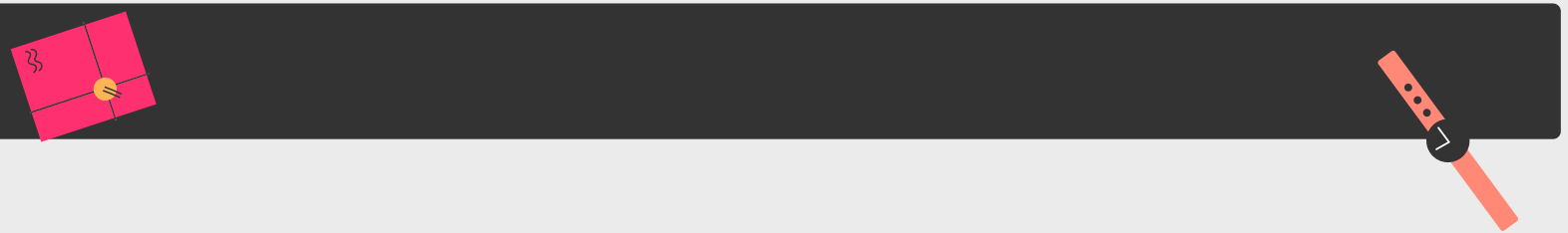
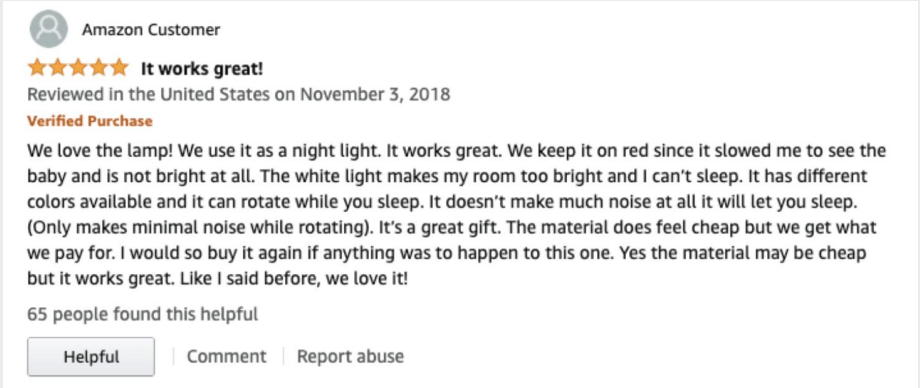
THE UNIVERSITY OF CHICAGO

Big Data Platform | Autumn 2023

Boya Zeng

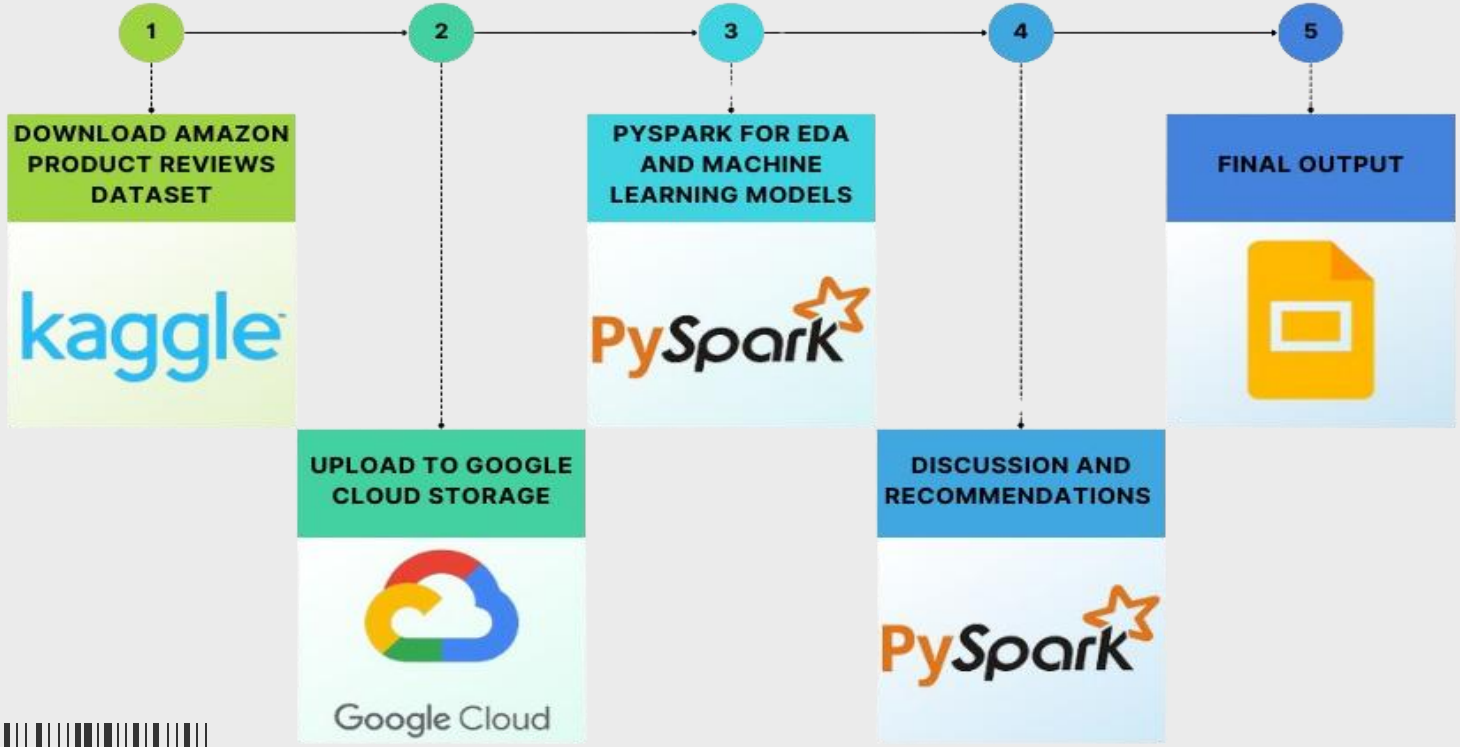


sentimentdl_use_twitter



- marketplace
- customer_id
- review_id
- product_id product_parent
- product_title product_category
- star_rating helpful_votes total_votes
- vine verified_purchase
- review_headline review_body review_date

https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-dataset/data?select=amazon_reviews_us_Apparel_v1_00.tsv

