*Amazon Food*

*Product Reviews Analysis*

**BANA 277 - Final Project Report**

**Group 2A:** Shefali Kumar, Martin Gui, Changyeol Lee, Vincentia Angelica, Kirti Bhalgat

**Table of Contents**

[**Summary**](#_mytuzpofilbe) **3**

[**Objectives**](#_6ehudotmyrdl) **3**

[**Data Preparation**](#_geybivbha6k6) **3**

[3.1 Data Description](#_qm21w8ldl46k) 3

[3.2. Data cleaning](#_k5x2fe6avetw) 4

[**Exploratory Data Analysis**](#_yd2qyrn0a9ec) **4**

[4.1. Distribution of review count](#_ayzog24am2cw) 5

[4.2 Trend in the number of reviews between 1999 and 2012](#_ad37hahcvagv) 6

[4.3 Top Reviewers analysis](#_c0ssuhxdu8sx) 6

[4.5 Textual Analysis](#_d2ghpiulaypr) 8

[**Sentiment Analysis**](#_ahi83tmpry9j) **9**

[**Logistic Regression**](#_e6gbddtfjlex) **13**

[**Interpretation and Key Takeaways**](#_eckhf71esjcn) **15**

[**Citations**](#_mhlx4jsazno5) **16**

# Summary

Amazon is one of the largest technology companies and the world’s largest online marketplace that sells books, movies, foods, household products and various other goods. Reviews play an important role in the buying decision of customers. It has become increasingly important to analyze customer behaviour through these reviews.

# Objectives

The objective of this project is to analyze reviews given on Amazon food products between October 1999 and October 2012 in order to observe/answer the following questions:

* Do we see any trends in the review count over the years?
* Are reviews from top reviewers more helpful?
* Conduct sentiment analysis.
* Conduct logistic regression to identify which features are important for a review to be helpful.
* Study the relationship between polarity and Subjectivity.

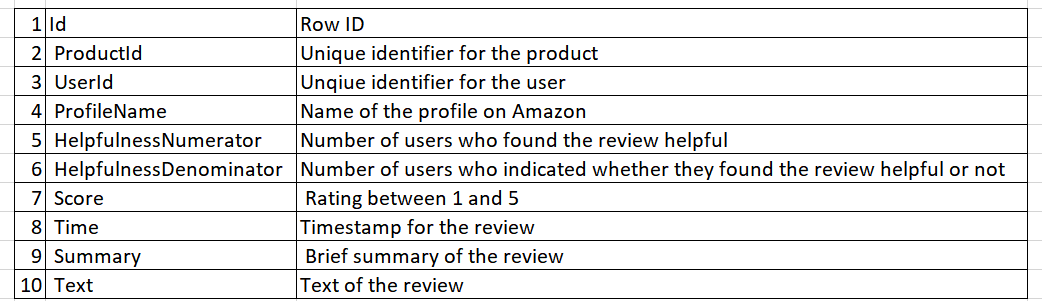
Doing so, we hope our analysis can be used as a tool to differentiate good and bad reviews, as well as bring awareness to what makes a review more beneficial to users.

# Data Preparation

## 

## 3[.1](https://docs.google.com/document/d/1hxO-U91JikqRElsEhRR4KyDDTKS2hwWUToPBK-rWjiM/edit#heading=h.kjorf0t4v7gh) [Data Description](https://docs.google.com/document/d/1hxO-U91JikqRElsEhRR4KyDDTKS2hwWUToPBK-rWjiM/edit#heading=h.y20r49ektq30)

The dataset we chose for this project is called reviews.csv, and has been collected between October 1999 up to October 2012. This data includes more than 500,000 reviews of Amazon food product reviews.There are 10 attributes and 568,454 records in the dataset. The following table briefly describes the attributes in the dataset:

****

The raw data without any cleaning and preprocessing includes:

