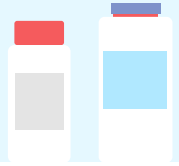
Pharmalytix

Intelligent Time Series Analysis, Sentiment Analysis and Drug Recommendation

Team-10:
Devyani Srivastava
Vamsi Gontu
Ashwathi Subramanian

Course: CSCI 5502-072 - Data Mining

Importance of Project



The pharmaceutical industry struggles with accurately determining drug prices for new and existing drugs, leading to price fluctuations and increased drug costs for patients. An forecasting system is needed to provide accurate drug pricing and enable healthcare providers to identify most-effective and cost-effective drugs for their patients. This would promote pricing transparency and ensure equitable access to necessary medications.

How did we reach our final findings

Project Roadmap

Week 5

Project Pitching

Week 8

Data Collection
Data Preprocessing
Data Manipulation
EDA

Week 10

Sales Forecasting
Dashboard MVP 1

Week 13

Sentiment Analysis
Recommendation Engine
Testing the Models
Dashboard MVP 2

Week 15

Final Presentation
App/Dashboard
Release

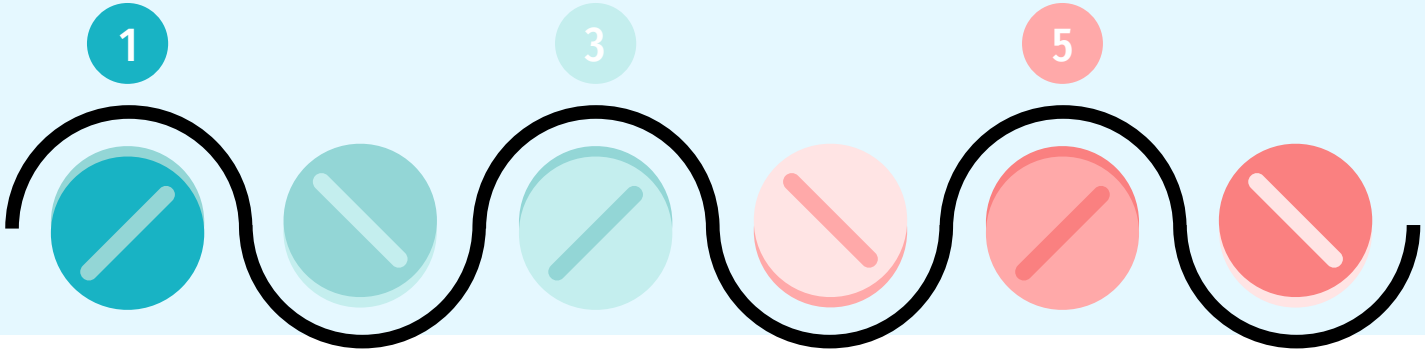
01 Time Series Analysis

Time Series Analysis and Forecasting

Time series forecasting is a technique used in statistics and data analysis to **predict future values of a series based on historical data points,** with the aim of identifying trends, patterns, and seasonal fluctuations.

is used in different fields for time-based predictions – like Weather Forecasting models, Stock market predictions, Signal processing, Engineering domain – Control Systems, and Communications Systems.

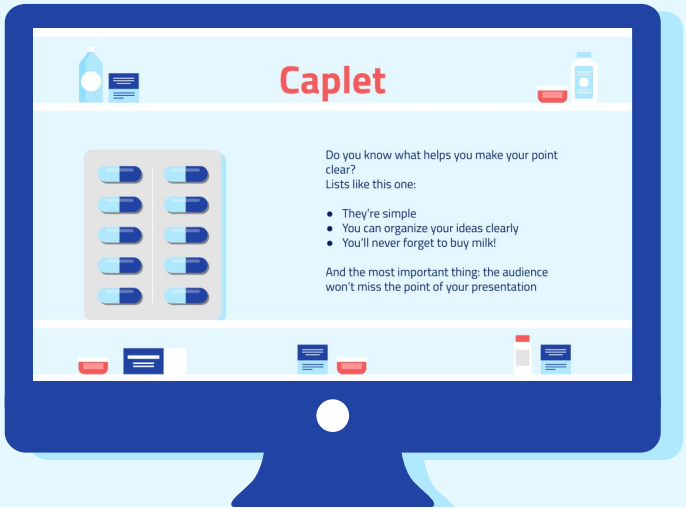
Time Series Analysis (TSA) Steps



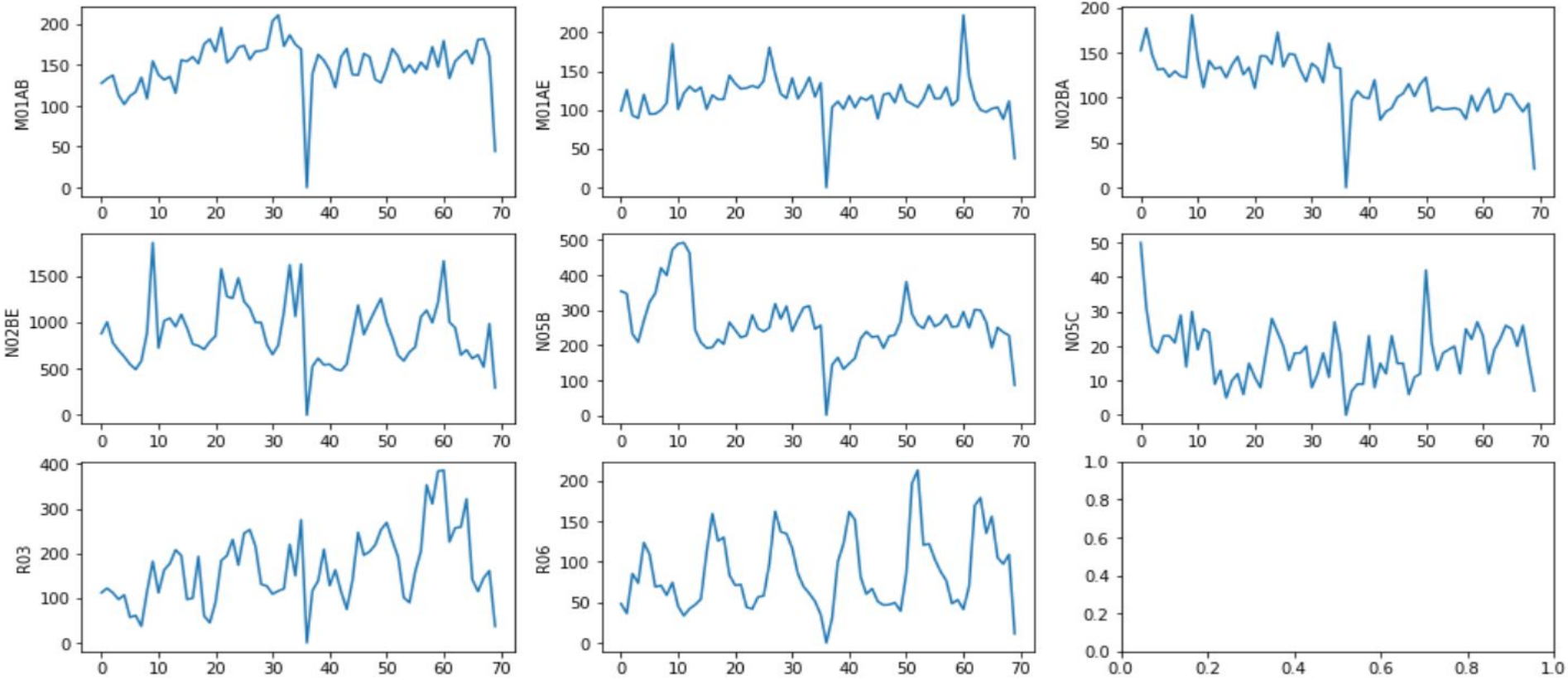
Data Source

For the purpose of sales forecasting due to limited availability of pharma data we used another dataframe which we fetched using kaggle API.

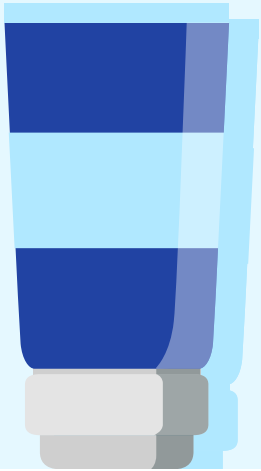
Source: https://www.kaggle.com/datasets/milanzdravkovic/pharma-sales-data



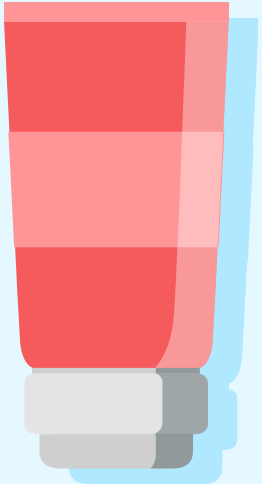
Features vs Time(monthly)



