**Exploratory Data Analysis**

**Importing Libraries**

#%%

import pandas as pd

import plotly.express as px

import matplotlib.pyplot as plt

import seaborn as sns

**DATA\_GATHERING**

#%% Load the data

drugs\_data = pd.read\_csv(r'D:\Datavs\drug200 - drug200.csv.csv')

**Basic data summaries**

#%%

# Get the head, tail, and describe summaries

head\_data = drugs\_data.head()

tail\_data = drugs\_data.tail()

describe\_data = drugs\_data.describe()

object\_data = drugs\_data.describe(include='object')

**PRINT\_SUMMARIES, DTYPES & COUNTS**

#%%

print("Head Data:\n", head\_data)

print("\nTail Data:\n", tail\_data)

print("\nDescribe Data:\n", describe\_data)

print("\nObject Data:\n", object\_data)

# **Print data types**

dtypes\_df = drugs\_data.dtypes.reset\_index()

dtypes\_df.columns = ['Column', 'Data Type']

print("\nData Types:\n", dtypes\_df)

# **Print value counts for categorical features**

print("\nSex Value Counts:\n", drugs\_data['Sex'].value\_counts())

print("\nBP Value Counts:\n", drugs\_data['BP'].value\_counts())

print("\nCholesterol Value Counts:\n", drugs\_data['Cholesterol'].value\_counts())

print("\nDrug Value Counts:\n", drugs\_data['Drug'].value\_counts())

**CHECK THE MISSING VALUES**

#%%  **Display the count of null values for each column**

print("\nCheck for Missing Values:\n", drugs\_data.isnull().sum())

**NUMERICAL AND CATEGORICAL FEATURES**

#%% **Print counts of numerical and categorical features**

numerical\_features = drugs\_data.select\_dtypes(include=['int64', 'float64'])

categorical\_features = drugs\_data.select\_dtypes(include=['object'])

print("\nCount of Numerical and Categorical Features")

print(f'Numerical Features Count: {numerical\_features.shape[1]}')

print(f'Categorical Features Count: {categorical\_features.shape[1]}')

**CHECK OUTLIERS**

#%% **Outlier detection function and outlier count**

def count\_outliers(column):

    Q1 = column.quantile(0.25)

    Q3 = column.quantile(0.75)

    IQR = Q3 - Q1

    lower\_bound = Q1 - 1.5 \* IQR

    upper\_bound = Q3 + 1.5 \* IQR

    outliers = column[(column < lower\_bound) | (column > upper\_bound)]

    return len(outliers)

#%% **Count outliers in numerical columns**

numerical\_df = drugs\_data.select\_dtypes(include=['number'])

outlier\_counts = numerical\_df.apply(count\_outliers)

#%% **Sort outliers by count in descending order**

sorted\_outliers = outlier\_counts.sort\_values(ascending=False)

print("\nOutlier counts per numerical feature:\n", sorted\_outliers)

**//Outlier counts per numerical feature:**

Na\_to\_K 8

Age 0

dtype: int64

**DISTRIBUTION AND VISUALIZATION**

# %%

plt.figure(figsize=(20, 120))

for n, feature in enumerate(categorical\_features):

    plt.subplot(22, 2, n + 1)

