EXPERIMENT ASSESSMENT

ACADEMIC YEAR 2025-26

Course: AIH Lab

Course code: CSDOL 7012

Year: BE SEM: VII

Experiment No. 2
Title: To perform EDA on healthcare data using Pandas and Matplotlib
Name: Ayush Gupta
Roll Number: 11
Date of Performance:21/7/25
Date of Submission:28/7/25

Evaluation

Performance Indicator	Max. Marks	Marks Obtained
Performance	5	
Understanding	5	
Journal work and timely submission.	10	
Total	**20**	

Performance Indicator	Exceed Expectations (EE)	Meet Expectations (ME)	Below Expectations (BE)
Performance	5	3	2
Understanding	5	3	2
Journal work and timely submission.	10	8	4

Checked by

Name of Faculty : Mrs.Kranti Gule

Signature :

Date :

Aim: To perform EDA on healthcare data using Pandas n Matplotlib

Objective: The objective of this analysis is to gain a comprehensive understanding of the healthcare dataset by employing Pandas and Matplotlib to visualize and summarize key aspects of the data. Through descriptive statistics, data visualization, and pattern identification, this EDA aims to uncover trends, anomalies, and correlations within the dataset, providing valuable insights for informed decision-making and potential areas of further investigation in the healthcare domain

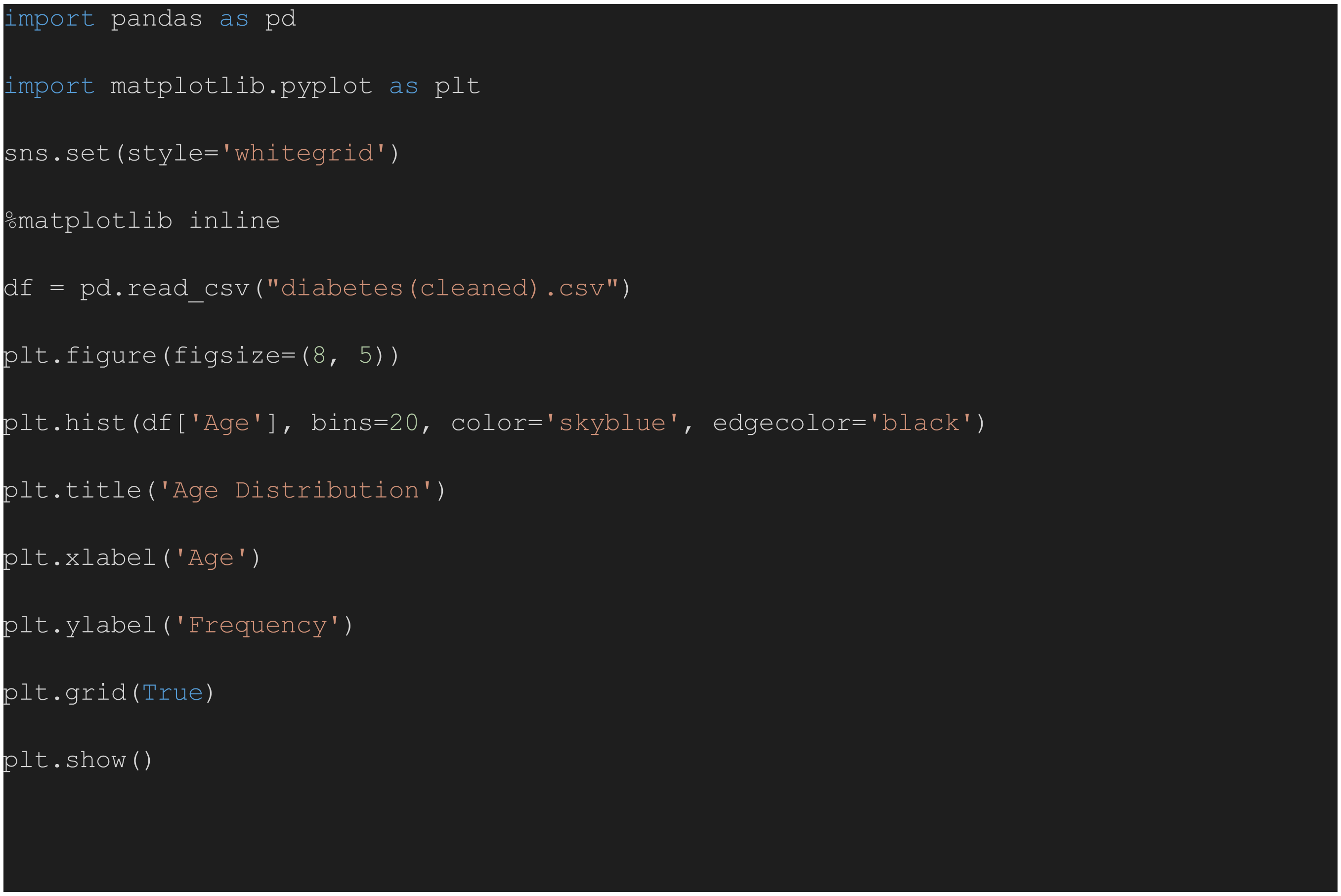
Theory: Exploratory Data Analysis (EDA) is a critical phase in the data analysis process that allows us to delve into the healthcare dataset using the powerful tools of Pandas and Matplotlib. EDA serves as a foundational step to unveil the inherent structure and characteristics of the data, paving the way for meaningful insights and actionable conclusions.

