

Amazon Customer Reviews Dataset

INFO7250 13375 Engineering Big-Data Systems SEC 01 Fall 2018 [BOS-2-TR] (INFO7250.13375.201910)

Author: Neha Jain (001237437)

Data Set: Amazon Customer Reviews Dataset

Source: <https://registry.opendata.aws/amazon-reviews/>

Inspiration:

Amazon Customer Reviews (a.k.a. Product Reviews) is one of Amazon's iconic products. In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. This makes Amazon Customer Reviews a rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others. This data helps to further research in multiple disciplines related to understanding customer product experiences. Specifically, this dataset represents a sample of customer evaluations and opinions, variation in the perception of a product across geographical regions, and promotional intent or bias in reviews.

About the Data:

The dataset contains the customer review text with accompanying metadata, consisting of three major components:

- A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015. This is intended to facilitate study into the properties (and the evolution) of customer reviews potentially including how people evaluate and express their experiences with respect to products at scale. (130M+ customer reviews)
- A collection of reviews about products in multiple languages from different Amazon marketplaces, intended to facilitate analysis of customers' perception of the same products and wider consumer preferences across languages and countries. (200K+ customer reviews in 5 countries)
- A collection of reviews that have been identified as non-compliant with respect to Amazon policies. This is intended to provide a reference dataset for research on detecting promotional or biased reviews. (several thousand customer reviews). This part of the dataset is distributed separately and is available upon request – please contact the email address below if you are interested in obtaining this dataset.

Technologies Used:

- Hadoop Map Reduce Framework
- Tableau Visualization
- HIVE

Future Scope:

- To use Amazon EMR to perform analysis on entire data
- Use complex Hive queries for performing ETL operations
- Use Classification techniques for A better Sentiment Analysis Score

Analysis:

This data consists of six important columns.

- Customer Id
- Product Id
- Rating
- Votes
- Helpful Votes
- Review Body

For this project, I have performed some basic map-reduce using Java & Hadoop Map-Reduce Framework.

My short-term aim from this project was to try and learn how to process and use the part of large data to get a better understanding of the dataset and problem statement.

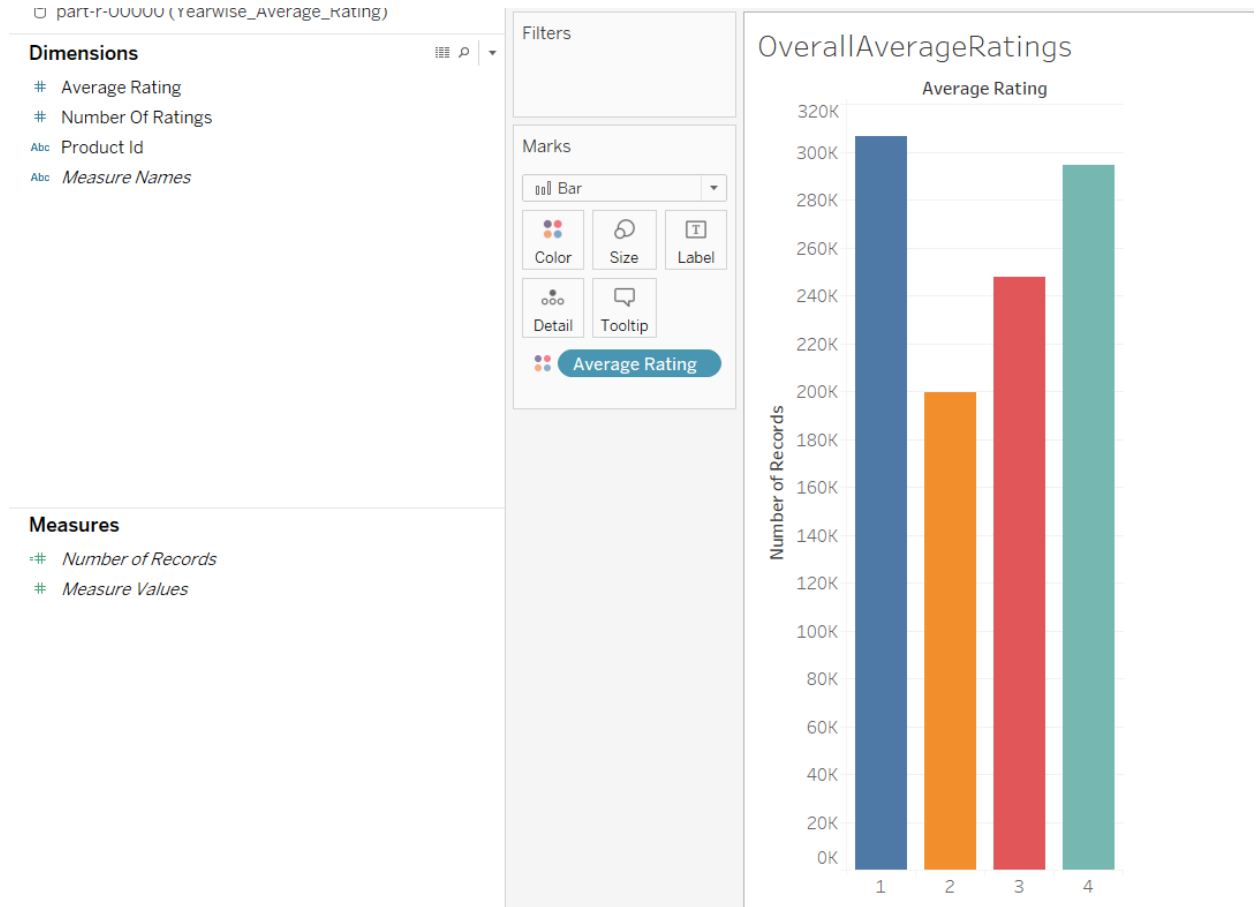
My long-term aim would be to perform some more high-level and complicated analysis and incorporate other available services like *Hive*, *Amazon EMR* for analysis on the entire dataset.

