Drug safety and pharmacovigilance

1. Objective

The goal is to explore the relationships between patient demographics and health indicators like blood pressure, cholesterol levels, and Na to potassium ratio. This exploration will help identify any patterns or correlations that could be important for drug safety and pharmacovigilance.

2. Dataset Overview

Let's assume you have a dataset with the following columns:

- Age: The age of the patient.
- Sex: The gender of the patient (e.g., Male/Female).
- Blood Pressure Levels (BP): Numerical or categorical values representing blood pressure.
- Cholesterol Levels: Numerical values representing cholesterol levels.
- Na to Potassium Ratio: The ratio of sodium to potassium in the body.

3. Data Loading and Initial Exploration

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import missingno as msno
```

In [44]:
```
df = pd.read_csv(r'C:\Users\user\OneDrive\Desktop\Data internship\Datasets\Proje
df.head(20)
```

Out[44]:

	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
5	22	F	NORMAL	HIGH	8.607	drugX
6	49	F	NORMAL	HIGH	16.275	DrugY
7	41	M	LOW	HIGH	11.037	drugC
8	60	M	NORMAL	HIGH	15.171	DrugY
9	43	M	LOW	NORMAL	19.368	DrugY
10	47	F	LOW	HIGH	11.767	drugC
11	34	F	HIGH	NORMAL	19.199	DrugY
12	43	M	LOW	HIGH	15.376	DrugY
13	74	F	LOW	HIGH	20.942	DrugY
14	50	F	NORMAL	HIGH	12.703	drugX
15	16	F	HIGH	NORMAL	15.516	DrugY
16	69	M	LOW	NORMAL	11.455	drugX
17	43	M	HIGH	HIGH	13.972	drugA
18	23	M	LOW	HIGH	7.298	drugC
19	32	F	HIGH	NORMAL	25.974	DrugY

In [3]:
```
df.shape
```

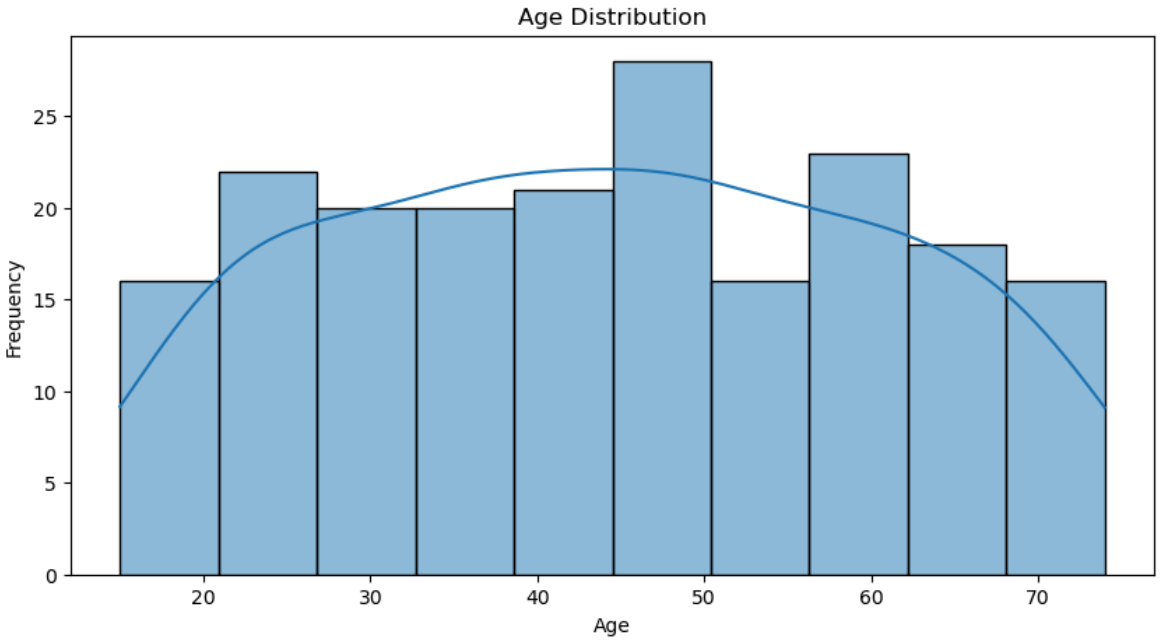
Out[3]:
```
(200, 6)
```

In [4]:
```
df.info()
```

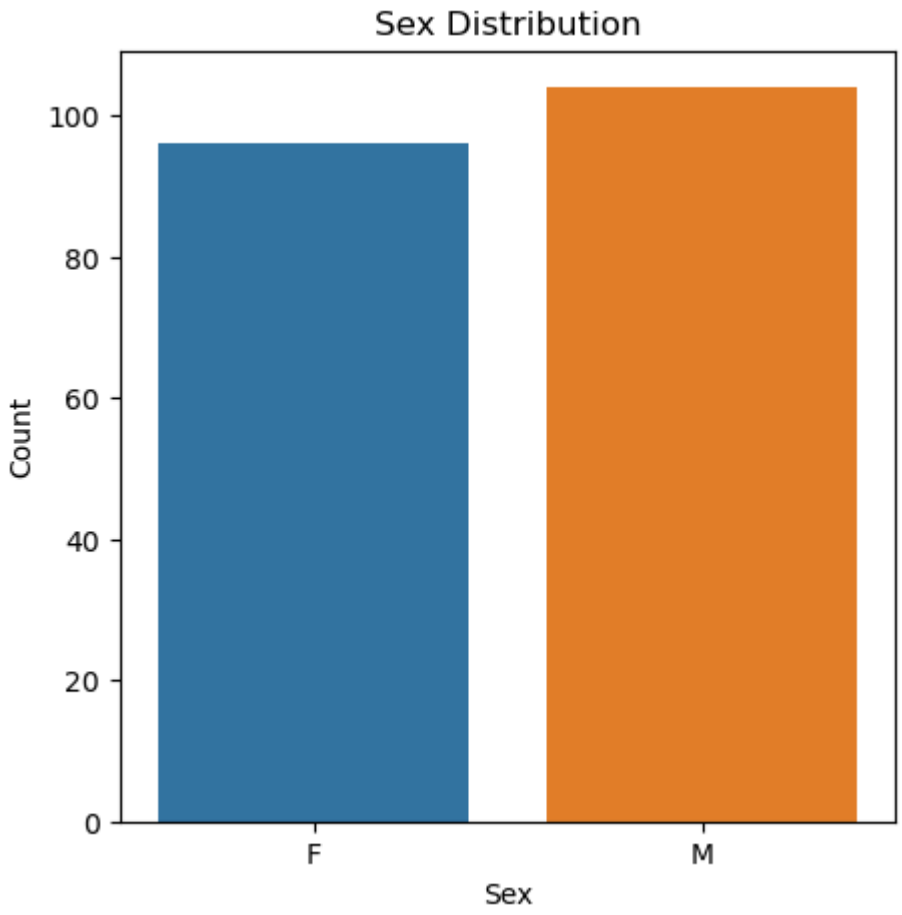
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Age          200 non-null    int64  
 1   Sex          200 non-null    object 
 2   BP           200 non-null    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
```

