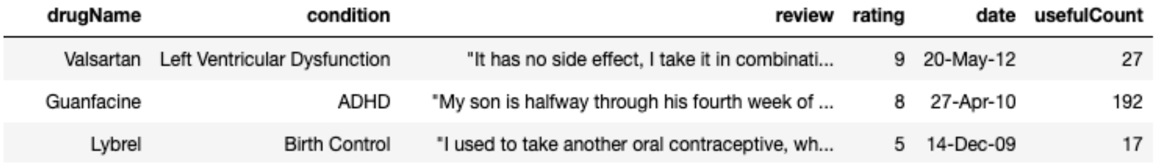
The growth of the internet and technology over the past two decades have allowed many humans to learn something that was not possible 30 years ago. The access to online information is unimaginable and medical patients have greatly benefited by utilizing web platforms to inquire about other patient’s experiences, views and suggestions about a particular medical condition or medical drugs. Patients are no longer limited to getting this information from known family members or a particular medical doctor allowing them to make a better-informed decision.

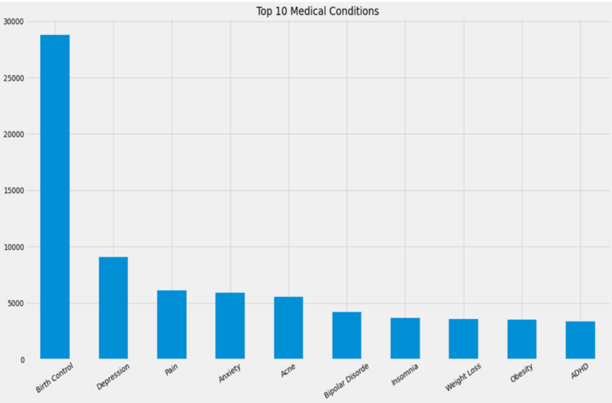
A medical drug is first tested and evaluated before its approval, but there are still instances where the drug must be withdrawn from the market due to some unexpected side effects. These unexpected side effects are often reported online review sites, healthcare web forums and discussion boards. However, the unstructured textual nature of the reviews was often time consuming for healthcare professionals and difficult to digest in a timely manner. The emergence of Natural Language Processing (NLP) and Sentiment Analysis has made big impact in the industry by allowing them to identify, extract and make use of subjective information.

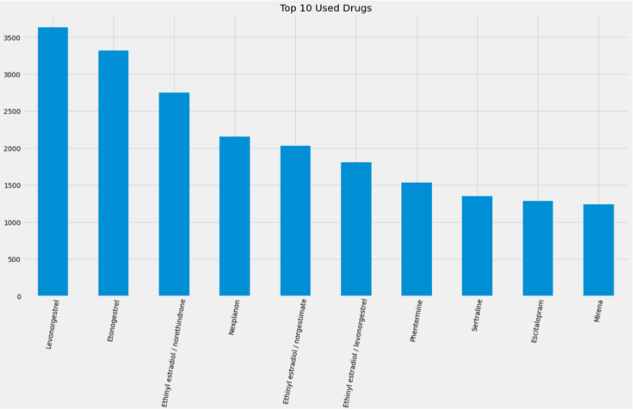
For this project, we will be using Natural Language Processing techniques to evaluate patient’s reviews and ratings of medical conditions and drugs. The data comes from UCI ML which contains over 160k reviews which was obtained by crawling online pharmaceutical review sites. Below is a sample of the data which has 0 to 10 ratings for a particular drug and condition as well as the comment left by the patient and the number of patients that found the drug useful.



