Day 3 – Diabetes Dataset Findings

1. Dataset Overview

Size: 200 patient records
Variables: 9 columns — Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome
Outcome: 0 = No diabetes, 1 = Diabetes

2. Data Cleaning Steps

1. Filled missing numeric values (Glucose, BloodPressure, BMI) with mean values. 2. Removed duplicates to ensure unique patient records. 3. Renamed BMI → Body_Mass_Index for clarity. 4. Verified all data types as numeric.

3. Key Patterns Found

- Patients with diabetes tend to have higher average glucose levels. - Older patients (50+) show a slightly higher prevalence of diabetes. - Higher BMI and Diabetes Pedigree Function values correlate with greater diabetes likelihood.

4. Visualizations

Figure 1 – Age Distribution

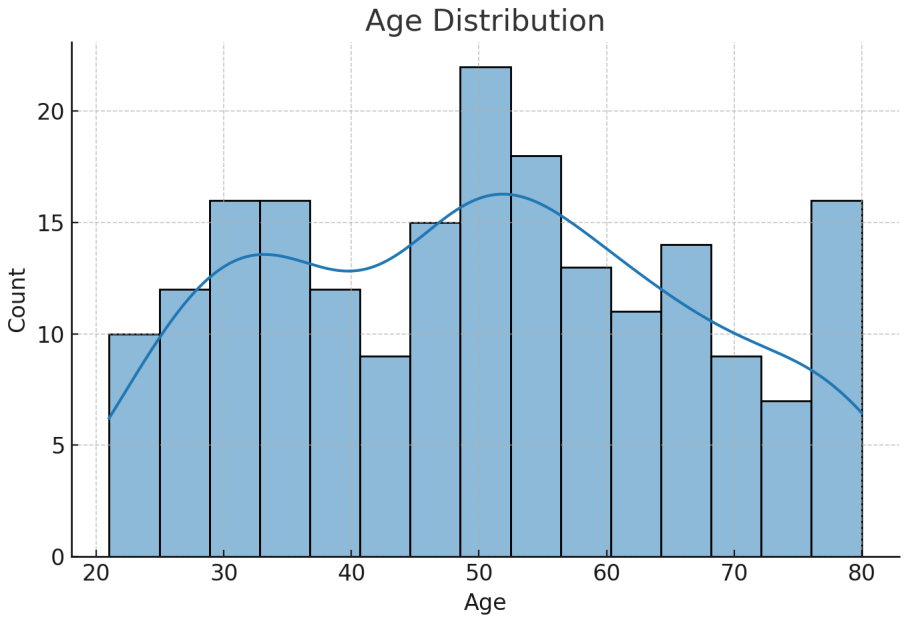

Figure 2 – Correlation Heatmap

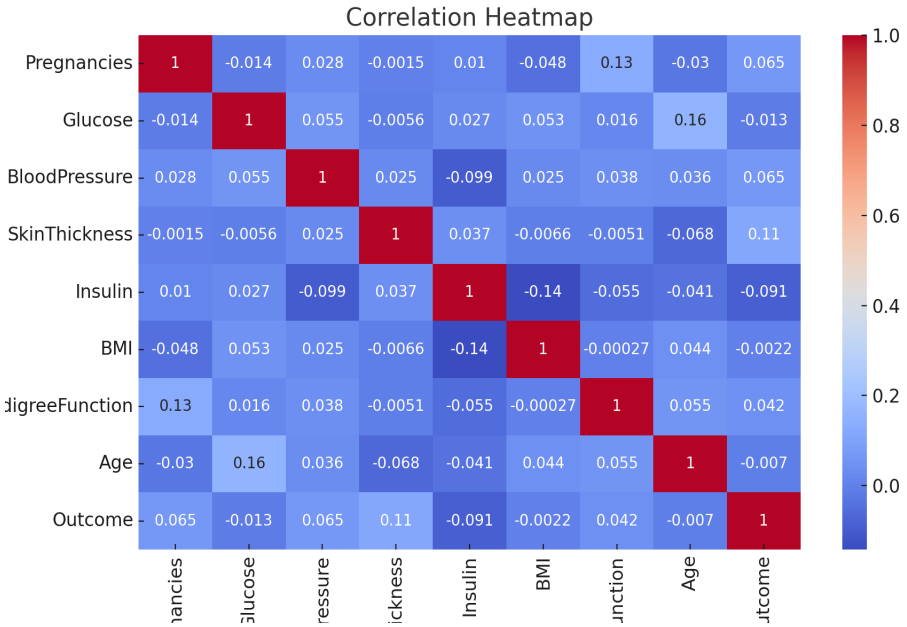

Figure 3 – Average Glucose by Outcome