MSDS 694 Distributed Computing

EDA on Amazon Review Data

By: Jih-Chin Chen, Wei He, Zhipeng Hong, Kaihang Zhao

- Data Description
- Analytics Goals and Content
- EMR Performance Discussion
    - Efficiency Improvement
    - Cluster Setting and Execution Time
- Conclusion

Overview of Amazon Review Data

11,095,239 observations in total
5 digital product categories
15 relevant columns
No extra data

Review content and other relevant features, including star rating, vote and verified purchase across five categories; ebook, video games, video downloads, music and software.

Analytical Goals

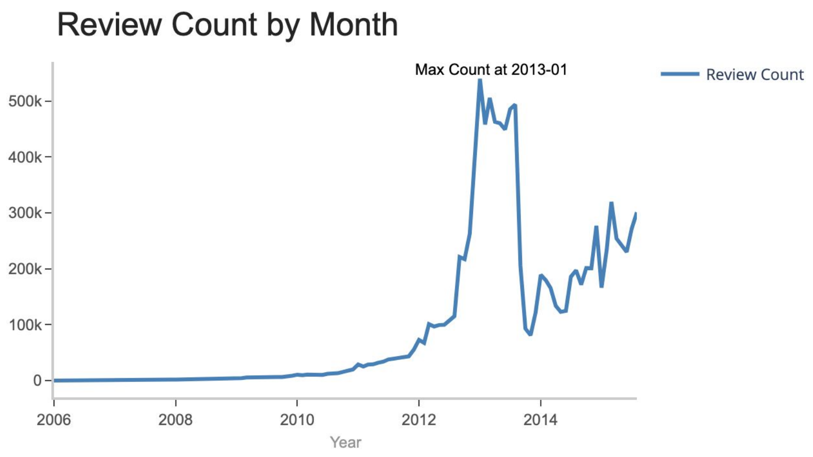
Find Interesting and Meaningful Patterns from different dimensions of Amazon Digital Products Review

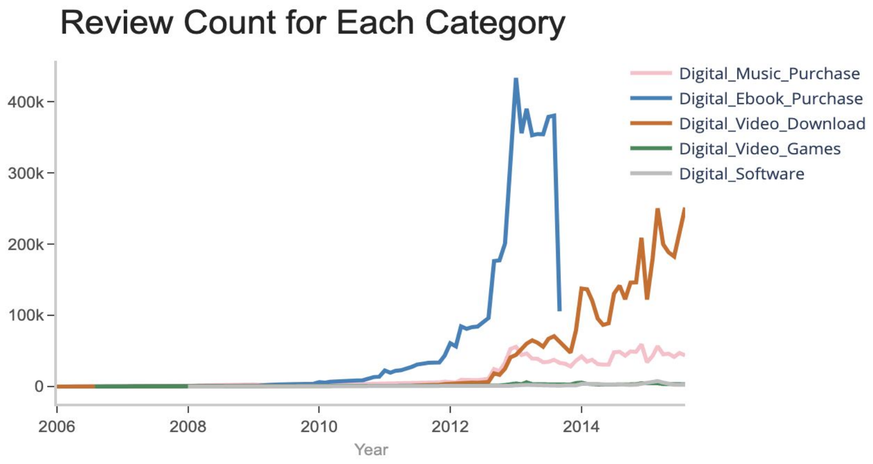
- Jin-Chin Chen: Review Content and Sentimental Score
- Wei He: Review with other features across category level
- Zhipeng Hong: Review with other features across time level
- Kaihang Zhao: Review with other features across time and category level

Our presentation will discuss features one by one, from univariate to multivariate level

