NETFLIX

PySpark Analysis Report

MovieLens Exploratory Data Analysis

Fictitious case on Netflix 2003

Germán Germán De Souza
MBD Section 2.1

Executive Summary

Becoming an active player in the entertainment industry isn't always an easy task. Fierce competitors such as Blockbuster are dangerous players which can be a potential threat to Netflix. Although it can be daunting competing with such a big player, at Netflix we want to revolutionize the entertainment industry. Over the last years, we have been able to position ourselves in a blue-ocean market which is substantially growing in the last couple of years. Our newest implementation of an online entertainment platform minimizes the users' pain points so that they can enjoy movies anywhere and at any time. At the end of 1999, Netflix began its journey on video on demand platforms, this platform caused such excitement that brought various innovators and early adopters to frequently use the app. As years came across, Netflix became very popular among different generations, it unexpectedly penetrated the entrainment industry. However, it became apparent to Netflix's top managers the importance of identifying their users and their different preferences to constantly improve their service offering and grant customers the most personalized VoD service in the industry. It was when the main managers from the company contacted the analytics department to conduct some research on this matter.

On behalf of the Netflix data analytics department, we are honored to present to the managers' board our main insights on the data that we have gathered in the last 4 years (2000-2003). In this report, we have analyzed the company's platform from a holistic approach perspective by implementing a multiangled exploratory data analysis. This analysis will help us understand the current situational analysis in which Netflix is currently in and propose some possible solutions to modify and improve our current strategy.

As we have mentioned, our report will cover different mutually exclusive main blocks in a structured and organized manner, covering the most important aspects of the platform for each block. The structure is as follows:

- Movies within the platform: In this block, we will be covering the movies genre distribution (Are the movies per genre well compensated or is there an uneven amount of movies per genre?) and analyzing if there is an actual correlation between the top viewed movies up to now in the platform and their corresponding genres.

- Movie Ratings: In this block, we will be covering the average rating distribution for the top 10 genres with most movies, how average ratings have been fluctuating throughout time per genre (This will help us identify the users' reactions for different genres with the movie uploads on a yearly basis) and the user's usability of the platform.

- Users profiling analysis: This block is purposely done to identify the different profiles of our users and the percentage split between the different categories. Also, we will focus on the top 15% of our users and identify which are our top profiles and which ones should the company focus its strategy on.

Netflix General Information

Netflix currently has more than 1 million movie ratings from 6,000 users and over 4,000 movies within the platform. Our data analysis covers data from April 1st, 2000 until February 31st, 2003. According to our data, this is a general overview of Netflix business situation:

- Out of the 4 years that we have analysed, the year 2000 has been the best year in terms of platform usability. The user movie view rate has been around 150 movies in 2000. As we can see, despite our small number of users within the platform we been able to catch some users attention and test our platform.

- The most watched movie has obtained around 2,300 ratings, almost a 0,2% of the total ratings at Netflix.

- The average rating within the platform over the past 4 years has been higher than the mean rating, approximately a 3,5/5 with a small variation between different user ratings. This means that our platform counts with a large variety of movies with relatively positive ratings, meaning that the content is highly valued by our audience.

Without disclosing more data, let's further analyse the blocks which we have explained above.

Exploratory Data Analysis

Analysis objectives

These are the business questions and the different sections which we will cover in this analysis:

Movies

- Movies Genre Distribution: (1) Which are the most frequent genres within the platform? (2) What is the percentage split between the Genres? Is it well compensated?
- 10 Most Popular Movies Genre Analysis: (1) Which are the top 10 most viewed movies within the platform? (2) Is there a specific genre pattern in the top 10 movies? Is there a specific Genre which is highly viewed?

Movie Ratings

- 10 Most Popular Genres Average Ratings Distribution: (1) Which genres has the best average rating score? How is the distribution of each genre? Is the dispersion higher in different Genre and which one is the most stable?
- 10 Most Popular Genres Average Ratings Distribution throughout time: (1) How have the average ratings been fluctuating along the years? Are they reasonably constant or have some of them peaked along these three years?
- Users usability of the platform: (1) How many movies each user watches per year and what is the overall trend along the years? Is the company increasing the users usage of the platform?

