Healthcare Data Exploration

Project Title: Healthcare Data Exploration
Name: Anushka Shakya
Roll No.: 202401100300060

Date: 11-03-2025

Introduction

Healthcare data plays a crucial role in improving patient care, medical research, and decision-making in the healthcare industry. With the increasing availability of healthcare datasets, analyzing and understanding the data becomes essential for identifying patterns, trends, and potential risks associated with diseases, treatments, and hospital management.

This project aims to perform an in-depth exploratory data analysis (EDA) of a healthcare dataset to uncover key insights. The analysis involves identifying missing values, understanding the distribution of numerical and categorical features, detecting potential outliers, and analyzing correlations between variables. By using visualization techniques such as histograms, box plots, and heatmaps, we can better interpret the data and make data-driven decisions.

Handling missing data is a crucial step in data preprocessing, as missing values can lead to biased or misleading results. In this project, missing values are addressed using appropriate imputation techniques, such as filling numerical columns with their mean values. Additionally, the correlation matrix provides insights into how different numerical features relate to one another, which can be valuable for predictive modeling and further machine learning applications.

The results of this analysis can assist healthcare professionals, researchers, and policymakers in making informed decisions, optimizing resource allocation, and improving patient outcomes. By applying data exploration techniques, we can extract valuable knowledge from raw data, setting the foundation for further predictive analytics and machine learning applications in healthcare.

