MovieLens Dataset Analysis With Spark

JOSEPH MOHR, UW - Madison, USA

There are many factors that contribute to the success of a movie or how well a person will like some movie. A few possible ideas will be looked into here, utilising the MovieLens Dataset along with Spark to help handle the amount of data contained across the files of this dataset. Other trends will also be examined, as well as a general analysis of the dataset.

ACM Reference Format:
Joseph Mohr. 2021. MovieLens Dataset Analysis With Spark. 1, 1 (May 2021), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This project dives into both the ratings and tags of movies as decided by users on MovieLens. The dataset features over 25 million ratings and over 1 million taggings. The data is spread across six files. Different facets of this dataset will be explored in this report, mainly focusing on how different genres are rated over time and on various aspects of tags.

I believe there may be interesting trends in how people enjoy different genres of movies, how much they enjoy movies at different times of the year, and how they perceive movies of different genres. This dataset provides a good starting point to investigate whether these trends realistically exist or not. While this few of features and rows is not enough to say anything conclusive, it can certainly give an idea of if these trends exist and if they're worth investigating further.

Some of these trends are things that movie companies could further look into to potentially help with getting better ratings for their own movies. While the scope of the data looked at in this analysis is rather narrow, expanding it to include more features and even more rows would help to get a better idea of how real these trends are and whether or not they may be able to be utilized by these companies.

The next two sections will introduce the tools used and then the dataset. It will make use of Spark's Python API, PySpark, to help perform this analysis. The graphs will be generated using Matplobtlib. Section 4 will go through the results followed by the conclusion.

The code can be found at;

```
https://github.com/jvmohr/cs784_final
```

.

Author's address: Joseph Mohr, UW - Madison, Madison, USA, jvmohr@wisc.edu.

XXXX-XXXX/2021/5-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 TOOLS USED

2.1 Spark

Apache Spark is an analytics engine focused on big data. It allows for easy intertwining of SQL, DataFrames, and more.

It has a high-level API for four different languages: Java, Python, R, and Scala. This project utilizes the Python API, PySpark. This API allows a Spark DataFrame to be easily converted into a Pandas DataFrame, which allows for great compatibility with the vast ecosystem for data analysis that Python has. In addition to this, user-defined functions through Pandas are also easy to create and are usable with Spark SQL.

2.2 Matplotlib

Matplotlib was used to create all of the graphs present in this paper. It generally works with Pandas DataFrames, so being able to convert was very helpful.

While there is a plotting library for PySpark, I found it harder to get used than I would have liked so I opted for what I'm very familiar with.

2.3 Pandas

I mainly used Pandas for two things in this analysis: to help with plotting and for something to compare PySpark's performance with.

2.4 Other

I used Jupyter Notebook as the environment I coded in. I used scikit-learn to help with one piece of the analysis (as well as one function from NumPy for that part). I also used the datetime module for Python to get the month out of the date for the Pandas part of the comparison.

3 MOVIELENS DATASET

In this project, the MovieLens dataset, specifically, the MovieLens 25M Dataset, is analyzed. [Harper and Konstan 2015] This version was released on November 21, 2019. It contains information from the MovieLens service, such as user's rating of movies, tags that they've applied to movies, and tag relevance scores. This information is spread across six different comma-separated values (csv) files, as noted in each of the following subsections. The data is from between January 09, 1995 and November 21, 2019, although the ratings in 1995 are generally disregarded in this analysis due to how few there are.

The scores for the tag genome were created using a machine learning algorithm. The tags and movies included in this are a subset of each. [Vig et al. 2012]

3.1 movies.csv

This file contains information about each movie: it's id, title, and genres. The genres are all in one column where each genre for a movie is separated by a vertical bar. The year the movie was released

is usually in parenthesis at the end of the title column, however, it is not consistently there. There is information about 62,423 movies.

3.2 ratings.csv

This file contains information about each time a user rated a movie. Each row contains the user id of the user responsible for this rating, the movie id of the movie they rated, the rating (out of 0.5 to 5.0 in 0.5 increments), and the timestamp of when the rating was made (the number of seconds since the Unix epoch). There are 25,000,095 rows in this file.

