INTERNSHIP: INTERIM PROJECT REPORT

Internship Project Title	RIO-210: Classification Model - Build a Model that Classifies the Side Effects of a Drug
Name of the Company	TCS iON
Name of the Industry Mentor	Himdweep Walia
Name of the Institute	Amity University Online

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
23-05-2024	21-06-2024	124	Jupyter Notebook	PYTHON Library: Pandas, NumPy, seaborn, matplotlib, pyplot, sklearn, etc

Milestone #	1	Milestone:	EXPLORATORY DATA ANALYSIS

TABLE OF CONTENT

1. ACKNOWLEDGEMENT

I extend my sincere gratitude to TCS iON for granting me the opportunity to undertake this internship.

I am highly indebted to TCS iON authorities for the facilities provided to accomplish this internship.

I am extremely grateful to my colleagues and friends who helped me in the successful completion of this internship.

I would like to thank my mentor, Mrs. Himdweep Walia for her insightful feedback and guidance throughout the internship. Her expertise significantly contributed to my growth.

I also would like to thank all the people who worked along with me to support me with their patience and openness fostering a positive working environment, making this experience enjoyable.

It is indeed with a great sense of pleasure and immense gratitude that I acknowledge the help and support of all these individuals.

2. OBJECTIVE

The primary objectives of this internship project are:

- **Data Collection and Preprocessing**: Collect and clean a detailed dataset including patient demographics, drug ratings, and reviews. This step involves sourcing data from reputable healthcare databases and ensuring its quality by addressing missing values, eliminating duplicates, and correcting inconsistencies. Effective data preprocessing is crucial for the success of subsequent analyses as it ensures that the dataset is accurate, complete, and ready for machine learning (ML) model development. This step may also include feature engineering to create new, meaningful variables from the existing data.

- **Exploratory Data Analysis (EDA)**: Analyze the data to find patterns and relationships among key variables. EDA involves using statistical techniques and visualization tools to gain insights into the dataset’s structure and distributions. This step helps in identifying trends, anomalies, and correlations that could influence the occurrence of drug side effects. It includes the use of descriptive statistics, histograms, box plots, scatter plots, and heat maps to summarize and visualize data characteristics, thus guiding the feature selection process for model development.

- **Model Development**: Create and compare various machine learning models to predict drug side effects. This objective involves selecting a range of different machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines (SVM), and K-nearest neighbor (KNN). The models are developed using training data and are designed to identify patterns and make predictions about the likelihood of drug side effects based on patient demographics and drug usage data. The diversity of models allows for a comprehensive comparison to find out the best-performing algorithm.

- **Model Evaluation**: Assess the models based on accuracy, precision, recall, and F1 score. Model evaluation is essential to determine the effectiveness and reliability of each ML model. This involves partitioning the dataset into training and testing subsets and using performance metrics to evaluate the models' predictive capabilities. Accuracy measures the overall correctness of the model, precision indicates the proportion of true positive predictions among all positive predictions, recall measures the ability to identify true positives, and F1 score provides a balance between precision and recall.

- **Insights and Recommendations**: Provide practical insights and advice for healthcare providers based on the analysis. This final objective focuses on translating the findings from the ML models and EDA into actionable recommendations for clinical practice. These insights can help healthcare providers make informed decisions about drug prescriptions, identify high-risk patients, and develop personalized treatment plans. Additionally, the study’s results could inform public health strategies and policy-making to enhance patient safety and improve medication adherence.

3. INTRODUCTION

Machine learning (ML) is a state-of-the-art approach that has extensive applications in categorization, prediction, and forecasting across diverse fields including medicine, engineering, education, manufacturing, weather forecasting, traffic management, robotics, and more. It is one of the most advanced concepts of artificial intelligence (AI), and provides a strategic approach to developing automated, intricate, and unbiased algorithmic techniques for multimodal and dimensional biomedical or mathematical data analysis. Machine learning has demonstrated significant potential in pharmaceuticals and medicine by enhancing the collection and utilization of diverse data types to improve analysis, prevention, and individualized treatment strategies.

Healthcare is an important industry that offers value-based care to millions of people. Machine learning (ML) applications integrated with real-time patient data from various healthcare systems across multiple countries can enhance the effectiveness of personalized treatment options. ML has extensive use in precision medicine and personalized treatments. Using ML techniques, both beneficial and adverse drug side effects can be categorized, enabling more informed decisions for precision medicine and personalized treatments. Drug classifiers based on side effects also serve as valuable resources to support licensed healthcare professionals in patient care.

A side effect is an unwanted drug reaction caused by a drug or medication, occurring alongside its intended therapeutic benefits. These effects can vary widely among individuals based on factors like their health condition, age, weight, gender, ethnicity, and overall well-being. They can occur when starting, decreasing/increasing dosages, or ending a medication and may lead to non-compliance with prescribed treatment. Side effects not only affect patient compliance and satisfaction but can also result in serious health issues. Severe side effects may require dose adjustments or prescription of alternative medications. Lifestyle changes or dietary adjustments may also help minimize side effects. Classifying side effects for each drug is a challenging task, but machine learning techniques have facilitated this task with improved accuracy. Pharmacogenetic research has revealed significant differences among racial and ethnic groups in how drugs are metabolized, their clinical effectiveness, and their side effect profiles, highlighting the importance of personalized medical approaches.

4. DESCRIPTION OF INTERNSHIP

Machine learning models have been developed to classify side effects of drugs based on age and gender using patient review dataset obtained from Kaggle. The dataset comprises 362,806 instances and includes 12 features containing categorical, numerical, and text data. It provides both demographic and clinical data and provides user reviews of specific drugs along with related conditions, side effects, age, sex, and ratings (ease of use, effectiveness, satisfaction) reflecting overall patient satisfaction. Each entry details user reviews for drugs based on purchased medications, illness conditions, ratings, and useful counts based on review helpfulness.

The insights obtained from these models can help healthcare providers in making informed decisions and personalize treatment plans to minimize adverse effects. By analyzing patterns in patient demographics and feedback, providers can tailor interventions more effectively, potentially enhancing patient outcomes and satisfaction. Moreover, the dataset's comprehensive nature facilitates a deeper understanding of how different factors influence drug efficacy and patient experiences across various demographic groups.

5. INTERNSHIP ACTIVITIES

Activity 1: Familiarize the topic and search for a dataset.

Familiarize with the TCSION platform and understand the needs of the project and dataset required. For that, I searched on various sites like Kaggle.

Activity 2: Finalize a dataset from Kaggle and do a basic analysis.

Finalized a WebMD dataset of 3,68,206 entries from Kaggle. After that, I performed some basic analysis on the dataset such as info, describe, columns, shape, null values, datatypes, unique count, unique values, etc.

Activity 3: Data preprocessing

Data preprocessing was done including Missing value handling, Cleaning irrelevant data, etc.

Activity4: Exploratory data analysis non-graphical

Performed some EDA like No. of users based on gender, Top 15 recommended drugs for each condition, Drug mostly used by the patient and most common condition, etc.

Activity 5: Exploratory data analysis graphical

Graphical EDA was done such as Most commonly used drugs, Top 10 drugs used for Pain, Most common conditions, Race Count plot, Gender Distribution by Race, Effectiveness of Drug, Side Effects by Age Group, etc.

6. APPROACH / METHODOLOGY

The dataset ("drugdata.csv") used for the project includes both demographic and clinical data. It contains both structured and unstructured data allowing to perform a multifaceted analysis. It contains 3,62,806 instances and 12 features including categorical, numerical, and text data. The dataset provides user reviews on specific drugs along with related conditions, side effects, age, gender, and ratings (ease of use, satisfaction, effectiveness) reflecting overall patient satisfaction.

The methodology involves rigorous data preprocessing to handle missing values, feature creation, and transforming skewed data. Exploratory Data Analysis (EDA) uses different visual techniques to discover patterns and relationships in the data. Machine learning models including Logistic Regression, SVM, KNN, Decision Trees, Random Forests, Naive Bayes, and Gradient Boosting will be developed to predict drug side effects effectively.

7. ASSUMPTIONS

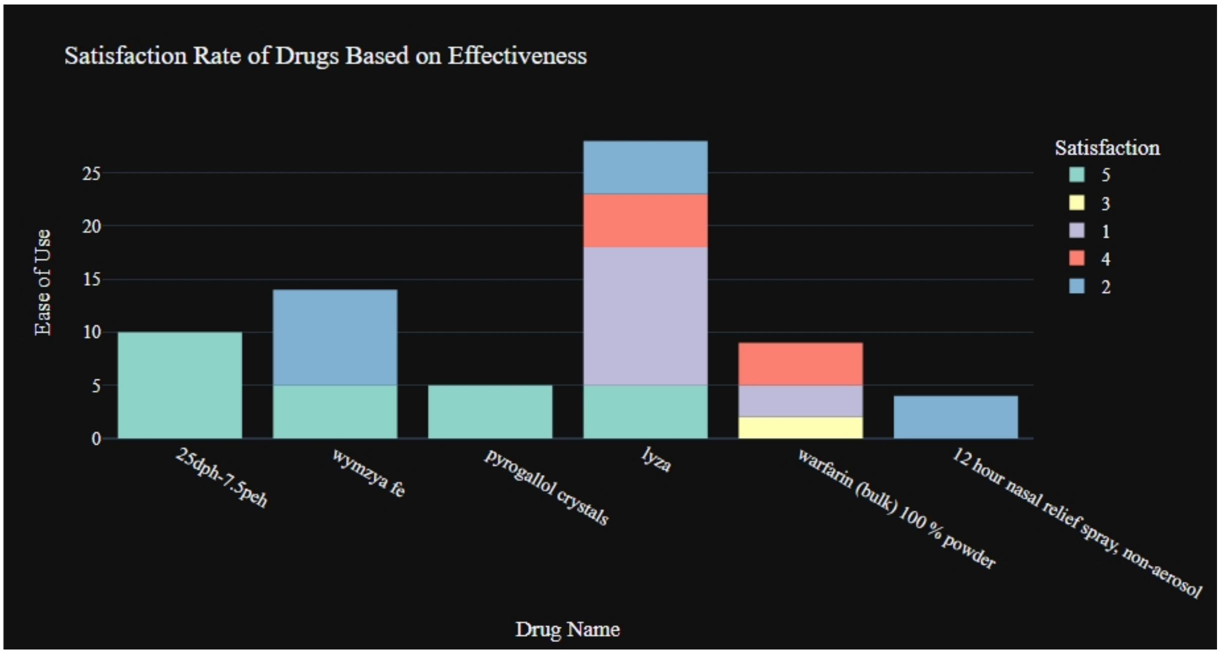
- The data collected by WebMD is representative of the population as a whole.
- The data collected is accurate and free from errors.
- The machine learning techniques used to classify side effects are appropriate for the dataset.
- The classification model developed is generalizable to other datasets.
- The age and gender information provided in the dataset is accurate and complete.

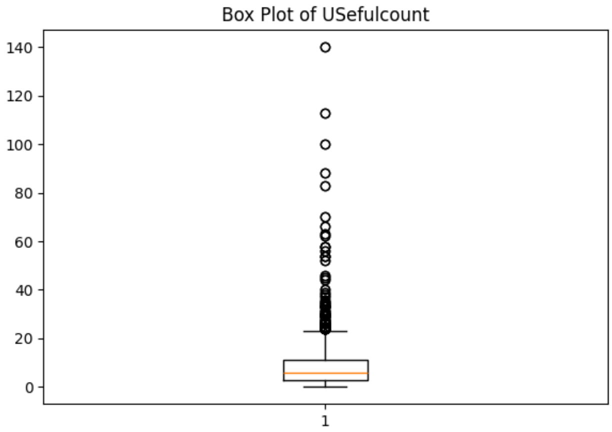
8. EXCEPTIONS / EXCLUSIONS

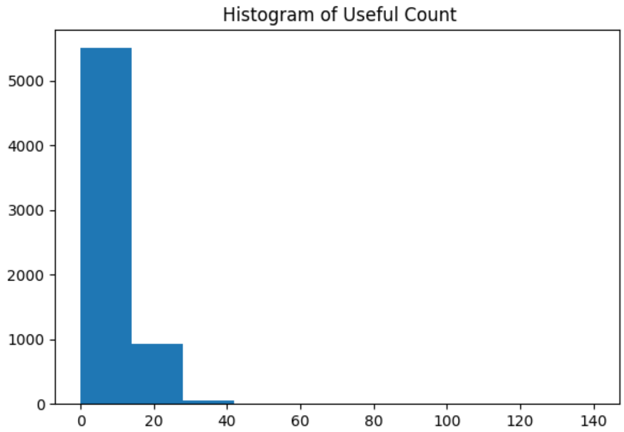
- The dataset may not include all possible side effects for each drug.
- The dataset may not include information on all drugs or medications.
- The dataset may not be up-to-date or accurate.
- The classification model developed may not be able to accurately predict side effects for all individuals.
- The classification model developed may not be generalizable to other datasets.
- The classification model developed may not be able to account for all the factors that can impact side effects.

