Lucas Reynolds
Springboard Data Science Career Track
January 2020 Cohort
Capstone Project 1

# Milestone Report

## Overview

Drug reviews provide important information to drug manufacturers, such as levels of customer satisfaction, successful outcomes, or negative side effects. The internet houses several sources of drug reviews that could potentially be scraped or used. However, the rating scale and information required often differs from site to site. If we can predict whether or not a rating will be good based on the review, we may be able to label reviews with missing ratings, standardize ratings across review sources, and better understand the factors that most influence a drug's rating.

## Dataset

Drugs.com provides a database of peer-reviewed information on various medicines as well as user ratings and reviews gathered from across the web. We will use a dataset including 215,063 rows of ratings and review data from Drugs.com available via the UCI Machine Learning Repository.[1]

The dataset contained 161,297 rows and the following 7 columns: *Unnamed* (review ID number), *drugName* (name of the drug reviewed), *condition* (diagnosed condition for which the drug was prescribed), *review* (given text of the review), *rating* (rating from 1.0 to 10.0), *date* (review data), *usefulCount* (number of website visitors who found the review useful).

The dataset was imported as a pandas dataframe and previewed to gain an understanding of overall structure and context. The *review* and *rating* columns were checked for any null values. No null values were present.

## Rating Context

Over 30% of reviews had a rating of 10 and ratings 1, 8, and 9 each held over 10% of reviews. I had some expectation that most reviews would be at one end or the other of ratings, but was surprised that the large majority had a rating of 10. Ratings 2 through 7 respectively held 6% of the overall reviews.

To understand general sentiment and context, I sampled 10 reviews from each rating score. Reviews sampled for ratings 1 through 4 generally had an overall negative sentiment, reviews sampled for ratings

[1] https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

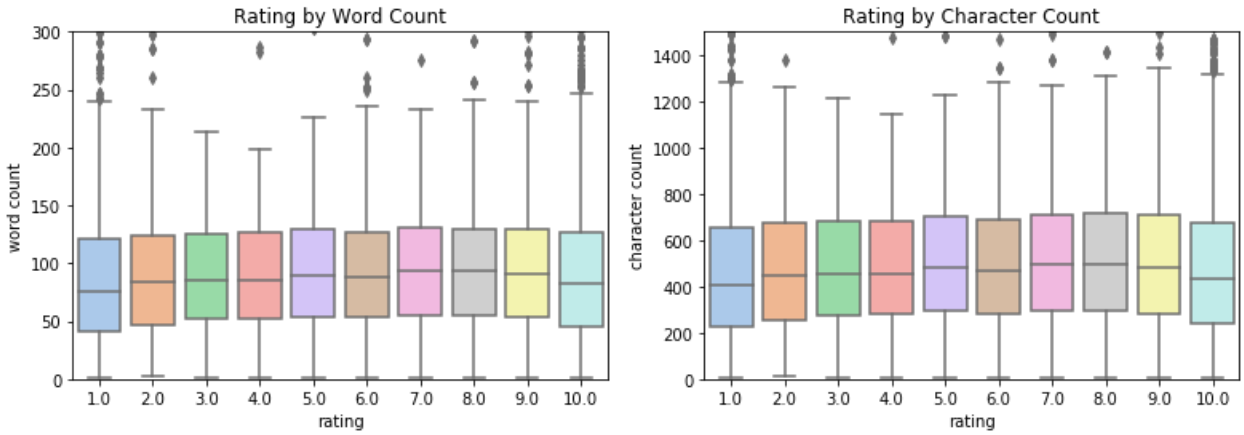
5 through 7 were more neutral or contained a mixture of positive and negative sentiment, and reviews sampled for ratings 8 through 10 had an overall positive sentiment.

# Cleaning Text

A bag-of-words approach was used for analyzing the review text. The text was cleaned and standardized by making all words lowercase and removing any punctuation, characters, and words that weren't essential to the meaning of the text. The remaining text was used to build a document-term matrix. A row is created for each review, every term that appears is given a column, and the matrix is then filled with counts for words that appear in each review. This process allows us to then evaluate text quantitatively.

# Review Length

The length of reviews were visually explored with box plots comparing word count versus rating and character count versus rating. There was not a clear relationship between review length and rating from the boxplots.




To better evaluate any relationship between review length and rating, I calculated the statistically significant correlation. Using a p-value of 0.05, I found there was a statistically significant correlation, but the correlation was very weak. The correlation coefficients of rating with both word count and character count were about 0.02 respectively.

