-

PROJECT TITLE :

Healthcare Data Exploration

NAME : ANEKA SRIVASTAVA

ROLL NO: 202401100300040

COURSE: BTECH(CSE-AI)

DATE : 11-03-25

INTRODUCTION

In this project, we aim to explore a healthcare dataset and perform basic data exploration and cleaning tasks. The dataset contains information about various healthcare parameters like Age, BloodPressure, SugarLevel, and Weight. The goal is to understand the distribution of numerical data, check for missing values, handle those missing values, and visualize the relationships between different features. These steps are essential to ensure that the data is clean and ready for more advanced analysis or modeling.

Methodology

1. Data Loading:

The dataset is loaded into a **pandas DataFrame** from a CSV file. This allows for easy manipulation and access to the data.

2. Data Exploration:

We explore the dataset by:

- Displaying the first few rows using head() to understand the structure of the data.
- Using describe() to generate summary statistics for numerical columns such as mean, standard deviation, and percentiles.

- Checking for missing values using isnull().sum() to identify columns with missing or null values.

3. Data Cleaning:

Missing values are filled with the mean of their respective columns using fillna(df.mean()). This approach ensures that the dataset is complete without any missing values, allowing for more accurate analysis.

4. Data Visualization:

Visualizations are created to understand the distribution of data and relationships between features:

- **Histograms** are used to visualize the distribution of numerical features like Age, BloodPressure, SugarLevel, and Weight.
- A **heatmap** is used to visualize the correlation between numerical variables, helping to identify any significant relationships between them.
- A **count plot** is created to visualize the distribution of values in the Age column (though this can be more useful for categorical data).

5. Saving the Cleaned Data:

After cleaning, the dataset is saved to a new CSV file for future use.

