Analysis of Drug review data using data analytics methods.

CIND 820 Big Data Analytics Final project

Student name: Vinay Gorrepati
Student ID: 500728854
OEN: 378147490
Supervisor: Dr. Ceni Babaoglu
Date: 2022-Feb -14

Ryerson University

Table of Contents

Introduction

Patient feedback on various pharmaceutical drugs in the form of ratings and comments can be utilized for clinical research. Patient drug review data can be obtained by web scraping or web crawling various pharmaceutical online websites. Such reviews contain a large amount of user sentiment related to a particular condition, which may be useful in the detection of side effects and understanding the efficacy of drugs. The analysis of this subjective data using natural language processing and sentiment analysis can provide powerful insights which in turn could support further research to improve the patient experience.

The UCI ML Drug Review dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating system reflecting overall patient satisfaction. (Felix Gräßer S. K., 2018) The data was obtained by crawling online pharmaceutical review sites. This data was published in a study on sentiment analysis of drug experience over various facets. (Felix Gräßer S. K., 2018)

This project will attempt to answer some research questions such as the following.

- What are the Top drugs for a given condition?
- Can the Drug rating be predicted based on the reviews?
- Can we determine if a review is positive or neutral or negative?

Sentiment Analysis will be conducted using the Python programming language.Various analytic tools will be utilized to conduct data analysis. Python programming language has powerful libraries such as NumPy, Pandas, Matplotlib which are utilized for high level mathematical functions, data manipulation and analysis. In addition, Python has Scikit-learn which is a

Machine Learning library. Python will primarily be used for this project via Jupyter Notebook. Apart from this Microsoft Azure platform will also be used. A GitHub repository will be created, and reports created for this project will be uploaded there.

Literature Review

Online review websites contain a plethora of information especially customer feedback and reviews. Almost all types of products and services available across diverse domains are reviewed by their users. This information can be leveraged to obtain valuable insights using data mining approaches such as sentiment analysis. One sub area within this are Online user reviews in the pharmaceutical field.

Although sentiment analysis has been applied to many diverse domains, in the pharmaceutical it has only gained more attention only recently. Early work in sentiment analysis of drug reviews mainly used rules and sentiment lexicons (such as SentiWordNet to detect the overall polarity (positive or negative) of a given drug review. (Esuli, 2006) (Goeuriot, 2012, January) (Na, 2012 November) (Wiley, 2014)

(Felix Gräßer S. K., 2018) created a dataset with drug reviews data which collected from the Drugs.com website. This dataset provides information on drugs which can be useful for patients and health care practitioners. This dataset is the source data for this project. Various data is collected in each review such as a score from 0 to 9, indicating the patient's degree of satisfaction with the drug. The reviews have been grouped into three classes according to their ratings: positive, negative, and neutral. The authors used a logistic regression to classify the drug reviews, achieving an accuracy of 0.9224. (Felix Gräßer S. K., 2018) (Colón-Ruiz, 2020)

(Bobicev, 2012 May) used a Bag of Word approach to conduct analysis on twitter messages containing personal health information. They evaluated different machine learning algorithms such as Naive Bayes, Decision trees, KNN and SVM. In (Ali, 2013, October), various algorithms such as Naive Bayes, SVM or Logistic Regression were investigated to estimate the polarity (positive or negative) of patients' posts in online health forums.

(Wilson, 2005, October) trained algorithms using sentiment analysis features such as the number of subjective words, the number of adjectives, positive ,negative, neutral words from the Subjectivity Lexicon.

(Mishra, 2015, November) developed a system for detecting polarity of drug reviews using Support Vector Machine (SVM). The system also performed sentiment analysis on drug reviews in order to predict ratings for conditions such as satisfaction, effectiveness and ease of use of the drug. Drug reviews were tokenized and SentiWordNet was used to assign the sentiment scores for each token. (Colón-Ruiz, 2020) (Mishra, 2015, November)

A word embedding model is a method where a mapping is developed between words and vectors capable to capture similarity between words. (Bengio, 2000) conducted seminal work in word embeddings. More recently, word embeddings have been used widely in conducting sentiment analysis of patient's online reviews and posts. Various Machine learning algorithms such as SVM, Naive Bayes and Random Forest have been explored. (de-Albornoz, 2018) (Colón-Ruiz, 2020)

Hence, it can be inferred that drug reviews data has been widely analysed using various Machine learning algorithms and the particular dataset used for this project has been popular in recent times. (Kaggle, 2018)

Research focus and contribution

As discussed, the objective of this project is to come up with data analytics methods to answer three specific research questions. While a lot of similar analysis has been conducted already and diverse methodologies have been explored, this project will firstly merge the two datasets provided to create a new dataset. Secondly, the coding will exclusively be done using Python to conduct Sentiment Analysis and other analysis. Various machine learning algorithms will be studied, and the outcomes analyzed, therefore demonstrating a strong understanding of Data Science methods.

Dataset description

The dataset was published on the UCI Machine Learning repository.
There are 215063 instances with 6 different columns and 1 unique id column.
The dataset was further differentiated into Test and Train data. (Felix Gräßer S. K., 2018)
To reiterate, the dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating system reflecting overall patient satisfaction. (Felix Gräßer S. K., 2018)

Header information

The header information for the Train Data is as follows.

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37

Figure 1: Header information and top 5 data rows in train data

The header information for the Test Data is as follows.

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	163740	Mirtazapine	Depression	"I've tried a few antidepressants over th...	10	28-Feb-12	22
1	206473	Mesalamine	Crohn's Disease, Maintenance	"My son has Crohn's disease and has done ...	8	17-May-09	17
2	159672	Bactrim	Urinary Tract Infection	"Quick reduction of symptoms"	9	29-Sep-17	3
3	39293	Contrave	Weight Loss	"Contrave combines drugs that were used for al...	9	5-Mar-17	35
4	97768	Cyclafem 1 / 35	Birth Control	"I have been on this birth control for one cyc...	9	22-Oct-15	4

Figure 2: Header information and top 5 data rows in test data

Train shape (# Rows, # Columns) : (161297, 7)

Test shape (# Rows, # Columns) : (53766, 7)

Train Set / Test Set 2.999981400885318

Here it can be seen that the both the Train and Test Data sets have 6 variables and 1 unique id
Train Data set is ~ 3 times larger than the Test Data set.

