Amazon Customer Review Sentiment Analysis

ACRSent

**Data Science Capstone Project   
Exploratory Data Analytics Report**

Date:

5/27/2021

Team Members:

Name: Kevin Fitzpatrick

Name: Tessa Karlovitz

Name: Dustin Luchmee

[The purpose of this report is to describe the exploratory data analytics. It includes five major sections:

1. Analyzing the basic metrics of variables: data types, size, descriptive statistics
2. Non-graphical and graphical univariate analysis: identifying unique value and counts, histogram, box plots, etc.
3. Missing value analysis and outlier analysis
4. Feature engineering and analysis: correlation analysis, dimensionality reduction, deriving new variables
5. Appendix]

**Analysis the basic metrics of variables**

[In this section, we identify all the variables in the dataset and conduct the basic metrics of the variables. What are the data types (numerical/categorical, discrete or continuous, ordinal or nominal) and size? Provide the descriptive statistics of the variables such as mean, standard deviation, min, max, percentiles, etc.]

**Book Data Dictionary (excluding columns not used):**

**Amazon:**

**product\_title** – Title of the book (string)

**review\_data** – Date of the review (date)

**review\_body** – body of text of the review (string)

**star\_rating** – star rating given in the review (integer)

**verified\_purchase** – yes or no for verified purchase or not (Boolean)

**helpful\_votes** – number of helpful votes for the review (integer)

**total\_votes** – total number of votes for the review (integer)

**Kaggle:**

**Year** – year for that top ranked book (integer)

**Rank** – rank for that book that year (integer)

**Book\_Title** – title of the book (string)

**Author** – Author of the book (string)

**Rating** – Average star rating (float)

**Num\_Customers\_Rated** – Number of customer ratings for the book (integer)

**Price** – Price of the book (float)

Descriptive Statistics for Amazon Review Data:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Star\_Rating** | **Total\_Votes** | **Length\_Review\_Body** |
| **count** | 2,668,988 | 2,668,988 | 2,668,988 |
| **mean** | 4.549 | 2.016 | 262.759 |
| **std** | 0.930 | 13.927 | 575.293 |
| **min** | 1 | 0 | 1 |
| **25%** | 4 | 0 | 68 |
| **50%** | 5 | 0 | 140 |
| **75%** | 5 | 1 | 262 |
| **max** | 5 | 4,756 | 49,521 |

\*Note the length of the review body is calculated by us.

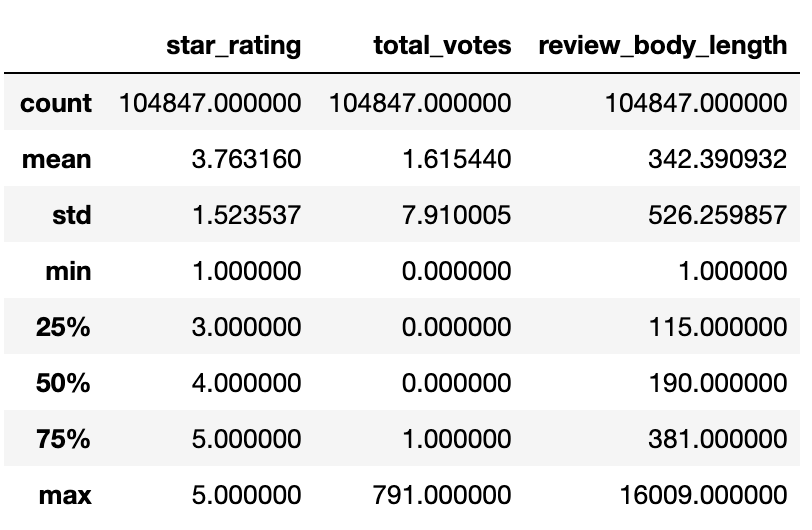
Descriptive Statistics for Kaggle Amazon Top 100 Ranked Books by Year Data:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Year** | **Rank** | **Rating** | **Num\_Customers\_Rated** | **Price** |
| **count** | 1094 | 1094 | 1094 | 1094 | 1094 |
| **mean** | 2015.002 | 50.379 | 4.631 | 14013.101 | 10.851 |
| **std** | 3.168 | 28.832 | 0.217 | 14303.567 | 11.866 |
| **min** | 2,010 | 1 | 3 | 4 | 0 |
| **25%** | 2,012 | 25 | 5 | 4,490 | 3 |
| **50%** | 2,015 | 50 | 5 | 10,378 | 9 |
| **75%** | 2,018 | 75 | 5 | 18,025 | 16 |
| **max** | 2,020 | 100 | 5 | 126,619 | 152 |

**Mobile Electronics Data Dictionary (excluding columns not used):**

|  |  |  |
| --- | --- | --- |
| **Field** | **Data Type** | **Description** |
| product\_title | string | Title of the product. |
| review\_headline | string | The title of the review. |
| review\_body | string | The review text. |
| review\_date | datetime | The date the review was written. |
| total\_votes | integer | Number of total votes the review received. |
| num\_words | integer | Number of words in the review body |
| num\_punctuation | integer | Number of punctuation characters in the review body |
| num\_capital\_words | integer | Number of capitalized words in the review body |
| num\_stop\_words | integer | Number of stop words used in the review body |
| num\_non\_stop\_words | integer | Number of non stop words used in the review body |
| num\_unique\_words | integer | Number of unique words used in the review body |
| star\_rating | integer | The 1-5 star rating of the review. |

Descriptive Statistics for Mobile Electronics:



**Amazon Review Stationary Electronics Data (excluding columns not used)**

Marketplace – country where product is available.

Customer ID – identification # of customer

Review ID – identification # of the review

Product ID – identification # of product

Product Parent - # corresponding to product type.

