MovieLens Data Analysis

Snehal Nair

Monday 23$^{\text{rd}}$ January, 2017

Abstract

The purpose of this study is to look at the distribution of ratings, movie and users over time, impact of user mood on average rating score and average rating score of genre over time. The analysis is divided into 4 different 5-year batches to run analysis on sections of data. It was found, the growth, trend and level are stable after the first 5 periods (i.e. after the year 2000). With frequency of rating showing high correlation to new movies and users added, trend for rating over time shows combining effect of growth in user and movie base . Further, weekday-weekend analysis show most of the ratings (approx.70%) are happening over the weekdays. For average rating score, a notable observation is, the shift in the rating pattern for the last batch(latest batch, 2011-2015). In this batch approximately 50% of the rating scores are average and the 25% each for poor and high rating scores in comparison to the other batches where it was 80-20 between average and high/poor rating scores. In the genre analysis it was found 9.4% times users rated genre below 3, 17.5% times for high and 70% times average.

1 Introduction

MovieLens is a recommender system and virtual community website that recommends movies for its users to watch, based on their film preferences using collaborative filtering. GroupLens Research, a research lab in the Department of Computer Science and Engineering at the University of Minnesota, created MovieLens in 1997 to gather research data on personalized recommendations. MovieLens 100M datatset is taken from the MovieLens website, which customizes user recommendation based on the ratings given by the user. There are 2 tuples, movies and ratings which contains variables such as MovieID::Genre::Title and UserID::MovieID::Rating::Timestamp respectively.

We are looking at a dataset of 22 years. Therefore the analysis is divided into 4 batches to study the ratings, movies, users and genres over time. Further analysis to be directed to answer the observations in the summary statistics for these batches. The study is primarily divided into three main sections. In the first section we are looking at the distribution of ratings, movies and users over time. For this analysis, we will look at the distribution over years to understand the individual effect thereafter run analysis involving both growth of movies and growth of users along with cumulative ratings to understand the combined impact. We will run correlations to study the relationship. Growth of ratings, movies and users are calculated on the cumulative number over the years. The formula is given as -

$$\frac{Growth(currentperiod)}{Growth(previousperiod)} - 1$$

Further we will look at average user rating pattern with respect to weekend, weekday and festivals.

In the next section we are running analysis on the mood of the user-base to understand the impact on average rating score for movies. For this analysis we will divide the dataset based on poor, average and high rating scores. Further we will study the average rating scores over different time-frames like day, month, year and hour. we will also look at the top 5 and bottom 5 movies based on average rating score for movies that received >20 ratings.

In the final section we will run analysis on genre over time. Genres are a pipe-separated entity in the movies tuple. There are 18 genre in all excluding "IMAX" and "no genres listed". These two genres will be excluded from analysis. Further, for this analysis movies classified with mutliple genres needs to be separated and classified under each genre, for e.g. a movie classified under 10 genres should be split and classified under all the genres it falls into. After we split the dataset based on genres, the total number of ratings received will see an increase from 100,004 to 262,343. We will thereafter run analysis in the subsections by breaking the genres based on poor, average and high rating scores received.

2 Exploratory Analysis

2.1 Summary Statistics

There are 9,125 movies in tuple for movies and 9,066 in the ratings tuple. After merging the movies dataset with ratings, the final dataset has 9066 movies. There are in all 3,840 days in the dataset with 100,004 ratings and 671 users. There is no missing data or NAs in the dataset.

Table 1: Summary of the variables in the MovieLens dataset

Movies	Users	Ratings	Period
9,066	671	100,004	1995 - 2016

In the Table 2 Summary, we have removed year-1995 data as it has higher dumps in the inception phase, causing bias in the summary. Thereafter to maintain 5 year batch slots for in depth analysis, year-2016 data is removed to have comparable figures. The min, mean and max are calculated based on the total ratings received in the years within respective 5-year batch, for e.g. in the first batch (1996-2000), the minimum number of ratings, 1,825 was received in the year 1998, the maximum 13,869 in the year 2000 and the mean,6,225 is the mean of the years in the batch.

