Milestone 1 Documentation: Personalized Medication Recommendation System

TABLE OF CONTENTS

1 Objective

The objective of this project is to develop a personalized medication recommendation system. The system utilizes multiple datasets containing patient reviews, pharmaceutical details, and prescription records. It employs advanced data-driven methodologies, including a conversational agent to guide patients on proper medication usage, predict possible side effects, and suggest safer alternatives. The overall aim is to enable informed decision-making, improve drug compliance, and enhance patient safety through personalized care.

2 Dataset Description

2.1 Datasets Used

Datasets (All are sourced from Kaggle.com/datasets):

1. shudhanshusingh/250k-medicines-usage-side-effects-and-substitutes

   Reason: Large dataset (250k entries) = rich information.

   Crucial for analyzing side effects and offering substitute recommendations, aligning perfectly with personalized recommendations.Supports risk analysis by matching patient reviews with common adverse effects.

1. prathamtripathi/drug-classification

   Reason: Helps in categorizing drugs based on therapeutic use, chemical composition, or pharmacological effects.Essential for building classification models that personalize recommendations based on patient history and drug categories.

2. tajuddinkh/drugs-prescriptions-with-providers

   Reason: Adds a real-world prescription dimension, showing how providers prescribe drugs.Helps incorporate a provider-based filtering feature in the recommendation engine (e.g., suggesting drugs commonly prescribed by top providers for similar conditions).

3. milanzdravkovic/pharma-sales-data

   Reason: Integrates popularity metrics through sales data, providing insights into widely accepted medications.Useful for trend analysis: Do highly-rated drugs correlate with high sales?

   Supports demand-based recommendations, balancing patient preferences with market trends.

4. mohneesh7/indian-medicine-data

   Reason: Introduces geographical diversity, allowing exploration of regional preferences.

   Supports location-specific recommendations, making the system adaptable for region-based personalization.Allows comparison with global datasets, adding depth to the model's adaptability.

Dataset Accessibility and Compliance:

- All datasets were accessed through Kaggle and verified for public availability.
- Licensing terms for each dataset were reviewed to ensure compliance with usage policies.
- The data is used strictly for research and analysis purposes, adhering to the specified licensing agreements.

Dataset Summaries:

1. **Medicine Prescription Records**

   File Name: medicine_prescription_records.csv

   Dimensions: 239,930 rows × 4 columns

   Description: Contains records of medications prescribed by healthcare providers across different specialties.

   Variables:

   specialty (String): Medical specialty of the provider (e.g., Nephrology, General Practice).

   years_practicing (Integer): Number of years the provider has been practicing.

   cms_prescription_counts (String): Comma-separated list of drugs prescribed by the provider.

2. **Medicine Dataset**

File Name: medicine_dataset.csv

Dimensions: 248,218 rows × 58 columns

Description: Comprehensive dataset containing drug information, substitutes, side effects, and classifications.

Variables:

name (String): Name of the medicine.

substitute0 - substitute4 (String): Suggested substitutes for the medicine.

sideEffect0 - sideEffect41 (String): Potential side effects associated with the medicine.

use0 - use4 (String): Primary medical uses of the drug.

Chemical Class (String): Chemical classification of the drug.

Therapeutic Class (String): Therapeutic classification.

Action Class (String): Pharmacological action class of the drug.

3. **Medicine Data**

File Name: medicine_data.csv

Dimensions: 195,605 rows × 8 columns

Description: Details on medicines including pricing, manufacturers, descriptions, and drug interactions.

Variables:

sub_category (String): Subcategory of the drug (e.g., Human Insulin Basal).

product_name (String): Name of the product.

salt_composition (String): Composition of the drug.

product_price (String): Price of the product.

product_manufactured (String): Manufacturer of the drug.

medicine_desc (String): Description of the medicine.

side_effects (String): Common side effects.

drug_interactions (JSON): List of known drug interactions.

