Amazon Product Recommender System

Jin Hui Xu

UMBC Data Science Capstone

Contents

Introduction

Introduction

Why recommender system is important?

- It can drive traffic through personalized email messages to the store site and increase average order value.
- It also enhances the shopping experience by delivering relevant content based on personalized preferences.
- It can reduce workload for inventory management and boost work effectiveness.
- It can create comprehensive reports to support making the right decision for business direction.
- Overall, product recommender systems not only boost the companies' revenue but also increase customer satisfaction and loyalty.

Research Questions

Research Questions

- What characteristics are useful to generate personalized recommendations?
- Which recommender systems algorithms/methods are most successful and practical?
- Can textual data improve recommender systems' performance?

Data Sources

Data Sources

The data for this project is the Amazon Review Data (2018) which is collected by the University of California San Diego (https://nijianmo.github.io/amazon/index.html).

The dataset includes reviews data and product metadata.

It contains a total number of 233.1 million real reviews with the size of 34 gigabytes from Amazon.

Due to the computing resource limitation, a subset in Appliances category will be used for this project.

Review Data

There are a total of 602,777 review records in the Appliances category, and the dataset has 12 different features.

The interested feature are overall, reviewTime, reviewerID, asin, reviewText, summary.

Feature	Data Type	Description
reviewerID	String	ID of the reviewer
asin	String	ID of the product
reviewerName	String	name of the reviewer
vote	Integer	helpful votes of the review
style	String	a dictionary of the product metadata, e.g., "Format" is "Hardcover"
reviewText	String	text of the review
overall	float	rating of the product
summary	String	summary of the review
unixReviewTime	Integer	time of the review (unix time)
reviewTime	Datetime	time of the review (raw)
verified	Boolean	verified review
image	Object	images that users post after they have received the product

Product Data

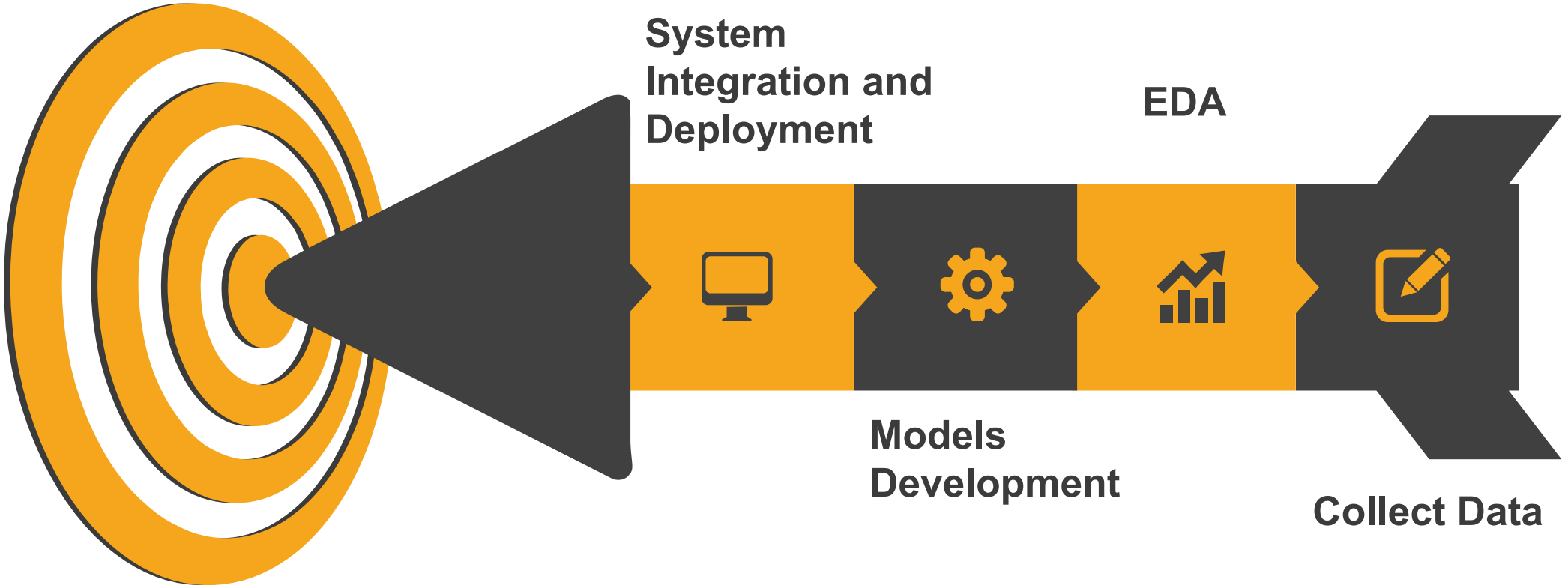
There are a total of 30,239 product records in this category, and the dataset has 19 different features.

The interested feature are category, description, title, brand, feature, main_cat, date, price, asin, imageURLHighRes.

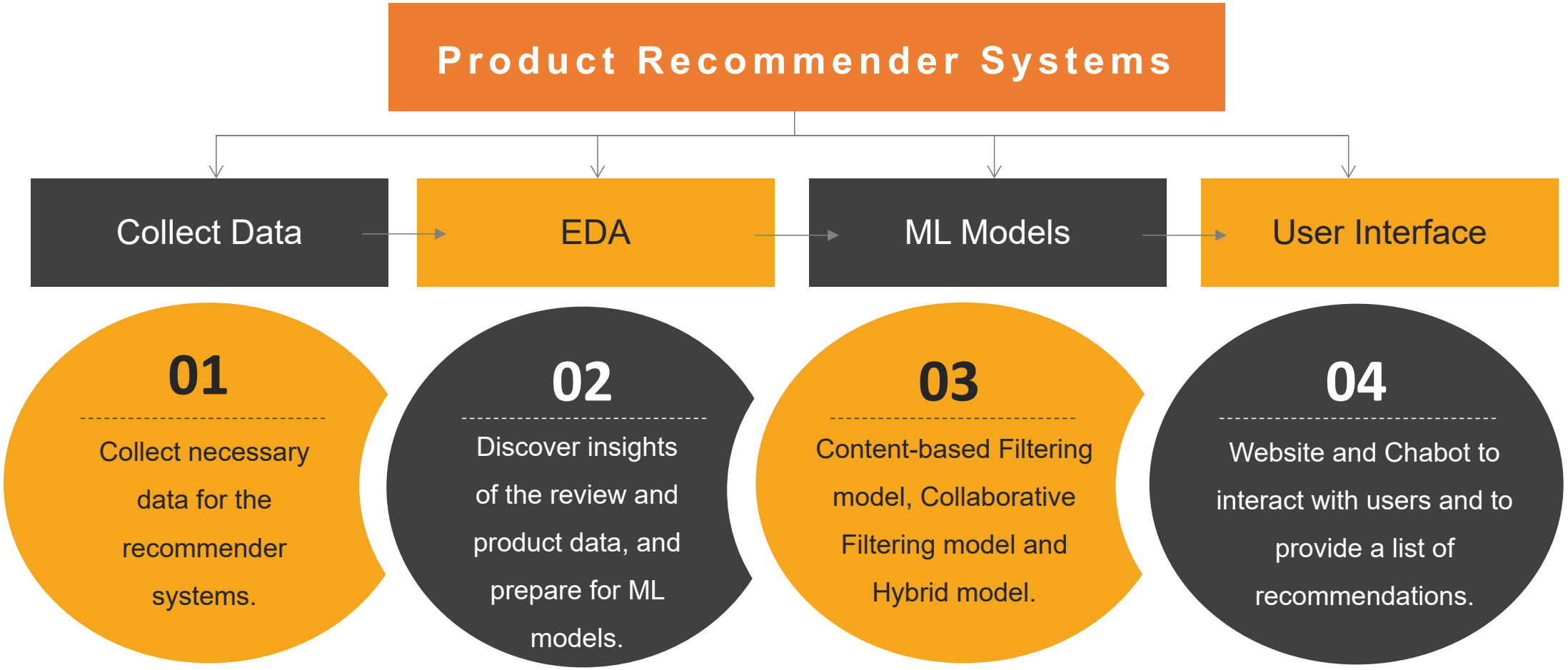
Feature	Data Type	Description
asin	String	ID of the product, e.g. 0000031852
title	String	name of the product
feature	String	bullet-point format features of the product
description	String	description of the product
price	String	price in US dollars (at time of crawl)
imageURL	Object	url of the product image
imageURLHighRes	Object	url of the high resolution product image
salesRank	String	sales rank information
brand	String	brand name
categories	String	list of categories the product belongs to
tech1	String	the first technical detail table of the product
tech2	String	the second technical detail table of the product
similar	String	similar product table
fit	String	size description of the product
also_buy	String	related products
also_view	String	related products
details	String	related product details
main_cat	String	main category the product belongs to
date	String	product release date

