Big Data on Amazon Reviews

Group 6:

Yifan Gao
Howard Lin
Ruoyun Zhang

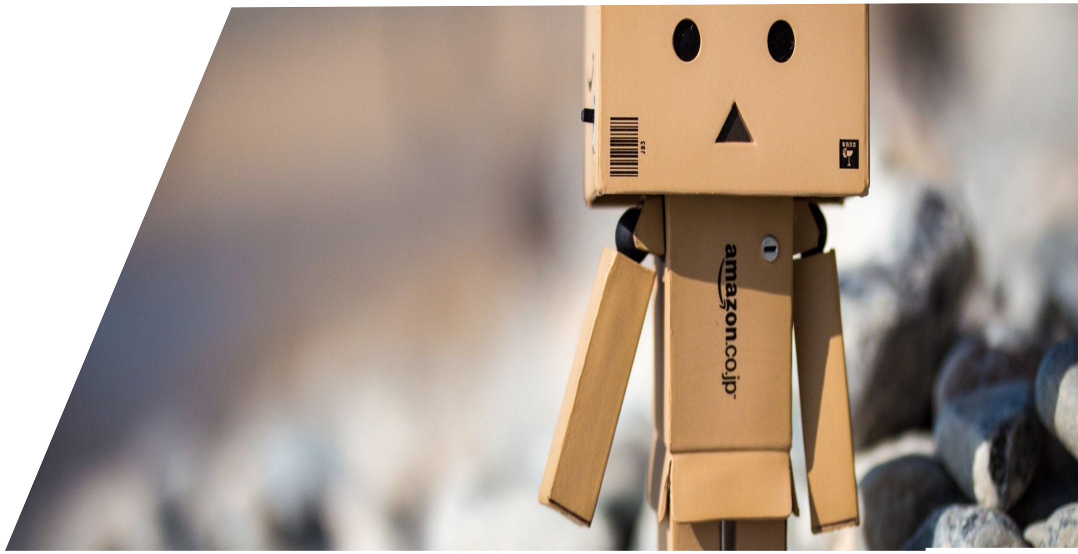
Agenda

01 Executive Summary & Business Objectives

02 Data Introduction and Explorative Data Analysis

03 Data Modeling

04 Q & A



Executive Summary & Business Objectives

Executive Summary

- Our project analyzes Amazon's data from 1995 to 2016 in order to glean insights from massive datasets. We employ natural language processing models, sentiment analysis, and ALS recommender system. The outcome can be used to measure customer interest/sentiment, optimize products, monitor reviews, and recommend products.

Business Objectives

- Extract, organize, analyze, and visualize data on big data platform, to review product popularity, customer satisfaction and review authenticity on Amazon.com from 1995 to 2016
- Pinpoint customer ID with suspicious review activities
- Perform time-based analysis to discover relationship among review volume, star rating and verified purchase ratio
- Build machine learning models for star rating prediction and sentiment analysis
- Build recommendation system to suggest products to customers

Data Introduction & Preprocessing

Data Overview

marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating
US	2975964	R1NBG94582SJE2	B00I01JQJM	860486164	GoPro Rechargeable Battery 2.0 (HERO3/HERO3+ o...	Camera	5
US	23526356	R273DCA6Y0H9V7	B00TCO0ZAA	292641483	Professional 58mm Center Pinch Lens Cap for CA...	Camera	5

- Source: https://www.kaggle.com/cynthiarempel/amazon-us-customer-reviews-dataset
- Description: customer review text written on Amazon.com
- Data Size: 54 GB (37 separate files)
- Format: tsv
- Shape: 110M rows, 15 columns
- Time Horizon: 1995 - 2015

Preprocessing

Raw Data	Clean Data
marketplace	customer_id
customer_id	product_id
review_id	product_parent
product_id	product_title
product_parent	product_category
product_title	star_rating
product_category	helpful_votes
star_rating	total_votes
helpful_votes	vine
total_votes	verified_purchase
vine	review_headline
verified_purchase	review_body
review_headline	review_date
review_body	**year**
review_date	**month**

Big Data Implementation

Google Cloud Platform: ~16G RAM * 8 nodes

- Configure cluster to ensure customized package such as SparkNLP is supported
- Mainly handles sentiment analysis which requires loading pre-trained models online
- Data & check points stored to Cloud Storage bucket
- Not as fast as RCC Midway2 due to quota constraint

RCC Midway2-compute node: 50G RAM * 8 nodes

- Spark and SparkNLP packages are sources from Prof. Igor Yakushin's folder (thanks for his help!)
- Mainly handles computational heavy jobs such as recommendation system and SparkNLP model
- Data & check points stored to scratch folder
- Faster than GCP but could suffer from inadequate compute nodes

RCC Midway3-compute

- Back up plan when midway2 is slow/down

Data Challenges

Challenges

- Data storage not enough at RCC individual folder
- **SparkNLP** not available on RCC midway2
- Multiple models take way **longer to execute**
- Have to rerun the pipeline when cluster/compute nodes shut down
- GCP: low compute capacity
- **Slow** shuffling data between worker nodes and serializing RDDs to disk
- **Buffer** limit exceeded
- Insufficient RAM in ALS factorization