Table 2: Summary of the total number of ratings received

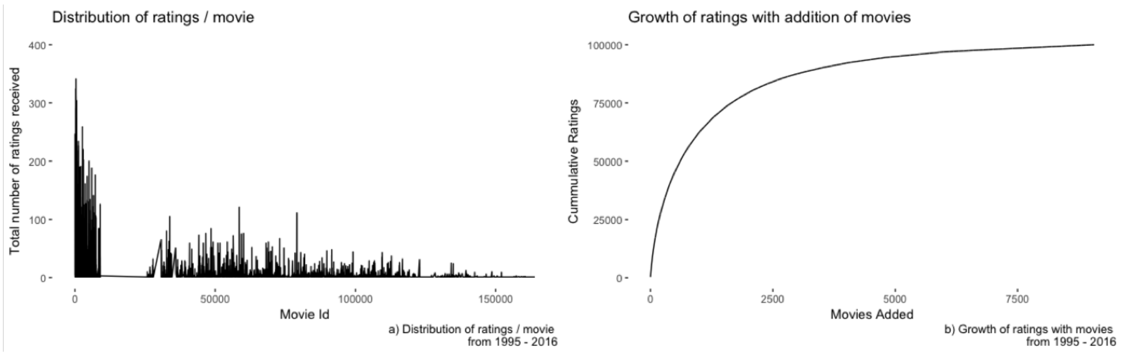
Number of ratings	1996 to 2000	2001 to 2005	2006 to 2010	2011 to 2015
Active Users	307	215	214	209
Min	1,825	3,938	1,548	1,969
Mean	6,225	4,975	3,733	3,820
Max	13,869	7,161	7,493	6,610
Total	31,128	24,877	18,669	19,102

The summary table shows highest number ratings received in the first, 5-year batch (1996-2000). Its maximum ratings, 13,869 is double the number of ratings received in the later batches. Second batch onwards we can see stability in the statistics. Although the total number of ratings received in the second batch is higher than the total number of ratings received in third and fourth batch. The total number of ratings received across first, second, third and fourth batches are 33.2%, 26.5%, 20% and 20.4% respectively. Last two batches clearly show drop in the ratings received caused by drop in mean due to lower number of ratings received, 1548 and 1969 in the year 2007 and 2013 respectively.

2.2 Distribution of Ratings per movie

Assignment Question - Compare distribution of ratings per movie (i.e. how many ratings are there per movie?).

In this section we will look at exploring the distribution of ratings per movie. We will first look at total number of ratings received by each movie and then compare the correlation between growth observed in ratings and growth observed in movies (new movie addition). In the subsection 2.2.1 we will compare the growth rate between ratings and movies from 1997 to 2016. We have removed the periods 1995 and 1996 in the growth analysis since the abnormal dump in the inception phase is affecting the scales in the graph.



a) Distribution of ratings / movie from 1995 - 2016

b) Growth of ratings with movies from 1995 - 2016

Figure 1: Distribution of ratings per movie from 1995 - 2016

Fig 1 a) shows distribution of ratings received per movie from 1995 to 2016. The small to large movie ID in the database represents old to new movies. From the graph it is clear, older the movie in the database more amount of exposure it has received to receive ratings from users. Therefore newer movies(as per the release date or as per the date of entry in the database) received lesser ratings. **Missing data**: The movie ID shows series of movie IDs, 9018 to 25737 are missing. There is no change in the naming convention after 25737, to skip these numbers. Further investigation reveals these movie IDs are reserved to enter movies released before 1995 as we can see movies older than 1995 are added in recent years using these reserved smaller Ids.

In Fig 2.1 b) Movies added date in the database is taken as the date it received its first rating. The figure, shows cumulative growth of ratings received with new movies added in the database. Cumulative ratings in the database shows logarithmic growth with respect to the movies added over 22 years period is given by $y = a + b * log(x)$. It has a period of rapid increase, followed by a period where the growth slows, but the growth continues to increase without bound proving the point that older movies will continue to received more ratings due to duration of exposure.

2.2.1 Comparison of Growth Rate of Ratings vs. Movie

This section focuses on growth rate of movies and ratings. We will also explore correlations between growth-rate between movies and ratings if any . The correlation coefficient between the growth of ratings and movies shows perfect correlation at 0.94. From the figure we can see the growth observed in the inception phase, i.e. the first batch (1995-2000) shows abnormally high growth with 80% growth in 1999. We can see the trend stabilizing after 2001, therefore the growth rate between periods is comparable.

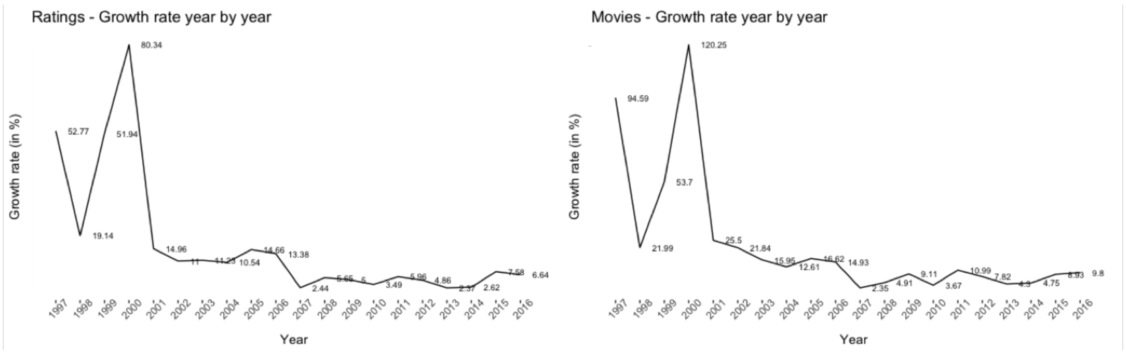


Figure 2: Comparison of growth between ratings in the database vs. the growth of users in the database by year.