Augmented Dickey-Fuller (ADF) Test

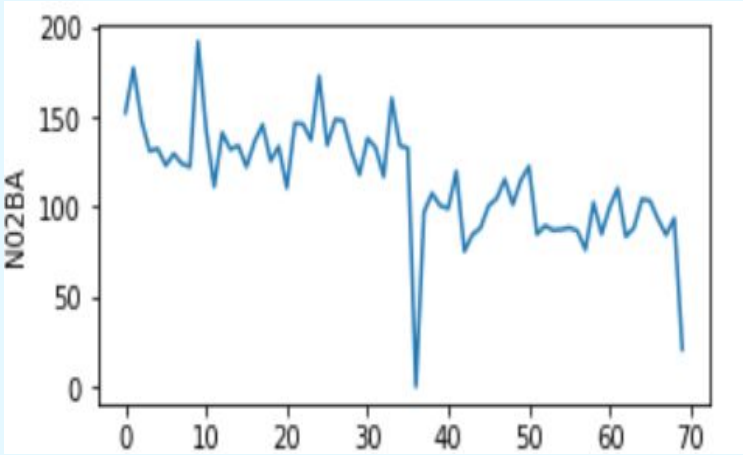


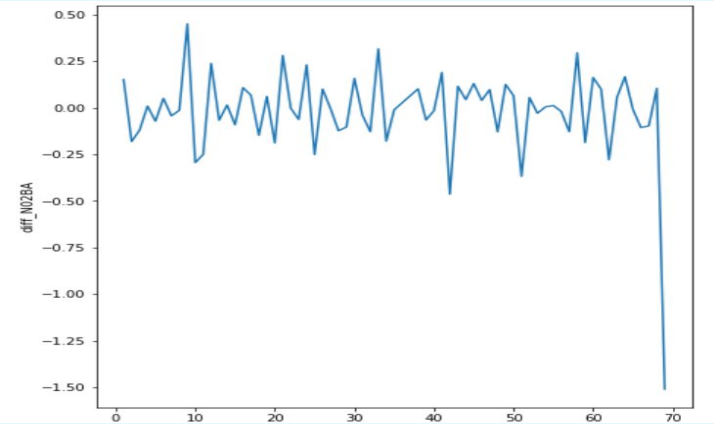
Is a commonly used statistical test in time series analysis to check for stationarity.

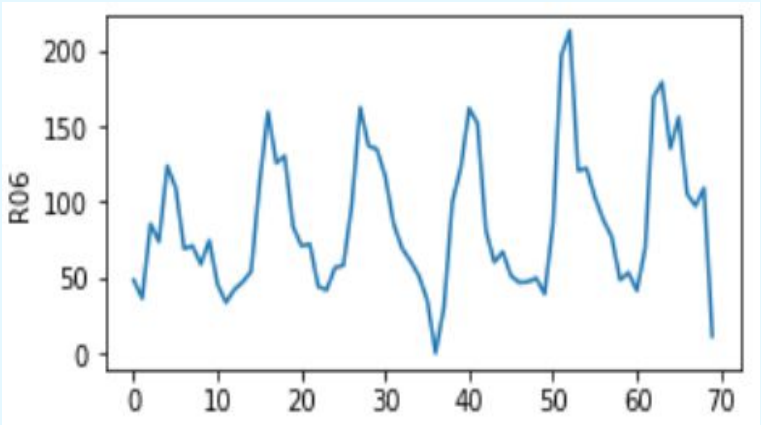


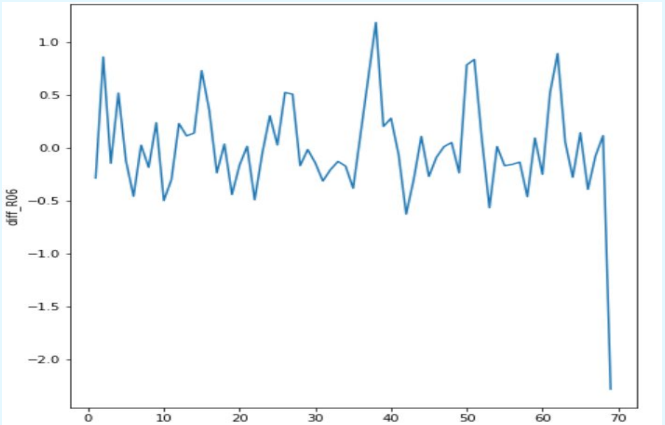
Null Hypothesis (H0): Series is non-stationary
Alternate Hypothesis (HA): Series is stationary
p-value >0.05 Fail to reject (H0)
p-value <= 0.05 Accept (H1)

N02BA and R06 are non Stationary. To change them to stationary values we performed log transformation and differencing