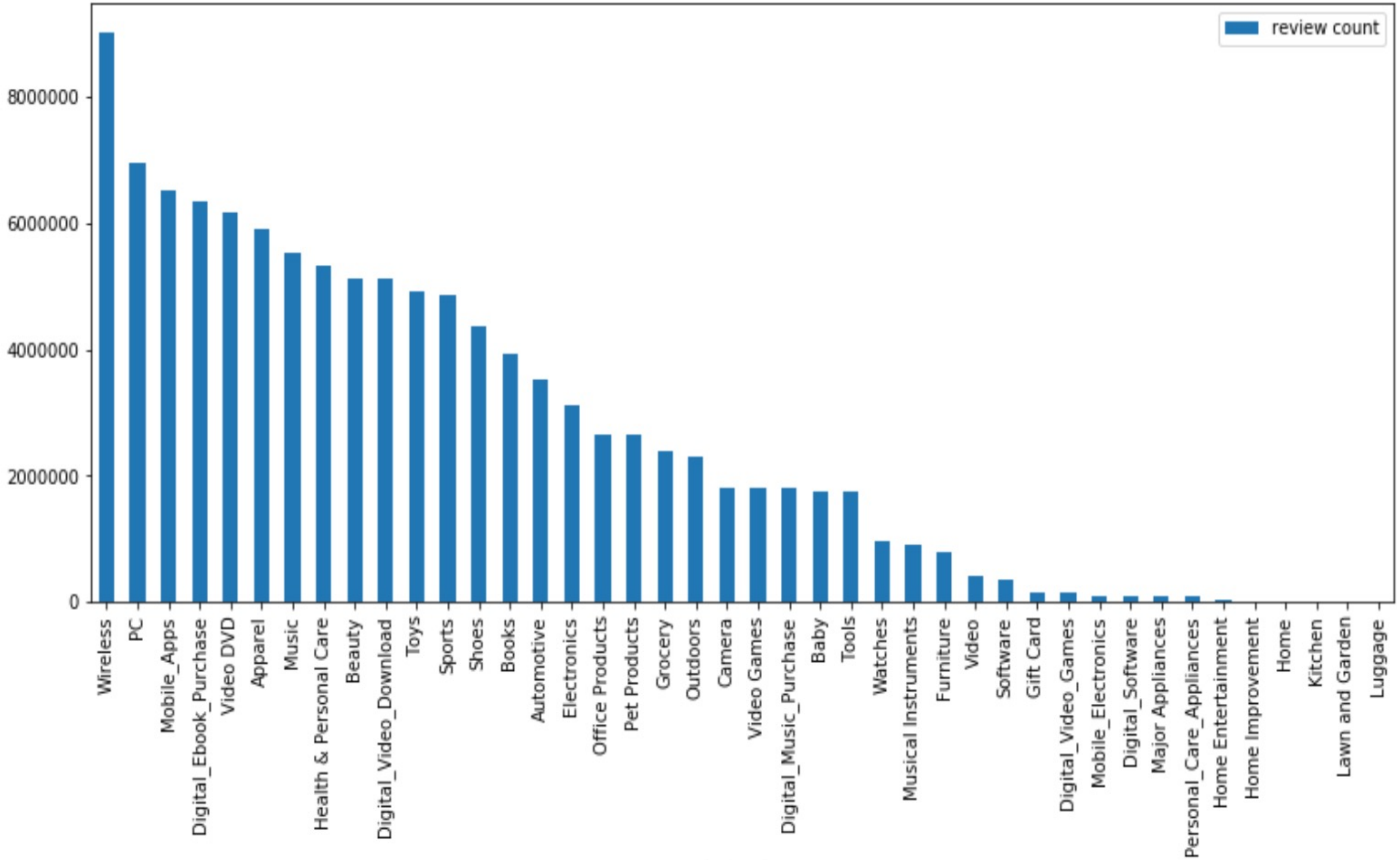
Solutions

- Store to scratch or tmp folder with more storage quota
- Consult/Work with Prof. Igor to resolve the package issue
- Try small sample first then scale up; split the work across multiple platforms (RCC and GCP)
- Checkpoint/save intermediate data and models
- Enable autoscaling
- Switch to KyroSerializer
- Increase buffer max in config
- Filter customers with more than 20 reviews to downsize data

EDA - Visualizing Category Distribution

Total Category Number: 42
Total Product Number: 15.2M

Product Category Ranked by Review Count



EDA - Average Rating by Product

Total Category Number: 42
Total Product Number: 15.2M

Top 20 Categories with Highest Rating

product_category	avg(star_rating)	rating_rank
Gift Card	4.731352294070298	1
Digital_Music_Pur...	4.638542111406816	2
Music	4.435098851371281	3
Video DVD	4.312607732385254	4
Grocery	4.312269146696672	5
Digital_Ebook_Pur...	4.262511660983939	6
Tools	4.262147816135473	7
Musical Instruments	4.251171329778752	8
Automotive	4.246277176356498	9
Shoes	4.241344241852507	10
Outdoors	4.239968733828497	11
Sports	4.22921513619547	12
Toys	4.2145692651318845	13
Digital_Video_Dow...	4.209598225894942	14
Books	4.20874830377662	15
Video	4.196926187784791	16
Beauty	4.187216275952948	17
Baby	4.1632115071652525	18
Health & Personal...	4.16175316573927	19
Pet Products	4.143630218299772	20

Top 20 Categories with Lowest Rating

product_category	avg(star_rating)	rating_rank
Digital_Software	3.5393869333934185	42
Software	3.5671616476491814	41
Major Appliances	3.716363223515812	40
Mobile_Electronics	3.7639697211761574	39
Digital_Video_Games	3.8531407942238265	38
Wireless	3.8921643092741736	37
Personal_Care_App...	3.9774617093281543	36
Kitchen	3.9934888768312535	35
Mobile_Apps	4.033717314526727	34
Electronics	4.035709742525166	33
Home Entertainment	4.036964021685559	32
Home	4.052316890881913	31
Video Games	4.060909568831088	30
Luggage	4.064102564102564	29
Office Products	4.07249061072986	28
Furniture	4.083964347539518	27
PC	4.087370531095652	26
Apparel	4.105200690420225	25
Lawn and Garden	4.128712871287129	24
Camera	4.128983751057	23

EDA - Customer Analysis

Total Customer Number: 27.5M

Average star rating and # of reviews by customer

customer_id	avg_star_rating	count	review_number_rank
50122160	4.99813456565057	23587	1
14539589	4.8867169462829	6497	7
20018062	4.809608540925267	6182	9
7080939	4.999822032390105	5619	12
22073263	4.7548108108108105	4625	17
53037408	4.911963390716932	4589	18
50199793	4.774026614095614	4058	22
50345651	4.979022704837117	4052	23
15725862	4.7483594864479315	3505	29
44731853	4.731501057082452	3311	34
49837360	4.72685609532539	3273	35
15536614	4.998916576381365	2769	46
53017806	4.883236994219653	2595	51
51591392	4.994428969359332	2513	57
12201275	4.932729007633588	2096	82
45070473	4.772481572481572	2035	87
34247947	4.8541153277476585	2029	88
39569598	4.8174924165824065	1978	97
47883385	4.997395833333333	1920	102
50776149	4.9912996193583465	1839	108

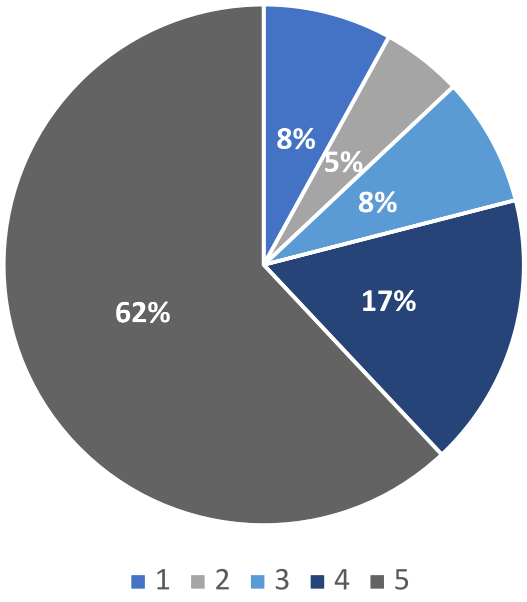
only showing top 20 rows

customer_id	avg_star_rating	count	review_number_rank
48608140	1.0	205	8142
16071656	1.2471264367816093	174	11839
44270361	1.049079754601227	163	13808
18853502	1.2352941176470589	153	15817
37141039	1.0	150	16672
47619896	1.2465753424657535	146	17472
30793307	1.036764705882353	136	20762
41542504	1.0	132	22034
42329785	1.0743801652892562	121	26745
40151153	1.1090909090909091	110	32949
24957250	1.2545454545454546	110	32650
20372208	1.2660550458715596	109	33583
39496978	1.2376237623762376	101	39274
13081743	1.0612244897959184	98	41512
17703766	1.0	89	50866
14241175	1.069767441860465	86	55498
34408569	1.0	82	60814
186275	1.0	80	64064
1960444	1.064102564102564	78	66527
36596648	1.1818181818181819	77	69012

