Sentiment Analysis and Product Recommendation on Amazon's Electronics Dataset Reviews

Uma Maheswari Raju

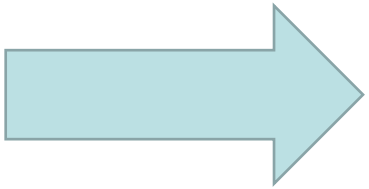
Overview

- Introduction
- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Modeling
- Summary
- Future Study

Introduction

Problem Statement

Can we predict whether a user liked a product or not based on their reviews?



Classification Model

Sentiment Analysis

Data Collection

Product Complete Reviews data:

- Electronics product reviews
- http://jmcauley.ucsd.edu/data/amazon/

	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

reviewerID : ID of the reviewer
asin : ID of the product
reviewerName : name of the reviewer
helpful : helpfulness of the review
reviewText : text of the review

overall : Rating (1,2,3,4,5)
summary : summary of the review
unixReviewTime : time of the review
reviewTime : time of the review

Data Collection

Product Metadata:

- **Electronics meta data**

	asin	imUrl	description	categories	title	price	salesRank	related	brand
0	0132793040	http://ecx.images-amazon.com/images/I/31JIPhp%...	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

asin: ID of the product

title: name of the product

price: price in US dollars

imUrl: url of the product image

related: related products

salesRank: sales rank information

brand: brand name

categories: list of categories

Data Wrangling

- *Merged Dataframes*
- *Extracted only headphones*
- *Brand name filled with first word of product description*
- *Dropped missing values*
- Concatenated “review text” and “summary”
- Helpful feature splitted into positive and negative
- Ratings greater than or equal to 3 categorized as “good” and less than 3 was classified as “bad”.
- ReviewTime converted to datetime '%m %d %Y format.

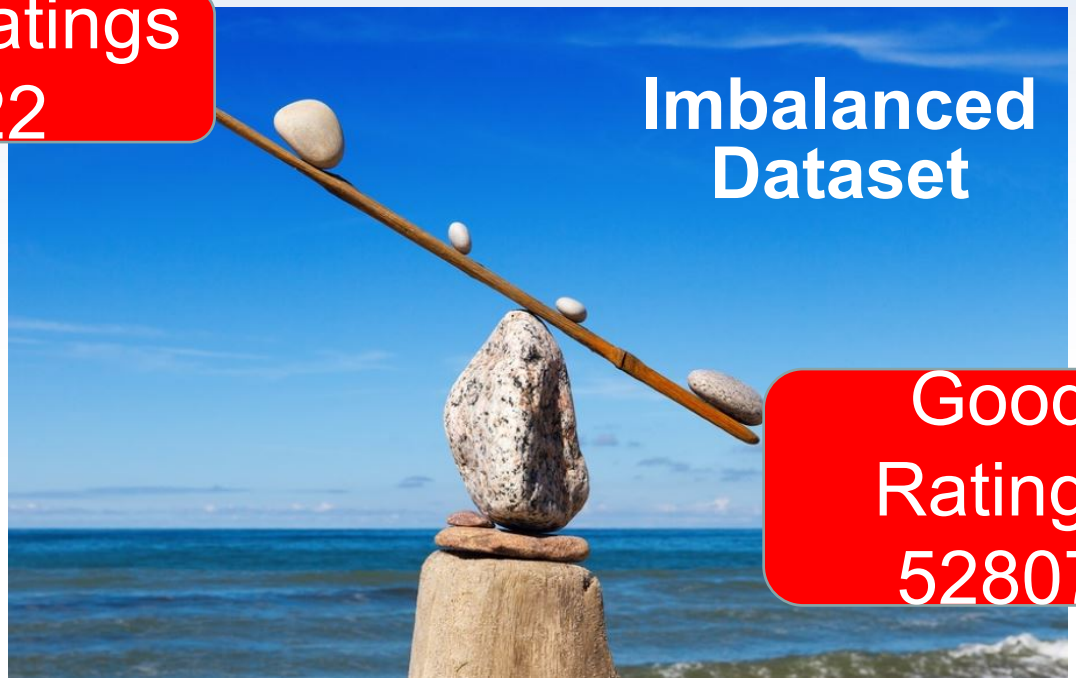
Data Wrangling

```
<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
```

Data Wrangling

Descriptive Statistics

Bad Ratings
8322



Good
Ratings
52807

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

Data Wrangling

Text Preprocessing

"review_text " Feature

- Remove tags
- Remove accented characters
- Expand contractions
- Remove special characters
- Lemmatisation versus Stemming
- Removing stopwords
- Tokenization
- Remove extra white space and digits
- Spelling corrections
- Grammatical Corrections
- Remove repeated characters
- Lower the text

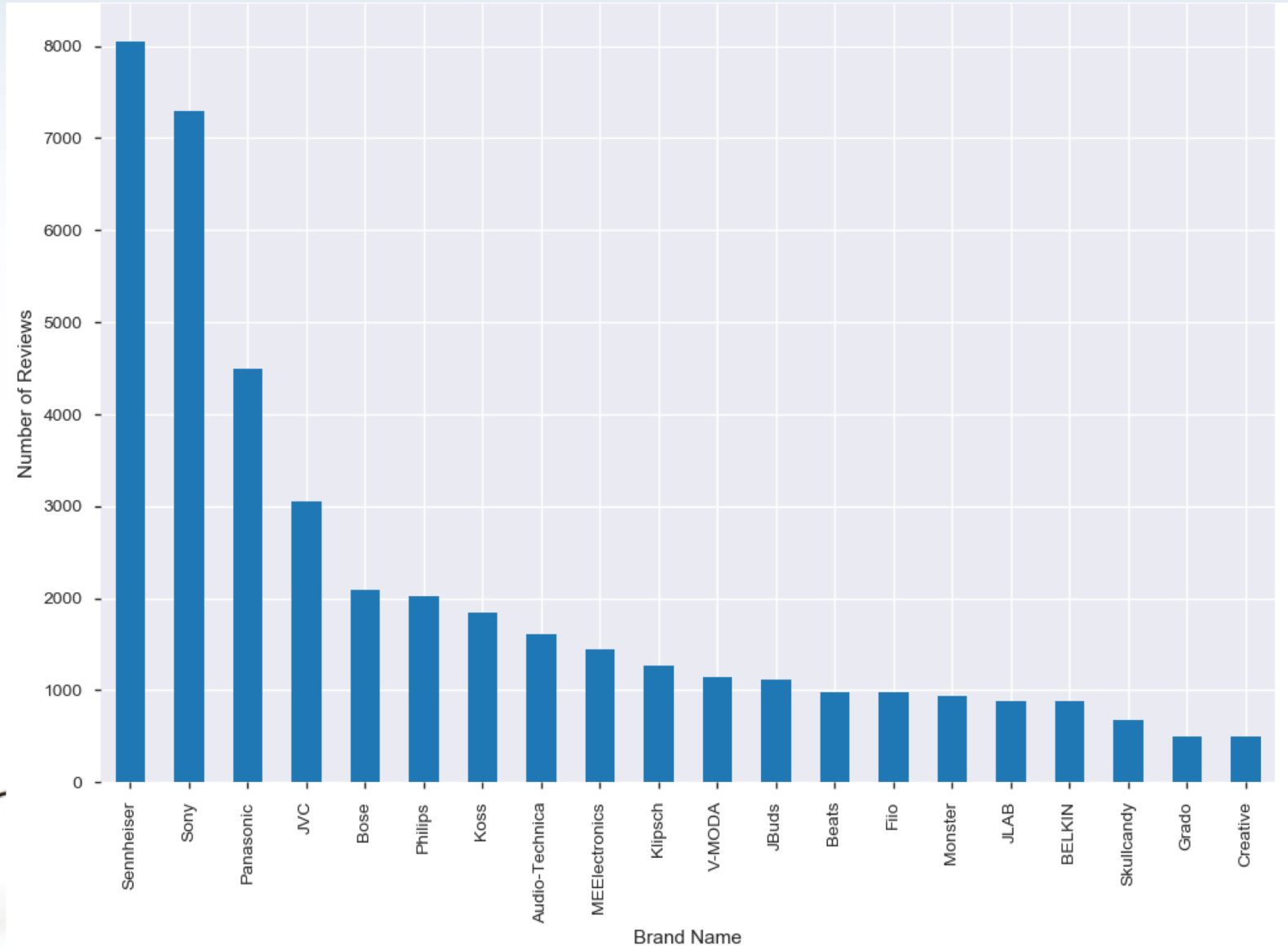
"clean_text" Feature

Exploratory Data Analysis (EDA)

- 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
- World cloud for different ratings, brand name etc