4. **Patient Drug Reviews**

File Name: drugsComTrain_raw.csv

Dimensions: 161,297 rows × 7 columns

Description: Patient-generated reviews for various medications along with ratings and conditions treated.

Variables:

uniqueID (Integer): Unique identifier for each review.

drugName (String): Name of the drug reviewed.

condition (String): Condition for which the drug was taken.

review (String): Patient's written review.

rating (Integer): Rating of the drug (1–10 scale).

date (String): Date when the review was posted.

usefulCount (Integer): Number of users who found the review helpful.

5. **Drug Classification Dataset**
   File Name: drug200.csv
   Dimensions: 200 rows × 6 columns
   Description: Dataset linking patient demographic and health characteristics to drug prescriptions.
   Variables:
   Age (Integer): Age of the patient.
   Sex (String): Gender of the patient (M/F).
   BP (String): Blood pressure category (HIGH, NORMAL, LOW).
   Cholesterol (String): Cholesterol level (HIGH, NORMAL).
   Na_to_K (Float): Sodium-to-Potassium ratio in the blood.
   Drug (String): Drug prescribed.

2.2 Dataset Summaries

Dataset	Description	Dimensions
Medicine Prescription Records	Prescriptions by healthcare providers across specialties	239,930 x 4
Medicine Dataset	Drug information, substitutes, side effects, classifications	248,218 x 58
Medicine Data	Pricing, manufacturers, drug descriptions, interactions	195,605 x 8
Patient Drug Reviews	Reviews, ratings, and conditions treated by patients	161,297 x 7
Drug Classification Dataset	Links patient demographics to prescriptions	200 x 6

3 Technology Stack

- Programming Language: Python
- Libraries: Pandas, NumPy, Matplotlib, Seaborn, WordCloud, SciPy, TextBlob
- Tools: Google Colab, Jupyter Notebook

4 Contributions

- Data Extraction: Asmitha Ramesh and Aslesha Sanjana Kodavali
- Preprocessing: Aslesha Sanjana Kodavali
- Exploratory Data Analysis: Asmitha Ramesh

5 Data Preprocessing

5.1 Handling Missing Values

- The condition column had missing values, which were removed.
- Missing values in product and price were cleaned and converted to numeric formats.

