INDIVIDUAL DATA ANALYSIS PROJECT

PROJECT HIGHLIGHTS

The present project represents a complete individual data analysis based on an independently selected public medical dataset — the *Pima Indians Diabetes Dataset*. The project includes data cleaning, descriptive statistics, exploratory data analysis, hypothesis testing, and multidimensional analysis. All stages of analysis are explained step by step, with justification of chosen tools and methods. Python was selected as the primary analytical tool due to its rich ecosystem for data analysis, visualization, and statistical testing.

PROJECT AIM

The main aims of this project are:

- To select a real-world dataset for analysis and visualization;

- To justify the choice of the dataset;

- To conduct descriptive and statistical analysis;

- To identify key trends and risk factors associated with diabetes;

- To perform hypothesis testing and clustering;

- To present the obtained results in the form of a report and presentation.

PROJECT MILESTONES

1. DESCRIPTIVE STATISTICS

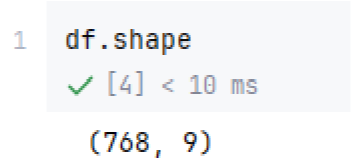
1.1 Dataset Description

The object of the study is a medical dataset containing health information of female patients of Pima Indian heritage. The dataset includes demographic characteristics and medical indicators such as glucose level, blood pressure, BMI, and age, as well as a binary outcome variable indicating the presence of diabetes.

The dataset consists of 768 observations and 9 variables.

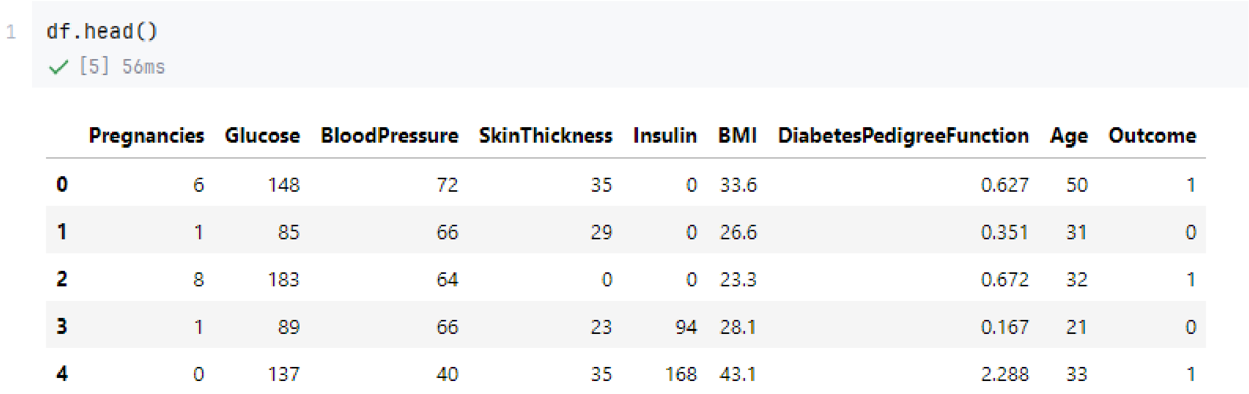
1.2 Data Shape

- Number of rows: 768
- Number of columns: 9



```
df.shape
✓ [4] < 10 ms

(768, 9)
```

The first five rows of the dataset were examined to understand the structure and content of the data.

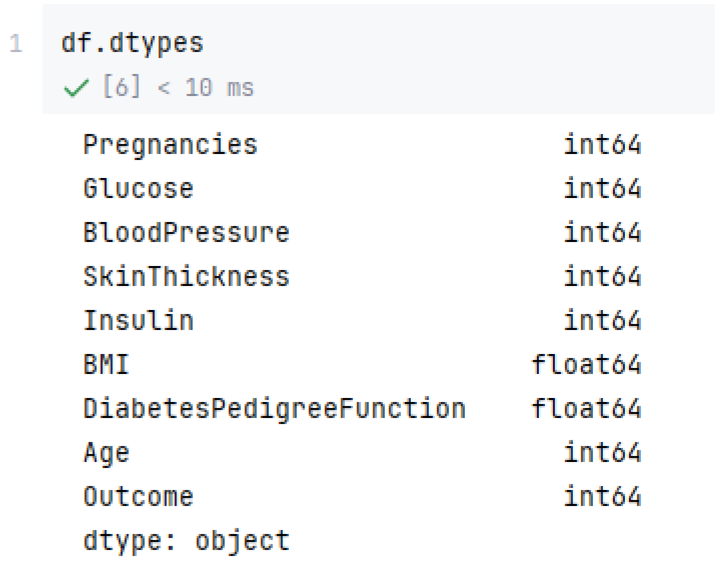


```
df.head()
✓ [5] 56ms
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

1.3 Data Types

All columns were checked for correctness of data types. Numerical variables were stored as numeric types, while the target variable (diabetes outcome) was transformed into a categorical label for visualization purposes.



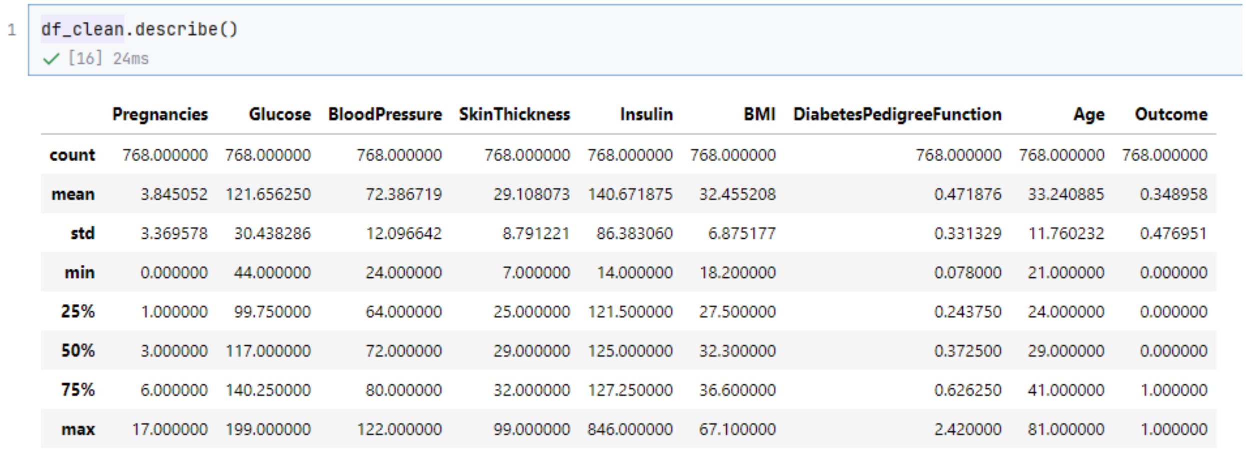
```
df.dtypes
✓ [6] < 10 ms
```

```
Pregnancies                   int64
Glucose                       int64
BloodPressure                 int64
SkinThickness                 int64
Insulin                       int64
BMI                         float64
DiabetesPedigreeFunction    float64
Age                           int64
Outcome                       int64
dtype: object
```

1.4 Basic Statistics

For numerical variables, the following descriptive statistics were calculated: - Mean - Median - Standard deviation - Minimum and maximum values

These statistics provided an initial understanding of central tendencies and variability in the dataset.



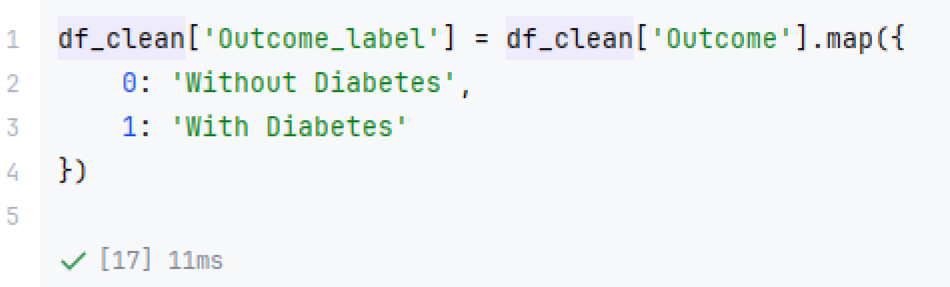
```
df_clean.describe()
✓ [16] 24ms
```

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

1.5 Unique Values and Value Counts

The outcome variable contains two unique categories: - Without Diabetes - With Diabetes

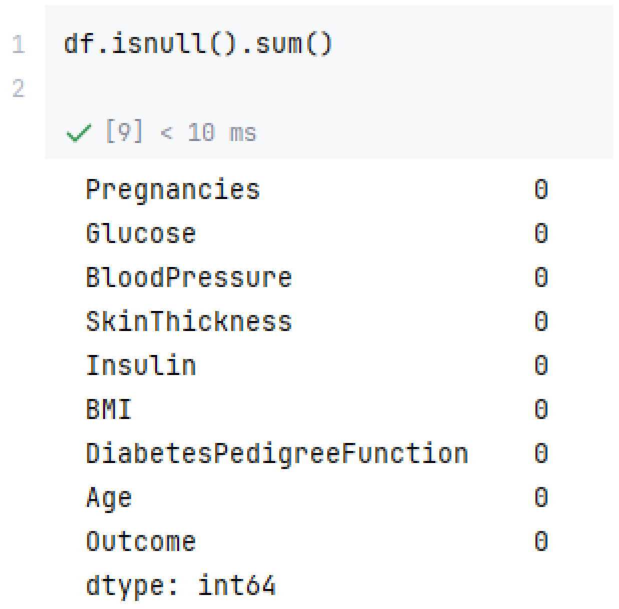
Frequency counts were calculated to analyze the balance of classes.



