# Amazon Magazine Subscription Recommendation System

How can Amazon leverage existing data to increase magazine subscriptions by generating personalized recommendations for target customers?

Simona Tso

June 21, 2021

# 1. Introduction

## 1.1 Problem Statement

Amazon is one of the leading e-commerce companies in the world. Amazon has millions of products listed on their site, this information overload makes it difficult for customers to find the right products. Recommendation systems solve this issue by suggesting the most relevant products to each customer. This in turn reduces churn and increases sales for Amazon, ensuring that Amazon stays ahead of competitors and becomes even more profitable.

Like most e-commerce companies, Amazon has the functionality that allows it's customers to review products they have purchased. I will utilize this data along with metadata for the magazine subscriptions category within Amazon to generate magazine recommendations (based on similar customers as well as similar products) for Amazon customers.

# 2. Data and pre-processing

## 2.1 Data Overview

Amazon Review Data relating to the 'Magazine Subscriptions' category provided by https://nijianmo.github.io/amazon/index.html was analyzed. The data contains Amazon reviews between May 1996 - Oct 2018.

### 2.1.1 Reviews Data

The reviews data for the 'Magazine Subscriptions' category contained 89,689 reviews in total. This dataset contained information regarding Amazon customers and their review for each magazine purchased. This dataset contains 11 features: image, overall, vote, verified, reviewTime, reviewID, sin, style, reviewerName, reviewText, summary and unixReviewTime.

### 2.1.2 Metadata

The metadata for the 'Magazine Subscriptions' category contained 3,493 magazines in total. This dataset contains product information regarding each magazine sold on Amazon. This dataset contains 13 features: asin, title, feature, description, price, imageURL, related, salesRank, brand, category, main category, tech1 and tech2.

## 2.2 Data cleaning and processing

### 2.2.1 Reviews Data

The reviews dataset contained 89,689 reviews in total, 1,193 of which are duplicated. These duplicates are removed, leaving 88,496 unique reviews. There is a total of 72,098 unique reviewers in the dataset.

For the reviews dataset, I will extract the 'asin' (productID), 'reviewerID' and 'overall' columns. There are 2,428 unique product IDs in the reviews dataset but there are only 2,320 unique product IDs in the metadata. This means that there are some reviewed products that are not listed in the products meta, I will remove these reviews. After removal, the reviews dataset contains 84,167 unique reviews.

#### 2.2.2.1 Category

This feature contains a list of categories the product belongs to. There are 178 unique categories in total. The largest category is 'Professional & Educational Journals', containing 497 magazines. The smallest categories only contain 1 magazine. 442 magazines are not categorized. Each magazine can belong to multiple categories. I cleaned this feature by unescaping special characters.

### 2.2.2 Metadata

The metadata contains 3,493 products in total, 1,173 of which are duplicates. These duplicates are dropped, leaving 2,320 unique products.

#### 2.2.2.1 Category

This feature contains a list of categories each product belongs to. There are 178 unique categories in total. The largest category is 'Professional & Educational Journals', containing 497 magazines. The smallest categories only contain 1 magazine. 442 magazines are not categorized. Each magazines can belong to multiple categories. I cleaned this feature by unescaping special characters.

#### 2.2.2.2 Main Category

This feature contains the main category the product belongs to. This feature was empty for the Magazine subscriptions dataset and the entire column is dropped.

#### 2.2.2.3 Technical Details Table

'Tech1' and 'Tech2' depict the first and second technical detail table of the product. Both features are empty for this dataset, they are dropped.

#### 2.2.2.4 Description

This feature contains a description of the product. There are 1666 unique product descriptions in the dataset. 354 magazines have empty descriptions.

#### 2.2.2.5 Fit

This feature describes the fit of the product. It is empty for the entire dataset, the column is dropped.

#### 2.2.2.6 Title

This feature depicts the name of the product. The feature is cleaned by unescaping special characters. There are 2,173 magazines with incorrect titles, I replaced these titles with an empty string. There are 147 unique magazine titles.

#### 2.2.2.7 Also Buy

This feature contains a list of products the reviewer also bought. The feature is dropped as it is not a product feature but instead describes customer buying behaviour.

#### 2.2.2.8 Brand

This feature contains the brand name of the product. The feature is cleaned by unescaping special characters. There are 147 magazines with missing brands, I replaced these titles with an empty string. There are 942 unique magazine brands. Hearst Magazines is the largest brand containing 255 unique magazines.

#### 2.2.2.9 Feature

This feature contains the features of the product in bullet-point format. The feature is empty for the entire dataset and the column is dropped.

#### 2.2.2.10 Rank

This feature contains the sales rank of the product. There are 1057 missing ranks (almost 50% of the dataset). Since there are a lot of missing values and we don't know how these rankings are calculated, this feature is dropped.

#### 2.2.2.11 Also View

This feature contains a list of products the reviewer also viewed. The feature is dropped as it is not a product feature but instead describes customer browsing behaviour.

#### 2.2.2.12 Details

This feature contains details (format, shipping, publisher & ASIN) of the product stored in a dictionary. Looking closely at each detail, we see that 2166 magazines have the format 'Print Magazine', 148 magazines have an empty format value and the remaining 6 magazines have formats 'Single Issue Magazine', 'DVD', 'Unknown

Binding' and 'Perfect Paperback'. Since most magazines are 'print magazine's, this detail is not very useful.

The details feature also contains Publisher information. Looking at the values, we can see that this is the same as the 'Brand' feature. All magazines have the same shipping information stored in the details feature, this is not useful at all. Finally the details feature also contains 'ASIN' values which we already have a separate feature for. As a result, the details column is dropped.

#### 2.2.2.13 Similar Item

This feature contains a list of similar products. The feature is empty for the entire dataset and the column is dropped.

#### 2.2.2.14 Date

This feature stores the date the product was added into Amazon's database. The feature is empty for the entire dataset and the column is dropped.

#### 2.2.2.15 Price

This feature stores the price of the product. The feature contains invalid values and the column is dropped.

#### 2.2.2.16 ImageURL

The 'imageURL' feature contains the URL of the product image and the 'imageURLHighRes' feature contains the url of the high resolution product image. Both of these columns are dropped as I will not be using this information for the recommendation systems I intend to build.

# 3. Exploratory Data Analysis

## 3.1 Data Overview

The cleaned reviews dataset now contains 84,167 unique reviews, whilst the cleaned metadata contains 2320 unique products.

## 3.2 Products and Reviews Data Exploration

### 3.2.1 Distribution of Ratings

#### 3.2.1.1 Distribution of Overall Ratings

Fig 1 shows the distribution of 84,167 magazine subscription ratings  We can see that 60.3% of all ratings given are 5-stars, followed by 4-star ratings (14.1%) and 1-star ratings (12.2%).  Very few ratings are 2 or 3-stars.  This suggests that reviewers tend to leave reviews when they are extremely satisfied with the product (4/5 stars) or when they are extremely dissatisfied with a product (1 star).

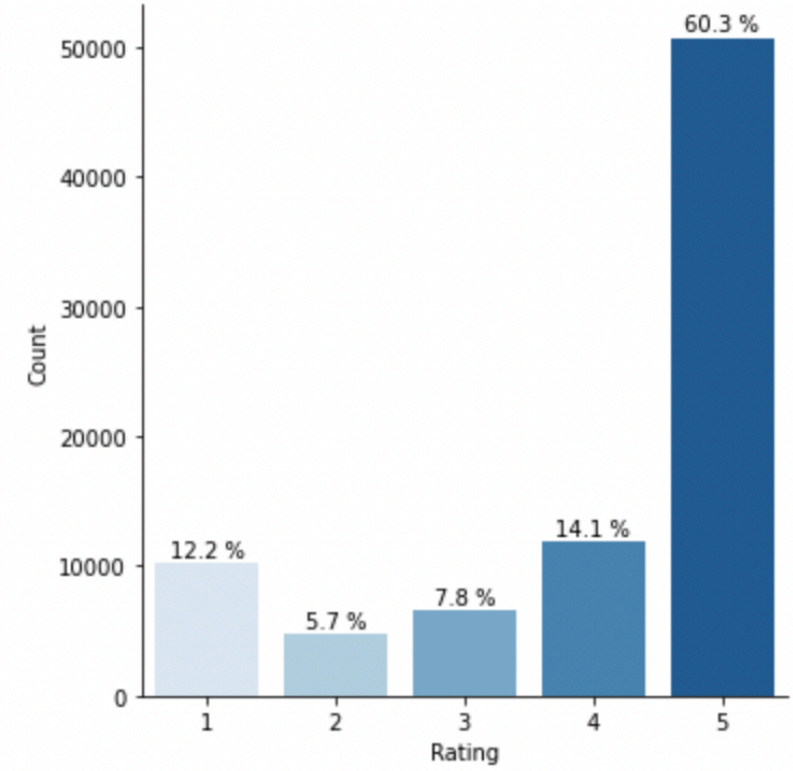



Fig 1. Distribution of all magazine subscription ratings.

#### 3.2.1.2 Distribution of average rating per reviewer

We can see that fig 2's distribution is very similar to the distribution of magazine subscription ratings (fig 1). There is a disproportionately high number of reviewers with an average rating between 4.75 - 5.00. In order to obtain such a high average rating, reviewers must be consistently giving out high ratings across all their purchases. The second most common average rating is between 1.00 - 1.25. In order to achieve such a low average rating, it suggests that these reviewers are consistently giving out low ratings across all their purchases.

