An Analysis of Game Rants and Raves

[redacted] and Jolie Zhou

CSE 163

Title and authors

An Analysis of Game Rants and Raves by Di Mao and Jolie Zhou

Summary of questions and results

1. What keywords in game reviews are correlated with recommending or not recommending a game?
   Keywords associated with recommended reviews include “fun”, “good”, “play”, “get”, “like”, while keywords associated with not recommended reviews include “get”, “play”, “like”, “dont”, “even”. Many keywords are popular for both kinds of reviews, although they differ in popularity.

2. How do keywords in game reviews and ratings vary depending on the hours played at the time of review?
   As the number of hours played increases, the popular keywords used in the reviews decrease.

3. Is there a relationship between review sentiment and reviews users find helpful?
   There is no association between qualities of the review such as recommended or not recommended sentiment, hours played, and character length of the review and reviews users find helpful.

4. How consistent are game reviews across different game review platforms?
   Game reviews, and specifically their usage of popular keywords, are similar across different game review platforms.

Motivation

Video games have increasingly become more relevant in our entertainment-driven society. Especially with the pandemic, more people spend time playing video games. More and more games have been introduced to the market and found success. A crucial part of this success is achieved through sharing users’ reviews, with positive reviews further increasing a game’s popularity. We aim to understand the relationships in these reviews, specifically on Steam, a popular video game hub.

Dataset

The dataset we will use is a CSV on Kaggle, Steam Reviews Dataset | Kaggle. It contains reviews from Steam games that were best selling as of February 2019, information about when the review was posted, how many thought it was funny, how many thought it was helpful, the hours played at the time of review, and other features about the review. There are little under 380 thousand reviews in the dataset.

Due to time limitations for this project and processing time, we will be working with a smaller subset of the reviews. We will select a random sample of 20,000 reviews, and for the ability to reproduce the results we found, we arbitrarily set the random seed to 1000.

Another dataset we will use is also a CSV on Kaggle, Metacritic critic games reviews 2011-2019 | Kaggle, which contains about 124 thousand game reviews from Metacritic from 2011 to 2019. This includes information about the magazine the review was published in, the review text, game name, platform, review score, and review date

Method

Research question 1: What keywords in game reviews are correlated with recommending or not recommending a game?

1. Count up the number of recommended or not recommended reviews keywords appear in, filtering out English stop words such as “a”, “the”, etc.
2. Visualize the information by plotting a bar chart of recommended words and the proportion to represent the relevance and popularity of the word in reviews. Similarly, graph a bar chart for not recommended words.
3. For visual effect and a pleasing visual representation, create a word map with the size of the word representing its popularity.

Research question 2: How do keywords in game reviews and ratings vary depending on the hours played at the time of review?

1. Parse the Steam dataset by hours (0-100, 100-1000, 1000-3000 hours)
2. Find the top 10 most popular keywords in each parsed dataset.
3. Count usage of popular keywords in each interval of play time (20, 50, 500 hours intervals).
4. Plot a line graph of popular keyword usage vs. each time interval.

Research question 3: Is there a relationship between review sentiment and reviews users find helpful? What review sentiment is most common in reviews users find most helpful?

1. Exclude observations in which no users find them helpful.
2. Filter out outliers in the data (where z-score > 3) so graphs can be interpreted.
3. Graph histograms for helpful reviews for recommended and not recommended (bin width = 10).
4. The two variables of interest in relation to reviews users find most helpful are hours played, character length of the review, and recommended or not recommended.
    a. For the first variable of hours played, graph a scatter plot of how many users find a review helpful against the hours played at the time of review, by if it's recommended or not.
    b. For the second variable of the length of the review, graph a scatter plot of helpfulness over character length, by if it's recommended or not.
5. Develop a machine learning model on features (hours played, character length, recommended or not) to predict labels (how many users will find the review helpful).

Research question 4: How consistent are game reviews across different game review platforms?

1. Parse the Steam dataset, find the top 20 most popular keywords and their respective usage proportion in all reviews.
2. Parse the Metacritic dataset, find the top 20 most popular keywords and their respective usage proportion in all reviews.
3. Graph a bar chart of most popular keywords from least to most popular for both Steam and Metacritic.
4. Determine keywords that are in the top 20 keywords of both Steam and Metacritic.
5. Graph a bar chart of shared keywords vs. proportion with Steam and Metacritic as different data.
6. Determine the usage proportion of popular keywords in the entire dataset.
7. Graph a pie chart of usage proportion.

Results

Research Question 1

To find the most popular keywords for the kinds of reviews, we plotted bar charts representing the counts of keywords in the reviews.

We found that the most popular keywords used in reviews that recommended a game are “game”, “fun”, “good”, “play”, “get”, “like”, “great”, “really”, and others, listed from most to least popular. Below is a graph showing the counts of each keyword in all the reviews that recommended a game:

