**Diabetes Dataset Findings**

**1. Dataset Overview**

* **Size:** 200 patient records
* **Variables:** 9 columns — Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome
* **Outcome:** 0 = No diabetes, 1 = Diabetes

**2. Data Cleaning Steps**

1. **Handled Missing Data:** Filled missing numeric values (Glucose, BloodPressure, BMI) with their respective mean values.
2. **Removed Duplicates:** Ensured each patient record was unique.
3. **Renamed Columns:** Changed BMI → Body\_Mass\_Index for clarity.
4. **Data Types Verified:** All variables confirmed in correct format (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 & Insights**

**Figure 1 – Age Distribution**  
  
*Younger adults make up a large proportion of the dataset, but diabetes prevalence rises in older groups.*

**Figure 2 – Correlation Heatmap**  
  
*Glucose has the strongest positive correlation with diabetes outcome.*

**Figure 3 – Average Glucose by Outcome**  
  
*Patients with diabetes (Outcome = 1) have significantly higher average glucose.*

✅ **Summary:** The analysis confirms known clinical patterns — higher glucose, higher BMI, and older age are key risk factors. This cleaned and analyzed dataset is now ready for use in predictive modeling or further statistical testing.