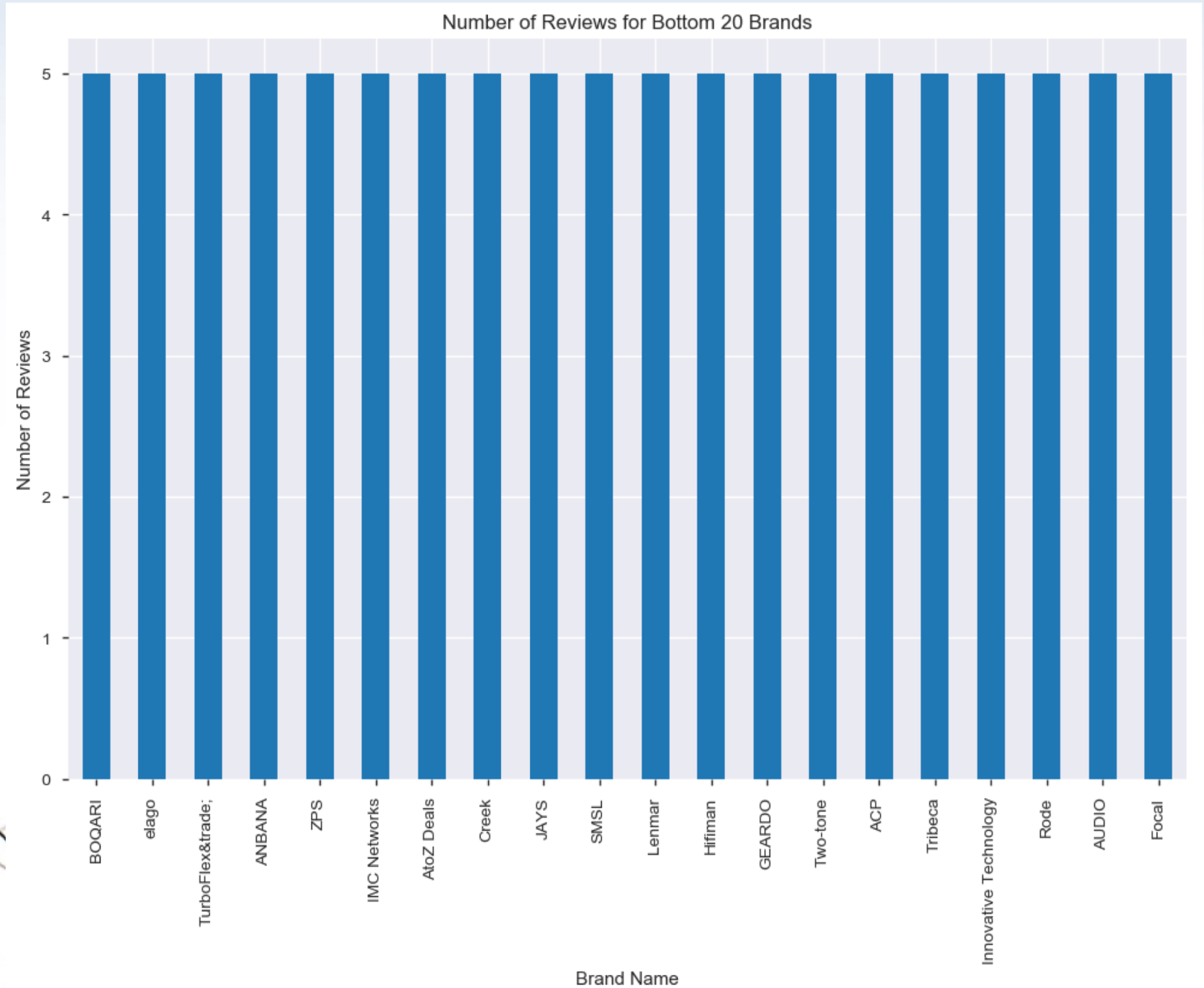
EDA

Top 20 Most Reviewed Brands



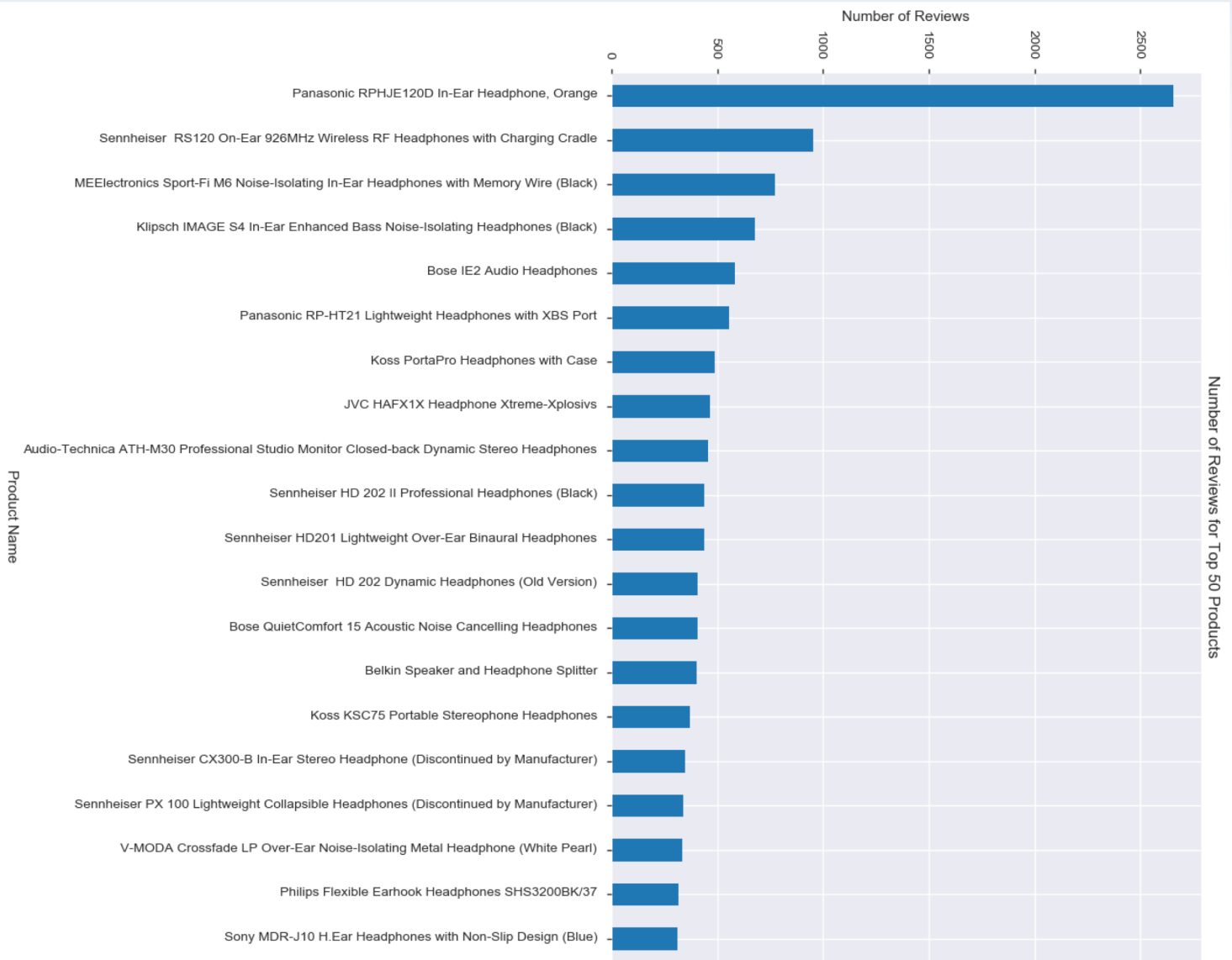
EDA

Top 20 Least Reviewed Brands



EDA

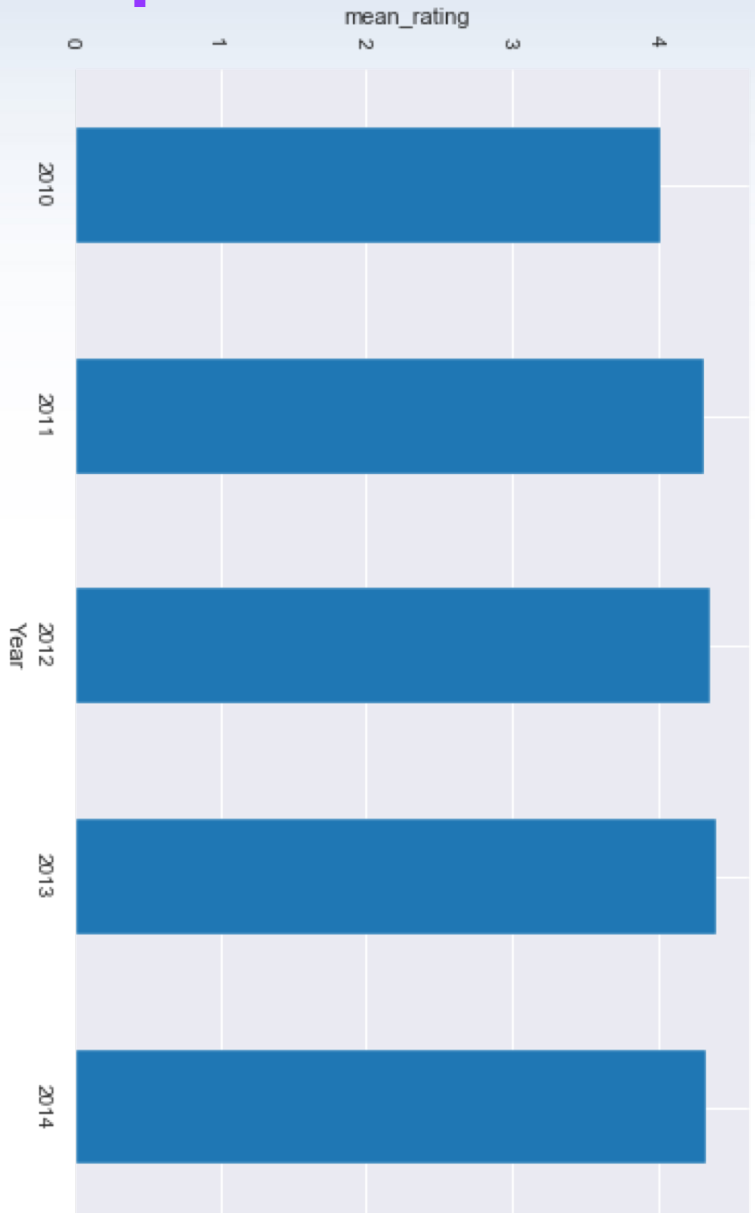
Top 20 Most Reviewed Products



EDA

Most Positively Reviewed Headphone

Panasonic ErgoFit In-Ear Earbud Headphones



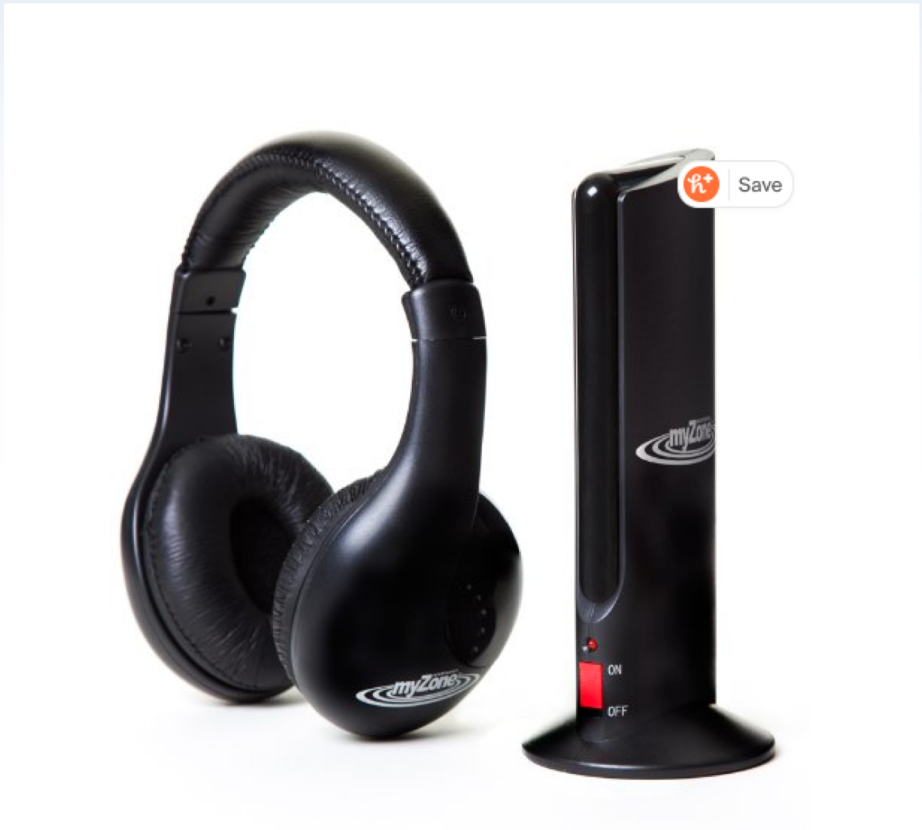
EDA

Top Positive Words

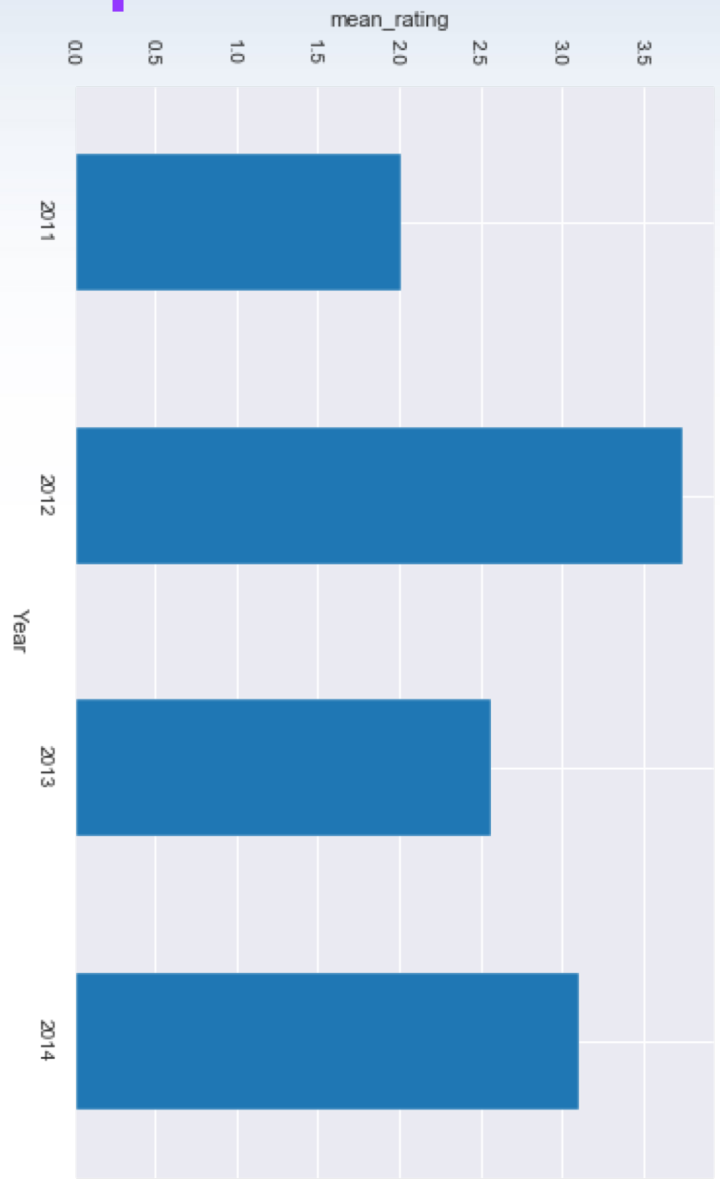