Pandas, a Python library, empowers us to efficiently manipulate and preprocess the healthcare data. We can employ Pandas functions to clean the dataset, handle missing values, and transform variables, ensuring the data is ready for analysis. By summarizing statistics, calculating measures of central tendency and dispersion, and categorizing data based on attributes such as age, gender, and health indicators, Pandas facilitates a comprehensive understanding of the dataset's basic attributes.

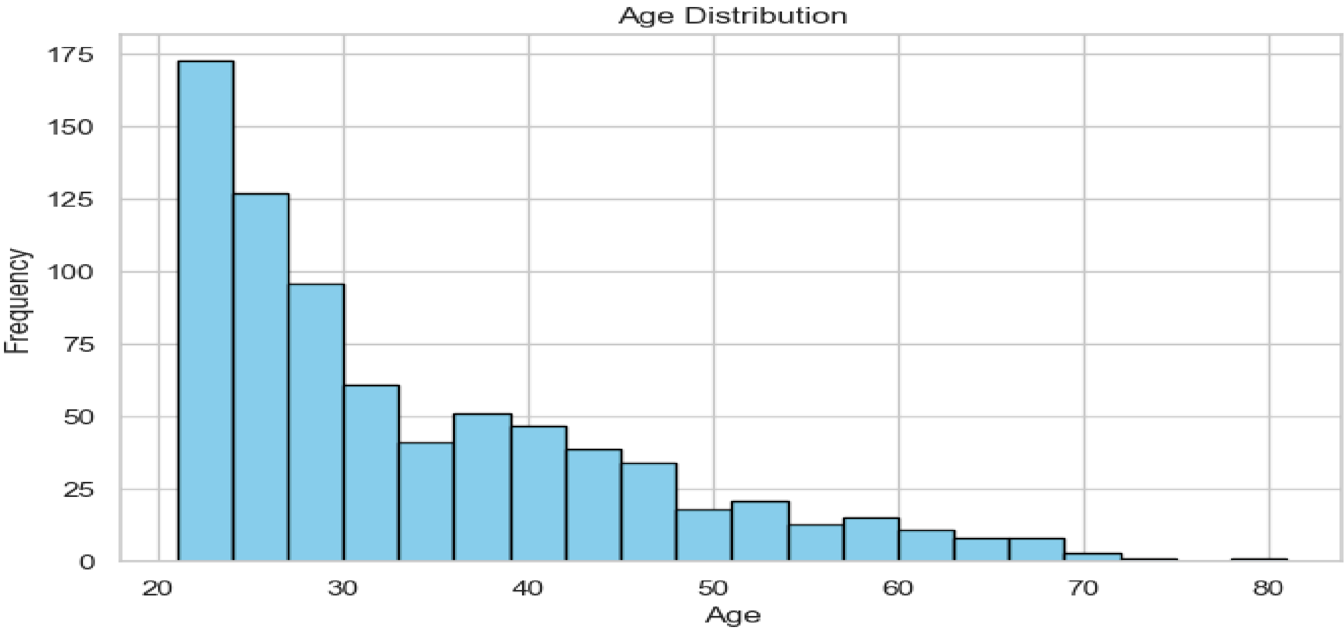
Matplotlib, on the other hand, equips us with an arsenal of visualization techniques. Through scatter plots, histograms, box plots, and correlation matrices, we can visually grasp the distribution, relationships, and variations within the healthcare data. These visualizations aid in identifying trends, outliers, and potential patterns that may warrant deeper investigation.