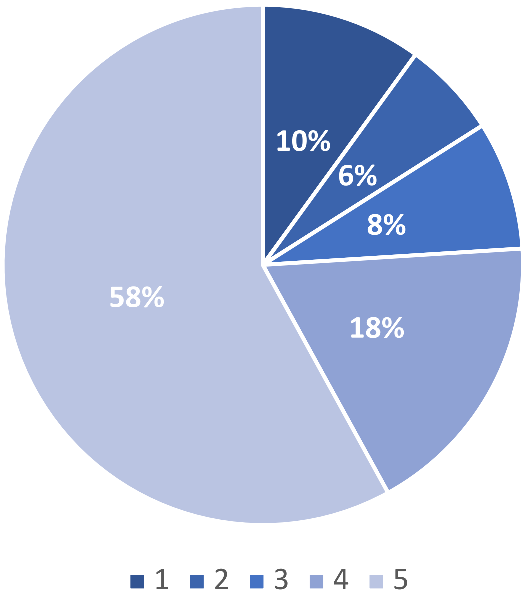
only showing top 20 rows

EDA - Verified Purchases

Star Rating of Verified Purchases

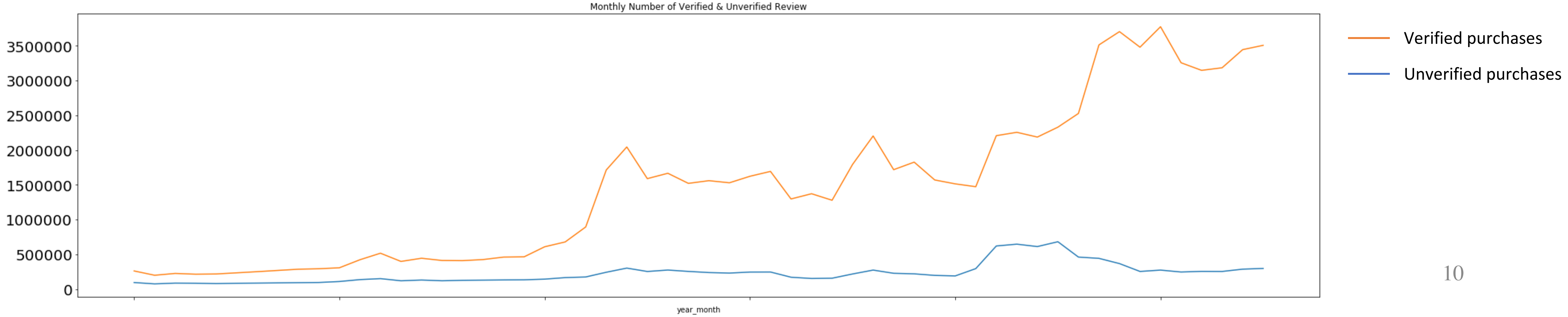


Star Rating of Unverified Purchases

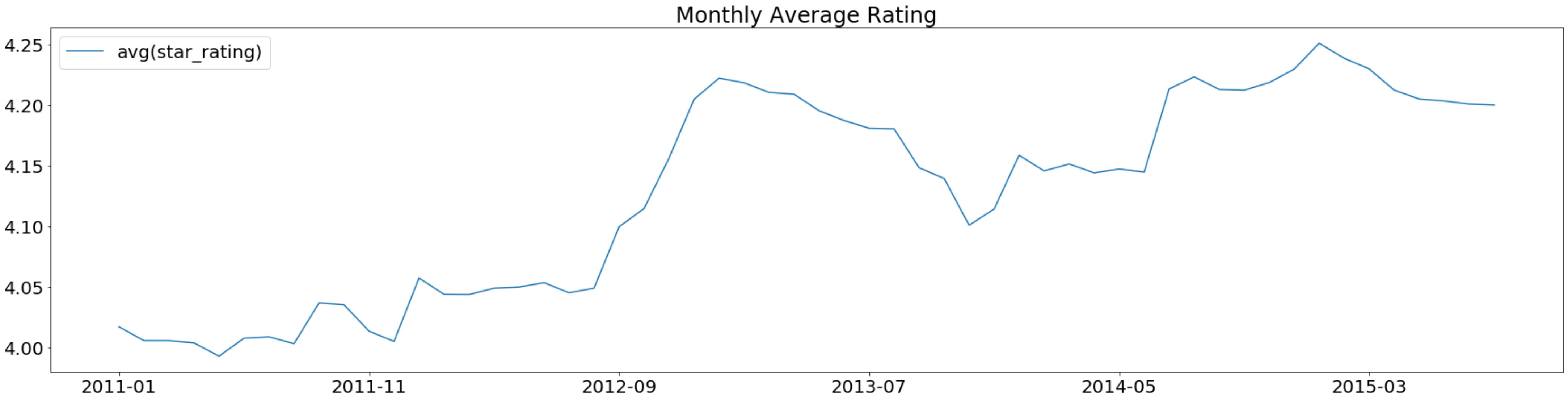


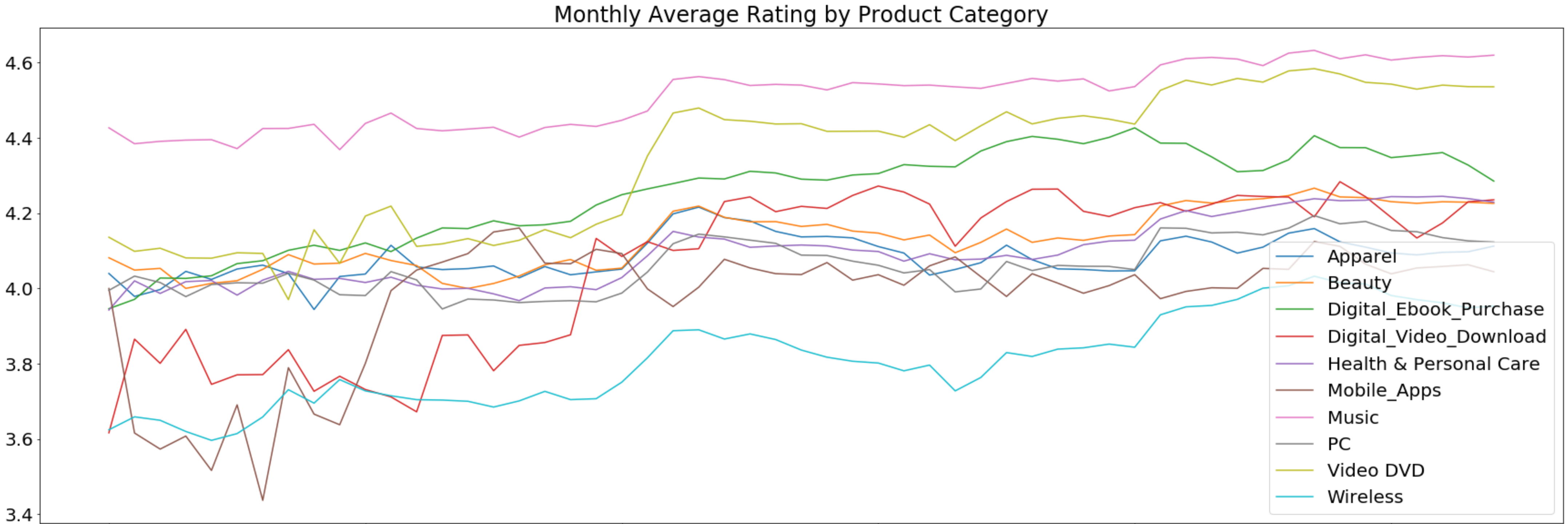
Correlation by each review:

	star_rating	helpful_votes	total_votes	verified_purchase
star_rating	1.000000	-0.020300	-0.045593	0.043456
helpful_votes	-0.020300	1.000000	0.987052	-0.055141
total_votes	-0.045593	0.987052	1.000000	-0.070857
verified_purchase	0.043456	-0.055141	-0.070857	1.000000

