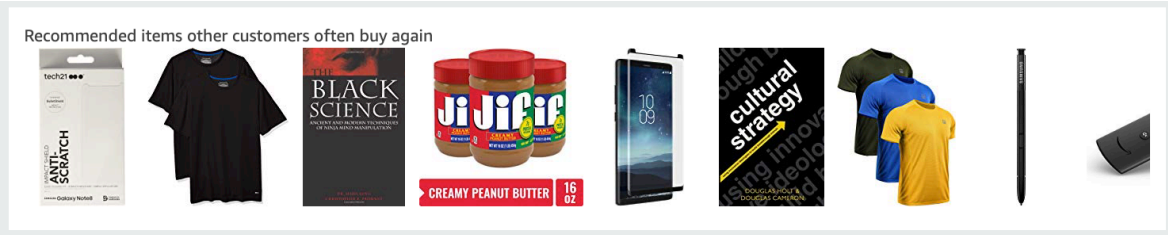
Amazon Hybrid Recommendation System

Dametreus Vincent – Capstone Project – Springboard Data Science – January 2020

Recommendation Systems

- **Recommendation Systems Are Used Everywhere Online**
  - Shopping – Best Buy, Walmart, Amazon, Giant
  - Entertainment – Netflix, Youtube, Pandora, Play/Apple Store
  - Referrals – Airbnb, Angie's List, Yelp, Travelocity
  - Relationships – Facebook, Twitter, Match.com
- **Amazon Uses Recommendation Systems to Sell Products**
  - Similar products, products you may like, products others like you like
  - Bundles, discounts, sales, last chance purchases
- **Our Goal**
  - Create a recommendation that change with the customer



Approach

- **Data Wrangling**
  - Amazon Customer Reviews Website > TSV Files > CSV Files > Pandas
  - Dataset > Product Parent > Product Category > Product ID > Product Title
- **Exploratory Data Analysis**
  - Customer Reviews & Rating Patterns
  - Purchasing Behavior
  - Changes in Behavior Over Time
- **Recommendation Systems**
  - Keyword Search > By category, name, or full title
  - Collaborative Filtering > Others like you, products similar to yours, because you rated this product.

Client

- **Amazon**
  - Increase market share, brand awareness, time-on-site, retention, sales, revenue and diversity of products.
- **Consumers**
  - People who like to shop online, stream or read content. People who may want different products tomorrow than they did today. Amazon Prime members.

Dataset

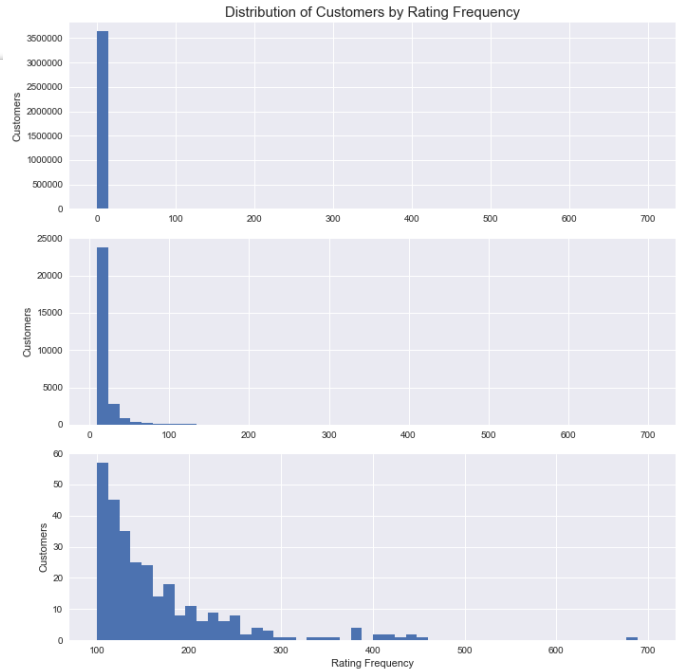
- **Amazon Customer Reviews Dataset**
  - https://s3.amazonaws.com/amazon-reviews-pds/readme.html
  - 20 years worth of amazon customer reviews on an international level.
- **Multilingual Reviews Dataset (US)**
  - https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt
  - US only customer reviews and ratings.

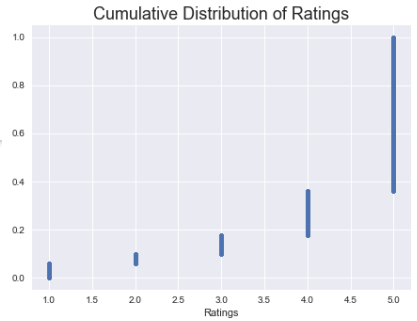
Data Wrangling

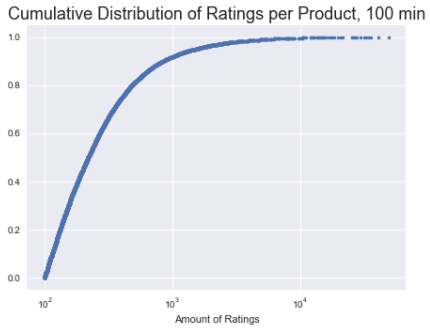
- **Clean Data**
  - Read in raw data
  - Standardize columns & product categories
  - Drop products with under 100 reviews (products under 100 were different variations of typos)
  - Discover difference between product parent, product title, product ID & variations in categories
  - Drop missing & duplicated data
  - Create new columns including purchased counts and date-time