5.2 Removing Duplicates

5.2.1 Duplicate rows were removed. The Medicine Details dataset had 84 duplicates.

5.3 Data Integration

5.2.2 The Medicine Name column was renamed to drugName.

5.2.3 A left join was performed on drugName across datasets.

5.4 Feature Engineering

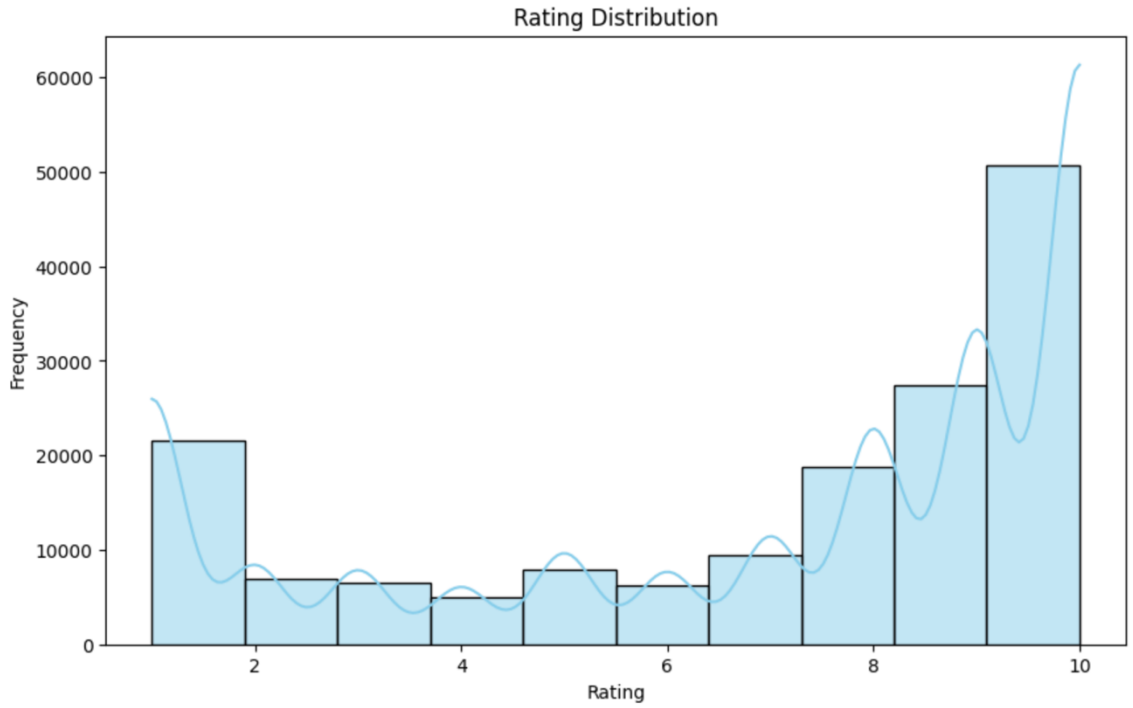
