Healthcare Data Exploration

Name: Tanishk Gupta

Roll No: 202401100300259

Class: CSEAI-D

Course: Introduction to AI

1. Introduction

Healthcare data often has mistakes, missing values, and duplicate records, which can make analysis difficult. Before using this data for machine learning or drawing conclusions, it is important to clean and organize it properly. This report explains a step-by-step process for exploring healthcare data using Python. It focuses on identifying and fixing errors, removing unnecessary data, and understanding patterns within the dataset. Additionally, it uses charts and graphs to visualize key trends, making the data easier to analyze and use for future decision-making.

2. Dataset Overview

The dataset used in this analysis contains information related to healthcare, such as patient details, diagnoses, treatments, and other medical records. The primary objective is to examine its structure, identify inconsistencies, and determine necessary preprocessing steps. The analysis includes:

• Checking for missing values
• Identifying duplicate records
• Understanding data types and distributions
• Summarizing key numerical attributes

3. Data Cleaning

Data cleaning is an essential step in preparing the dataset for further analysis. The program executes the following cleaning tasks:

• Identifies missing values and handles them by either filling them with appropriate values (mean/median/mode) or removing incomplete records.
• Detects and removes duplicate entries to prevent biased analysis.
• Ensures data consistency by standardizing formats and correcting any discrepancies.
These steps ensure that the dataset remains structured, accurate, and reliable for further processing.

4. Data Exploration & Analysis

Once the dataset is cleaned, exploratory data analysis (EDA) is performed to derive meaningful insights. This includes:

• Generating descriptive statistics to understand the distribution of numerical variables.
• Visualizing key attributes using histograms to observe frequency distributions.
• Creating correlation heatmaps to identify relationships between different features, which can help in understanding how variables influence each other.
• Identifying patterns and anomalies that may affect future predictions or decision-making processes.

5. Conclusion

This exploratory analysis demonstrates a systematic approach to cleaning and analyzing healthcare data. By removing inconsistencies and visualizing trends, the dataset becomes more suitable for predictive modeling and decision-making in healthcare applications. Such techniques are fundamental in medical data processing, ensuring better accuracy in diagnoses and treatments when combined with advanced analytics and machine learning models.

CODE

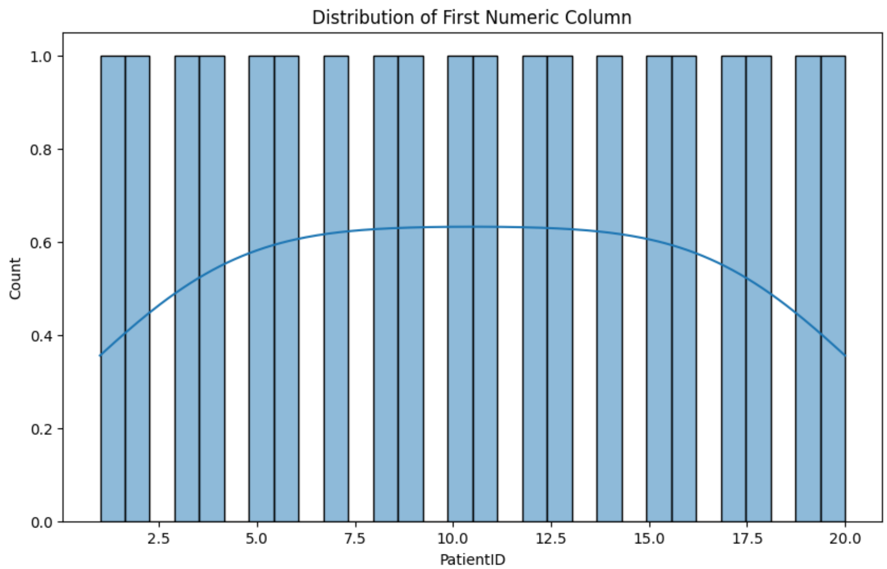
```
#Import library
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Step 1: Load the dataset
file_path = '/content/healthcare_data.csv'  #
Update if needed
df = pd.read_csv(file_path)

# Step 2: Display basic dataset information
print("Basic Information About the Dataset:")
df.info()
print("\nFirst 5 rows:")
df.head()

# Step 3: Checking for missing values
print("\nMissing Values in Each Column:")
df.isnull().sum()

# Step 4: Handling missing values (Option: fill
with mean/median/mode or drop)
df_cleaned = df.dropna()  # Dropping rows with
missing values
print("\nDataset after removing missing
values:")
df_cleaned.info()
```

```
Dataset after removing missing values:
<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
```

