250k Medicines Usage, Side Effects, and Substitute Analysis Report

short line

Opeyemi Fayipe  
17th September, 2024

# Introduction

This project focuses on analyzing a dataset containing information on over 248,000 drugs. The key objectives are:

* To understand the distribution of side effects among drugs.
* To explore the availability of substitutes.
* To analyze the habit-forming potential of drugs across various therapeutic classes.

The insights from this analysis aim to inform healthcare professionals and pharmaceutical stakeholders about drug safety and potential risks.

# Data Description

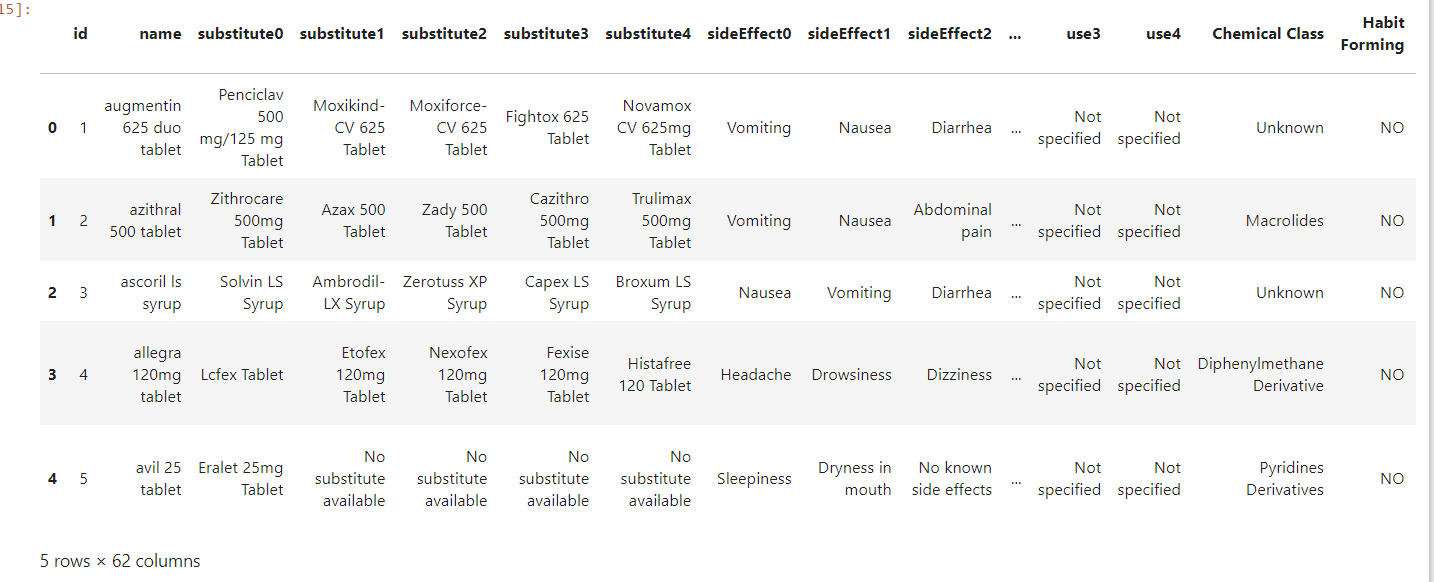
The dataset consists of various columns representing the drugs' names, side effects, therapeutic classes, chemical composition, and whether they are habit-forming. The main columns include:

1. **Drug Name**: The commercial name of the drug.
2. **Substitute**: Alternative drugs with similar compositions.
3. **Side Effects:** Information on adverse reactions.
4. **Therapeutic Class:** It is classifying medical drugs according to their functions. Each therapeutic class is a group of similar medications classified together because they are intended to treat the same medical conditions. For instance, respiratory is one of the classes, so all medications treating this illness are in the same class.
5. **Action Class:** It is the way of classifying medications based on actions they perform such as "H2 Receptor Blocker" It Block H2 receptors in parietal cells of the stomach → decreases gastric acid secretion. So drugs with similar action are grouped under "H2 Receptor Blocker".
6. **Chemical Class**: As the name suggests, it is grouped based on the chemical compound used.
7. **Habit Forming:** Indicates whether the drug is habit-forming.

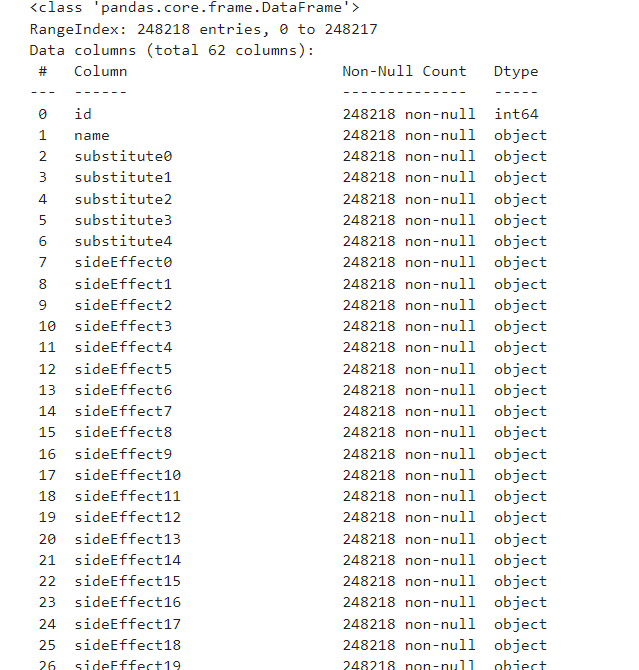
# Data Loading and Initial Exploration

Load the Dataset, to get a basic overview

| import pandas as pd  # Load the dataset file\_path = (r"C:\Users\Admin\OneDrive\Desktop\Unified Mentor Projects\250k Medicines Usage, Side Effects and Substitutes.csv") df = pd.read\_csv(file\_path)  df.head() |
| --- |



