Which drugs are the most recommended for a specific condition?

Sepehr Rezaee
Department of Computer Science
Shahid Beheshti University
sepe.rezaee@mail.sbu.ac.ir

Mahdi Firouz
Department of Computer Science
Shahid Beheshti University
mahdi.firouz@mail.sbu.ac.ir

Abstract

This study presents a recommendation system for suggesting drugs for specific medical conditions. The system leverages the BERT (Bidirectional Encoder Representations from Transformers) model, a state-of-the-art Deep Neural Network, to understand the context of various features of drugs, including 'EaseOfUse', 'Effective', 'Price', 'Reviews', and 'Satisfaction'. The system generates BERT embeddings for these combined features and calculates the cosine similarity between different drugs to recommend the most suitable ones for a specific condition. The performance of the recommendation system is evaluated using Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) metrics. The results demonstrate the effectiveness of the proposed system in recommending drugs for specific conditions, showcasing the power of BERT in handling such tasks.

1 Introduction

Implementing a drug recommender system is a complex task that involves several aspects, including medical, technological, and patient perspectives. This document aims to examine each of these aspects in detail.

1.1 Medical Perspective

1.1.1 Accuracy

The accuracy of a drug recommender system is paramount. Inaccurate recommendations could lead to ineffective treatment or even harm the patient. The system should be trained and tested on a large and diverse dataset of patient information and drug responses to ensure its accuracy. This involves collecting data from a wide range of patients with different health conditions, demographics, and drug responses. The data should be cleaned and preprocessed to remove any errors or inconsistencies. The system should be trained using machine learning algorithms that can learn from this data and make accurate predictions. The performance of the system should be evaluated using appropriate metrics such as precision, recall, and F1 score. The system should also be regularly updated as new data becomes available.

1.1.2 Side Effects

Every drug has potential side effects. The system should consider these and aim to recommend drugs with the least harmful side effects. This involves having access to up-to-date and comprehensive

(LateX template borrowed from NeurIPS 2019) information about all potential drug side effects. The system should be able to analyze this information and weigh the potential benefits of a drug against its potential side effects. The system should also consider the patient's health profile and any potential interactions between the recommended drug and other drugs the patient is taking. The system should provide clear and understandable information about potential side effects to the patient and healthcare provider.

1.1.3 Personalized Recommendations

The system should consider the patient's previous health profile, lifestyle, and habits to make personalized recommendations. This requires a sophisticated algorithm capable of processing and making sense of large amounts of data. The algorithm should be able to analyze the patient's health history, genetic information, lifestyle factors such as diet and exercise, and other relevant information. The algorithm should also be able to learn from the patient's past responses to different drugs and use this information to make more accurate recommendations. The system should also allow the patient and healthcare provider to provide feedback on the recommendations, which can be used to further improve the system.

1.2 Technological Perspective

1.2.1 Data Privacy

Patient data used by the system must be kept private and secure. This involves implementing strong data encryption methods and ensuring that data is stored and transmitted securely. The system should comply with all relevant data privacy laws and regulations. The system should also provide clear and understandable information to the patient about how their data is being used and protected.

1.2.2 System Performance

The system should provide quick and reliable predictions. This requires efficient algorithms and high-performance computing resources. The system should be designed to handle large amounts of data and make predictions in real-time. The system should also be scalable, able to handle increasing amounts of data and users without a decrease in performance.

1.2.3 Algorithm Fairness

The system should be fair and not biased towards certain treatments. This involves careful algorithm design and regular audits of system performance to check for any signs of bias. The system should be transparent, providing clear and understandable explanations of how it makes its recommendations. The system should also be accountable, providing mechanisms for patients and healthcare providers to provide feedback and challenge the system's recommendations.

1.3 Patient Perspective

1.3.1 Trust

Patients need to trust the system. This can be achieved by making the system transparent and explainable. Patients should be able to understand how the system makes its recommendations. The system should also be accountable, providing mechanisms for patients to provide feedback and challenge the system's recommendations.

1.3.2 Accessibility

The system should be easy to use and accessible to all patients. This involves user-friendly interface design and ensuring that the system is available on multiple platforms. The system should also provide clear and understandable information to the patient about how to use the system and what to expect from it.

1.3.3 Personalization

Patients may appreciate a system that considers their personal circumstances and preferences. This involves designing the system to take into account individual patient characteristics and needs. The

system should allow patients to provide input on their preferences and goals, and use this information to make more personalized recommendations.
In conclusion, while a drug recommender system can be a powerful tool in healthcare, it's crucial to carefully consider these aspects during its implementation to ensure it's beneficial and safe for patients. It's also important to regularly evaluate and update the system as new data and research become available. Please note that while such systems can assist healthcare professionals, they are not intended to replace professional medical advice. Always consult with a healthcare provider for medical advice.

2 Related Work/Background

Several recent works have proposed different approaches to overcome the challenges associated with drug recommendation systems. These works have found solutions for different settings of this problem, such as providing personalized drug recommendations, enabling shared decision-making, and ensuring fair and safe usage of drugs.

2.1 Recommender Systems in the Healthcare Domain

This research provides a systematic overview of existing research on healthcare recommender systems, including drug recommendation systems. The authors discuss various recommendation scenarios and approaches, such as food recommendation, drug recommendation, health status prediction, healthcare service recommendation, and healthcare professional recommendation. They also develop working examples to give a deep understanding of recommendation algorithms. This work highlights the need for recommender systems in the healthcare domain to help both end-users and medical professionals make more efficient and accurate health-related decisions (1).

2.2 A Pharmaceutical Therapy Recommender System Enabling Shared Decision-Making

This work demonstrates an exemplary Clinical Decision Support System (CDSS) which provides individualized pharmaceutical drug recommendations to physicians and patients. The core of the proposed system is a neighborhood-based collaborative filter (CF) that yields data-based recommendations. CFs are capable of integrating data at different scale levels and a multivariate outcome measure. The authors provide a detailed literature review, a holistic comparison of various implementations of CF algorithms, and a prototypical graphical user interface (GUI). They show that similarity measures, which automatically adapt to attribute weights and data distribution perform best (2).

2.3 A Fair and Safe Usage Drug Recommendation System in Medical

This work discusses four types of recommender systems, including content-driven filtering, collaborative filtering, knowledge-driven recommender systems, and hybrid recommender systems. The authors highlight that since the drug recommendation framework includes medical terminology, such as infection names, side effects, and synthetic names, only a limited number of papers are available (3).

2.4 Future Directions

While significant progress has been made in the field of drug recommendation systems, there are still many challenges and opportunities for future research. For instance, the integration of real-time patient data, such as electronic health records and wearable device data, could further personalize drug recommendations. Moreover, the application of advanced machine learning techniques, such as deep learning, could potentially improve the accuracy of these systems. Lastly, addressing issues related to data privacy and security will be crucial as these systems continue to evolve. (4)

3 Proposed method

This section is based on a comprehensive dataset that provides performance metrics for drugs associated with 37 common medical conditions. The dataset includes a variety of variables that offer insights into the drug's characteristics and its perceived effectiveness based on customer reviews.

The variables in the dataset are as follows:

1. **Condition**: This is a categorical variable that represents the medical condition associated with the drug.
2. **Drug**: This is a categorical variable that represents the name of the drug.
3. **EaseOfUse**: This is a numerical variable that represents the ease of use of the drug based on customer reviews.
4. **Effective**: This is a numerical variable that represents the effectiveness of the drug based on customer reviews.
5. **Indication**: This is a categorical variable that represents the purpose of the drug.
6. **Reviews**: This is a numerical variable that represents the number of reviews associated with the drug.
7. **Satisfaction**: This is a numerical variable that represents the satisfaction level of the drug based on customer reviews.
8. **Type**: This is a categorical variable that represents the type of drug (generic or brand).
9. **Form**: This is a categorical variable that represents the form of the drug (e.g. tablet, capsule, etc.).
10. **Price**: This is a numerical variable that represents the average price of the drug.

The objective of the analysis is to derive meaningful insights from these variables, which could potentially inform healthcare professionals and patients about the performance of different drugs for various medical conditions.

3.1 Descriptive Statistics

In this section, we perform a detailed statistical analysis of the dataset. The analysis is divided into two parts: numerical and categorical data analysis.

3.1.1 Numerical Data Analysis

For the numerical variables, we calculate the following statistical measures:

- **EaseOfUse**: Mean = 3.92, Median = 4.05, Mode = 5.0, Standard Deviation = 0.89
- **Effective**: Mean = 3.52, Median = 3.6, Mode = 5.0, Standard Deviation = 0.95
- **Reviews**: Mean = 82.64, Median = 10.35, Mode = 1.0, Standard Deviation = 273.28
- **Satisfaction**: Mean = 3.20, Median = 3.2, Mode = 5.0, Standard Deviation = 1.03
- **Price**: Mean = 174.21, Median = 49.99, Mode = 11.99, Standard Deviation = 667.74

We also generate visualizations for these numerical variables:

- **Histograms**: These provide a visual representation of the data distribution for each numerical variable. The x-axis represents the range of values and the y-axis represents the frequency of occurrence.
- **Box Plots**: These are used to visually represent the dispersion and skewness of the numerical data. They show the median, quartiles, and potential outliers in the data.
- **Pair Plots**: These are used to visualize the relationship between each pair of numerical variables. They can help identify correlations and patterns in the data.

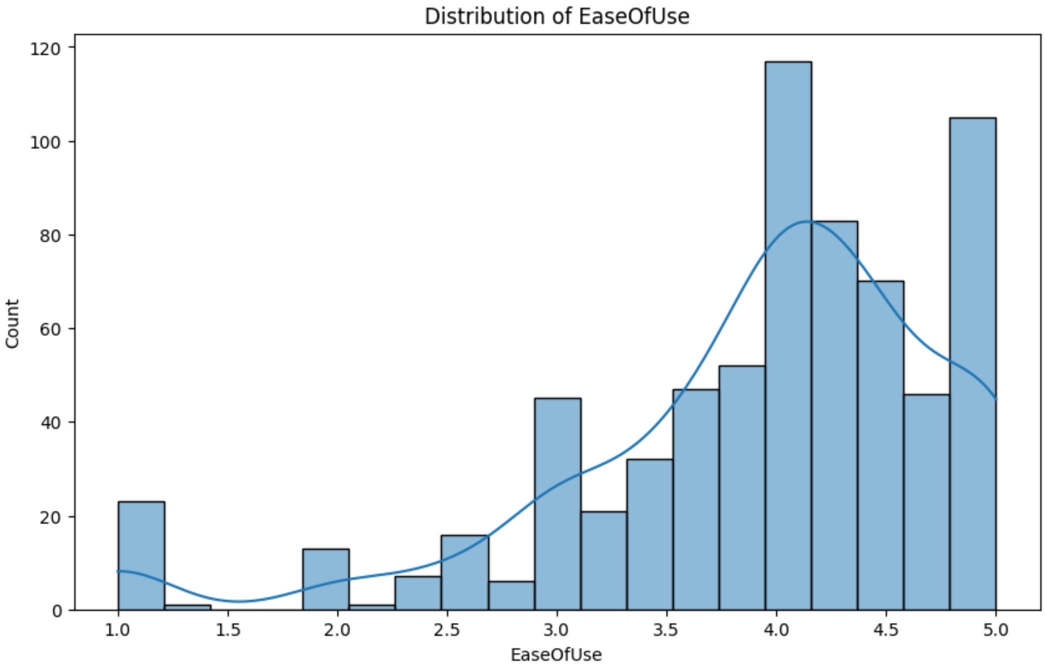

Figure 1: Histogram with Kernel Density Estimate Overlay Demonstrating the Bimodal Distribution of ‘EaseOfUse’ Ratings. The distribution exhibits two modes around ratings of 3.5 and 5.0, indicating a significant proportion of users find the product either moderately easy or extremely easy to use.

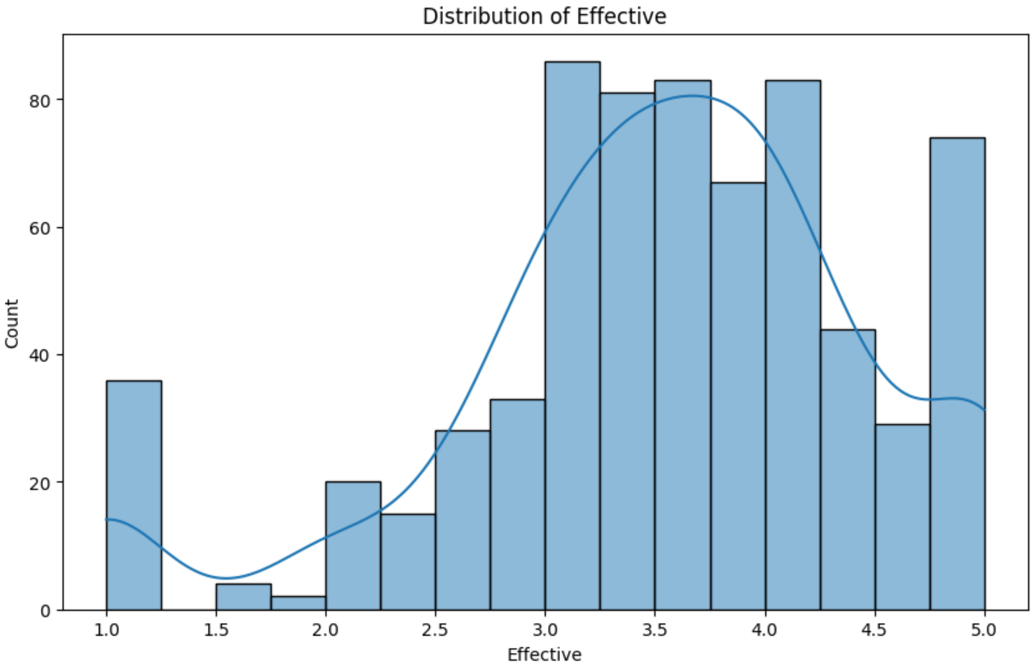

Figure 2: Histogram and Density Plot Depicting the Distribution of ‘Effective’ Ratings. The distribution demonstrates a range of effectiveness ratings from 0 to 5, with the highest frequency observed around the rating of 3. The overlaid density plot provides a smooth estimate of the distribution, highlighting the overall pattern of ‘Effective’ ratings.