In the Table 2 summary statistics in section 2.1 we saw the number of ratings received in second and third batches have dropped. From the figure we can see it is attributed by steep level drop in movie growth from 13.4% in 2006 to 2.4% in 2007. Thereafter the growth trend maintained single digit growth. Comparing the growth in 2015 and 2016, we can see movies showed growth by 1% where as ratings dropped by 1% which means ratings is correlated with some other factor. We will investigate this further the next section for distribution of ratings per user.

2.3 Distribution of Ratings per user

Assignment Question - Compare distribution of ratings per user (i.e. how many ratings are there per user?)

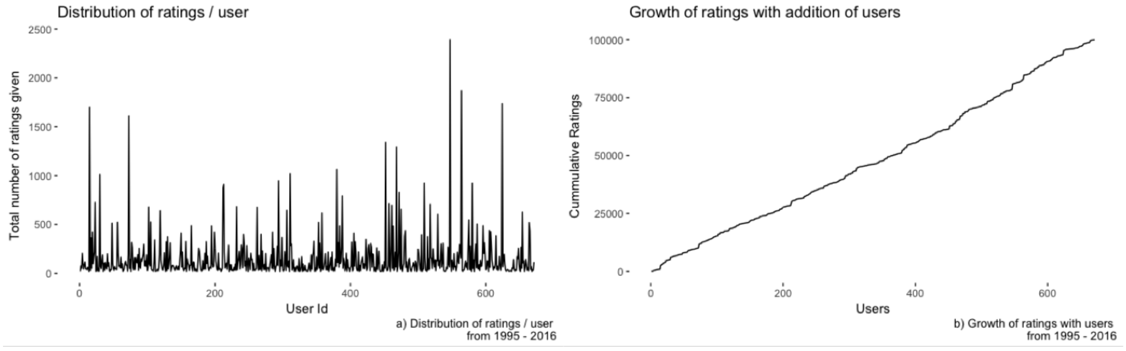


Figure 3: Distribution of ratings per user from 1995 - 2016

Fig 3 a) shows distribution of ratings per user added from 1995 to 2016. Like movies Id, small to large user ID in the database represents old to new users. Ideally we will expect older users to give larger number of ratings as they have had more exposure to movies over the years. However we see a stable trend and level. In-depth investigation shows that many users remain dormant after entering the dataset. Some of those users became active in the later years and some others still remained dormant, while some of the new users showed high levels of activity from the the time they were entered in the data. For e.g. UserId "2" existed in the database from 1996 therefore UserId 1 is expected in the similar timeframe however it appears in the dataset by giving its first rating to "Dangerous Minds (1995)" in 2009. The maximum number of ratings given by a user is 2391 and the minimum is 20 as the data is cleaned by GroupLens company for users with 20 ratings only.Fig 1 b) shows the cumulative growth of number of ratings with respect to the growth of user database. Both are highly correlated with coefficient 0.9976.

Let us run further investigation on the influence of growth-rate of users on the growth-rate of ratings.

2.3.1 Comparison of Growth Rate of Ratings vs. Users

In this section let us investigate the influence and correlation between the growth rates for user and ratings. The growth rate of ratings and the growth rate of users added is directly correlated by 0.95.

We observed in the section 2.2.1 that the while the movies growth dropped by 1% ratings showed positive growth by 1% in 2016. From the below graph we can see the growth for users dropped by 4% in 2016. Since rating growth is highly correlated with both movies and user growth, the drop in user by 4% got balanced by the growth in movies by 1%. Therefore growth for ratings was not affected to a larger extent.