| # Display basic info about the dataset df.info() |
| --- |



| # Display basic info about the dataset  df.describe() |
| --- |

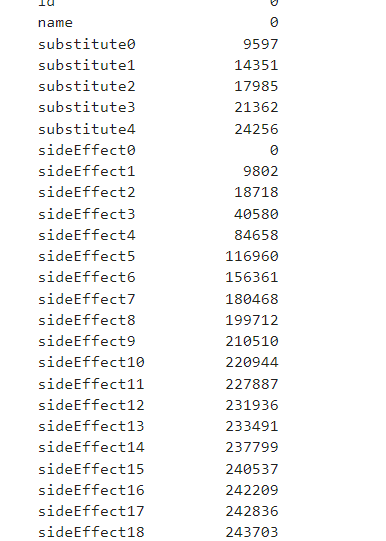
## 

# 

# Data Cleaning and Preparation

The dataset contained missing values, particularly in the Chemical Class, Therapeutic Class, and Action Class columns. Missing data was handled as follows:

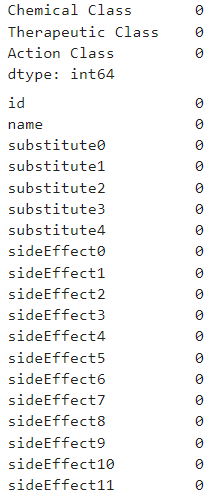
| # Check for missing values print(df.isnull().sum()) |
| --- |



### Handling Missing Values

* **Substitute Columns:** Missing values were replaced with 'No substitute available'.
* **Side Effects Columns:** Missing values were replaced with 'No known side effects'.
* **Habit Forming Column:** Missing values were filled with 'NO'.
* Chemical, Therapeutic, and Action Classes: Missing values were imputed with 'Unknown'.

| # Fill missing values for substitutes and side effects substitute\_cols = [f'substitute{i}' for i in range(5)] side\_effect\_cols = [f'sideEffect{i}' for i in range(42)] df[substitute\_cols] = df[substitute\_cols].fillna('No substitute available') df[side\_effect\_cols] = df[side\_effect\_cols].fillna('No known side effects')  # Fill missing values in 'Habit Forming' column df['Habit Forming'] = df['Habit Forming'].fillna('NO')  # Fill missing usage columns with 'Not specified'  usage\_cols = [f'use{i}' for i in range(5)]  df[usage\_cols] = df[usage\_cols].fillna('Not specified')  # Handle missing values in 'Chemical Class', 'Therapeutic Class', and 'Action Class' with 'Unknown'  df['Chemical Class'] = df['Chemical Class'].fillna('Unknown')  df['Therapeutic Class'] = df['Therapeutic Class'].fillna('Unknown')  df['Action Class'] = df['Action Class'].fillna('Unknown')  # Verify that there are no more missing values in these columns  missing\_values\_summary = df[['Chemical Class', 'Therapeutic Class', 'Action Class']].isnull().sum()  print(missing\_values\_summary)  # Check if missing values are handled properly  missing\_summary = df.isnull().sum()  missing\_summary |
| --- |



# Exploratory Data Analysis (EDA)

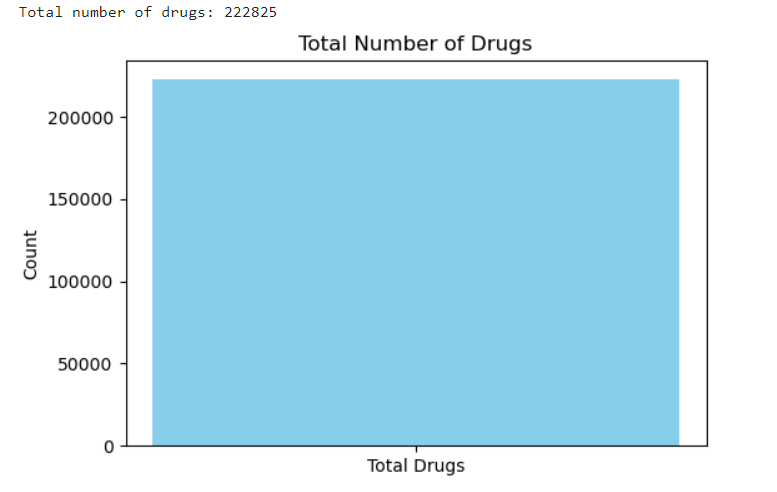
**Python Code Snippet:**

| # Total number of unique drugs total\_drugs = df['name'].nunique() print(f"Total number of drugs: {total\_drugs}")  import matplotlib.pyplot as plt plt.figure(figsize=(6,4)) plt.bar(['Total Drugs'], [total\_drugs], color='skyblue') plt.title('Total Number of Drugs') plt.ylabel('Count') plt.show() |
| --- |

### Total Number of Drugs

* There are 222,825 unique drugs in the dataset.