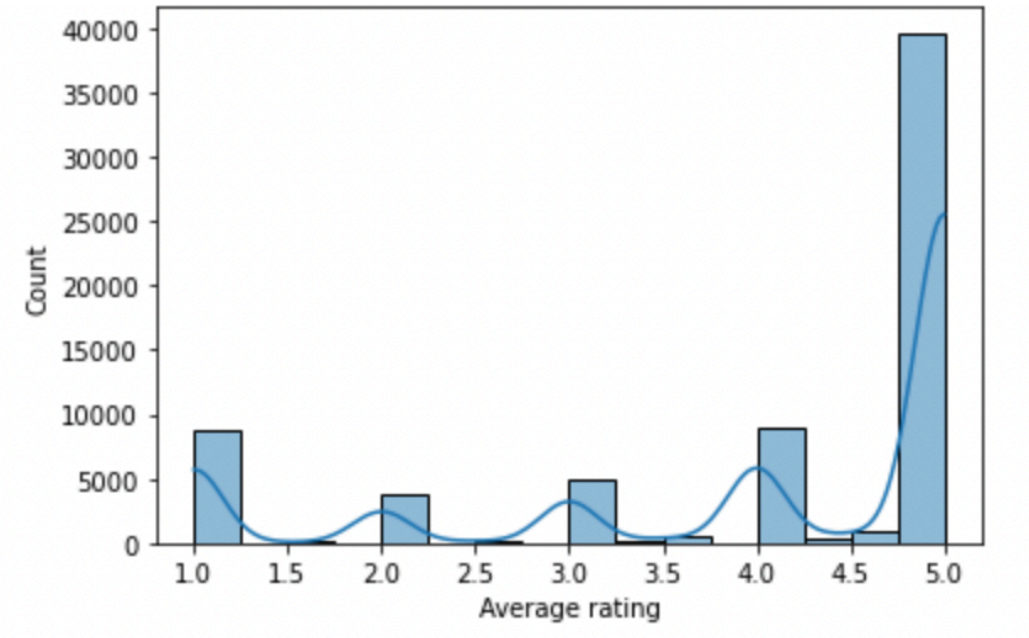



Fig 2. Distribution of average ratings per reviewer.

Fig 3 shows the same distribution with only reviewers that reviewed more than 1 magazine. We can see that the average rating is highly left skewed. When compared to fig 2, we can see that the scale on the y-axis is 160 times smaller, this indicates that most reviewers only reviewed 1 magazine. We can also see that the proportion of average ratings between 1.00-1.25 dropped significantly, this means

that most reviewers with an average rating of 1.00-1.25 only reviewed 1 magazine. The proportion of reviewers with an average rating of between 4.75 - 5.00 is still the highest, indicating that many reviewers consistently gave high ratings to most of the magazines they rated.

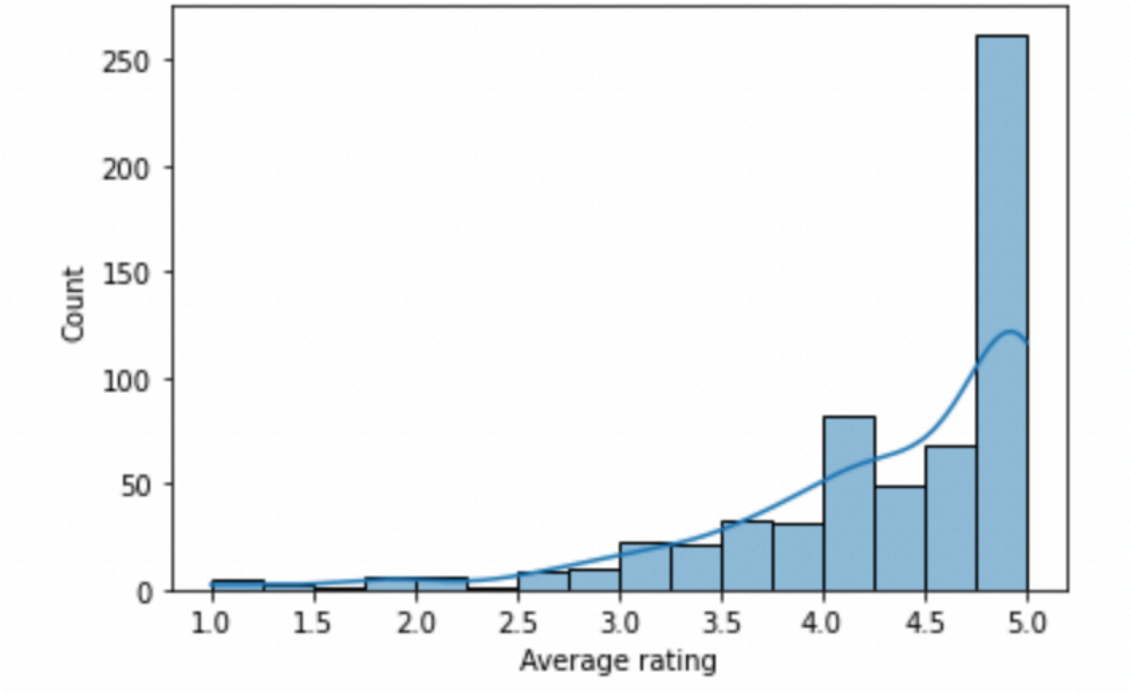



Fig 3. Distribution of average ratings per reviewer (for reviewers who reviewed more than 1 magazine).

From fig 4, we can see that very few reviewers gave many reviews. Based on the density of the scatterplot and opacity of each datapoint, it is evident that most reviewers gave less than 10 reviews. Reviewers that gave a high number of reviews generally have higher average ratings. This suggests that reviewers tend to give more high ratings than low ratings on average. There is an outlier where a reviewer gave a total of 54 reviews.

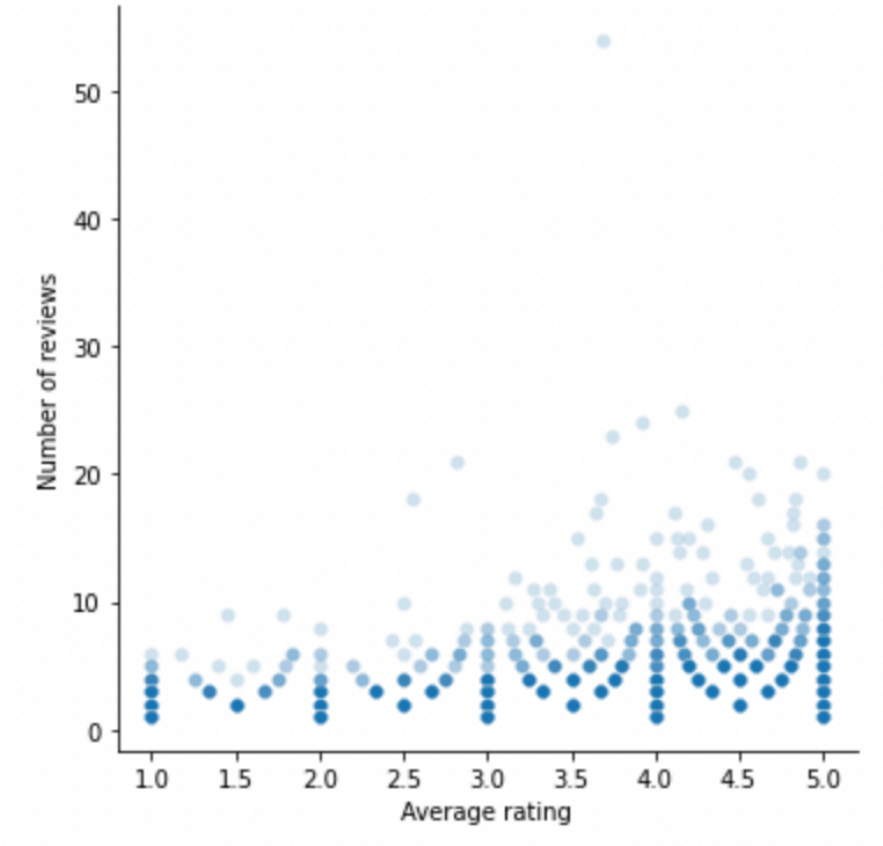



Fig 4. Distribution of average ratings and number of reviews per reviewer.

#### 3.2.1.3 Distribution of Average Rating Per Magazine

From fig 5 we can see that over 500 magazines have an average rating of between 4.9 - 5.0. The second highest bar is found at average rating 3.9 - 4.0. Aside from the small peaks at average rating 1.0 - 1.1 and 2.9 - 3.0, there doesn't seem to be many magazines with average ratings below 3.5, this indicates that Amazon carries highly rated magazines that it's customer enjoys.