Research Process

Research Process

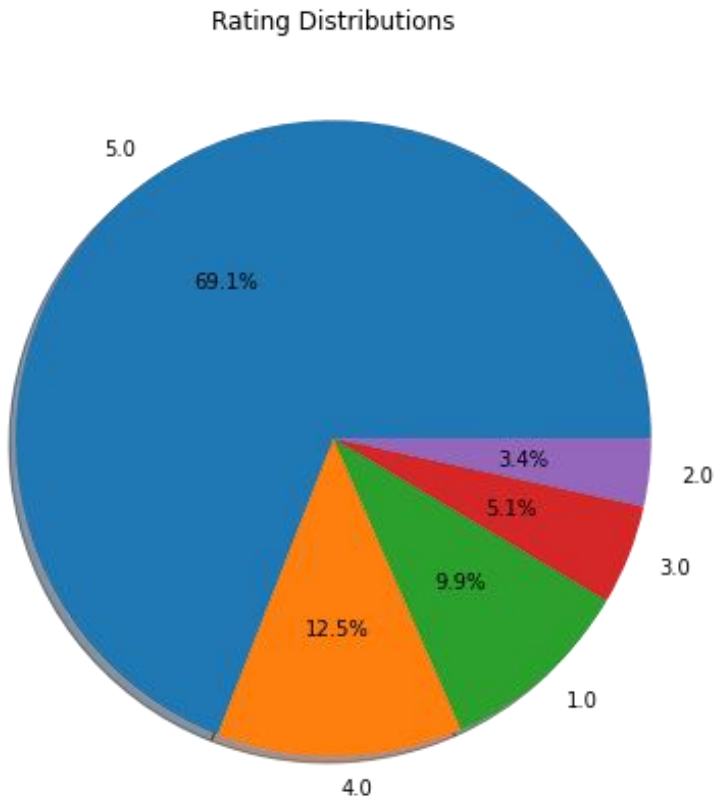


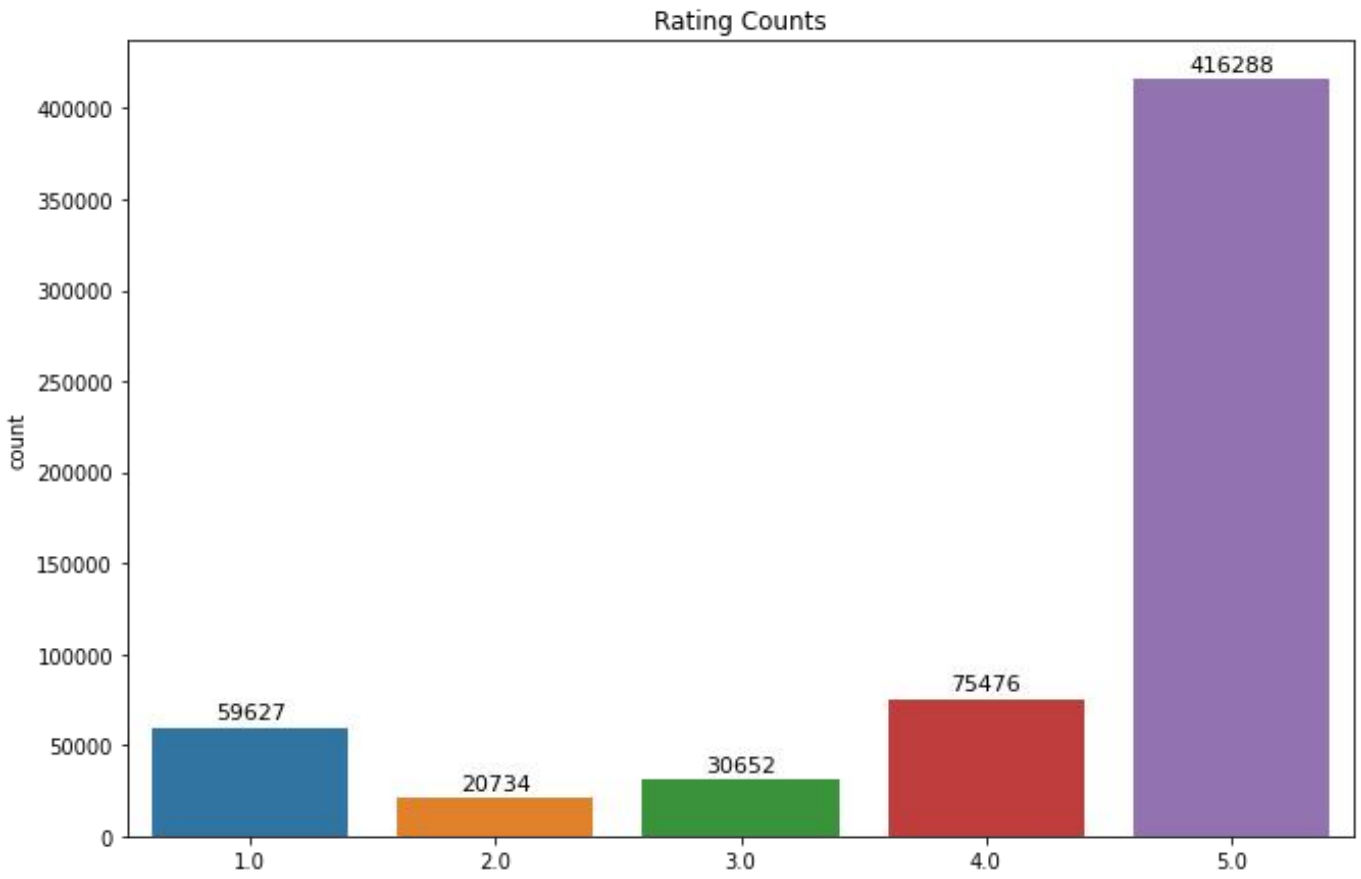
Research Process



Exploratory Data Analysis

EDA – Review Data





The rating distribution graphs show that the overall ratings in this review data set are highly imbalanced, which contains more than 69% of 5 stars rating.

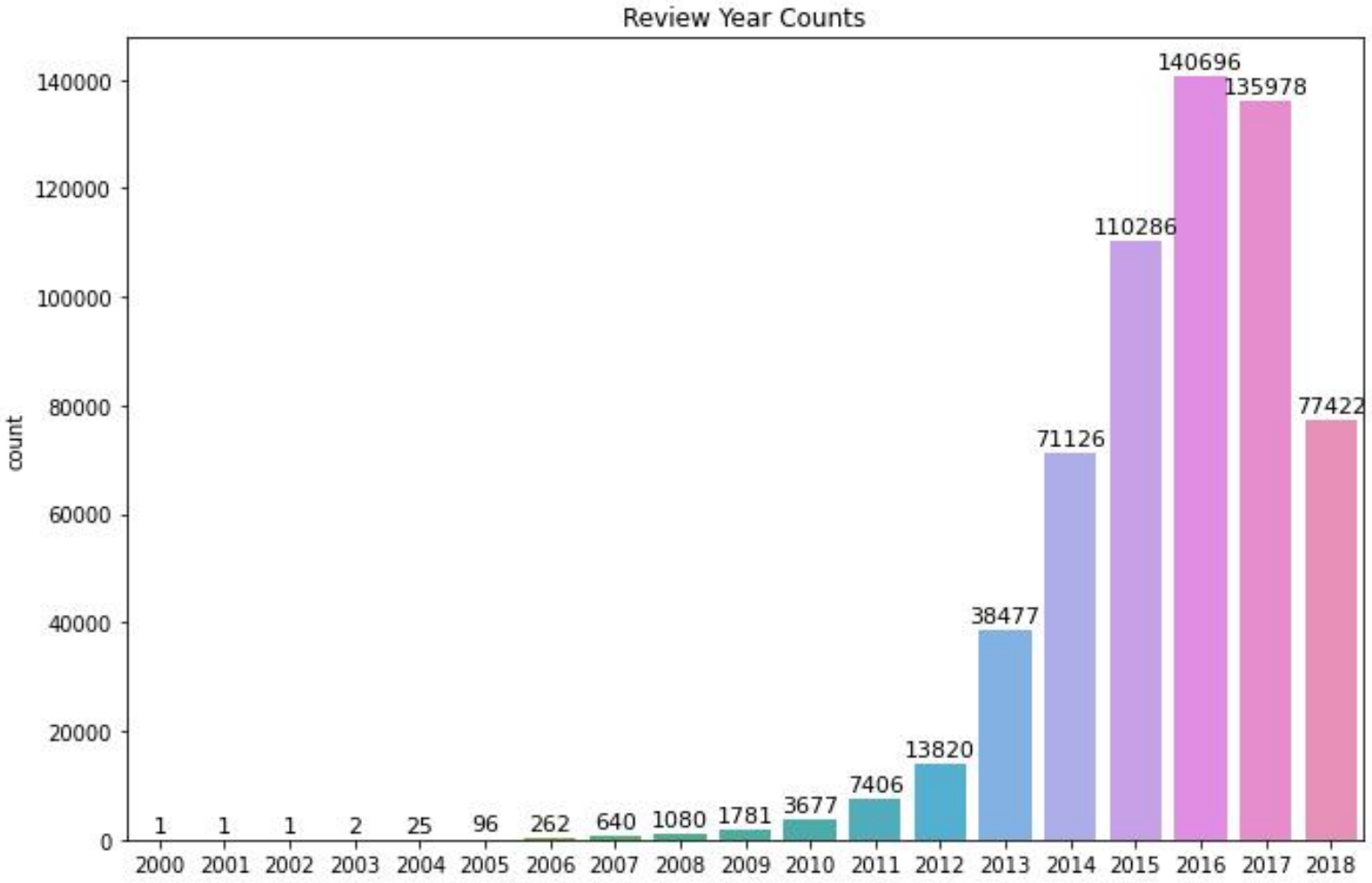
EDA – Review Data

	reviewerID	counts	rating_mean
412749	A8WEXFRWX1ZHH	208	4.980769
71295	A1IT56MV1C09VS	207	4.995169
142776	A21TPY9BVC9IKZ	206	5.000000
156061	A25C30G90PKSQA	206	3.000000
384058	A3TMNU7VNK5JJE	206	3.000000
...	...	...	...
183327	A2CH9B6K2QJS6Z	1	5.000000
183326	A2CH8ZFJWN1R60	1	5.000000
183325	A2CH7FPVP7H0XX	1	5.000000
183323	A2CH75VNS4T0GV	1	4.000000
515649	AZZZY1W55XHZR	1	3.000000