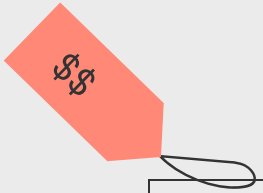

1
2
3
4
5
DOWNLOAD AMAZON PRODUCT REVIEWS DATASET
kaggle
UPLOAD TO GOOGLE CLOUD STORAGE
Google Cloud
PYSPARK FOR EDA AND MACHINE LEARNING MODELS
PySpark
DISCUSSION AND RECOMMENDATIONS
PySpark
FINAL OUTPUT

Buckets > msca-bdp-student-gcs > group10-amazon-review

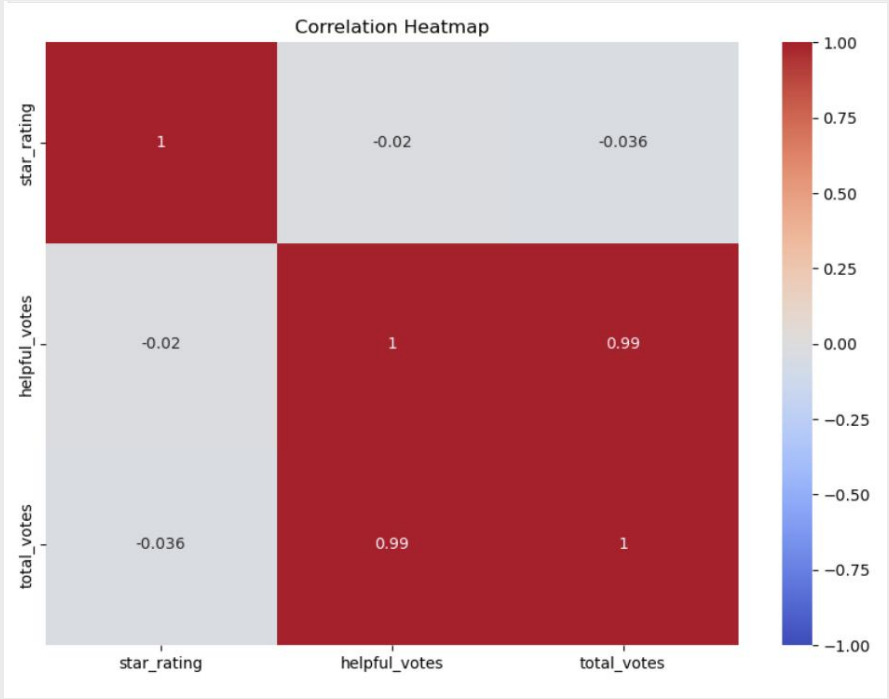
Name	Size
amazon_reviews_us_Baby_v1_00.tsv	831.9 MB
amazon_reviews_us_Beauty_v1_00.tsv	2 GB
amazon_reviews_us_Camera_v1_00.tsv	1 GB
amazon_reviews_us_Electronics_v1_0...	1.6 GB
amazon_reviews_us_Furniture_v1_00....	350 MB

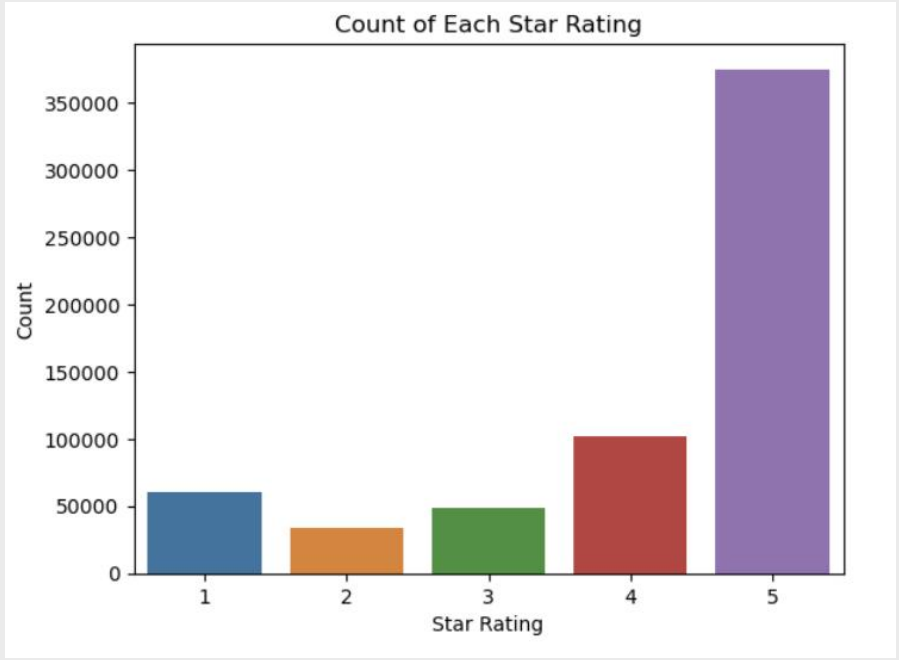
- Remove any unwanted characters
  - #, %, ^, //, etc.
- Convert to all lowercase text for consistency
- Check for duplicates based on a column
- Remove missing values