Data Overview

A patient with a unique ID purchases a drug that meets his or her condition and subsequently writes a review, provides a rating for the drug he/she purchased on the date. later on, if the other users read that review and find it helpful, they will click usefulCount, which will add 1 for the variable hence increasing the usefulCount.

Data Types

drugName (categorical): name of the drug

condition (categorical): name of the condition

review (text): patient's review

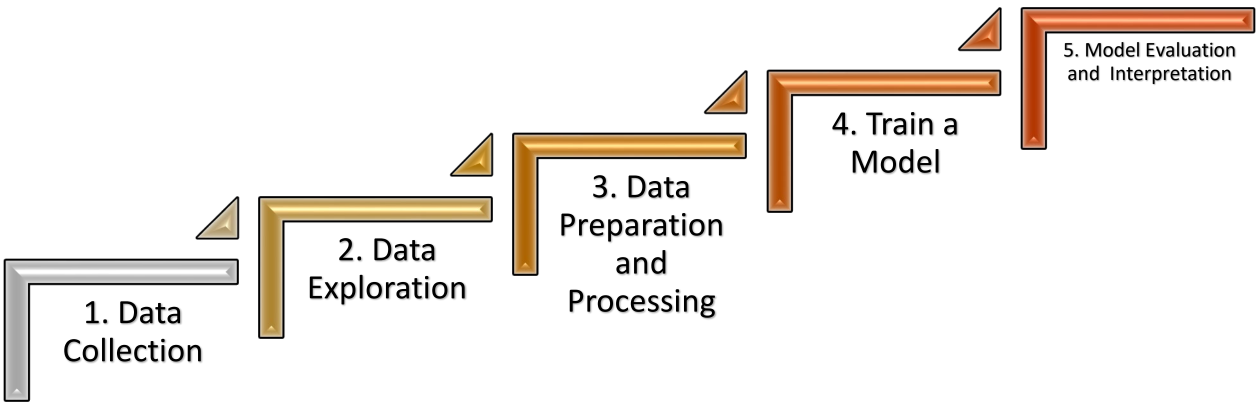
rating (numerical): 10 star patient's rating

date (date): date of review entry

usefulCount (numerical): number of users who found review useful

Data Approach and Methodology

Currently in this project the following methodology will be used. (tentative)



Github repository

The GitHub repository for this project is located at the following link.

https://github.com/vgorrepa/Big_Data_Final_Project

The repository will be updated periodically till the closure of the project.

Exploratory Analysis

Here we will use various data visualization tools to better understand the data.

<u>Top health conditions by count</u>

```
Birth Control                28788
Depression                    9069
Pain                          6145
Anxiety                       5904
Acne                          5588
Bipolar Disorde               4224
Insomnia                      3673
Weight Loss                   3609
Obesity                       3568
ADHD                          3383
Diabetes, Type 2              2554
Emergency Contraception       2463
High Blood Pressure           2321
Vaginal Yeast Infection       2274
Abnormal Uterine Bleeding     2096
Name: condition, dtype: int64
```

Figure 3: Top 15 health conditions by count

It can be inferred that Birth Control followed by Depression and Pain are the top 3 health conditions experienced by the users.

