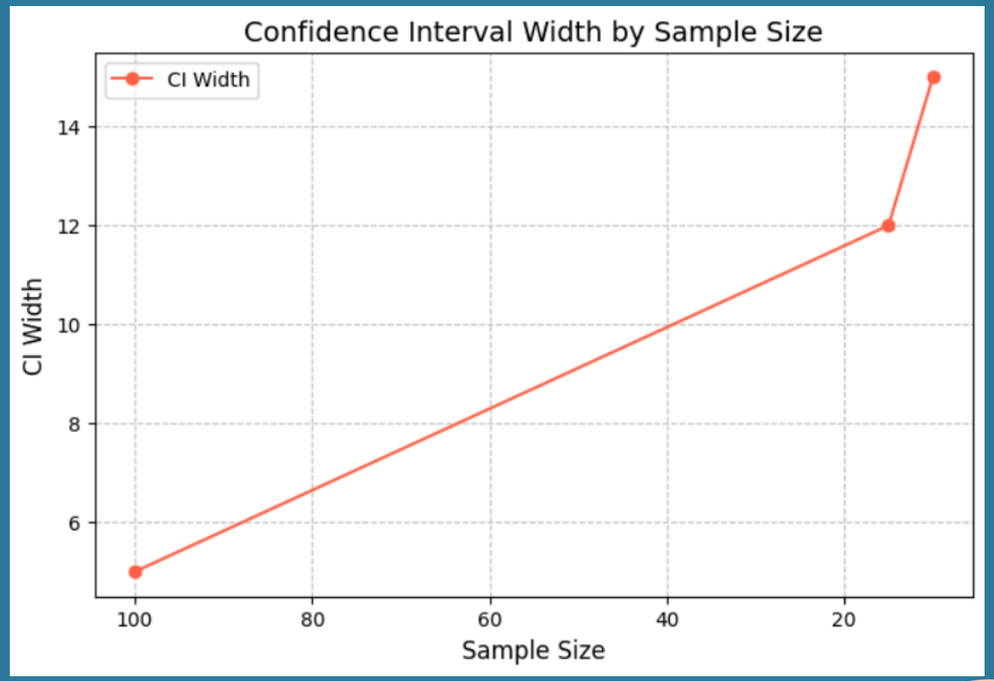
- **Approach:**
  - Randomly sampled glucose levels from the dataset for different sample sizes.
  - Calculated 95% confidence intervals for the sample mean in each case.
  - Compare the proportion of CIs containing the true population mean.



- **Scenarios:**
  - 25 samples of size 15: Moderate CI width; reasonable accuracy.
  - 25 samples of size 100: Narrower CIs; higher accuracy.
  - 20 samples of size 10: Wider CIs; lower accuracy.

- **Findings:**
  - Larger sample sizes result in narrower confidence intervals and better coverage of the true population mean.
  - Small sample sizes lead to less precision in estimating the true mean.

- **Key Insight:**
  - Increasing sample size improves the reliability of confidence interval estimates.



PREBARED BY:

- Mazen Ahmed 202202073
- Mahmoud Abdelrahman 202201819
- Ahmed Sameh 202201124
- Ahmed Barakat 2022200168
- Jilan Ismail 202201997