**Healthcare Data Cleaning and Analysis**

**Title Page**

**Project Title:** Healthcare Data Cleaning and Analysis  
**Author:** Kartikey Jaiswal

**Submitted to:** Mr.Shivansh Prasad  
**Date:** 11-03-2025

**1. Introduction**

Data cleaning is a crucial step in data analysis and machine learning. In this project, we process healthcare data by removing inconsistencies, handling missing values, and identifying outliers. The goal is to prepare high-quality data for further analysis and predictive modeling.

**2. Methodology**

**Step 1: Data Loading**

* Read the dataset using Pandas.
* Inspect the data structure, including column names, types, and missing values.

**Step 2: Data Cleaning**

* Remove duplicate records.
* Handle missing values (if any) appropriately.
* Detect and remove outliers using the IQR (Interquartile Range) method.

**Step 3: Data Visualization**

* Use boxplots to visualize outliers in key variables (Blood Pressure, Sugar Level, and Weight).

**Step 4: Summary Statistics**

* Compute descriptive statistics to understand the distribution of data.

**3. Code Implementation**

The following Python script was used to clean and analyze the healthcare dataset:

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

def load\_data(file\_path):

return pd.read\_csv(file\_path)

def check\_missing\_values(df):

return df.isnull().sum()

def remove\_duplicates(df):

return df.drop\_duplicates()

def remove\_outliers(df, column):

Q1 = df[column].quantile(0.25)

Q3 = df[column].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

return df[(df[column] >= lower\_bound) & (df[column] <= upper\_bound)]

def visualize\_data(df):

fig, axes = plt.subplots(1, 3, figsize=(15, 5))

sns.boxplot(y=df['BloodPressure'], ax=axes[0], color="skyblue")

axes[0].set\_title("Blood Pressure")

sns.boxplot(y=df['SugarLevel'], ax=axes[1], color="lightcoral")

axes[1].set\_title("Sugar Level")

sns.boxplot(y=df['Weight'], ax=axes[2], color="lightgreen")

axes[2].set\_title("Weight")

plt.tight\_layout()

plt.show()

def main():

file\_path = "healthcare\_data.csv" # Update with actual path

df = load\_data(file\_path)

print("Initial Data Info:")

print(df.info())

print("\nMissing Values:")

print(check\_missing\_values(df))

df = remove\_duplicates(df)

for col in ['BloodPressure', 'SugarLevel', 'Weight']:

df = remove\_outliers(df, col)

print("\nCleaned Data Info:")

print(df.info())

print("\nBasic Statistics:")