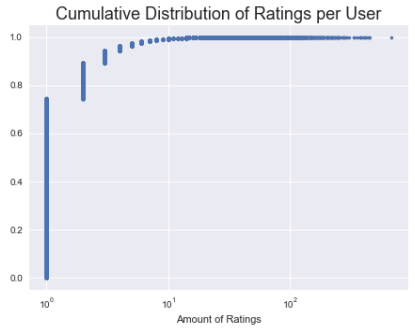
Exploratory Data Analysis

- **Distribution of Ratings Among Products & Customers**
  - 80% positive ratings, 15% negative ratings, 5% neutral
  - 50% products average over 4.5 stars, 90% average over 3.5 stars
  - 75% of customers have only one review, 24% have up to 10
  - 90% of products have up to 1000 reviews, 9% between 1000 & 10000



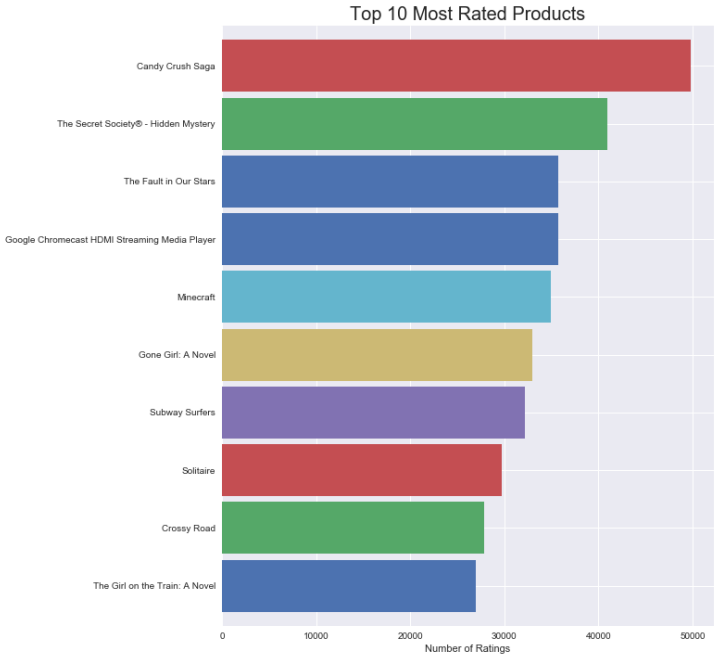