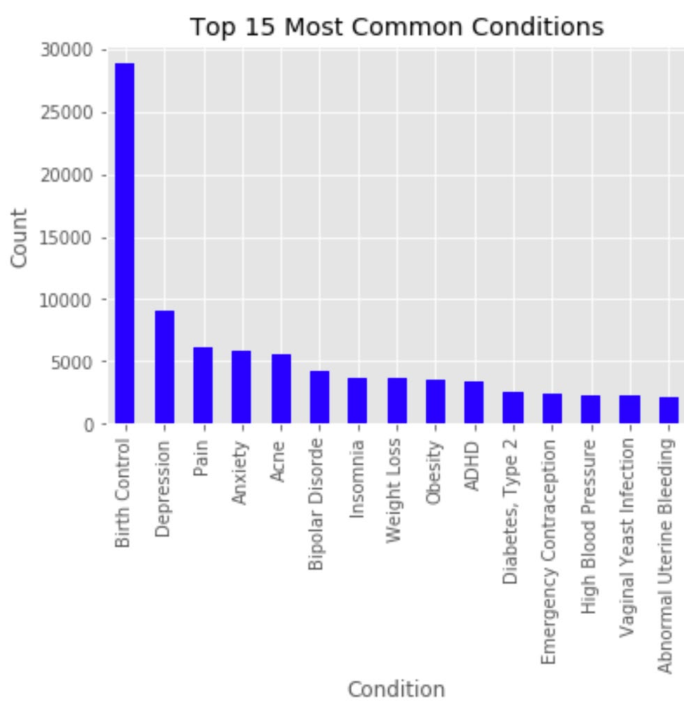


Figure 4: Histogram of Top 15 health conditions by count

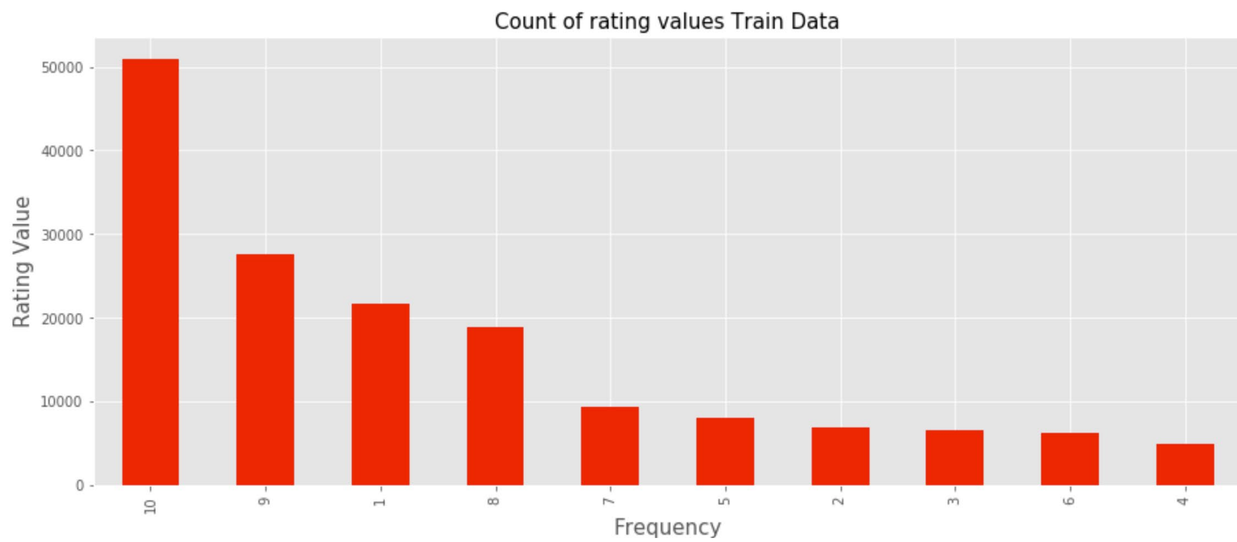


Figure 5: Histogram of count of rating values for Train Data set.

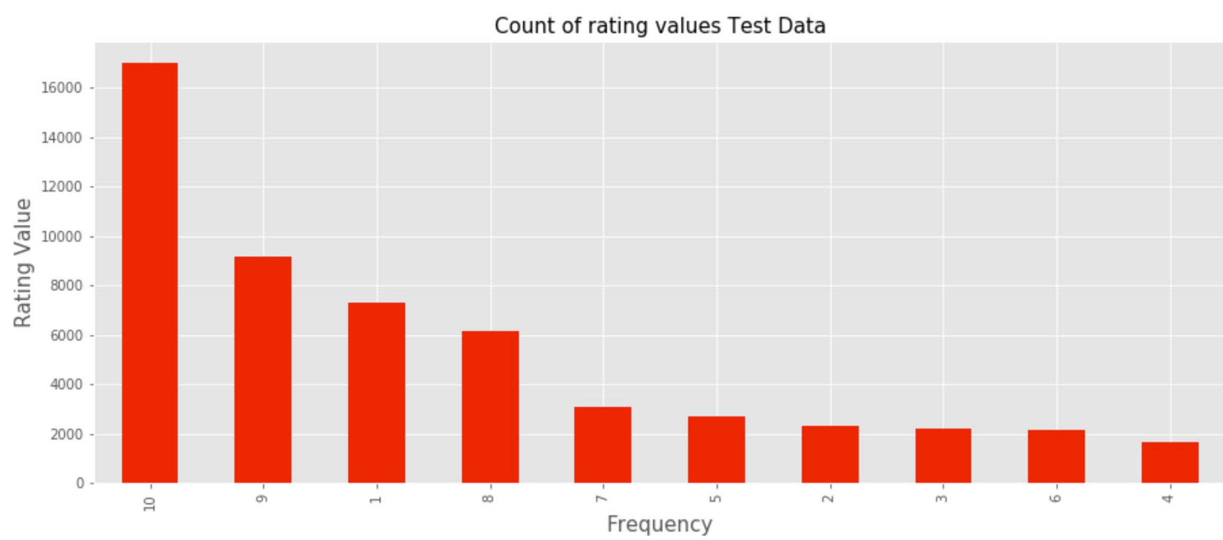


Figure 5: Histogram of count of rating values for Test Data set.

Comparing the histograms of the ratings it can be deduced that the users giving ratings and reviews are mostly either very satisfied (10,9) or highly unsatisfied (1,2).

Also, the distribution seems to be similar in both the Test and Train data sets.

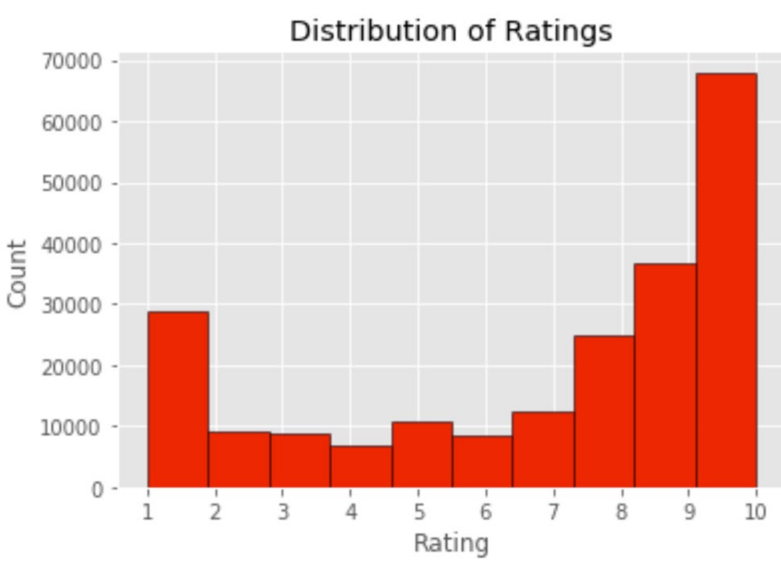
<u>Merging the Data sets together</u>

As can be observed from the above analysis both the dataset contains same columns hence, we can combine them for better analysis using a larger data set.

