

Sentiment Analysis and Product Recommendation on
Amazon's Electronics Dataset Reviews
Data Wrangling and EDA - Capstone Project 2

INTRODUCTION:

The Internet has revolutionized the way we buy products. In the retail e-commerce world of online marketplace, where experiencing products are not feasible. Also, in today's retail marketing world, there are so many new products are emerging every day. Therefore, customers need to rely largely on product reviews to make up their minds for better decision making on purchase. However, searching and comparing text reviews can be frustrating for users. Hence we need better numerical ratings system based on the reviews which will make customers purchase decision with ease.

Goal:

During their decision making process, consumers want to find useful reviews as quickly as possible by rating system. Therefore, models able to predict the user rating from the text review are critically important. Getting an overall sense of a textual review could in turn improve consumer experience. Also, it can help businesses to increase sales, and improve the product by understanding customer's needs.

In this project, the amazon review dataset for electronics products will be considered. The reviews and ratings given by the user to different products as well as reviews about user's experience with the product(s) will be considered.

The main goal for this project is to develop a model to predict user rating, usefulness of review and recommend most similar items to users based on collaborative filtering.

Client:

Amazon is our client. The company wants to develop a software tool that will identify the positive and negative words which customers use when they write reviews for the home and kitchen products as their purchase inclination.

Getting an overall sense of a textual review could in turn improve better consumer decision making experience. Also, it can help businesses to increase sales, and improve the product by understanding customers' needs and problems.

DATA COLLECTION:

For this project, the electronics dataset consist of reviews and product information from amazon were collected. This dataset includes reviews (ratings, text, helpfulness votes) and product metadata (descriptions, category information, price, brand, and image features).

1. Product Complete Reviews data:

This dataset includes electronics product reviews such as ratings, text, helpfulness votes. This dataset was obtained from <http://jmcauley.ucsd.edu/data/amazon/>. The original data was in json format. The json was imported and decoded to convert json format to csv format. The sample dataset is shown below:

	asin	helpful	Rating	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
0	0528881469	[0, 0]	5	We got this GPS for my husband who is an (OTR)...	06 2, 2013	AO94DHGC771SJ	amazdnu	Gotta have GPS!	1370131200
1	0528881469	[12, 15]	1	I'm a professional OTR truck driver, and I bou...	11 25, 2010	AMO214LNFCEI4	Amazon Customer	Very Disappointed	1290643200
2	0528881469	[43, 45]	3	Well, what can I say. I've had this unit in m...	09 9, 2010	A3N7T0DY83Y4IG	C. A. Freeman	1st impression	1283990400
3	0528881469	[9, 10]	2	Not going to write a long review, even thought...	11 24, 2010	A1H8PY3QHMQQA0	Dave M. Shaw "mack dave"	Great grafics, POOR GPS	1290556800
4	0528881469	[0, 0]	1	I've had mine for a year and here's what we go...	09 29, 2011	A24EV6RXELQZ63	Wayne Smith	Major issues, only excuses for support	1317254400

Fig 1: Sample Product Reviews Dataset

Each row corresponds to a customer review and includes the following variables:

reviewerID : ID of the reviewer, e.g. A2SUAM1J3GNN3B - type: object

asin : ID of the product , e.g. 0000013714 – type: object

reviewerName : name of the reviewer – type: object

helpful : helpfulness of the review, e.g. 2/3 – type: object

reviewText : text of the review – type: object

overall : Rating (1,2,3,4,5)– type: float64

summary : summary of the review – type: object

unixReviewTime : time of the review (unix time) – type: int64

reviewTime : time of the review (raw) – type: object

2. Product Metadata:

This dataset includes electronics product metadata such as descriptions, category information, price, brand, and image features. This dataset was obtained from <http://jmcauley.ucsd.edu/data/amazon/>. The json was imported and decoded to convert json format to csv format. The sample product meta dataset is shown below:

	asin	imUrl	description	categories	title	price	salesRank	related	brand
0	0132793040	http://ecx.images- amazon.com/images/I/31JlPhp%...	The Kelby Training DVD Mastering Blend Modes i...	[[Electronics, Computers & Accessories, Cables...	Kelby Training DVD: Mastering Blend Modes in A...	NaN	NaN	NaN	NaN
1	0321732944	http://ecx.images- amazon.com/images/I/31uogm6Y...	NaN	[[Electronics, Computers & Accessories, Cables...	Kelby Training DVD: Adobe Photoshop CS5 Crash ...	NaN	NaN	NaN	NaN
2	0439886341	http://ecx.images- amazon.com/images/I/51k0qa8f...	Digital Organizer and Messenger	[[Electronics, Computers & Accessories, PDAs, ...	Digital Organizer and Messenger	8.15	{'Electronics': 144944}	{'also_viewed': ['0545016266', 'B009ECM8QY', '...']}	NaN

Fig 2: Sample Product Meta Dataset

Each row corresponds to product and includes the following variables:

asin: ID of the product, e.g. 0000031852

title: name of the product

price: price in US dollars (at time of crawl)

imUrl: url of the product image

related: related products (also bought, also viewed, bought together, buy after viewing)

salesRank: sales rank information

brand: brand name

categories: list of categories the product belongs to

DATA WRANGLING:

Merging Dataframes:

Product reviews and meta datasets in json files were saved in different dataframes and two dataframes were merged together using left join and “asin” was kept as common merger. Final merged data frame description is shown below:

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1689188 entries, 0 to 1689187
Data columns (total 17 columns):
asin                1689188 non-null object
helpful            1689188 non-null object
Rating             1689188 non-null int64
reviewText         1689188 non-null object
reviewTime         1689188 non-null object
reviewerID         1689188 non-null object
reviewerName       1664458 non-null object
summary            1689188 non-null object
unixReviewTime     1689188 non-null int64
imUrl              1687975 non-null object
description         1655511 non-null object
categories          1689188 non-null object
title              1643686 non-null object
price              1639882 non-null float64
salesRank           810070 non-null object
related            1662142 non-null object
brand              954251 non-null object
dtypes: float64(1), int64(2), object(14)
memory usage: 232.0+ MB

```

Fig 3: Merged dataframe info

In order to reduce time consumption for running models, only headphones products were chosen and the following method was adopted.

1. Out of 1689188 rows, 45502 rows were null values in product title. Those rows were dropped.
2. Dataset with product title named “Headphones”, “Headphones”, ”headphones”, ”headphone” were extracted from merged dataframe. Final headphones dataset was 64305 rows (observations).

Handling Duplicates, Missing Values:

1. 22699 rows in brand column were observed as null values. To solve this, brand name was extracted from title and replaced null values in brand.
2. Dropped missing values in “reviewerName”, ”price”, ”description”, ”related” were dropped.
3. “reviewText” and “summary” were concatenated and was kept under review_text feature
4. Helpful feature was split into positive and negative feedback.

5. Ratings greater than or equal to 3 was categorized as “good” and less than 3 was classified as “bad”.
6. Helpfulness ratio was calculated based on pos feedback/total feedback for that review
7. Dropped duplicates based on “asin”, “reviewerName”, “unixReviewTime”. After dropping duplicates, the dataset consisted 61129 rows and 18 features as shown in Fig 4.
8. ReviewTime was converted to datetime '%m %d %Y' format.
9. Columns were renamed for clarity purpose.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 61129 entries, 1260 to 1689187
Data columns (total 18 columns):
product_id      61129 non-null object
rating          61129 non-null int64
reviewer_id     61129 non-null object
reviewer_name   61129 non-null object
unix_review_time 61129 non-null int64
url             61129 non-null object
description     61129 non-null object
categories      61129 non-null object
product_title   61129 non-null object
price           61129 non-null float64
related         61129 non-null object
brand_name      61129 non-null object
review_text     61129 non-null object
pos_feedback    61129 non-null int64
neg_feedback    61129 non-null int64
rating_class    61129 non-null object
help_prop       61129 non-null float64
review_time     61129 non-null datetime64[ns]
dtypes: datetime64[ns](1), float64(2), int64(4), object(11)
memory usage: 8.9+ MB
```

Fig 4: Description of headphones dataset

Descriptive Statistics:

The following summary statistics was obtained

Number of reviews: 61129

Number of unique reviewers: 42062

Number of unique products: 1878

Average rating score: 4.056

Average helpful ratio score: 0.351

Number of positive feedback: 25222

Number of negative feedback: 14202

Number of good ratings: 52807

Number of Bad ratings: 8322

Ratings Number of Reviews

1 3870

2 4452

3 7102

4 14639

5 31066

The summary statistics for headphones dataset is shown below:

	rating	pos_feedback	neg_feedback	help_prop
count	61129.000000	61129.000000	61129.000000	61129.000000
mean	4.056438	3.082269	0.641709	0.350887
std	1.217506	31.516073	2.777990	0.442493
min	1.000000	0.000000	0.000000	0.000000
25%	3.000000	0.000000	0.000000	0.000000
50%	5.000000	0.000000	0.000000	0.000000
75%	5.000000	1.000000	0.000000	0.947368
max	5.000000	3800.000000	144.000000	1.000000

Fig 5. Summary Statistics

Preprocessing Text:

Since, text is the most unstructured form of all the available data, various types of noise are present in it and the data is not readily analyzable without any pre-processing. The entire process of cleaning and standardization of text, making it noise-free and ready for analysis is known as text preprocessing. In this section, the following text preprocessing were applied.

1. Removing HTML tags

HTML tags which typically does not add much value towards understanding and analyzing text. HTML words were removed from text.

2. Removing accented characters

Accented characters/letters were converted and standardized into ASCII characters.

3. Expanding Contractions

Contractions are shortened version of words or syllables. They exist in either written or spoken forms. Shortened versions of existing words are created by removing specific letters and sounds. In case of English contractions, they are often created by removing one of the vowels from the word.

By nature, contractions do pose a problem for NLP and text analytics because, to start with, we have a special apostrophe character in the word. Ideally, we can have a proper mapping for contractions and their corresponding expansions and then use it to expand all the contractions in our text.

4. Removing Special Characters

One important task in text normalization involves removing unnecessary and special characters. These may be special symbols or even punctuation that occurs in sentences. This step is often performed before or after tokenization. The main reason for doing so is because often punctuation or special characters do not have much significance when we analyze the text and utilize it for extracting features or information based on NLP and ML.

5. Lemmatization

The process of lemmatization is to remove word affixes to get to a base form of the word. The base form is also known as the root word, or the lemma, will always be present in the dictionary.

6. Removing stopwords

Stopwords are words that have little or no significance. They are usually removed from text during processing so as to retain words having maximum significance and context. Stopwords are usually words that end up occurring the most if you aggregated any corpus of text based on singular tokens and checked their frequencies. Words like a, the, me, and so on are stopwords.