Users Profiling

- Identification of different user profiles within the platform: (1) How many users do we have working, studying or unemployed? Is the split equally distributed or is there an unequal distribution between the different categories? (2) How is the age distribution between the different occupation categories?
- Active workers occupation category analysis: (1) What is the percentage split of the "Active workers" in the occupation category? Which occupation category is highly interested in the platform?
- Identifying the top 15% of users occupation: (1) Which occupation do our top 15% most frequent users and to which age range do they correspond?

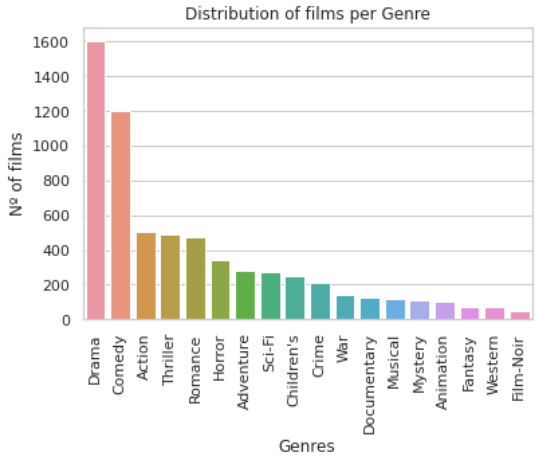
Insights

Platform Movies

(I) Movies Genre Distribution

According to our data analysis, the most frequent genres are Drama and Comedy holding around a 44% of all the films on the platform. The distribution of films in the platform is not evenly distributed and makes sense¬. Films production depends on genre popularity, meaning that movie producers would focus more on those segments that can target more people. Drama and Comedy are highly viewed by most of our users and other less frequent genres such as "Film-Noir" or "Western" are specially selected by users on specific occasions.

Therefore, we can conclude that film distribution is properly assigned based on popularity and the movie producers industry.



(II) 10 Most Popular Movies Genre Analysis

We can clearly state that there is no clear pattern or correlation between the most frequent genres and the top 10 most viewed films within the platform. 7/10 of the top 10 most viewed movies are Action or Sci-Fi movies, genres that aren't that significant within the platform in terms of quantity (Sci-Fi movies hold approximately 4% of the total movies offered in the platform and 7 movies in that category are the most viewed in the platform). This is an actual interesting insight, as although the quantity of some genres indeed increases the likelihood of being in the top 10 charts, it won't ever bet the quality of some movies. This has been presented in this section analysis, in which Sci-Fi movies which were barely noticeable on the platform (due to the number of movies) were ranked in the top charts as the most viewed movies. It is therefore that users don't want a large variety of movies, but what they want are high-quality movies that are worth watching.

Chart 2 - Top 10 most viewed movies (Genre classification)

Genres	Top10RatedMoviesGenres
Action	7
Sci-Fi	7
Adventure	4
Drama	4
War	3
Thriller	3
Comedy	2
Fantasy	1
Romance	1

Chart 3 - Top 10 most viewed movies (2000-2003)

Title	MovieID	Top10RatedMoviesViews
American Beauty (...	2858	3428
Star Wars: Episod...	260	2991
Star Wars: Episod...	1196	2990
Star Wars: Episod...	1210	2883
Jurassic Park (1993)	480	2672
Saving Private Ry...	2028	2653
Terminator 2: Jud...	589	2649
Matrix, The (1999)	2571	2590
Back to the Futur...	1270	2583
Silence of the La...	593	2578