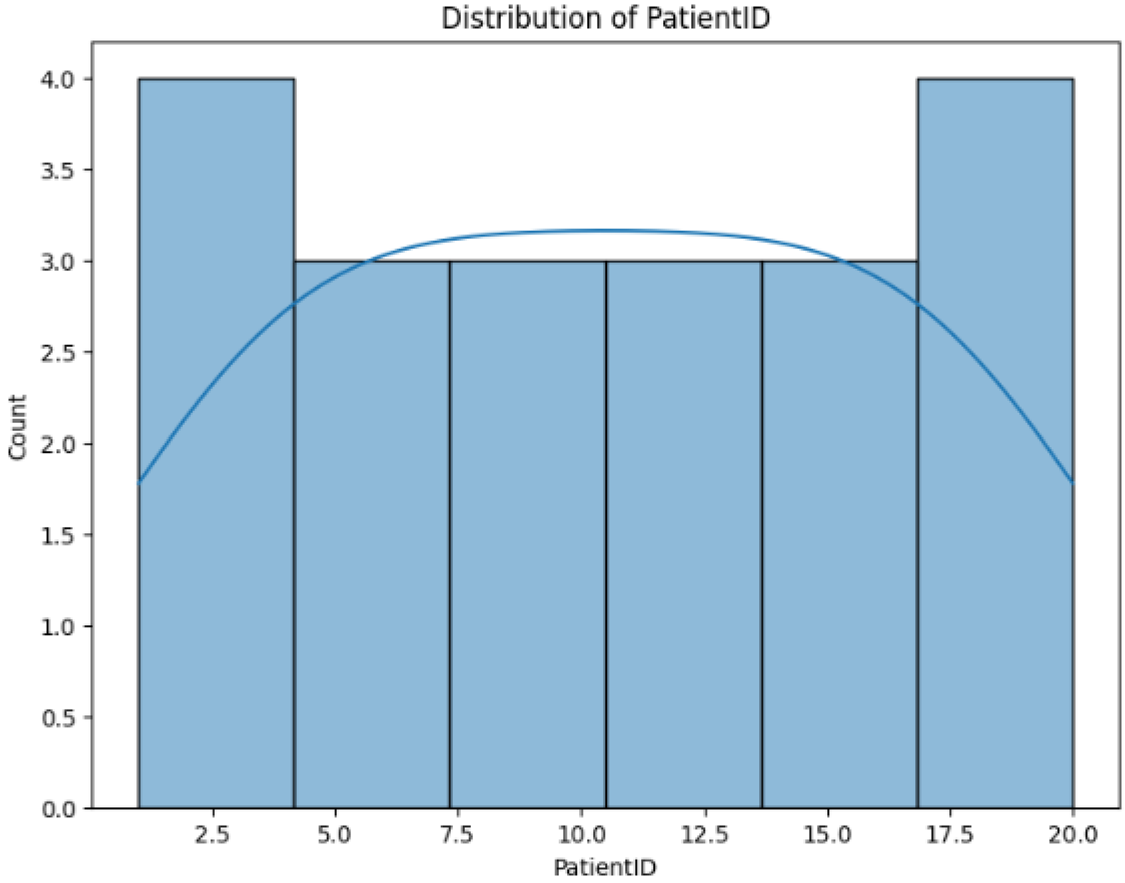
Methodology

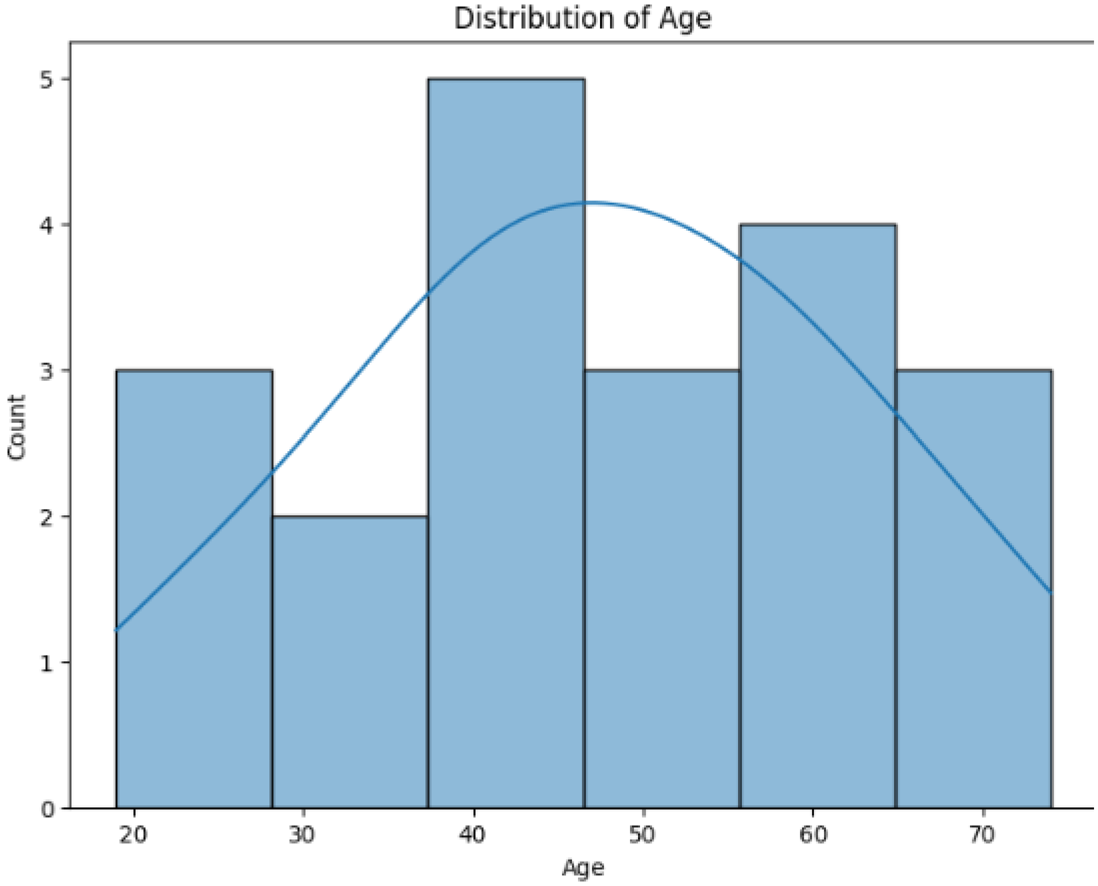
1. **Data Loading:** The dataset is read using pandas.read_csv().
2. **Data Overview:** The structure of the dataset is examined using .head(), .info(), and .describe().
3. **Missing Values Handling:**
   - Missing values are detected using .isnull().sum().
   - Numerical columns with missing values are filled using the column mean.
4. **Data Distribution Analysis:**
   - Histograms (sns.histplot()) are used for numerical features.
   - Bar plots (sns.countplot()) are used for categorical features.
5. **Correlation Analysis:** A heatmap (sns.heatmap()) is used to visualize feature correlations.