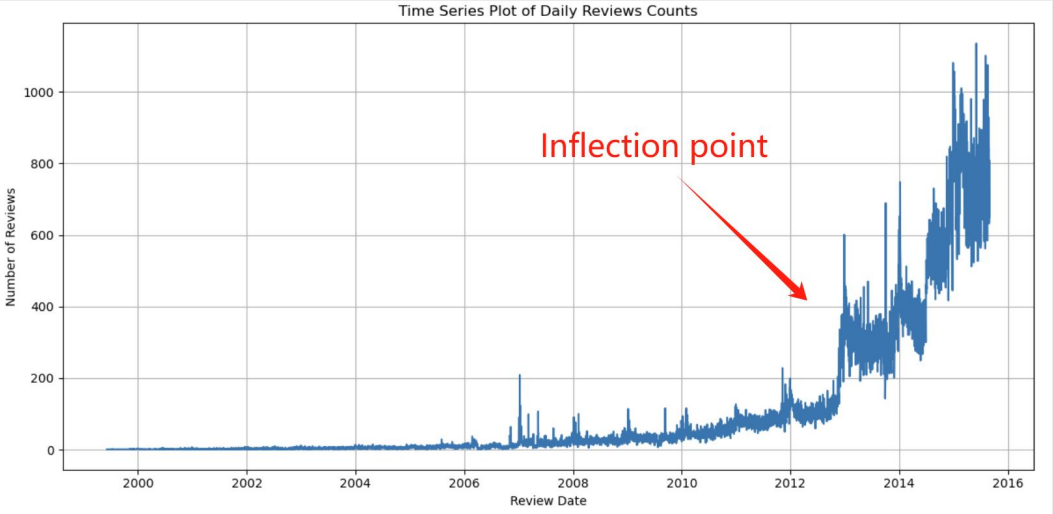
```
Missing values in 'marketplace': 0
Missing values in 'customer_id': 0
Missing values in 'review_id': 0
Missing values in 'product_id': 0
Missing values in 'product_parent': 0
Missing values in 'product_title': 0
Missing values in 'product_category': 0
Missing values in 'star_rating': 0
Missing values in 'helpful_votes': 0
Missing values in 'total_votes': 0
Missing values in 'vine': 0
Missing values in 'verified_purchase': 0
Missing values in 'review_headline': 1
Missing values in 'review_body': 201
Missing values in 'review_date': 57
```

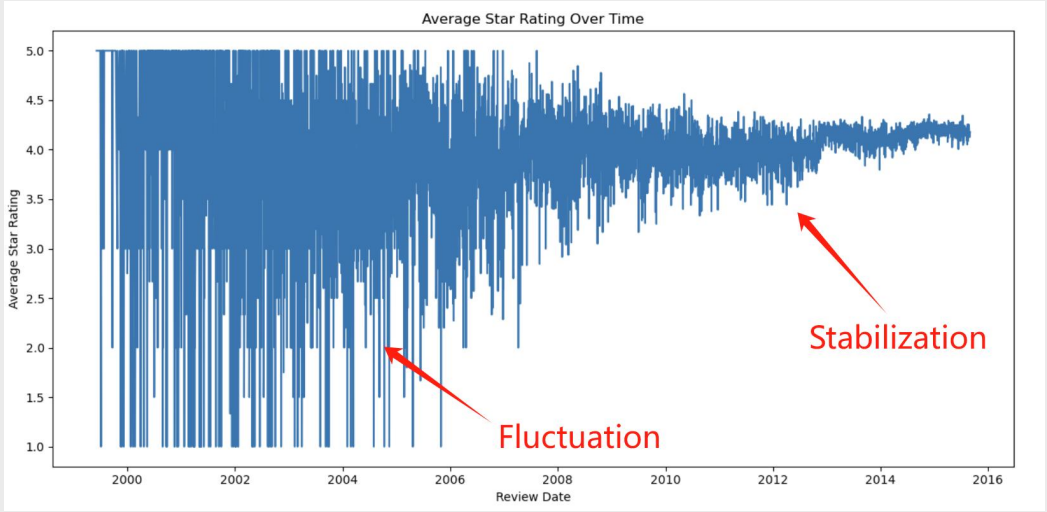


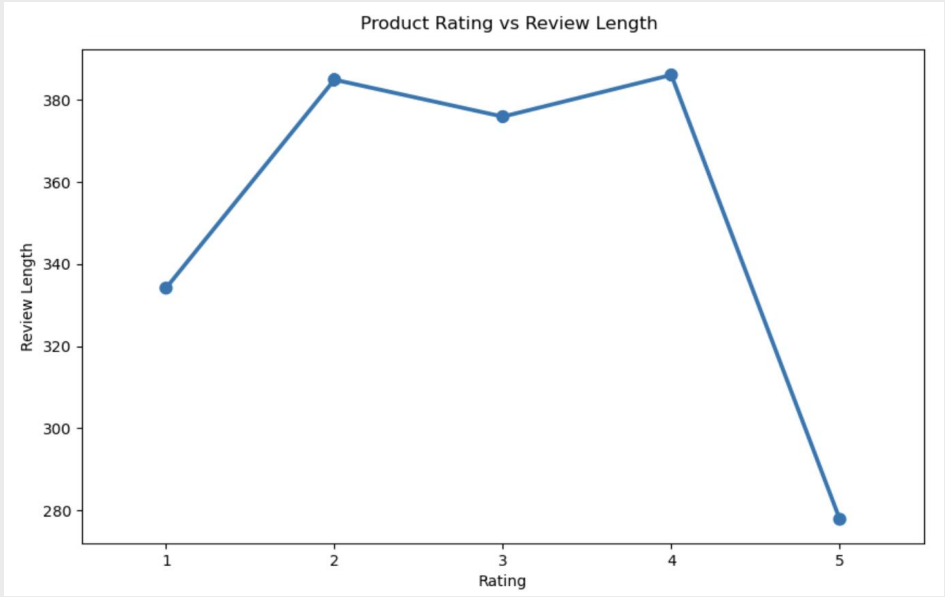
count	619630	619630	619630
mean	4.13	1.97	2.47
std	1.33	21.39	22.61
min	1	0	0
max	5	8937	9072


