Healthcare Data Exploration

Report

Name: Shourya Ojha
Class Roll No : 19
Univ Roll No : 202401100300239
Branch: CSE AI
Library ID: 2428CSEAI1182

Date: 11 March 2025

Introduction

Healthcare data is crucial for analysing trends, detecting diseases, and improving patient care. This project explores a healthcare dataset to understand patterns, detect missing values, visualize distributions, and identify outliers. By analysing various attributes such as age, blood pressure, and correlations, we gain insights that can be valuable for healthcare professionals.

The growing digitization of healthcare records has made it easier to collect vast amounts of patient data. However, raw data is often messy, containing missing values and potential inconsistencies. Therefore, it is essential to preprocess and analyse the dataset efficiently to extract meaningful insights. This report focuses on data exploration techniques that help in understanding the dataset, identifying crucial variables, and uncovering hidden patterns that could aid in better medical decision-making.

This study aims to:

- Identify and handle missing values in the dataset.
- Visualize key variables to understand their distributions.
- Detect outliers that could indicate data anomalies or medical conditions.
- Establish correlations between different healthcare parameters to support predictive analytics

Methodology

Data Loading & Exploration:
The dataset is loaded using Pandas for analysis.
Basic dataset information, including column types, data types, and missing values, is displayed to understand data quality.

Data Visualization:
A histogram is plotted to visualize the distribution of the Age column and detect skewness.
A boxplot is used for BloodPressure to check for potential outliers and extreme values.
A heatmap is generated to identify correlations between numerical variables, which helps in understanding patterns in the dataset.

Outlier Detection:
The Interquartile Range (IQR) method is used to detect and quantify outliers in numerical columns.
This helps in identifying extreme values that might affect the analysis and decision-making.
Summary Statistics:
The dataset is summarized using statistical measures such as mean, median, standard deviation, and percentiles.
These statistics provide insights into the central tendency and variability of key attributes.

Observations & Insights:
The combination of visualizations and statistical analysis aids in drawing meaningful conclusions.
Identifying missing values and extreme data points allows for better data preprocessing and model preparation in future steps.

CODE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv("healthcare_data.csv")

# Display basic information
print("Dataset Info:\n")
df.info()  # Prints summary of dataset including column types and non-null counts
print("\nFirst 5 rows:\n", df.head())  # Displays first five rows of the dataset

# Check for missing values
print("\nMissing Values:\n", df.isnull().sum())  # Checks for missing values in each column

# Summary statistics
print("\nSummary Statistics:\n", df.describe())  # Displays statistical summary of numerical columns

# Visualizing distributions
plt.figure(figsize=(10, 5))
sns.histplot(df['Age'], bins=10, kde=True)  # Plots histogram with density estimate for Age column
plt.title("Age Distribution")
plt.show()
```

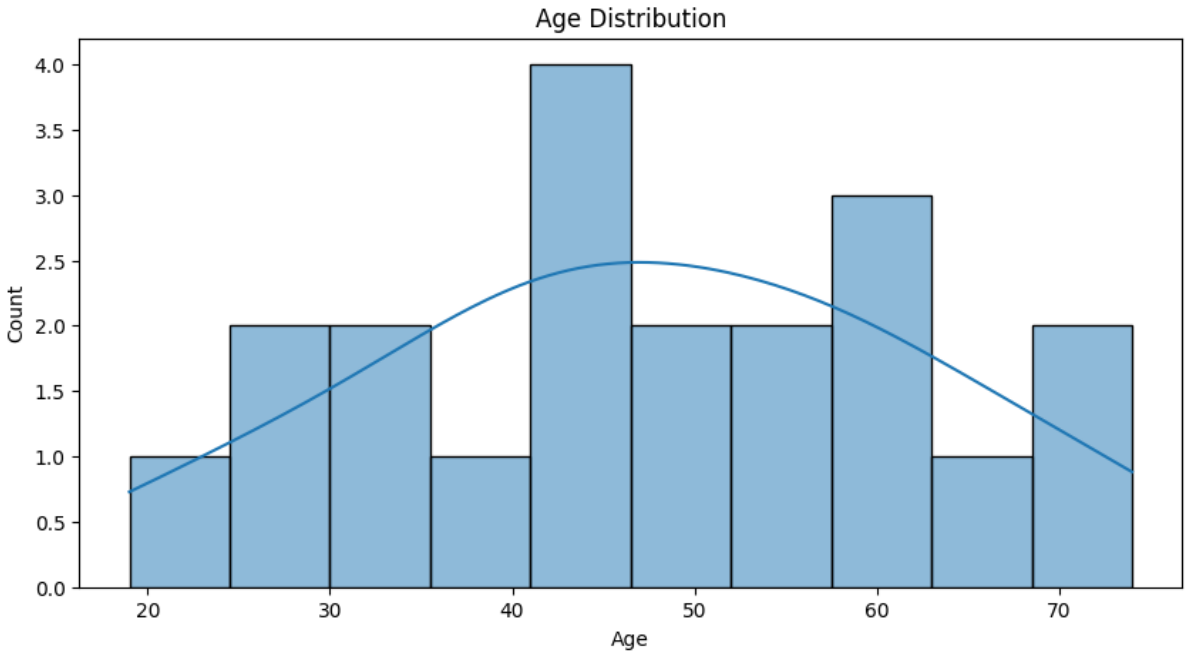
```
plt.figure(figsize=(10, 5))
sns.boxplot(x=df['BloodPressure'])  # Creates a boxplot for
BloodPressure column to identify outliers
plt.title("Blood Pressure Boxplot")
plt.show()