My initial analysis on the dataset are as follows.

1. Calculate the overall rating of all the products

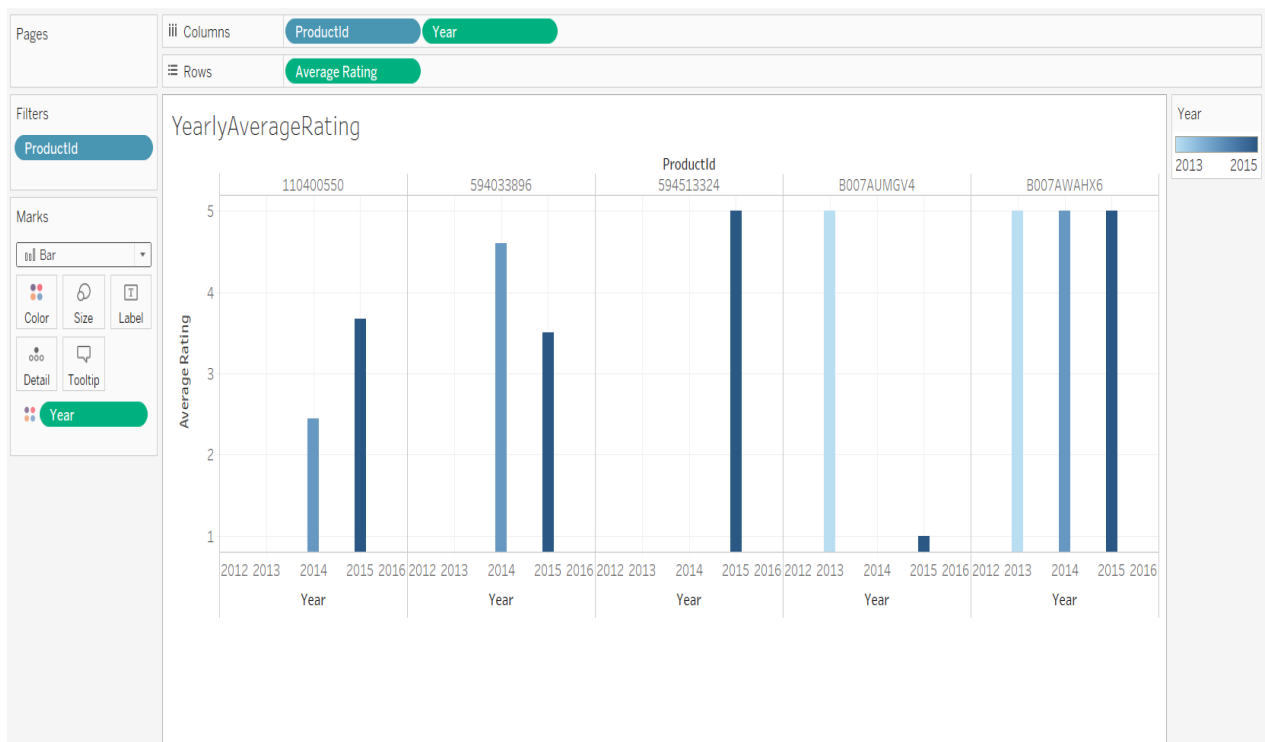
- a. I started with performing a basic map-reduce for calculating the average rating of the products.
- b. Here, Mapper output key was a **Writable class** named ProductWritable and a CountAverageTuple.
- c. This ProductWritable class contains product details and the CountAverageTuple class consists the count and average to be emitted by every mapper.
- d. The Reducer accepts the output from Map and after processing the overall average of a particular product emits the ProductWritable class as Key and CountAverageTuple as Value.
- e. The output of this analysis gives us a high-level idea of how many times a product has been rated and what is its average rating.