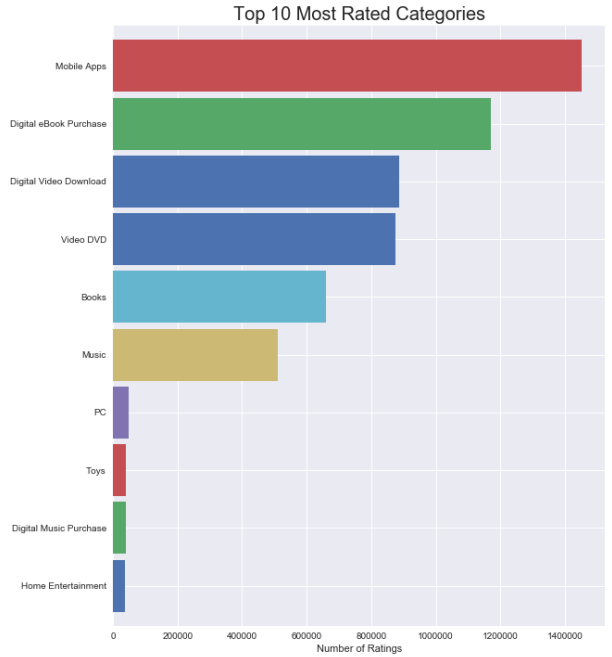




Exploratory Data Analysis

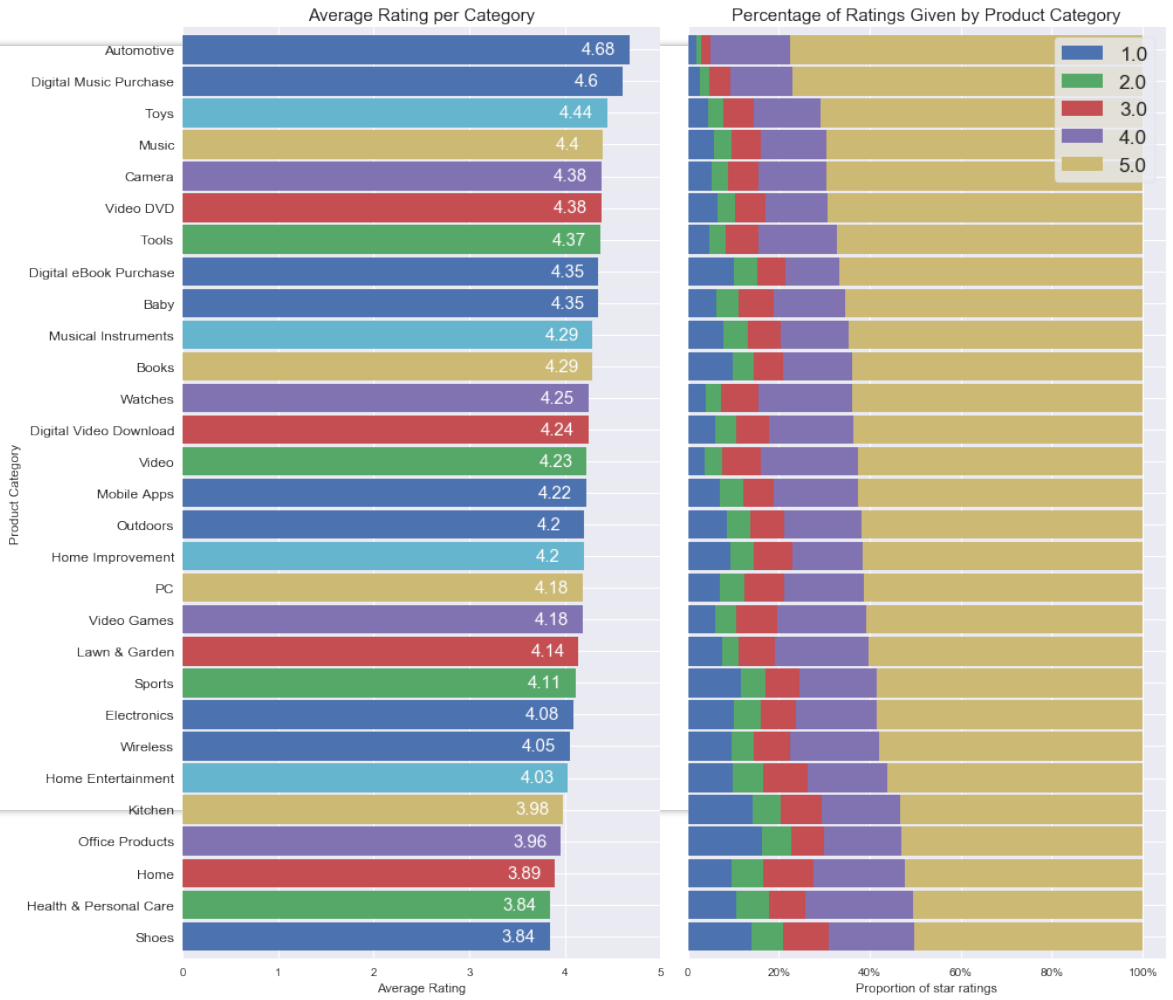
- **Most Popular Products & Categories**
  - Candy Crush Saga is the most popular product reviewed (mobile app)
  - Top 10 products reviewed have over 25000 reviews each, Candy Crush 40000+
  - 6 of top 10 categories dominate amazon with 90+% of total reviews
  - Top 3 categories are all digital (mobile apps, ebook, videos)