Correlation Heatmap
star_rating
helpful_votes
total_votes
1
-0.02
-0.036
-0.02
1
0.99
-0.036
0.99
1
star_rating
helpful_votes
total_votes
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

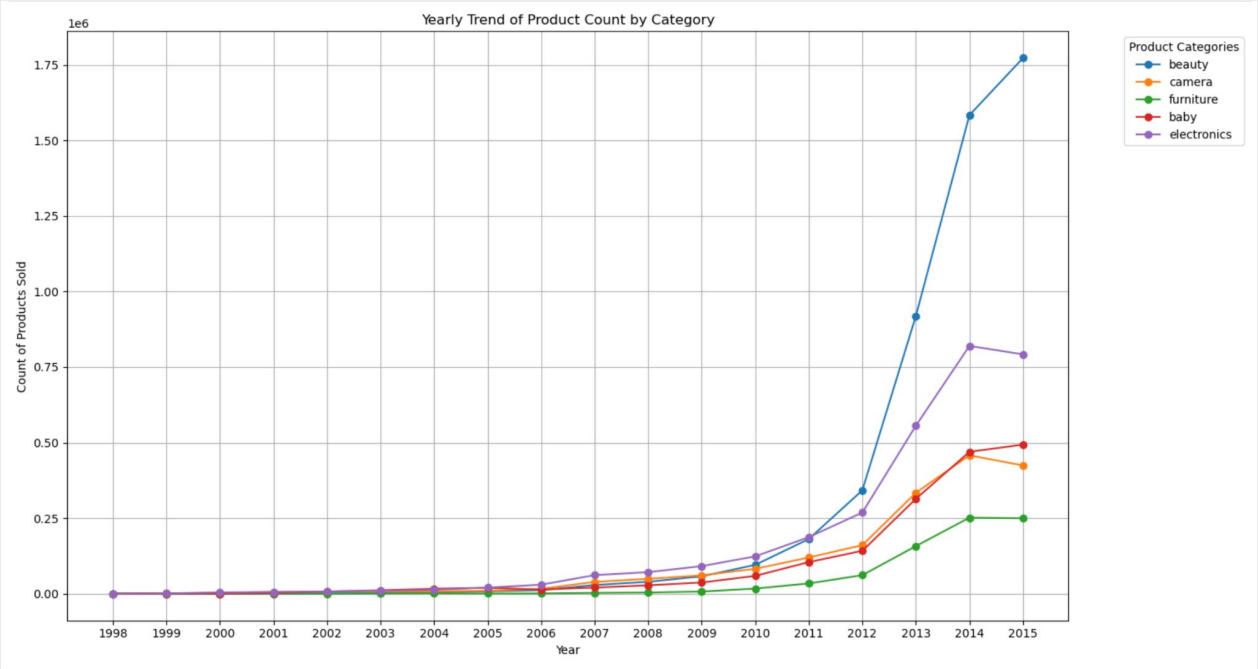

Count of Each Star Rating
350000
300000
250000
200000
150000
100000
50000
0
Count
1
2
3
4
5
Star Rating

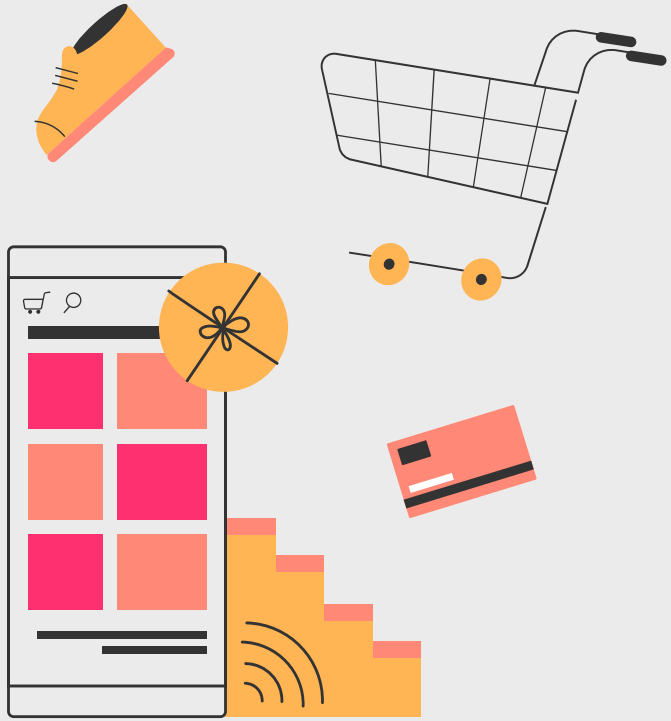

Time Series Plot of Daily Reviews Counts
Inflection point
Number of Reviews
Review Date
0
200
400
600
800
1000
2000
2002
2004
2006
2008
2010
2012
2014
2016


Average Star Rating Over Time
Average Star Rating
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
2000
2002
2004
2006
2008
2010
2012
2014
2016
Review Date
Stabilization
Fluctuation


Product Rating vs Review Length
Review Length
380
360
340
320
300
280
1
2
3
4
5
Rating

Word clouds for review headline


Yearly Trend of Product Count by Category
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Count of Products Sold
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Year
Product Categories
beauty
camera
furniture
baby
electronics