Review Count and Star Rating Analysis

Review Count Across Time and Category

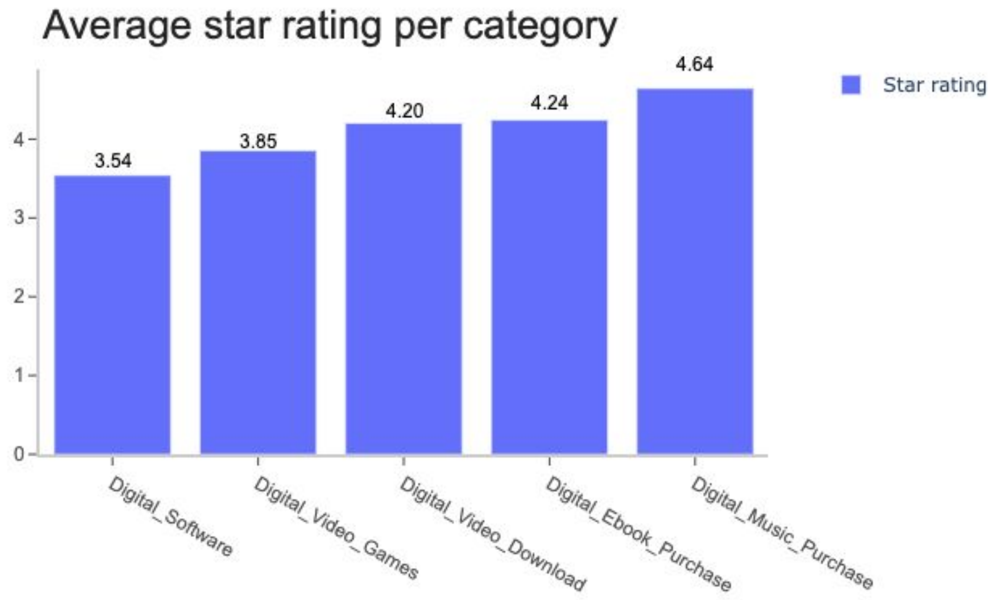




- This plot shows that the number of review increase sharply at from 2012-2013

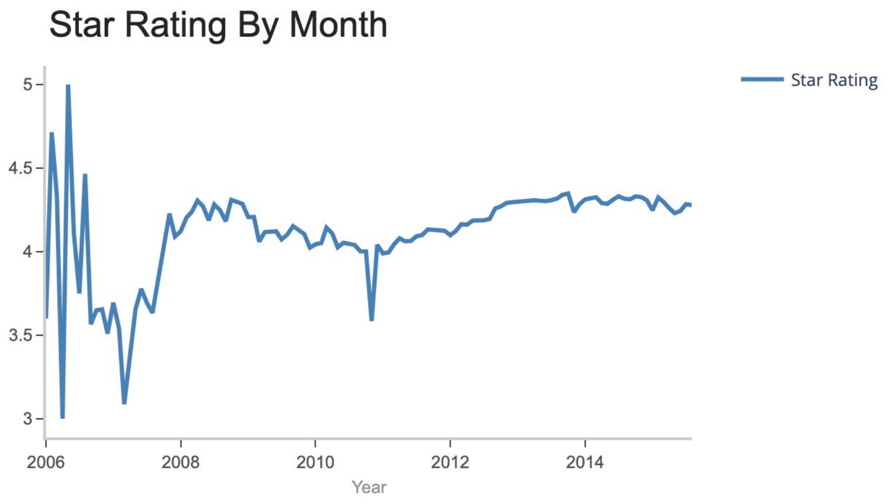
- Boom of Ebook Review
- Video Download Review took over the main trend after 2013

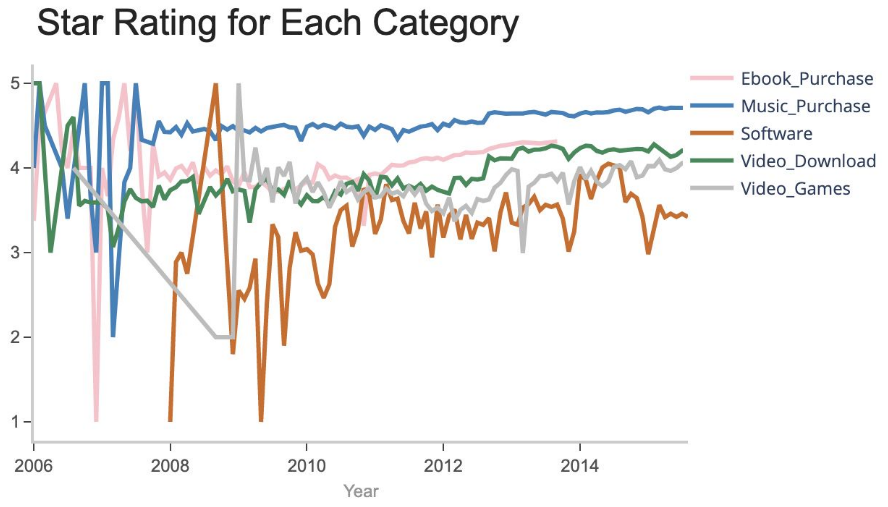
Star Rating



This is plot of star rating for each category. We calculate each average star rating of each category. Music category has highest rating, software has the lowest.

