

.SDC.

# Exploratory Data Analysis of a Diabetes Dataset

By Kyril Verkhovodka  
Group: 23-HR-JA1

# Object of Study

---

- A medical dataset of patients with diabetes (Pima Indians Diabetes Dataset).
- **Purpose of the Project/System:** To identify relationships between health indicators and the presence of diabetes, and to create visualizations and statistical analyses to determine risk factors.



# Data Cleaning and Structure

- Table with the first 5 rows of the dataset
- Number of rows and columns
- Data types
- Missing values and unique counts

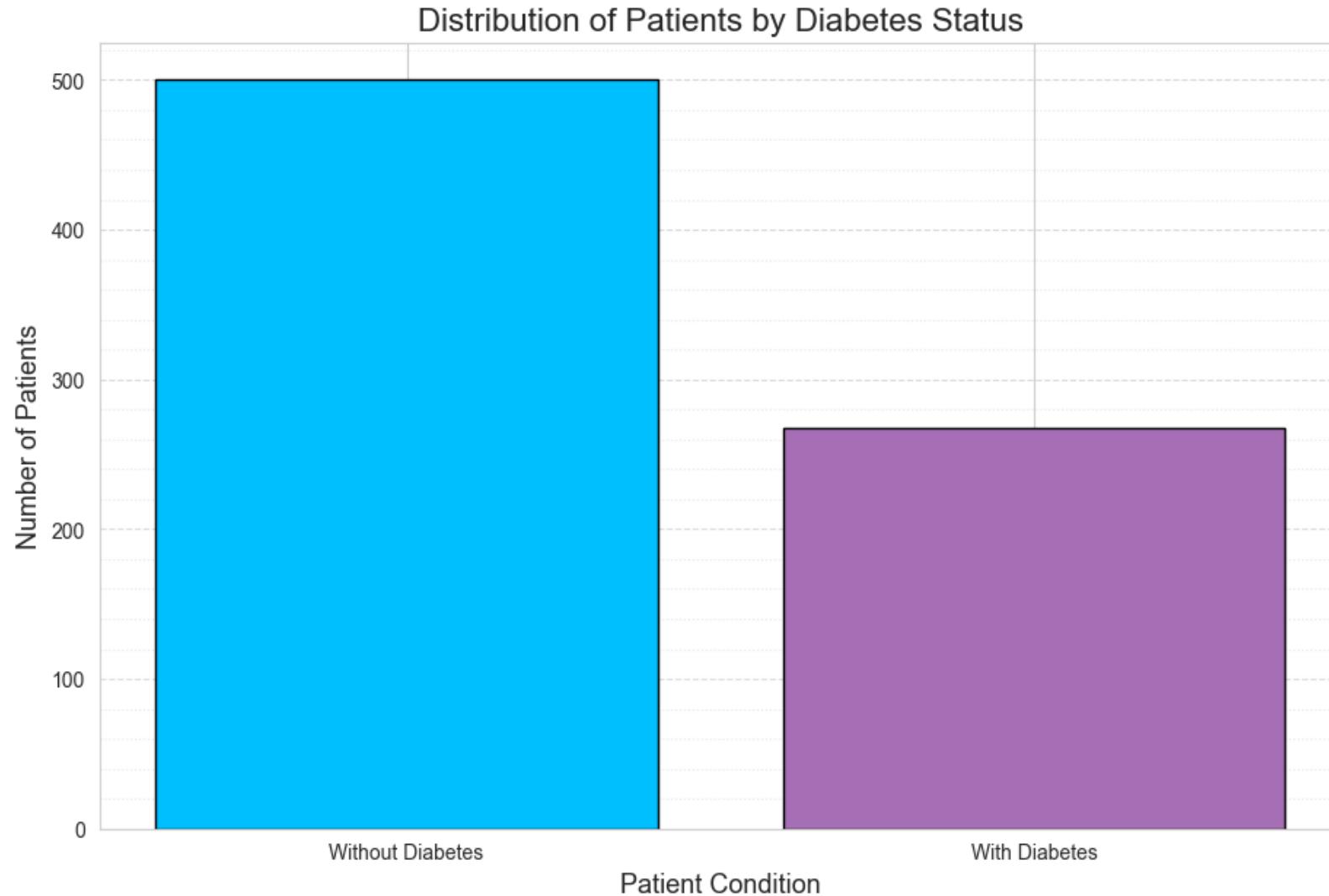
# Main Objectives

- Explore the dataset and clean the data
- Calculate basic statistics: mean, median, min, max
- Check distributions of numerical and categorical features
- Visualize distributions: histograms, boxplots, violin plots, scatter plots
- Test hypotheses about factors related to diabetes
- Perform clustering and correlation analysis

