**Drugs, Side Effects & Medical Conditions — Final Report**

**Author:** Manasa Shetty

**Summary**

In this project, I analyzed a dataset of drugs, their associated medical conditions, side effects, and user ratings. The goal of my analysis was to explore how users perceive drug effectiveness, what common side effects are reported, and how ratings and reviews vary across different drugs.

The dataset contained 2,931 records and initially 17 columns, which I expanded to 20 after feature engineering. Some key takeaways from my work:

* The mean user rating (non-zero) was about 6.93, with a median of 7.05, showing that most users rate drugs moderately positively.
* Highly reviewed drugs included phentermine (2,934 reviews, rating 8.7), contrave (1,939 reviews, rating 6.6), and escitalopram (1,471 reviews, rating 7.4).
* Common side effects reported included hives, nausea, vomiting, itching, dizziness, and diarrhea. However, I also found parsing artifacts like “lips” and “tongue,” which came from splitting free-text descriptions.
* I discovered that the number of reviews and ratings are weakly correlated (r ≈ 0.218), meaning popular drugs don’t always get the highest scores.

**Dataset Description**

The dataset I worked on came from **Drugs.com** user-contributed reviews.

The dataset contains the following columns:

* **drug\_name**: Name of the drug.
* **medical\_condition**: The condition the drug is used to treat.
* **side\_effects**: Common side effects of the drug.
* **generic\_name**: The generic name of the drug.
* **drug\_classes**: The class of the drug (e.g., antibiotic, antihistamine).
* **brand\_names**: Brand names under which the drug is sold.
* **activity**: The activity of the drug (e.g., active, inactive).
* **rx\_otc**: Indicates if the drug is prescription (Rx) or over-the-counter (OTC).
* **pregnancy\_category**: The drug's pregnancy risk category.
* **csa**: Controlled Substances Act schedule, if applicable.
* **alcohol**: Interactions with alcohol.
* **related\_drugs**: Other drugs related to the primary drug.
* **medical\_condition\_description**: A brief description of the medical
* condition.
* **rating**: Average user rating of the drug.
* **no\_of\_reviews**: Number of user reviews.
* **drug\_link**: URL link to more information about the drug.
* **medical\_condition\_url**: URL link to more information about the medical
* condition.

After cleaning and feature engineering, my working dataset contained:

* **2,931 rows**
* **20 columns** including new fields such as review\_score, side\_effect\_count, and effectiveness.

**Tools Required**

* **Python**: The primary programming language for data analysis.
* **Pandas**: For data manipulation and analysis.
* **Matplotlib/Seaborn**: For data visualization.
* **Jupyter Notebook**: To write and run Python code.

**Data Cleaning**

When I first examined the dataset, I noticed a lot of missing values:

* brand\_names (1,213 missing)
* related\_drugs (1,469 missing)
* alcohol (1,554 missing)
* rating and no\_of\_reviews (1,345 missing each)

To address this, I filled missing values with placeholders such as "Unknown", "None", or "Not Specified" depending on the column. For ratings and number of reviews, I filled missing values with **0** to keep all rows.

I also standardized text columns (lowercased and stripped whitespace), and converted categorical columns like rx\_otc, pregnancy\_category, and csa into categorical datatypes for easier analysis.

**Feature Engineering**

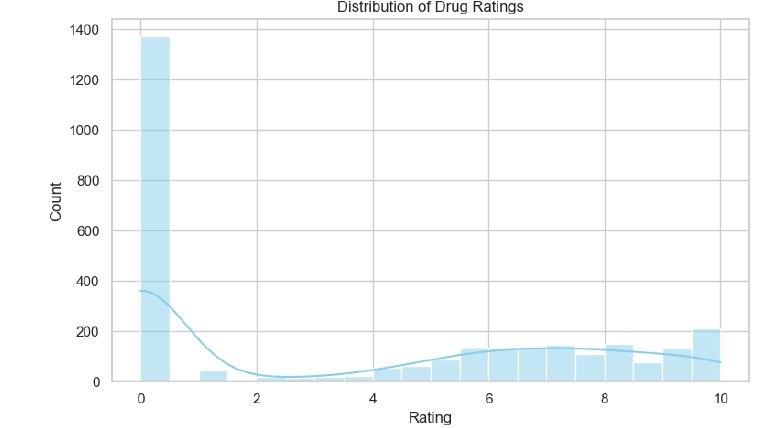
To enhance the analysis, I created three new features:

1. **Review Score** = rating \* no\_of\_reviews --- to capture both popularity and quality.
2. **Side Effect Count** = number of side effects listed--- to approximate how many issues are reported per drug.
   * Median count was about 27, but one record had as many as 1,394 listed effects.
3. **Effectiveness** → categorical grouping of ratings:
   * High (≥ 8)
   * Medium (≥ 5)
   * Low (> 0, < 5)
   * Unrated (0)

**Exploratory Data Analysis**

**Rating Distribution**

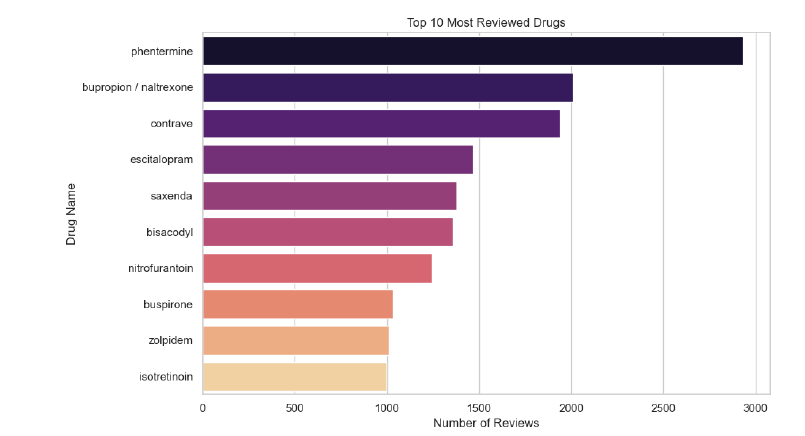
Among **1,560 non-zero ratings**, the mean was **6.93** and median **7.05**. Most drugs fell in the 6–9 range.



**Top Reviewed Drugs**

The most-reviewed drugs were:

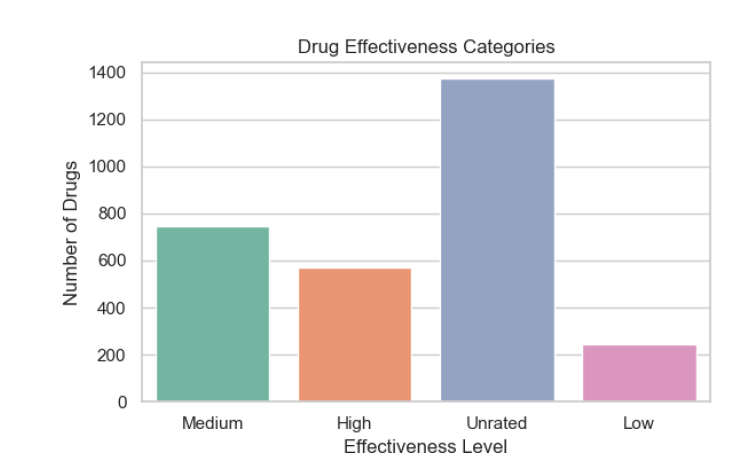
* **Phentermine** — 2,934 reviews, rating 8.7
* **Bupropion/Naltrexone** — 2,013 reviews, rating 6.6
* **Contrave** — 1,939 reviews, rating 6.6
* **Escitalopram** — 1,471 reviews, rating 7.4
* **Saxenda** — 1,377 reviews, rating 7.5