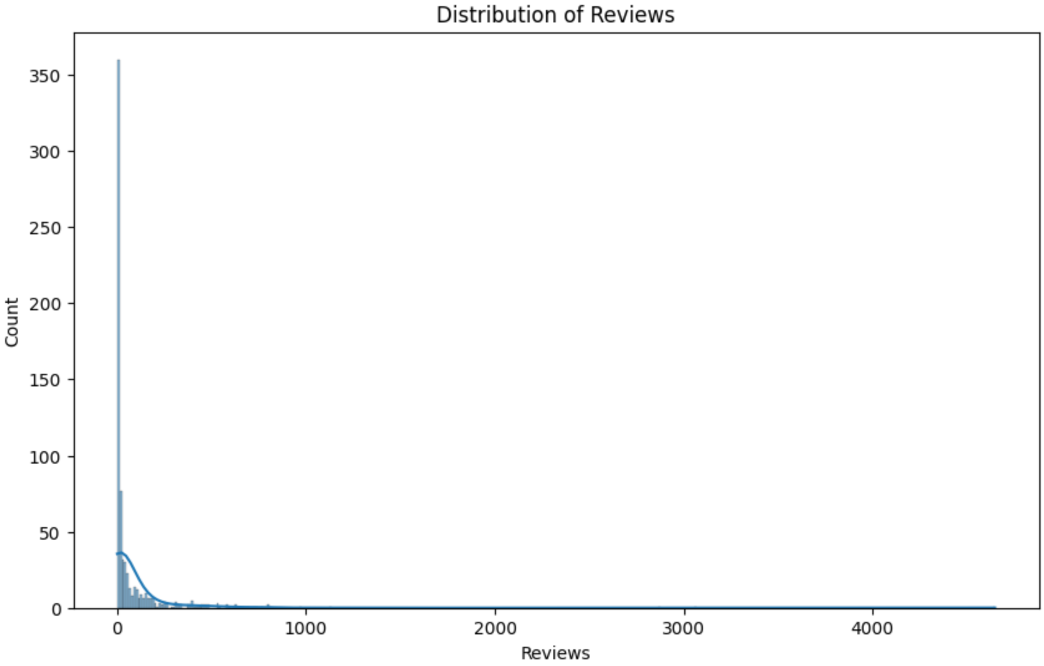

Figure 3: Frequency Distribution Plot of ‘Reviews’. The plot illustrates a highly skewed distribution of reviews, with a majority clustered near zero. This indicates that a significant proportion of items have a low number of reviews. The x-axis represents the number of reviews, ranging from 0 to 4000, while the y-axis represents the count, ranging from 0 to 350. The rapid decline in count as the number of reviews increases underscores the prevalence of items with fewer reviews.

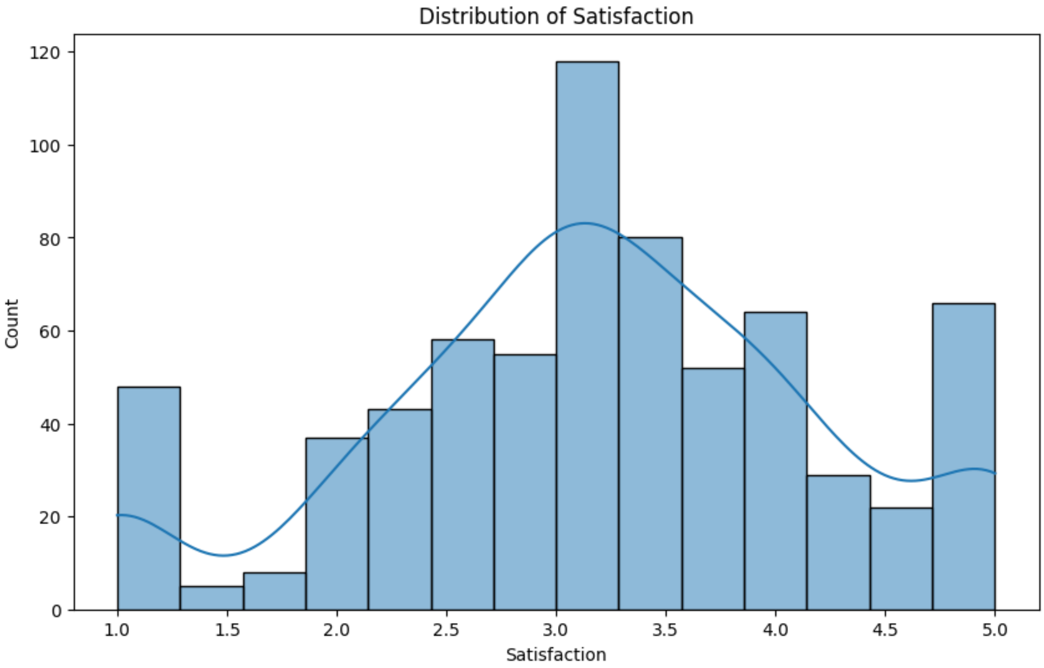

Figure 4: Histogram with Frequency Polygon Depicting the Distribution of ‘Satisfaction’ Scores. The histogram illustrates the frequency of satisfaction scores on a scale from 1 to 5, with the count of respondents on the y-axis. A notable peak is observed at a satisfaction score of approximately 3.5, indicating a significant proportion of respondents with this score. The overlaid frequency polygon provides a smooth estimate of the distribution, highlighting the overall pattern of ‘Satisfaction’ scores.

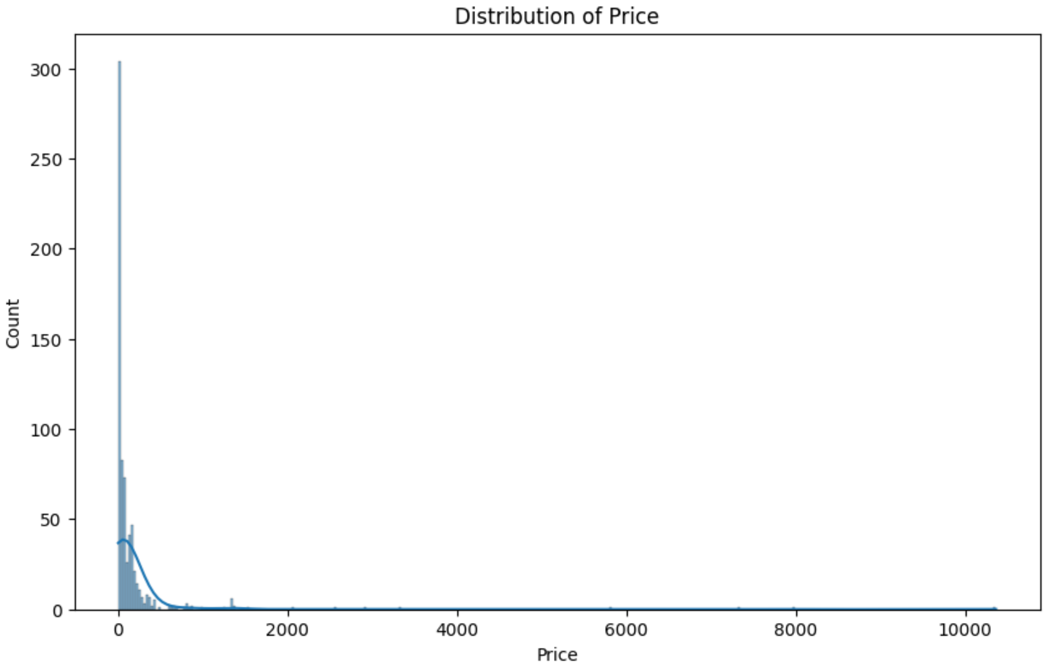

Figure 5: Histogram Depicting the Distribution of ‘Price’. The histogram illustrates a pronounced skewness towards the lower end of the price spectrum, with a significant concentration of data around lower prices. The x-axis represents various price points, ranging from 0 to 10,000, while the y-axis indicates their corresponding counts. This visualization provides critical insights into pricing trends and patterns.

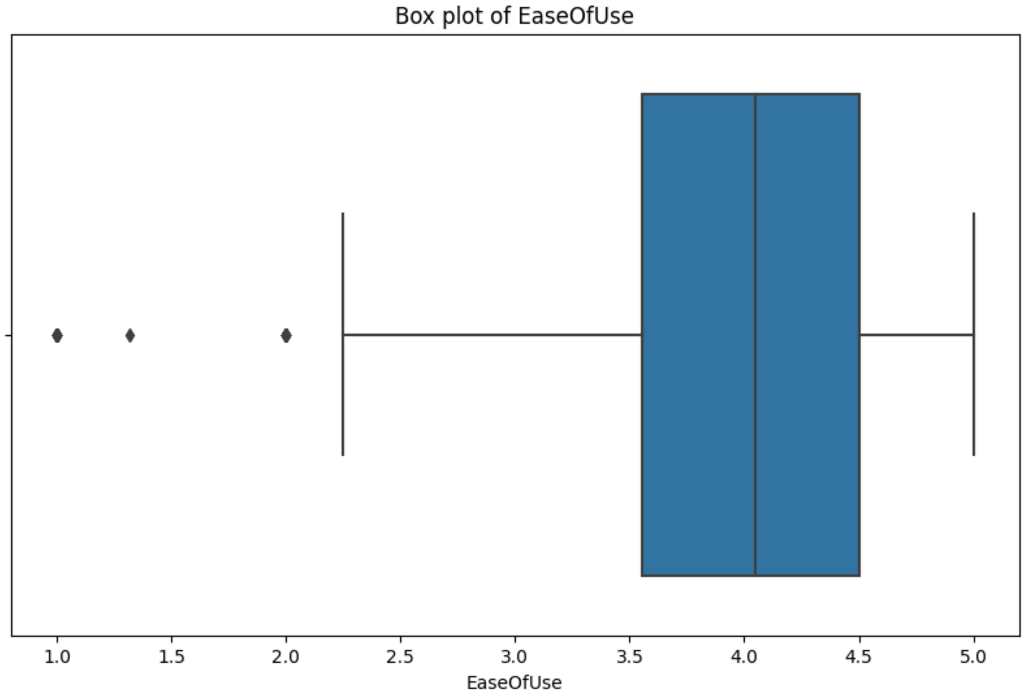

Figure 6: Box Plot of ‘EaseOfUse’. This box plot delineates the distribution of user ratings concerning the ‘Ease of Use’ of a particular interface. The x-axis represents the ‘EaseOfUse’ scores ranging from 1.0 to 5.0. The interquartile range, shaded in blue, indicates where the bulk of the data points lie, approximately between 3.5 and 4.5. The presence of three outliers on the left side, with values approximately between 1.0 and 2.0, suggests a few instances of low ‘Ease of Use’ ratings. The absence of upper or lower whiskers suggests limited variance among these particular data points.

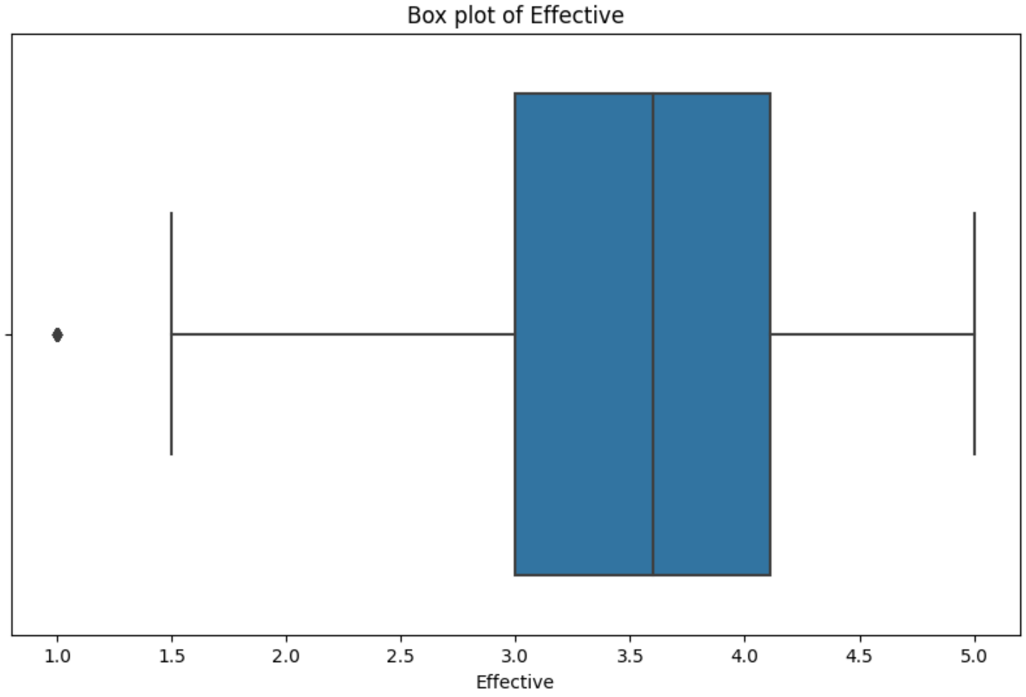

Figure 7: Box Plot of ‘Effective’ Ratings. This box plot illustrates the distribution of ‘Effective’ ratings. The x-axis represents the ‘Effective’ scores ranging from 1.0 to 5.0. The interquartile range, shaded in blue, indicates where the bulk of the data points lie, approximately between 3.0 and 4.5. The median line, depicted in white, divides the blue box near its center. The horizontal lines extending from either side of the blue box represent the range within which the bulk of the values fall. A single black dot to the left indicates an outlier in the data at around 1.5 on the ‘Effective’ scale.

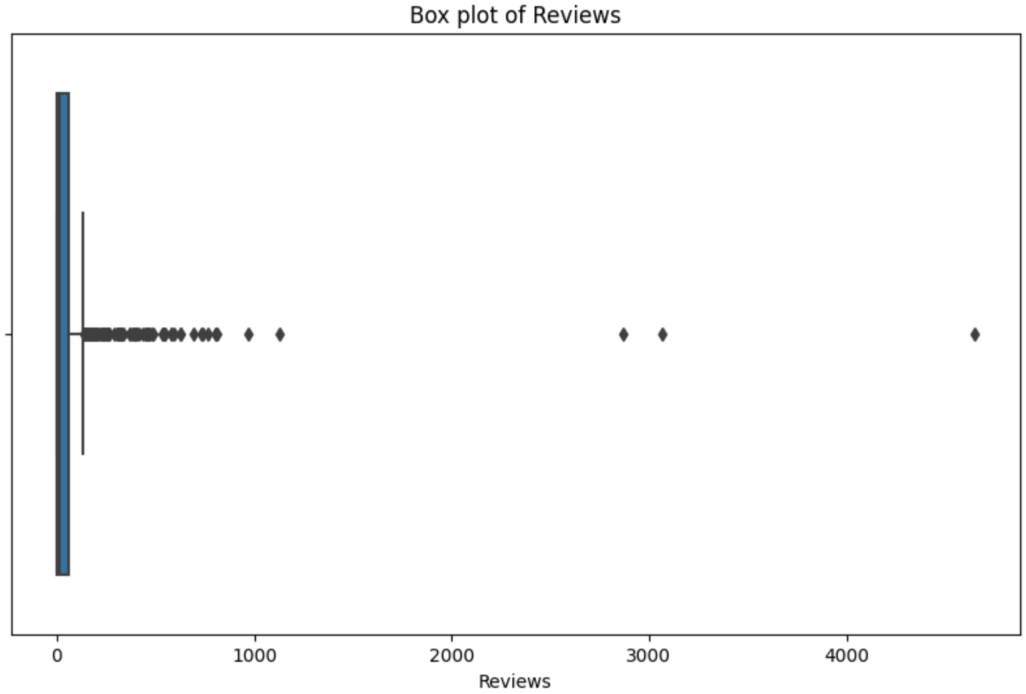

Figure 8: Box Plot of ‘Reviews’. This box plot illustrates the distribution and dispersion of reviews. The x-axis represents the ‘Reviews’ with markers at intervals of 1000 up to 4000. The interquartile range (IQR), located near the y-axis, represents the middle 50% of the data. The median is represented by a line inside the IQR box, indicating the middle value of the data. Several data points plotted as small circles to the right of the IQR box indicate potential outliers in the review data, suggesting a few products or services have significantly more reviews than others.