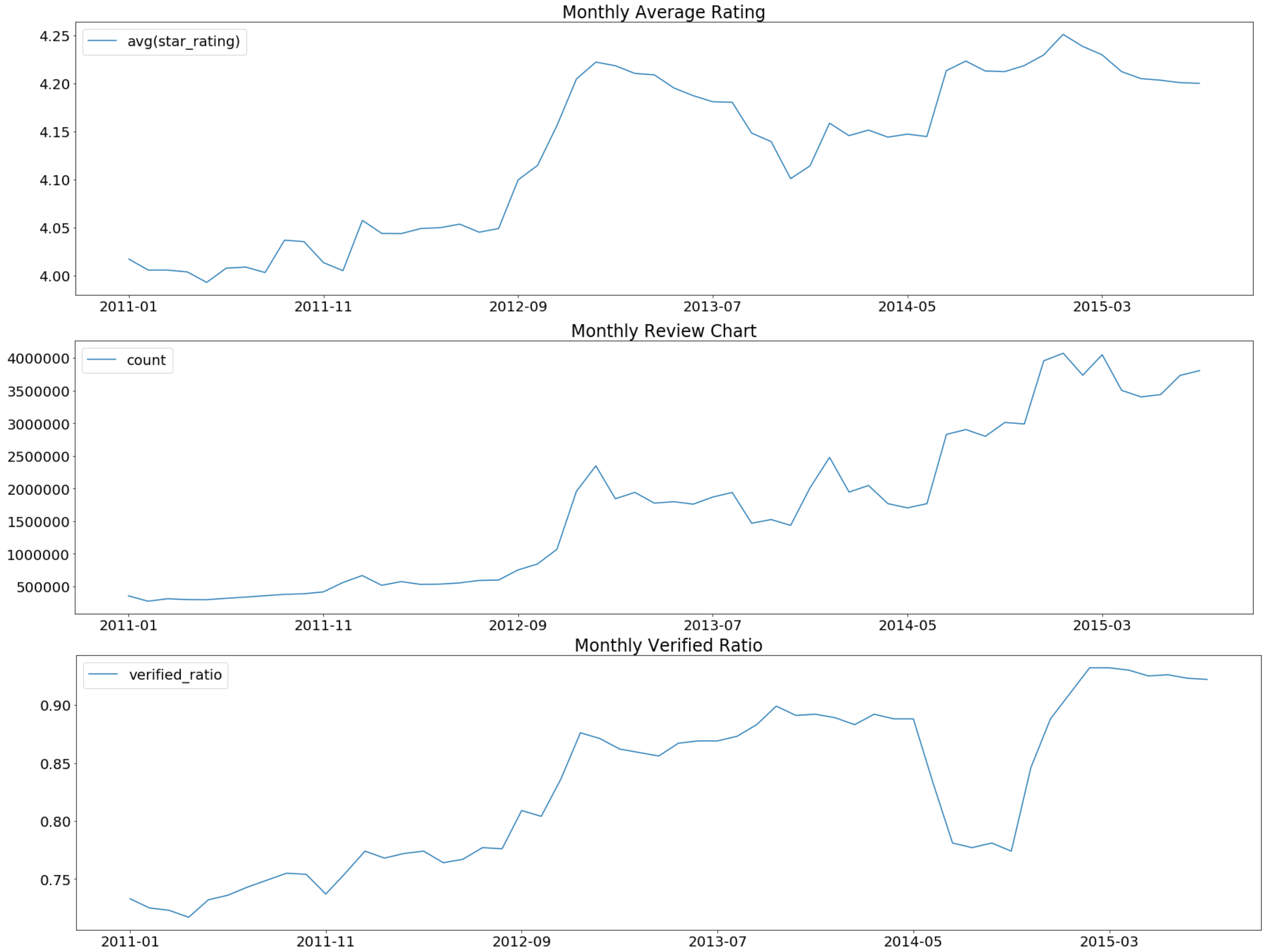


EDA - Time-Based Analysis

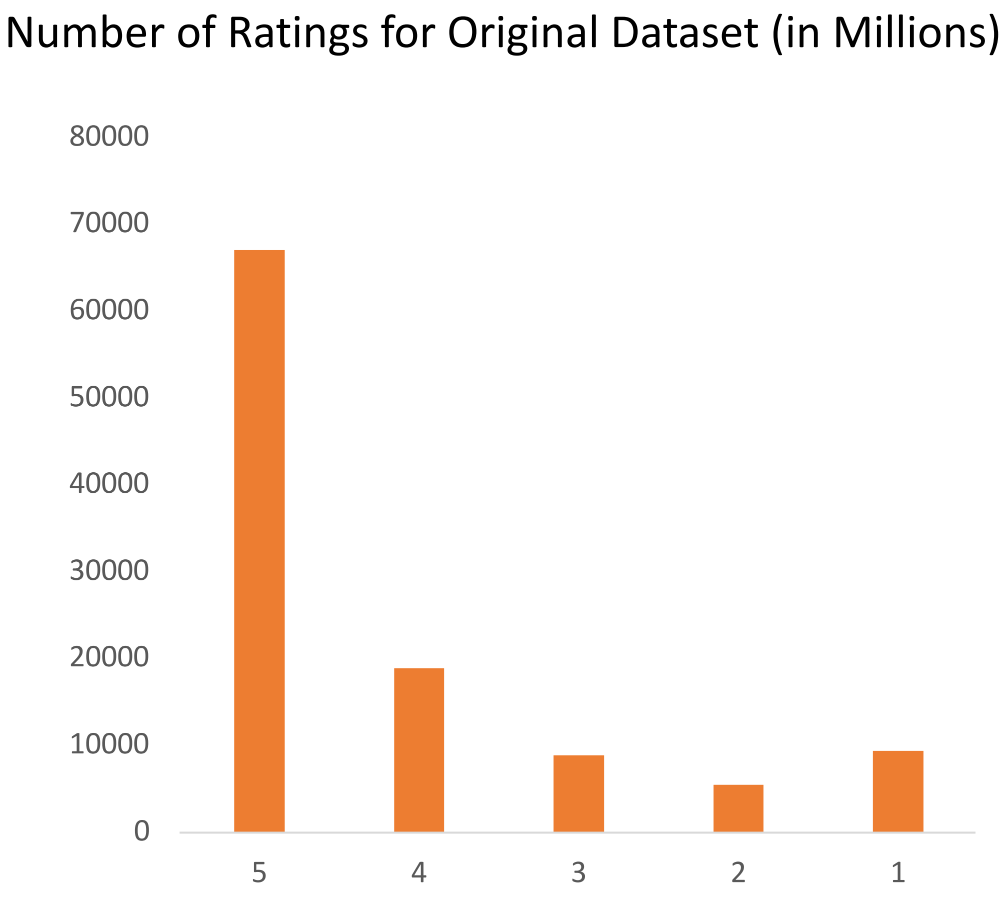


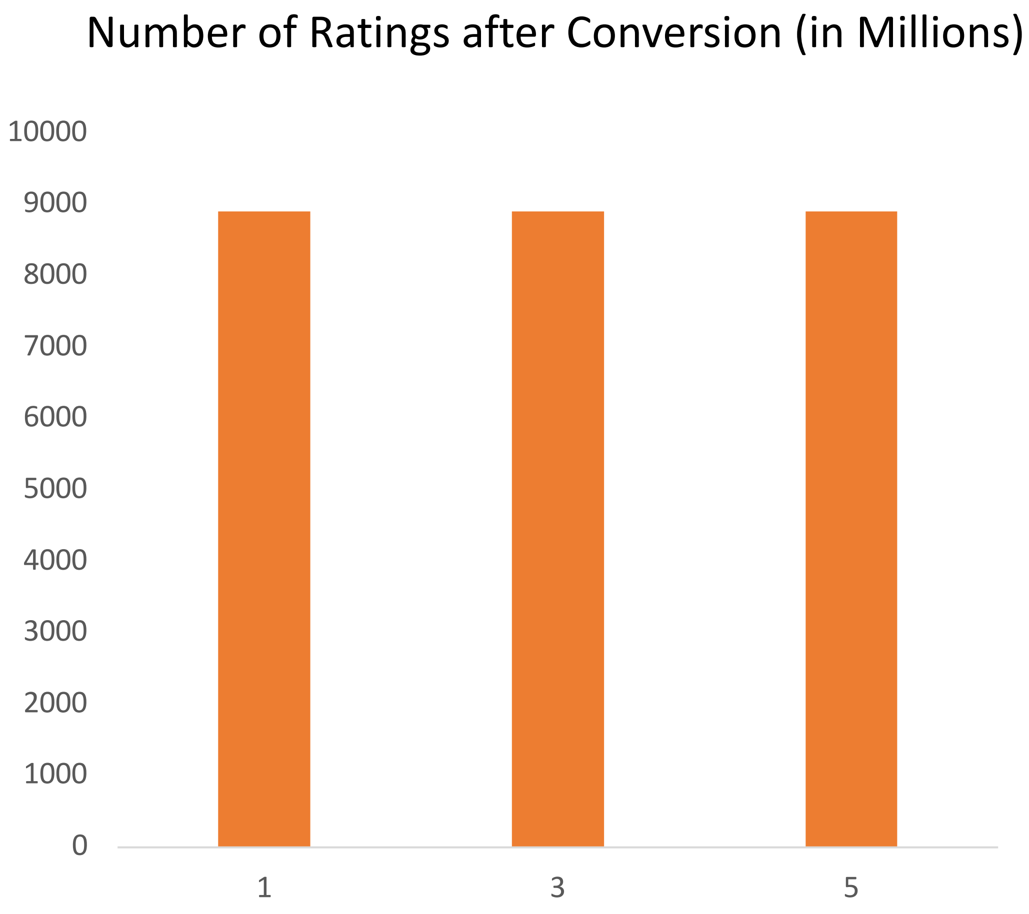


EDA - Time-Based Analysis



Handling with imbalance dataset





Modeling - Predict Star Rating with NLP

Deploy Models

Base model

```
pipeline1 =
Pipeline(stages=[tokenizer,remover,hashingTF,idf])
```

SparkNLP model

```
pipeline2 = Pipeline(stages = [document_assembler,
tokenizer, normalizer, stemmer, finisher, hashingTF)
```

- Input is review_body and output is star_rating so it's a supervised classification problem.
- For base model, we tokenize the string and remove words that do not provide much depth to the meaning. Then, we apply HashingTF to convert terms to fixed-length feature vectors and IDF to decrease the weights of frequently occurring words. Data will then be fitted by the learning model and the best model will be used for further evaluation.
- For SparkNLP model, we add additional stages such as normalizer, stemmer and finisher to improve model performance.

Modeling - Predict Star Rating with NLP

Deploy Models

SparkNLP sentiment model

```
pipeline3 = Pipeline(stages = [documentAssembler, use, sentimentdl])
```

- Input is review_body and no output is used to fit the model so it's an unsupervised learning.
- We will solely rely on review_body variable and pretrained sentiment model to analyze the sentiment of each user, which is categorized as negative, neutral and positive.