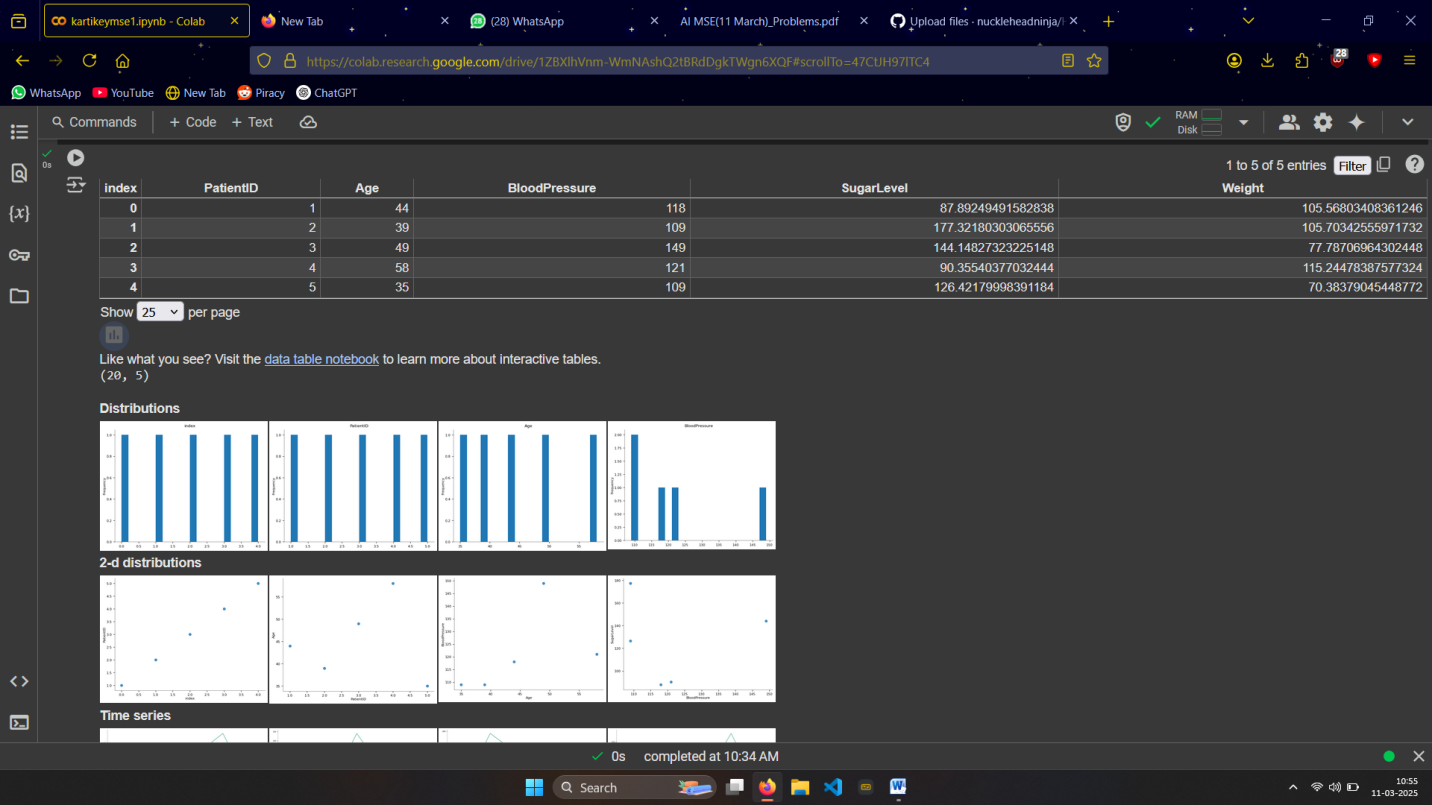
print(df.describe())