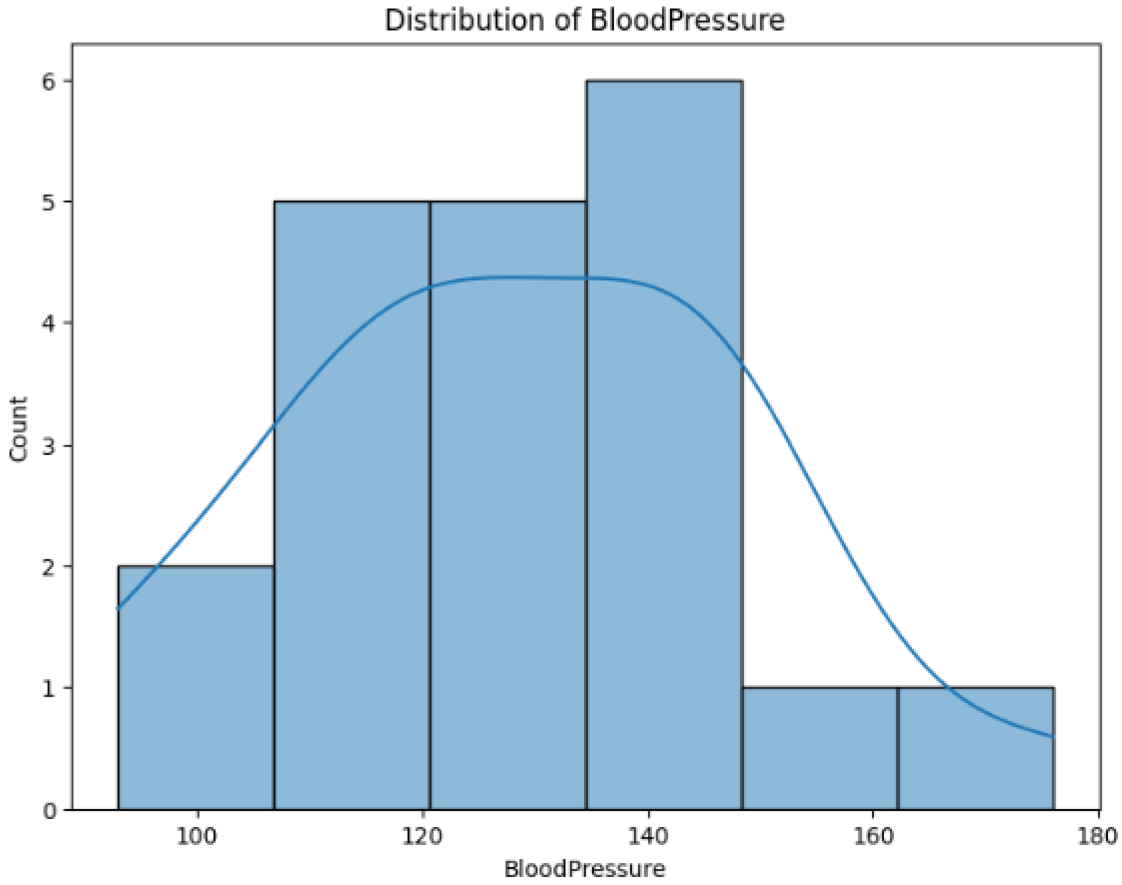
Code

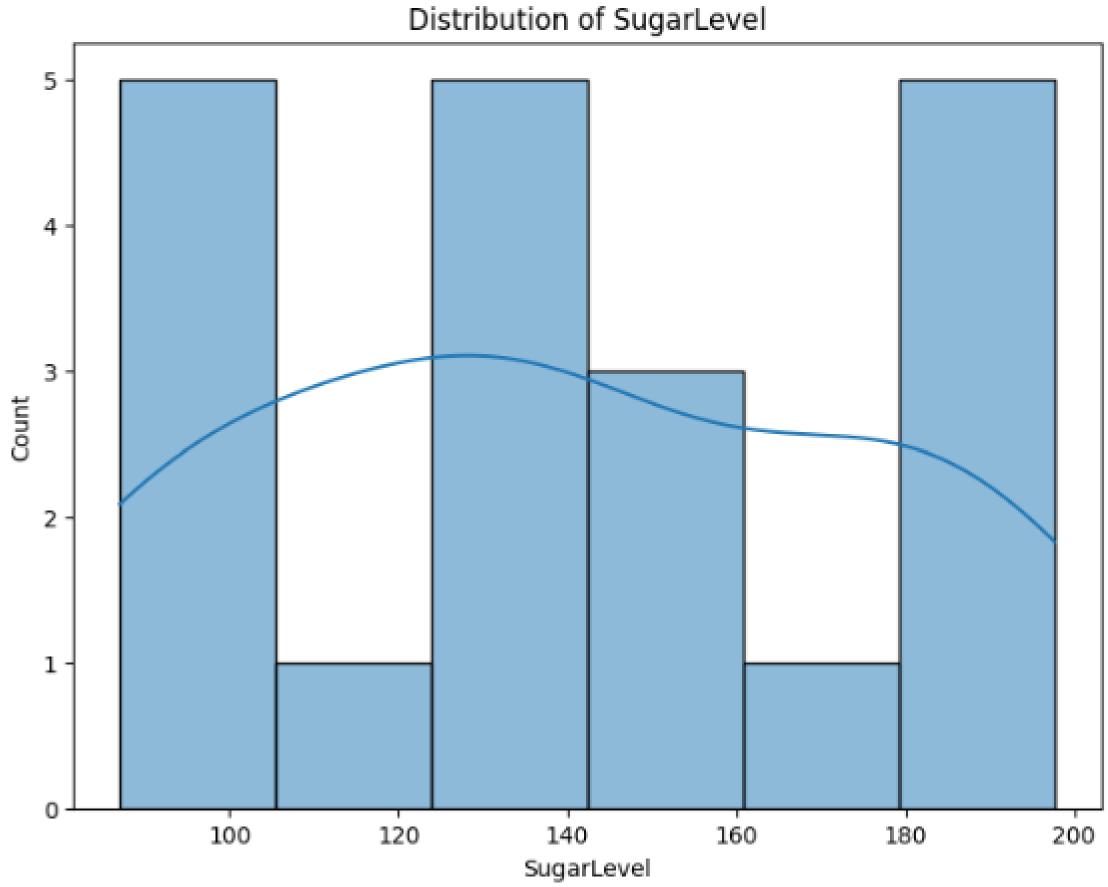
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sys

# Load the dataset
file_path = '/content/healthcare_data (1).csv'
try:
    df = pd.read_csv(file_path)
except FileNotFoundError:
    print(f"Error: '{file_path}' not found. Please upload the file or provide the correct path.")
    sys.exit(1)

# Data Overview
print(df.head())
print(df.info())
print(df.describe())

# Missing Values
print("Missing values per column:\n", df.isnull().sum())

# Visualizing missing values
plt.figure(figsize=(10, 6))
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.title('Missing Values Heatmap')
plt.show()

# Handle missing values
for col in df.select_dtypes(include=['number']).columns:
    if df[col].isnull().all():
        df.drop(columns=[col], inplace=True)
    elif df[col].isnull().any():
        df[col] = df[col].fillna(df[col].mean())

# Data Distribution - Numerical Features
for col in df.select_dtypes(include=['number']).columns:
    plt.figure(figsize=(8, 6))
    sns.histplot(df[col], kde=True, bins=30)
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel("Count")
    plt.show()