Each user included in this dataset rated at least twenty movies and the average number of ratings is over 153.

3.3 links.csv

This file contains more information about each movie. Each row contains a movie id, the id of the movie used by IMDB, and the id of the movie used by TMDB. There is a row for each movie in movies.csv.

This is the only file that wasn't used in the analysis portion.

3.4 tags.csv

This file contains information about tags applied to any movie. Each row contains the user id of the user that applied the tag to a movie, the movie id of said movie, the tag that was applied, and the timestamp of when the tag was applied. There are 1,093,360 rows in this file, one per tagging. 73,050 different tags are present in this file.

3.5 genome-tags.csv

This file contains information about each tag. Each row contains a tag id and the actual tag. The tag id is only shared with genome-scores.csv. There are 1128 rows in this file, one for each tag included in the scoring in genome-scores.csv.

3.6 genome-scores.csv

This file contains information about the relevance of each of the 1128 tags in genome-tags.csv for 13,816 movies. Each row contains a movie id, a tag id, and a relevance (from 0 to 1) of how relevant that tag is to that movie. There are 15,584,448 rows in this file, as each of the 13,816 movies has a row for each of the 1128 tags.

4 ANALYSIS RESULTS

In this section I will present the results of my analysis, ranging from table-based to graphs. This was mostly done using Spark SQL along with Pandas at the end to allow for use of Matplotlib.

4.1 Top Rated

As a simple start, I'll quickly go through the top rated movies, genres, and tags.

4.1.1 Movies. Table 1 shows the top 10 rated movies, which contains a lot of movies that aren't surprising to see there, like the nature documentaries. Obviously a movie like "Obsession" is a standout due to the low number of ratings, but I think it's good to include to help show the variance among how many times each movie has been rated.

Table 1. Top 10 Movies by Average Rating

Rank	Title	Avg Rating	# Ratings
1	Planet Earth II (2016)	4.483	1124
2	Planet Earth (2006)	4.465	1747
3	Shawshank Redemption, The (1994)	4.414	81482
4	Band of Brothers (2001)	4.399	1356
5	Cosmos	4.327	277
6	Godfather, The (1972)	4.324	52498
7	Blue Planet II (2017)	4.290	659
8	Usual Suspects, The (1995)	4.284	55366
9	Obsession (1965)	4.278	36
10	Twin Peaks (1989)	4.267	288

Table 2. Genres Ranked by Average Rating

Rank	Genre	Avg Rating	Number of Ratings
1	Film-Noir	3.927	247227
2	War	3.791	1267346
3	Documentary	3.705	322449
4	Crime	3.685	4190259
5	Drama	3.677	10962833
6	Mystery	3.670	2010995
7	Animation	3.615	1630987
8	IMAX	3.604	1063279
9	Western	3.586	483731
10	Musical	3.555	964252
11	Romance	3.543	4497291
12	Thriller	3.523	6763272
13	Adventure	3.517	5832424
14	Fantasy	3.512	2831585
15	Sci-Fi	3.478	4325740
16	Action	3.467	7446918
17	Children	3.433	2124258
18	Comedy	3.424	8926230
19	(no genres listed)	3.326	26627
20	Horror	3.294	1892183

4.1.2 Genres. Table 2 has the genres ranked by average rating. This is computed by taking the average of every rating that rates a movie that has that genre.

Also of note is the "Number of Ratings" column. There does seem to be a decent correlation between average rating and number of ratings as most of the 10 genres that have the most ratings are in the bottom 10 for average rating.

Already here it is possible to see some groups of genres, such as "Crime" and "Drama" being close together, along with "Sci-Fi" and "Action." We'll explore this more later on.

4.1.3 Tags. Table 3 contains the tags with the highest average rating (that have been tagged at least 20 times) when the user tags and rated the same movie. We will take a better look into best rated tags later, This section is mainly here for symmetry with the top

Table 3. Top 5 Tags by Average Rating

Rank	Tag	Avg Rating	Number of Ratings
1	FAVORITE	4.970	33
2	delights	4.967	30
3	rewatch worthy	4.900	45
4	best movie ever	4.897	34
5	5 stars	4.864	22

rated movies and genres and to show that the fifth highest average rated tag in this table, "5 stars," does not have a 5.0 average rating.