plt.figure(figsize=(10, 5))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm',
fmt=".2f")  # Displays correlation heatmap
plt.title("Correlation Heatmap")
plt.show()

# Detecting outliers using IQR
Q1 = df.quantile(0.25)  # First quartile (25th percentile)
Q3 = df.quantile(0.75)  # Third quartile (75th percentile)
IQR = Q3 - Q1  # Interquartile range
outliers = ((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 *
IQR))).sum()  # Count of outliers in each column
print("\nOutlier Count:\n", outliers)

print("\nData Exploration Completed.")
```

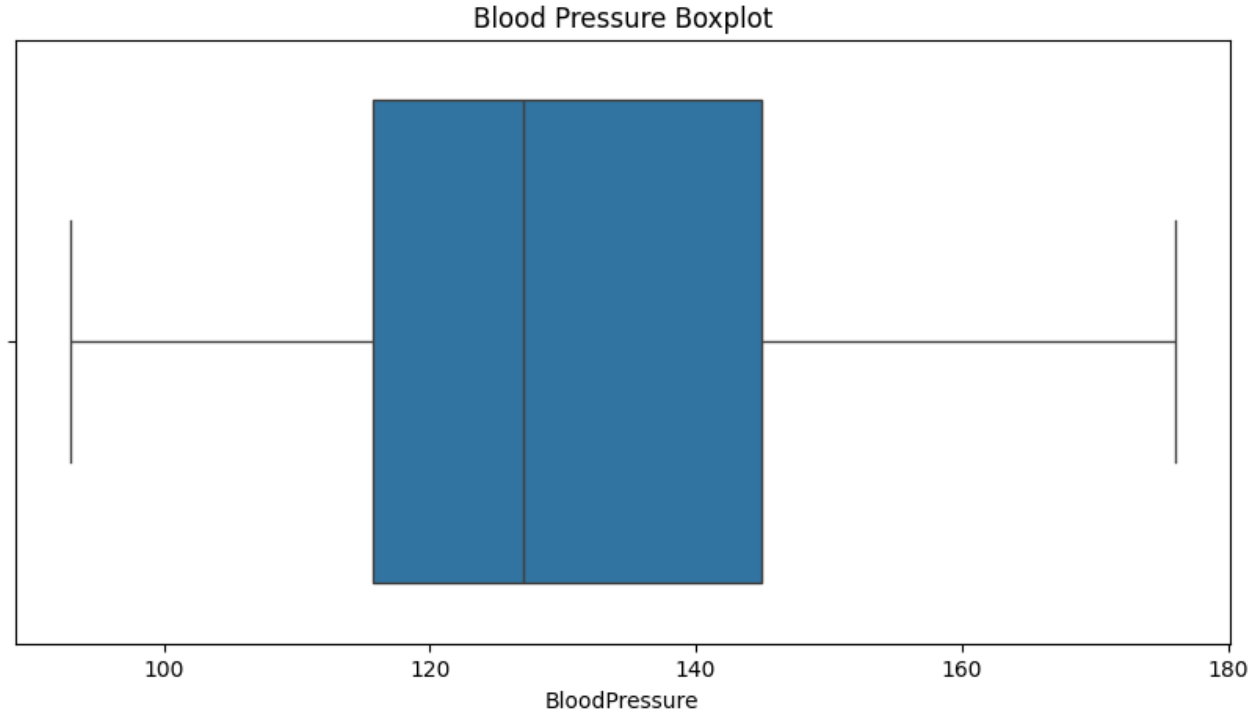
Output/Result

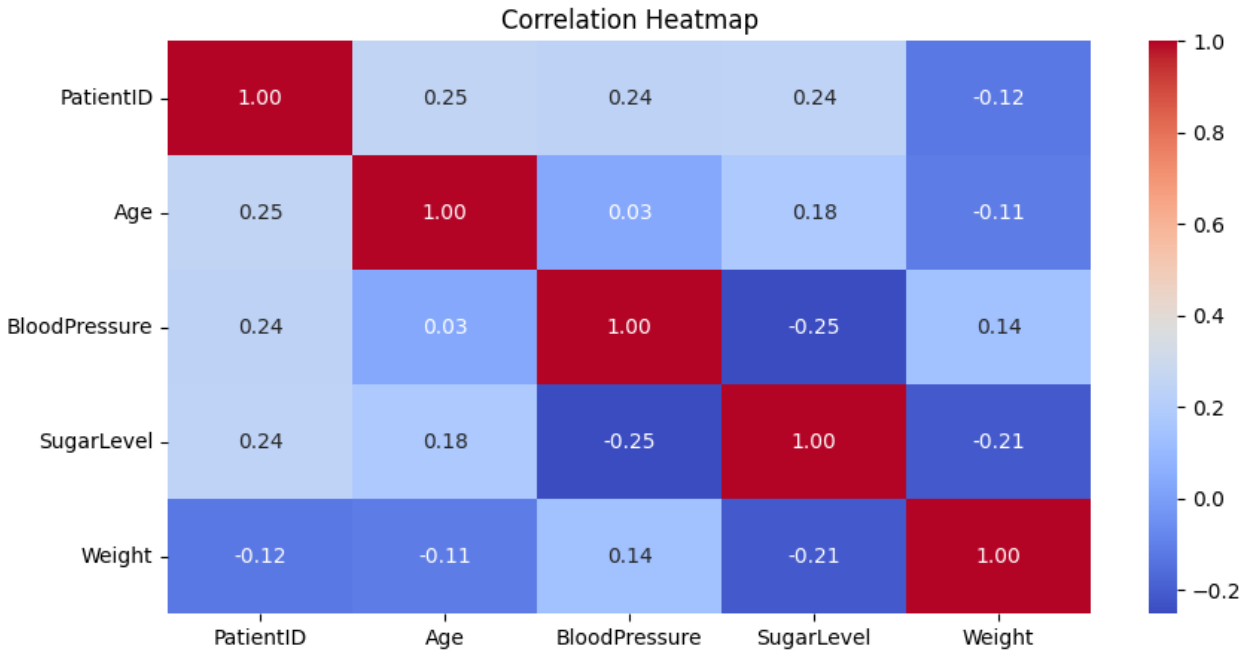
```
Dataset Info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   PatientID      20 non-null     int64
 1   Age            20 non-null     int64
 2   BloodPressure  20 non-null     int64
 3   SugarLevel     20 non-null     float64
 4   Weight         20 non-null     float64
dtypes: float64(2), int64(3)
memory usage: 932.0 bytes
```