- We will do a trick to convert the sentiment to the three class star_ratings: negative to 1, neutral to 3 and positive to 5. This way we can calculate the metrics and compare with the previous models.
- For SparkNLP sentiment model, we load two pretrained models.
    - The Universal Sentence Encoder encodes text into high-dimensional vectors that can be used for text classification, semantic similarity, clustering.
    - The sentiment model "sentimentdl_use_imdb", an english sentiment analysis trained on the IMDB dataset.

Modeling - Predict Star Rating with NLP

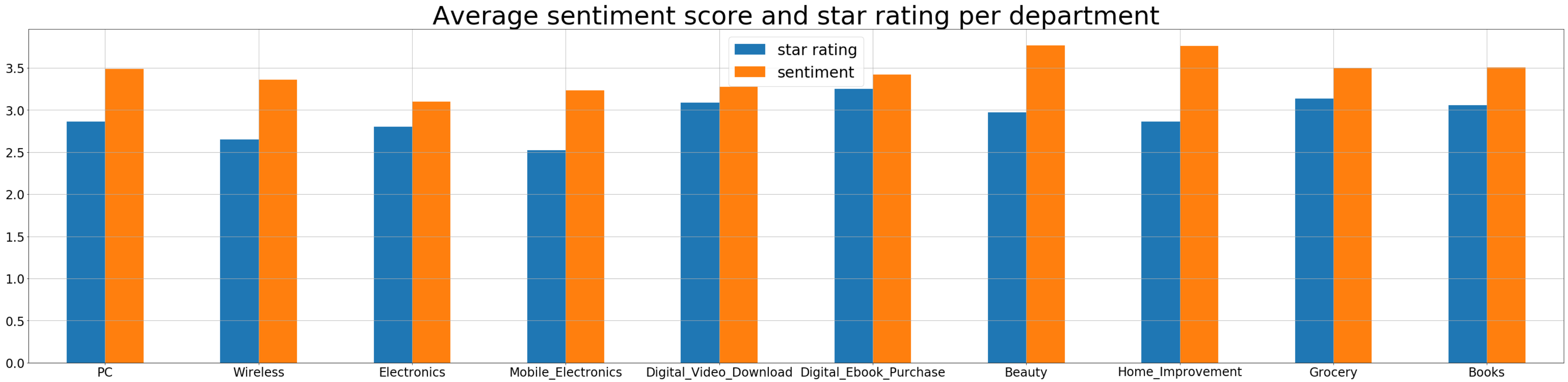
Model Performance Evaluation

Model	F1	Accuracy	Platform	Time taken
Base model	0.46	0.47	GCP	~3 hours
SparkNLP model	0.51	0.52	RCC Midway2	~10 hours
SparkNLP sentiment model	0.42	0.52	GCP	~2 hours

- In terms of metrics (F1 and accuracy), our best model is SparkNLP model, regardless of computer resources and time cost.
- If we need to factor in time and computer costs, then base model is a good alternative, though it's metrics are around 0.05 lower than SparkNLP ones.
- We also tried running the three models on imbalanced dataset. We got much higher scores (~0.8 ~0.7). However, we still prefer balanced dataset as it won't skew towards a specific class so the prediction is more reliable.