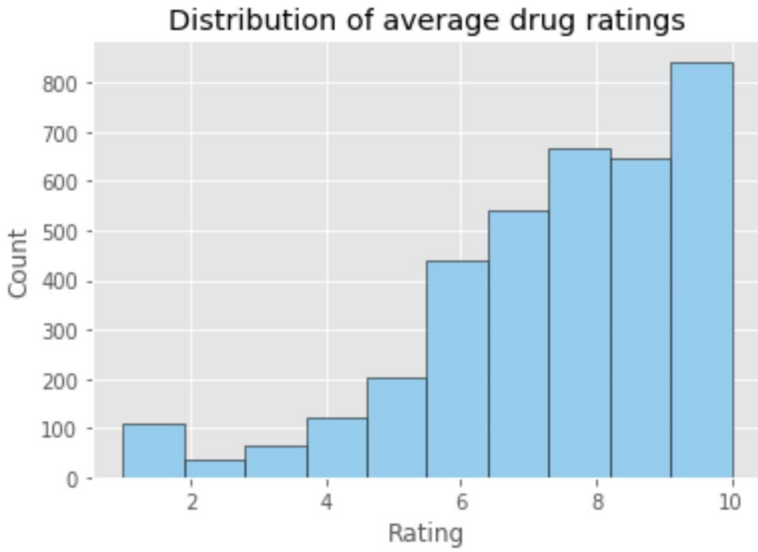
<u>Top health conditions by count full data set</u>

```
Birth Control                38436
Depression                   12164
Pain                          8245
Anxiety                       7812
Acne                          7435
Bipolar Disorde               5604
Insomnia                      4904
Weight Loss                   4857
Obesity                       4757
ADHD                          4509
Diabetes, Type 2              3362
Emergency Contraception       3290
High Blood Pressure           3104
Vaginal Yeast Infection       3085
Abnormal Uterine Bleeding     2744
Name: condition, dtype: int64
```

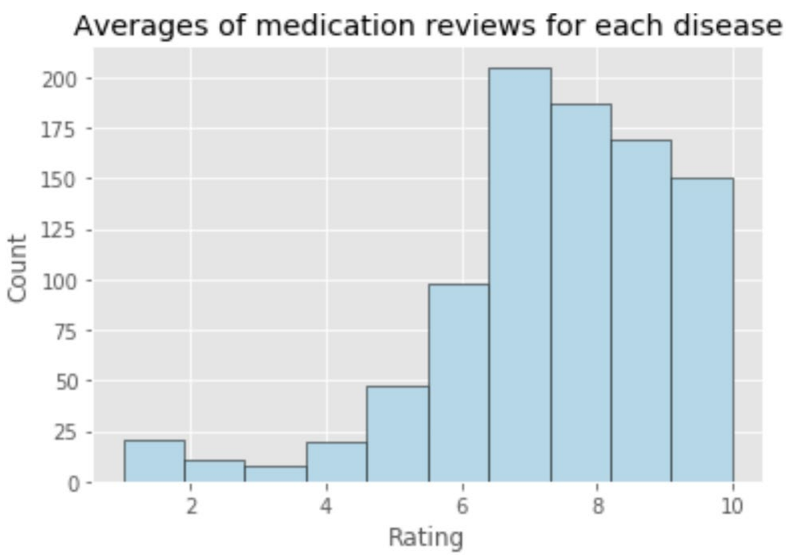
<u>Distribution of Ratings</u>



It can be inferred from above that majority of the ratings are positive (> 5).



It can be inferred from above that majority of the drugs have an average ratings >= 5 and hence positive.



It can be inferred from above that for a given medical condition the reviews generally have a average ratings >= 5 and hence positive.

<u>Descriptive statistics of the data</u>

	uniqueID	rating	usefulCount
count	215063.000000	215063.000000	215063.000000
mean	116039.364814	6.990008	28.001004
std	67007.913366	3.275554	36.346069
min	0.000000	1.000000	0.000000
25%	58115.500000	5.000000	6.000000
50%	115867.000000	8.000000	16.000000
75%	173963.500000	10.000000	36.000000
max	232291.000000	10.000000	1291.000000

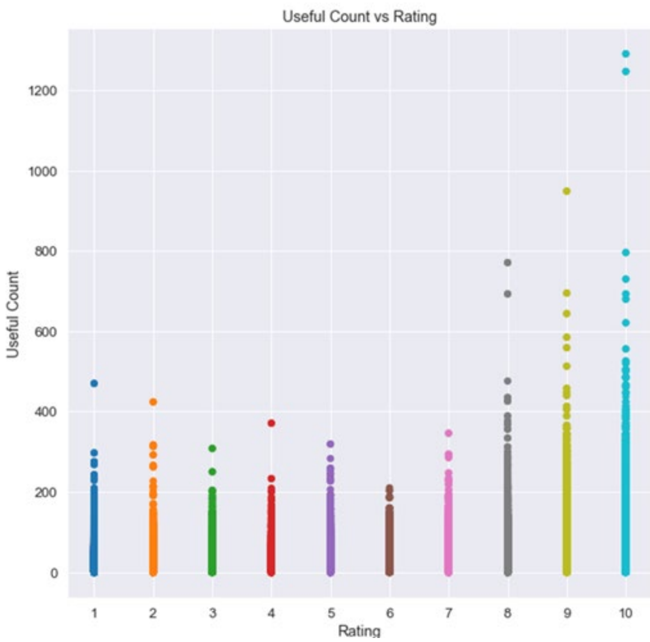
The average rating is 6.99, while the median rating is 8.

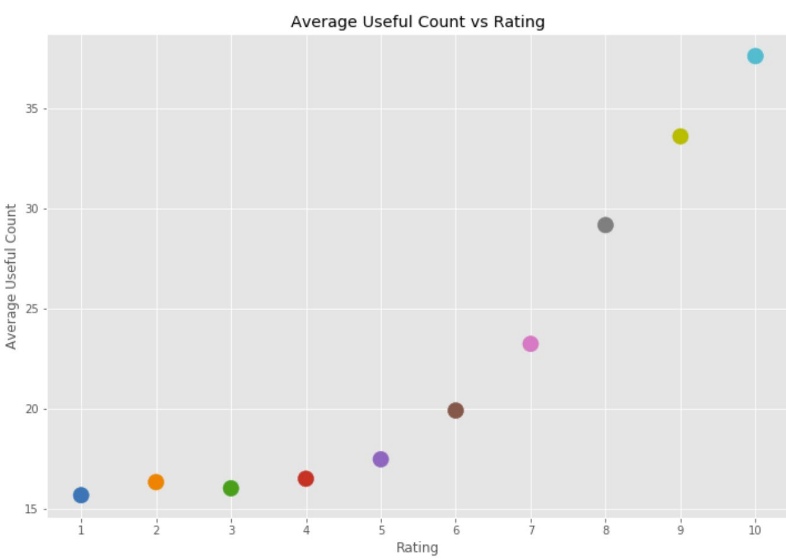
As the mean is less than the median, the distribution of ratings is negatively skewed.