7. Building a Text Normalizer

Based on the functions which we have written above and with additional text correction techniques (such as lowercase the text, and remove the extra newlines, white spaces, apostrophes), we built a text normalizer in order to help us to preprocess the new_text document.

After applying text normalizer to 'the review_text' document, we applied tokenizer to create tokens for the clean text. As a result of that, we had 3070479 words in total. Eventually, after completing all data wrangling and preprocessing phases, we save the dataframe to csv file as a 'Cleaned_Reviews_Home_and_Kitchen.csv'. After cleaning, we have 25276 observations.

A clean dataset will allow a model to learn meaningful features and not overfit on irrelevant noise. After following these steps and checking for additional errors, we can start using the clean, labelled data to train models in modeling section.

EXPLORATORY DATA ANALYSIS (EDA):

After collecting data, wrangling data then exploratory analyses were carried out. The following insights were explored through exploratory analyses.

- Predicting ratings based on reviews
- Usefulness on large volume of reviews
- Rating vs number of reviews
- Rating vs proportion of reviews
- Helpful proportion vs Number of reviews
- Rating vs helpfulness ratio
- Top 20 most reviewed products
- Bottom 20 reviewed products
- Positive and negative words
- Word cloud for different ratings, brand name etc

1. Top 20 Most Reviewed Brands

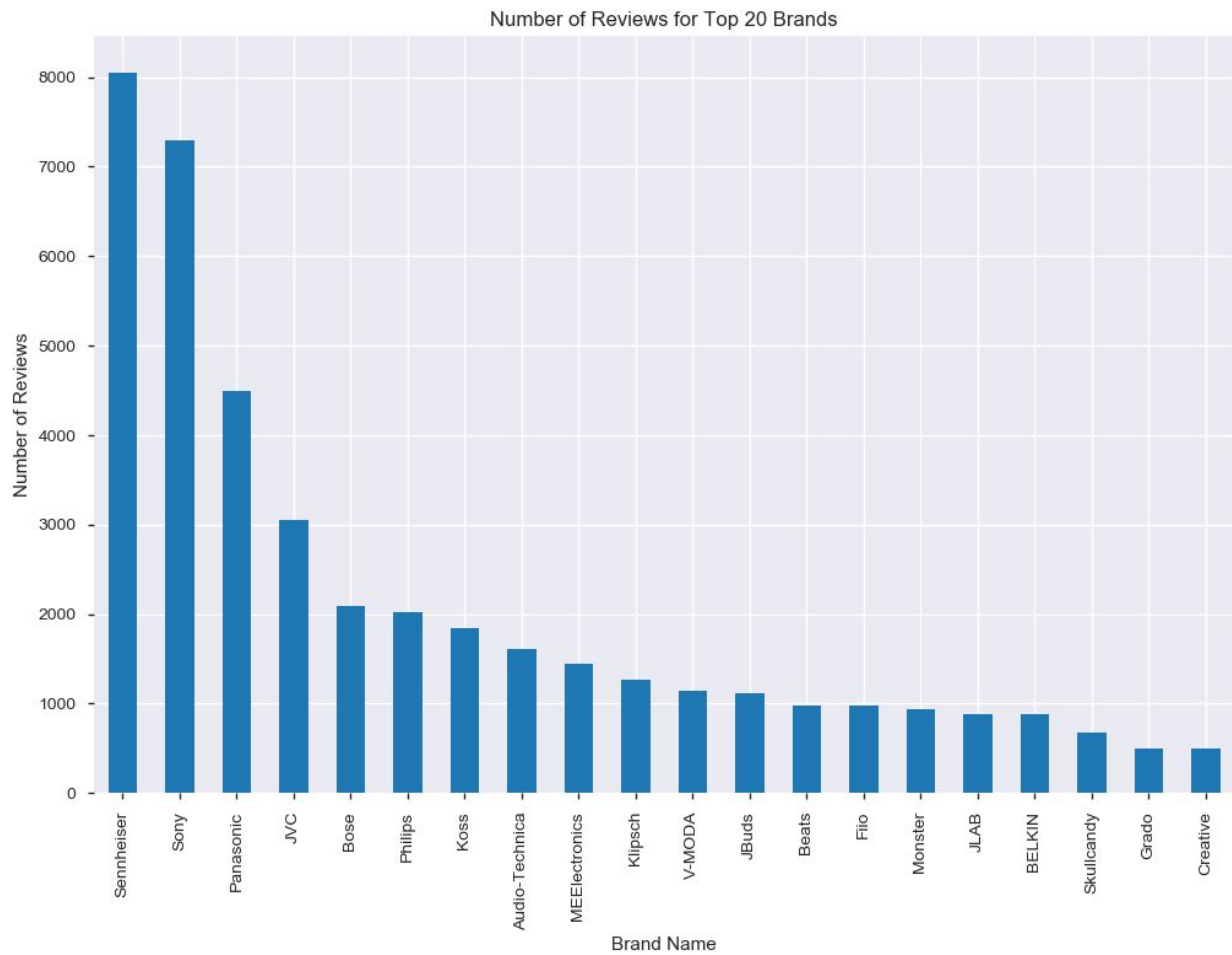


Fig 6. Top 20 most reviewed brands

2. Top 20 Least Reviewed Brands

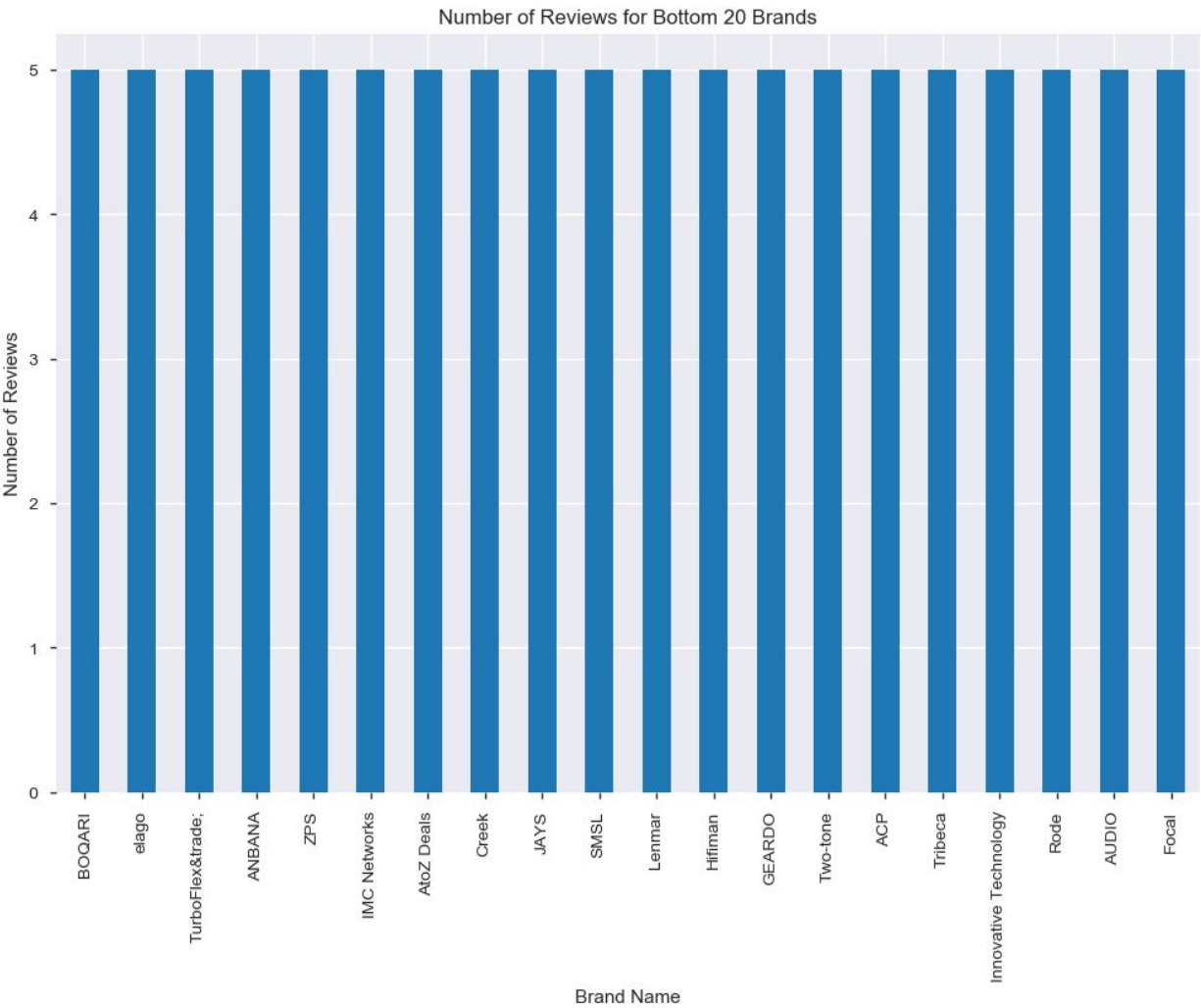


Fig 7. Top 20 least reviewed brands

3. Top 20 Most Reviewed Products

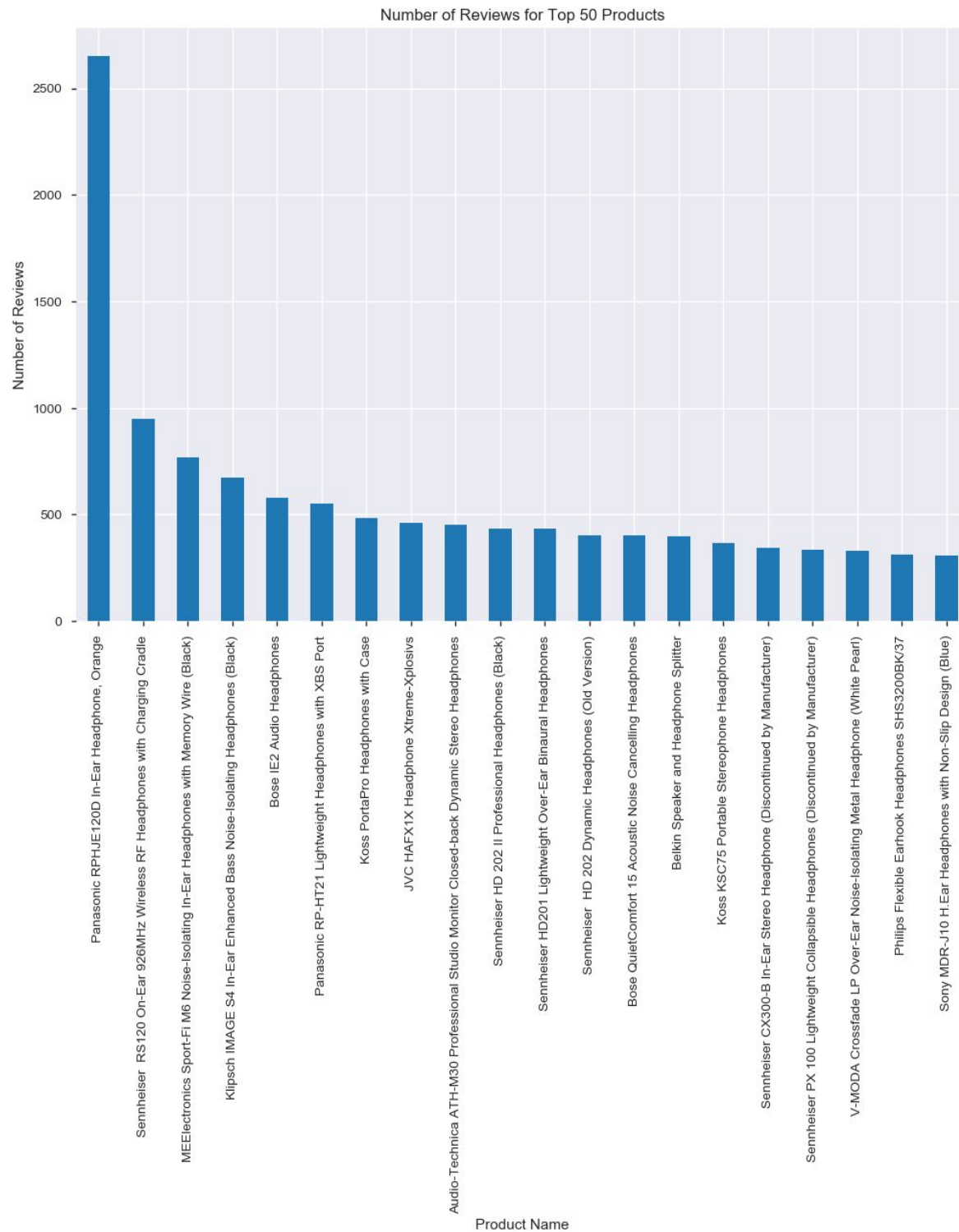


Fig 8. Top 20 most reviewed products

4. Most Positively Reviewed Headphone