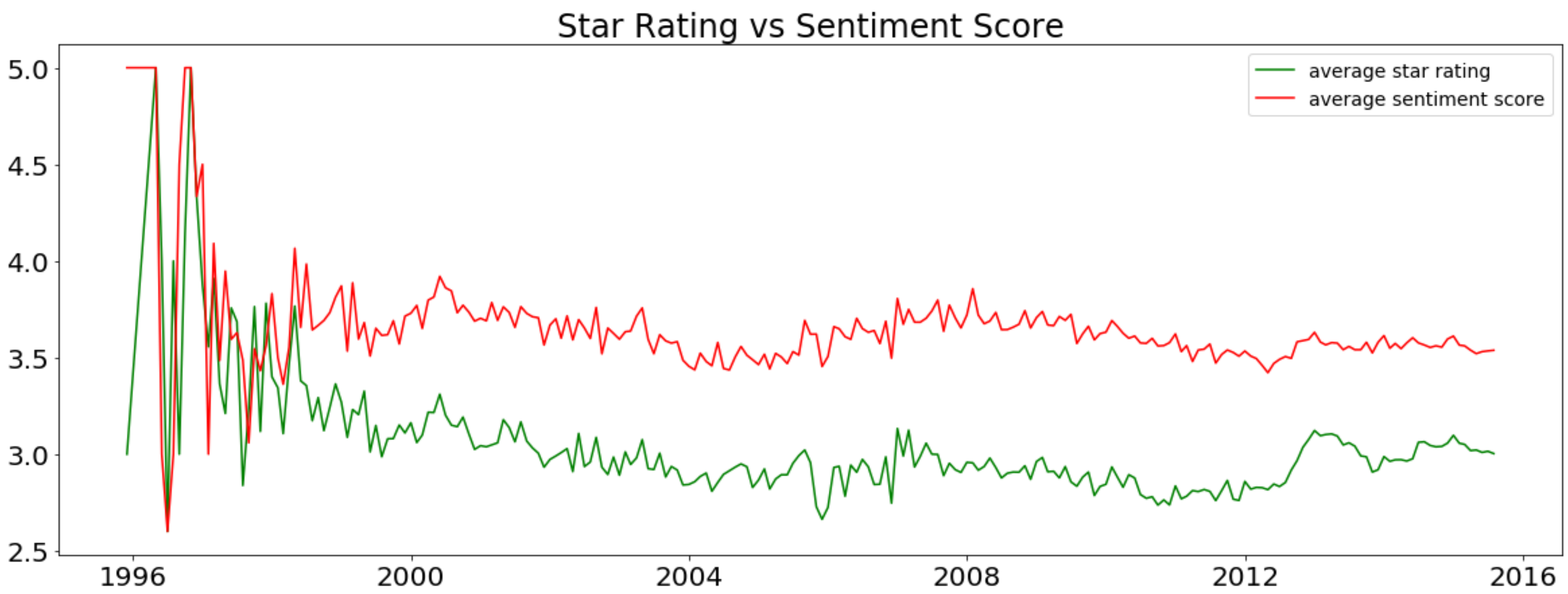
Modeling – Sentiment Analysis

review_body	truth_sentiment	sentiment	prediction
I bought this product 3 years ago and it's definitely worth it!!!	positive	pos	5.0
The case look nice and very well designed. But quality is bad!!	neutral	neg	1.0
Amazon is selling this more expensive than Ebay. Not recommending.	negative	neg	1.0



- Overall, the average sentiment score is higher than star rating. This indicates people tends to write slightly more positive reviews while giving a lower rating for the product, which means people have a higher standard on the product.
- Digital products seem to have a smaller differences in the two scores compared with others. This makes sense as the delivery of these products is done immediately and people get exactly what they expect so little difference between expectation and what's delivered.