Product Title – name of the product

Product Category – category of product (all electronic)

Star Rating – Rating 1-5 of the product with 1 being poor and 5 being excellent.

Helpful Votes – number of people who found the review helpful.

Total Votes – total number of votes for the review.

Verified Purchase – Y or N, whether amazon has verified that this product was sold to this client.

Review Headline – The title of review

Review Body – Actual review, total content

Review Date – Date the review was made (converted to year).

**Descriptive Statistics for Electronic Data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **customer\_id** | **product\_parent** | **star\_rating** | **helpful\_votes** | **total\_votes** |
| **count** | 3.091024e+06 | 3.091024e+06 | 3.091024e+06 | 3.091024e+06 | 3.091024e+06 |
| **mean** | 2.879196e+07 | 5.102057e+08 | 4.035415e+00 | 1.860874e+00 | 2.372440e+00 |
| **std** | 1.542984e+07 | 2.868327e+08 | 1.387460e+00 | 2.133715e+01 | 2.249700e+01 |
| **min** | 1.000500e+04 | 6.478000e+03 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| **25%** | 1.503881e+07 | 2.623618e+08 | 3.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| **50%** | 2.806679e+07 | 5.085523e+08 | 5.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| **75%** | 4.328128e+07 | 7.632433e+08 | 5.000000e+00 | 1.000000e+00 | 1.000000e+00 |
| **max** | 5.309658e+07 | 9.999982e+08 | 5.000000e+00 | 1.278600e+04 | 1.294400e+04 |

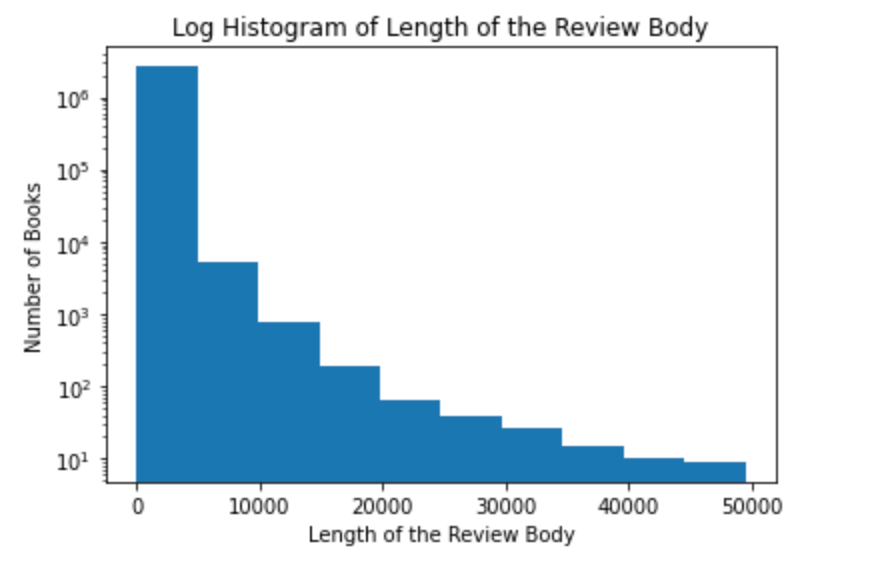
1

**Non-graphical and graphical univariate analysis**

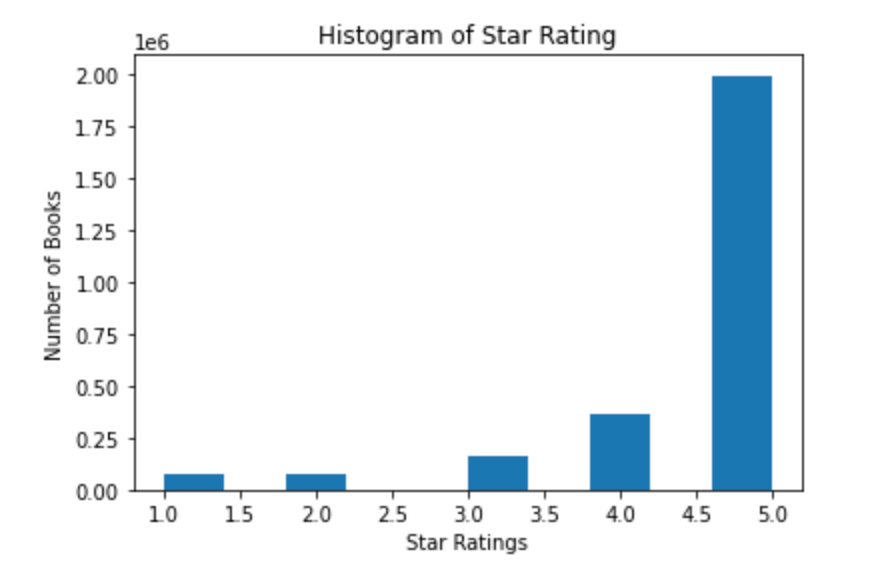
[In this section, we identify the list and number of unique values for each variable and provide the histogram and box plots to understand the distribution of the data.]

**Book Data**

Log Histogram of the Length of the Review Body for each book in the final cleaned dataset for amazon customer reviews. Log is used here since most reviews are short but there are a few long ones.



Histogram of the Star Rating for each book in the final cleaned amazon customer review dataset



Monthly Average Star Rating for all the books in the final cleaned amazon customer review dataset

Chart, line chart

Description automatically generated

Monthly Average Review Body Length for all the books in the final cleaned amazon review dataset. This shows a fairly dramatic change in the length of the reviews over the 3 year span which is quite interesting. Perhaps the length was shortened as “text-language” became more acceptable.

Chart, line chart

Description automatically generated

Monthly Average Total Votes for all the books in the final cleaned amazon review dataset. This also shows quite a dramatic drop in the use of total votes.

