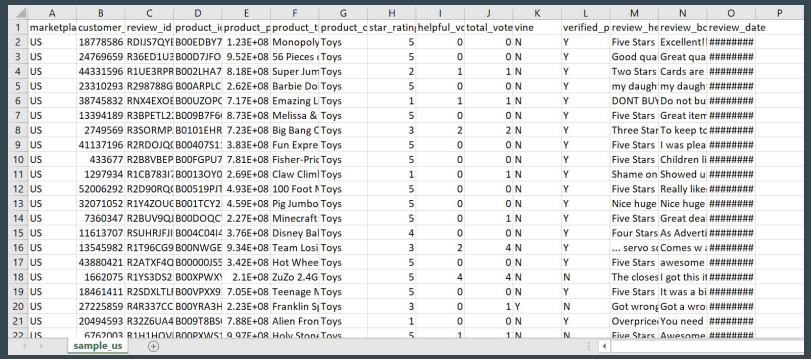
Amazon Customer Reviews

Lili Zhang
Min Yang
Jiaqi Ling

Introduction



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	marketpla	customer_	review_id	product_i	product_p	product_t	product_c	star_ratin	helpful_vc	total_vote	vine	verified_p	review_he	review_bc	review_date	
2	US	18778586	RDIJS7QY(	B00EDBY7	1.23E+08	Monopoly	Toys	5	0	0	N	Y	Five Stars	Excellent!	########	
3	US	24769659	R36ED1U3	B00D7JFO	9.52E+08	56 Pieces (	Toys	5	0	0	N	Y	Good qua	Great qua	########	
4	US	44331596	R1UE3RPR	B002LHA7	8.18E+08	Super Jum	Toys	2	1	1	N	Y	Two Stars	Cards are	########	
5	US	23310293	R298788G	B00ARPLC	2.62E+08	Barbie Do	Toys	5	0	0	N	Y	my daugh	my daugh	########	
6	US	38745832	RNX4EXOE	B00UZOPC	7.17E+08	Emazing L	Toys	1	1	1	N	Y	DONT BU	Do not bu	########	
7	US	13394189	R3BPETL2	B009B7F6	8.73E+08	Melissa &	Toys	5	0	0	N	Y	Five Stars	Great iten	########	
8	US	2749569	R3SORMP	B0101EHR	7.23E+08	Big Bang C	Toys	3	2	2	N	Y	Three Star	To keep tc	########	
9	US	41137196	R2RDOJQ(	B00407S1	3.83E+08	Fun Expre	Toys	5	0	0	N	Y	Five Stars	I was plea	########	
10	US	433677	R2B8VBEP	B00FGPU7	7.81E+08	Fisher-Pric	Toys	5	0	0	N	Y	Five Stars	Children li	########	
11	US	1297934	R1CB783I	B0013OY0	2.69E+08	Claw Climl	Toys	1	0	1	N	Y	Shame on	Showed u	########	
12	US	52006292	R2D90RQ(	B00519PJT	4.93E+08	100 Foot N	Toys	5	0	0	N	Y	Five Stars	Really like	########	
13	US	32071052	R1Y4ZOU(	B001TCY2	4.59E+08	Pig Jumbo	Toys	5	0	0	N	Y	Nice huge	Nice huge	########	
14	US	7360347	R2BUV9QJ	B00DOQC	2.27E+08	Minecraft	Toys	5	0	1	N	Y	Five Stars	Great dea	########	
15	US	11613707	RSUHRJFJI	B004C04I4	3.76E+08	Disney Bal	Toys	4	0	0	N	Y	Four Stars	As Adverti	########	
16	US	13545982	R1T96CG9	B00NWGE	9.34E+08	Team Losi	Toys	3	2	4	N	Y	... servo sc	Comes w i	########	
17	US	43880421	R2ATXF4Q	B00000J5S	3.42E+08	Hot Whee	Toys	5	0	0	N	Y	Five Stars	awesome	########	
18	US	1662075	R1YS3DS2	B00XPWX\	2.1E+08	ZuZo 2.4G	Toys	5	4	4	N	N	The closes	I got this it	########	
19	US	18461411	R2SDXLTL	B00VPXX9	7.05E+08	Teenage M	Toys	5	0	0	N	Y	Five Stars	It was a bi	########	
20	US	27225859	R4R337CC	B00YRA3H	2.23E+08	Franklin S	Toys	3	0	1	Y	N	Got wrong	Got a wro	########	
21	US	20494593	R32Z6UA4	B009T8BS	7.88E+08	Alien Fron	Toys	1	0	0	N	Y	Overprice	You need	########	
22	US	6762003	R1H1HQV	B00PXWS1	9.97E+08	Holy Stone	Toys	5	1	1	N	N	Five Stars	Awesome	########	

