
Drugs, Side Effects & Medical Conditions — Analytics Report

Role: Data Analyst / Data Scientist
Tools: Python, Pandas, NumPy, Scikit-Learn, MLxtend, Matplotlib

---

1. Executive Summary

This project analyzes a real-world pharmaceutical dataset containing **2,931 drug entries** sourced from Drugs.com. The dataset includes drug profiles, side effects, usage conditions, user ratings, pregnancy categories, and controlled substance classifications.

The objective was to perform **end-to-end data cleaning, feature engineering, exploratory data analysis (EDA), and data transformation** to prepare the dataset for predictive modeling and healthcare insights.

Key outcomes include:

- A fully cleaned and standardized dataset (0 missing values)
- Identification of major side effects, drug classes, and conditions
- Categorization of high-risk drugs based on pregnancy safety, alcohol interactions, and CSA schedule
- Conversion of unstructured text (side effects, drug classes) into useful analytical features
- Visual insights on drug effectiveness, user engagement, and therapeutic trends

This report provides a structured view of the dataset and highlights data-driven observations relevant to healthcare analytics, pharmaceutical research, and patient-safety decision systems.

---

2. Dataset Overview

The dataset contains **17 attributes**, including:

- **Drug information:** drug_name, generic_name, drug_classes
- **Medical usage:** medical_condition, medical_condition_description
- **Safety indicators:** pregnancy_category (A–X), CSA schedule, alcohol interaction
- **User feedback:** rating, number_of_reviews
- **Unstructured fields:** side_effects, related_drugs (URLs and text)

The records span treatments for **Pain, Acne, Flu, Hypertension, Diabetes, Asthma, Mental Health disorders, Skin conditions**, and more. Several columns featured large amounts of missing data, text irregularities, and mixed data types requiring extensive preprocessing.

---

3. Data Cleaning & Preprocessing

3.1 Missing Value Handling

A total of **6,192 missing values** were present initially.
Actions taken:

- Filled clinical text columns (side_effects, related_drugs) with "Unknown"
- Converted numerical fields (rating, no_of_reviews) to float and filled missing entries with 0
- Normalized activity values by removing % and converting to 0–1
- Replaced alcohol interaction: X → 1, missing → 0
- Removed inconsistent brand_names column
- Ensured all categorical fields were encoded consistently

After cleaning, the dataset contains **0 missing values**.

---

4. Exploratory Data Analysis (EDA)

4.1 Medical Condition Distribution

Top medical conditions treated:

Condition	Count
Pain	264
Colds & Flu	245
Acne	238
Hypertension	177
Osteoarthritis	129

Insight: Common general-population conditions dominate the dataset, reflecting broad drug usage patterns.

4.2 Most Frequent Side Effects

Top reported side effects include:

- **Hives** – 1,788 occurrences
- **Difficult breathing** – over 1,500 combined entries
- **Itching** – 275 occurrences
- **Dizziness, swelling, fever symptoms** appear across several drug classes

Insight: Reactions are largely allergy-related, suggesting hypersensitivity is a major driver of adverse effects.

4.3 Drug Class Analysis

Most common drug classes:

- Upper respiratory combinations
- Topical acne agents
- Topical steroids
- Antihistamines
- Various antimicrobial classes

Insight: Respiratory and dermatology treatments account for the highest percentage of drugs.

4.4 Safety Categories

- **Pregnancy Category X & D drugs** are heavily represented in acne, psychiatric, and cancer treatments.
- Over **1,500 drugs show alcohol interaction**, reflecting significant patient safety implications.
- CSA Schedule 2–4 drugs appear primarily in neurological, psychiatric, and pain-management categories.

---

5. Feature Engineering & Data Transformation

To prepare the dataset for machine learning:

- **Label Encoding** applied to categorical columns (medical_condition, generic_name, etc.)
- **StandardScaler** applied to rating, no_of_reviews, activity, CSA codes, etc.
- **Boolean flags created** for major side effects and drug classes (Hives, Itching, Upper respiratory drugs, etc.)
- Text fields were parsed and split to extract structured attributes
- Cleaned dataset exported as version2.csv and version3.csv for ML pipelines

The transformed dataset is now fully suitable for:

- Classification
- Clustering
- Recommendation engines
- Association rule mining

---

6. Key Insights

1. User ratings correlate with review counts

Drugs with more reviews tend to score higher, indicating stronger acceptance and efficacy perception.

2. Respiratory and dermatological drug classes dominate

Cold/flu, acne, and allergy medications represent the largest segments.

3. Side effects are highly concentrated around allergic responses

Most frequent side effects include generalized allergic reactions such as hives and difficulty breathing.

4. High-risk drug categories require careful handling

Pregnancy Category X and CSA-controlled drugs appear across major therapeutic classes.

5. The dataset is ideal for predictive modeling

Encoded, scaled, and structured features support multiple ML applications.

---

7. Conclusion

This project delivers a complete analytical and ML-ready representation of a complex pharmaceutical dataset. Through robust data cleaning, structured feature engineering, and exploratory visual analytics, the data has been transformed into a form suitable for clinical insights, drug recommendation systems, and healthcare decision support tools.

The findings highlight significant trends in drug usage, side effects, safety categories, and medical condition prevalence. These insights can support pharmaceutical planning, regulatory oversight, clinical research, and patient-safety initiatives.