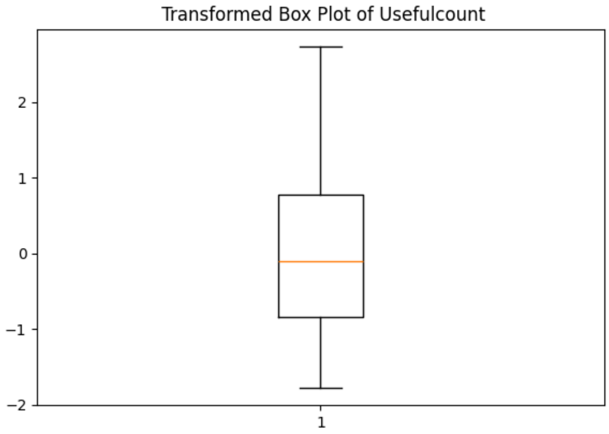
9. CHARTS, TABLES, DIAGRAMS

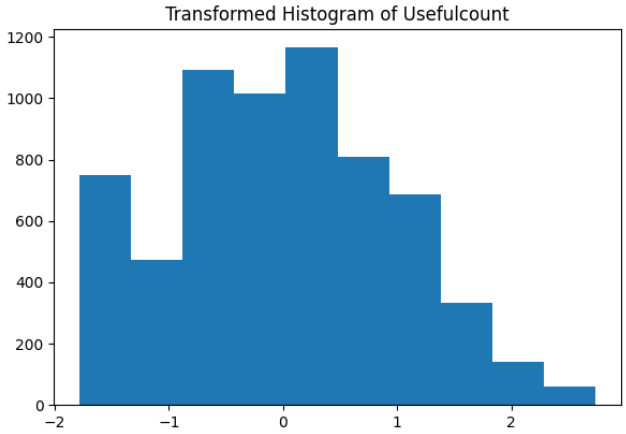
Exploratory Data Analysis

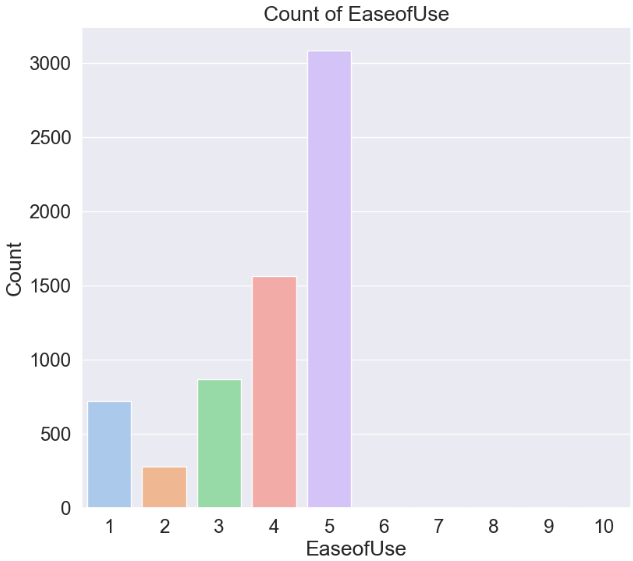


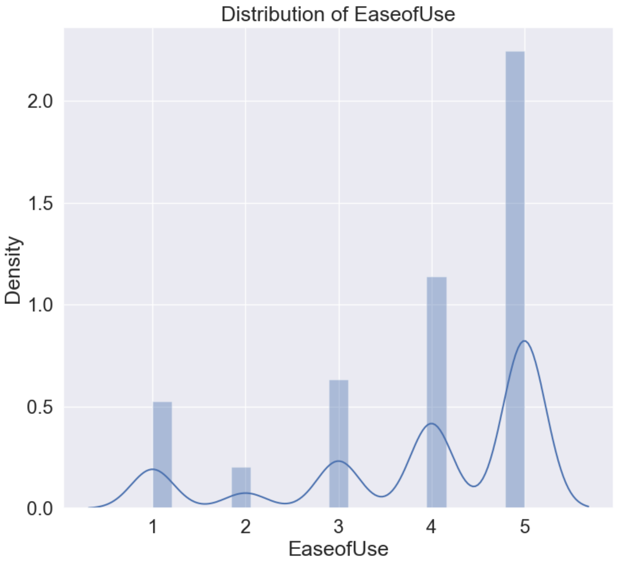


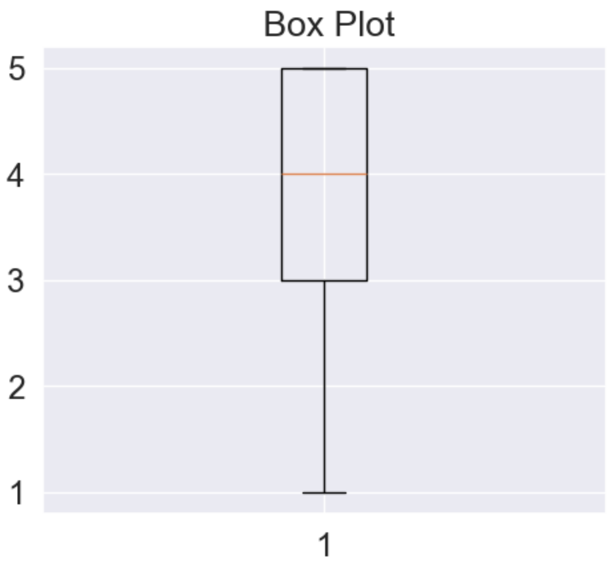


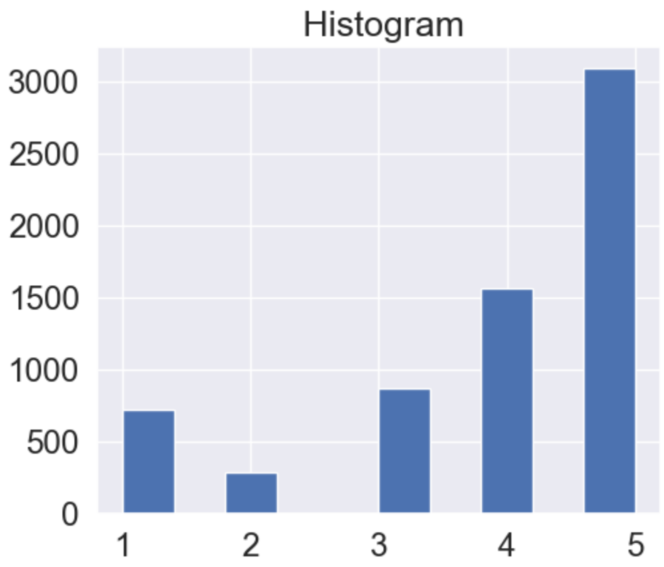


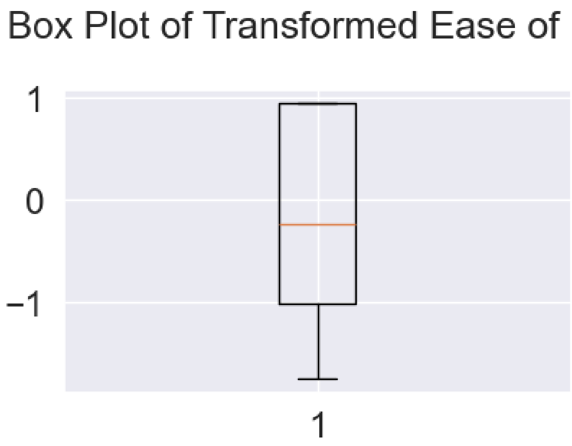


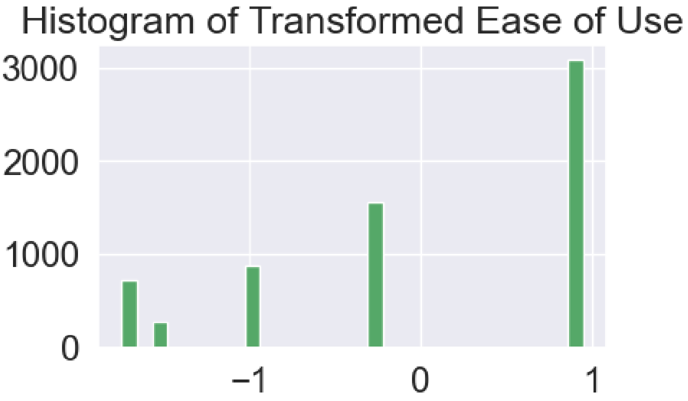

Count of EaseofUse
Count
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10
EaseofUse

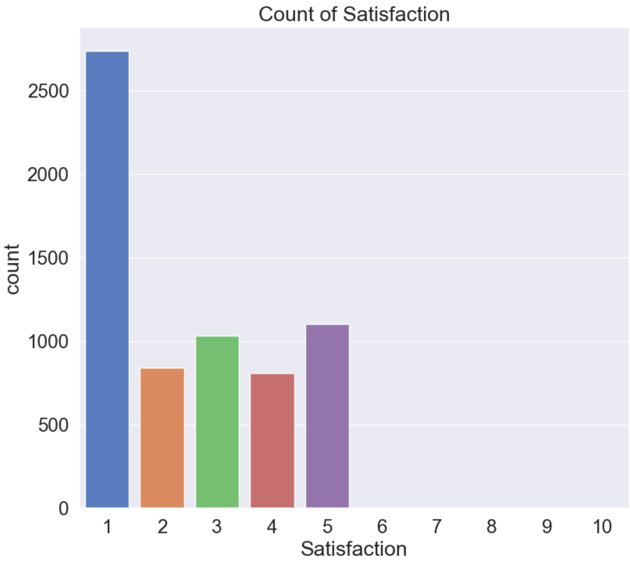

Distribution of EaseofUse
Density
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
EaseofUse