Top Critique Words

EDA

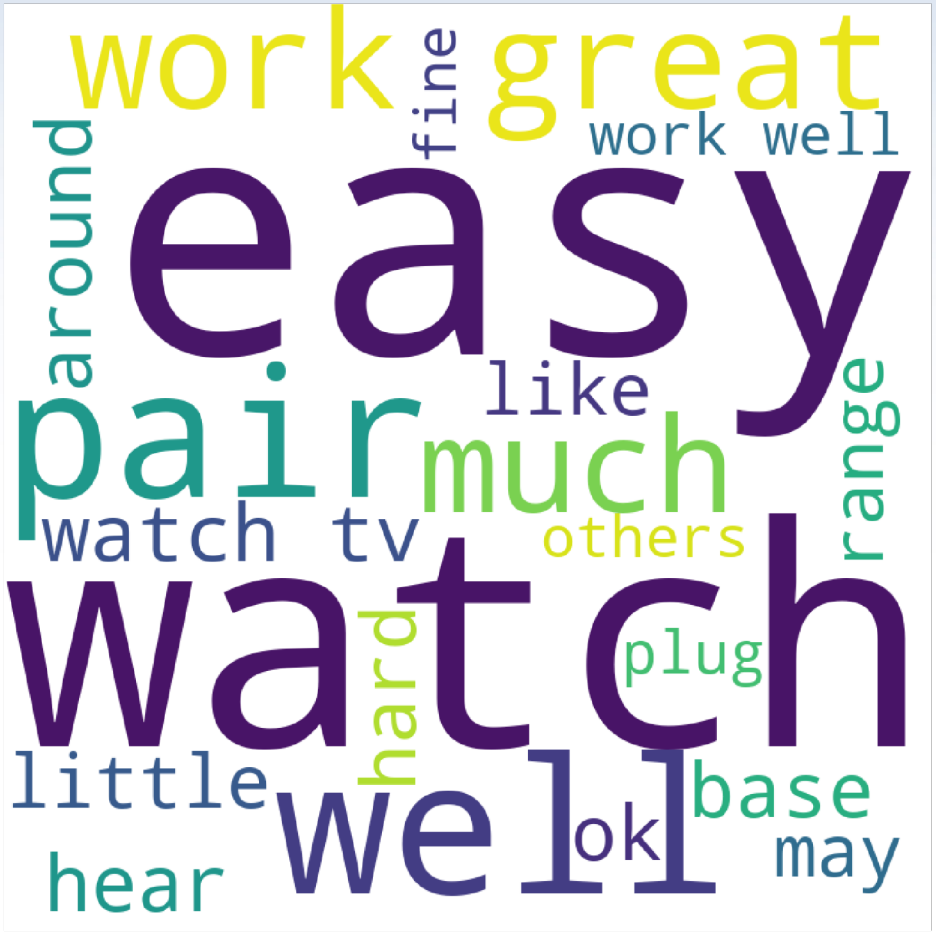
Most Negatively Reviewed Headphone

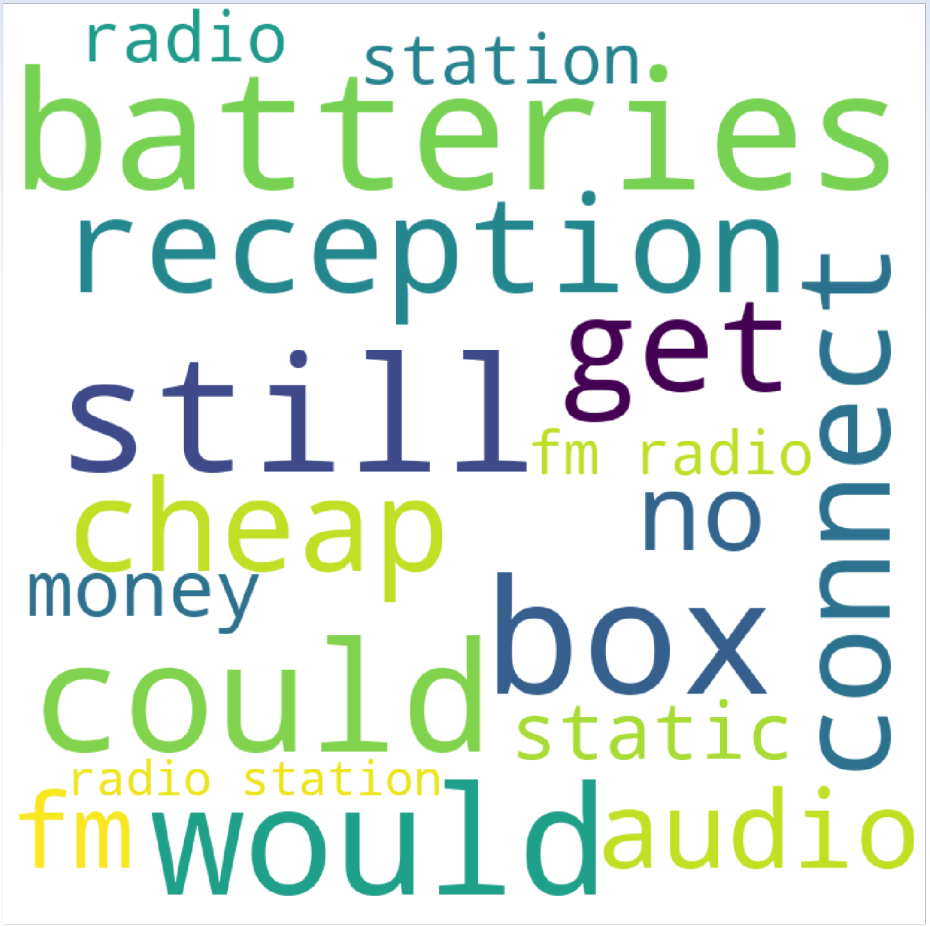


My Zone Wireless headphone



EDA



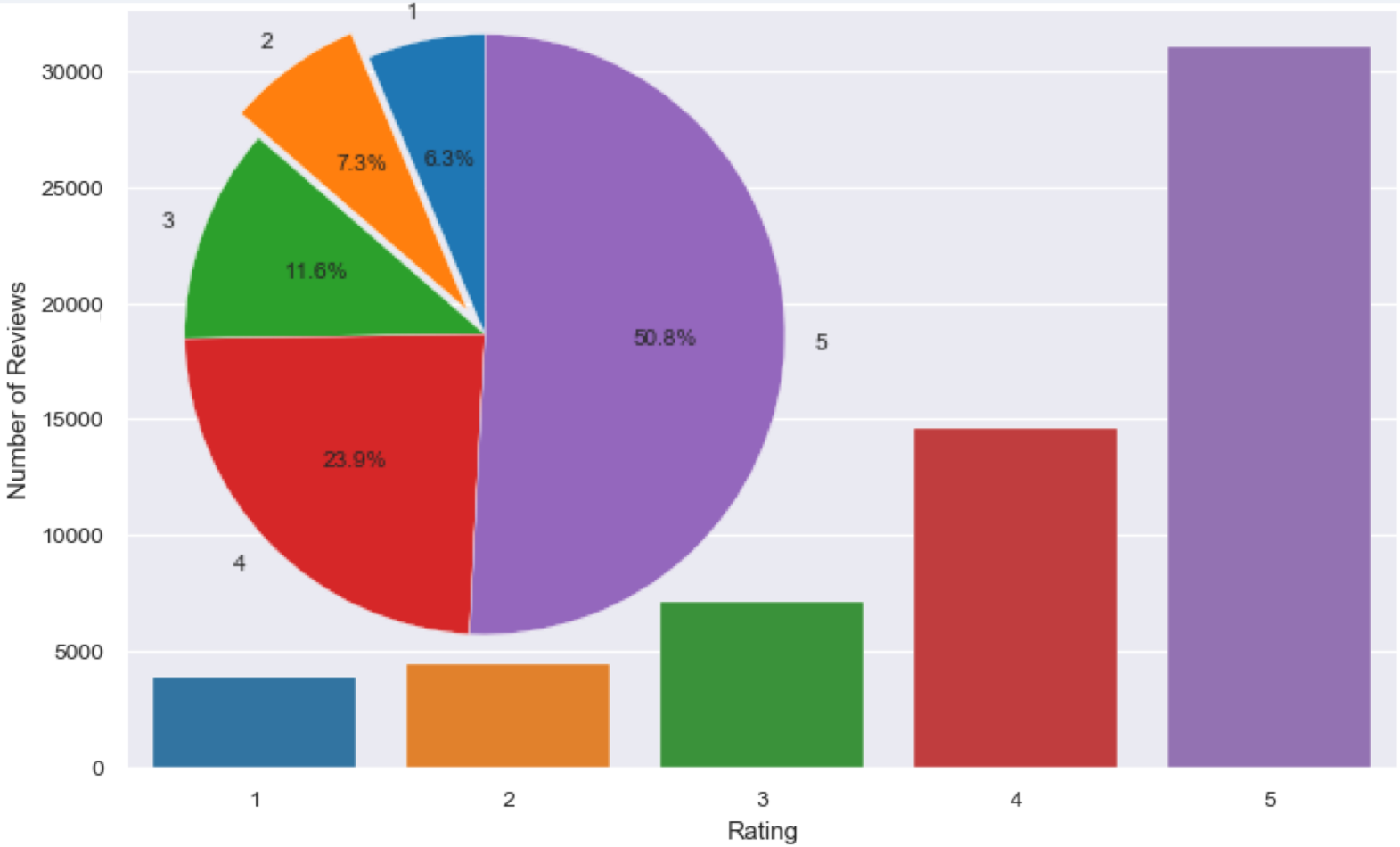


Top Positive Words

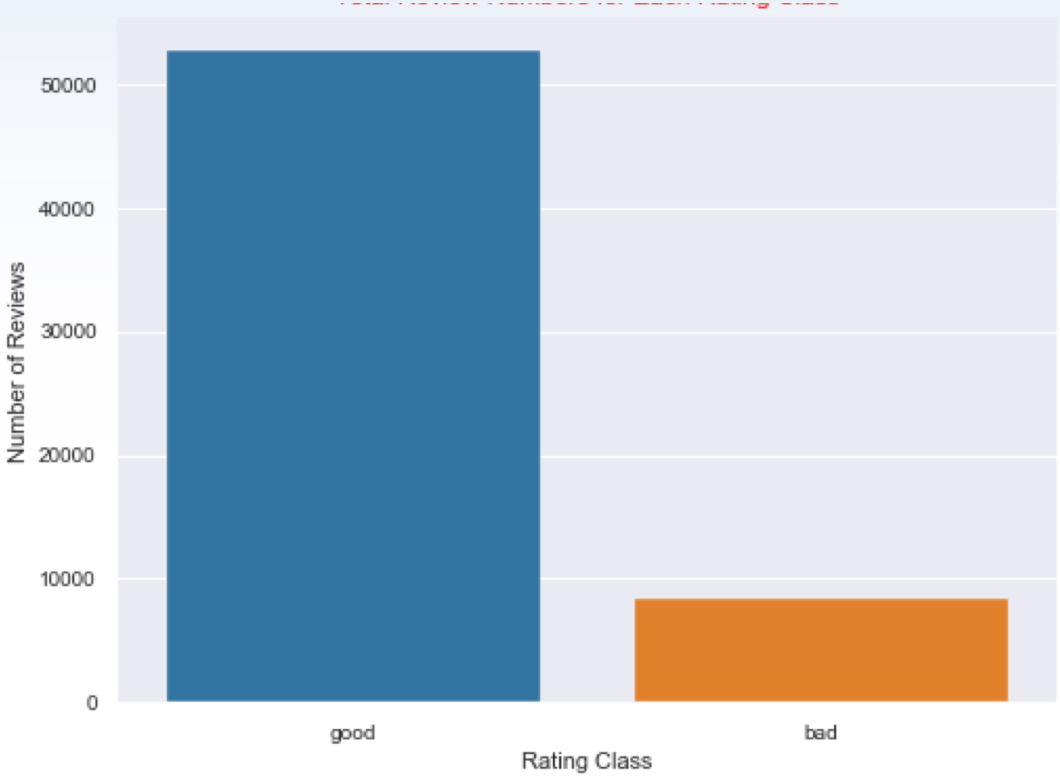
Top Critique Words

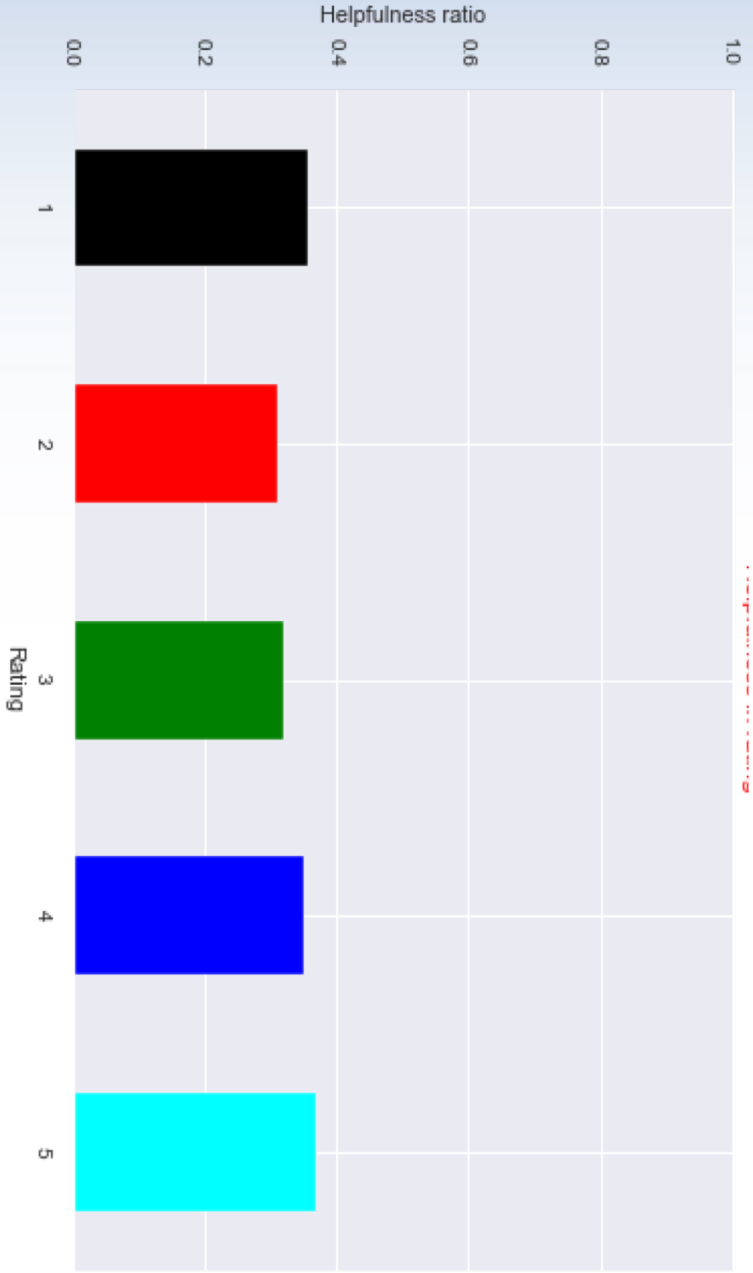
EDA

Which Rating got Highest Number of Reviews?