Star Rating Across Time and Category

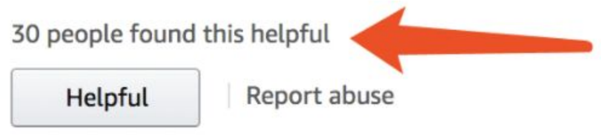




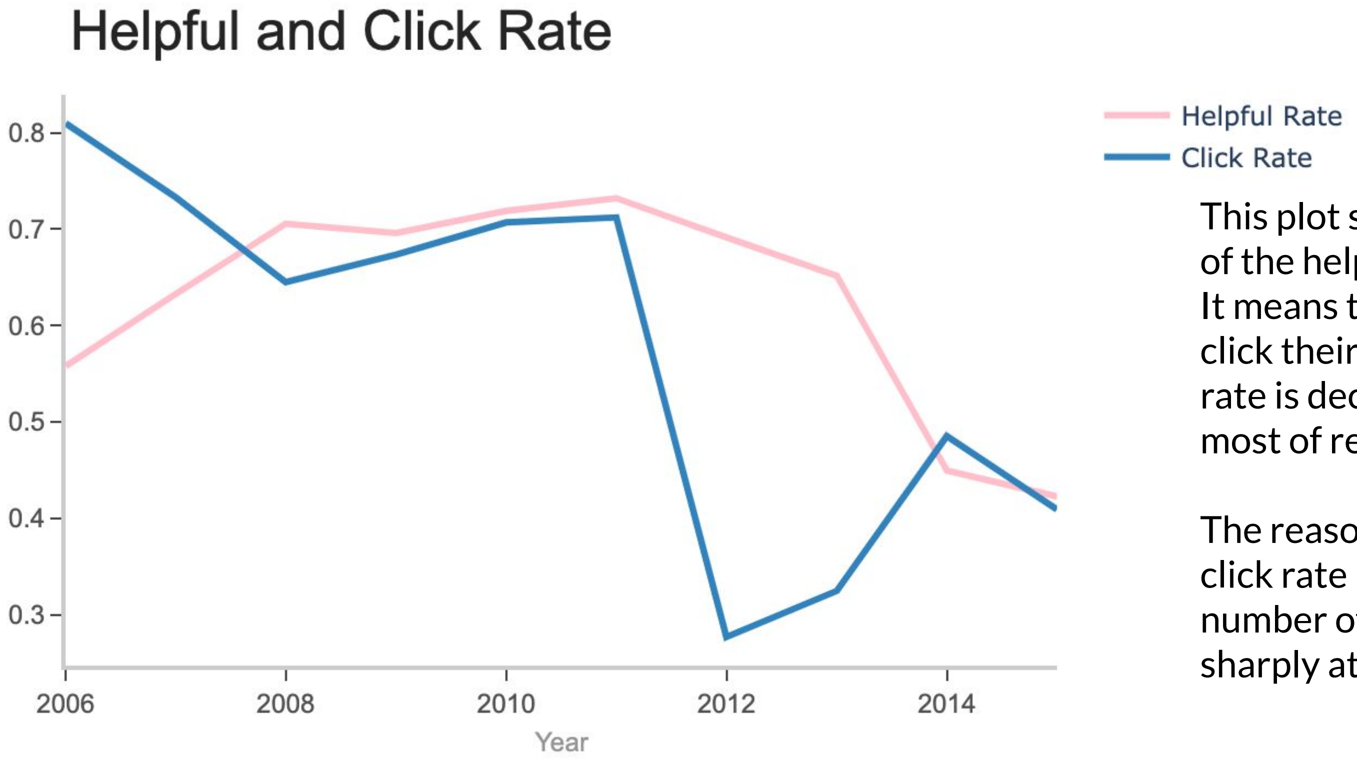
- These two plots show the star rating change by month.
- The sharp change of star rating from 2006 to 2008 is caused by small amount of data. With exponential growth afterwards, star rating is around 4 to 4.5.

Ratio Analysis

Helpful Rate and Vote Rate



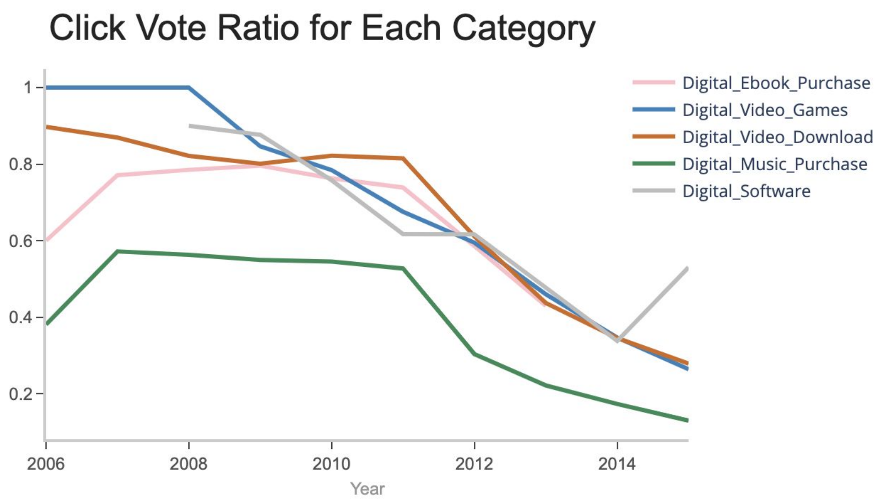
Helpful Vote Rate and Click Vote Rate

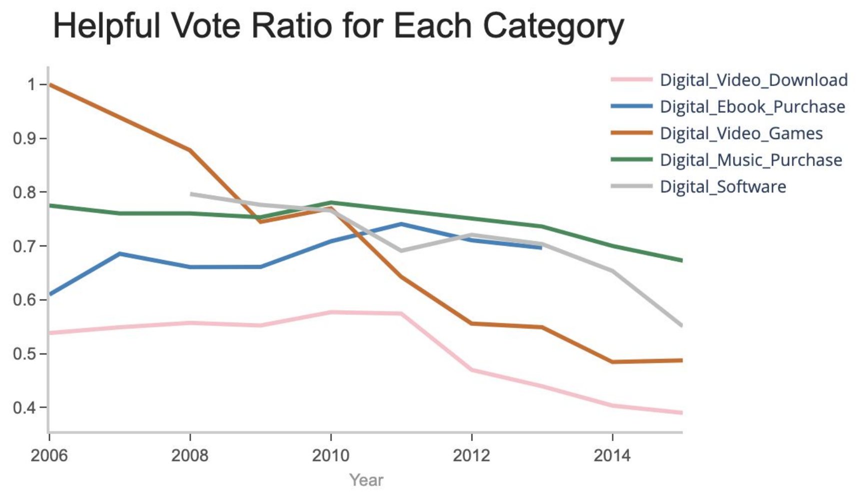


This plot shows that the decrease of the helpful rate and click rate. It means that people unlikely to click their votes. And since helpful rate is decreasing, people think most of review is unhelpful.

The reason of sudden decrease of click rate in 2012 is because the number of review decrease sharply at this time.