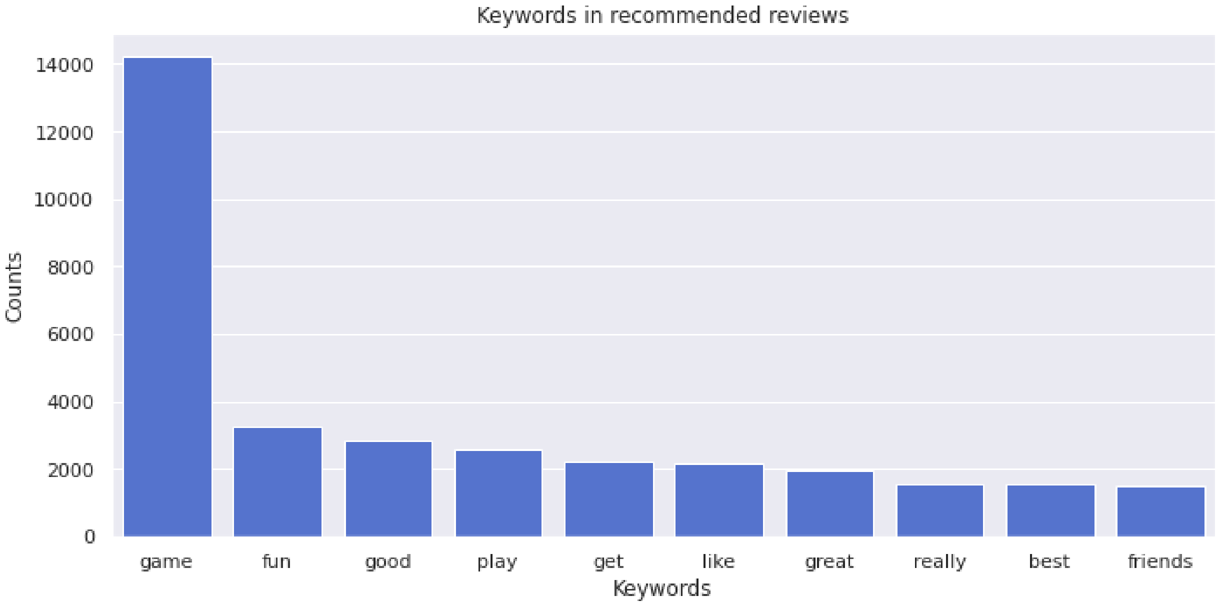


Figure 1. Bar chart of keywords for recommended reviews.

We found that for reviews that did not recommend a game, similar keywords were popular, although in a different order. These keywords listed from most to least popular include “game”, “get”, “play”, “like”, “dont”, “money”, “even”, and others. The following figure shows the counts of the keywords:

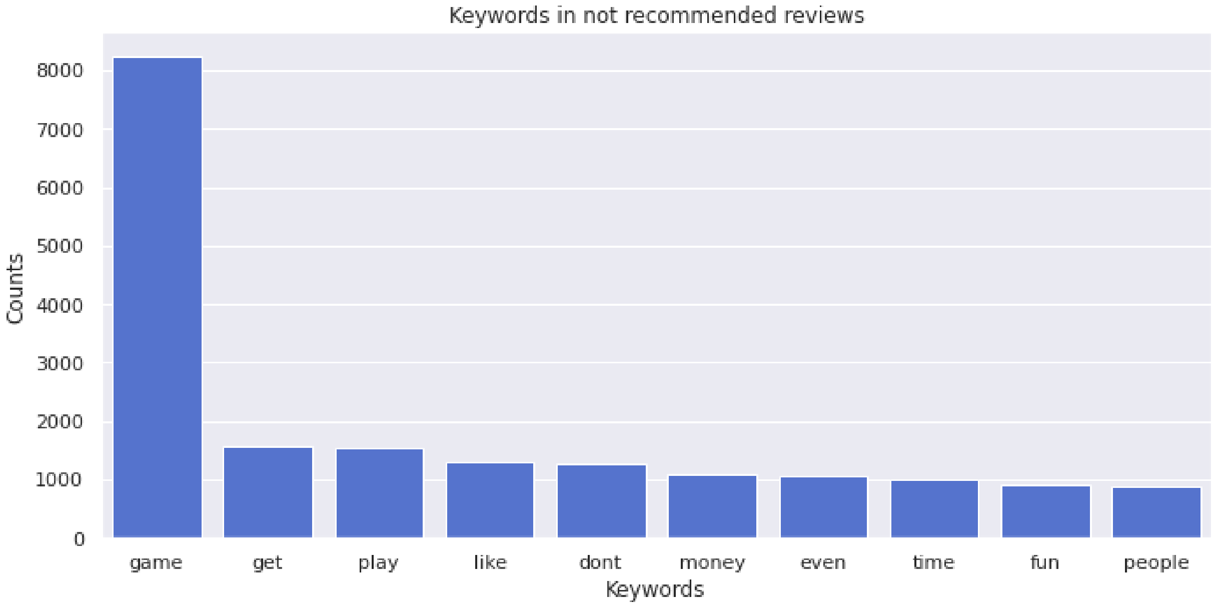


Figure 2. Bar chart of keywords for not recommended reviews

It’s interesting how similar keywords are popular for both kinds of reviews (recommended and not recommended). This may be because reviews that do not recommend a game pair together keywords with words that convey an opposing feeling, such as “don’t”. That explains why similar keywords are popular for both and also why “don’t” is a unique keyword that is popular in not recommended reviews. People probably put these words with phrases, e.g. saying “It was fun” compared to “It was not fun”.

More of the keywords for not recommended reviews also refer to money, with “money”, “time”, and “buy”, suggesting more of a focus on whether or not a game is worth its monetary value. On the other hand, more of the keywords for recommended reviews are superlatives, with “good”, “great”, and “best”, suggesting more of a focus on the quality of the game. This suggests a difference in the values reviewers have.

Some topics of further research would be analyzing the popularity of particular phrases in these reviews, to further explore why keywords can be popular for both kinds of reviews.

And for fun, we created word clouds of the keywords!