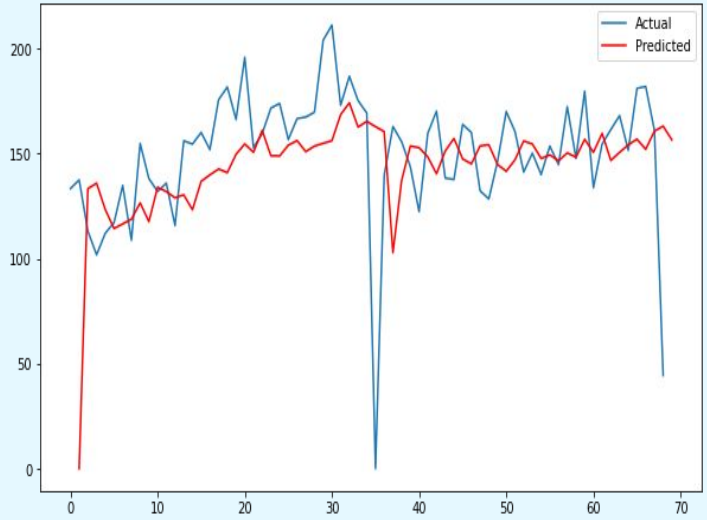


Before and after transformation

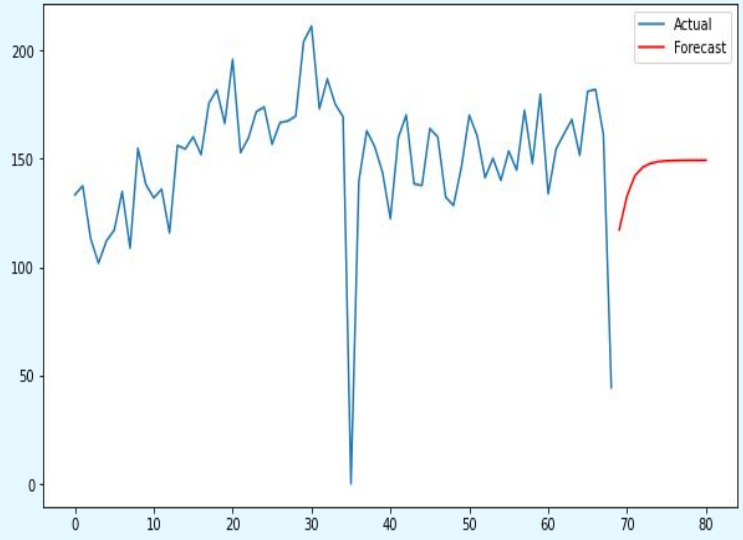
Model Results

We used ARIMA (AutoRegressive Integrated Moving Average) for sales forecasting

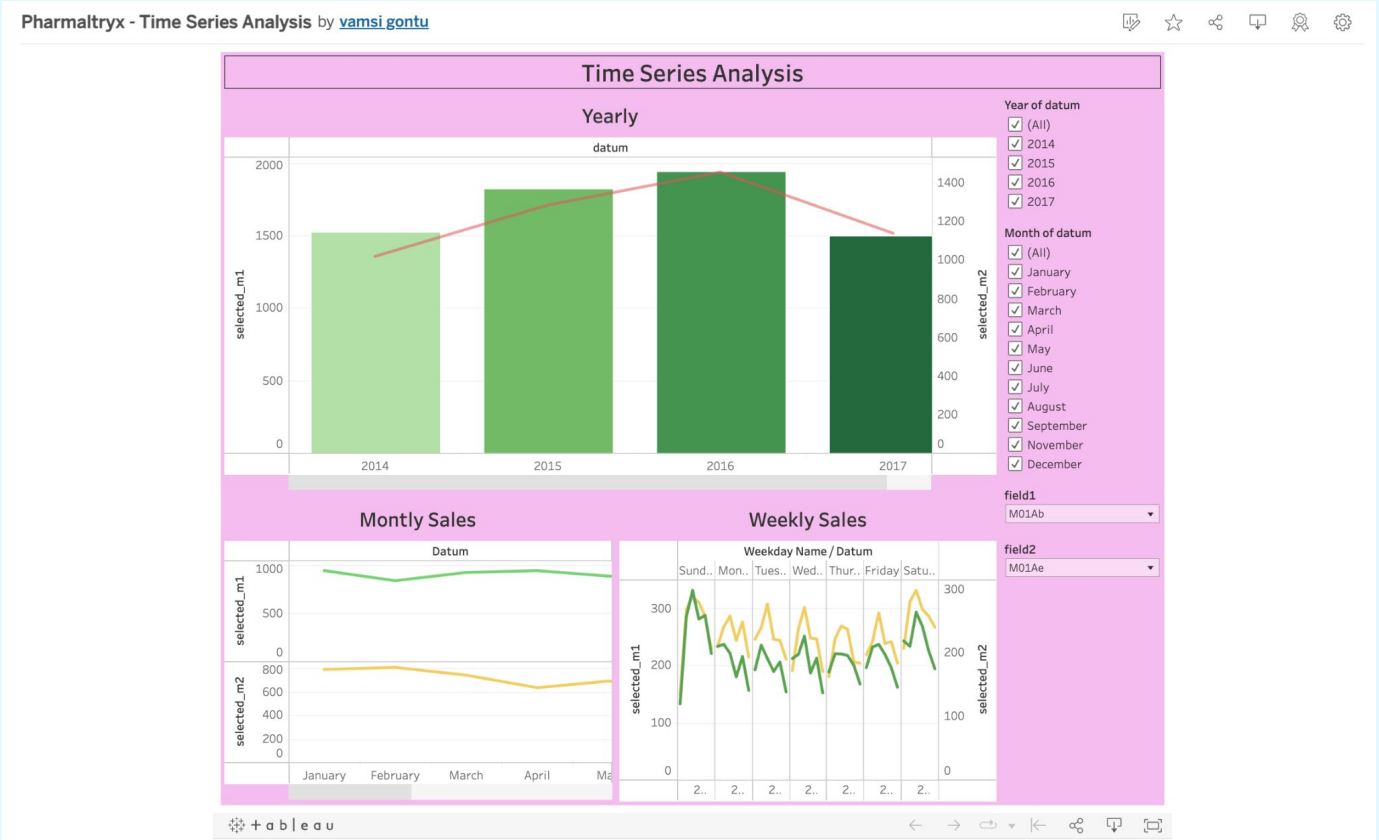
Actual vs Predicted Plot



Forecast Plot



Dashboard Result



Source: Link

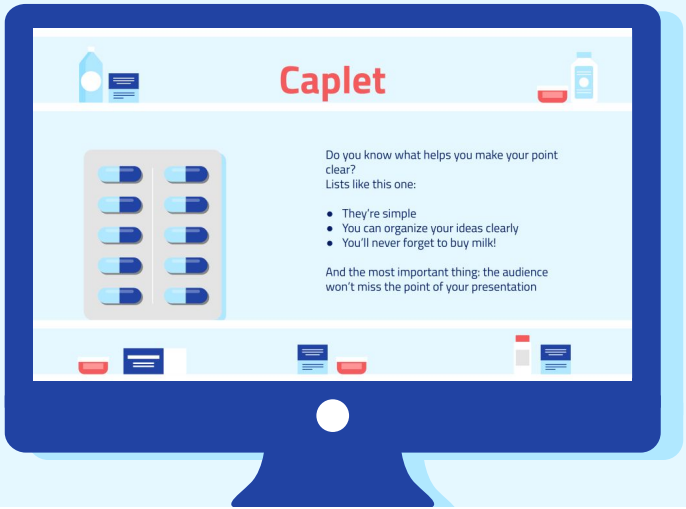
02

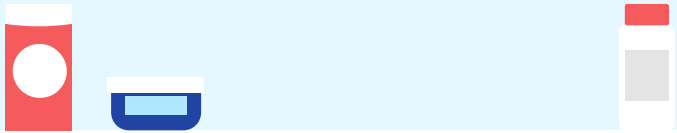
Sentiment Analysis of Drug

Data Source

we will us kaggle dataset

https://www.kaggle.com/datasets/thedevastator/drug-performance-evaluation

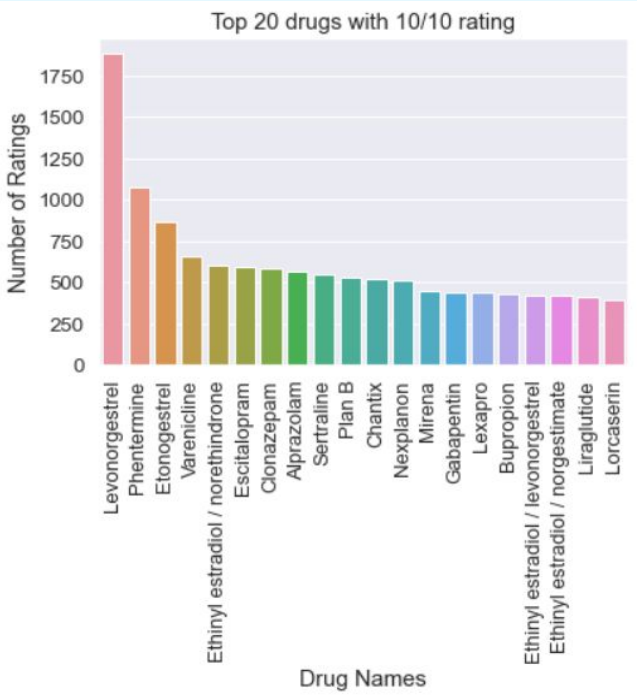


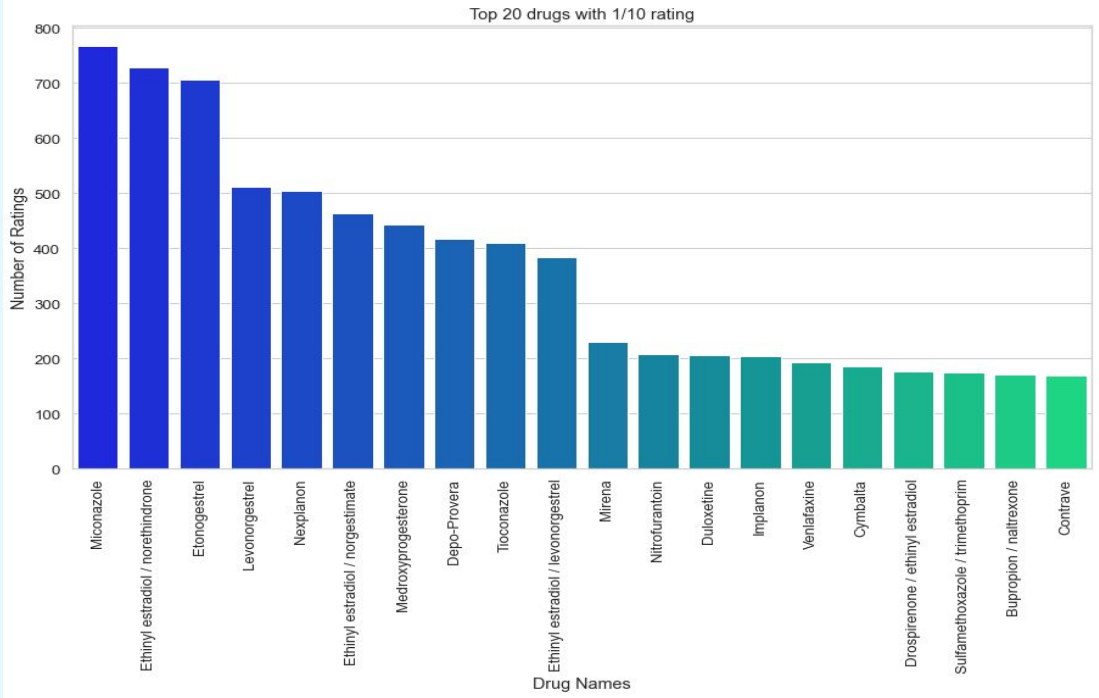


Sentiment Analysis

1. Drug sentiment analysis is important because it can help pharmaceutical companies and healthcare providers to understand how patients perceive drugs and their side effects.
2. By analyzing drug sentiment, companies can identify areas for improvement in their products, and healthcare providers can tailor their treatment plans to address patient concerns.
3. Additionally, drug sentiment analysis can aid in identifying potential safety issues and adverse events, which can help improve patient safety and outcomes.

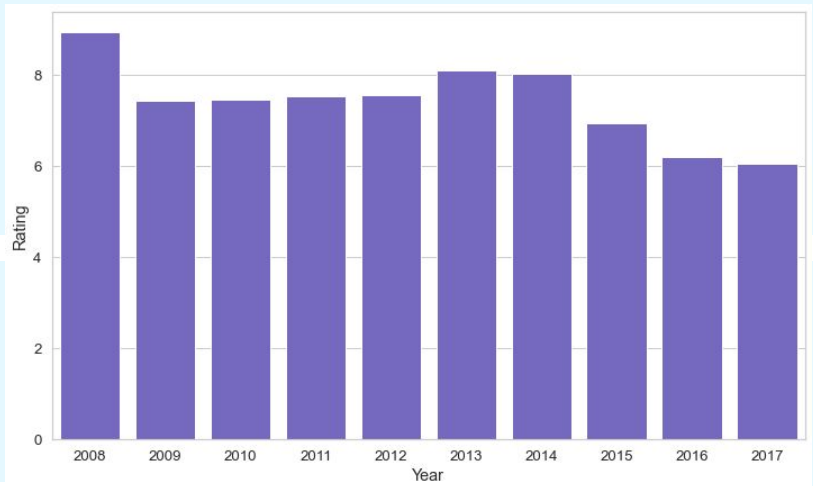
EDA

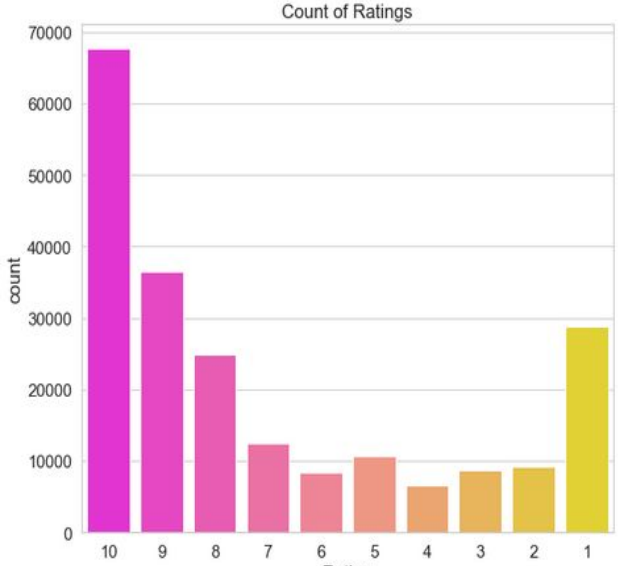

Top 20 drugs with 10/10 rating
Number of Ratings
1750
1500
1250
1000
750
500
250
0
Levonorgestrel
Phentermine
Etonogestrel
Varenicline
Ethinyl estradiol / norethindrone
Escitalopram
Clonazepam
Alprazolam
Sertraline
Plan B
Chantix
Nexplanon
Mirena
Gabapentin
Lexapro
Bupropion
Ethinyl estradiol / levonorgestrel
Ethinyl estradiol / norgestimate
Liraglutide
Lorcaserin
Drug Names