# Categorical Features
for col in df.select_dtypes(include=['object', 'category']).columns:
    if df[col].nunique() > 20:
        print(f"Skipping {col}, too many unique categories: {df[col].nunique()}")
        continue
    plt.figure(figsize=(8, 6))
    sns.countplot(x=col, data=df, order=df[col].value_counts().index)
    plt.title(f'Distribution of {col}')
    plt.xticks(rotation=45, ha='right')
    plt.show()

# Correlation Analysis
correlation_matrix = df.corr(numeric_only=True)
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f",
linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()
```

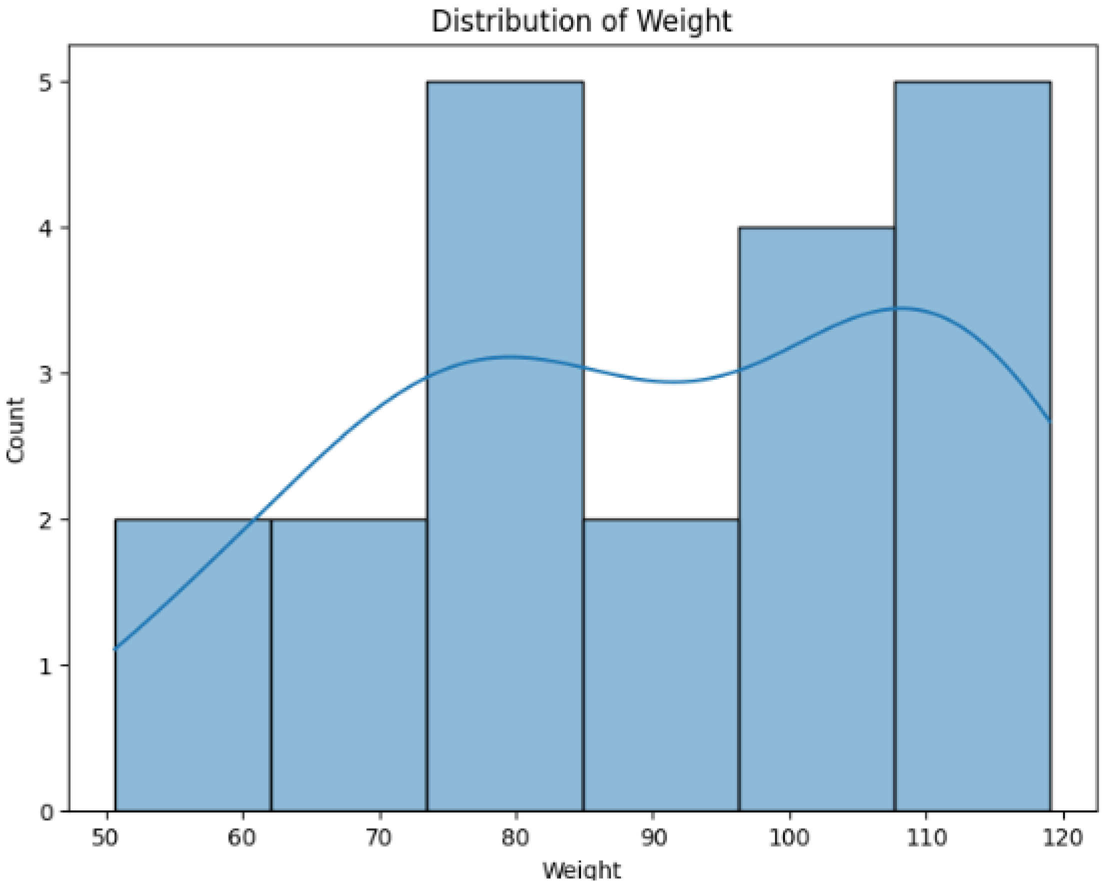