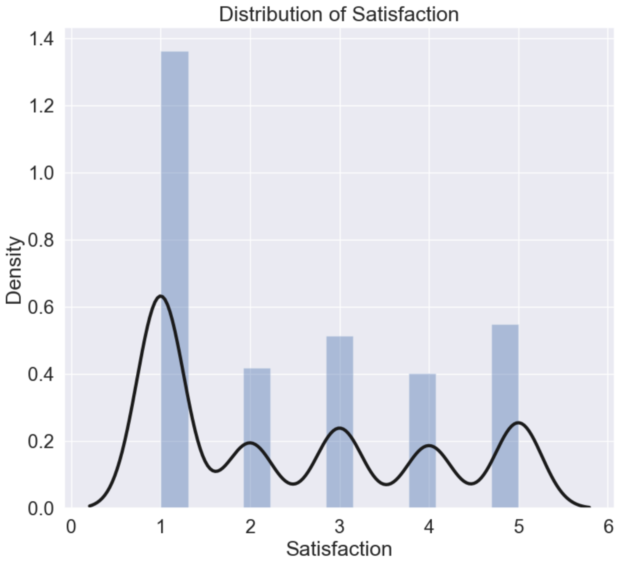

Box Plot
5
4
3
2
1
1

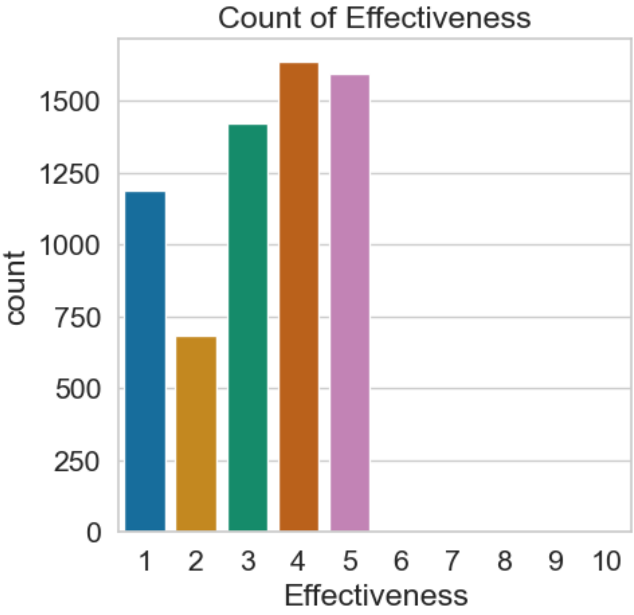

Histogram
3000
2500
2000
1500
1000
500
0
1
2
3
4
5

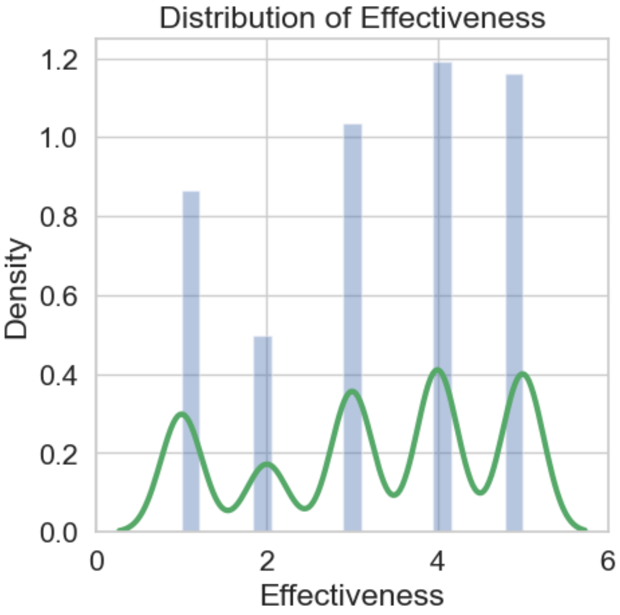

Box Plot of Transformed Ease of Use
1
0
−1
1

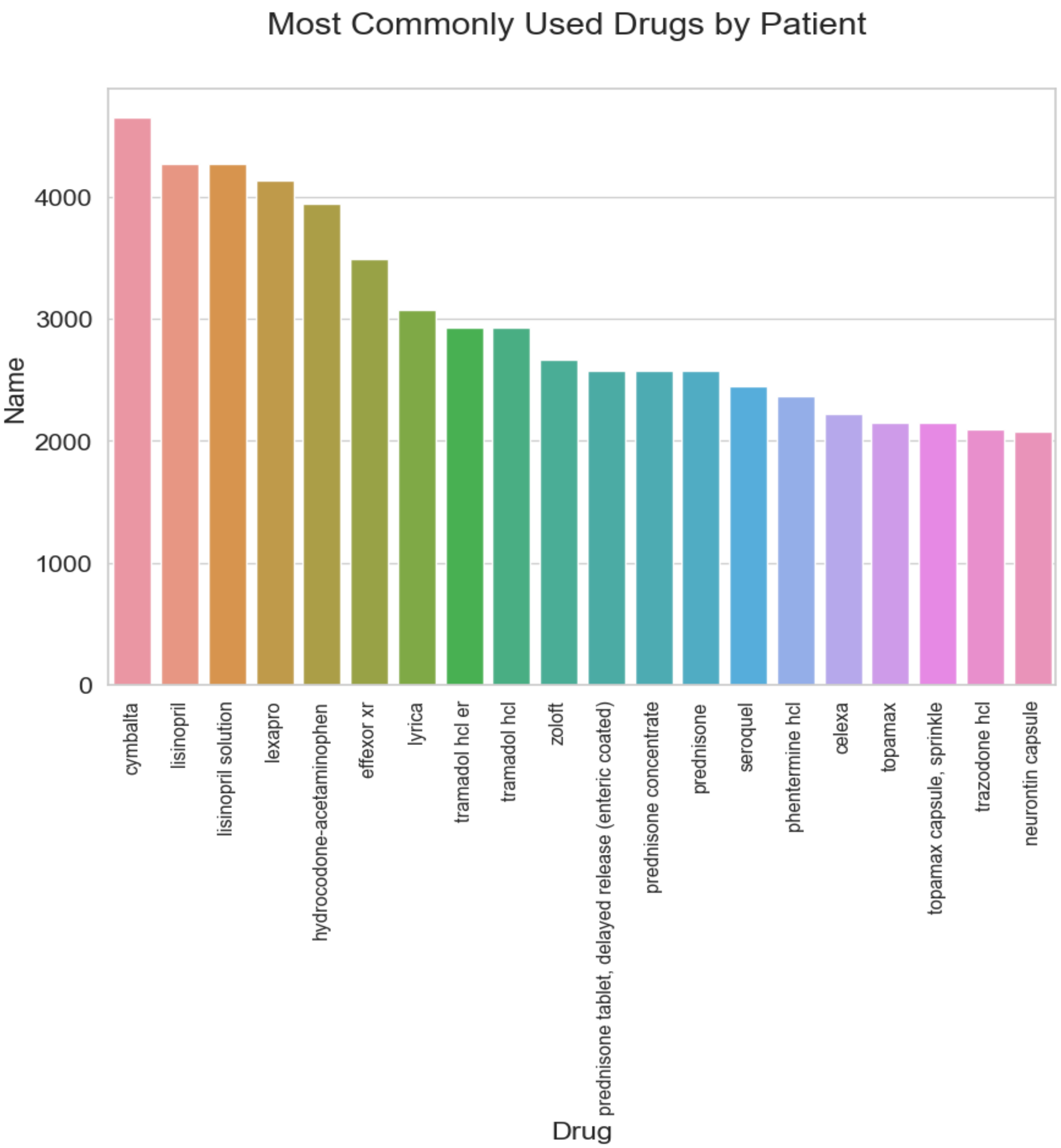

Histogram of Transformed Ease of Use
3000
2000
1000
0
−1
0
1

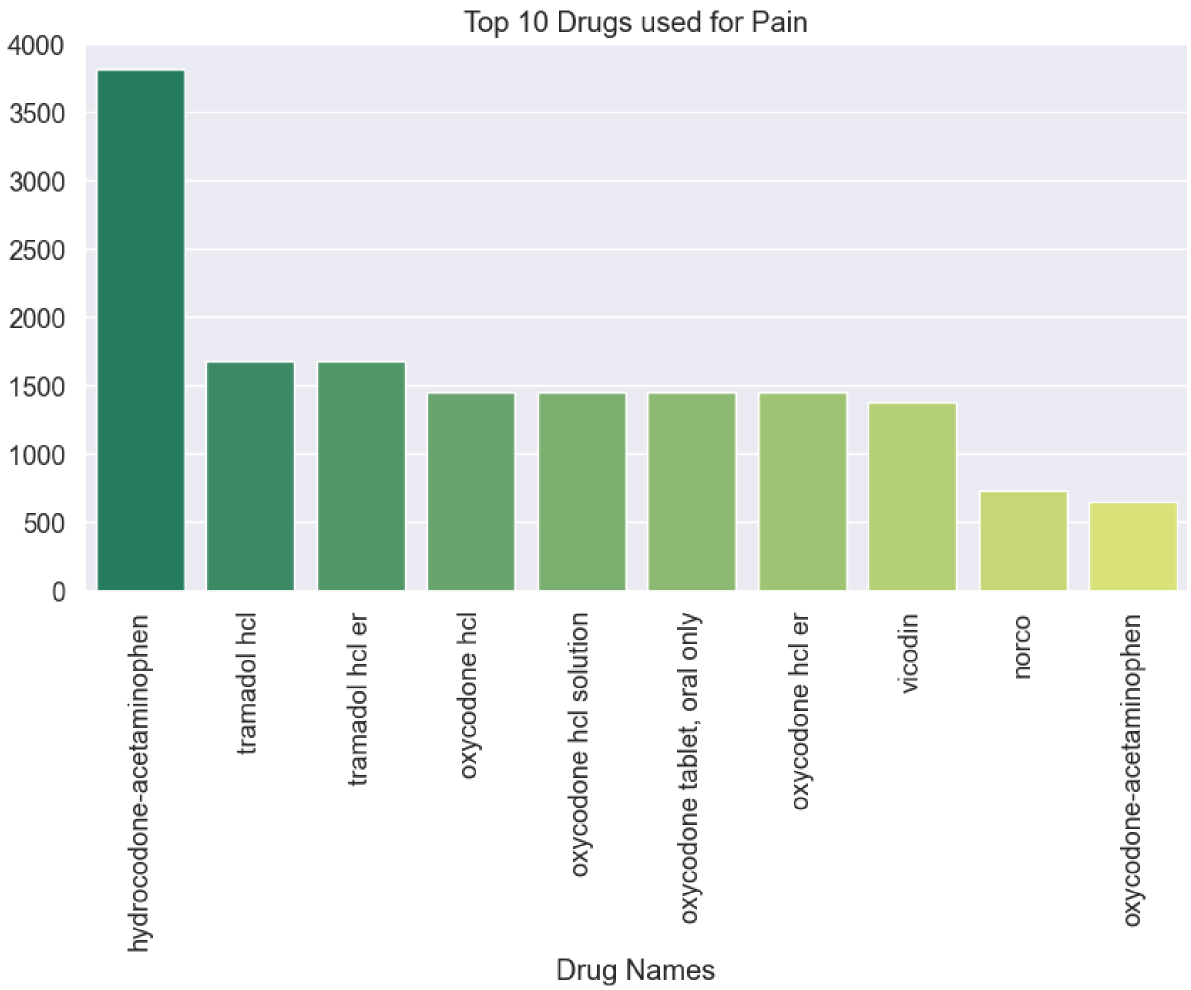

Count of Satisfaction
count
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10
Satisfaction

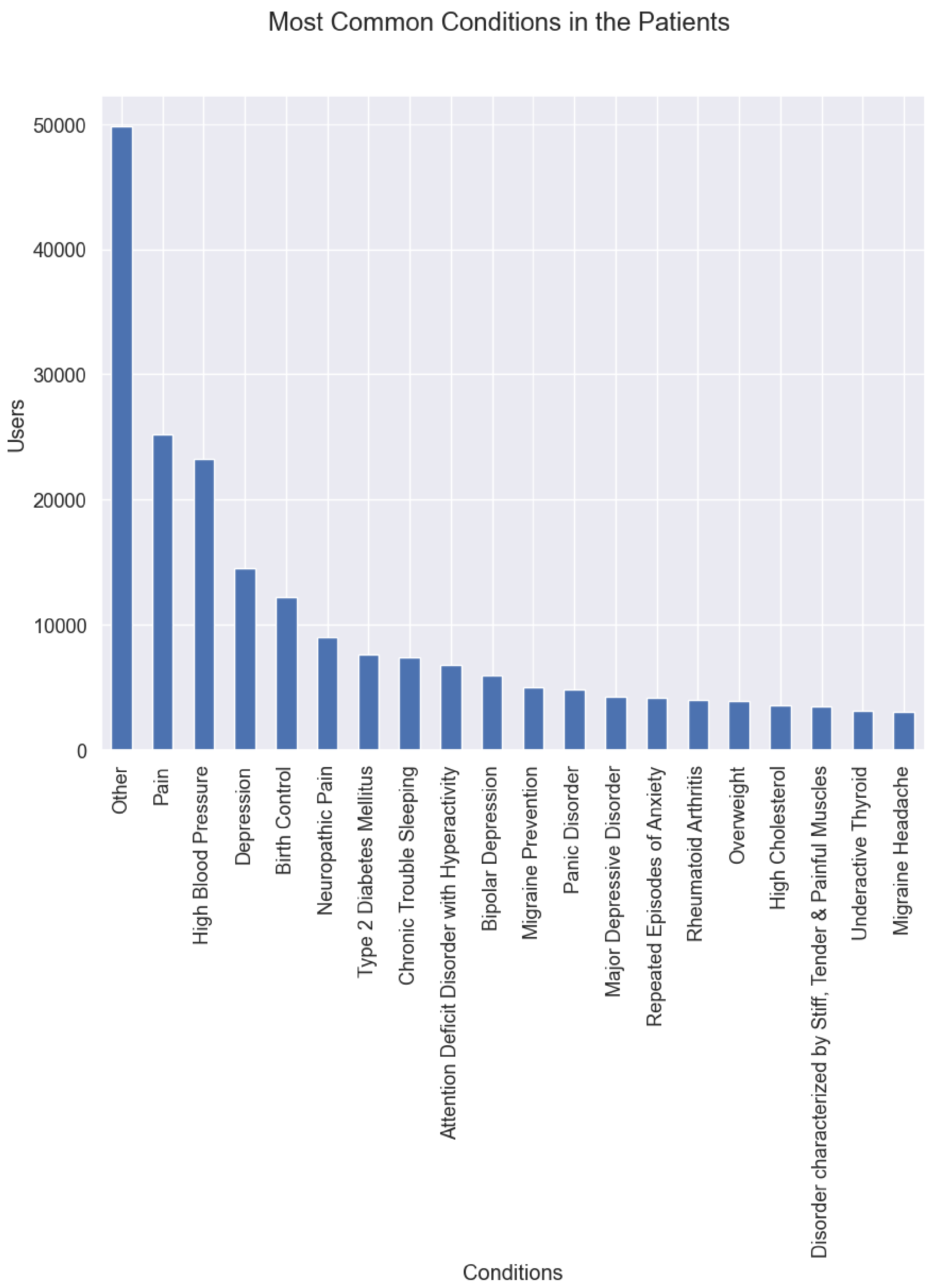

Distribution of Satisfaction
Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Satisfaction