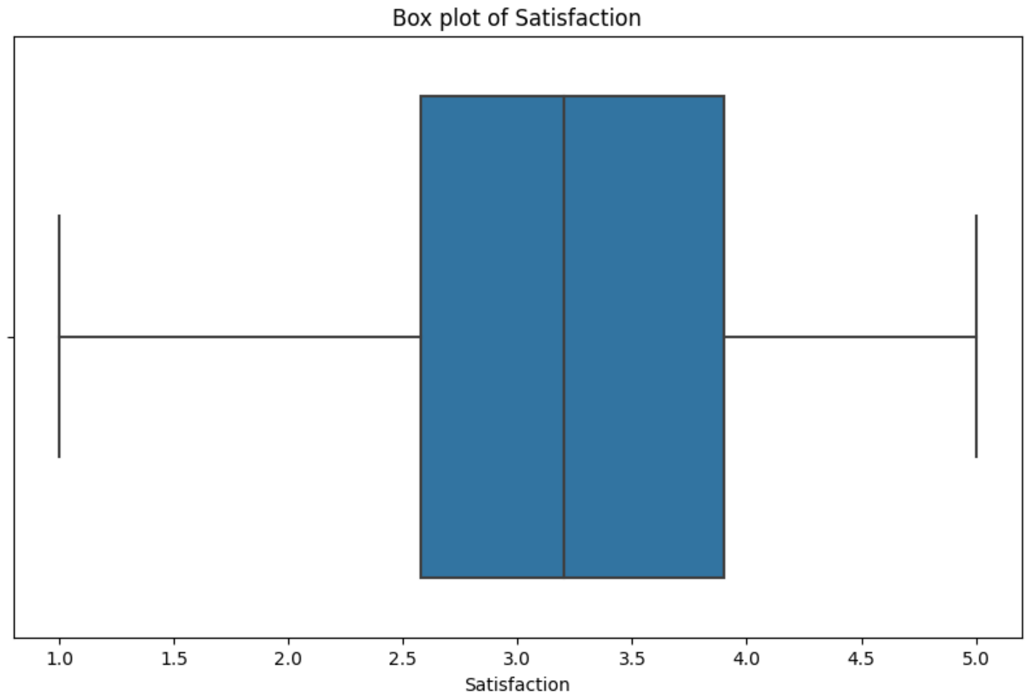

Figure 9: Box Plot of ‘Satisfaction’ Scores. This box plot illustrates the distribution and dispersion of satisfaction scores, quantitatively measured on a scale from 1 to 5. The interquartile range, highlighted in blue, represents the middle 50% of the data, offering insights into the central tendency and variability of satisfaction levels among the surveyed population. The absence of visible outliers suggests a relatively uniform satisfaction level within the surveyed population.

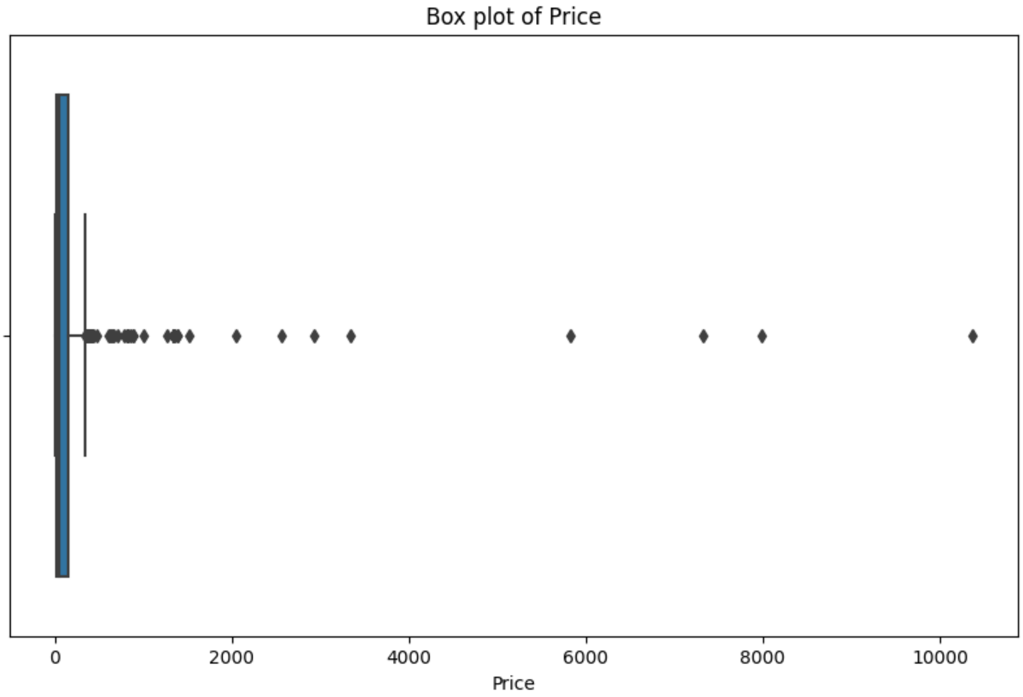

Figure 10: Box Plot of ‘Price’. This box plot illustrates the distribution and dispersion of prices. The x-axis represents the ‘Price’ with markers at intervals of 2000 up to 10000. The interquartile range, highlighted in blue, represents the middle 50% of the data, offering insights into the central tendency and variability of prices among the surveyed population. The black line within the blue box indicates the median price. Multiple black dots to the right of the IQR represent outliers in price, suggesting a few products or services have significantly higher prices than others.

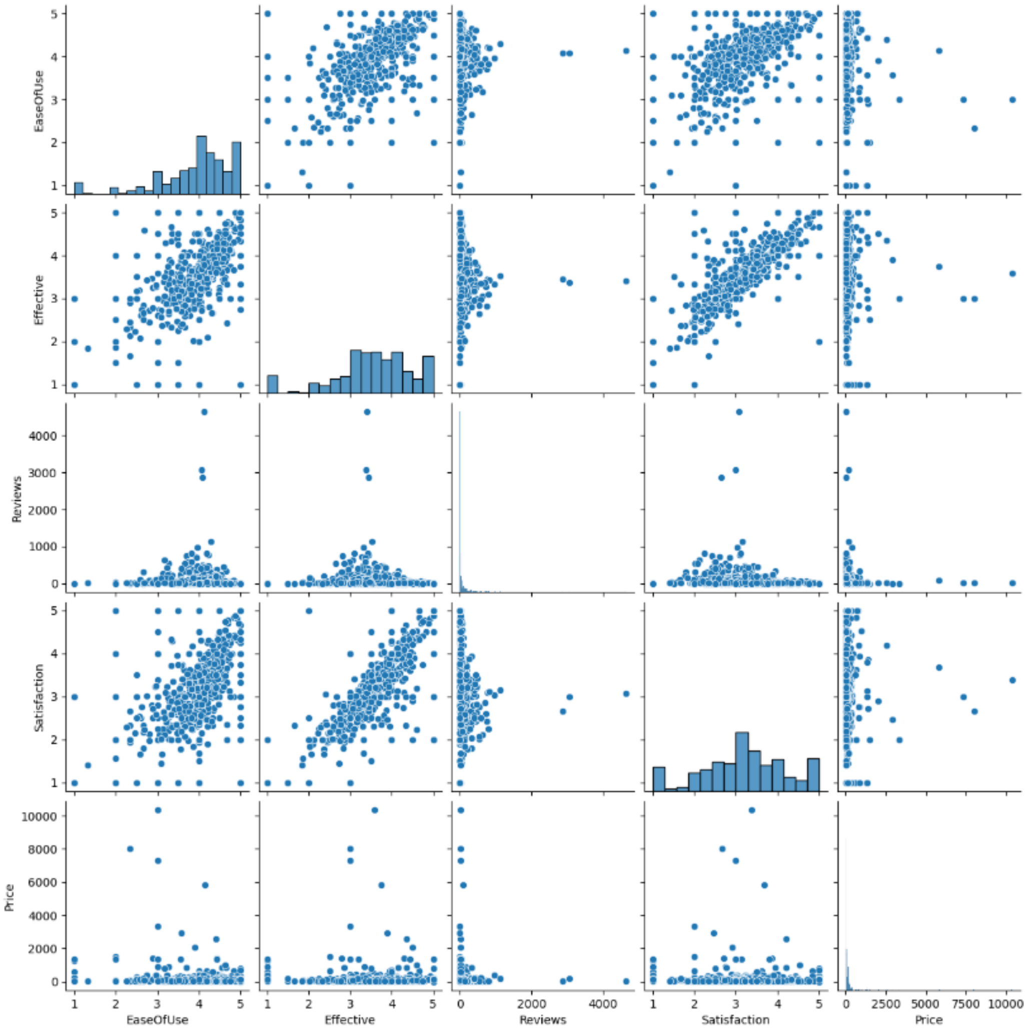

Figure 11: Scatterplot Matrix of Excuse, Effective, Reviews, Satisfaction, and Price Variables. This comprehensive scatterplot matrix illustrates the correlations and distributions amongst Excuse, Effective, Reviews, Satisfaction, and Price variables. Each cell in the 5x5 grid represents a plot between two variables, with the diagonal from the top left to bottom right containing histogram plots for each variable respectively. The scatterplots show relationships between pairs of variables, such as a positive correlation between 'Effective' and 'Satisfaction', and a concentration of data points at lower values for both 'Excuse' and 'Effective'. The absence of a clear correlation between 'Price' and 'Reviews' is also noteworthy.

3.1.2 Categorical Data Analysis

For the categorical variables, we calculate the following:

- **Condition**: Mode = Hypertension, Unique Values = 37
- **Drug**: Mode = Niacin, Unique Values = 470
- **Indication**: Mode = On Label, Unique Values = 3
- **Type**: Mode = RX, Unique Values = 4
- **Form**: Mode = Tablet, Unique Values = 6

We also generate count plots for these categorical variables, which provide a visual representation of the frequency of each category. This can help identify the most common categories and any imbalances in the data.

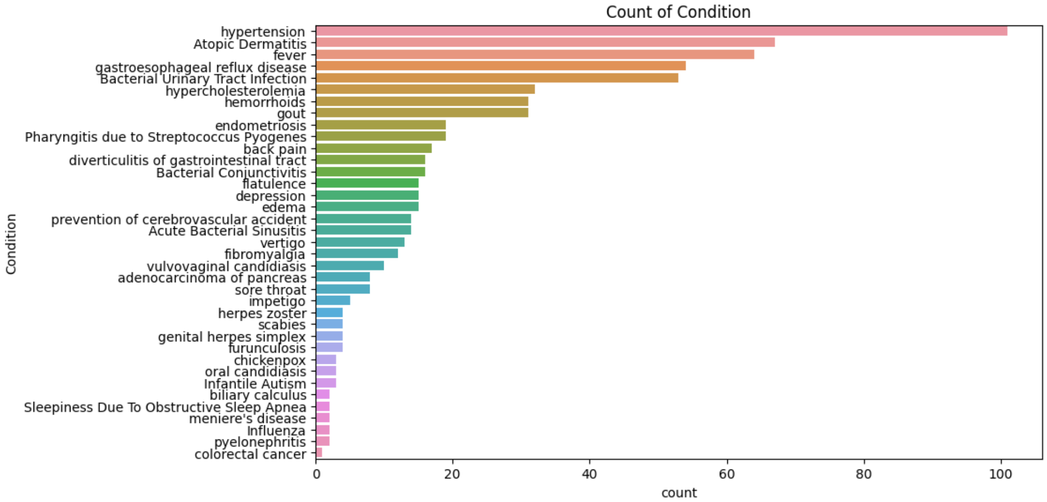

Figure 12: Bar Graph of the Prevalence of Various Medical Conditions. This bar graph elucidates the count of distinct medical conditions, offering a quantitative insight that is instrumental for healthcare professionals and researchers in understanding the prevalence and prioritizing healthcare resources accordingly. The x-axis represents the count ranging from 0 to 100, while the y-axis lists various medical conditions including hypertension, atopic dermatitis, gastroesophageal reflux disease, and many others totaling to 39 different conditions. Each condition is represented by a colored bar indicating its count, with hypertension having the highest count, followed by atopic dermatitis and gastroesophageal reflux disease.

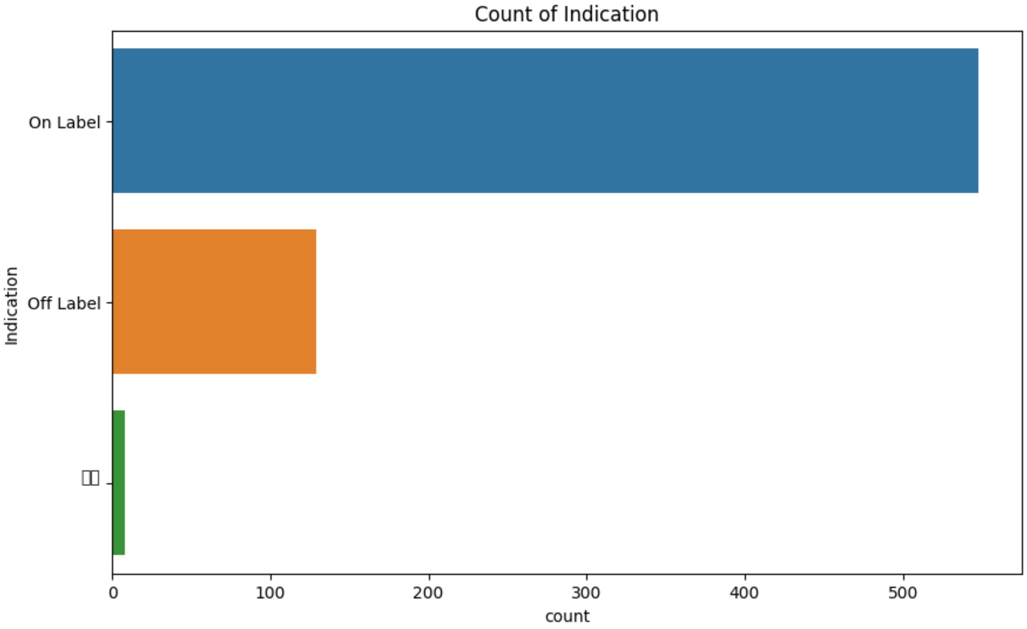

Figure 13: Comparative Analysis of On-Label and Off-Label Indications. This bar graph represents the count of indications for ‘On Label’ and ‘Off Label’ categories. The x-axis represents the count, ranging from 0 to over 500, while the y-axis represents the indication categories. The ‘On Label’ category, represented by a blue bar, extends beyond 500 counts, indicating a higher prevalence. The ‘Off Label’ category, represented by an orange bar, extends to approximately 150 counts. A third category, represented by a green bar, is unlabeled and has a minimal count close to zero.