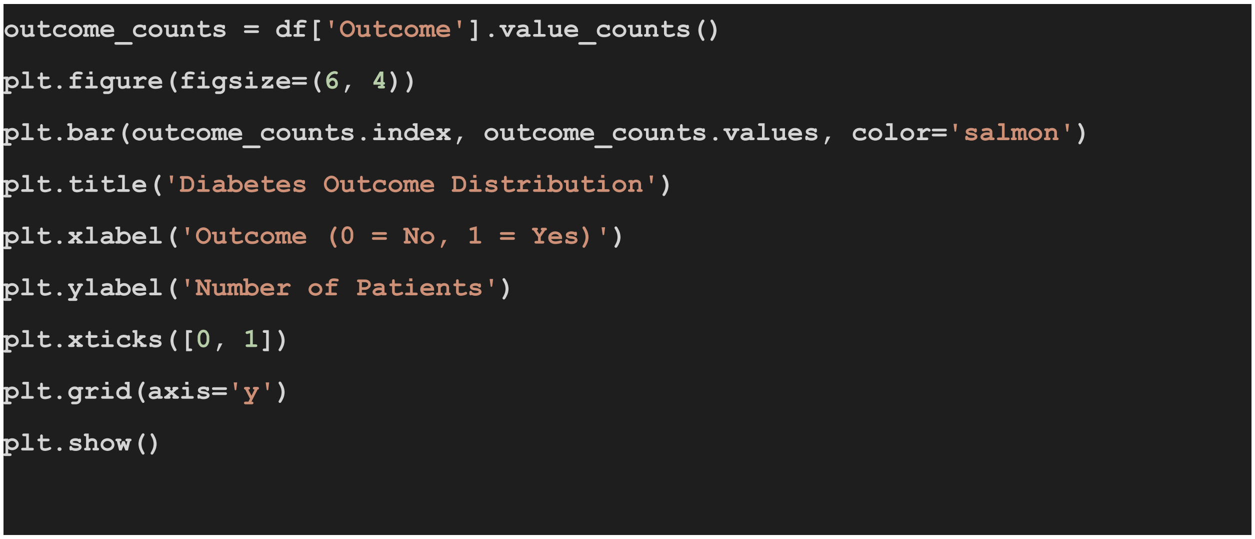
The objective of this EDA is to leverage the synergy of Pandas and Matplotlib to extract actionable insights from the healthcare dataset. By combining statistical analysis with compelling visuals, we aim to uncover meaningful relationships between symptoms, demographics, and health indicators. These insights can guide informed decision-making, influence healthcare policies, and spark new research directions, ultimately contributing to improved patient care and outcomes. As we embark on this journey of exploration, the union of Pandas and Matplotlib serves as our compass, guiding us toward a deeper understanding of the intricate landscape of healthcare data.

Program and output:

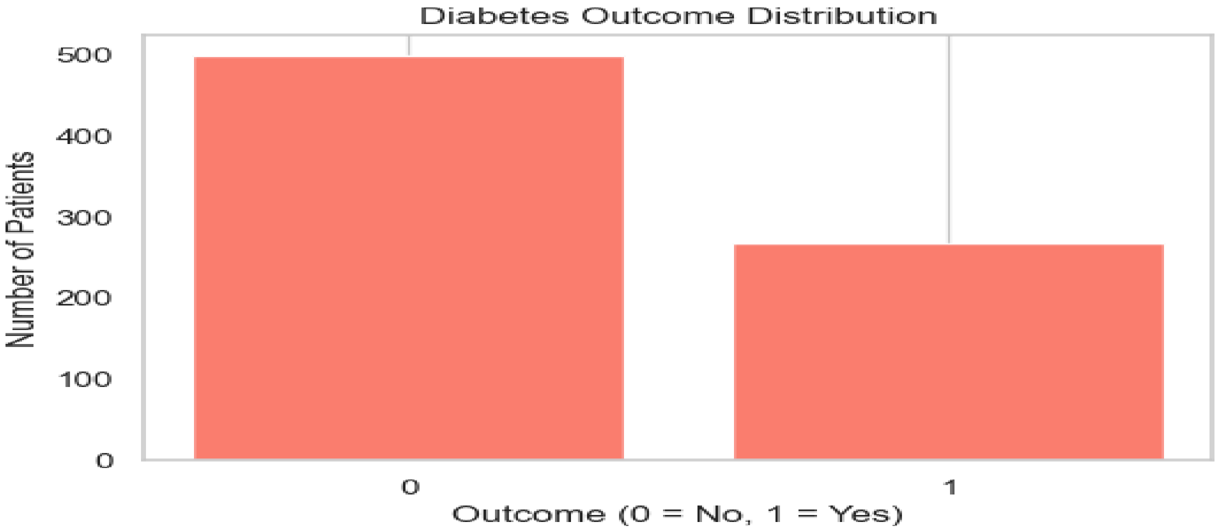


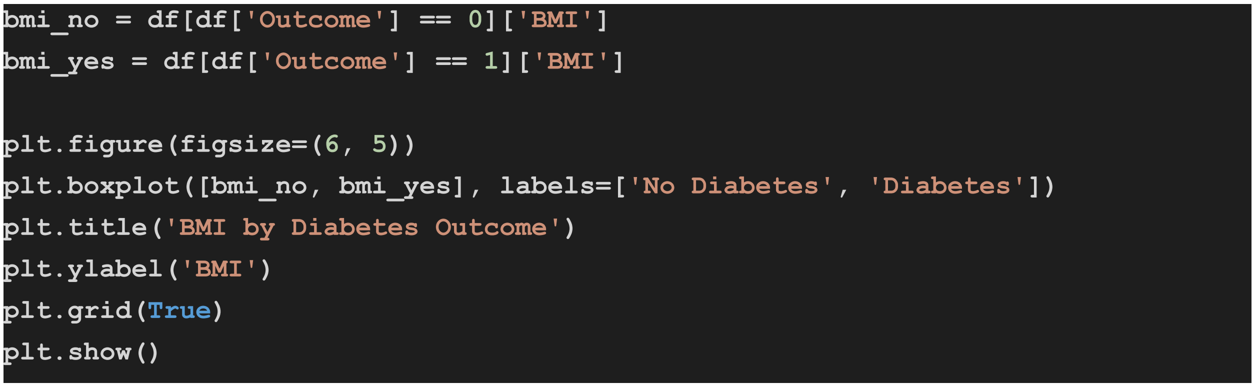
```
import pandas as pd

import matplotlib.pyplot as plt

sns.set(style='whitegrid')

%matplotlib inline

df = pd.read_csv("diabetes(cleaned).csv")

plt.figure(figsize=(8, 5))

plt.hist(df['Age'], bins=20, color='skyblue', edgecolor='black')

plt.title('Age Distribution')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.grid(True)

plt.show()
```