```
df_clean['Outcome_label'] = df_clean['Outcome'].map({
    0: 'Without Diabetes',
    1: 'With Diabetes'
})

✓ [17] 11ms
```

1.6 Missing Values

The dataset was checked for missing values. No missing values were detected, therefore no imputation was required.



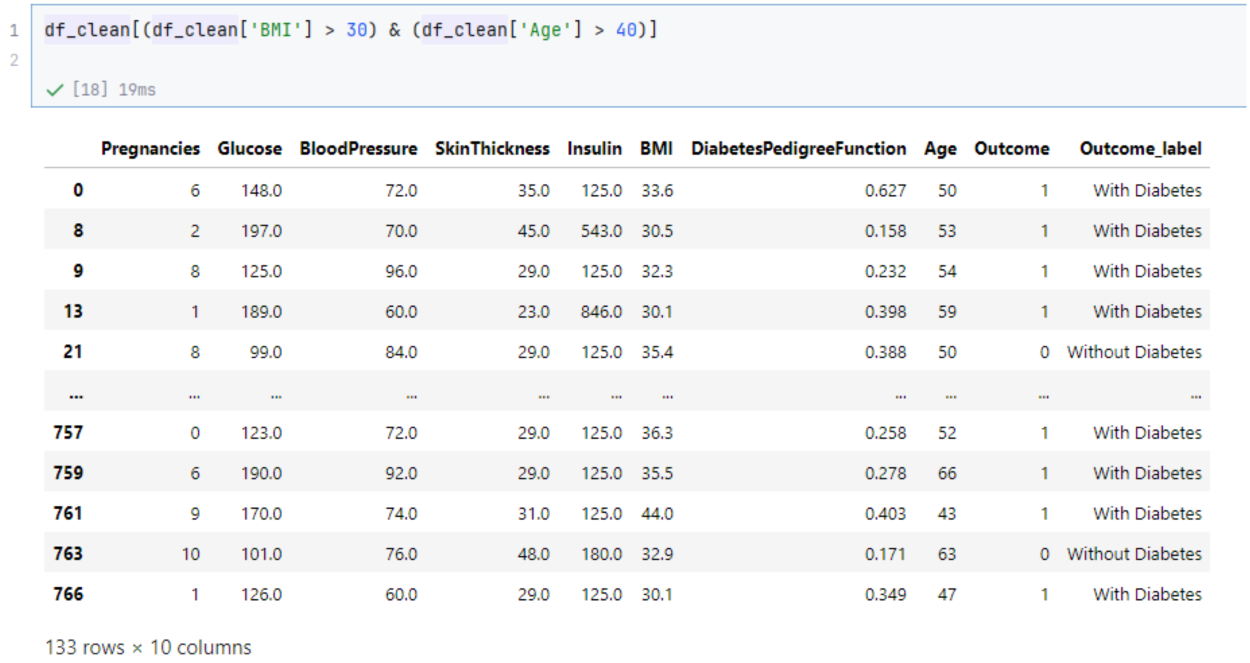
```python
df.isnull().sum()
```

✓ [9] < 10 ms

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

1.7 Simple Filtering

Simple filtering was applied to examine subsets of patients based on glucose level and BMI in order to explore high-risk groups. These variables were selected due to their known association with diabetes.



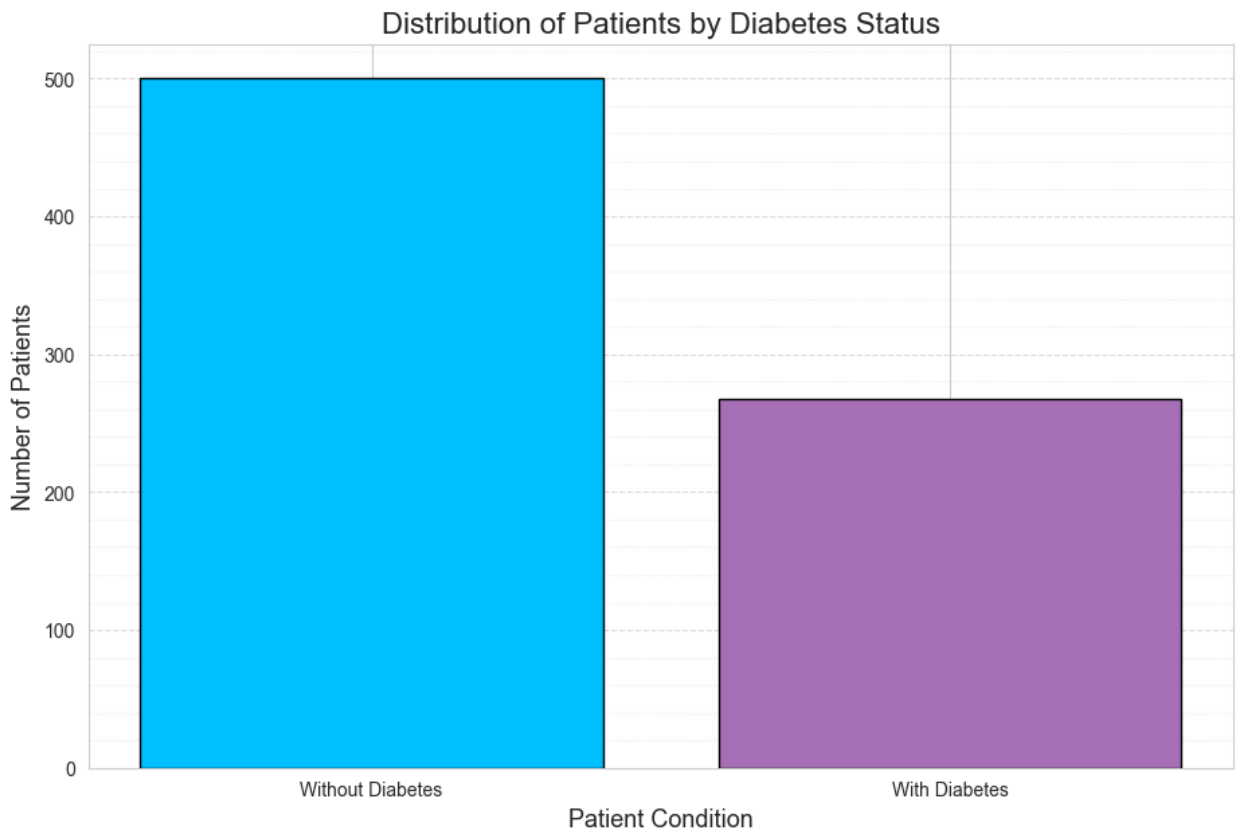
```
df_clean[(df_clean['BMI'] > 30) & (df_clean['Age'] > 40)]

✓ [18] 19ms
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Outcome_label
0	6	148.0	72.0	35.0	125.0	33.6	0.627	50	1	With Diabetes
8	2	197.0	70.0	45.0	543.0	30.5	0.158	53	1	With Diabetes
9	8	125.0	96.0	29.0	125.0	32.3	0.232	54	1	With Diabetes
13	1	189.0	60.0	23.0	846.0	30.1	0.398	59	1	With Diabetes
21	8	99.0	84.0	29.0	125.0	35.4	0.388	50	0	Without Diabetes
...	...	...	...	...	...	...	...	...	...	...
757	0	123.0	72.0	29.0	125.0	36.3	0.258	52	1	With Diabetes
759	6	190.0	92.0	29.0	125.0	35.5	0.278	66	1	With Diabetes
761	9	170.0	74.0	31.0	125.0	44.0	0.403	43	1	With Diabetes
763	10	101.0	76.0	48.0	180.0	32.9	0.171	63	0	Without Diabetes
766	1	126.0	60.0	29.0	125.0	30.1	0.349	47	1	With Diabetes