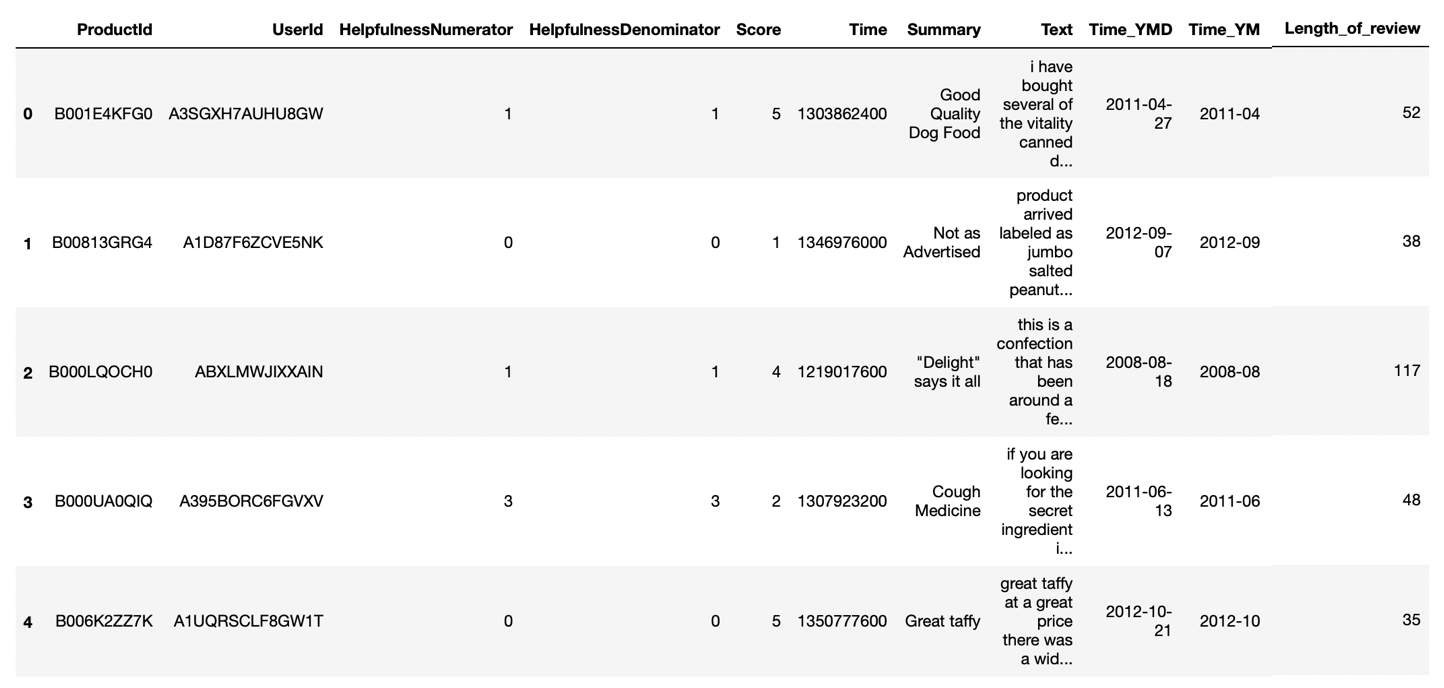
Number of reviews: 568,454

Number of users: 256,059

Number of products: 74,258

## 3[.2.](https://docs.google.com/document/d/1hxO-U91JikqRElsEhRR4KyDDTKS2hwWUToPBK-rWjiM/edit#heading=h.nje1xzvvw7y0) Data cleaning

Before working on our review analysis, we start data cleaning by removing the attributes ID and ProfileName since those attributes are irrelevant in our analysis, and removing rows with missing values. We proceeded to remove newline tabs, white spaces and punctuation in our Text attribute. Additionally, we changed all the words in the reviews to lowercase. Lastly, we replaced all common apostrophe words into their full form. The final data and its dimensions are shown below:



Rows: 568,427

Attributes: 8

Number of reviews: 393,556

Number of users: 256,056

Number of products: 74,258

# Exploratory Data Analysis

## 4.1. Distribution of review count

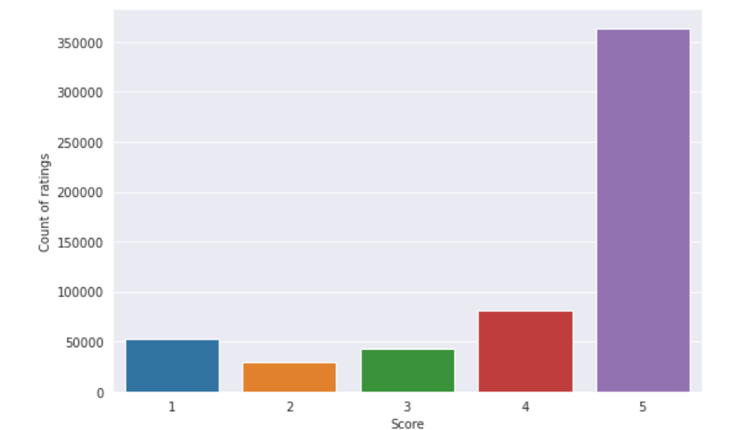


Fig 1. Review Score Vs Count of reviews

The bar chart displays that users tend to give a score of 5 on the majority of reviews. There is an imbalance in the dataset with most reviews inclined towards positive.