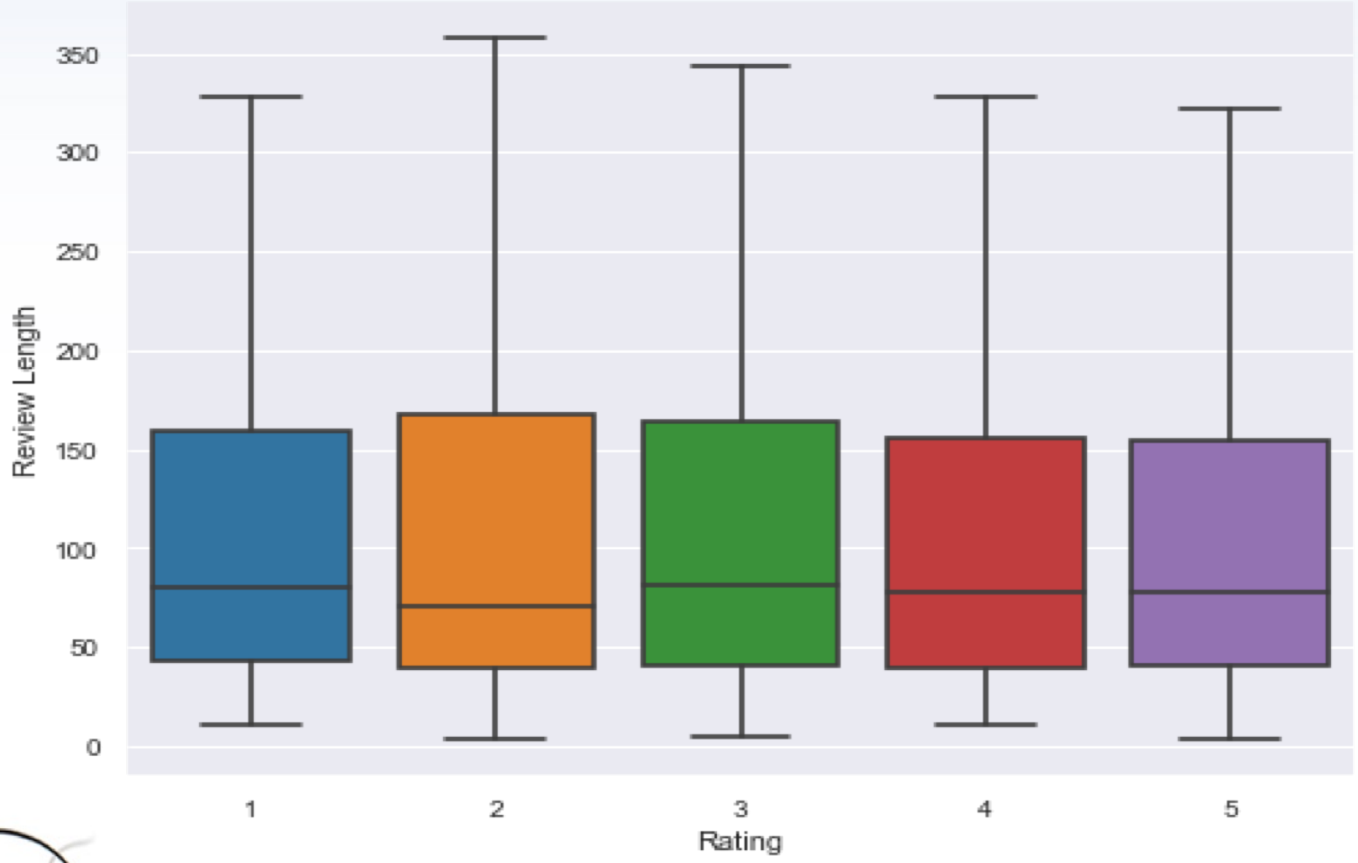
EDA


50000
40000
30000
20000
10000
0
Number of Reviews
good
bad
Rating Class


Helpfulness ratio
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Rating

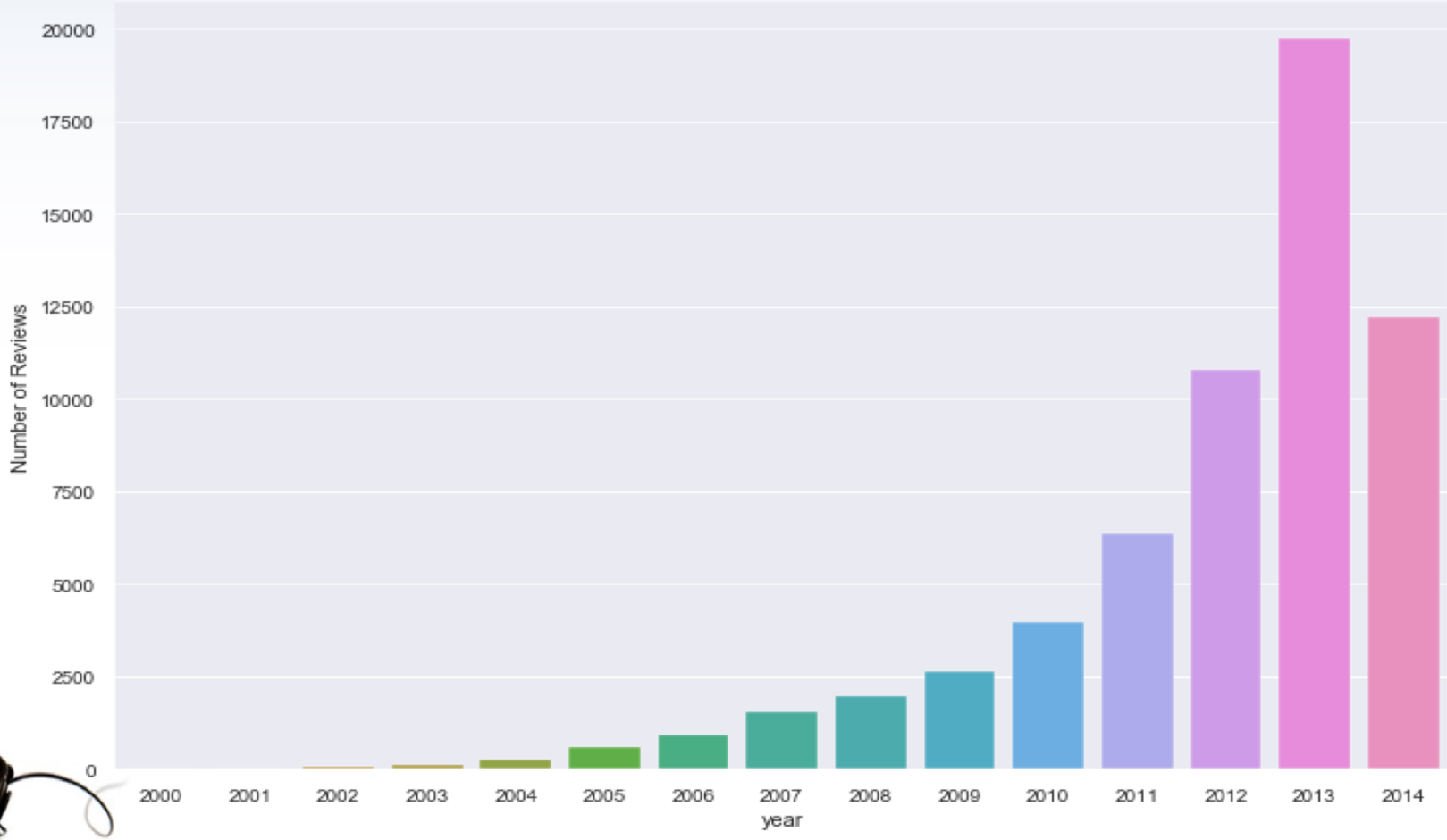
EDA

Which Rating got Highest Number of Review Length?



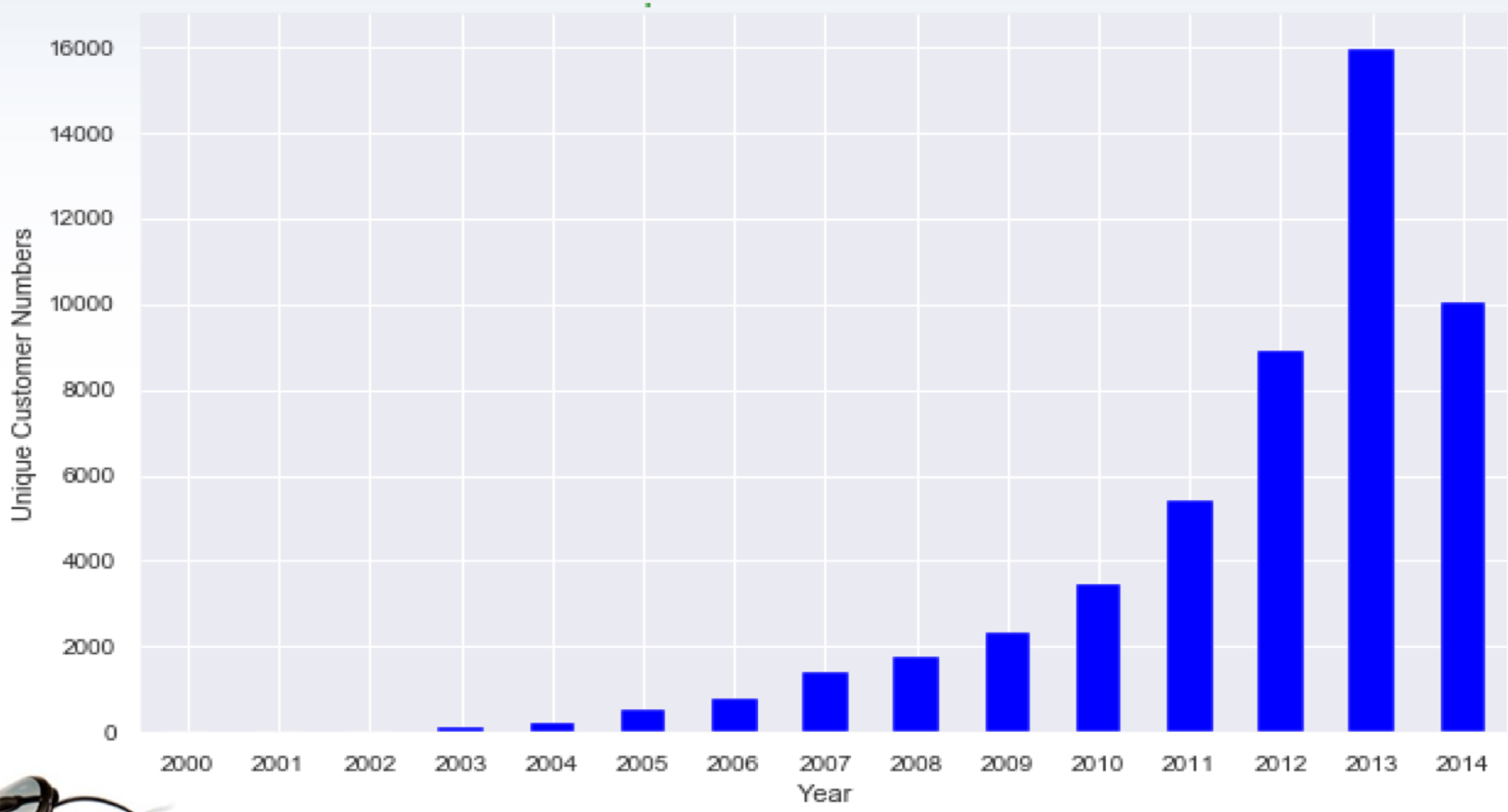
EDA

Which Year has the Highest Number of Reviews?