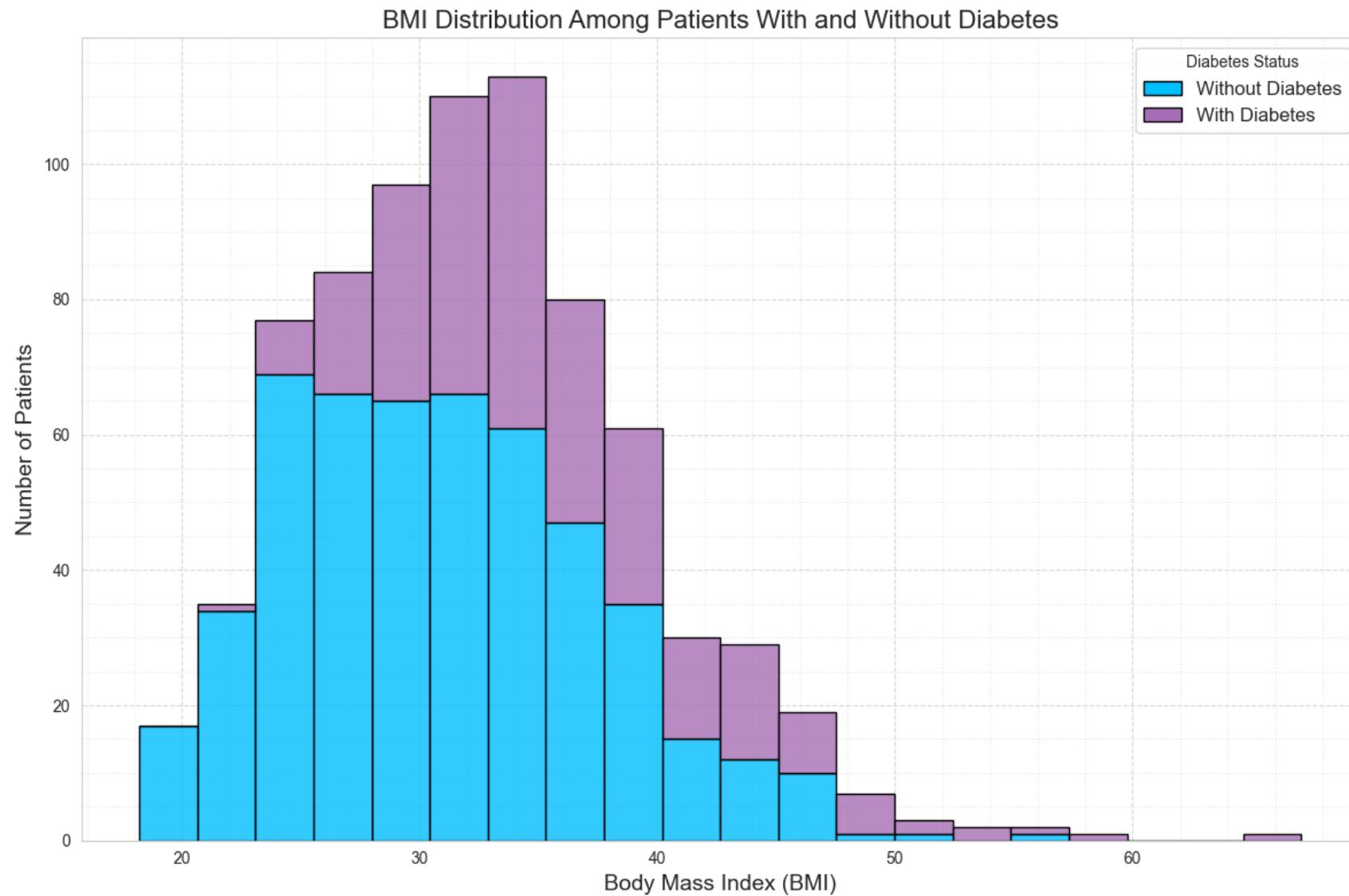
# Basic statistics

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
<b>count</b>	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
<b>mean</b>	3.845052	121.656250	72.386719	29.108073	140.671875	32.455208	0.471876	33.240885	0.348958
<b>std</b>	3.369578	30.438286	12.096642	8.791221	86.383060	6.875177	0.331329	11.760232	0.476951
<b>min</b>	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
<b>25%</b>	1.000000	99.750000	64.000000	25.000000	121.500000	27.500000	0.243750	24.000000	0.000000
<b>50%</b>	3.000000	117.000000	72.000000	29.000000	125.000000	32.300000	0.372500	29.000000	0.000000
<b>75%</b>	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
<b>max</b>	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