515650 rows × 3 columns

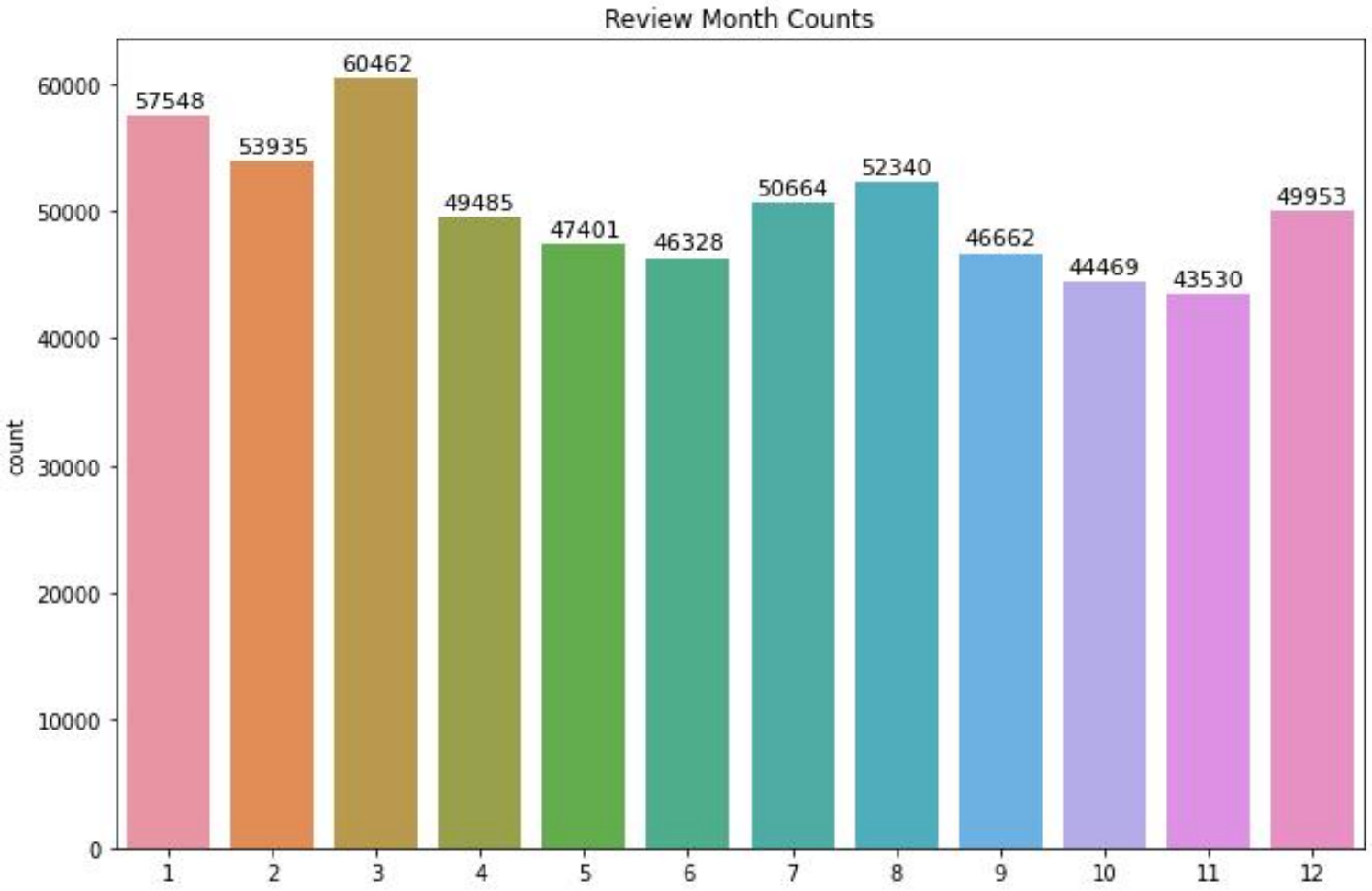
There are a total of 515,650 distinct reviewers in this dataset, and the most active reviewer had reviewed 208 products with an average 4.98 rating score.

EDA – Review Data



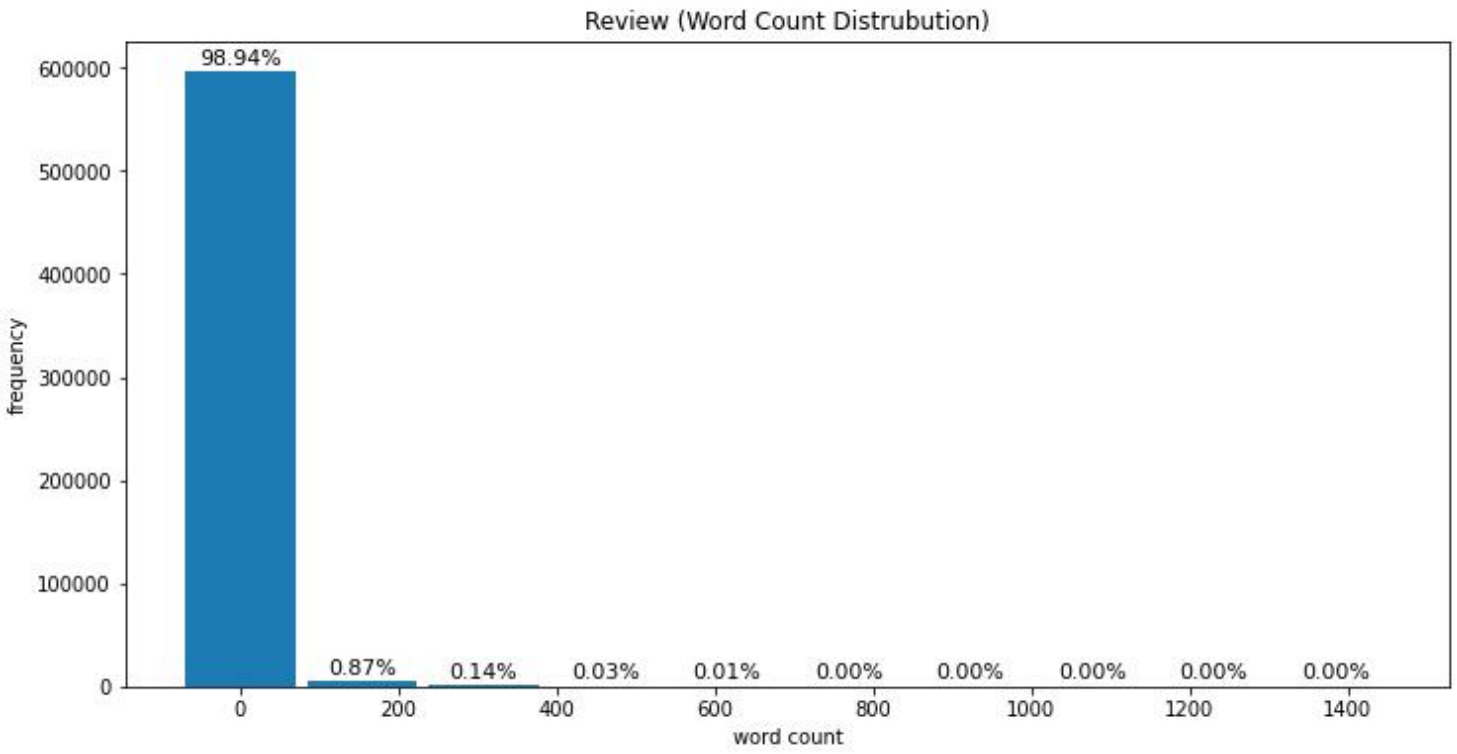
The review year distribution graphs show that the reviews in this dataset are heavily collected after the year 2013, which can quite well represent the current generation customers' preferences.