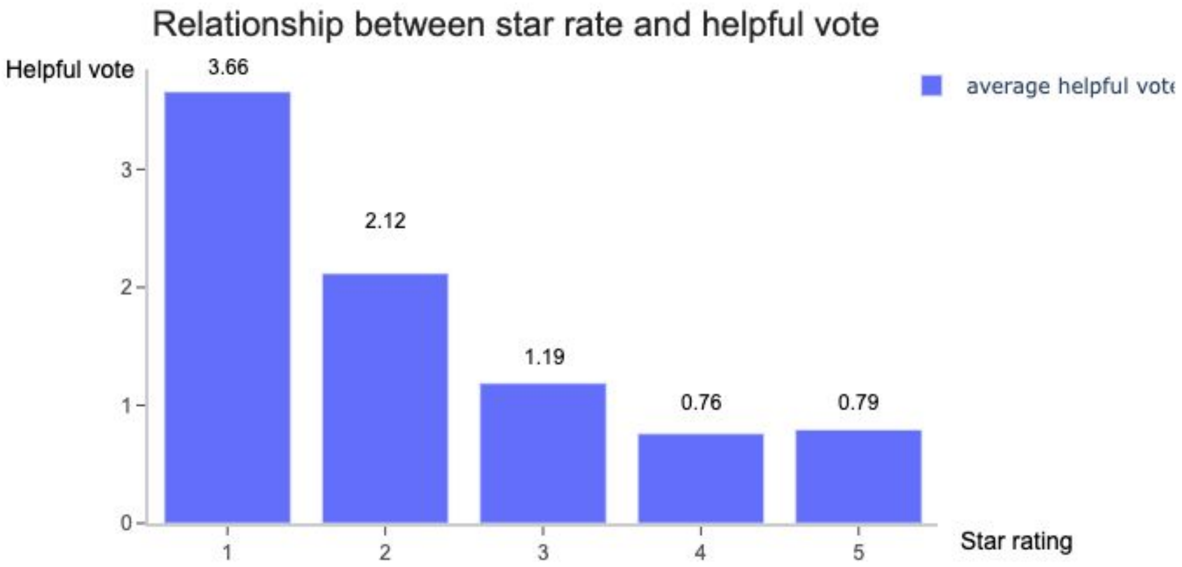
Helpful Vote Rate and Click Vote Rate

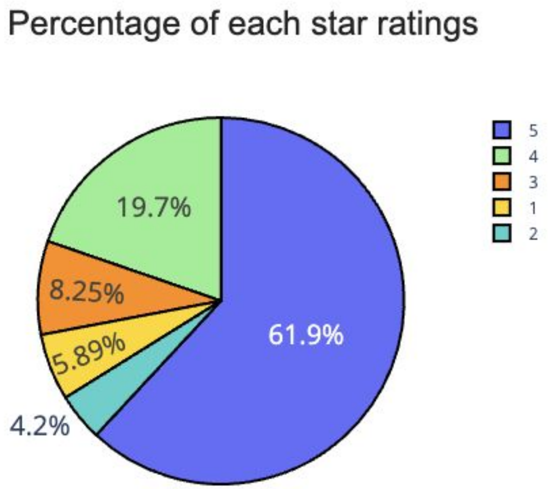




- These two plot shows the click rate and helpful rate between 5 category.
- Helpful Rate and Click Rate is consistent decreasing

Helpful Vote VS Star Rating

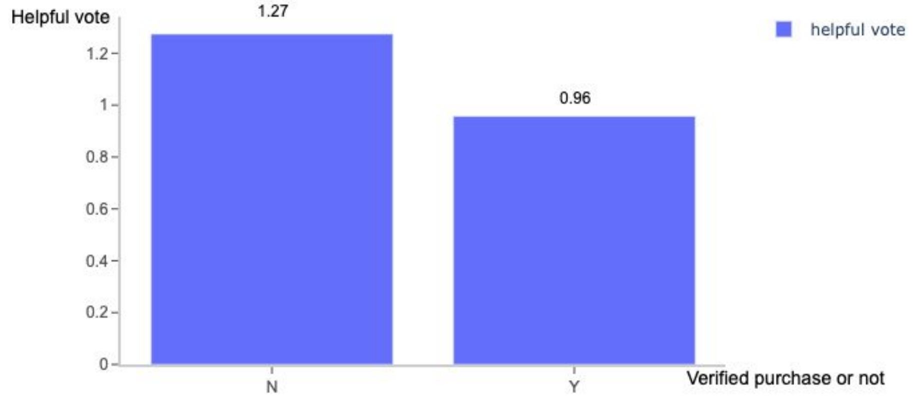




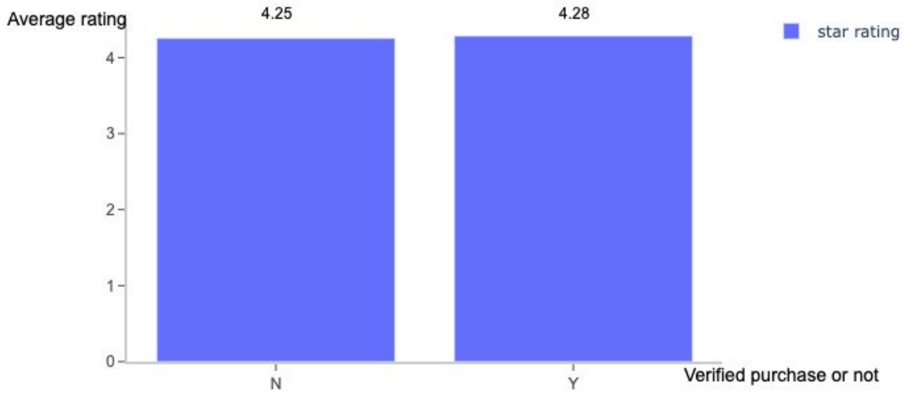
Helpful Vote, Star Rating and Verified Purchase

Amazon Verified Purchase review: Amazon have verified that the person writing the review purchased the product at Amazon and didn't receive the product at a deep discount.