Below product was the most reviewed product in Amazon under headphones category. Product name is “Panasonic ErgoFit In-Ear Earbud Headphones RP-HJE120-D (Orange) Dynamic Crystal Clear Sound, Ergonomic Comfort-Fit”. This product had overall good rating more than 3.



Fig 9a. Most reviewed headphone

5. How the above product will go up after some years?

The panasonic earbud headphone had overall positive review from 2010 onwards. Fig 9b shows the distribution of rating over a period of time. This product had overall good mean rating more than 4.

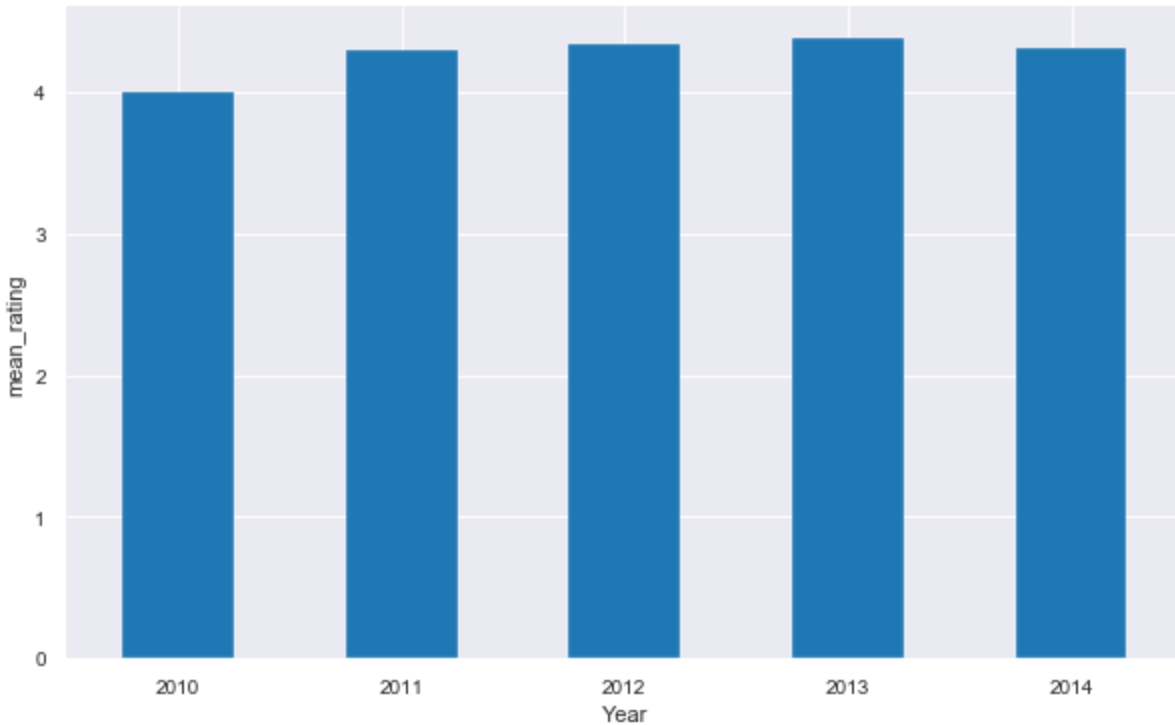


Fig 9b. Year vs mean rating for Panasonic earbud headphone

Fig 9c shows the word cloud from good rating reviews for the above product. It shows major insight in terms of sellers perspective. **It indicates most of the positive customers agree with “great fit”, “good price” and least with “sound quality”.** Similarly, Fig 9d shows The shows the word cloud from bad rating reviews for the above product. **It indicates most of the customers agree with “poor quality” and “terrible sound”. From the sellers perspective, this product needs to be updated with “better sound” and “quality” in order to get positive feedback from customers.**



Fig 9c. Insight words from good rating reviews for Panasonic headphone

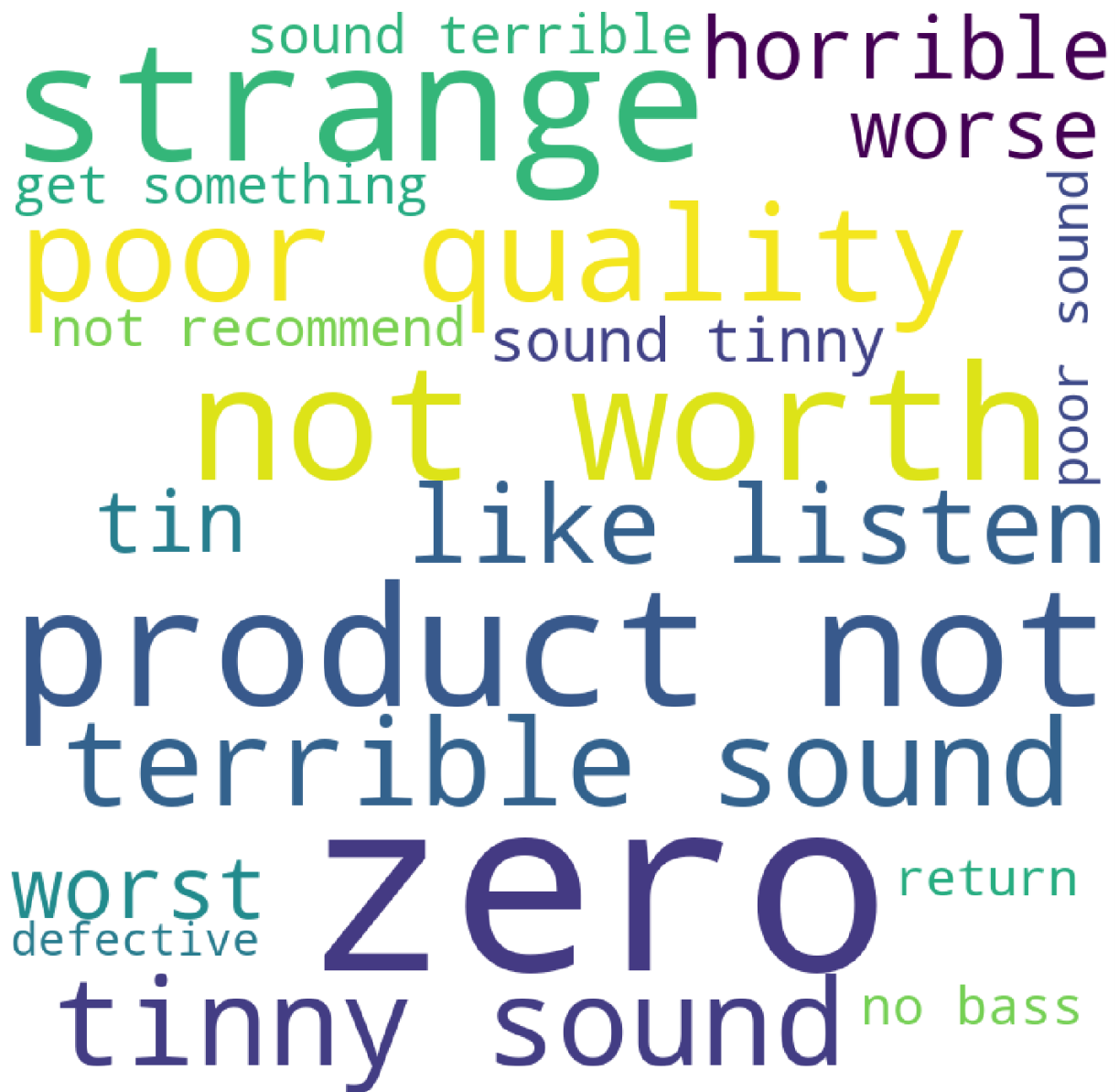


Fig 9d. Insight words from bad rating reviews for Panasonic headphone

6. Most Negatively Reviewed Headphone

Below product was the most negatively reviewed product in Amazon under headphones category. Product name is “My Zone Wireless Headphones”. This product had overall bad rating less than 3.



Fig 10a. Least reviewed headphone

7. How the above product will go up after some years?

My zone wireless headphone had overall negative review from 2010 onwards except 2012. Fig 10b shows the distribution of rating over a period of time. This product had overall bad mean rating of around 2.5.