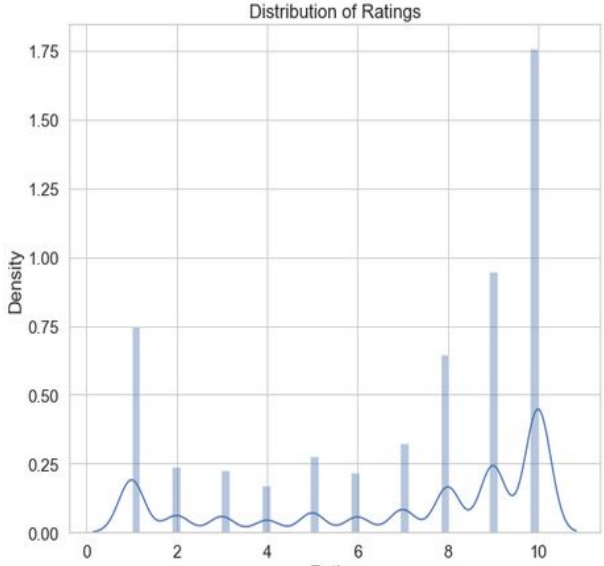

Top 20 drugs with 1/10 rating
Number of Ratings
800
700
600
500
400
300
200
100
0
Miconazole
Ethinyl estradiol / norethindrone
Etonogestrel
Levonorgestrel
Nexplanon
Ethinyl estradiol / norgestimate
Medroxyprogesterone
Depo-Provera
Tioconazole
Ethinyl estradiol / levonorgestrel
Mirena
Nitrofurantoin
Duloxetine
Implanon
Venlafaxine
Cymbalta
Drospirenone / ethinyl estradiol
Sulfamethoxazole / trimethoprim
Bupropion / naltrexone
Contrave
Drug Names