**Visualization:**



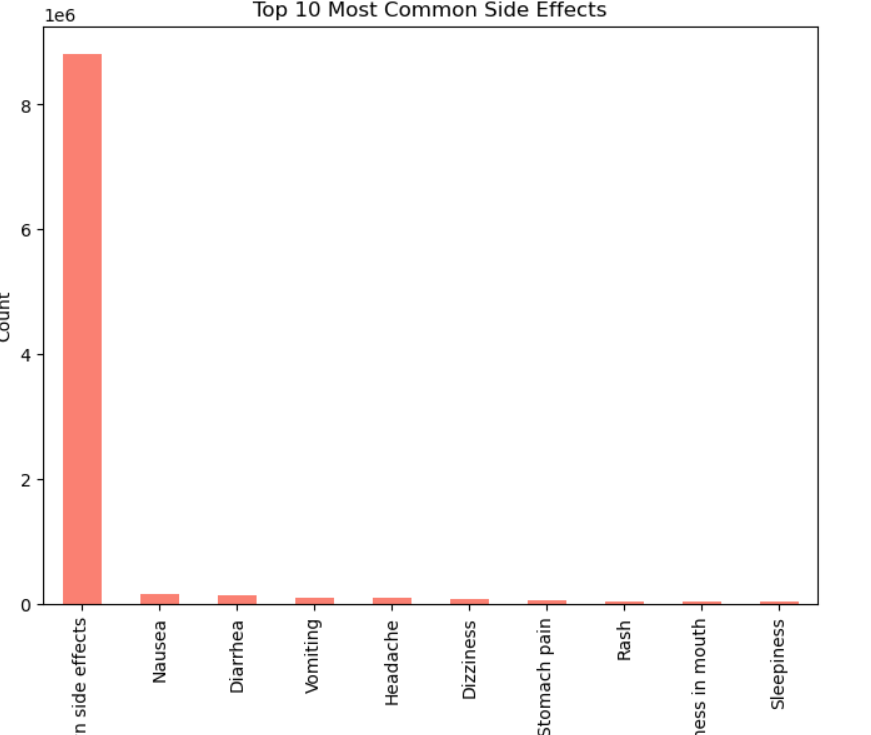
### Most Common Side Effects

**Python Code Snippet:**

| # Combining all side effect columns side\_effect\_columns = [col for col in df.columns if 'sideEffect' in col] side\_effects = pd.Series(df[side\_effect\_columns].values.ravel()).dropna()  # Getting the top 10 most common side effects common\_side\_effects = side\_effects.value\_counts().head(10)  # Plotting the most common side effects plt.figure(figsize=(8,6)) common\_side\_effects.plot(kind='bar', color='salmon') plt.title('Top 10 Most Common Side Effects') plt.xlabel('Side Effect') plt.ylabel('Count') plt.xticks() plt.show() |
| --- |

**Visualization:**

The top 10 most common side effects are **Nausea**, **Vomiting**, and **Headache**, among others. With the highest being “No known side effects”.



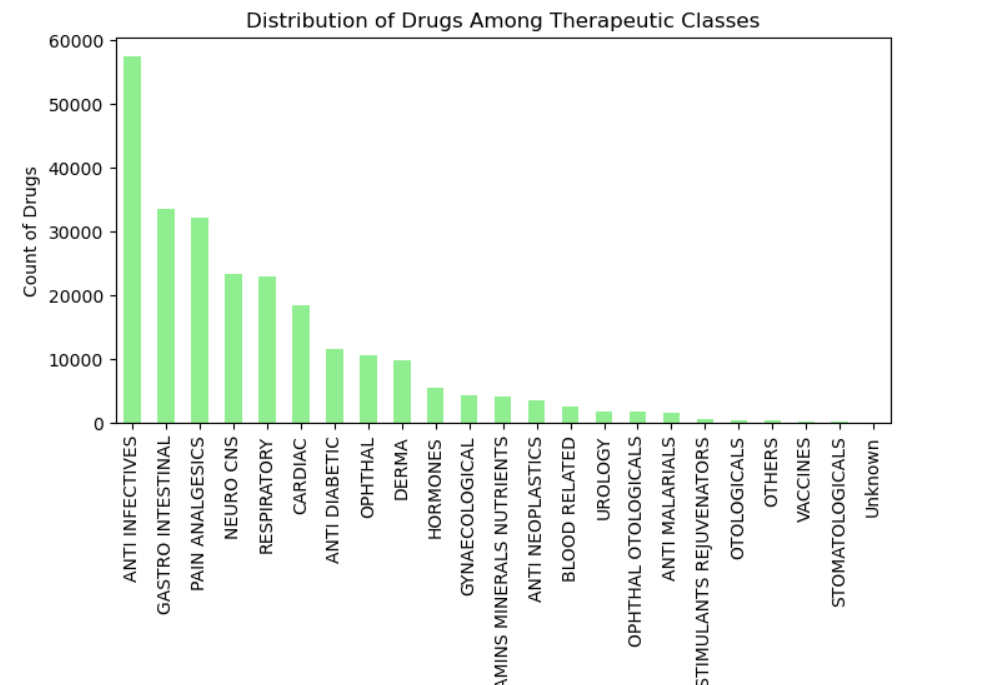
### 

# Distribution of Drugs Among Therapeutic Classes

**Python Code Snippet:**

| # Distribution of drugs among therapeutic classes # Distribution of drugs among therapeutic classes therapeutic\_class\_distribution = df['Therapeutic Class'].value\_counts()  # Plotting the distribution plt.figure(figsize=(12,8)) therapeutic\_class\_distribution.plot(kind='bar', color='lightgreen') plt.title('Distribution of Drugs Among Therapeutic Classes') plt.xlabel('Therapeutic Class') plt.ylabel('Count of Drugs') plt.xticks(rotation=90) plt.show() |
| --- |