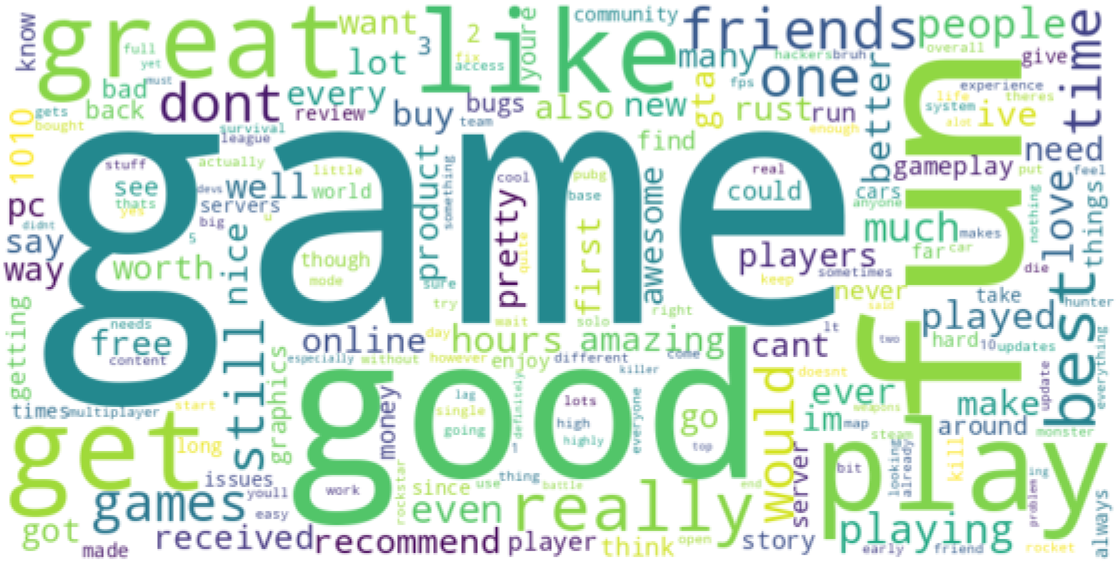


Figure 3. Word cloud of popular keywords in recommended reviews

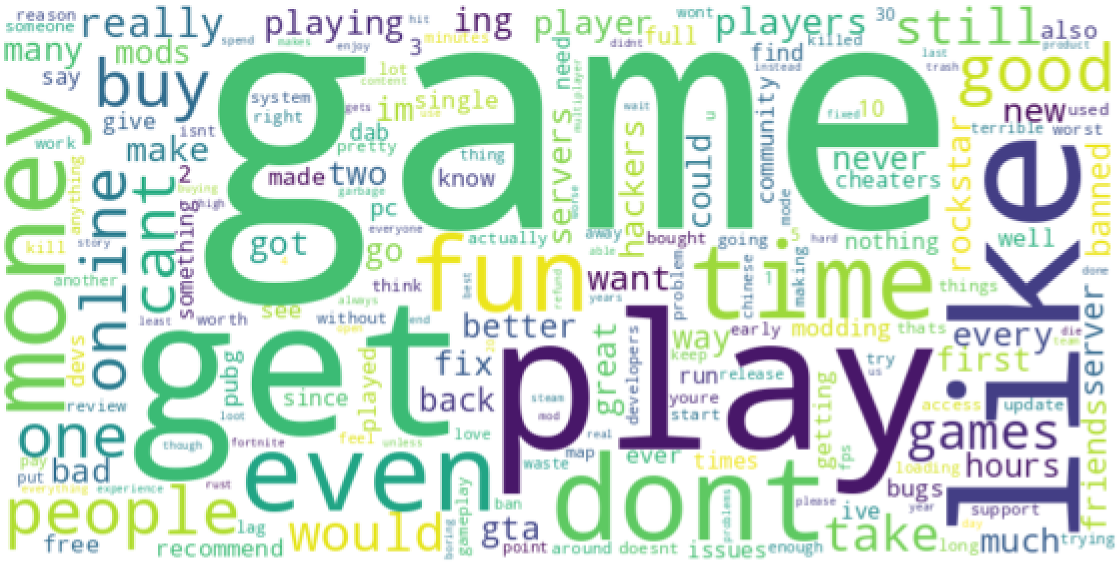


Figure 4. Word cloud of keywords for not recommended reviews

Research Question 2

While overviewing our Steam dataset, we recognized that there was a ‘hour_played’ column. Because of this, we decided to explore whether the number of hours played on a game impacts the popular keywords used. We ask, how do keywords in game reviews and ratings vary depending on the hours played at the time of review?

To approach this question, we planned to split the dataset into hour intervals, say 20 hours each. One quick issue we found with this approach was that while we increased the number of hours played, this hour interval should also increase. For instance, for hours played under 100 hours, the hour interval used for analysis would be 20 hours. For hours 100 to 1000, we set this interval arbitrarily to 50, for hours 1000 to 3000 (about the highest amount of hours played), we set this interval arbitrarily to 500. To accomplish this, we split the dataset into three smaller datasets filtered by hours played.

For each smaller dataset, we first find the top 10 most popular keywords. Then, we found the count of each popular keyword in the smaller datasets. Finally, we plot these counts by hour interval in a histogram, with each column being a different popular keyword.