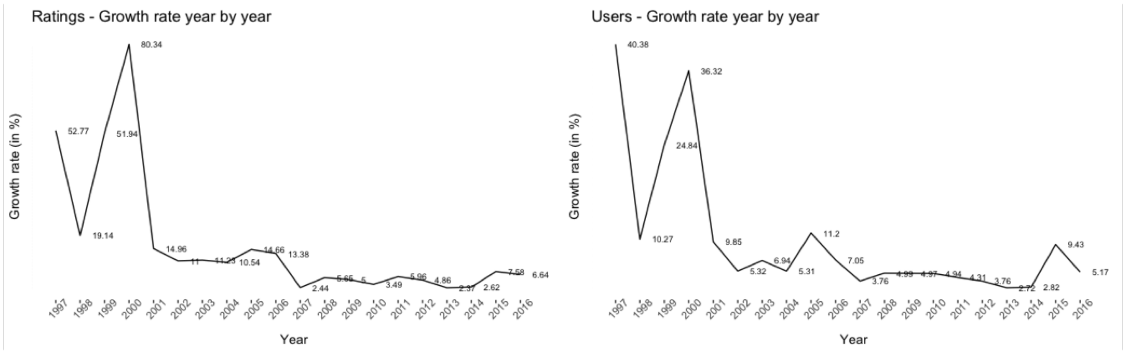


Figure 4: Shows comparison of growth between ratings in the database vs. the growth of users in the database by year.

2.4 Distribution of Ratings Over-Time

Assignment Question - The number of ratings over time (i.e. how many ratings produced over time?)

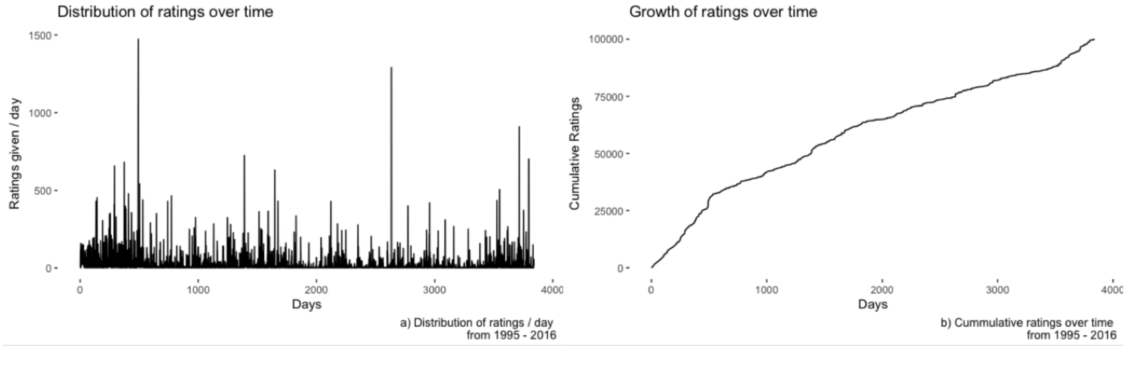


Figure 5: Distribution of ratings over-time from 1995 - 2016

Based on the analysis in the previous sections, it is clear that growth of ratings over time is a combined effect of growth observed for movies and users in the database. We can observe a steady growth for ratings over the years. The highest number of ratings received in the last 20 years was 13869 in the year 2000 and the minimum ratings 1548 in 2007. In the first three years the movies database growth is flat, whereas the user database growth is very high. After 4000 days we can observe steady growth for both movies and users database growth.

2.4.1 Weekend vs. Weekday Ratings Over-Time

In this section let us explore for patterns in the ratings given over weekends and weekdays. For this analysis we chose a random period of year-2010. The graph below compares overall ratings given per movie ID with weekend, weekday and festival (Christmas). The figure shows most of the ratings are given in the weedays which accounts for 70%. 15% ratings are given on Christmas (25th Dec) and the rest of the 15% over weekdays.