**Visualization**:

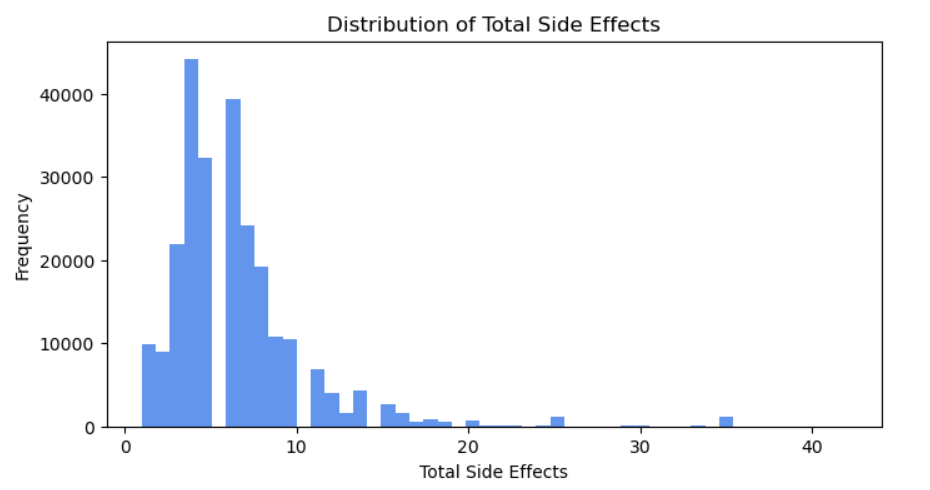
The majority of drugs belong to the Anti Infectives, Gastrointestinal, and Respiratory therapeutic classes.  


# Total Side Effects

The dataset shows a wide range of side effects per drug, with most drugs having fewer than 10 side effects.

**Python Code Snippet:**

| # List of side effect columns side\_effect\_cols = [col for col in df.columns if 'sideEffect' in col]  # Creating the 'side\_effect\_count' column by counting non-'No known side effects' entries df['side\_effect\_count'] = df[side\_effect\_cols].apply(lambda row: row[row != 'No known side effects'].count(), axis=1)  # Distribution of total side effects plt.figure(figsize=(8,4)) plt.hist(df['side\_effect\_count'], bins=50, color='cornflowerblue') plt.title('Distribution of Total Side Effects') plt.xlabel('Total Side Effects') plt.ylabel('Frequency') plt.show() |
| --- |

**Visualization:**

# Habit-Forming Drugs Analysis

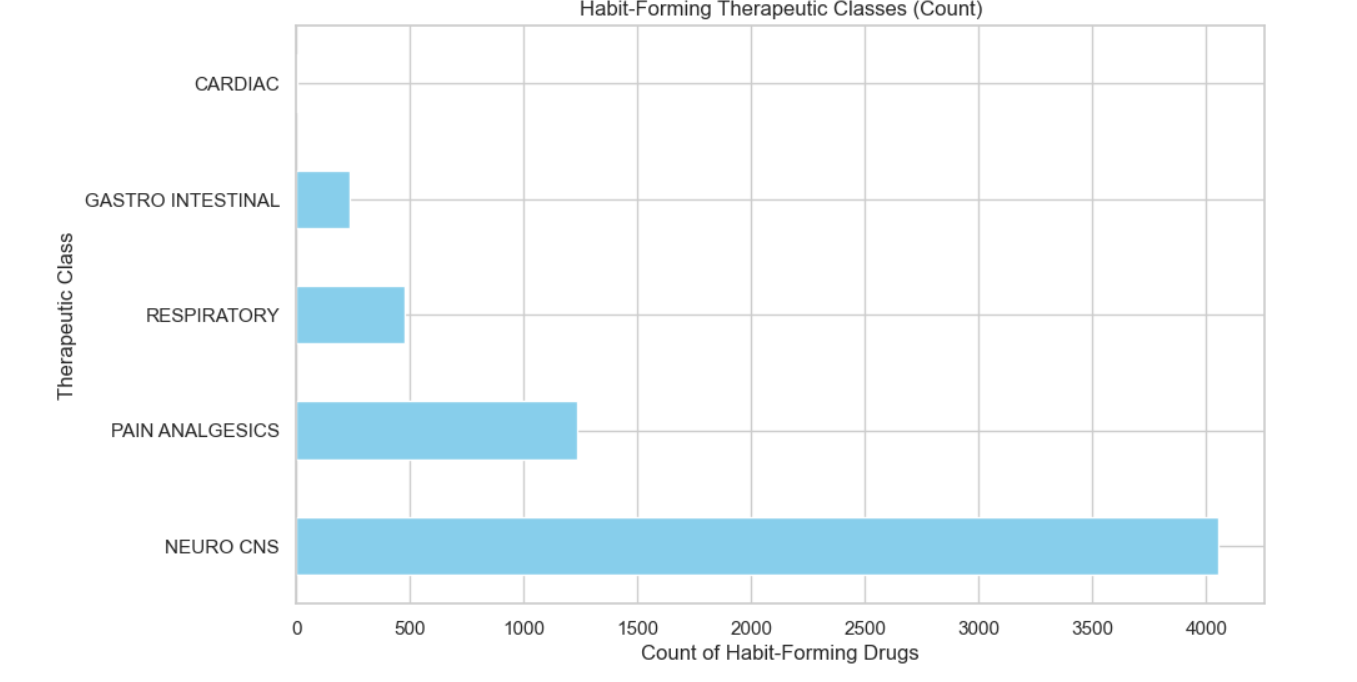
### Habit-Forming Drugs by Therapeutic Class

The Neuro CNS and Pain Analgesics therapeutic classes have the highest number of habit-forming drugs, with 4054 and 1233 drugs, respectively.