Helpful vote for verified purchase or not



Average star rating for verified purchase or not



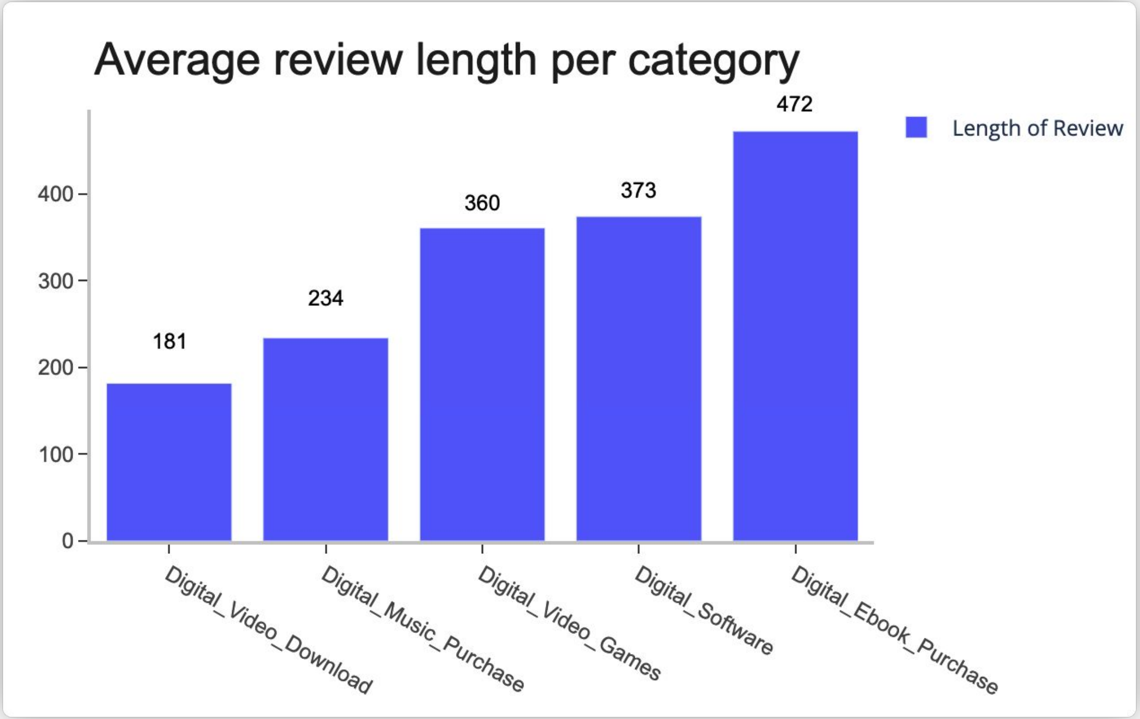
Review Analysis

Review Text

Description:
This plot shows the average length of review across different categories.

Phenomenon:
People who purchase digital ebook are more likely to share their review on product.

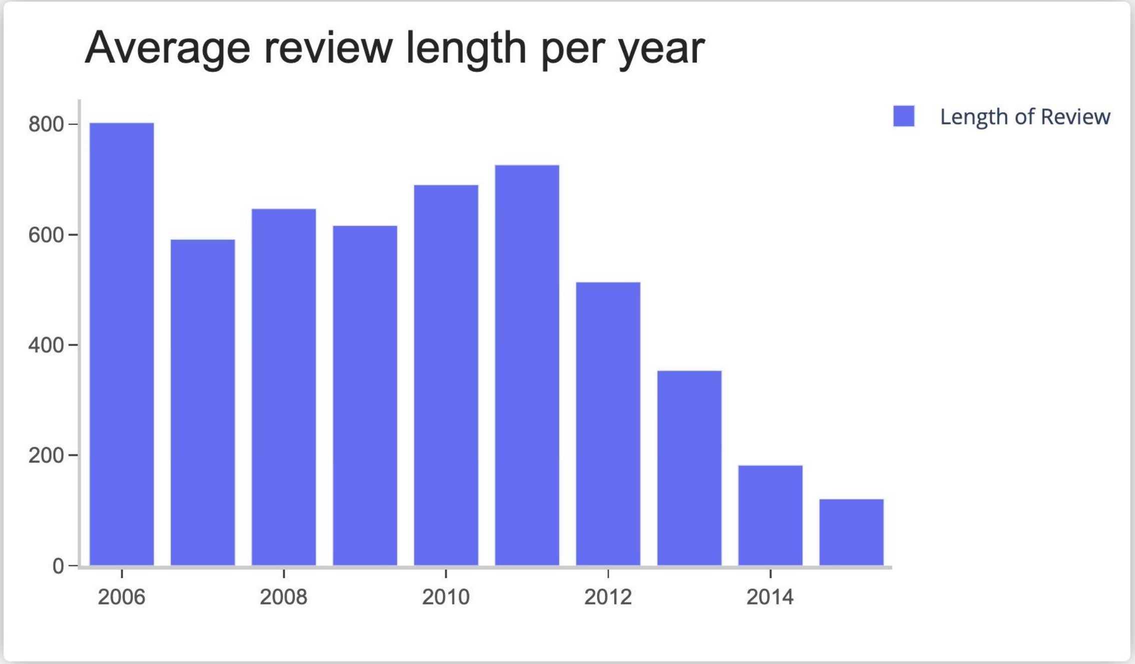
People who purchase digital video are less likely to share their review.



