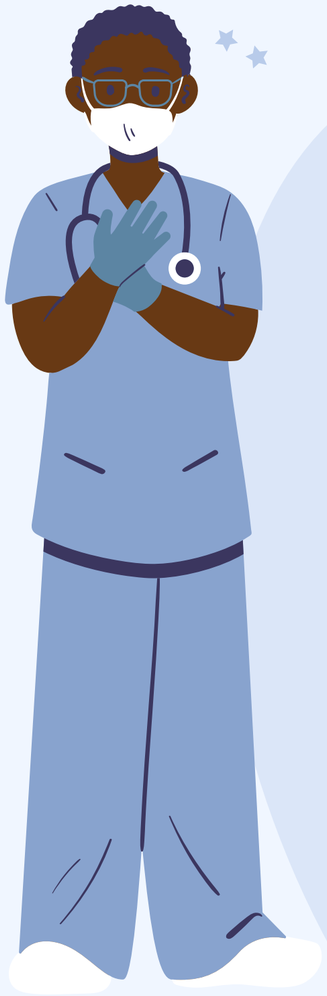
Project

Drug safety and pharmacovigilance

Name: Ayushi

Introduction to the Project



The project explores data related to health metrics like cholesterol levels, sodium-to-potassium ratio, age, and blood pressure.

The goal is to uncover insights that contribute to drug safety and pharmacovigilance by identifying relationships between health variables.

The analysis is performed using Python, a versatile language used extensively in data analysis and signal detection in the pharmaceutical industry.

Importing Required Libraries & reading file

Libraries for Data Analysis

:

- Pandas: For data manipulation and cleaning.
- Matplotlib & Seaborn: For data visualization.
- Numpy: For numerical operations.

```
]: import pandas as pd
   import seaborn as sns
   import matplotlib.pyplot as plt

   data = pd.read_csv('Drug safety and pharmacovigilance.csv')
```

Reading the CSV File:

- The dataset is read using pd.read_csv() to load it into a Pandas DataFrame.
- This step allows you to begin working with the data and apply various analyses.

Dataset Overview

```
print(data.head())
print(data.tail())
print(f"Dataset shape: {data.shape}")
print(data.info())
print(data.isnull().sum())  # Check for missing values
print(data.describe())
```

```
   Age Sex      BP Cholesterol  Na_to_K   Drug
0   23   F    HIGH        HIGH   25.355  DrugY
1   47   M     LOW        HIGH   13.093  drugC
2   47   M     LOW        HIGH   10.114  drugC
3   28   F  NORMAL        HIGH    7.798  drugX
4   61   F     LOW        HIGH   18.043  DrugY
     Age Sex      BP Cholesterol  Na_to_K   Drug
195   56   F     LOW        HIGH   11.567  drugC
196   16   M     LOW        HIGH   12.006  drugC
197   52   M  NORMAL        HIGH    9.894  drugX
198   23   M  NORMAL      NORMAL   14.020  drugX
199   40   F     LOW      NORMAL   11.349  drugX
Dataset shape: (200, 6)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Age          200 non-null    int64
 1   Sex          200 non-null    object
 2   BP           200 non-null    object
 3   Cholesterol  200 non-null    object
 4   Na_to_K      200 non-null    float64
 5   Drug         200 non-null    object
dtypes: float64(1), int64(1), object(4)
memory usage: 9.5+ KB
None
Age            0
Sex            0
BP             0
Cholesterol    0
Na_to_K        0
Drug           0
dtype: int64
                Age     Na_to_K
count  200.000000  200.000000
mean    44.315000   16.084485
std     16.544315    7.223956
min     15.000000    6.269000
25%     31.000000   10.445500
50%     45.000000   13.936500
75%     58.000000   19.380000
```

Dataset Description:

- The dataset contains 200 rows and 6 columns.
- Columns include: Age, Sex, BP (Blood Pressure), Cholesterol, Na_to_K (Sodium-to-Potassium ratio), and Drug.

  - data.tail(): Displays the last few rows of the dataset for a quick preview.
  - data.shape: Provides the dimensions of the dataset (200 rows, 6 columns).
  - data.info(): Displays information about the data types and missing values.
  - data.describe(): Shows summary statistics of numerical columns.