The average Useful count is 28 while the median Useful count is 16.

As the mean is greater than the median, the distribution of the Useful count is positively skewed.

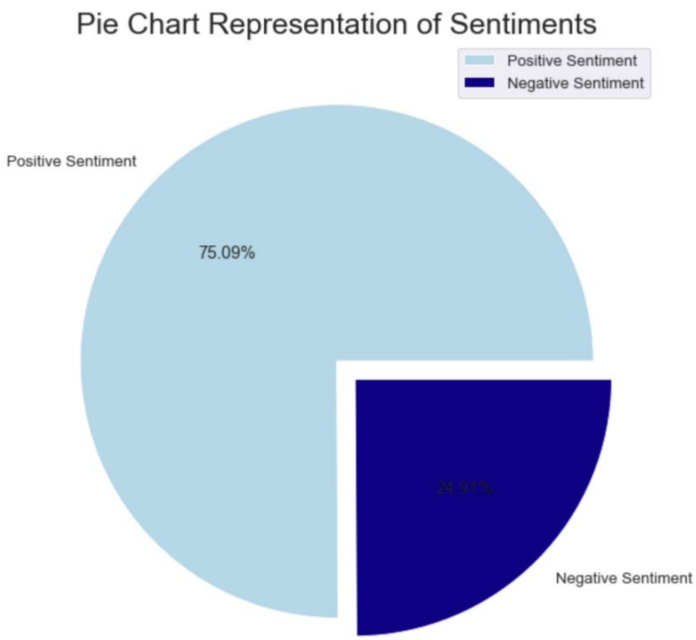
<u>Usefulness vs Rating</u>





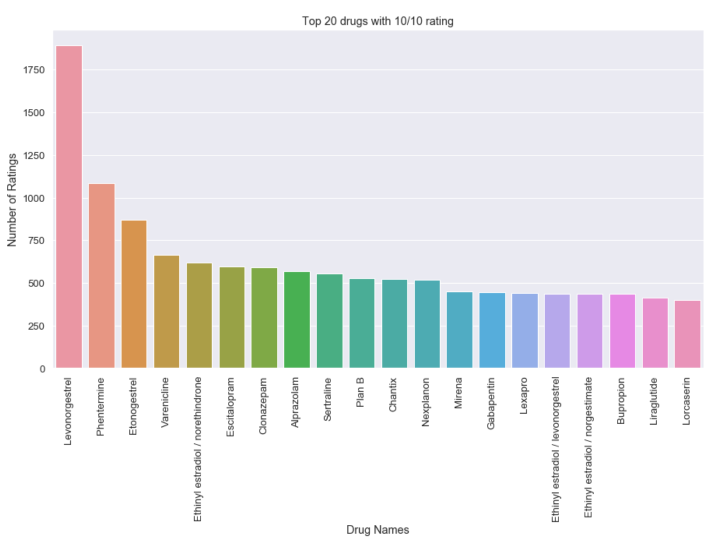
It can be inferred that people found reviews with higher scores to be more useful. The reviews with high ratings received more 'useful' tags than reviews with low ratings.

A pie chart to represent the sentiments of the patients

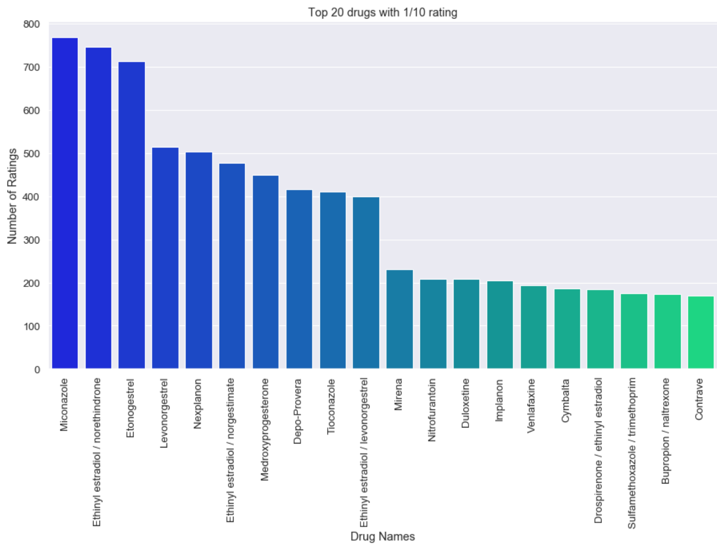


It can be inferred that the majority of the patients have a positive sentiment.