Review Text

Description:
This plot shows the average length of review across different years.

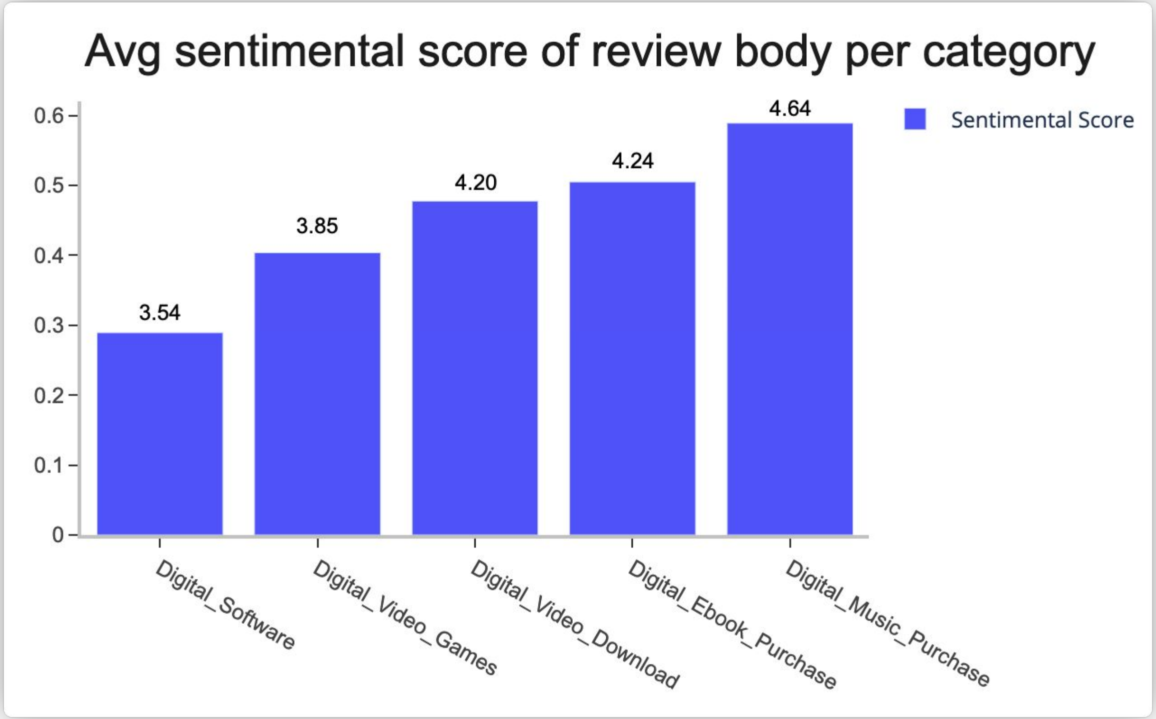
Phenomenon:
There is a obviously decrease trend in this plot after 2011.



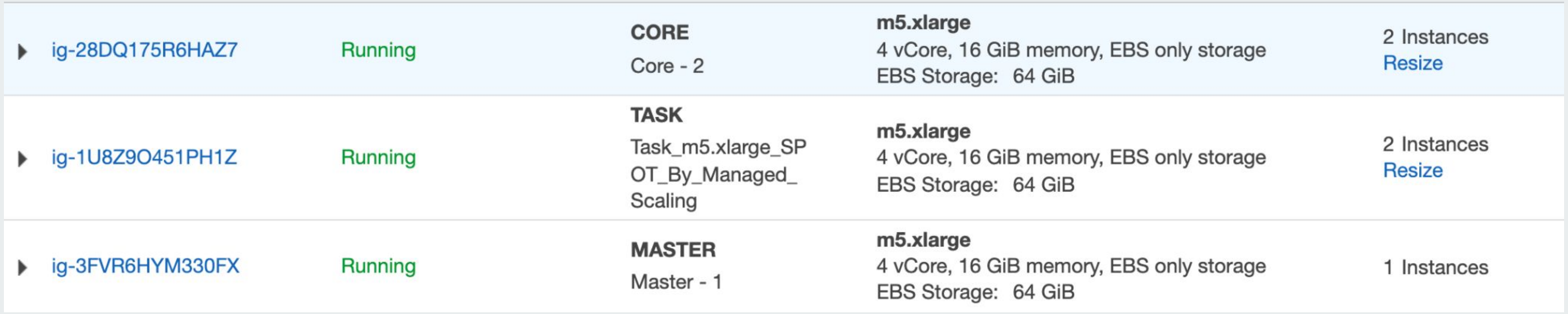
Review Text

Description:
This plot shows the average sentimental score of review across different categories. (the number above the bar is star rating)

Phenomenon:
There is a perfect match between the distribution of sentimental score of review and the distribution of star rating on product.



