Data Scraping Amazon Reviews

David Geng
COS 482
12/19/24
Final Project

Introduction

My idea with the project was to scrape user reviews off of sites like Amazon to collect user ratings and reviews. Products sold online can easily face changes in quality especially over time where the company may not update or announce that the product's quality has changed. This could be due to multiple reasonings, like cost cutting, change in supplier, delivery issues, material change. I wanted to analyse the change in rating over time to possibly find if a drop occurred, if it was an item, brand, or distributor change, and what words were commonly associated with positive or negative reviews.

Methodology

- **Obtaining**

Data was obtained by using Beautifulsoup for data scraping, and Selenium for automation of scrolling, and next paging. The code for this is in FinalProject_COS482.py. I selected 3 items from 3 brands, and 1 from Amazon. This decision was primarily due to increasing Captcha Issues as it began to increase the amount required to answer each time, but otherwise I wanted to get a sample size from several brands to get a wider pool of data points. Amazon only had two sort by options, which were by highest ratings, and recency, along with limiting how many reviews it brings up at 100 for each product. I decided to scrape by recency as otherwise the reviews would be massively biased. I stored all the results into their own csvs for processing

- **Processing**

This mostly takes place in the FinalProect2_COS482.py file, some more processing occurs in Megafile.py. I cleaned and processed the data by importing and reading them. They were split into several variables for their Text, Month, Rating, and Word Count which required stripping and cleaning of extra data off of the data. Additionally the text of each review was further broken down to record how many instances of a word occurred in all of the reviews and the rating associated with them. The average rating of all the words was saved to Avg.csv it had columns for the word itself, the average of its ratings it was associated with, and how many times it occurred, and the rest of the processed data was inserted into a table on postgresql.

- **Visualizing**

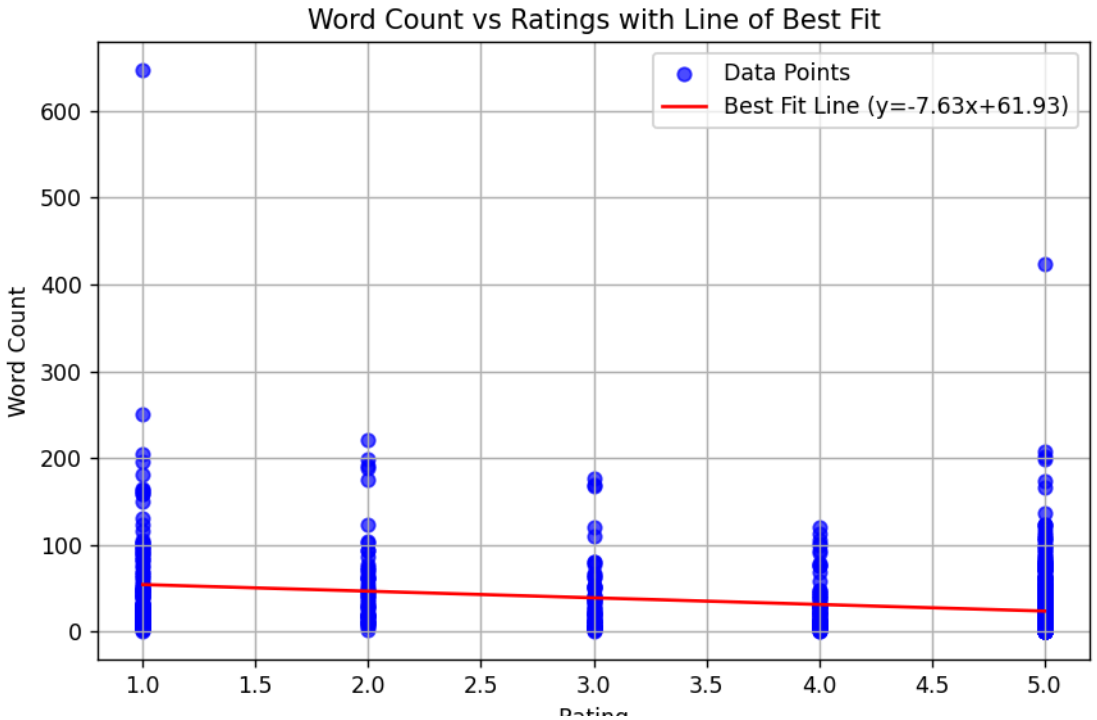
The code for this is in Plotstuff.py. For visualizing the data I used numpy, scipy, and matplotlib to create graphs and charts for the data. I created charts comparing the relation of Word Count and Review Rating, Average Rating by Month, Average Rating by Brand, Average Ratings over months for each of the brands, The words that had the highest and lowest ratings associated with them on average (That were longer than 3 letters, and occurred at least 40 times), and the results of the ANOVA test for difference between the rating of different months. Some of the data used for visualizing was from the SQL databases I initialized earlier and some were from the master list of reviews in the form of csvs (Avg.csv and CombinedReviews2.csv)

