Healthcare Data Analysis – Diabetes Risk Report

1. Introduction

This report presents a comprehensive analysis of diabetes risk using statistical exploration, graphical visualizations, and insights derived from multiple health indicators. The goal is to identify the strongest predictors of diabetes and understand how patient characteristics such as glucose levels, BMI, age, and pregnancies influence diabetes outcomes.

2. Dataset Summary

The dataset contains **768 patient records**, each with the following features:

- **Pregnancies**
- **Glucose**
- **BloodPressure**
- **SkinThickness**
- **Insulin**
- **BMI**
- **DiabetesPedigreeFunction**
- **Age**
- **Outcome** (0 = No Diabetes, 1 = Diabetes)

Outcome distribution:

- **268 patients** tested *positive* for diabetes
- **500 patients** tested *negative*
- **Diabetes prevalence: ~34.9%**

3. Data Cleaning & Preprocessing

To ensure accuracy:

- Zero values in columns like *Glucose, BMI, BloodPressure, SkinThickness,* and *Insulin* were treated as missing.
- Missing values were imputed using the **mean strategy**.
- Numeric variables were kept in their original units to preserve interpretability.

This cleaning process ensured that the dataset was suitable for statistical and visual analysis.

4. Key Risk Factors Identified

1. Glucose

Glucose has the *strongest correlation* with diabetes outcome (**0.49 correlation**).
Higher glucose levels significantly increase the likelihood of diabetes.

2. BMI

BMI has a moderate correlation (**0.31 correlation**) with diabetes.
Higher BMI values are strongly associated with diabetes risk.

3. Age

Age shows a positive correlation (**0.24 correlation**) with diabetes.
Older patients have a higher probability of having diabetes.

4. Pregnancies

A correlation of **0.22** indicates that women with more pregnancies are more at risk.
These findings are supported by the correlation heatmap.

5. Visual Analysis

A. Diabetes Outcome Count

The majority of patients (**500**) do not have diabetes, while **268** are diabetic.
This imbalance shows moderate skewness but is acceptable for analysis.

B. Glucose Level Distribution

Glucose levels follow a near-normal distribution, peaking around 100–130 mg/dL.
Patients with higher glucose levels (above ~140 mg/dL) are more likely to be diabetic.

C. Age vs BMI Scatter Plot

The scatter plot shows:

- Higher BMI is common across all age groups.
- Diabetic patients tend to cluster at **higher BMI** and **higher age** ranges.
- Younger, low-BMI individuals rarely show positive diabetes outcomes.

D. Correlation Heatmap

Strongest relationships observed:

- **Glucose ↔ Outcome (0.49)**
- **BMI ↔ SkinThickness (0.54)**
- **Age ↔ Pregnancies (0.54)**

The heatmap supports the conclusion that glucose is the top predictor.

6. Insights & Interpretation

1. **Glucose is the dominant predictor**
   High glucose levels strongly indicate diabetes risk.
2. **BMI is a major health factor**
   Overweight individuals show higher diabetes probability.
3. **Age increases risk**
   Older adults have compromised metabolic efficiency.
4. **Pregnancy history matters**
   Multiple pregnancies may contribute to long-term metabolic stress.
5. **Insulin and SkinThickness correlations are weaker**
   These indicators alone are not sufficient predictors.