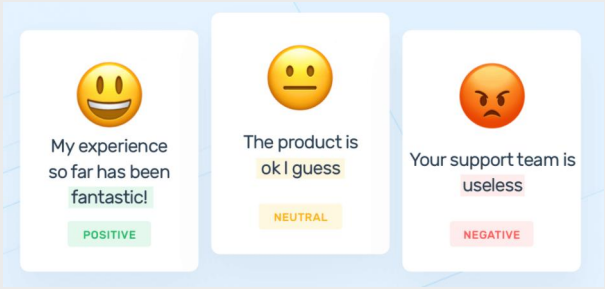


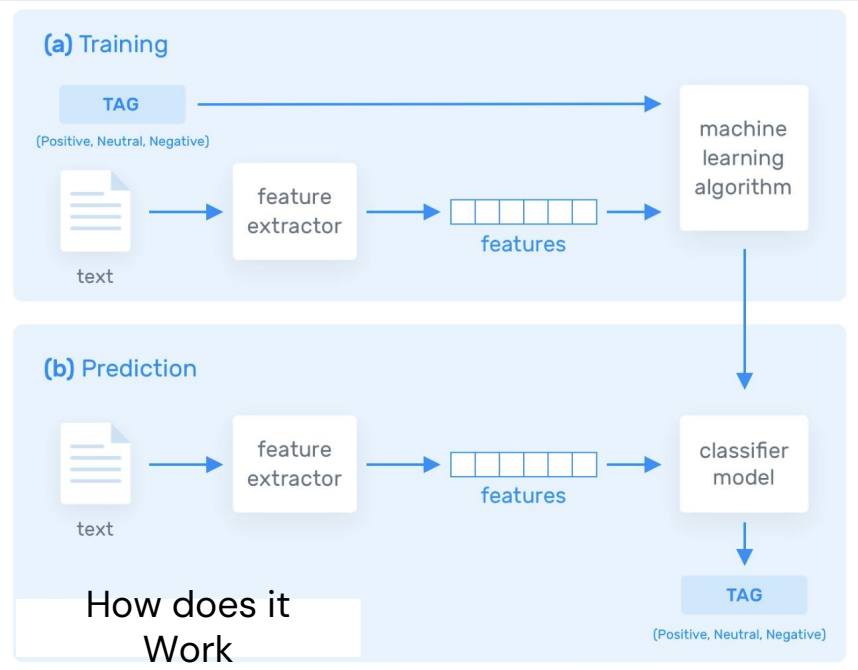
1. Sentiment Analysis
2. Customer Segmentation
3. Recommendation System

- The process of using NLP and machine learning to classify sentiments (positive, negative, neutral) in text data

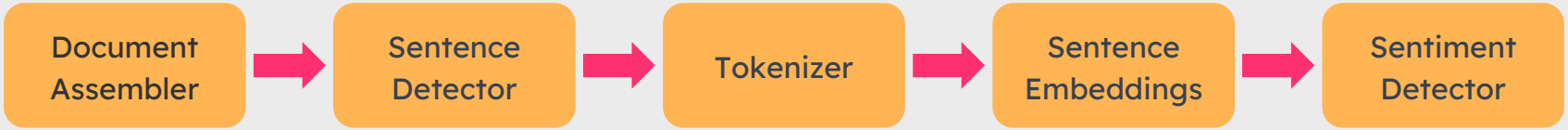
Our Problem

- Classify the sentiment of an Amazon product review as Positive, Neutral, or Negative.

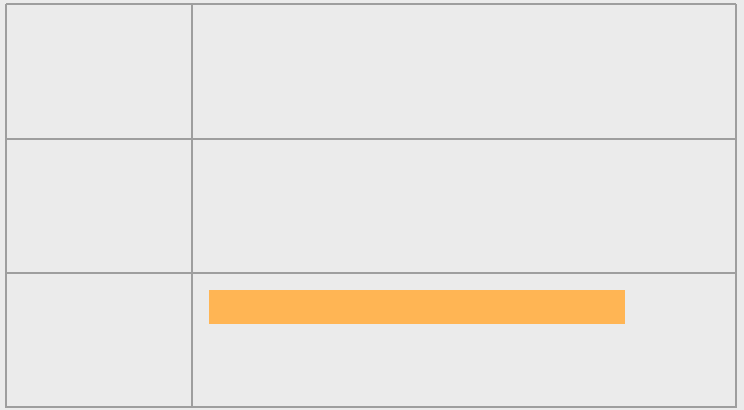




How does it Work

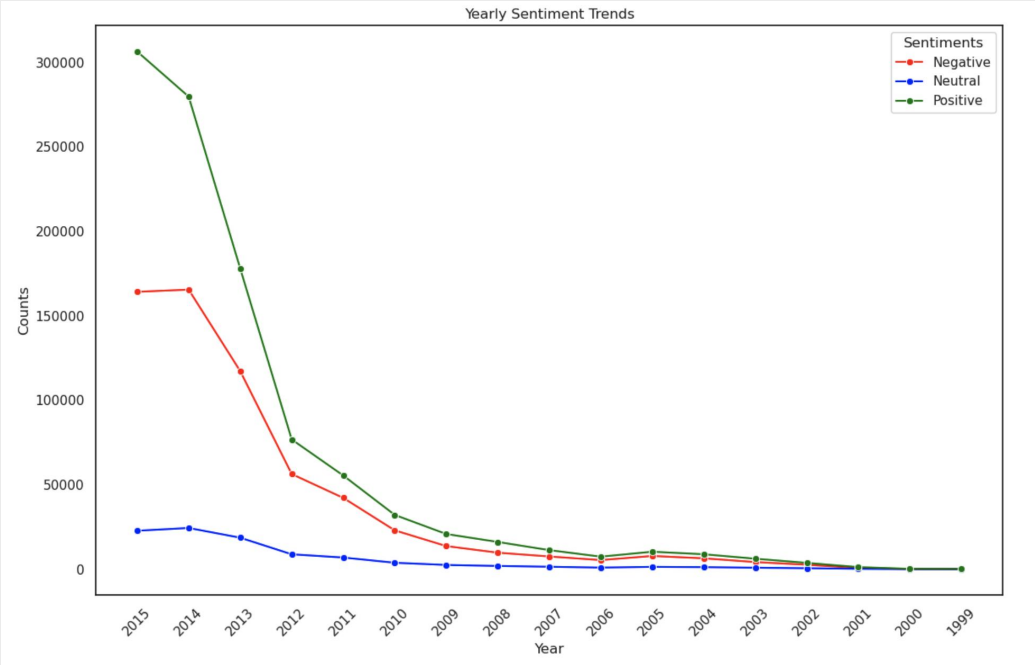

Document Assembler
Sentence Detector
Tokenizer
Sentence Embeddings
Sentiment Detector