- **Analysis**

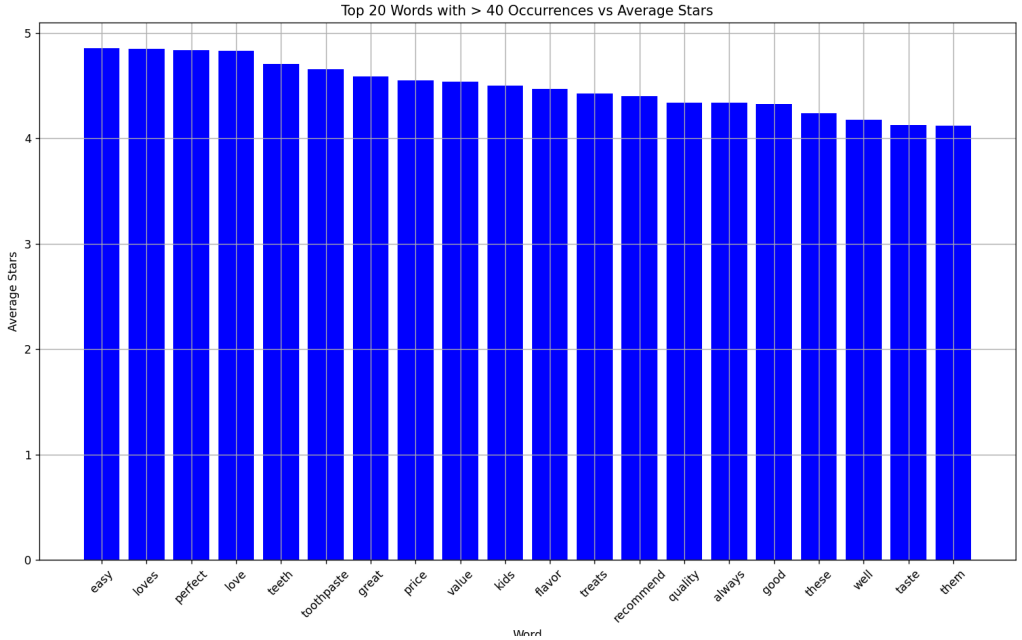
Analysis was done through an ANOVA test and lines of best fit, along with visual examination of the charts. Charts and test results are in the results section.

