Name: Fardeen Bablu
Dataset: https://doi.org/10.24432/C5SK5S
Title: Exploring User Engagement Patterns and Correlations in Drug Reviews Dataset

Introduction
Drug efficacy is a critical aspect of pharmaceutical development, influencing treatment outcomes and patient well-being. To optimize drug selection and administration, healthcare providers rely not only on clinical trials but also on real-world patient feedback. This feedback, often captured through drug reviews, provides valuable insights into subjective efficacy, which is the perceived effectiveness and satisfaction experienced by patients. However, analyzing vast amounts of unstructured text data from online review platforms poses significant challenges.

The dataset under examination originates from a 2018 study that did large-scale aspect-based sentiment analysis of drug reviews and applied that model for cross-data learning [1]. Essentially, it used aspect-based sentiment analysis, a technique that dissects reviews into smaller phrases or words (aspects), enabling precise sentiment analysis. This method demonstrated the utility of this approach but also highlighted the scarcity of annotated data, which is where data is labeled and classified to then be put in a model.

This project will extend these findings using the study's publicly available dataset sourced from Drugs.com [2]. The problem is to understand and find the relationships and correlations between drug efficacy and user satisfaction. The ultimate goal is to enhance understanding of subjective efficacy and its relationship with objective drug performance, which may assist in more informed decision-making in healthcare.

Building upon the insights from the original study, I hypothesize that there exist correlations between user engagement metrics, such as review frequency and usefulCount, and the perceived performance or satisfaction associated with each unique drug. Specifically, I predict that a higher drug effectiveness rating and longer drug review text will correlate with a higher usefulCount.

Data Exploration
The dataset comprises two subsets, totaling 215,063 reviews, partitioned into training (75%) and testing (25%) sets. There are 6,345 unique drugs within the dataset, and the variables of interest are as follows:

1. drugName (categorical): The name of the drug.
2. condition (categorical): The medical condition associated with the drug.
3. review (text): Patient reviews, to be transformed into numerical data using sentiment analysis techniques.
4. rating (numerical): Patient ratings on a 10-star scale.
5. date (date): The date of each review entry.
6. usefulCount (numerical): The count of users who found a particular review helpful.

When downloading the dataset, there are two files, one that was used to train the study's sentiment analysis model, and another that was used to assess the model. For the sake of simplicity, the two files were combined in Excel. After combining, it was confirmed that there are 215,063 total reviews, compared to the assumed 218,614. After doing this, I imported and printed the dataset into an R Markdown file.

The data was first sorted by drugName and the head was printed to check accuracy. In order to go any further, more variables are needed so that a proper model can be created. This is because this project requires additional numerical variables to do key analysis techniques (plotting, regressions, confidence intervals, etc). The following were additional numerical variables that were added:

- avgRating (numerical): The average patient rating on a 10-star scale, with a higher value indicating higher perceived effectiveness or satisfaction with drug
- avgCharCount (numerical): The average total character count of patient reviews.
- reviewCount (numerical): The total number of reviews for each unique drug.
- avgUsefulCount (numerical): The average number of users who found each review useful.
- avgWordCount (numerical): The average word count of patient reviews.
- avgReadability (numerical): The average readability score of patient reviews based on the Flesch Reading Ease score.

Moving on, after finding these averages, the singular data points from the original data frame were largely removed. This was done to narrow down the data into more relevant variables that can serve as predictors for later regression models. After doing so, I printed a summary of the data frame. From the summary of the data frame, we can observe the following:

Figure 1: drugSumm Summary Table

drugName	avgRating	avgCharCount	reviewCount	avgUsefulCount	avgWordCount	avgReadability
Length	3671					
Class	character					
Mode	character					
Min.	1.000	3.0	1.00	0.00	1.00	-21.81
1st Qu.	6.286	232.8	2.00	6.00	42.00	74.25

Median	7.778	345.0	6.00	14.00	62.69	79.09
Mean	7.441	342.3	58.58	20.16	62.31	77.78
3rd Qu.	9.000	447.1	30.00	28.73	82.00	83.23
Max.	10.000	2008.0	4930.00	166.23	371.50	121.22
NA's						43

Average Rating (avgRating)

- The mean rating is 7.441, suggesting that drugs, on average, are rated relatively high by users in this dataset.