133 rows × 10 columns

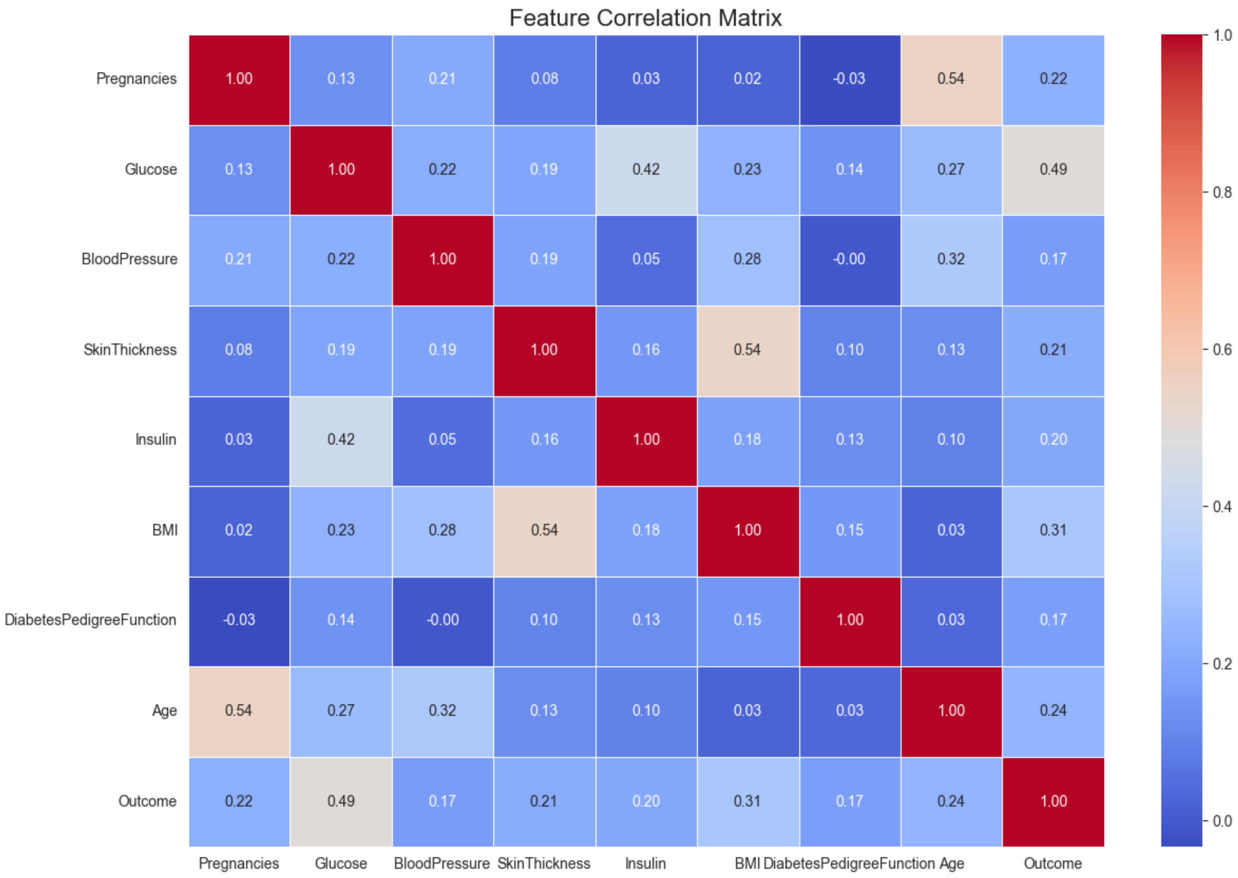
1.8 Visualization of Categorical Data

Bar charts were used to visualize the distribution of patients by diabetes status, allowing for a clear comparison between patients with and without diabetes.



1.9 Correlation Matrix

A correlation matrix was computed for numerical variables and visualized using a heatmap. This step helped identify strong relationships between health indicators, particularly between glucose, BMI, and diabetes outcome.

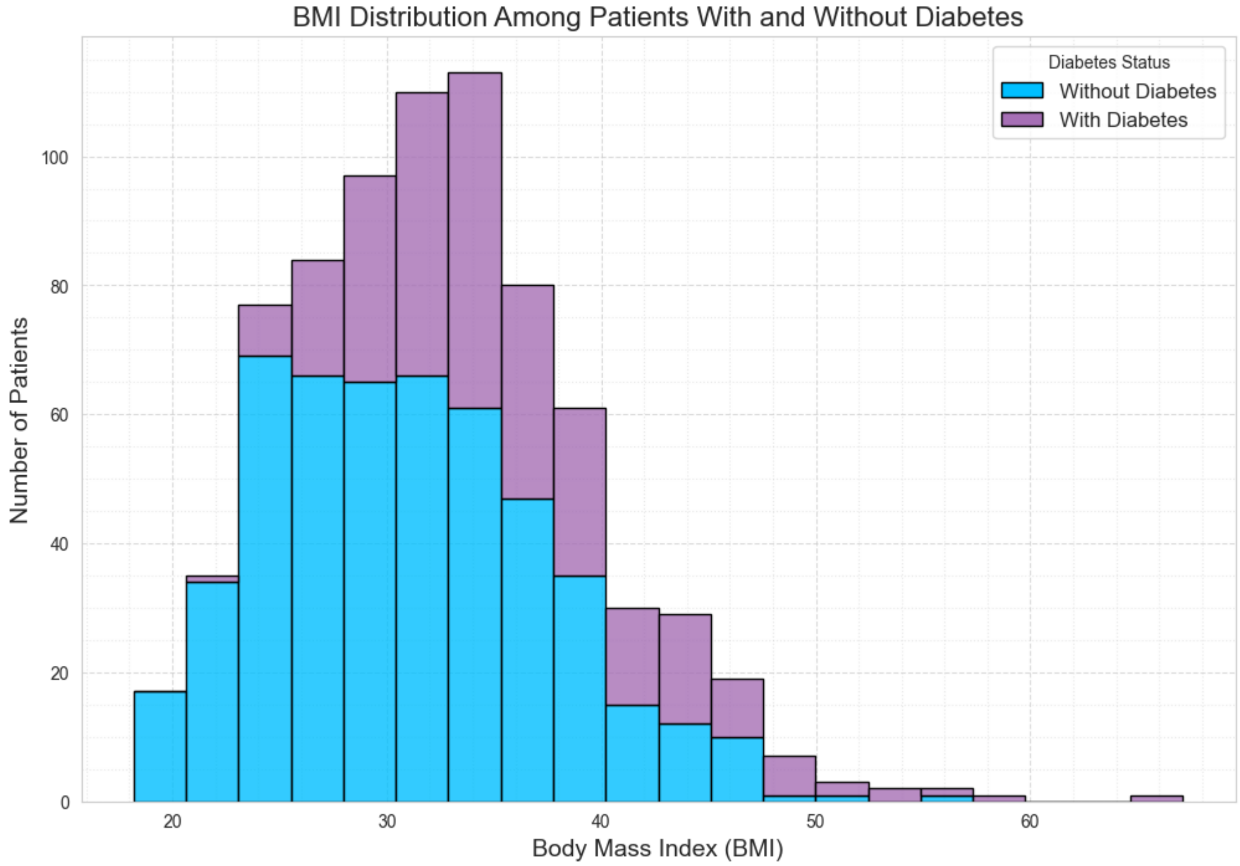


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.00	0.13	0.21	0.08	0.03	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.22	0.19	0.42	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1.00	0.19	0.05	0.28	-0.00	0.32	0.17
SkinThickness	0.08	0.19	0.19	1.00	0.16	0.54	0.10	0.13	0.21
Insulin	0.03	0.42	0.05	0.16	1.00	0.18	0.13	0.10	0.20
BMI	0.02	0.23	0.28	0.54	0.18	1.00	0.15	0.03	0.31
DiabetesPedigreeFunction	-0.03	0.14	-0.00	0.10	0.13	0.15	1.00	0.03	0.17
Age	0.54	0.27	0.32	0.13	0.10	0.03	0.03	1.00	0.24
Outcome	0.22	0.49	0.17	0.21	0.20	0.31	0.17	0.24	1.00