EDA

Distribution of ratings varying from 1 to 10

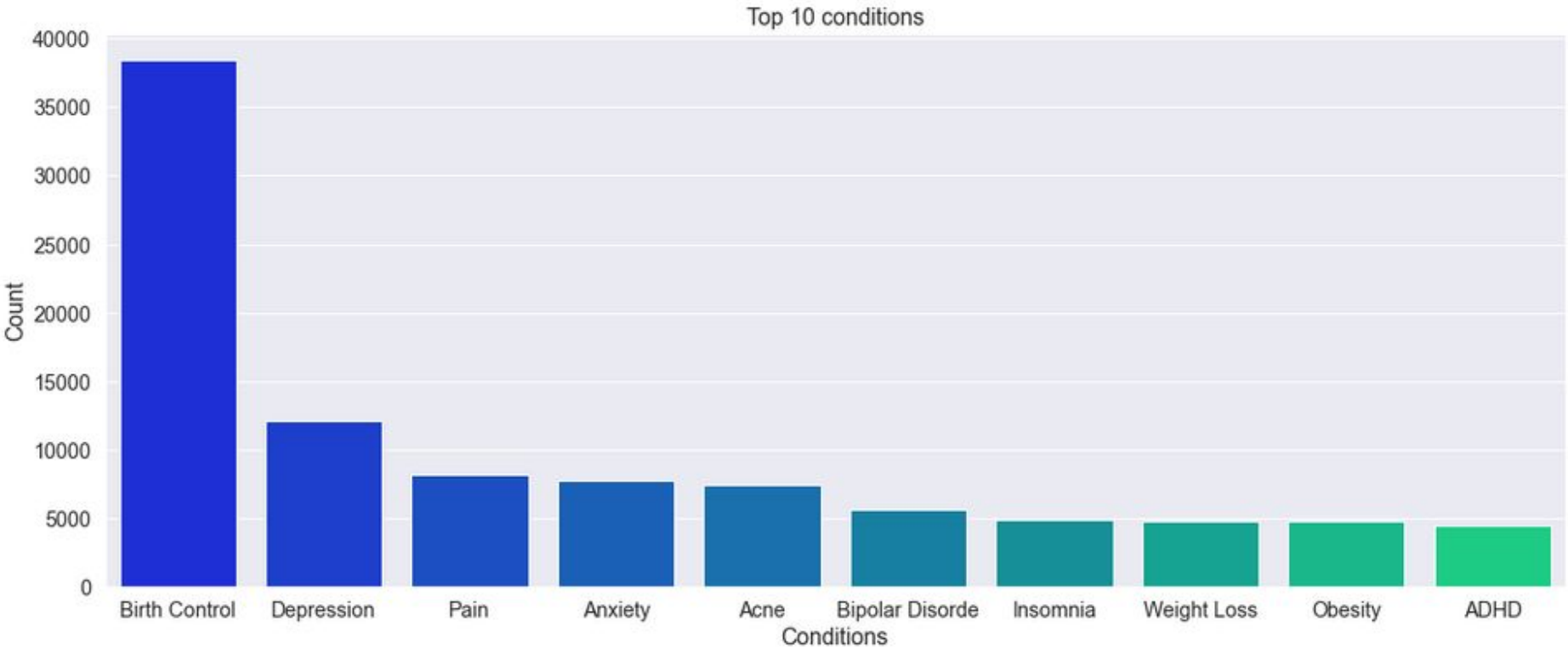




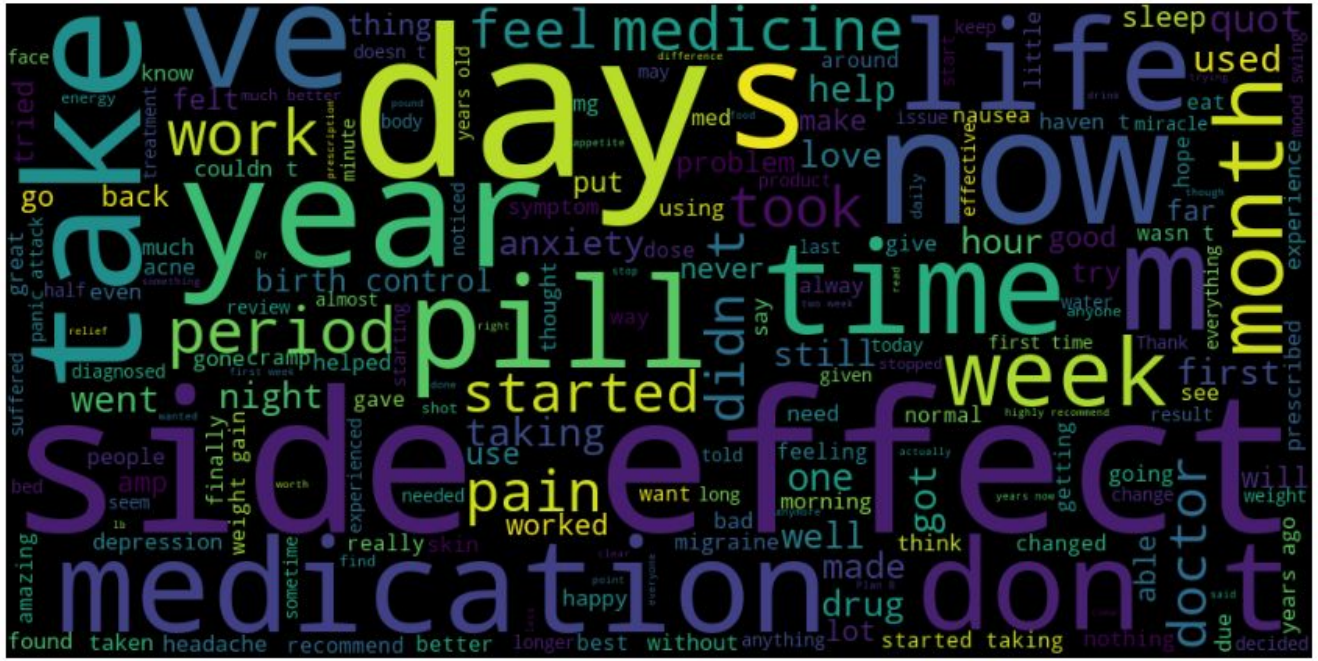


Total Count of Ratings Year-wise

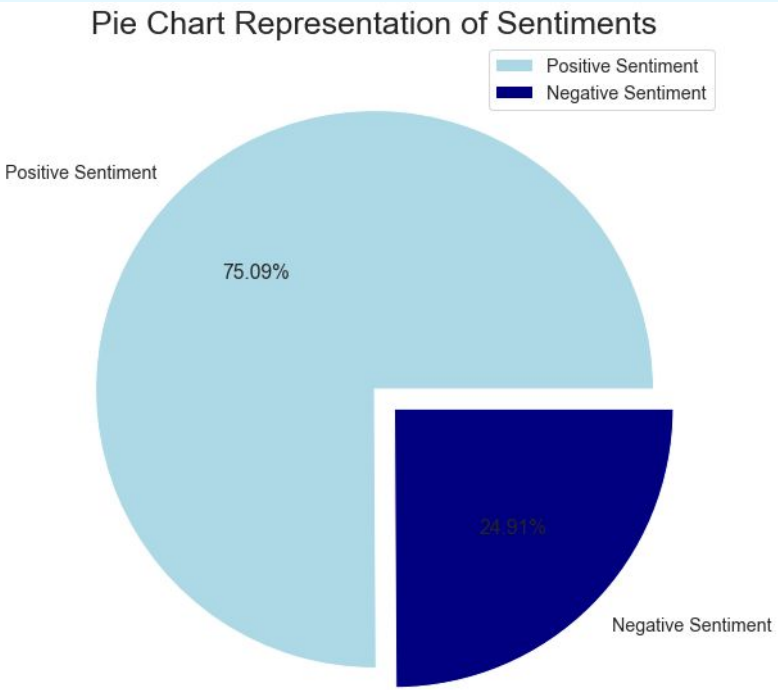
EDA - Top 10 Ailments Count

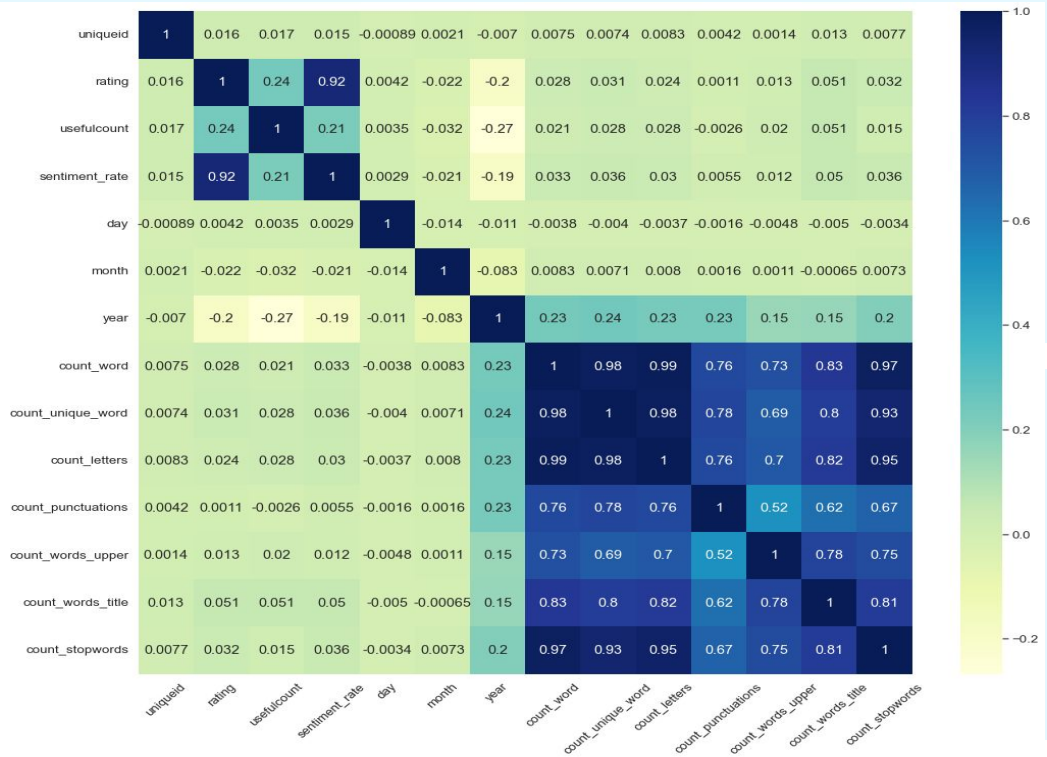


Wordcloud - Positive Reviews

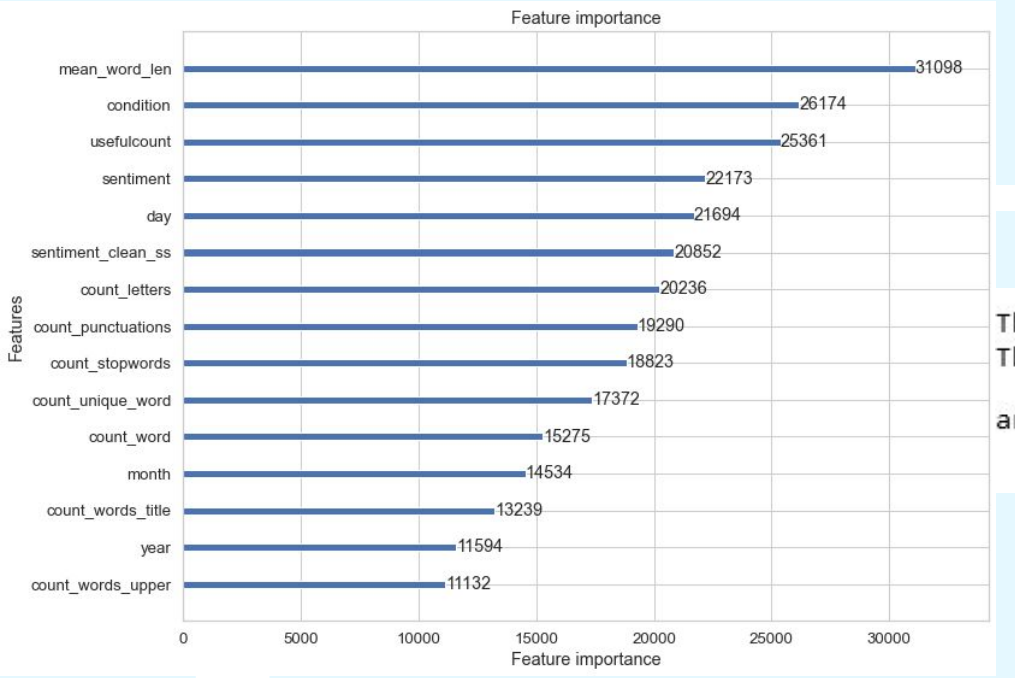


Feature Engineering



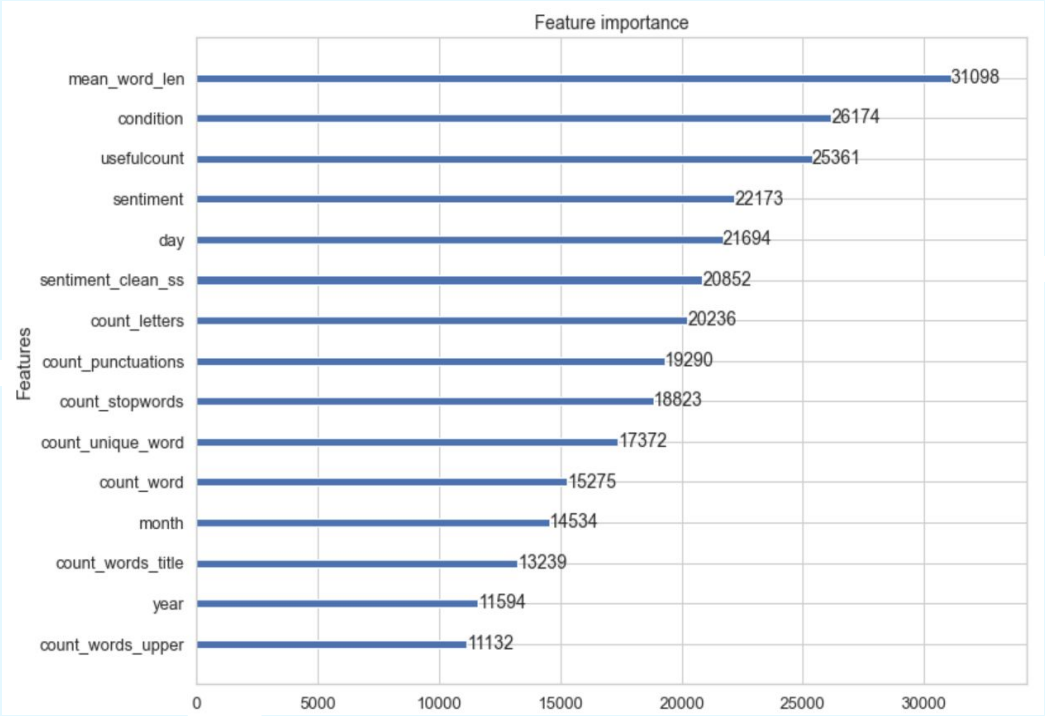


Modelling XG-Boost




```
The Accuracy of the model is :  0.7685042315425258
The confusion Matrix is

array([[ 8197, 11148],
       [ 3705, 41111]], dtype=int64)
```

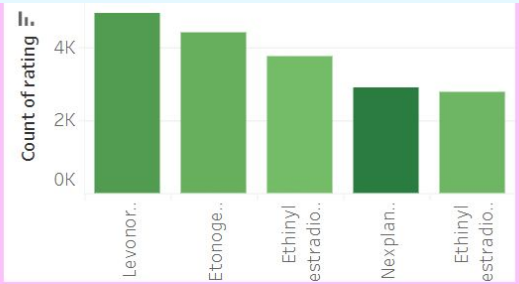
Modelling Light-GBM

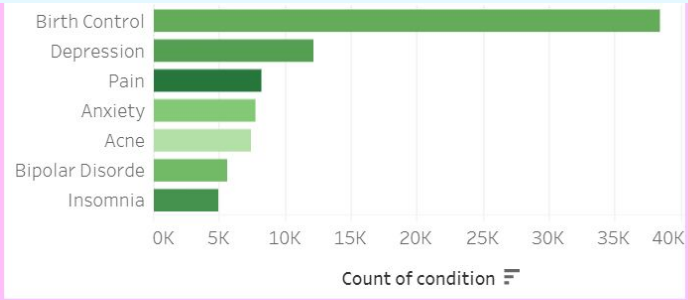



```
array([[14800,  4545],
       [ 2581, 42235]], dtype=int64)
```

Modelling CatBoost

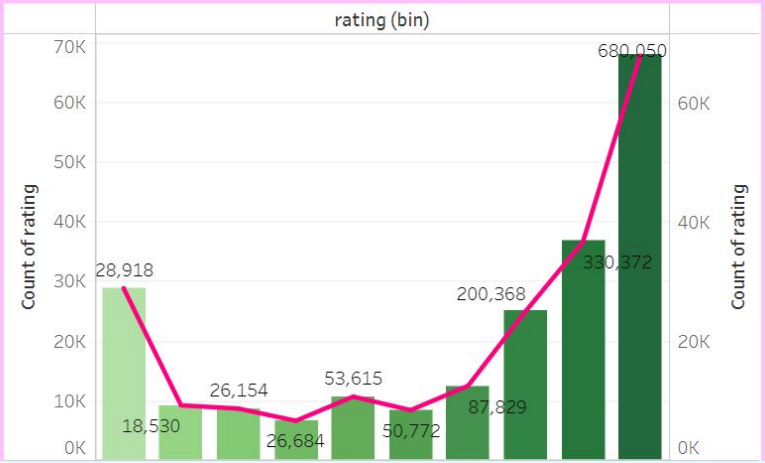
```
The Accuracy of the model is :  0.8843066660432349
The confusion Matrix is

array([[15007,  4338],
       [ 3085, 41731]], dtype=int64)
```