Chart, line chart

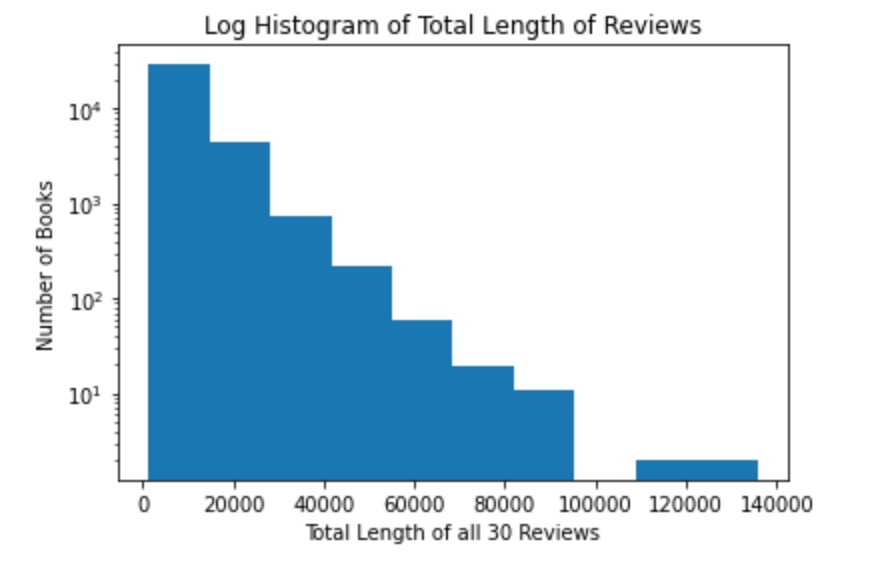
Description automatically generated

Word Cloud of the review bodies of all the books in the final cleaned Amazon Customer Review Dataset

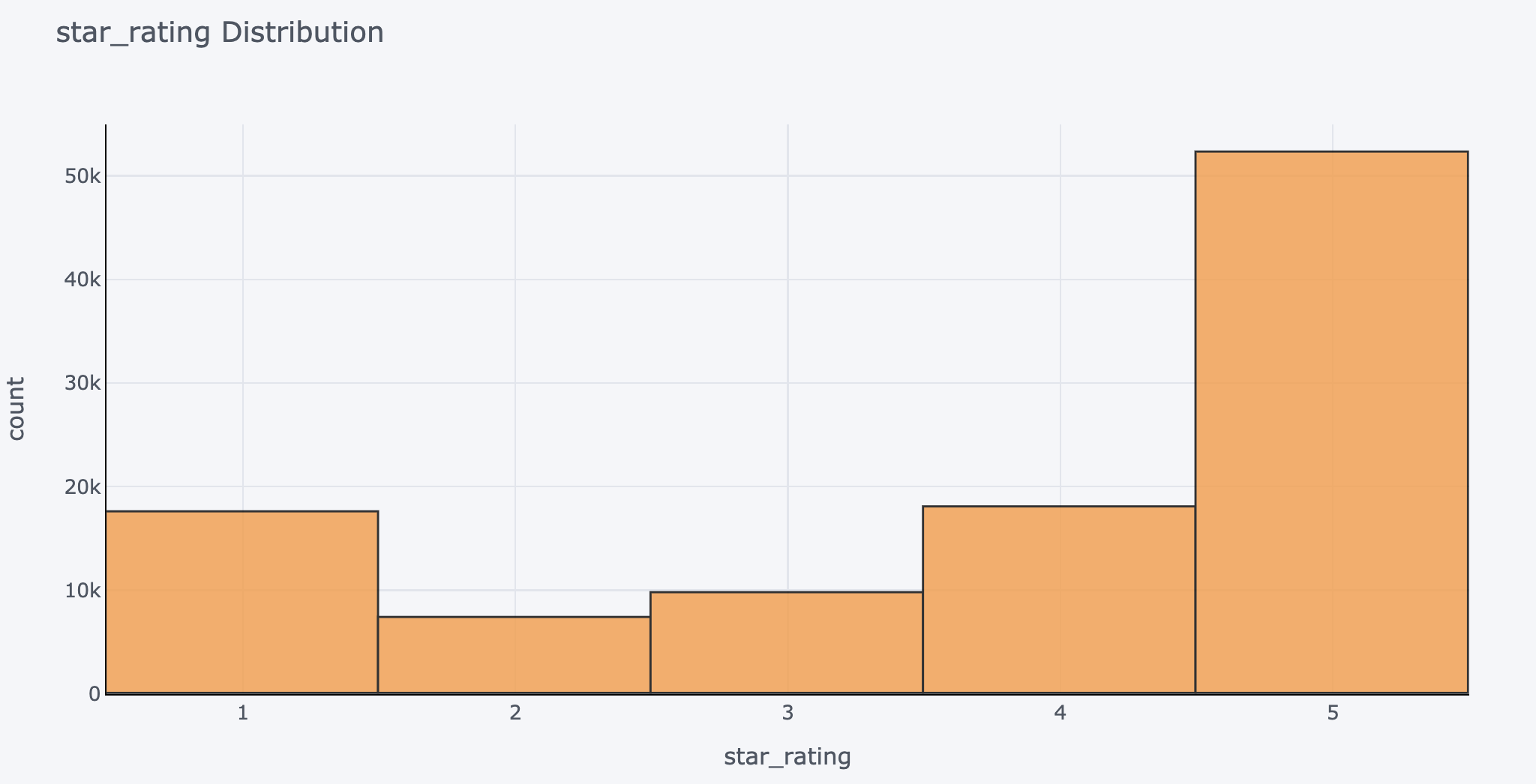
Text, calendar

Description automatically generated

Log Histogram of the Total Length of Reviews for 30 Reviews in the final dataset for combined amazon customer reviews and top 100 ranked books on amazon. Like above we’re applying a log function since there are many short reviews but a few long ones.