	Logistic Regression Model
Star Rating	
Sentiment	

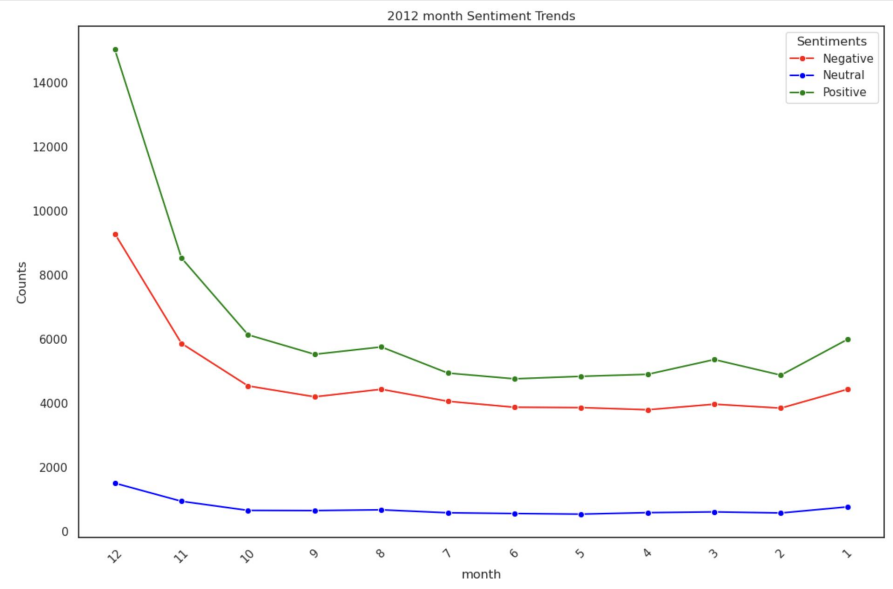


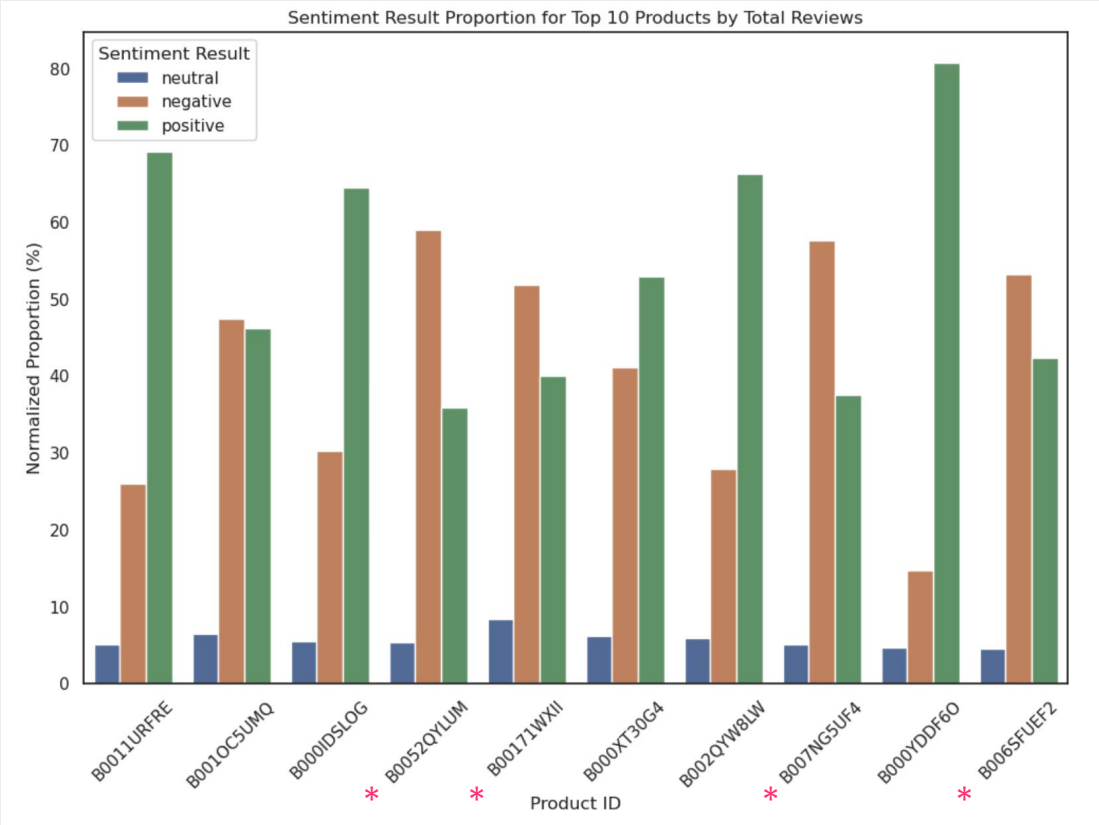
- Drop Neutral,

Trend Analysis



Trend Analysis



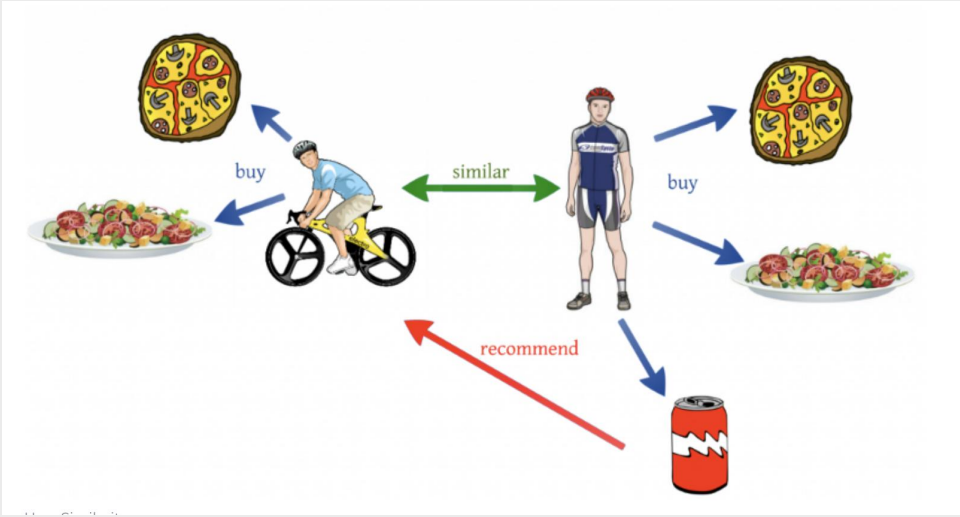


	product_id	avg_star_rating
1	B0052QYLUM	3.874745228830832
2	B00171WXII	4.656218402426694
3	B006SFUEF2	3.8251057827926656
4	B007NG5UF4	3.9677570093457946