5.2.4 Average Rating per drug

5.2.5 Sentiment Score derived from reviews

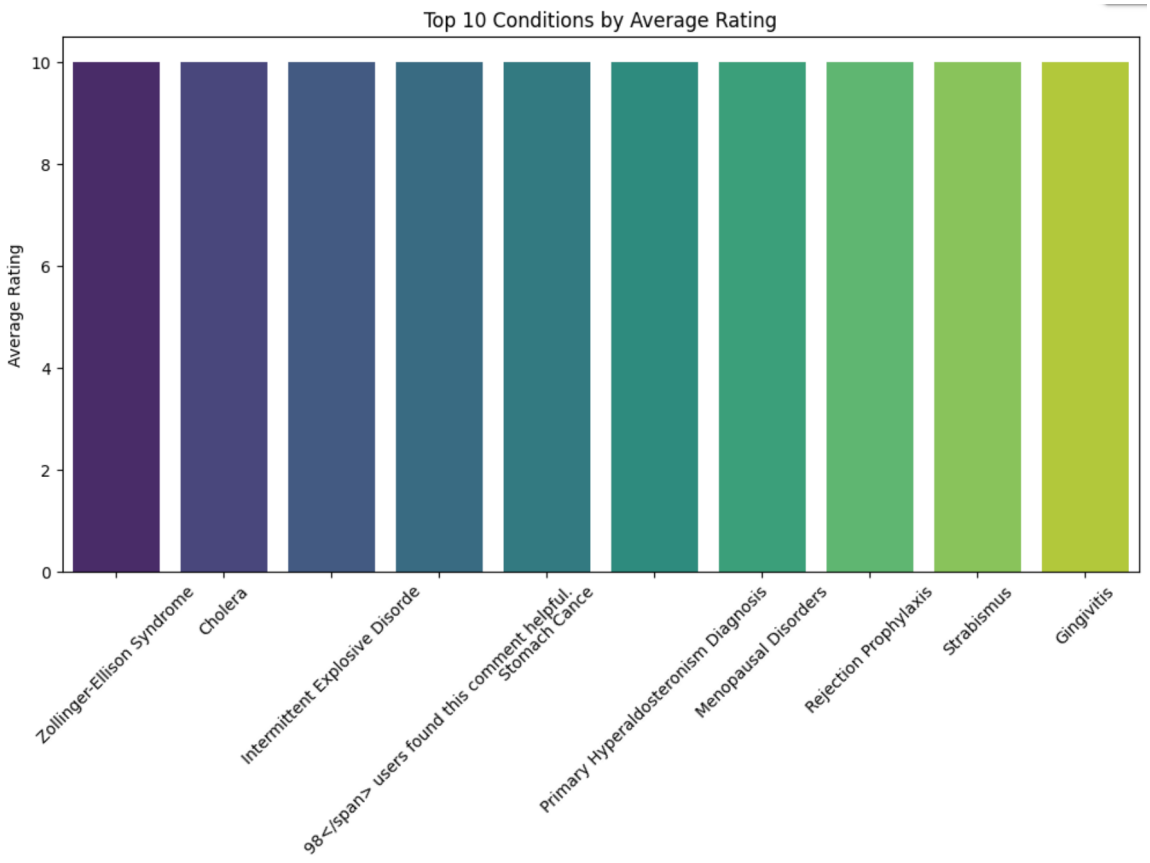
5.2.6 Drug interaction counts

6 Exploratory Data Analysis (EDA)

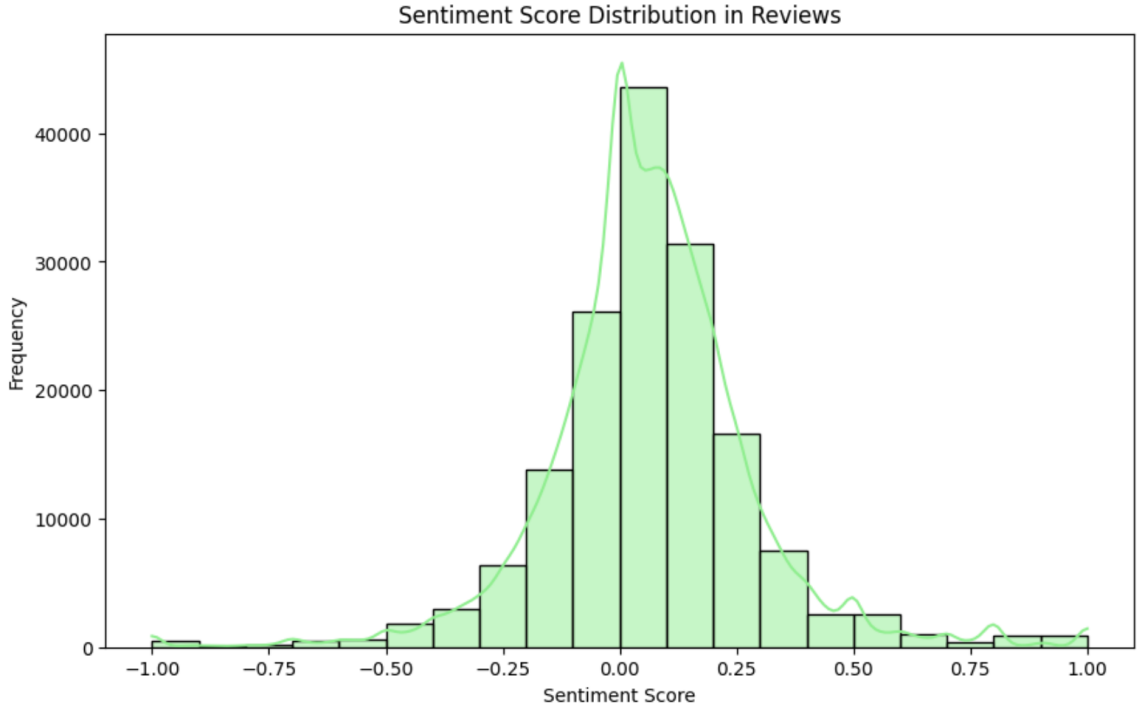
6.1 Rating Distribution: Majority between 7 and 10, showing right skew.