| **Therapeutic Class** | **Values** |
| --- | --- |
| NEURO CNS | 4054 |
| PAIN ANALGESICS | 1233 |
| RESPIRATORY | 474 |
| GASTROINTESTINAL | 236 |
| CARDIAC | 6 |

**Python Code Snippet:**

| **# Cross-checking the count of habit-forming drugs per therapeutic class habit\_forming\_classes = df[df['Habit Forming'] == 'YES']['Therapeutic Class'].value\_counts() print(habit\_forming\_classes)   # Plot for Habit-Forming drugs plt.figure(figsize=(10, 6)) habit\_forming\_classes.plot(kind='barh', color='skyblue') plt.title('Habit-Forming Therapeutic Classes (Count)') plt.xlabel('Count of Habit-Forming Drugs') plt.ylabel('Therapeutic Class') plt.show()** |
| --- |

****

## 

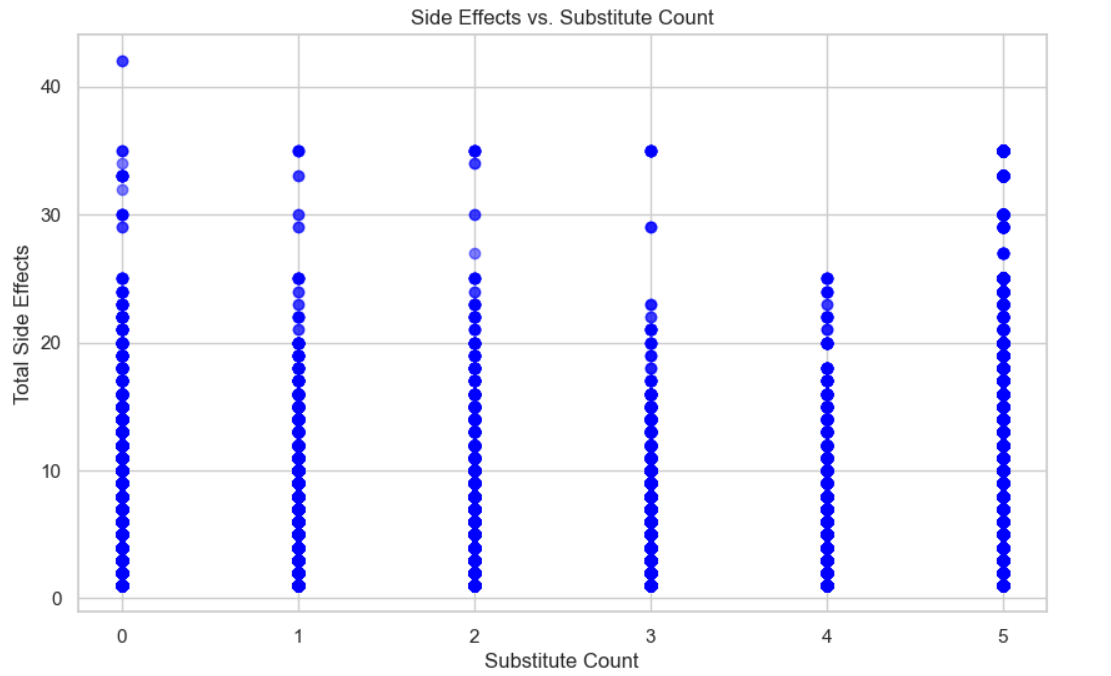
# Substitutes and Side Effects

### Substitutes and Side Effects Relationship

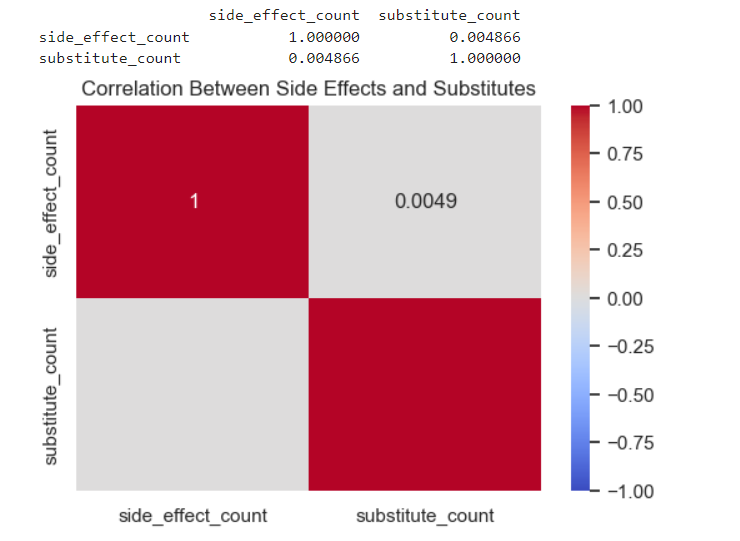
An analysis of the relationship between the number of substitutes and side effects suggests that there is minimal correlation between these two factors.

Python Code Snippets

| # Relationship between side effect count and substitutes plt.figure(figsize=(10,6)) plt.scatter(df['substitute\_count'], df['side\_effect\_count'], alpha=0.5, color='blue') plt.title('Side Effects vs. Substitute Count') plt.xlabel('Substitute Count') plt.ylabel('Total Side Effects') plt.show() |
| --- |

****

| **# Calculate the correlation between numerical columns correlation\_matrix = df[['side\_effect\_count', 'substitute\_count']].corr()  # Display the correlation matrix print(correlation\_matrix)  # Plot the correlation heatmap plt.figure(figsize=(6, 4)) sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1) plt.title('Correlation Between Side Effects and Substitutes') plt.show()** |
| --- |



# Habit-Forming Drugs with Substitutes

About **5756 habit-forming drugs** have substitutes, while **247 habit-forming drugs** do not.

