Descriptive Analysis

1. Data Structure

- Total records: **128 rows**
- Number of variables: **11**
- Data types:
    - Numerical: **Age, BMI, FBS, HbA1c**
    - Categorical: **Gender, Blood Pressure, Family History, Smoking, Diet, Exercise, Diagnosis**

2. Missing Values

- BMI: **3 missing**
- Family History of Diabetes: **2 missing**
- Smoking: **2 missing**
- Diet: **2 missing**
- Exercise: **1 missing**
- After removing rows with missing values, around **125 records remain.**

3. Categorical Summary

3.1 Gender

- Male: 61
- Female: 57

The gender distribution is fairly balanced.

3.2 Blood Pressure

- High: 73
- Normal: 35
- Low: 10

Most participants have high blood pressure.

3.3 Family History of Diabetes

- No: 71
- Yes: 47

Most people do not have a family history of diabetes.

3.4 Smoking

- Yes: 74
- No: 44

The number of smokers is almost twice that of non-smokers.

3.5 Diet

- Poor: 74
- Healthy: 44

More than half of the participants have a poor diet.

3.6 Exercise

- No: 74
- Regular: 44

A clearly larger group does not exercise regularly.

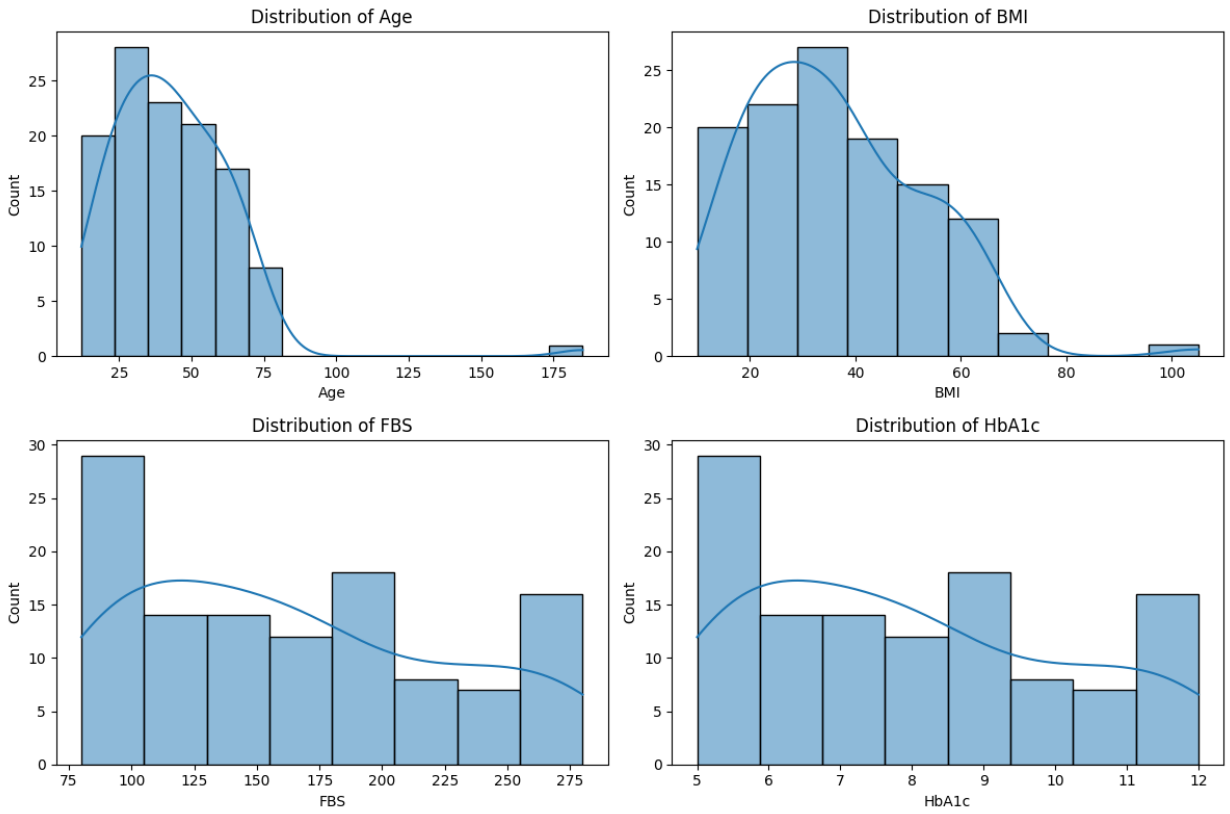
3.7 Diagnosis (Diabetes result)

- 0 = Not diabetic: 88
- 1 = Diabetic: 30

About 23% of the sample are diabetic.