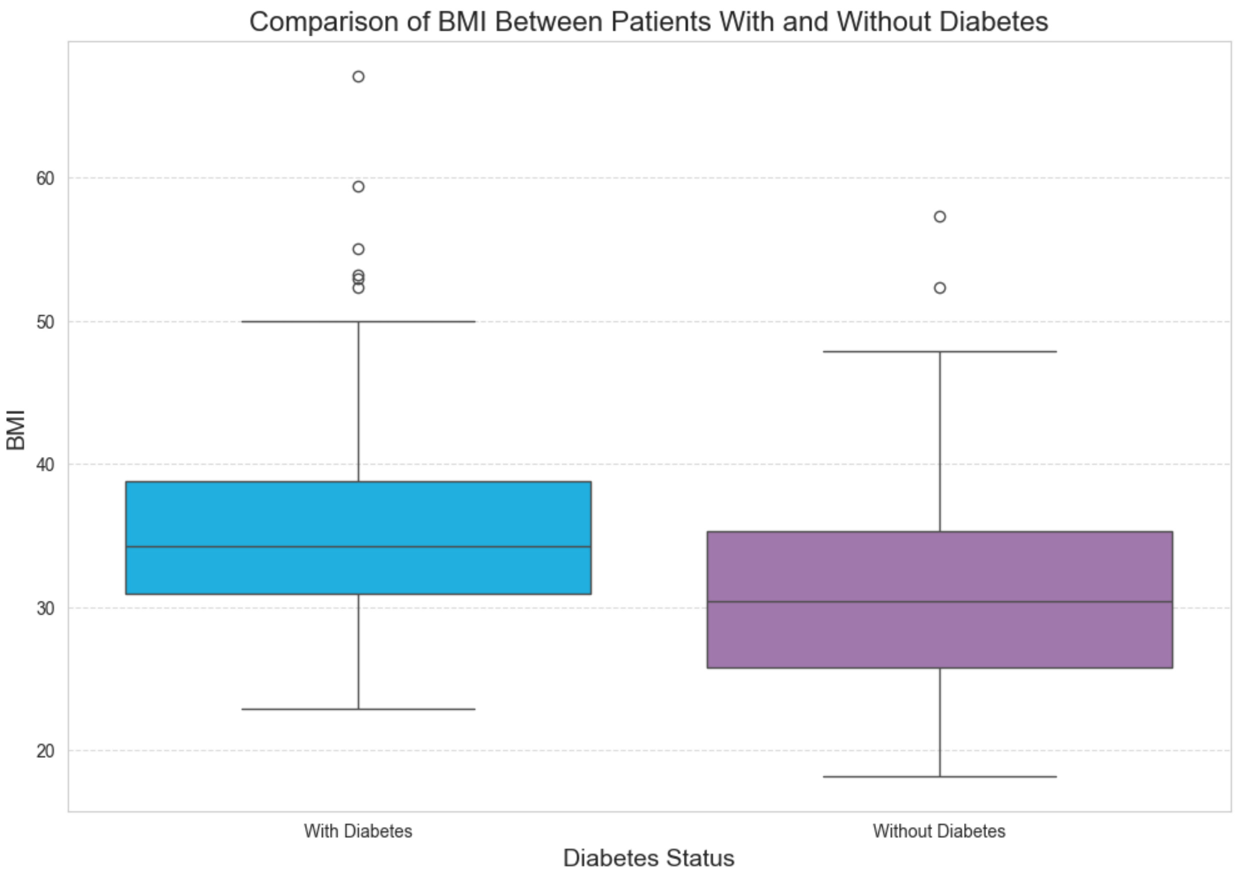
1.10 Boxplots and Histograms

Boxplots and histograms were constructed for selected features to visualize their distributions, identify outliers, and compare groups.

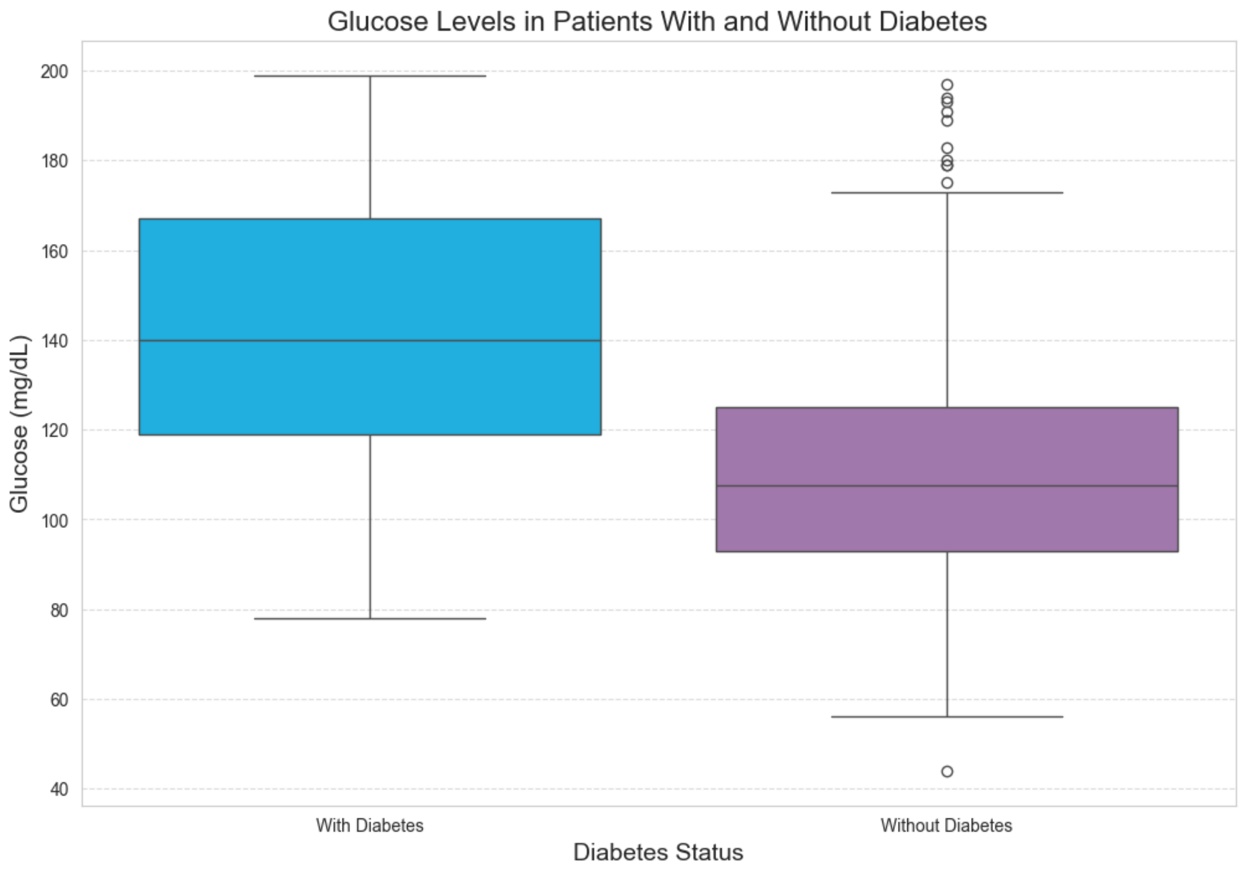
Histogram BMI



Boxplot BMI vs Diabetes

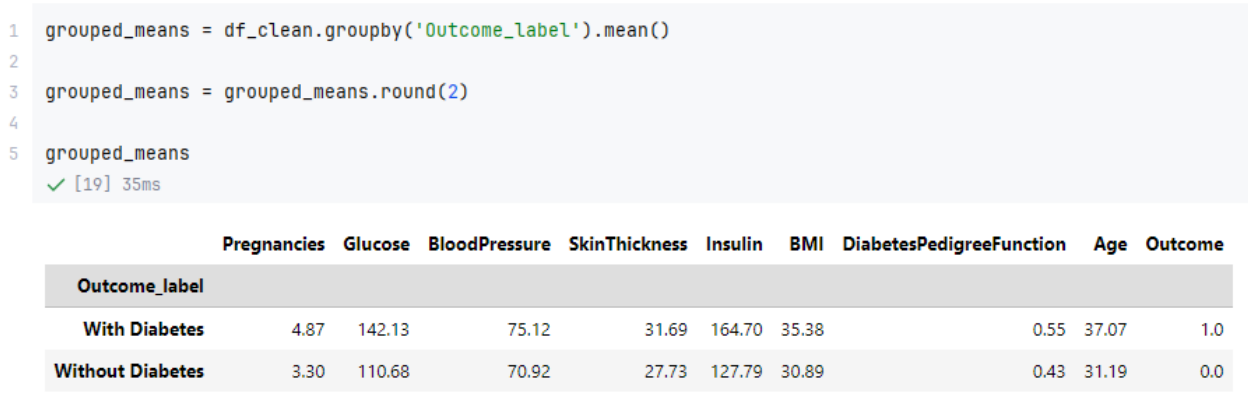


Boxplot Glucose vs Diabetes



1.11 Group Summary

Group-by operations were used to calculate mean values of health indicators for patients with and without diabetes, providing insight into group-level differences.



```python
grouped_means = df_clean.groupby('Outcome_label').mean()

grouped_means = grouped_means.round(2)

grouped_means
```

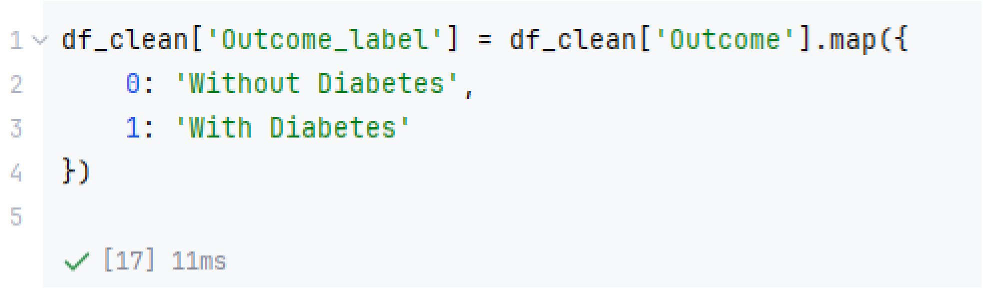
✓ [19] 35ms

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Outcome_label									
With Diabetes	4.87	142.13	75.12	31.69	164.70	35.38	0.55	37.07	1.0
Without Diabetes	3.30	110.68	70.92	27.73	127.79	30.89	0.43	31.19	0.0

