# Healthcare Data Cleaning

**Project Title**: AI-Based Healthcare Data Cleaning and Visualization
**Student Name**: Anushka Rajput
**Roll Number**: 59
**Course**:Computer Science and Engineering(Artificial Intelligence)
**Date**: 11-03-2025

---

**Introduction**

The AI-Based Healthcare Data Cleaning and Visualization project focuses on processing and analyzing healthcare data to identify missing values, perform data cleaning, and generate insightful visualizations. The project utilizes Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn to preprocess and explore the dataset effectively. This ensures that the data is clean and ready for further analysis or machine learning applications.

---

**Methodology**

1. **Data Import**: The dataset is loaded from a CSV file.
2. **Data Inspection**: The structure, missing values, and summary statistics of the dataset are analyzed.
3. **Data Cleaning**:
    - Missing values are identified.
    - Numerical missing values are filled with the column mean.
4. **Visualization**:
    - Histogram plots are generated to understand the distribution of numerical features.
    - A correlation heatmap is created to identify relationships between variables.

---

**Code Implementation**

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

# Uploading the dataset
from google.colab import files
uploaded = files.upload()
```

# Healthcare Data Cleaning

```
# Reading the CSV file
df = pd.read_csv('/content/healthcare_data (2).csv')

# Displaying dataset information
df.info()
print("\nDataset Summary:")
print(df.describe())

# Displaying first few rows
print("First 5 rows of the dataset:")
print(df.head())

# Checking for missing values
print("\nMissing Values Before Cleaning:")
print(df.isnull().sum())

# Handling missing values
df.fillna(df.mean(numeric_only=True), inplace=True)

print("\nMissing Values After Cleaning:")
print(df.isnull().sum())

# Histogram for numerical features
df.hist(figsize=(10, 8), bins=20, edgecolor="black")
plt.suptitle("Distribution of Numerical Features", fontsize=14)
plt.show()

# Correlation heatmap
plt.figure(figsize=(8, 5))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap="coolwarm", linewidths=0.5)
plt.title("Correlation Heatmap")
plt.show()

print("\nData Cleaning & Visualization Completed!")
```

---

**Output/Result**

# Healthcare Data Cleaning

```
df.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
```

```
df.head()
```

| | 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 |

```
[19] print("\nDataset Summary:")
     print(df.describe())
```

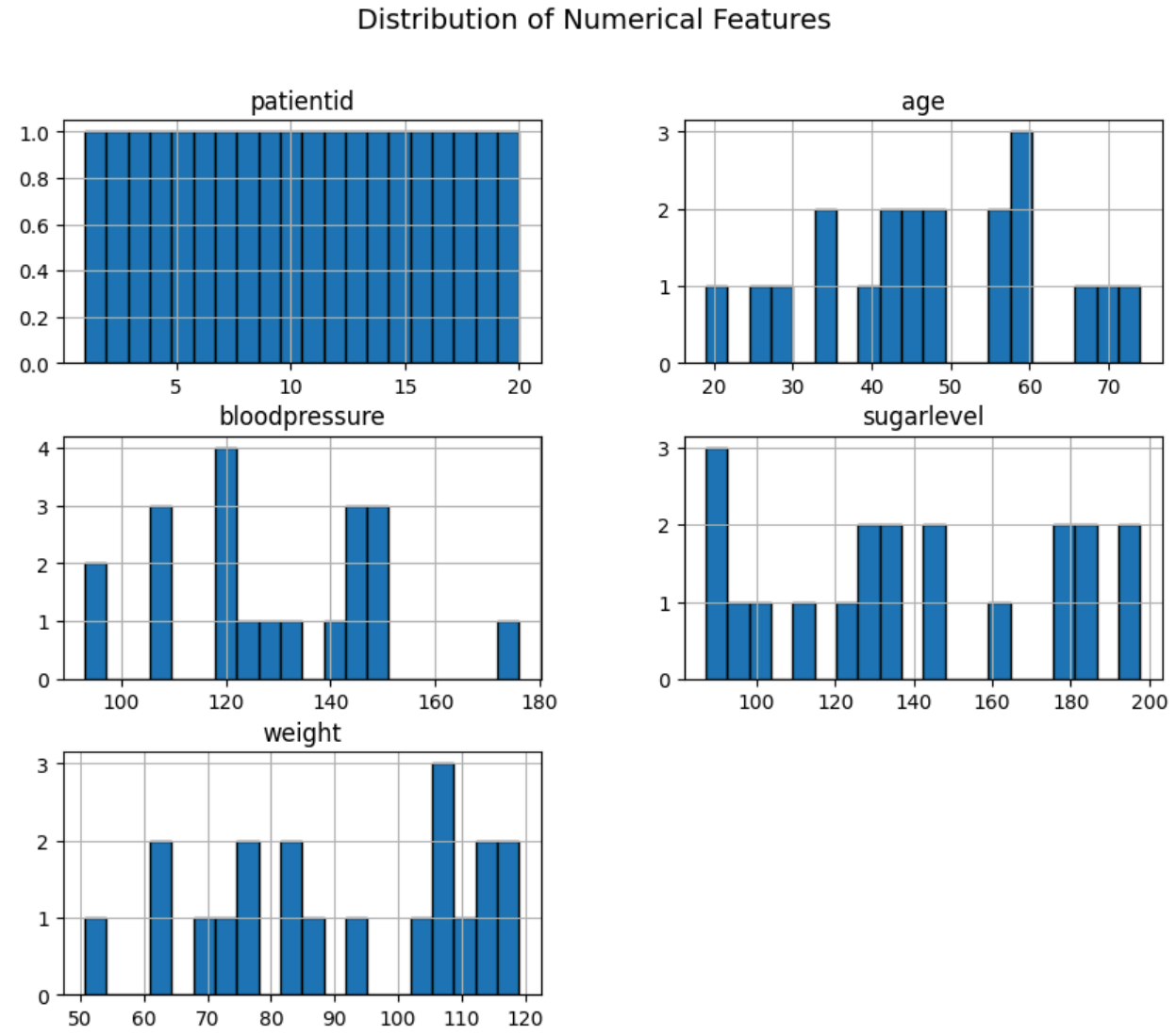
```
Dataset Summary:
       patientid        age  bloodpressure  sugarlevel      weight
count   20.00000  20.000000      20.000000   20.000000   20.000000
mean    10.50000  47.500000     128.650000  139.412236   90.916368
std      5.91608  14.968388      20.893905   37.010795   21.124021
min      1.00000  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
```

# Healthcare Data Cleaning