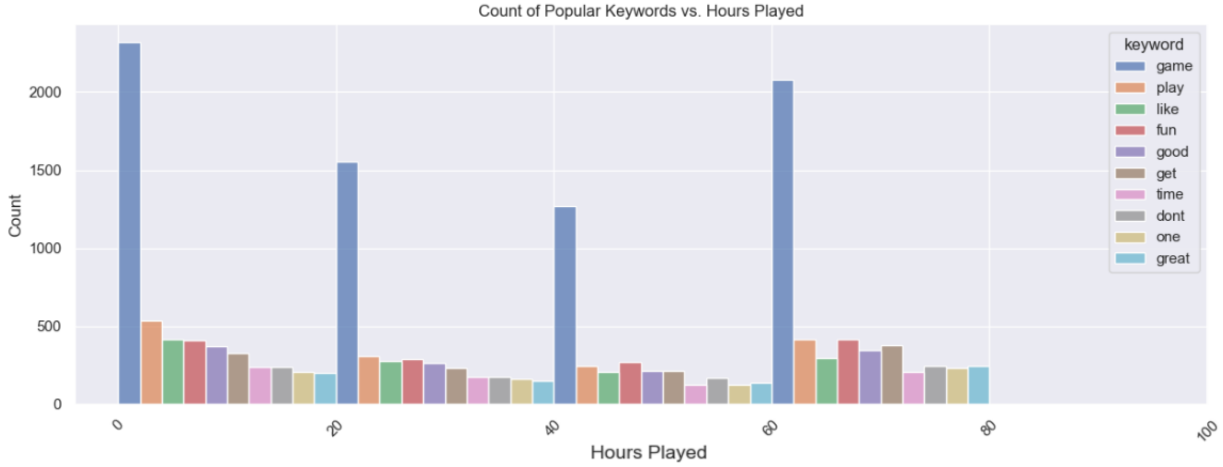


Figure 5. Histogram of popular keyword counts for 0-100 hours played

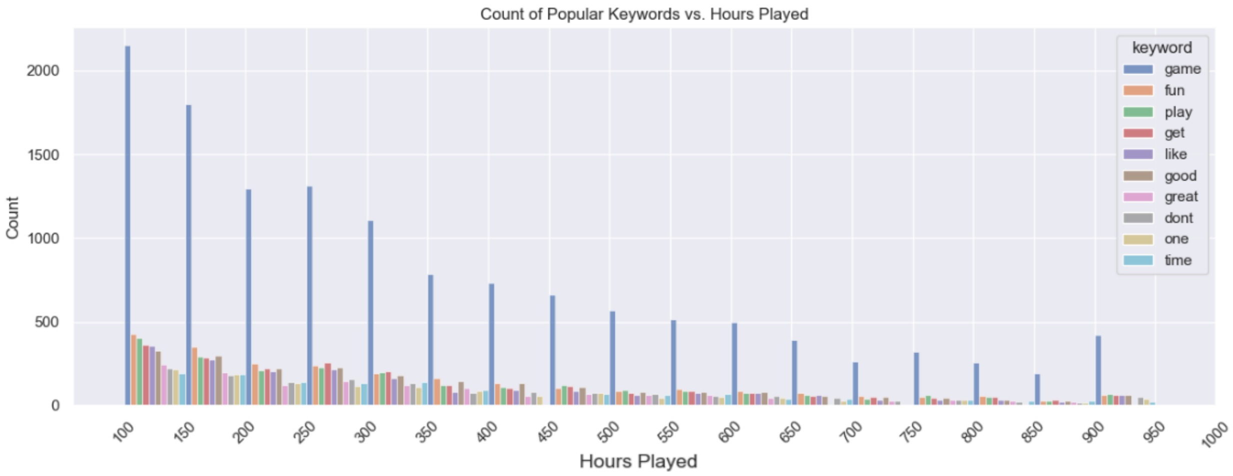


Figure 6. Histogram of popular keyword counts for 100-1000 hours played

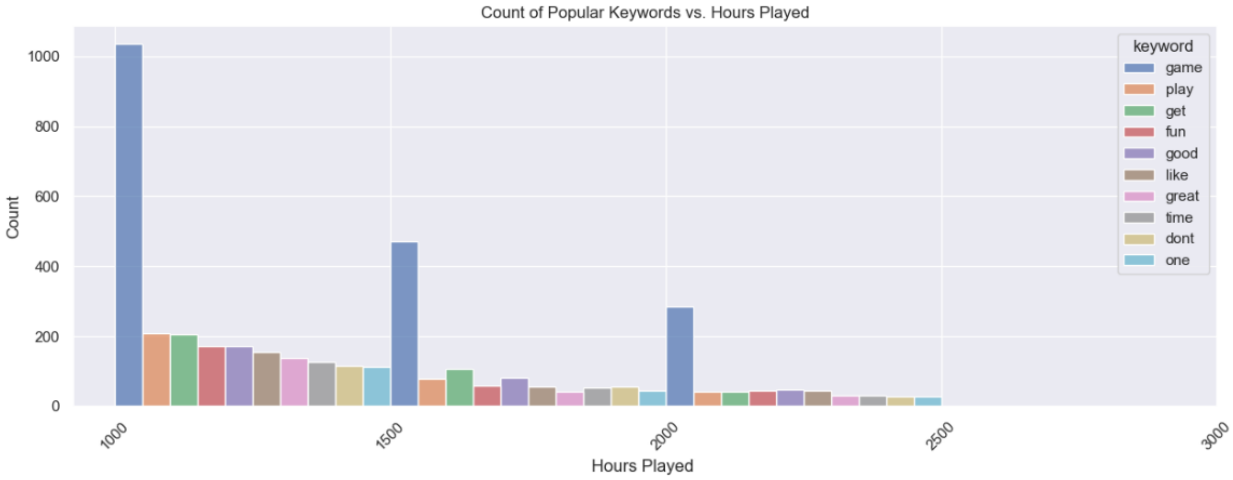


Figure 7. Histogram of popular keyword counts for 1000-10000 hours played

By looking over these graphs, we recognize that the most popular keyword across all time intervals is, by a large amount, ‘game’. This is unsurprising, especially when discussing in context with Research Question 1, which also found ‘game’ to be a popular keyword over the reviews. Furthermore, the last interval of each dataset is missing, and this is likely due to limitations in our code.

For the graph displaying the count of keywords from 0-100 hours played, we can determine that the counts of keywords did not change much. However, where it did, the entire keywords seemed to follow this trend. For instance, in the interval 40-60, there is a relatively big dip in counts, as shown most clearly by the counts for the ‘game’ keyword. Similarly, the other popular keywords also have a dip in the count. One surprising finding is that no popular keywords were found in the interval of 80-100 hours. This could be because of some unique reviews not containing any popular keywords.

For the graph displaying keyword counts from 100-1000 hours played, we see an apparent decrease in keywords as the number of hours played increases. We see the same downward count trend for 1000-3000 hours played. Because of this finding, we could make some guesses as to why this may be. For one, the reviews could have become shorter as the number of hours played increased. This would be surprising because one would expect reviews to become longer as the number of playing hours increased. We would expect players who spend a long time on the game to leave longer reviews. Another reason this finding may be the case is that the reviews could have been more detailed. Specifically, they make less use of popular keywords. On one hand, this would make sense using similar logic as before. We may expect players who spent longer on a game to leave more complex reviews. However, from another perspective, this is a surprising finding. No evidence shows that players who spend longer on a game would use words different from general popular keywords.

Overall, our analysis shows that popular keywords are used less in reviews as the number of hours played by the reviewer increases.

Research Question 3

To better understand the distribution of recommended and not recommended reviews, we first plotted histograms of the distribution of helpful reviews for each category. We removed outliers (using the calculation that a z-score > 3 is considered an outlier) and set the binwidth to 10.

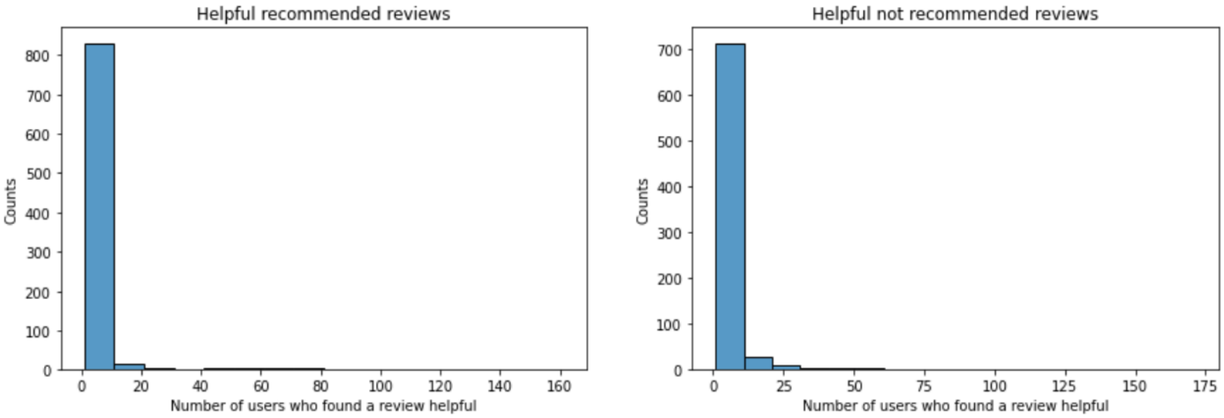


Figure 8. On the left, histogram of recommended reviews, on the right, not recommended

In the sample selected, it appears there are more recommended reviews than not recommended reviews. The reviews with z-score < 3 are distributed unimodal and with a slight right skew. Most reviews only have about one to ten users who found the review helpful. The other bins are mostly undetectable on the histogram because of the great difference in counts between the bins.