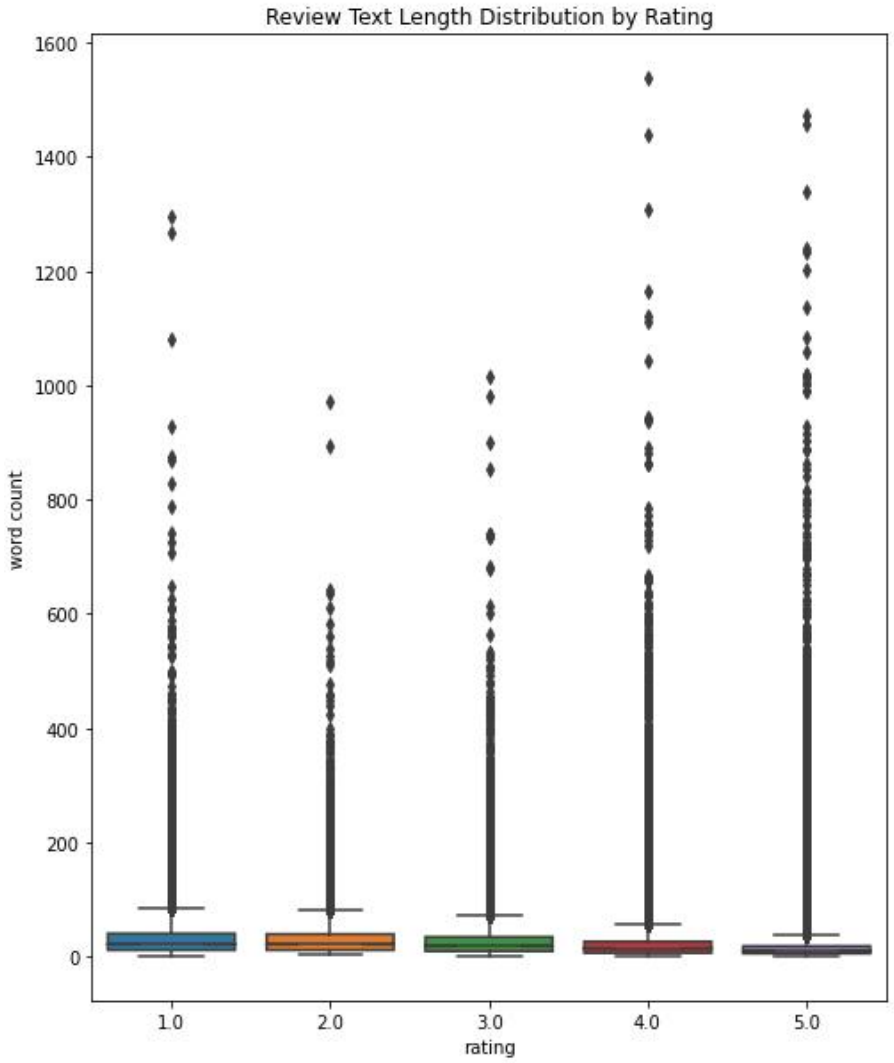
EDA – Review Data



The review month distribution graphs show that the months are quite evenly distributed in the dataset, which we can conclude that the season doesn't play a significant role in the influence of the purchase of the appliances.

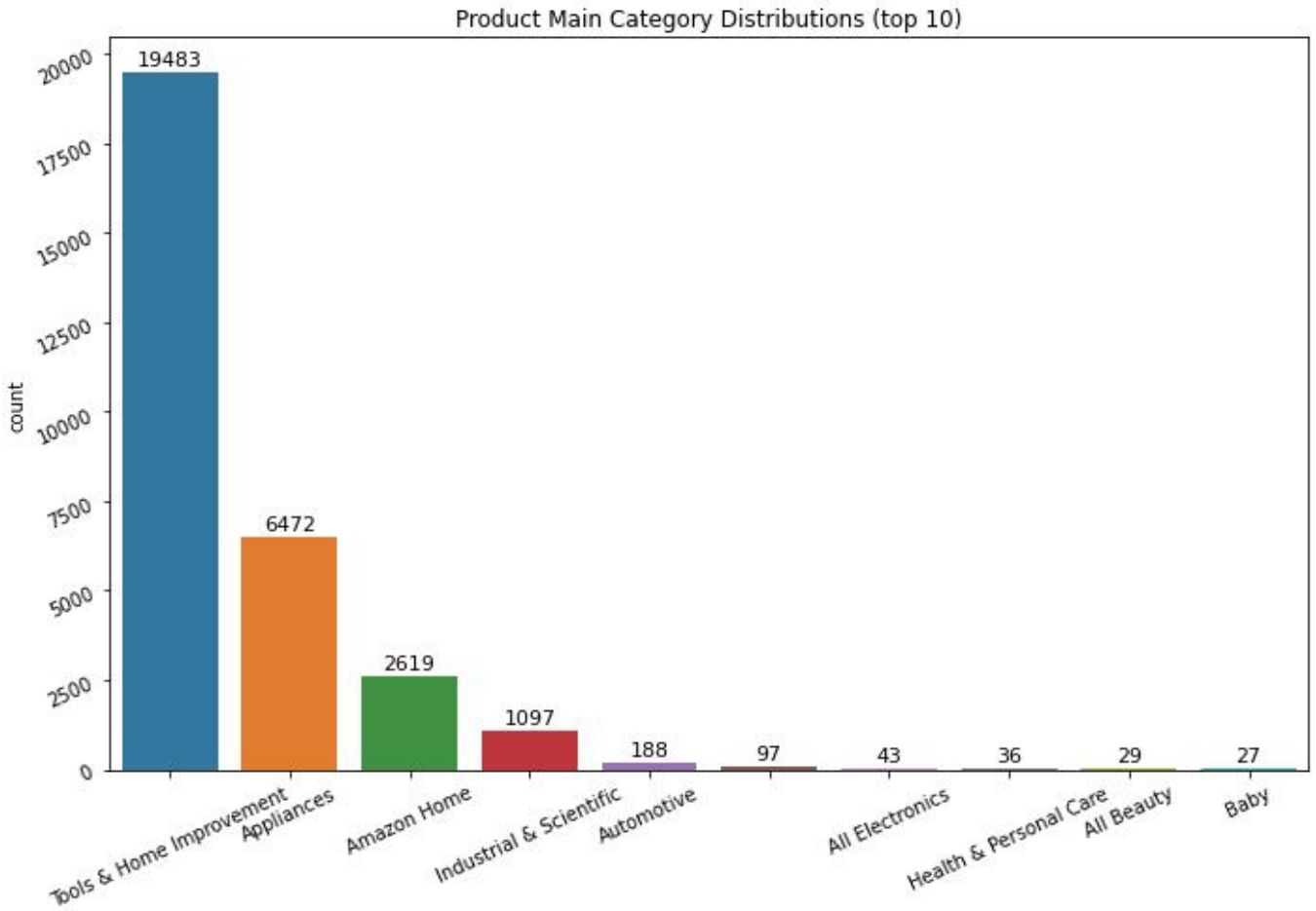
EDA – Review Data

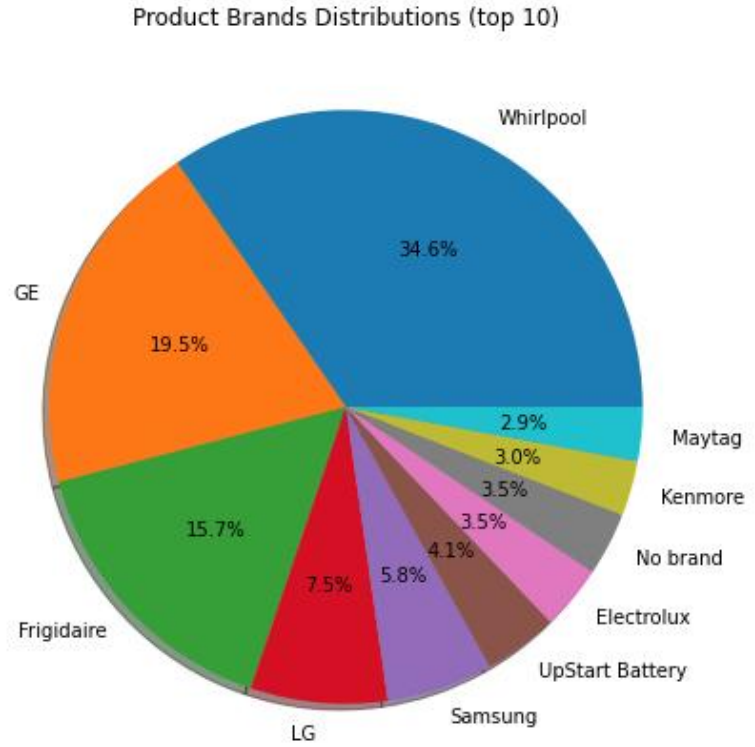




- Most of the reviews contain less than 100 words.
- The word counts distributions for each star rating review are similar. The box plot shows that the 5 stars rating reviews have the lowest interquartile range (IQR) compared to the other 4 ratings, which implies that it has average the shortest review text.

EDA – Product Data





- The majority of the products (64.7%) are in the Tools & Home Improvement category, and the Appliances category also holds 21.5%.
- There are a total of 2,762 brands, and Whirlpool is at the rank 1 position of amount of products.