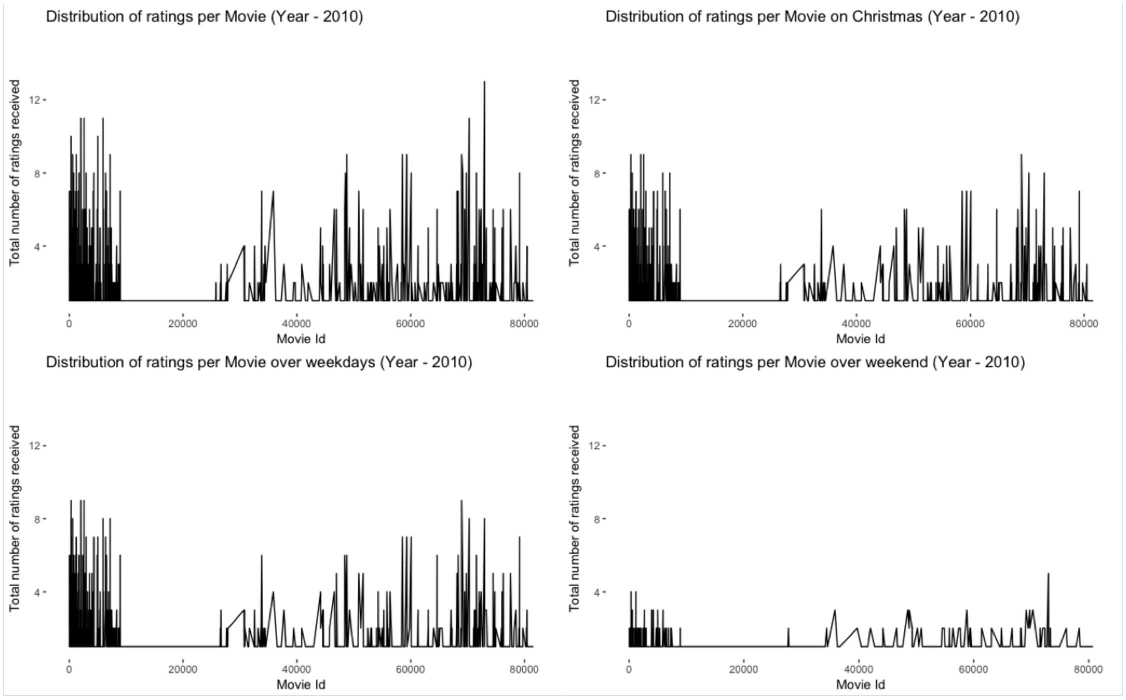


Figure 6: Rating distribution analysis over weekend , weekdays and christmas for the year 2010.

2.4.2 Average Rating per User Over-Time

In this section let us explore seasonality, if any in the user behaviour in giving ratings. From the graphs below we can see the initial years received higher ratings. From section 2.1 Table 2 Summary, we know that the active users in this batch slot (1996-2000) was 307 in comparison to the other batches with 200 active users which explains higher levels of rating activity in this batch. After the year 2000, the level of average ratings moves to a new level and maintains a stable trend.

Monthly average ratings shows a minor trough from June to September this can be attributed to relatively fewer movie releases before festival movie releases. We can therefore observe ratings pick up after September. But relatively the trend is stable. Months do not seem to have major influence over average ratings per user.

Weekly rating averages establishes our observation in the section 2.4.1. It shows the rating is more active in the weekdays in comparison to weekends. It also shows clear pattern where the rating activity pick towards middle of the week and again drops towards the end of the week.

Days of month rating averages shows peaks and troughs a pattern similar to weekend and weekday. There is no monthly trend, where the rating is active in the initial days of the month or towards the end, the rating looks stable across month except few peaks and troughs attributed by weekend and weekday.

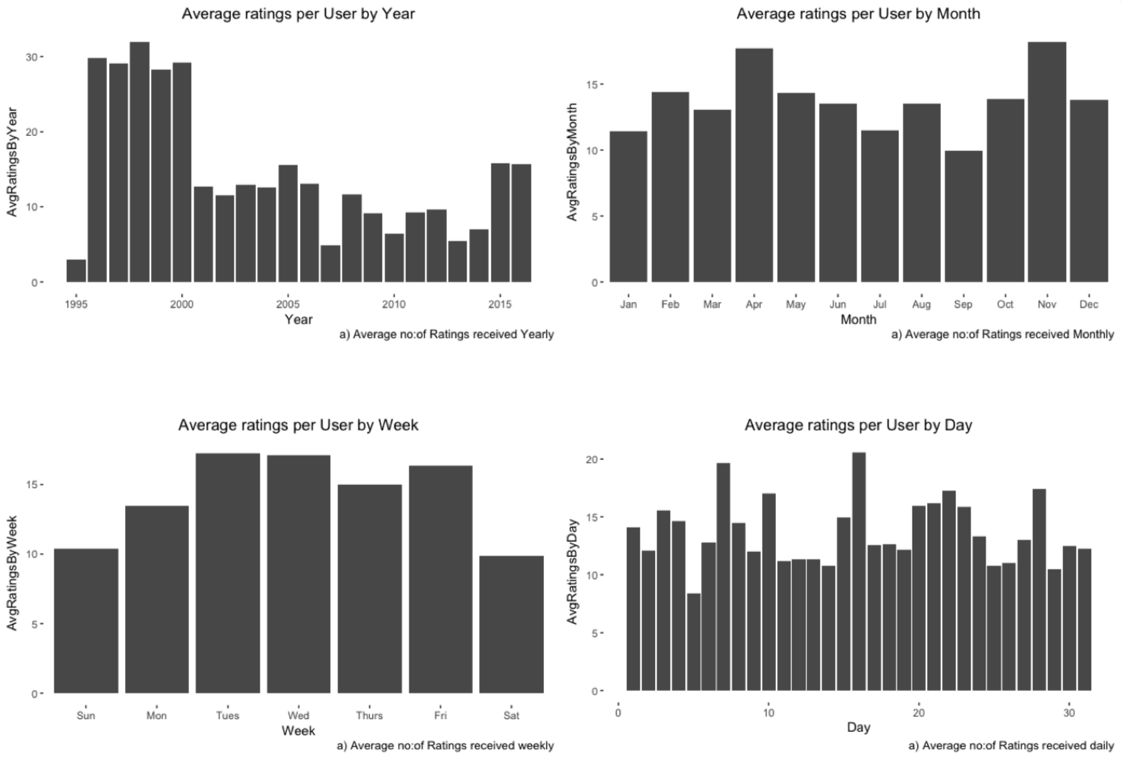


Figure 7: Rating distribution analysis over weekend , weekdays and christmas for the year 2010.