Then we plotted scatterplots of how helpful a review was over the hours played at time of review and over the character length of the review. There are few reviews with over 500 users who found the review helpful.

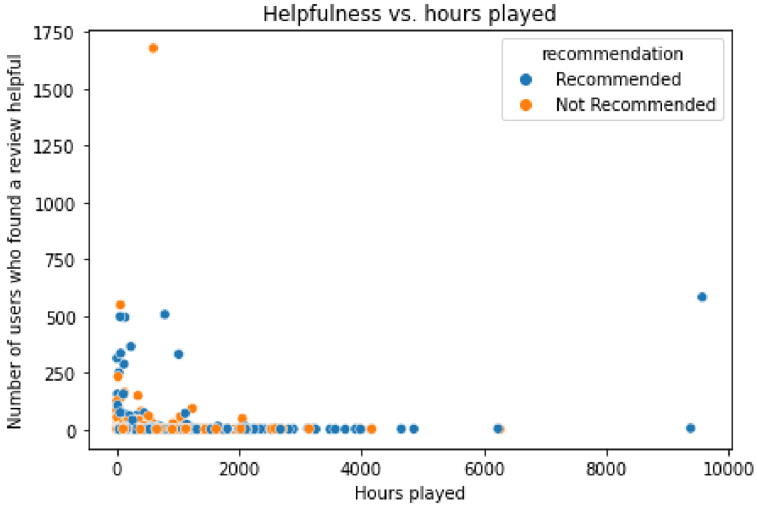


Figure 9. Scatter plot of the number of users who found a review helpful vs. hours played at the time of review

There is no linear correlation between the number of users who found a review helpful and the hours played at the time of review. Interestingly, at least for this sample, more people found some recommended reviews helpful compared to not recommended reviews.

Interestingly, there are also two reviews in which the reviewer had over 8000 hours played at the time of review, which is almost a year's worth of hours. This seems unrealistic so these are likely outliers. It's very interesting how the reviewer with about 10000 hours played had a review that many users found helpful.

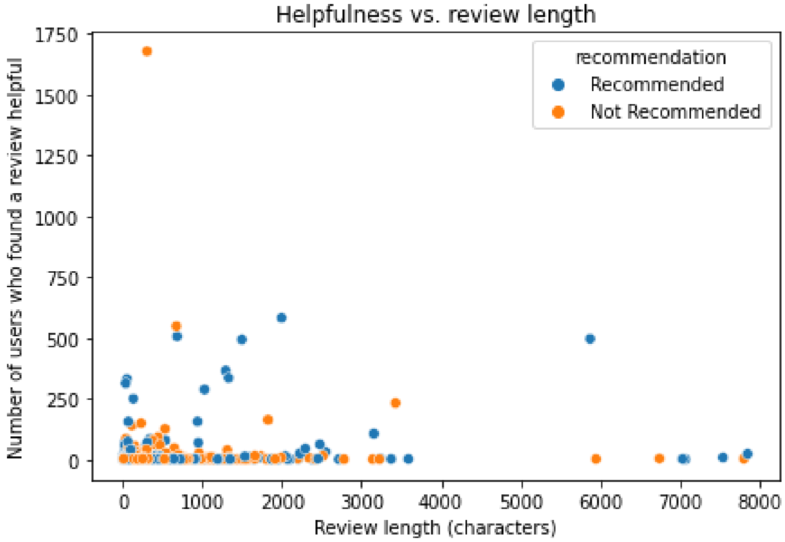


Figure 10. Scatter plot of the number of users who found a review helpful vs. hours play at the time of review

There is no linear correlation between the number of users who found a review helpful and the character length of the review. although there appears to be some weak association with increased review length and increased helpfulness for recommended reviews with character lengths between 1000 and 2000 characters. This is likely an interesting quality of the sample selected and not indicative of any significant insight between the two variables.

We also created a machine learning decision tree regressor model on the data, training on features of 'funny', 'hour_played', 'recommendation', and 'review_length', using these features to predict the label 'helpful'. The mean squared error on the training data was about 0.001523 and about 1337.021277 on the testing data, which is quite high. The machine learning model likely has such a high error because there is no correlation between the variables of interest, as demonstrated by the scatter plots above. Additionally, there are more recommended reviews than not recommended reviews in the sample, so this may impact the decisions made by the model.

Research Question 4

While analyzing the reviews left on Steam, we grew increasingly curious about reviews outside of Steam and from other game review platforms. From previous knowledge, we knew of such a website by name, Metacritic. Metacritic is a website that compiles reviews of various media from professionals such as top critics and from media such as publications. So, we decided to perform an analysis on Metacritic reviews and compare it to an analysis of Steam reviews. We ask, how consistent are game reviews across different game review platforms?

First, we analyzed our Steam reviews dataset with an approach similar to what has been used in previous Research Questions. We found the top 20 most popular keywords in the dataset and counted how often they occurred. Then, we graphed these counts as a bar graph with decreasing keyword count.