4.2 Rating Distribution

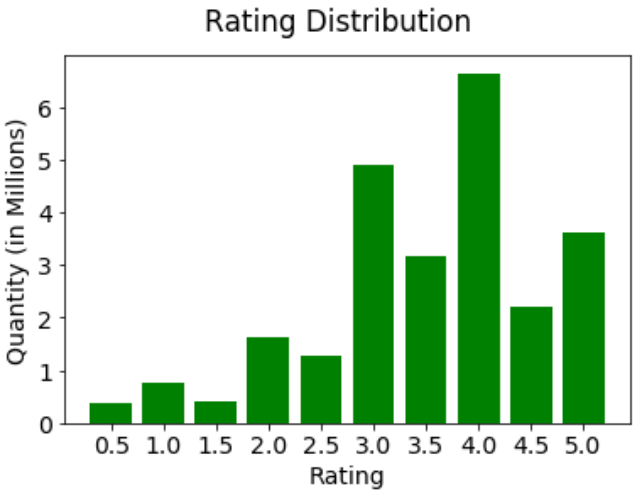

Fig. 1. The distribution of all the ratings in the dataset

The user-created ratings are one of the central pieces of this dataset, so they're worth looking at from a broad level at first. Figure 1 shows the distribution of all the just over twenty-five million ratings. Better ratings do seem to be more popular with a 4 being the most popular rating. Also of interest is the differences in the bars of whole numbers and those in-between. People are apparently more inclined to give a movie an integer rating rather than a half-way point in-between.

The average rating is just over 3.53.

4.3 Average Rating Per Genre By Year

Figure 2 shows the average rating each genre received. This was simply done by taking the average of each rating applied to a movie in each genre. As the genres listed for each movie were all in the same column, this proved to be trickier than originally thought. While probably not the best way, I eventually settled on expanding the movies.csv such that each movie has a row for each of its genres. So if a movie has four genres, it has four rows where the only difference is the genre. This became a new csv file that then became a new table.

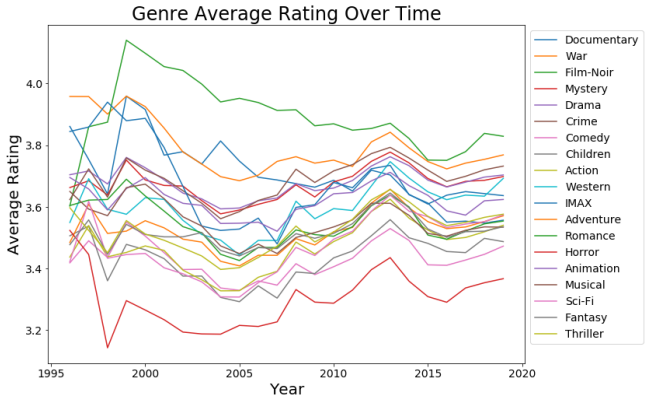

Fig. 2. The average rating of each genre

As a lot of the genres seem to be clustered together in the plot, I decided to try to actually cluster them together using KMeans.

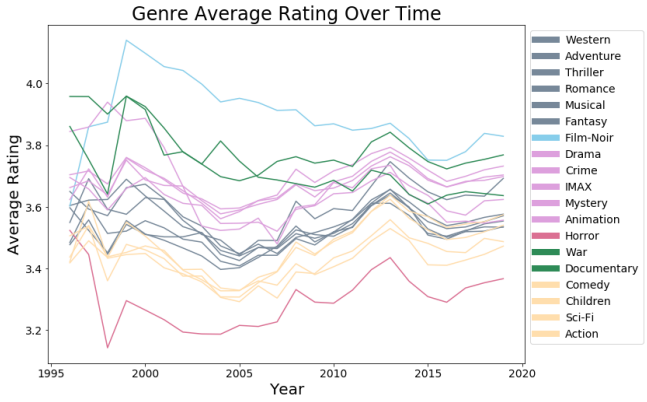

Fig. 3. The average rating of each genre, colored using KMeans results

4.3.1 Using KMeans. Figure 3 is nearly the same as Figure 2, except now the genres are colored according to their cluster. Each genre was assigned to one of six clusters based on their average rating each year. This not only makes it easier to look at, but it makes it easier to pick out the clusters of similar genres (according to average rating over the years).