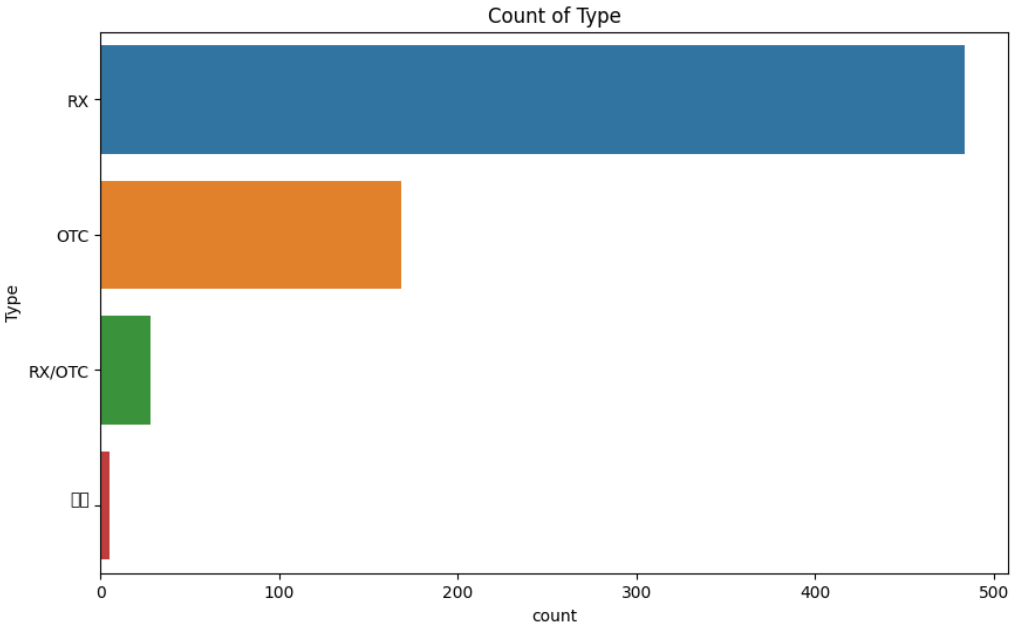

Figure 14: Comparative Analysis of Prescription (RX), Over-The-Counter (OTC), and Combined RX/OTC Medications. This horizontal bar graph illustrates the count of each type of medication. The x-axis represents the count, ranging from 0 to 500, while the y-axis represents the type of medication with categories RX, OTC, and RX/OTC. The RX category, represented by a blue bar, extends past 400 on the count axis, indicating a higher prevalence. The OTC category, represented by an orange bar, extends just beyond 100 on the count axis. The RX/OTC category, represented by a green bar, barely reaches 50 on the count axis, suggesting a lower prevalence of combined RX/OTC medications.

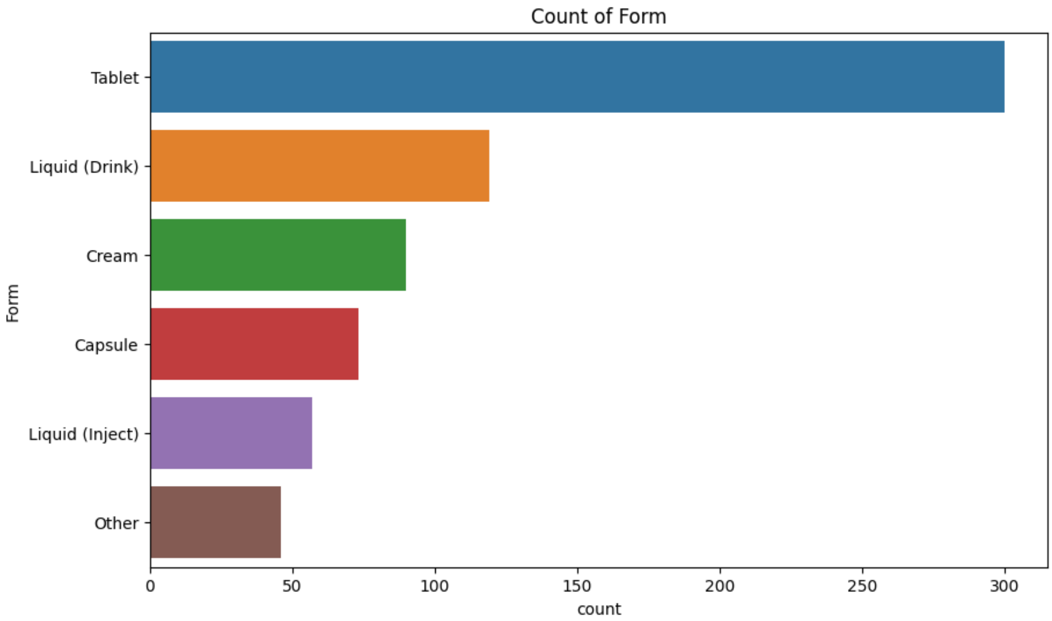

Figure 15: Comparative Analysis of Various Forms of Pharmaceuticals. This horizontal bar graph represents a quantitative comparison of various forms of pharmaceuticals, categorized by their physical state and method of administration. The x-axis denotes the count, ranging from 0 to 300, while the y-axis categorizes the data into Tablet, Liquid (Drink), Cream, Capsule, Liquid (Inject), and Other. The graph illustrates that tablets are the most prevalent form, followed by Liquid (Drink), Cream, Capsule, Liquid (Inject), and Other forms respectively.

This comprehensive analysis provides a deep understanding of the data and forms the basis for further analysis and hypothesis testing.

3.2 Data Grouping

In this section, we group the data based on different categorical variables and calculate the mean for the numerical variables within each group. This allows us to understand the data at a more granular level and identify patterns within specific groups.

- **Grouped by Condition**: The data is grouped by the ’Condition’ variable, and the mean of the numerical variables (’EaseOfUse’, ’Effective’, ’Reviews’, ’Satisfaction’, and ’Price’) is calculated for each unique condition.
- **Grouped by Drug**: The data is grouped by the ’Drug’ variable, and the mean of the numerical variables is calculated for each unique drug.
- **Grouped by Type**: The data is grouped by the ’Type’ variable, and the mean of the numerical variables is calculated for each unique type.
- **Grouped by Form**: The data is grouped by the ’Form’ variable, and the mean of the numerical variables is calculated for each unique form.
- **Grouped by Condition, Drug, Type, and Form**: The data is grouped by multiple variables (’Condition’, ’Drug’, ’Type’, and ’Form’), and the mean of the numerical variables is calculated for each unique combination of these variables.

We also perform additional operations on the grouped data:

- **Sorting**: The groups can be sorted based on a specific column. For example, the ’Drug’ groups are sorted by the ’Price’ column to identify the drugs with the highest and lowest average prices.

- **Filtering**: The groups can be filtered based on a condition. For example, we filter the 'Drug' groups to only include those where the average 'Satisfaction' is greater than 4.

Table 1: Summary of User Ratings Grouped by Medical Condition

Condition	EaseOfUse	Effective	Reviews
Acute Bacterial Sinusitis	3.19	3.15	136.17
Atopic Dermatitis	3.96	3.62	32.45
Bacterial Conjunctivitis	3.70	2.99	11.12
Bacterial Urinary Tract Infection	3.43	3.23	75.55
Infantile Autism	3.85	3.30	117.08
Influenza	3.08	2.83	196.50
Pharyngitis due to Streptococcus Pyogenes	3.47	3.17	70.33
Sleepiness Due To Obstructive Sleep Apnea	4.41	4.05	258.25
adenocarcinoma of pancreas	2.89	3.25	20.48
back pain	4.19	4.07	7.73
biliary calculus	4.45	4.27	30.00
chickenpox	4.13	4.46	99.33
colorectal cancer	1.00	3.00	5.00
depression	4.10	3.51	124.06
diverticulitis of gastrointestinal tract	2.97	2.87	116.63
edema	3.73	3.19	50.34
endometriosis	4.21	3.72	40.71
fever	4.01	3.58	23.04
fibromyalgia	4.20	3.58	876.31
flatulence	4.45	3.75	2.62
furunculosis	4.56	3.27	22.88
gastroesophageal reflux disease	4.21	3.74	50.21
genital herpes simplex	3.74	3.50	72.33
gout	4.32	3.79	107.60
hemorrhoids	3.82	3.58	6.00
herpes zoster	3.94	3.95	90.50
hypercholesterolemia	3.72	3.54	113.00
hypertension	4.00	3.61	123.14
impetigo	4.07	3.18	50.80
meniere's disease	4.87	4.57	4.10
oral candidiasis	3.99	3.18	83.88
prevention of cerebrovascular accident	4.04	3.51	204.85
pyelonephritis	3.54	2.34	199.14
scabies	4.08	2.96	18.75
sore throat	4.05	3.50	2.89
vertigo	4.43	3.60	34.53
vulvovaginal candidiasis	4.22	3.29	45.29

Table 2: Summary of User Ratings Grouped by Drug

Drug	EaseOfUse	Effective	Reviews	Satisfaction
ASA-Acetaminophen-Salicyl-Caff	1.99	2.08	9.50	2.05
ASA-Acetaminophen-Salicyl-Caff, Mg Salicylat-Ac...	3.00	2.60	6.00	2.80
Acebutolol	4.33	3.75	14.50	4.20
Acetaminophen	3.80	3.22	8.49	3.20
Acetaminophen-Caffeine	4.88	5.00	11.00	4.88
Vit E-Glycerin-Dimethicone	5.00	4.00	1.00	5.00
Vit E-Glycerin-Dimethicone, Glycerin-Dimethicon...	3.00	2.00	3.00	2.50
Vit E-Grape-Hyaluronate Sodium	4.54	3.69	22.00	3.77
Zanamivir	2.35	2.13	10.00	2.00
Zinc Oxide	1.00	1.00	1.00	1.00

Table 3: Summary of User Ratings Grouped by Drug, Sorted by Price

Drug	EaseOfUse	Effective	Reviews	Satisfaction
Cefazolin	2.33	2.44	7.0	1.94
Al Hyd-Mg Tr-Alg Ac-Sod Bicarb, Magaldrate, Alu...	5.00	5.00	1.0	5.00
Cefazolin In Dextrose (Iso-Os)	3.00	2.00	1.0	1.00
Al Hyd-Mg Tr-Alg Ac-Sod Bicarb	2.00	2.00	1.0	1.00
Magaldrate	5.00	5.00	1.0	5.00
Nafarelin	3.00	3.00	5.0	2.00
Capecitabine	4.14	3.75	83.0	3.69
Hydrocortisone-Iodoquinl-Aloe2	3.00	3.00	14.0	3.00
Tigecycline	2.33	3.00	9.0	2.67
Paclitaxel-Protein Bound	3.00	3.59	14.0	3.39

Table 4: Summary of User Ratings and Prices for Drugs with High Satisfaction

Drug	EaseOfUse	Effective	Reviews	Satisfaction	Price
Acebutolol	4.33	3.75	14.50	4.20	24.49
Acetaminophen-Caffeine	4.88	5.00	11.00	4.88	9.99
Acetaminophen-Pamabrom, Ibuprofen	5.00	5.00	2.00	5.00	13.49
Al Hyd-Mg Tr-Alg Ac-Sod Bicarb, Magaldrate, Alu...	5.00	5.00	1.00	5.00	5.99
Alum-Mag Hydroxide-Simeth, Calcium Carbonate-Si...	5.00	5.00	2.00	5.00	11.99
Telmisartan-Amlodipine	4.45	4.27	14.00	4.32	155.99
Trolamine Salicylate-Aloe Vera	5.00	5.00	1.00	5.00	11.99
Ursodiol	4.45	4.27	30.00	4.15	84.89
Verapamil	4.71	4.13	42.33	4.03	124.36
Vit E-Glycerin-Dimethicone	5.00	4.00	1.00	5.00	14.50

This grouped analysis provides a deeper understanding of the data and can reveal insights that may not be apparent from the overall statistics.

3.3 Correlation Analysis

In this section, we perform a correlation analysis to understand the relationships between different variables in the dataset. The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. The range of values for the correlation coefficient is -1.0 to 1.0. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables.

3.3.1 Correlation Matrix

We first calculate the correlation matrix for the numerical variables, which include ’EaseOfUse’, ’Effective’, ’Reviews’, ’Satisfaction’, and ’Price’. The correlation matrix is a table showing correlation coefficients between many variables. Each cell in the table shows the correlation between two variables.

Table 5: Correlation Matrix of User Ratings and Price

	EaseOfUse	Effective	Reviews	Satisfaction	Price
EaseOfUse	1.000000	0.659237	0.011962	0.650156	-0.107480
Effective	0.659237	1.000000	-0.035802	0.864863	-0.017532
Reviews	0.011962	-0.035802	1.000000	-0.084216	-0.024927
Satisfaction	0.650156	0.864863	-0.084216	1.000000	-0.024800
Price	-0.107480	-0.017532	-0.024927	-0.024800	1.000000

3.3.2 Heatmap Visualization

To better visualize the correlation matrix, we use a heatmap. A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. This provides an immediate visual summary of the information.

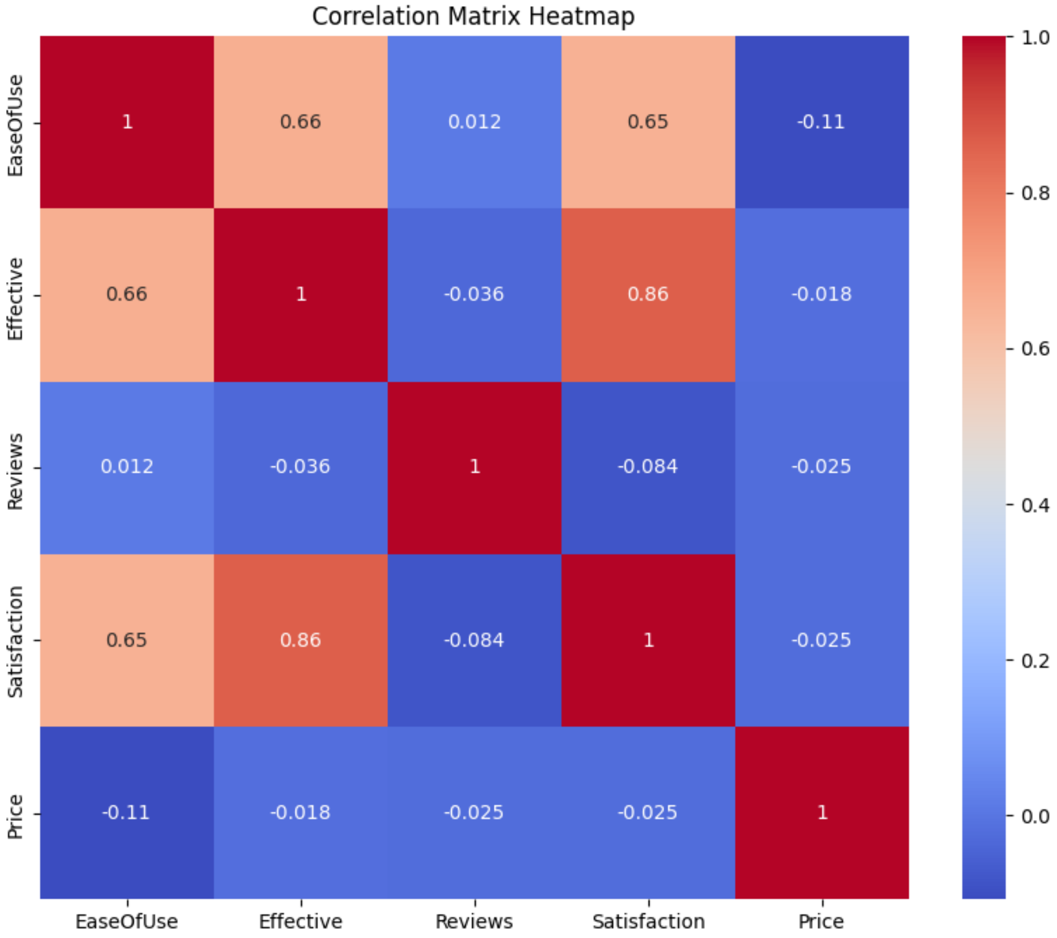

Figure 16: Correlation Matrix Heatmap of User Experience Metrics. This heatmap visualizes the correlations amongst five different variables: Ease of Use, Effective, Reviews, Satisfaction, and Price. Each cell in the matrix displays the correlation coefficient between two variables, with colors ranging from red (positive correlation) to blue (negative correlation). Notable correlations include a strong positive correlation between Effectiveness and Satisfaction (0.86), and Ease of Use and Satisfaction (0.65), and a mild negative correlation between Ease of Use and Price (-0.11).