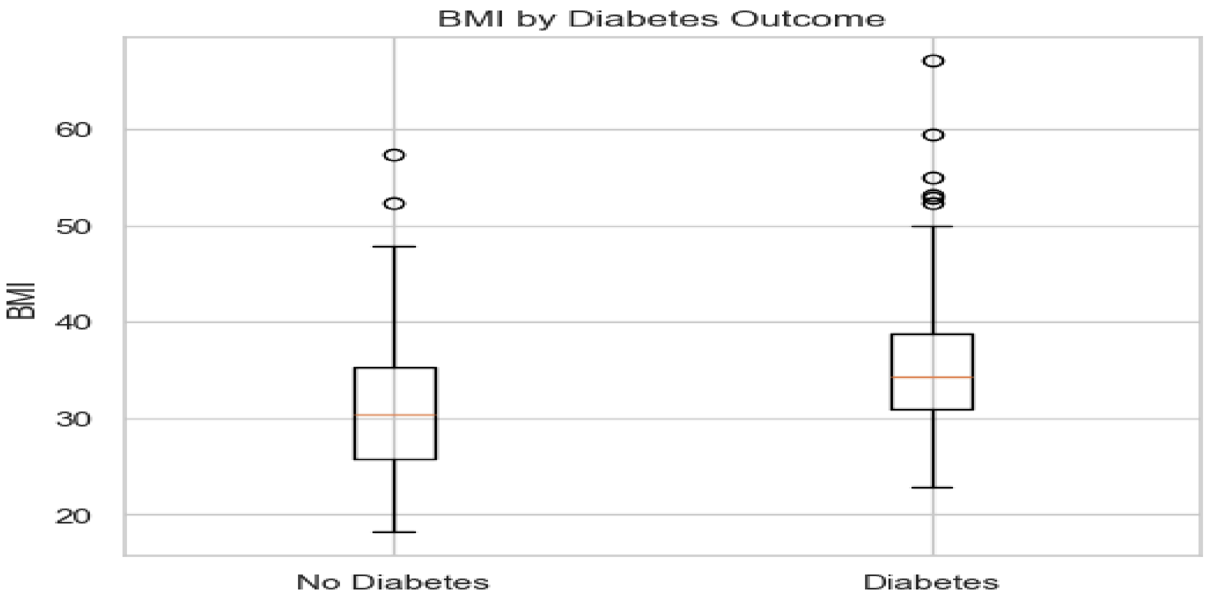


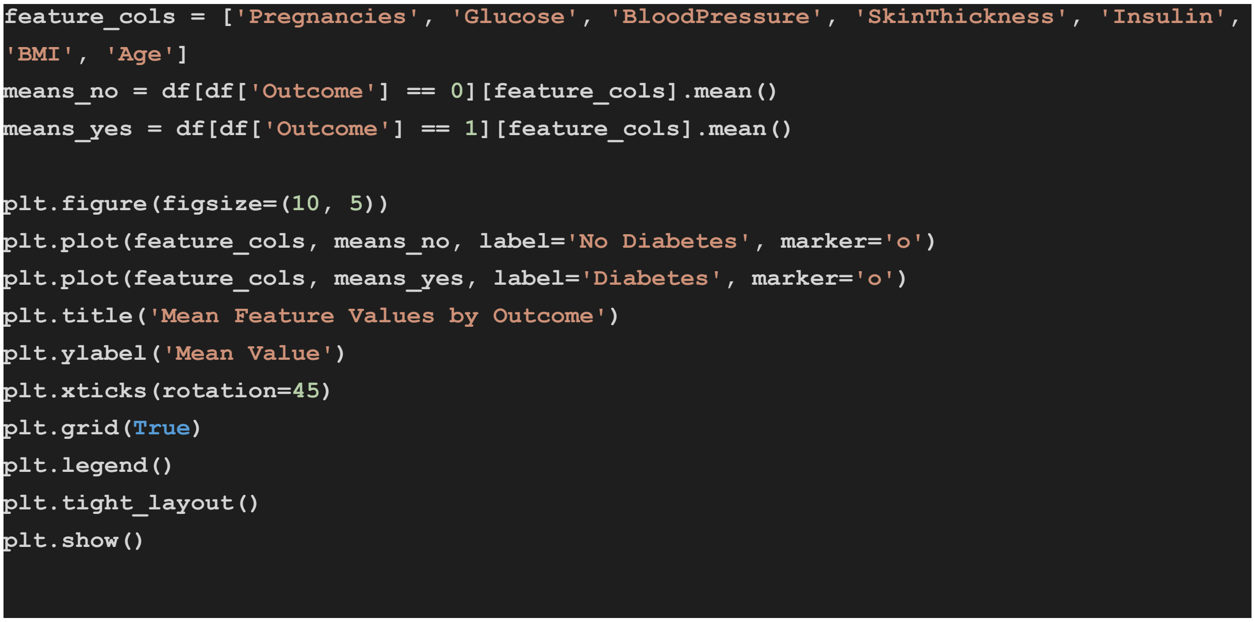
```
outcome_counts = df['Outcome'].value_counts()
plt.figure(figsize=(6, 4))
plt.bar(outcome_counts.index, outcome_counts.values, color='salmon')
plt.title('Diabetes Outcome Distribution')
plt.xlabel('Outcome (0 = No, 1 = Yes)')
plt.ylabel('Number of Patients')
plt.xticks([0, 1])
plt.grid(axis='y')
plt.show()
```