EMR Performance Discussion



▸	ig-28DQ175R6HAZ7	Running	**CORE** Core - 2	**m5.xlarge** 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB	2 Instances Resize
▸	ig-1U8Z9O451PH1Z	Running	**TASK** Task_m5.xlarge_SPOT_By_Managed_Scaling	**m5.xlarge** 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB	2 Instances Resize
▸	ig-3FVR6HYM330FX	Running	**MASTER** Master - 1	**m5.xlarge** 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB	1 Instances

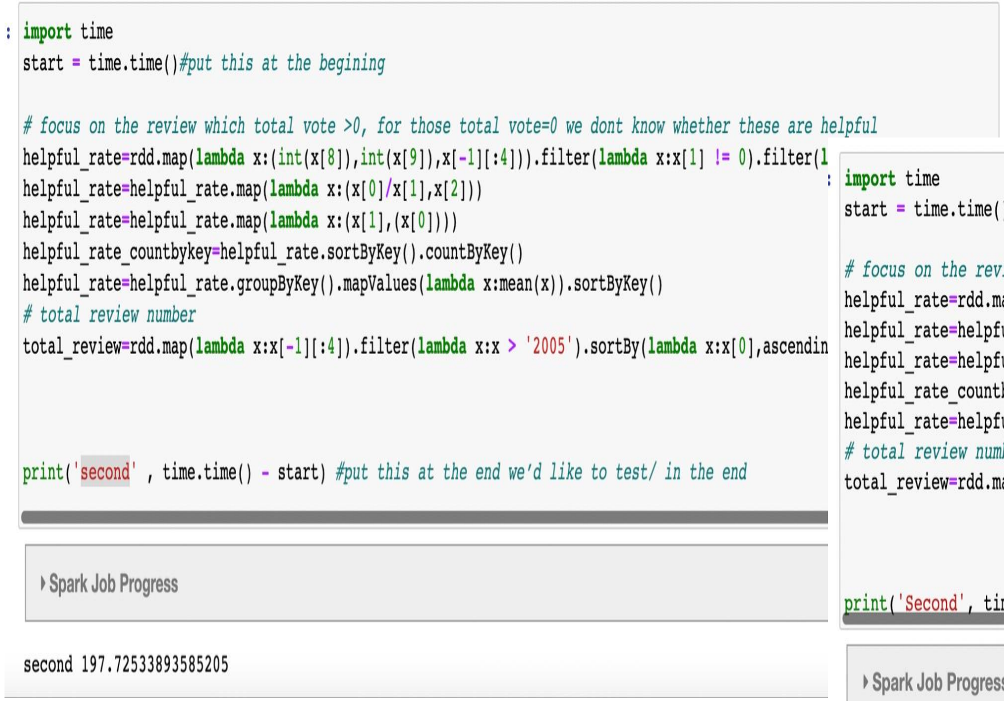
3 instance group in EMR
5 instances m5.xlarge total

EMR Performance Discussion: Efficiency

Use rdd.cache() to persist our initial rdd,

Before 197 seconds; After 82 seconds

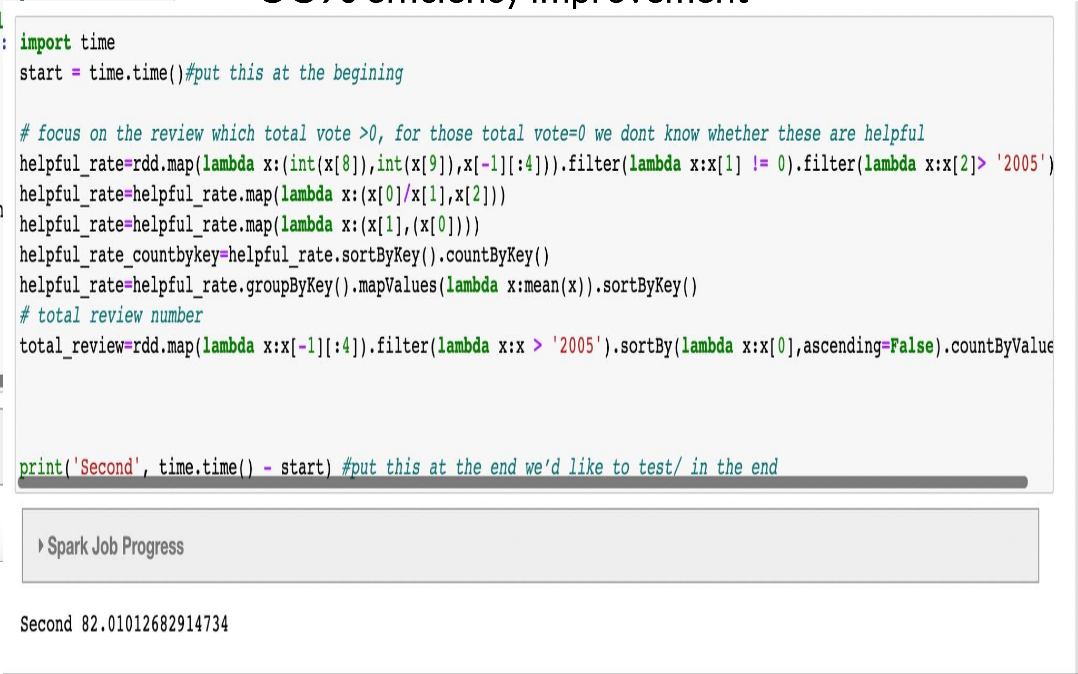
58% efficiency improvement



```
: import time
start = time.time()#put this at the begining

# focus on the review which total vote >0, for those total vote=0 we dont know whether these are helpful
helpful_rate=rdd.map(lambda x:(int(x[8]),int(x[9]),x[-1][:4])).filter(lambda x:x[1] != 0).filter(l
helpful_rate=helpful_rate.map(lambda x:(x[0]/x[1],x[2]))
helpful_rate=helpful_rate.map(lambda x:(x[1],(x[0])))
helpful_rate_countbykey=helpful_rate.sortByKey().countByKey()
helpful_rate=helpful_rate.groupByKey().mapValues(lambda x:mean(x)).sortByKey()
# total review number
total_review=rdd.map(lambda x:x[-1][:4]).filter(lambda x:x > '2005').sortBy(lambda x:x[0],ascendin



print('second' , time.time() - start) #put this at the end we'd like to test/ in the end
```