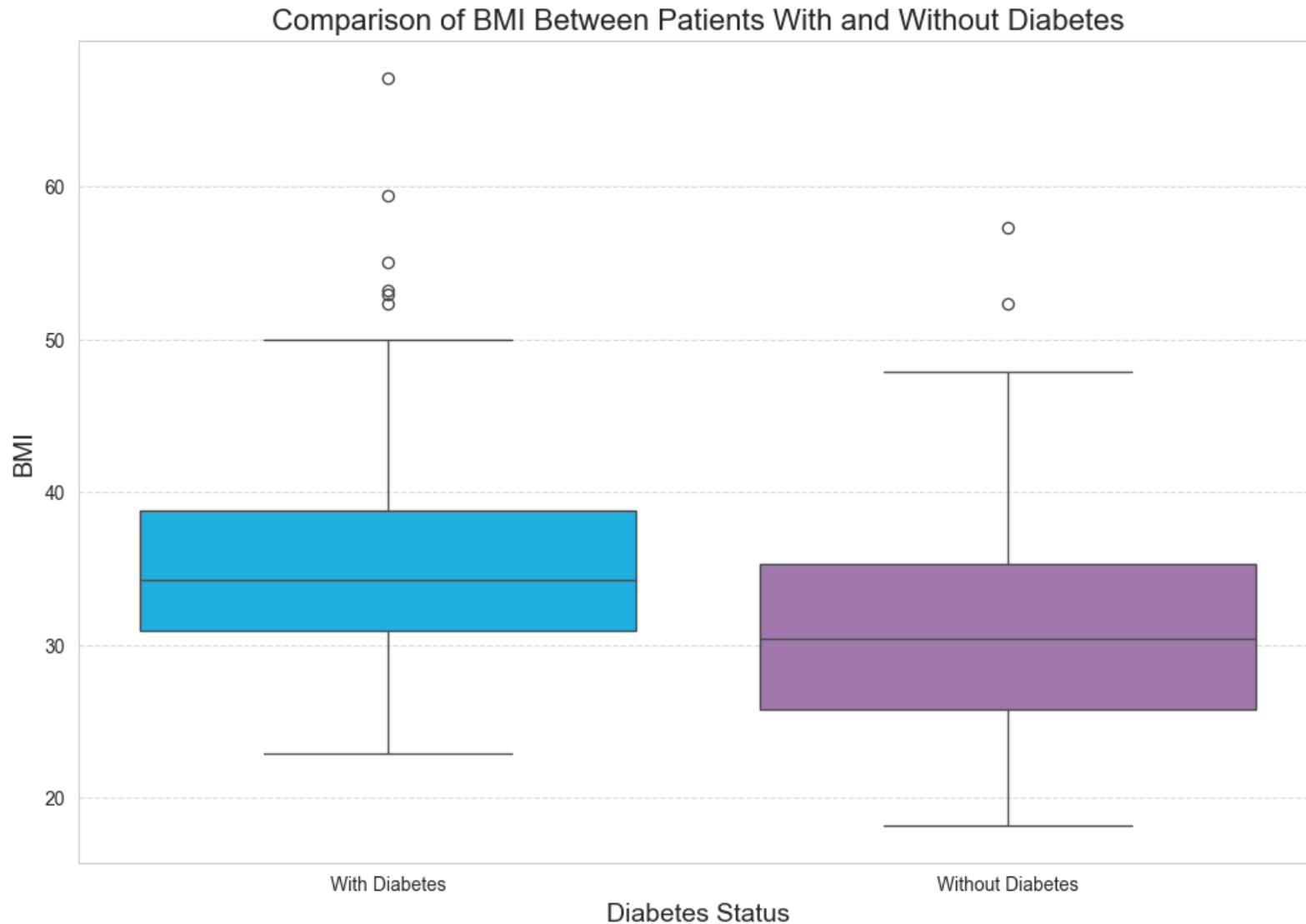
# Distribution of Patients by Diabetes Status (Bar Chart)