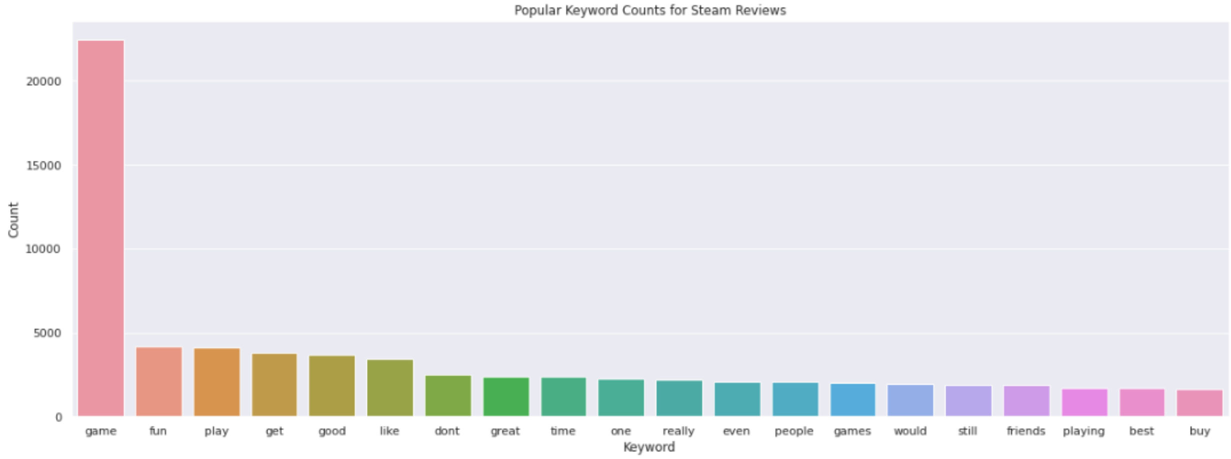


Figure 11. Bar chart for top 20 most popular keywords and their counts in Steam reviews

Next, we performed the same analysis for the Metacritic game reviews.

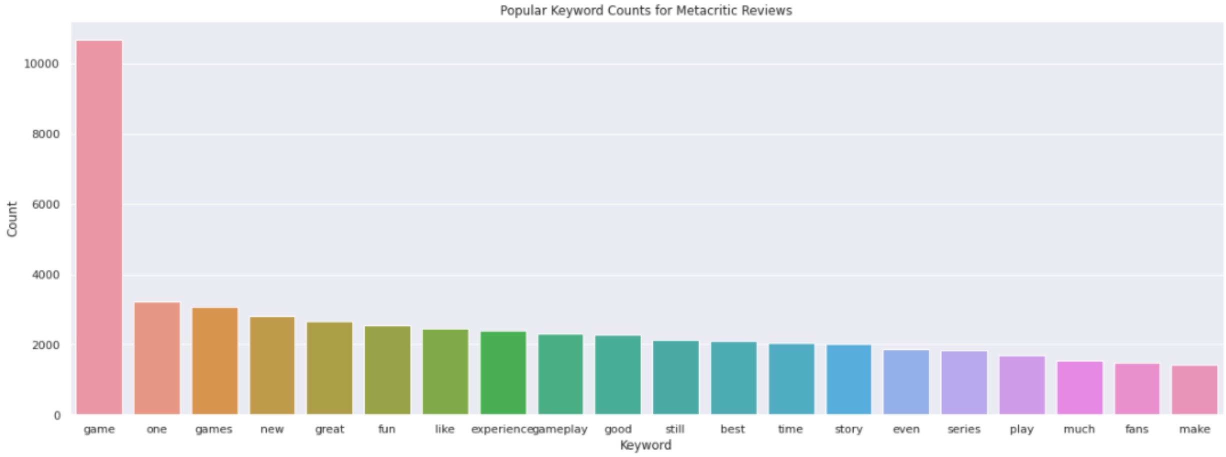


Figure 12. Bar chart for top 20 most popular keywords and their counts in Metacritic reviews

From briefly looking at these two graphs, we already see that the keyword ‘game’ appears frequently in both platforms. Other keywords, such as “fun”, and “good” are also shared as popular keywords. Because overlapping popular keywords exist, we decided that one way we would compare the datasets is to graph the respective usage proportion of the shared popular keywords in each dataset.

To begin this comparison, we first found the shared keywords by using the lists of top 20 keywords we found for each review platform before. Next, we calculated the usage proportion, or ratio, of these keywords in relation to the entire dataset. This means, dividing the count of the popular keyword in the review platform by the total keyword count. Finally, we graphed a bar chart representing this comparison.

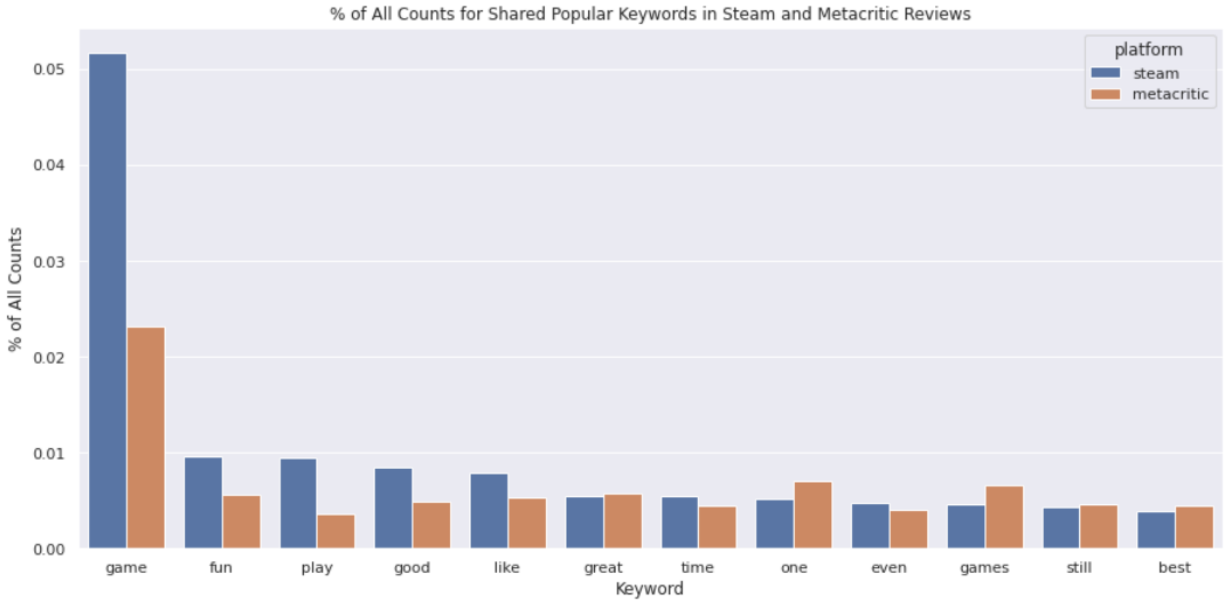


Figure 13. Bar chart for shared popular keywords between Steam and Metacritic reviews and their ratios

By comparing the Steam reviews dataset with the Metacritic reviews dataset, we found that there are lots of popular words shared in both review platforms. Out of the top 20 keywords in both game review platforms, 12 of the keywords, or 60% of the popular keywords, are shared as popular keywords. This is surprising because we thought Metacritic, written by professionals, would have more formal words. Therefore, the two review platforms would not have many shared keywords. However, our data analysis shows otherwise. Another unexpected result of this analysis is how little each popular keyword takes up the entire dataset. The largest percentage is 5% for the keyword ‘game’ in Steam reviews.