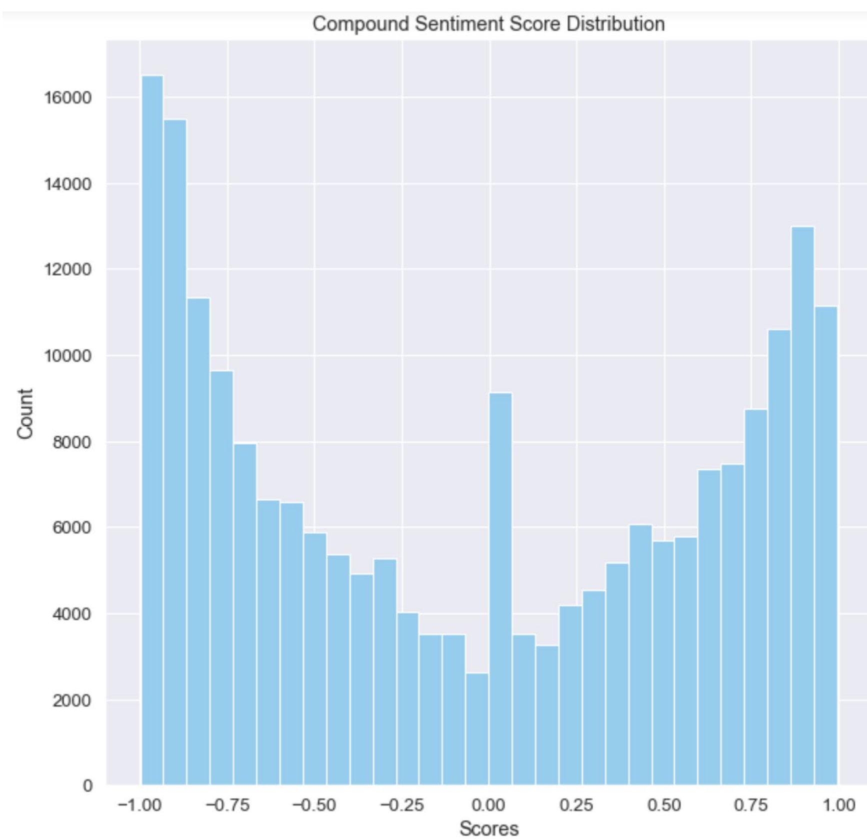
Top 20 Drugs with a 10/10 Rating

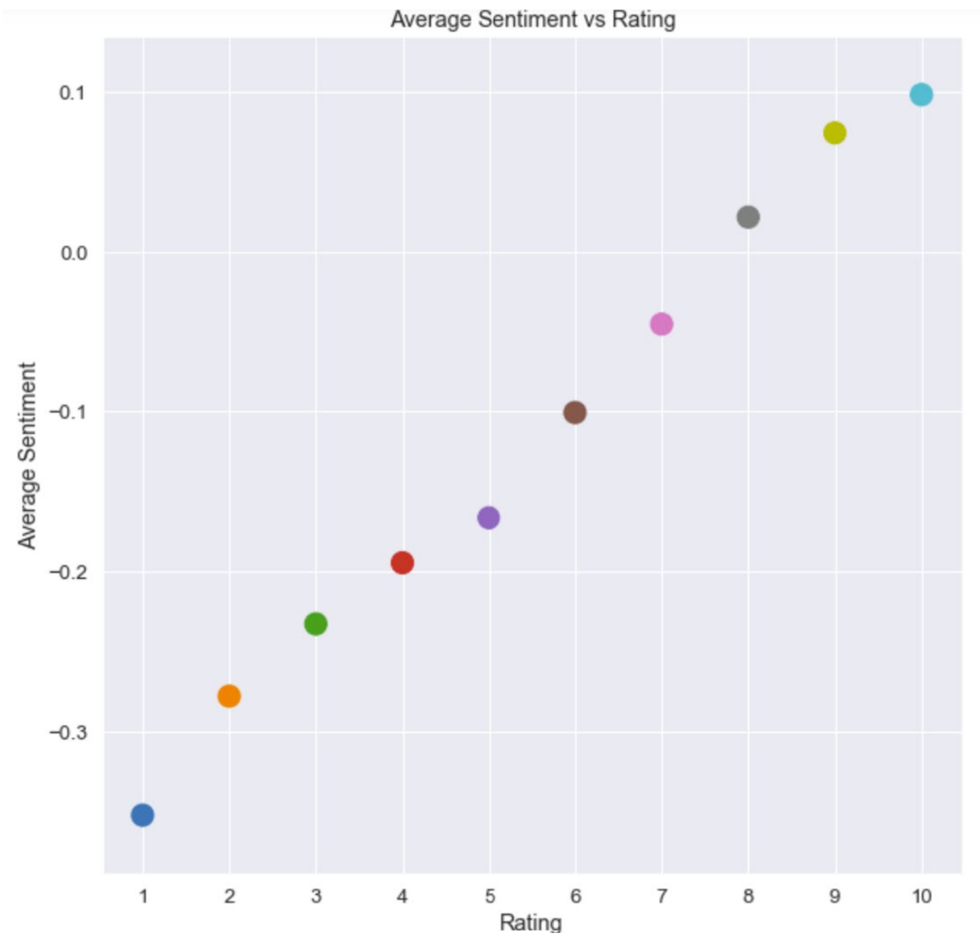


Top 20 Drugs with a 1/10 Rating



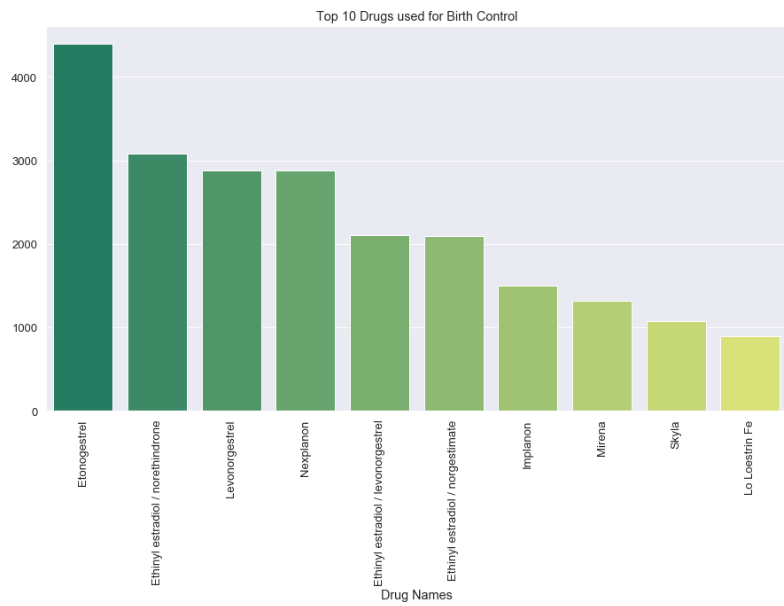
Sentiment Analysis of the Reviews



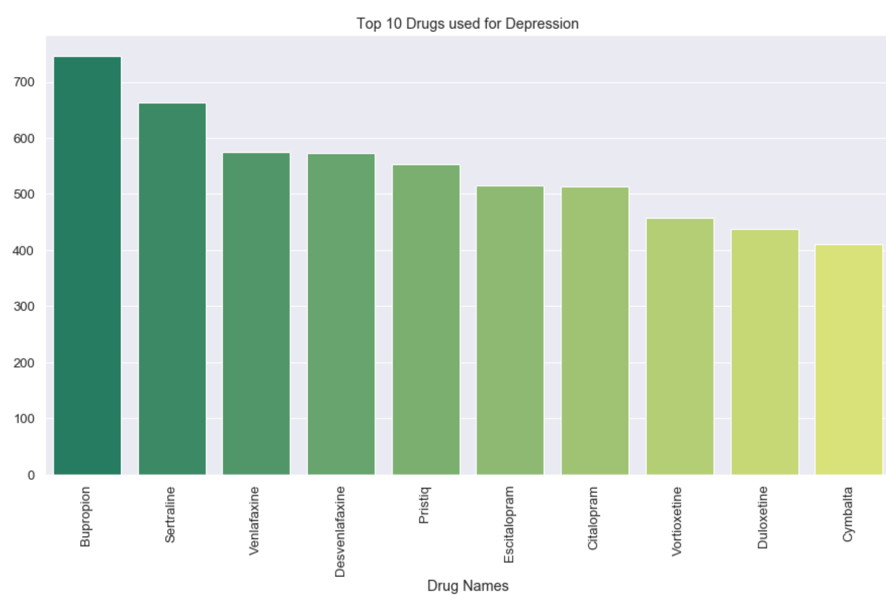


Hence, from above it can be inferred that a drug with a higher average rating generally has a higher ,more positive average sentiment.

Top 10 Drugs for Birth Control



Top 10 Drugs for Depression



Prediction: Feature Engineering and Model building

We will try the Naive Bayes Classifier, Logistic Regression , Random Forest Classifier, eXtreme Gradient Boosting algorithms and select the algorithm with the best accuracy.

Here we attempt to predict the Drug rating (Dependent Variable) based on the review

(Independent Variable).

TF-IDF : Term frequency-inverse document frequency

Natural language data is in the form of raw text, so that the text needs to be transformed into a vector. The process of transforming text into a vector is commonly referred to as text vectorization. Text vectorization algorithm namely TF-IDF vectorizer can help in transforming text into vectors. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF). (Ramadhan, 2022)

The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document. Term frequency represents every text from the data as a matrix whose rows are the number of documents and columns are the number of distinct terms throughout all documents. Document frequency is the number of documents containing a specific term. Document frequency indicates how common the term is.

Inverse document frequency (IDF) is the weight of a term, it aims to reduce the weight of a term if the term’s occurrences are scattered throughout all the documents. (Ramadhan, 2022)

The Naive Bayes Classifier

```
Training time: 0.09275031089782715
Accuracy on training set: 0.7791688462656204
Accuracy on test set: 0.7748587636296004
Confusion Matrix
[[ 1073  9570]
 [  114 32256]]
              precision    recall  f1-score   suppor

         0.0       0.90      0.10      0.18     1064
         1.0       0.77      1.00      0.87     3237