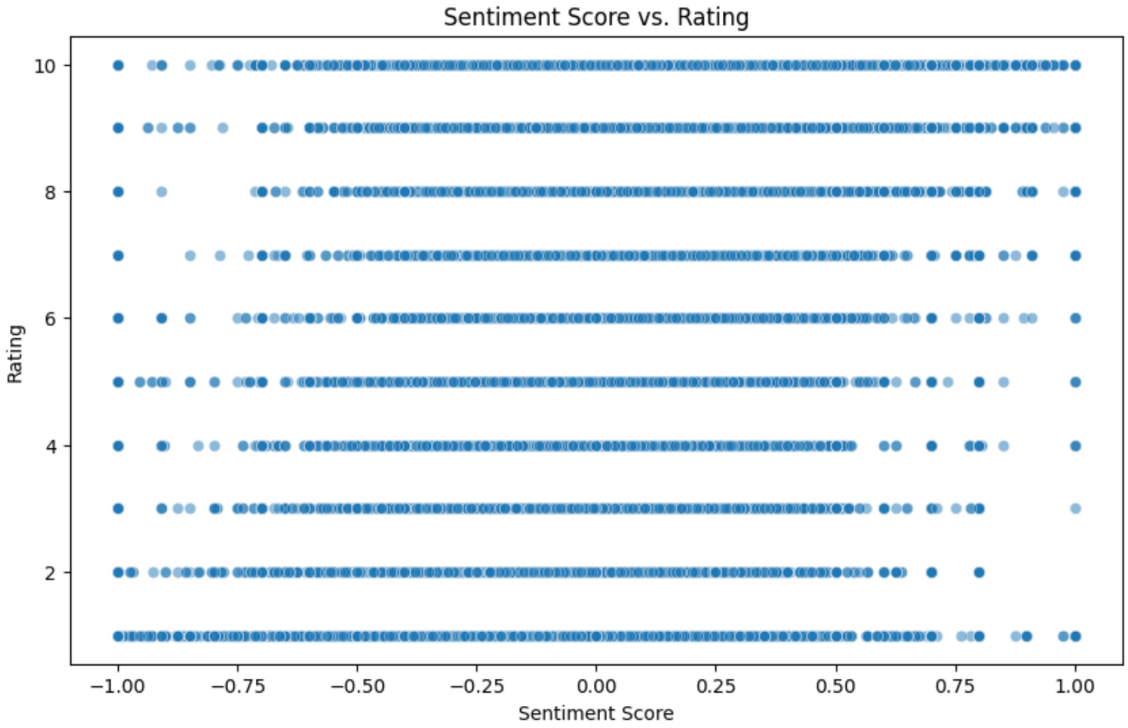
6.2 Top Conditions: Zollinger-Ellison Syndrome, Cholera, and Depression.