EDA – Product Data

	asin	counts	rating_mean
421	B000AST3AK	6510	4.422427
5891	B004UB1O9Q	5702	4.341810
1634	B0014CN8Y8	4048	4.676383
17054	B00KJ07SEM	3200	4.409063
5289	B0045LLC7K	2936	4.403270
...	...	...	...
15012	B00GMJ0QCU	1	5.000000
15013	B00GMJ1IDQ	1	5.000000
15014	B00GMJ1XYU	1	5.000000
15018	B00GMJ5SGY	1	4.000000
30251	B01HJHHQM6	1	5.000000

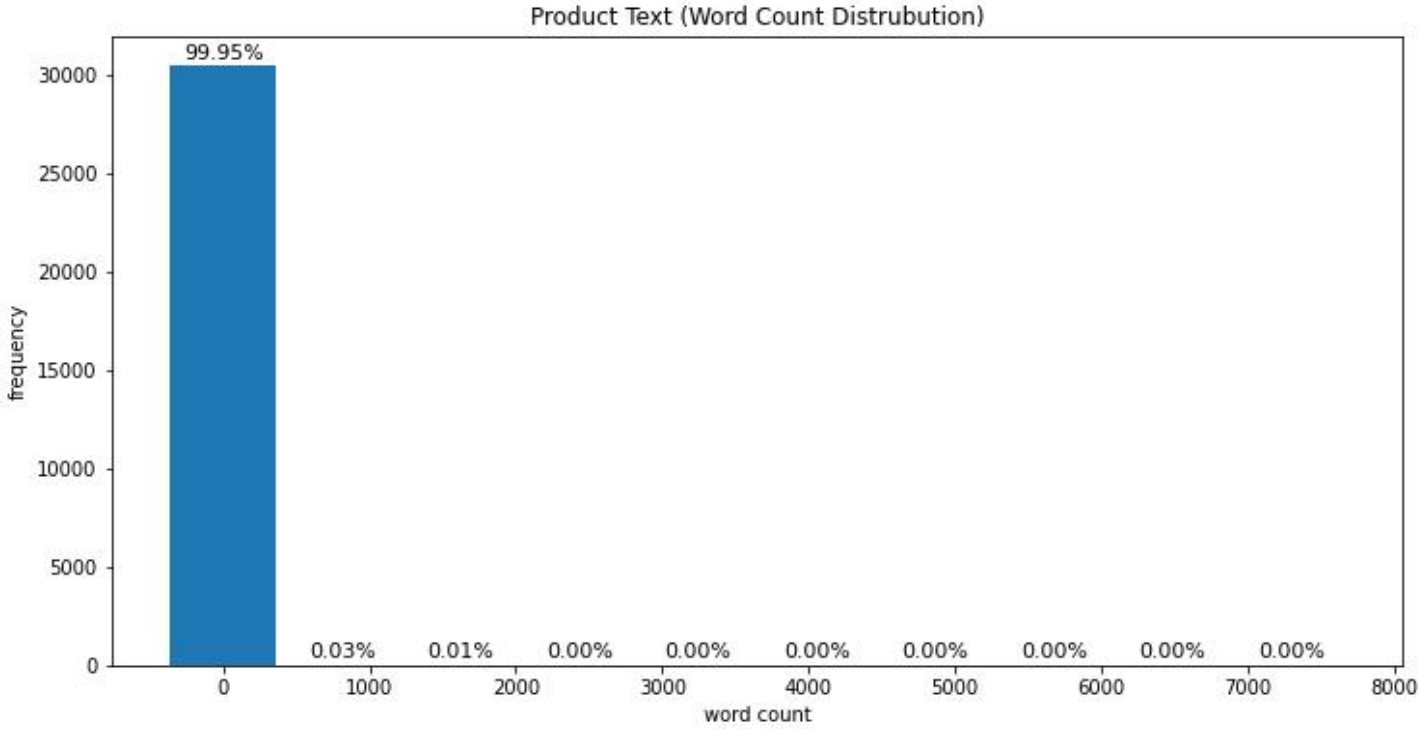
30252 rows × 3 columns

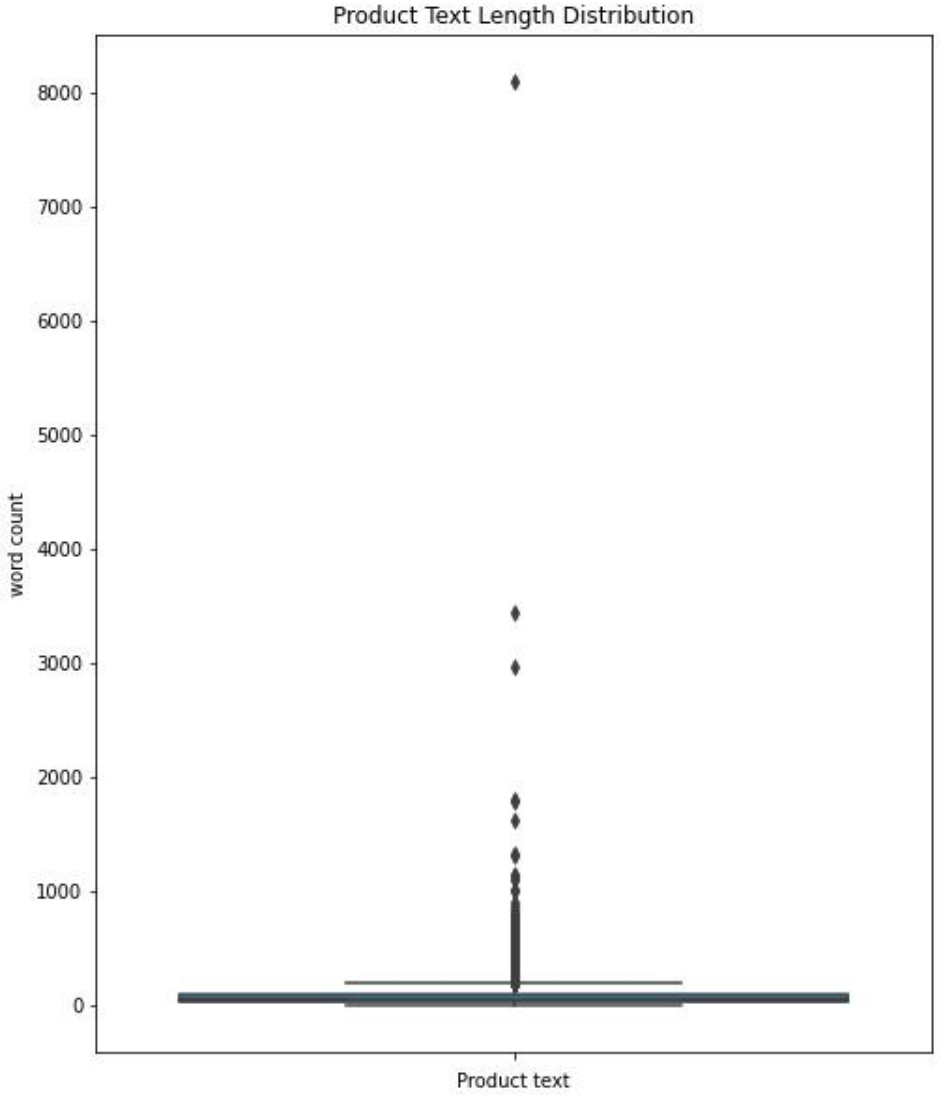
This is the list of the ranking of most reviewed products and their average ratings. Among 30,239 Appliances products, there are 30,252 products were reviewed. So there are some products are not included in the product dataset.

Within this list, the most reviewed product is General Electric MWF Refrigerator Water Filter, and the second most reviewed product is Samsung Genuine DA29-00020B Refrigerator Water Filter, 3 Pack. Both of them are Refrigerator Water Filters.

Most reviewed product: General Electric MWF Refrigerator Water Filter Second most reviewed product: Samsung Genuine DA29-00020B Refrigerator Water Filter, 3 Pack

