Data Science Project | Report-1

Submitted by: Aditya Girdhar, Bhavya Narnoli, Hardik Singh, Dhruv Sood

Describe Data

Dataset URL: https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com

Dataset Sample:

	Unnamed: 0	drugName	condition	review	rating	date	usefulCount
0	163740	Mirtazapine	Depression	"I've tried a few antidepressants over th...	10.0	February 28, 2012	22
1	206473	Mesalamine	Crohn's Disease, Maintenance	"My son has Crohn's disease and has done ...	8.0	May 17, 2009	17
2	159672	Bactrim	Urinary Tract Infection	"Quick reduction of symptoms"	9.0	September 29, 2017	3
3	39293	Contrave	Weight Loss	"Contrave combines drugs that were used for al...	9.0	March 5, 2017	35

Attributes:

1. **Patient/review ID** *(int data)*
   This is a unique id assigned to each review.
2. **Drug Name** *(categorical)*
   This attribute states the name of the Drug whose review and effects are being described.
3. **Condition** *(categorical)*
   This attribute states the condition the patient is being treated for.
4. **Review** *(text)*
   This attribute contains the main review given by the patient that we are going to use to analyze the data.
5. **Rating** *(numerical)*
   This attribute gives a direct idea about the effectiveness of the drug and would be used to give an idea about the drug. Data report (Drug Review Dataset) 3
6. **Date** *(date)*
   This attribute gives the date of the review.
7. **Useful Count** *(numerical)*
   This states the number of people who have found the review helpful and gives an Idea about the correctness of the review.

Feature Grouping: We can create a group of the attributes Review, Rating and Useful Count. This group can be called the “User Feedback” group as all these attributes re ⊥ect the feedback of users who have used these drugs.

Existing Analysis

- Drugs-Recommendation-using-Reviews/DrugsAnalysis.ipynb at master https://github.com/sharmaroshan/Drugs-Recommendation-using-Reviews/blob/master/DrugsAnal ysis.ipynb
- Starter: UCI ML Drug Review dataset 08ecc634-6 | Kaggle https://www.kaggle.com/code/toiyeuem/starter-uci-ml-drug-review-dataset-08ecc634-6
- Exploratory Data Analysis and Sentiment Analysis of Drug Reviews https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462022000301191
- Drug-Review | Kaggle https://www.kaggle.com/datasets/mohamedabdelwahabali/drugreview
- SENTIMENT ANALYSIS OF DRUG REVIEWS USING WIT.AI https://www.irjmets.com/uploaded ıles/paper/volume_3/issue_12_december_2021/17663/ ınal/ın_ irjmets1639553975.pdf
- Classifying Drug Ratings Using User Reviews with Transformer-Based Language Models - PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9744636/
- Drug Review Dataset (druglib.com) https://archive.ics.uci.edu/dataset/461/drug+review+dataset+druglib+com
- DRAW — Drug Review Analysis Work https://medium.com/sfu-cspmp/draw-drug-review-analysis-work-96212ed98941
- Exploring Online Drug Reviews using Text Analytics, Sentiment Analysis, and Data Mining Models https://support.sas.com/resources/papers/proceedings20/4809-2020.pdf
- Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning - Abstract - Europe PMC https://europepmc.org/article/ppr/ppr343039

A detailed review of each of these links has been attached in the appendix.

Problem Statement

Preparing the dataset for analysis:

1. The reviews in the dataset contained unformatted HTML entity codes, we replaced them with their UTF equivalents to allow for semantically accurate language analysis.
2. Dropped rows with missing values. Only the conditions column had missing values.
3. The dataset was split into train and test, we merged the two datasets to ınd trends and perform analysis on the overall data.
4. There is a date feature which is not very useful on its own. We split the date feature into day, month and year to explore seasonal/periodic relationships.
5. Reviews (strings) are also not very useful on their own, hence we added a new sentiment polarity feature for each review.
6. We are also exploring the possibility of adding composite features by combining two or

more features.

For example: We can combine rating and usefulCount to generate weightedRating.

$$weightedRating^{(i)} = (usefulCount)^{(i)} \cdot (rating)^{(i)}$$

Another example: We can combine rating, usefulCount and reviewSentimentPolarity to generate a more relevant metric called reviewScore. An appropriate function can be determined after deeper analysis.

$$reviewScore^{(i)} = f(rating^{(i)}, reviewSentimentPolarity^{(i)}, usefulCount^{(i)})$$

Hypothesis we anticipate in data:

1. **Rating and Sentiment Polarity**: It is anticipated that there exists a positive correlation between the drug rating and the sentiment polarity of the reviews. In other words, reviews with more positive sentiment are likely to result in higher ratings, while negative sentiment may lead to lower ratings.
2. **Seasonal Variations in Reviews**: Depending on the condition and drug type, you might anticipate that reviews and ratings could vary seasonally. For instance, drugs for seasonal allergies might receive more reviews during speci ıc months.
3. **Textual correlation**: We can perform sentiment analysis on the review text to investigate whether certain words or phrases are associated with higher or lower ratings. For example, do reviews containing words like "side effects" tend to have lower ratings?
4. **Review Length and Useful Count**: It is hypothesized that there is a positive correlation between the length of drug reviews and the count of users ınding them useful (usefulCount). Longer reviews may provide more information, potentially aiding more users and resulting in higher useful counts.
5. **Number of Drug Reviews by Condition and Time**: We expect to ınd correlations between the number of drug reviews for a speci ıc medical condition and the temporal aspect, possibly identifying trends such as an increase in reviews over time for certain conditions due to factors like emerging treatments or disease outbreaks.
6. **Review Sentiment Polarity and Useful Count**: There is an expectation of a positive correlation between the sentiment polarity of reviews and the number of users ınding those reviews useful (usefulCount). Reviews with more polarized sentiments (very positive or very negative) might draw more attention and, therefore, accumulate higher useful counts.
7. **Temporal Trends in Ratings**: We can investigate whether there are temporal trends in drug ratings. Do newer drugs tend to have higher ratings than older ones? Are there certain months or years when drugs receive higher or lower ratings?