The average ratings users give to products on Amazon is 3.979, as described in the summary statistics.

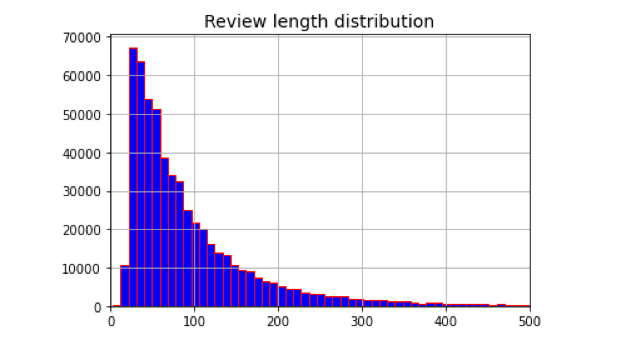


Fig 2. Distribution of review length

The histogram above shows the distribution of the length of reviews. On average, people write reviews with a length of a review with 99 words.

## 4.2 Trend in the number of reviews between 1999 and 2012

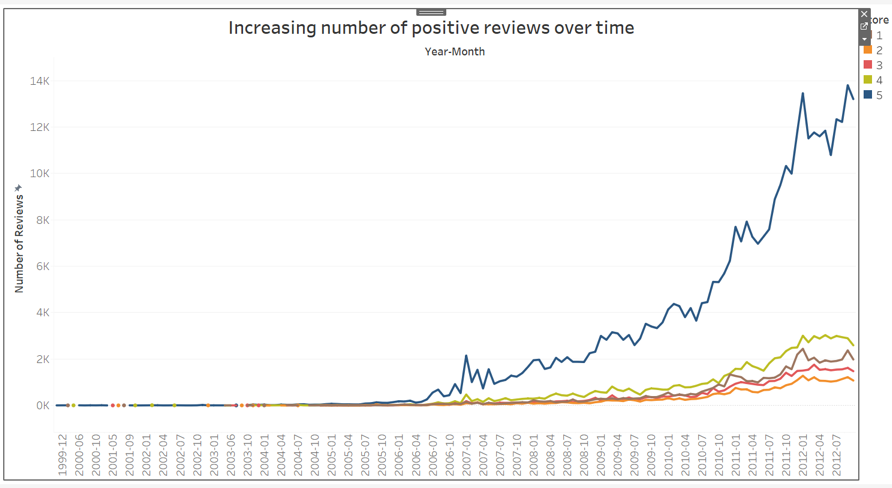


Fig 3. Trend in the review count over time

As popularity of Amazon and its sales increased, there has been a gradual increase in the number of reviews for score 1-4. But as seen in the blue line graph above, there is an exponential increase in the number of reviews with a score 5. Possible explanation is that after 2006, many fake accounts have been created that give a fake positive review for the product. Another reason could be that people generally follow a “herd mentality”. As they see some positive reviews, more people follow the trend giving 5 scoring reviews.

## 4.3 Top Reviewers analysis

We define top reviewers in our dataset as the users that have provided more than 10 reviews on the website. Top reviewers spend more time on the website and may have higher engagement with the community. Let us conduct some exploratory data analysis to understand the behaviour of the top reviewers.

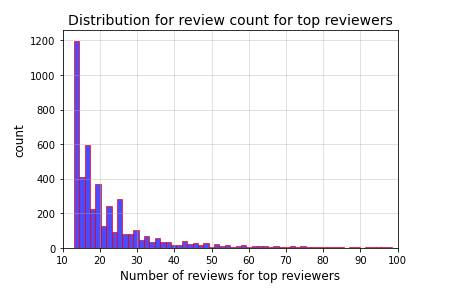


Fig 4. Distribution of review count for top reviewers

The histogram above shows the distribution for the number of reviews in the case of top reviewers. From the graph we can see that most of the top reviewers have review counts between 10 and 30 reviews, and the average number of reviews is 20.

|  |  |
| --- | --- |
| Fig 5. Review Helpfulness Ratio Vs Review Count | Fig 6. Review Helpfulness Ratio Vs Review Count |
|  |  |

The scatterplot in Fig 5. shows the review helpfulness ratio versus the review count. We see a higher number of reviews to be on the right hand side of the graph which tells us that the users have upvoted and found useful those reviews from the top reviewer category. The scatter plot in Fig.6 shows the relationship between review helpfulness score and the review length. We see from the plot that higher review lengths have a higher helpfulness ratio. From this we can infer that users tend to find lengthy reviews helpful.