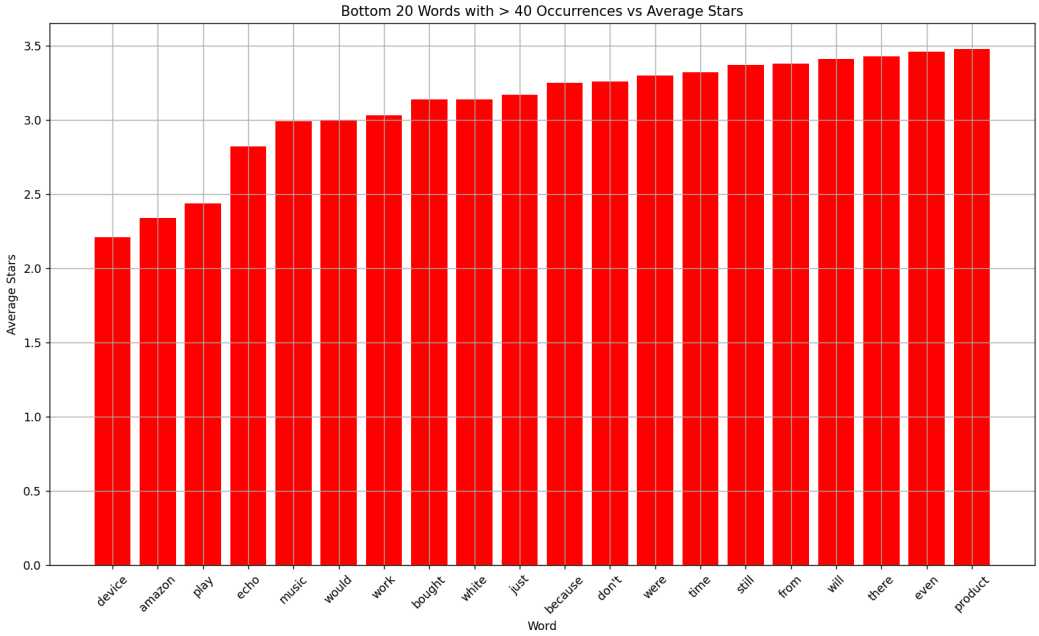
Results

This chart shows that as the rating increases the odds it is a negative review decrease and vice versa where the lower a review it is the more likely to be longer in length