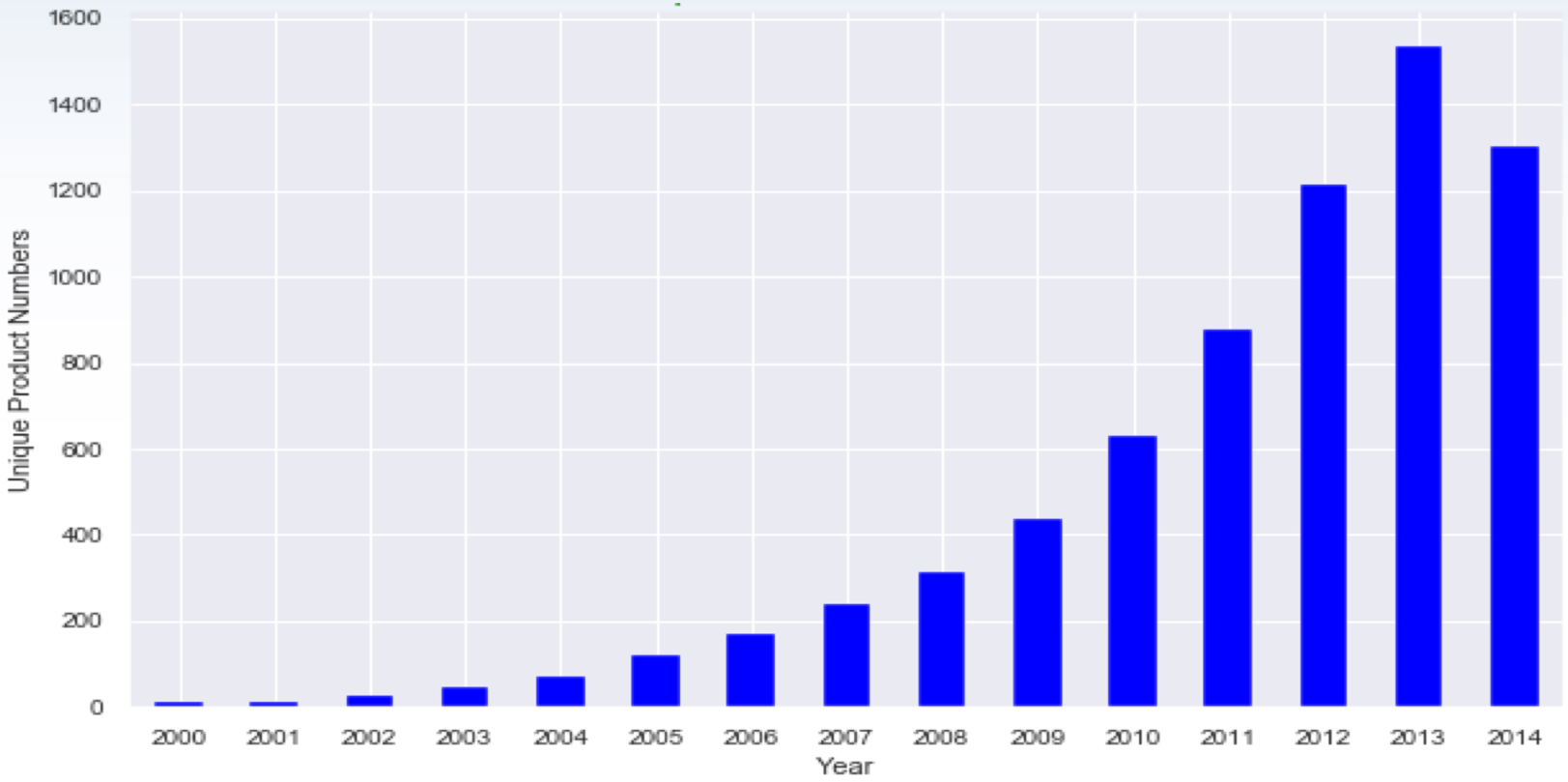
EDA

Which Year has the Highest Number of Customers?



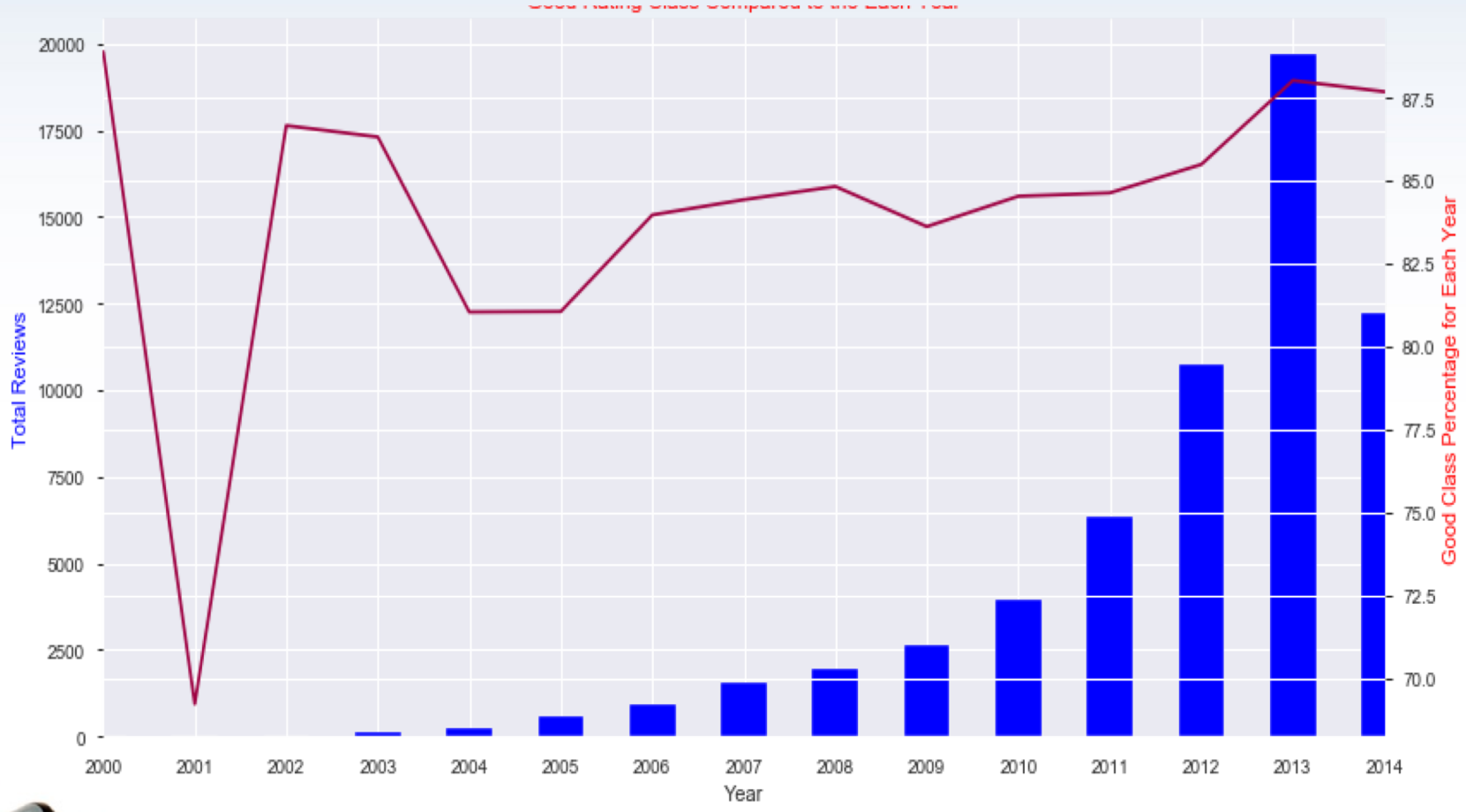
EDA

Which Year has the Highest Number of Product?



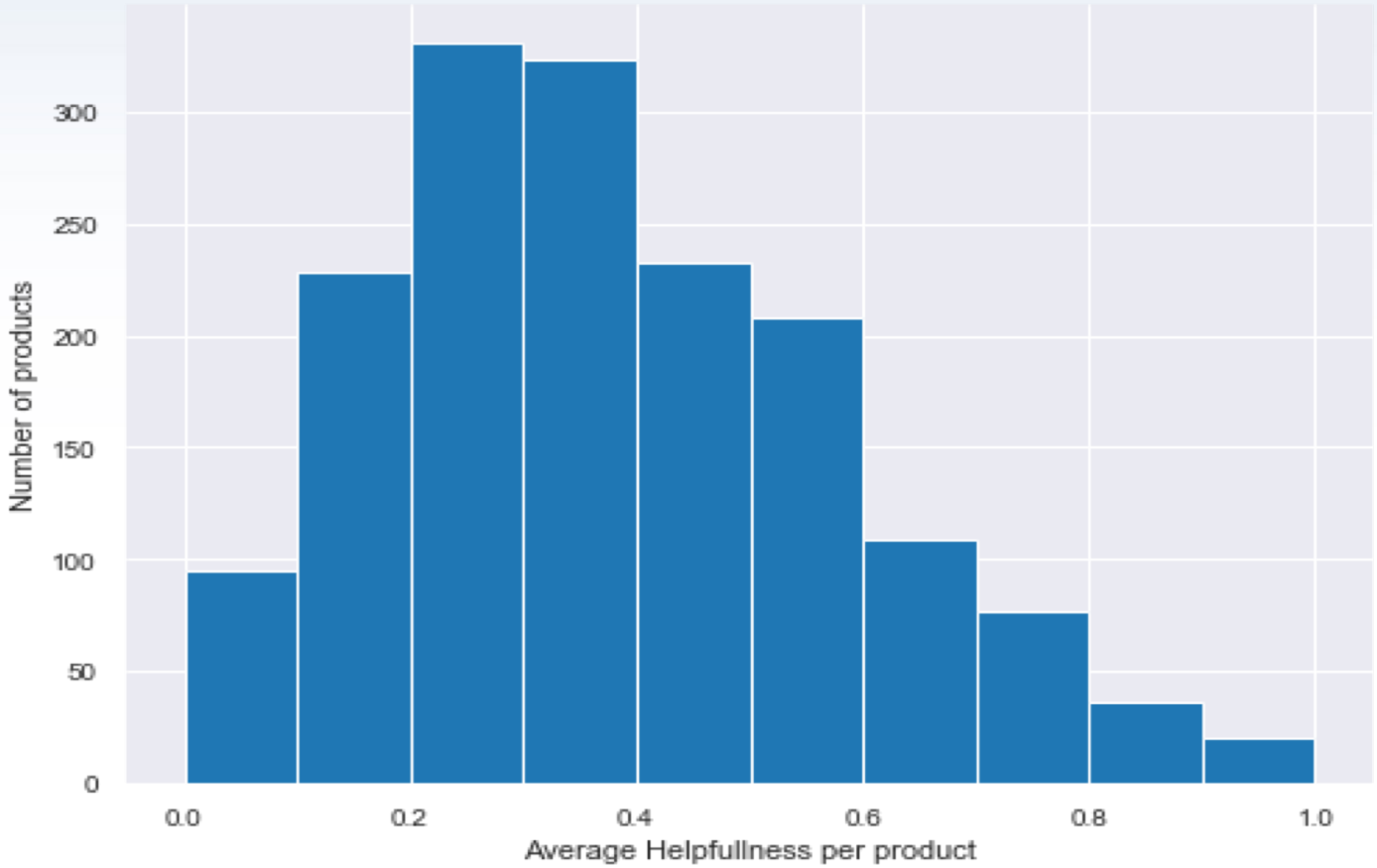
EDA

Which Year has the Highest Number of Product?