3 Assess the mood of the user base

Assignment Question - In order to assess the mood of the user base, explore how the average rating score per day changes over time

In this section let us explore the user behaviour in respect to the rating score given. Here we will look at the summary statistics across the four different 5 year batches that we built in the beginning of this report. We will compare the rating scores given across periods. In Table 3 below for poor rating we have considered any rating below 3, for average 3/3-4, for high ratings 4/4-5.

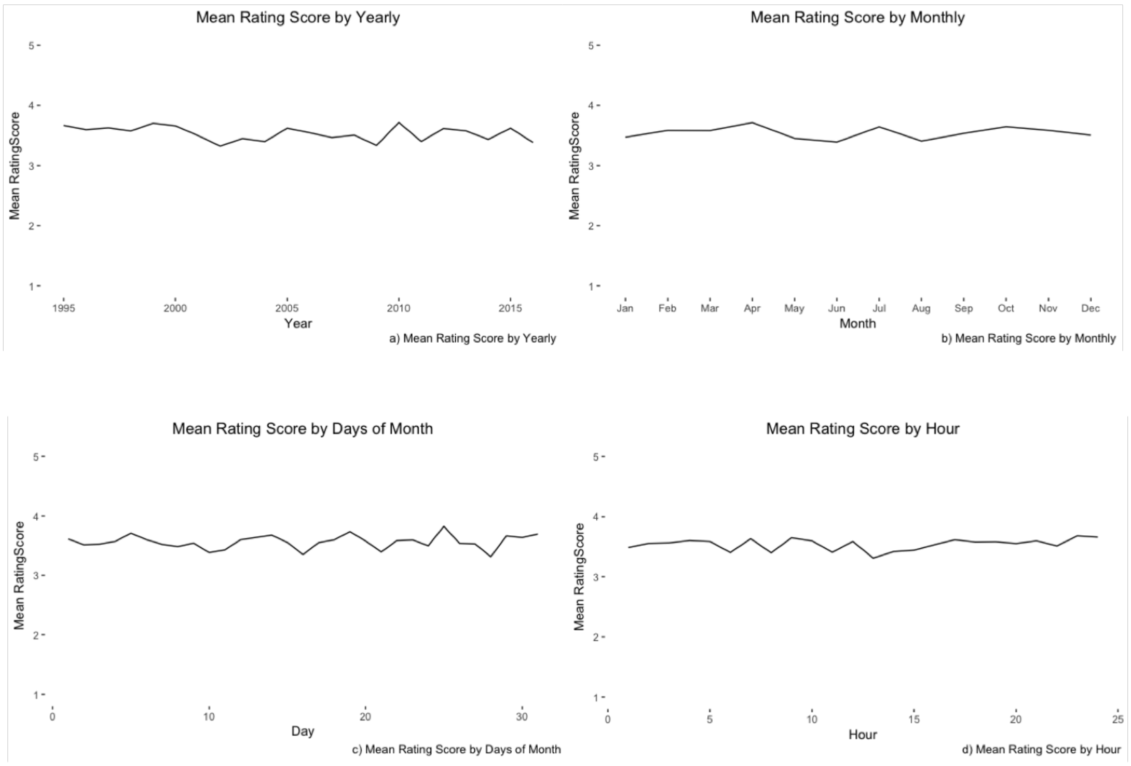
Table 3: Summary Rating Scores

Avg rating score	1996 to 2000	2001 to 2005	2006 to 2010	2011 to 2015
Min	3.57	3.32	3.33	3.39
Mean	3.62	3.46	3.51	3.53
Max	3.70	3.62	3.71	3.62
Std	1.04	1.07	1.00	1.04
# Poor Rating (< 3)	1,336 (4.3%)	3,140 (12.6%)	1,631 (8.7%)	4,363 (22.8%)
# Average Rating (3 - 4)	24,900 (80.0%)	18,954 (76.2%)	13,775 (73.8%)	10,293 (53.9%)
# High Rating (>=4)	4,892 (15.7%)	2,783 (11.2%)	3,263 (17.5%)	4,446 (23.3%)
Total Ratings	31,128	24,877	18,669	19,102

From table 3 above, we can see first batch (1995-2000) shows users showed a bend towards average ratings scores with 80% going to average rating scores. Very few, 4% given to poor rating scores and 15.7% went to higher rating scores. And the trend is followed in the next two batches.