**Mobile Electronics**



**Stationary Electronics**

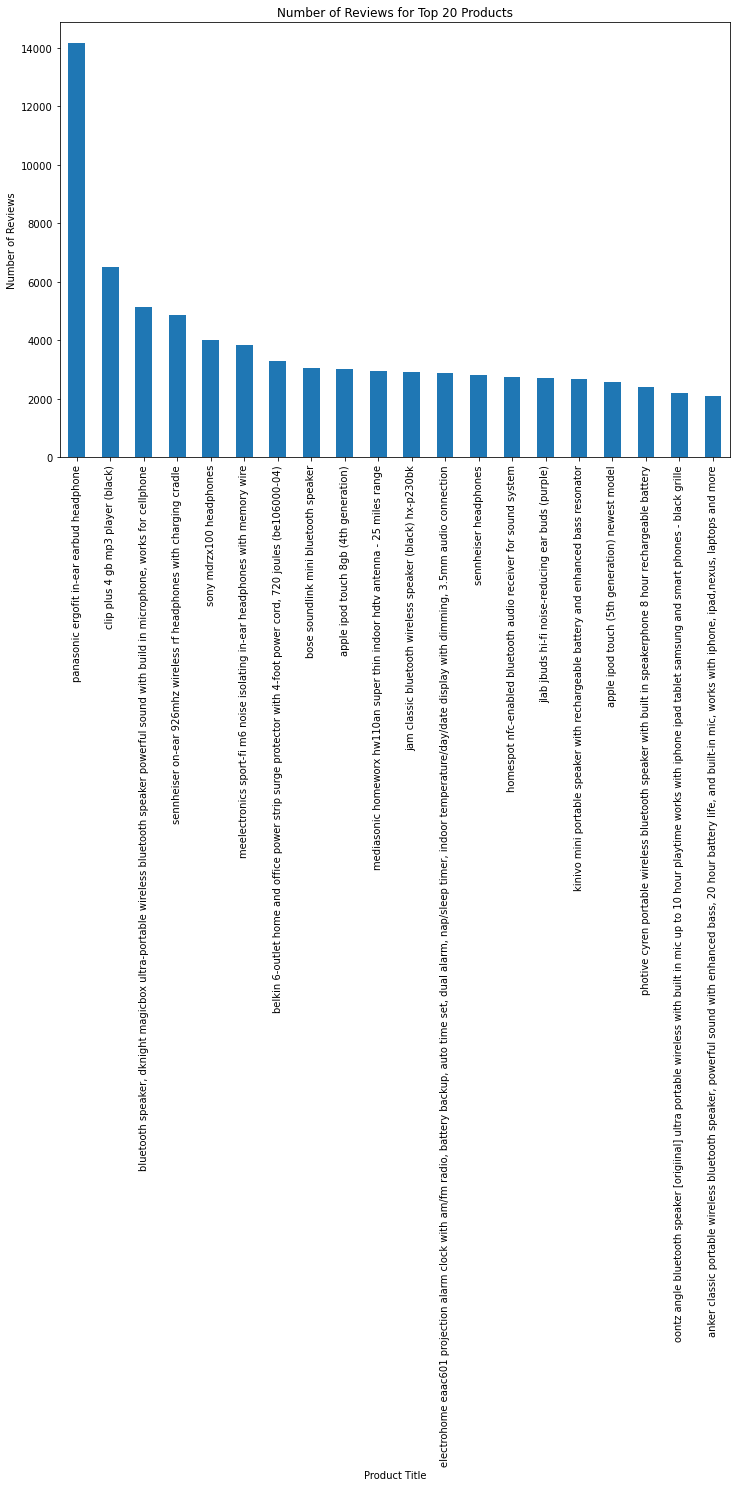
The dataset originally is quite large, consisting of 3,091,024 data entries. Not all entries are going to be useful for data analysis. Any rows with null values were dropped, as were rows that had empty titles for products. For my analysis, I took only products that had 500 reviews or more, because these products are likely to be well known, have been on the market a long time, and meet the needs of people seeking a particular electronic device. In addition, items such as sleeves, cases, cleaning products, cables, chargers, adapters, and mounts that might be sold in an electronics department but are not devices themselves were removed. All items come from the US marketplace and are verified Amazon purchases. Lastly, the dataset was converted into all lowercase text. After removal of undesirable items and random sampling of 60% of the dataset, 679 unique products remained for analysis.

**Missing Value Analysis**

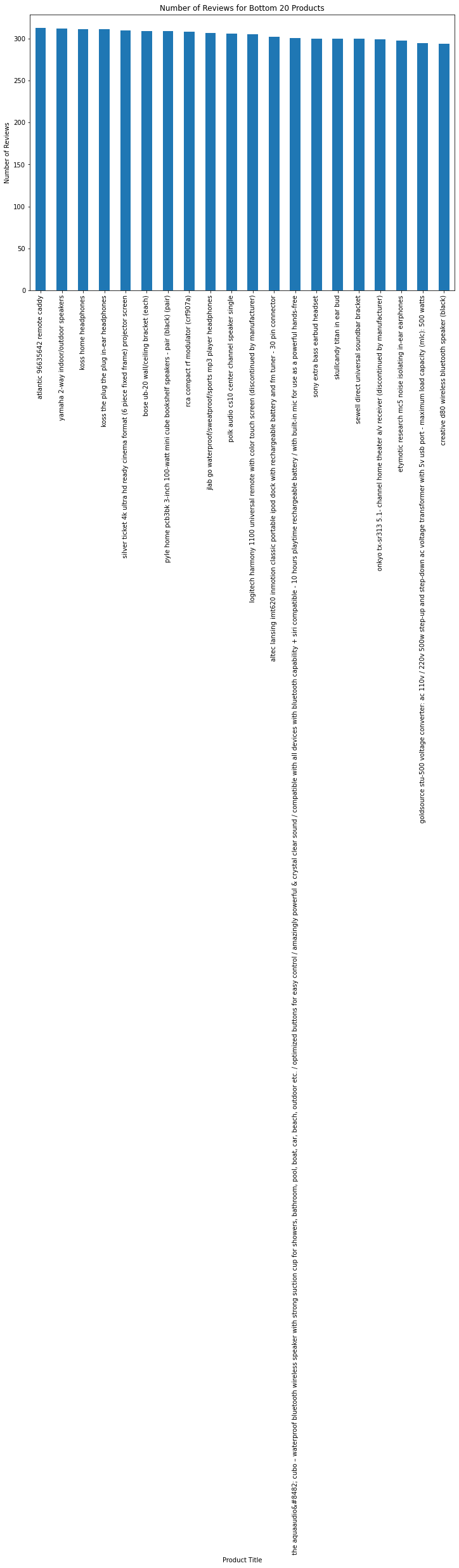
Missing values were removed from the dataset. What I had found in the dataset was that some product titles were blank: “Create a generic Title per Amazons guidelines.” These rows were removed, as no specific product title is available, and these reviews could refer to anything.