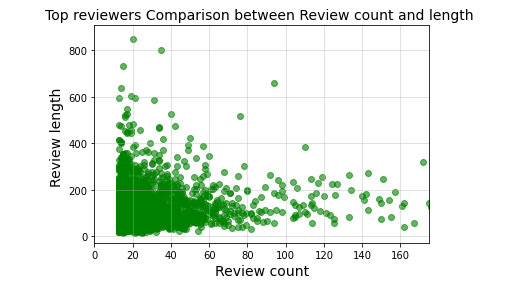


Fig 7. Review Count Vs Review Count

In the scatterplot in Fig.7 we try to find the relationship between review length and review count for the top reviewers category. We noticed two behaviours are common within the top reviewer category. Top reviewers with a very high review count (more than 100) tend to write shorter reviews compared to top reviewers that write lengthy reviews (more than 200 words). The reviewers who write lengthy reviews seem to have lower review counts.

## 4.5 Textual Analysis

According to the original dataset, we have a variable “Score” range from score 1 to score 5. Therefore, in order to implement further textual analysis, we split up the data and assumed score 1 or score 2 will be negative review, conversely, score 4 or score 5 will be assumed as positive review. We utilized the Wordcloud package to visually and aesthetically showcase these words that are highly associated with certain categories that fall under either positive or negative review. The bigger size the word, the more value and closeness it represents. Furthermore, as clear as we can see from the screenshot below, the lower and higher ratings evidently have some similar words ( good, great, taste, flavor). Thus, we do believe that these words are not a good indicator to show what words will matter the most for positive and negative reviews. This is why we agreed and came to the conclusion that it is imperative to conduct further more thorough sentiment analysis in order to explore more information and value out of this dataset.





Fig 8. Word Clouds of Reviews with Positive and Negative Scores

# Sentiment Analysis

For Amazon food review analysis, we gathered and understood the sentiment and insightful information by conducting a sentiment analysis. In order to perform sentiment analysis on the reviews, we used the TextBlob python package, which generates the polarity and subjectivity of each review. The sentiment analyzer in this package uses the Naive Bayes classifier to train the model. We used the sentiment analyzer on our reviews and obtained two scores: polarity and subjectivity. The polarity score is a float within the range [-1.0, 1.0], where -1.0 indicates a very negative review and 1.0 indicates a very positive review. Subjective sentences refer to personal opinions and beliefs, while objective sentences refer to factual information. The subjectivity score is also a float within the range [0.0, 1.0], where 0.0 indicates the review is very objective and 1.0 indicates the review is very subjective. We first start by analyzing the significance of attribute Score and polarity, and observing the significance of each. When purchasing products on Amazon or other sites, many customers tend to only look at the ratings given instead of considering other factors. However, this should not be the case. For this analysis, we took a sample of 10 food products with 5,419 reviews in total. These are our top 10 products based on the number of reviews written for each. We conducted a sentiment analysis based on Score (rating), which has a rating from 1 to 5.