Data Cleaning and Handling Missing Value

- **Checking for Missing Values:**
- **We use data.isnull().sum() to check for any missing values.**

```
print(data.isnull().sum())  # Check for missing values
```

```
Age            0
Sex            0
BP             0
Cholesterol    0
Na_to_K        0
Drug           0
dtype: int64
```

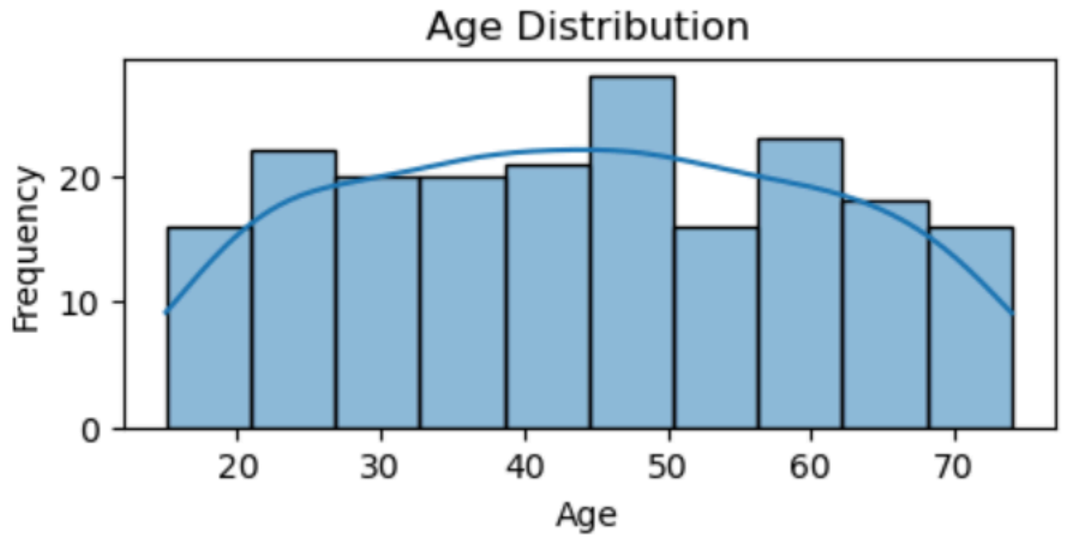
- Handling Missing Data:
  - Missing or NaN values were identified in some columns like Cholesterol.
  - We can choose to either fill or remove the rows with missing values.

Data Visualization - Distribution of Age

Age Distribution:

- A histogram was plotted to understand the distribution of the Age variable.

```
# Age distribution
plt.figure(figsize=(5,2))
sns.histplot(df['Age'], kde=True, bins=10)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



The histogram shows the distribution of age in the dataset, allowing us to identify the most common age ranges.

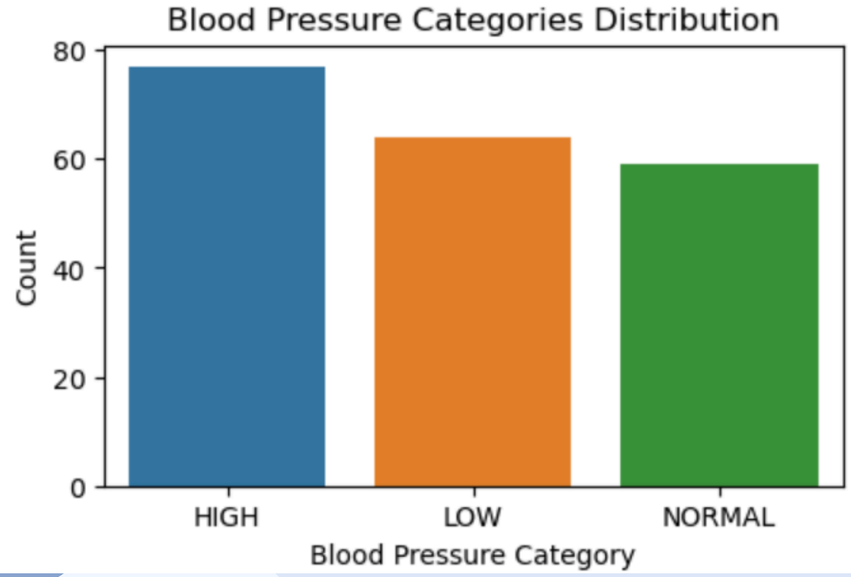
Blood Pressure (BP) Categories

Understanding BP Categories:

- The BP column indicates the blood pressure categories: HIGH, NORMAL, and LOW.
- A countplot was used to visualize the distribution of patients across these categories.

- This countplot shows how many patients fall into each blood pressure category, providing insights into the data distribution.