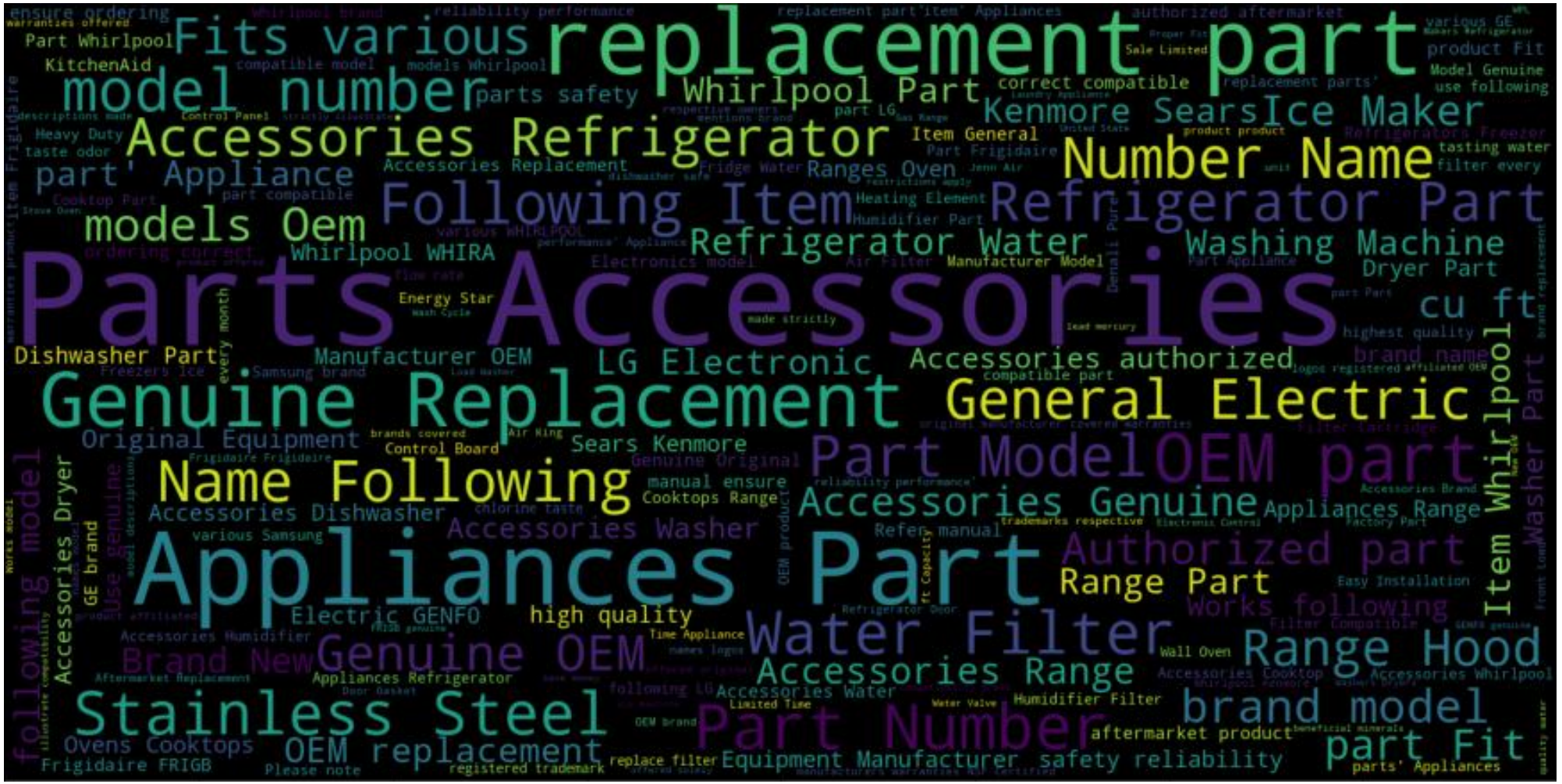
EDA – Product Data





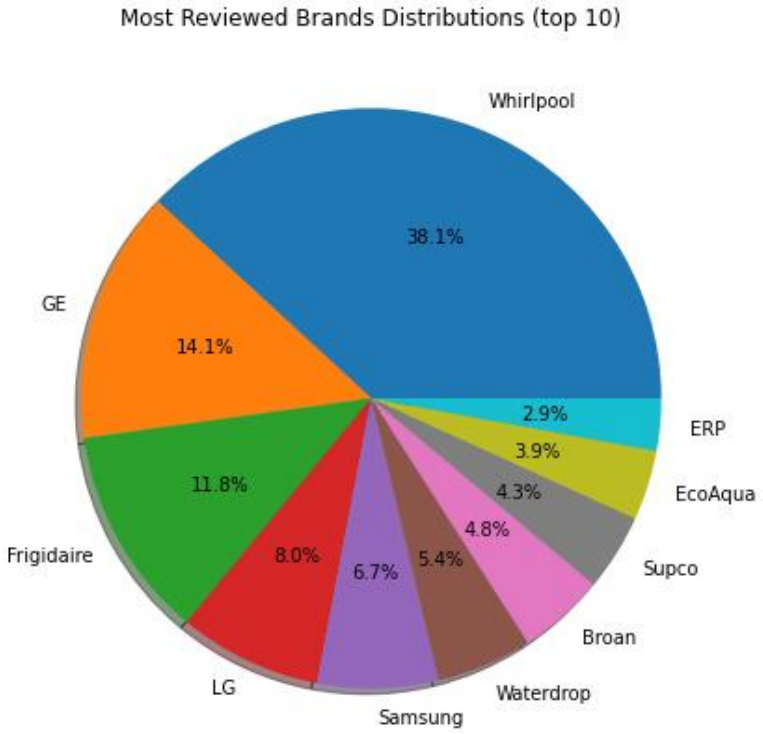
The product text distribution histogram and box plot show that majority of the product text is less than 1000 words. There are only a few outliers that are greater than 2000 words, so for future NLP model development, in order to reduce the padding size, we can consider a smaller number instead.

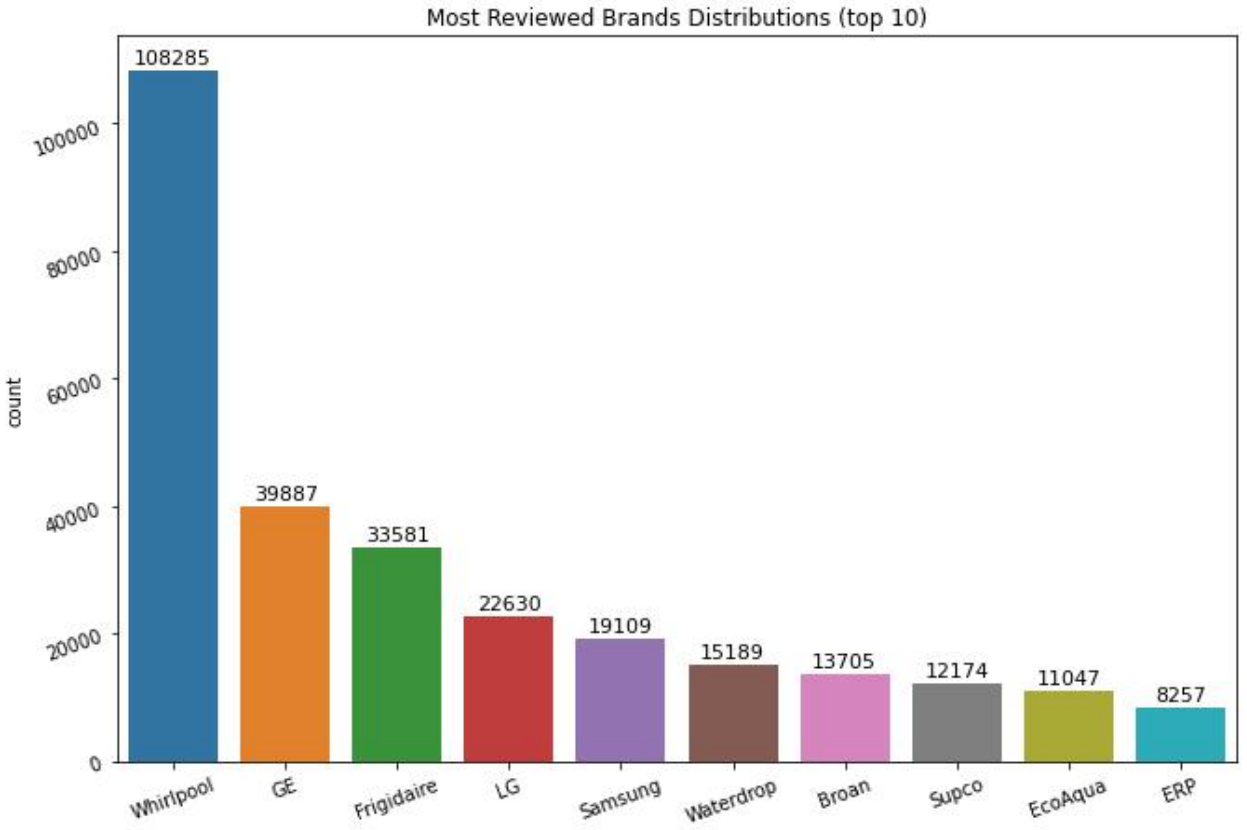
EDA – Product Data



The word cloud shows that the most frequently used words for Appliances products are related to replacement, part, and model number.

EDA – Merged Data





- Most Reviewed Brands Distributions (top 10) graphs show that Whirlpool products have the rank 1 position of amount of reviews.
- There are some other brands in the list that are not in the list of top 10 product numbers, which means offering more products doesn't imply more sales and revenue.