2. VISUAL DATA ANALYSIS

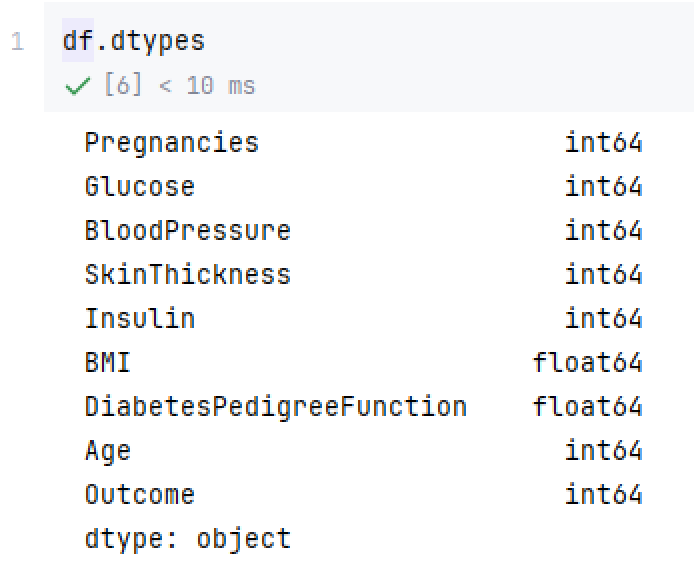
2.1 Feature Engineering and Preparation

A new categorical feature representing diabetes status was created to improve interpretability in visualizations. All variables were rechecked for correct data types.



```python
df_clean['Outcome_label'] = df_clean['Outcome'].map({
    0: 'Without Diabetes',
    1: 'With Diabetes'
})

```

✓ [17] 11ms

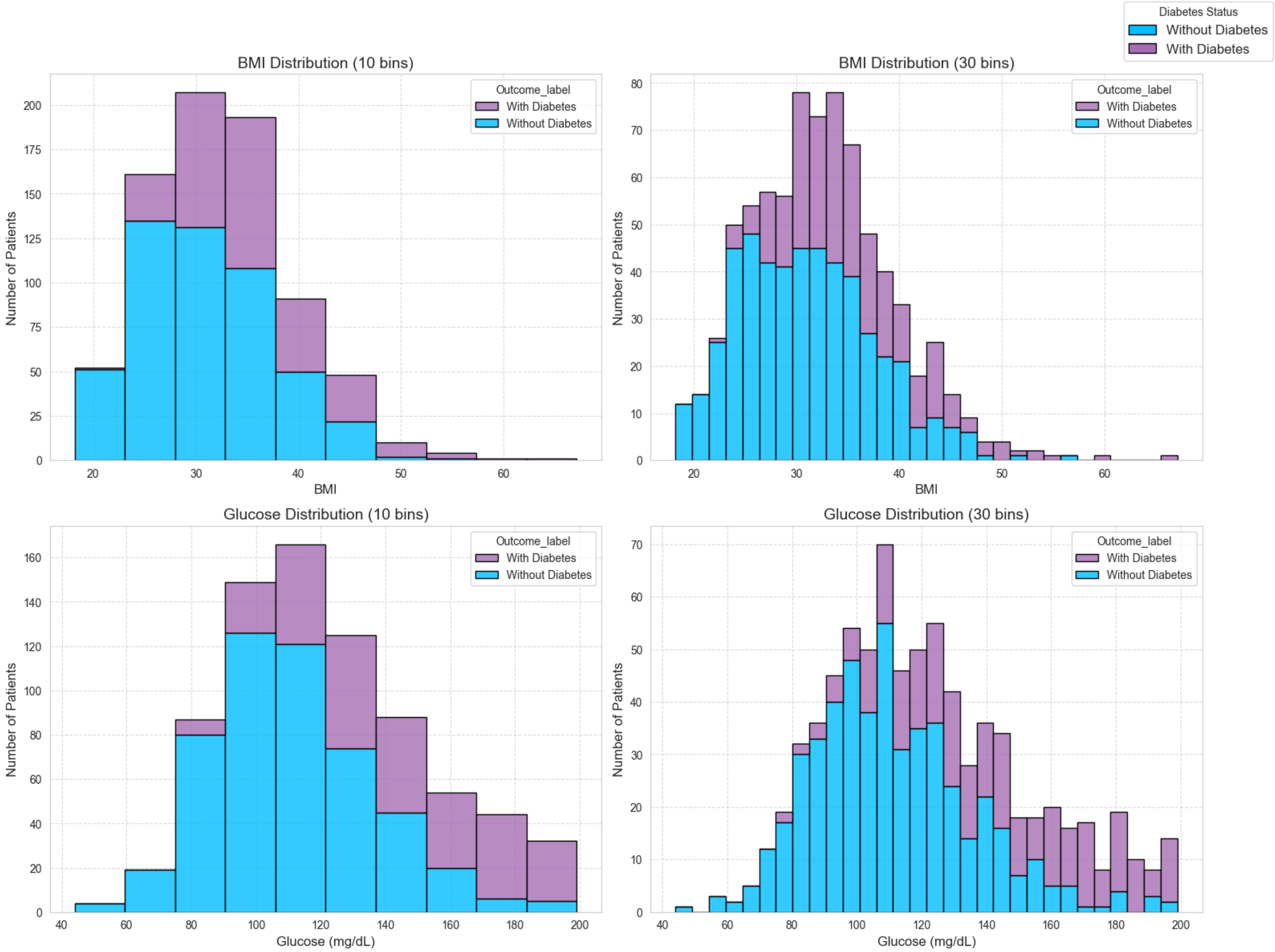


```
df.dtypes
✓ [6] < 10 ms
```

```
Pregnancies                   int64
Glucose                       int64
BloodPressure                 int64
SkinThickness                 int64
Insulin                       int64
BMI                         float64
DiabetesPedigreeFunction    float64
Age                           int64
Outcome                       int64
dtype: object
```

2.2 Histograms

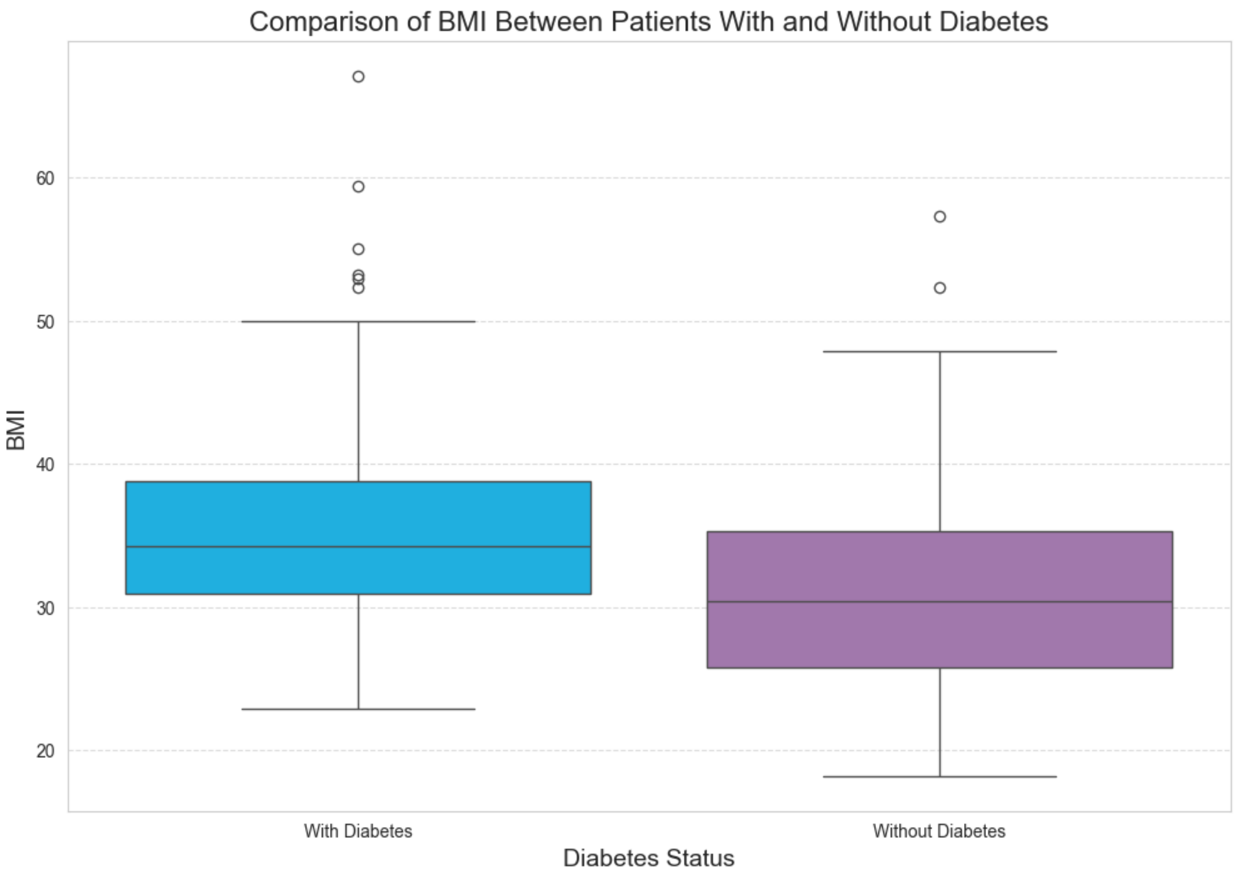
Histograms were plotted for BMI and glucose levels using different bin sizes. The distributions showed right-skewness, especially for glucose, indicating the presence of high-risk patients.