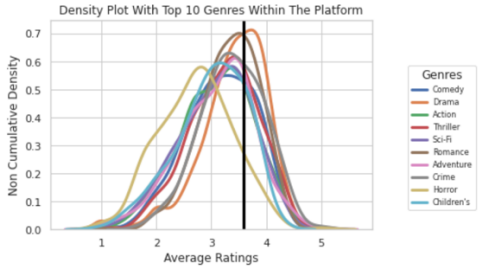
Movie Ratings

(I) 10 Most Popular Genres Average Ratings Distribution

As we did mention earlier, the average rating distribution is relatively high with a low variation. However, we decided that it would be a good idea to deep dive into each category and see the different distributions. As we can observe in the graph, the genre with the best overall ratings is Drama. This genre has the highest rating relative to other top 10 popular genres and is the genre with the highest number of movies on the platform (making it a highly stable genre with certainty). Conversely, Horror movies have the worst average ratings and are the genre with the highest variability in ratings out of all the 18 genres. It might be recommendable to have a closer look at the movies in the horror category and upload high-quality movies to increase this category's average ratings.

Chart 4 - Summary Statistics of each Genre Category

	Genres	count	mean	std	min	25%	50%	75%	max
7	Drama	1493	3.41533	0.610467	1	3.1	3.5	3.84127	5
4	Comedy	1163	3.15992	0.65199	1	2.73071	3.20202	3.66667	5
0	Action	495	3.0983	0.65157	1.28571	2.6538	3.15108	3.5635	4.6087
15	Thriller	485	3.22594	0.62408	1	2.78125	3.26876	3.67103	4.52055
13	Romance	460	3.33914	0.530114	1	3	3.3879	3.74148	4.5
10	Horror	339	2.72705	0.666278	1	2.23802	2.7451	3.18146	5
1	Adventure	281	3.09966	0.67649	1	2.65909	3.19863	3.57143	5
14	Sci-Fi	274	3.07347	0.657751	1.05882	2.61607	3.16525	3.54524	4.45369
3	Children's	250	3.00633	0.637329	1	2.60497	3.08206	3.4922	4.24796
5	Crime	201	3.38513	0.576922	1.71429	3	3.38889	3.81338	5

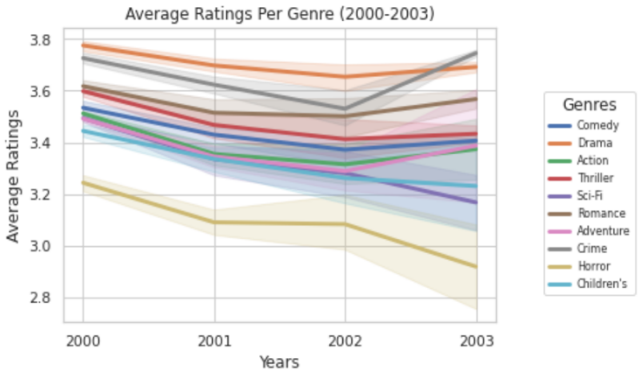


(II) 10 Most Popular Genres Average Ratings Distribution throughout time

While analysing the average ratings gives a good indicator signal that there is a problem in certain genre categories, we are not looking at the entire picture throughout time. Genre's might fluctuate in an oscillating manner over time, by looking at the average rating per year and the different uploads per genre we could identify if there is an actual problem with different movies. It is, therefore, that we bring we introduce the graph below to our analysis. As we can observe, the overall trend in all genres has experienced a downhill trend in their ratings for the past 3 years. Genres such as Adventure and Crime are gaining traction, their ratings have increased during the past year. Conversely, all of the other genres have remained stable over the years except the horror genre which has experienced an 11% decrease in its average ratings in the past 3 years.

Chart 5 - Average ratings per genre category (2000-2003)

Genres	2000	2001	2002	2003
Crime	3.73	3.62	3.53	3.75
Romance	3.62	3.51	3.5	3.57
Thriller	3.6	3.47	3.41	3.43
Adventure	3.49	3.34	3.29	3.39
Drama	3.78	3.7	3.65	3.69
Children's	3.44	3.34	3.26	3.23
War	3.91	3.79	3.73	3.94
Documentary	3.97	3.93	3.88	3.8
Fantasy	3.47	3.34	3.27	3.36
Mystery	3.7	3.63	3.61	3.62
Musical	3.68	3.62	3.63	3.63
Animation	3.7	3.53	3.56	3.42
Film-Noir	4.09	4.03	3.97	4.28
Horror	3.24	3.09	3.08	2.92
Western	3.63	3.61	3.59	3.55
Comedy	3.53	3.43	3.37	3.41
Action	3.51	3.35	3.31	3.37
Sci-Fi	3.49	3.33	3.28	3.17