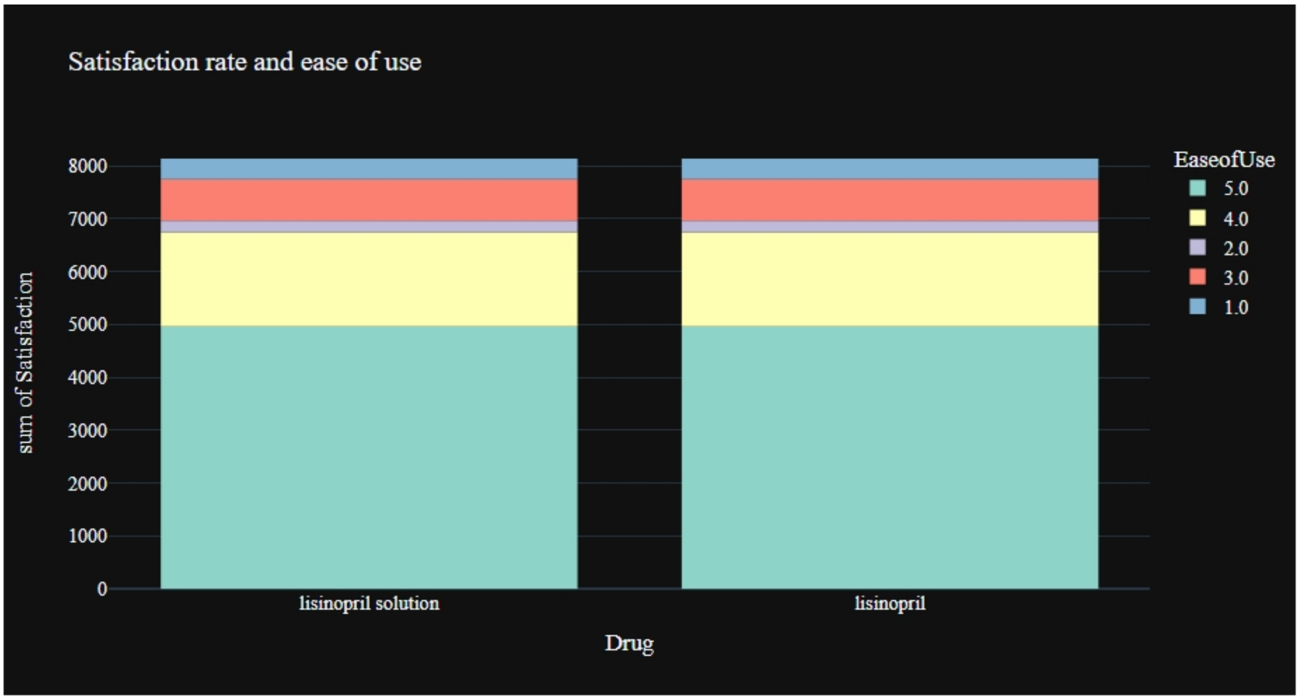

Count of Effectiveness
count
1500
1250
1000
750
500
250
0
1
2
3
4
5
6
7
8
9
10
Effectiveness

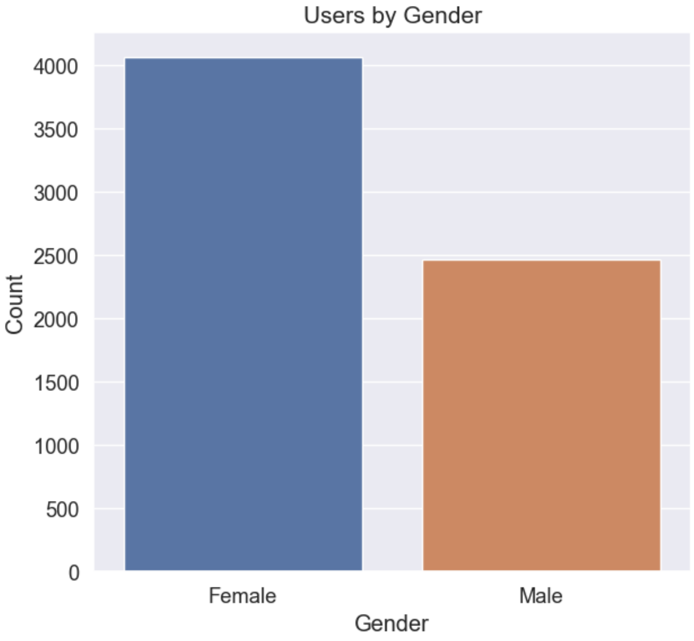

Distribution of Effectiveness
Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
Effectiveness

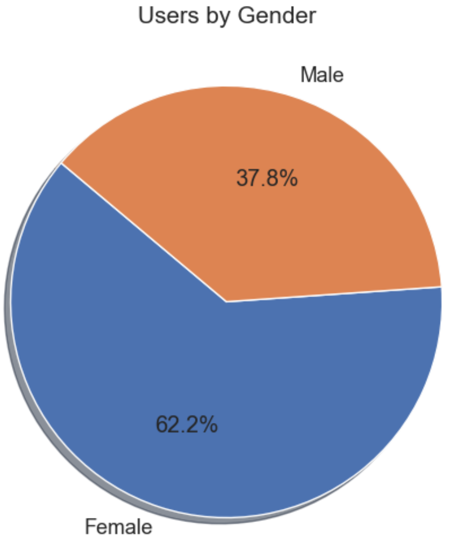

Most Commonly Used Drugs by Patient
Name
0
1000
2000
3000
4000
Drug
cymbalta
lisinopril
lisinopril solution
lexapro
hydrocodone-acetaminophen
effexor xr
lyrica
tramadol hcl er
tramadol hcl
zoloft
prednisone tablet, delayed release (enteric coated)
prednisone concentrate
prednisone
seroquel
phentermine hcl
celexa
topamax
topamax capsule, sprinkle
trazodone hcl
neurontin capsule

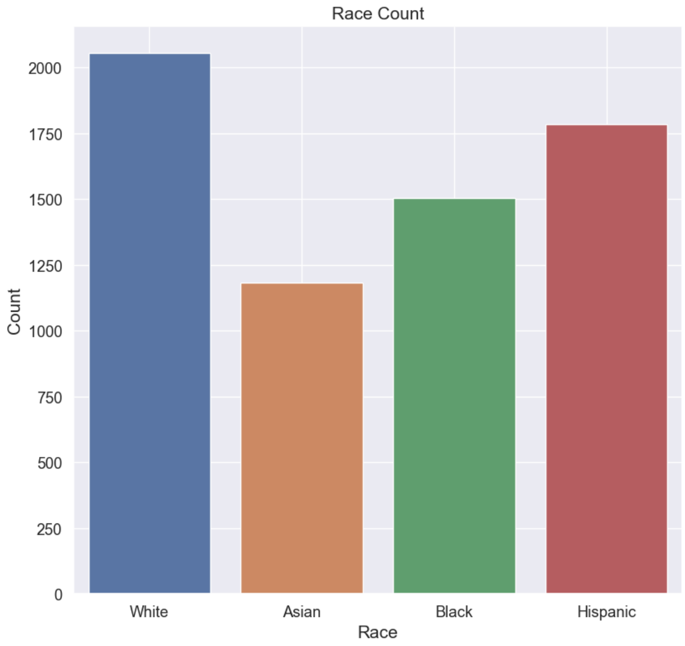

Top 10 Drugs used for Pain
4000
3500
3000
2500
2000
1500
1000
500
0
hydrocodone-acetaminophen
tramadol hcl
tramadol hcl er
oxycodone hcl
oxycodone hcl solution
oxycodone tablet, oral only
oxycodone hcl er
vicodin
norco
oxycodone-acetaminophen
Drug Names

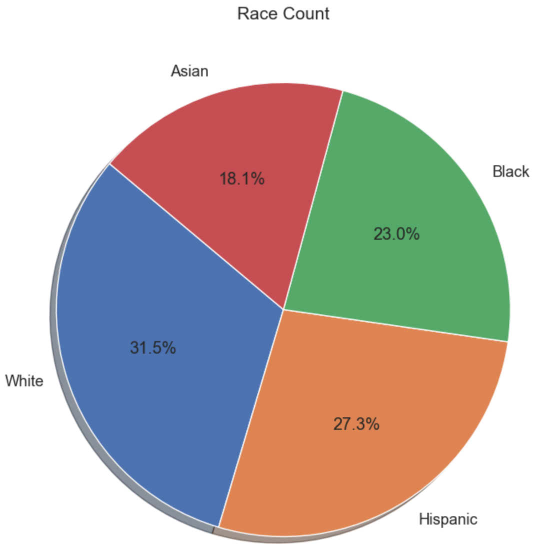

Most Common Conditions in the Patients
Users
50000
40000
30000
20000
10000
0
Other
Pain
High Blood Pressure
Depression
Birth Control
Neuropathic Pain
Type 2 Diabetes Mellitus
Chronic Trouble Sleeping
Attention Deficit Disorder with Hyperactivity
Bipolar Depression
Migraine Prevention
Panic Disorder
Major Depressive Disorder
Repeated Episodes of Anxiety
Rheumatoid Arthritis
Overweight
High Cholesterol
Disorder characterized by Stiff, Tender & Painful Muscles
Underactive Thyroid
Migraine Headache
Conditions

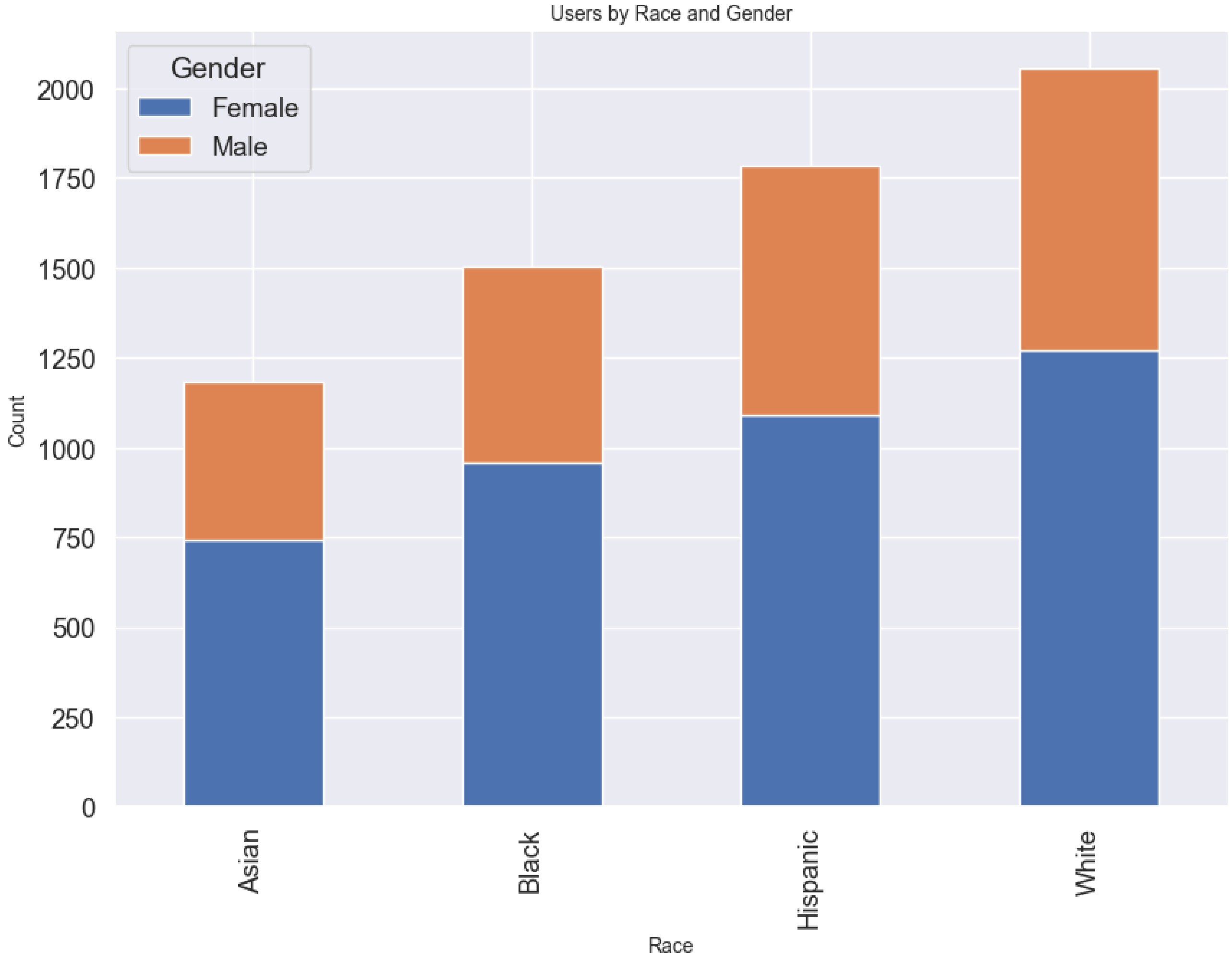

Satisfaction rate and ease of use
sum of Satisfaction
0
1000
2000
3000
4000
5000
6000
7000
8000
lisinopril solution
lisinopril
Drug
EaseofUse
5.0
4.0
2.0
3.0
1.0