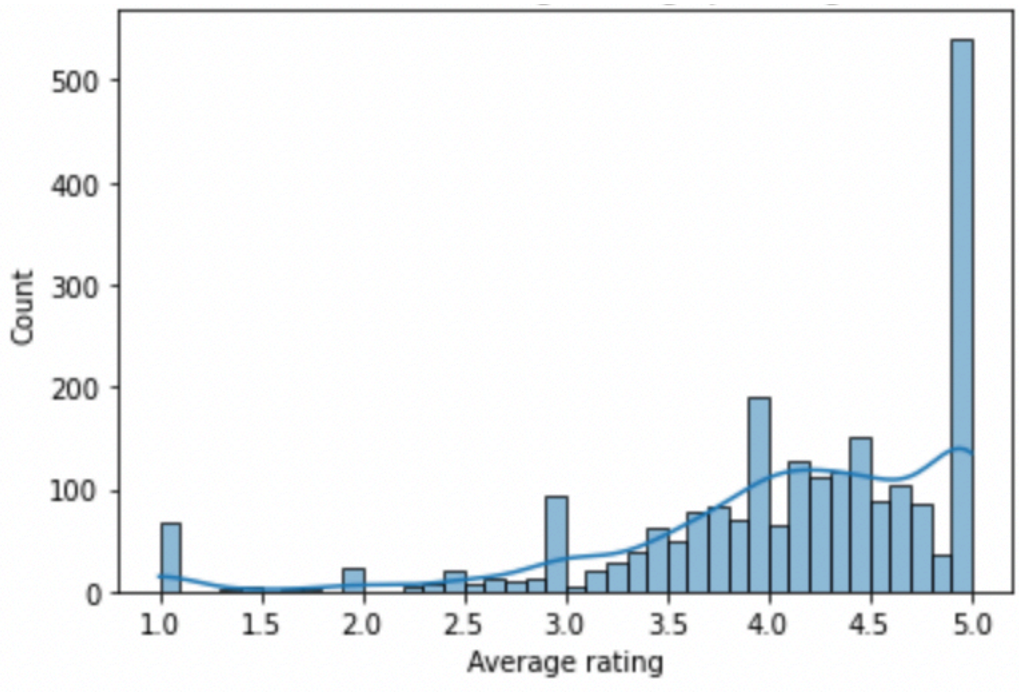



Fig 5. Distribution of average ratings per magazine.

Fig 6 shows the same distribution as fig 5 but omitting magazines with only 1 review. We can see that the scale on the y-axis is approximately 4 times smaller, indicating that a very high proportion of magazines only received 1 rating. The distribution appears to be much more smooth here, the peaks at 1, 2 and 3 are no longer present. This indicates that most magazines with average reviews 1, 2 and 3 only had 1 review. The new peak for average rating per magazine is between 4.1 - 4.2, with over 120 magazines. The peak at 5 dropped significantly compared to fig 5, this also shows that a lot of magazines with an average rating of 5 only had 1 review.

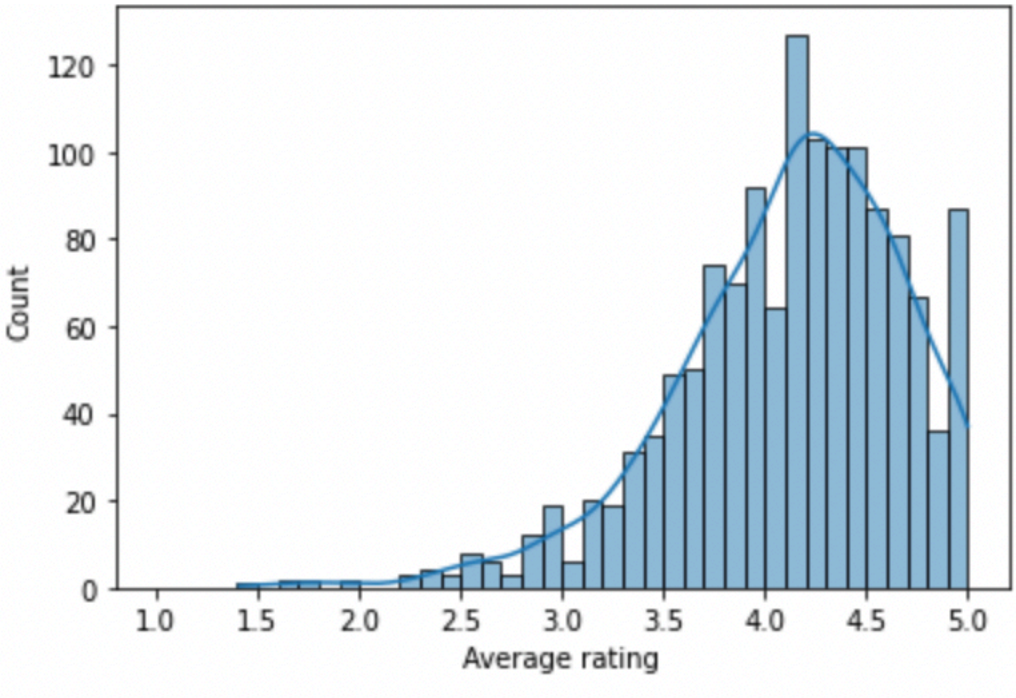



Fig 6. Distribution of average ratings per magazine (for magazines with more than 1 review).

Based on the density of fig 7 and the opacity of each individual datapoint, we can see that most magazines have very few reviews. In fact, 75% of magazines have less than 20 reviews. Magazines with over 50 reviews have average ratings above 2.5. This suggests that magazines that receive more reviews tend to have high average ratings. There are several outliers where magazines received thousands of reviews.

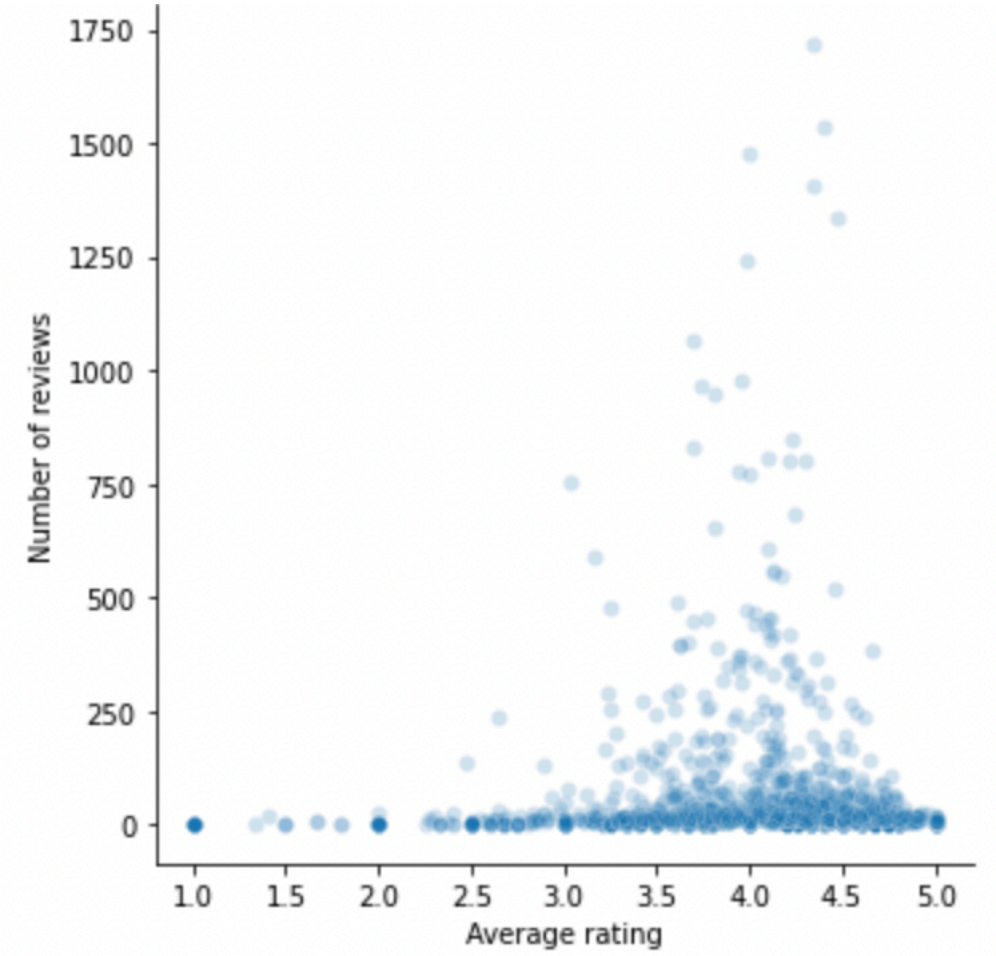



Fig 7. Distribution of average ratings and number of reviews per magazine.

### 3.2.2 Distribution of Reviews

#### 3.2.2.1 Distribution of Number of Reviews Per Magazine

Fig 8 contains a highly right skewed distribution. Most magazine subscriptions received only 1 review. Very few magazine subscriptions have many ratings, 50% of magazines have 6 or less reviews. The most reviewed magazine subscription received 1,718 reviews.

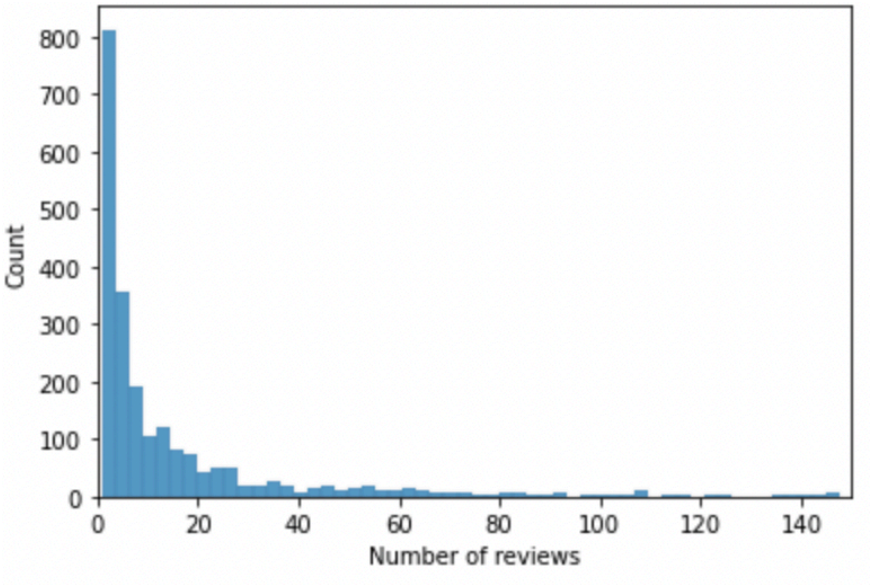



Fig 8. Distribution of number of reviews per magazine (clipped at 150).