# Drug Names

To determine which drugs had the largest effect on rating, I calculated statistically significant average differences with a p-value of 0.05. To minimize the effect of outliers and drugs with small sample size, I filtered out drugs that appeared less than 50 times. 233 drugs had statistically significant average differences in rating ranging from an average decrease in rating of 4.87 to an average increase of about

3.00. Below are the top 5 drugs with largest significant average increase and largest significant average decrease in rating.

| 5 Drugs with Largest Avg. Increase in Rating | | |
|---|---|---|
| | **Drug Name** | **Avg. Difference (Rating)** |
| 1 | Privine | 3.00 |
| 2 | Zinc Oxide | 3.00 |
| 3 | Astelin | 3.00 |
| 4 | Acetaminophen / pseudoephedrine | 3.00 |
| 5 | Chlorpheniramine / phenylephrine | 3.00 |

| 5 Drugs with Largest Avg. Decrease in Rating | | |
|---|---|---|
| | **Drug Name** | **Avg. Difference (Rating)** |
| 1 | Systane | -4.87 |
| 2 | Succinylcholine | -4.62 |
| 3 | Trimethoprim | -4.25 |
| 4 | Chloraseptic Sore Throat Spray | -4.00 |
| 5 | Monistat 7 | -3.97 |

The 5 drugs with the largest average increase in rating are all over the counter (OTC) options for common conditions like nasal congestion, allergies, pain, and rashes. The 5 drugs with the largest average decrease in rating included some OTC options for eye drops, sore throat, and yeast infection along with a muscle relaxant and an antibiotic for urinary tract infection that require prescriptions.

# Medical Condition

Similar to drug names, I explored which medical conditions had a significant average difference in rating. To minimize the effect of outliers and conditions with small sample size, conditions that appeared less than 50 times were not included. Again, a p-value of 0.05 was used. There were 75 conditions with significant average differences ranging from an average decrease in rating of about 3.25 to an average increase in rating of about 3.00. Below are the top 5 conditions with the largest significant average increase and the top 5 conditions with the largest significant average decrease.

| 5 Conditions with Largest Avg. Increase in Rating | | |
|---|---|---|
| | **Medical Condition** | **Avg. Difference (Rating)** |
| 1 | Costochondritis | 3.00 |
| 2 | Menopausal Disorders | 3.00 |
| 3 | B12 Nutritional Deficiency | 3.00 |
| 4 | Von Willebrand's Disease | 3.00 |
| 5 | Gingivitis | 3.00 |

| 5 Conditions with Largest Avg. Decrease in Rating | | |
|---|---|---|
| | **Medical Condition** | **Avg. Difference (Rating)** |
| 1 | Bronchospasm Prophylaxis | -3.25 |
| 2 | Sore Throat | -3.11 |
| 3 | Herpes Zoster, Prophylaxis | -3.00 |
| 4 | Macular Edema | -2.85 |
| 5 | Pancreatic Exocrine Dysfunction | -2.50 |

The 5 conditions with the largest significant average increase in rating include pain-inducing inflammation of the rib cage, menopausal disorders, a nutritional deficiency that results in tiredness and weakness, hemophilia, and gingivitis. The 5 conditions with the largest significant average decrease in rating include include asthma, sore throat, shingles, build up of fluid in the retina, and a deficiency of pancreatic enzymes that leads to an inability to digest food properly

# Words

I explore the 10 most frequent words in reviews for each rating. At first there was a lot of overlap of certain words across the top 10 words of each rating. I revisited my stop words and removed words with the most overlap. The top 10 words for each rating are listed below.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| like | months | months | months | months | first | first | effects | years | years |
| months | like | like | first | first | months | effects | first | effects | effects |
| took | first | first | like | effects | effects | months | years | first | first |
| never | pill | effects | effects | like | like | like | months | months | life |
| would | effects | pill | pill | pill | years | years | like | like | like |
| first | would | would | period | feel | weight | feel | feel | feel | feel |
| pill | back | period | month | years | feel | also | also | would | months |
| effects | got | also | feel | period | period | weight | weeks | back | would |
| doctor | period | feel | also | also | also | pill | would | one | one |
| one | weeks | month | weight | weight | week | back | weight | weeks | back |

Using a p-value of 0.05, I compared ratings with and without each word and determined words with a significant average difference in rating. There were 342 words with significant average differences in rating. The significant average differences ranged from an average decrease in rating of about 2.15 to an average increase in rating of about 1.76. Below are the 5 words with the largest significant average increase in rating and the 5 words with the largest significant average decrease in rating.

| 5 Words with Largest Avg. Increase in Rating | | |
|---|---|---|
| | **Words** | **Avg. Difference (Rating)** |
| 1 | Radio | 1.76 |
| 2 | Coworker | 1.53 |
| 3 | Levoxyl | 1.41 |
| 4 | Aspergers | 1.40 |
| 5 | Laminectomy | 1.39 |

| 5 Words with Largest Avg. Decrease in Rating | | |
|---|---|---|
| | **Words** | **Avg. Difference (Rating)** |
| 1 | Moist | -2.15 |
| 2 | Consists | -1.50 |
| 3 | Separation | -1.48 |
| 4 | Tumour | -1.43 |
| 5 | Wire | -1.43 |