```
[22] print("First 5 rows of the dataset:")
     print(df.head())
```

```
First 5 rows of the dataset:
   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
```

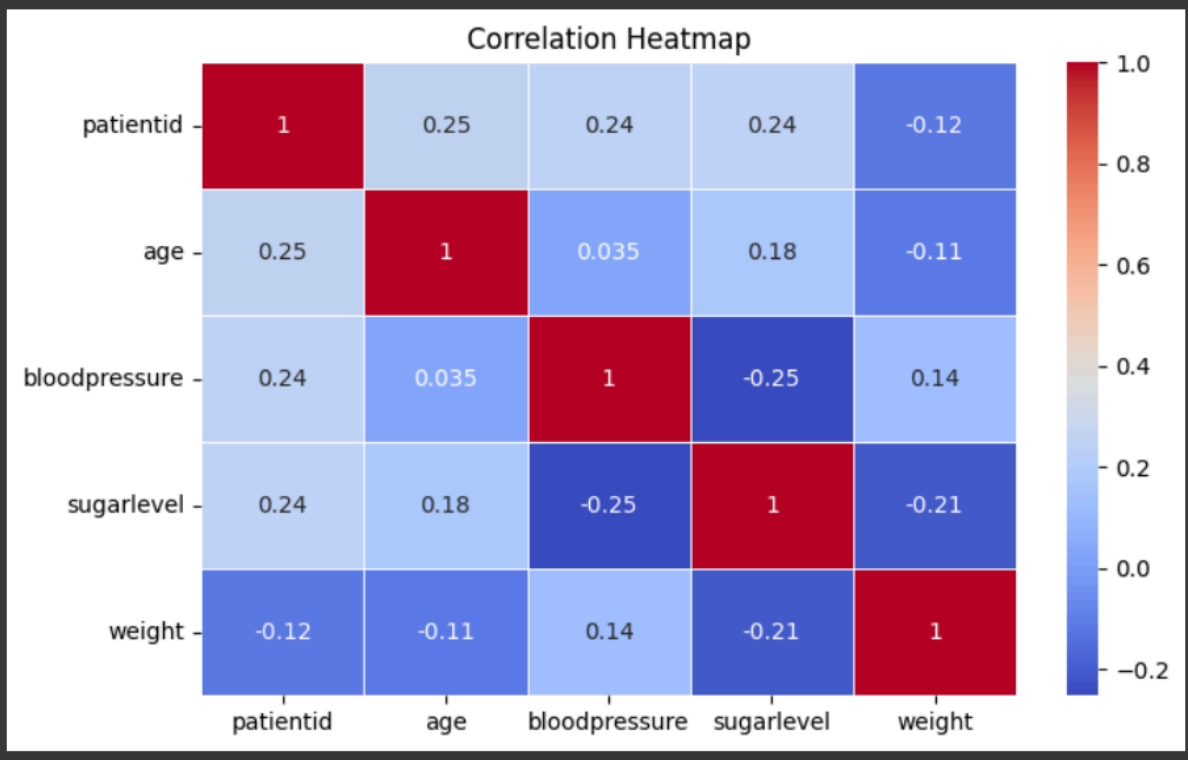
```
df.hist(figsize=(10, 8), bins=20, edgecolor="black")
plt.suptitle("Distribution of Numerical Features", fontsize=14)
plt.show()
```



# Healthcare Data Cleaning

```
plt.figure(figsize=(8, 5))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap="coolwarm", linewidths=0.5)
plt.title("Correlation Heatmap")
plt.show()

print("\nData Cleaning & Visualization Completed! ")
```




---

**References/Credits**

- Dataset Source:"C:\Users\anush\Downloads\healthcare_data.csv"
- Code developed by: Anushka Rajput