4. Distribution of Age and Sex:

In [9]:
```
# Age distribution
plt.figure(figsize=(10,5))
sns.histplot(df['Age'], kde=True, bins=10)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

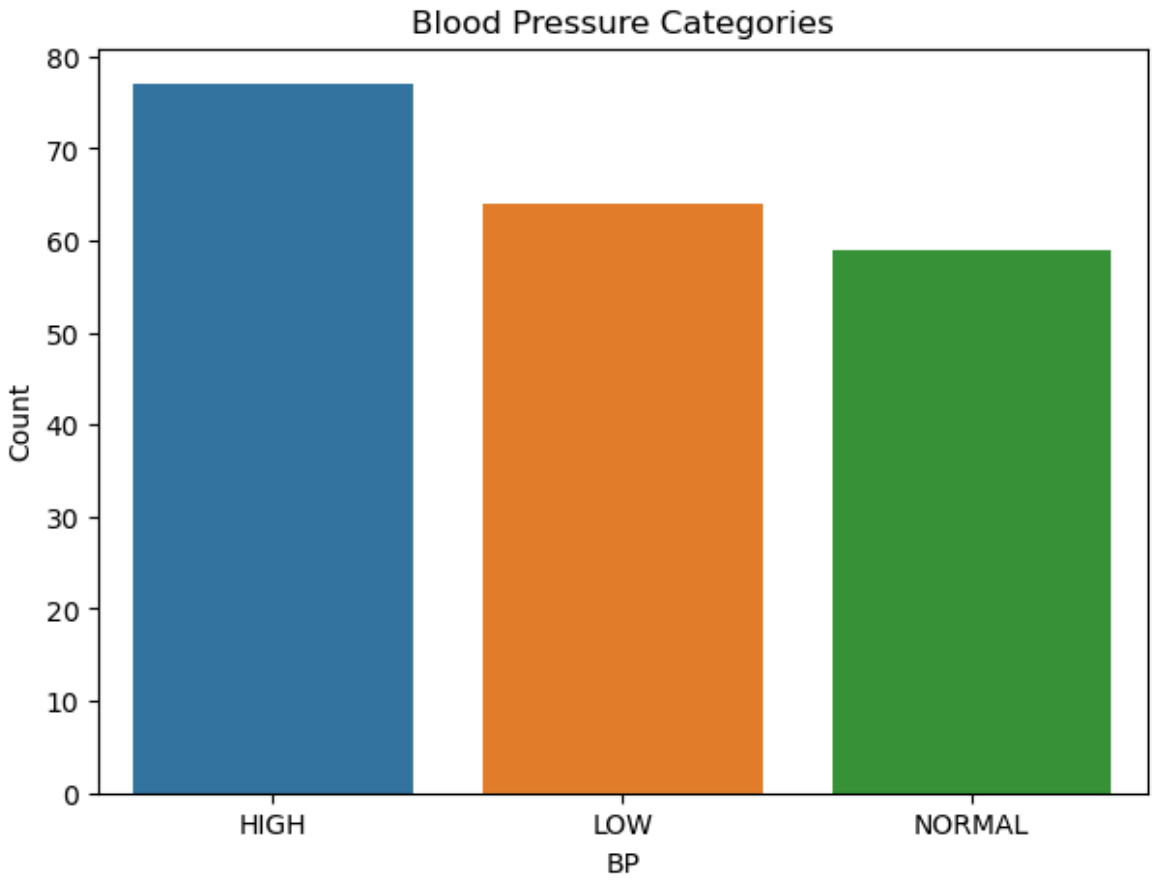


In [10]:
```
# Sex distribution
plt.figure(figsize=(5,5))
sns.countplot(x='Sex', data=df)
plt.title('Sex Distribution')
plt.xlabel('Sex')