    accuracy                           0.77     4301
   macro avg       0.84      0.55      0.53     4301
weighted avg       0.80      0.77      0.70     4301
```

Logistic Regression

```
Training time: 75.55984711647034
Accuracy on training set: 0.9343156059285092
Accuracy on test set: 0.8840815567386604
Confusion Matrix
[[ 7753  2890]
 [ 2096 30274]]
              precision    recall  f1-score   support

         0.0       0.79      0.73      0.76     10643
         1.0       0.91      0.94      0.92     32370

    accuracy                           0.88     43013
   macro avg       0.85      0.83      0.84     43013
weighted avg       0.88      0.88      0.88     43013
```

<u>Random Forest Classifier</u>

```
Training time: 399.81198167800903
Accuracy: 0.8274940134377978
Confusion Matrix
[[ 3264  7379]
 [   41 32329]]
              precision    recall  f1-score   support

         0.0       0.99      0.31      0.47     10643
         1.0       0.81      1.00      0.90     32370

    accuracy                           0.83     43013
   macro avg       0.90      0.65      0.68     43013
weighted avg       0.86      0.83      0.79     43013
```

<u>eXtreme Gradient Boosting</u>

```
Training time: 4.380842447280884
Accuracy on training set: 0.7732752106945655
Accuracy on test set: 0.7734638365145421
Confusion Matrix
[[ 1186  9457]
 [  287 32083]]
              precision    recall  f1-score   support

         0.0       0.81      0.11      0.20     10643
         1.0       0.77      0.99      0.87     32370

    accuracy                           0.77     43013
   macro avg       0.79      0.55      0.53     43013
weighted avg       0.78      0.77      0.70     43013
```

<u>Algorithm comparison</u>

Algorithm	Accuracy	Time Taken in seconds
The Naive Bayes Classifier	0.77 or 77%	0.093
Logistic Regression	0.88 or 88%	75.56
Random Forest Classifier	0.83 or 83%	399.82
eXtreme Gradient Boosting	0.77 or 77%	4.38

Hence, Logistic Regression has higher accuracy and while it takes a relatively long time than the other Algorithms, it is not worst performer. Random Forest Classifier takes the longest time.