(III) Users usability of the platform

As we were supposing from the previous blocks of analysis, although the company had in the 2000s a positive acceptance in the entertainment industry, along the last 3 years the platform usability has plummeted. According to the charts below, the average number of movies per user per year has decreased around a 249% in the last 3 years. This has been due to the decrease in the accumulated number of movies watched per year and the number of different users using the platform. We can state that there is something wrong with the platform, however, we would need to gather more data to identify the root cause for the decrease in average movies per user. If the team had to suppose the possible reasons for this decrease they would be the following:

- There are no more new movies on the platform, causing people to leave: If this is the cause we would need all the uploads of movies on the platform and their respective dates. Identifying the movie uploads frequency is crucial to decrease user's churn rate.
- The streaming company is doing something wrong in their infrastructure that is affecting their customers: user experience and interactivity with the platform are key. If this is affected it could be threatening and users may switch to other alternatives in the entertainment industry.

Chart 6 - Descriptive analysis on movie count per year

Year	MovieCount	DistinctMovieCount	DistinctUserIDs	AvgMoviesPerUser(PerYear)
2000	904721	3678	6034	150.0
2001	68094	3289	1070	64.0
2002	24046	2971	565	43.0
2003	3348	1601	178	19.0

Users Profiling

(I) Identification of different user profiles within the platform

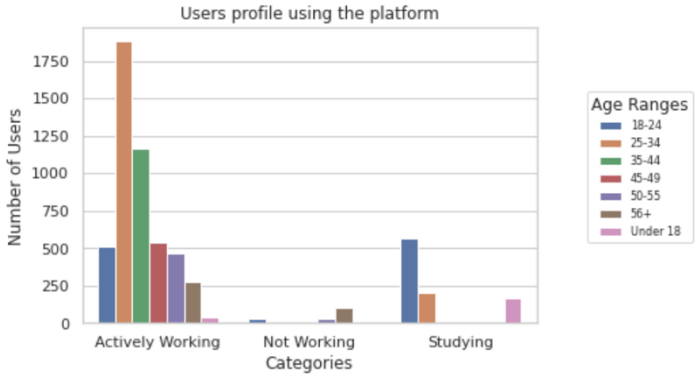
Our users are relatively young, a 73% of our audience is between 18-44 years old. Almost 50% of the platform users are between the ages 25-44, approximately 88% are active workers and a small segment is currently studying as a college/grad student. In this section we found it fascinating the small proportion of students using the platform. It was therefore our motivation to further explore if this segment was frequently using the platform or not. We know that our customer base is between 25-44, however it would be interesting if we should focus our marketing efforts to target the younger segment. Therefore, let's analyze the most frequent users in the platform and check if there are various users in the 18-24 age bracket.

Chart 8 - Occupation Categories division in the platform

OccupationCategory	UsersCount	PercentageSplit
Actively Working	4872	81.0
Studying	954	16.0
Not Working	214	4.0

Chart 9 - Age Distribution in the platform

Age	UsersCount	PercentageSplit
25-34	2096	35.0
35-44	1193	20.0
18-24	1103	18.0
45-49	550	9.0
50-55	496	8.0
56+	380	6.0
Under 18	222	4.0



(II) Active workers occupation category analysis

As active workers comprise 81% of our platform users, it might be worthwhile to analyze the different occupations and identify if there is a pattern of our target buyer persona. Interestingly, we can state that those professions with higher leisure time are Netflix user targets and those that have professions with lower leisure time rarely use the platform.