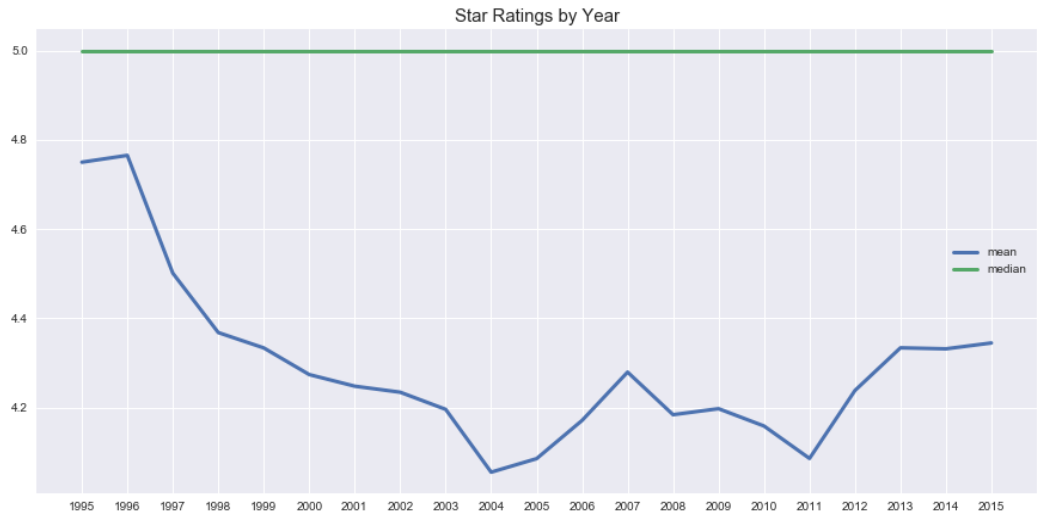
Exploratory Data Analysis

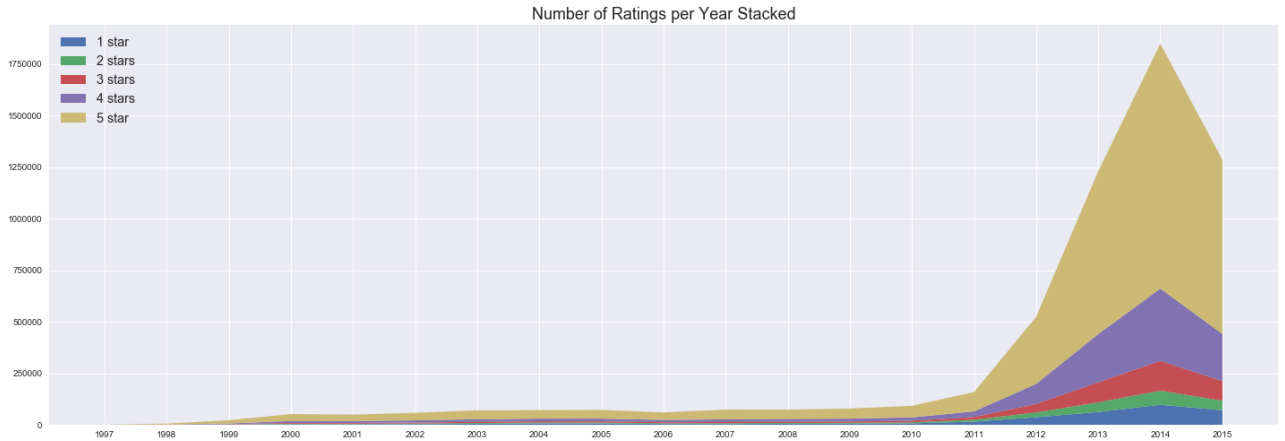
- **Ratings Among Product Categories**
  - Range in average ratings from highest to lowest is 0.84, from 3.84 to 4.68
  - Low rated categories are those where customers don't know what they're getting in quality
  - Highly rated categories have less variability in what will be received
  - Determining rating is highly influenced by 1, 4 & 5 stars



Exploratory Data Analysis

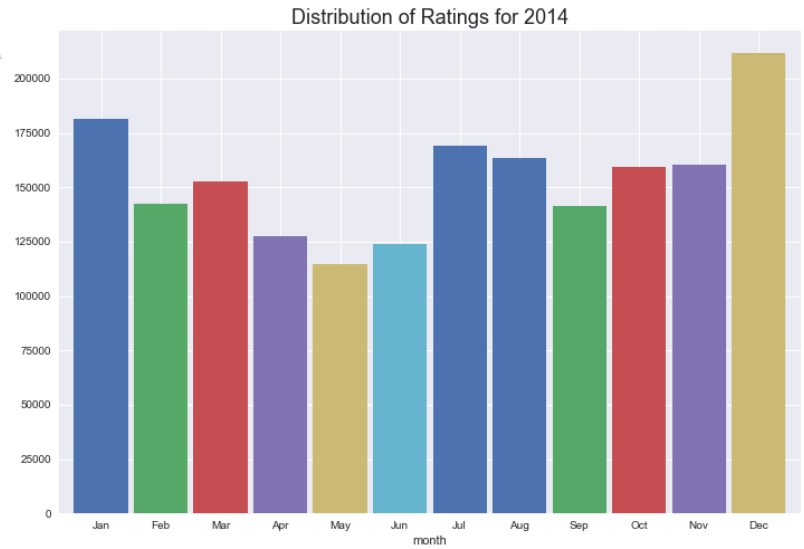
- **Ratings by Year**
  - Media rating is 5 stars (most products rated 5 stars)
  - Average annual product rating has stayed between 4.1 and 4.4
  - Initial average annual product rating was highest
  - Marketplace gained popularity in 2010
  - Exploded exponentially since then
  - Drop off from 2014 to 2015 is due to last review being in August