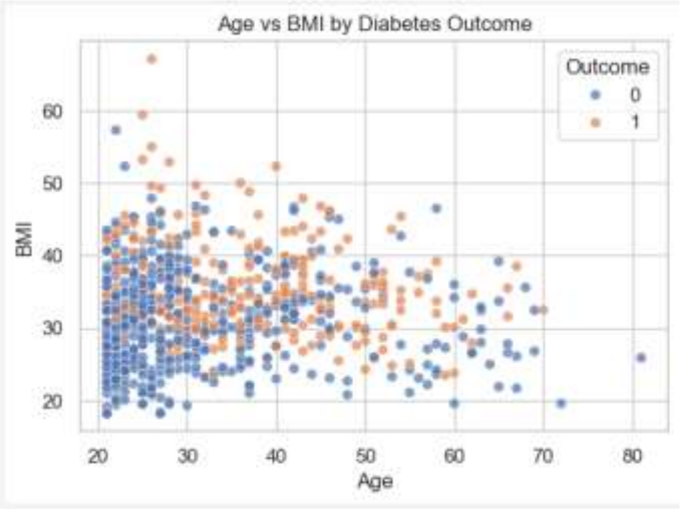
7. Conclusion

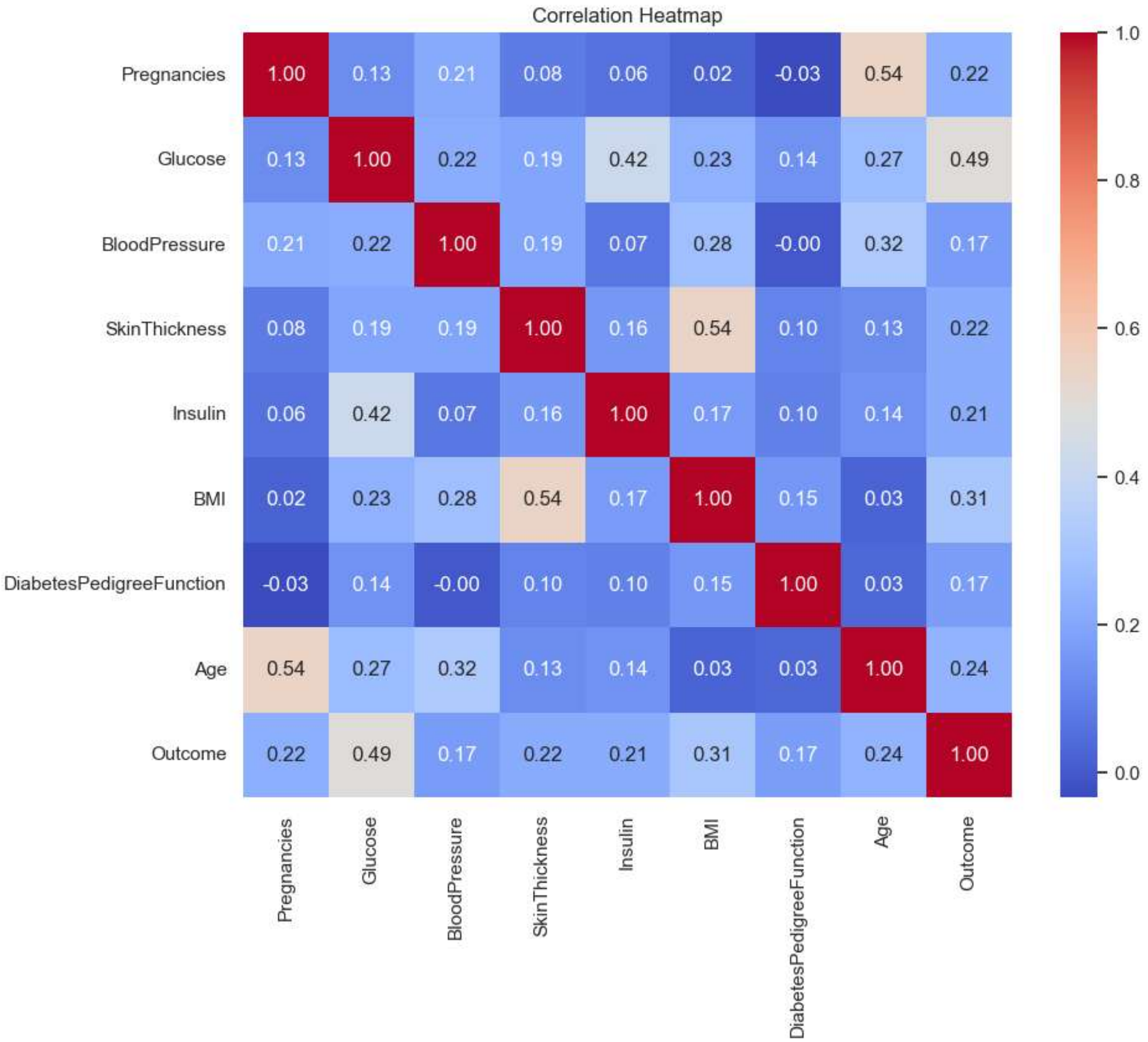
This analysis highlights that:

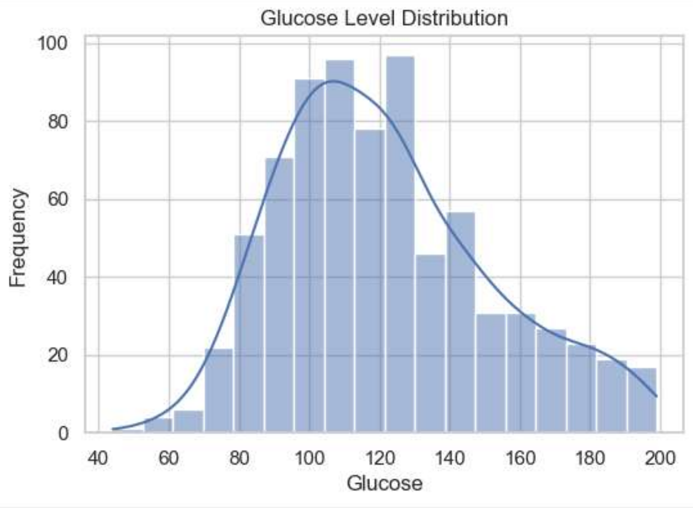
- **Glucose**, **BMI**, **Age**, and **Pregnancies** are the top predictors of diabetes.
- Glucose has the **strongest statistical relationship** with diabetes.
- Preventive healthcare should focus on lifestyle interventions targeted at high-BMI and older individuals.
- Early screening for high-glucose individuals can significantly reduce long-term risks.

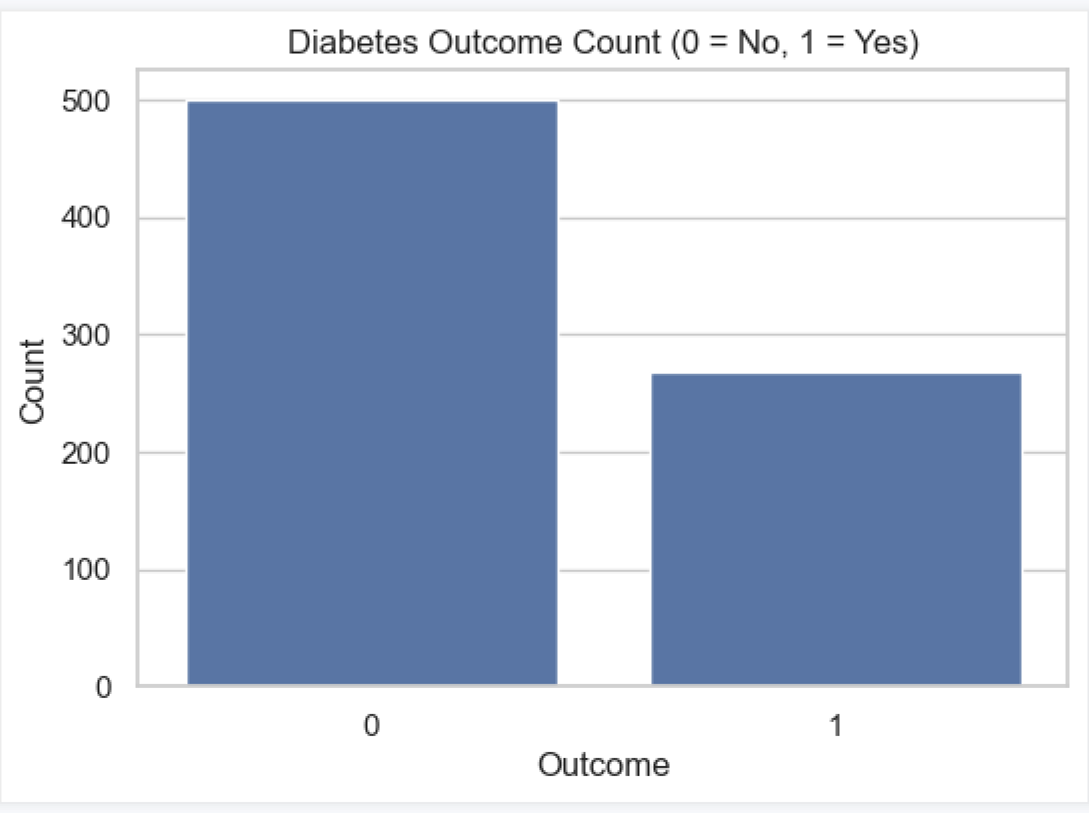
The results provide valuable insights for developing early detection systems and improving health programs.

