Analysis of Movie and TV Reviews: A Study of Amazon Data 2005 – 2015

Jasmine Diep, William Lam, Nina Roberts, Taron Sarksyan, Harutyun Sepetjyan

Abstract:

In this paper, we analyzed top Amazon movie and TV reviews rated by Amazon users based on the data collected from Amazon Web Services for the period of 2005 to 2015. The Amazon dataset is 5.33GB in total and consists of three different types of reviews including DVD, digital download, and VHS. Since the raw data gives a numerical score of 1 to 5, we used different technological platforms to create visualizations of the most popular 100 movie and TV titles based on star count ratings and the top 20 movie and TV titles based on total reviews.

Introduction:

During 2020, the world experienced a global pandemic. Many of us were strongly advised to stay home since April except for essential activities. This led to a widespread search for entertainment, including movies and TV shows from Amazon.com. Our group sought to find Amazon titles to explore from five or even ten years ago. What would be a good indicator of quality or high entertainment value for Amazon? We considered which titles had a high average rating for a particular title over the decade. Titles that were popular over the years might also be good content for viewers. So that began our analysis for the top movies and TV titles on Amazon for the years 2005 to 2015. The analysis determined the top 100 titles from average star ratings and top 20 titles from total reviews. It is also important to know how many reviews were submitted from 2005 to 2015 as well. Since this multinational technology platform is the number one in e-commerce, more people are likely to use Amazon Prime to watch movies and TV shows as it offers access to unlimited streaming of thousands of movies and TV episodes.

Related Work:

In order to explain our project more efficiently, we had to take a look at other research projects that had its similarities and differences to ours. One research paper, “Text Mining on Amazon Reviews to Extract Feature Based Feedback” by Sinduja Balasubramanian, explained the process of going through Amazon reviews to know specifically which aspect of the product is lacking customer satisfaction. To ease the process of going through thousands of reviews, the author stated that a system needed to be created to show a statistical report on the reviews that are negative for a product. On our project, the dataset released Amazon Movies and TV shows with low star ratings as well. However, our project placed emphasis on products more with high star ratings. The bigger difference is that the author focused on a variety of products even when the intent was to obtain Amazon reviews. The visualizations created by the author are from Tableau which look almost similar to those that presented later on this paper. The database used for the research is MongoDB compared to this project in which we used the Oracle Cloud to connect to the Hadoop cluster along with Apache Hive.

The second paper, “Ratings vs. Reviews in the Recommender Systems: A Case study on the Amazon Movies Dataset" by Stratigi et al., explains how using recommender systems can discover potential preferences through ratings and reviews. The preferences are emulated by their online behaviors for particular items. One way to show recommendation in our project was to display the most reviewed Movies and TV titles in order to see whether they were in Amazon's best sellers. The recommendation model contains a mathematical formula that is not included in this project. Also, this paper includes a sentiment analysis that is not included on our project as well. The Amazon Movies Reviews Dataset is from 1997-2012, which is five more years than this project (2005-2015). The figures on the research paper place emphasis on the total count of ratings which is very similar to our approach of displaying data as well.

The third paper “A Brief Analysis of Amazon Online Reviews” by Shadi AlZu’bi et al., explains the importance of helpful reviews on Amazon and to display the data about it. Reviews that impact a person's decision are considered to be influential reviews according to the authors. This is similar to our project in which we showed which titles have the most reviews. This would interest people to watch that specific title and influence them to watch other titles within the same genre. The research paper lists the fields like we did that were used to display the data. Some fields were similar to ours, including the helpful votes. However, the authors used all of the fields provided from the dataset they obtained. Also, the research paper focuses on different categories while ours was just one. The charts displayed on the paper focus on the Number of Reviews which is identical to our project.

Data Source:

This dataset is a collection of metadata of reviews of Amazon entertainment products published by Amazon. Users are able to query reviews via Amazon Web Services. The dataset used for this particular project consists of three files of reviews – DVD, digital downloads, and VHS. The time span is from 2005 to 2015. The data is available to the public and Amazon publishes instructions on creating a connection. The instructions can be found here: https://s3.amazonaws.com/amazon-reviews-pds/readme.html.

Using the instructions from Amazon we were able to create a table in Amazon Web Services – Athena, connect to the reviews database and query the most recent dates for reviews (Figure 1). Indeed, 2015 was the last date available for movie and TV reviews (Figure 2).