Average Review Length (avgCharCount)

- The mean review length is 342.3 characters, indicating the average number of characters per review in the dataset.

Average Review Count (reviewCount)

- The mean review count is 58.58, indicating the average number of reviews per drug in the dataset.

Average Usefulness Count (avgUsefulCount)

- The mean usefulness count is 20.16, representing the average number of users who found a review helpful.

Average Word Count (avgWordCount)

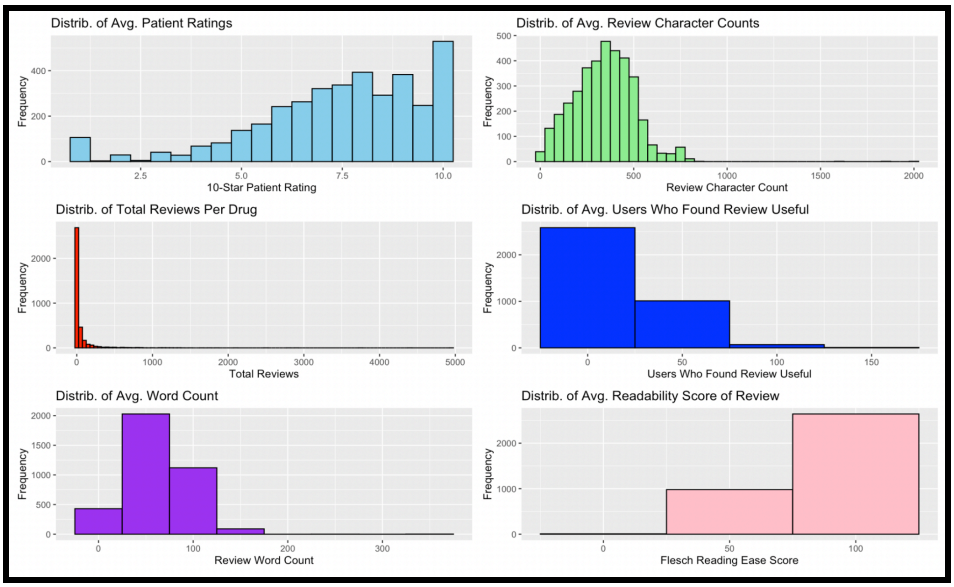
- The mean word count is 62.31, indicating the average number of words per review.

Average Readability Score (avgReadability)

- The mean readability score is 77.78, indicating the average readability level of user-generated reviews.

After observing this, I made some histogram visualizations to notice any patterns and correlations that can later be extrapolated.