2008	6,785
2009	15,642
2010	11,227
2011	15,454
2012	13,382
2013	16,359
2014	16,104
2015	36,192
2016	46,607
2017	37,311

Count of Ratings



% Positive & Negative Reviews

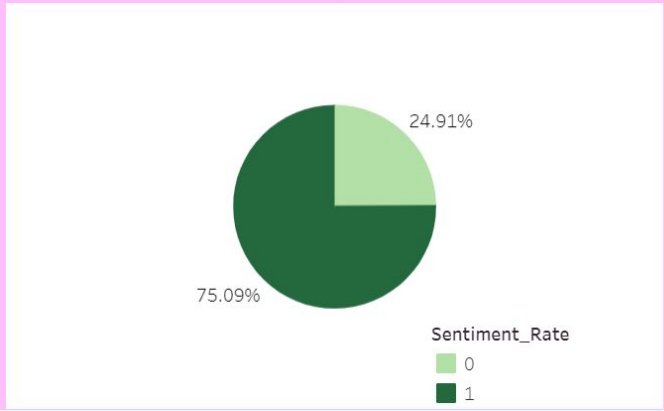
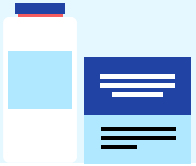
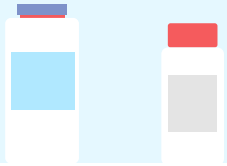


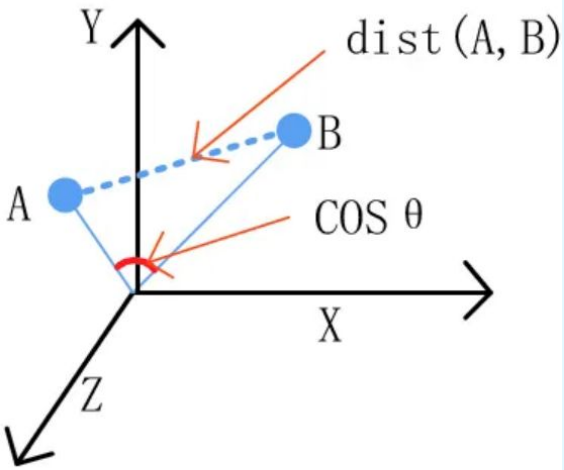
Tableau Dashboard published on https://public.tableau.com/views/drug-analysis-dashboard/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link



03 Recommendation Engine

Recommendation System



Difference with Euclidean metrics

A recommendation system is a type of information filtering system that suggests or recommends items to users based on their preferences, behaviors, or past interactions.

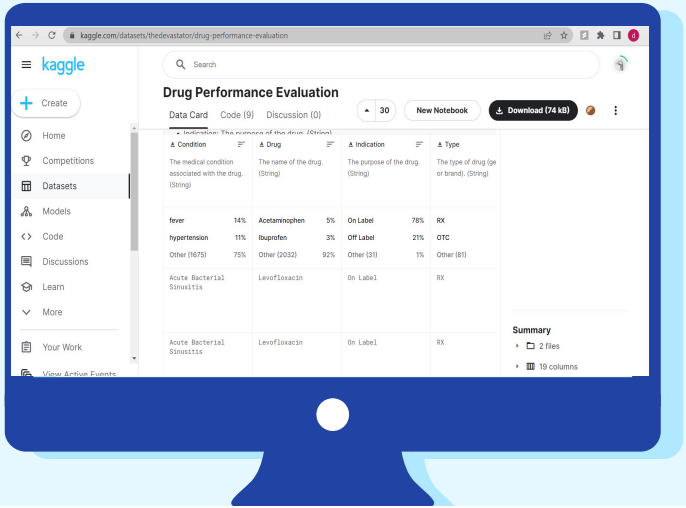
cosine similarity is used to compare the attributes or features of items to recommend similar items to users.

For our project we have used content based recommendation system

Data Source

we will us kaggle dataset collected using Kaggle API.

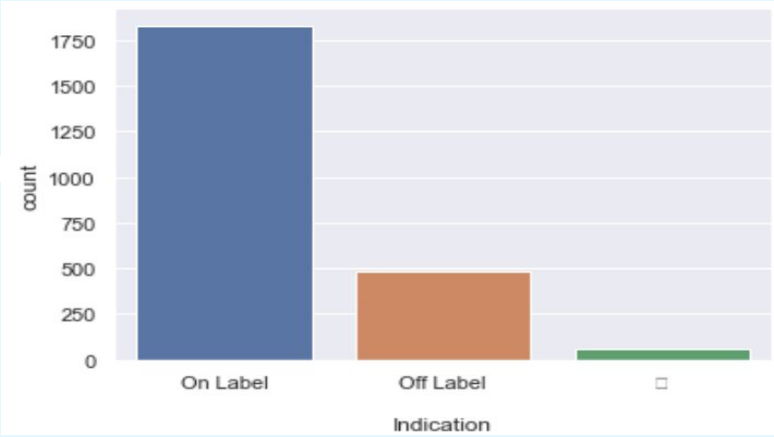
https://www.kaggle.com/datasets/thedevastator/drug-performance-evaluation

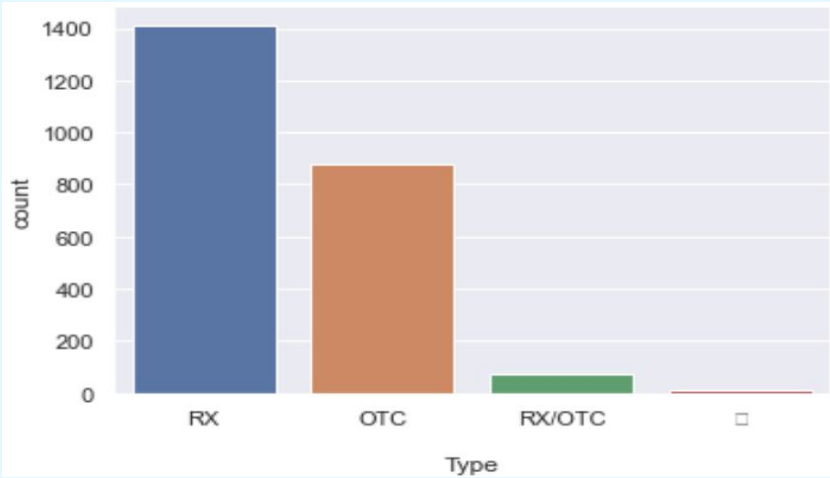


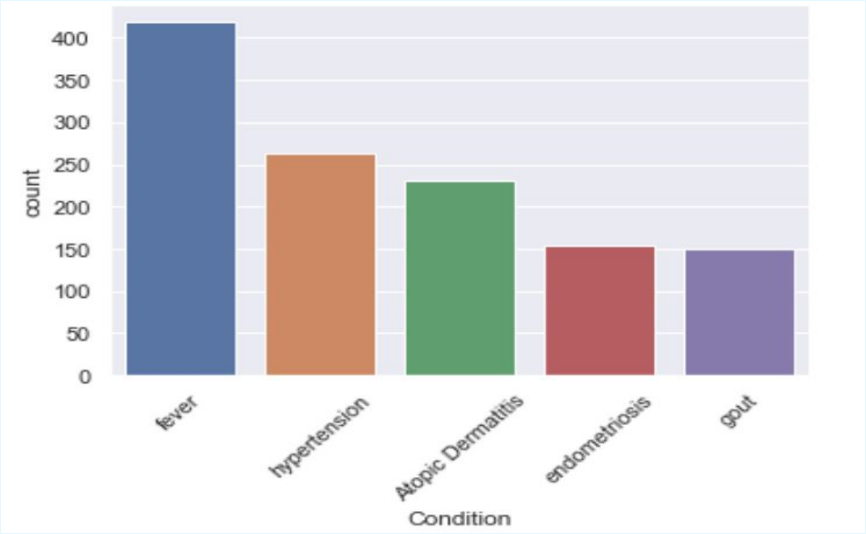
Exploratory Data Analysis

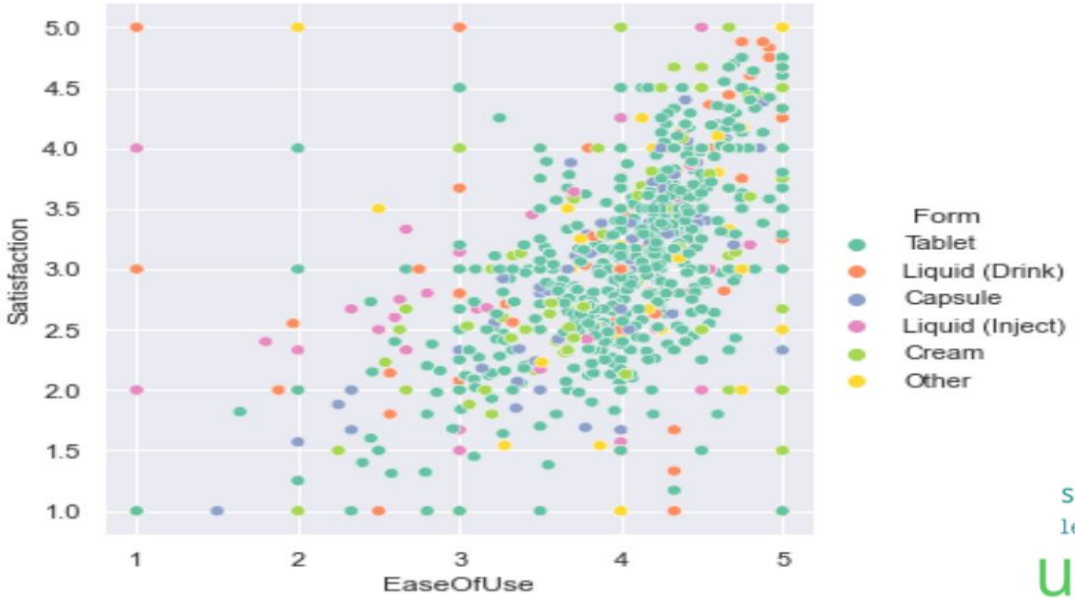
We can infer that most people buy medicine on prescription and that most information is on table.