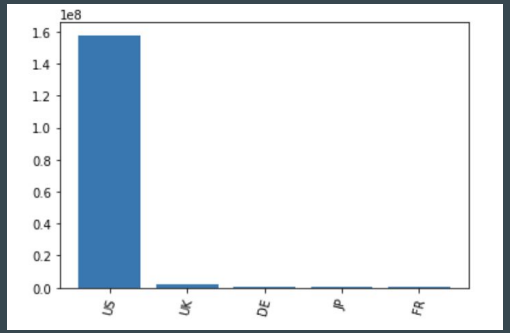
sample_us

Screenshot of dataset sample

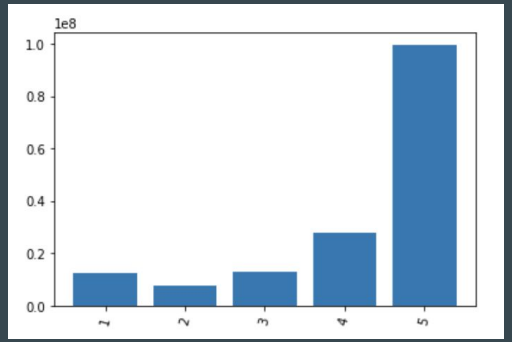
- Over 130+ million customer reviews from 1995 to 2015
- Data size: Around 50 GB
- 200K+ customers in 5 countries
- 15 variables include star_rating, review_body, helpful_votes, vine, etc.
- Available both in TSV and Parquet format in S3 bucket

Data link:https://s3.amazonaws.com/amazon-reviews-pds/readme.html

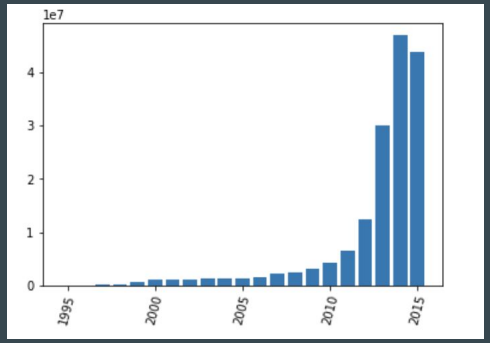
Distribution of the dataset



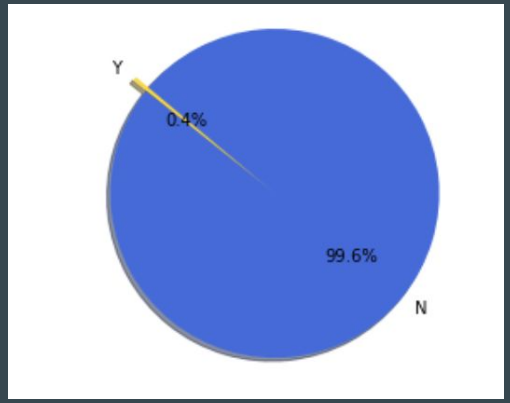
Marketplace



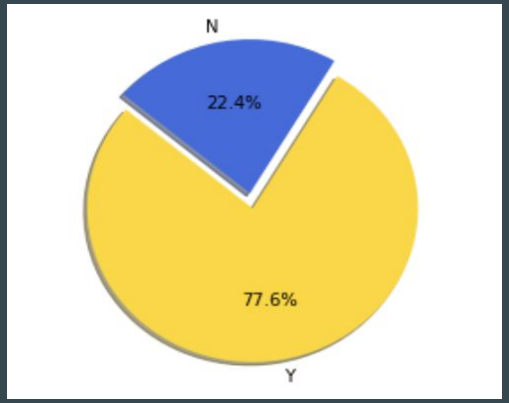
Star Rating



Year



Vine Membership



Verified Purchase

Methodology

- **Basic data prep**

  Tidy and Clean in general;