3.3.3 Encoding Non-numerical Columns

Before we can calculate the correlation matrix for the entire dataset, we need to convert the non-numerical columns into numerical values. This is because the correlation coefficient can only be calculated between numerical variables. We use label encoding to convert the categorical values into numerical values.

3.3.4 Correlation Matrix for Entire Dataset

After encoding the non-numerical columns, we calculate the correlation matrix for the entire dataset and visualize it using a heatmap.

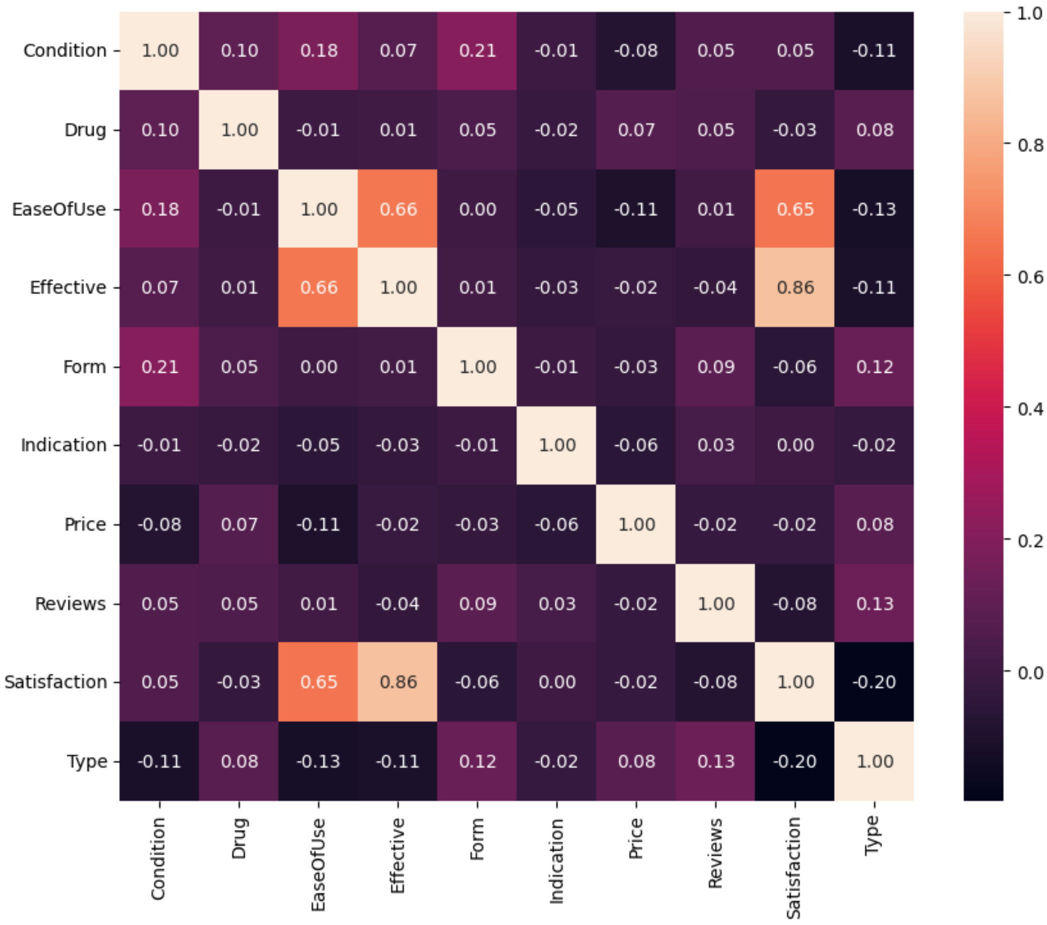

Figure 17: Correlation Heatmap of Various Factors Influencing Drug Evaluation Metrics. This heatmap illustrates the relationships among various factors including Condition, Drug, EaseOfUse, Effective, Form, Indication, Price, Reviews, Satisfaction, and Type. Each cell in the heatmap represents the correlation coefficient between two variables, with color intensity indicating the strength and direction of the relationship. Dark purple indicates a strong negative correlation, while dark orange indicates a strong positive correlation. Neutral correlations are represented in lighter shades.

This correlation analysis provides valuable insights into the relationships between different variables in the dataset and can inform further analysis and modeling.

3.4 Visualization

In this section, we generate a variety of visualizations to better understand the data and the relationships between different variables.

3.4.1 Bar Chart of Average Satisfaction Rating for Each Condition

We create a bar chart that shows the average 'Satisfaction' rating for each 'Condition'. This visualization helps us understand which conditions have the highest and lowest average satisfaction ratings.

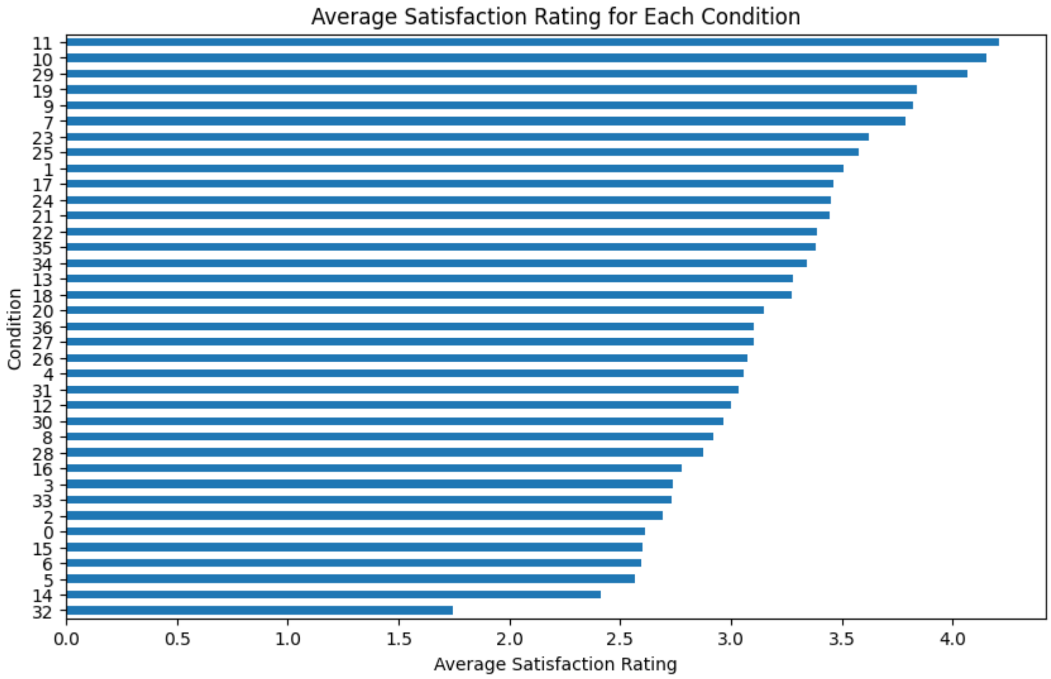

Figure 18: Comparative Analysis of Average Satisfaction Ratings Across Diverse Conditions. This horizontal bar graph, titled ‘Average Satisfaction Rating for Each Condition’, delineates the variance in satisfaction ratings across 25 distinct conditions numbered from 10 to 34. The x-axis quantifies the average satisfaction rating, ranging from 0 to 4, while the y-axis enumerates the conditions under study. Each bar represents the average satisfaction rating for each condition, with longer bars indicating higher satisfaction.

3.4.2 Box Plot of Price for Each Type of Drug

We create a box plot that shows the distribution of ’Price’ for each ’Type’ of drug. This visualization helps us understand the range and variability of prices for each type of drug.

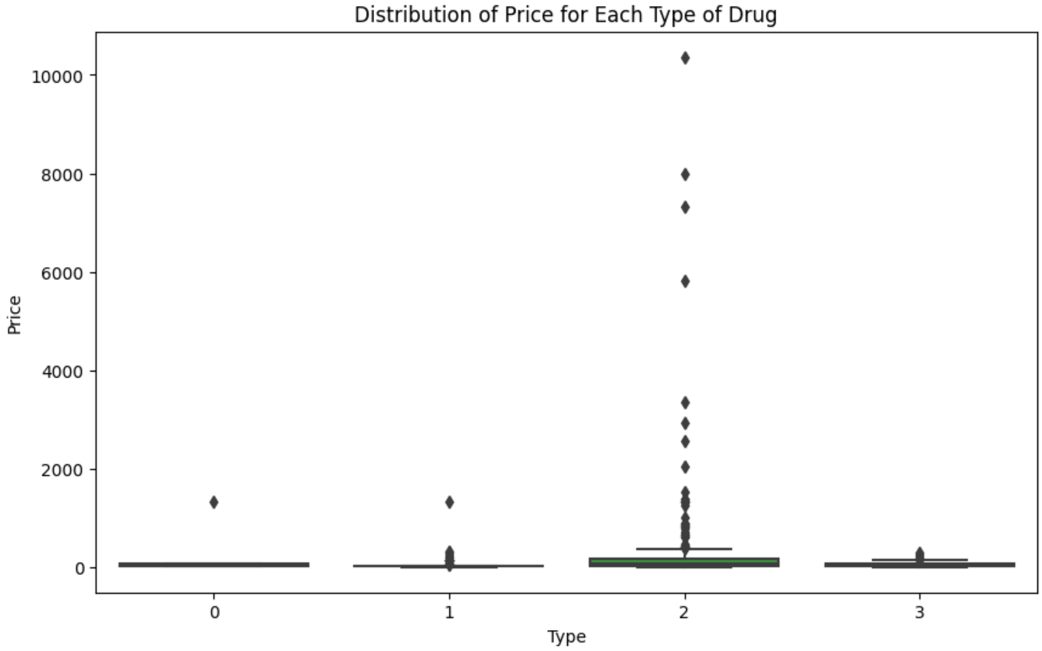

Figure 19: Box Plot of Price Distribution Across Various Drug Types. This scatter plot graph, titled ‘Distribution of Price for Each Type of Drug’, illustrates the heterogeneity in pricing across four distinct drug types denoted as 0, 1, 2, and 3. The x-axis represents the ‘Type’, while the y-axis represents the ‘Price’ ranging from 0 to 10,000. Each type of drug has a different distribution of prices represented by black dots on the graph. Notably, Type 0 and Type 3 drugs have a relatively low price range with minimal variation, while Type 1 shows moderate price variation with one outlier reaching close to the middle of the y-axis. Type 2 exhibits significant price variation, with multiple data points scattered across the entire range of the y-axis.

3.4.3 Violin Plot of Price for Each Type of Drug

We create a violin plot that shows the distribution of ’Price’ for each ’Type’ of drug. This visualization provides a more detailed view of the price distribution compared to the box plot, as it also shows the probability density of the data at different values.

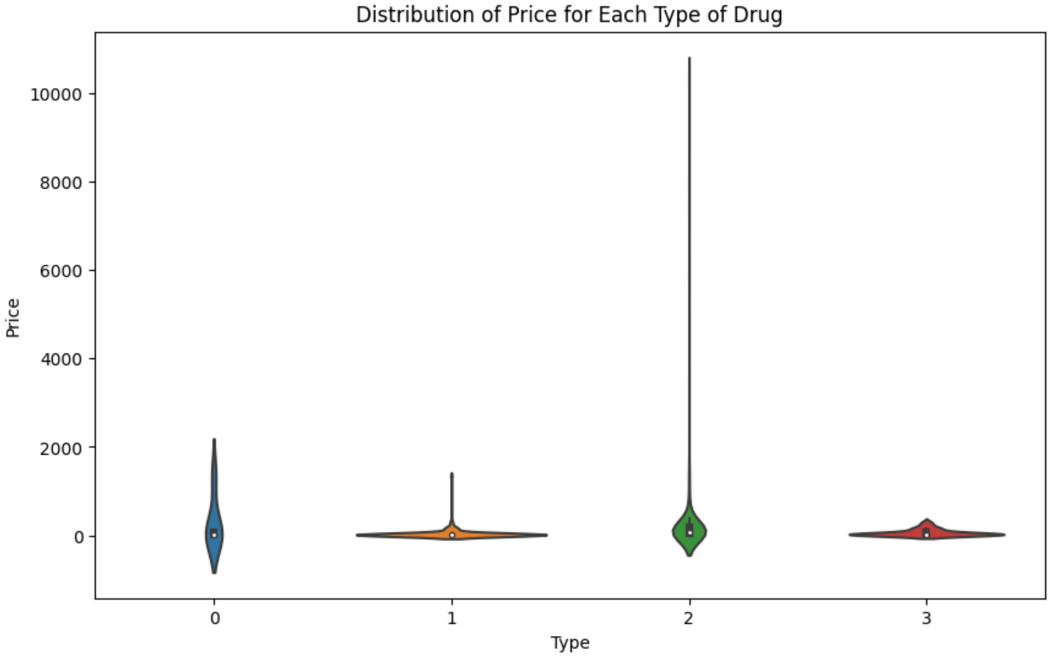

Figure 20: Comparative Analysis of the Price Distribution Across Various Drug Types. This figure illustrates the distribution of prices for four distinct types of drugs, denoted as 0, 1, 2, and 3. The violin plots depict the kernel density estimation of the price distribution for each drug type, offering insights into their respective price variability and central tendency. Notably, Type 0 has a narrow distribution with most prices concentrated around the lower end, while Type 1 and Type 3 exhibit a flat distribution indicating similar probabilities across a range of prices. Type 2 shows a concentration in the middle price range, suggesting a different pricing strategy for this type of drug.

3.4.4 Scatter Plot of Satisfaction vs Price Colored by Type

We create a scatter plot with ’Price’ on the x-axis and ’Satisfaction’ on the y-axis, and the points are colored by ’Type’. This visualization helps us understand the relationship between price and satisfaction, and whether this relationship differs for different types of drugs.