Figure 1:

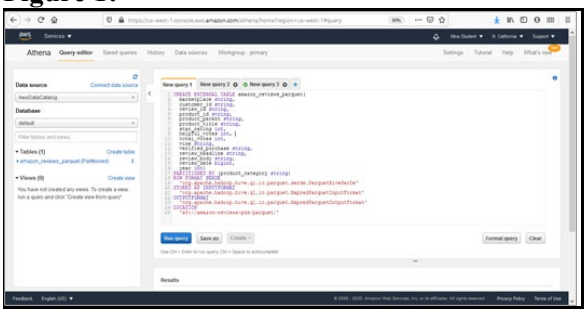
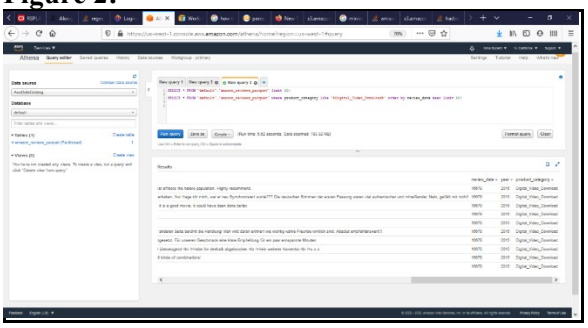


Figure 2:



These are the source data files that were used to begin with the analysis from the link above:

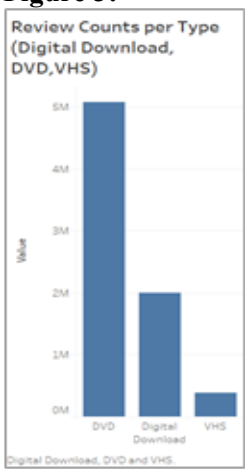
- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Digital_Video_Download_v1_00.tsv.gz
- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_DVD_v1_00.tsv.gz
- https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_v1_00.tsv.gz

Raw Data Review Distribution:

The distribution of reviews consists of (Figure 3):

1. DVD Reviews (1.2GB) - 5,069,149 reviews
2. Digital Downloads (3.7GB) - 2,003,323 reviews
3. VHS (337MB) - 380,604 reviews

Figure 3:



Data Format:

There were 15 fields provided in the Amazon AWS link (See link of first paragraph above). Before creating a table in the database, you would need a marketplace, customer ID, review ID, product ID, product parent, product title, product category DVD, star rating, helpful votes, total votes, verified purchase, review headline, review body, and review data. This would be needed before creating queries needed to show your main results. The example below is what this analysis represents:

marketplace_US (all data is from US market)
customer_id 27288431
review_id R33UPQQUZQEM8
product_id B005T4ND06
product_parent 465011673
product_title Footloose
product_category Digital_Video_Download
star_rating int 5 scale of 1 – 5(best)
helpful_votes 17fraction of users who found the review helpful
total_votes 21
vine Y
verified_purchase N
review_headline "Bravo!!!"

review_body “I love this movie!!!”
review_date 01-15-15

Data Implementation:

Since the data files combined total to 5.33 gigabytes, the team knew that a “Big Data” solution was required. Using Oracle Cloud to connect to a Hadoop cluster was required in order to execute Hadoop and Hive commands. Oracle Cloud was available via our Professor and mentor, Dr. Jongwook Woo, Big Data AI Center (BigDAI): High Performance Information Computing Center (HiPIC) at California State University, Los Angeles. We connected to a node of Oracle BDCE (Big Data Compute Edition) cluster using a terminal and ssh command and performed Hive and Hadoop commands at the terminal level. The team performed a gzip at the terminal to decompress the .gz files and saved the .tsv files produced to HDFS for further implementation.

Once the (3) files were transferred to HDFS (Hadoop Distributed File System), the team merged the files together using the -getmerge file command at the terminal. The command was then made for HIVE to create a table that would contain all the combined data. Please see the laboratory workbook entitled "Lab-Tutorial_Group1_CIS5200.pdf" for specific details on this implementation. With the HIVE table in place, the team was able to ascertain that approximately 284k reviews had been discarded or shed in the implementation phase. This was due to characters that did not match the table schema and were placed in a log for further research. This team did not find any relevant reviews in the error logs.

Data Scrub:

The next step was a discovery session of the data’s characteristics. For instance, “Does a title have one Product_ID or many?” How should we group products together to determine ratings? We found that by querying specific titles, using HIVE QL, the team could identify the standards for the data. In the case of the movie “Footloose,” the team found many product_ids associated with this title and we were able to confirm that this standard was repeated with most other product_titles. The implications were that since multiple product_IDs were associated with a single product_title, it was determined that grouping must be done on the product_title field.