A drastic shift in the rating pattern is observed for the last batch(latest batch, 2011-2015). We can see a clear shift in the user behaviour in the rating scores given. In this batch approximately 50% of the rating scores are average and the 25% each divided between poor and high rating scores.

Let us now take a look at the impact of time over rating scores given by users. For this we will look at four different time-frames like year, month, days of month and hour. The figure below exhibits stable rating score trend across different time-frames therefore we can conclude, rating score is not affected by time of frame.


Mean Rating Score by Yearly

a) Mean Rating Score by Yearly

Mean Rating Score by Monthly

b) Mean Rating Score by Monthly

Mean Rating Score by Days of Month

c) Mean Rating Score by Days of Month

Mean Rating Score by Hour

d) Mean Rating Score by Hour

Figure 8: Assessing mood of user base based on the ratings given in different periods of time.

4 Top and Bottom 5 movies

Assignment Question - What are the top-5 and bottom-5 rated movies, based on their average scores, that have at least 25 ratings?

Table 4: Top 5 Movies based on average scores that have atleast 25 ratings

Top-5	$TotalRatingsReceived$	$AverageRatingScore$
$Godfather, The(1972)$	200	4.49
$ShawshankRedemption, The(1994)$	311	4.49
$OntheWaterfront(1954)$	29	4.45
$AllAboutEve(1950)$	38	4.43
$Ran(1985)$	26	4.42

Table 5: Bottom 5 Movies based on average scores that have atleast 25 ratings

Bottom-5	$TotalRatingsReceived$	$AverageRatingScore$
$Anaconda(1997)$	28	2.02
$WildWildWest(1999)$	47	2.03
$LostinSpace(1998)$	26	2.13
$Batman\&Robin(1997)$	47	2.15
$Godzilla(1998)$	28	2.18

5 Genre Analysis

Assignment Question - Explore how the average rating score of each genre changes over time in the dataset

Let us analyze the rating scores given for each genre in this section. Genres are a pipe-separated entity in the movies tuple. There are 18 genre in all excluding ”IMAX” and ”no genres listed”. These two genres are excluded from analysis. Further, for this analysis movies classified with mutliple genres were separated and classified under each genre, for e.g. a movie classified under 10 genres were split and classified under all the genres it falls into. After we split the dataset based on genres, the total ratings increased from 100,004 to 262,343. We will thereafter run analysis in the subsections by breaking the genres based on poor, average and high rating scores received.

From the below figure 9, we can see the ratings across genre is stable and fluctuates mostly between 3 and 4. Action, adventure, fantasy and film-noir shows stable trend line across different periods. Rest all other genre shows slightly declining rating trend except western genre. Most of the genre shows declining trough in the middle periods between 2000 and 2010. Let us look at the genre under different rating scales in the below sections.

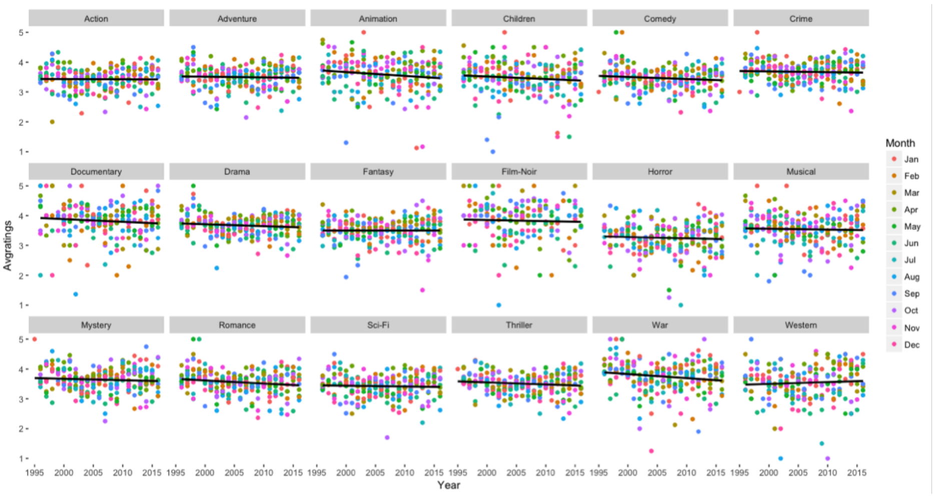


Figure 9: Average rating score by genre over time..