**Code Snippet:**

| **# Habit-forming drugs with and without substitutes habit\_with\_substitutes = df[(df['Habit Forming'] == 'YES') & (df['substitute\_count'] > 0)].shape[0] habit\_without\_substitutes = df[(df['Habit Forming'] == 'YES') & (df['substitute\_count'] == 0)].shape[0]  print(f"Habit-forming drugs with substitutes: {habit\_with\_substitutes}") print(f"Habit-forming drugs without substitutes: {habit\_without\_substitutes}")** |
| --- |

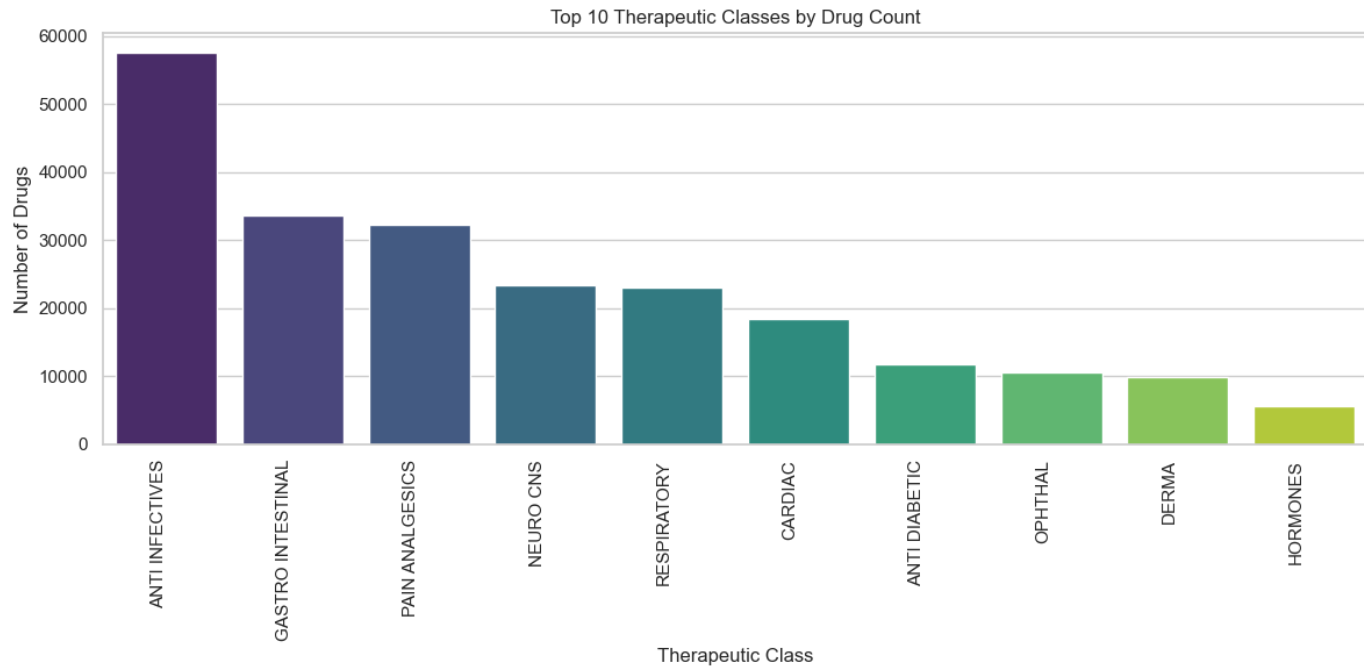
****

# Top 10 Therapeutic Classes by Drug Count

Python Code Snippet:

| import seaborn as sns  # Set plot styles for better visuals sns.set(style="whitegrid") plt.figure(figsize=(12, 6))  # Step 1: Distribution of drugs by Therapeutic Class therapeutic\_class\_counts = df['Therapeutic Class'].value\_counts().head(10) # Top 10 therapeutic classes sns.barplot(x=therapeutic\_class\_counts.index, y=therapeutic\_class\_counts.values, palette="viridis") plt.title('Top 10 Therapeutic Classes by Drug Count') plt.ylabel('Number of Drugs') plt.xlabel('Therapeutic Class') plt.xticks(rotation=90, ha='right') plt.tight\_layout() plt.show() |
| --- |