***Feature Engineering and Analysis***

To get a sense of the number of reviews, I pulled the top 20 products within the dataset and examined the total number of reviews for the top 20 products.

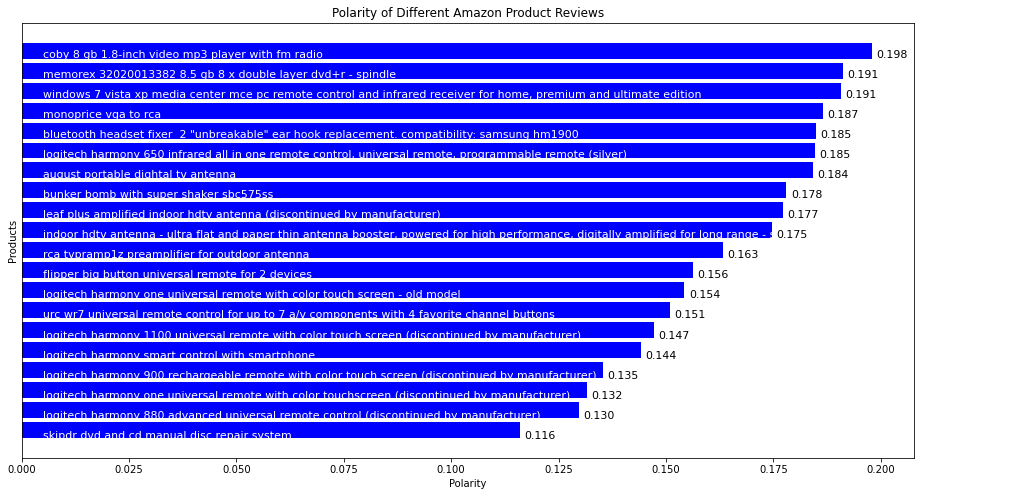


This same process was repeated for the bottom 20, or worst products as well:



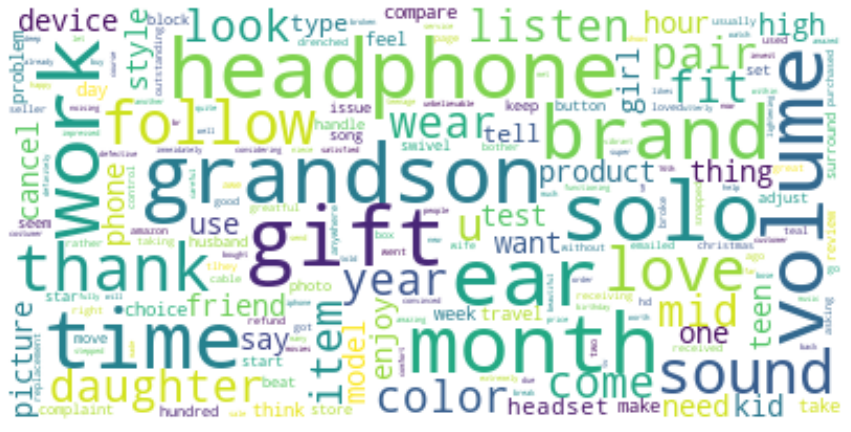
Something critical to understanding about reviews is polarity. Polarity tells you how positive or negative a review is. What I had done was take the polarity of the top and bottom 20 products.



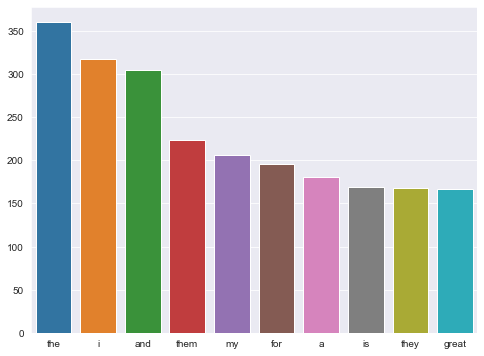


From this data, it looks like the beats solo hd over-ear headphones is the hottest product and the skindr dvd and cd manual disc repair system is the least liked.