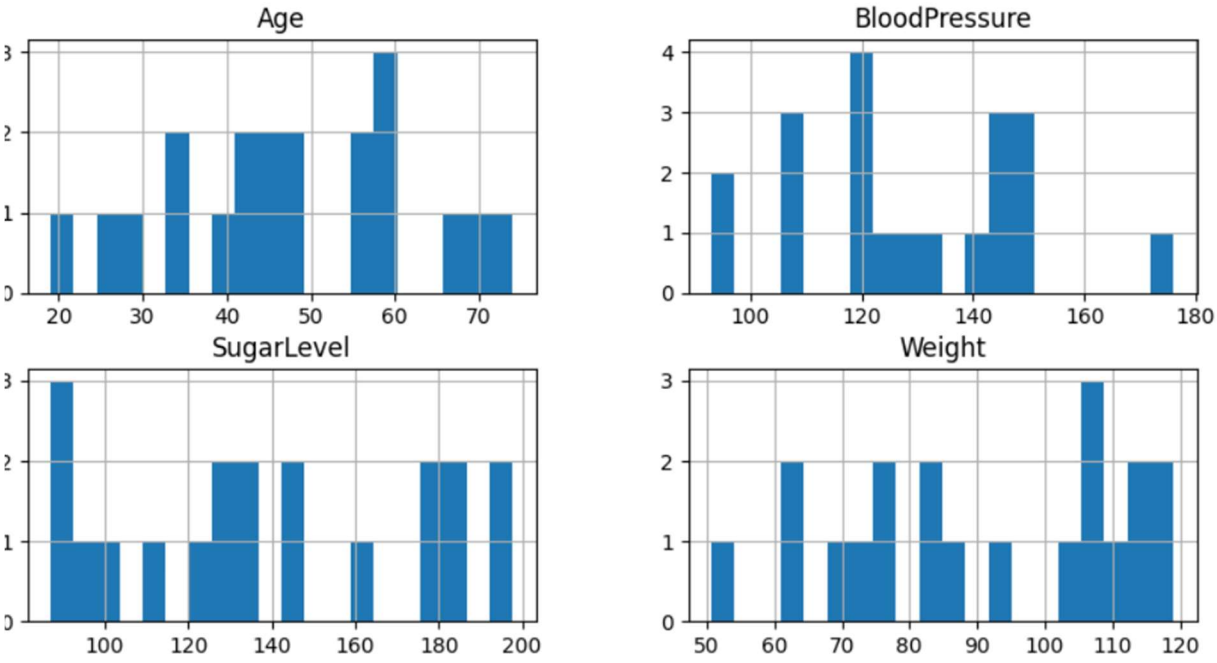
CODE

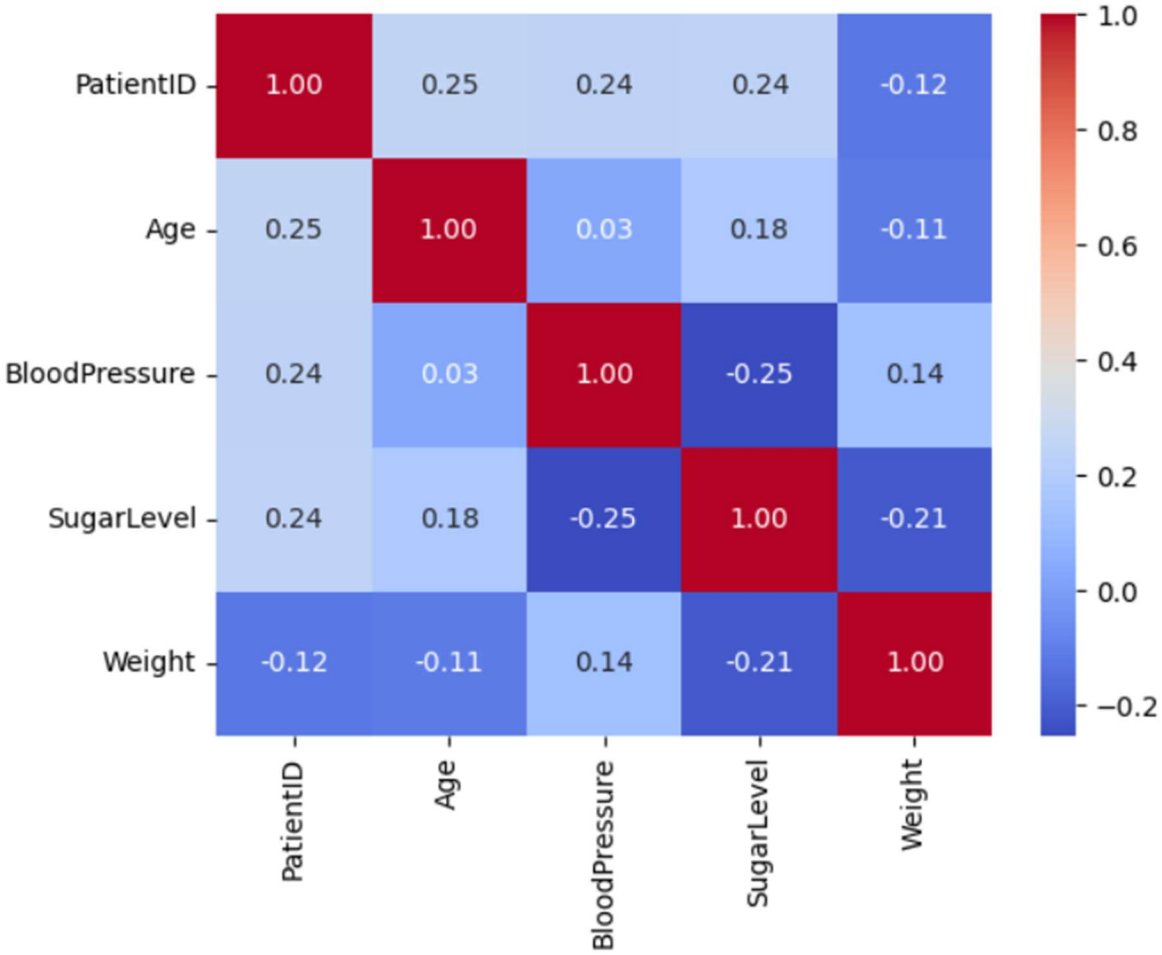
```
import pandas as pd
```

```
import matplotlib.pyplot as plt
import seaborn as sns

# Load dataset (replace with your actual file path)
df = pd.read_csv('/content/healthcare.csv')

# Show first 5 rows of the dataset
print(df.head())

# Basic statistics for numerical columns
print(df.describe())

# Check for missing values in the dataset
print(df.isnull().sum())

# Fill missing values with the column mean
df.fillna(df.mean(), inplace=True)

# Verify if there are still missing values
print(df.isnull().sum())


# Plot histograms for Age, BloodPressure, SugarLevel, Weight
df[['Age', 'BloodPressure', 'SugarLevel',
'Weight']].hist(figsize=(10, 5), bins=20)
plt.show()

# Show correlation between numerical features using a heatmap
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt='.2f')
```

```
plt.show()

# Plot the count of each category in the 'Age' column
sns.countplot(x='Age', data=df)
plt.show()

# Save the cleaned data to a new CSV file
df.to_csv('cleaned_healthcare_data.csv', index=False)
```

OUTPUT





REFERENCES

- **Pandas Documentation**:
https://pandas.pydata.org/pandas-docs/stable/

- **Matplotlib Documentation**:
https://matplotlib.org/stable/contents.html