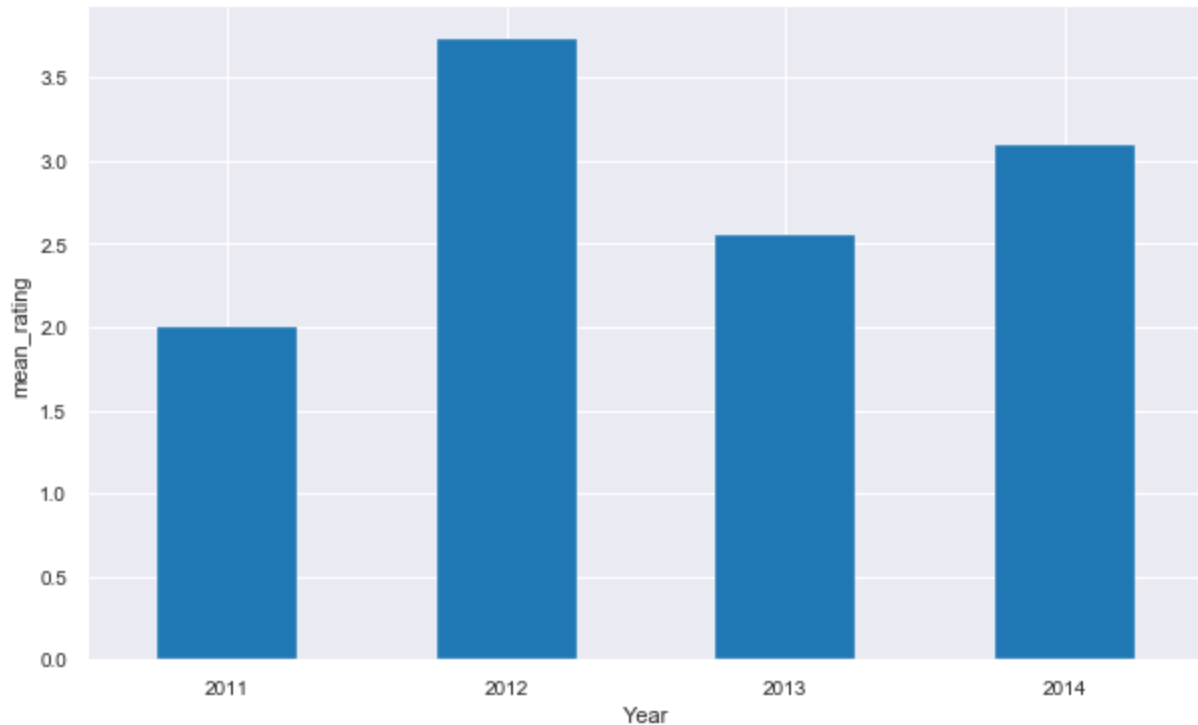


Fig 10b. Year vs mean rating for My zone wireless headphone

Fig 10c shows the word cloud from good rating reviews for the above product. It shows major insight in terms of sellers perspective. **It indicates most of the positive customers agree with “easy setup”, “work with TV” and least agree with “work great”**. Similarly, Fig10d shows The shows the word cloud from bad rating reviews for the above product. **It indicates most of the customers agree with “battery issue” and “horrible reception” and “static interference”**. From the sellers perspective, this product needs to be updated with “good quality battery”, “reception issue” and “static issue” in order to get positive feedback from customers.

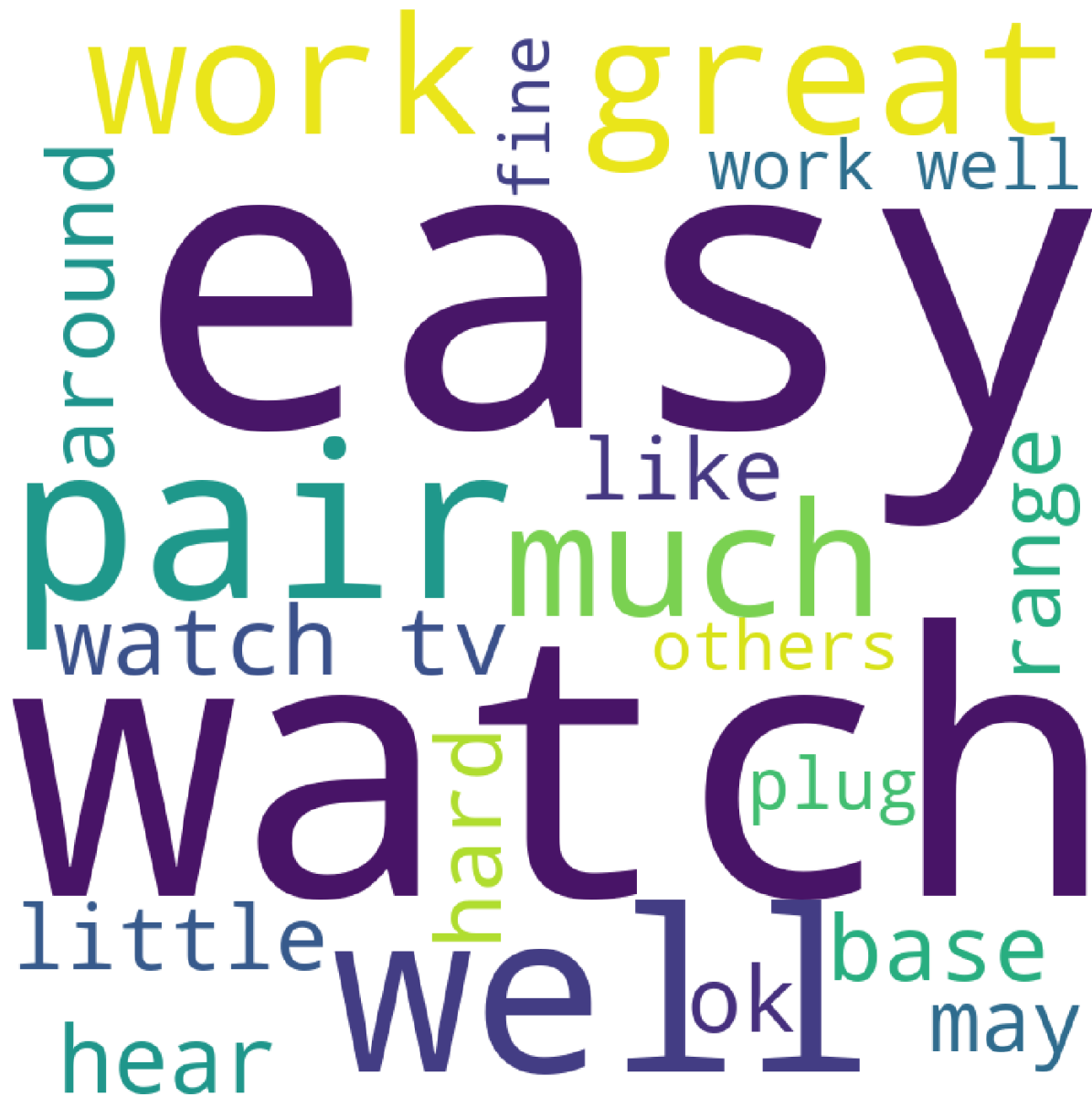


Fig 10c. Insight words from good rating reviews for My zone wireless headphone

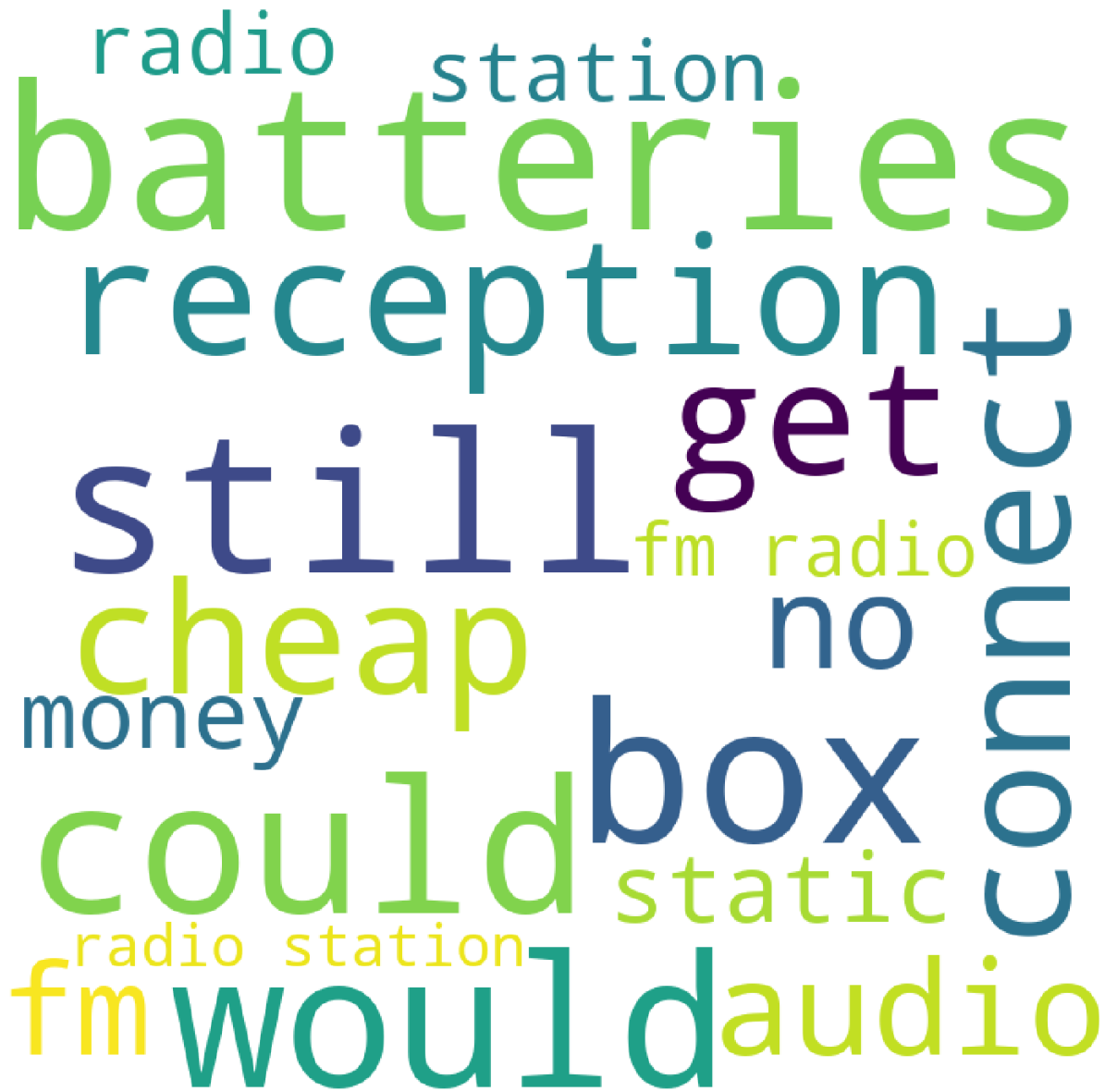


Fig 10d. Insight words from bad rating reviews for My zone wireless headphone

8. Which Rating got Highest Number of Reviews?

Customers have written reviews and ratings were given from 1 to 5 for headphones they bought from Amazon between 2000 to 2014. The distribution and percentage of ratings vs number of reviews is shown in Fig.11. Number of reviews for rating 5 were high compared to other ratings. Overall, customers were happy about the products they purchased. About 50% customers gave 5 rating for the products they purchased. Only 15% customers gave ratings less than 3 (Fig 12).