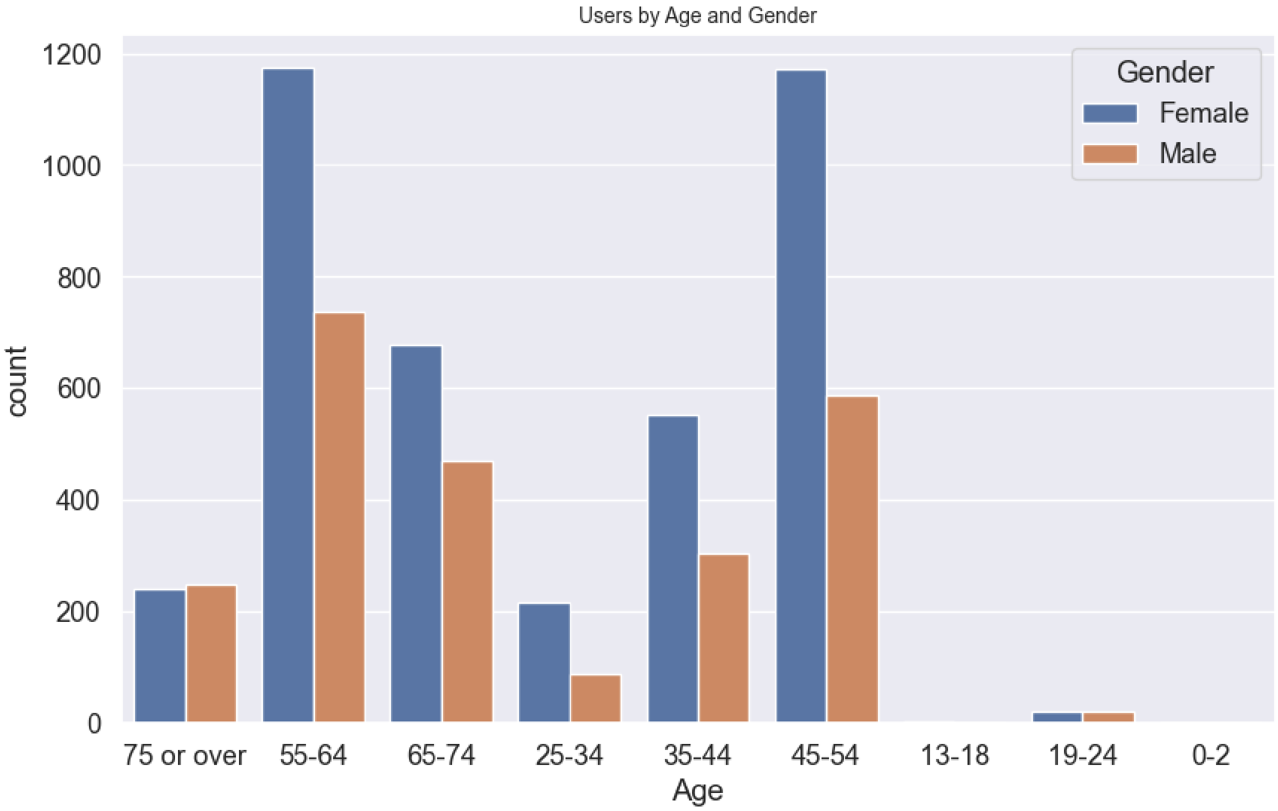

Users by Gender
Count
0
500
1000
1500
2000
2500
3000
3500
4000
Female
Male
Gender

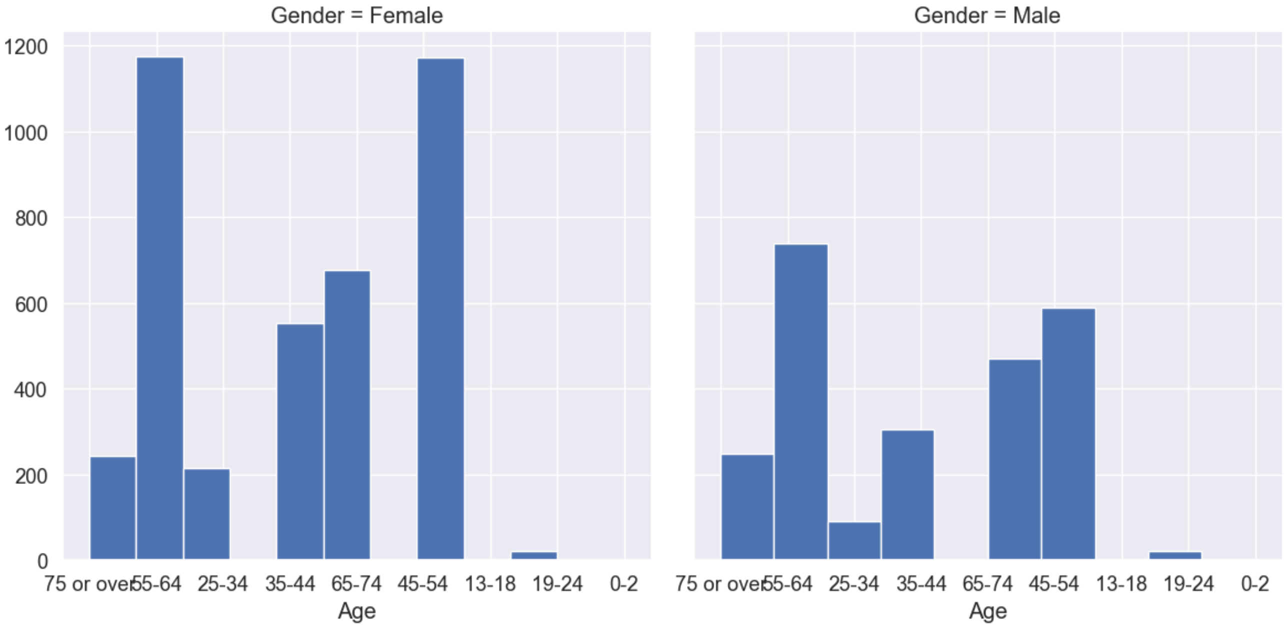

Users by Gender
Male
37.8%
62.2%
Female

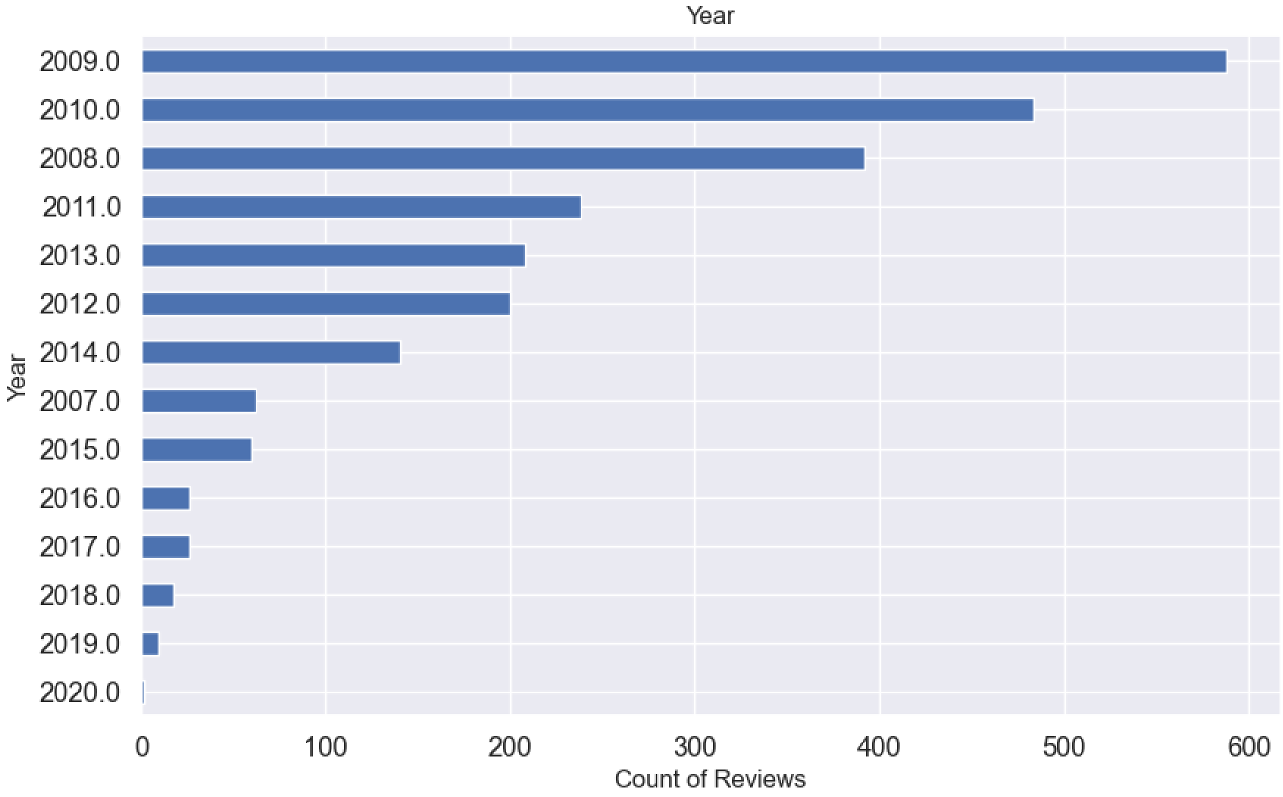

Race Count
2000
1750
1500
1250
1000
750
500
250
0
Count
White
Asian
Black
Hispanic
Race