▸ Spark Job Progress

```
second 197.72533893585205
```

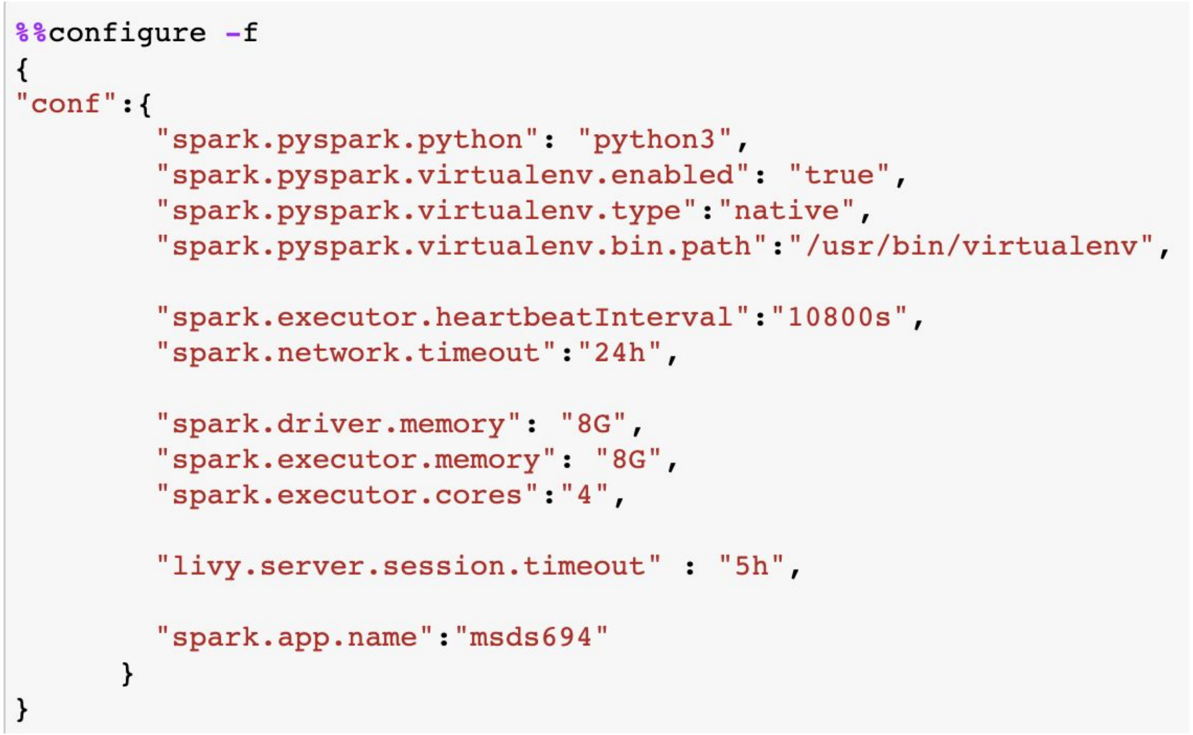


```
: import time
start = time.time()#put this at the begining

# focus on the review which total vote >0, for those total vote=0 we dont know whether these are helpful
helpful_rate=rdd.map(lambda x:(int(x[8]),int(x[9]),x[-1][:4])).filter(lambda x:x[1] != 0).filter(lambda x:x[2]> '2005')
helpful_rate=helpful_rate.map(lambda x:(x[0]/x[1],x[2]))
helpful_rate=helpful_rate.map(lambda x:(x[1],(x[0])))
helpful_rate_countbykey=helpful_rate.sortByKey().countByKey()
helpful_rate=helpful_rate.groupByKey().mapValues(lambda x:mean(x)).sortByKey()
# total review number
total_review=rdd.map(lambda x:x[-1][:4]).filter(lambda x:x > '2005').sortBy(lambda x:x[0],ascending=False).countByValue



print('Second', time.time() - start) #put this at the end we'd like to test/ in the end
```

▸ Spark Job Progress

```
Second 82.01012682914734
```

EMR Performance Discussion: Cluster Setting



```
%%configure -f
{
"conf":{
        "spark.pyspark.python": "python3",
        "spark.pyspark.virtualenv.enabled": "true",
        "spark.pyspark.virtualenv.type":"native",
        "spark.pyspark.virtualenv.bin.path":"/usr/bin/virtualenv",

        "spark.executor.heartbeatInterval":"10800s",
        "spark.network.timeout":"24h",

        "spark.driver.memory": "8G",
        "spark.executor.memory": "8G",
        "spark.executor.cores":"4",

        "livy.server.session.timeout" : "5h",

        "spark.app.name":"msds694"
      }
}
```

Final Execution Time:

Jih-Chin Chen:	63 minutes 30 seconds
Wei He:	2 minutes 1 seconds
Zhipeng Hong:	3 minutes 35 seconds
Kaihang Zhao:	3 minutes 29 seconds

Conclusion

- Review Count: mainly from ebooks and video downloads
- Review Rate: Ebooks highest around 4.5; Software only 3.5
- Helpful rate and Click rate are decreasing.
- Review Length: Start to decrease in 2011
- Helpful rate: People are more caring about the weakness or shortcomes of the products.
- Review Sentimental Score: perfectly matches with the star rating