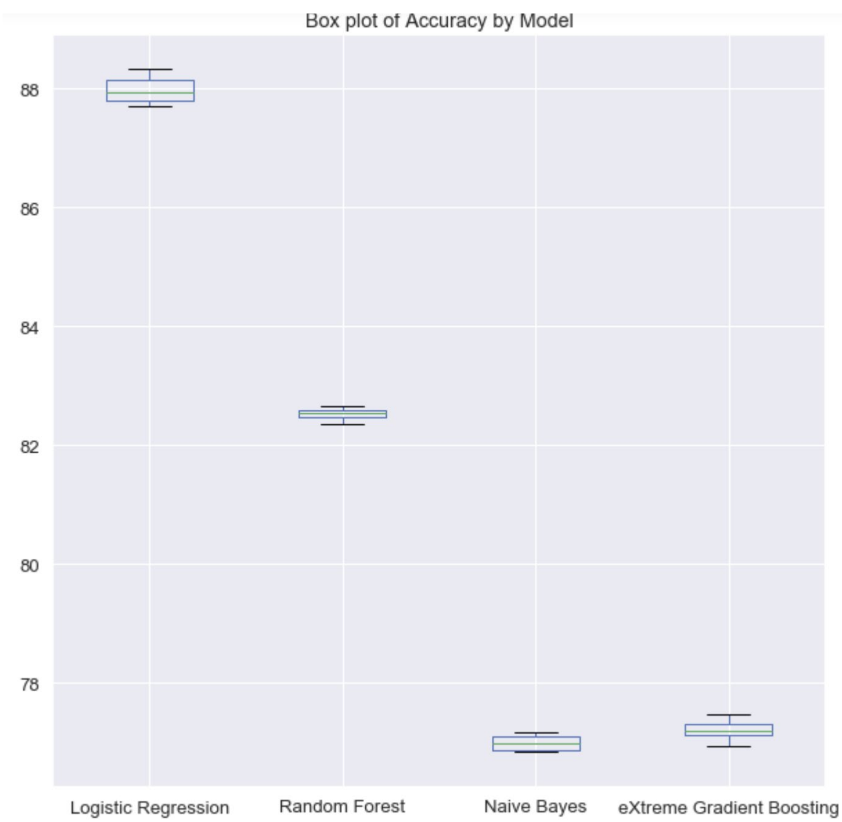
Evaluation of Results

K-Fold Cross-validation

K-fold (folds = 10) Cross validation was carried out on the data for the four algorithms above.

Accuracy for the 10 folds

	Logistic Regression	Random Forest	Naive Bayes	eXtreme Gradient Boosting
k-fold				
0	87.759372	82.609706	76.925312	77.465853
1	88.090671	82.441151	76.896251	77.221738
2	87.70125	82.551584	77.122929	77.204301
3	87.951177	82.662017	77.181052	76.948561
4	88.177855	82.563208	77.029933	77.349608
5	88.253415	82.359779	76.843941	76.942749
6	87.922116	82.551584	77.029933	77.448416
7	87.782621	82.452775	76.86719	77.111305
8	88.334786	82.534147	76.849753	77.163615
9	87.898867	82.64458	77.122929	77.186864
Average	87.987213	82.5370531	76.9869223	77.204301



Comparing the four models using the Accuracy metric, it becomes clear that Logistic Regression gives the most accurate predictions.

Conclusion

In conclusion, an attempt has been made to answer the three research questions using various analytical and Machine Learning Techniques, primarily using the Python software.

This project attempted to answer the following 3 research questions.

1. What are the Top drugs for a given condition?

   Top conditions in the data were identified and for any of the top conditions the top drugs by rating were identified. Comparing the histograms of the ratings it can be deduced that the users giving ratings and reviews are mostly either very satisfied or highly unsatisfied. Generally, the majority of the drugs have an average rating >= 5 and hence positive and also for a given medical condition the reviews generally have a average ratings >= 5 and hence positive.
2. Can the Drug rating be predicted based on the reviews?

   An attempt was made to predict the Drug rating (Dependent Variable) based on the review (Independent Variable). Four algorithms were explored namely, Naive Bayes Classifier, Logistic Regression , Random Forest Classifier, eXtreme Gradient Boosting and based on Cross-validation results the Logistic Regression algorithm was selected as it has the best accuracy.
3. Can we determine if a review is positive or neutral or negative?

Sentiment Analysis was conducted on the reviews to identify if a review is positive, neutral, negative and it was inferred that a drug with a higher average rating generally has a higher, more positive average sentiment. Next, a direct binary classification of the ratings was utilized such that a >=5 is considered positive review while <5 is considered a negative review. This showed that the majority (75.09%) of the reviews were positive while 24.01% of the reviews were negative.

Short comings

One of the main draw backs is that more work can be done to build a prediction model to predict the condition based on the reviews. Another drawback is that only four types of Machine Learning algorithms were explored while there are a lot more algorithms which might perhaps deliver better accuracy and performance. Finally, a better drug recommendation tool could be built from this data and perhaps it could be connected directly to the drugs.com website to take in live data instead of a static dataset as used in this project.

References

Ali, T. S. (2013, October). Can i hear you? sentiment analysis on medical forums. *The sixth international joint conference on natural language processing*, (pp. 667-673).

Bengio, Y. D. (2000). A neural probabilistic language model. *Advances in Neural Information Processing Systems, 13.*

Bobicev, V. S. (2012 May). Learning sentiments from tweets with personal health information. *Canadian conference on artificial intelligence* (pp. 37-48). Berlin, Heidelberg: Springer.

Colón-Ruiz, C. &.-B. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics, 110*, 103539-103539. Retrieved from https://doi.org/10.1016/j.jbi.2020.103539

de-Albornoz, J. C. (2018). Feature engineering for sentiment analysis in e-health forums. *PLOS ONE, 13(11)*, (pp. 1-25).

Esuli, A. &. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. . *Proceedings of the fifth international conference on language resources and evaluation.* (LREC'06).

Felix Gräßer, S. K. (2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. *In Proceedings of the 2018 International Conference on Digital Health (DH '18)* (pp. 121-125). New York, NY, USA: ACM.

Felix Gräßer, S. K. (2018, 10 04). *Drug Review Dataset (Drugs.com) Data Set.* Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

Goeuriot, L. N. (2012, January). Sentiment lexicons for health-related opinion mining. *In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, (pp. 219-226).

Kaggle. (2018). *Kaggle hackathon UCI ML Drug Review dataset.* Retrieved from Kaggle: https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018

Mishra, A. M. (2015, November). Towards automatic pharmacovigilance: analysing patient reviews and sentiment on oncological drugs. *IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 1402-1409). IEEE.

Na, J. C. (2012 November). Sentiment classification of drug reviews using a rule-based linguistic approach. *International conference on asian digital libraries* (pp. 189-198). Spring.

Ramadhan, L. (2022, 04 01). *Natural Language Processing : TF-IDF Simplified, a short introduction to TF-IDF vectorizer*. Retrieved from Towards Datascience: https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530

Wiley, M. T. (2014). Pharmaceutical drugs chatter on online social networks. . *Journal of biomedical informatics, 49*, 245-254.

Wilson, T. W. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. *Human language technology conference and conference on empirical methods in natural language processing*, (pp. 347-354).