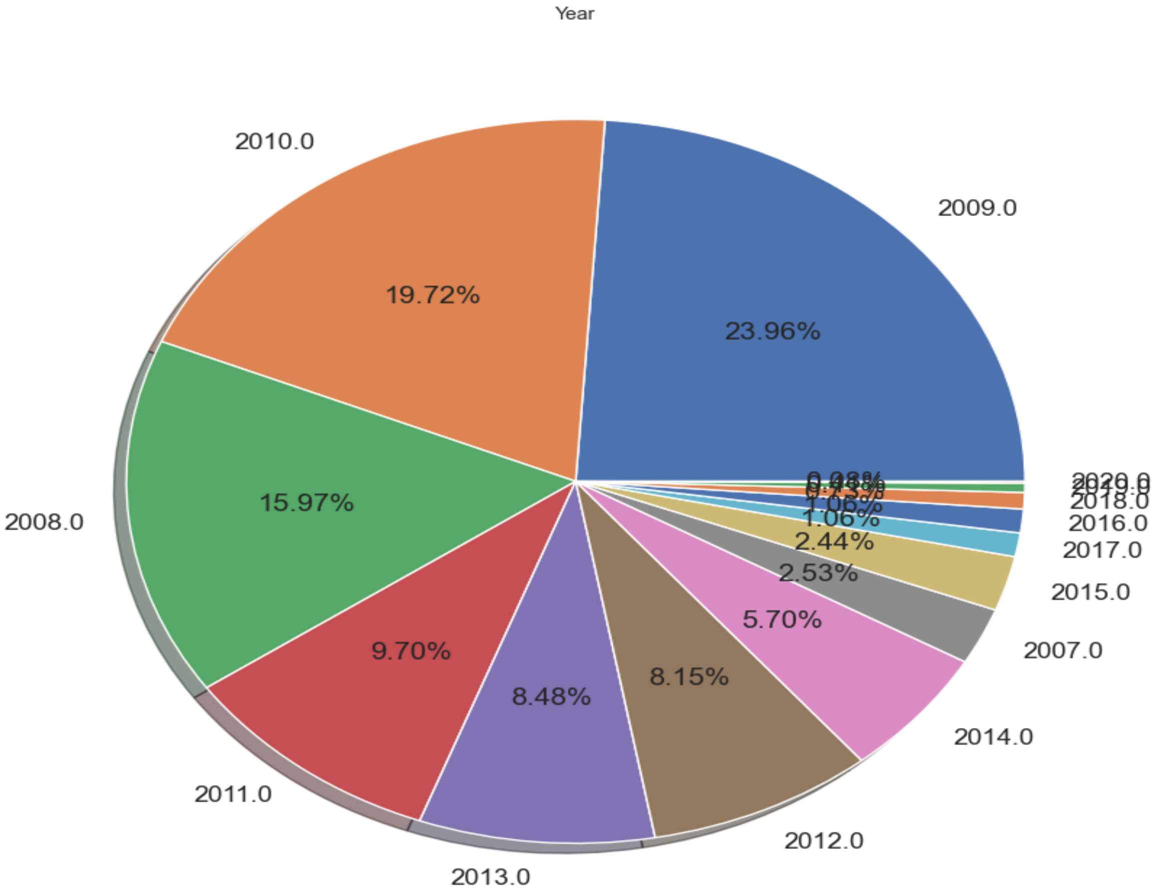

Race Count
Asian
18.1%
Black
23.0%
White
31.5%
27.3%
Hispanic

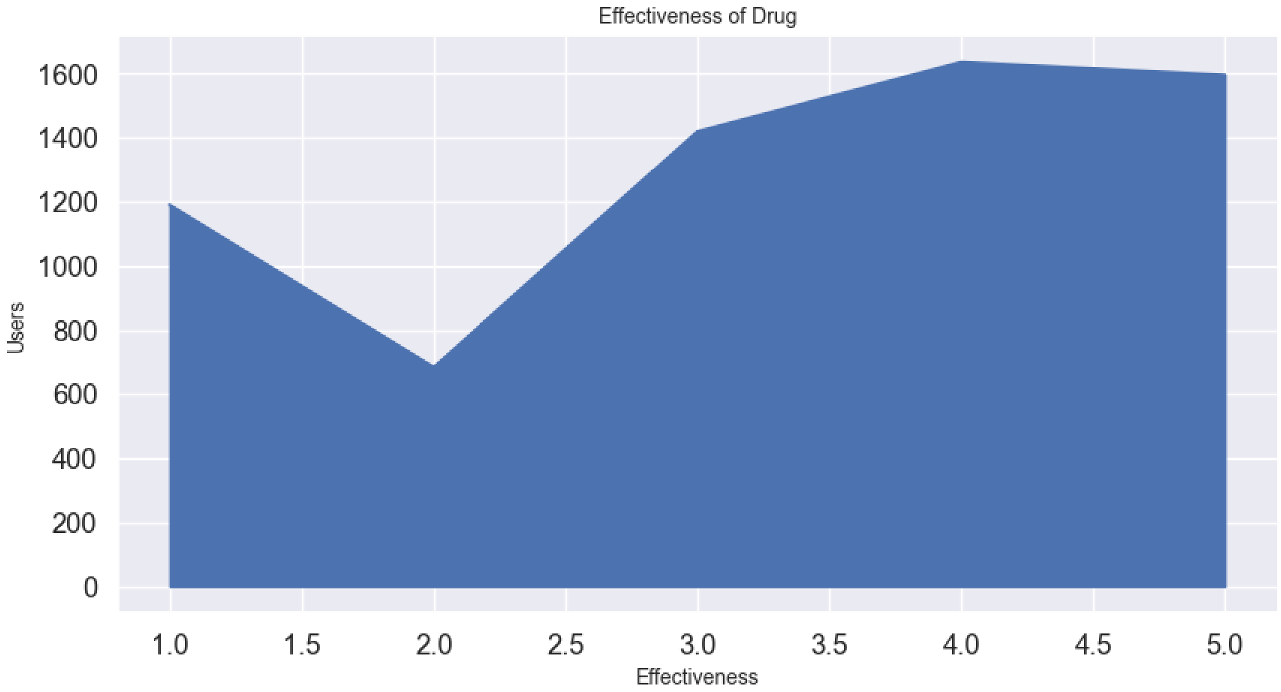

Users by Race and Gender
Gender
Female
Male
2000
1750
1500
1250
1000
750
500
250
0
Count
Asian
Black
Hispanic
White
Race

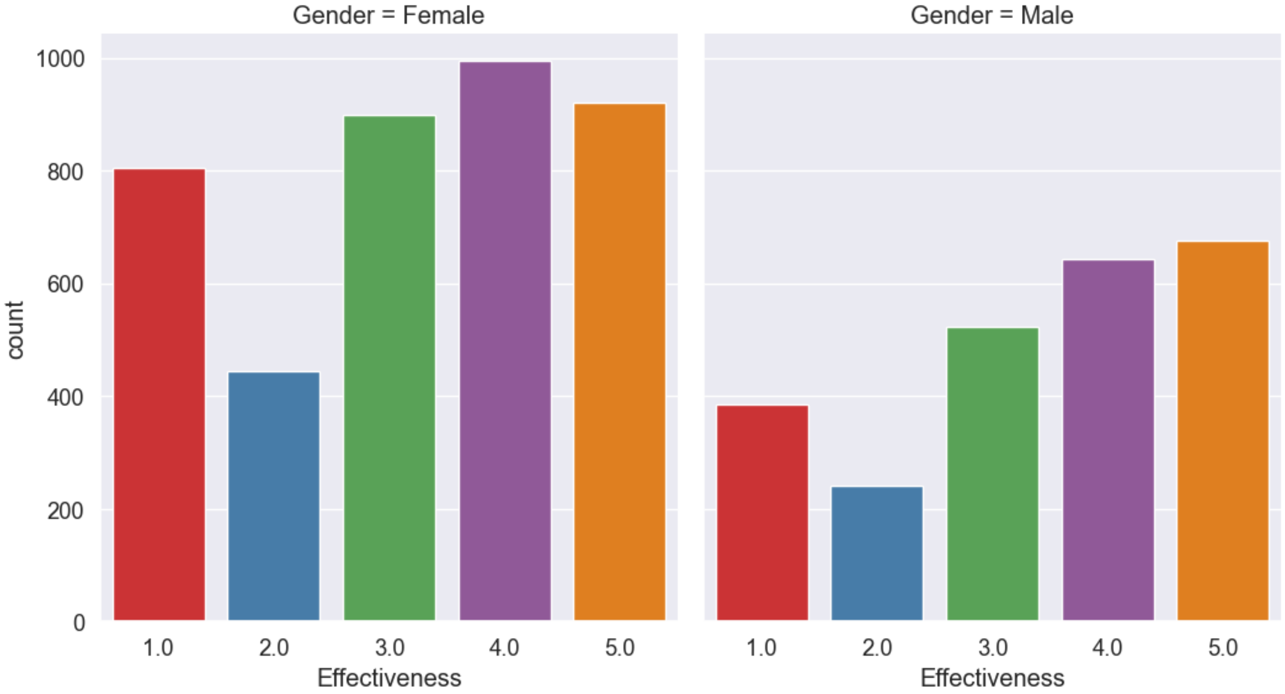

Users by Age and Gender
1200
1000
800
600
400
200
0
count
Gender
Female
Male
75 or over
55-64
65-74
25-34
35-44
45-54
13-18
19-24
0-2
Age

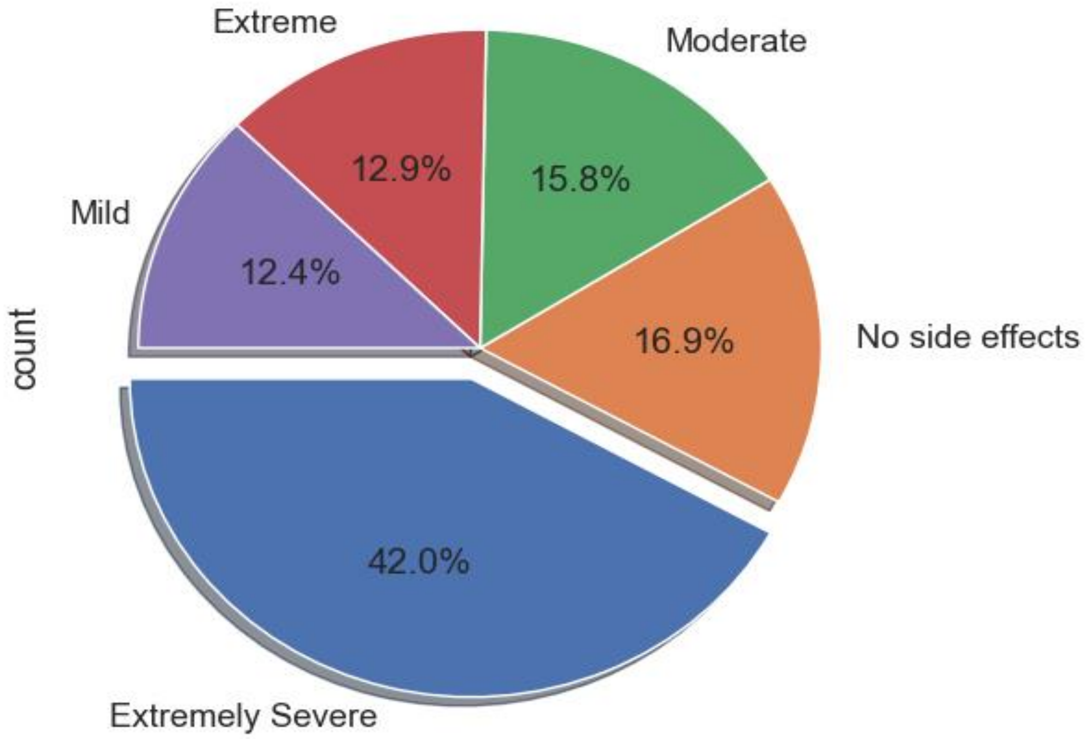

Gender = Female
Gender = Male
1200
1000
800
600
400
200
0
75 or over
55-64
25-34
35-44
65-74
45-54
13-18
19-24
0-2
Age

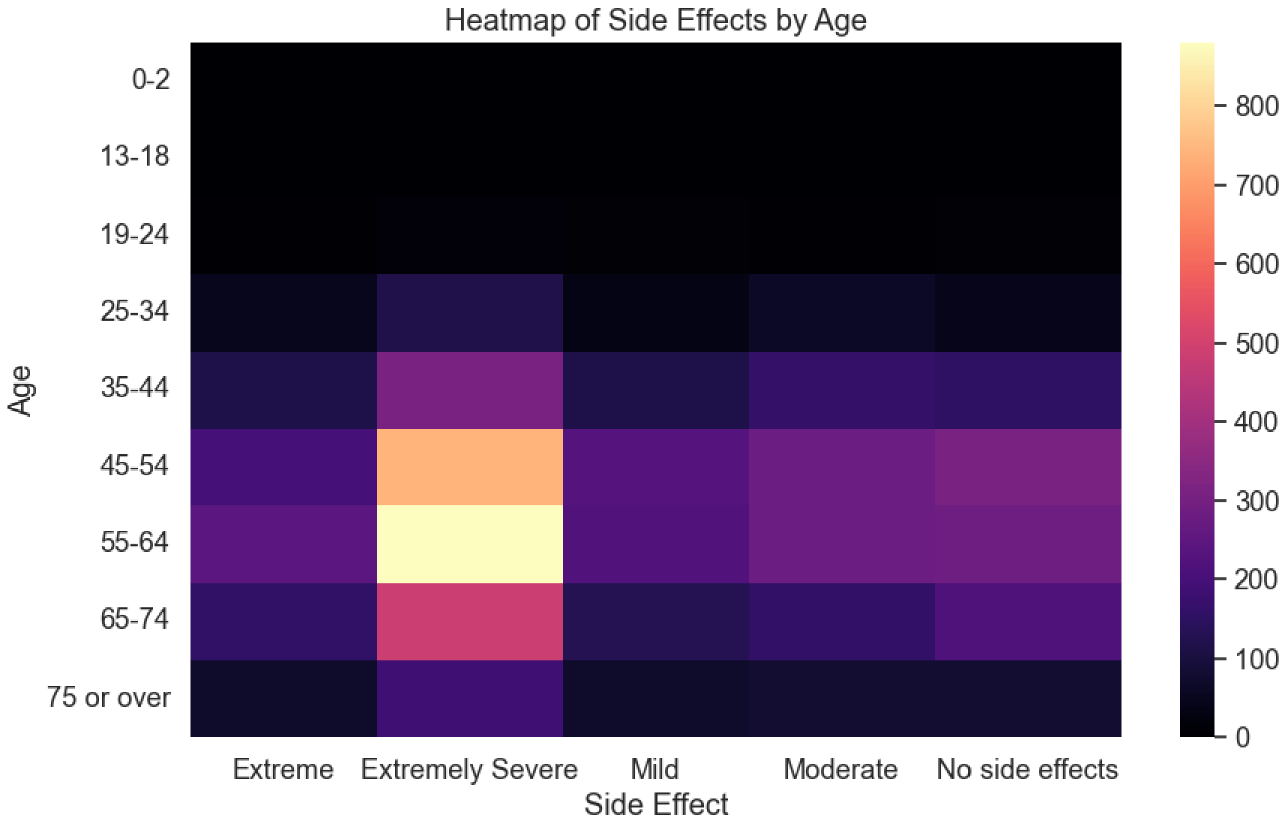

Year
2009.0
2010.0
2008.0
2011.0
2013.0
2012.0
2014.0
2007.0
2015.0
2016.0
2017.0
2018.0
2019.0
2020.0
Year
0
100
200
300
400
500
600
Count of Reviews