```
In [74]: # Plot blood pressure categories
         plt.figure(figsize=(5,3))
         sns.countplot(x='BP', data=data)
         plt.title('Blood Pressure Categories Distribution')
         plt.xlabel('Blood Pressure Category')
         plt.ylabel('Count')
         plt.show()
```



Boxplots to Analyze Distribution by Sex (Na to Potassium Ratio)

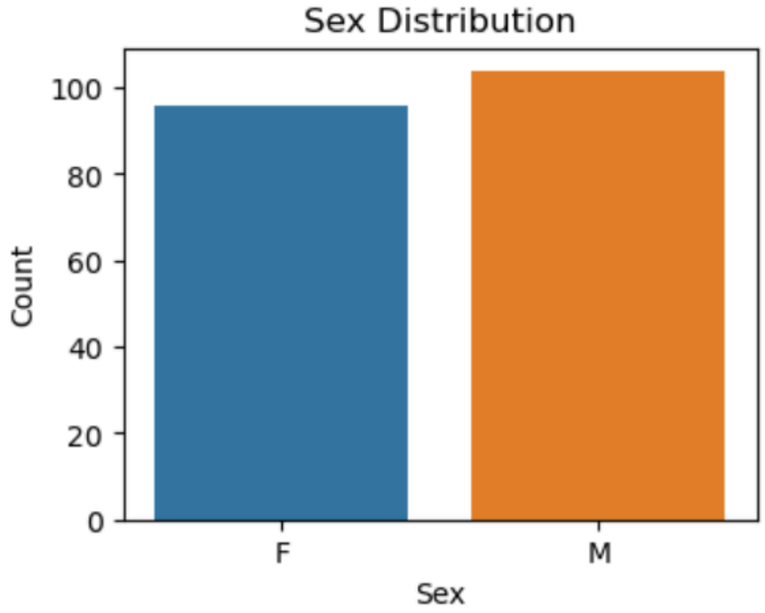
Na to Potassium Ratio by Sex:

This boxplot examines how the Na to Potassium ratio differs between males and females in the dataset.

Plot Details:

- X-axis: Gender (Sex)
- Y-axis: Na to Potassium Ratio (Ratio of Sodium to Potassium levels)
- Key Insights:
- The boxplot displays the distribution of the Na to Potassium ratio for both males and females, showing the median, interquartile range (IQR), and potential outliers.
- This can help detect any gender-based differences in the sodium-to-potassium balance, which may be useful in pharmacovigilance for understanding how gender influences drug safety outcomes.

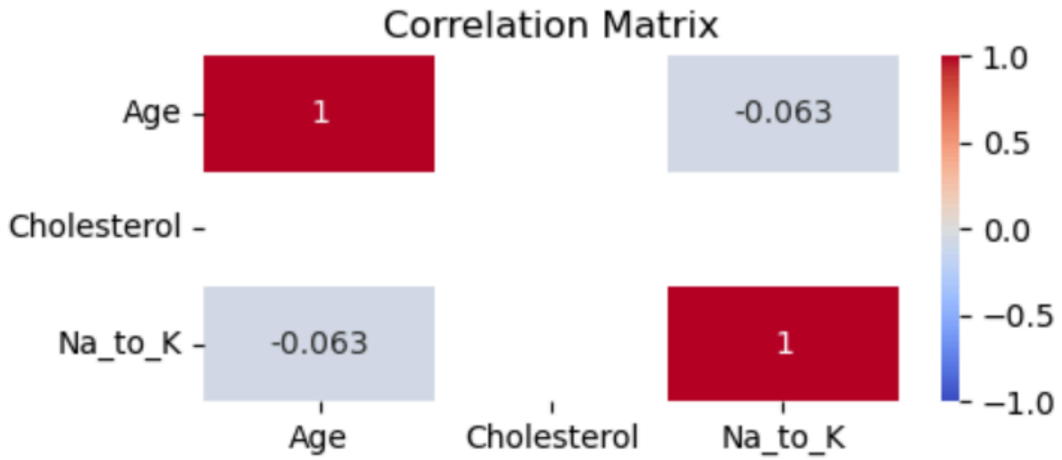
```
# Sex distribution
plt.figure(figsize=(4,3))
sns.countplot(x='Sex', data=df)
plt.title('Sex Distribution')
plt.xlabel('Sex')
plt.ylabel('Count')
plt.show()
```



Correlation Analysis