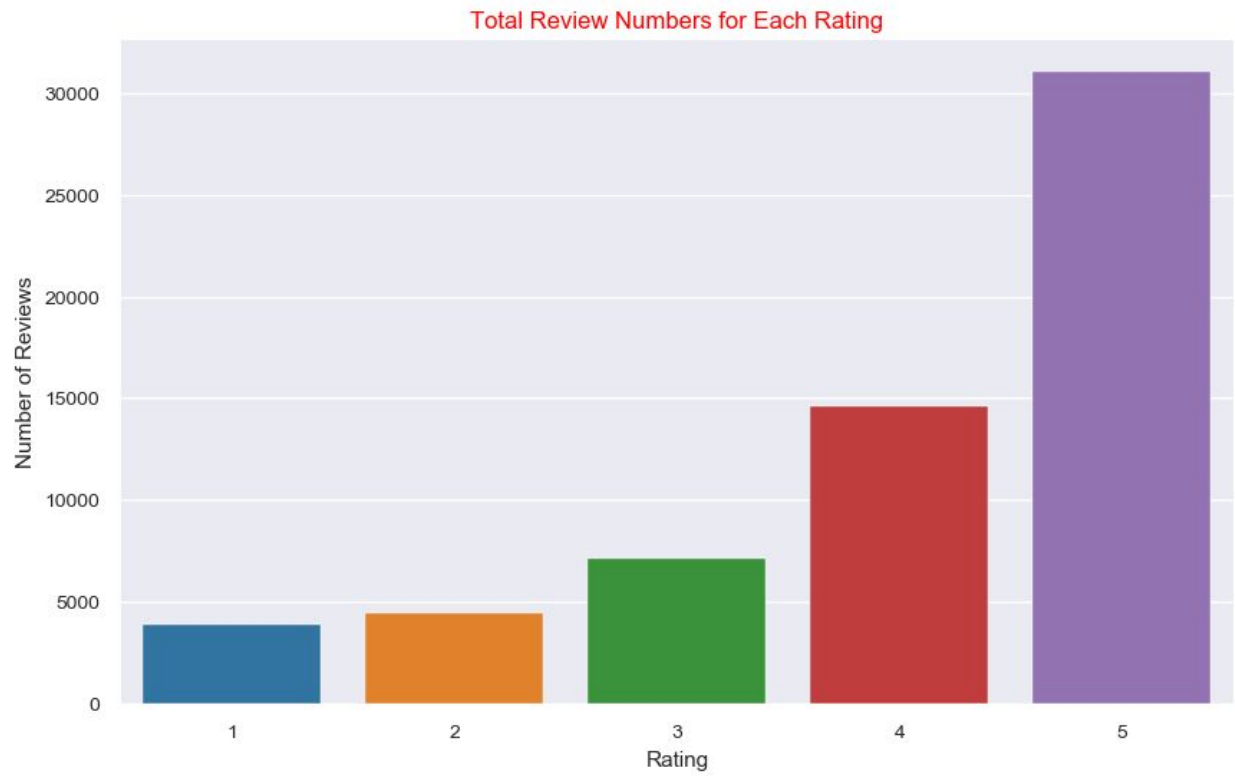


Fig 11. Rating Vs Number of reviews

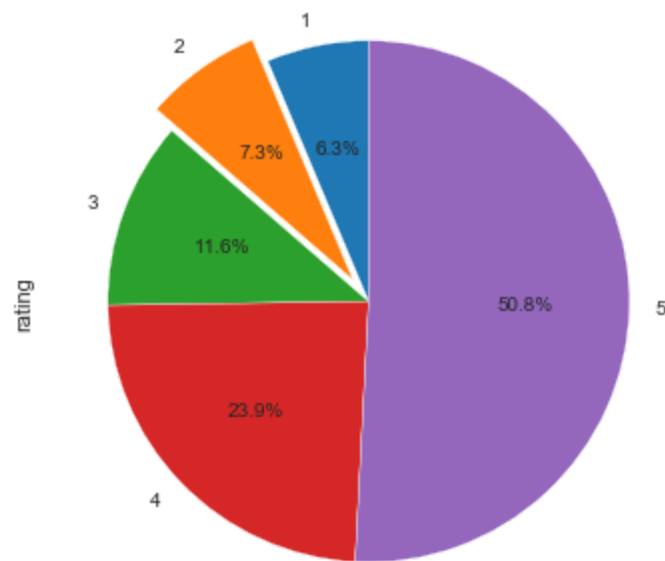


Fig 12. Distribution of ratings Vs Number of reviews

The ratings were divided into two categories. The rating below 3 were classified as “bad” and the remaining ratings were grouped as “good”. The distribution of rating class vs number of reviews is shown in Fig 13. It indicates about 50000 reviews were identified as good rating.

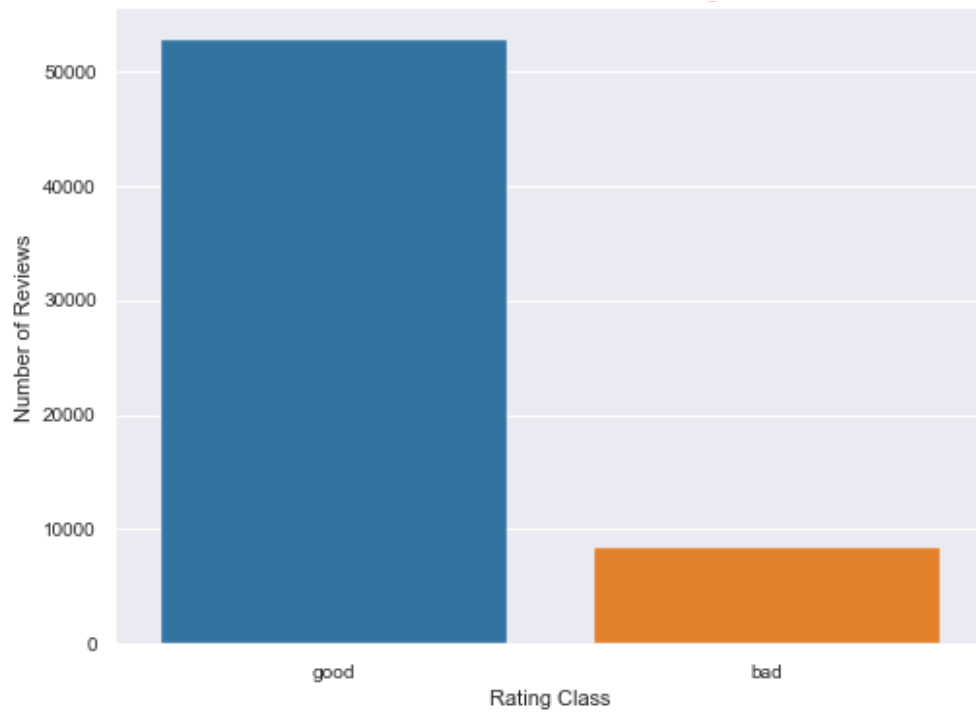


Fig 13. Total review numbers for each rating class

The distribution of ratings vs helpfulness ratio is shown in Fig 14. It indicates that all ratings have same helpfulness ratio.

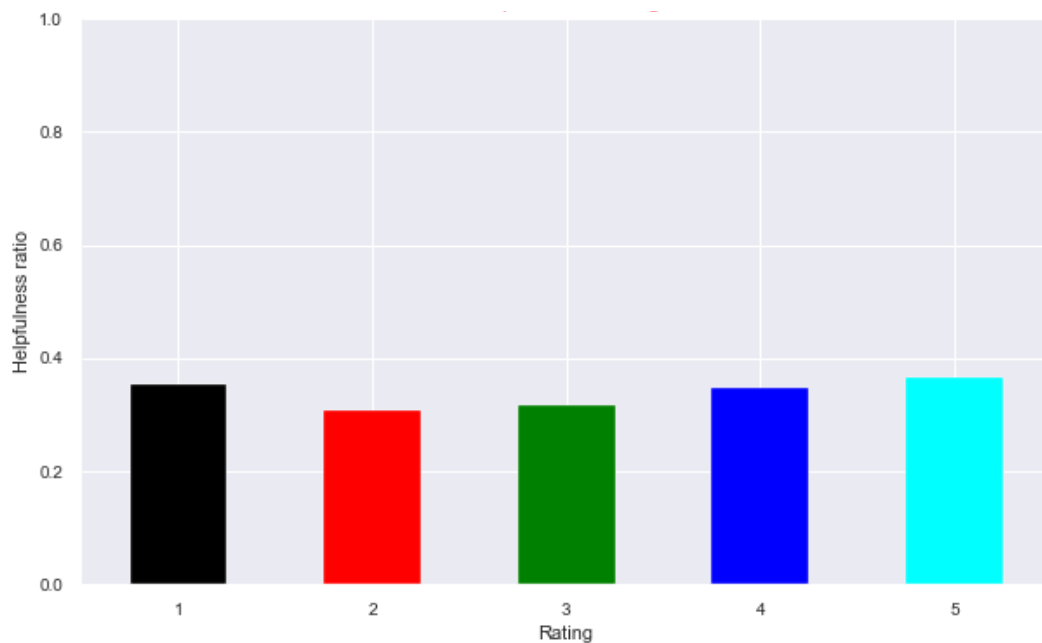


Fig 14. Helpfulness in rating

9. Which Rating got Highest Number of Review Length?

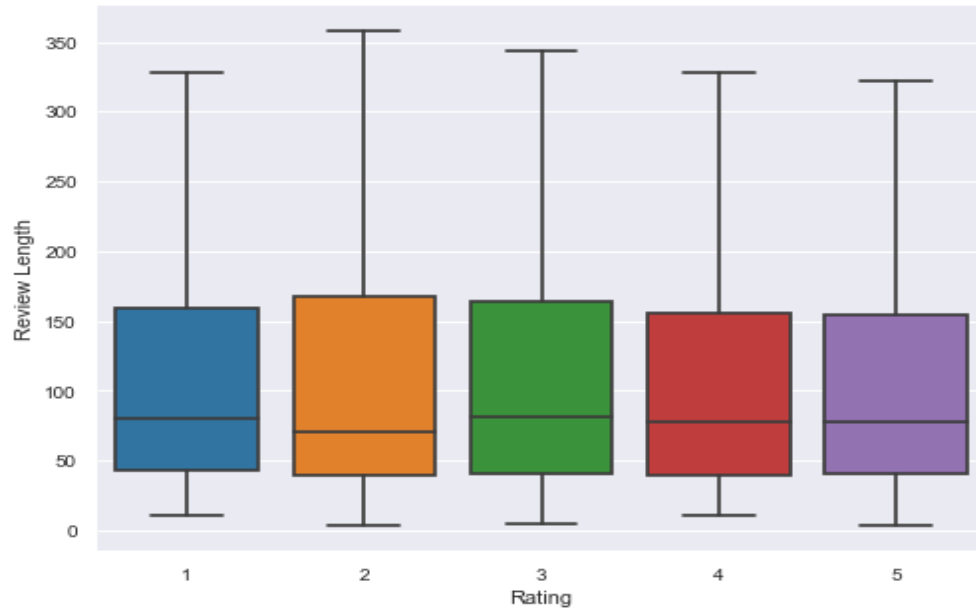


Fig 15. Rating vs Review length

10. Which Year has the Highest Number of Reviews?

Total review numbers for each year is shown in Fig 16. Number of reviews were low during 2000-2010. 2013 has the highest number of reviews.