Exploratory Data Analysis (EDA)

1. Distribution of Numerical Variables



1.1 Age

- Most values fall between 30-70 years.
- An outlier was found at around 180 years.

  This should be reviewed or removed before modelling.

1.2 BMI

- The average BMI is high, mostly between 25-35 ⟶ Overweight / Obese range.
- An Outlier above 100 is clearly incorrect.

1.3 FBS (Fasting Blood Sugar)

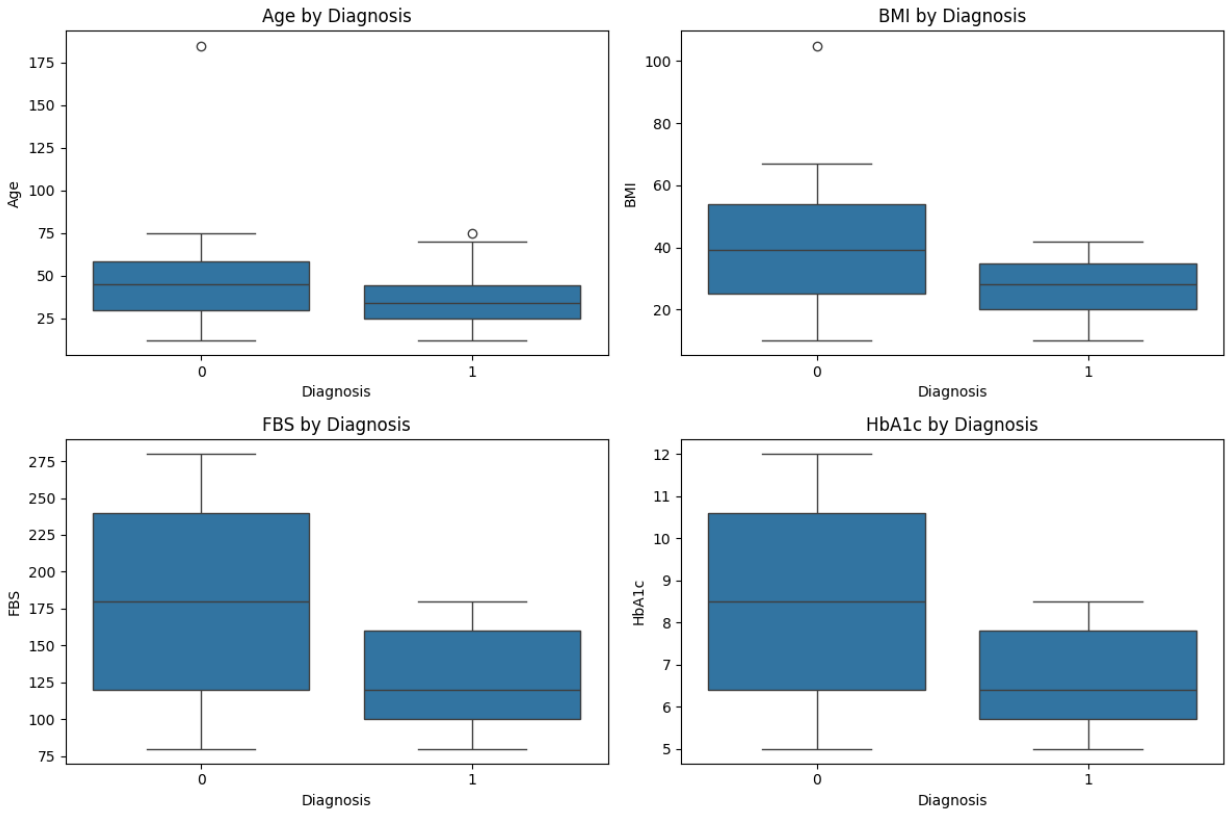
- Most values are above 120 mg/dL.

  This indicates many participants are at risk of diabetes.

1.4 HbA1c

- Values range from 5% to 12%.
- Those diagnosed with diabetes (Diagnosis = 1) have noticeably higher HbA1c levels.

2. Boxplots by Diagnosis



2.1 Age

- Diabetic participants do not appear significantly older.

  Age may not be a key factor in this dataset.

2.2 BMI

- The diabetic group has a clearly higher average BMI.

  BMI is an important variable.

2.3 FBS

- Non-diabetic participants show a wider spread of FBS values.
- Diabetic participants have higher FBS overall.