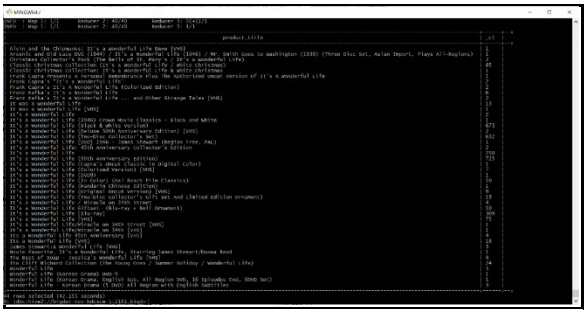
Next, we would have to research the degree of standardization of the product_title field. Our research yielded confirmation that most titles also were amended with phrases such as “[VHS]” or “(Special Edition)”. These phrases would have to be removed to normalize the data. The following is a small list of some of the various phrases that amended titles in the database.

We can see that the product_title field in these two examples has these types of notations which will prevent us from grouping successfully on product_title:

-[Blu-Ray]
-[VHS]
-[Theatrical Release]
-[DVD + Digital Copy + UltraViolet]
-[DVD + Digital]
-[Region 2]
-[Ultra HD]
-[PAL]
-[Blu-ray + Digital Copy]
-(Blu-ray/DVD Combo)

Figure 4: Example 1 – Title = “Footloose”

Figure 5: Example 2 – Title = “It’s A Wonderful Life”



Other assumptions could be made as well after our research. For the most part, TV listings do not seem problematic for popular titles. For movies between 2005 and 2015, Amazon saved the movie year in the product_title field. We discovered that if you paste the product_id into Amazon.com it will link you to the title. Also, the following would prevent accurate grouping.

- product_title contains both “A” and “a”
- product_title contains both “The” and “the”
- product_title contains both “Its” and “It’s”

The group decided on a systemic solution to solve these issues. A regexp_replace command would be used to replace the parentheses and its content in the product_title field. Similarly, a regexp_replace command would replace the brackets and those contents in the product_title field. The code is below:

```
SELECT product_title =
regexp_replace(product_title,'\s*\[[^()]*\]', '')
FROM amazon;

SELECT product_title =
regexp_replace(product_title,'\s*\([^()]*\)', '')
FROM amazon;
```

There will be some titles that will erroneously be grouped together. This was the pitfalls of the method. These examples were identified and manually fixed:

- Little Women
- Midway

Data Analysis Methodology:

The original methodology was to query in HIVE QL for any title that had a 5.0 on the scale of 1-5. We would order this data by average star rating and review count. This methodology yielded a list of unfamiliar titles.

The next methodology we settled on was to set a requirement of a certain threshold of review count. We noticed that there was a correlation to recognizable titles once we set the requirement to 3,000 reviews. At 3,000 reviews and above a 4.5 average rating over all reviews during 2005 to 2015, we yielded a list of entertainment titles that we felt were representative of “top titles”.

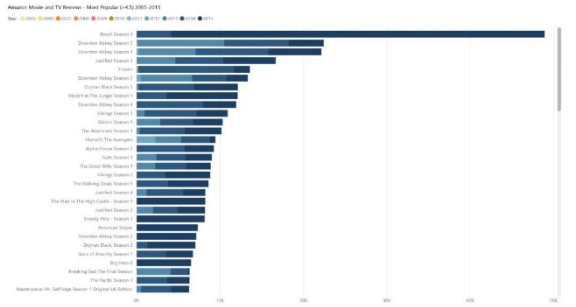
Figure 6: Top 100 Using Average Star Ratings and Review Count

Product_title	Average Star_rating	Review Count
Bosch Season 1	4.61	47,940
Downton Abbey Season 3	4.87	22,457
Downton Abbey Season 1	4.86	22,155
Justified Season 1	4.68	16,698
Frozen	4.67	12,697
Downton Abbey Season 2	4.89	13,356
Orphan Black Season 1	4.68	12,149
Mozart in The Jungle Season 1	4.51	12,116
Vickings Season 1	4.68	10,926
Grimm Season 1	4.65	10,322

Visualizations:

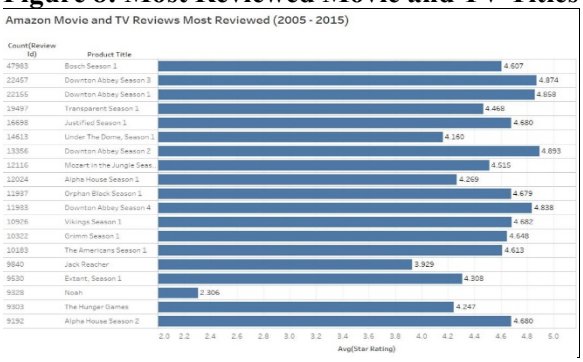
The top 100 titles were uploaded and created using Power BI. Bosch Season 1 had the most amount of reviews. Rome season 2 had the lowest amount of reviews (Figure 7).