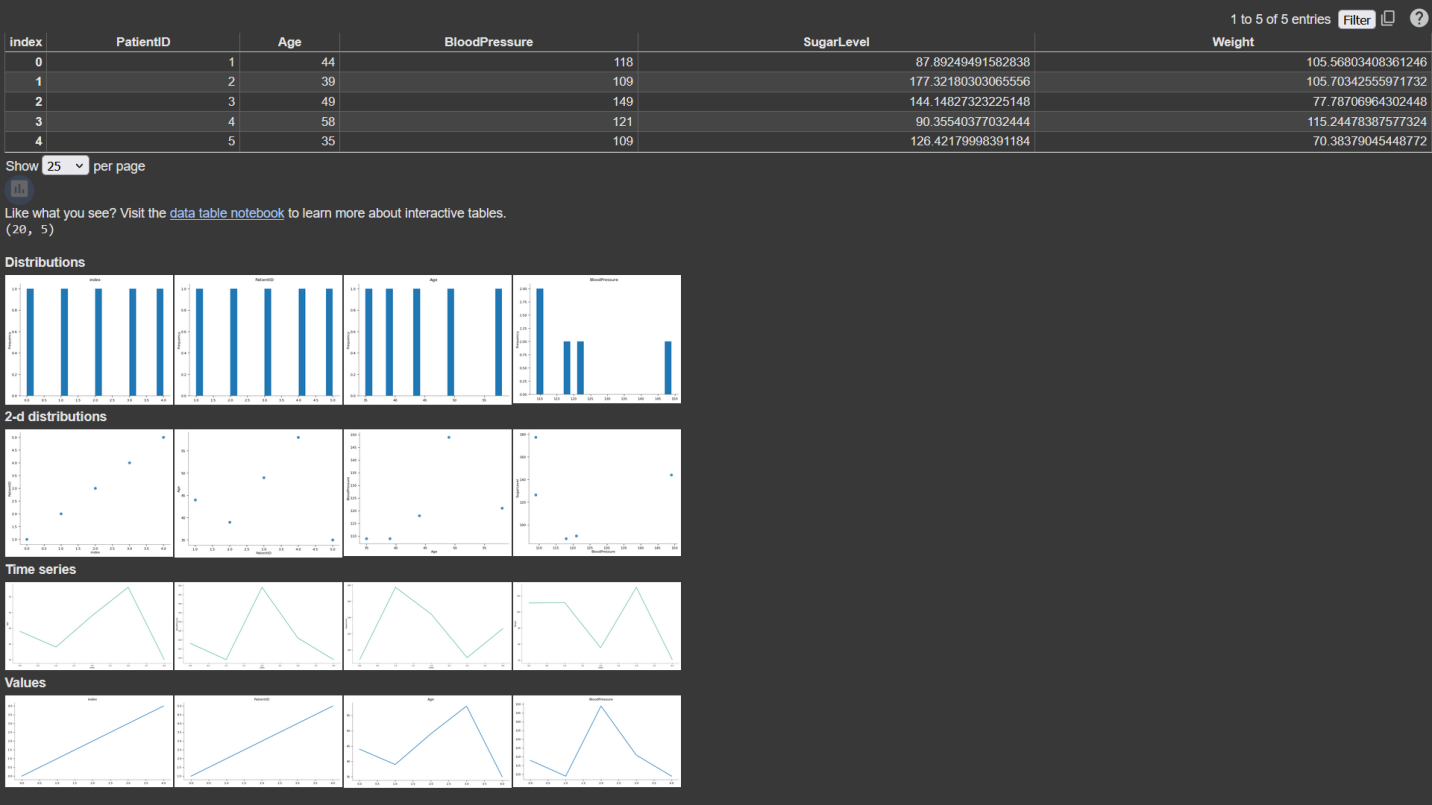
visualize\_data(df)