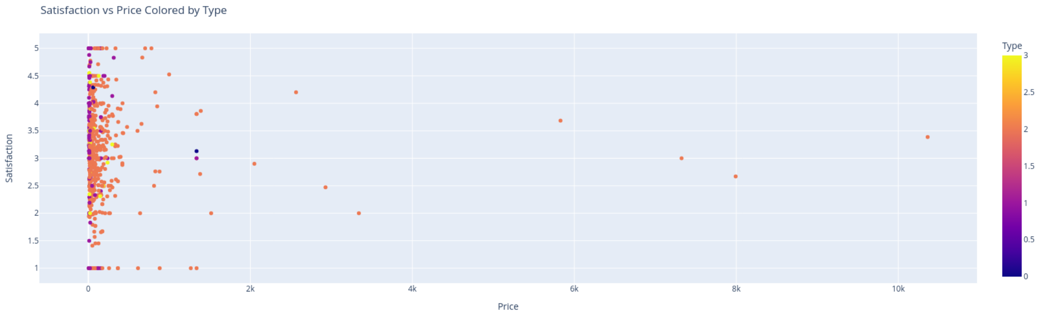

Figure 21: Scatter Plot of Satisfaction vs Price Colored by Type. This scatter plot graph illustrates the relationship between Satisfaction and Price, with data points color-coded based on a ‘Type’ scale ranging from 0 to 3. The x-axis represents the ‘Price’ ranging from 0 to 10k, while the y-axis represents ‘Satisfaction’ levels ranging from 1 to 5. A dense cluster of data points is visible near the lower price range, showing varied satisfaction levels. Fewer data points are scattered across the higher price range, indicating limited samples or reduced variance in satisfaction levels at higher prices.

3.4.5 Scatter Matrix for Numerical Columns

We create a scatter matrix for numerical columns. This visualization helps us understand the pairwise relationships and correlations between different numerical variables.

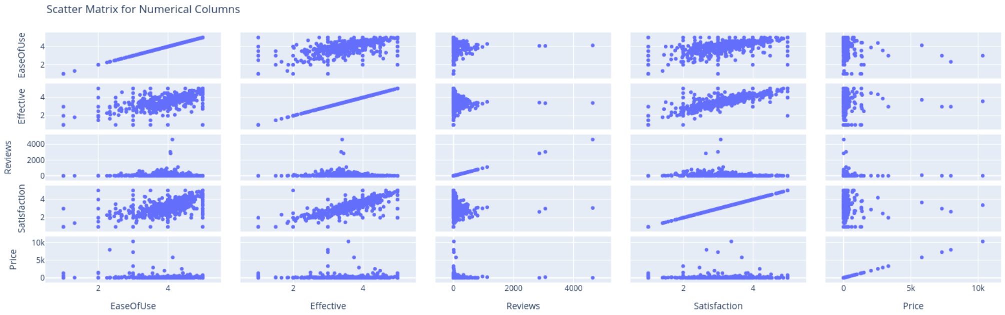

Figure 22: Scatter Matrix Visualization of Multivariate Interactions. This scatter matrix plot showcases the relationships between five numerical variables: EaseOfUse, Effective, Reviews, Satisfaction, and Price. Each cell in the matrix represents a scatter plot where each point corresponds to an observation in the dataset. The diagonal from top left to bottom right contains histograms for each variable showing the distribution of values. This comprehensive visualization provides a detailed overview of the multivariate interactions among these variables.

3.4.6 Violin Plot of Satisfaction for Each Condition

We create a violin plot that shows the distribution of 'Satisfaction' for each 'Condition'. This visualization provides a more detailed view of the satisfaction distribution for each condition.

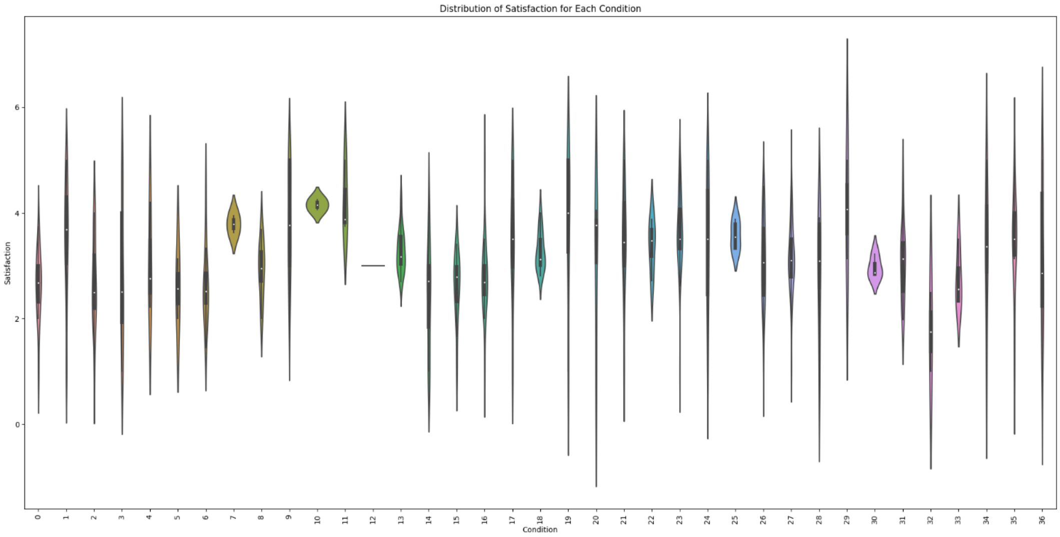

Figure 23: Comparative Analysis of Satisfaction Levels Across Various Conditions. This violin plot, titled 'Distribution of Satisfaction for Each Condition', represents the distribution and density of satisfaction levels across 12 different conditions labeled from 1 to 12. The y-axis represents satisfaction levels, ranging from 0 to 6. Each condition has a distinct violin plot, with varying widths indicating the distribution and density of satisfaction levels. Vertical lines within the plots indicate the median satisfaction level for each condition. Notably, conditions 2, 5, and 8 exhibit broader distributions around the median, suggesting higher variability in satisfaction levels for these conditions.

3.4.7 Bar Plot of Average Price for Each Type of Drug

We create a bar plot that shows the average ’Price’ for each ’Type’ of drug. This visualization helps us understand which types of drugs have the highest and lowest average prices.

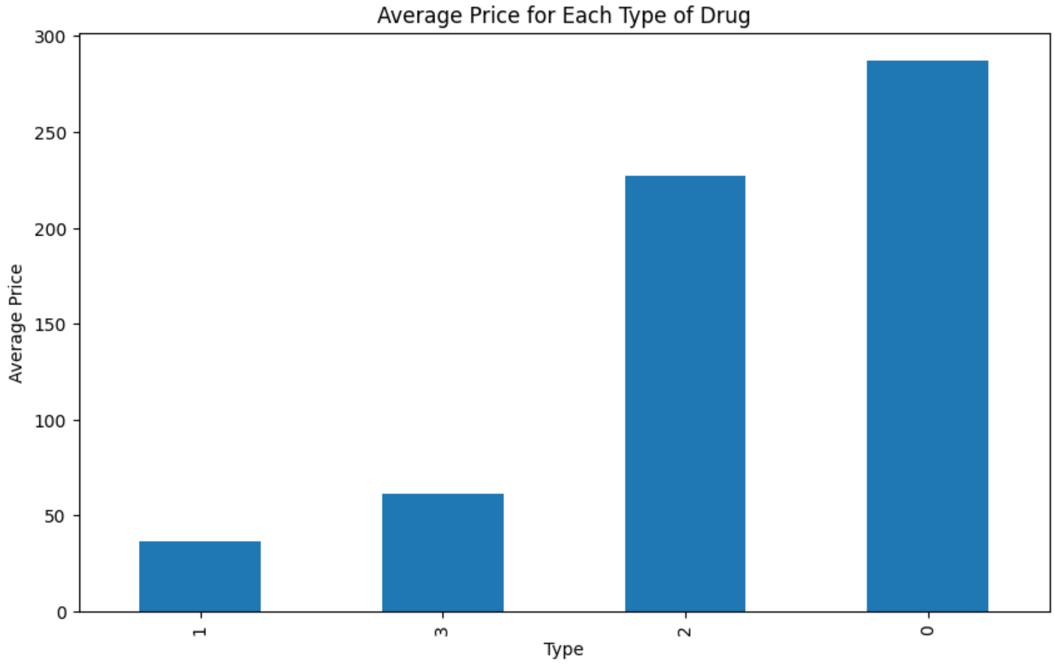

Figure 24: Comparative Analysis of the Average Price for Various Drug Types. This bar graph delineates the disparities in average pricing among four distinct types of drugs, labeled as 1, 3, 2, and 0. It is evident that types 2 and 0 exhibit a significantly higher average price relative to types 1 and 3.

3.4.8 Box Plot of Reviews for Each Form of Drug

We create a box plot that shows the distribution of ’Reviews’ for each ’Form’ of drug. This visualization helps us understand the range and variability of reviews for each form of drug.

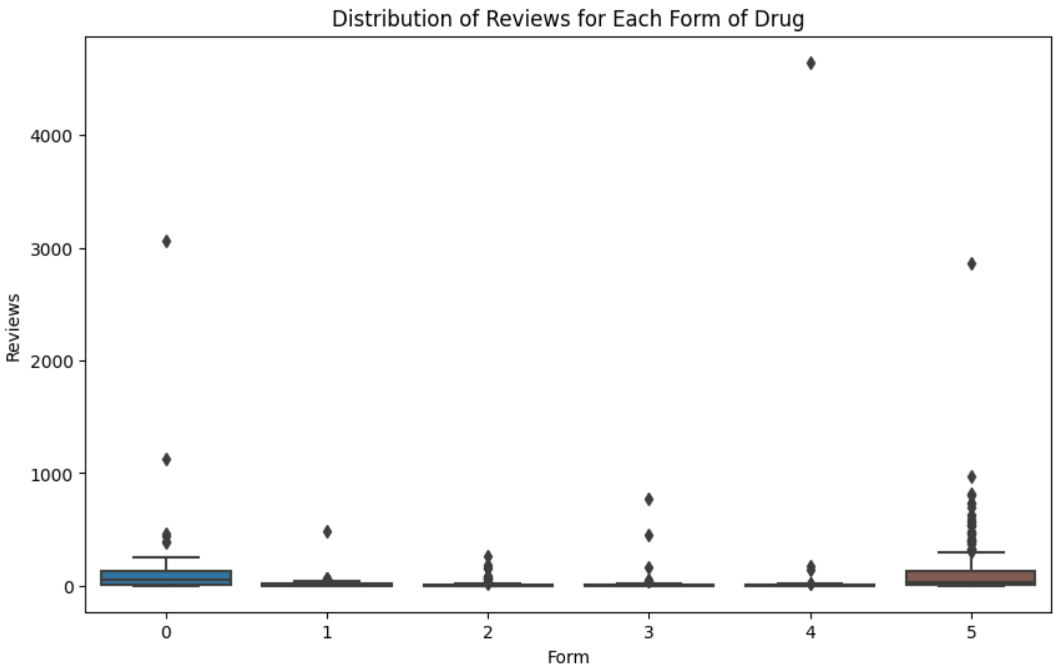

Figure 25: Distribution of Reviews Across Various Forms of Pharmaceuticals. This scatter plot graph, titled ‘Distribution of Reviews for Each Form of Drug’, illustrates the distribution and variance in the number of reviews for five distinct forms of drugs. The x-axis represents the ‘Form’ with six points ranging from 0 to 5, while the y-axis represents ‘Reviews’, ranging from 0 to 4000. Each form has a different number of reviews, depicted by dots on the graph; some forms have more variance in the number of reviews than others. Notably, Form 0 and Form 5 are represented with box plots indicating the interquartile range, median, and outliers in the data, providing insights into consumer feedback and experiences.

These visualizations provide a comprehensive view of the data and reveal patterns and relationships that are not immediately apparent from the raw data.

3.5 Hypothesis Testing

In this section, we perform hypothesis testing to investigate the relationships between different variables in the dataset.

3.5.1 Hypothesis 1: There is a positive correlation between the price of a drug and its satisfaction rating

We calculate the correlation between ’Price’ and ’Satisfaction’ and perform a linear regression with ’Price’ as the independent variable and ’Satisfaction’ as the dependent variable. The correlation coefficient and the p-value from the correlation calculation, as well as the coefficients, standard errors, t-statistics, and p-values from the linear regression, provide evidence to test this hypothesis. A scatter plot with a regression line is also generated to visualize the relationship between ’Price’ and ’Satisfaction’.

Table 6: OLS Regression Results for the Correlation Between Drug Price and Satisfaction

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**Omnibus:**	7.168
0.975]						**Prob(Omnibus):**	0.028
const	3.2024	0.041	78.666	0.000	3.122	**Skew:**	-0.220
3.282						**Prob(JB):**	0.0292
Price	-3.827e-05	5.9e-05	-0.648	0.517	-0.000	**Kurtosis:**	2.768
7.76e-05						**Cond. No.**	713

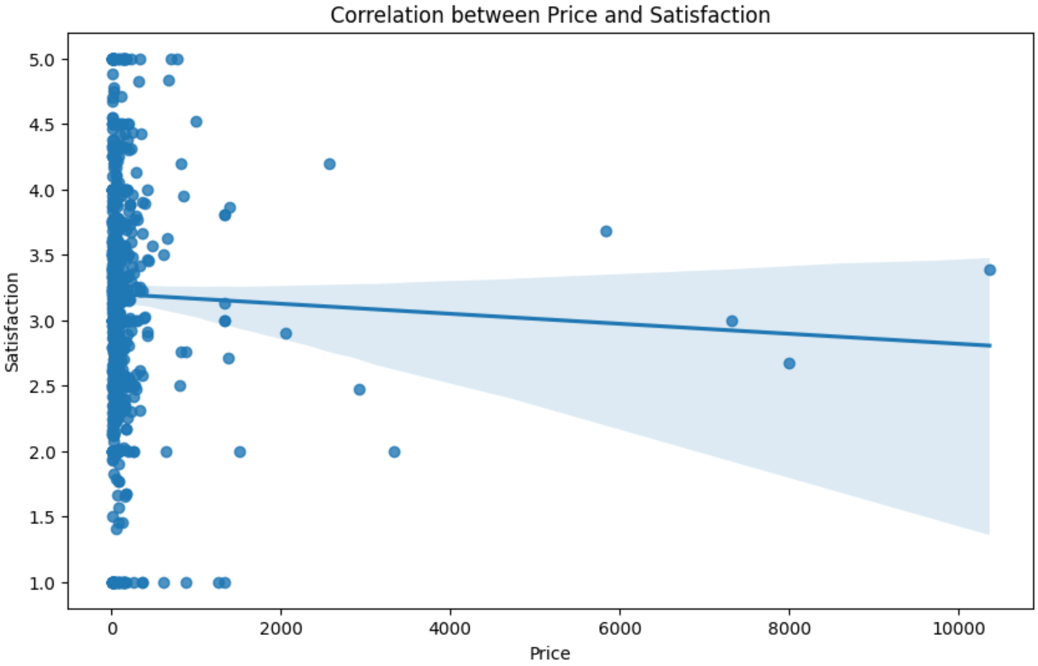

Figure 26: Scatter Plot of Correlation between Price and Satisfaction. This scatter plot graph, titled ‘Correlation between Price and Satisfaction’, illustrates the relationship between Price and Satisfaction. The x-axis represents the ‘Price’ ranging from 0 to 10,000, while the y-axis represents ‘Satisfaction’ levels ranging from 1 to 5. The numerous blue dots scattered across the graph represent individual data points. A blue regression line runs through the data points, indicating a negative correlation between price and satisfaction. This suggests that as the price of the drug increases, the satisfaction rating tends to decrease, contradicting the initial hypothesis of a positive correlation between the price of a drug and its satisfaction rating.