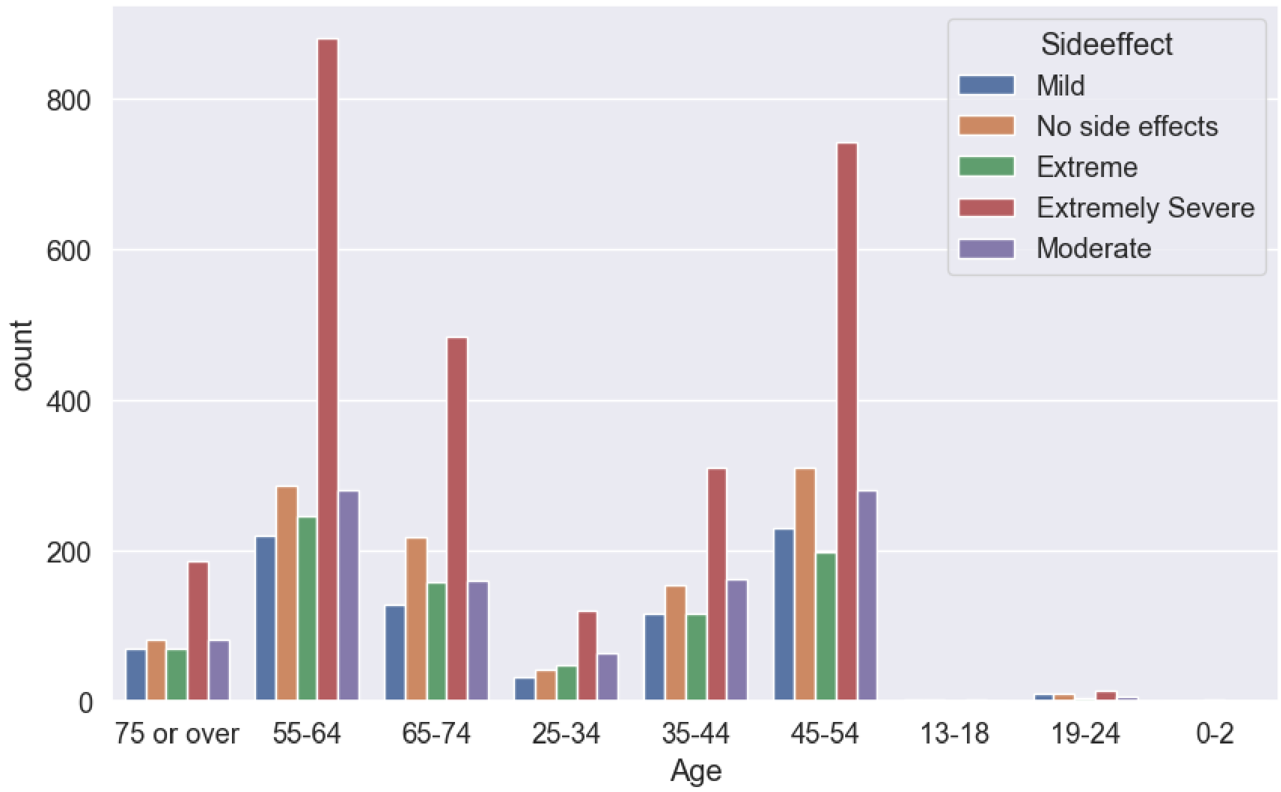

Year
2010.0
19.72%
2009.0
23.96%
2008.0
15.97%
2011.0
9.70%
2013.0
8.48%
2012.0
8.15%
2014.0
5.70%
2007.0
2.53%
2015.0
2.44%
2017.0
1.06%
2016.0
1.06%
2018.0
0.73%
2019.0
2020.0
0.08%

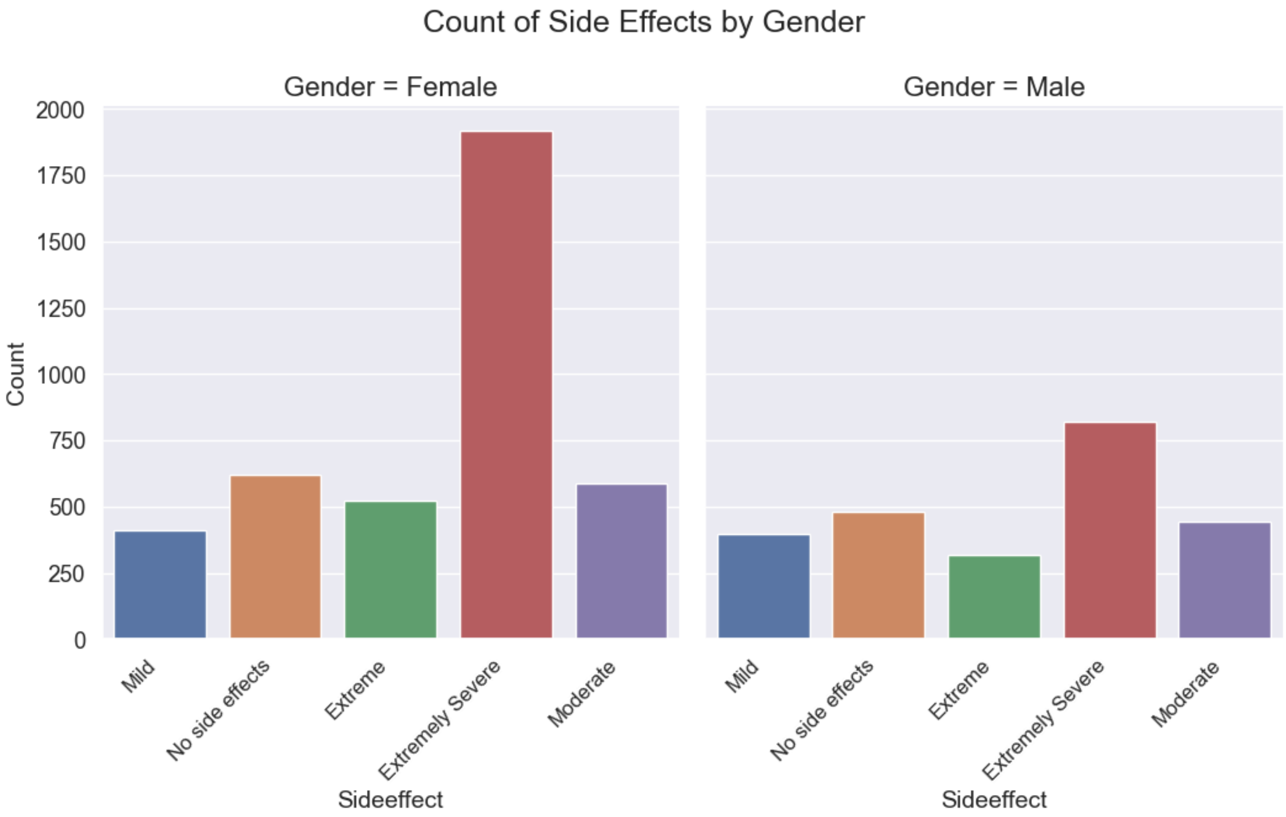

Effectiveness of Drug
Users
1600
1400
1200
1000
800
600
400
200
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Effectiveness

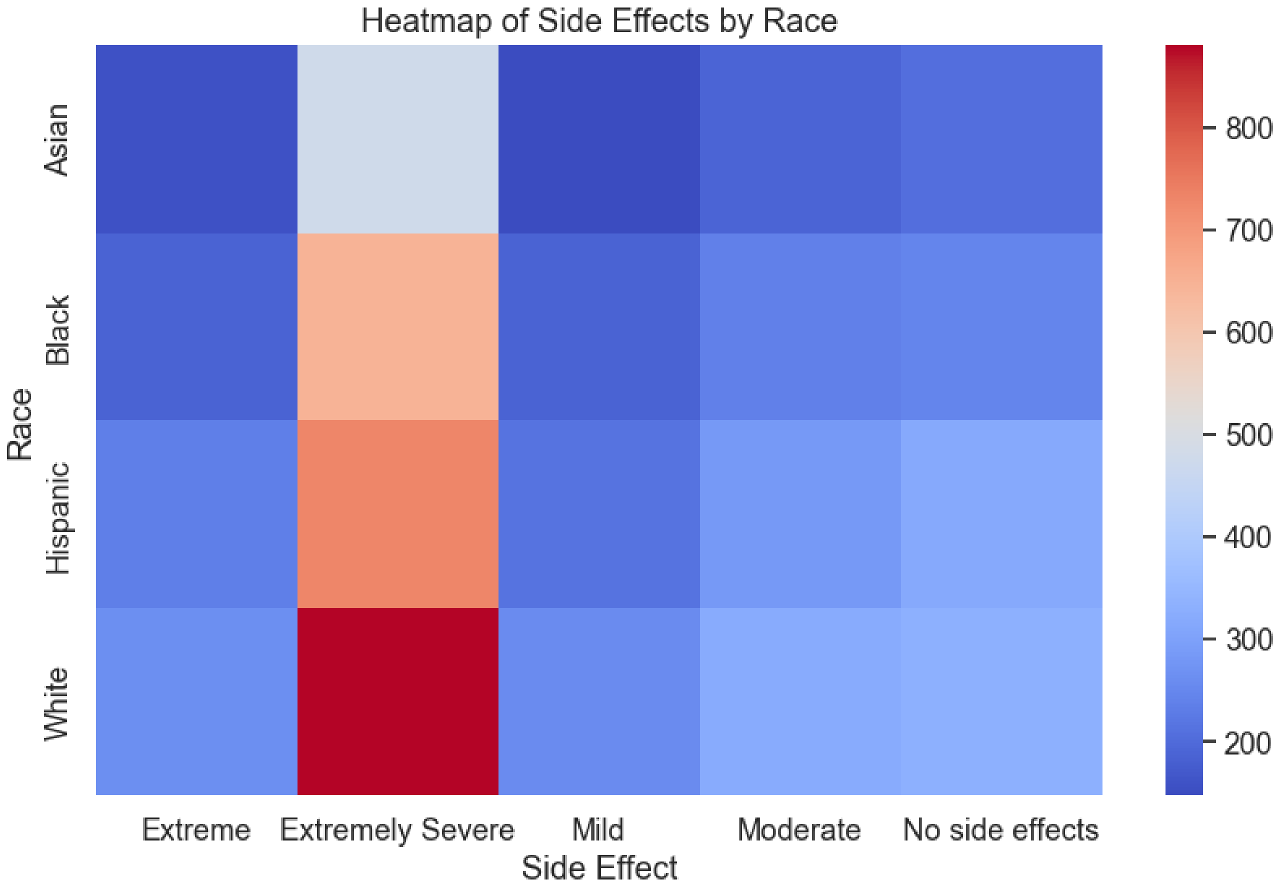

Gender = Female
Gender = Male
count
1000
800
600
400
200
0
1.0
2.0
3.0
4.0
5.0
Effectiveness
1.0
2.0
3.0
4.0
5.0
Effectiveness

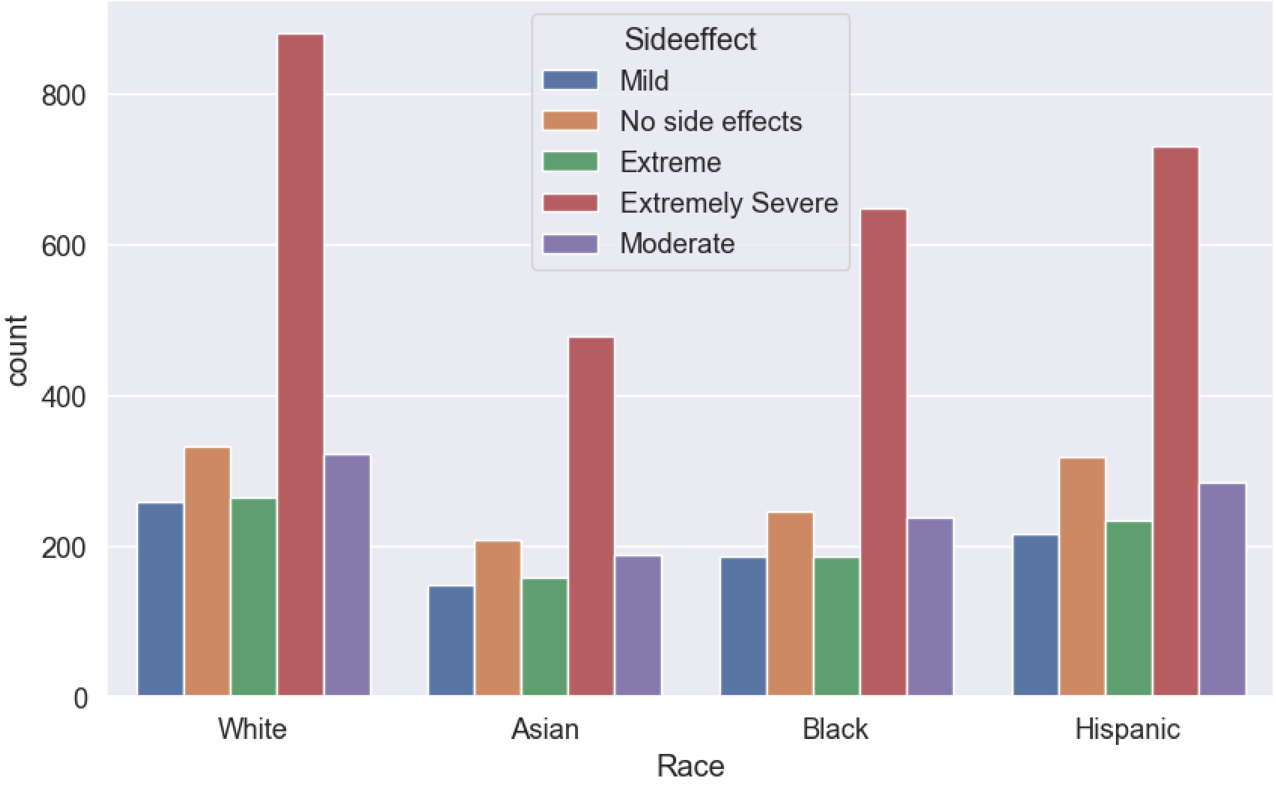

Extreme
Moderate
Mild
No side effects
Extremely Severe
count
12.9%
15.8%
12.4%
16.9%
42.0%