**Drug Effectiveness Categories**

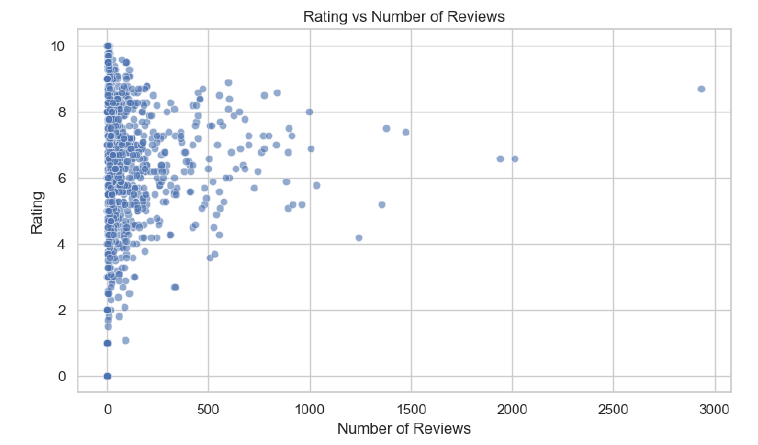
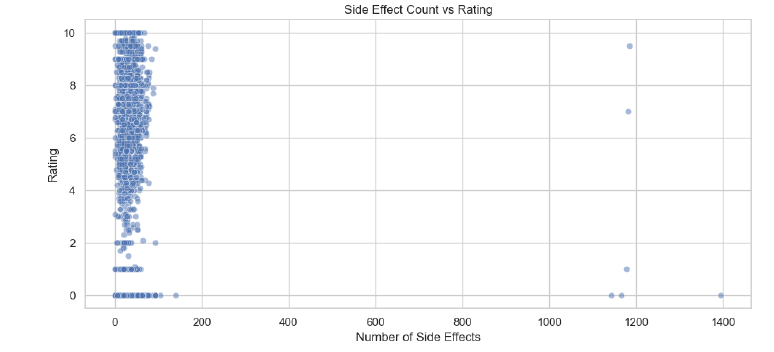
After categorizing ratings:

* **Unrated:** 1,371
* **Medium:** 746
* **High:** 570
* **Low:** 244



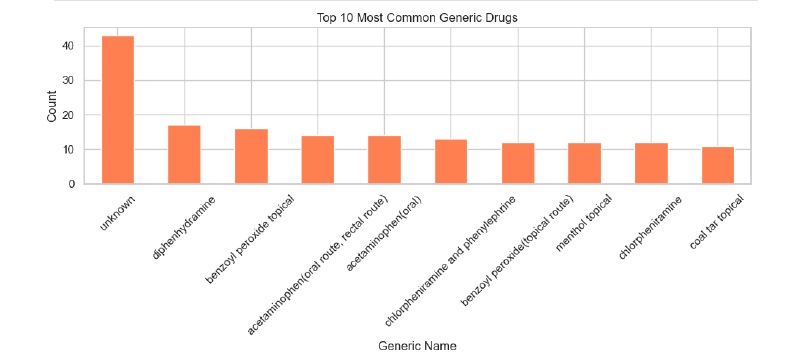
**Relationship Between Reviews, Ratings, and Side Effects**

* I found a **weak positive correlation (r ≈ 0.218)** between number of reviews and ratings.
* Scatterplots showed that drugs with more side effects did not always receive lower ratings.