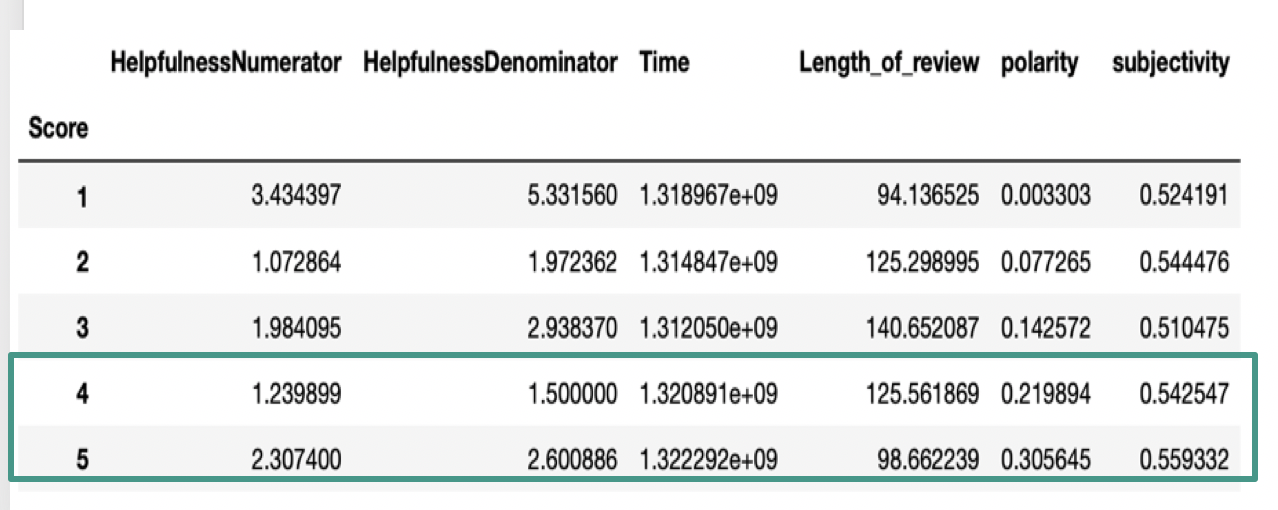


Fig 9. The average polarity and subjectivity for each rating

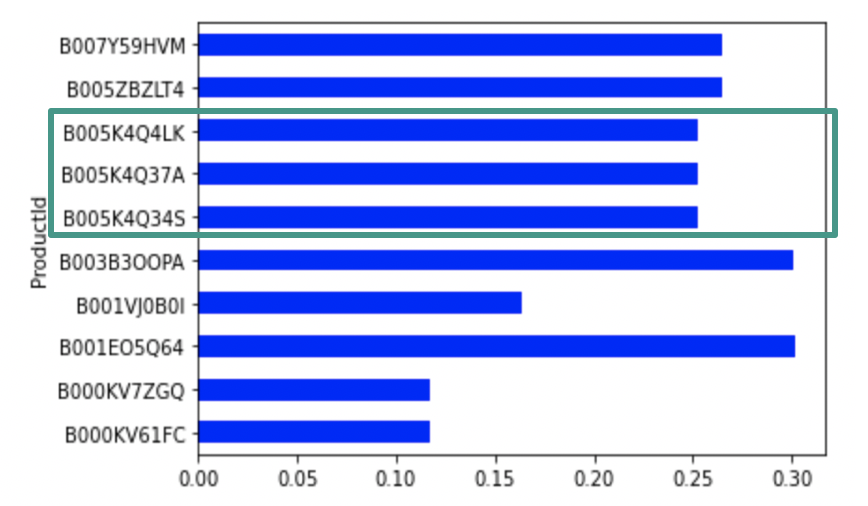
The graph below shows the average polarity of the top 10 products. Here we will be observing three specific products, which are Science Diet, Juniper Ridge Mint Tea, and Sea Salt Vinegar Chips, respectively. All three products have an average polarity score of 0.25. Now comparing with Figure 9, we can see that products that have an average polarity score of 0.25 should have a rating score of between 4 and 5. But when we cross check this with Figure 11, we observe these products have a rating below 4. Therefore, we can conclude that Score (rating) shouldn’t be the only factor considered in determining a product’s quality or when purchasing a product.

Fig 10. The average polarity of the top 10 products

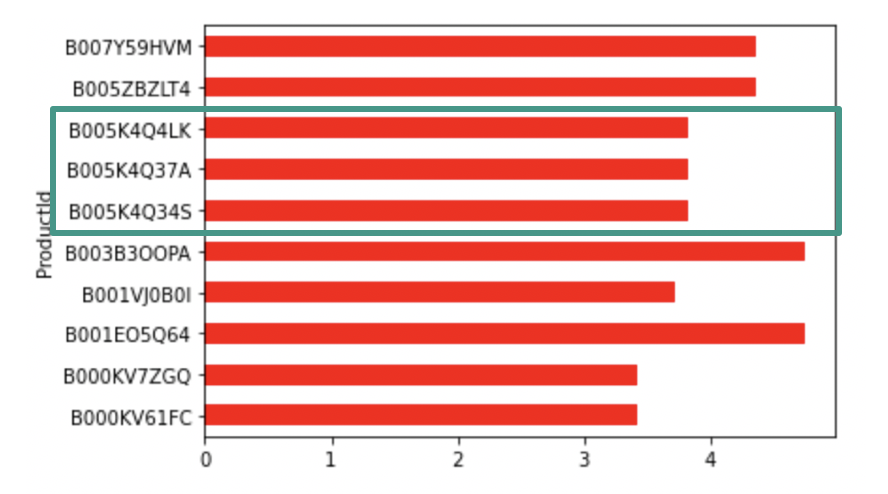


Fig 11. The average rating of the top 10 products

Our findings from working with the top 10 products display that rating may not be a good predictor of product quality since it may not accurately reflect people’s opinions about the products. But if we had used more products to conduct our analysis, we could have obtained a much more accurate observation.

Due to limitations in processing power of our system, we decided to analyze the reviews of the top 300 food products, which is based on the number of reviews written for each product. We will be working with 300 products and 43,891 reviews. After we ran the sentiment analysis, we have 38,188 positive reviews, which are greater than zero, and we have 5,703 negative reviews.