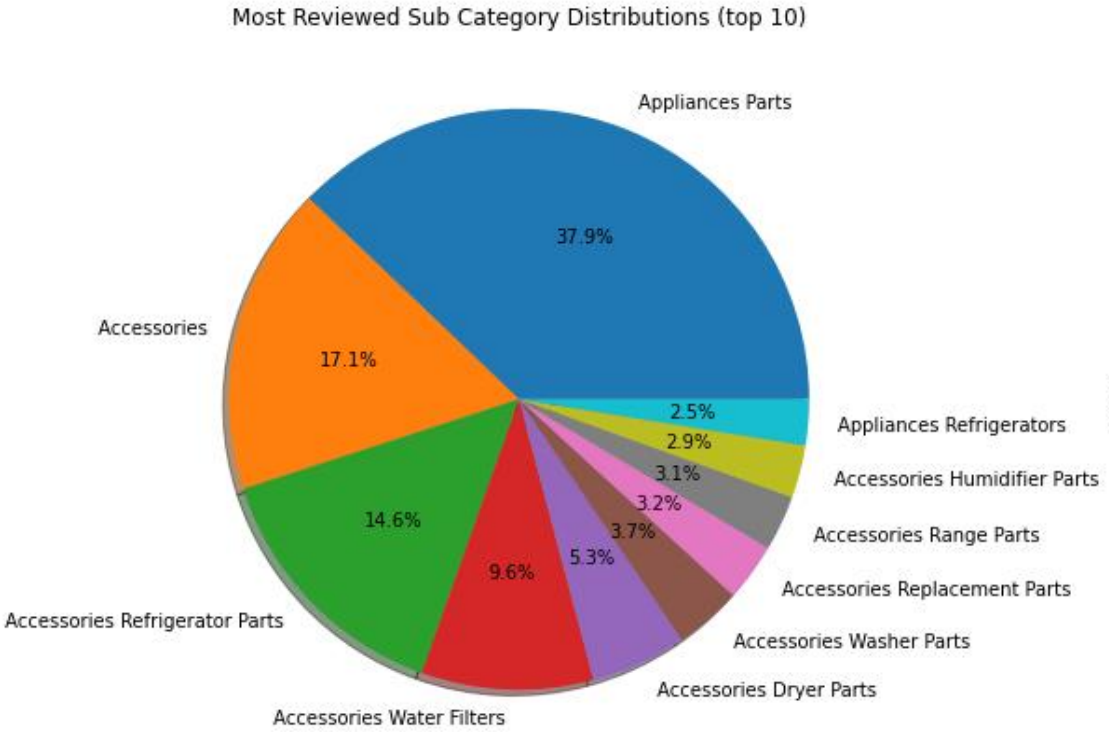
EDA – Merged Data

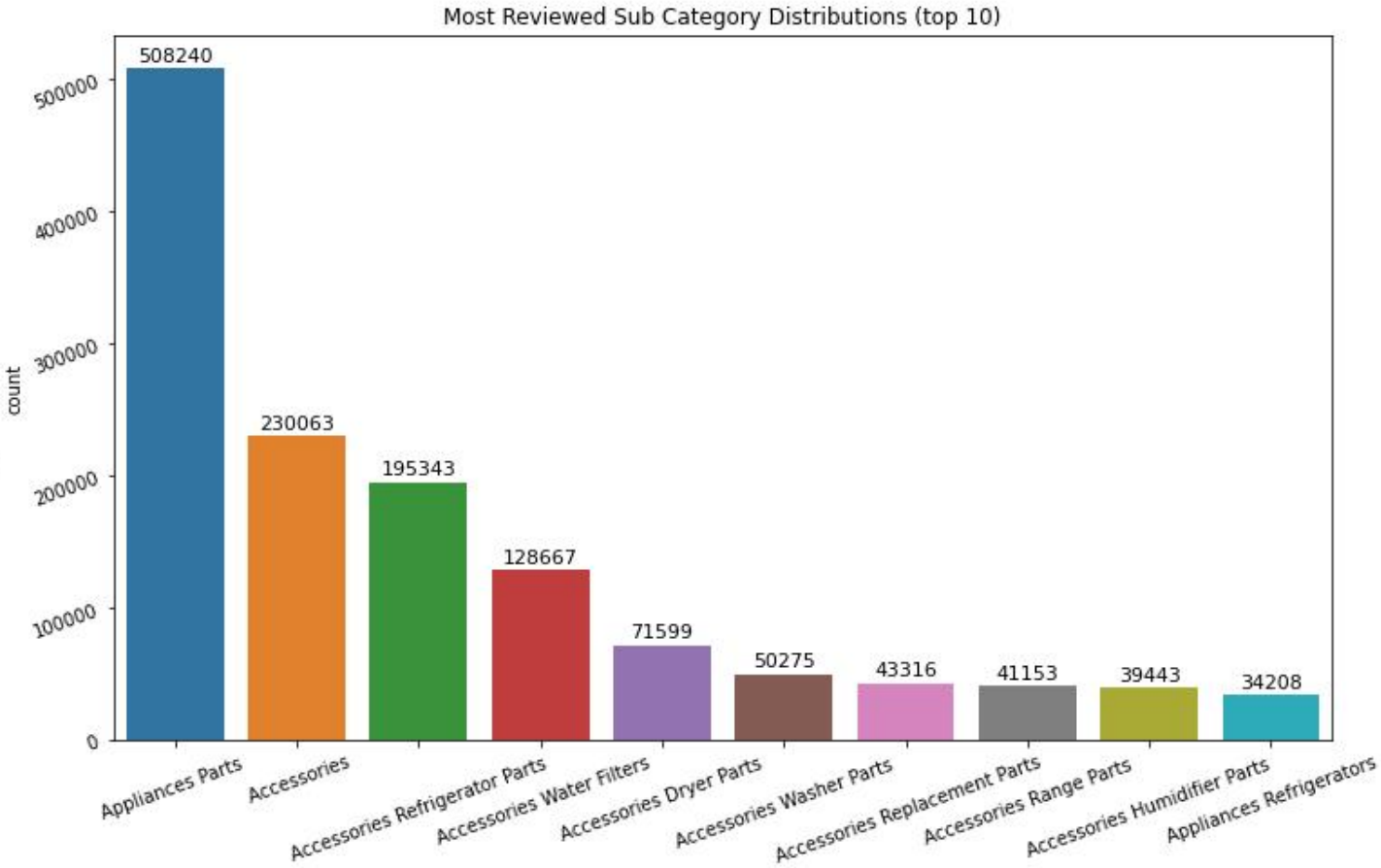
	brand	review_counts	rating_mean
1536	LintEater	6088	4.617280
2570	Waterdrop	15189	4.525183
2294	Supco	12231	4.494808
2591	Whirlpool	108295	4.477280
1419	Kenmore	5490	4.389253
1491	LG	22630	4.382501
769	ERP	8257	4.378225
793	EcoAqua	11047	4.374038
967	Frigidaire	33581	4.367738
991	GE	40213	4.357049

This table shows the top 10 average rating brand (reviews > 5000) in the dataset.

The result show that brand LintEater has the highest average rating 4.62 with over 6,000 reviews. Whereas Whirlpool has the rank 4 in this list.

EDA – Merged Data

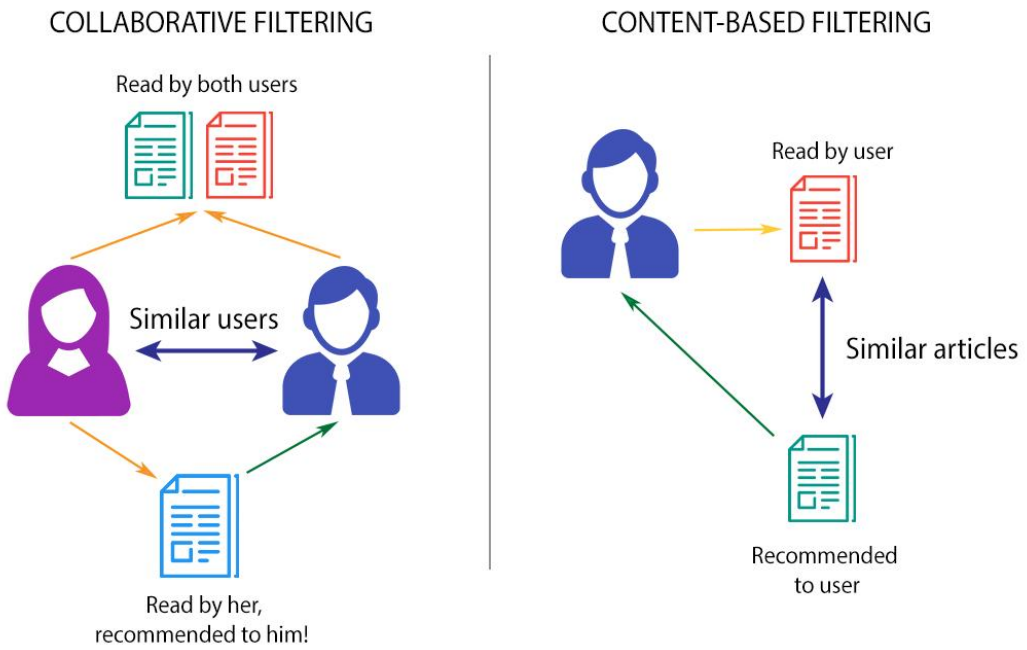




- Most Reviewed Sub Category Distributions (top 10) graphs show that 37.9% of the reviews are in the Appliances Parts sub category, and the Accessories sub category also holds 17.1% in the dataset.

Machine Learning Models

ML Models



Use matrix factorization method against the review data, then apply KNN and Singular value decomposition (SVD) to create clusters for product recommendation.

Use cosine similarity method against the product metadata to identify the similar products for the given one. Apply NLP techniques TF-IDF and LDA topic modeling to output similarity scores.