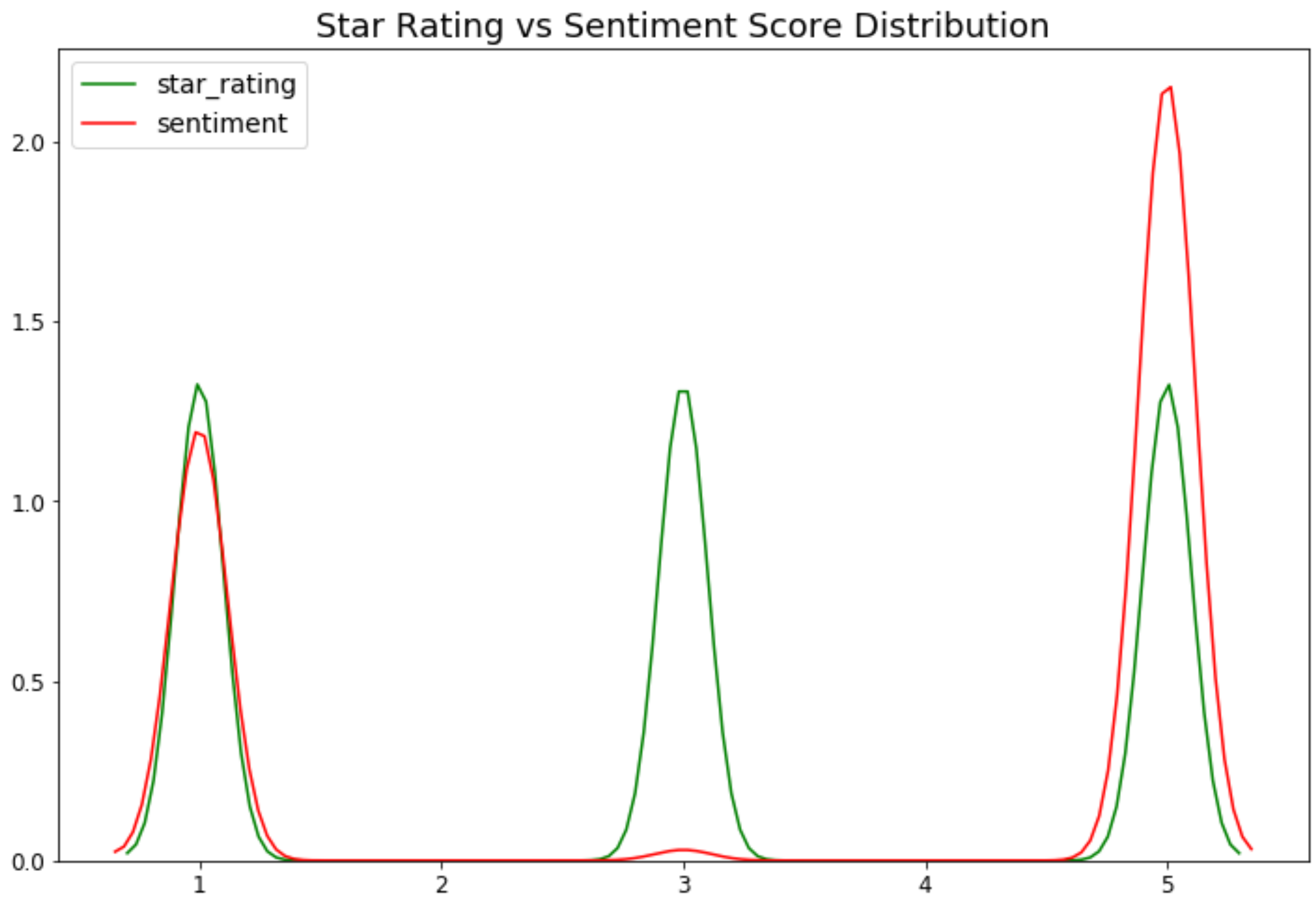
Modeling – Sentiment Analysis



- Overall, the average sentiment score is higher than star rating, and they are high correlated.
- In the late 1990s, there is large variance in both star rating and sentiment score, reflecting customers less confidence in the product Amazon offered or maybe the company itself, possibly due to Dot-com bubble.
- There is seasonality in both star rating and sentiment score.

Modeling – Sentiment Analysis

- As expected, the class distribution for star rating is well balanced, as we resampled the data.
- The sentiment score predicts very well when the rating is 1, but underrepresents when the rating is 3 and overrepresents when the rating is 5.



Modeling - ALS Recommendation System

Deploy Models

Alternating Least Squares (ALS) Model

- Model input: customer_id (more than 20 reviews), product_id and star_rating
- Train test split: 80:20
- Hyperparameter: maxIter=10, regParam=0.1, coldStartStrategy="drop", nonnegative = True
- Model output: user factor (for customer), item factor (for product)

Obstacle

- The whole dataset contains 110M reviews, making the factorization memory consuming
- We only include customers posting more than 20 reviews to downsize the data
- We set checkpoint and save models for further prediction and evaluation

Result

- Prediction RMSE on test dataset: 1.36

Modeling - ALS Recommendation System

Product recommendation for customer ID 44983593

Recommendation

product_title	product_category
Pure Spirulina Powder (5 lbs) Protein Superfoo...	Health & Personal Care
A Picture Book of Thomas Jefferson (Picture Bo...	Books
Reiko Magnetic Closure Flip Case for Samsung G...	Wireless
Quilted Purse, Handbag, Wallet - Black, Pink, ...	Shoes
Sparkle Wide Headband	Sports
Genuine Apple iMac Power Cord - 922-7139 922-9...	PC
Blondo Women's Marcia Knee-High Boot	Shoes
Jensen Shower Radio - JWM125	Electronics
Pert Plus 2 in 1 Shampoo + Conditioner Dandruf...	Beauty
Snowflake Thank You Cards (24 Foldover Cards a...	Office Products

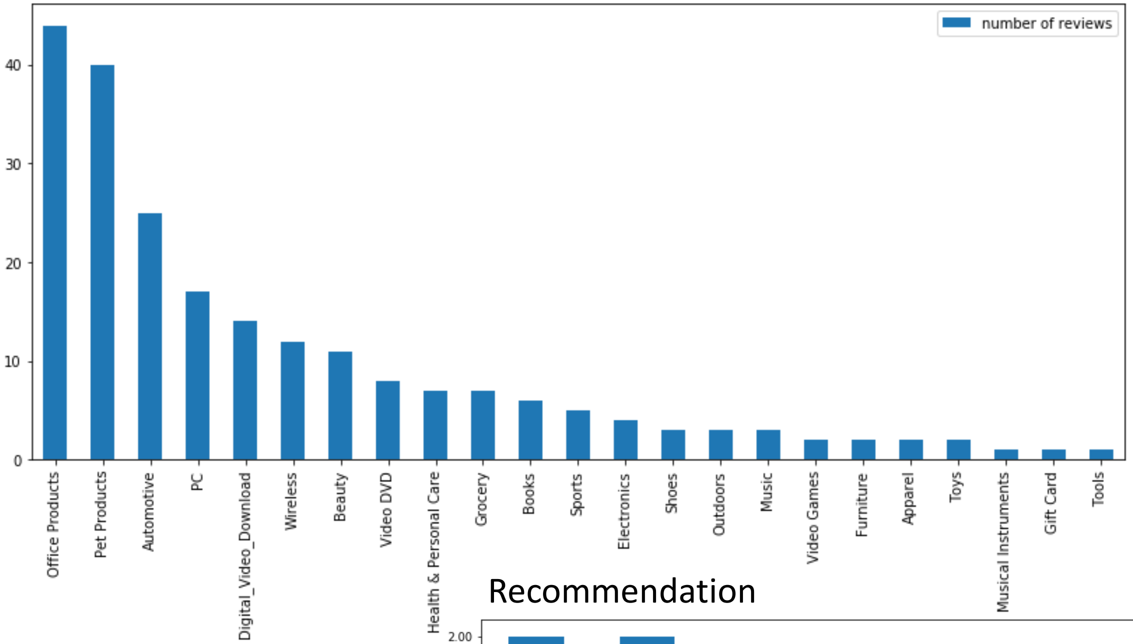
History (above 5 star)