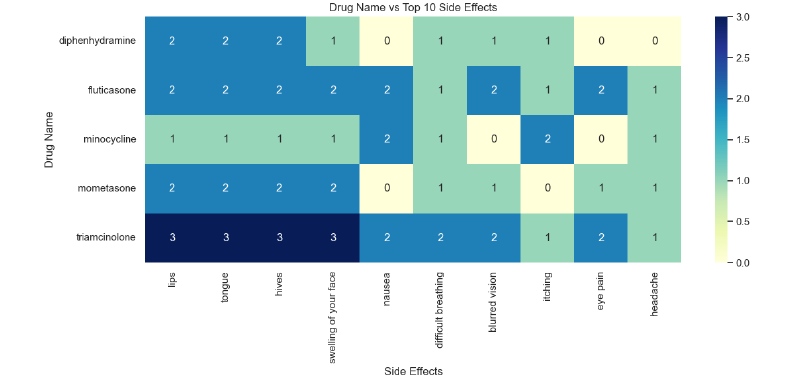
**Top Drugs by Review Score**

The most common generic name recorded in the dataset is **“Unknown”**, reflecting missing or unlisted information. Among identified drugs, **diphenhydramine** and **benzoyl peroxide (topical)** appear most frequently, followed by different formulations of **acetaminophen** and **chlorpheniramine**, which are widely used over-the-counter medications



**Side Effect Heatmap for Top Drugs**

I created a heatmap of the top 5 most-reviewed drugs versus their most frequent side effects. Drugs like **phentermine, saxenda, escitalopram** showed overlapping side effects such as hives, nausea, and dizziness.



**Key Observations**

* Many fields in the dataset were missing, and my approach of filling them with placeholders allowed me to retain all records but introduced a large “Unknown” category.
* The most common ratings were in the 6–9 range, showing moderate user satisfaction.
* Weight-loss and antidepressant drugs dominated the review counts.
* Free-text parsing of side effects created noisy terms, underlining the need for more advanced NLP preprocessing.
* Drugs with more reviews only had a slightly higher rating correlation, which means popularity doesn’t always equal better perception.

**Limitations**

* Free-text parsing of side effects was simplistic, producing artifacts like “lips” and “tongue.”
* Filling missing ratings with **0** inflated the count of “Unrated” drugs.
* Data is user-reported and unverified, so results should be treated cautiously.

**Future Work**

* Apply advanced NLP to clean and normalize side-effect text.
* Group drugs by class or indication for fairer comparisons.
* Perform sentiment analysis on user reviews (if available).
* Use advanced models (Random Forest, Gradient Boosting) to predict drug ratings or effectiveness.

**Conclusion**

Working on this project has given me a deeper understanding of how real-world drug data can be structured, analyzed, and interpreted. By cleaning and exploring the dataset, I was able to uncover important trends about drug ratings, reviews, side effects, and effectiveness. One of my main takeaways was that while the average rating of around 7/10 suggests overall moderate satisfaction, the distribution of reviews and side effects shows that user experience varies widely across drugs. Popular drugs like phentermine and escitalopram receive thousands of reviews and relatively high ratings, while others show lower effectiveness or attract complaints about side effects.

Another important finding is that the dataset is noisy and incomplete. Many fields contained missing information, especially for brand names, alcohol interactions, and related drugs. I addressed this by filling missing values with placeholders, which helped me keep the dataset intact, but I also realized this introduced categories like “Unknown” that are not clinically meaningful. Similarly, the side effect text field required far more advanced natural language processing (NLP) than I was able to implement in this phase. The raw string-splitting approach produced artifacts such as “lips” and “tongue,” which do not represent true standalone side effects. This taught me how critical text preprocessing and domain-specific knowledge are when working with medical data.

From a broader perspective, this project highlighted the potential value of combining quantitative metrics (ratings, review counts, review scores) with qualitative information (side effect descriptions). By doing so, I was able to see not just how drugs are perceived overall, but also the specific challenges patients face with different treatments. At the same time, I recognized the limitations of user-reported data, which can be biased, incomplete, or inconsistent.

Overall, this project reinforced my skills in data cleaning, feature engineering, exploratory data analysis, and interpretation, while also showing me areas for growth such as advanced NLP and predictive modeling. If I continue developing this project, I would like to focus on grouping drugs by class or medical condition, applying more sophisticated NLP to side effects, and testing machine learning models that could predict drug effectiveness or risk of adverse events. Through this project, I not only practiced technical data analysis but also gained valuable insight into the importance of data quality and thoughtful interpretation in healthcare-related analytics.

**Code & Output**