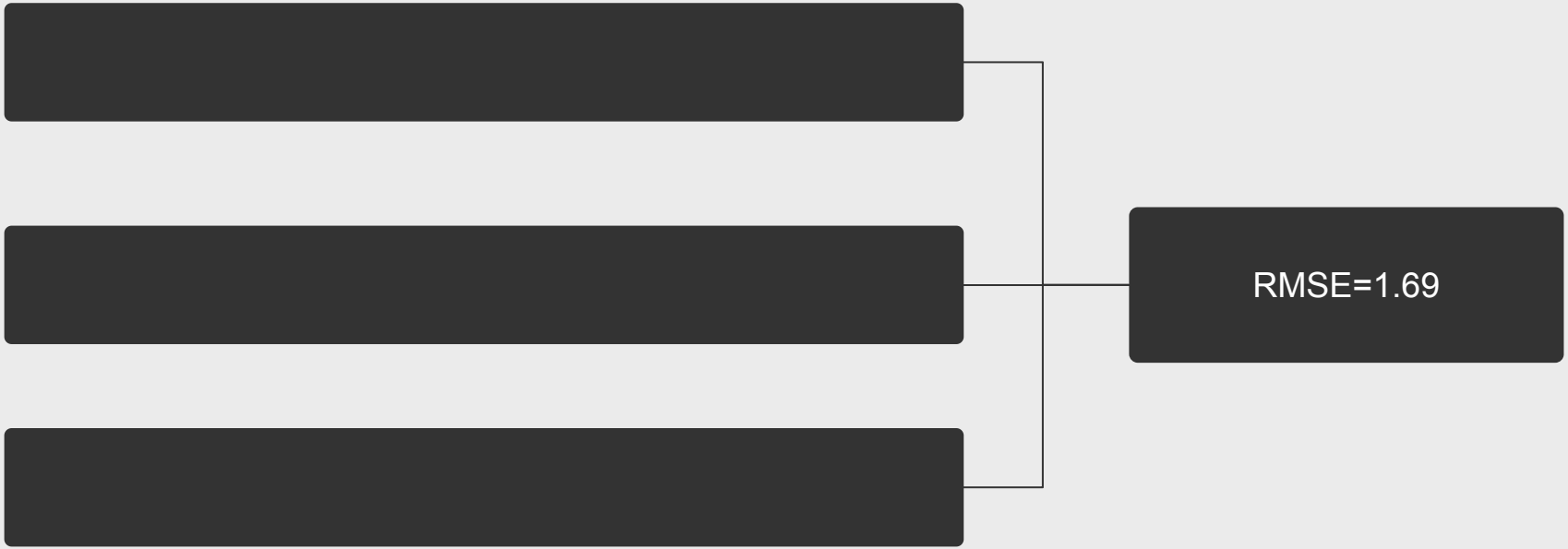
- The process of recommending products to users based on their **past reviews and ratings.**

Our Problem

- Recommend products to users using ALS (Alternating Least Squares



Why ALS?


RMSE=1.69

customer_id	product_id	star_rating	userIndex	productIndex	prediction
14553407	B0018CWAPC	4.0	2366.0	17773.0	3.3290687
19446147	B00M6N6OSG	4.0	25517.0	46717.0	3.0405002
40657719	B00E6UMJBS	5.0	38311.0	12368.0	3.272395
14693899	B004A2ZCA2	5.0	4190.0	13691.0	3.869201
15314315	B002L3T9ZG	5.0	168235.0	2142.0	3.468196
43991184	B000K4YSVI	4.0	29075.0	274.0	3.7592874
84161	B0051BOEGE	5.0	277780.0	1685.0	3.8804202
7469708	B00B507D7C	5.0	276613.0	2387.0	4.230116
19186502	B00RX4XSGY	1.0	181716.0	36378.0	0.7524073
21625369	B000FT7NRO	5.0	34488.0	3631.0	4.133099

only showing top 10 rows

userIndex	recommendations
54153	[{130316, 4.146166}, {157300, 4.070095}, {87688, 4.0627165}, {120423, 3.9332132}, {106345, 3.9102447}, {137490, 3.8619552}, {121945, 3.859049}, {137083, 3.8465044}, {102506, 3.8437939}, {102510, 3.843584}]
185454	[{107133, 4.3452725}, {113525, 4.2993264}, {127194, 4.2806015}, {143549, 4.275772}, {103862, 4.2616043}, {122801, 4.250621}, {140351, 4.2365813}, {121085, 4.2060156}, {145834, 4.2051806}, {116915, 4.204182}]
830365	[{140351, 1.0298496}, {103862, 1.0269122}, {116915, 1.019562}, {120940, 1.0094104}, {136497, 1.0090052}, {142031, 1.004876}, {143549, 0.99919}, {158770, 0.99863714}, {107133, 0.99649}, {50515, 0.99398077}]