ML/DL Models

1. **User-Based Recommender**: Group users based on their review patterns and preferences. Recommend drugs to users who have reviewed drugs for similar conditions or have given similar ratings to drugs.
2. **Collaborative Filtering**: Collaborative ıltering techniques can be applied to recommend drugs to users based on their historical preferences and the preferences of similar users.
3. **Temporal Recommender**: Take into account the temporal aspect of the data by recommending drugs that are currently popular for certain conditions or have received recent positive reviews.

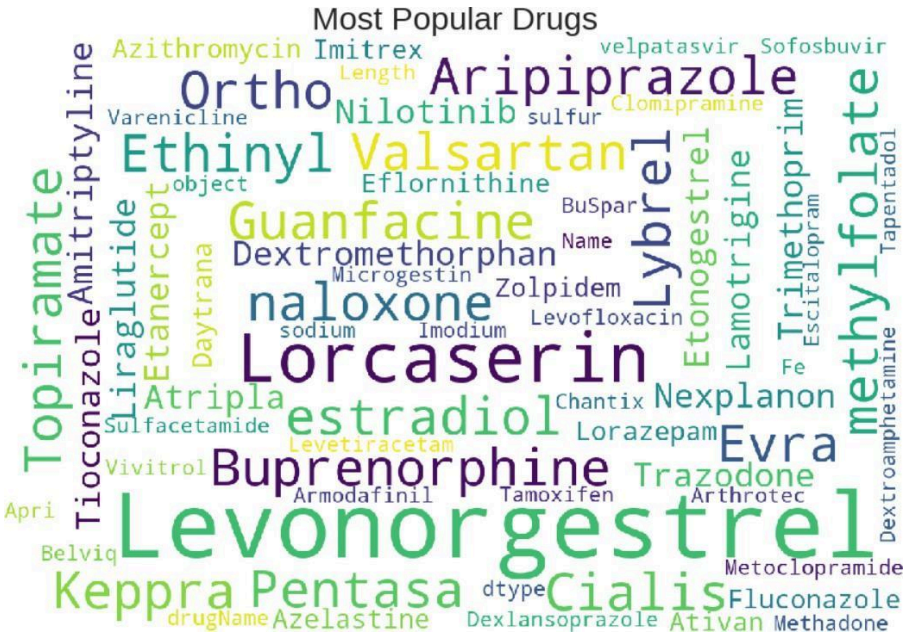
Appendix

Describe Data

2.0.1 Missing Values

Condition values are missing in the dataset for some rows, it's about the problem of the patient. 1194 rows have missing data rows, while analyzing reviews to not over ıt deleted those rows. [In the ırst link]