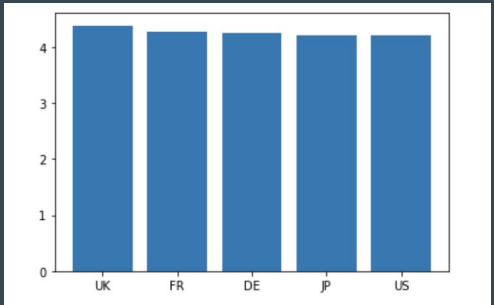
  Drop records with missing or abnormal values

- **Query and EDA**: Spark DataFrames, SparkSQL, MapReduce
- **Data visualization:** matplotlib, wordcloud packages
- **Sentiment Analysis:** Sentiment Lexicon
- **Modeling:** pyspark.ml

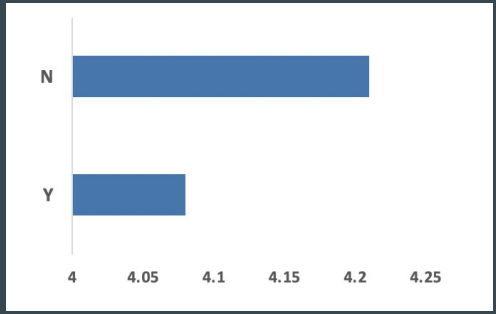
Hypothesis

1. How do ratings vary with verified purchase, Vine membership, Marketplace and over time?
2. How do level of helpfulness vary by consumer identity?
3. What's the common words in different product categories?
4. Do Amazon star ratings reflect sentiment in reviews?
5. What type of customers tend to give positive reviews?

1. How do star ratings vary?



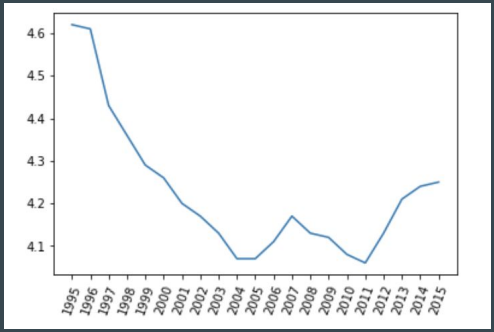
Marketplace



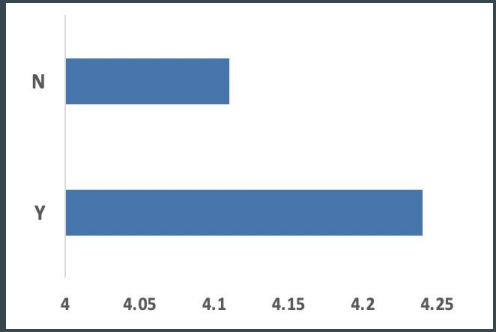
Vine

product_category	avg_rating
Gift_Card	4.73
Digital_Music_Pur...	4.64
Music	4.44
Books	4.34
Grocery	4.31
Video_DVD	4.31
Digital_Ebook_Pur...	4.31
Tools	4.26
Musical_Instruments	4.25
Automotive	4.25