### 3.2.2.2 Distribution of Number of Reviews Per Reviewer

Fig 9 shows that almost all reviewers reviewed less than 10 magazines. In fact, over 75% of reviewers only reviewed 1 magazine. There is an outlier where a single reviewer reviewed 54 magazines.

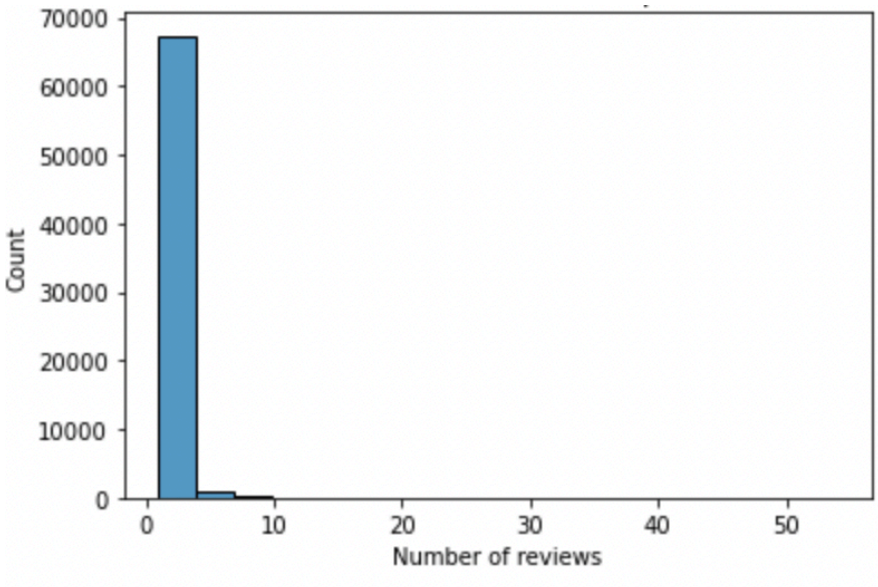



Fig 9. Distribution of number of reviewers per reviewer.

## 3.2.3 Most and Least Popular Magazines

### 3.2.3.1 Top 5 Most Reviewed Magazines

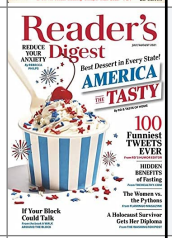
| | Magazine | Title | Brand | Categories | Number of Reviews |
|---|---|---|---|---|---|
| 1 |  | National Geographic | National Geographic Partners LLC | • Professional & Educational Journals<br>• Professional & Trade<br>• Humanities & Social Sciences<br>• History | 1718 |
| 2 |  | The Family Handyman | Trusted Media Brands, Inc | • Home & Garden<br>• How-to & Home Improvements | 1537 |
| 3 |  | Popular Science | Bonnier Corporation | • Science<br>• History & Nature<br>• Essays & Commentary | 1480 |
| 4 |  | Reader's Digest | Trusted Media Brands, Inc | • Literary<br>• Sci-Fi & Mystery<br>• Literary Magazines & Journals | 1409 |
| 5 |  | Food Network Magazine | Hearst Magazines | • Cooking<br>• Food & Wine<br>• Recipes & Techniques | 1339 |

Table 1. Top 5 most reviewed magazines.

Table 1 contains the top 5 most reviewed magazines. The most reviewed magazine is National Geographic which has 1718 reviews. These magazines span across multiple different categories and brands, indicating that there isn't a single brand or category that dominates the top charts.

### 3.2.3.2 Top 5 Magazines with the Highest Average Rating

| | Magazine | Title | Brand | Categories | Average Rating | Number of Reviews |
|---|---|---|---|---|---|---|
| 1 |  | Mystery Scene | Kbs Communications | • Literary<br>• Sci-Fi & Mystery<br>• Literary Magazines & Journals | 5.0 | 22 |
| 2 |  | House & Home | Canadian Home Publishers | • Home & Garden | 5.0 | 18 |
| 3 |  | Gray's Sporting Journal | Morris Communications Corp | • Sports<br>• Recreation & Outdoors<br>• Sports & Leisure<br>• Hunting & Firearms | 5.0 | 16 |
| 4 |  | B.O.S.S | Clay & Clay LLC | | 5.0 | 15 |
| 5 |  | Dirt Bike | Hi-Torque Publications | • Sports<br>• Recreation & Outdoors<br>• Sports & Leisure<br>• Cycling | 5.0 | 12 |

Table 2. Top 5 most magazines with the highest average rating.

Table 2 contains the top 5 magazines with the highest average rating and number of reviews. Since a lot of magazines received very few ratings, a threshold was set so that the top magazines must have at least 5 reviews. The highest rated magazine is Mystery Scene which has 22 5-star reviews. There are two 'Sports' related magazines in the top 5, this suggests that Amazon's customers may have a preference for sports related magazines. The top 5 magazines span across

multiple different categories and brands, indicating that there isn't a single brand or category that dominates the entire top charts.

### 3.2.3.3 Top 5 Magazines with the Lowest Average Rating

| | Magazine | Title | Brand | Categories | Average Rating | Number of Reviews |
|---|---|---|---|---|---|---|
| 1 |  | American Baby | Meredith | | 1.41 | 17 |
| 2 |  | Elle Decor - Italian Ed | Hearst Magazines Italia Spa | • Professional & Education Journals<br>• Professional & Trade<br>• Arts<br>• Decorative Arts | 1.67 | 6 |
| 3 |  | Variety | Penske Business Media, LLC | • Professional & Education Journals<br>• Professional & Trade<br>• Entertainment & Media | 1.67 | 6 |
| 4 |  | Elle - French Ed | Hachette Filipacchi | • Fashion & Style<br>• Women | 1.80 | 5 |
| 5 |  | V Magazine | Visionaire Publishing | • Professional & Education Journals | 1.80 | 5 |

Table 3. Top 5 magazines with the lowest average rating.

Table 3 contains the top 5 magazines with the lowest average rating and highest number of reviews. A threshold of 5 was also set for this analysis. 3 out of 5 of the lowest rating magazines are categorized as 'Professional & Educational Journals',

this indicates that Amazon's customer base tend to dislike this type of content. The magazine with the lowest average rating is 'American Baby', which has an average rating of 1.41 over 17 reviews.

### 3.2.4 Magazines Categories

#### 3.2.4.1 Top Categories by Number of Magazines

Fig 10 shows the number of magazines in the top 10 categories. 'Professional & Education Journals' and 'Professional & Trade' magazines are most common, with over 400 magazines in each category. All other magazine categories contain less than 250 magazines.

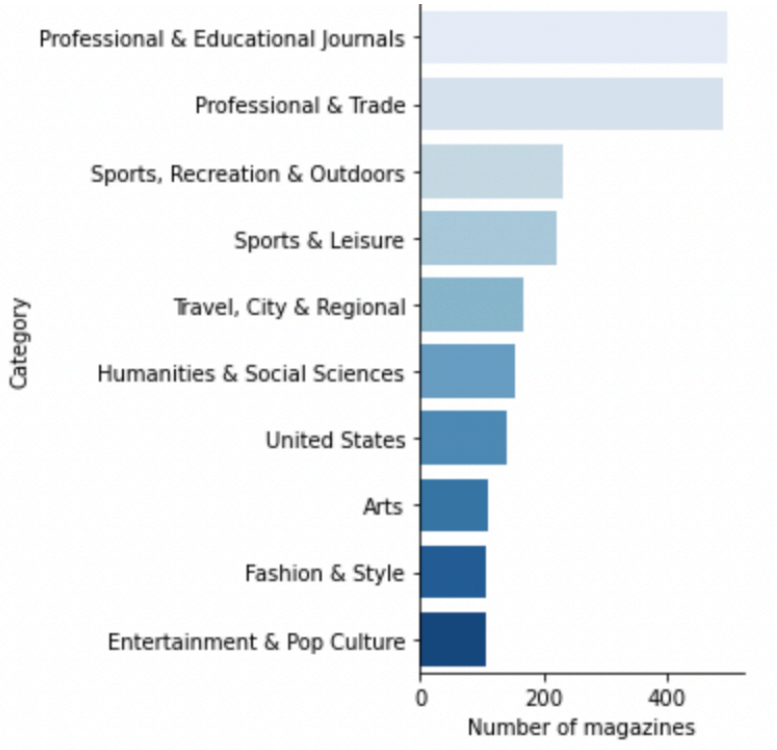



Fig 10. Distribution of the number of magazines per category.

### 3.2.4.2 Top Categories by Number of Reviews

Fig 11 shows the number of reviews in the top 10 categories. 'Professional & Education Journals' and 'Professional & Trade' magazines are not only the most common (fig 10), but also the most reviewed, with over 10,000 reviews each. Uncategorized magazines are also among the top 10 most reviewed categories.