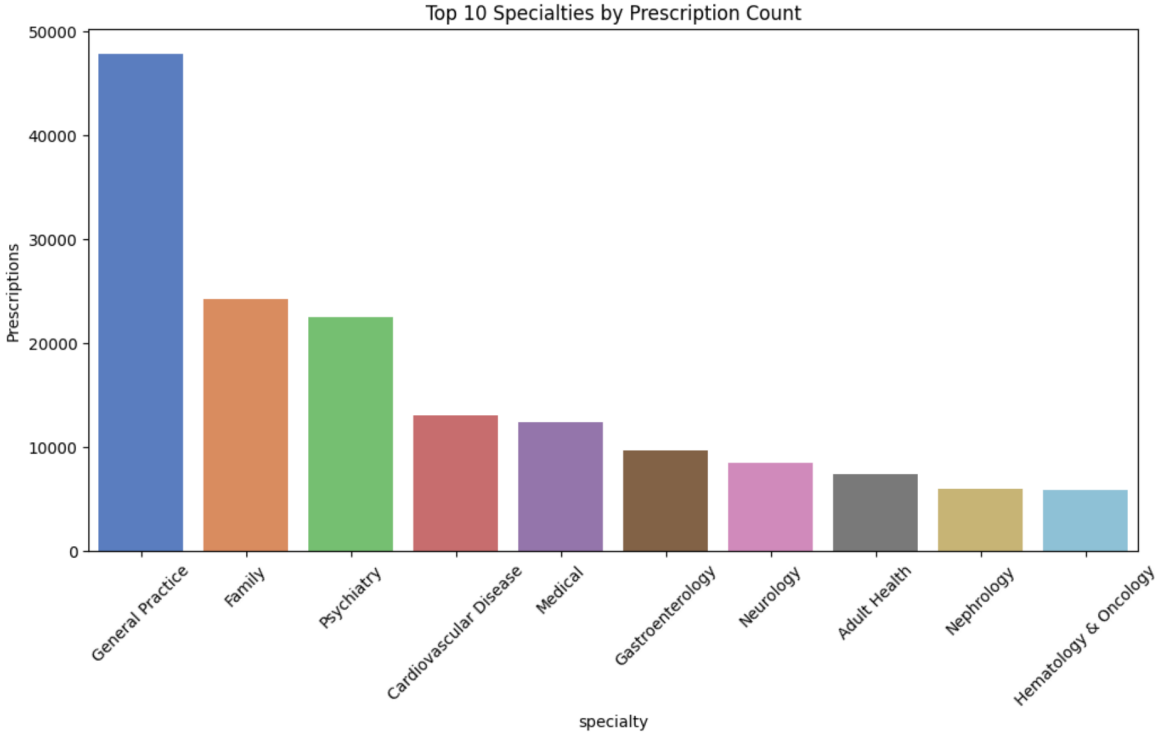
6.3 Sentiment Analysis: Positive sentiment aligned with higher ratings.



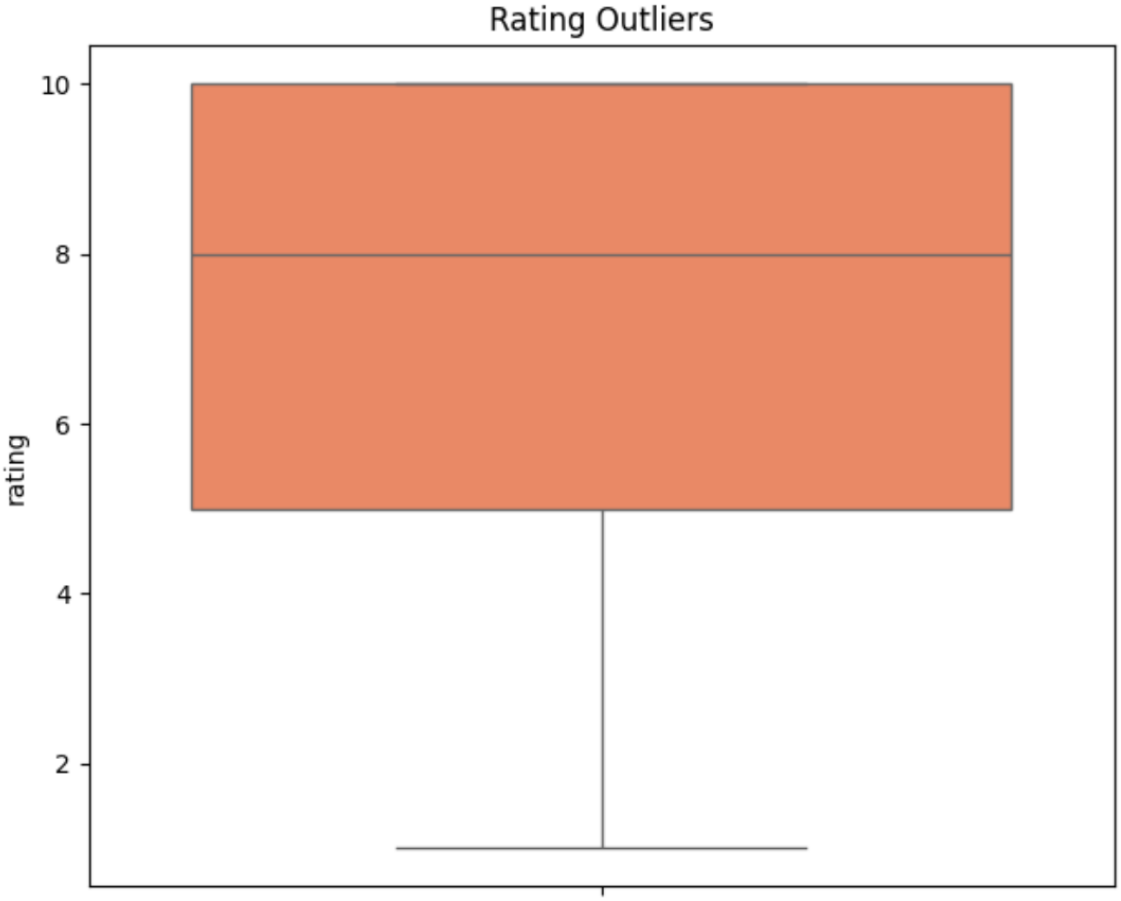
6.4 Correlation Analysis: Weak correlation between rating and sentiment score.