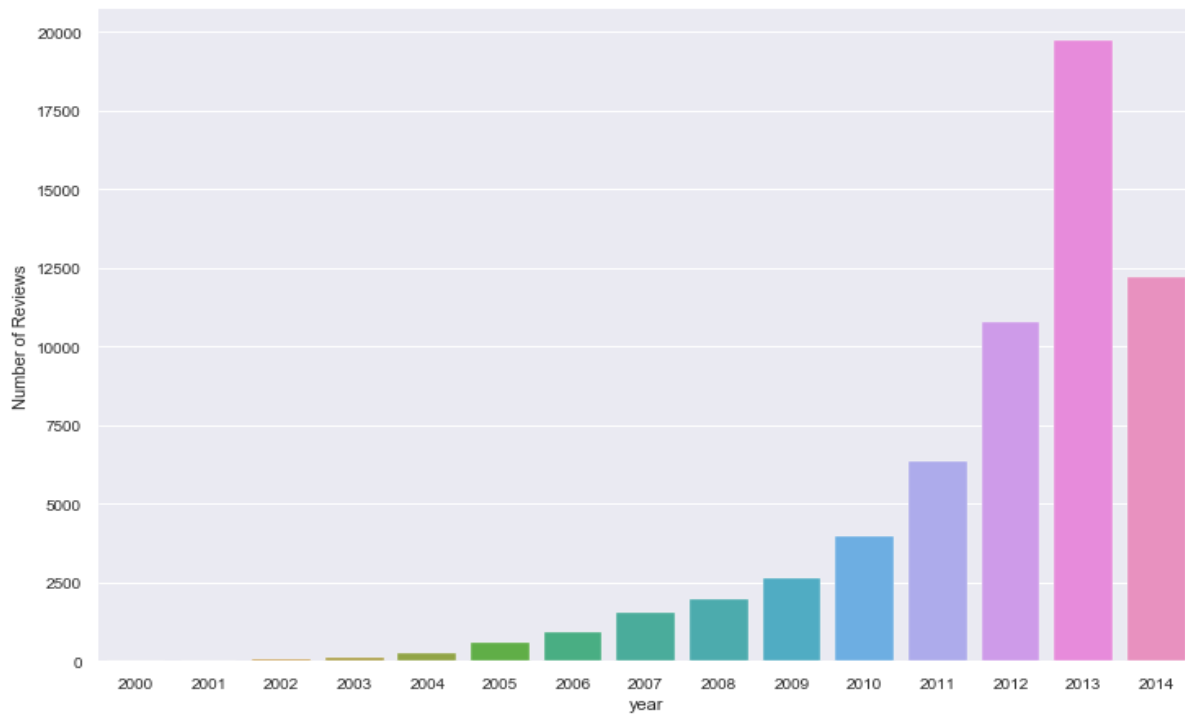


Fig 16. Rating vs Review length

11. Which Year has the Highest Number of Customers?

Total unique customers for each year is shown in Fig 17. Number of unique customers were low during 2000-2010. 2013 has the highest number of customers.

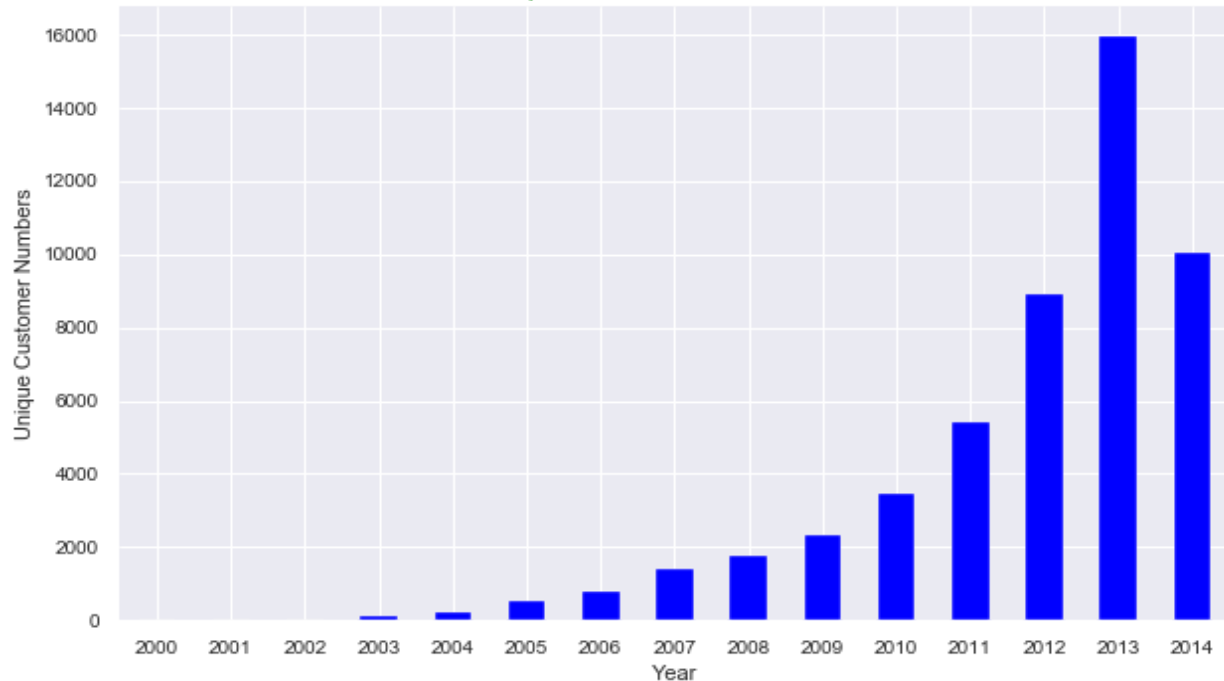


Fig 17. Unique customers in each year

12. Which Year has the Highest Number of Product?

Total unique product numbers for each year is shown in Fig 18. Number of unique products were low during 2000-2010. 2013 has the highest number of products.

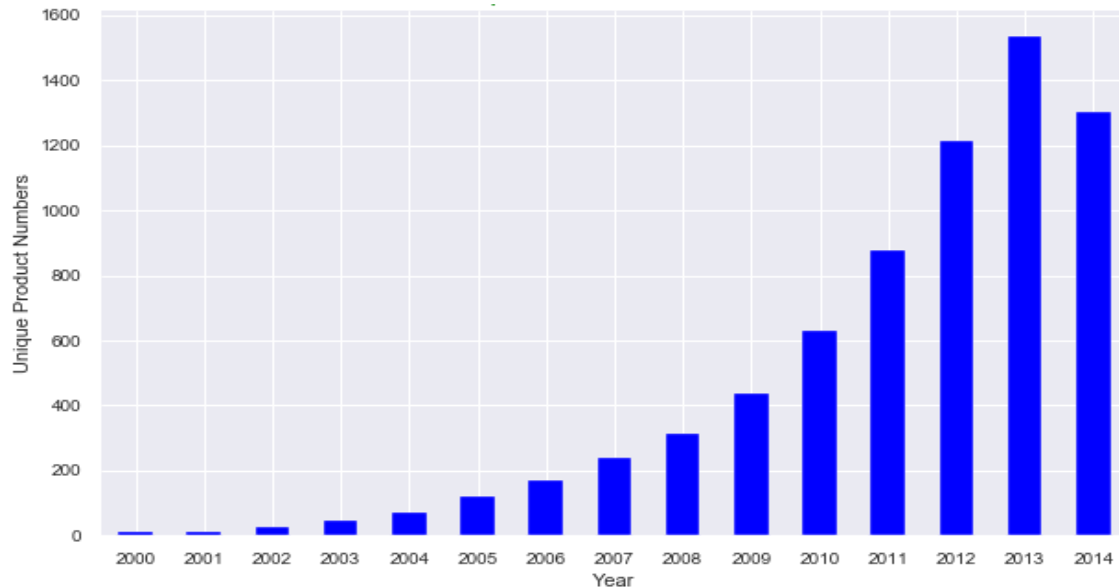


Fig 18. Unique products in each year

13. Which Year has the Highest Number of Product?

Except 2001, 'good ratings' percentage is progressing over 80%. 2001 has the lowest good ratings with 69% overall. 'good ratings' percentage is 90% in 2000. As it might be seen in the graph, the overall good rating is progressing between 81% and 90% in headphones products.