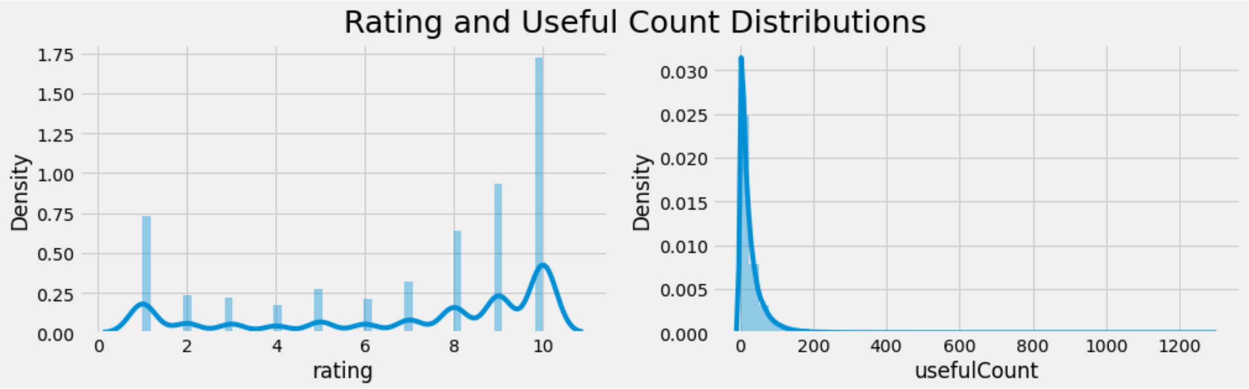
drugName	condition	review	rating	date	usefulCount
Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17

The dataset contains 884 unique medical conditions and 3,431 medical drugs. The next thing we want to look at is what types of conditions our patients are dealing with and what drugs are most often used. The 10 most used drugs and conditions are:

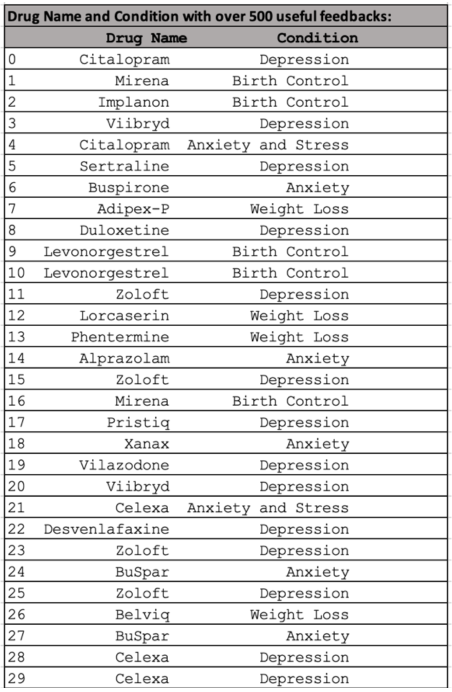




The other numerical data we can see are the distributions of the ratings and the usefulCount columns. Both features will allow us to see how satisfied the patients are with their medical drug prescriptions.



Even though we have over 800 medical conditions, it looks like most of the conditions and drugs did not get good feedback from their users or there was not enough data. To better understand this issue, we will look at cases where we had over 500 useful counts.

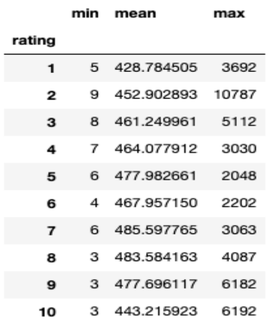


Drug Name and Condition with over 500 useful feedbacks:

	Drug Name	Condition
0	Citalopram	Depression
1	Mirena	Birth Control
2	Implanon	Birth Control
3	Viibryd	Depression
4	Citalopram	Anxiety and Stress
5	Sertraline	Depression
6	Buspirone	Anxiety
7	Adipex-P	Weight Loss
8	Duloxetine	Depression
9	Levonorgestrel	Birth Control
10	Levonorgestrel	Birth Control
11	Zoloft	Depression
12	Lorcaserin	Weight Loss
13	Phentermine	Weight Loss
14	Alprazolam	Anxiety
15	Zoloft	Depression
16	Mirena	Birth Control
17	Pristiq	Depression
18	Xanax	Anxiety
19	Vilazodone	Depression
20	Viibryd	Depression
21	Celexa	Anxiety and Stress
22	Desvenlafaxine	Depression
23	Zoloft	Depression
24	BuSpar	Anxiety
25	Zoloft	Depression
26	Belviq	Weight Loss
27	BuSpar	Anxiety
28	Celexa	Depression
29	Celexa	Depression