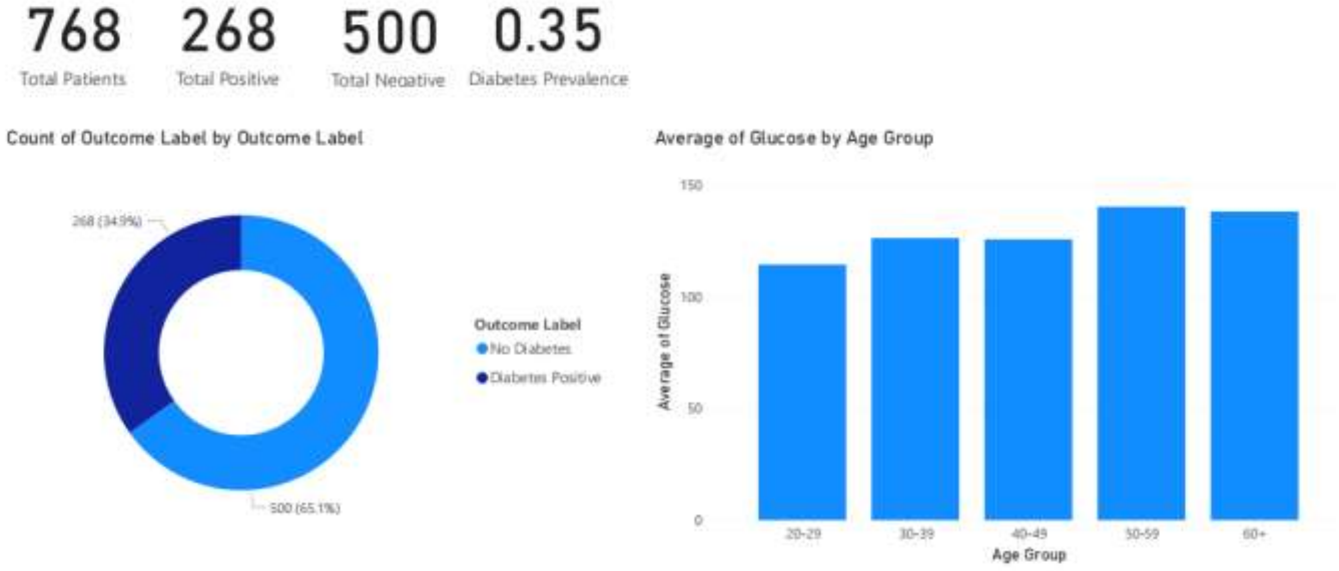
OUTPUTS:




Correlation Heatmap
Pregnancies 1.00 0.13 0.21 0.08 0.06 0.02 -0.03 0.54 0.22
Glucose 0.13 1.00 0.22 0.19 0.42 0.23 0.14 0.27 0.49
BloodPressure 0.21 0.22 1.00 0.19 0.07 0.28 -0.00 0.32 0.17
SkinThickness 0.08 0.19 0.19 1.00 0.16 0.54 0.10 0.13 0.22
Insulin 0.06 0.42 0.07 0.16 1.00 0.17 0.10 0.14 0.21
BMI 0.02 0.23 0.28 0.54 0.17 1.00 0.15 0.03 0.31
DiabetesPedigreeFunction -0.03 0.14 -0.00 0.10 0.10 0.15 1.00 0.03 0.17
Age 0.54 0.27 0.32 0.13 0.14 0.03 0.03 1.00 0.24
Outcome 0.22 0.49 0.17 0.22 0.21 0.31 0.17 0.24 1.00
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
1.0
0.8
0.6
0.4
0.2
0.0


Glucose Level Distribution
Frequency
100
80
60
40
20
0
40
60
80
100
120
140
160
180
200
Glucose