```
n [81]: # Calculate the correlation matrix using the correct column names
corr_matrix = df[['Age', 'Cholesterol', 'Na_to_K']].corr()

plt.figure(figsize=(5,2))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Correlation Matrix')
plt.show()
```



Correlation Matrix:

A correlation matrix was plotted to see the relationships between age, cholesterol, and sodium-to-potassium ratio.

The heatmap helps us understand how strongly the numerical variables are correlated with each other, which is essential for identifying potential factors impacting drug safety.

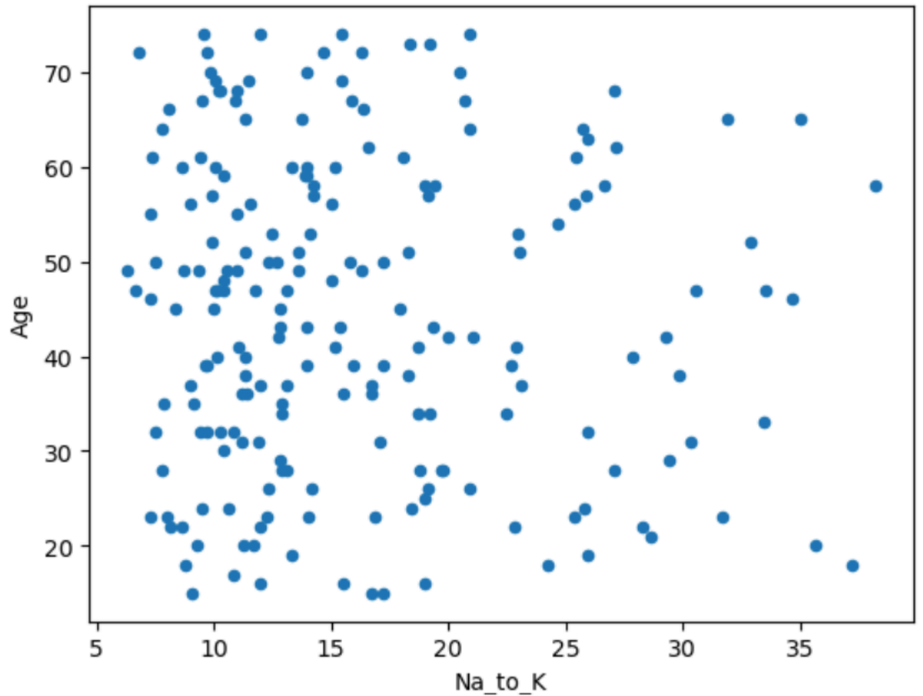
Scatter Plot - Na/K Ratio vs Age

Scatter Plot:

A scatter plot was created to check the relationship between the sodium-to-potassium ratio and age.

- This scatter plot helps visualize the relationship between age and sodium-to-potassium ratio, indicating if there's any noticeable trend or outliers.

```
In [54]: df.plot.scatter(x='Na_to_K',y='Age') # to check variable scatter with respect to other variable
Out[54]: <AxesSubplot:xlabel='Na_to_K', ylabel='Age'>
```



Conclusion

Key Findings:

- The dataset provides important insights into cholesterol levels, blood pressure, and sodium-to-potassium ratio across different age groups and sexes.
- Visualizations such as histograms, boxplots, and scatter plots have been effective in understanding the distribution and relationships between variables.
- These insights can contribute to further analysis in drug safety and pharmacovigilance by identifying potential risk factors.

Thank you for your attention