We conducted a box plot of polarity by each score. We first noticed that when the score goes up, then the polarity goes up, as well. However, we also found something interesting that the differences of polarities among scores do not vary significantly. The boxplot below shows that score 1 has median polarity of 0, while the score 5 has median polarity of 0.25. As the polarity for lower scores does not differ significantly from the higher and remains positive or zero, we infer that users may not want to be too aggressive when they write reviews.

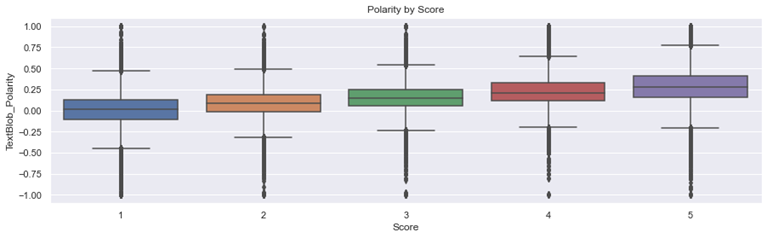


Fig 12. Polarity by Score

In order to verify our analysis, we extracted the frequency of terms with NMF (Non-negative Matrix Factorization) and TfidfVectorizer. Within this method, we extracted the most common terms from the reviews by score.

NMF is a one of the most used unsupervised topic modeling techniques. NMF obtains low-rank matrices with non-negative elements. Given a matrix A, which is the review for our dataset, NMF estimates matrices W and H. W is known as a basis matrix and finds terms from the review. H is known as a coefficient matrix and finds the weights for those terms.

TfidfVectorizer converts a collection of documents to a matrix of TF-IDF (Term Frequency – Inverse Document Frequency) features. TfidfVectorizer computes the term counts, IDF values, and TF-IDF scores, all at once.

The TF of TF-IDF summarizes how often given terms appear in a document, while IDF downsizes terms, which appear more frequently through a corpus or a collection of documents. TF-IDF quantifies terms in the document, which compute weights for each term, indicating an importance of terms in the document or the corpus.

The matrix below shows that users, who gave score 5, usually used positive terms, such as ‘highli remend (highly recommend)’, ‘tast great (taste great)’, or ‘much better’. So, we could clarify that the score 5 has the highest polarity. Regarding the score1, users usually used some negative terms, such as ‘waste money’ or ‘never buy’. However, some terms like ‘would remend (would recommend)’ have been repeated in the list of words for score 1 and score 5. This can explain why the polarity does not vary that much in the dataset.

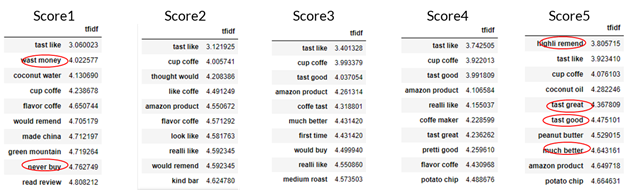


Fig 13. Most CommonTerms by Score

# [Lo](https://docs.google.com/document/d/1hxO-U91JikqRElsEhRR4KyDDTKS2hwWUToPBK-rWjiM/edit#heading=h.cgt3515wf2mn)gistic Regression

After the sentiment analysis, we performed a linear regression analysis to see the relationship between helpfulness and the other variables (Score/rating,length of review, polarity, and subjectivity of the review). We initially create a helpfulness score by dividing the number of helpfulness numerator to the denominator. For every helpfulness score above or equal to 0.5, it is considered as helpful reviews and all reviews with helpfulness score below 0.5 is considered as not helpful. Below is the results of the logistic regression model.