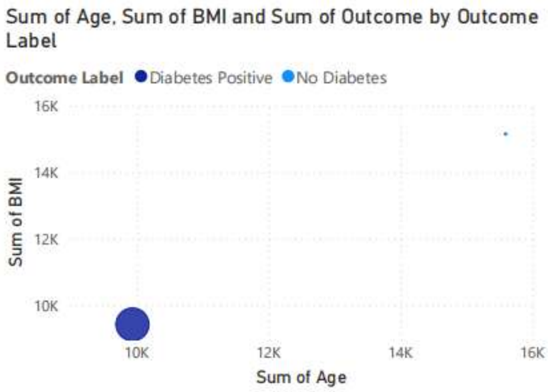


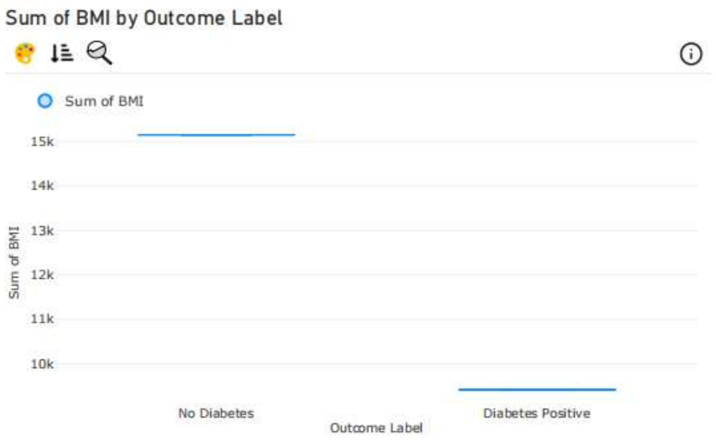


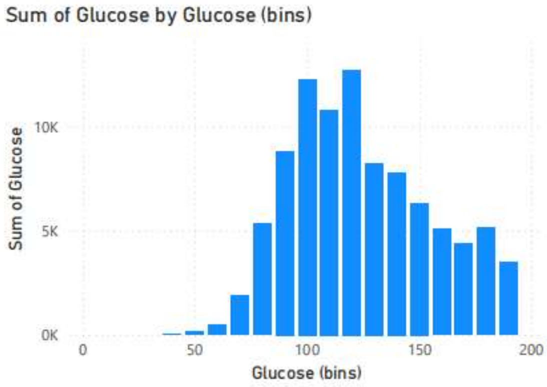
768	268	500	0.35
Total Patients	Total Positive	Total Negative	Diabetes Prevalence