Figure 7: Top 100 Titles Based on Review Count (First screenshot of top 100)



The most reviewed movie and TV titles are represented as the top 20 for Figure 8. Each title shows its average star rating from 2005-2015. The title that stands out is Noah with a very low average rating of 2.3 (Figure 8).

Figure 8: Most Reviewed Movie and TV Titles



The charts below were made in Tableau to show a detailed comparison of the top 20 titles from 2009-2015. What makes these charts more detailed compared to Figure 7 is by understanding which title was successful in a specific year. Charts on Tableau are best to be simplified for the top 10 results, since it is featured in the filter option when you modify an enabled field. Hence why the two charts below are split apart for a more simplified view of the results. For Figure 9, Bosch Season 1 had the highest number of reviews in 2015. For Figure 10, The Man in the High Castle Season 1 had the highest number of reviews in 2015.

Figure 9: Top 10 Titles Based on Review Count

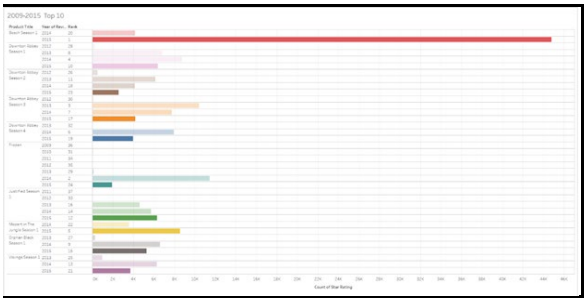
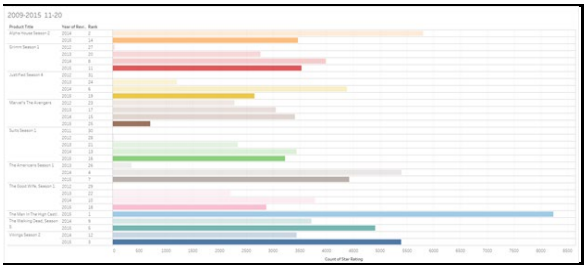


Figure 10: Top 11-20 Titles Based on Review Count



Quality Assurance:

Titles were queried from Amazon.com's "Best Sellers in Movie and TV" listing from 2020.
We validated some of our popular titles from 2005 to 2015 were present (Figure 11):

- Mean Girls
- Star Trek
- National Lampoon's Christmas Vacation
- A Christmas Carol
- A Christmas Story

Figure 11:

Conclusion:

In summary, based on the datasets collected from Amazon Web Services for the period of 2005 to 2015, we were able to analyze the top Amazon movie and TV reviews based on the highest ratings provided by Amazon users. We also created visualizations in Power BI and Tableau, which were interactive tools that generated tables and charts for us to visually analyze movies, and shows to see their overall ratings. Based on our extensive analysis of the Amazon AWS datasets, we have concluded that Bosch Season 1 is recommended from the 'Top 100 Reviews', and for 'Most Reviewed Movie and TV titles', Downton Abbey Season 2 is the most reviewed based on the average.

References:

- AlZu'bi S., Alsmadiv A., AlQatawneh S., Al-Ayyoub M., Hawashin B., and Jararweh Y. (October, 2019) A Brief Analysis of Amazon Online Reviews. Retrieved from https://ieeexplore.ieee.org/document/8931816
- Balasubramanian S. (December, 2017) Text Mining On Amazon Reviews To Extract Feature Based Feedback. Retrieved from http://dspace.calstate.edu/bitstream/handle/10211.3/198628/sinduja_project_report_readers_approved.pdf?sequence=1
- Stratigi M., Li X., Stefanidis K., Zhang Z. (September, 2019) Ratings vs. Reviews in Recommender Systems: A Case Study on the Amazon Movies Dataset. Retrieved from https://doi.org/10.1007/978-3-030-30278-8_9