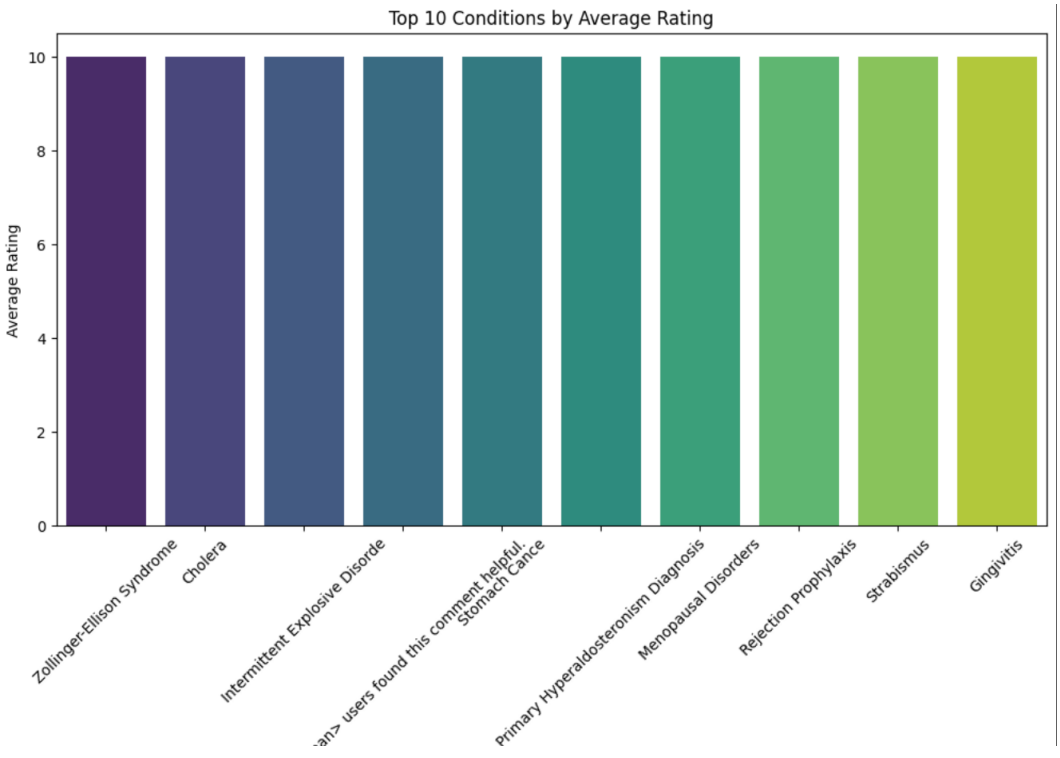
6.4.1 Prescription Based on Specialty: Drugs usage trends varied significantly among various specialities.