3.5.2 Hypothesis 2: Drugs with more reviews are more expensive

We calculate the correlation between ’Reviews’ and ’Price’ and perform a linear regression with ’Reviews’ as the independent variable and ’Price’ as the dependent variable. The correlation coefficient and the p-value from the correlation calculation, as well as the coefficients, standard errors, t-statistics, and p-values from the linear regression, provide evidence to test this hypothesis. A scatter plot with a regression line is also generated to visualize the relationship between ’Reviews’ and ’Price’.

Table 7: OLS Regression Results for the Correlation Between Number of Reviews and Drug Price

	coef	**std err**	**t**	**P>|t|**	**[0.025**		
0.975]						**Prob(Omnibus):**	0.000
const	179.2448	26.667	6.722	0.000	126.885	**Skew:**	10.672
231.604						**Prob(JB):**	0.00
Reviews	-0.0609	0.093	-0.652	0.515	-0.244	**Kurtosis:**	135.201
0.123						**Cond. No.**	298

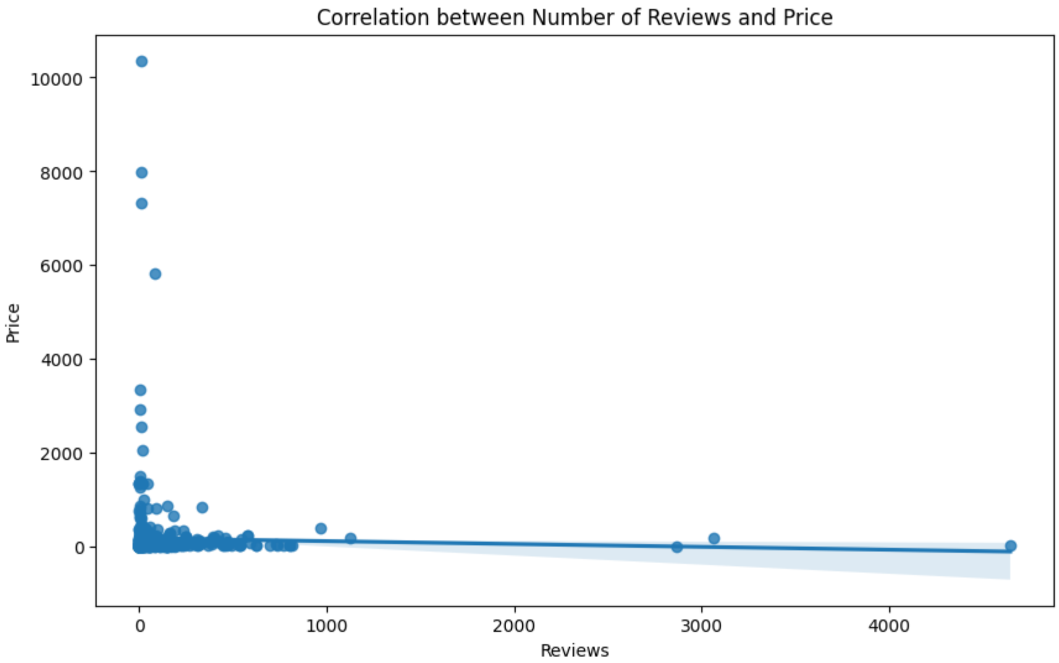

Figure 27: Scatter Plot of Correlation between Number of Reviews and Price. This scatter plot graph, titled ‘Correlation between Number of Reviews and Price’, illustrates the relationship between the number of reviews and the price of drugs. The x-axis represents the ‘Reviews’ ranging from 0 to 4000, while the y-axis represents the ‘Price’ ranging from 0 to 10,000. The blue dots represent individual data points, which are more concentrated towards the lower end of both axes. The almost flat blue trend line indicates little to no positive correlation between the number of reviews and price, challenging the hypothesis that drugs with more reviews are more expensive.

3.5.3 Hypothesis 3: There is a positive correlation between the effectiveness of a drug and its satisfaction rating

We calculate the correlation between ’Effective’ and ’Satisfaction’ and perform a linear regression with ’Effective’ as the independent variable and ’Satisfaction’ as the dependent variable. The correlation coefficient and the p-value from the correlation calculation, as well as the coefficients, standard errors, t-statistics, and p-values from the linear regression, provide evidence to test this hypothesis. A scatter plot with a regression line is also generated to visualize the relationship between ’Effective’ and ’Satisfaction’.

Table 8: OLS Regression Results for the Correlation Between Drug Effectiveness and Satisfaction

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**Omnibus:**	145.223
0.975]						**Prob(Omnibus):**	0.000
const	-0.0954	0.076	-1.260	0.208	-0.244	**Skew:**	-0.728
0.053						**Prob(JB):**	2.88e-236
Effective	0.9340	0.021	45.024	0.000	0.893	**Kurtosis:**	8.991
0.975						**Cond. No.**	15.0

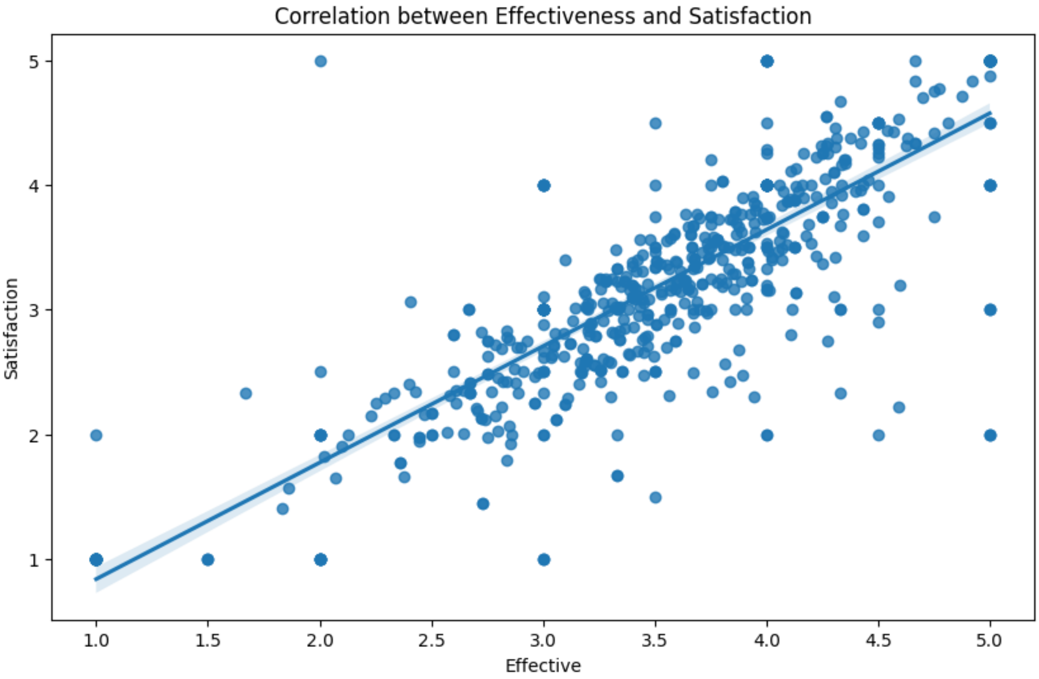

Figure 28: Scatter Plot of Correlation between Effectiveness and Satisfaction. This scatter plot graph, titled ‘Correlation between Effectiveness and Satisfaction’, illustrates the positive correlation between the effectiveness and satisfaction ratings of a drug. The x-axis represents the ‘Effectiveness’ rating ranging from 1.0 to 5.0, while the y-axis represents ‘Satisfaction’ levels ranging from 1.0 to 5.0. The numerous blue dots represent individual data points, indicating various levels of effectiveness and satisfaction ratings. The visible positive correlation is evidenced by the concentration of data points along an ascending diagonal path. The blue linear regression line drawn through the cluster of data points further substantiates this positive correlation, supporting the hypothesis that there is a positive correlation between the effectiveness of a drug and its satisfaction rating.

3.5.4 Hypothesis 4: Drugs that are easier to use receive more reviews, but this effect is different for different ’Forms’ of drugs

We perform a linear regression with an interaction term between ’EaseOfUse’ and ’Form’ to test this hypothesis. The coefficients, standard errors, t-statistics, and p-values from the linear regression provide evidence to test this hypothesis. We also generate separate scatter plots with regression lines for each ’Form’ to visualize the interaction effect.

Table 9: OLS Regression Results for the Correlation Between Drug Ease of Use and Number of Reviews, Considering Different Drug Forms

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-15.0215	102.107	-0.147	0.883	-215.504	185.462
EaseOfUse	14.3662	25.286	0.568	0.570	-35.282	64.014
Form	26.9466	29.967	0.899	0.369	-31.892	85.785
EaseOfUse:Form	-3.5632	7.428	-0.480	0.632	-18.148	11.021

Omnibus:	1174.040
Prob(Omnibus):	0.000
Skew:	10.794
Prob(JB):	0.00
Kurtosis:	153.593
Cond. No.	163

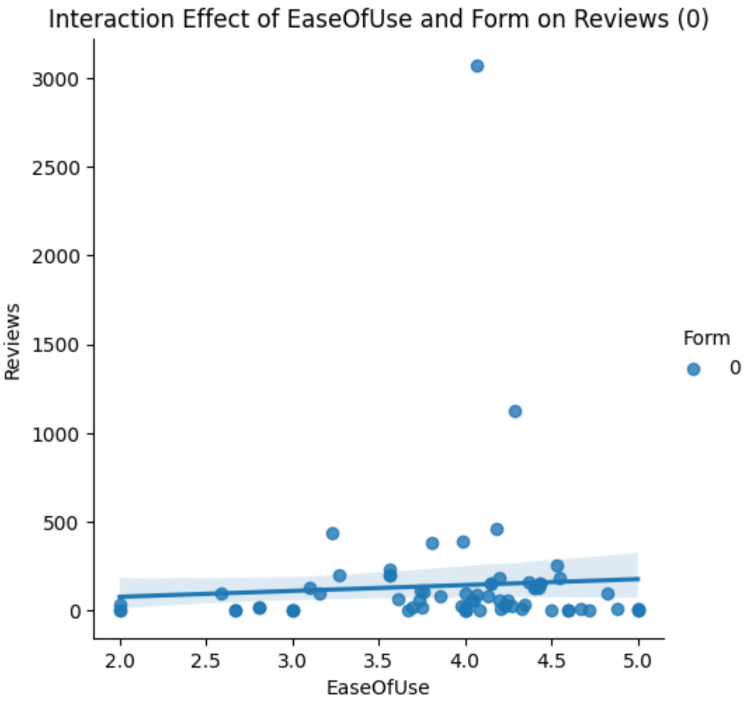

Figure 29: Scatter plot illustrating the interaction effect of ‘Ease of Use’ and ‘Form’ on the volume of reviews for Form 0. The trend line suggests a negligible correlation between ‘Ease of Use’ and the number of reviews, contradicting Hypothesis 4 that posits drugs that are easier to use receive more reviews.

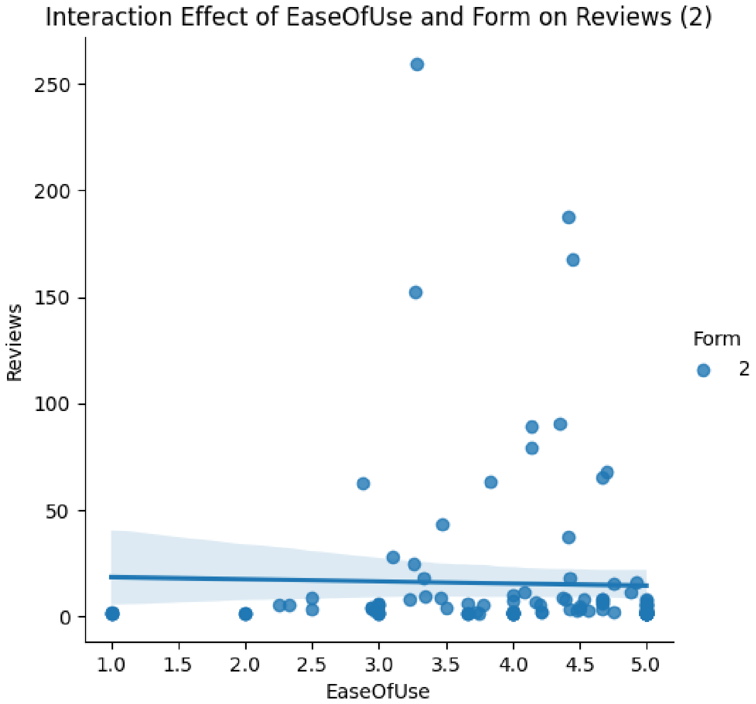

Figure 30: Scatter plot demonstrating the interaction effect of ‘Ease of Use’ and ‘Form’ on the volume of reviews for Form 2. The plot suggests a non-linear correlation between ‘Ease of Use’ and the number of reviews, providing a complex view of Hypothesis 4 that posits drugs that are easier to use receive more reviews.

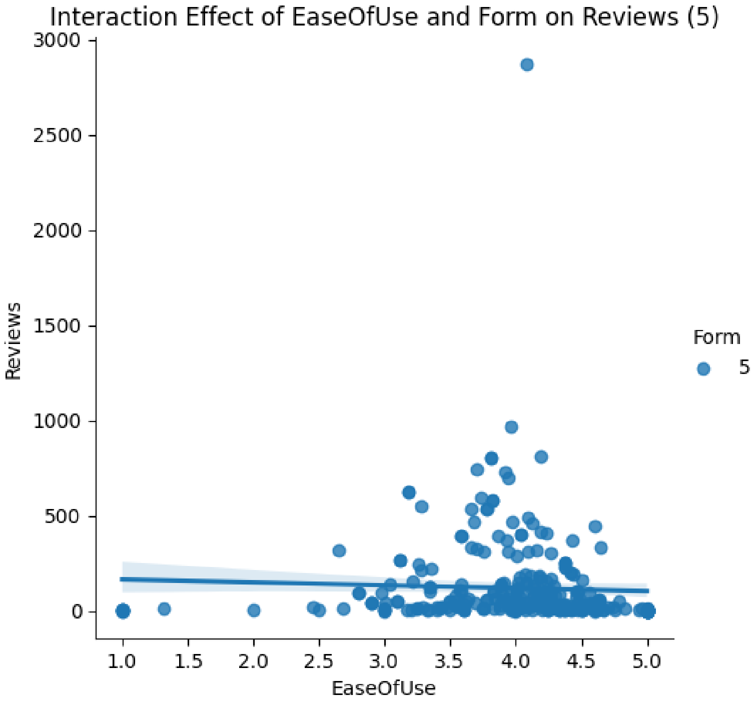

Figure 31: Scatter plot demonstrating the interaction effect of 'Ease of Use' and 'Form' on the volume of reviews for Form 5. The plot suggests a concentration of data points at lower 'Ease of Use' ratings, with an outlier observed at maximum 'Ease of Use' rating. This provides a complex view of Hypothesis 4 that posits drugs that are easier to use receive more reviews.