I examined a random review for the beats solo hd over-ear headphones. I was curious to see what a review for a highly liked product might look like graphically:

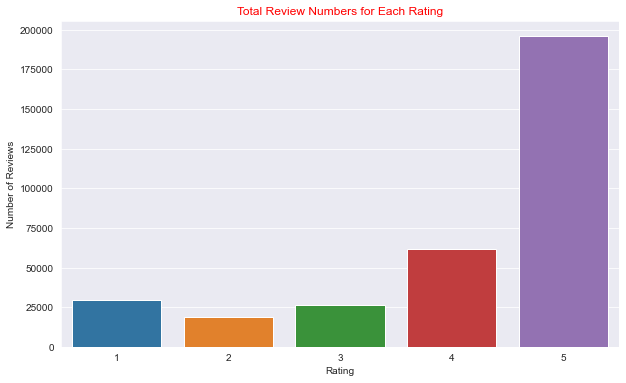


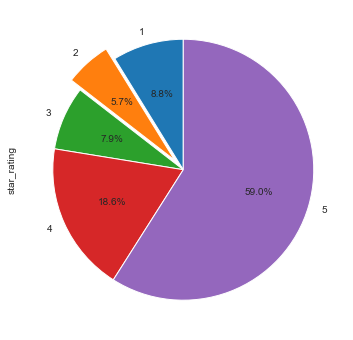
Interestingly, the brand is not even mentioned. The review contains 1308 words. I also looked at the frequency of the words within the reviews of this product that were above a 3-star rating.



It seems the more positive a review is, the less likely that the product name is even mentioned directly. It makes sense because often when someone issues a complaint about a product, they express anger or frustration when referencing the product.

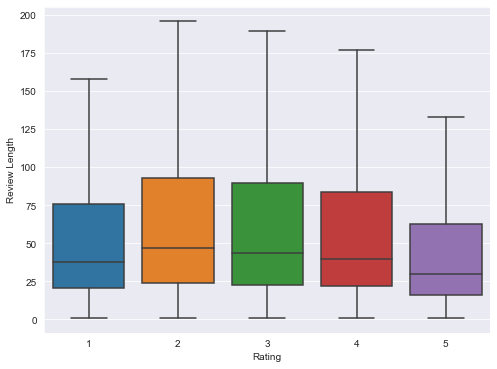
It is also important to understand the distribution of star ratings within the dataset:





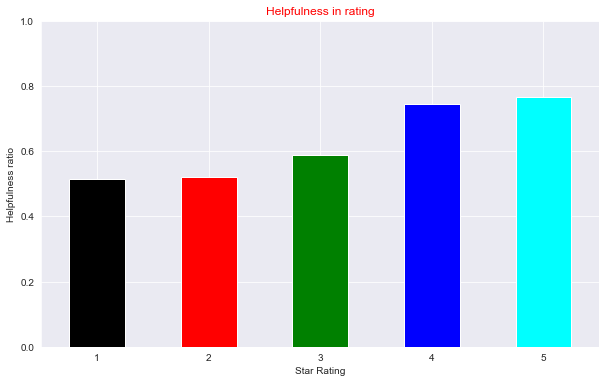
From this data, we learn that if your product is spectacular, you will get more reviews which can encourage potential buyers to add your product to their checkout cart. 5-star ratings are the most common, followed by 4-, 1-, 3-, and 2-star reviews.

It may also be of interest to look at the review length in regard to the star rating.



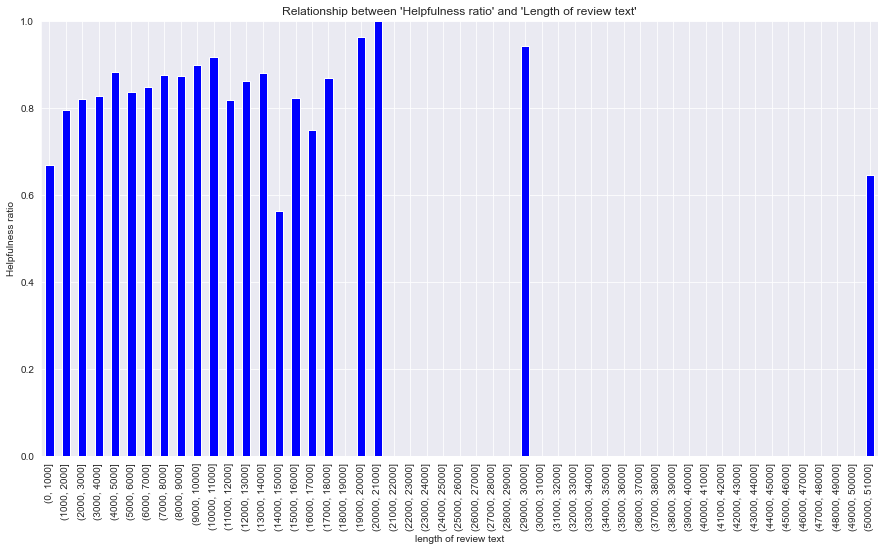
On average, 5-star reviews are shorter than 2-star reviews. When people are happy with a product, there is usually less to critique. Super fans of a product may add more length or descriptions to their reviews.

Do people find reviews helpful? A helpfulness ratio for reviews was calculated by taking the helpful votes / total votes for each review. This was plotted in relation to the star rating.