Product Category

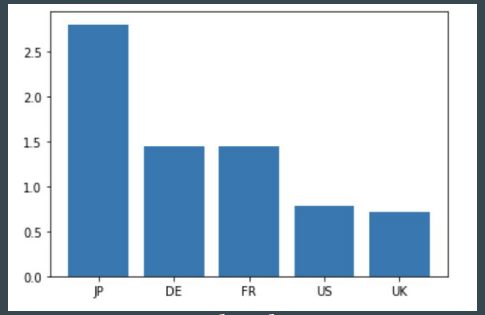


Year

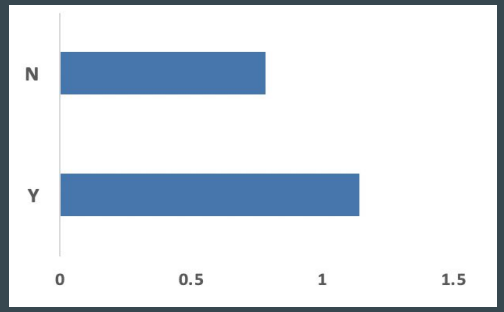


Verified Purchase

2. How do review helpfulness vary?



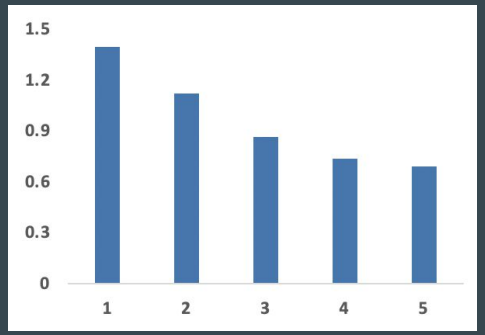
Marketplace



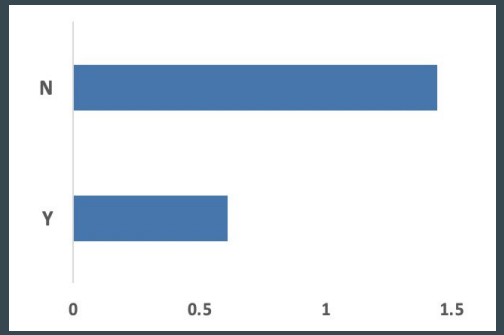
Vine

product_category	avg_help	count
Video	2.07	437408
Music	1.50	6177622
Software	1.42	342133
Books	1.40	20725905
Major_Appliances	1.36	96894
Video_DVD	1.21	7135754
Personal_Care_App...	1.17	86686
Home_Entertainment	1.01	743684
Video_Games	0.92	1808434
Furniture	0.92	792111
Camera	0.89	1838692
Musical_Instruments	0.88	920676
Health_&_Personal...	0.87	5332715
Lawn_and_Garden	0.82	2559115
Luggage	0.80	349108