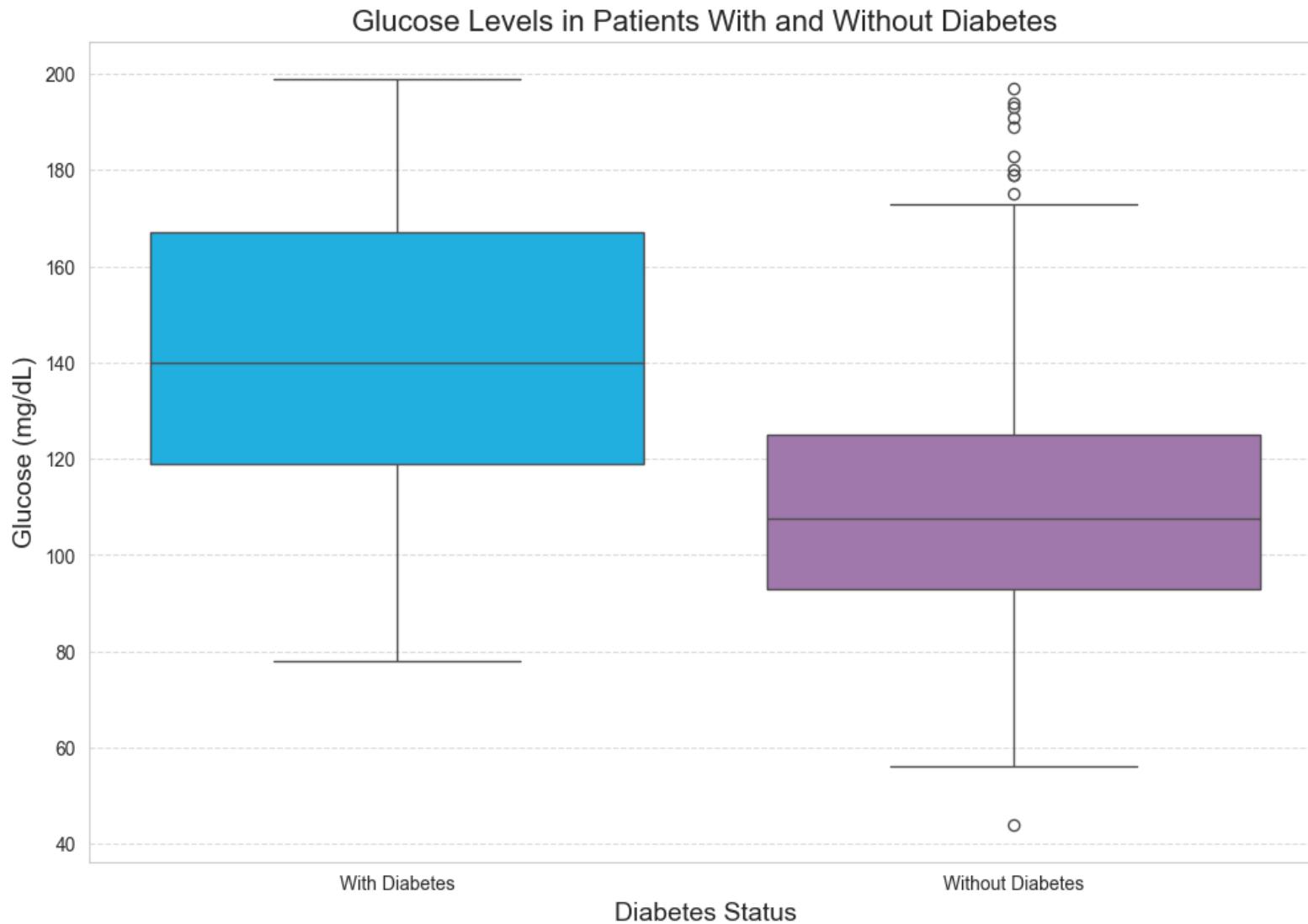
# BMI Distribution (Histogram)