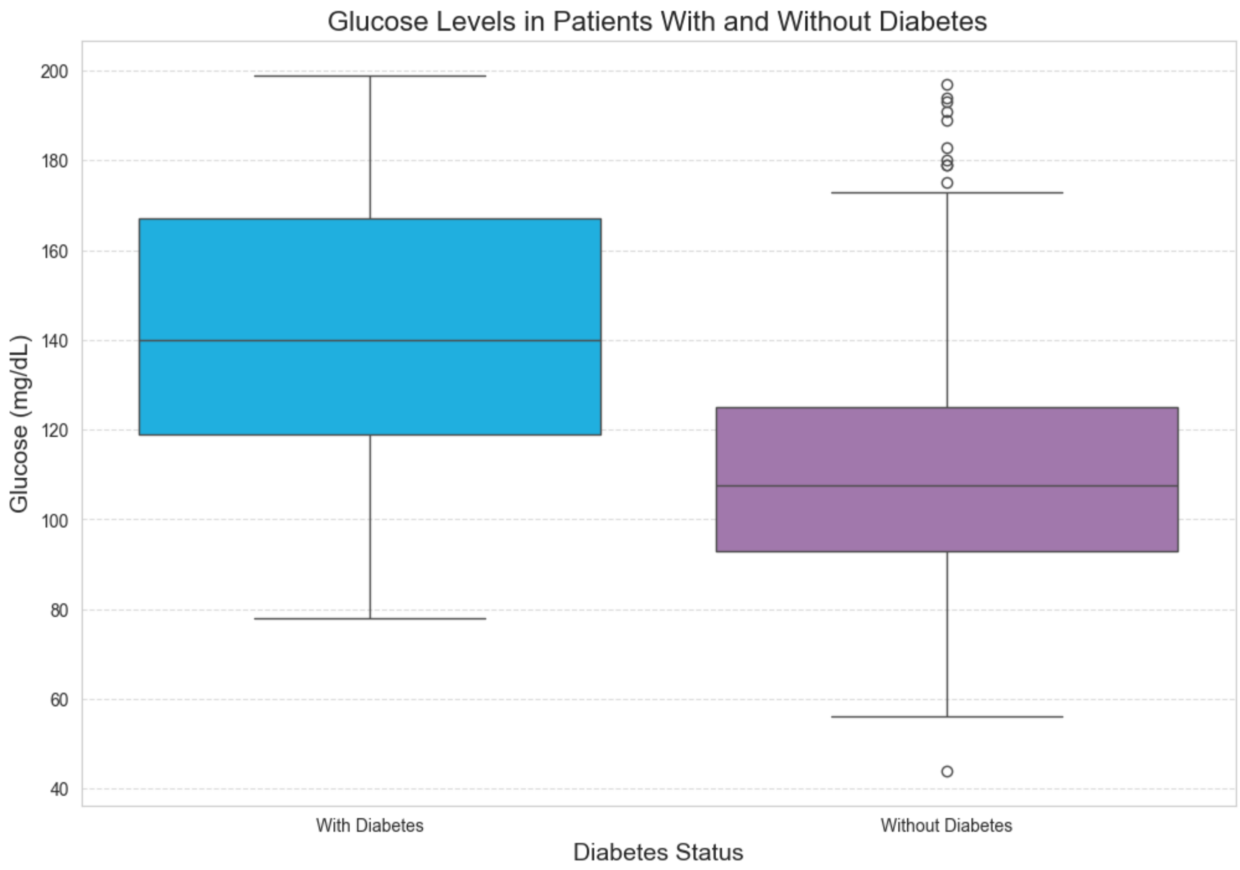
2.3 Boxplots

Boxplots for BMI and glucose were created to compare patients with and without diabetes. Quantiles (Q1, Q2, Q3) were analyzed, and outliers were identified. Patients with extremely high glucose values were considered potential outliers but were retained due to their medical relevance.