- f. Code Link:
- g. Output Link:
- h. Visualization Results:



2. Calculate the Yearly Average Ratings of Products

- It is very important to know the trend. It is one of the most important business factors. Hence, analyzing yearly trends will give us a better picture of how the trend changes or how many products are even liked by the customers year after year.
- To analyze this data, I used the **secondary-sort technique** to order by year of review.
- I used a Composite Key Writable class, a Group Comparator class, Mapper, Reducer and the Secondary Sort Comparator class for performing this analysis.
- The Natural key is productid and the sorting key is the Year.
- Code Link:
- Output Link:
- Visualization:



3. Counting with counters to find Products per Rating Bucket

- One more important analysis would be to understand the number of products which are low rated and number of high rated products.
- Determining the number of low rated products can help business to gain an insight about the customer choices and can help to define different business and sale strategies to make better sales and pickup poorly rated product reviews to understand why the products not doing well.
- For performing the initial analysis on this, I used the **Counting with Counters Summarization Pattern** to put the products into the different ratings buckets and then pick up poorly rated products for further analysis.
- Code Link:
- Output Link:
- Visualization:

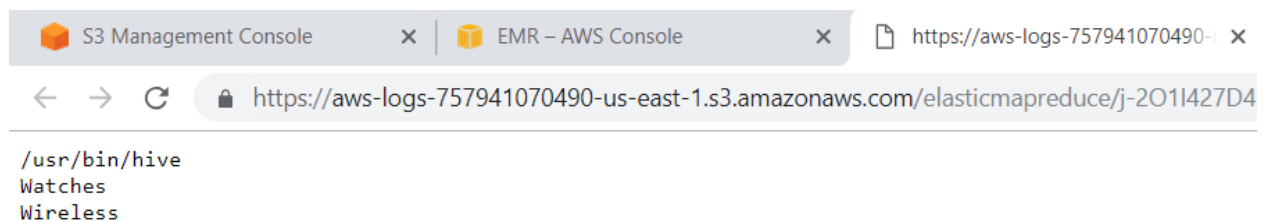
```
Number of Products with Rating 1 = 1262376 products.  
Number of Products with Rating 2 = 598330 products.  
Number of Products with Rating 3 = 815205 products.  
Number of Products with Rating 4 = 1501327 products.  
Number of Products with Rating 5 = 4824783 products.
```

4. Hive Query to get how many distinct category of products

- Since, the data is too large to be processed locally by mapreduce and I want to determine the categories of the products in the dataset, I wrote a Hive Query to extract this information.
- I ran **my Hive Query on AWS EMR** for this processing.

Query Link:





























Output Link:



5. Category-Wise Binning of products









- After getting an overview about the products as a whole, we then can move on to some analysis on specific data.