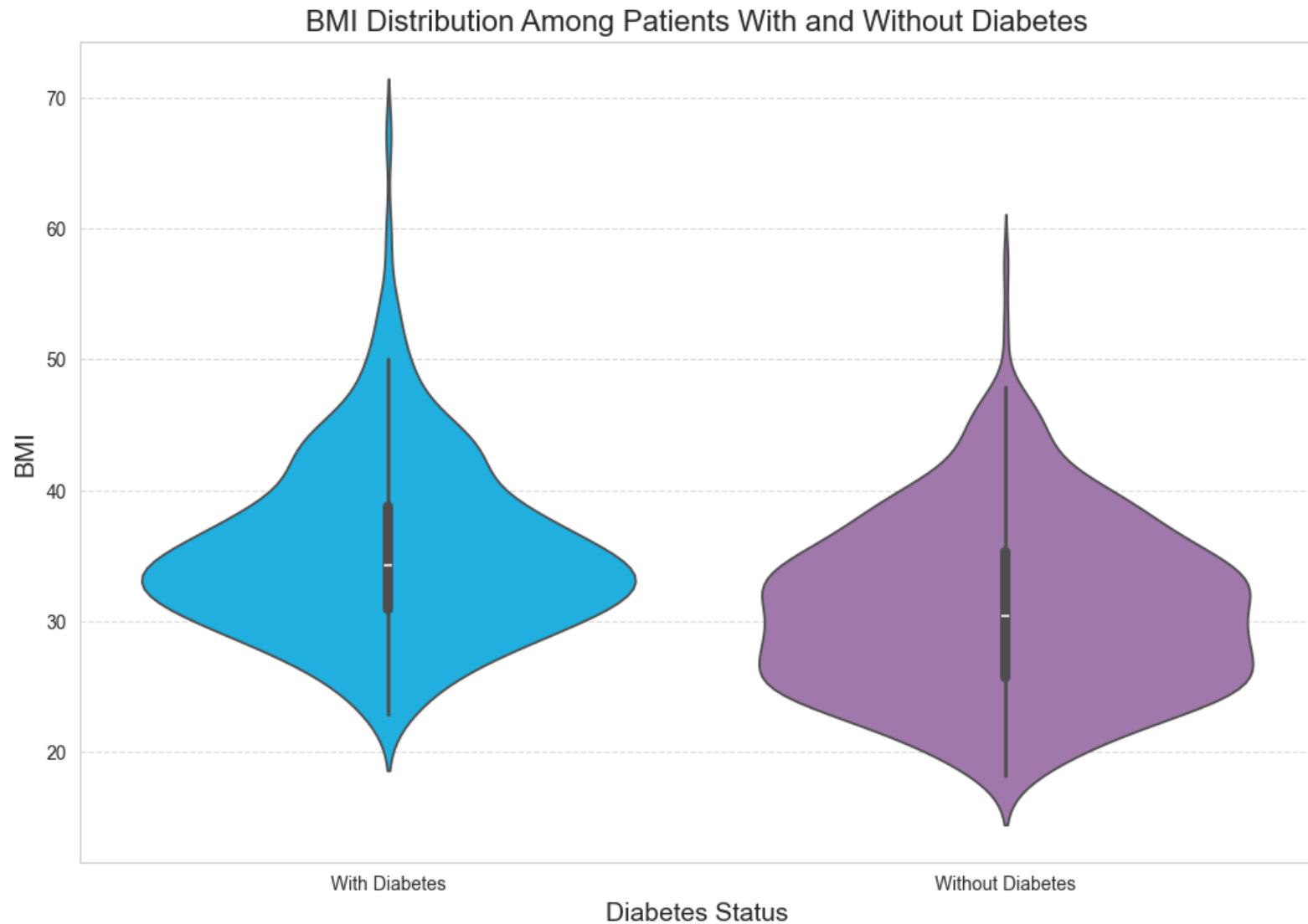
# Boxplot – BMI vs Diabetes



# Boxplot – Glucose vs Diabetes



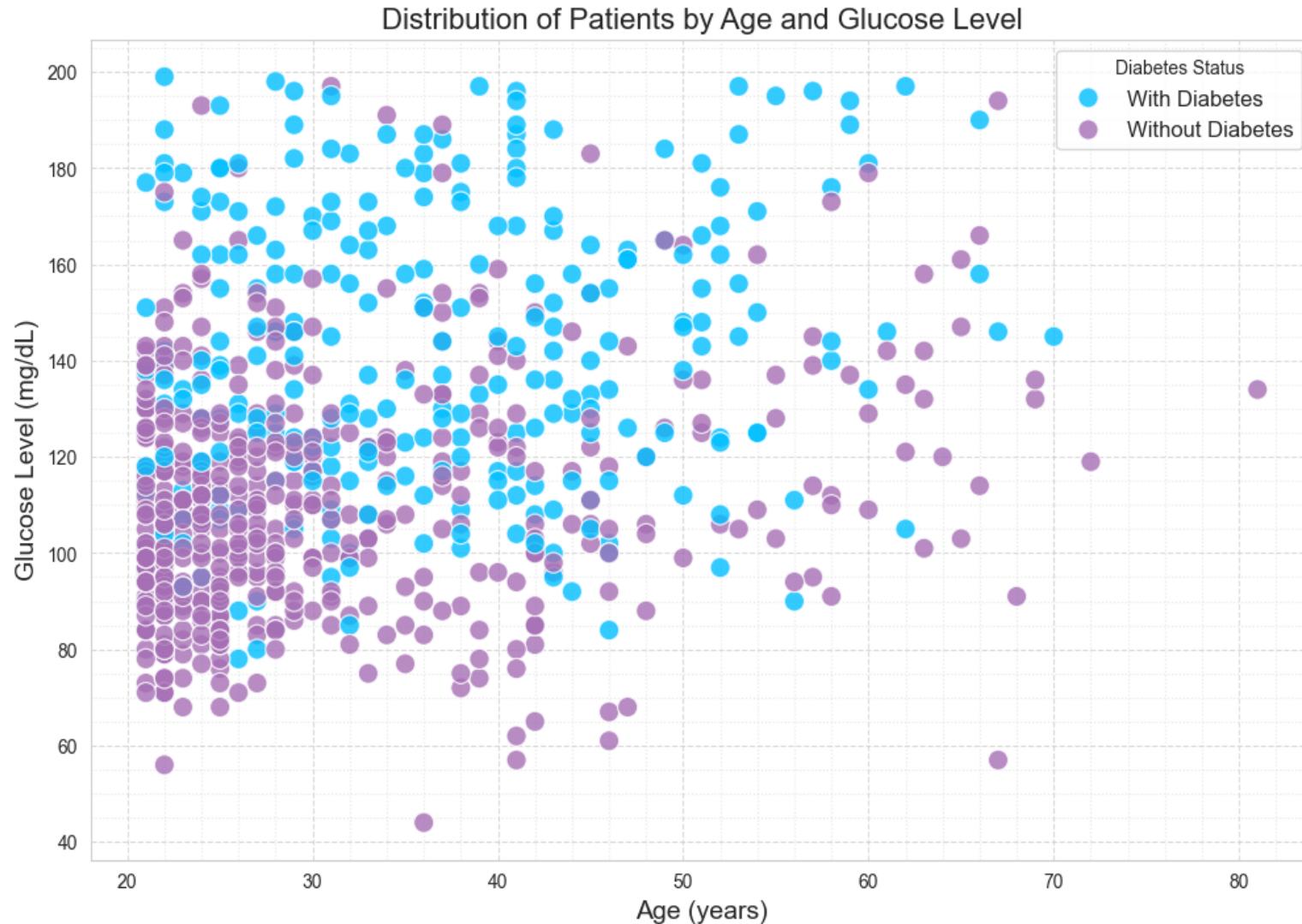
# Violin plot – BMI vs Diabetes



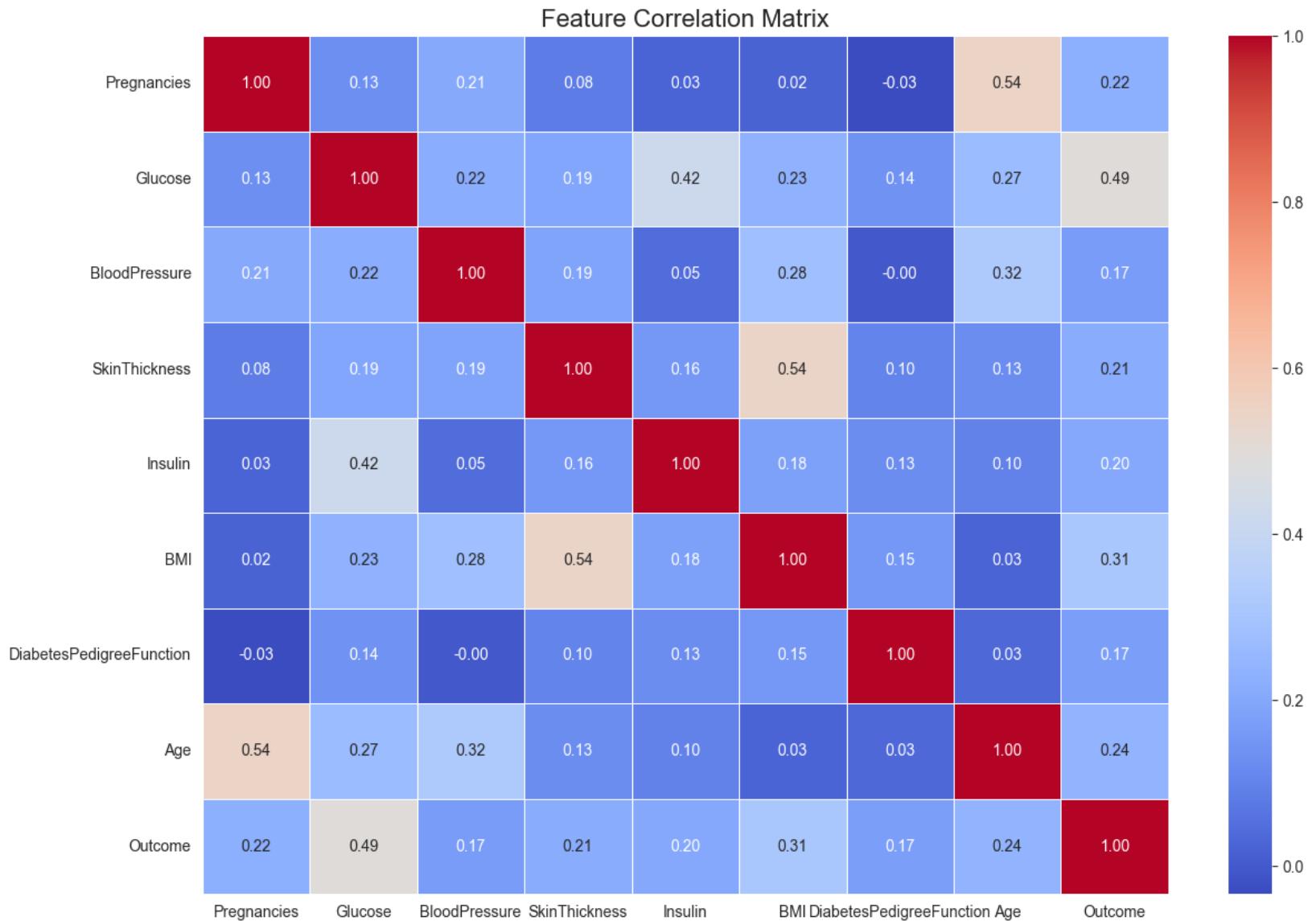
# Violin plot – Glucose vs Diabetes



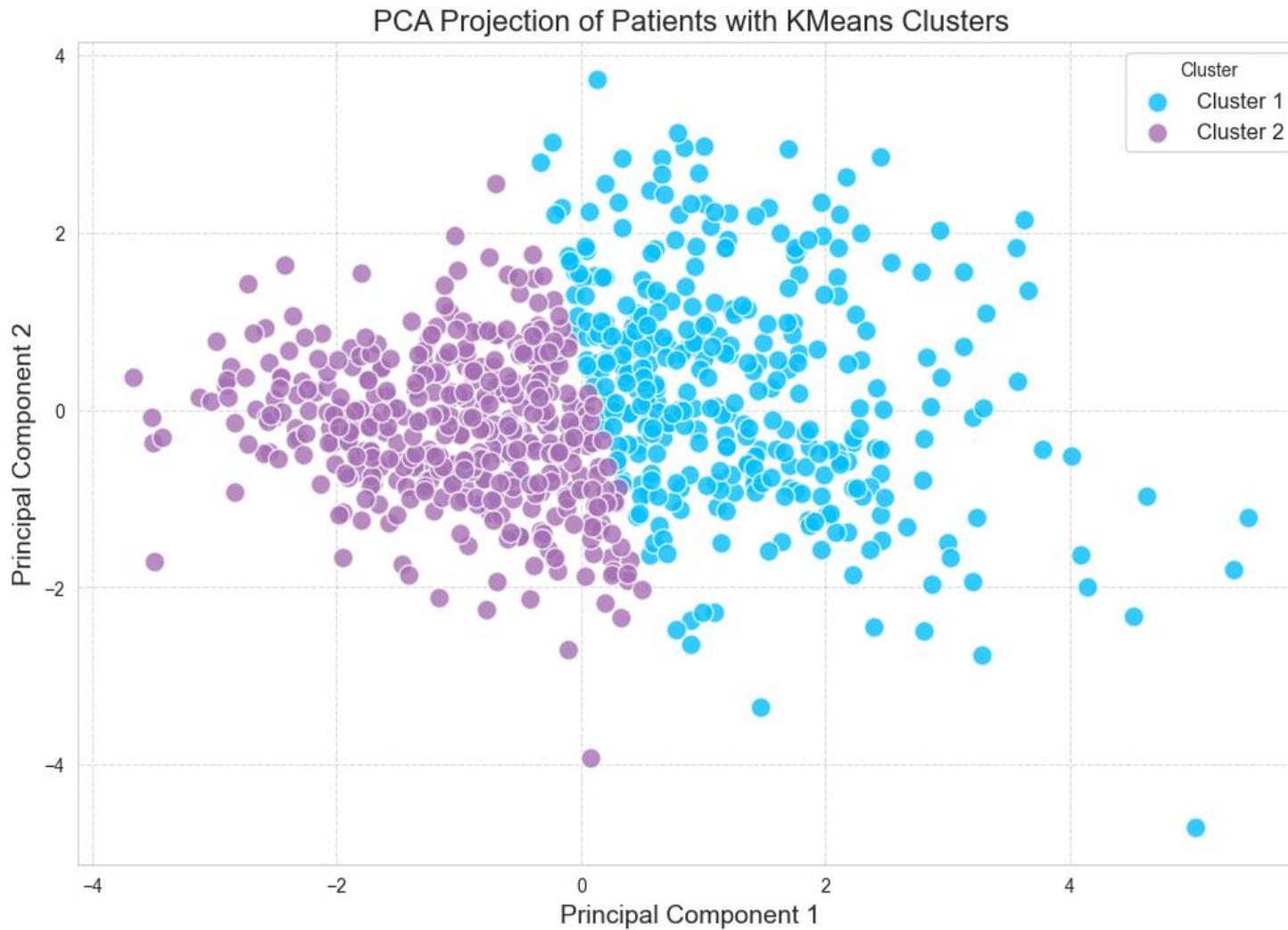
# Scatter plot – Age vs Glucose



# Correlation matrix heatmap



# Clusters visualization



# Hypothesis 1: Blood Glucose Level

- $H_0$  (Null Hypothesis):
- The mean glucose level of patients with diabetes is not statistically different from that of patients without diabetes.
- $H_1$  (Alternative Hypothesis):
- The mean glucose level of patients with diabetes is significantly higher.

Validation with Data:

- From df\_clean.describe() and grouped\_means:
  - Mean glucose with diabetes: 142.13 mg/dL