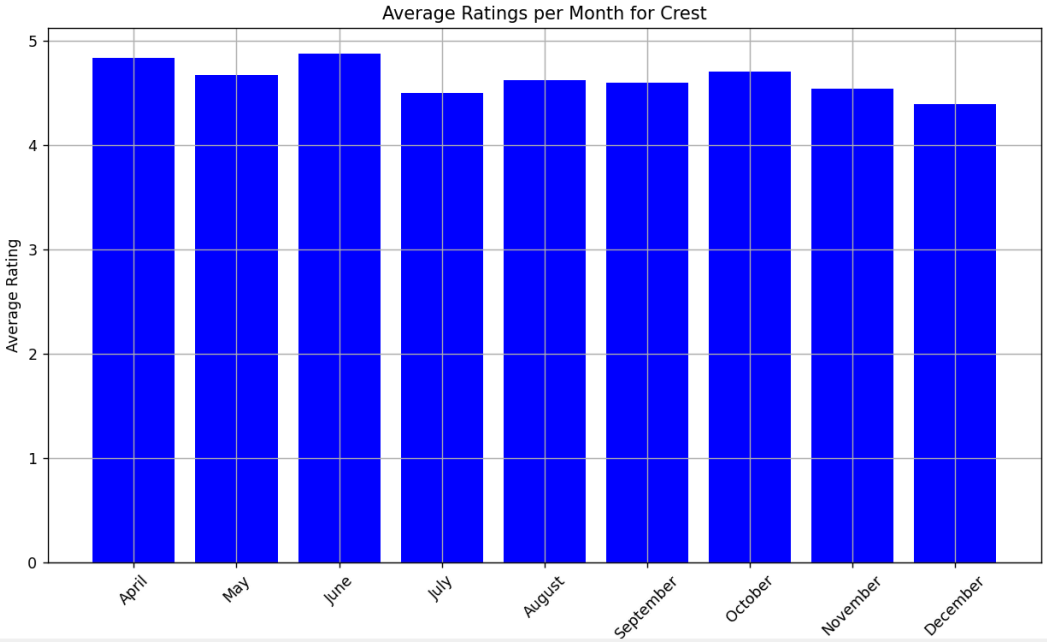


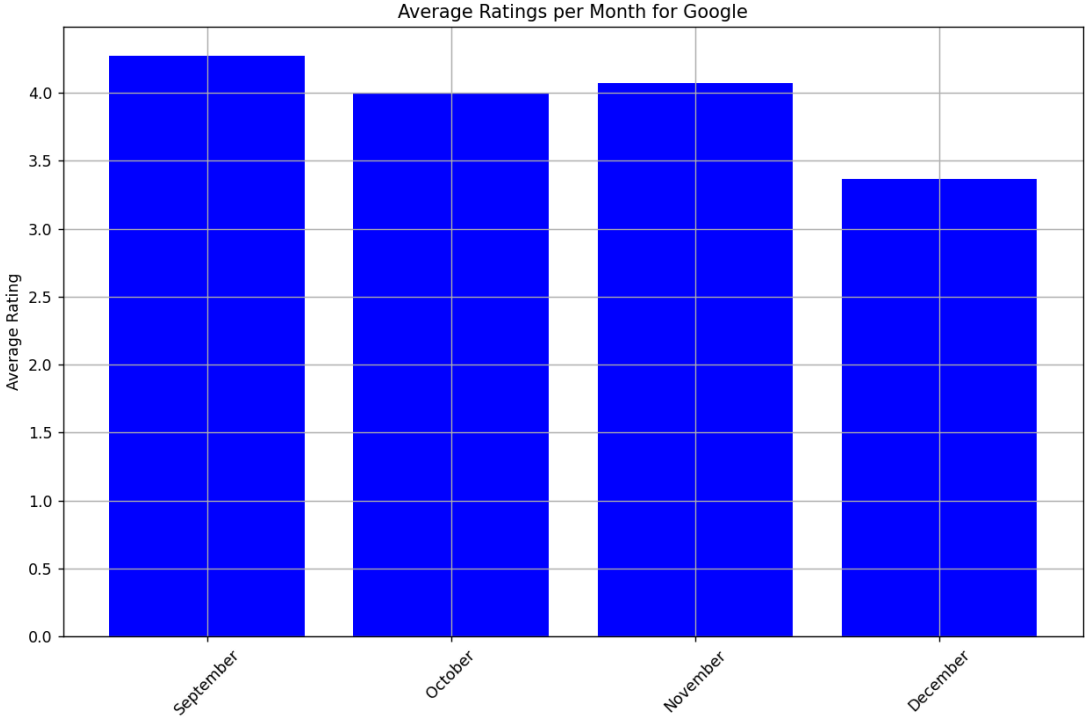
These charts show the words that were most associated with positive and negative reviews with the criteria of being at least 4 letters long and occurred more than 40 times to avoid prepositions and other words I was trying to avoid while ensuring that the words that did show up had a reasonable sample size.