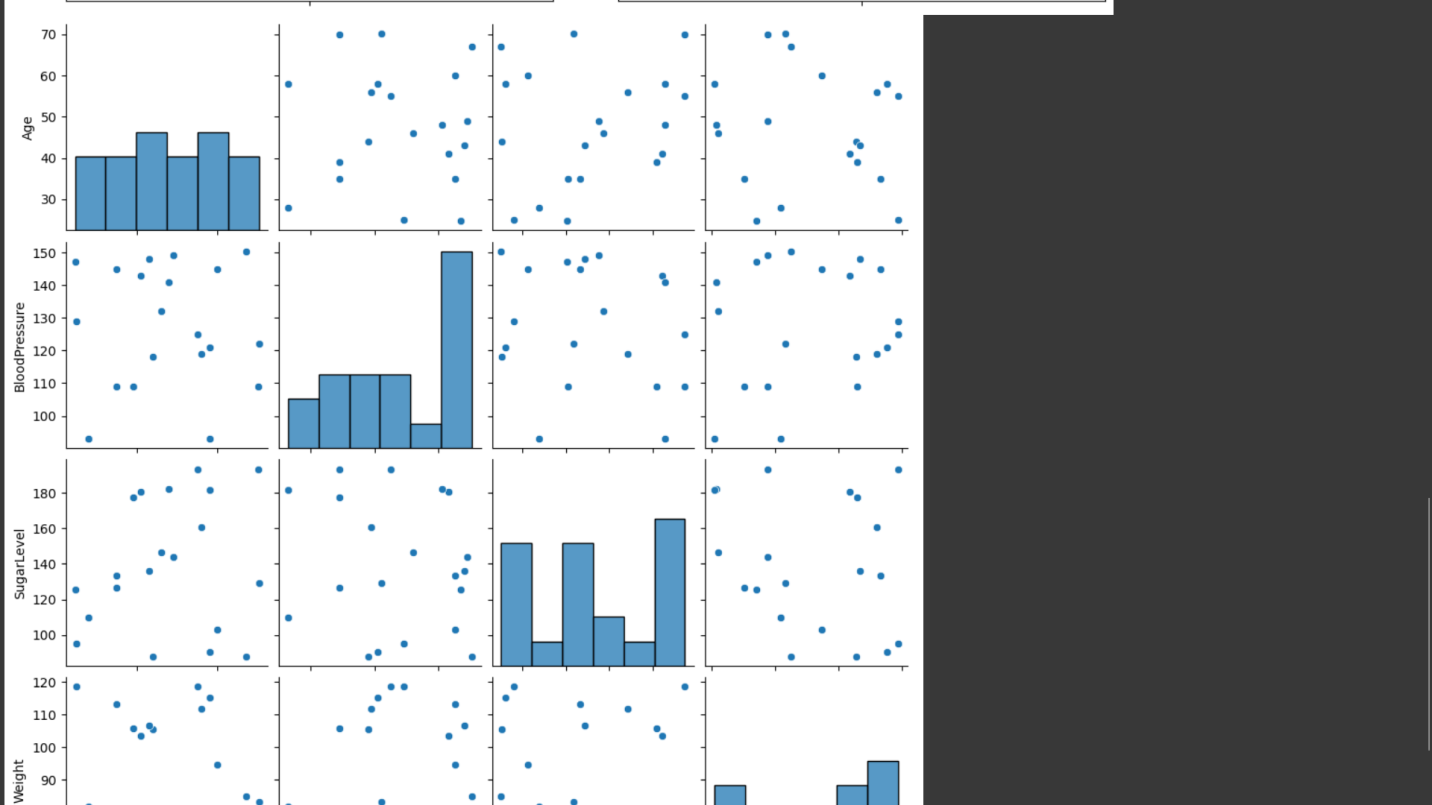
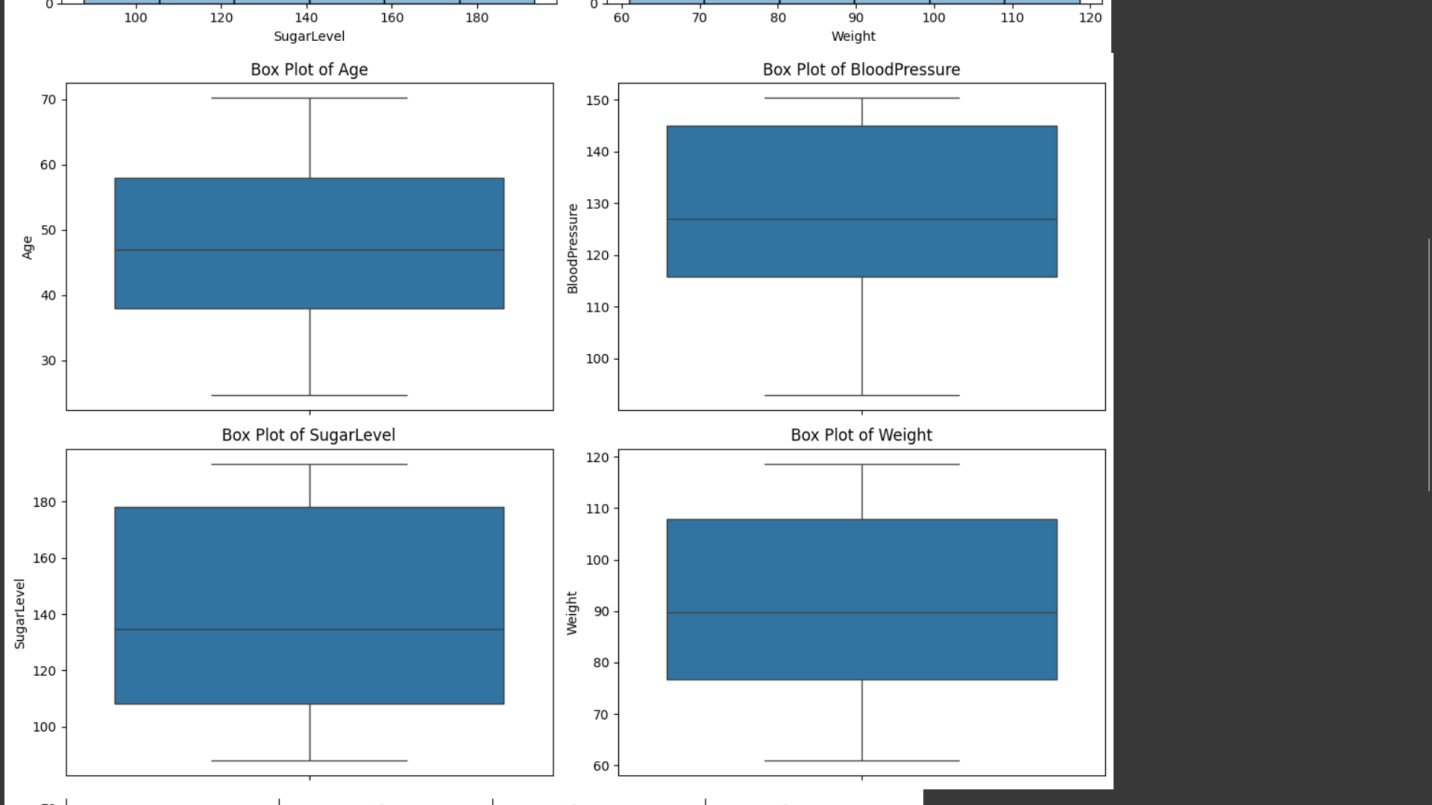
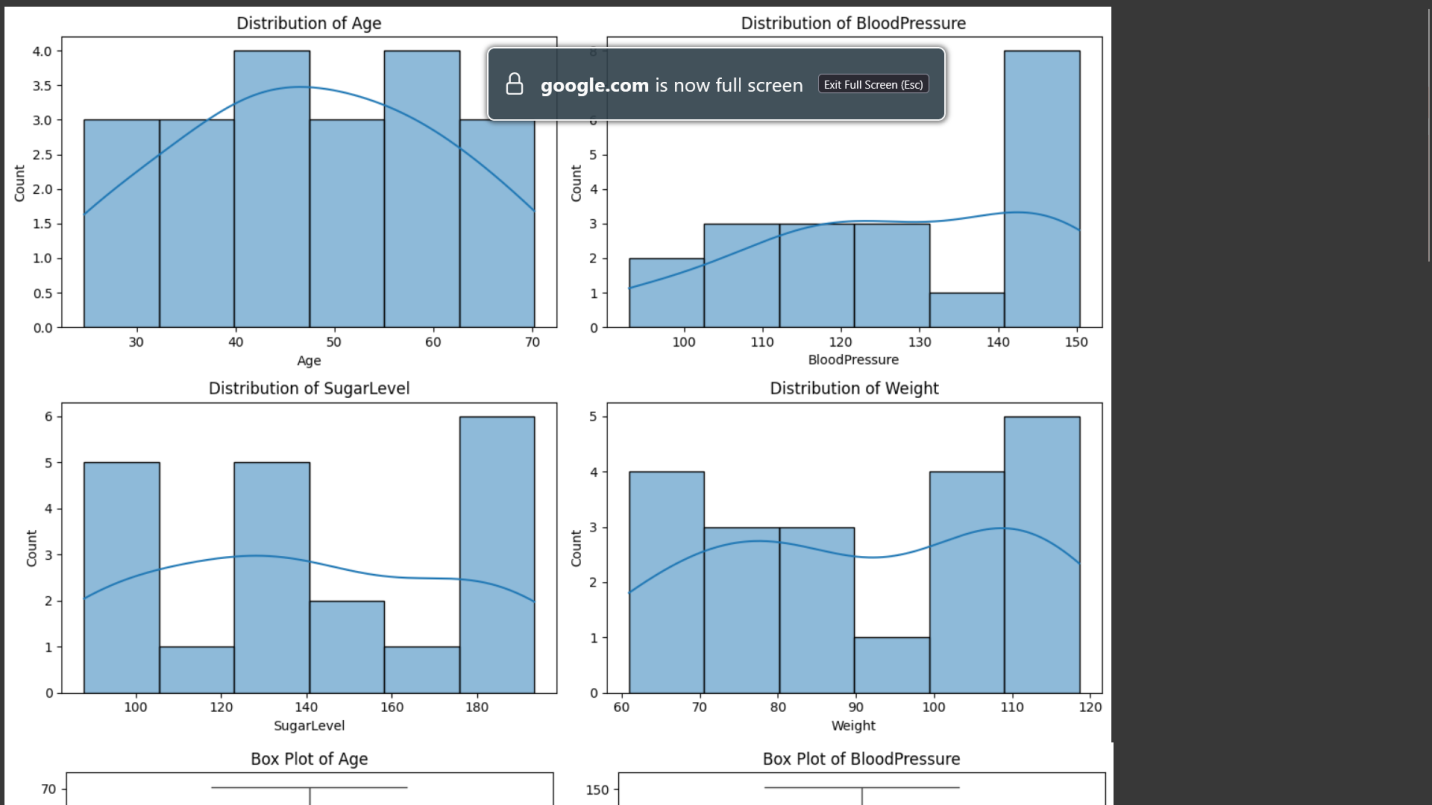
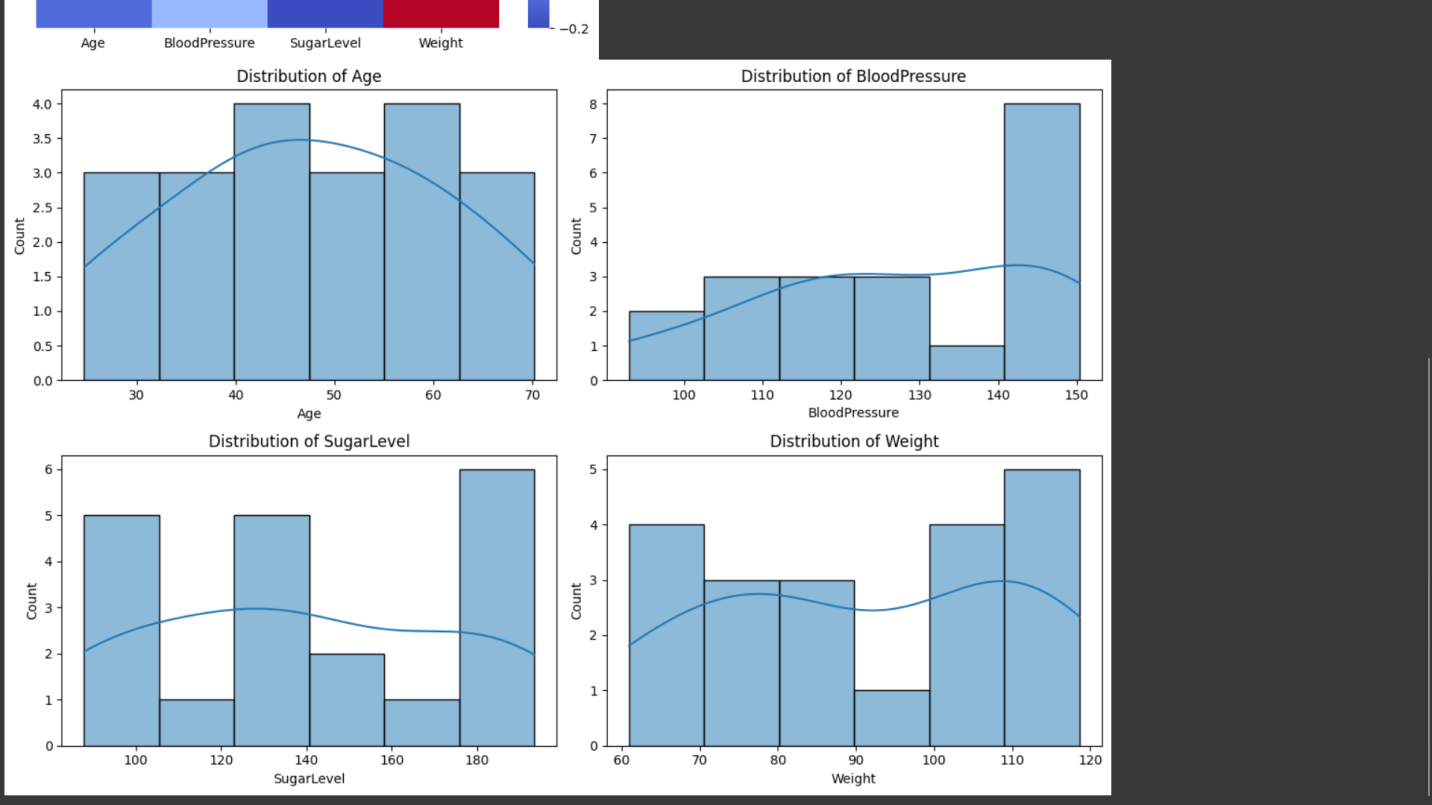
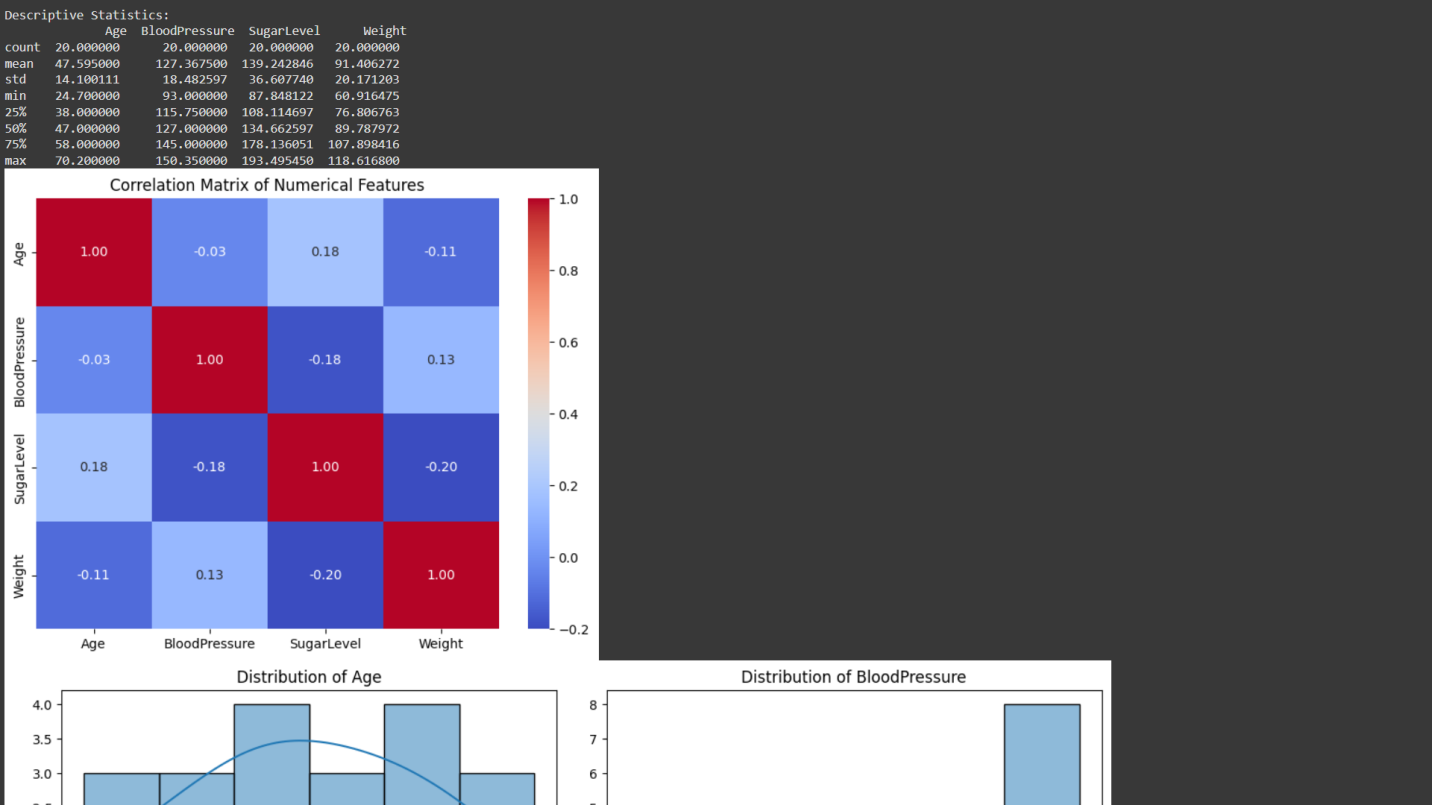
if \_\_name\_\_ == "\_\_main\_\_":

main()

**4. Screenshots of Output**







**5. Conclusion**

This project successfully cleaned and analyzed healthcare data by addressing missing values, duplicates, and outliers. The cleaned dataset can now be used for predictive modeling or further research.

**References**

**Journal Articles & Conference Papers**

1. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2*(1), 1-10.
2. Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics, 114*, 57-65.

**Books**

1. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques.* Elsevier.
2. Dey, N., Ashour, A. S., & Borra, S. (2018). *Healthcare data analytics and management.* Academic Press

**Reports & Online Resources**

1. IBM. (2021). The role of AI in healthcare data management. Retrieved from <https://www.ibm.com>
2. World Health Organization (WHO). (2022). Digital health interventions and data quality. Retrieved from <https://www.who.int>
3. National Institutes of Health (NIH). (2021). AI in healthcare data cleaning: Challenges and best practices. Retrieved from <https://www.nih.gov>

THANK YOU FOR READING