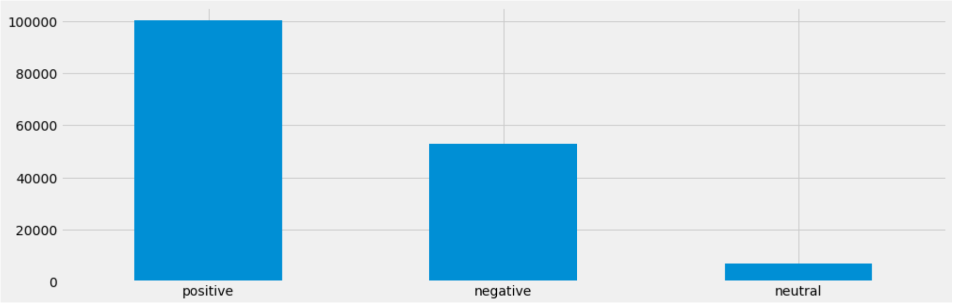
We can see that Anxiety, Birth Control, Depression, Stress and Weigh Loss are the conditions that occur the most with several different medical drugs.

Now, we have a strong understanding of our data so we can move on to the reviews. Before performing sentiment analysis, we want to see if the length of the review comments have any relationship to the ratings given by the patient.

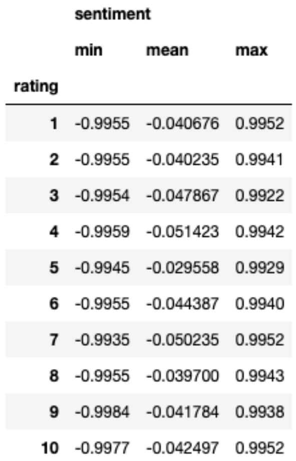


	min	mean	max
rating			
1	5	428.784505	3692
2	9	452.902893	10787
3	8	461.249961	5112
4	7	464.077912	3030
5	6	477.982661	2048
6	4	467.957150	2202
7	6	485.597765	3063
8	3	483.584163	4087
9	3	477.696117	6182
10	3	443.215923	6192

Looking at the mean length of the reviews doesn't tell us any useful information. Therefore, we can move on to perform sentiment analysis. Our results from the sentiment analysis can be seen below:



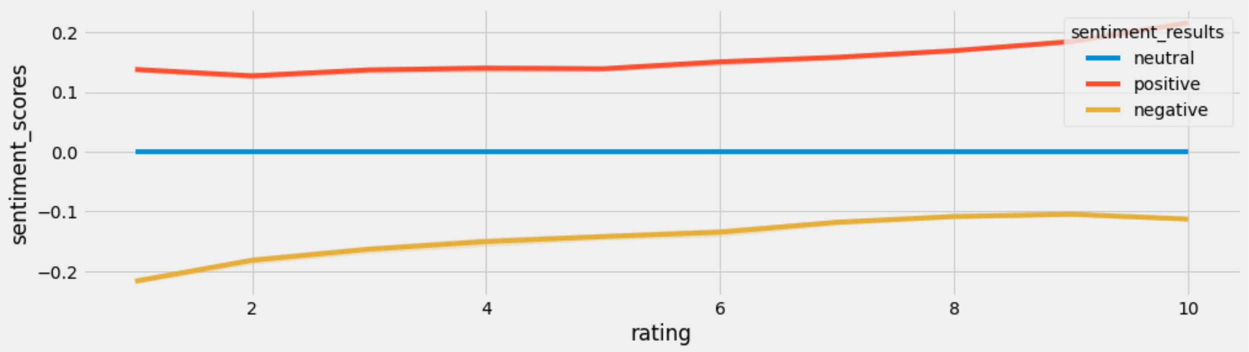
Before we call it a day, let's look at the sentiment analysis with a little more detail to understand the results. Like the length of the reviews, we will use the min, mean and max of the compound polarity scores.



	sentiment		
	min	mean	max
rating			
1	-0.9955	-0.040676	0.9952
2	-0.9955	-0.040235	0.9941
3	-0.9954	-0.047867	0.9922
4	-0.9959	-0.051423	0.9942
5	-0.9945	-0.029558	0.9929
6	-0.9955	-0.044387	0.9940
7	-0.9935	-0.050235	0.9952
8	-0.9955	-0.039700	0.9943
9	-0.9984	-0.041784	0.9938
10	-0.9977	-0.042497	0.9952

Unfortunately, the results are unreliable because similar scores are spread throughout all the ratings. Therefore, we will have to create our own scoring system to get a better understanding.

Instead of just looking at the sentiment compound results, let's look at the scores by negative, positive and neutral scores.



This gives us a better picture of what we are looking at, but we would conclude that additional work is need to validate the results.