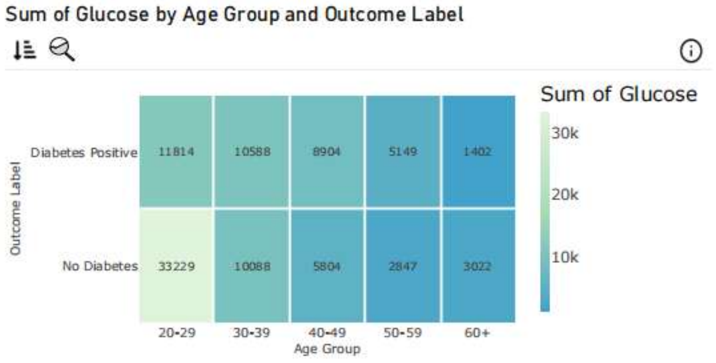
Count of Outcome Label by Outcome Label

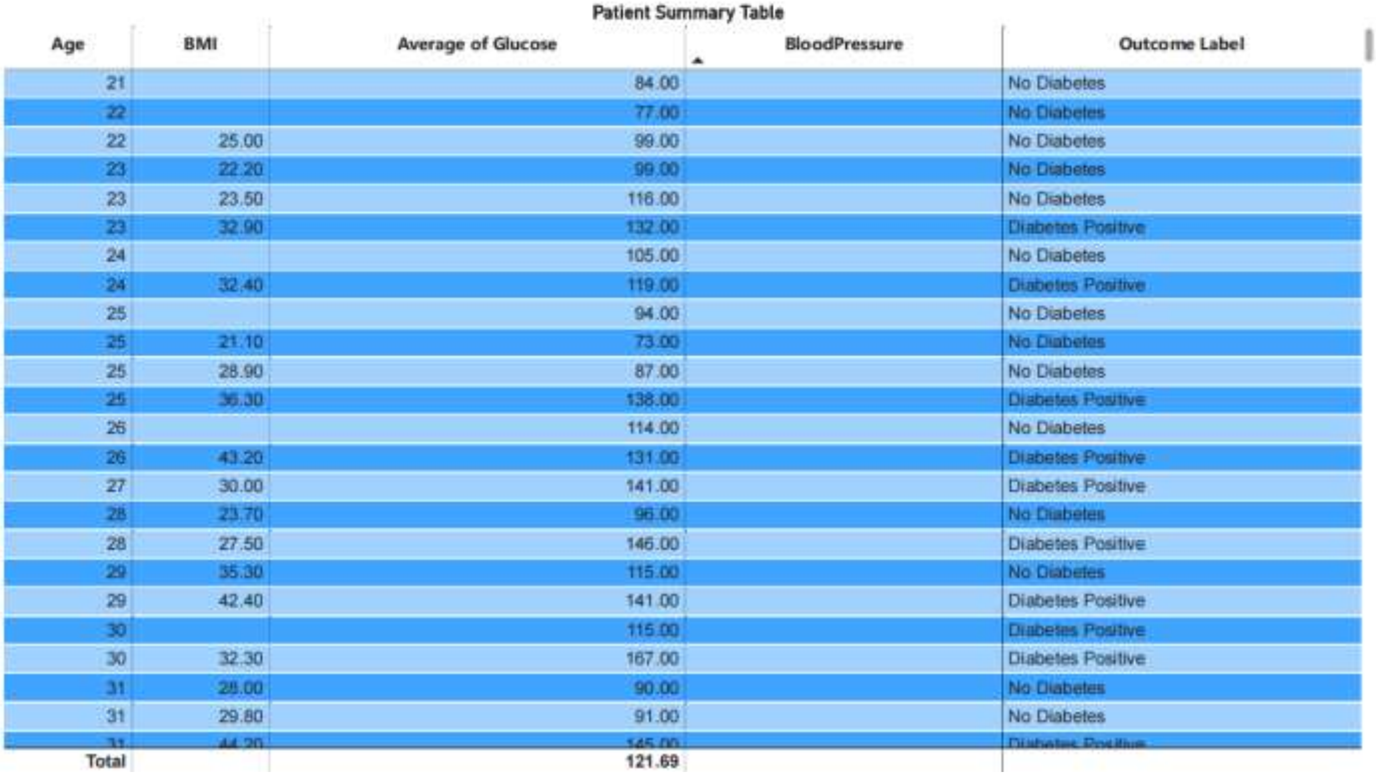
Average of Glucose by Age Group


Sum of Age, Sum of BMI and Sum of Outcome by Outcome Label


Sum of BMI by Outcome Label


Sum of Glucose by Glucose (bins)


Sum of Glucose by Age Group and Outcome Label


Patient Summary Table

Age	BMI	Average of Glucose	BloodPressure	Outcome Label
21		84.00		No Diabetes
22		77.00		No Diabetes
22	25.00	99.00		No Diabetes
23	22.20	99.00		No Diabetes
23	23.50	116.00		No Diabetes
23	32.90	132.00		Diabetes Positive
24		105.00		No Diabetes
24	32.40	119.00		Diabetes Positive
25		94.00		No Diabetes
25	21.10	73.00		No Diabetes
25	28.90	87.00		No Diabetes
25	36.30	138.00		Diabetes Positive
26		114.00		No Diabetes
26	43.20	131.00		Diabetes Positive
27	30.00	141.00		Diabetes Positive
28	23.70	96.00		No Diabetes
28	27.50	146.00		Diabetes Positive
29	35.30	115.00		No Diabetes
29	42.40	141.00		Diabetes Positive
30		115.00		Diabetes Positive
30	32.30	167.00		Diabetes Positive
31	28.00	90.00		No Diabetes
31	29.80	91.00		No Diabetes
31	44.20	145.00		Diabetes Positive
Total		121.69		