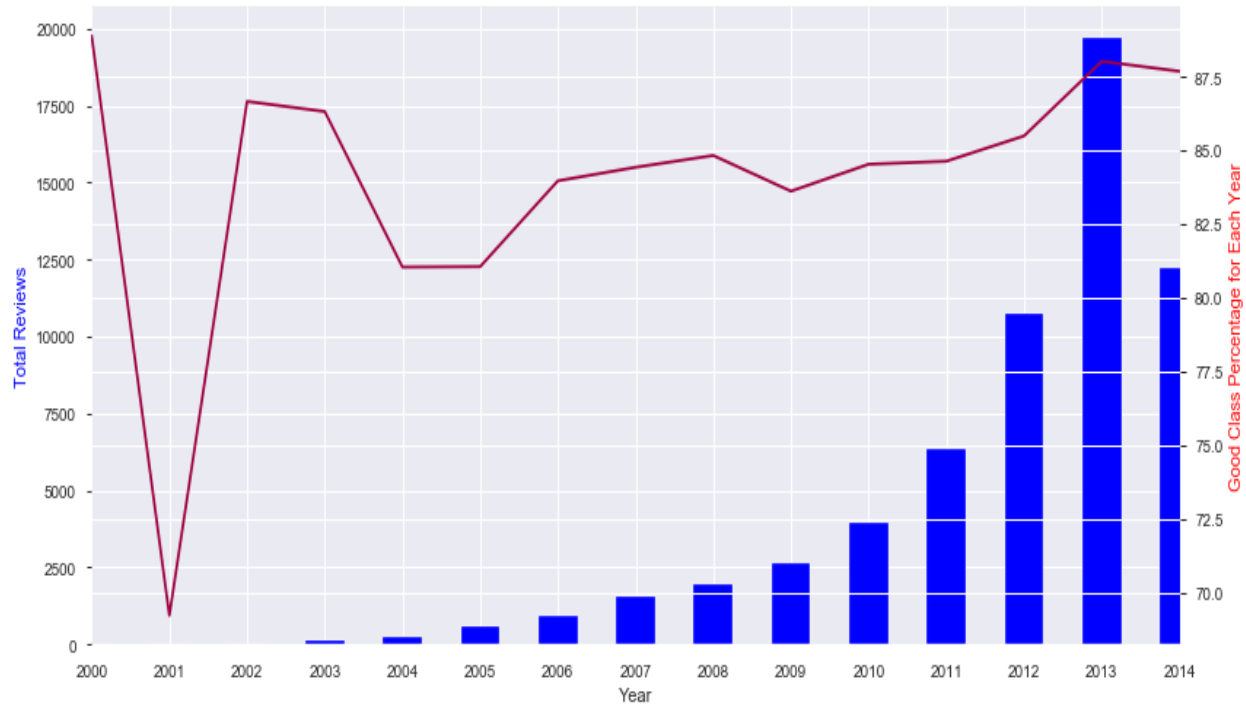


Fig 19. Distribution of good rating class over the period of time

14. Average Helpfulness of the Product

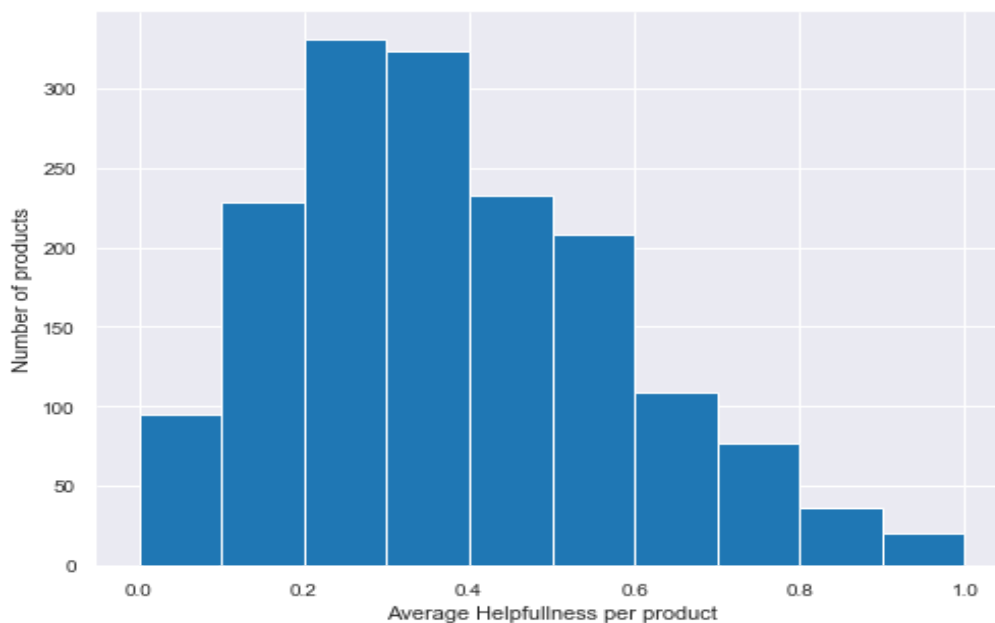


Fig 20. Distribution of average helpfulness per product

15. Distribution of Review Length

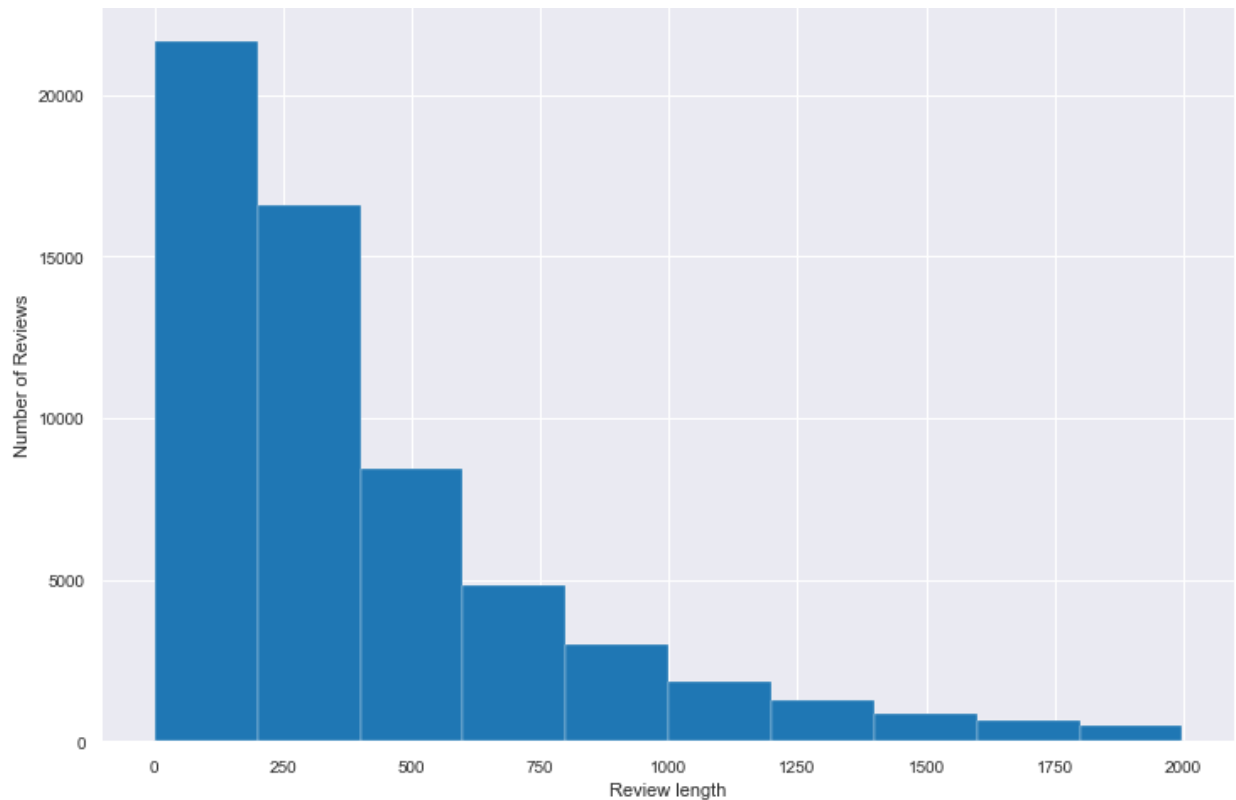


Fig 21. Distribution of review length

16. Which review length bin has the Highest Number of Good Rating?

As it might be seen in the Fig 22, the highest percentage of good rating reviews lies between 0-1000 words with 96 % whereas lowest percentage of good rating review lies between 1700-1800 words with 80%. As the review length extends, the good rating tends to increase. Generally, the customers who have write longer reviews (more than 1900 words) tends to give good ratings.

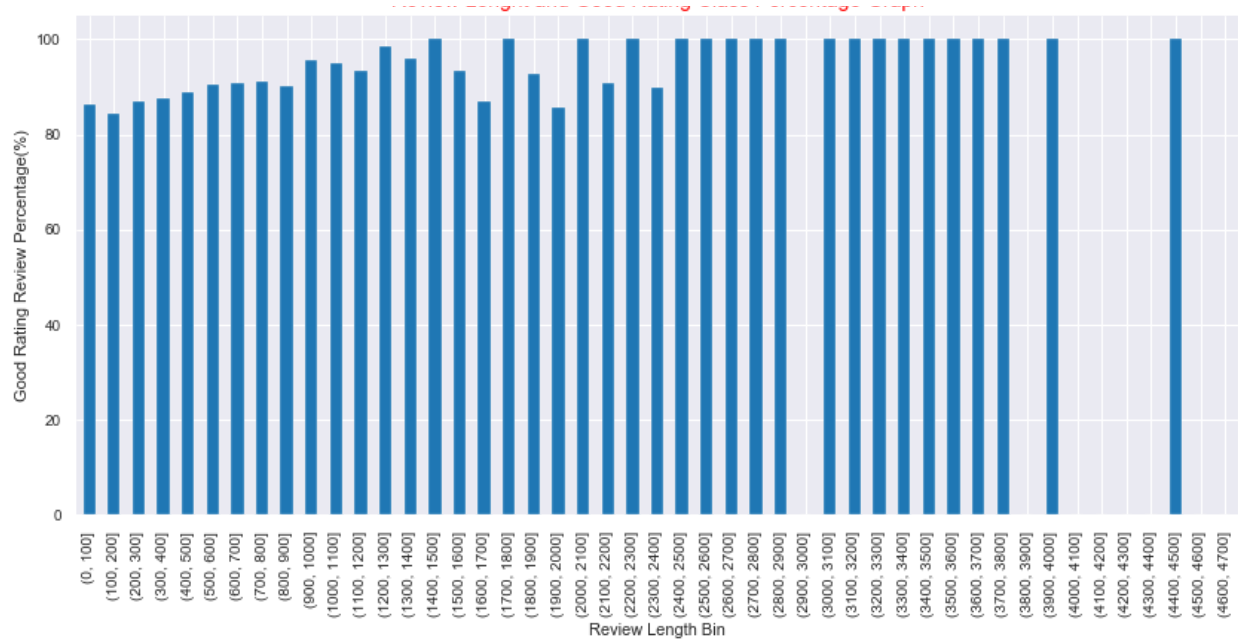


Fig 22. Distribution of review length vs good rating

17. Which review length bin has the Highest Number of Helpfulness Ratio?

As it might be seen in the Fig 23, the highest helpfulness ratio lies between 0-1200 words with 0.8 whereas lowest helpfulness ratio lies between 1200-1300 words with 0.6. As the review length extends, the helpfulness ratio tends to increase. Generally, the customers who have write longer reviews (more than 1300 words) tends to have high helpfulness ratio.