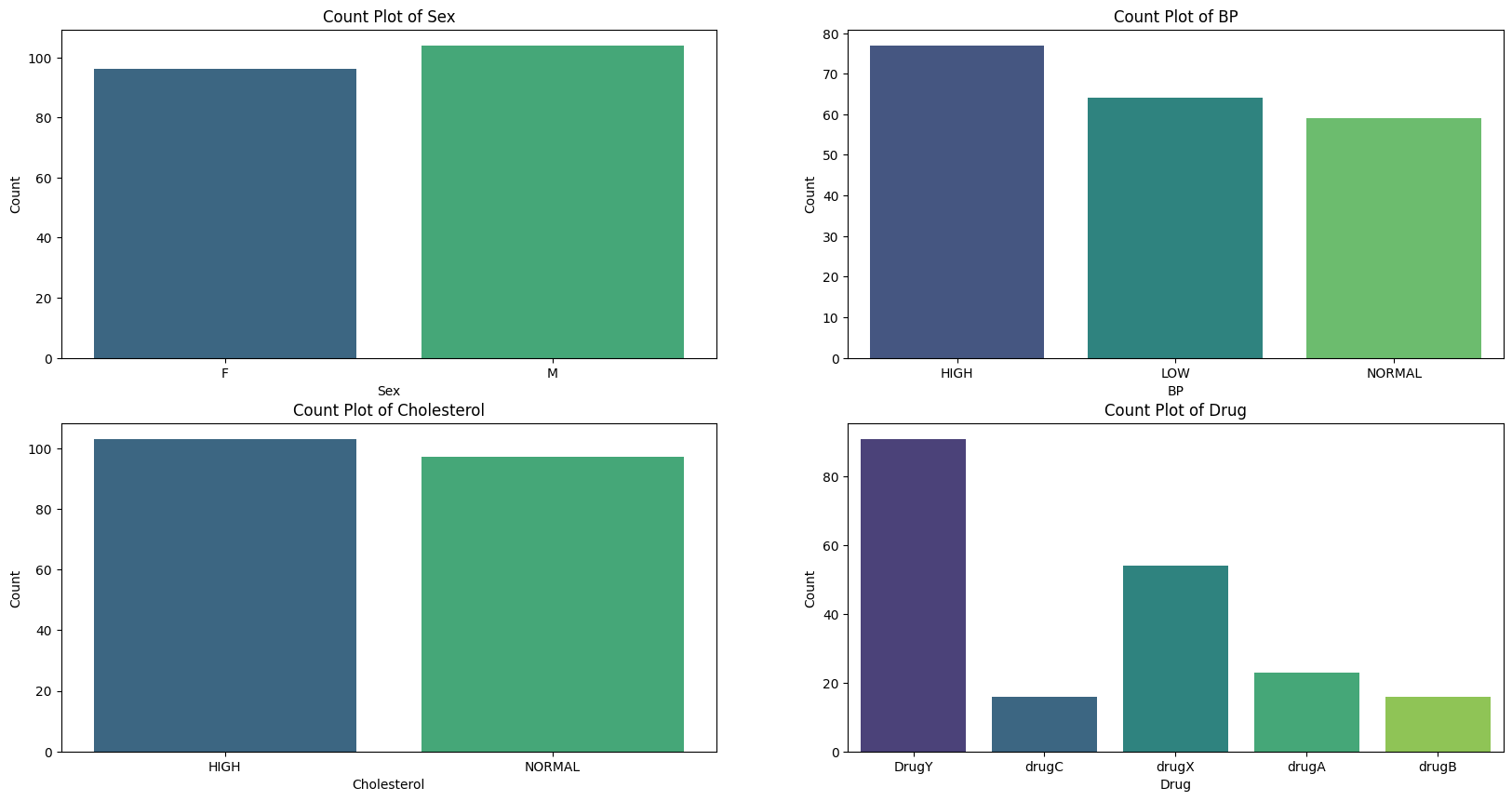
    sns.countplot(x=feature, data=drugs\_data, palette='viridis')

    plt.title(f'Count Plot of {feature}')

    plt.xlabel(feature)

    plt.ylabel('Count')

plt.show()



These count plots summarize the distribution of key categorical variables in the dataset:

1. **Sex**: The distribution between females (F) and males (M) is balanced, with a slight majority of males.
2. **Blood Pressure (BP)**: "HIGH" BP is most common, followed by "LOW" and then "NORMAL." This imbalance suggests a prevalence of high BP in the dataset.
3. **Cholesterol**: "HIGH" cholesterol is slightly more prevalent than "NORMAL," indicating a skew toward higher cholesterol levels.
4. **Drug**: "DrugY" is the most frequently used, followed by "drugX," while "drugC," "drugA," and "drugB" have lower counts.