Most occurring condition are hypertension and fever.



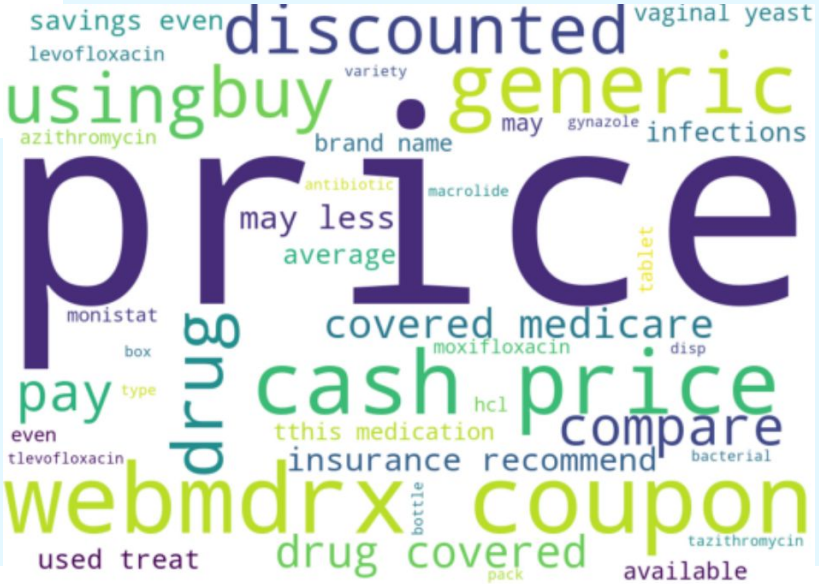




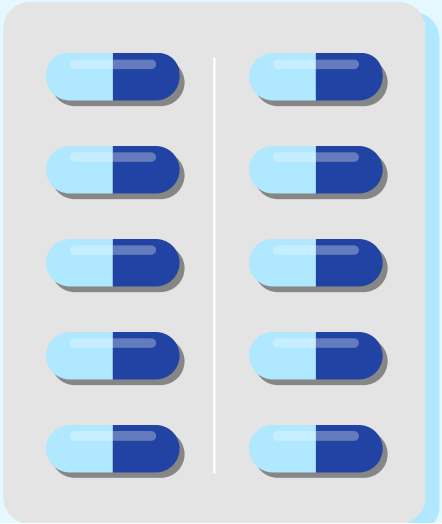


Most used medicine type was tablet and satisfaction is positively correlated to ease of use

The most occurring word on any drug was price



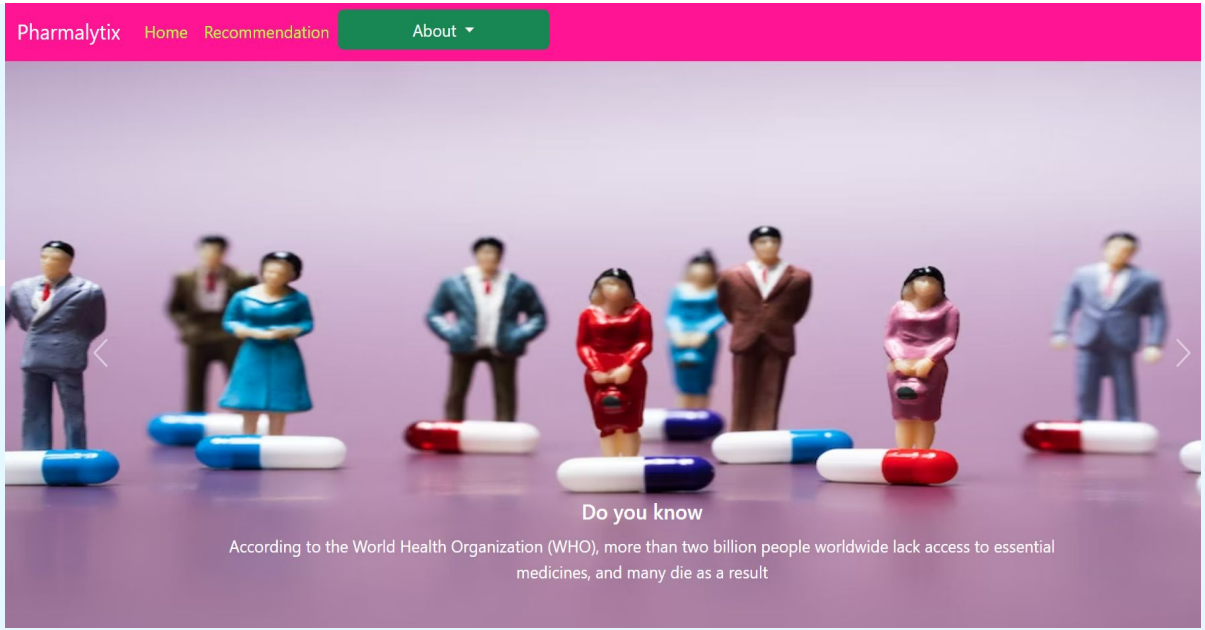
Design of Recommendation System



We constructed two recommendation system

1) Suggest drug based on condition for which effectiveness, ease of use, and patient satisfaction scores.
2) Same columns were used to recommend conditions based on drugs.

The principle behind it to calculate cosine distance between data points and 5 elements with least cosine distance were output