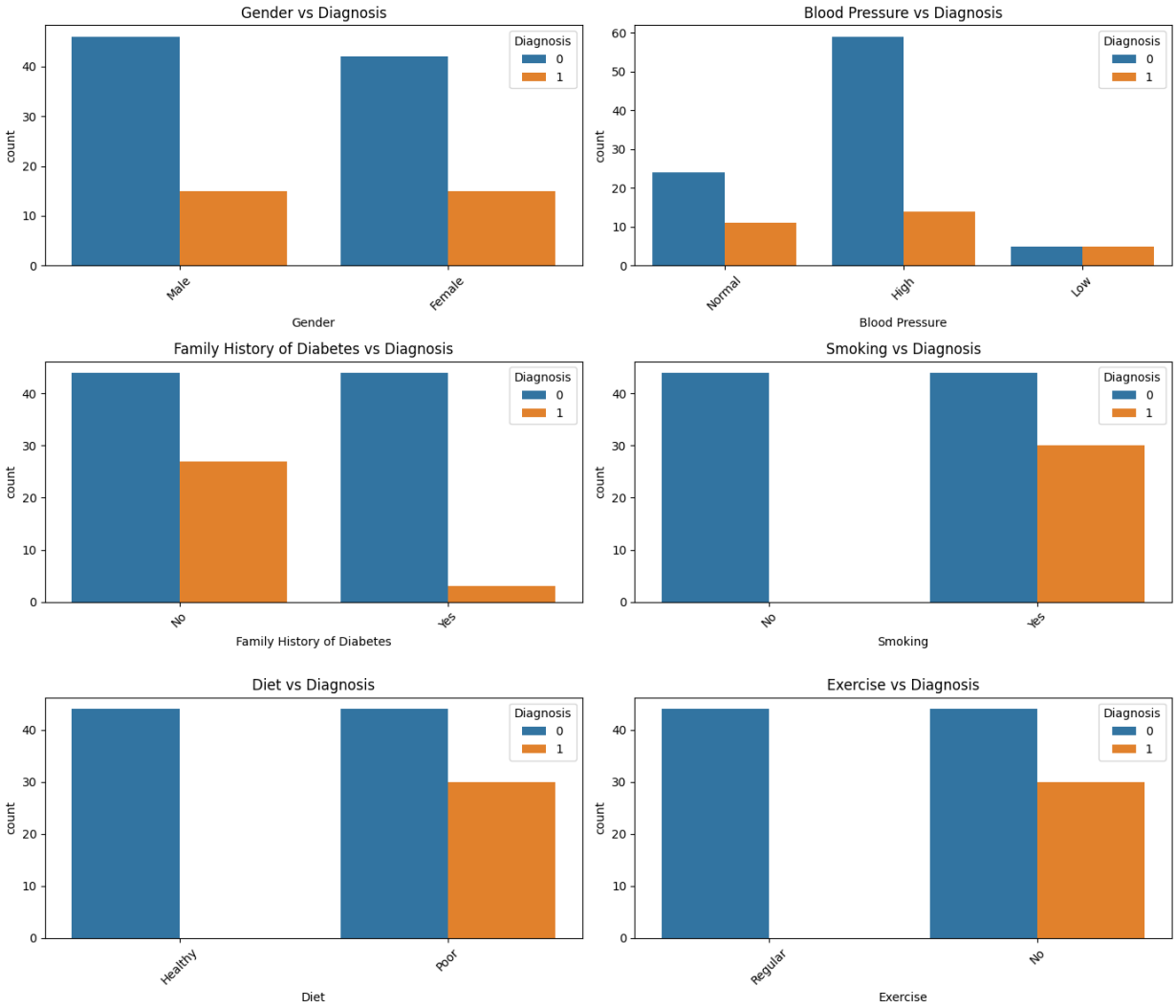
  There is a positive relationship between FBS and diabetes.

2.4 HbA1c

- Higher HbA1c values in the diabetic group support medical expectations.

  HbA1c strongly distinguishes diabetic cases.

3. Count Plots (Categorical Variables vs Diagnosis)



3.1 Gender vs Diagnosis

- No major difference between males and females.

  Gender is not a strong factor.

3.2 Blood Pressure vs Diagnosis

- The High BP group includes more diabetic cases.

  High blood pressure is linked to higher diabetes risk.

3.3 Family History vs Diagnosis

- Those with a family history have a higher rate of diabetes.

  Genetics plays an important role.

3.4 Smoking vs Diagnosis

- Smokers have more cases of Diagnosis = 1.

  Smoking may increase the risk.

3.5 Diet vs Diagnosis

- Poor diet is strongly associated with higher diabetes rates.

  Eating habits play a significant role.

3.6 Exercise vs Diagnosis

- Those who do not exercise show a higher rate of diabetes.

  Exercise clearly affects diabetes risk.