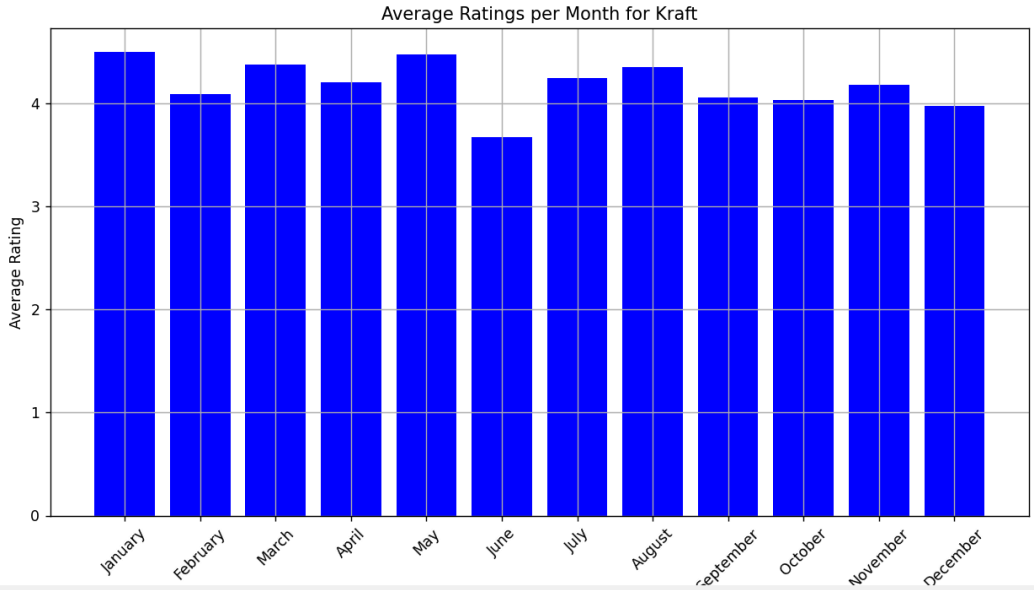


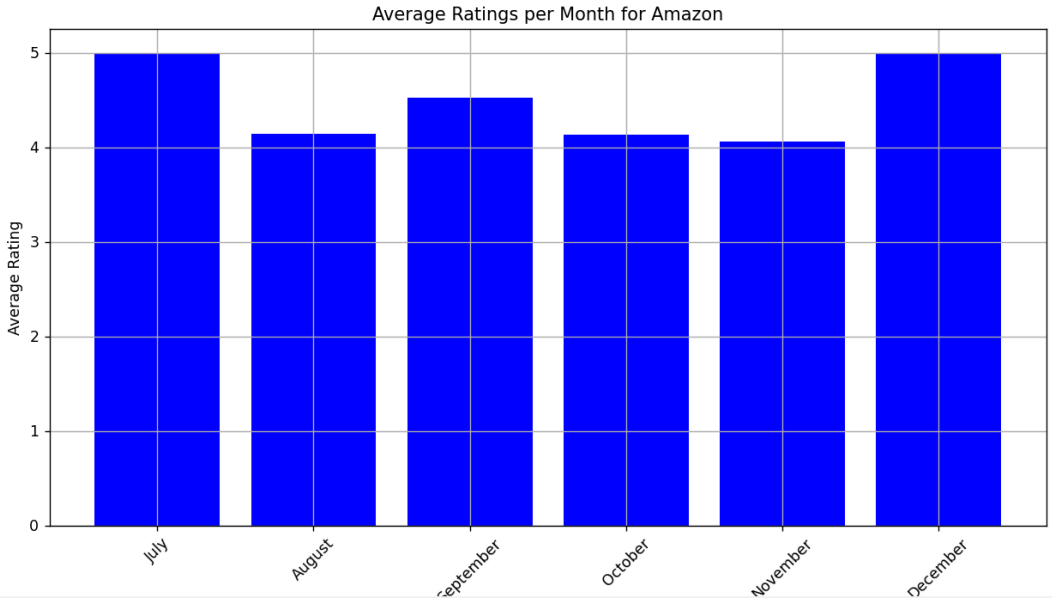


These graphs show the average ratings per month using the 3 sample products from each company. What months are included vary due to Amazon limiting the scraping to the most recent 100 reviews which unfortunately means some are more concentrated than others.