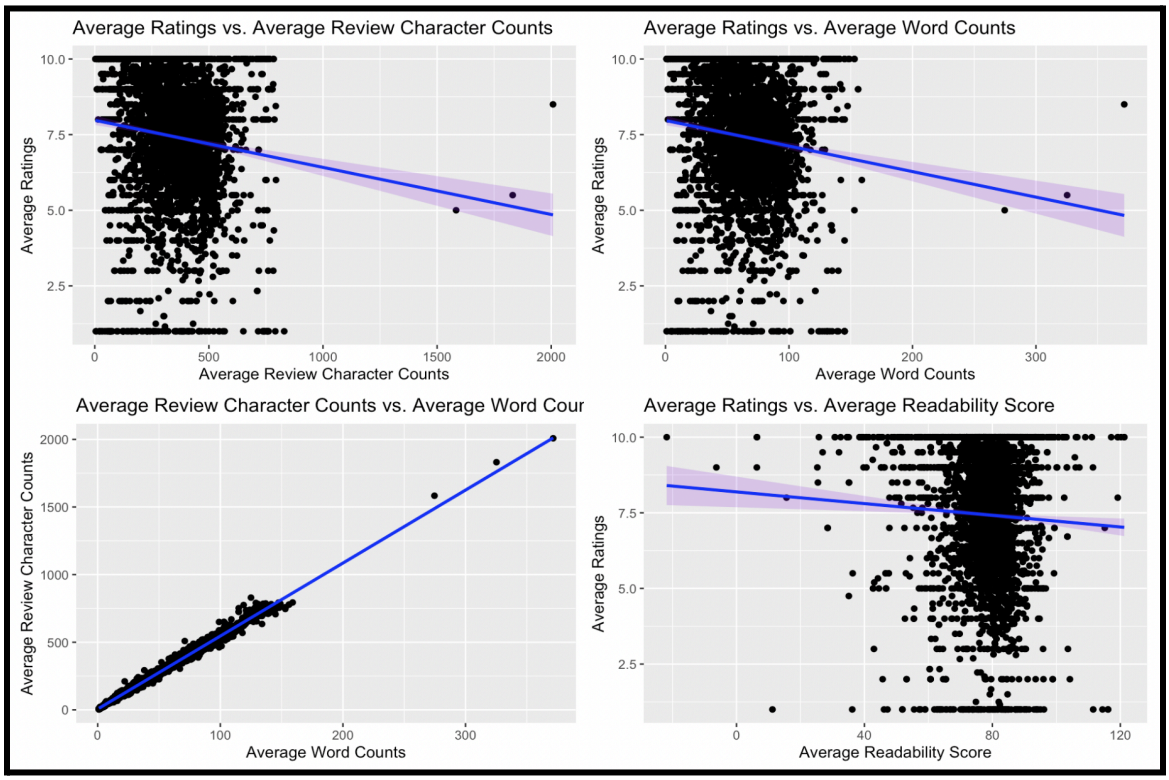
Figure 2: Histogram Visualizations of Variables



Looking through these visualizations, we can see the bigger picture. The most relevant graph to be noted is the Distribution of Average Patient Ratings, which is a left-skewed distribution towards higher reviews. This is interesting, and fits logically with the fact that many patients who review these drugs are either passionately against or for the drug. In this case, it shows that patients that are reviewing drugs are more likely to give a higher rating, as it may have provided better patient satisfaction, and thus they would like to share their positive experience online.