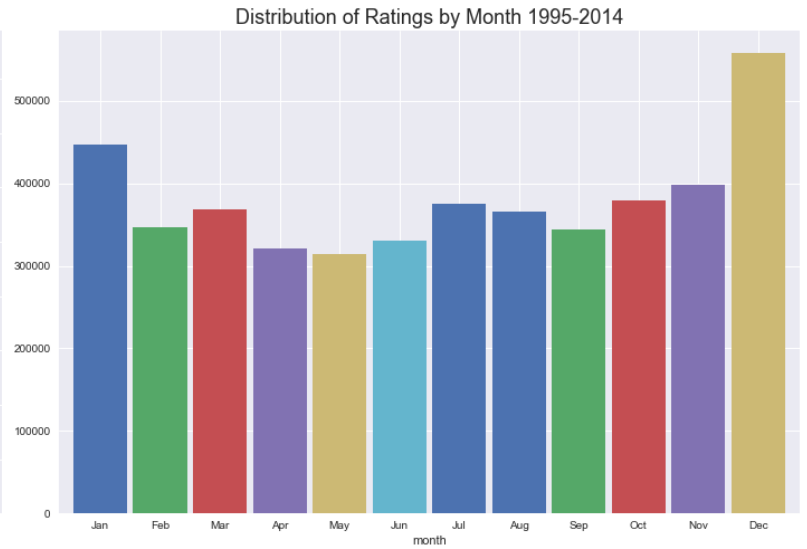




Exploratory Data Analysis

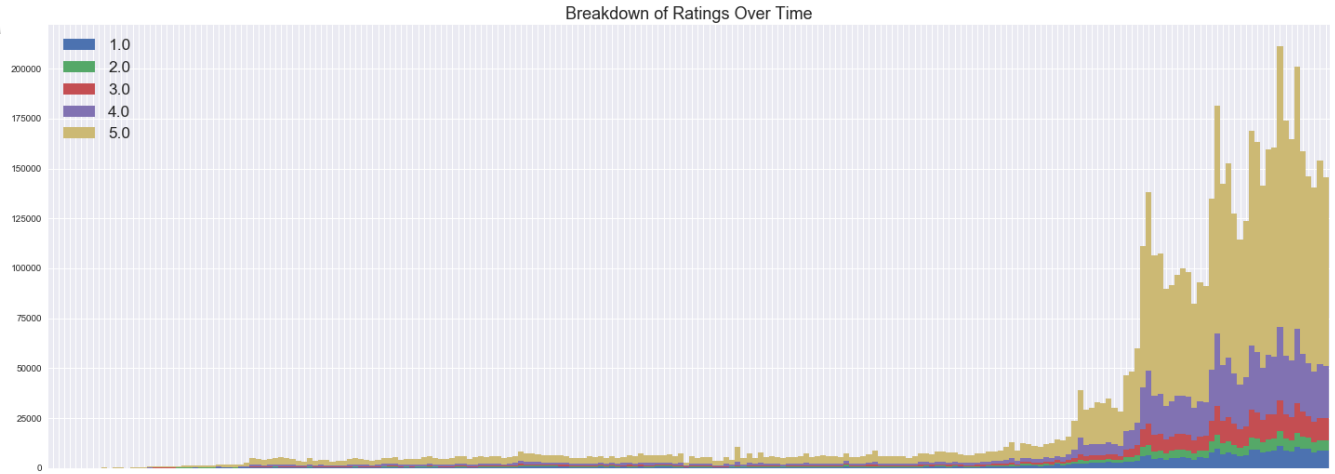
- **Ratings by Month**
  - December has the most reviews, followed by January
  - Likely due to holiday season and purchasing gifts
  - Least reviews are May & June
  - Likely due to starting summer vacations & travel

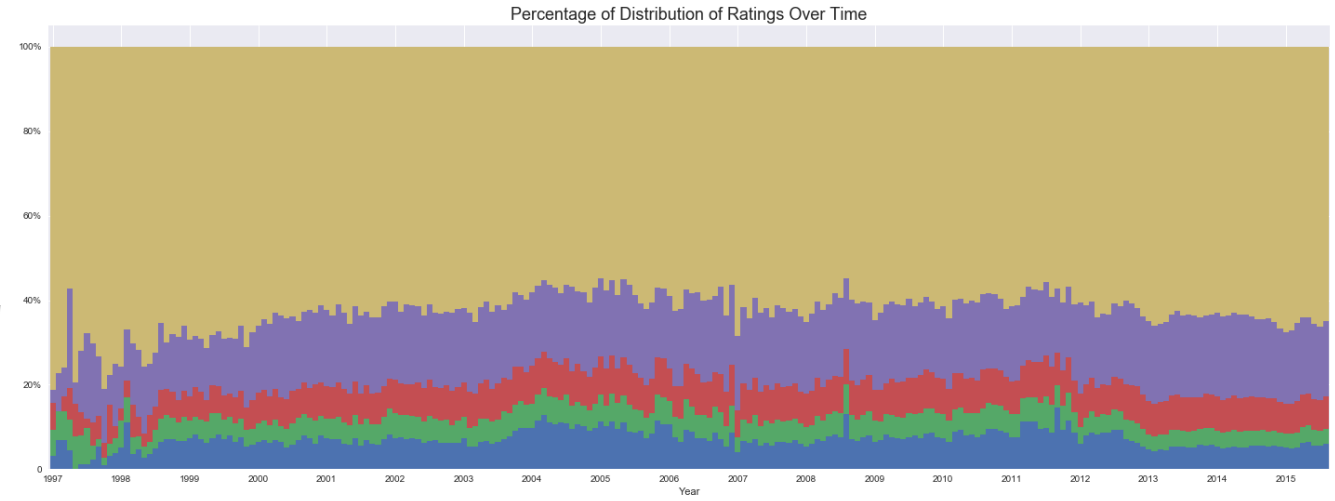




Exploratory Data Analysis

- **Ratings Over Time**
  - The breakdown of ratings over time changes slightly over time with no clear spike in changes to rating behavior
  - We can see the spike in reviews at the beginning of each year starting in 2011