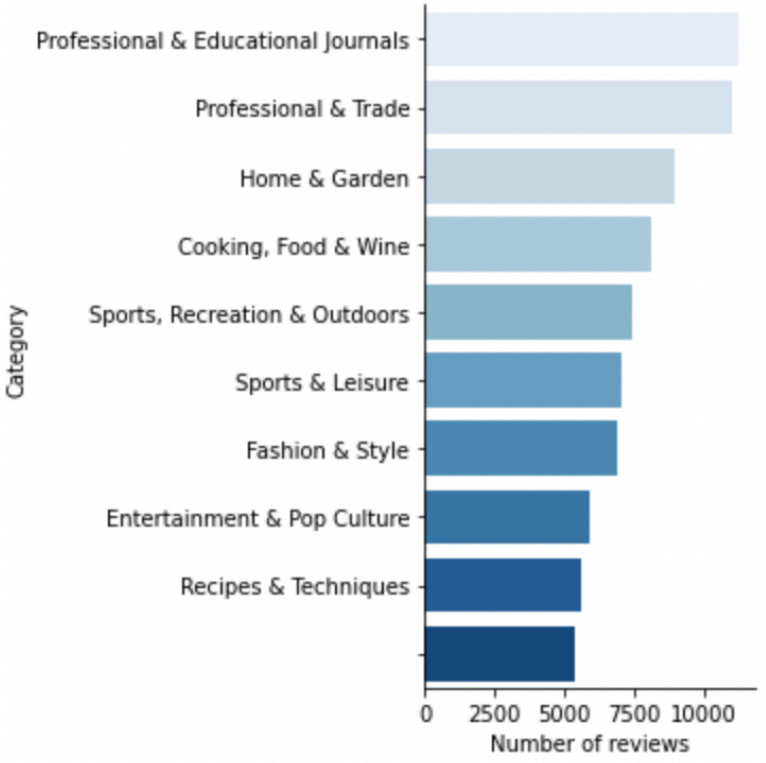



Fig 11. Distribution of the number of reviews per category.

### 3.2.4.3 Top Categories by Average Rating

Fig 12 shows the average ratings in the top 10 categories. The top 10 categories each have average ratings above 4.6 and at least 8 reviews. 'Advertising' and 'Geography' have the highest average rating of 5.0 and 18 and 11 number of reviews respectively. Out of the top 10 categories, 'Education' has the highest number of reviews (180), followed by History of Education (72). This indicates that customers

who purchase an education related magazine are likely to leave a review (most likely a positive one too!).

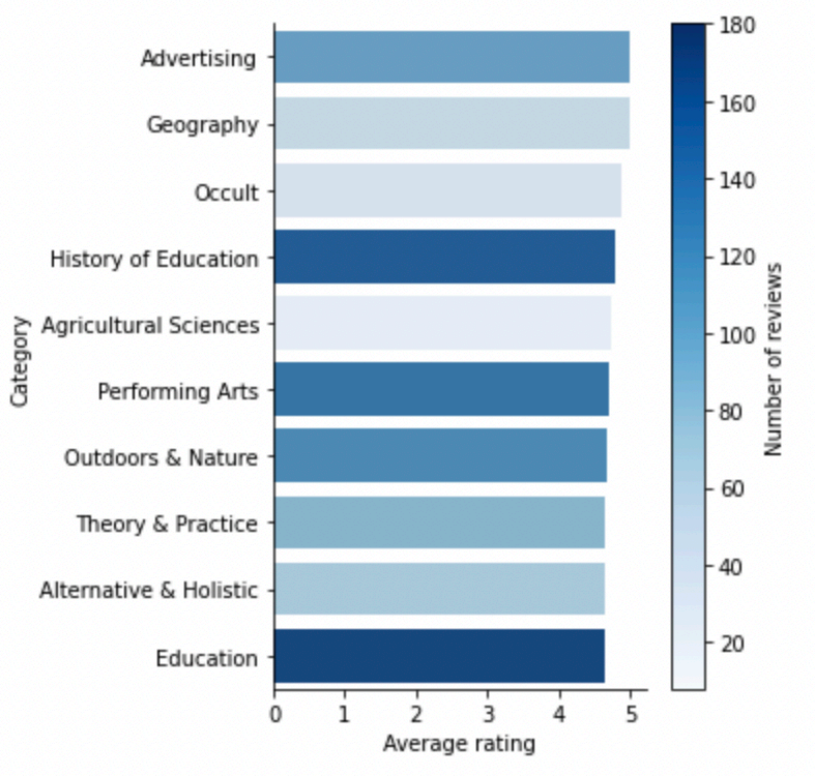



Fig 12. Distribution of average ratings and number of reviews per category (Clipped at top 10 highest average rating).

### 3.2.4.4 Bottom Categories by Average Rating

Fig 13 shows the average ratings in the bottom 10 categories. The 10 categories with the lowest average ratings all have ratings below 3.6 and at least 10 reviews. The lowest rated category is 'Finance', with an average rating of 3.07. This category has a total of 2625 reviews which means that the low rating represents a large proportion of customers.

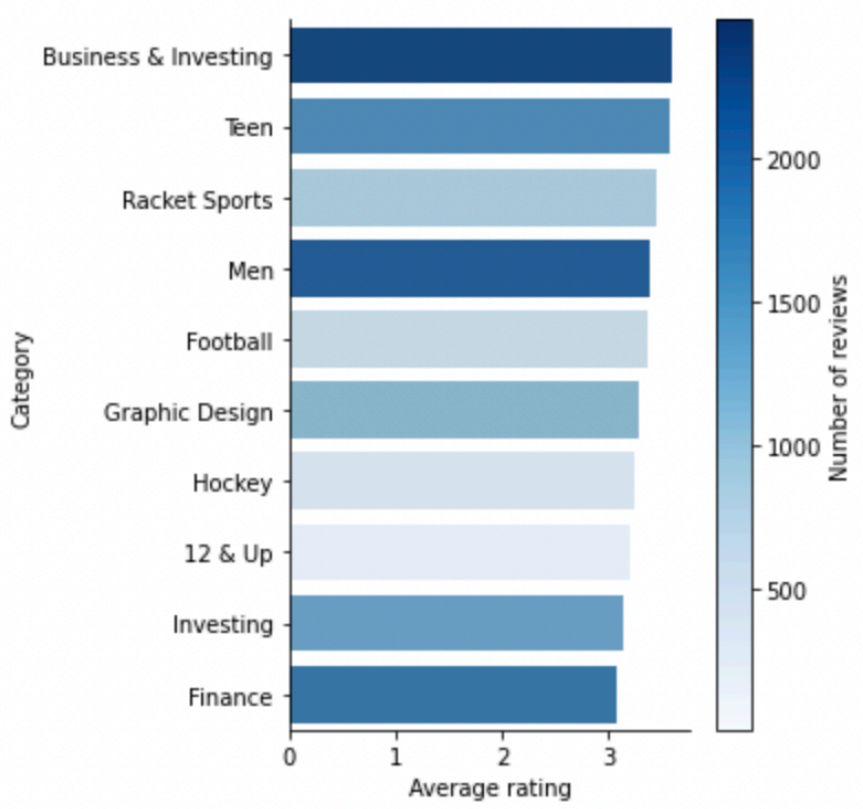



Fig 13. Distribution of average ratings and number of reviews per category (Clipped at top 10 lowest average rating).

## 3.2.5 Magazines Brands

### 3.2.5.1 Top Brands by Number of Magazines

Fig 14 shows the number of magazines each brand contains. Hearst Magazine carries the largest number of magazines (255), followed by Conde Nast (184). All other brands have less than 50 magazines available on Amazon.

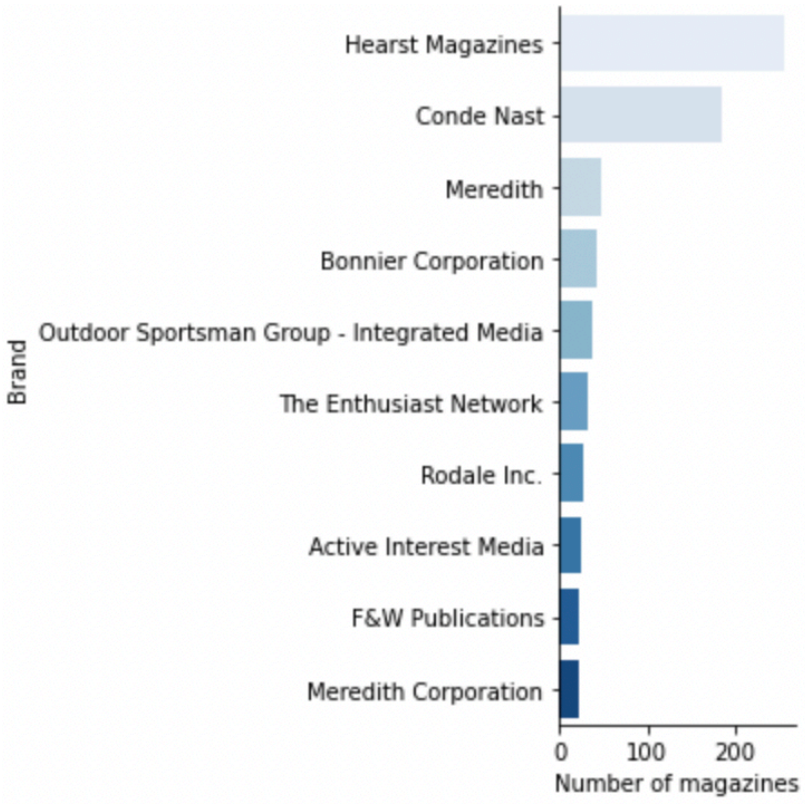



Fig 14. Distribution of the number of magazines per brand.