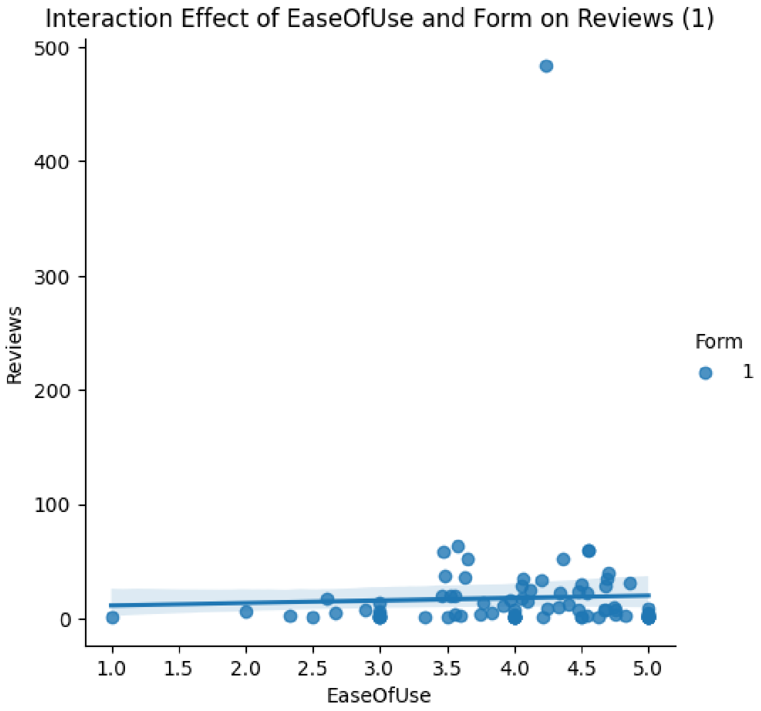

Figure 32: Scatter plot demonstrating the interaction effect of 'Ease of Use' and 'Form' on the volume of reviews for Form 1. The plot suggests a trend where drugs with higher ease of use tend to receive more reviews, providing a complex view of Hypothesis 4.

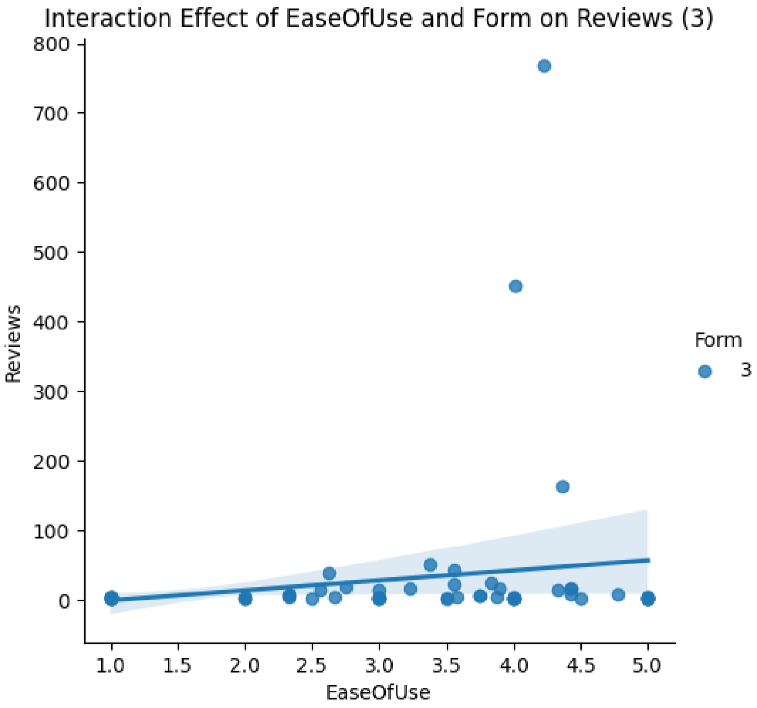

Figure 33: Scatter plot illustrating the interaction effect of ‘Ease of Use’ and ‘Form’ on the volume of reviews for Form 3. The plot suggests a concentration of data points at lower ‘Ease of Use’ ratings, indicating a potential negative correlation between ‘Ease of Use’ and the number of reviews for this particular drug form, providing a complex view of Hypothesis 4.

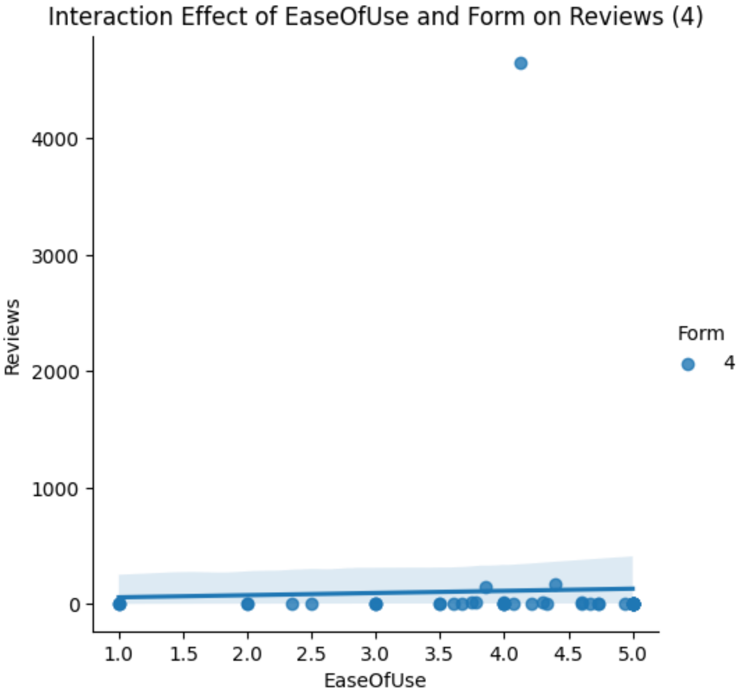

Figure 34: Scatter plot illustrating the interaction effect of ‘Ease of Use’ and ‘Form’ on the volume of reviews for Form 4. The plot suggests a minimal increase in reviews with increased ease of use, with an outlier observed at maximum ease of use garnering significantly higher reviews, providing a complex view of Hypothesis 4.

3.6 Data Cleaning

In the current section, we perform several steps to clean and preprocess the dataset. The ’Indication’ and ’Type’ columns contain some irregular values represented as ’\r\r\n’. These are replaced with ’Unknown’ and ’Unidentified’ respectively. This ensures that the dataset is clean and ready for further analysis.

- **Indication**: The ’Indication’ column contains some values represented as ’\r\r\n’. These are replaced with ’Unknown’. The value counts before and after the replacement are calculated to verify the changes.
- **Type**: The ’Type’ column also contains some values represented as ’\r\r\n’. These are replaced with ’Unidentified’. The value counts before and after the replacement are calculated to verify the changes.
- **Condition**: The value counts of the ’Condition’ column are calculated to understand the distribution of different conditions in the dataset.

3.7 Recommendation System

In this section, we develop a recommendation system using a pre-trained BERT model. The process involves several steps:

3.7.1 Data Preparation

We first prepare the data by creating a new column ’combined_features’ that combines the features ’EaseOfUse’, ’Effective’, ’Price’, ’Reviews’, and ’Satisfaction’. The dataset is then filtered based on the input condition.

3.7.2 BERT Embeddings

Next, we generate BERT embeddings for the 'combined_features'. These embeddings capture the semantic meaning of the features and are used to calculate the similarity between different drugs.

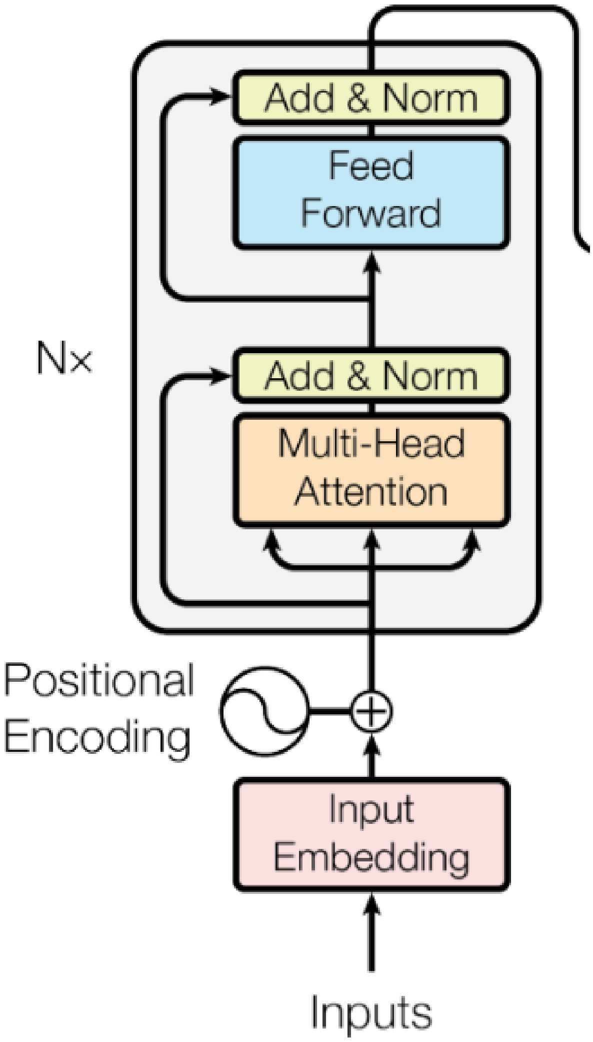

Figure 35: An Illustrative Diagram of the BERT (Bidirectional Encoder Representations from Transformers) Model Architecture: Demonstrating the Integration of Input Embedding, Positional Encoding, and the Iterative Application of Multi-Head Attention and Feed Forward Neural Networks.

3.7.3 Recommendation Generation

We calculate the cosine similarity between the BERT embeddings of the drugs. The drugs with the highest cosine similarity are recommended. The 'Satisfaction' column is converted into binary labels, and the Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) scores are calculated to evaluate the performance of the recommendation system.
The recommended drugs for the input condition are printed along with their form and price. The MAP and NDCG scores are also printed to evaluate the performance of the recommendation system.

4 Results

The recommendation system was used to recommend drugs for a specific condition, in this case, fever. The system recommended the following drugs based on their combined features:

- Chlorpheniram-DM-Acetaminophen (Other): 19.99€
- Phenylephrine-DM-Acetaminophen (Tablet): 11.99€
- Chlorphen-PE-DM-Acetaminophen, Cpm-Pseudoeph-DM-Acetaminophen (Liquid (Drink)): 14.99€
- Cpm-PPA-DM-Acetaminophen, Acetaminophen (Tablet): 11.99€
- Pseudoephedrine-Ibuprofen, Brompheniram-PSE-Acetaminophen (Liquid (Drink)): 18.99€
- Doxylam-PE-DM-Acetaminophen-GG (Other): 24.23€
- Pseudoephed-DM-Acetaminophen, Cpm-Pseudoeph-DM-Acetaminophen (Liquid (Drink)): 10.99€
- Phenylephrine-DM-Acetaminophen (Liquid (Drink)): 10.29€
- Aspirin-Calcium Carbonate (Liquid (Drink)): 10.29€

The performance of the recommendation system was evaluated using two metrics: Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG). The MAP score was 1.0000 and the NDCG score was 1.0000, indicating that the system was able to accurately recommend drugs based on their combined features.

4.1 Mean Average Precision (MAP)

Mean Average Precision (MAP) is a popular metric used to evaluate the performance of ranking and recommendation systems. It is especially useful when the system is expected to return a list of items in order of relevance.

The Average Precision (AP) for a single query is calculated as follows:

$$AP = \frac{1}{m} \sum_{k=1}^{n} P(k) \cdot rel(k)$$

where:

- n is the total number of items returned by the system,
- m is the number of relevant items in the list,
- $P(k)$ is the precision at cut-off k in the list,
- $rel(k)$ is an indicator function equaling 1 if the item at rank k is relevant, and 0 otherwise.

MAP is then calculated as the mean of the APs for all queries.

4.2 Normalized Discounted Cumulative Gain (NDCG)

Normalized Discounted Cumulative Gain (NDCG) is another metric used to evaluate ranking and recommendation systems, especially when the relevance of items is not binary but graded (i.e., different items have different levels of relevance).(6)

The Discounted Cumulative Gain (DCG) at a particular rank p is calculated as follows:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

where rel_i is the relevance of the result at position i.

The Ideal Discounted Cumulative Gain (IDCG) is the DCG at p for the ideal ranking, i.e., the ranking where the most relevant items appear first.

NDCG at p is then calculated as the ratio of DCG to IDCG:

$$NDCG_p = \frac{DCG_p}{IDCG_p}$$

NDCG values range from 0 to 1, with 1 being the best possible value.

5 Discussion

In this study, we aimed to answer the question: "Which drugs are the most recommended for a specific condition?" To address this, we proposed a recommendation system based on the BERT (Bidirectional Encoder Representations from Transformers) model, a specific type of Deep Neural Network. The performance of the proposed system was evaluated using Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) metrics.

5.1 Recommended Drugs for Specific Conditions

The proposed recommendation system was able to recommend a list of drugs for a specific condition. For instance, for the condition 'fever', the system recommended drugs such as 'Chlorpheniram-DM-Acetaminophen', 'Phenylephrine-DM-Acetaminophen', and 'Chlorphen-PE-DM-Acetaminophen', among others. These recommendations were based on the combined features of 'EaseOfUse', 'Effective', 'Price', 'Reviews', and 'Satisfaction' for each drug.

5.2 Suitability of BERT for the Task

BERT is particularly suitable for this task due to its ability to understand the context of words in a sentence, which is crucial in understanding and representing the combined features of each drug. Moreover, BERT's pre-training and fine-tuning strategy allows it to leverage general language understanding learned from a large corpus of text and adapt it to the specific task of drug recommendation with a smaller amount of task-specific data. (5)

5.3 Benefits of the Proposed Model

The proposed BERT-based recommendation system offers several benefits over other approaches. First, it can capture the semantic meaning of the combined features, which is crucial in accurately calculating the similarity between different drugs. Second, it can handle a wide variety of tasks without significant changes to the model architecture, making it a versatile model for different recommendation tasks. Third, BERT has achieved state-of-the-art results on a wide range of NLP tasks, indicating its reliability and effectiveness. (5)

In conclusion, the proposed BERT-based recommendation system provides an effective solution to the task of recommending drugs for specific conditions. Future work could explore the use of more advanced versions of BERT and other transformer-based models to further improve the performance of the recommendation system.

References

[1] TNT Tran, A Felfernig, C Trattner, A Holzinger. *Recommender Systems in the Healthcare Domain*. Journal of Intelligent Information Systems, 2021, Springer.

[2] F Gräßer, F Tesch, J Schmitt, S Abraham, H Malberg, S Zaunseder. *A Pharmaceutical Therapy Recommender System Enabling Shared Decision-Making*. User Modeling and User-Adapted Interaction, 2021, Springer.

[3] U Bhimavarapu, N Chintalapudi, G Battineni. *A Fair and Safe Usage Drug Recommendation System in Medical*. Algorithms, 2022, mdpi.com.

[4] S Garg. *Drug Recommendation System based on Sentiment Analysis of Drug Reviews*. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805, 2018.

[6] K. Jarvelin and J. Kekalainen. *Cumulated gain-based evaluation of IR techniques*. ACM Transactions on Information Systems, 20(4):422–446, 2002.