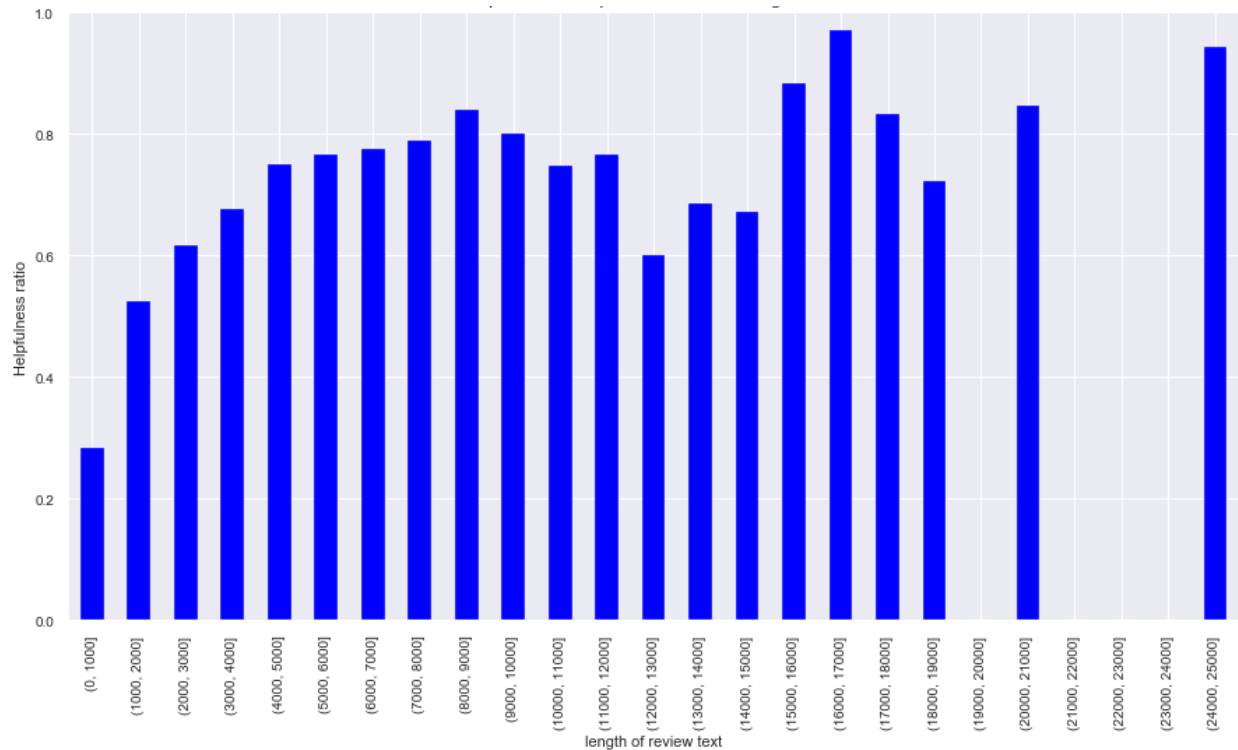


Fig 23. Distribution of review length vs helpfulness ratio

The frequency of review length for helpfulness and unhelpfulness is shown in Fig. 24. It indicates that overall helpfulness and unhelpfulness ratio were the same for larger review length. Unhelpfulness ratio were high in case of small length review.

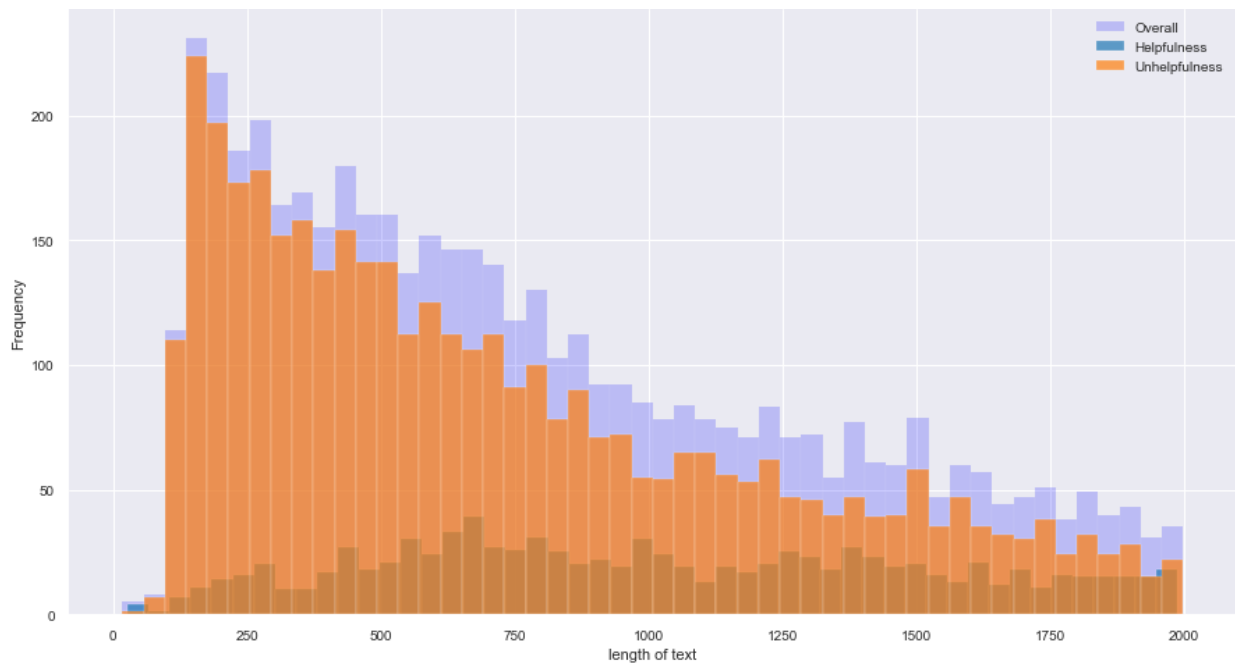


Fig 24. Distribution of review length vs helpfulness and unhelpfulness

18. Top Words for Good Rating?

The most common 50 words, which belong to good rating class, are shown in Fig 25. It shows the all good words from customers about the products.



Fig 25. Good rating words

19. Top Words for Bad Rating?

Similarly, the most common words, which belong to bad rating class, are shown in Fig 26. It shows all bad rating words from customers about the products.

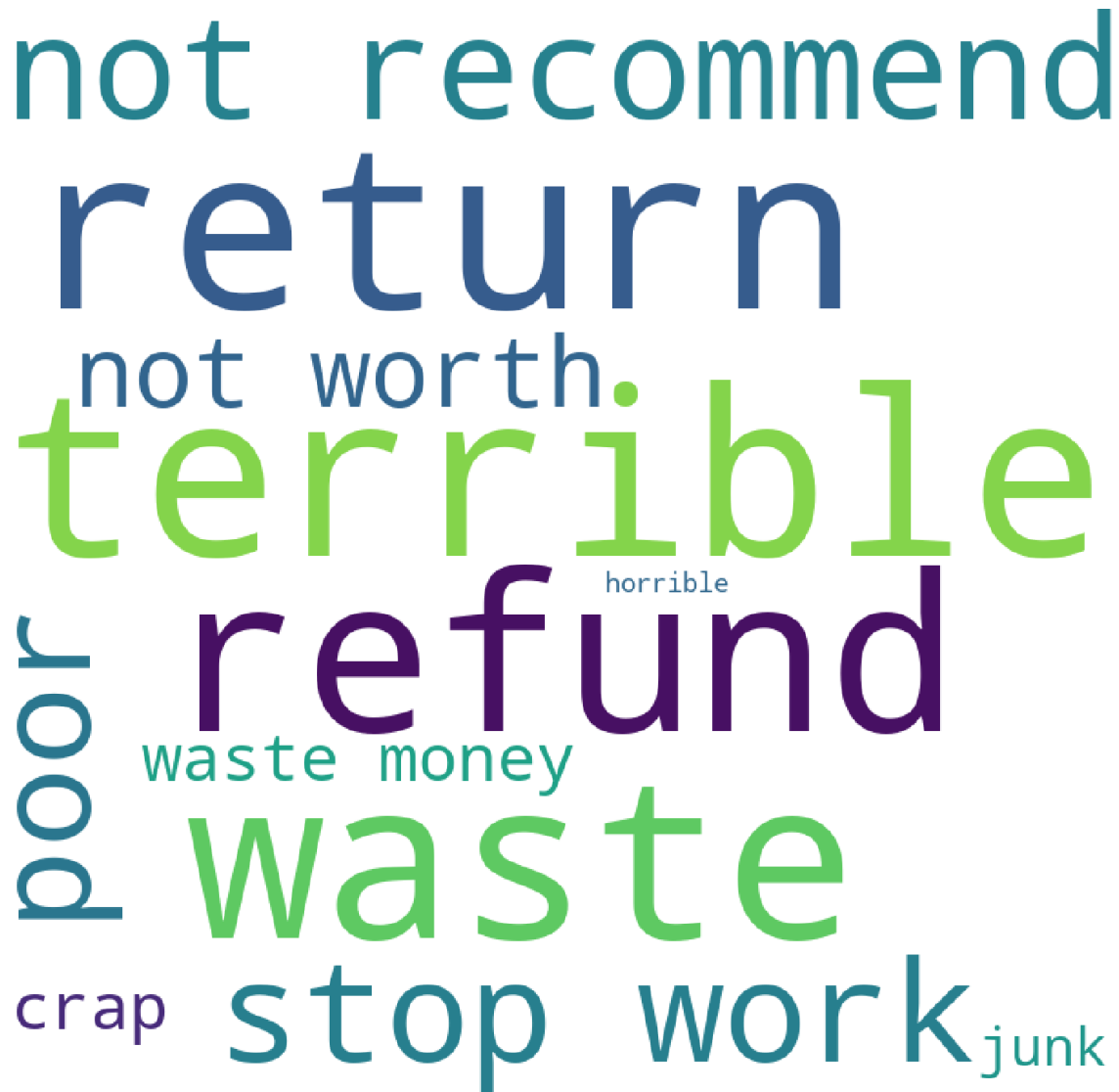


Fig 26. Bad rating words

Code:

Data Wrangling:

https://github.com/umaraju18/Capstone_project_2/blob/master/code/Amazon-Headphones_data_wrangling.ipynb

EDA:

https://github.com/umaraju18/Capstone_project_2/blob/master/code/Amazon-headphones_EDA.ipynb