**Visualization:**

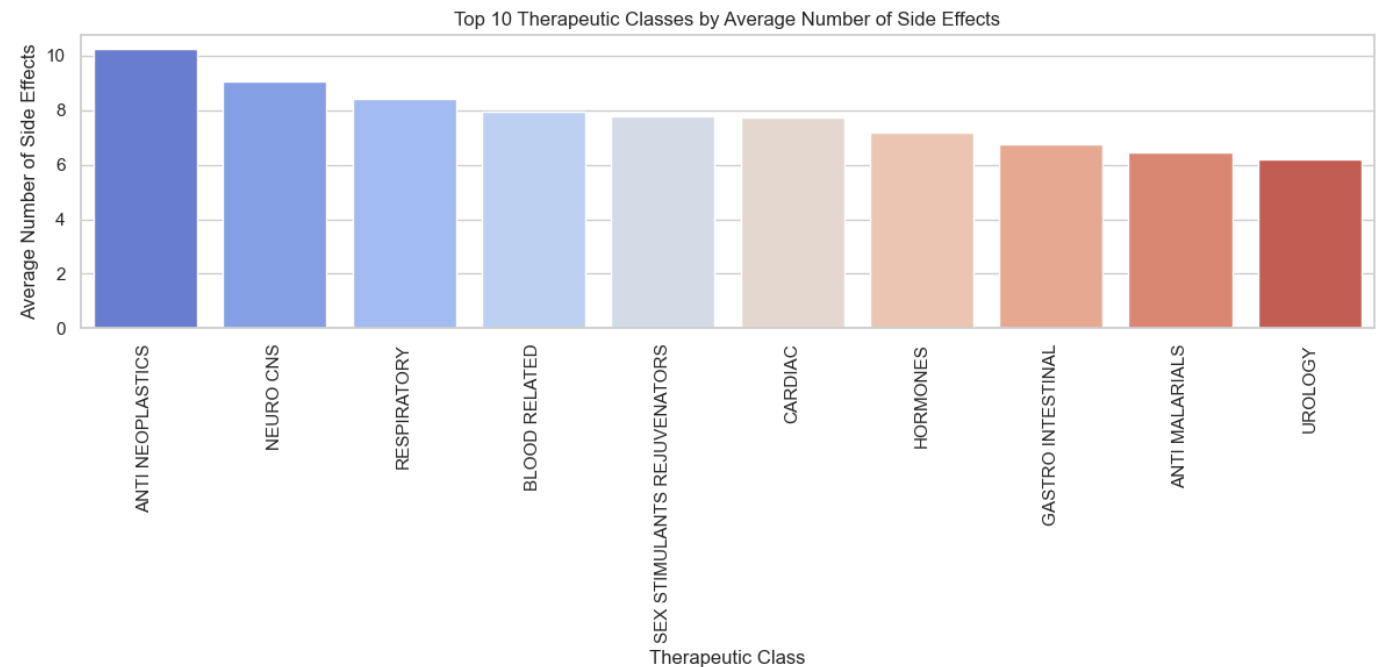


# Top 10 Therapeutic Classes by Average Number of Side Effects

Python Code Snippets:

| # Analyzing Average Side Effects Per Therapeutic Class # Count the number of non-empty side effects for each drug df['side\_effect\_count'] = df[side\_effect\_cols].apply(lambda row: row[row != 'No known side effects'].count(), axis=1)  # Calculate the average side effects per therapeutic class avg\_side\_effects\_per\_class = df.groupby('Therapeutic Class')['side\_effect\_count'].mean().sort\_values(ascending=False)  # Plot the result plt.figure(figsize=(12, 6)) sns.barplot(x=avg\_side\_effects\_per\_class.index[:10], y=avg\_side\_effects\_per\_class.values[:10], palette="coolwarm") plt.title('Top 10 Therapeutic Classes by Average Number of Side Effects') plt.ylabel('Average Number of Side Effects') plt.xlabel('Therapeutic Class') plt.xticks(rotation=90) plt.tight\_layout() plt.show() |
| --- |

Visualization:

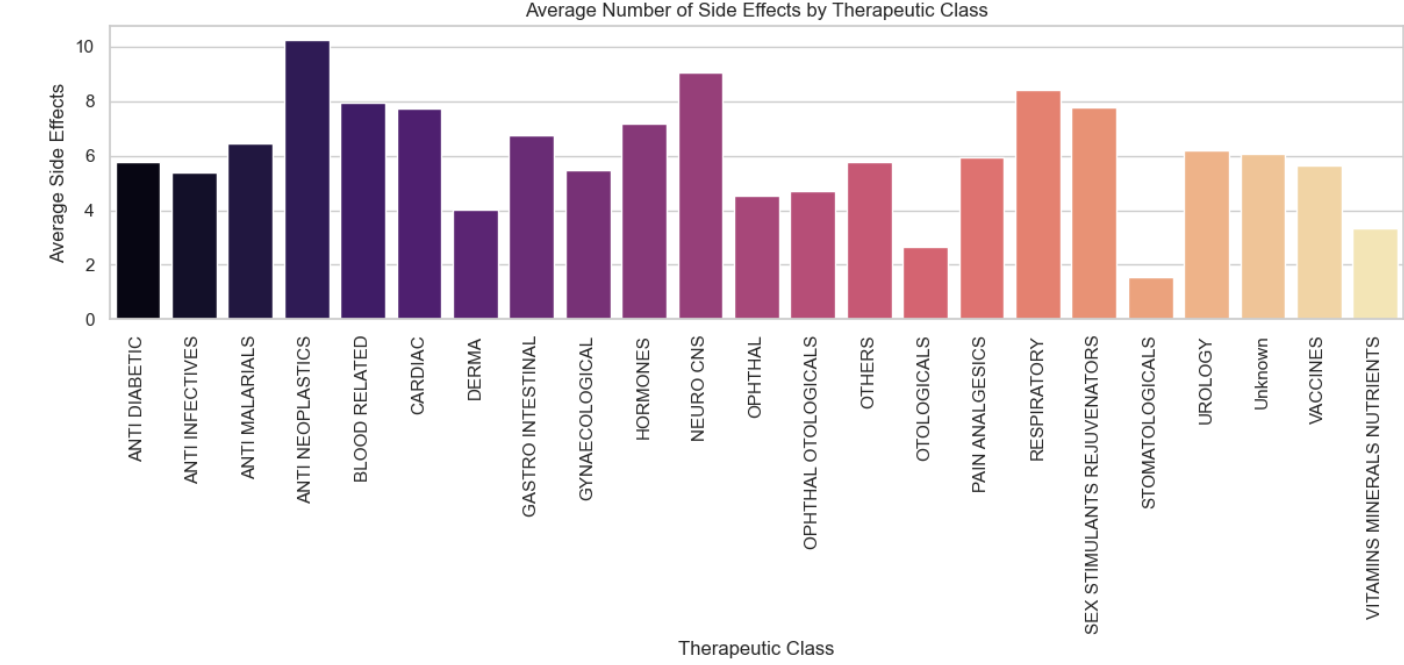


# Average Number of Side Effects by Therapeutic Class

Python Code Snippet:

| # Group by Therapeutic Class and calculate the average side effects avg\_side\_effects\_by\_class = df.groupby('Therapeutic Class')['side\_effect\_count'].mean()  # Plot the results plt.figure(figsize=(12, 6)) sns.barplot(x=avg\_side\_effects\_by\_class.index, y=avg\_side\_effects\_by\_class.values, palette="magma") plt.title('Average Number of Side Effects by Therapeutic Class') plt.xticks(rotation=90) plt.ylabel('Average Side Effects') plt.tight\_layout() plt.show() |
| --- |

Visualization:



# 

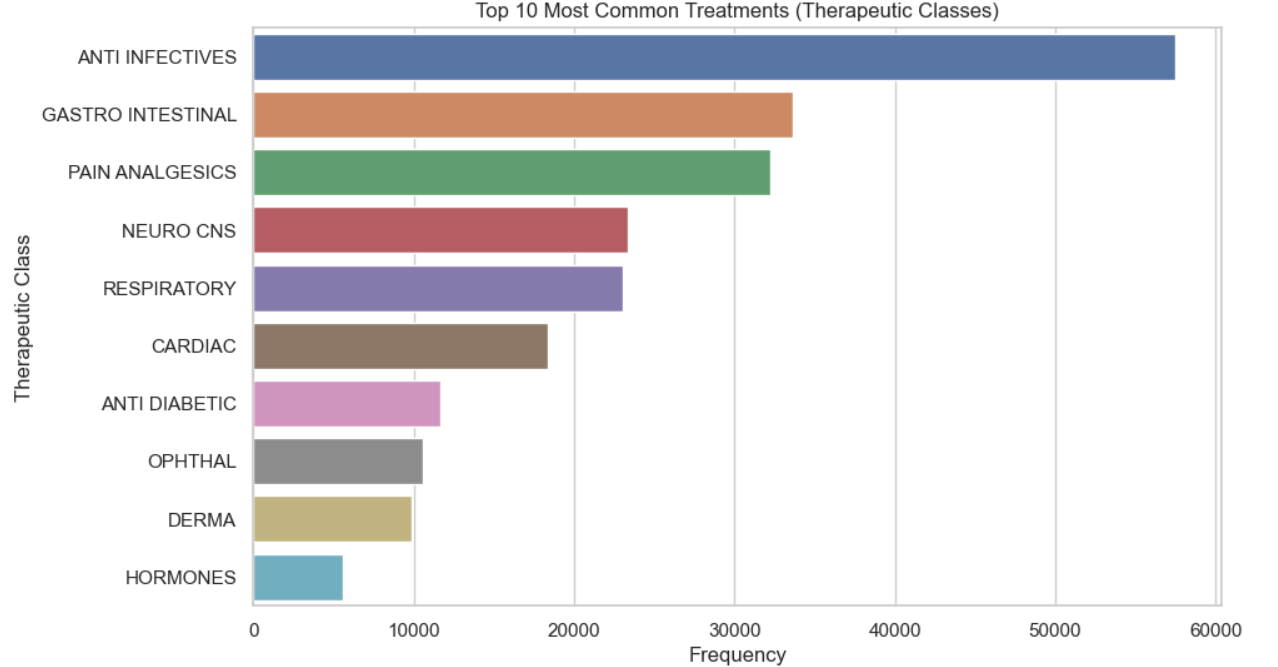
# Top 10 Most Common Treatments By Therapeutic Classes

Anti-infectives, Gastro-intestinal, Pain Analgesics are the most common treatments.

Python Code Snippet:

| # Popular Treatment # Count the occurrences of each therapeutic class therapeutic\_class\_counts = df['Therapeutic Class'].value\_counts()  # Plot popular therapeutic classes plt.figure(figsize=(10,6)) sns.barplot(x=therapeutic\_class\_counts[:10].values, y=therapeutic\_class\_counts[:10].index) plt.title('Top 10 Most Common Treatments (Therapeutic Classes)') plt.xlabel('Frequency') plt.ylabel('Therapeutic Class') plt.show() |
| --- |

Visualization:



# Conclusion

### Key Insights:

* Neuro CNS and Pain Analgesics therapeutic classes have the highest number of habit-forming drugs.
* Drugs with many substitutes do not necessarily have fewer side effects, indicating that substitutes do not always offer safer alternatives.
* Most drugs in the dataset are not habit-forming, with only 6003 out of 248,218 drugs labeled as habit-forming.

### Correct and Precise Interpretation:

The data reveals that certain therapeutic classes, particularly those dealing with neurological and pain management drugs, are more prone to being habit-forming. Additionally, having substitutes does not guarantee that a drug will have fewer side effects, suggesting that further investigation into drug alternatives is required.

# 

# Recommendations:

1. Healthcare Providers: Should be cautious when prescribing drugs from habit-forming classes like Neuro CNS and Pain Analgesics.
2. Pharmaceutical Companies: Should focus on developing safer alternatives, particularly for drugs with known side effects and no substitutes.

### Future Analysis:

* Side Effect Severity: Further research could explore the severity of side effects across therapeutic classes.
* Dosage Analysis: Investigating how dosage impacts side effects could yield actionable insights.
* Habit-Forming Prediction: Developing models to predict habit-forming tendencies based on chemical and therapeutic classes.