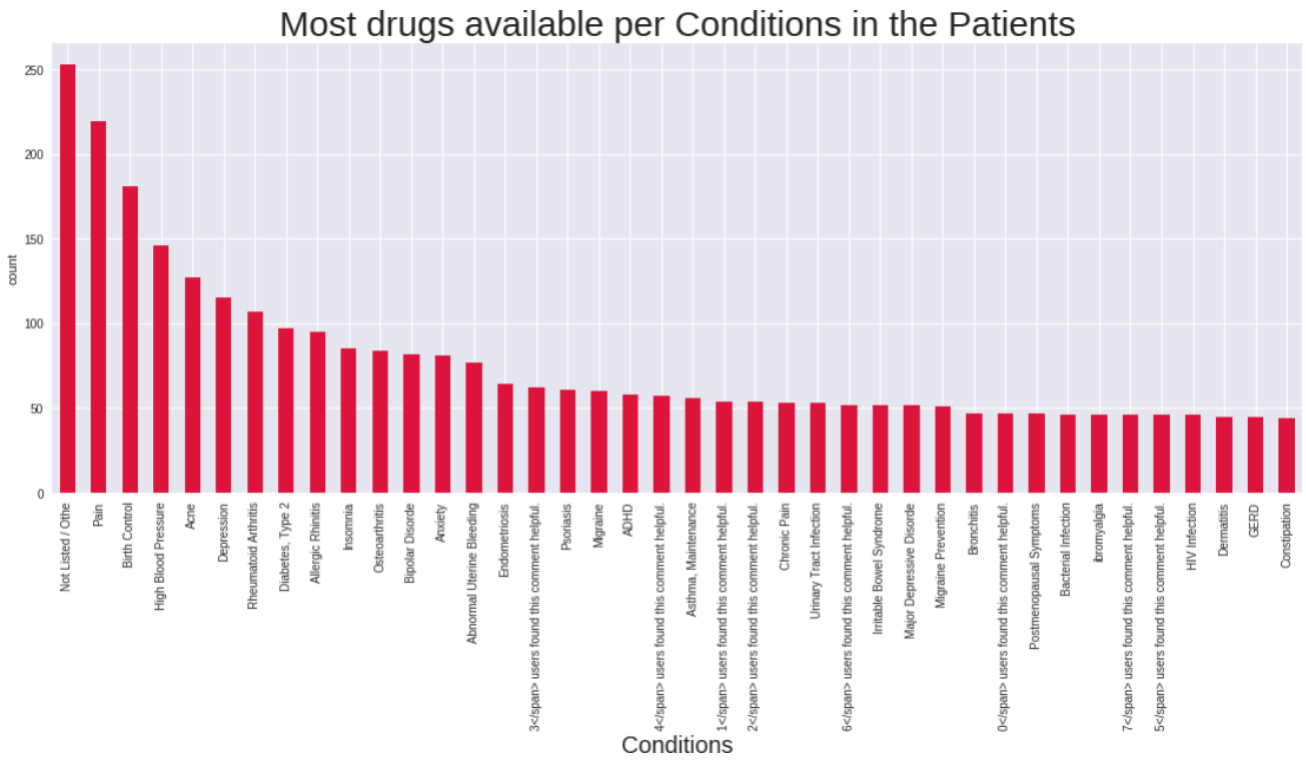
2.0.2 Most popular drugs



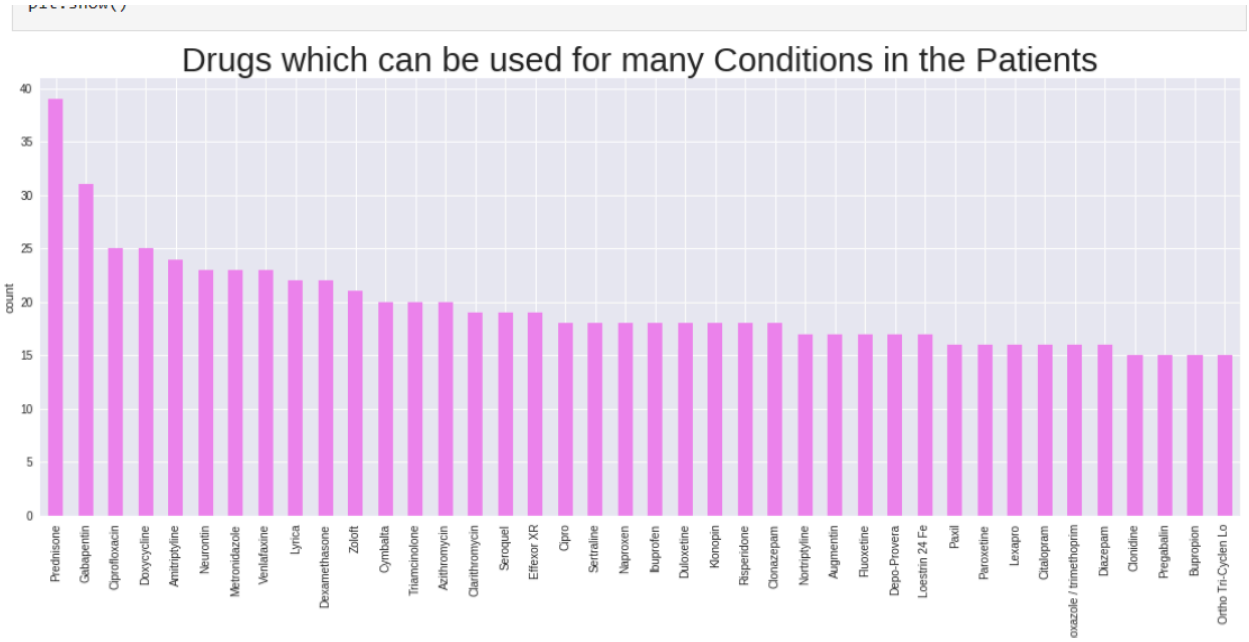
Levonorgestrel, Lorcaserin is the most used drug among all whereas sodium is relatively less

2.0.3 No. of drugs per condition

Plot made for no. of drugs available for each condition A analysis can be done for most used drug in each condition (the name and why it is the most used drug for that condition)