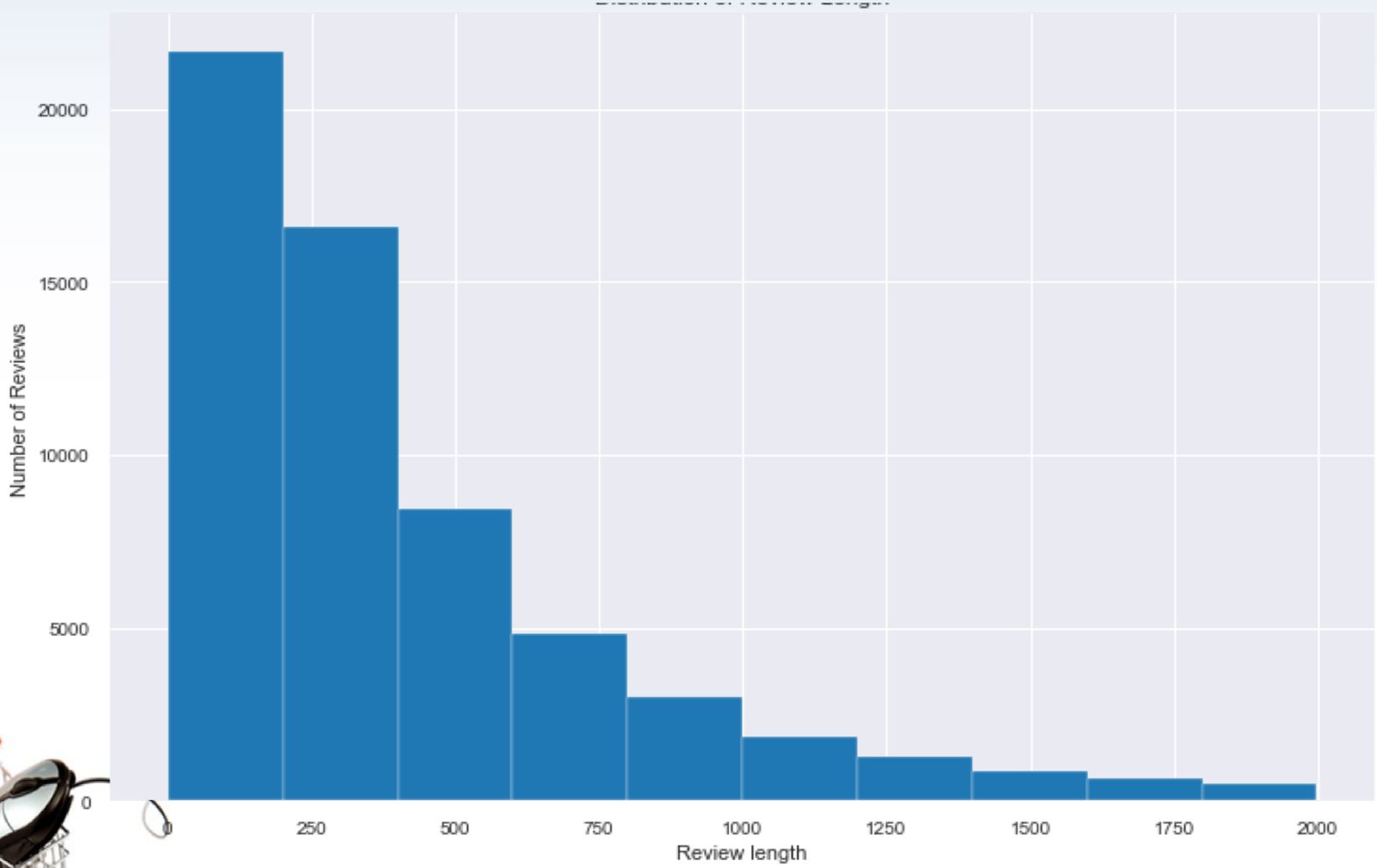
EDA

Average Helpfulness of the Product



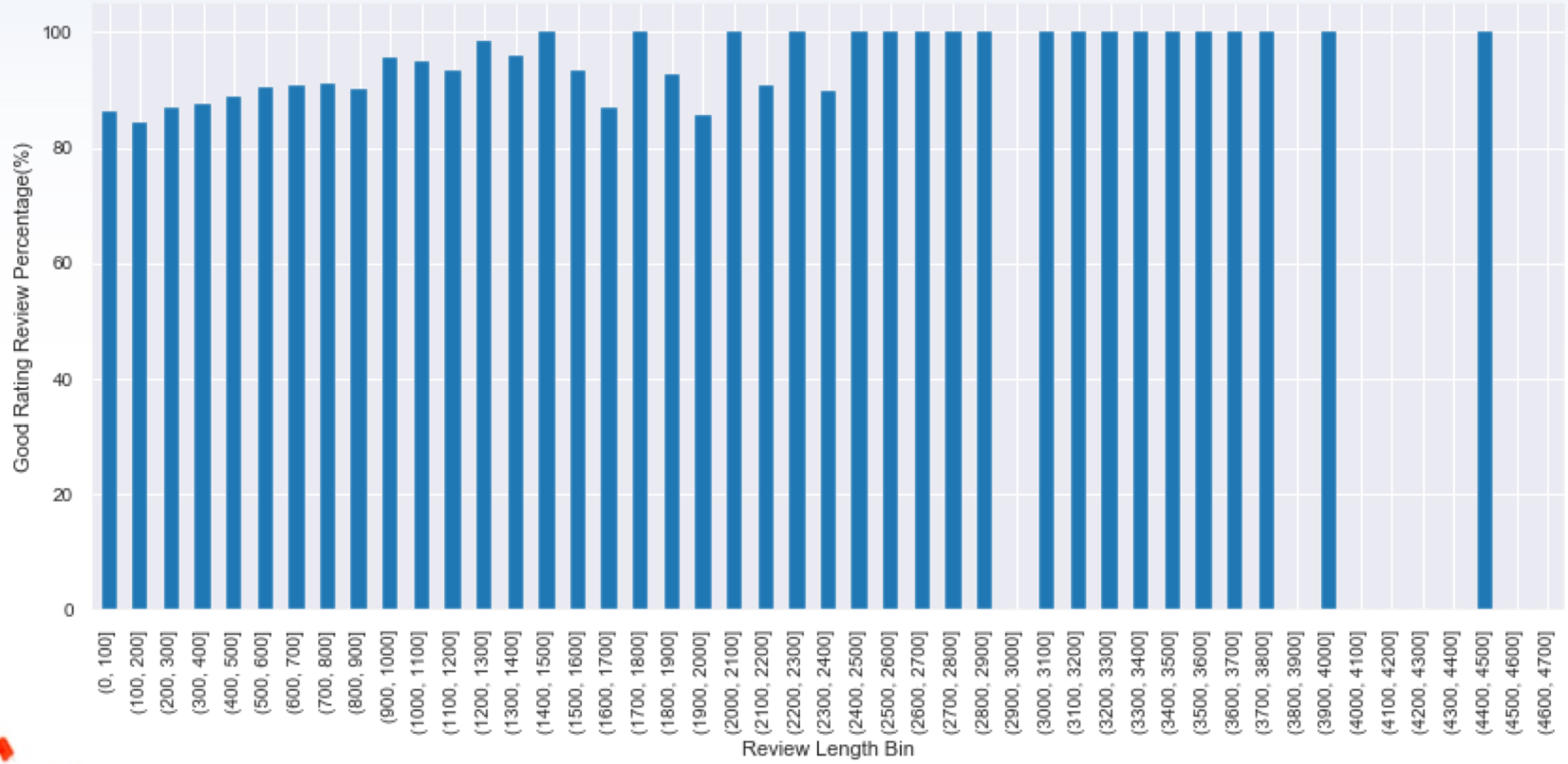
EDA

Distribution of Review Length



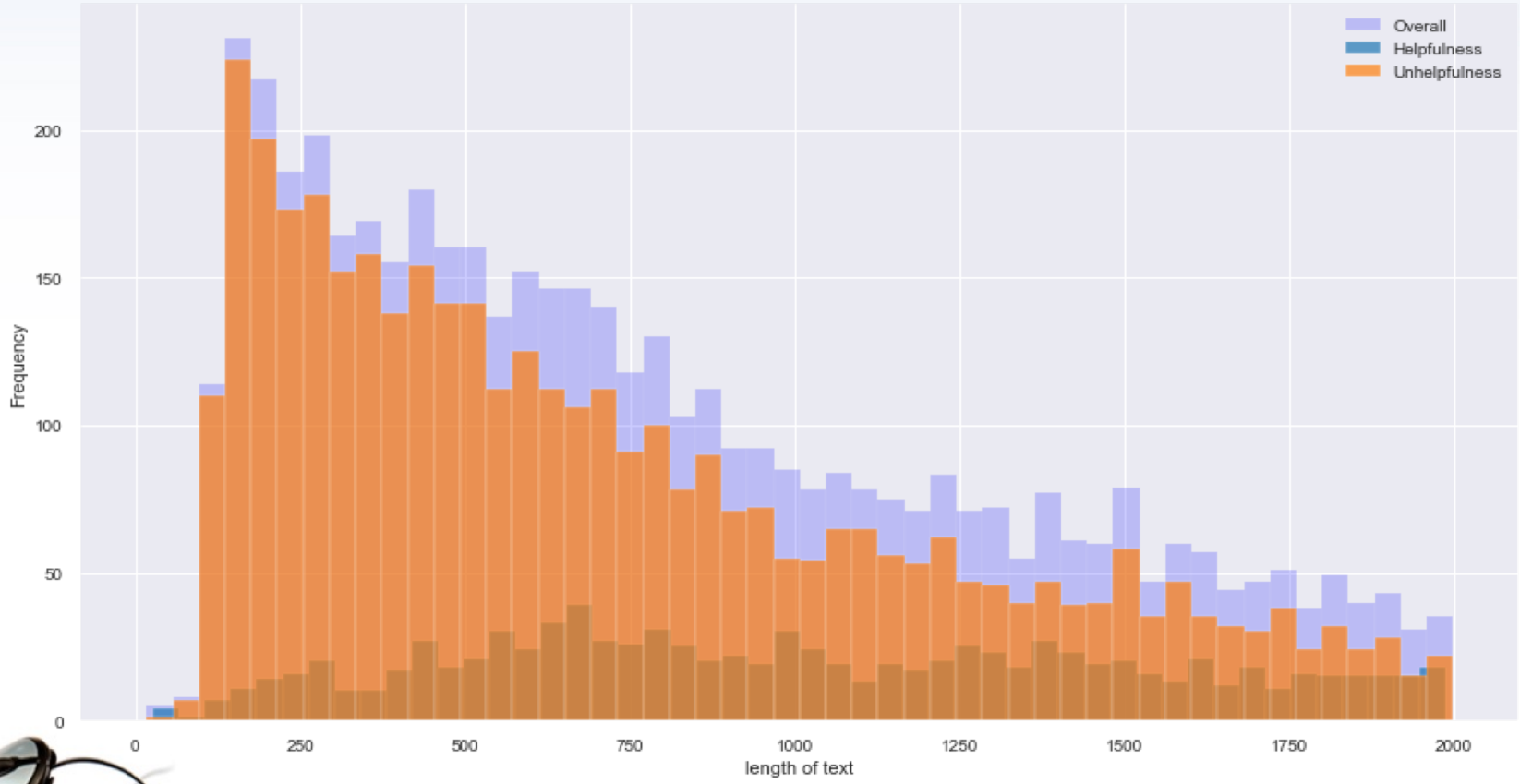
EDA

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



EDA

Which review length has the Highest Number of Helpfulness Ratio?

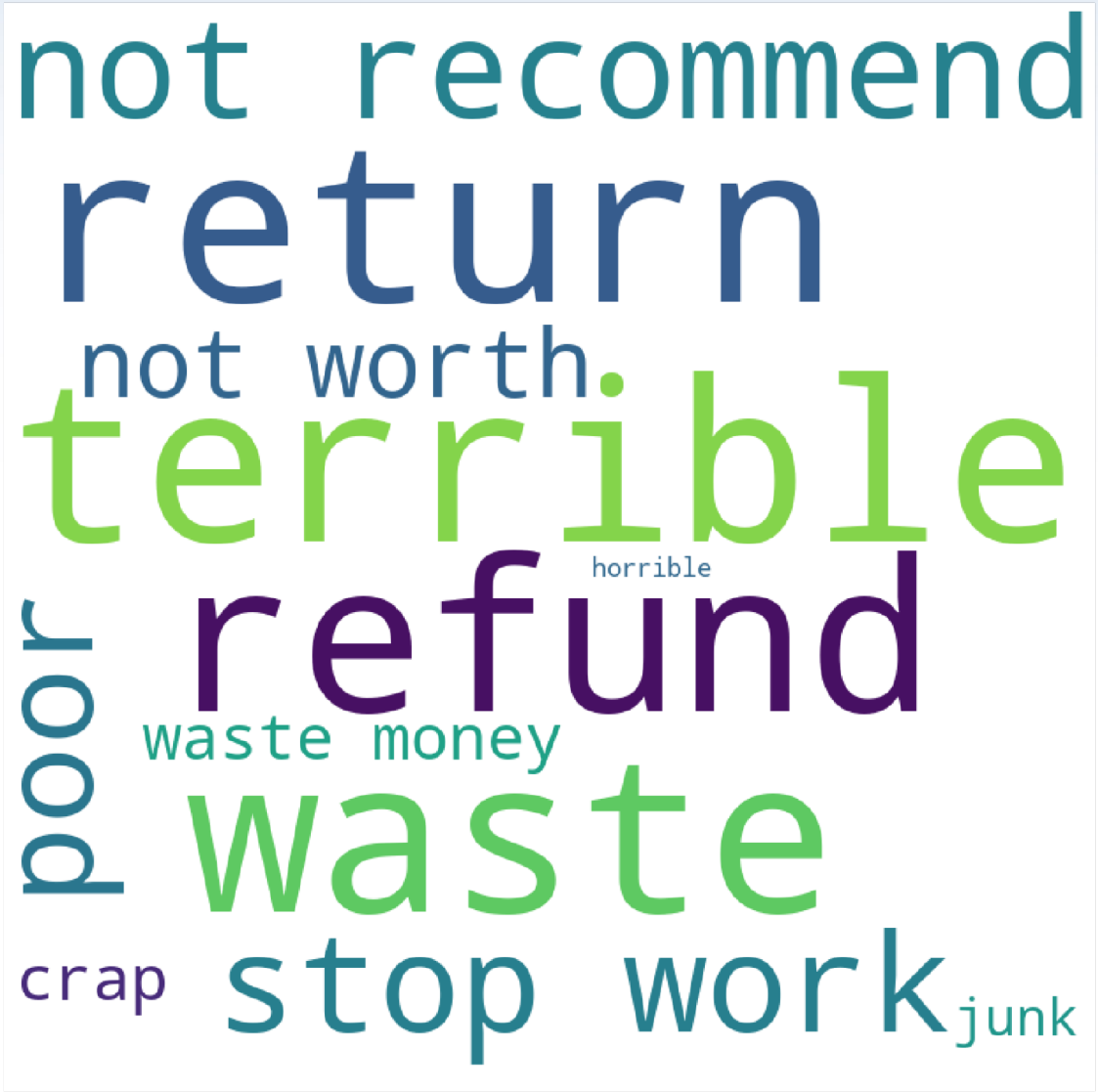


EDA

Top Words for Good Rating?

EDA

Top Words for Bad Rating?



Modeling

Data Preprocessing and Evaluation Metrics

Separating Response Variable and Feature

```
X = df4['clean_text']
y = df4['rating_class']
```

Splitting Dataset into Training and Test Set

```
Test Size = 25%
Stratified Sampling
```

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Predicted Value (predicted by the test)	Actual Value (as confirmed by experiment): positives	Actual Value: negatives
positives	**TP** True Positive	**FP** False Positive
negatives	**FN** False Negative	**TN** True Negative

```
Train Set Shape : (11250,)
Test Set Shape  : (3750,)
```

Modeling

Vectorizing

Turning a collection of text documents into numerical feature vectors.