plt.ylabel('Count')
plt.show()
```



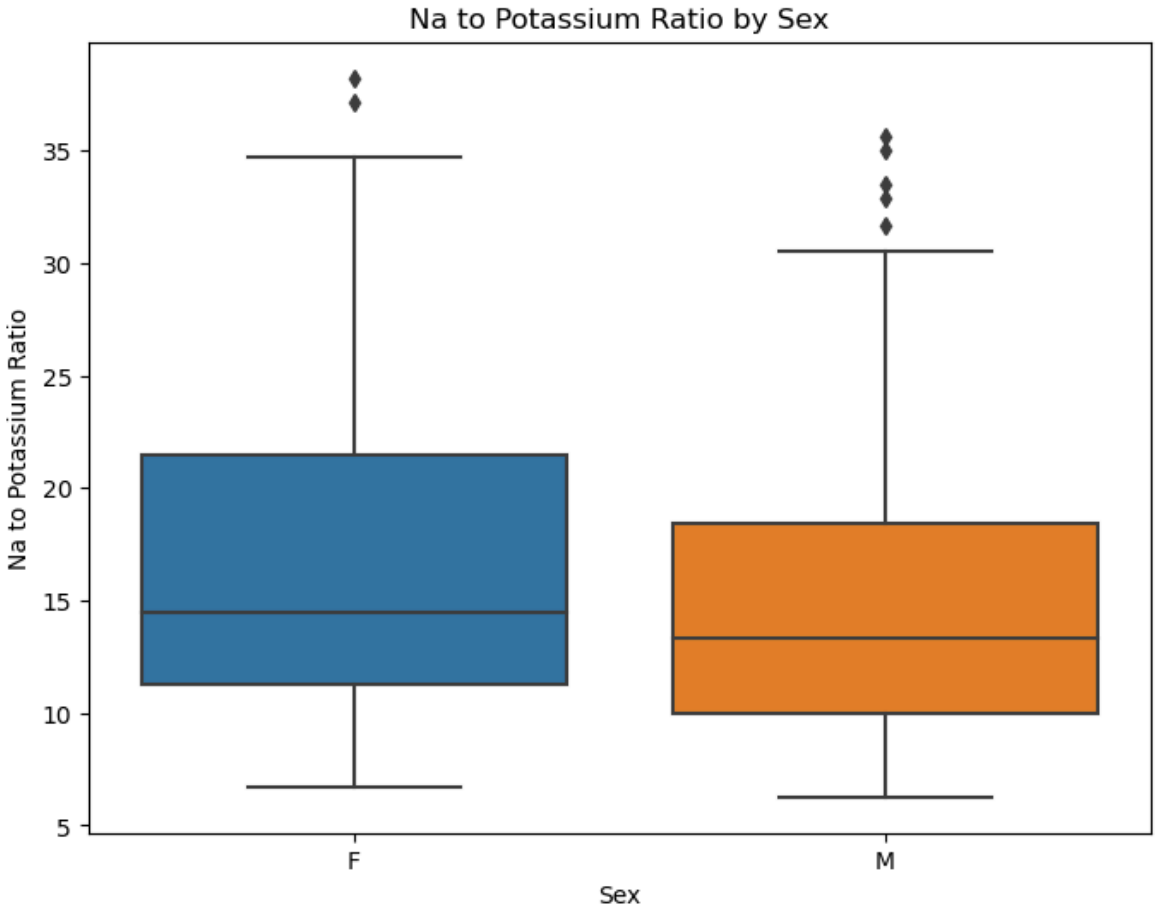
5. Blood Pressure Analysis:

In [23]:
```
# Plot BP categories
plt.figure(figsize=(7,5))
sns.countplot(x='BP', data=df)
plt.title('Blood Pressure Categories')
plt.xlabel('BP')
plt.ylabel('Count')
plt.show()
```

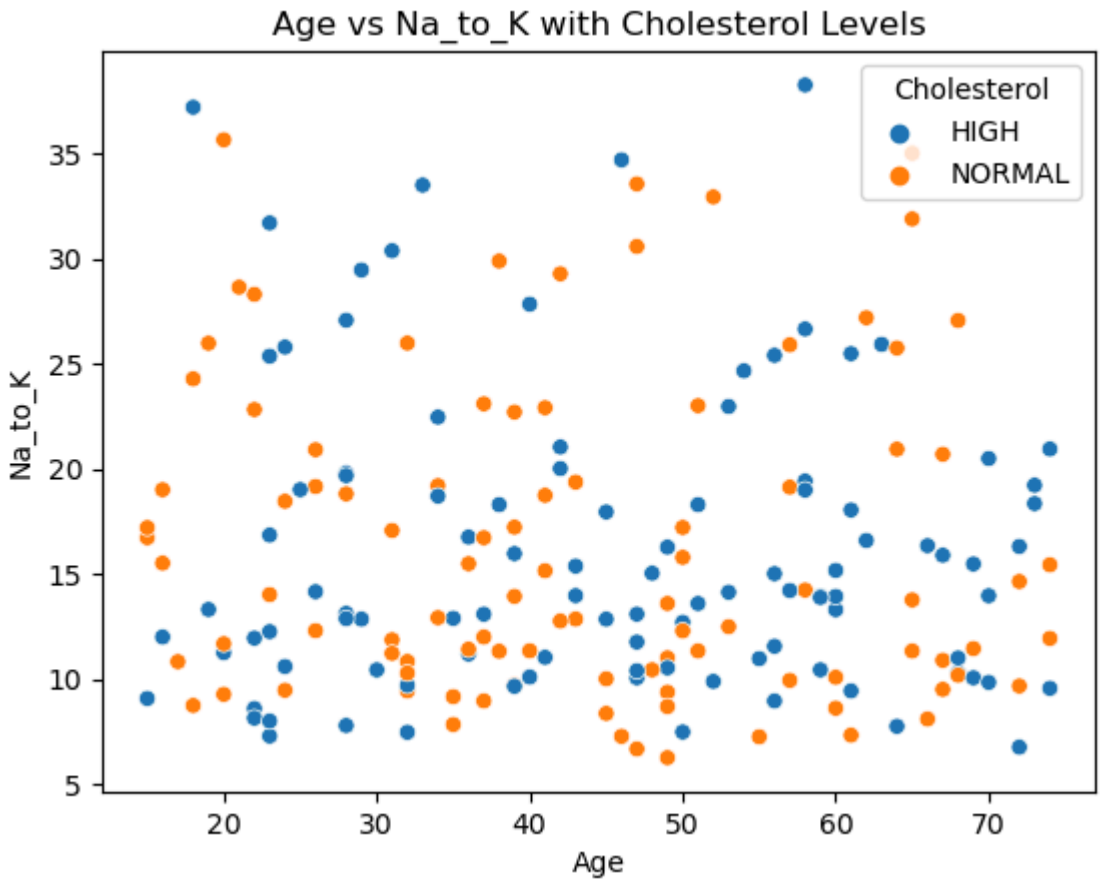


6. Boxplots to Analyze Distribution by Sex:

In [26]:
```
# Na to Potassium Ratio by Sex
plt.figure(figsize=(8,6))
sns.boxplot(x='Sex', y='Na_to_K', data=df)
plt.title('Na to Potassium Ratio by Sex')
plt.xlabel('Sex')
plt.ylabel('Na to Potassium Ratio')
plt.show()
```



In [37]:
```
# Scatter plot
sns.scatterplot(x='Age', y='Na_to_K', hue='Cholesterol', data=df)
plt.title("Age vs Na_to_K with Cholesterol Levels")
plt.show()