Recommendation Filtering

- **Simple Keyword Search**
  - Adjusted star ratings
  - Returns 10 highest rated products matching words
  - Search by product category
  - Search by product title

```
# Test 6
print('\n-------------------\nTest 6: one word')
kw.keyword(product_title='fire')

# Test 7
print('\n-------------------\nTest 7: more detail in search')
kw.keyword(product_title='played with fire')
```

```
-------------------
Test 6: one word
Top 10 recommended products for you:
      product_id                                      product_title  \
3465   B0000AQS0F                      Firefly: The Complete Series
6807   B004XUMMFE                                   Firefly Season 1
6035   B003O86FMW       Catching Fire (Hunger Games Trilogy, Book 2)
10066  B00EFWU91E  City of Heavenly Fire (The Mortal Instruments ...
4021   B000FBFN1U  A Storm of Swords (A Song of Ice and Fire, Boo...
8602   B00957T6X6  George R. R. Martin's A Game of Thrones 5-Book...
5008   B001EN71CW             Firefly: The Complete Series [Blu-ray]
8098   B008673ZAQ               Reflected in You (Crossfire, Book 2)
4442   B000QCS8TW  A Game of Thrones (A Song of Ice and Fire, Boo...
4130   B000H4TX9I                                   Firefly Season 1

           product_category  star_rating  adjusted_rating  purchased_counts
3465                Video DVD         4.86             4.77           4960.00
6807   Digital Video Download         4.86             4.70           2428.00
6035   Digital eBook Purchase         4.69             4.66          12781.00
10066  Digital eBook Purchase         4.74             4.62           2558.00
4021   Digital eBook Purchase         4.72             4.60           2224.00
8602   Digital eBook Purchase         4.65             4.58           3900.00
5008                Video DVD         4.82             4.58           1091.00
8098   Digital eBook Purchase         4.62             4.57           4399.00
4442   Digital eBook Purchase         4.62             4.56           3977.00
4130   Digital Video Download         4.87             4.52            557.00

-------------------
Test 7: more detail in search
Top 8 recommended products for you:
      product_id                                      product_title  \
5180   B001NLKT60  The Girl Who Played with Fire (Millennium Seri...
7453   B0067AC5KA  Girl with the Dragon Tattoo Trilogy Bundle: Th...
338    0307595579  Stieg Larsson's Millennium Trilogy Deluxe Boxe...
5570   B002RHGYOA  The Girl Who Played With Fire (Millennium Series)
364    0307949508  The Girl Who Played with Fire (Millennium Series)
337    0307594777  Stieg Larsson's Millennium Trilogy Bundle: The...
284    0307269981        The Girl Who Played with Fire (Millennium )
313    030745455X  The Girl Who Played with Fire (Millennium Series)

           product_category  star_rating  adjusted_rating  purchased_counts
5180   Digital eBook Purchase         4.56             4.47           1569.00
7453   Digital eBook Purchase         4.64             4.44            496.00
338                     Books         4.72             4.41            244.00
5570   Digital eBook Purchase         4.66             4.38            152.00
364                     Books         4.44             4.36            328.00
337                     Books         4.51             4.36            128.00
284                     Books         4.32             4.33            455.00
313                     Books         4.30             4.32            763.00
```

Recommendation Filtering

- **Simple Collaborative Filtering**
  - Takes in dataset, user & product
  - Filters to all other users that purchased same product
  - Filters to all users that gave same rating as original user
  - Filters to all products other users purchased
  - Filters to 5 star products
  - Filters top 10 most purchased products
  - Returns the list
  - "Others that rated product A the same as you like this:"

	customer_id	product_id	product_title	product_category	star_rating	review_headline	purchased_counts
22981	502696	B0001NBMBC	Vol. 3: The Subliminal Verses	Music	5.0	Awsome	1000

```
# Test specific customer and product among different thresholds
collab(apr, 502696, 'B0001NBMBC')
collab(multiple, 502696, 'B0001NBMBC')
collab(several, 502696, 'B0001NBMBC')
collab(many, 502696, 'B0001NBMBC')
```

```
----------------------------------------
497 other customers purchased this product.
Similar customers purchased 2266 other products.

We recommend these products from those similar customers:
0   Aenima
1   Ashes Of The Wake
2   Slipknot - Disasterpieces
3   Toxicity
4   All Hope Is Gone
5   Vol. 3: The Subliminal Verses
6   Iowa
7   Slipknot (EX)
8   Master of Puppets
9   Mezmerize
```