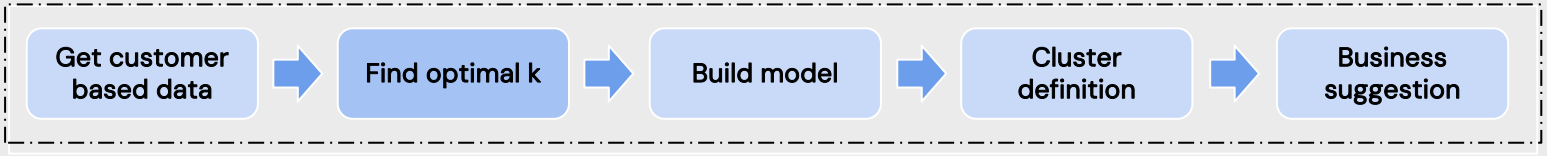
- The process of dividing a customer base into groups with similar characteristics or behaviors

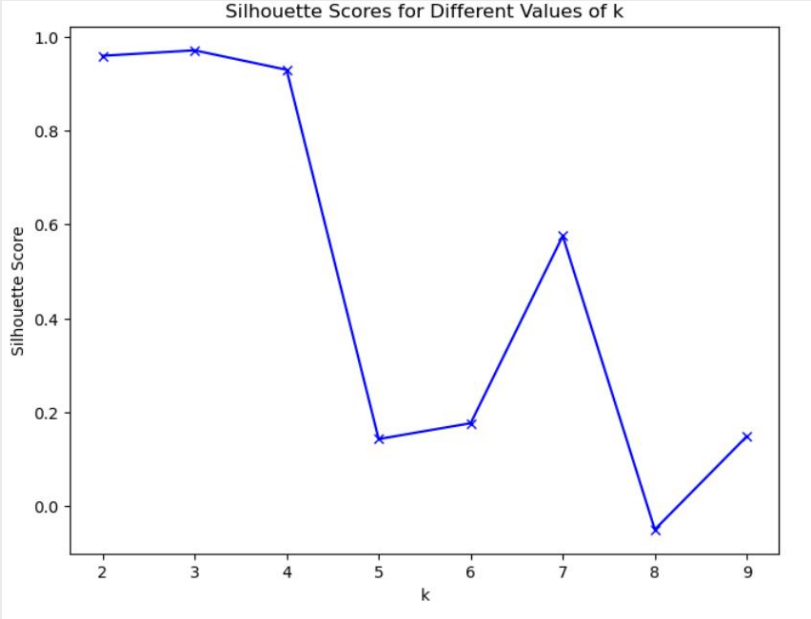
Our Problem

- Cluster Amazon customers based on their reviewing behavior using **K-means Clustering**


```
+----------+----------+--------------+---------+--------------+--------------------+---------------+----------+------------+----------+----+-----------------+--------------------+--------------------+-----------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_headline|         review_body|review_date|
+----------+----------+--------------+---------+--------------+--------------------+---------------+----------+------------+----------+----+-----------------+--------------------+--------------------+-----------+
|         US|    9970739| R8EWA1OFT84NX|B00GSP5D94|     329991347|Summer Infant Swa...|           Baby|         5|           0|         0|   N|                Y|Great swaddled bl...|Loved these swadd...| 2015-08-31|
|         US|   23538442|R2JWY4YRQD4FOP|B00YYDDZGU|     646108902|Pacifier Clip Gir...|           Baby|         5|           0|         0|   N|                N|Too cute and real...|These are adorabl...| 2015-08-31|
|         US|    8273344| RL5ESX231LZ0B|B00BUBNZC8|     642922361|Udder Covers - Br...|           Baby|         5|           0|         0|   N|                Y|         Five Stars|          Great gift| 2015-08-31|
+----------+----------+--------------+---------+--------------+--------------------+---------------+----------+------------+----------+----+-----------------+--------------------+--------------------+-----------+
```

```
+-----------+-----------+------------+-----------------+----------------+--------------+-------------+-------------+--------------------------+--------------------------+
|customer_id|review_nums|product_nums|  avg_star_rating|sum_helpful_votes|sum_total_votes|vine_Y_counts|vine_N_counts|verified_purchase_Y_counts|verified_purchase_N_counts|
+-----------+-----------+------------+-----------------+----------------+--------------+-------------+-------------+--------------------------+--------------------------+
|   47914293|          2|           2|              4.5|               1|             1|            0|            2|                         2|                         0|
|    9884235|          6|           6|4.166666666666667|               2|             2|            0|            6|                         5|                         1|
|   49447827|          2|           2|              4.0|               0|             1|            0|            2|                         2|                         0|
|   28505016|          1|           1|              5.0|               0|             0|            0|            1|                         1|                         0|
|   14927295|          1|           1|              5.0|               0|             0|            0|            1|                         1|                         0|
+-----------+-----------+------------+-----------------+----------------+--------------+-------------+-------------+--------------------------+--------------------------+
```


Get customer based data
Find optimal k
Build model
Cluster definition
Business suggestion


Silhouette Scores for Different Values of k
Silhouette Score
k
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
7
8
9

cluster	avg_review_nums	avg_product_nums	avg_star_rating	avg_helpful_votes	avg_total_votes	avg_vine_Y_counts	avg_vine_N_counts	avg_verified_purchase_Y_counts	avg_verified_purchase_N_counts
1	1.2108916935452279	1.2103775186408807	4.812399341420913	1.2110614528787265	1.2110614528787265	0.002198709829065...	1.2086929837161624	0.979356285665783	0.23153540787944477
3	12.641511873464639	12.632590613282765	4.237341555008436	23.59069947851571	23.59069947851571	0.1290781364478731	12.512433737016766	10.406499159591432	2.235012713873206
2	32.67901234567901	32.6327160493827 15	4.190675855829574	697.9351851851852	697.9351851851852	15.709876543209877	16.969135802469136	8.290123456790123	24.38888888888889
0	2.1658136769140492	2.163871307851376	2.743378677940839	4.185184022994637	4.185184022994637	0.008368817916236511	2.157444858997813	1.6702334294581136	0.4955802474559358


Get customer based data
Find optimal k
Build model
Cluster definition
Business suggestion