- b. First step towards that analysis would be to cluster the different category of products for further comparison.
- c. I thought of segregating the products based upon their categories using **the Binning Data Organization Pattern**.
- d. This included a Mapper class, which would categorize the products into the different category bins
- e. Code Link:
- f. Output Link:
- g. Visualization:

 Wireless-m-00001	12/10/2018 2:44 A...	File	1,751 KB
 Wireless-m-00002	12/10/2018 2:44 A...	File	1,743 KB
 Wireless-m-00003	12/10/2018 2:44 A...	File	1,744 KB
 Wireless-m-00004	12/10/2018 2:44 A...	File	1,739 KB
 Wireless-m-00005	12/10/2018 2:44 A...	File	1,711 KB
 Wireless-m-00006	12/10/2018 2:44 A...	File	1,762 KB
 Wireless-m-00007	12/10/2018 2:44 A...	File	1,759 KB
 Wireless-m-00008	12/10/2018 2:44 A...	File	1,773 KB
 Wireless-m-00009	12/10/2018 2:44 A...	File	1,748 KB
 Wireless-m-00010	12/10/2018 2:44 A...	File	1,738 KB
 Wireless-m-00011	12/10/2018 2:44 A...	File	1,766 KB
 Wireless-m-00012	12/10/2018 2:44 A...	File	1,765 KB
 Wireless-m-00013	12/10/2018 2:44 A...	File	1,785 KB
 Wireless-m-00014	12/10/2018 2:44 A...	File	1,768 KB
 Wireless-m-00015	12/10/2018 2:44 A...	File	1,788 KB
 Wireless-m-00016	12/10/2018 2:44 A...	File	1,787 KB
 Wireless-m-00017	12/10/2018 2:44 A...	File	1,809 KB
 Wireless-m-00018	12/10/2018 2:44 A...	File	1,782 KB
 Wireless-m-00019	12/10/2018 2:44 A...	File	1,818 KB
 Wireless-m-00020	12/10/2018 2:44 A...	File	1,763 KB
 Wireless-m-00021	12/10/2018 2:44 A...	File	1,841 KB
 Wireless-m-00022	12/10/2018 2:44 A...	File	1,845 KB
 Wireless-m-00023	12/10/2018 2:44 A...	File	1,862 KB
 Wireless-m-00024	12/10/2018 2:44 A...	File	1,853 KB
 Wireless-m-00025	12/10/2018 2:44 A...	File	1,911 KB
 Wireless-m-00026	12/10/2018 2:44 A...	File	1,913 KB
 Wireless-m-00027	12/10/2018 2:44 A...	File	1,928 KB
 Wireless-m-00028	12/10/2018 2:44 A...	File	1,926 KB

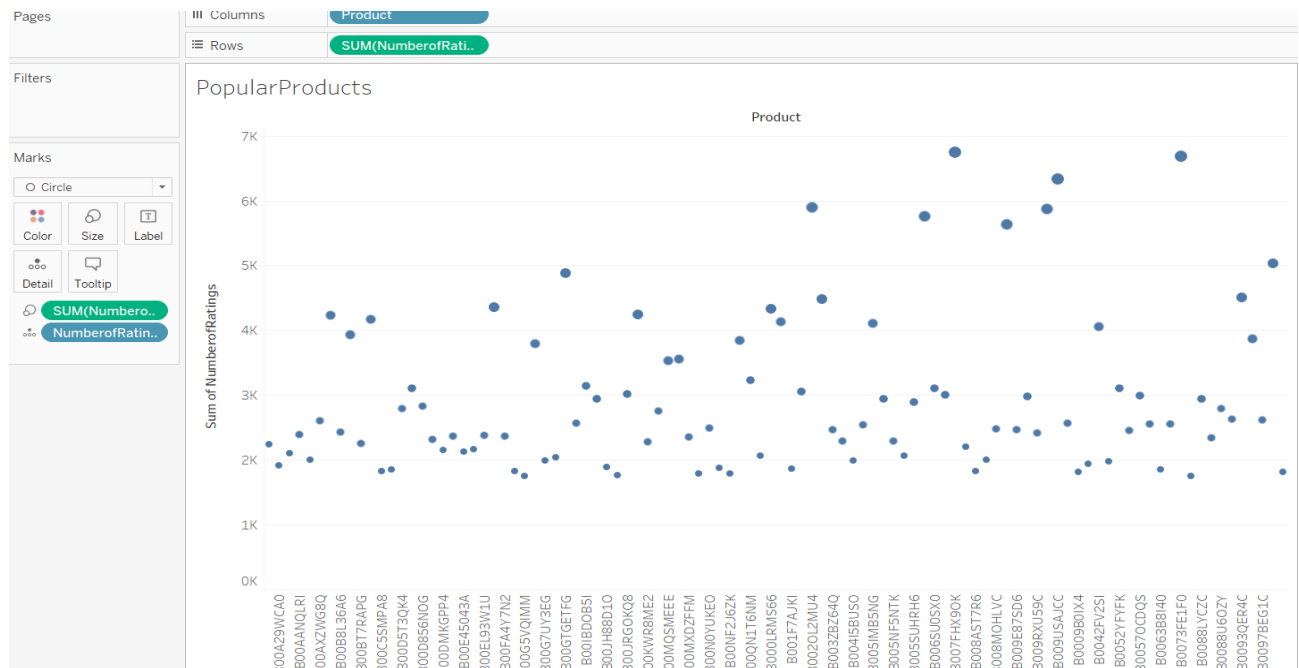
6. Average Rating-wise Binning of Products

