by Martin Sanchez

DRUG DATASET

CS-131- 61
Aug 7, 2024

The original dataset provides patient reviews on specific drugs alongside related conditions and a 10 star rating reflecting overall satisfaction.

OBJECTIVE:
Identify different methods to measure drug effectiveness using ratings scales

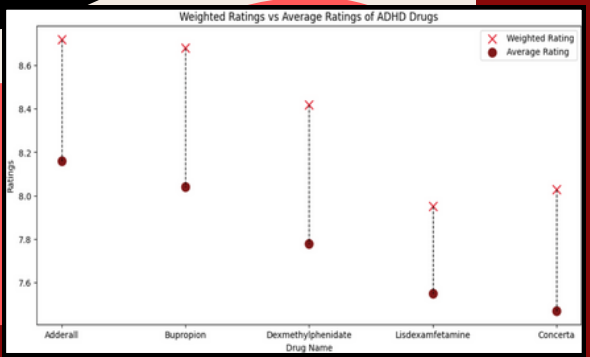
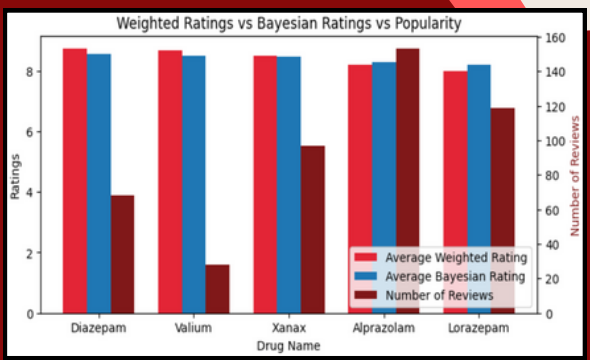
METADATA:

10 most common conditons

Birth Control *Depression*
Pain *Anxiety*
Bipolar D. *Acne*
Insomnia *Weight Loss*
Obesity *ADHD*

Variable Name	Role	Type	Description	Units	Missing Values
id	ID	Integer	Unique identifier for each review	-	No
drugName	Feature	Categorical	Name of the drug	-	No
condition	Feature	Categorical	Medical condition the drug is used for	-	No
rating	Feature	Numerical	User rating of the drug on a scale of 1-10	-	No
date	Feature	Date	Date the review was posted	-	No
usefulCount	Feature	Numerical	Number of users who found the review useful	-	No

SUMMARY:

By manipulating data, I found that weighted ratings offer a more meaningful measure of effectiveness compared to standard mean ratings, due to their product of average ratings and the useful count per review. I first applied these ratings to find the top 5 highest-rated ADHD drugs, where I unexpectedly found that the only non-stimulant drug, Bupropion, is a much more effective drug compared to the following 3 stimulants. Additionally, I found a third method of scaling drug effectiveness, by using the Bayesian formula, which favors drugs with a higher quantity of reviews.

Before these discoveries, I found the 10 most common conditions in my .csv dataset using the ***cut*** command and copied every review of these conditions into a separate .tsv file with ***awk***. Then I calculated the weighted and non-weighted ratings for each drug, with two separate shell scripts. In the ***wr.sh*** file, I specified a condition, multiplied each review's *rating* by its *usefulCount*, and divided the sum of these *weighted ratings* by the sum of all *usefulCounts*. For the ***ar.sh*** file, I simply averaged the ratings for each drug. I also found the Bayesian rating through a combination of the mean rating across all drugs for a condition, the number of reviews for each drug, and the weighted ratings. Finally, I visualized the Weighted Ratings vs Average Ratings on a scatter plot for ADHD in Google Colab. As well as a bar graph showing the comparison between Weighted and Bayesian ratings.