We wanted to further understand this surprising result regarding the usage of shared popular keywords in the two different datasets. The previous analysis showed that a large percentage of popular keywords in the different game review platforms are shared. So, we wanted to analyze how much these popular keywords are used in the entire review data. To do so, we calculated the total counts of the top 20 keywords for both game review platforms and divided them by the total counts of all keywords. We then graphed these findings in pie charts.

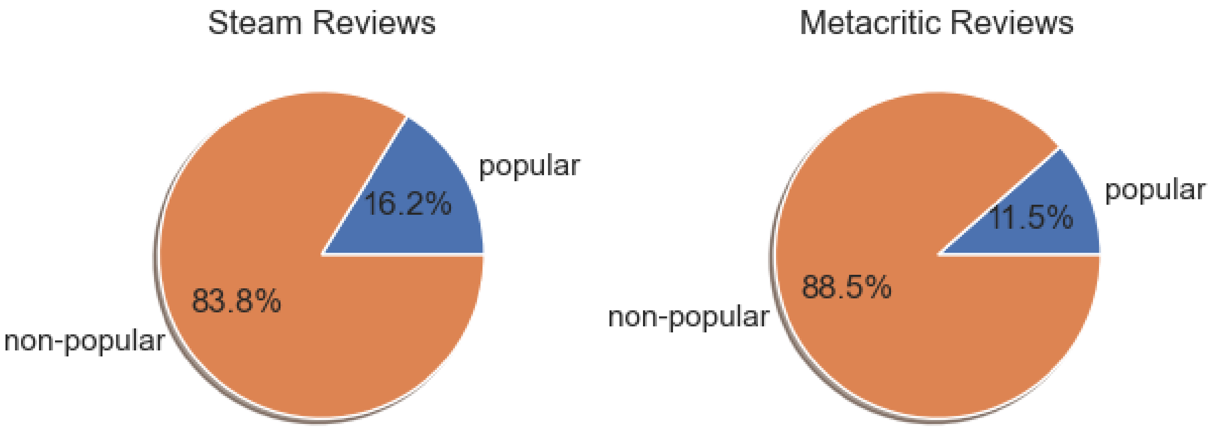


Figure 14. Pie charts for popular keywords usage ratio in Steam and Metacritic reviews

Unsurprisingly, the Metacritic game reviews had a lower proportion of popular keywords (11.5%) in comparison to Steam game reviews (16.2%). Since Metacritic is considered to be written by professionals while Steam game reviews are written by the general public, this is expected. One would expect the general user to use more popular keywords than top critics. Furthermore, similarly to before, we find that the usage of popular keywords is quite small, with both usage percentages found to be less than 20%. One would think that popular keywords would take a larger portion of the reviews.

Overall, we found that, surprisingly, popular keywords are about the same across platforms targeting different reviews, and that popular keywords only take up a small portion of keywords of all reviews.

Impact and Limitations

Our analyses of the data were limited by the qualities of the population. Since we were working with game reviews on Steam, limitations of the review format impacted how significant our results were. For example, users can type free text in their interviews so some users write reviews with words that convey sentiment opposite to their actual recommendation on the game. To a person, the review can be funny but to a computer, the inconsistent information can be confusing to draw conclusions. There is a lot of noise in our data because of how diverse the format of the data is.

Another limitation also relates to the free text quality of reviews. Users can type their reviews using casual English or even a different language. Our analysis only considers English keywords and filters out properly-spelled English stopwords, so we are possibly implicitly excluding data with different spellings of words, slang for words, foreign keywords, etc. People who'd benefit from our analysis would be users typing full and proper English words while those who are excluded from our analysis would be users who do not follow proper English conventions.

Additionally, in particular with Research Question 3, many review features such as "funny" and "helpful" are uncommonly used, skewing the data to few reviews with many users who found it "funny" or "helpful", and a vast majority of reviews with one or no users who found it "funny" or "helpful". This significantly impacted results particularly with Research Question 3, and some features are not as commonly used and so the machine learning model developed had a large error.

A group that can benefit from this analysis could be companies. They can understand the popular keywords in reviews and change their marketing strategy and language to match the popular language used by reviewers. The analysis helps us understand what words people like to use, but can also impact decisions companies make to sell and advertise their games to customers.

Groups that can be excluded from this analysis include gamers that do not leave reviews or use other platforms besides Steam to share opinions about games. Although Steam is a widely used and popular platform for games, not all users will use the features on Steam equally and may turn to other platforms such as Reddit to discuss. Connecting back with companies who can benefit from the analysis, this missing source of data would bias the interpretations and decisions they make towards users of Steam rather than other platforms.

Another concern to take into consideration in our analysis is the idea of user consent and privacy of information. To make a review on Steam, users have to create a Steam account. Users can only write reviews for games on their Steam account, and when writing the review their playtime will also show up with the review. In writing these reviews users are implicitly consenting to this release of information, but as we analyze this data and observe trends in language, users are not clearly communicated about the use of their review data.