### 3.2.5.2 Top Brands by Number of Reviews

It is not surprising that Hearst Magazine also received the highest number of ratings since there is a total of 255 available magazines to review. Hearst Magazine received a total of 13,788 reviews, which is over double that of the second most reviewed magazine brand. The second most reviewed magazine brand is 'Meredith Corporation', it was ranked 10th based on its number of magazines (22). It is interesting that Meredith Corporation magazines were reviewed 6,941 times and have an above average rating of 4.04.

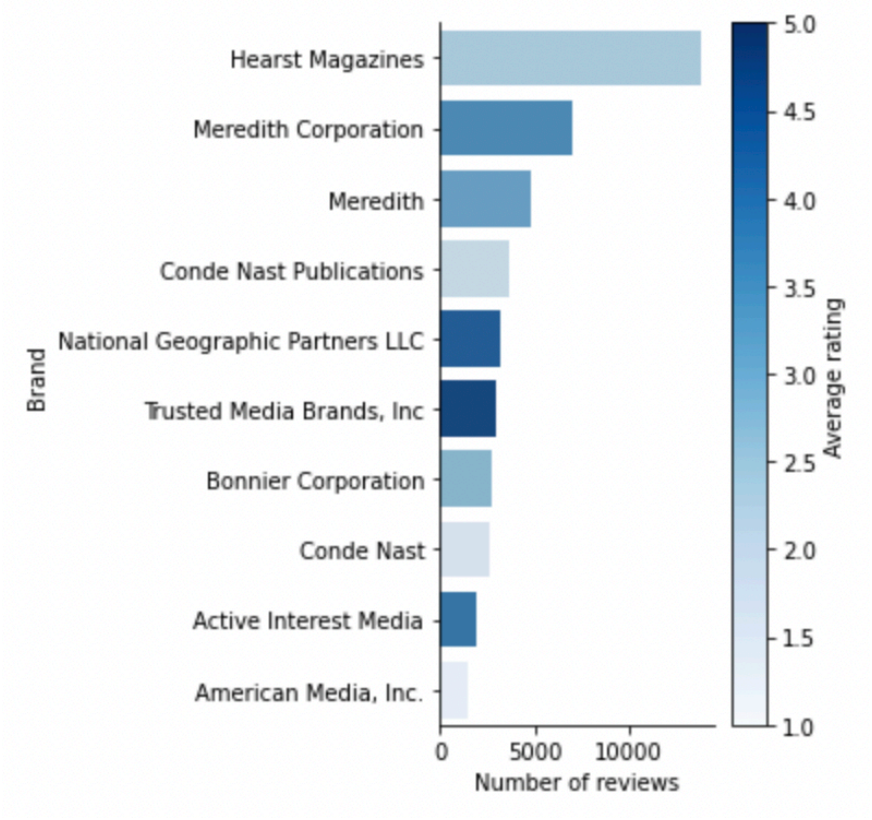



Fig 15. Distribution of the number of reviews and average rating per brand (Clipped at top 10 highest number of reviews).

#### 3.2.5.3 Top Brands by Average Rating

The top 10 brands sorted by average rating all have 5 star ratings and at least 9 reviews. Fig 16's brands don't coincide with fig 14 or fig 15, this shows that large brands and brands that receive a lot of reviews are not necessarily the most loved.

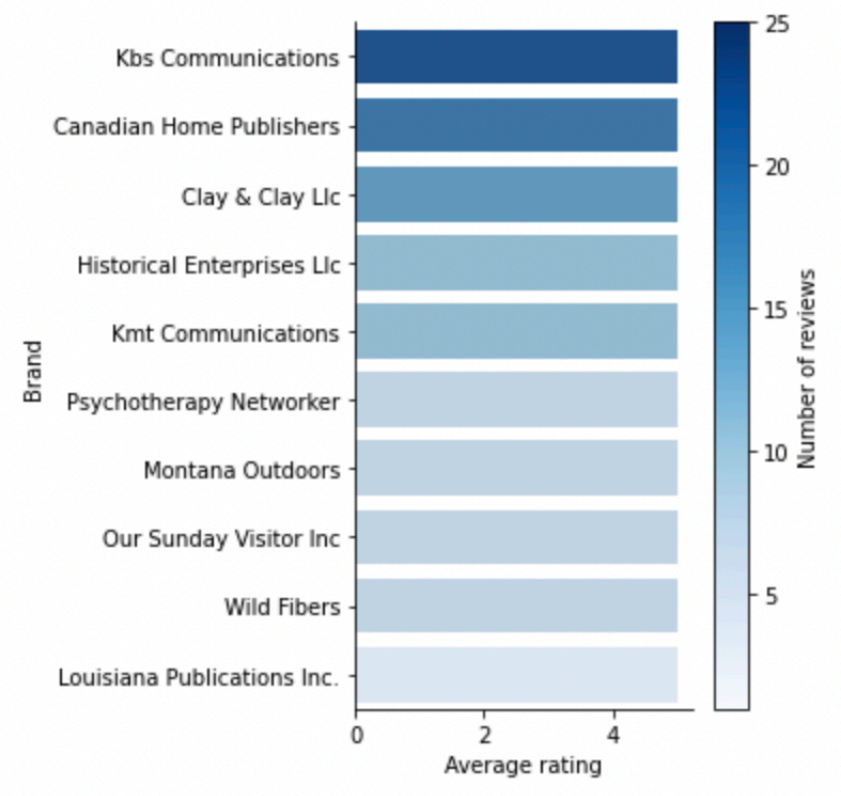



Fig 16. Distribution of average rating and number of reviews per brand (clipped at top 10 highest rated brands).

### 3.2.5.4 Bottom Brands by Average Rating

Fig 17 shows that the top magazine brands with the lowest average ratings all have ratings less than 2.7. We recognize 'Visionaire Publishing' which is the publisher of one of the top 5 magazines with the lowest average rating (V magazine) and 'Penske Business Media, LLC' which is the publisher of 'Variety', another magazine with one of the lowest average ratings. Further investigation shows that these two brands only have 1 magazine each available on Amazon, indicating that this low brand rating is completely attributed to the individual magazines.

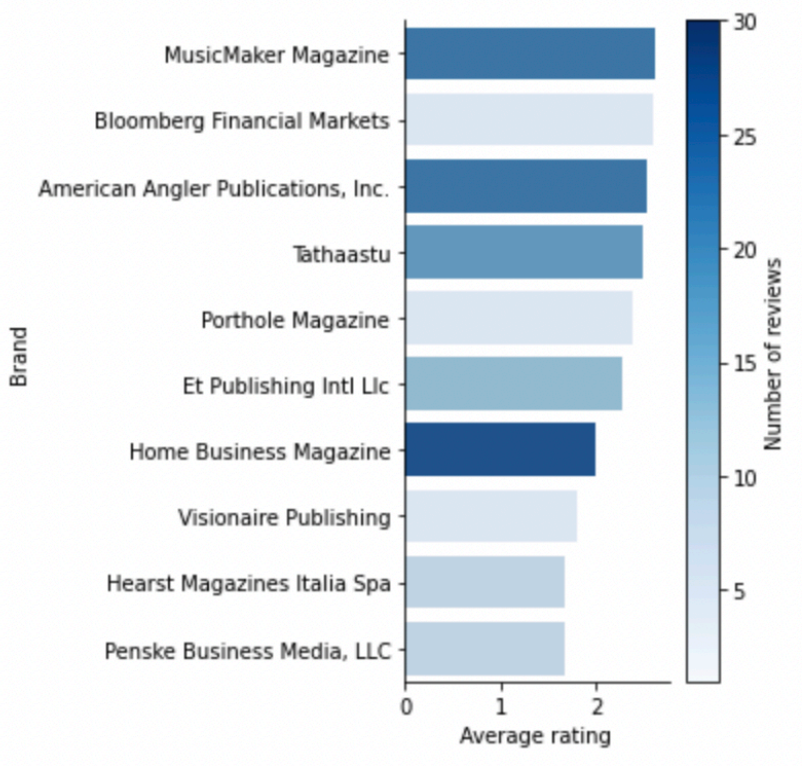



Fig 17. Distribution of average rating and number of reviews per brand (clipped at 10 lowest rated brands).

## 3.3 Case Study: Amazon Customer

In this report, we will look at an individual customer closely and generate recommendations for the customer from the recommendation engines built in order to assess the sensibility and relevancy of the recommendations. The chosen customer is reviewerID: 'A5RHZE7B8SV5Q.'