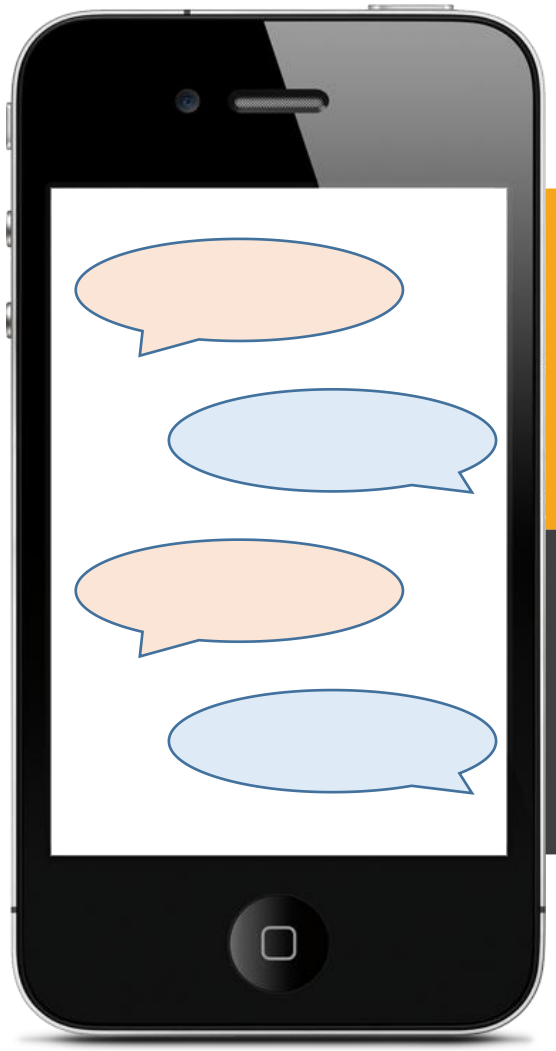
ML Models

Content-based Filtering Model

Collaborative Filtering Model

Hybrid Recommenders Model

ML Models – Integration/Deployment

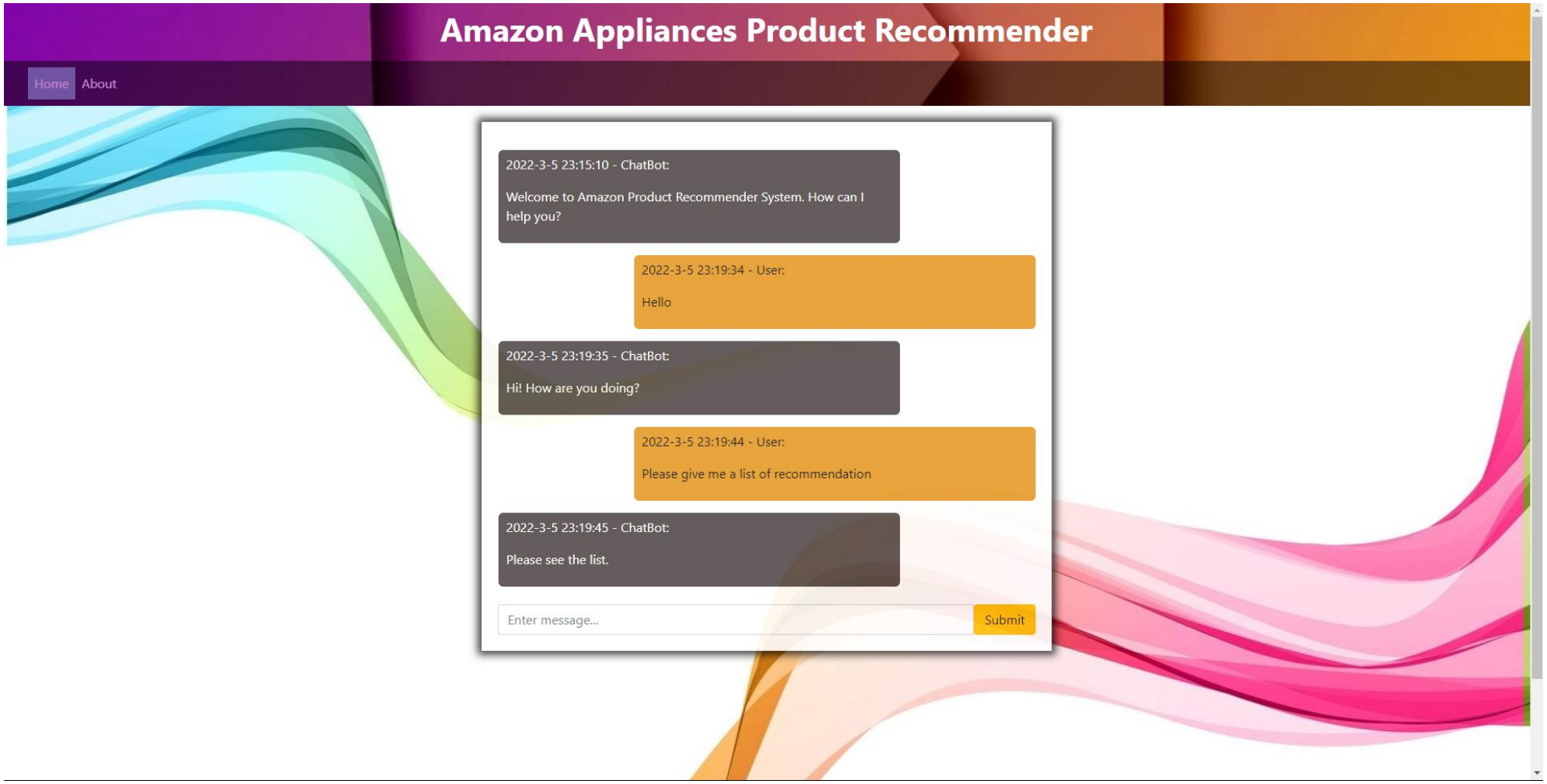


Develop a Flask website and use PythonAnywhere web hosting service to host our recommender system.

Develop a Chabot using DialogFlow platform to assist users for product recommendation and integrate it with the Flask website.

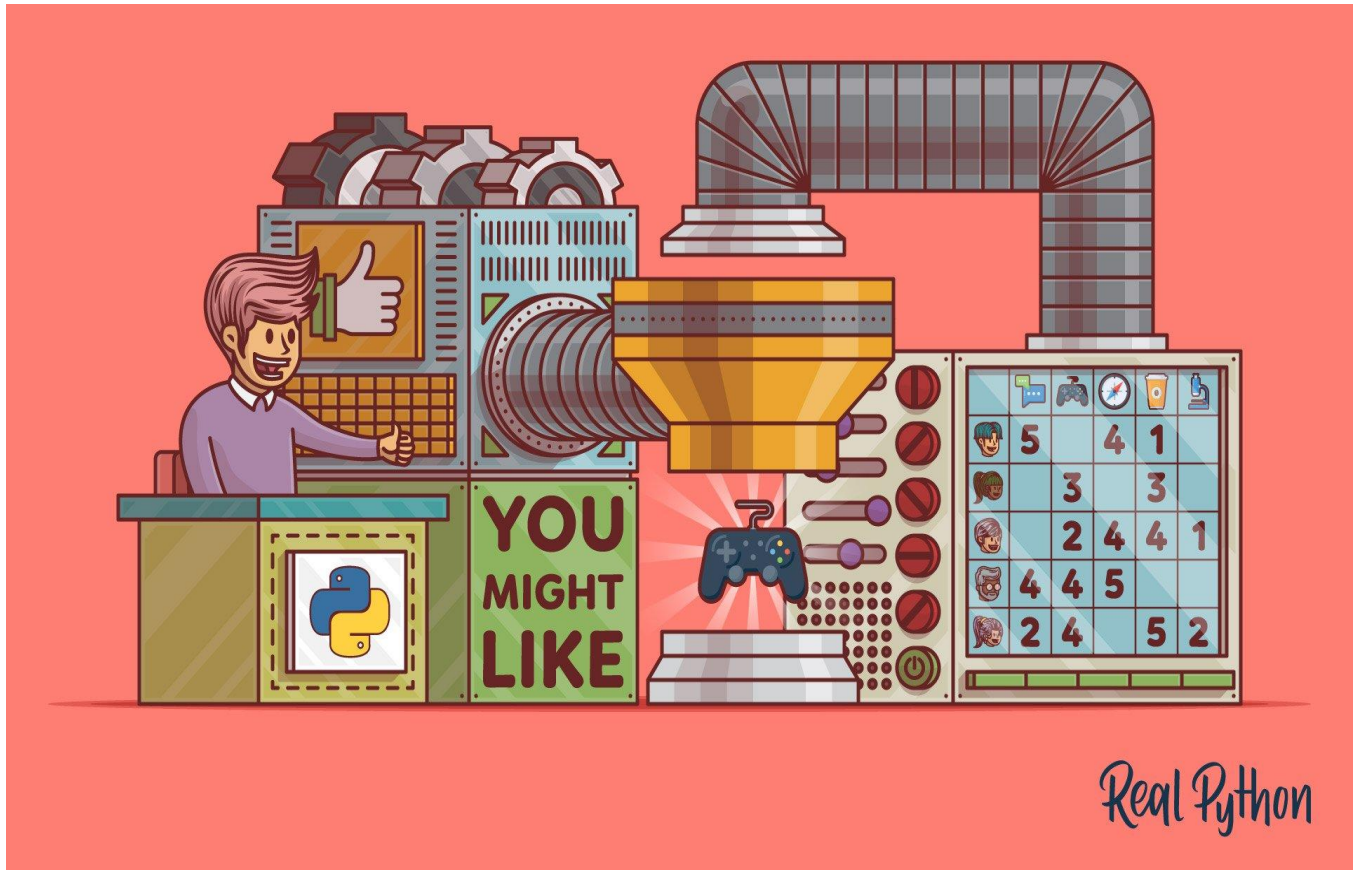
ML Models – Web App Prototype

The flask app prototype is hosted at the url: https://data606project.pythonanywhere.com/



Expected Outcomes

Expected Outcomes



- Develop product recommender systems/models that can accurately predict customers' preferences
- Identify the most useful characteristics to promote certain products to customers
- Understand the role of text data in recommender systems
- Provide a website and Chatbot to assist amazon users to make purchase decisions.
- Provide a comprehensive report of recommender systems for the business owners.

References

- Jianmo Ni, Jiacheng Li, Julian McAuley Empirical Methods in Natural Language Processing (EMNLP), 2019 http://cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19a.pdf
- Doshi, S. (2019, February 20). Brief on Recommender Systems. Medium. Retrieved February 13, 2022, from https://towardsdatascience.com/brief-on-recommender-systems-b86a1068a4dd
- Engineering@ZenOfAI. (2019, August 7). Creating chatbot with Webhooks using python (FLASK) and dialogflow. Medium. Retrieved March 5, 2022, from https://medium.com/zenofai/creating-chatbot-using-python-flask-d6947d8ef805
- BANIK, R. O. U. N. A. K. (2018). Hands-on recommendation systems with Python: Start building powerful and personalized, ... recommendation engines with python. PACKT Publishing Limited.


Thanks