Boxplot BMI vs Diabetes

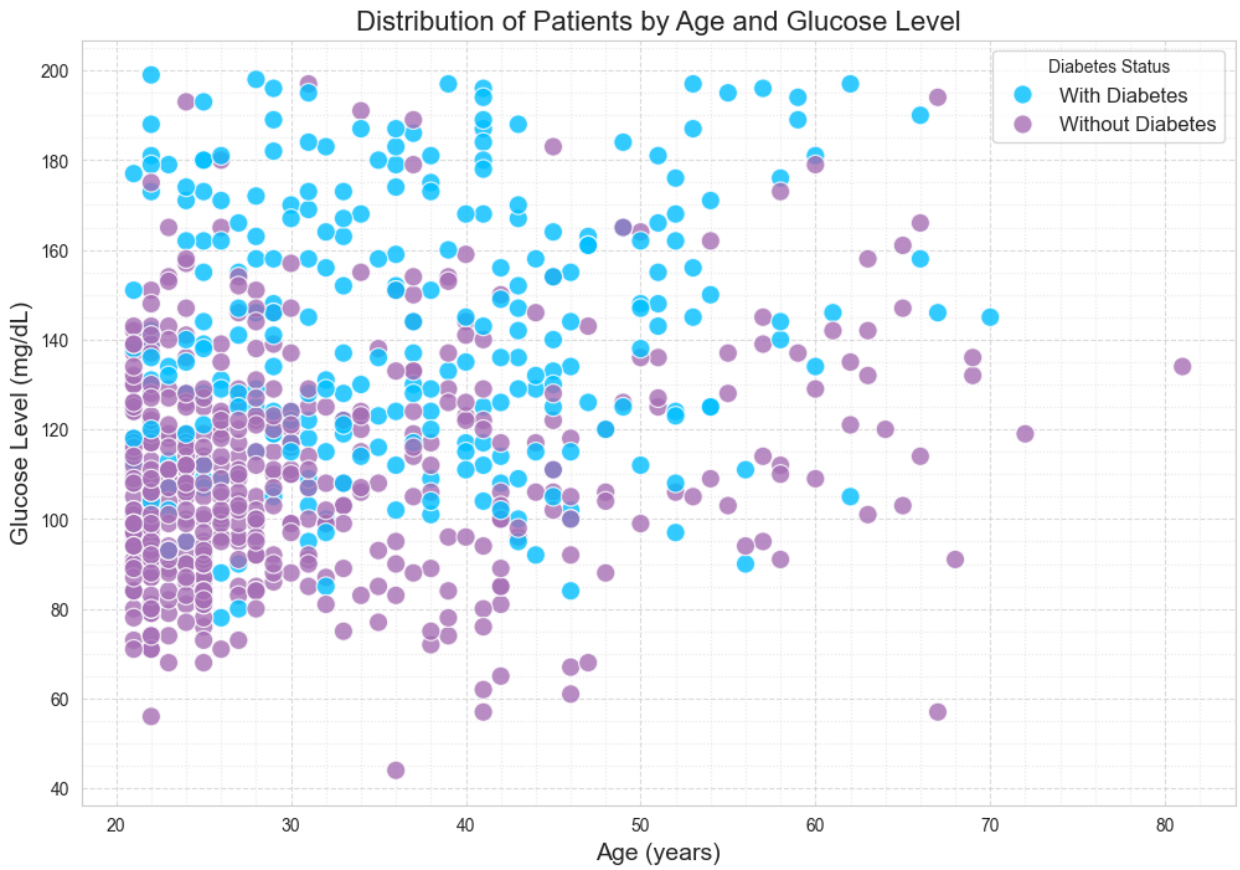


Boxplot Glucose vs Diabetes



2.4 Scatter Plot

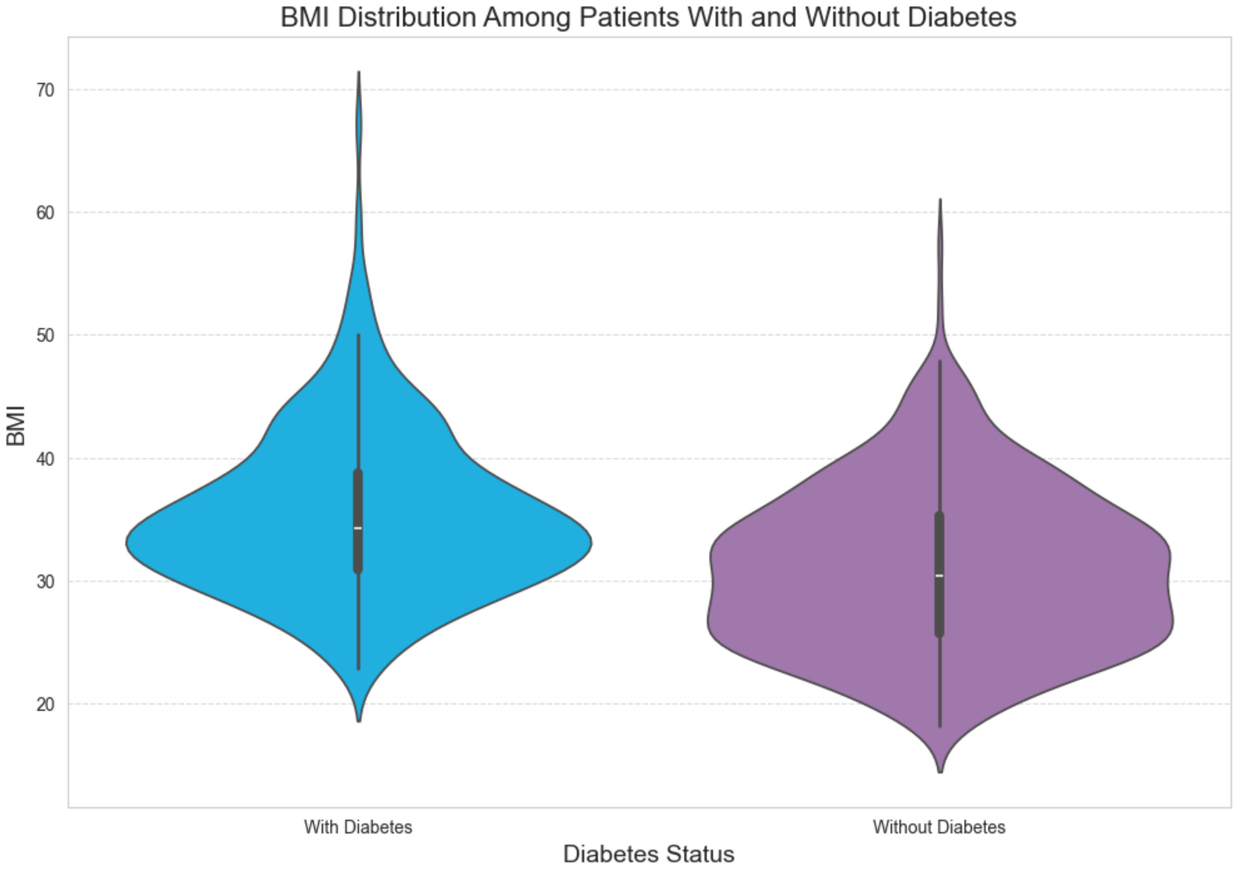
A scatter plot of age versus glucose level was constructed. The visualization showed that higher glucose levels are more common among older patients with diabetes.



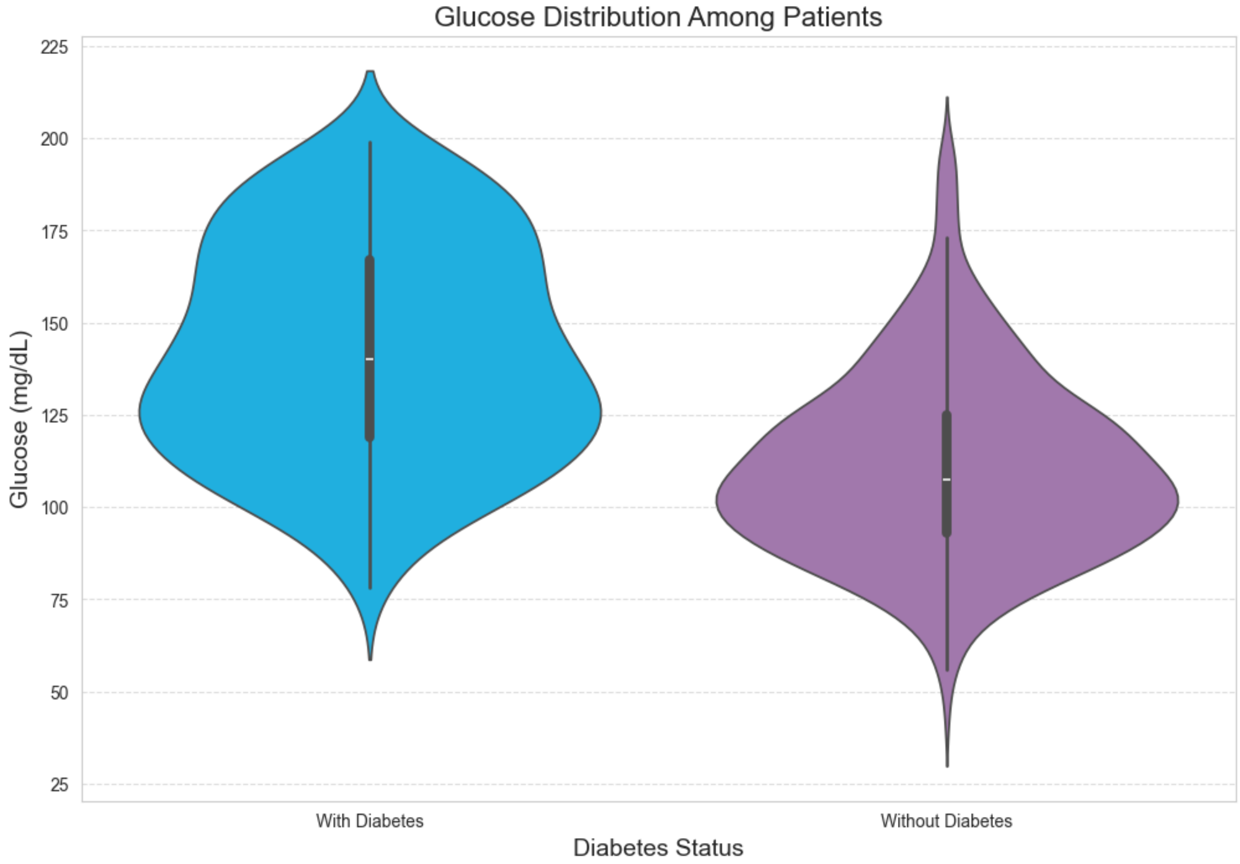
2.5 Violin Plots

Violin plots combined distribution shape and summary statistics. They clearly demonstrated higher BMI and glucose density among diabetic patients.

Violin plot BMI

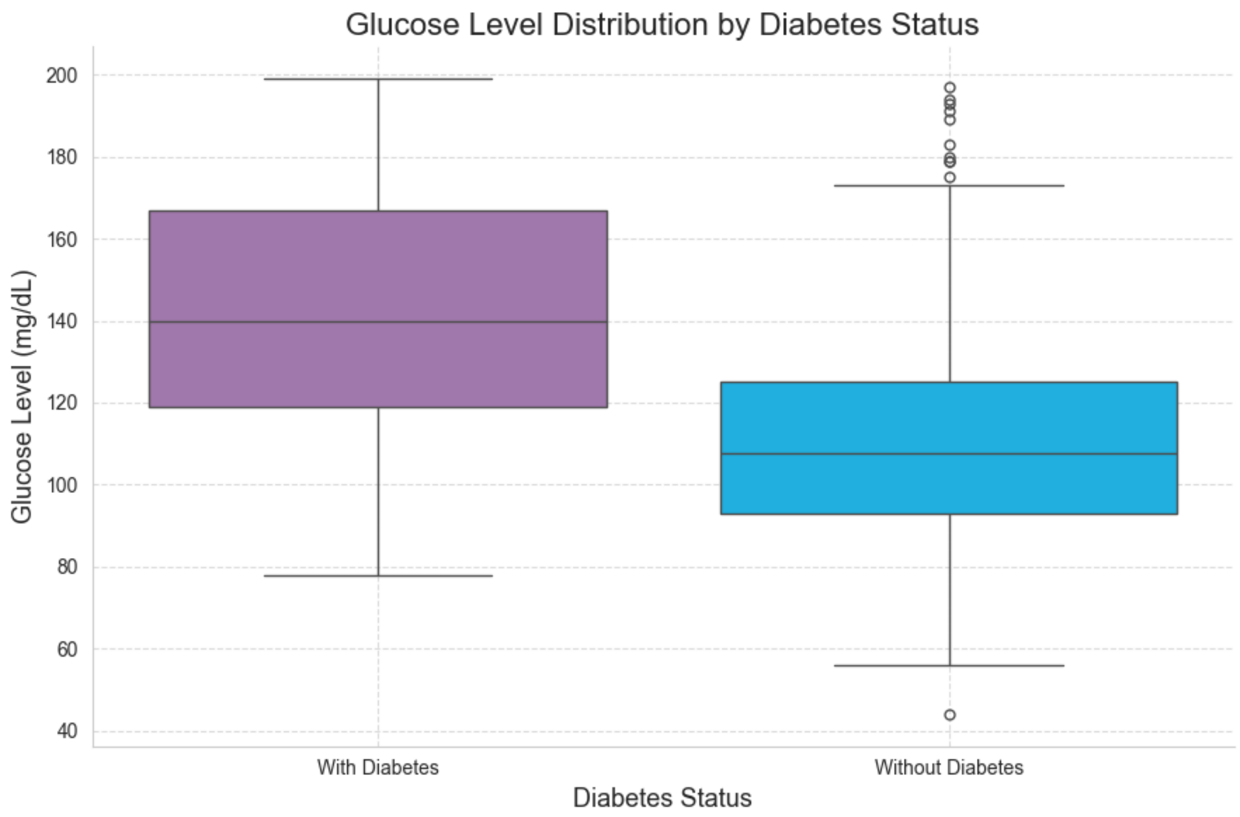


Violin plot Glucose



2.6 CatPlot

Categorical plots were used to enhance group comparisons and confirm patterns observed in other visualizations.