### 3.3.1 Preferences

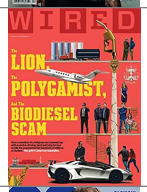
| | Magazine | Title | Brand | Categories | Average Rating |
|---|---|---|---|---|---|
| 1 |  | Maximum PC | Future US, Inc. | • Technology<br>• Computers & Internet | 1.0 |
| 2 |  | Vanity Fair | Conde Nast Publications | • Professional & Educational Journals<br>• Professional & Trade<br>• Entertainment & Media | 2.0 |
| 3 |  | PC World | | • Technology<br>• Computers & Internet | 4.0 |
| 4 |  | Details | Conde Nast Publications | • Fashion & Style Men | 5.0 |
| 5 |  | Wired | Conde Nast Publications | • Science<br>• History & Nature<br>• Technology | 5.0 |
| 6 |  | GQ | Conde Nast Publications | • Fashion & Style Men | 5.0 |
| 7 |  | Bon Appétit | Conde Nast Publications | • Cooking<br>• Food & Wine<br>• Recipes & Techniques | 5.0 |
| 8 |  | Xbox | Future US, Inc. | • Literary<br>• Sci-Fi & Mystery<br>• Science Fiction & Fantasy<br>• Fantasy | 5.0 |

Table 4. Magazines reviewed by reviewer A5RHZE7B8SV5Q.

Table 4 shows reviewer A5RHZE7B8SV5Q's preferences. This reviewer reviewed 8 magazines in total, gave 5 magazines 5-stars, 1 magazine 4-stars, 1 magazine 2-stars and another magazine 1-star. The reviewer appears to be male, judging from the men's fashion magazines he purchased. The general interests for this reviewer include: men's fashion, cooking, technology, sci-fi, history & nature, science and literary. Out of the 8 magazines he reviewed, 5 of them are published by the brand 'Conde Nast' and 2 are published by 'Future US, Inc'. We can deduct that this reviewer really enjoys magazines by Conde Nast as he gave 4 out of 5 of their magazines a 5-star rating.

# 4. Machine Learning

The goal of this project is to build a recommendation system that can generate the most relevant magazine subscription recommendations for Amazon's customer base. I will utilize customer ratings to build a collaborative filtering recommendation systems as well as utilize product metadata to build a content-based filtering recommendation system. We will evaluate the performances of each recommendation system in order to arrive at the best solution for recommending magazine subscriptions.

## 4.1 Model Building

### 4.1.1 Collaborative Filtering

Collaborative filtering is a technique that generates recommendations for a user based on similar users (those who like similar products). In order to determine if two users are similar or not, their preference for different products are recorded and compared.

There are two types of user preference data: explicit and implicit. Explicit data refers to data collected by explicitly asking the customer, e.g. asking a customer to rate the product from 1 - 5. Implicit data is data that is not provided intentionally but can be easily gathered from available data streams e.g. a list of products a customer has purchased. Implicit data is easier to collect but is often less accurate e.g. a person who purchased a product may not 100% like the product, however a person who rated a product 5-stars definitely likes the product.

For this project, since we have access to explicit ratings, I will use this to build an explicit collaborative filtering recommendation system. I will build the

collaborating filtering recommendation system in two different environments: Pyspark and Python. This is because both environments have different packages that I would like to use.

#### 4.1.1.1 Collaborative Filtering using Various Prediction Algorithms in Python

Using Python's surprise library, I modelled my data using various different prediction algorithms. I used 3-fold cross validation to select the best algorithm and then I used grid search with 3-fold cross validation to fine tune the selected algorithm. I adopted root mean square error (RMSE) as the evaluation metric. I chose RMSE over mean absolute error (MAE) because RMSE gives more weight to large errors and this is a desirable trait for our predicts as we will be generating predictions on a scale of 1-5.

From Table 5, we can see that the KNNBaseline has the lowest RMSE of 1.349. It is interesting to note that KNNBaseline also had one of the highest fit and test times. If our dataset increases in size, this may not be scalable. The best hyper parameters for this algorithm are: {reg_i: 10, reg_u: 5, n_epochs: 10}, producing an RMSE of 1.337. Reg_i refers to the regularization parameter for items and reg_u refers to the regularization parameter for users. N_epochs refers to the number of iterations of the ALS procedure.

| Algorithm | Test RMSE | Fit Time | Test Time |
|---|---|---|---|
| **KNNBaseline** | 1.349 | 82.513 | 1.395 |
| **SVD** | 1.351 | 2.669 | 0.124 |
| **BaselineOnly** | 1.363 | 0.127 | 0.112 |
| **SVDpp** | 1.369 | 3.758 | 0.139 |
| **CoClustering** | 1.379 | 2.907 | 0.079 |

| Algorithm | Test RMSE | Fit Time | Test Time |
|---|---|---|---|
| **KNNWithMeans** | 1.382 | 83.052 | 1.288 |
| **KNNWithZScore** | 1.383 | 84.478 | 1.304 |
| **SlopeOne** | 1.388 | 0.405 | 0.130 |
| **KNNBasic** | 1.398 | 83.255 | 1.243 |
| **NMF** | 1.404 | 4.925 | 0.096 |
| **NormalPredictor** | 1.803 | 0.048 | 0.115 |

Table 5. Prediction Algorithm Comparisons.

#### 4.1.1.2 Collaborative Filtering using Alternating Least Squares in PySpark

I built a recommendation system using Alternating Least Squares (ALS) in PySpark. I used 5-fold cross validation and RMSE as the evaluation metric to choose the best hyper parameters. The best performing model had an RMSE of 1.215 and the best hyper parameters are: {Rank: 1, MaxIter: 10, RegParam: 0.25}. Rank refers to the number of latent factors in the model, maxIter refers to the maximum number of iterations to run and regParam specifies the regularization parameter in ALS.

### 4.1.2 Content-Based Filtering using Cosine Similarity

I built a content-based recommendation system using cosine similarity in Python. I selected 'category', 'description' and 'brand' as features to describe each product. I combined these three features into a single feature. I then generated the cosine similarity matrix (2320 x 2320) between every pair of products in my dataset.

## 4.2 Model Comparison

Since collaborative filtering and content-based recommendation systems are fundamentally different, the only way I will be able to compare the two is if I performed AB testing to see which recommendations generate more sales.

Between the two different implementations of the collaborative filtering recommendation systems, the PySpark implementation has a slightly lower RMSE. For this reason, it is preferred over the Python model.

Both the content based recommendation system and collaborative filtering recommendation system (PySpark) will be used for further analysis in subsequent sections of the report.

# 5. Recommendation System Recommendations

We will make recommendations for user 'A5RHZE7B8SV5Q', which we have already analyzed their preferences in section 3.3.1.

## 5.1 Collaborative Filtering Recommendation System (PySpark)

### 5.1.1 Recommendations

Table 6 shows our collaborative filtering recommendation system's top 5 recommendations for our target customer.

| | Magazine | Title | Brand | Category | Summary |
|---|---|---|---|---|---|
| 1 |  | Gourmet News | Oser Communications Group Inc. | • Cooking<br>• Food & Wind<br>• Reference | • Gourmet industry<br>• New products<br>• Trade news |
| 2 |  | Diapason | Scranton Gilette Commun Inc | • Professional & Educational Journals<br>• Professional & Trade<br>• Entertainment & Media<br>• Music | • Organ, harpschord & church music |
| 3 |  | Architectural Digest | Conde Nast | • Arts<br>• Music & Photography | • Features work of top architects & interior decorators |
| 4 |  | Allure | Conde Nast | • Fashion & Style | • Beauty expert advice & tips |
| 5 |  | Professional Artist | Turnstile Publ | • Music & Photography<br>• Art & Art History | • Guide artists on their journey toward making a living with their art |

Table 6. Top 5 recommendations generated for reviewer 'A5RHZE7B8SV5Q' by the collaborative filtering recommendation system (PySpark).

### 5.1.2 Analysis

Based on our analysis in section 3.3.1, we know that our target customer is a huge fan of Conde Nast magazines as he has already subscribed to 5 of them and given 4/5 of them 5-star reviews. It is no surprise that our recommendation system has recommended 2 magazines from Conde Nast.

We also know that our customer is interested in cooking magazines as he rated 'Bon Appétit' 5-stars. We can see that our recommendation system has recommended another cooking magazine 'Gourmet News' to the customer. We also know that the customer is highly interested in men's fashion & style, although 'Allure' doesn't specialize in men's fashion & style, it is categorized as 'Fashion & Style' which encompasses men's fashion & style.

As for 'Diapason' and 'Professional Artist', (the other two magazines recommended to our customer by the recommendation system) our recommendation system is telling us that these two magazines are commonly enjoyed by users similar to our target customer. Recommendation systems can often uncover relevant recommendations that we may not understand. These two magazines could be highly popular for men in our target customer's demographic.

## 5.2 Content-Based Filtering Recommendation System

### 5.1.1 Recommendations

One of the top rated magazines by our target customer is 'Details', a men's fashion & style magazine by Conde Nast Publications. We will generate the top 5 most similar magazines to 'Details'. Table 7 shows our content-based filtering

recommendation system's top 5 recommendations for a customer who likes 'Details' magazine.

| | Magazine | Title | Brand | Category | Summary |
|---|---|---|---|---|---|
| 1 |  | Vanity Fair | Conde Nast | • Professional & Educational Journals<br>• Professional & Trade<br>• Entertainment & Media | • People, personalities & power |
| 2 |  | Cowboys & Indians | Cowboys & Indians | • Travel<br>• City & Regional<br>• United States<br>• South | • Luxury western lifestyle<br>• Western & American Indian art, interior, fashion and jewelry |
| 3 |  | Esquire | Reader's Digest Association | • Fashion & Style Men | • Men's lifestyle<br>• Style & clothes, cars, culture & entertainment and advice on money |
| 4 |  | Coastal Living | Meredith Corporation | • Home & Garden<br>• Design & Decoration | • Home & destination on coastline of US, Canada & Caribbean<br>• Recipes, coastal news, products & events |
| 5 |  | Glamour | Conde Nast | • Fashion & Style Women | • Hair & beauty<br>• Relationship & sex questions<br>• Monthly horoscopes<br>• Health & diet |

Table 7. Top 5 recommendations generated for a customer who enjoys Details magazine.