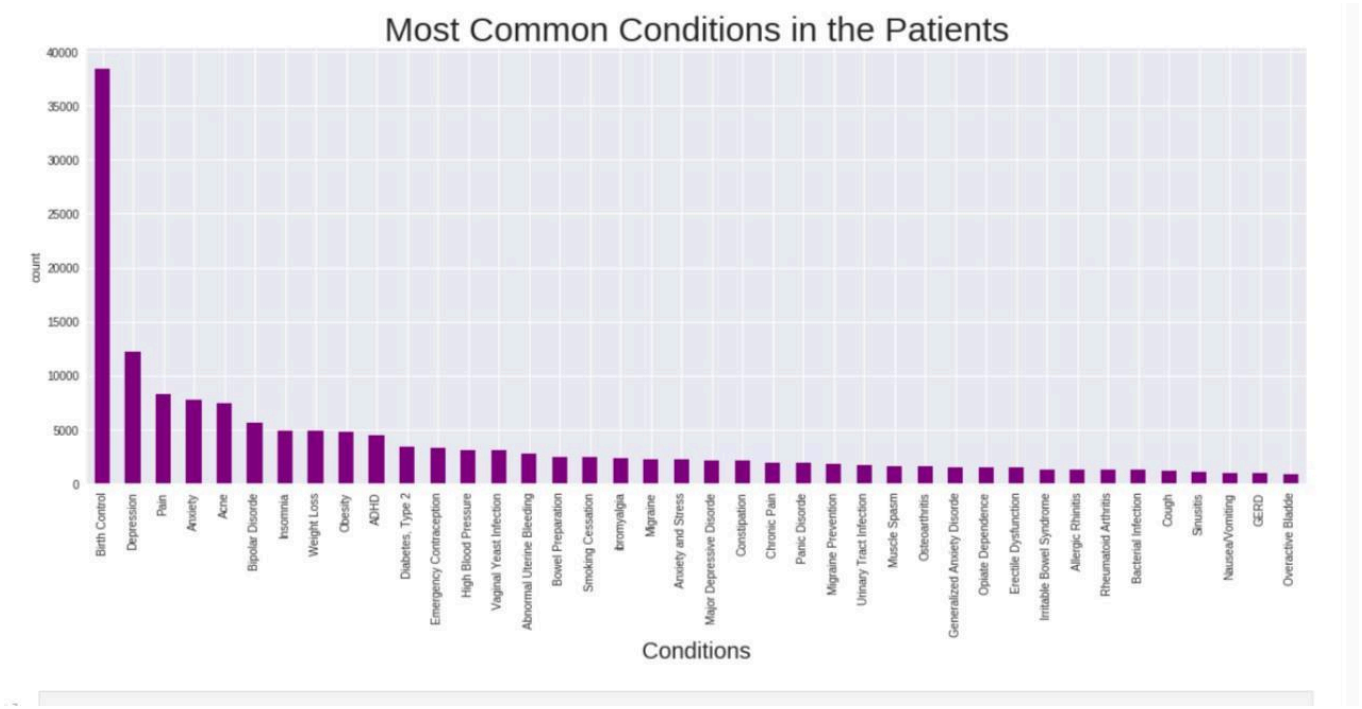


2.0.4 Most used drugs in most of the conditions



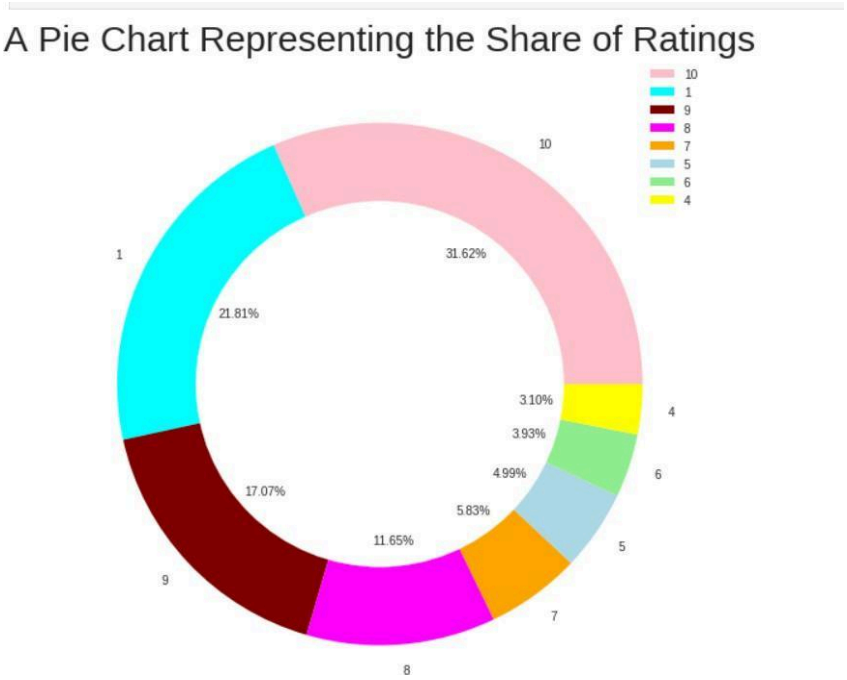
Checking which is the most used drug commonly for most conditions