Challenge Goals

Multiple Datasets

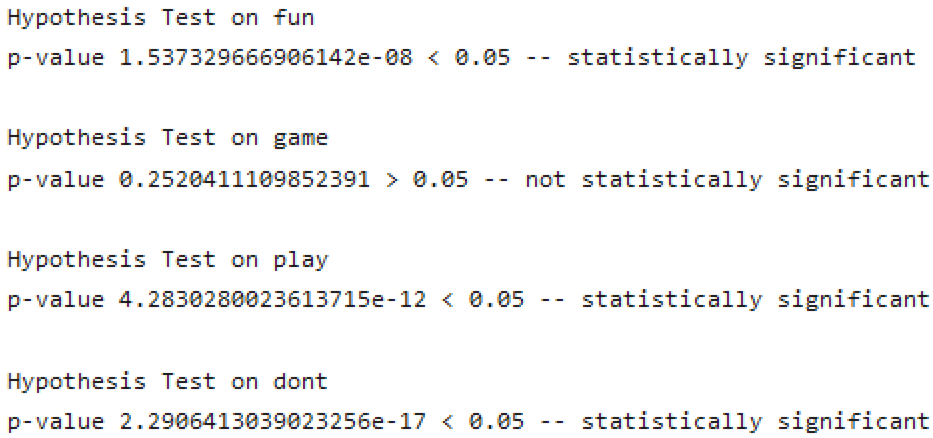
In Research Question 4, we analyzed the consistency of game reviews across platforms. To do so, in addition to the Steam reviews dataset, we introduced another dataset to analyze, namely the Metacritic reviews dataset. We chose the Metacritic reviews dataset because of the people it targets to write reviews. While reviewers on Steam are general users, reviewers on Metacritic are professionals. Because of this difference of users, by combining these two datasets in our analysis, we were able to make conclusions in our analysis of game reviews.

Result Validity

In our proposal, we imagined that there would be an association between the number of people who find a review helpful and the number of hours the reviewer has played at the time of review. We planned to verify this linear association, but after analyzing the scatter plot of the data, it is clear there is no association between helpfulness and the number of hours played.

Instead, we changed our focus to verify the results of popular keywords for recommended compared to not recommended reviews. As an extension to our analysis for Research Question 1, we explored whether or not the proportion of keywords used in recommended reviews was statistically significant and different from that of not recommended reviews. This extension would verify and corroborate our findings in Research Question 1 about why particular keywords were more or less commonly used between the two kinds of reviews.

We conducted a hypothesis test on the difference of proportions for the following keywords: “fun”, “game”, “play”, and “dont” (sic). We used a significance level of 0.05, setting the null hypothesis that the proportions of reviews using the keyword for recommended and not recommended reviews are equal, and the alternative hypothesis is that they are not equal (two-tailed hypothesis test). We found the following:



```
Hypothesis Test on fun
p-value 1.537329666906142e-08 < 0.05 -- statistically significant

Hypothesis Test on game
p-value 0.2520411109852391 > 0.05 -- not statistically significant

Hypothesis Test on play
p-value 4.2830280023613715e-12 < 0.05 -- statistically significant

Hypothesis Test on dont
p-value 2.2906413039023256e-17 < 0.05 -- statistically significant
```

Similar to our findings and analysis in Research Question 1, for keywords “fun”, “play”, and “dont” (sic), the p-value is less than 0.05, so we reject the null hypothesis. That means the difference is statistically significant and there is a difference in the proportion of the use of the keyword between recommended and not recommended reviews. This makes sense because the ordering of these keywords in the bar charts for recommended and not recommended reviews were different, indicating that their level of popularity and usage was different between the two kinds of reviews.

Meanwhile, for “game”, the p-value is greater than 0.05, so we do not reject the null hypothesis at the 5% significance level. The difference is not statistically significant, suggesting that the proportion of reviews using the word “game” is equal for both kinds of reviews. This also makes sense since “game” was the keyword most popular for both kinds of reviews, so there was no significant difference in its popularity for recommended and not recommended reviews.

Work Plan Evaluation

Task 1: Download the datasets and set up the coding environment

- Estimate: 1 day
- Reality: 1 hour

Task 2: Analyze and graph our data and create the ML model

- Estimate: 4 days
- Reality: Over a week (~15 hours combined)

Task 3: Analyze visualizations, answer research questions, and determine the accuracy, result validity and significance of our results.

- Estimate: 4 days
- Reality: 3 Days (~3 hours combined)

Task 4: Write the report

- Estimate: N/A (was not on our work plan in the proposal)
- Reality: Over 3 days (~10 hours combined)

Overall, we spent less time on task 1 than we originally thought, more on task 2 than we originally thought, and about the same for task 3 than we originally thought. We think that our estimates were both close and far from reality because of our varied experience level with the different aspects of the final project. Determining and setting up the coding environment took less time than we anticipated because we had a good idea of what we wanted our environment to provide (ex. Live coding collaboration, fast and flexible production of visualizations). However, coding and creating visualizations took more time than we expected because we ran into many errors, bugs, and issues we did know how to code. So, we had to spend more time on research on how to fix these issues. Determining the accuracy of our results took about how much time we expected, and this is likely because the process behind determining these values are generally the same and only the result differs. Finally, in addition to the methods explained in the report, we completed another task during this project, which was writing the report that took place over three days.

Testing

While coding our project, we created test datasets. These test datasets were smaller versions of our larger datasets and were modified to test each research question. For instance, there were two test datasets for Research Question 1. One included the first 50 rows with recommended reviews, and the second included the first 50 rows with recommended reviews combined with the first 50 rows with non recommended reviews. Another example is Research Question 4, where there are also two test datasets. We used the same dataset with 100 rows (with both recommended and not recommended reviews) to test the Steam dataset and another with the first 100 rows of the Metacritic review dataset.

We first used these smaller datasets to test that our code produced output that seemed reasonable and that we could expect. Before scaling to our final code for the analysis, we also ran our code against small samples (unseeded random) to observe that the results were similar to previous runs of our code. This code is not included in our final files because we scaled up to our final larger sample of 20000 reviews and used an arbitrarily seeded random so we can reproduce our results.

Furthermore, we determined that the results from our tests matched the results using the final datasets. Because of these matches, our results are likely accurate.

Additionally, for the hypothesis test used for result validity, we used assert statements to confirm that large sample estimation and claim requirements were met so that our results from the test can be considered valid.

Collaboration

We would like to thank our mentor, Suh Young, for providing us with specific and helpful feedback throughout our project creation :).

While creating this project, we consulted numerous online resources, including our weekly Learning Reflections, StackOverflow, GeeksForGeeks, and the specification for Python libraries, including Seaborn and TextBlob.