There a lot of common trends here, such as average rating going up noticeably in 2013 for every genre and then down for a few years afterwards. Each genre follows the same trends, for the most part, some are just higher-rated than others.

As for the clusters, there are several here that one wouldn't find too surprising, such as "War" and "Documentary" being in the same genre and "Sci-Fi" and "Action" also being together. It would be interesting to see how well genres could be clustered together with additional information such as plot keywords, length, and number of main and minor characters. Incorporating some information about the tags like their relevance scores would've been interesting to do as well, but I haven't thought of it earlier.

Also of note is "Film-Noir" and "Horror" not only being the highest-rated and lowest-rated genres, respectively, but they're each in a cluster of their own due to how far they are from other genres.

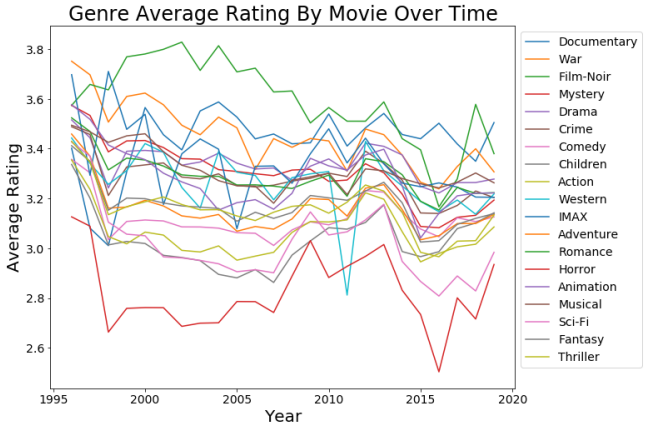

Fig. 4. The average rating of the average ratings of each movie in each genre

4.3.2 By Taking Average Rating of Each Movie in Each Genre. Figure 4 has the same idea as the previous graphs, except the average rating for each movie was first computed and then the average rating for each genre was then taken as the average of each movie that has that genre. This has overall dropped the averages down by a few tenths. While some of the same trends as before are still visible, there are a few genres that buck them by having more drastic movements from year-to-year than other genres. "Film-Noir" and "Horror" still take the highest-rated and lowest-rated awards, respectively.

4.4 Average Rating Per Genre By Month

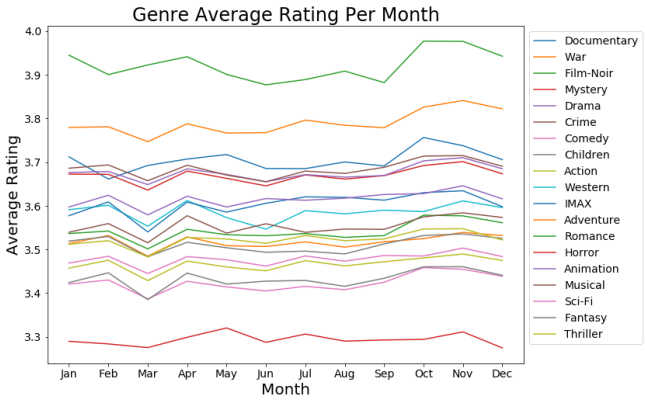

Fig. 5. The average rating of each genre in each month

Figure 5 is similar to the previous plots, except this one distinguishes by month instead of by year. Average ratings here are, again, just the average of each rating that fits the month and genre criteria.

The three most noticeable things to be are the drop in March for most genres, the rise for October and November, and the clearer difference among the top two and the bottom genres compared to when we were looking at the averages over the years.