	product_title	product_category
0	Day Runner Nature Weekly Planner Refill 2015, ...	Office Products
1	WORLD OF WARCRAFT HORDE PVP - WOW - Vinyl Car ...	Automotive
2	ProDen PlaqueOff Dental Powder	Pet Products
3	Prevue Pet Products 62605 Calypso Creations Sh...	Pet Products
4	Prestige Medical 607 Fluoride Coated Scissor, ...	Health & Personal Care
5	MUJI Aluminum Body Fountain Pen - Fine Nib - w...	Office Products
6	Constructive Anatomy (Dover Anatomy for Artists)	Books
7	RCA ANT111Z Durable FM Antenna, Rabbit Ears	Electronics
8	Logitech LX7 Cordless Optical Mouse	PC
9	Erase Markers	Office Products
10	Not Another Christmas Album: An Alternative Ch...	Music
11	Smittybilt 769541 First Aid Storage Bag	Automotive
12	Philips Sonicare HX6013/64 Proresults Brush He...	Beauty
13	LG WH16NS40 Super Multi Blue Internal SATA 16x...	PC
14	SiriusXM Snap XM radio reciever	Wireless
15	Evolution Undercoat Rake	Pet Products
16	Figure Drawing for All It's Worth	Books
17	Galaxy S4 Glass Screen Protector, Tech Armor P...	Wireless
18	Rampage Jeep 595001 Freedom Top Storage Bag	Automotive
19	Tough By Nature Hol-ee Roller, Assorted	Pet Products
20	Twilight Forever: The Complete Saga [Blu-ray +...	Video DVD
21	Anker AK-B2105121 PowerIQ Technology 40W 5-Por...	Wireless
22	Tuffy Barnyard Dog Toy	Pet Products
136	OneTigris Tactical Dog Training Molle Vest Har...	Pet Products
137	Maldon Sea Salt Flakes	Grocery
138	WEN by Chaz Dean Lavender Re-Moist Intensive H...	Beauty
139	Pentel Super Hi-Polymer Lead Refills, 0.5 mm, ...	Office Products
140	Corsair Obsidian Series 750D Performance Full ...	PC
141	magicJack GO Digital Phone Service, Includes 1...	Office Products
142	DUX Pencil and crayon Sharpener made of brass ...	Office Products
143	AntennaX Off-Road (13-inch) Antenna for (07 th...	Automotive
144	Philosophy, Science, and Technology Finger Pup...	Toys
145	Epson DURABrite XL T127120 Ultra 127 Extra Hig...	Office Products
146	MSI ATX DDR3 2600 LGA 1150 Motherboards Z97-G4...	PC
147	SABRE RED Pepper Gel Spray - Police Strength -...	Sports
148	Ballistix Sport 8GB Kit (4GBx2) DDR3 1600 MT/s...	PC
149	Doggles ILS Flames Dog Glasses	Pet Products
150	Tactical Gear Clip - Multipurpose Fastener For...	Sports
151	Flipside Wallets Men's RFID Blocking Flipside ...	Apparel
152	Dead Rising 2: Off The Record	Video Games
153	SanDisk Cruzer 8GB USB 2.0 Flash Drive (SDCZ36...	PC
154	Premier ECO Gentle Leader Head Dog Collar	Pet Products
155	Anker PowerCore+ mini 3350mAh Lipstick-Sized P...	Wireless
156	Ethical Plush Skinneeez Fox 24-Inch Stuffingle...	Pet Products
157	Leslie Sansone: Walk Away the Pounds Ultimate ...	Video DVD
158	Filofax Ruled Pink Paper (B133007)	Office Products

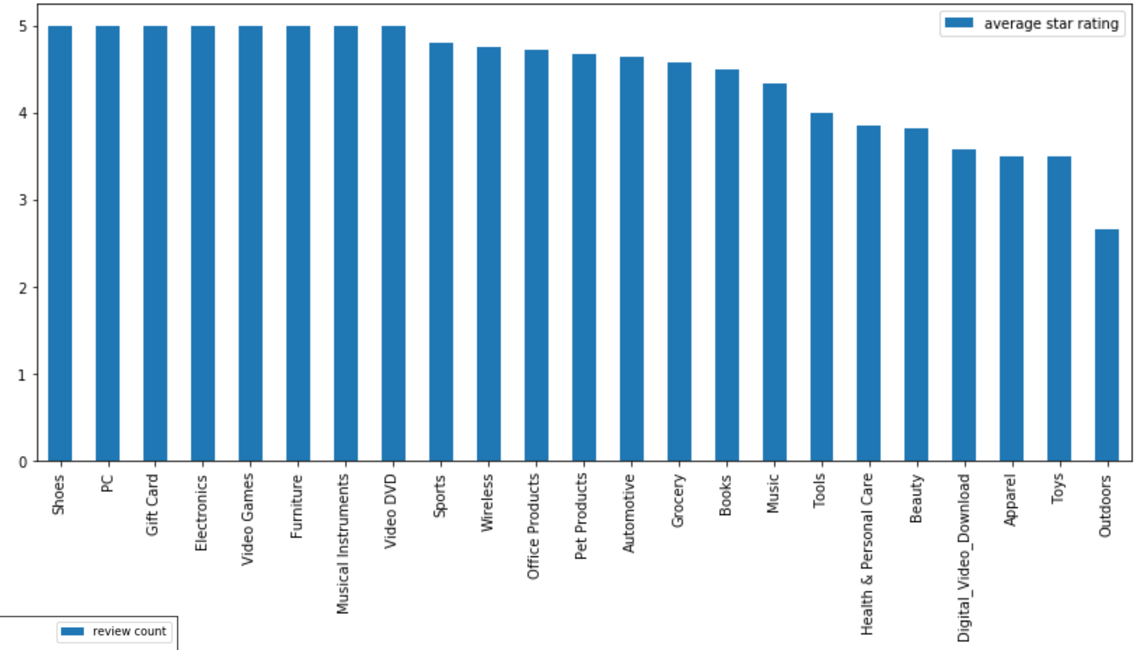
Modeling - ALS Recommendation System

Product recommendation for customer ID 44983593

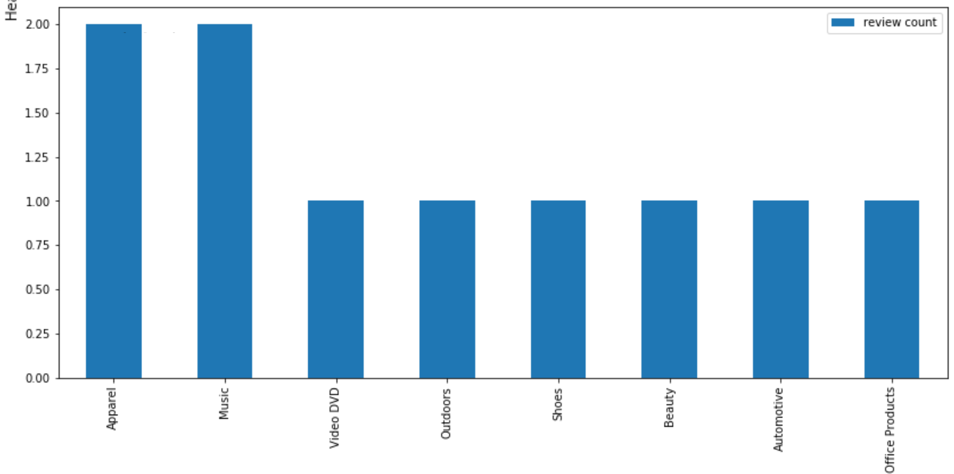
Review number by category



Average rating by category



Recommendation



Conclusion and Recommendations

01 Highest number of review in tech-related categories such as wireless device & PC. Highest rating in gift card. Lowest rating in software.

02 Customer Analysis can be used to detect fraudulent activities and monitor reviews

03 SparkNLP is our go-to model but it takes longer to train; If time & computer resources are concerns we could switch to base model.

04 The time series analysis on sentiment and star rating could provide some insights on how the company performs.

Future Work

Big Data

- Try other big data technologies such as repartitioning, compression, cache to improve the efficiency
- Our data is outdated so we could incorporate the most recent data source to get a better sense of the reviews

EDA

- For customer analytics, we only considered number of reviews and average rating score, we can try to evaluate review text as well.

Modeling

- For NLP models, we pick linear regression to predict due to computer constraint. We could pick more advanced models when spinning up more nodes.
- For sentiment analysis, we compared two pretrained models but could try more models to increase our prediction power.

Q & A