2.0.5 Most common conditions in the patients

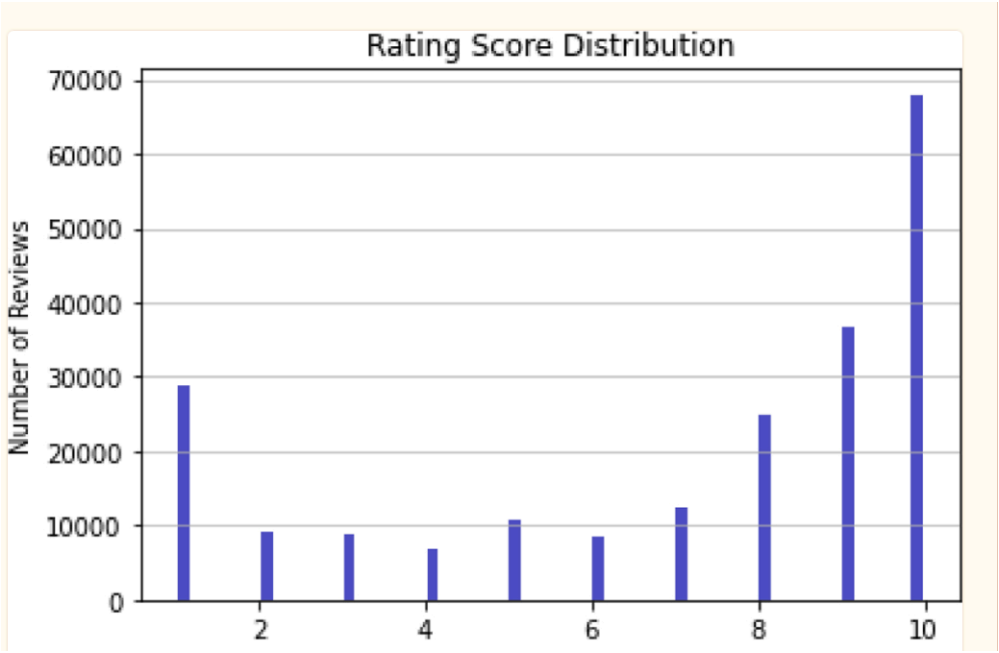


Graph for no. of people vs diseases in descending format

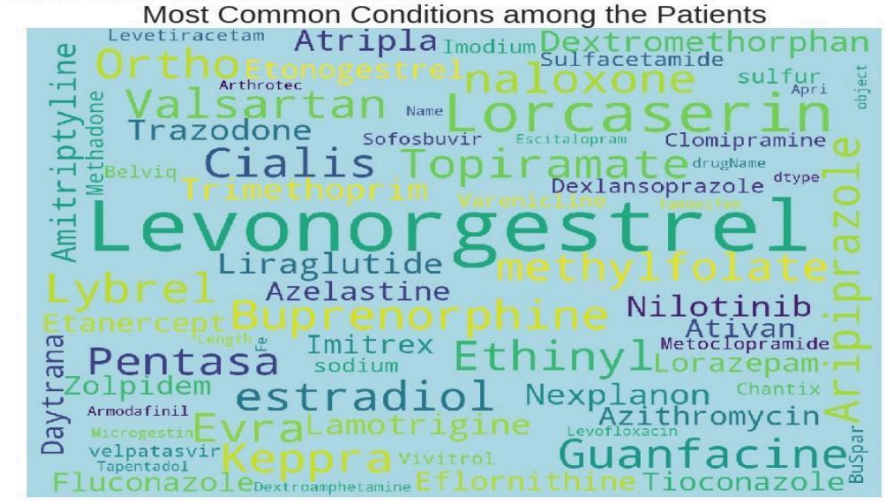
2.0.6 Most popular drugs



2.0.7 Rating Score Distribution

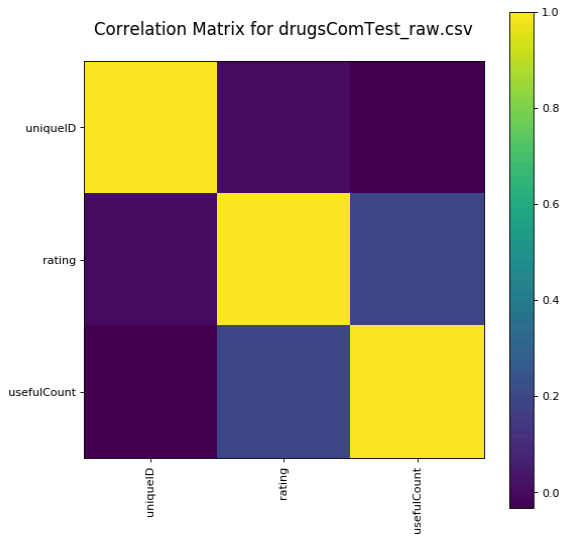


2.0.8 Most common Conditions in patients

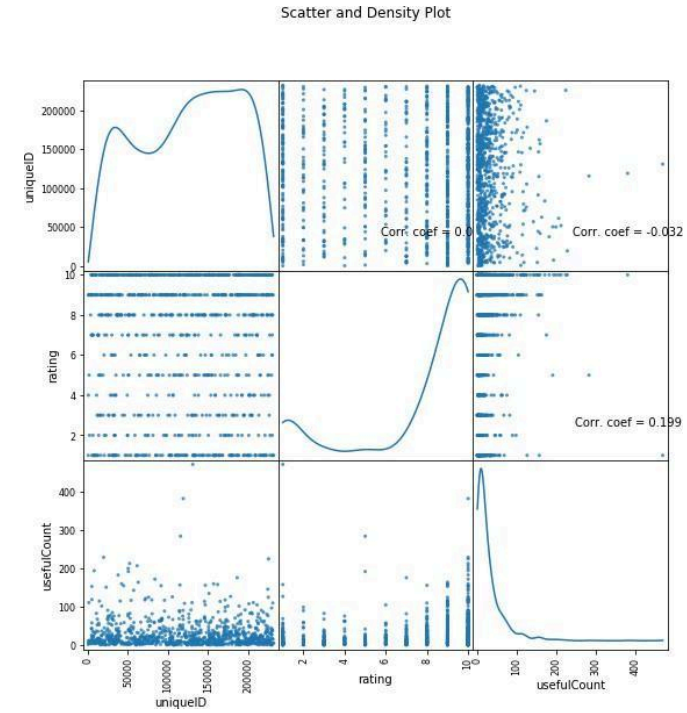


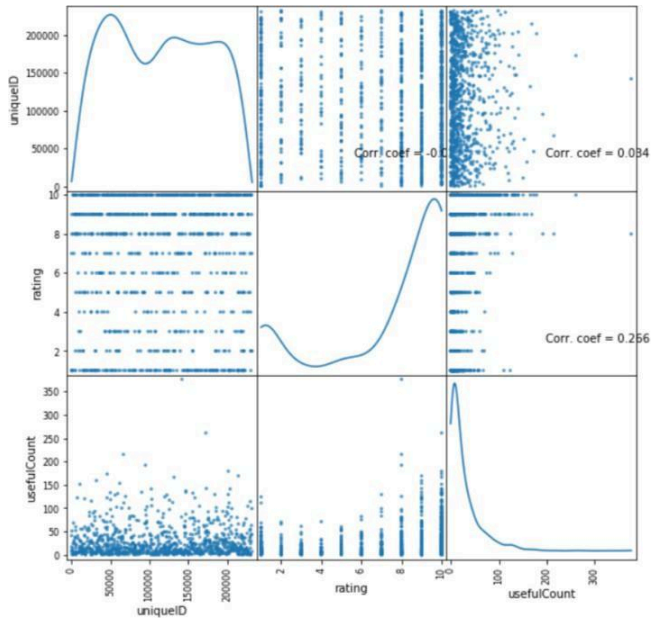
Correlation Matrix for the dataset

```
plotCorrelationMatrix(df1, 8)
```