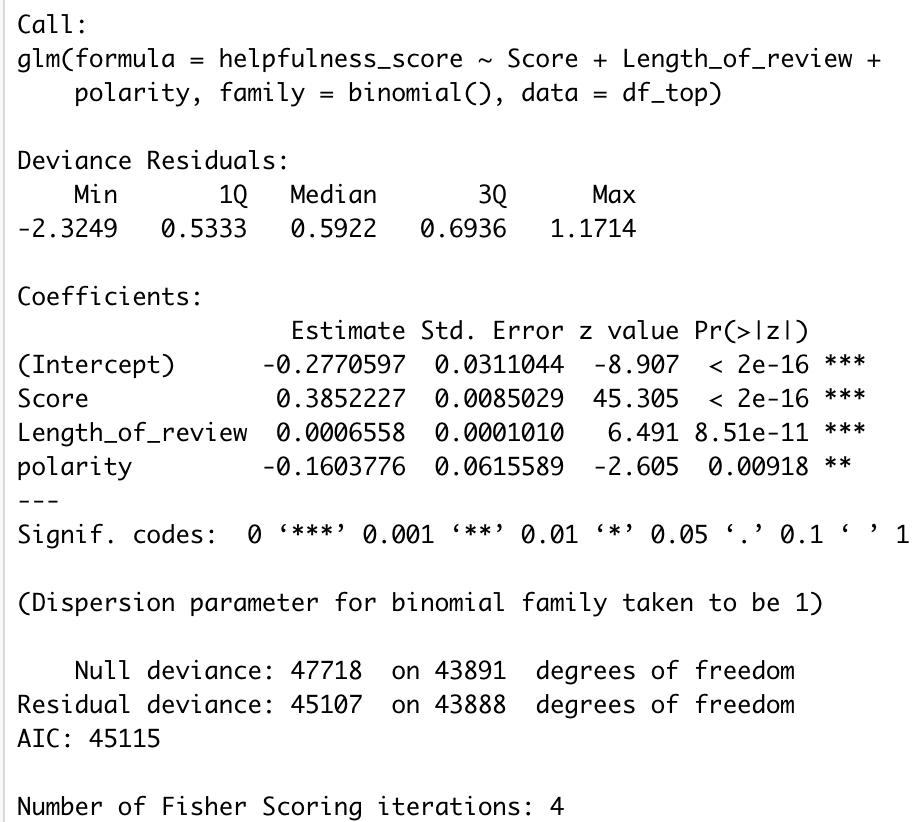
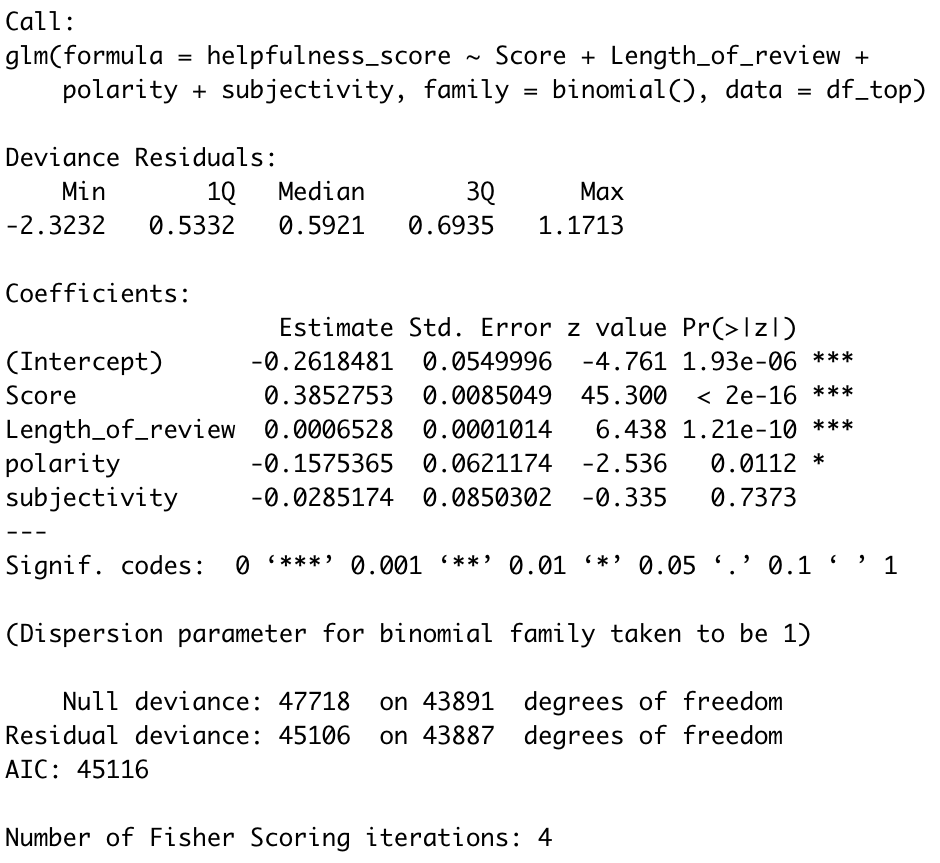
****

Fig 14. Logistic Regression- All variables Fig 15. Logistic Regression - Significant variables

The result on Fig 14. is the first model with all the variables. The result shows that only score, length of review, and polarity are significant. Afterward, We re-run the model with only the significant variables (Fig 15). If we look at the coefficients, we know that the higher the rating and length of the review, the more helpful the reviews are. The smaller the polarity, the more helpful the reviews, which means reviews with “negative” sentiment are more helpful compared to the reviews with positive sentiment.