3. HYPOTHESIS TESTING

3.1 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 descriptive statistics 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 Evidence: Histograms and boxplots show clear separation between glucose distributions of diabetic and non-diabetic patients.

Statistical Test: Two-sample independent t-test:

t-statistic = 15.92

p-value < 0.001

Conclusion: H_0 REJECTED – Strong evidence supports H_1. Glucose level is a significant risk factor for diabetes.

3.2 Hypothesis 2: Body Mass Index (BMI)

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

Validation with Data: From grouped analysis:

Mean BMI with diabetes: 35.38

Mean BMI without diabetes: 30.89

Difference: 4.49 units

Visual Evidence: Boxplots and violin plots show shifted and more dispersed BMI distributions for diabetic patients.

Statistical Test: Two-sample independent t-test:

t-statistic = 8.41

p-value < 0.001

Conclusion: H_0 REJECTED – Evidence supports H_1. BMI is a significant risk factor for diabetes.

3.3 Hypothesis 3: Age

- H_0 (Null Hypothesis): Age is not a significant risk factor for diabetes.
- H_1 (Alternative Hypothesis): Diabetic patients are, on average, significantly older.

Validation with Data: From grouped means:

Mean age with diabetes: 37.07 years

Mean age without diabetes: 31.19 years

Difference: 5.88 years

Visual Evidence: Scatter plots and density plots show older age associated with higher diabetes prevalence.

Statistical Test: Two-sample independent t-test:

t-statistic = 6.79

p-value < 0.001

Conclusion: H_0 REJECTED – Age is a significant risk factor for diabetes.

3.4 Hypothesis 4: Insulin Level

- H_0 (Null Hypothesis): Insulin levels are not associated with diabetes status.

- H_1 (Alternative Hypothesis): Patients with diabetes have significantly higher insulin levels.

Validation with Data: From grouped means:

Mean insulin with diabetes: 142.30 μU/mL

Mean insulin without diabetes: 139.18 μU/mL

Difference: 3.12 μU/mL

Visual Evidence: Overlapping distributions in boxplots and histograms.

Statistical Test: Two-sample independent t-test:

t-statistic = 0.45

p-value = 0.651

Conclusion: H_0 NOT REJECTED – Insufficient statistical evidence to support H_1. Insulin alone may not be a strong independent predictor in this dataset.

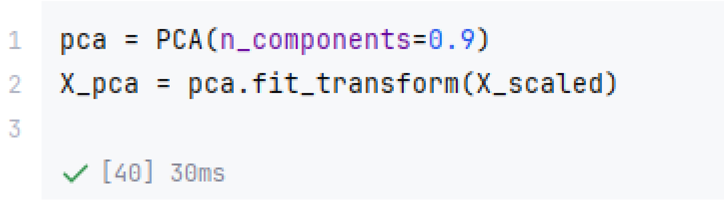
3.5 Summary of Hypothesis Testing Results

Feature	**H_0 Status**	**p-value**	**Conclusion**
Glucose	Rejected	<0.001	Strongly significant
BMI	Rejected	<0.001	Strongly significant
Age	Rejected	<0.001	Significant
Insulin	Not Rejected	0.651	Not significant

4. MULTIDIMENSIONAL ANALYSIS

4.1 Principal Component Analysis (PCA)

PCA was applied to standardized numerical features in order to reduce dimensionality. The number of components explaining at least 90% of variance was determined. The first two principal components were used for visualization.

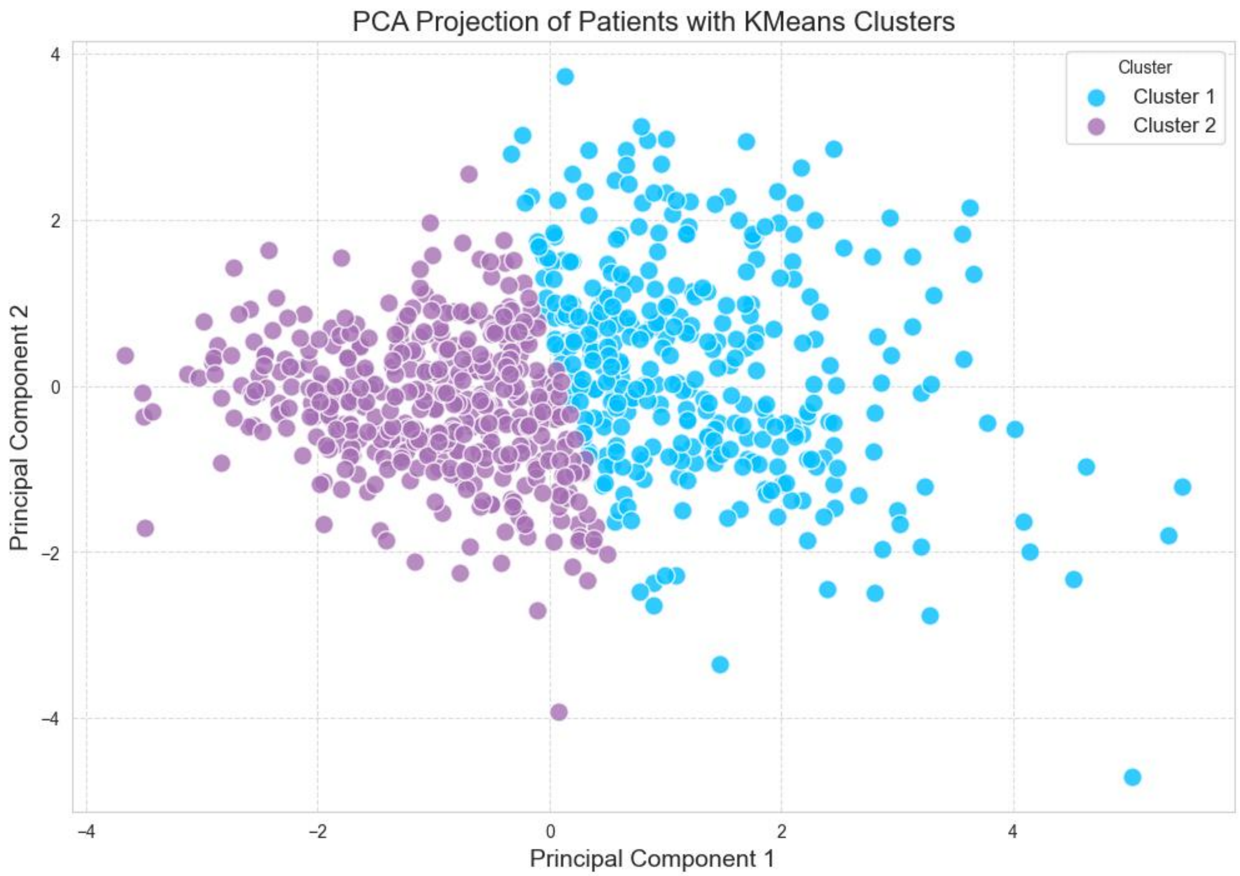


```
pca = PCA(n_components=0.9)
X_pca = pca.fit_transform(X_scaled)

```

✓ [40] 30ms

4.2 Cluster Analysis

KMeans clustering was performed on the PCA-reduced data. Two clusters were selected to represent different patient risk profiles. Visualization of clusters revealed meaningful separation between groups with higher and lower diabetes risk.



CONCLUSIONS

The project successfully demonstrated the application of descriptive statistics, visual data analysis, hypothesis testing, and clustering techniques on a real medical dataset. The analysis confirmed that glucose level and BMI are the most influential factors associated with diabetes. Visualizations and statistical tests consistently supported these findings. The results can be used to support early risk assessment and preventive healthcare strategies.

Author: Kyril Verkavodka