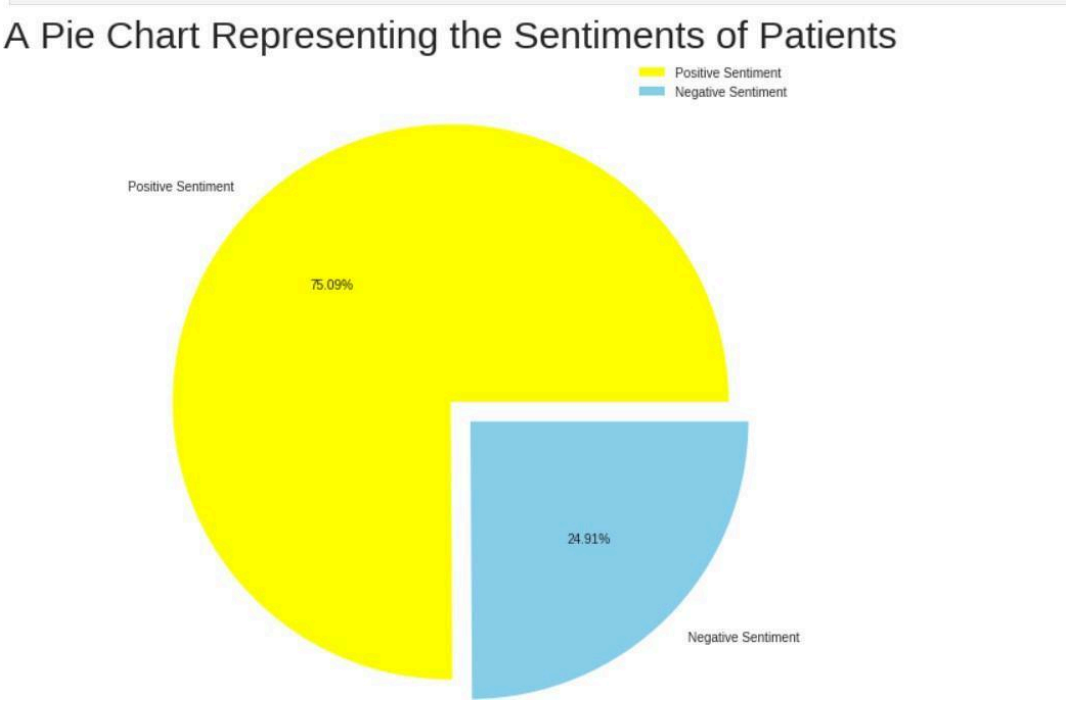


- The Distribution of Sentiments for each of the Reviews month-wise positive and negative



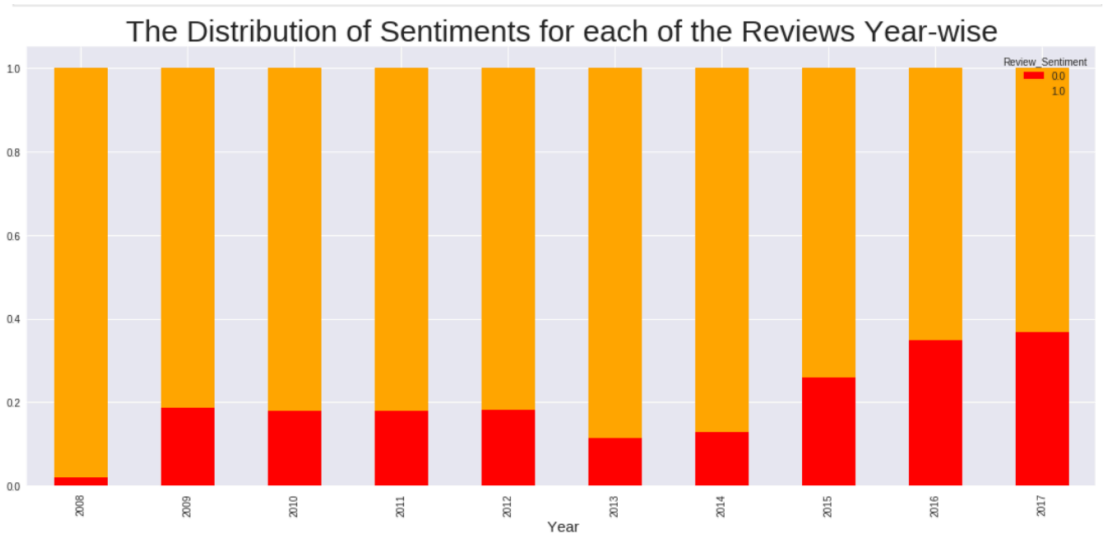


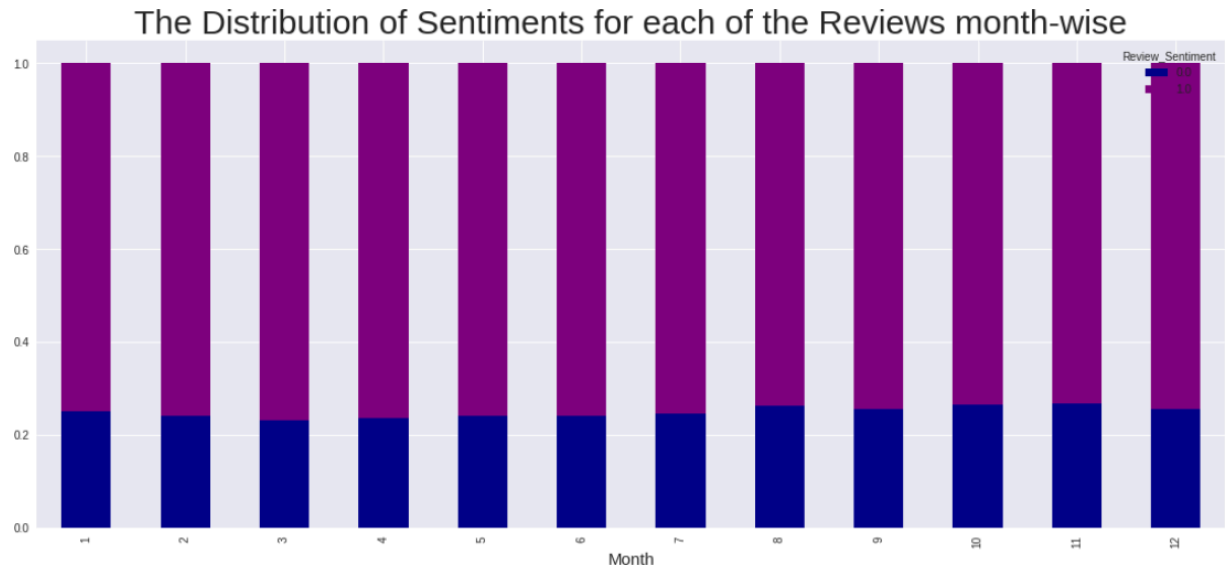
- Scatter And Density Plot




```
A Pie Chart Representing the Sentiments of Patients
```