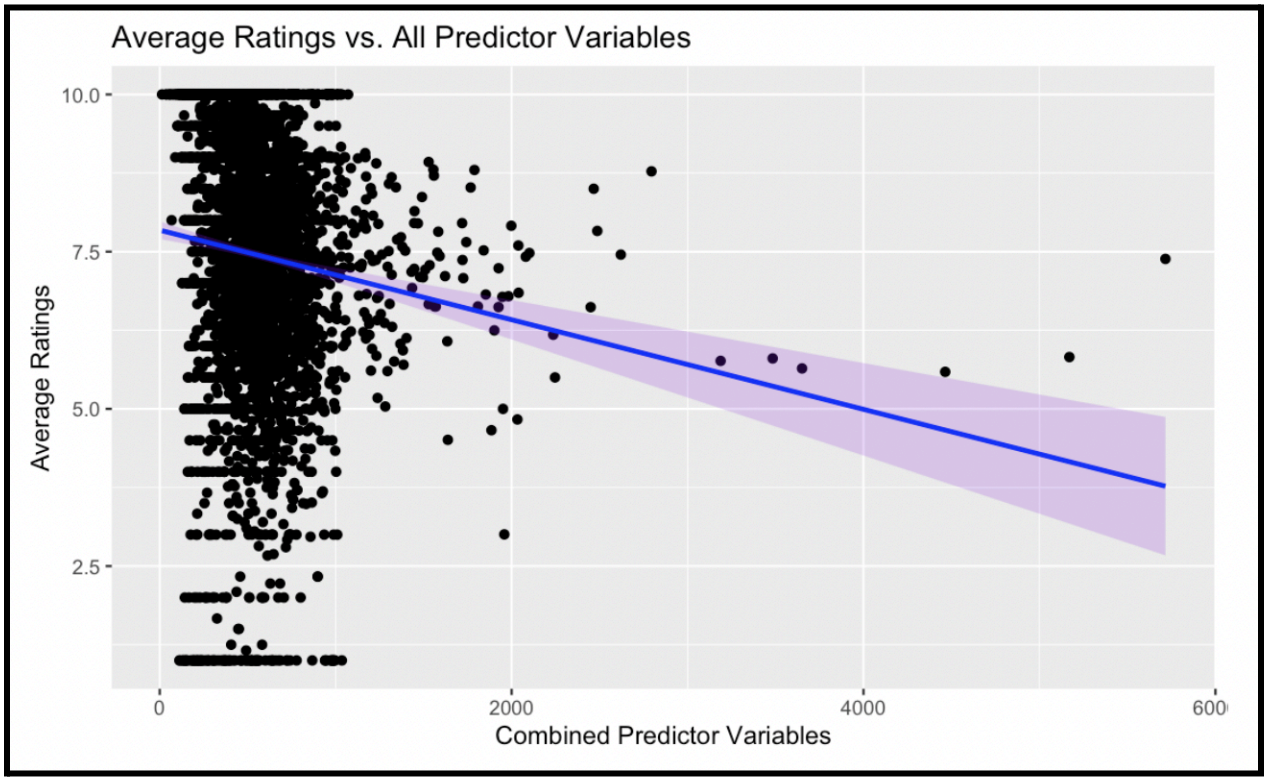
Going further, I looked at scatter plots with linear regression models and confidence intervals to see any relationships between these variables.

Figure 3: Scatterplot Visualizations



With this, it is a little bit difficult to see some meaningful relationships. Although the data is verbose, relationships between variables are not necessarily corollary. Regardless, after noticing these plots, I decided to create a larger model that included all variables excluding the average rating per drug against the average rating per drug. Despite the unclear visualizations from the scatterplot, I considered the average ratings metric to be a significant dependent variable that can be placed in the y-axis for further investigation.

Figure 3: Combined Variables vs Average Ratings Plot



The regression model (mainModel in code) suggests that the adjusted R-squared value of the model is 0.02362, indicating that the model explains about 2.362% of the variance in the average drug rating. The p-values of avgUsefulCount and avgReadability are statistically significant at the 0.05 level, suggesting that these predictors are associated with average drug ratings.

Conclusion

This project did comprehensive analysis of drug efficacy and user satisfaction leveraging real-world patient feedback from online drug reviews. Through extensive data exploration, including the examination of various numerical variables and their relationships, key insights were gleaned. The hypothesis posited correlations between user engagement metrics, such as review frequency and usefulness, and perceived drug performance. Despite challenges in visualizing clear relationships between variables, the regression model constructed shed light on

significant predictors impacting average drug ratings. However, despite the significance levels (p-values from mainModel summary printout) indicating otherwise, I cannot conclude that a higher drug effectiveness rating and longer drug review text will correlate with a higher usefulCount. Regardless, I can still make some conclusions about the data.

The findings of the regression analysis revealed that the average usefulness count and average readability score significantly influence average drug ratings. This implies that drugs with higher perceived usefulness, as indicated by user reviews, tend to receive higher ratings. Additionally, drugs associated with more readable reviews, likely indicating clearer and more informative feedback, also tend to fare better in terms of ratings. These results underscore the importance of user engagement metrics and review quality in shaping perceptions of drug efficacy and satisfaction.

While the model explained only a modest proportion of the variance in average drug ratings, the statistically significant predictors identified offer valuable insights for healthcare decision-makers. By considering factors such as user engagement and review readability, healthcare providers and pharmaceutical stakeholders can better gauge and understand patient experiences and perceptions of drug efficacy. Ultimately, this enhanced understanding can inform more informed decision-making processes in healthcare, potentially leading to improved patient outcomes and satisfaction.