- a. Implemented exactly similar to the Category-wise binning.
- b. Just categorized the products according to ratings as well because this output data can be surely useful when we need to analysis like which products to put on sale etc.
- c. Code Link:
- d. Output Link:
- e. Visualization:

<input type="checkbox"/> Name	Date modified	Type	Size
 _SUCCESS	12/10/2018 3:12 A...	File	0 KB
 1.0-m-00000	12/10/2018 3:12 A...	0-M-00000 File	16,498 KB
 2.0-m-00000	12/10/2018 3:12 A...	0-M-00000 File	10,739 KB
 3.0-m-00000	12/10/2018 3:12 A...	0-M-00000 File	5,532 KB
 3.0-m-00001	12/10/2018 3:12 A...	0-M-00001 File	7,786 KB
 4.0-m-00001	12/10/2018 3:12 A...	0-M-00001 File	17,484 KB
 5.0-m-00001	12/10/2018 3:12 A...	0-M-00001 File	7,499 KB
 5.0-m-00002	12/10/2018 3:12 A...	0-M-00002 File	25,099 KB

7. Top 100 Products

- It is very important for us to know the most trending products and most popular products.
- This top 100 product analysis will help us to gain this information.
- I used the **Top 10 Filter Pattern** in Map Reduce to find out the most popular products.
- The popularity of the products is not only based on its average ratings but also how many people have rated it.
- For eg. If there is only one person who has rated it and has given a 5-star, even when the average rating is high, it is not well evaluated by the customers yet.
- Hence, top products are a result of both measures – no of ratings and ofcourse the ratings.
- This implementation includes a Mapper using a sortedmap for storing the top most rated products.
- The mapper emits this list to the code and every mapper emits its top 100 list to the reducer. The reducer then again filters out the top 100 amongst these lists and then emits the top 100 Products
- Code Link:
- Output Link:
- Visualization:



8. Distinct Customers details

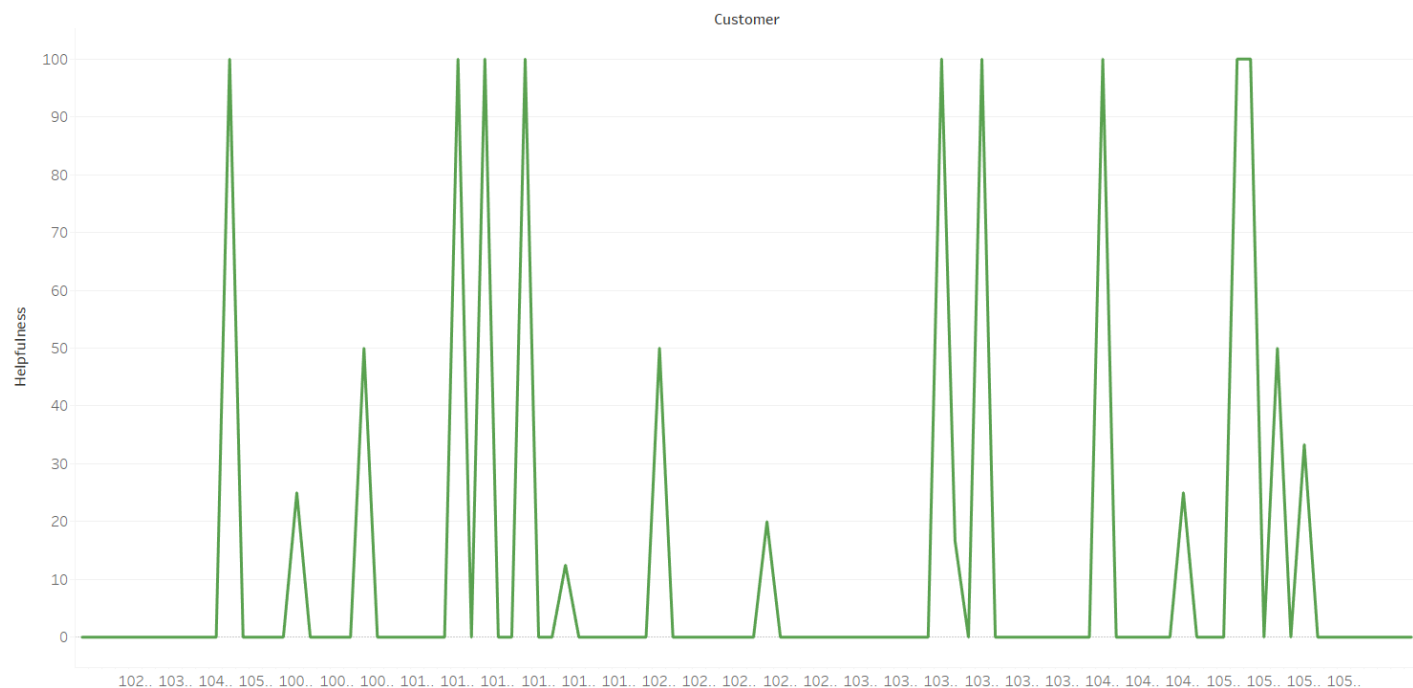
- Now to perform some further analysis, we found the count of customers who have been part of reviews is found.
- I found the distinct customers using the **Distinct Pattern in Mapreduce**.
- This gave me a list of customer ids which can be later used to perform any analysis using JOINS if we have separate dataset for customers
- Code link:
- Output link:
- Visualization:

```
part-uuuuu
1 10000000
2 10000016
3 10000002
4 10000023
5 10000034
6 10000062
7 10000065
8 10000068
9 10000071
10 10000093
11 10000096
12 10000098
13 10000099
14 1000010
15 10000110
16 10000122
17 10000124
18 10000139
19 1000014
20 10000144
21 10000149
22 10000152
23 10000165
24 10000176
25 10000179
26 1000018
```


9. Customer Helpfulness Ratio

- a. This is again an important analysis. It is very important to understand which reviews and ratings from the customers are helpful and which are not. Because, not all of them will be legitimate and can be used for the product analysis.
- b. I have calculated the customer helpfulness ratio by dividing the number of helpful votes/total number of votes.
- c. Customer feedback is very important but Customer's Right Feedback is more important.
- d. Hence, calculating the helpfulness of the customers can help us to categorize customers and use this information in the future when we want to send some special surveys for improvements.
- e. In this implementation, I have used a Mapper, Reducer and a Driver class. The ratio is the number of helpful votes the customer has got divided by the total number of votes he has done.
- f. Code link:
- g. Output Link:
- h. Visualization:

CustomerAnalysis



10. Top 100 customers based upon Helpfulness Ratio

- a. After determining the helpfulness ratio of the customers, I have filtered out the top 100 customers who have high helpfulness ratio.
- b. I did that this so that while performing the sentiment analysis of the reviews, I can use the reviews from these customers as my training dataset. Since, their helpfulness ratio is high, we can assume that their reviews MUST be having high sentiments. Hence, training such data would yield better results.
- c. Code Link:
- d. Output Link:
- e. Visualization:

11. Data Preprocessing

- a. This is the again one of the important parts of my Analysis.
- b. Since it is a reviews dataset, I have to perform a sentiment analysis.
- c. Now sentiment analysis involves a lot of preprocessing like Stopwords removal, stemming, tokenizing etc.
- d. I have used a Java program to eliminate the stopwords and stem the words, remove punctuations and special characters to get a set of clean words as the reviews.
- e. This file can then be input to my Sentiment Analysis Mapreduce program.
- f. Code Link:
- g. Output Link:
- h. Visualization:

1	B007FHX9OK	mount works droid razr clamps great job holding phone tightly don effect volume rockers base ur
2	B007FHX9OK	product designed incredible suction month con person big hands bit problem making adjustments i
3	B007FHX9OK	numerous iphone dashboard mounts ve sticks dashboard strongly touch release lock perfect easy s
4	B007FHX9OK	wife samsung galaxy case phone wife case mount works gs case tight remove phone arms stay open
5	B007FHX9OK	experience product product description oversells stick dashboard days find ground back car made
6	B007FHX9OK	put phone everytime car calling phone hands free perfect price perfect highly recommend windshi
7	B007FHX9OK	love product sits corolla real nice adjust turn screen easy doesn shake doesn drop device thing
8	B007FHX9OK	iottie touch mount primarily hold fire department pager responding calls pager belt coat diffic
9	B007FHX9OK	easy snack difficult plugin strong suck secure fit iphone disadvantage portrait mode block heac
10	B007FHX9OK	ve settled stating ve holds dash surface additional steps easy easy weeks love br br months lov
11	B007FHX9OK	hot humid florida summer product stayed dashboard ve iphone speck candyshe ll card case pretty k
12	B007FHX9OK	good hasn detached holds evo firmly easy position reposition product
13	B007FHX9OK	ve iottie touch galaxy iii phone weeks happy times suction cup stayed stuck day fell time dew c
14	B007FHX9OK	versatility device amazing sticky suction cup stick hardest mount device side radio dashboard v
15	B007FHX9OK	started wave iphone navigation traffic information constantly problem pig battery plugged makes
16	B007FHX9OK	great product bought cars iottie thought design works windshield dash
17	B007FHX9OK	satisfied product works perfectly iphone son galaxy performs
18	B007FHX9OK	good br super easy clamp release br amazing suction cup br great options adjusting tilt angle k
19	B007FHX9OK	sturdy priced dash mount holds fits size phones rotate action works recommend
20	B007FHX9OK	point satisfied arm holds steady position set holds driving rough roads staying firmly dash
21	B007FHX9OK	suction cup holder doesn budge great item hold phone wit extended battery
22	B007FHX9OK	things great mount stuck windshield fear falling damaging phone mp player put ve driven pothole
23	B007FHX9OK	iottie easy flex shake crazy barely stick window mount suction cup stick surface touch works g
24	B007FHX9OK	ordered unit hold iphone driving pro fit brackets things didn phone passenger spot iottie holde
25	B007FHX9OK	received car mount holder works great universal vehicle cup holder adapter suction mount surfac
26	B007FHX9OK	ordered verizon galaxy works mount problems holding case touch basically spring loaded button t
27	B007FHX9OK	mount works ve grips window securely mounting iphone quick secure matters

12. Sentiment Analysis:

- This is the most important analysis which we have to perform on the dataset.
- Sentiment Analysis can be done by many ways from a high level to going deep into NLP techniques.
- For the purpose of this project, I have used a simple sentiment analysis technique to calculate sentiment score of each review. This may not have a very high accuracy rate but can give us a high-level overview of the customer sentiments.
- Sentiment Score is calculated using the existing **Afinn** Library.
- Code link:
- Output Link:
- Visualization:

Average and Median Rating graphs:

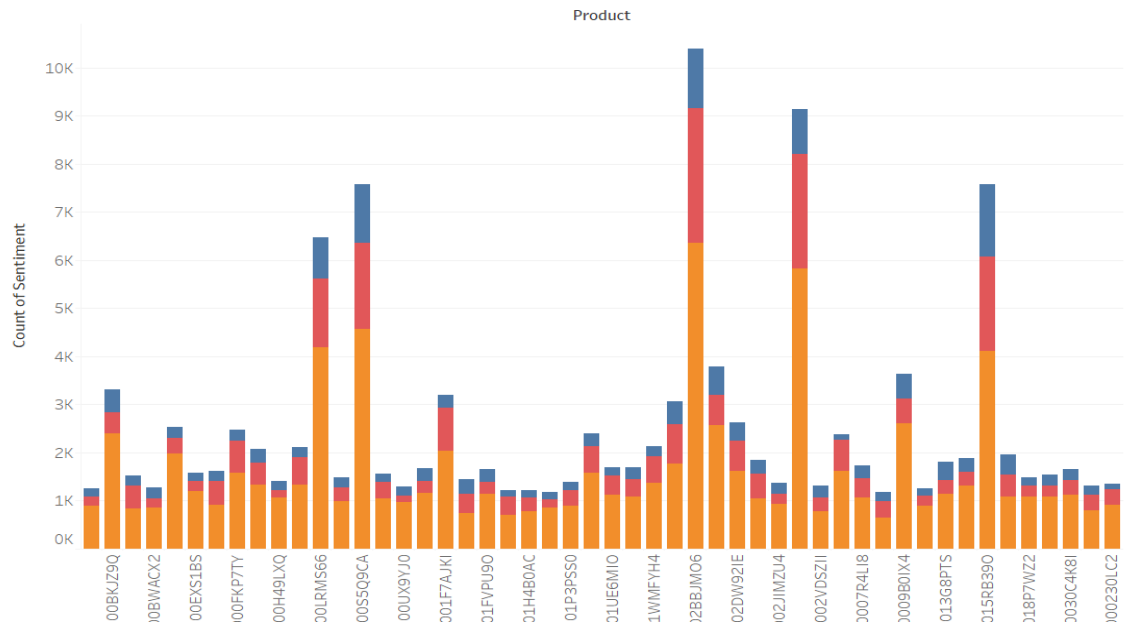
AverageMedianRatingAnalysis



Sentiment Analysis Graph

Columns	Product
Rows	CNT(Sentiment)

SentimentsForTopRatedProducts



13. User Recommendation:

- a. It is mandate that an e-commerce website provides some recommendations. This is a better UX and expected by all customers today.
- b. In this analysis, I have used the output of the previous analysis of sentiment score and rating of users against products to determine which user likes which products and dislikes which products. I have processed this data to be valid input to the **Mahout Recommender**.
- c. I have used the **Mahout User-Recommendation** Framework for determining the recommendations for a particular-user.
- d. Code Link:
- e. Output Link:
- f. Visualization:

```
Customer Id: 2355
No recommendations for this user.
Customer Id: 2356
No recommendations for this user.
Customer Id: 2357
No recommendations for this user.
Customer Id: 2358
No recommendations for this user.
Customer Id: 2359
No recommendations for this user.
Customer Id: 2360
Recommended Item Id 621973. Strength of the preference: 5.000000
Recommended Item Id 340608. Strength of the preference: 5.000000
Recommended Item Id 424505. Strength of the preference: 5.000000
Recommended Item Id 324686. Strength of the preference: 5.000000
Recommended Item Id 613794. Strength of the preference: 4.883039
Customer Id: 2361
No recommendations for this user.
Customer Id: 2362
No recommendations for this user.
Customer Id: 2363
No recommendations for this user.
Customer Id: 2364
No recommendations for this user.
Customer Id: 2365
No recommendations for this user.
Customer Id: 2366
No recommendations for this user.
Customer Id: 2367
```

14. Filter top 5 product's all reviews using Simple Filter:

- a. This is an analysis which is a pre-requisite to my next analysis.
- b. I have used the output from my top 100 product list, to filter out the top 10 products
- c. Then, I have prepared **a filter consisting these 10 products.**
- d. While streaming my input data, I filtered out the reviews only for those products which are in my filter.
- e. This can also be done using **the Replicated Join**. But, I used a normal simple filtering technique for my analysis.
- f. The output of this analysis gave me a file containing the reviews only for top 5 products.
- g. Code Link:
- h. Output Link:
- i. Visualization:

15. Word Count

- a. To get a more idea about the product reviews, one of the analysis is to extract the high frequent words used against the products.
- b. This can even help us to understand which words are adding to the products sentiment score.
- c. **I used the sanitized output for top few products for counting the word frequency for products.**
- d. Code Link
- e. Output Link
- f. Visualization:

Dimensions

- # Frequency
- ABC Product
- ABC Word
- ABC Measure Names

Measures

- # Number of Records
- # Measure Values

Filters

Frequency

Marks

Text

Color Size Text

Detail Tooltip

Word

Word

Frequency

ATTR(Frequen...

Word Cloud

B007FHX9OK

Product

[illegible]

Conclusion:

Discovered various map-reduce techniques and learnt to apply various summarization, organization and filtering patterns.

Better understanding of data by visualization using tableau.