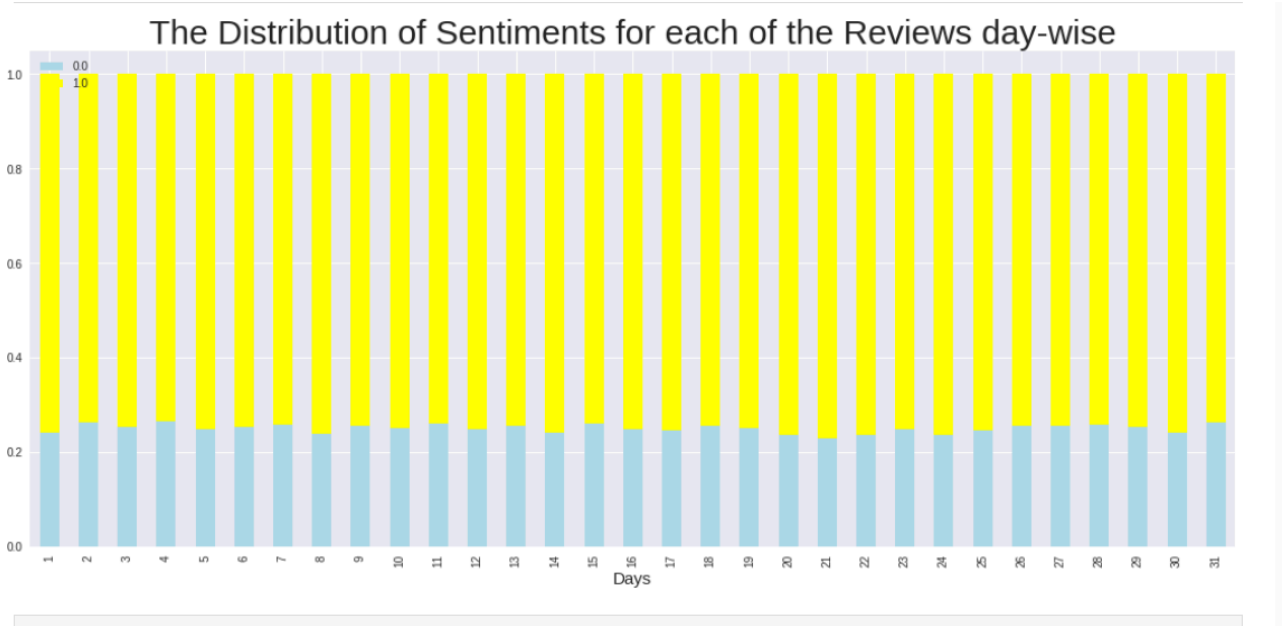
- The Distribution of Sentiments for each of the Reviews year-wise





- The Distribution of Sentiments for each of the Reviews month-wise

- The Distribution of Sentiments for each of the Reviews day-wise



- Sentime Score prediction of each drug

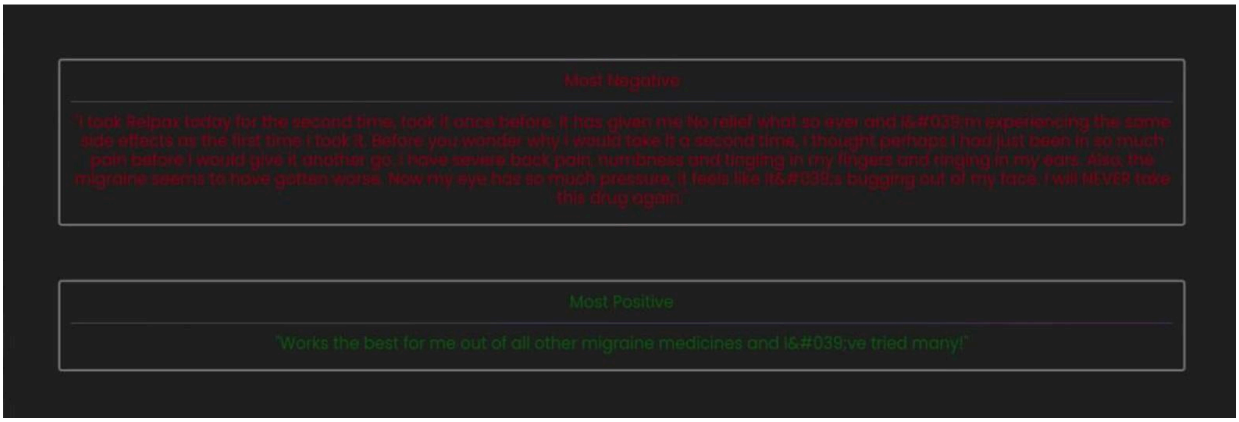
100%|██████████| 159332/159332 [01:55<00:00, 1377.71it/s]