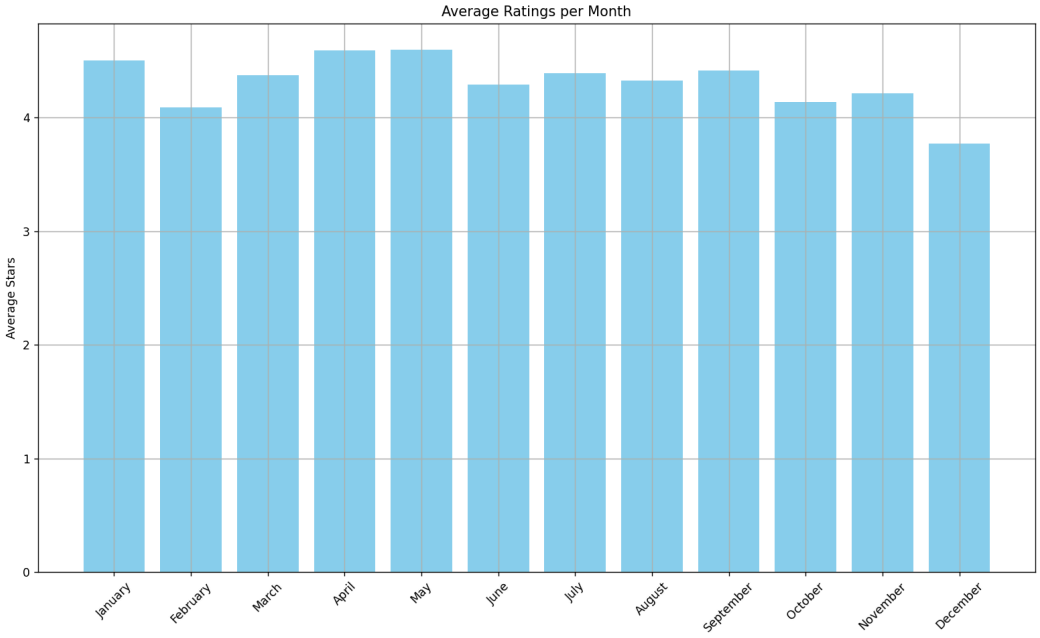


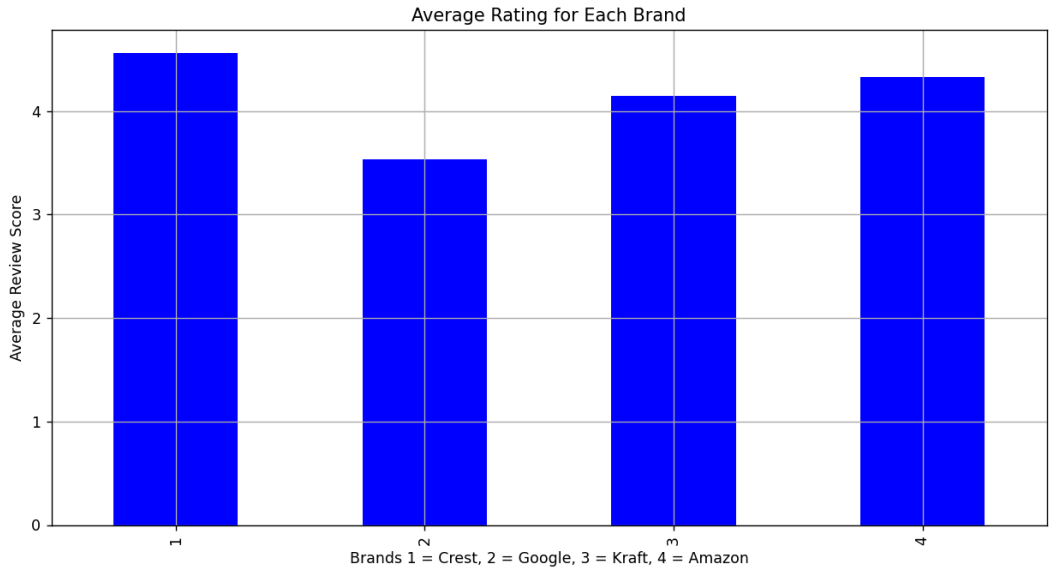



Average Ratings per Month for Kraft
Average Rating
0
1
2
3
4
January
February
March
April
May
June
July
August
September
October
November
December


Average Ratings per Month for Amazon
Average Rating
0
1
2
3
4
5
July
August
September
October
November
December

The next two graphs show the average of ratings of all reviews from all brands and products over the year, and the other graph shows the average rating from their respective products.





This image is for the results of the ANOVA test which compares the ratings between months showing the calculated F-statistic and P-value. Not sure if I totally implemented that correctly as that result seems to suggest that it is very significant, but I don't think the data backs that up?

```
F-statistic: 3.7654422465577713
P-value: 2.651192085075962e-05
There is a significant difference between the ratings of different months.
```

The results from all the data suggest several things, one is that a bad review is more likely to be longer in length and a good review is likely to be shorter in length on average, the words device, amazon, and play typically have a negative review associated with them, easy, loves, and perfect have a predictably positive review associated with them. The brands by month graph unfortunately have a small sample size from Amazon limiting reviews, but Kraft seems to have had an issue in June leading to a noticeable decrease in review averages, the rating average typically fluctuates throughout the year, but has a noticeable drop towards the end of the year, and that Crest seems to be declining in reviews over the course of the year and in comparison to the other brands sampled

Conclusion

In conclusion, while the exact reasoning for the change in reviews over time is somewhat unclear due to a limited data set it suggests that Crest has been decreasing product quality, and review averages over all have been decreasing over the year, possibly due to delivery being affected by weather/holiday season traffic or relating to Amazon as it does not seem brand specific, besides Crest.