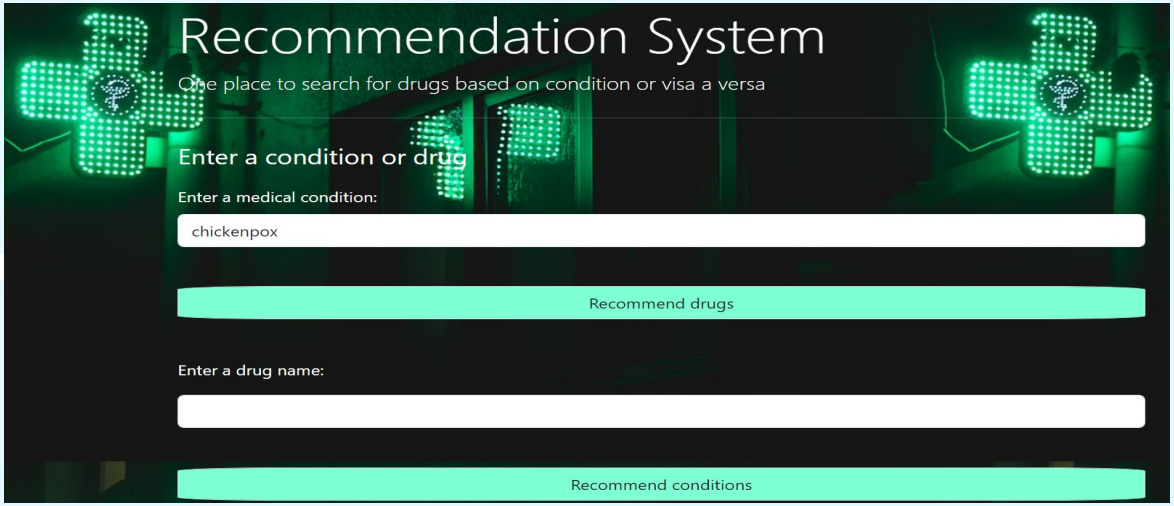
Website

-Pharmalytix

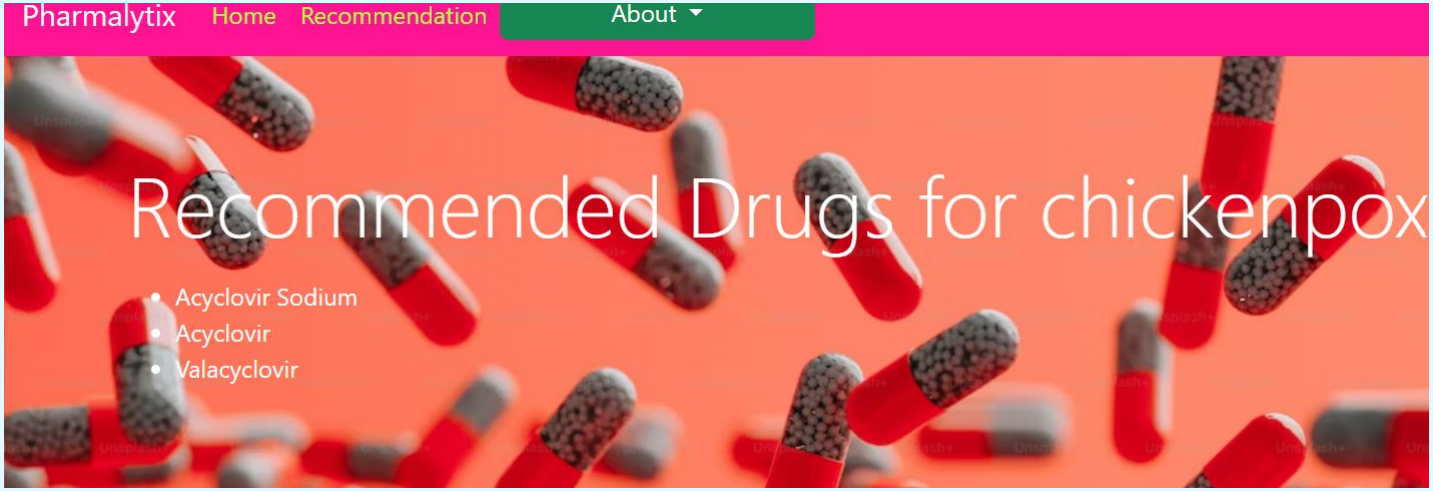
Using Flask we constructed our website

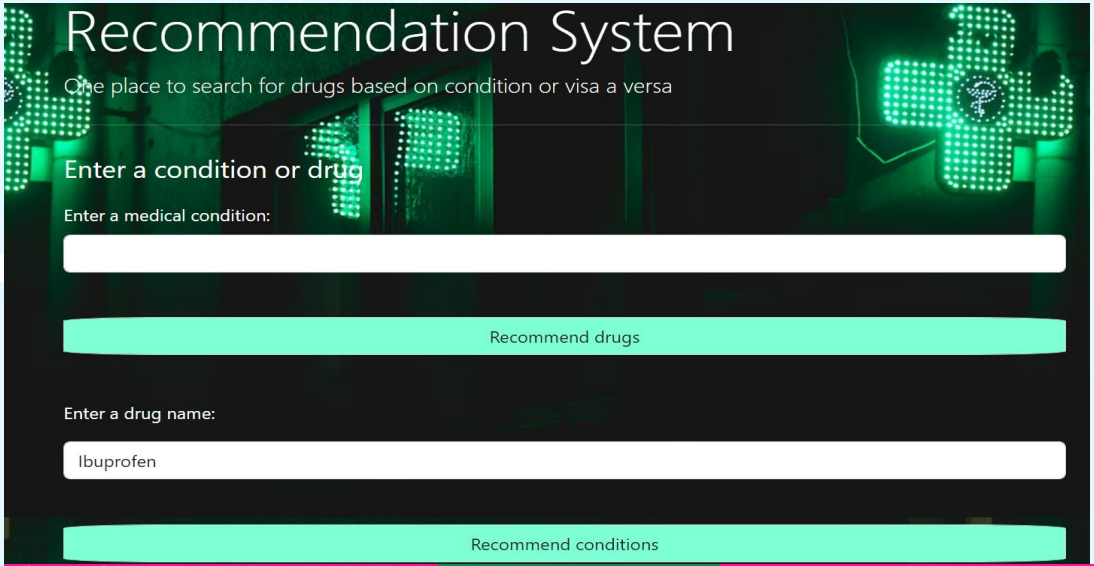
In App.py we imported dataset and defined methods for routing through the website which calls the various html pages .

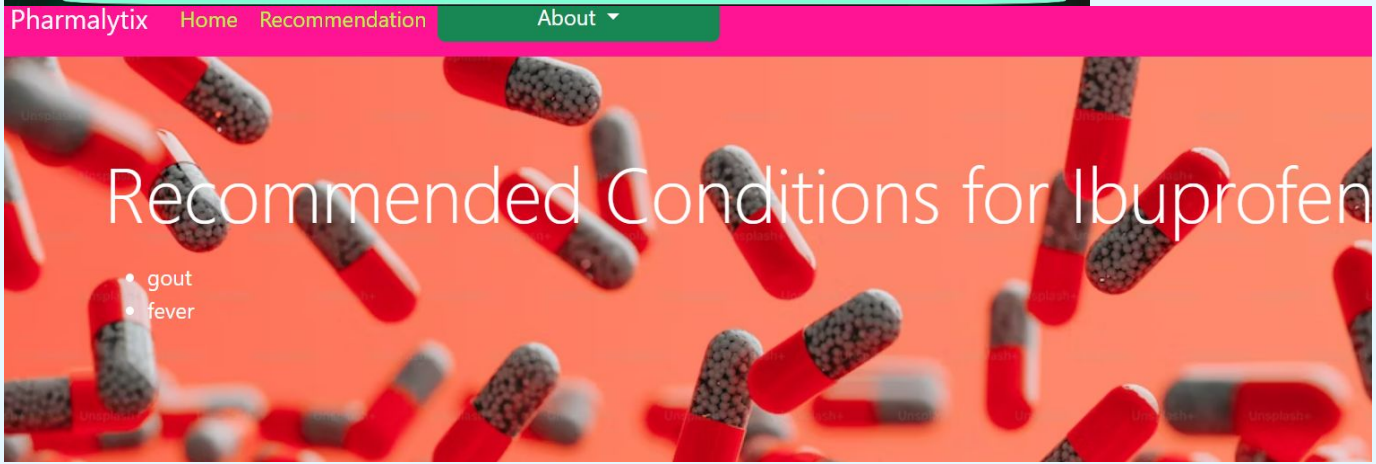
Condition and drug is variable for two methods . Results page uses binding to post the recommended values



When searching drug for chickenpox







Searching what Ibuprofen is used for

Summary of Findings

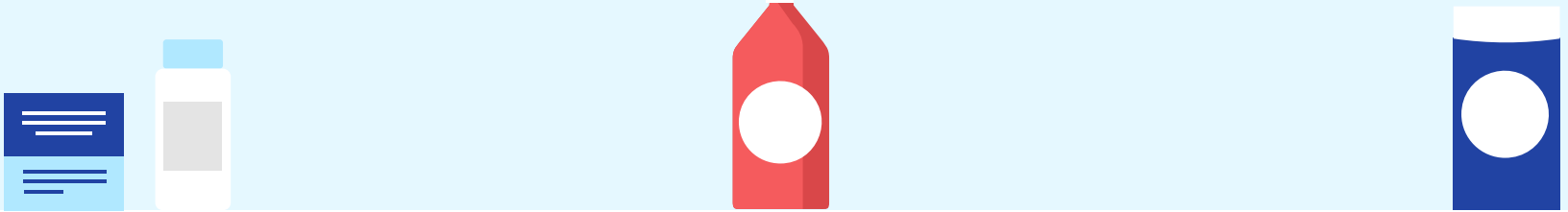
For **Time Series Analysis:**

- 70% of the drug sales are high during the year 2016-2017
- Saturdays are the best sale date for most of the drugs.(almost 75%)
- July, August months are having the highest sales of the drugs.

For **Sentiment Analysis:** Most of the reviews were positive. 3 models were used to categorize the sentiments as positive or negative, namely **XGBoost, CatBoost and LightBGM**, and the **best performing model is LightBGM** with an accuracy of **88%**. The feature engineering was performed and various new features like **word_count, avg_word count and polarity score** contributed towards increasing the accuracy of the model.

For **Recommendation Engine:**

The recommendation system was able to suggest top 5 drugs based on condition and top 5 conditions based on drug input. Website application was working well using flask.

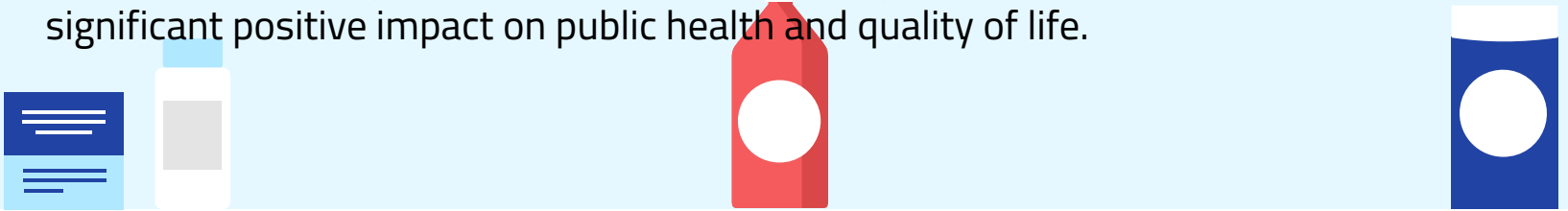


Where will your project impact the future?

- Sales forecasting aspect of your project can help pharmaceutical companies better understand market demand for their products, enabling them to optimize production and distribution, reduce waste, and improve profitability.

- Sentiment analysis of drugs can help pharmaceutical companies identify potential side effects or issues with new drugs before they are released to the market. This can help to improve drug safety and reduce the risk of adverse events for patients.

- Recommendation system can improve the accuracy and efficiency of drug prescriptions, potentially reducing medical errors and improving patient outcomes. This can lead to a significant positive impact on public health and quality of life.





04

Backstage

Some questions need to be answered :))

Which part is most interesting of your project?

Our website and dashboards

What is the most valuable thing you learned from conducting the project?

Implementing all the domains was itself a learning experience

Q1 Q2 Q3 Q4

Which part is most difficult of your project?

Data Collection, Integration, Sales forecasting

If you could start it over, what would you modify?

We would like to do data collection again and using develop frontend such that it pull data dynamically to show charts

Our Team

Architect of

Time Series Analysis

Specialize in data engineering, Time series forecasting domain

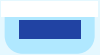
Sentiment Analysis

Former business analyst, expert in NLP

Recommendation Engine

Knows frontend, ML, NLP, Computer vision

Special Thanks!!!!!!!!!!!!!!



Thanks!!!!