- CountVectorizer
- TF-IDF
- Hashing Vectorizer

Modeling

Classification Models

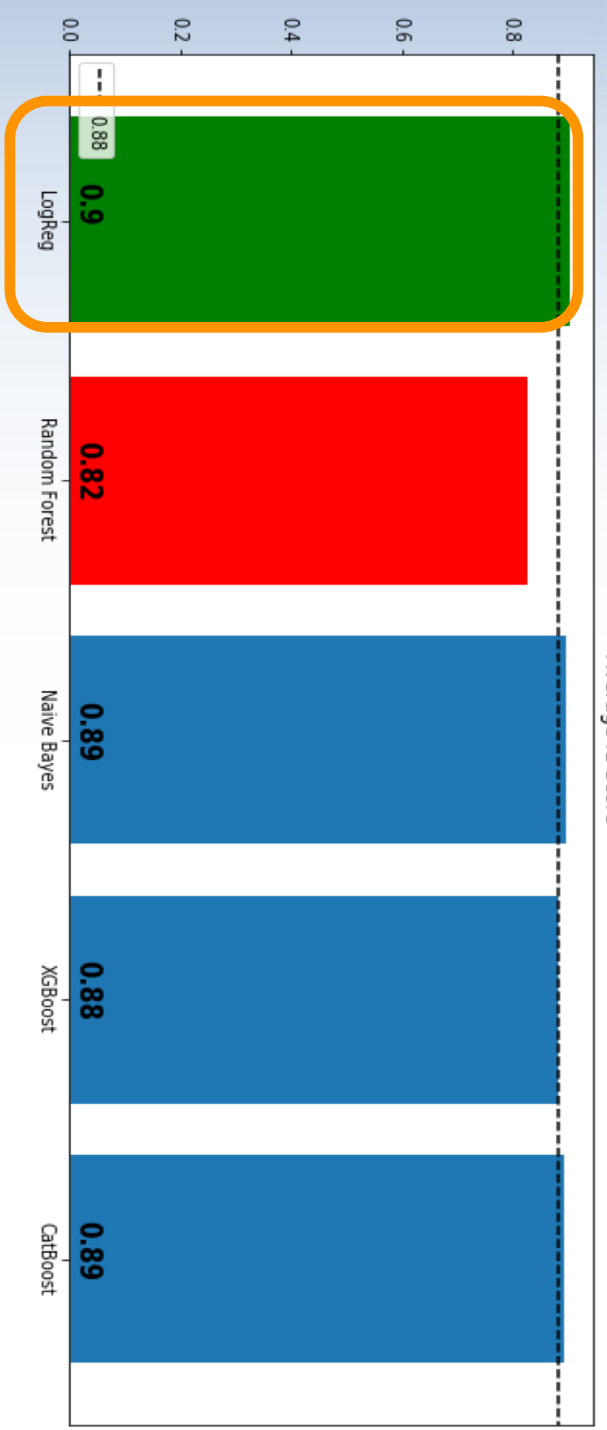
- Logistic Regression
- Random Forest
- Naïve Bayes
- XGBoost
- CatBoost

Modeling

Bag of Words Model

vectorizer	model	accuracy	class	precision	recall	f1-score	support
CountVect	LogReg	0.896267	bad	0.688581	0.833799	0.754264	716.0
			good	0.958724	0.911009	0.934257	3034.0
			average	0.907144	0.896267	0.899891	3750.0
	Random Forest	0.857867	bad	0.969231	0.263966	0.414929	716.0
			good	0.851758	0.998022	0.919108	3034.0
			average	0.874188	0.857867	0.822843	3750.0
	Naive Bayes	0.898400	bad	0.790295	0.636872	0.705336	716.0
			good	0.918059	0.960119	0.938618	3034.0
			average	0.893664	0.898400	0.894077	3750.0
	XGBoost	0.890933	bad	0.880893	0.495810	0.634495	716.0
			good	0.892142	0.984179	0.935903	3034.0
			average	0.889994	0.890933	0.878355	3750.0
	CatBoost	0.896800	bad	0.818182	0.590782	0.686131	716.0
			good	0.909372	0.969018	0.938248	3034.0
			average	0.891961	0.896800	0.890111	3750.0

Logistic Regression ➔ 0.896267

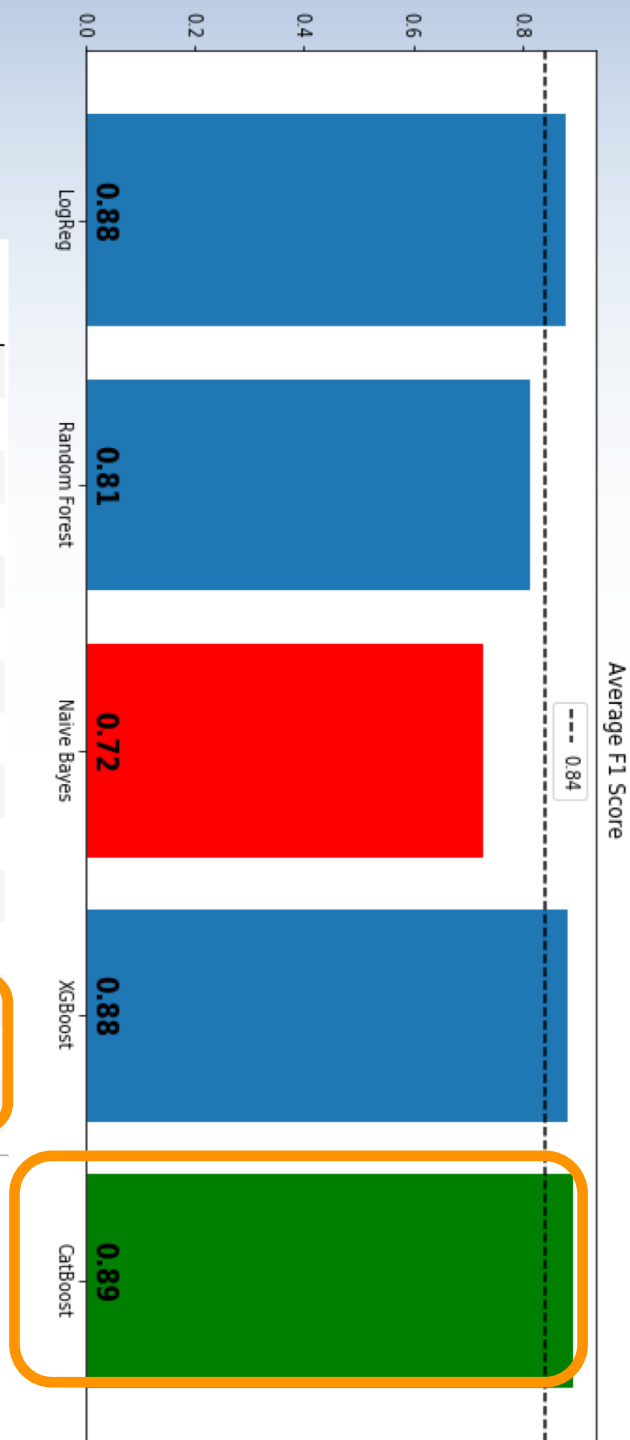


Modeling

TF-IDF Model

vectorizer	model	accuracy	class	precision	recall	f1-score	support
CountVect	LogReg	0.866933	bad	0.602070	0.893855	0.719505	716.0
			good	0.971716	0.860580	0.912777	3034.0
			average	0.901138	0.866933	0.875875	3750.0
	Random Forest	0.852533	bad	0.988024	0.230447	0.373726	716.0
			good	0.846218	0.999341	0.916427	3034.0
			average	0.873294	0.852533	0.812808	3750.0
	Naive Bayes	0.809600	bad	1.000000	0.002793	0.005571	716.0
			good	0.809498	1.000000	0.894721	3034.0
			average	0.845872	0.809600	0.724953	3750.0
	XGBoost	0.892267	bad	0.902062	0.488827	0.634058	716.0
			good	0.891136	0.987475	0.936836	3034.0
			average	0.893222	0.892267	0.879025	3750.0
	CatBoost	0.896533	bad	0.805970	0.603352	0.690096	716.0
			good	0.911637	0.965722	0.937900	3034.0
			average	0.891461	0.896533	0.890586	3750.0

CatBoost ➔ 0.896533

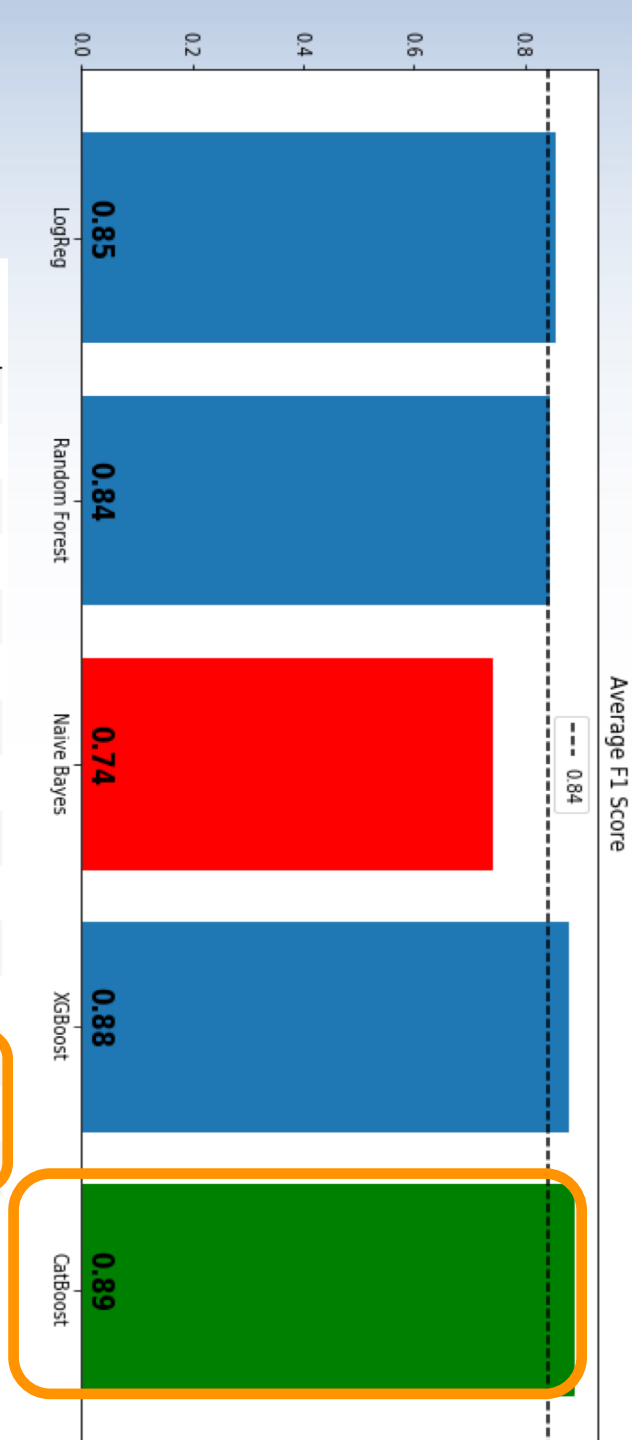


Modeling

Hashing Vectorizing

vectorizer	model	accuracy	class	precision	recall	f1-score	support
CountVect	LogReg	0.842400	bad	0.556256	0.863128	0.676519	716.0
			good	0.962865	0.837508	0.895822	3034.0
			average	0.885229	0.842400	0.853950	3750.0
	Random Forest	0.870667	bad	0.971429	0.332402	0.495317	716.0
			good	0.863623	0.997693	0.925830	3034.0
			average	0.884207	0.870667	0.843630	3750.0
	Naive Bayes	0.816000	bad	0.964286	0.037709	0.072581	716.0
			good	0.814884	0.999670	0.897869	3034.0
			average	0.843410	0.816000	0.740294	3750.0
	XGBoost	0.889867	bad	0.912807	0.467877	0.618652	716.0
			good	0.887378	0.989453	0.935640	3034.0
			average	0.892233	0.889867	0.875116	3750.0
	CatBoost	0.894133	bad	0.812133	0.579609	0.676447	716.0
			good	0.907070	0.968359	0.936713	3034.0
			average	0.888943	0.894133	0.887019	3750.0