Below are some visualizations from the data to show some of the points we make from the logistic regression analysis.

**Subjectivity vs. Polarity**

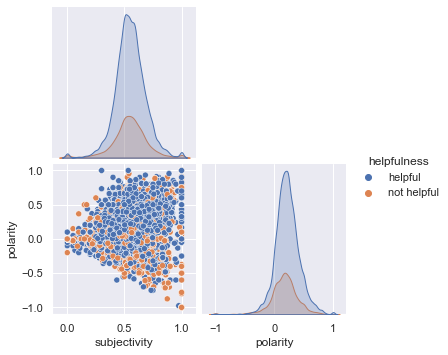
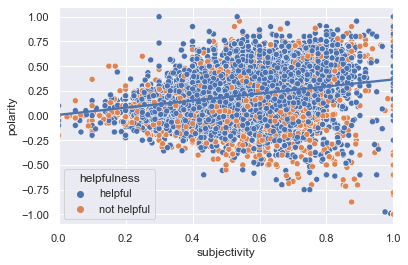
****

Fig 16. Subjectivity vs. Polarity

From the figure above, we can see a positive correlation between subjectivity and polarity. As the polarity increases,the higher the subjectivity as well. This shows that most of the positive reviews are pretty subjective. While reviews with negative sentiment tend to be more objective.

**Length of review vs. Helpfulness**

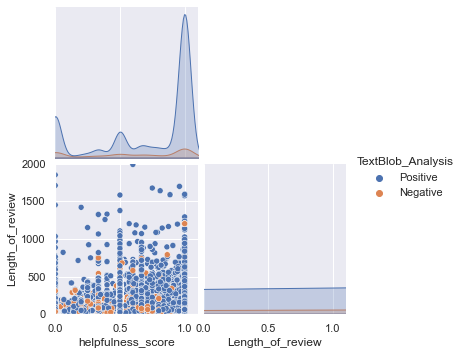
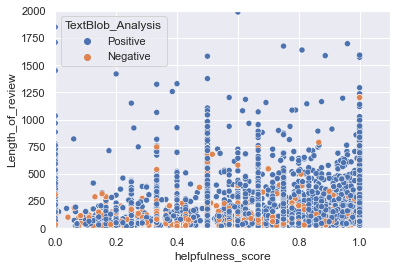
****

Fig 17. Length of review vs. Helpfulness

The graph above shows that as the length of review increases,the higher the helpfulness score. Most of the longer reviews fall in the higher range of helpfulness score.

**Subjectivity vs. Helpfulness**

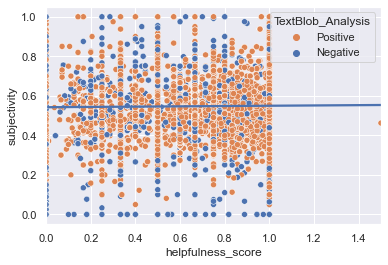
****

Fig 18. Subjectivity vs. Helpfulness

This graph above shows how subjectivity has no relationship with helpfulness. All the dots are scattered and we do not see any clear pattern.This shows that we couldn’t find any significant relationship between subjectivity and helpfulness. It doesn’t matter if the review is subjective or objective, a review could be helpful or not helpful.

# [Interpretation and Key Takeaways](https://docs.google.com/document/d/1hxO-U91JikqRElsEhRR4KyDDTKS2hwWUToPBK-rWjiM/edit#heading=h.dqhcv59at584)

After all the analysis we conducted, we derived two key takeaways for the reader of reviews and the reviewers. First, buyers/the reader need to perform higher due diligence especially for highly rated products. Second, reviewers need to be aware of what makes a review beneficial to the users. Based on our analysis, here is some guidance to differentiate between good or bad reviews.

1. The top reviewer usually either writes a large number of reviews but in shorter length or less reviews that are lengthy. This shows how reviewers do put effort in what they write and it usually shows the quality of the reviews as well.
2. Some factors to consider to decide whether a review is good or bad are the length of the review, the rating score, and the polarity. Those are the significant factors that determine the helpfulness of a review.
3. Users should not only consider score rating as a factor in determining product’s quality. Because not all reviews are good reviews, some might be fake.
4. Users need to try to read a few of the negative reviews before making a purchase decision. This is because negative reviews tend to be more helpful.

# 

# 

# Citations

1. McAuley, Julian, and Jure Leskovec. *From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews*. <http://i.stanford.edu/~julian/pdfs/www13.pdf>
2. “Web Data: Amazon Fine Foods Reviews.” *SNAP*, <https://snap.stanford.edu/data/web-FineFoods.html>