```
First 5 rows:
    PatientID  Age  BloodPressure  SugarLevel      Weight
0          1   44            118   87.892495  105.568034
1          2   39            109  177.321803  105.703426
2          3   49            149  144.148273   77.787070
3          4   58            121   90.355404  115.244784
4          5   35            109  126.421800   70.383790

Missing Values:
 PatientID        0
Age               0
BloodPressure     0
SugarLevel        0
Weight            0
dtype: int64
```

```
Summary Statistics:
        PatientID        Age  BloodPressure  SugarLevel      Weight
count   20.00000  20.000000      20.000000   20.000000   20.000000
mean    10.50000  47.500000     128.650000  139.412236   90.916368
std      5.91608  14.968388      20.893905   37.010795   21.124021
min      1.00000  19.000000      93.000000   87.005027   50.684835
25%      5.75000  38.000000     115.750000  108.114697   76.806763
50%     10.50000  47.000000     127.000000  134.662597   89.787972
75%     15.25000  58.000000     145.000000  178.136051  107.898416
max     20.00000  74.000000     176.000000  197.726356  119.050356
```


Age Distribution
Count
Age
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
20
30
40
50
60
70


Blood Pressure Boxplot
BloodPressure
100
120
140
160
180




```
Outlier Count:
 PatientID        0
Age               0
BloodPressure     0
SugarLevel        0
Weight            0
dtype: int64

Data Exploration Completed.
```

References/Credits

1. **Dataset Source**:
   - *Healthcare dataset sourced from KIET Group of Institutions.*

2. **Libraries & Tools Used**:
   - **Pandas**: For data loading, exploration, and preprocessing.
   - **NumPy**: For numerical computations and handling array operations.
   - **Matplotlib & Seaborn**: For data visualization, including histograms, boxplots, and heatmaps.

3. **Image Credits**:
   - Output images *from Google Collab Output.*

4. **Acknowledgments**:
   - Thanks to professors, mentors, or peers who provided guidance or assistance in completing the project.
   - *Special thanks to Mr. Abhishek Shukla for guidance on data analysis concepts.*