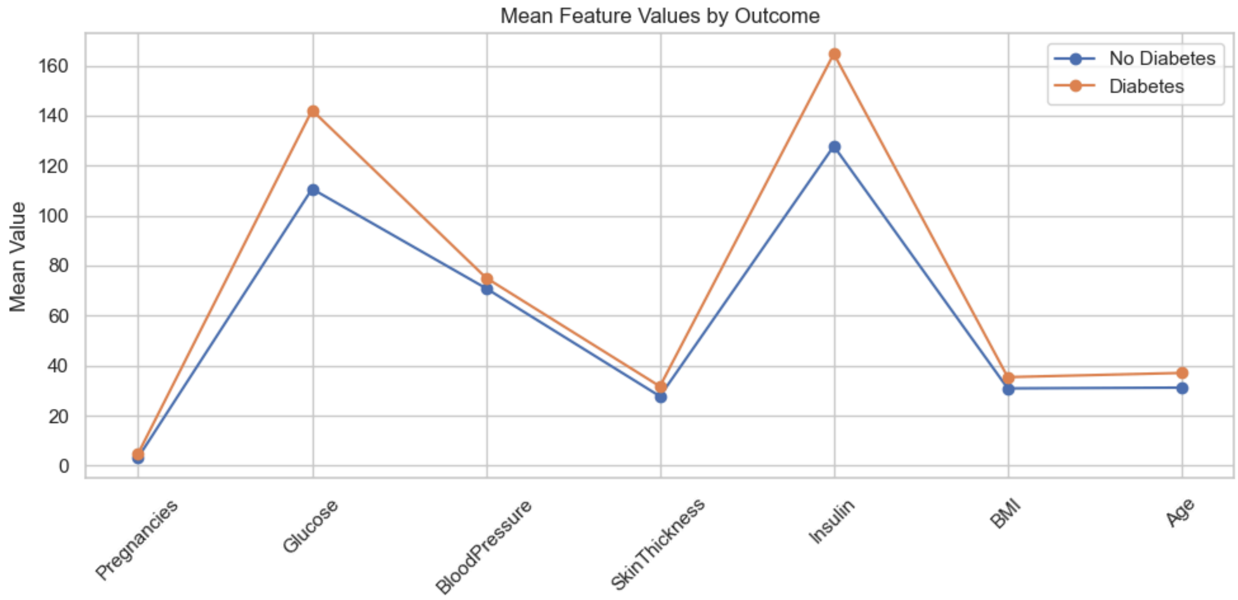


```
bmi_no = df[df['Outcome'] == 0]['BMI']
bmi_yes = df[df['Outcome'] == 1]['BMI']

plt.figure(figsize=(6, 5))
plt.boxplot([bmi_no, bmi_yes], labels=['No Diabetes', 'Diabetes'])
plt.title('BMI by Diabetes Outcome')
plt.ylabel('BMI')
plt.grid(True)
plt.show()
```





```
feature_cols = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
'BMI', 'Age']
means_no = df[df['Outcome'] == 0][feature_cols].mean()
means_yes = df[df['Outcome'] == 1][feature_cols].mean()

plt.figure(figsize=(10, 5))
plt.plot(feature_cols, means_no, label='No Diabetes', marker='o')
plt.plot(feature_cols, means_yes, label='Diabetes', marker='o')
plt.title('Mean Feature Values by Outcome')
plt.ylabel('Mean Value')
plt.xticks(rotation=45)
plt.grid(True)
plt.legend()
plt.tight_layout()
plt.show()
```



Conclusion-

1.What insights were obtained from the healthcare dataset through EDA using Pandas and Matplotlib?

EDA using Pandas and Matplotlib helped us understand the structure and key features of the healthcare data. We found that higher glucose levels, BMI, and age were strongly linked to diabetes. The data showed more non-diabetic patients, but diabetic cases increased in older age groups. Box plots and scatter plots helped spot these patterns clearly.

2. How were missing values and outliers identified and handled in the dataset?
Missing values were identified using `.isnull().sum()` and handled by either filling them with forward fill (`fillna()`) or dropping them if necessary. Outliers were detected using box plots, especially in features like BMI and glucose. Some extreme or unrealistic values were removed or capped to avoid skewing the results.

3.What trends or patterns were observed in the healthcare data after visualization?
Visualizations revealed that diabetes is more common in people with high glucose, high BMI, and older age. Age groups above 40 showed a higher number of diabetes cases. Glucose and BMI had a strong impact on the outcome, showing a clear trend in the data related to health risks.