These distributions highlight potential health trends and imbalances, suggesting that further analysis may need to address class imbalances for accurate modeling.

# %%  ***Creating Distribution of Numerical Features***

for col in numerical\_features:

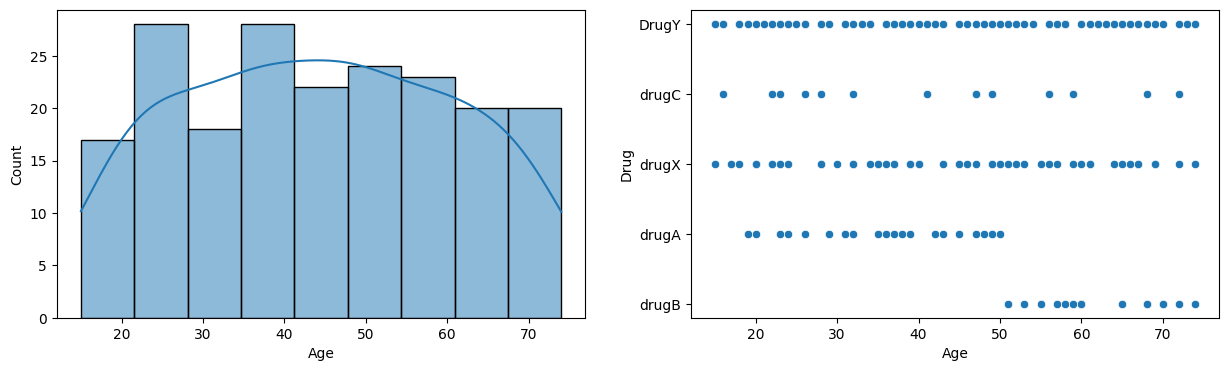
    plt.figure(figsize=(15, 4))

    plt.subplot(1, 2, 1)

    sns.histplot(drugs\_data[col], kde = True)

    plt.subplot(1, 2, 2)

    sns.scatterplot(data = drugs\_data, x = col, y='Drug')



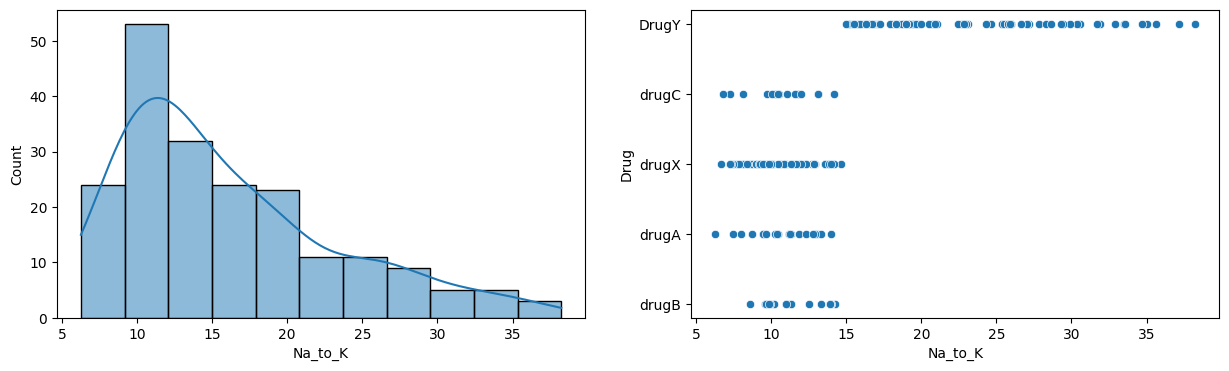
The plots suggest a **categorical relationship** between **Age** and **Drug type**, giving insight into the age distribution for each drug type, but they don't show a direct numerical correlation (like you would see with two continuous variables). Here’s what each plot implies about the relationship:

**Age Distribution  
The histogram shows a fairly balanced age distribution across different age groups, indicating that all age groups are well represented in the data.**