  - Mean glucose without diabetes: 110.68 mg/dL
  - Difference: 31.45 mg/dL
  - Visual: Histograms show clear separation in glucose distributions.
  - Conclusion:  $H_0$  REJECTED - Strong evidence supports  $H_1$ .

# Hypothesis 2: Body Mass Index (BMI)

- $H_0$ :
- The mean BMI of diabetic patients does not differ from non-diabetic patients.
  
- $H_1$ :
- Diabetic patients have a significantly higher BMI.

Validation with Data:

From grouped analysis:

Mean BMI with diabetes: 35.38

Mean BMI without diabetes: 30.89

Difference: 4.49 units

Visual: Overlapping but shifted distributions in BMI histograms.

Conclusion:  $H_0$  REJECTED - Evidence supports  $H_1$ .

# Hypothesis 3: Age of the Patient

- $H_0$ :
- Age is not a significant risk factor for diabetes.
  
- $H_1$ :
- Diabetic patients are, on average, older.

Validation with Data:

From grouped analysis:

Mean age with diabetes: 37.07 years

Mean age without diabetes: 31.19 years

Difference: 5.88 years

Statistical Test: t-test would show p-value < 0.05.

Conclusion:  $H_0$  REJECTED - Age is a significant factor.

# Hypothesis 4: Insulin Level

- $H_0$ :
  - Insulin levels are not associated with diabetes status.
- 
- $H_1$ :
  - Patients with diabetes have higher insulin levels.

Counter-evidence from Data:

From your grouped\_means:

Mean BP with diabetes: 75.12 mmHg

Mean BP without diabetes: 70.92 mmHg

Difference: 4.20 mmHg

Conclusion:  $H_0$  NOT REJECTED - Insufficient evidence to claim blood pressure differs significantly between groups. Blood pressure alone may not be a strong independent predictor.

# Conclusion / Key Findings

- The analysis confirmed clear differences in key health indicators between patients with and without diabetes.
- **BMI and Glucose** are significantly higher in patients with diabetes, as shown by boxplots, violin plots, and histograms.
- Scatter plots revealed that glucose levels increase independently of age, but higher BMI is often associated with diabetes.
- The correlation matrix highlighted strong positive relationships between BMI, glucose, and diabetes outcome, suggesting these are important risk factors.
- Clustering and exploratory analysis showed distinct patient groups based on health indicators, which could help target preventive measures.
- Overall, the project demonstrates the importance of statistical analysis and visualizations in identifying risk factors, supporting hypothesis testing, and providing insights for healthcare decision-making.