Out[143…

gName	condition	review	rating	date	usefulCount	Review_Sentiment	Year	month	day	review_clean	Predict_Sentiment
alsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	2012-05-20	27	1.0	2012	5	20	no side effect take combin bystol mg fish oil	0.000000
nfacine	ADHD	"My son is halfway through his fourth week of ...	8	2010-04-27	192	1.0	2010	4	27	son halfway fourth week intuniv becam concern ...	0.114583
Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	2009-12-14	17	1.0	2009	12	14	use take anoth oral contracept pill cycl happi...	0.105000
ho Evra	Birth Control	"This is my first time using any form of birth...	8	2015-11-03	10	1.0	2015	11	3	first time use form birth control glad went pa...	0.300000
orphine aloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	2016-11-27	37	1.0	2016	11	27	suboxon complet turn life around feel healthie...	0.147037

It also displays most positive and most negative review

Figure 7: Result



2.0.8 The two papers linked have mentioned about

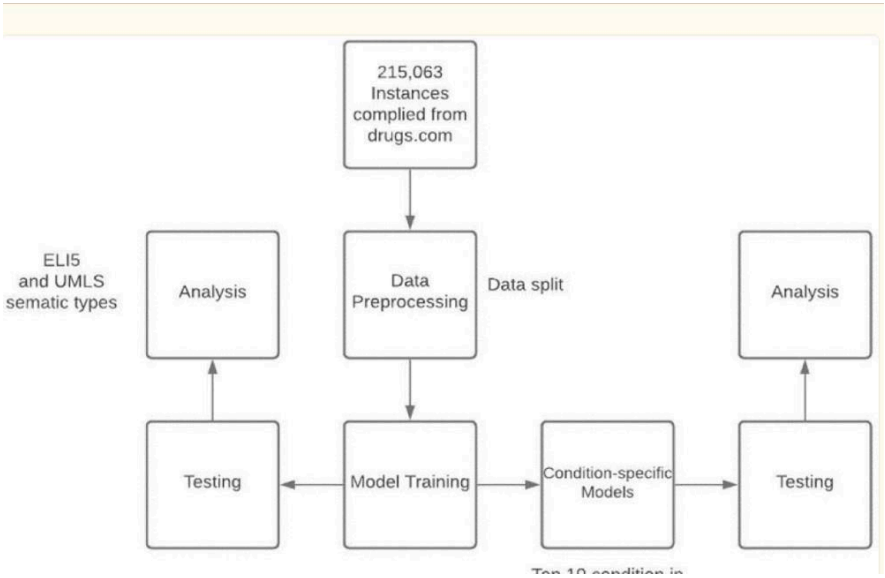
- This paper talks about building a ML/DL model to predict the type of drug in positive/negative/neutral • Each drug review was rated on a scale of 0 to 9, with 0 representing the least satis ıed patients and 9 representing the most satis ıed patients.
- Based on the review's rating, the dataset was categorized into three classes, namely, negative (rating less than 4), neutral (rating greater than 4 and less than 7), and positive (rating greater than 7).
- Finally made a model to calculate Rp which denotes the polarity score range of given drug.
- Made a model to categorize them in either of three

2.0.9 Research Paper 3 [link]

This research paper demonstrates that transformer-based classi ıcation models can be used to classify drug reviews and identify reviews that are inconsistent with the ratings.link

- Graph for no of reviews vs rating of the drug no.
- Users tend to reduce the effort required in reporting values by rating all qualities as highly important, thus resulting in overly positive ratings [16].
- This could lead to an unintended positive view of the overrated drugs by the general public, albeit less effective for certain population subgroups.
- Prior research has found that web-based reviews have the potential to be viewed as an applicable source of information for analysis, but review scoring biases may exist.
- For example, addictive drugs have been observed to be typically numerically highly rated in comparison to other drugs which have treated the same condition, even if these additive drugs underperformed based on experience [17].
- Prediction of user review score directly from user input may provide a method to limit this issue [18].
- The numeric ratings had a mean of 7.00 with a standard deviation of 3.27. There are 836 classi ıed medical conditions in the dataset.
- Ratings 8 or above were considered above average and below 8 were considered below average

• The graph shows top ıve most popular drugs on the basis of their usefulness count where we found



Sertraline, Escitalopram, Citalopram, Bupropion, Venlafaxine are top ıve most popular drugs

Overall condition validation from the test dataset for the minimized loss for the top-performing models.

Model	Above Average F1	Below Average F1	Accuracy
BERT	0.84	0.84	0.84
RoBERTa	0.83	0.83	0.84
XLNet	0.84	0.84	0.84
Bio_ClinicalBERT	**0.87**	**0.87**	**0.87**
ELECTRA	0.85	0.87	0.86
ALBERT	0.75	0.81	0.78
Random Forest (BOW)	0.77	0.45	0.68
Naïve Bayes (BOW)	0.76	0.03	0.61

• This research outlined a potential process to identify consumer drug review bias

• Users may rate all qualities as highly important, leading to overtly positive ratings, potentially resulting in a negative view of overrated drugs.

• Classify user ratings based on their textual review using machine learning and natural language processing.

• It explores the use of advanced natural language processing and machine learning techniques, particularly transformer-based models, to classify drug ratings based on user-generated textual reviews.

- It aims to overcome challenges related to biases in numerical ratings and provide a more nuanced and informative approach to evaluating drugs based on user experiences.
- the use of interpretation tools and semantic analysis adds depth to the research's ındings and insights.

The ratings of the reviews pertaining to the Birth Control, Depression, Pain drugs were classi ıed with high accuracy than the ratings of the drugs for other conditions. The conditions with lower instances had lower accuracy than the conditions with higher instances. However, there are many notable deviations present such as the pain and obesity models' lower accuracy or the higher accuracy for the ADHD model.