**Age vs. Drug Type (Dot Plot)  
The dot plot reveals age-based preferences for drug types:**

* **DrugY spans all age groups, suggesting it’s widely prescribed.**
* **DrugA is more common among middle-aged individuals (40-60).**
* **DrugB is mainly used by older individuals, implying it may be targeted for this age group.**
* **DrugC and DrugX appear across age groups but with narrower ranges than DrugY.**

**These plots suggest potential age-specific preferences for certain drugs, though a chi-square test could further validate this association.**



**Histogram of Na\_to\_K Levels**  
The histogram shows a right-skewed distribution of Na\_to\_K (sodium to potassium ratio) levels, with most values between 10 and 15, peaking around 10. Higher Na\_to\_K levels are less common in the dataset.

**Na\_to\_K vs. Drug Type (Dot Plot)**  
The dot plot illustrates the association between Na\_to\_K levels and drug types. **DrugY** is prescribed across a wide Na\_to\_K range, especially at higher levels, while **DrugA**, **DrugX**, and **DrugC** cluster around lower levels (10-20). **DrugB** is almost exclusively linked to the lowest Na\_to\_K values (5-10). This pattern suggests specific drugs are preferred at particular Na\_to\_K levels, potentially due to their effectiveness in different Na\_to\_K conditions.

Here is a summary of the key insights and relationships uncovered during the exploratory data analysis (EDA):

**1. Distribution Insights**

* **Age Distribution**: The data shows a fairly balanced distribution across age groups, meaning that no single age group dominates the dataset. This allows us to explore drug prescriptions and Na\_to\_K ratios across a wide age range.
* **Na\_to\_K Ratio Distribution**: The Na\_to\_K ratio is right-skewed, with most values concentrated in the lower range (10-15). This indicates that a higher Na\_to\_K ratio is less common, which might be important for understanding dosage or health conditions related to electrolyte levels.

**2. Drug Type vs. Age**

* **DrugY** is prescribed across all age groups, making it a broadly applicable drug in this dataset. Its wide distribution may suggest general applicability across varying health conditions.
* **DrugA** is more common among middle-aged individuals, indicating it may target conditions prevalent in this age group.
* **DrugB** is associated with older age groups, implying it could be used for age-related health issues.
* **DrugC** and **DrugX** have scattered but narrower age ranges, suggesting limited or specific age-group applicability.

**3. Drug Type vs. Na\_to\_K Ratio**

* **DrugB** is primarily associated with the lowest Na\_to\_K values, hinting that it might be prescribed to patients with lower sodium-to-potassium ratios.
* **DrugY** covers a broader range of Na\_to\_K ratios, particularly in higher values, possibly indicating its use in conditions with elevated Na\_to\_K ratios.
* **Drugs A, C, and X** tend to be clustered around moderate Na\_to\_K values (10-20), suggesting these drugs may target patients with balanced or mildly elevated Na\_to\_K ratios.

**4. Categorical Relationships and Potential Associations**

* The visualization suggests a potential relationship between age groups and specific drug types, with certain drugs being more prevalent in particular age ranges.
* The distribution of Na\_to\_K ratios among drugs suggests that electrolyte levels may influence the type of drug prescribed, which could be related to specific health conditions.

**5. Further Statistical Analysis**

* Visual analysis suggests possible categorical associations (age groups with drug types, Na\_to\_K levels with drug types). To confirm these relationships, further statistical tests (e.g., chi-square tests for categorical data) could quantify these associations.

**Summary**

This EDA reveals that both **age** and **Na\_to\_K levels** may play significant roles in drug prescription patterns. Drugs seem to target specific age groups and Na\_to\_K ranges, indicating age and electrolyte balance as potentially important factors in choosing a suitable drug. These insights could help guide further analysis or support clinical decisions.

Project Link: https://www.dropbox.com/scl/fi/ved6pq5owjlsc4bdgryio/Datavs.zip?rlkey=x3a8lt987wau77acpdri82k1s&st=i2bnppj9&dl=0