# Yaş grupları analizi
df['Age_Group'] = pd.cut(df['Age'], bins=[0, 30, 50, 70, 100], labels=['0-30', '
print(df.groupby(['Age_Group', 'Cholesterol']).size())
```

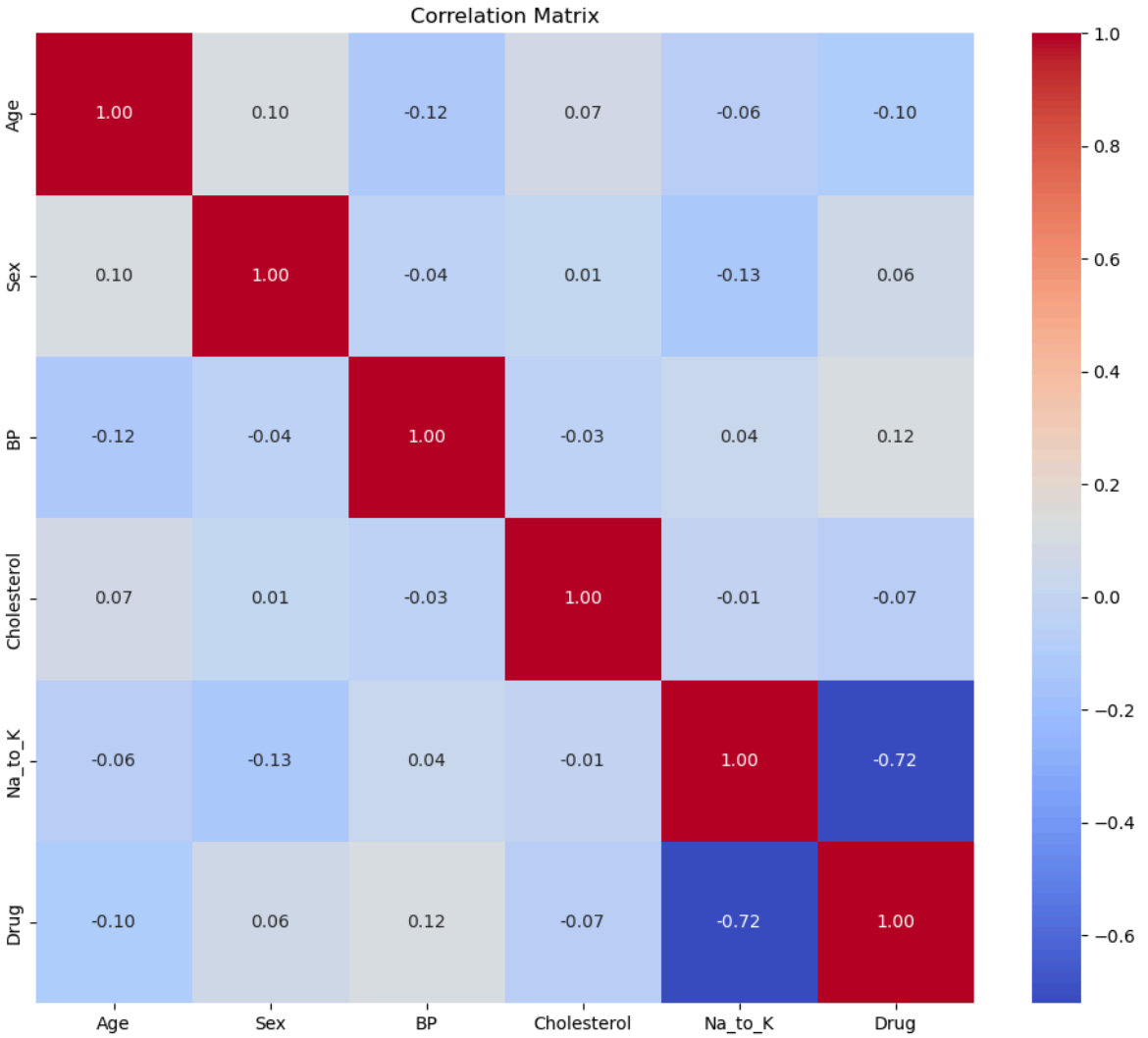