	customer_id	product_id	product_title	product_category	star_rating	review_headline	purchased_counts
29	10285	B00AREIAI8	My Horse	Mobile Apps	5.0	Great game!	26826

```
%%time

# Threshold: 1 or more purchases per customer
collab(apr, 10285, 'B00AREIAI8')
```

```
----------------------------------------
19124 other customers purchased this product.
Similar customers purchased 29287 other products.

We recommend these products from those similar customers:
0   Minecraft
1   Flappy Wings (not Flappy Bird)
2   Candy Crush Saga
3   Crossy Road
4   My Horse
5   Subway Surfers
6   Minion Rush: Despicable Me Official Game
7   Guess The Emoji
8   Temple Run 2
9   Temple Run
```

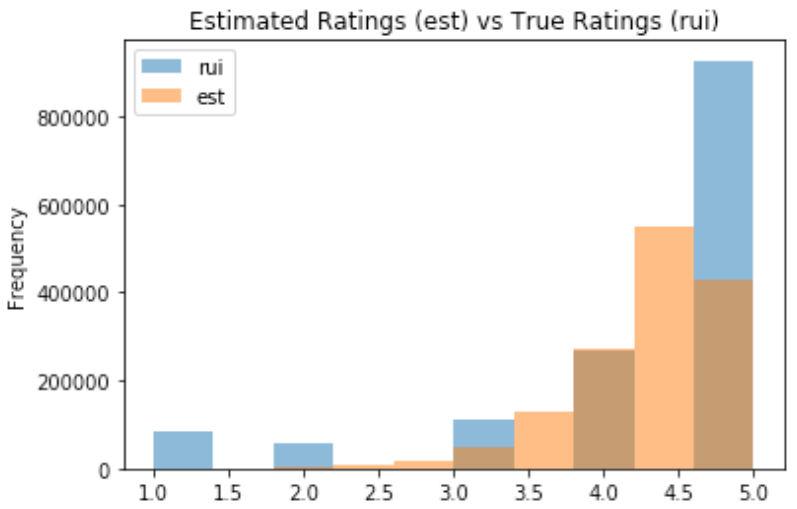
Recommendation Filtering

- **Matrix Factorization**
  - Customers > Rows, Products > Columns, Ratings > Values
  - Sparse matrix (99.98% empty)
- **Surprise Scikit**
  - Recommendation Module
  - Predicts Star Ratings & Recommends Products
  - Benchmark Various Algorithms
  - SVD as made famous by Netflix contest
- **Evaluate**
  - RMSE
  - Precision@k
  - Recall&k

Algorithm	test_rmse	fit_time	test_time
SVD	1.113691	1.899741	0.128023
KNNBaseline	1.114587	57.526784	0.207765
BaselineOnly	1.115938	0.176921	0.093586
SVDpp	1.116863	3.660672	0.100597
KNNBasic	1.155447	56.116376	0.166689
SlopeOne	1.160438	1.202418	0.129246
KNNWithMeans	1.160772	60.187209	0.326530
CoClustering	1.161342	4.473748	0.090783
NMF	1.165692	4.658261	0.096936
NormalPredictor	1.469713	0.048221	0.153851

$$\text{Precision@k} = \frac{|\{\text{Recommended items that are relevant}\}|}{|\{\text{Recommended items}\}|} \quad \text{Recall@k} = \frac{|\{\text{Recommended items that are relevant}\}|}{|\{\text{Relevant items}\}|}$$

product_id	B005ZOBNOI	B0063IH60K	B006LSZECO	B0094BB4TW	B00992CF6W	B00BAXFECK	B00DJFIMW6	B00E8KLWB4	B00FAPF5U0	B00L9B7IKE
customer_id										
5291529	NaN	NaN	NaN	4.0	NaN	NaN	NaN	4.0	NaN	NaN
11877567	NaN	3.0	NaN	NaN	NaN	NaN	NaN	NaN	3.0	NaN
14535682	NaN	5.0	NaN	5.0	5.0	NaN	4.0	2.0	5.0	NaN
27626904	NaN	5.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
31612235	NaN	NaN	NaN	2.0	NaN	NaN	5.0	NaN	5.0	NaN
40079136	NaN	NaN	NaN	NaN	4.0	NaN	NaN	NaN	NaN	NaN
42418272	NaN	NaN	NaN	NaN	NaN	NaN	4.0	NaN	4.0	NaN
44834233	NaN	1.0	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN
46671309	NaN	5.0	NaN	NaN	NaN	NaN	3.0	NaN	2.0	NaN
46823467	NaN	NaN	NaN	3.0	NaN	NaN	2.0	NaN	5.0	NaN
47959143	5.0	NaN	NaN	NaN	NaN	5.0	NaN	NaN	NaN	NaN
48233483	NaN	NaN	4.0	NaN	NaN	4.0	NaN	NaN	NaN	4.0
48417139	NaN	NaN	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50605810	NaN	NaN	NaN	4.0	NaN	NaN	NaN	NaN	NaN	5.0
52139318	NaN	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN	NaN
53092367	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.0