Product Category



Star_rating



Verified Purchase

3. What's the common words in reviews?

High Rating

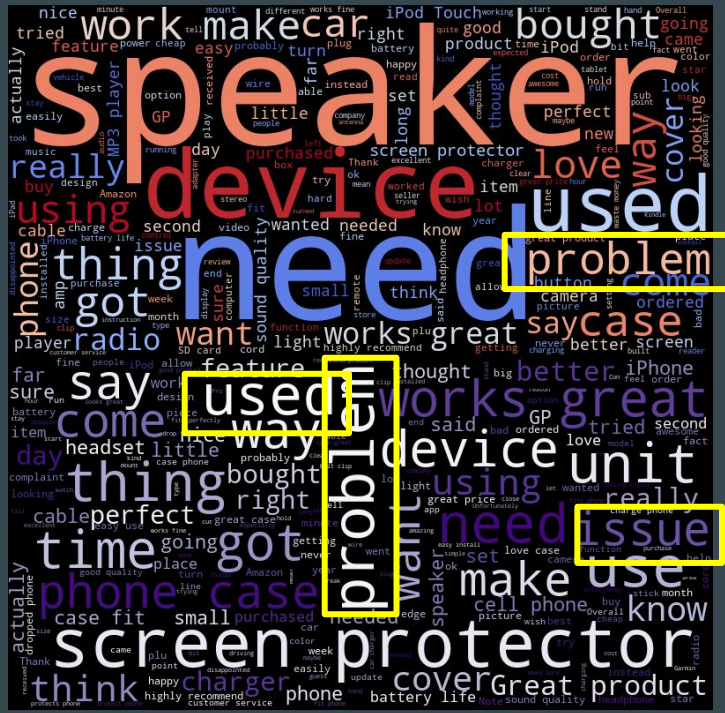
Low Rating

Gift Card

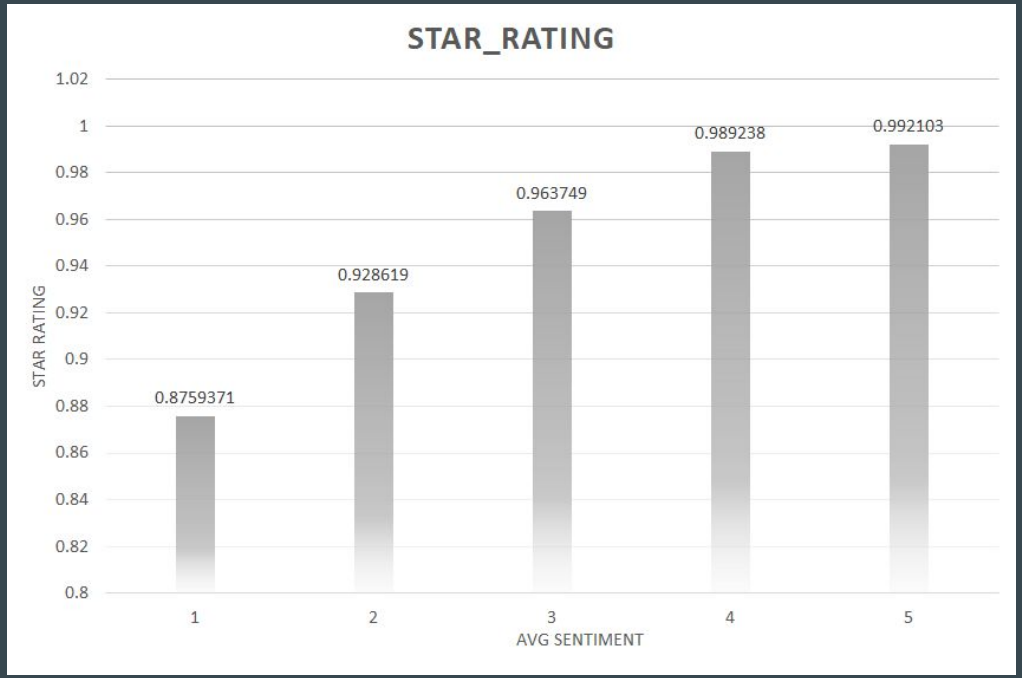
Mobile Electronics

Book

Wireless



4. Do Amazon star ratings reflect sentiment in reviews?



Average sentiment score of different star rating level

Average star rating of different type of customer

positive_or_not	avg
0.0	2.420764494862878
1.0	4.147445398809962

5. What type of customers tend to give positive reviews ?

Verified purchase distribution of giving positive reviews

verified_purchase	count	percent
Y	25182499	66.4595602038264
N	12708963	33.54043979617361

Verified purchase distribution of giving negative reviews

verified_purchase	count	percent
Y	738586	79.5994732066069
N	189292	20.400526793393098

Vine membership distribution of giving positive reviews

vine	count	percent
Y	8817	0.023269094235529895
N	37882645	99.97673090576447

Vine membership distribution of giving negative reviews

vine	count	percent
N	927878	100.0

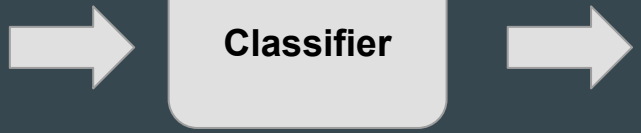
Classification Model

- **Purpose**

By classifying the helpfulness of a newly posted review, Amazon can optimize the website by putting helpful reviews on the top of section to better notify fellow customers with the product experience

Features

Review_sentiment
star_ratings
vine_membership
....



Label: help_pct

Model	Accuracy
Logistic Regression	0.726
Random Forest	0.745
GBTClassifier	0.742

Future Work

1. **More work on Modeling:**

Refine information by featuring engineering to improve operational efficiency
(10 m4.xlarge core ~ 1-2 hours/model)

2. **Detecting fake reviews:**

Include a collection of reviews that have been identified as non-compliant with respect to Amazon policies for research on detecting promotional or biased reviews

Thank you!