```
Age_Group  Cholesterol
0-30       HIGH           27
           NORMAL         21
30-50      HIGH           33
           NORMAL         46
50-70      HIGH           37
           NORMAL         26
70+        HIGH            6
           NORMAL          4
dtype: int64
```

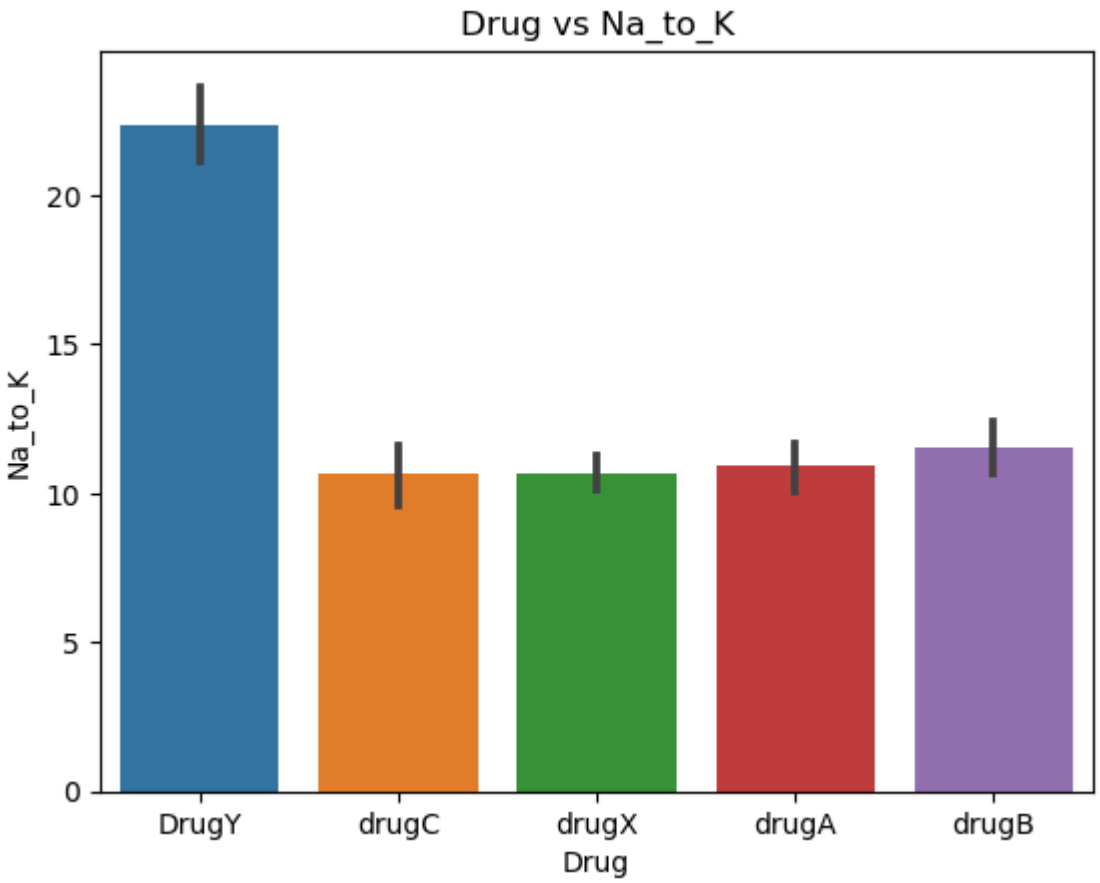
7. Correlation Analysis:

In [46]:
```python
# Convert specific categorical variables to numerical
df_encoded = df.copy()