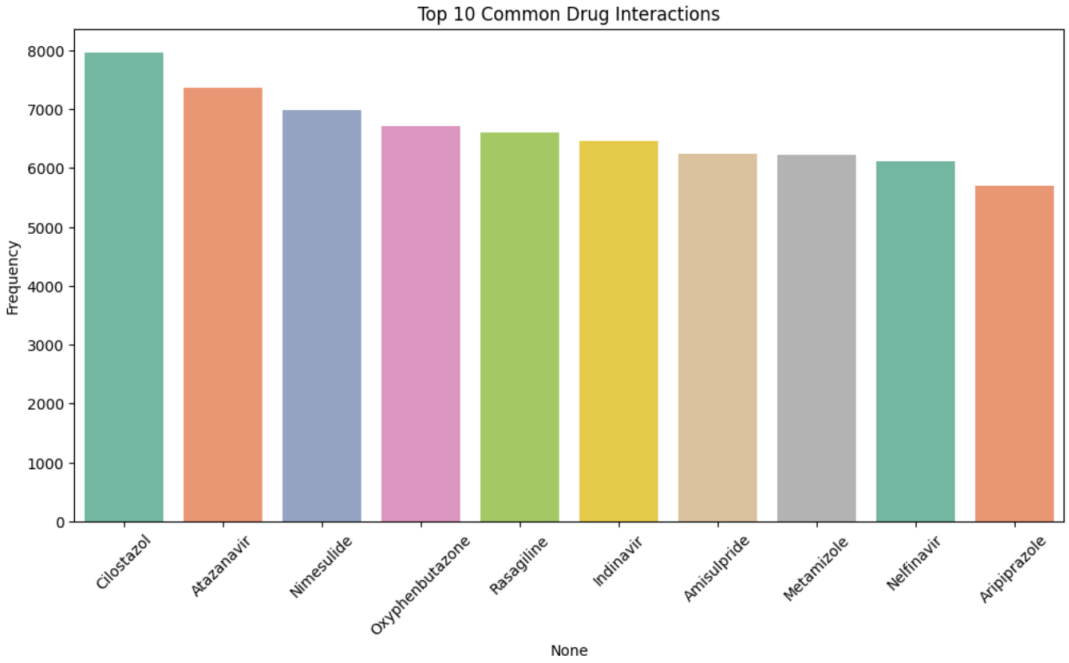
6.5 **Outlier Detection**: Outliers in product pricing and sentiment scores were analyzed.



6.6 **Top Conditions by Rating**: Top conditions experienced by patients were analyzed.



- **Top Drug Interactions by Rating**: Top drugs used by patients were analyzed.



7 Project Timeline

Milestone	Task	Timeline
Milestone 1	Data Collection, Preprocessing, EDA	Feb 5 - Feb 23, 2025
Milestone 2	Feature Engineering, Selection, Data Modeling	Feb 23 - Mar 21, 2025
Milestone 3	Evaluation, Tool Development, Presentation	Mar 21 - Apr 23, 2025

8 Key Insights and Next Steps

a. Sentiment analysis shows that positive reviews do not always correlate with high ratings.

b. Drug substitutes should be prioritized based on fewer side effects and pricing.

c. Manufacturer pricing strategies can inform cost-effective recommendations.

d. Future work includes advanced modeling techniques and NLP analysis.

9 GitHub Repository

• Repository: cap5771sp25-project
• Contents:
– Reports: Milestone1.pdf
– Notebooks: Preprocessing and EDA
– Scripts: Python data handling scripts
– README.md: Reproduction instructions
• Collaborators: TA Jimmy (@JimmyRaoUF), Grader Daniyal (@abbasidaniyal), Dr. Cruz (@lcruz-cas), Dr. Grant (@cegme)

10 Conclusion

Milestone 1 is complete. All deliverables are prepared for submission. The next phase will focus on data modeling and optimization to develop a robust and effective personalized medication recommendation system.