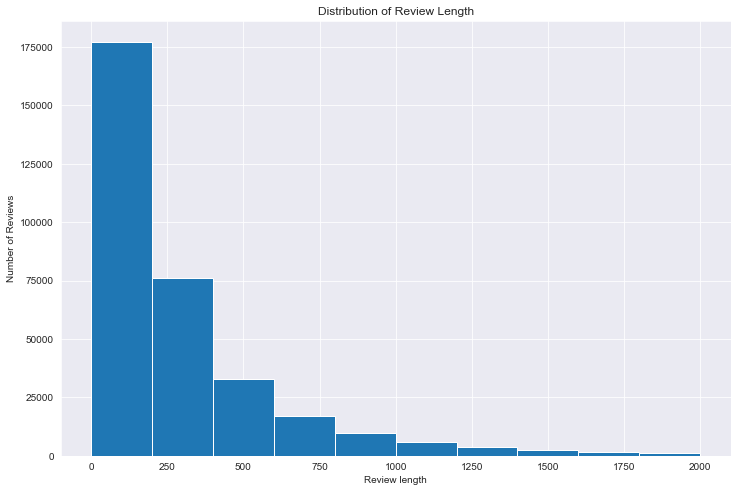


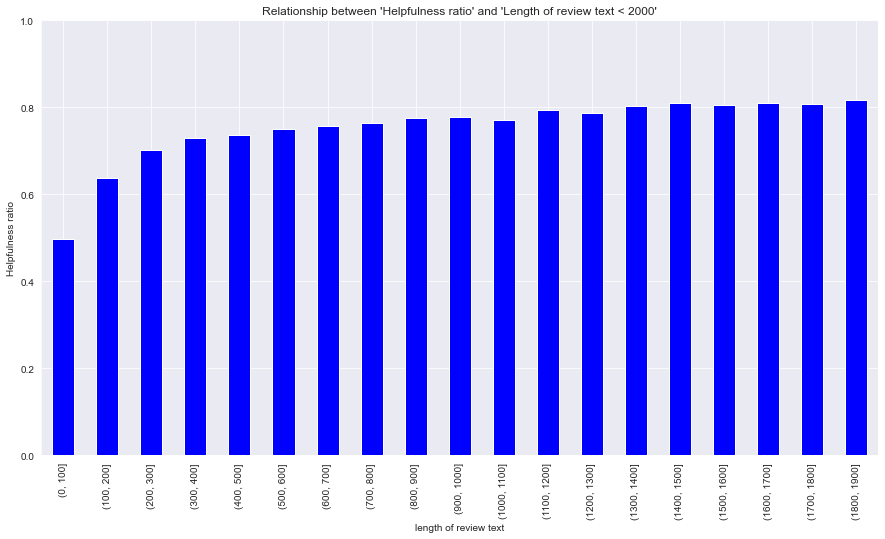
Positive reviews are more helpful than negative reviews. This may be due to the content, as complaints may involve whining behavior versus helpful insights into the product.

How does helpfulness impact review length?

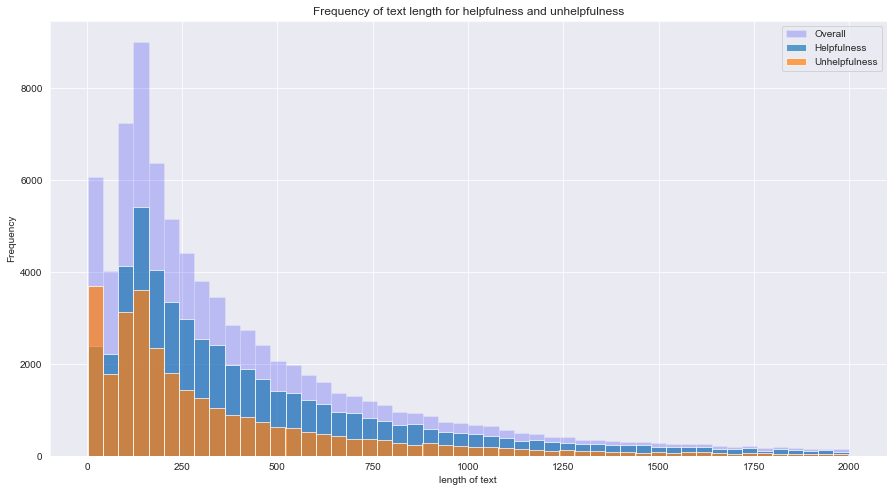


If a review is going to be helpful, it will be about 1900-2100 words long. I also looked at the distribution of review length in relation to the number of reviews and the relationship between helpfulness ratio and review length.



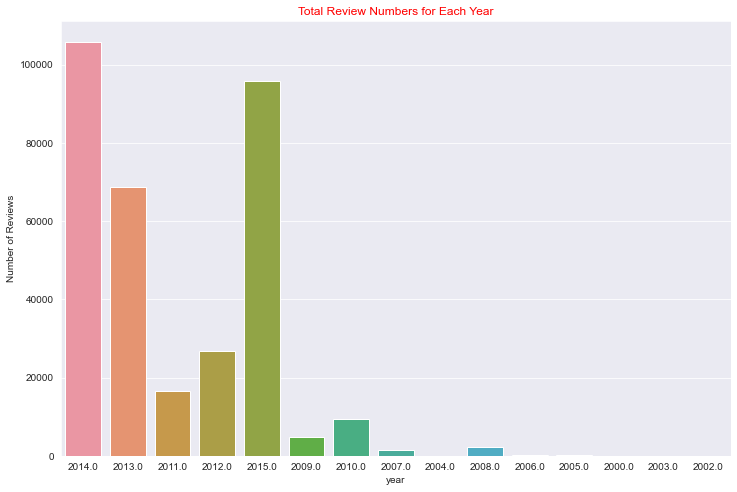


How long are reviews that are considered helpful or unhelpful? How frequent do these review lengths come up?

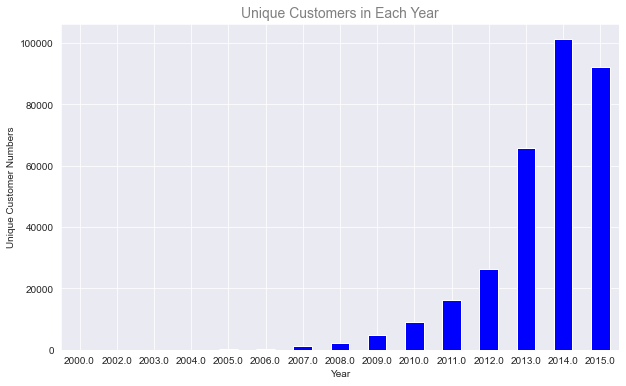


Whether helpful or unhelpful, people like to be concise and to the point. Some reviews in the dataset were “five stars” or “awful. Don’t bother.”.