# Convert BP to numerical
bp_mapping = {"LOW": 0, "NORMAL": 1, "HIGH": 2}
df_encoded["BP"] = df_encoded["BP"].map(bp_mapping)

# Convert Cholesterol to numerical
cholesterol_mapping = {"NORMAL": 0, "HIGH": 1}
df_encoded["Cholesterol"] = df_encoded["Cholesterol"].map(cholesterol_mapping)

# Convert Sex to numerical
sex_mapping = {"F": 0, "M": 1}
df_encoded["Sex"] = df_encoded["Sex"].map(sex_mapping)

# Convert Drug to numerical
drug_mapping = {"DrugY": 0, "drugC": 1, "drugX": 2, "drugA": 3}
df_encoded["Drug"] = df_encoded["Drug"].map(drug_mapping)
```

```
# Calculate correlation matrix
corr_matrix = df_encoded.corr()

# Plot the correlation matrix
plt.figure(figsize=(12, 10))
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap='coolwarm')
plt.title("Correlation Matrix")
plt.show()
```



In [49]:
```
# Scatter plot
sns.barplot(x='Drug', y='Na_to_K', data=df)
plt.title("Drug vs Na_to_K")
plt.show()
```



8. Interpretation of Results

● Age Distribution: An understanding of which age groups are most represented in the dataset.

● Sex Distribution: The ratio of male to female patients.

● Blood Pressure Categories: Insights into how many patients fall into high, normal, or elevated blood pressure categories.

● Correlation Matrix: Whether there's a significant correlation between the variables.

● Boxplots: Differences in cholesterol levels and Na to Potassium ratio between males and females.

9. Conclusion

Based on the exploratory analysis, you can summarize key findings such as:

● Whether certain age groups are more prone to high blood pressure or high cholesterol.

● How sodium and potassium balance varies with age and sex.

● Initial insights into how these variables might impact drug safety. This project provides a solid foundation for understanding basic data analysis concepts in the context of Drug

Safety and Pharmacovigilance.