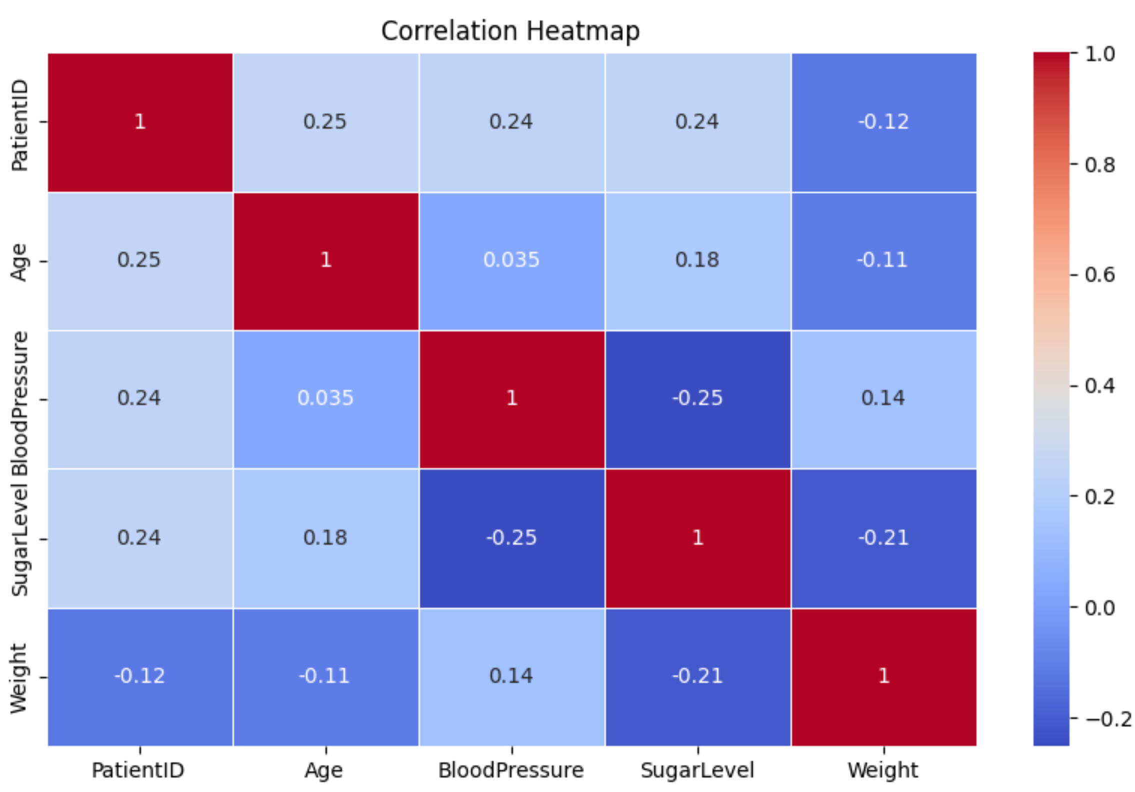
```
# Step 5: Checking for duplicate entries
duplicates = df_cleaned.duplicated().sum()
print(f"\nNumber of duplicate rows: {duplicates}")
df_cleaned = df_cleaned.drop_duplicates()
```

```
Statistical Summary of Numerical Columns:
       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
```

```
# Step 7: Visualizing key insights
plt.figure(figsize=(10, 6))
sns.histplot(df_cleaned.select_dtypes(include=['number']).iloc[:, 0], bins=30, kde=True)
plt.title("Distribution of First Numeric Column")
plt.show()
```




```python
# Step 8: Correlation Heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(df_cleaned.corr(), annot=True,
cmap='coolwarm', linewidths=0.5)
plt.title("Correlation Heatmap")
plt.show()
```



References

The codes and methodologies used in this report were generated with the assistance of AI-based tools such as **ChatGPT** and other AI-powered platforms. These tools helped in data exploration, cleaning, and visualization techniques. Additional information and best practices were sourced from publicly available documentation and standard data science practices.