It may be of interest to better understand review and customer behavior through time. For example, how has review behavior changed over the years? Are people reviewing more than they were before?

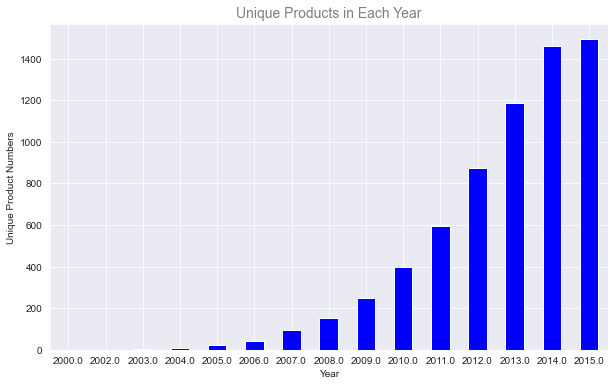


Yes, with the internet and advent of social media, it is likely that people became more comfortable expressing their thoughts online. 2014 seems to be the year the most people reviewed Amazon products. I wonder if this may be the case for 2020 as many more people ordered things online and were home, so they had the time to write more product reviews.



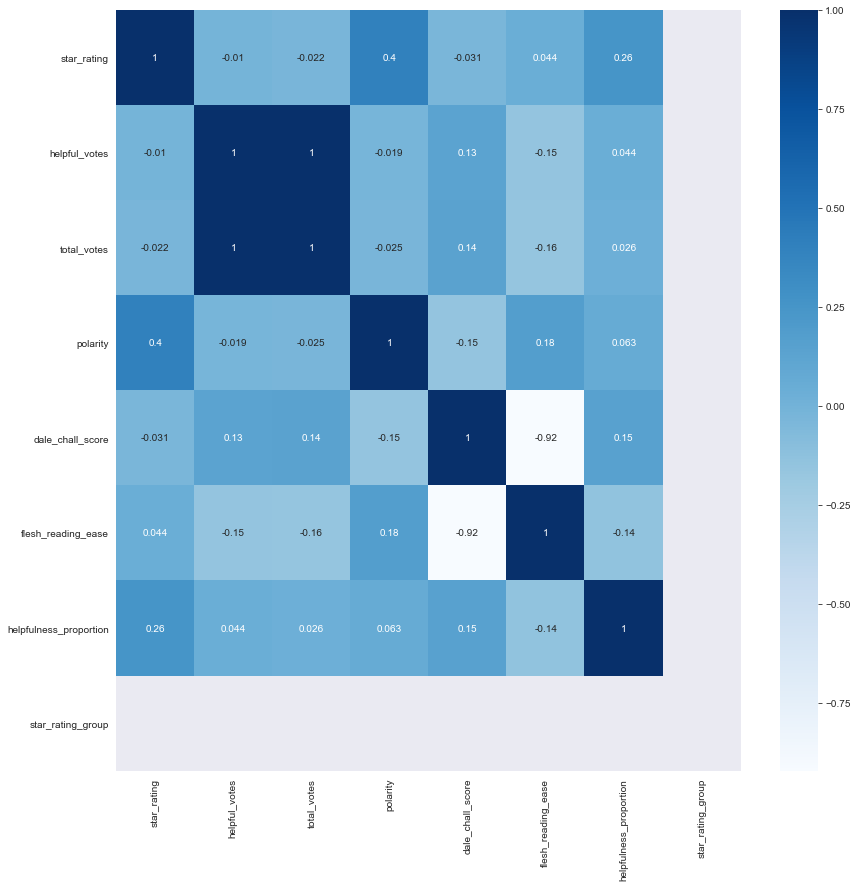
It is also helpful to know the total number of unique customers each year. 2014 was also the year that there were the most unique customers reviewing electronic products on Amazon.

Amazon is a huge marketplace. How many unique products are there? Has the number of unique products changed year-to-year?



It looks like in 2014 and 2015, amazon had more unique electronic products than any other year. That is due to technological advancements and changing tastes and needs of consumers as smart technology became more advanced.

Lastly, I did a correlation analysis to see which of the numerical values are correlated with one another.



Reading scores are added but take a long time to process. Am currently pulling the Dale-Chall, Flesch Kinkaid, and Gunning Fog Index. These metrics were chosen to understand the complexity and grade level of reviews. I am trying to see whether positive vs negative reviews are typically more complex or simple.

Other metrics that may be of interest are subjectivity vs polarity of a review, to see whether people inject their opinions more into a review when it is positive or negative.

My plan is to run sentiment analysis using machine learning models such as logistic regression, naïve bayes, random forest, XGBoost, and Cat Classifier models to see which one may be best to predict sentiment of product reviews. My aim is to be able to recommend products and take a better look at products that are not doing well and be able to let makers know what features seem to not be working out for their customers.

**Missing value analysis and outlier analysis**

[In this section, we identify the missing values and outliers and determine how we handle these values before analysis.]

**Book Data**

The way we dealt with missing values in the book data was by removing them. Very few lines had missing values so in this case it made sense to remove the lines than work with other techniques.

**Feature engineering and analysis**

[In this section, we identify the variables that are useful for predictive modeling and machine learning through correlation analysis. You may also reduce the dimension or derive new variables so that the predictive modeling can be more efficient and effective.]

**Appendix**

[Provide the code or pseudo code, and any other information in the appendix here.]

Table of Contributions

The table below identifies contributors to various sections of this document.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Analysis the basic metrics of variables** | **All** | **All** |
| **2** | **Non-graphical and graphical univariate analysis** | **All** | **All** |
| **3** | **Feature engineering and analysis** | **All** | **All** |
| **4** | **Appendix** | **All** | **All** |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.