CatBoost ➔ 0.894133

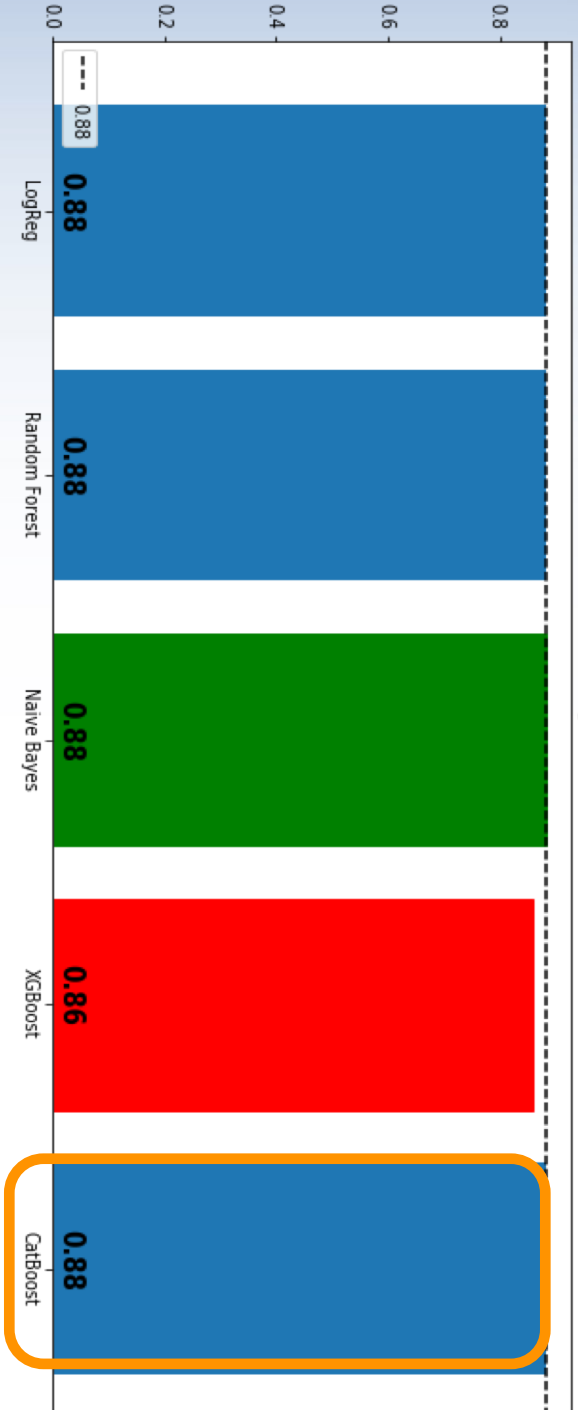


Modeling

Count-Vectorizing - Adding Most and Least Common Words to Stopwords List

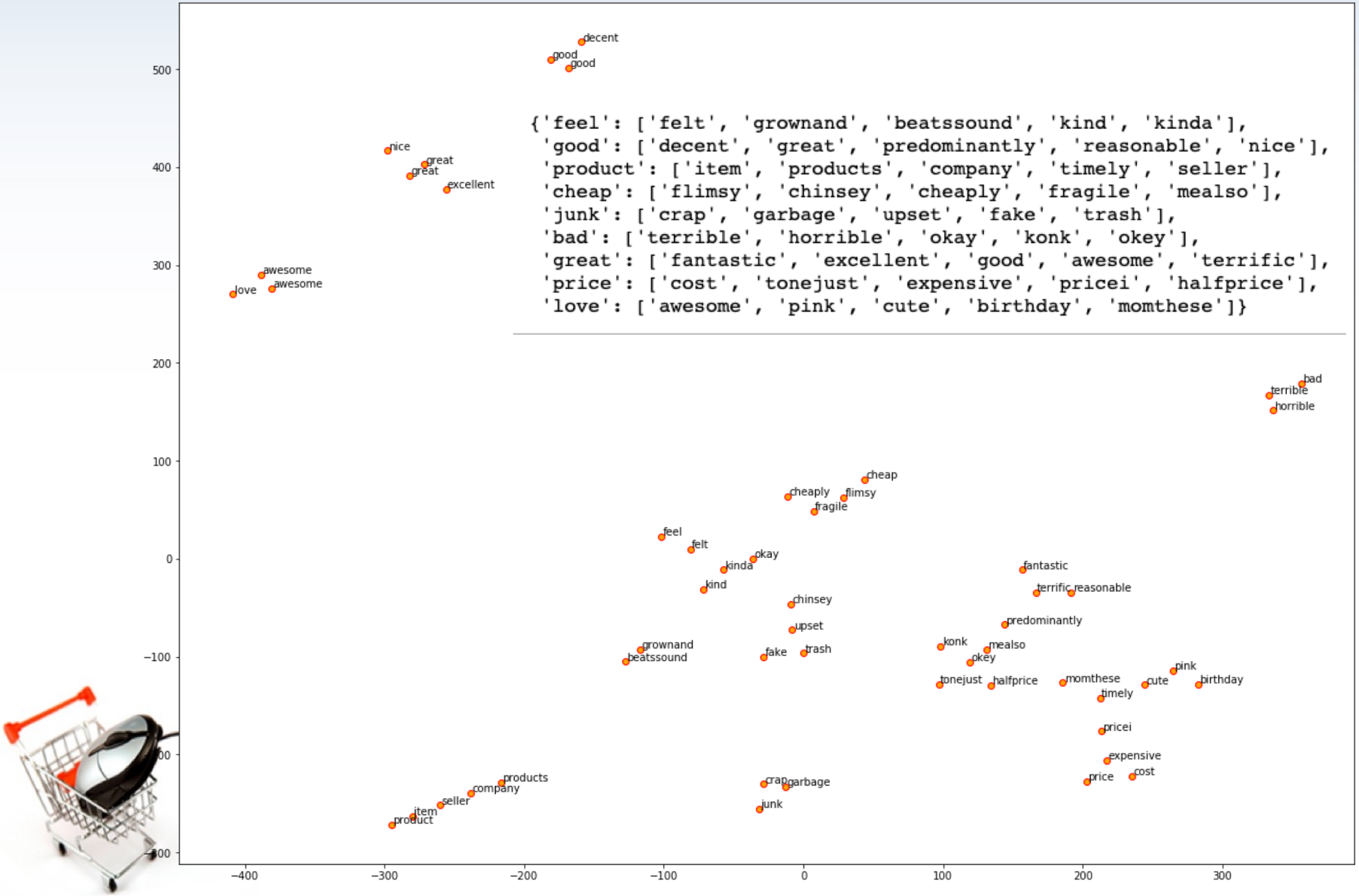
vectorizer	model	accuracy	class	precision	recall	f1-score	support
CountVect	LogReg	0.875200	bad	0.643519	0.776536	0.703797	716.0
			good	0.944560	0.898484	0.920946	3034.0
			average	0.887081	0.875200	0.879485	3750.0
	Random Forest	0.889333	bad	0.841270	0.518156	0.641314	716.0
			good	0.895739	0.976928	0.934574	3034.0
			average	0.885339	0.889333	0.878580	3750.0
	Naive Bayes	0.881600	bad	0.675258	0.731844	0.702413	716.0
			good	0.935440	0.916941	0.926099	3034.0
			average	0.885763	0.881600	0.883389	3750.0
	XGBoost	0.877600	bad	0.890578	0.409218	0.560766	716.0
			good	0.876352	0.988134	0.928892	3034.0
			average	0.879068	0.877600	0.858605	3750.0
	CatBoost	0.890133	bad	0.819328	0.544693	0.654362	716.0
			good	0.900428	0.971655	0.934686	3034.0
			average	0.884943	0.890133	0.881163	3750.0

CatBoost ➔ 0.890133

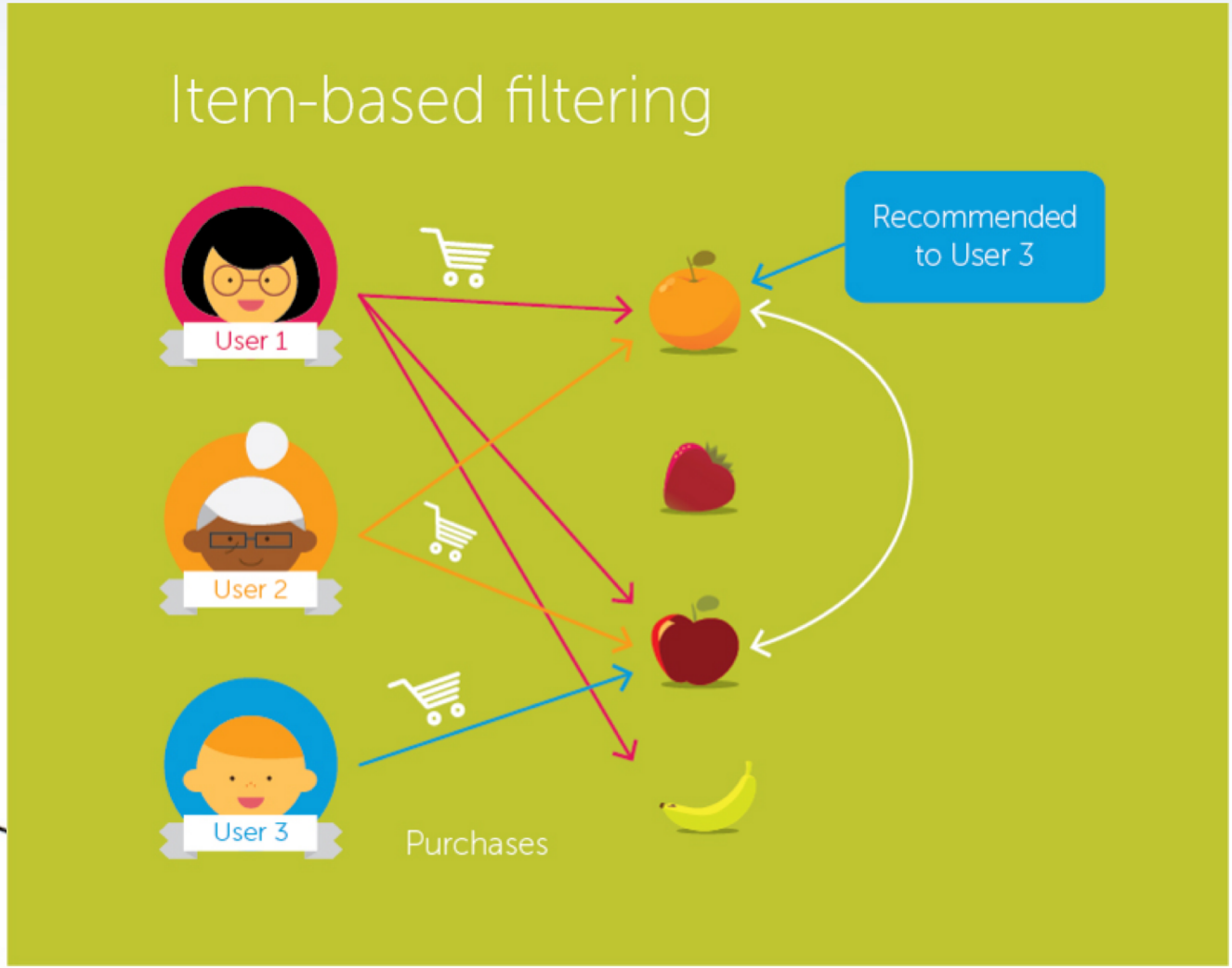


Modeling

Word2Vec and Simple Neural Network



Item-Item based Collaborative Filtering



Product Recommendation

	prod_ID	prod_name	ratings_sum
3313486	B003ES5ZUU	AmazonBasics High-Speed HDMI Cable - 15 Feet (...	846.0
6104981	B0088CJT4U	TP-LINK TL-WDR4300 Wireless N750 Dual Band Rou...	814.0
1198226	B000N99BBC	TP-LINK TL-SG1005D 10/100/1000Mbps 5-Port Giga...	755.0
5941380	B007WTAJTO	SanDisk Ultra 64GB MicroSDXC Class 10 UHS Memo...	741.0
6031403	B00829TIEK	Seagate Backup Plus 3TB USB 3.0 Desktop Extern...	626.0
6028195	B00829THK0	Seagate Backup Plus 1TB Desktop External Hard ...	560.0
6190152	B008DWCRQW	D-Link Wireless AC 1750 Mbps Home Cloud App-En...	524.0
4024382	B004CLYEDC	Micra Digital CAT5e Snagless Patch Cable, 5 Fe...	517.0
2796338	B002R5AM7C	Flip MinoHD Video Camera - Brushed Metal, 8 GB...	514.0
2883526	B002V88HFE	eneloop SEC-CSPACER4PK C Size Spacers for use ...	475.0

Top 10 popular products by sum user ratings

```
User A100WO06OQR8BQ has already purchased 91 items.
Recommending the highest 5 predicted  items not already purchased.
```

Product recommendation for user A100W006oQR8BQ

	prod_ID	prod_name
11745	B004CLYEDC	Micra Digital CAT5e Snagless Patch Cable, 5 Fe...
5649	B000N99BBC	TP-LINK TL-SG1005D 10/100/1000Mbps 5-Port Giga...
5012	B004CLYEFK	Micra Digital USB A to USB B Cable (6 Feet)
1169	B00829THK0	Seagate Backup Plus 1TB Desktop External Hard ...
656	B00834SJSK	Seagate Expansion 500GB Portable External Hard...

Summary

- **CatBoosting with TF-IDF Vectorizing (f1 score = 0.890586) or Logistic Regression with Bag of Words model (f1 score = 0.899891) are best models.**

- **Adding most and least common words to the stop word list did not have impact on model's performance.**

Future Study

- Using different methods in order to minimize the effect of the matching words
- Using different AutoML tools.
- Implementation of Dask library for parallel processing to decrease run time.