Table 4. Top 10 Tags According to User Ratings

Tag	Average Rating	Times Tagged
10/10	4.875	123
awesome	4.708	121
masterpiece	4.585	501
must see	4.560	131
good dialogue	4.531	406
good science	4.521	121
epic adventure	4.514	152
prospect preferred	4.5	406
powerful ending	4.491	568
great story	4.481	110

4.5 Best Tags According to Average Rating

Table 4 shows the top 10 tags according to their average rating in movies where the user applied the tag and also rated the movie. Only tags having that condition and still have over 100 occurrences were kept.

Some of these tags aren't surprising to see here (and aren't really helpful), like the first four. However, the rest do tell more about what users like to see in movies. Information like this (over much more data) could be helpful to movie companies to see what aspects people rate higher than others.

4.6 Most Relevant Tag Per Genre

Table 5 attempts to show the most relevant tag for each genre. Some of these are some variation of the genre's name. Others are more interesting, such as the "great ending" tag for "Crime" and "Mystery." Others just make sense, like "gunfight" for "Western." Digging deeper into these tag relevancy scores could also be helpful for movie companies to know which aspects of a movie to focus on.

5 COMPARISON BETWEEN PYSPARK AND PANDAS

A quick comparison of how PySpark - both its DataFrames and SQL - performed compared to Pandas' DataFrames. As Table 6 shows, Within these few tests, PySpark performed much better on joins, while Pandas performed much better on a group by. In a problem combining the two (the last row), PySpark came out on top by being about 4x quicker. I think a more rigorous comparison between the two, involving more problems and datasets would be interesting.

6 CONCLUSION

I found PySpark very easy to use; at least with the parts of it that I explored. While I tried using Plotly, that one seemed more daunting to learn, so I switched to Matplotlib to be able to quickly visualize the results of what I had been doing and I just never switched back. PySpark performed very well, especially with some of the very large joins that I asked of it. I have no regrets about using it. I just wish I had spent more time either finding or scraping a dataset

Table 5. Most Relevant Tag Per Genre

Genre	Tag	Average Relevance
Crime	great ending	0.544
Romance	relationships	0.493
Thriller	suspense	0.568
Adventure	adventure	0.539
Drama	drama	0.500
War	war	0.644
Documentary	documentary	0.803
Fantasy	fantasy world	0.565
Mystery	great ending	0.581
Musical	musical	0.686
Animation	animation	0.846
Film-Noir	film noir	0.790
(no genres listed)	storytelling	0.558
IMAX	action	0.640
Horror	horror	0.714
Western	gunfight	0.767
Comedy	comedy	0.514
Children	family	0.642
Action	action	0.677
Sci-Fi	sci fi	0.623

Table 6. Top 10 Movies by Average Rating

Problem	Pandas	PySpark - DataFrames	PySpark - SQL
“Small” Join	3.35 sec	0.234 sec	0.224 sec
“Big” Join	9.32 sec	0.28 sec	0.263 sec
Group By and Mean	0.491 sec	10.2 sec	10.4 sec
Genre Avg Per Month	64 sec	-	17.2 sec

that I would’ve enjoyed more so I would’ve been more inclined to spend much more time on it throughout the semester (such as an extensive baseball dataset). This would’ve allowed for a larger variety of questions, whereas with this data, I was mostly stuck looking into something that involved rating and then either genres or tags.

As for the data analysis, there were a few interesting takeaways. March seems to be a bad month for movie ratings while later on in the year can be kinder. Nature movies, such as "Planet Earth," are indeed very highly regarded and rated. At least according to ratings, genres can be clustered together. While most of this is likely due to the myriad of movies that share some of these genres, it would be interesting to investigate with additional features whether these clusters hold together. Lastly, apparently 5 stars doesn’t mean a rating of 5.

Expanding on these points with additional data (think Netflix) could lead to some interesting trends that movie companies could take advantage of to help their movies garner better ratings than they otherwise would.

REFERENCES

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

Jesse Vig, Shilad Sen, and John Riedl. 2012. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Trans. Interact. Intell. Syst.* 2, 3, Article 13 (Sept. 2012), 44 pages. https://doi.org/10.1145/2362394.2362395