Recommendation Filtering

- **Recommending Products to a Customer**
  - Returns top-n products by estimated highest rating
  - "Products you might like:"

	customer_id	product_id	product_title	product_category	star_rating	review_headline	purchased_counts
4518678	44894569	B00006RU5B	Audioslave	Music	4.0	Great CD - if one accepts that it is not RATM ...	969
4518679	44894569	B003OF3R0S	Nightmare	Music	5.0	A culmination of their previous four albums	236

```
products_recommended(44894569)
```

```
The top 10 product recommendations for user 44894569 is:

0                                Me Before You: A Novel
1                     Alarm Clock Xtreme Free + Timer
2                              Downton Abbey Season 3
3                                         Crossy Road
4                                              Frozen
5                                        Suits Season 1
6         Catching Fire (Hunger Games Trilogy, Book 2)
7                              Three Days of the Condor
8    The Life-Changing Magic of Tidying Up: The Jap...
9      The Hunger Games (Hunger Games Trilogy, Book 1)
Name: 44894569, dtype: object
```

Recommendation Filtering

- **Recommending Similar Products**
  - Maps k nearest neighbors of product
  - "Products similar to product A:"

```
The 10 most similar products to Minecraft are:

0: Mobile Apps -  Bloons TD 5
1: Mobile Apps -  Angry Birds Epic RPG
2: Mobile Apps -  Twitter
3: Mobile Apps -  Farming Simulator 14
4: Mobile Apps -  Goat Simulator
5: Mobile Apps -  Head Soccer
6: Mobile Apps -  Asphalt 8: Airborne
7: Mobile Apps -  Hungry Shark Evolution
8: Mobile Apps -  The Dark Knight Rises (Kindle Tablet Edition)
9: Mobile Apps -  Dragon City
```

```
The 10 most similar products to Google Chromecast HDMI Streaming Media Player are:

0: Mobile Apps -  Twitter
1: Mobile Apps -  TubeMate YouTube Downloader
2: Video DVD -  The Dark Knight Trilogy (Batman Begins / The Dark Knight / The Dark Knight Rises) [Blu-ray]
3: Video DVD -  The Hunger Games: Catching Fire [Blu-ray + DVD + Digital HD]
4: Electronics -  FiiO E6 Portable Audio Headphone Amplifier
5: Mobile Apps -  Adobe Acrobat Reader- PDF Reader and more
6: PC -  SanDisk Ultra 16GB UHS-I/Class 10 Micro SDHC Memory Card With Adapter- SDSDQUAN-016G-G4A [Old Version]
7: Digital eBook Purchase -  11/22/63: A Novel
8: PC -  Thunderbolt to Gigabit Ethernet Adapter
9: Mobile Apps -  OfficeSuite Professional
```

Summary

- **In our analysis we were able to:**
  - Look at the distribution of ratings among products and customers.
  - Determine the Observe how ratings change over time.
  - Determine the point where the Amazon marketplace exploded in growth.
  - most popular products and categories.
  - Look at the highest rated categories and see what affects rating averages.
  - Observe how ratings change over time.
  - Determine the point where the Amazon marketplace exploded in growth.

- **We have found that:**
  - Over 60% of all reviews receive 5 star ratings.
  - Most products have between 100 and 1,000 reviews.
  - Most customers give between 1 and 10 reviews.
  - Out of 11,500+ products, less than 250 have an average rating under 3.0.
  - 90% of the Amazon marketplace revolves around books, music, movies and mobile apps.
  - The top 3 categories have intangible products (digital).
  - The range in the difference of average categorical ratings is 0.84, from 3.84 to 4.68.
  - Over time the average annual ratings stay between 4.1 and 4.4.
  - The marketplace has grown exponentially since 2011.
  - Average product ratings are mostly affected by 5, 4 and 1 star reviews.
  - Most reviews occur in January and December, likely because of the holidays.

Summary

- **We use 3 different methods for recommending products:**
  - The first method suggests products that other users like to a specific user based on their rating of a specific product.
  - The second method predicts what products a user may like based on past rating history and the rating history of other users.
  - The third method predicts products similar to a specific product by finding the nearest neighbors to that product.
- **We evaluate our methods by:**
  - Testing the recommendation systems to see what products they return.
  - Understanding estimated ratings vs. true ratings.
  - Performing precision@k and recall@k tests.

Limitations

- We feel that our limited time and computational power prevented us from making the best predictions in selecting and tuning our model as well as making product predictions for specific customers. In the future, we would like to work on training the complete dataset instead of a very small percentage (10% and less).
- As stated in the conclusion of the milestone report, we wanted to also build a content based and hybrid recommendation system as well as work on the cold start problem. Unfortunately, because of time constraints, we were unable to do that. Hopefully, in the future, we can add to that in this project.

Questions?

https://github.com/dametreusv/amazon_hybrid_recommendation_system