Output/Result

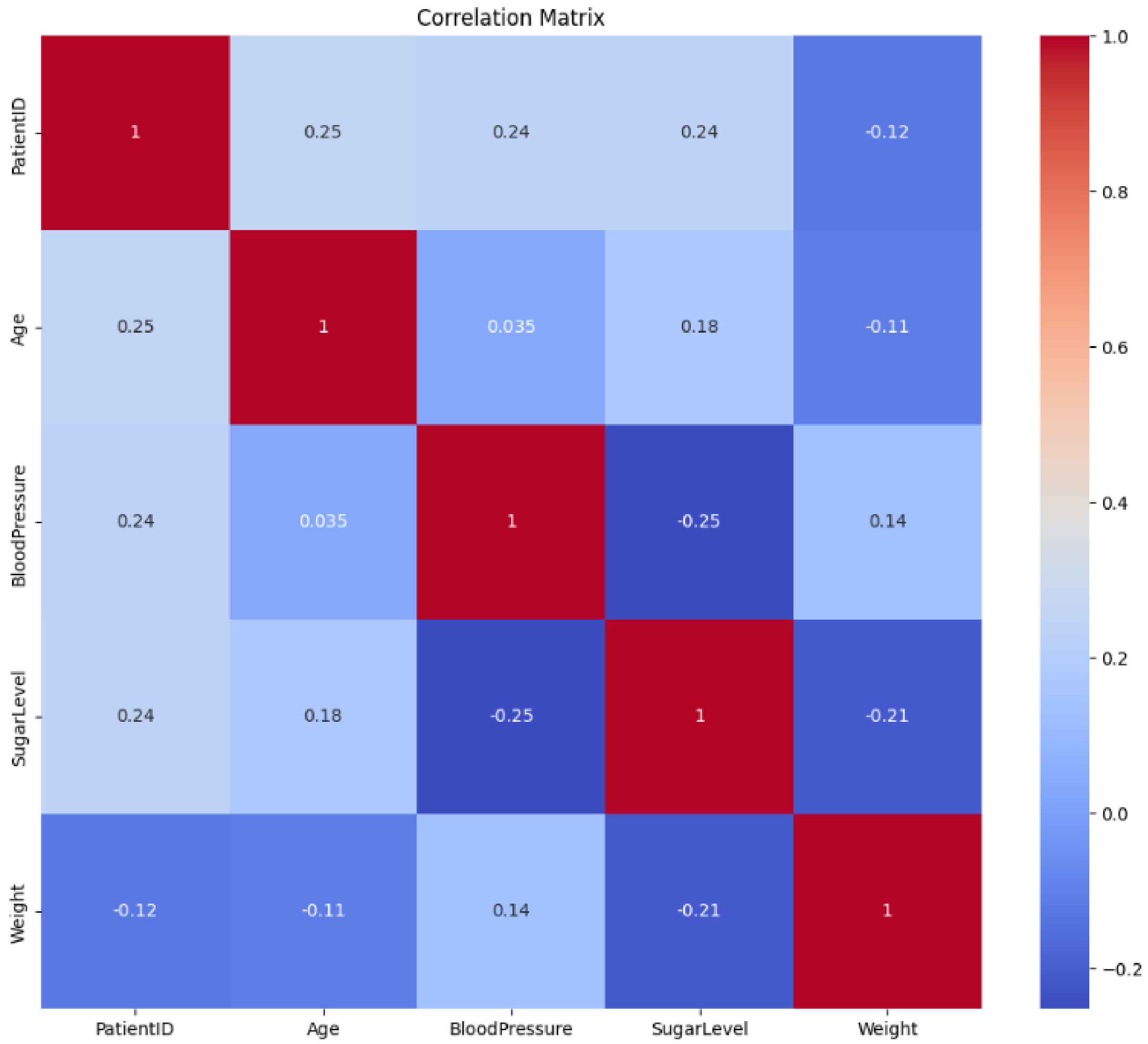

Distribution of PatientID
Count
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
PatientID


Distribution of Age
Count
5
4
3
2
1
0
20
30
40
50
60
70
Age


Distribution of BloodPressure
Count
0
1
2
3
4
5
6
100
120
140
160
180
BloodPressure


Distribution of SugarLevel
Count
0
1
2
3
4
5
100
120
140
160
180
200
SugarLevel


Distribution of Weight
Count
0
1
2
3
4
5
50
60
70
80
90
100
110
120
Weight



References/Credits

- Dataset: [Provided]
- Libraries used: pandas, matplotlib, seaborn
- Concept References: Data science tutorials, online documentation