### 5.1.2 Analysis

The similarities between the magazines in table 7 and 'Details' are quite easy to identify because the similarity matrix was built on the magazine's brand, category and descriptions. 'Vanity Fair' and 'Glamour' are both Conde Nast magazines, which is the same brand is Details. Vanity Fair mainly discusses celebrities which is

also prominently discussed in Details magazine.  Glamour is a magazine filled with lifestyle content which is also very similar to Details.

Cowboys & Indians is a magazine celebrating the luxury western lifestyle which is a type of men's lifestyle.  Coastal living is also a lifestyle magazine that explores lifestyles by the coastline.  Finally Esquire is a men's fashion & style magazine exactly like Details.

# 6. Conclusion

The best recommendation systems built for this project are the content-based recommendation system and the collaborative filtering recommendation system (PySpark) which has an RMSE of 1.215. We chose the PySpark implementation over the Python implementation of collaborative filtering because it had a slightly lower RMSE. Besides its superior RMSE, the PySpark implementation is also much more scalable. As Amazon continues to expand, we expect more customers and more products to be added, thus it is vital to adopt a recommendation system that is highly scalable.

The recommendations generated by both algorithms in section 5 are sensible and even uncover some hidden magazine similarities that we would not have been able to pick up on without the algorithms. A combination of both algorithms will be able to put the most relevant products in front of customers and drive more sales. Further testing e.g. AB testing can be used to determine the best strategy for each of the algorithms, in terms of placement on the website.

# 7. Recommendations

Our content-based filtering and collaborative filtering recommendation systems are generating recommendations based on different factors. I believe that both recommendation systems should be used in conjunction on different parts of Amazon's website.

## 7.1 Content-Based Filtering Recommendation System

I would suggest displaying recommendations generated from the content-based recommendation system whenever a customer has clicked into a product's page.

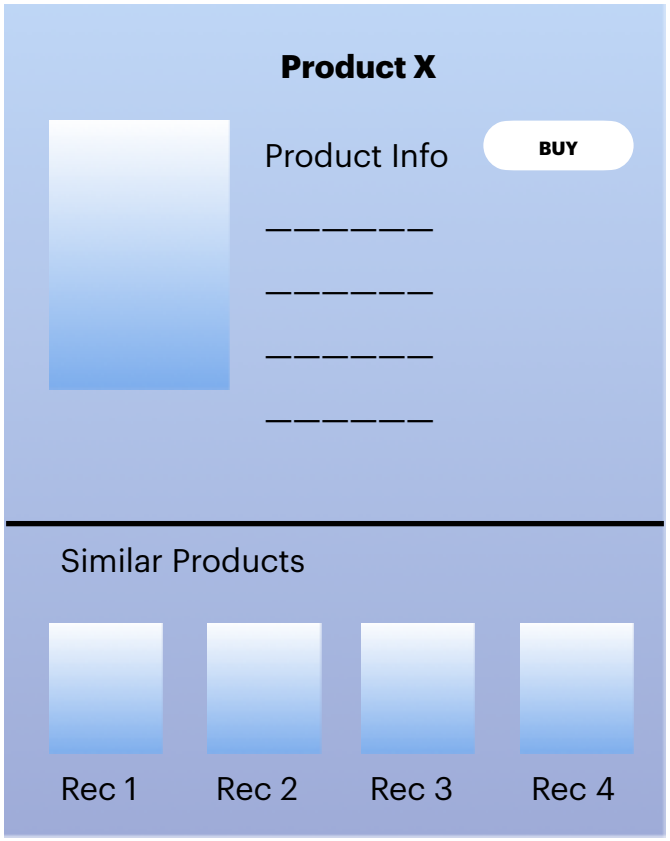



Fig 18. Amazon's product page mock up.

## 7.2 Collaborative Filtering Recommendation System

For the collaborative filtering model, I would suggest placing it on a customer's home page once they are signed into Amazon. Since we are not sure what the user is looking for, we can display the next best thing - what similar users have been purchasing.

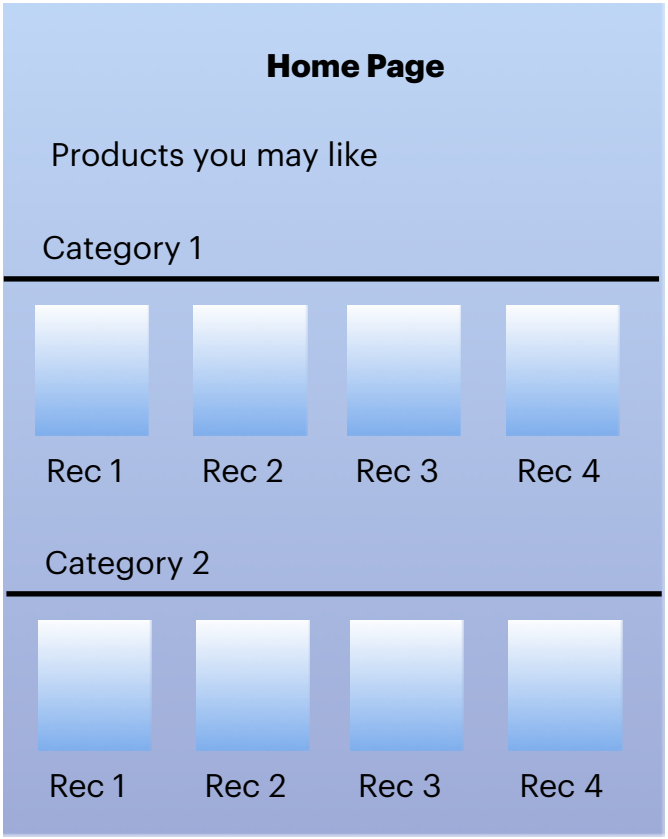



Fig 19. Amazon's user homepage mock up.

## 7.3 Cold Start Problem

The cold start problem is a phenomenon that affects new users. The recommendation systems we've built assumes that each user has rated at least 1 product. When new users sign up for Amazon, we don't have any ratings information so these recommendation systems cannot be utilized. In order to address this issue, we will begin by recommending the most popular items (identified in section 3.2.3). Once a user has reviewed 1 item, we can use our recommendation systems to generate more personalized recommendations.

The most popular magazines can also be displayed on Amazon's main homepage (before a user signs-in), these magazines are highly rated and we expect a large portion of the general public to be interested. This will attract new customers for Amazon.

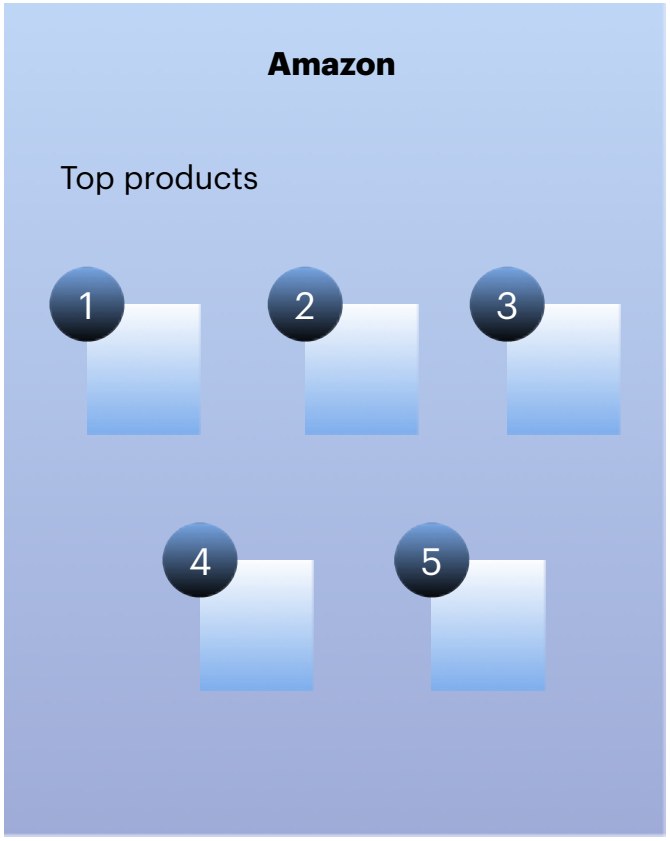



Fig 20. Amazon's public homepage mock up.

# 8. Future Scope

In the future, I would like to use image recognition to include features from magazine covers into my content-based filtering recommendation system. I would also like to use sentiment analysis to include review descriptions in conjunction with ratings to improve my collaborative filtering recommendation system. Furthermore, I would like to use AB testing to determine which recommendation system works best and the best locations to display these recommendations on Amazon's website.

I think adding additional datasets can also improve recommendations, datasets such as audience demographic would be extremely useful as we've previously discussed in the report. Given more computing power, I would like to train my models on the entire Amazon Reviews dataset which consists of 233.1 million reviews. This will allow recommendations to span across multiple different categories e.g. if a user bought a fashion magazine, they are likely to be interested in new fashion trends and with our model trained on the entire dataset, we can recommend trending fashion pieces for the customer.