Chart 10 - Occupation split in the 'actively working' category

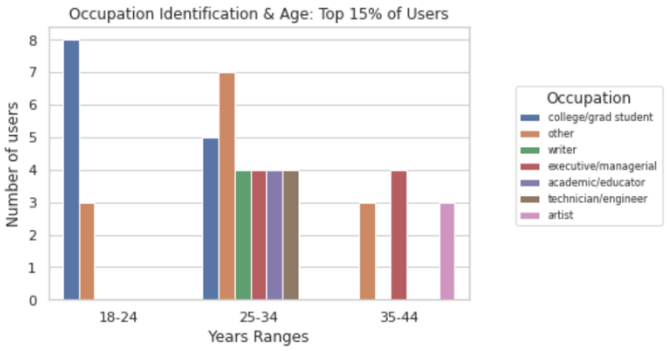
OccupationCategory	Occupation	UsersCount	PercentageSplit
Actively Working	other	711	15.0
Actively Working	executive/managerial	679	14.0
Actively Working	academic/educator	528	11.0
Actively Working	technician/engineer	502	10.0
Actively Working	programmer	388	8.0
Actively Working	sales/marketing	302	6.0
Actively Working	writer	281	6.0
Actively Working	artist	267	5.0
Actively Working	self-employed	241	5.0
Actively Working	doctor/health care	236	5.0
Actively Working	clerical/admin	173	4.0
Actively Working	scientist	144	3.0
Actively Working	lawyer	129	3.0
Actively Working	customer service	112	2.0
Actively Working	homemaker	92	2.0
Actively Working	tradesman/craftsman	70	1.0
Actively Working	farmer	17	0.0

(III) Identifying the top 15% of users occupation

Looking at the data in chart 11 and the graph below, we can confirm that although the 18-24 age bracket isn't the highest, it would be a great strategy to focus on them and diversify the risk of focussing solely on the 25-44 bracket. This young segment has more time and are medium-term financially reliant on their relatives, building a brand image from an initial stage could potentially increase their brand loyalty towards Netflix and become a long-term user in the future. Heavily investing in the 25-44 bracket is very attractive, but introducing strategies to capture younger clients might not be a bad idea (this is since this segment have a higher screen time than the average user).

Chart 11 - Occupation Identification & Age: Top 15% of Users

	Occupation	Age	UsersCount
0	college/grad student	18-24	8
1	other	25-34	7
2	college/grad student	25-34	5
3	writer	25-34	4
4	executive/managerial	35-44	4
5	academic/educator	25-34	4
6	executive/managerial	25-34	4
7	technician/engineer	25-34	4
8	other	18-24	3
9	artist	35-44	3
10	other	35-44	3



Conclusions & Recommendations

Taking into consideration all the different blocks, we can state that Netflix is facing a difficult situation. User's screening time and average ratings have dramatically decreased in the last 3 years. Also, the number of users hasn't increased that much on a yearly basis. It is therefore that we propose Netflix start targeting new clients (preferably users between 18-44 and actively working or studying), increase the quality of movies in different

genres (For instances: Horror movies or less popular genres), increase the frequency of uploads and users awareness on the new uploads.
These are some of the following recommendations which we strongly encourage Netflix to execute to minimize these issues:

- To increase the quality of uploads, Netflix should analyse the best movies over the years and analyse the most popular ones. Also, if Netflix manages to identify some trends in the film industry, they would be able to anticipate user's wants and frequently upload different content adapted to relevant actual trends. If we manage to do this, we would attract more new users and create an expectation on our current users. This will hopefully increase the average screening time and the number of users would decrease their churn rate in the long term.
- Netflix fixed monthly subscription works, however, to attract new users we should set different subscription rates that would be adapted to different customer segments. We have analysed the different user profiles within Netflix. Maybe it would be recommendable to propose a new pricing strategy to incentivize this new blue ocean market segment.

Netflix infrastructure is relatively new and can grow, however, the implementation of these recommendations are crucial for the future growth of the company and their future survival within the industry. Our team considers that It is time that Netflix effectively reformulates its strategy, we hope that this analysis enables the other segments within the company to work cohesively towards making Netflix one of the most successful companies in the entertainment industry.