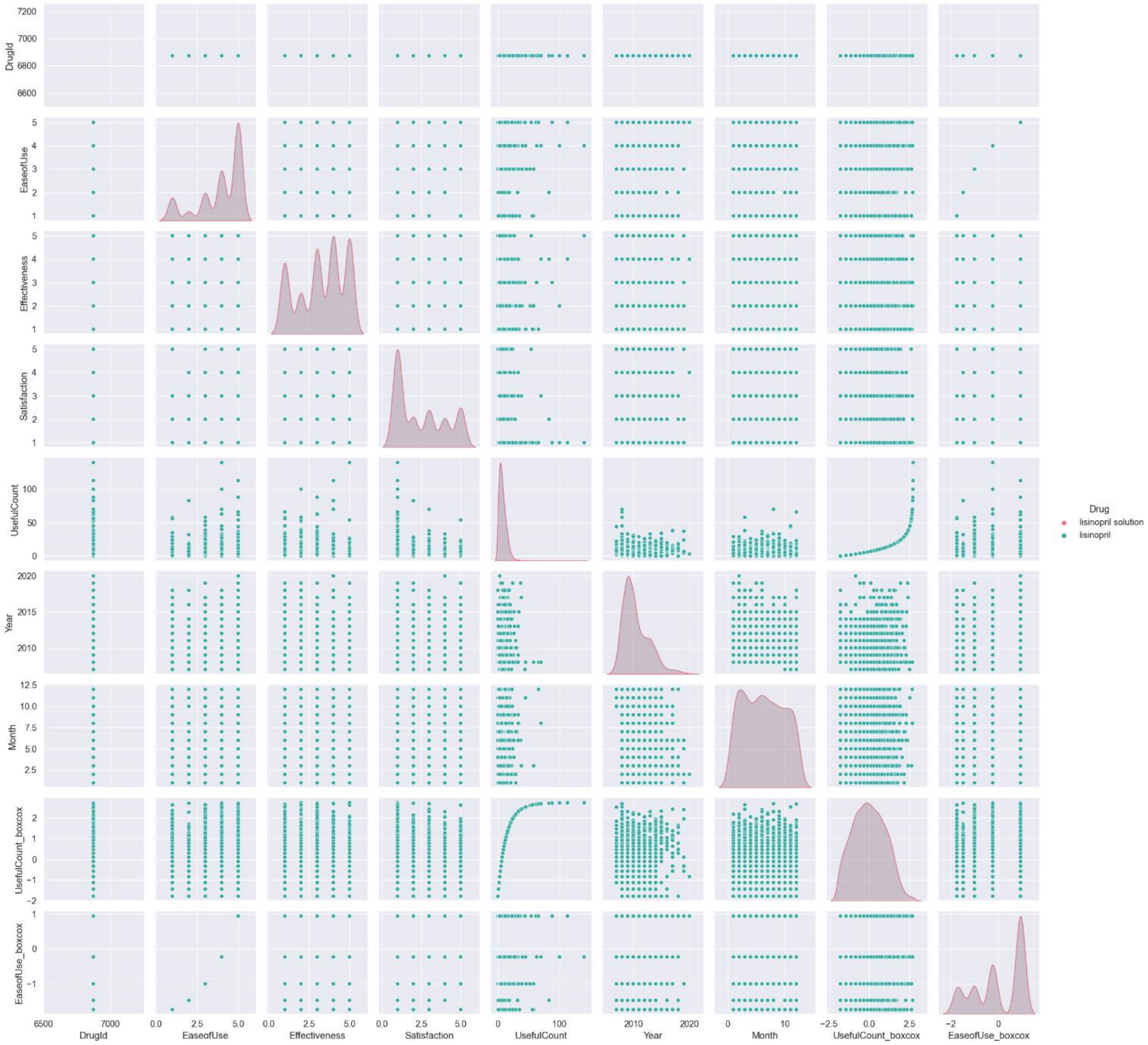

Heatmap of Side Effects by Age
Age
0-2
13-18
19-24
25-34
35-44
45-54
55-64
65-74
75 or over
Extreme
Extremely Severe
Mild
Moderate
No side effects
Side Effect
800
700
600
500
400
300
200
100
0


Sideeffect
Mild
No side effects
Extreme
Extremely Severe
Moderate
count
800
600
400
200
0
75 or over
55-64
65-74
25-34
35-44
45-54
13-18
19-24
0-2
Age


Count of Side Effects by Gender
Gender = Female
Gender = Male
Count
2000
1750
1500
1250
1000
750
500
250
0
Mild
No side effects
Extreme
Extremely Severe
Moderate
Sideeffect
Mild
No side effects
Extreme
Extremely Severe
Moderate
Sideeffect


Heatmap of Side Effects by Race
Race
Asian
Black
Hispanic
White
Side Effect
Extreme
Extremely Severe
Mild
Moderate
No side effects
800
700
600
500
400
300
200


Sideeffect
Mild
No side effects
Extreme
Extremely Severe
Moderate
count
0
200
400
600
800
Race
White
Asian
Black
Hispanic



Lisinopril, the most commonly used drug shows side effects but its effectiveness and satisfaction rate are good. The proportion of female users is consistently higher across all racial and age groups. Females reported more severe side effects as well as higher drug effectiveness compared to male users. The race distribution of users revealed that White individuals were the most predominant users. The Heatmap analysis also revealed that the 60+ age group reported the most side effects. The number of reviews increased until 2014, then slightly declined, possibly indicating users' adaptation to side effects or a focus on treating illness with less concern for side effects.

10. ALGORITHMS

For this project, various machine learning (ML) algorithms will be developed and then models will be compared for predicting adverse drug reactions (ADRs). Machine learning algorithms to be used for this project are:

1. Logistic Regression

A linear model used for binary classification tasks. It estimates the probability of a binary outcome (e.g., the presence or absence of side effects) by combining input features linearly. The model applies the logistic function to convert linear combinations into probabilities.

2. Decision Tree

Non-linear models that partition data into subsets based on feature values to make predictions. They recursively split the data into branches, aiming to increase the purity of the resulting subsets.

3. Random Forest

An ensemble learning method that improves the accuracy and robustness of predictions by constructing multiple decision trees and combining their outputs. Each tree is trained on a random subset of the data and a random subset of features.

4. Support Vector Machine (SVM)

SVM is a powerful classification algorithm that finds the optimal hyperplane to separate classes by maximizing the margin between the closest points of different classes. SVMs can handle non-linear relationships using kernel functions.

5. K-Nearest Neighbors (KNN)

KNN is a simple, instance-based learning algorithm that predicts the class of a sample based on the majority class of its closest neighbors. The distance between samples is typically measured using Euclidean distance.

6. Naive Bayes

A probabilistic classifier based on Bayes' theorem, assuming independence between features. Despite its simplicity, Naive Bayes is efficient and often performs well in high-dimensional datasets, particularly in text classification.

7. Gradient Boosting

An ensemble technique that builds models sequentially, each one correcting the errors of its predecessor. It combines weak learners to form a strong predictive model.

These algorithms will collectively provide a robust framework for predicting drug side effects, each contributing unique strengths to handle the complexity and variability of the dataset.

11. CHALLENGES & OPPORTUNITIES

CHALLENGES:-

- The data collected may not be accurate or complete.
- The dataset may not represent the entire population.
- The age and gender information provided in the dataset may not be accurate or complete.
- The dataset may not include all possible side effects for each drug.
- The dataset may not include information on all drugs or medications.
- The dataset may not be up-to-date or accurate.
- Developing a classification model that can accurately predict side effects for all individuals is a challenging task.
- The classification model developed may not be generalizable to other datasets.

OPPORTUNITIES:-

- Despite inherent limitations, such as the necessity of considering all factors affecting side effects and ensuring dataset accuracy, developing such a model can help healthcare professionals make more informed decisions about prescribing medications and help patients make more informed decisions about their health.
- Another opportunity is that the model can be used to identify previously unknown side effects of drugs, which can lead to new discoveries and better treatment options.
- The model can be used to identify patterns in side effects across different drugs and patients, which can help researchers better understand the underlying mechanisms of drug side effects.
- The model can be used to develop personalized treatment plans based on an individual's age, gender, and other factors that may impact their risk of experiencing side effects.
- Enhanced classification models can support healthcare professionals in making more informed medication decisions and enable patients to make better health decisions.

12. RISK VS REWARD

Risk	Reward
As a risk, developing a classification model that can accurately predict side effects for all individuals is a challenging task. It's crucial to account for all factors that can impact side effects, and the dataset may not include all possible side effects for each drug. The dataset may also not include information on all drugs or medications, and it may not be up-to-date or accurate. It is important to carefully consider the risks and rewards of any treatment or intervention thoughtfully before making a decision. Healthcare providers should work closely with their patients to ensure that they understand the potential benefits and possible adverse reactions of different treatments and can make well-informed decisions about their healthcare.	As a reward, Healthcare professionals can make more informed decisions while prescribing medications and assist patients make more informed decisions regarding their health Moreover, it enables the identification of previously undisclosed drug side effects, potentially uncovering new insights and improving treatment options. Additionally, the model can be used to identify patterns in side effects across various drugs and patient groups, which can help researchers better understand the underlying mechanisms of drug side effects.

13. REFLECTIONS ON THE INTERNSHIP

Developing a classification model that can accurately predict drug side effects based on age and gender. The project also highlighted the importance of data preprocessing and cleaning to ensure data accuracy and completeness. Additionally, the project demonstrated the potential of machine learning techniques to identify adverse drug reactions and patterns in side effects across different medications and patient demographics.

By analyzing a dataset containing patient demographics, drug usage, and ratings, we aim to develop a reliable predictive model. Future research could focus on developing more accurate classification models to comprehensively address factors influencing side effects and their generalization to other datasets. The insights obtained from these models could empower healthcare providers to make well-informed decisions and personalize treatment plans to minimize adverse effects.

14. RECOMMENDATIONS

- **Data Collection**: Collect data from multiple sources to increase the diversity of the dataset and improve the generalizability of the model.
- **Feature Engineering**: Explore different feature engineering techniques to extract meaningful information from the dataset. Consider incorporating additional features such as drug dosage, treatment duration, and patient medical history to improve the model's performance.
- **Model Selection**: Experiment with various machine learning algorithms. Consider ensemble methods or deep learning models to further enhance the model's predictive capabilities.
- **Hyperparameter Tuning**: Optimize the hyper parameters of the selected model using techniques such as grid search or random search. This can help improve the model's generalization ability and prevent overfitting.
- **Validation and Testing**: Use appropriate validation techniques such as cross-validation to estimate the model's performance on unseen data. Perform rigorous testing on an independent test set to assess the model's real-world performance.
- **Interpretability**: Use interpretable machine learning models such as decision trees or logistic regression to provide insights into the factors influencing side effects.
- **Continual Improvement**: Regularly update the model with new data to ensure its relevance and accuracy over time. Monitor its performance in real-world scenarios and refine it based on feedback from healthcare professionals and end-users.

15. OUTCOME / CONCLUSION

In conclusion, from the EDA analysis, I concluded that the lisinopril drug has side effects but its effectiveness and satisfaction rate are good. The proportion of female users is consistently higher across all racial and age groups. Females reported more severe side effects as well as higher drug effectiveness compared to male users. The gender distribution indicates that females are more proactive in managing their health and sharing their review of medications. The race distribution of users revealed that White individuals were the most predominant users, followed by Hispanic, Black, and Asian individuals. The stacked bar plot indicated a higher prevalence of female users across all races, with the highest user count observed in the White community. The bar plot analysis revealed that female users generally reported higher drug effectiveness compared to male users. A lower effectiveness rating could potentially signal the presence of side effects. The side effects varied between genders, with female users reporting more extreme side effects compared to male users. The data indicated that female users dominate across all age groups, except in the ‘0-17’ age group where male users were slightly more in number. The analysis revealed that the 60+ age group reported the most side effects, suggesting that older individuals experience more side effects than other age groups. The distribution of the pie chart provides valuable insights into the side effects profile of the drug. 42% of users reported extreme side effects. Analysis done on the reviews collected over the years revealed an increasing trend in the number of reviews until 2014, followed by a slight decline. This trend indicates an increasing awareness and use of online platforms for reporting drug experiences during the early 2010s, followed by a stabilization or shift in user behavior in subsequent years.

Despite inherent limitations, such as the necessity of considering all factors affecting side effects and ensuring dataset accuracy, there are significant opportunities to enhance patient care and advance medical research. Enhanced classification models can support healthcare professionals in making more informed medication decisions and enable patients to make better health decisions.

16. ENHANCEMENT SCOPE

Regularly update the model with new data to ensure its relevance and accuracy over time. Monitor its performance in real-world scenarios and refine it based on the feedback received from healthcare professionals and patients.

17. LINK TO THE CODE AND EXECUTABLE FILE

- https://github.com/Frazahmed98/Side-Effects-Classification