5.1 Genre Analysis with Average Rating Score less than 3

In this section we are only looking at genres that received poor rating scores (rating score below 3). From the figure, it is evident not many movies received poor ratings except horror which received 58 poor rating scores. Sci-fi, children, action musical, comedy and western received 25-30 poor rating scores. Overall 9.4% times genres received poor ratings across periods.

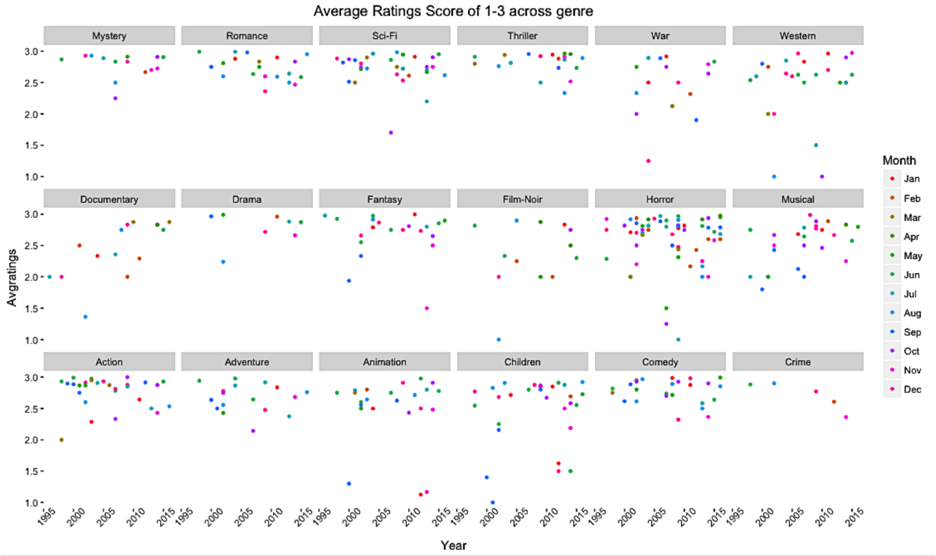


Figure 10: Average rating score(1-3) by genre over time.

5.2 Genre Analysis with Average Rating Score between 3 and 4

Let us take a look at the genre performance under the average rating score scale (rating score between 3 and 4, excluding 4). The below figure shows, this break-up has highest number of ratings with 70% of the ratings scores. The most popular genres in this section are action, adventure, comedy and thriller which received more than 200 ratings scores over the periods.

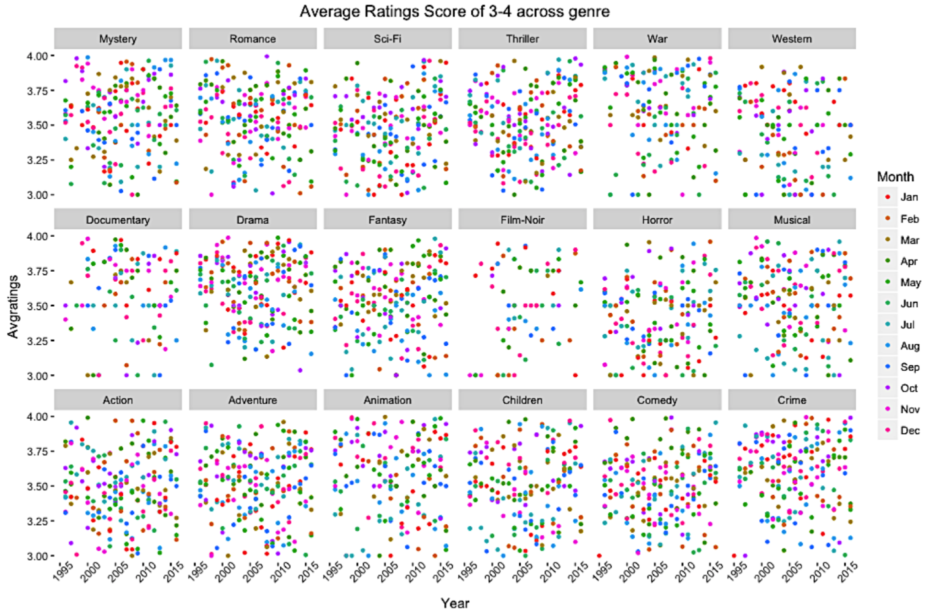


Figure 11: Average rating score (3-4) by genre over time.

5.3 Genre Analysis with Average Rating Score between 4 and 5

This section focuses on high rating scores (between 4 and 5 including 4). From the figure it is apparent some genres are highly and distinctly favoured. Genres like war, film Noir and documentary received more than 75 rating scores. 17.5% of the times genres received rating scores between 4-5 across periods.

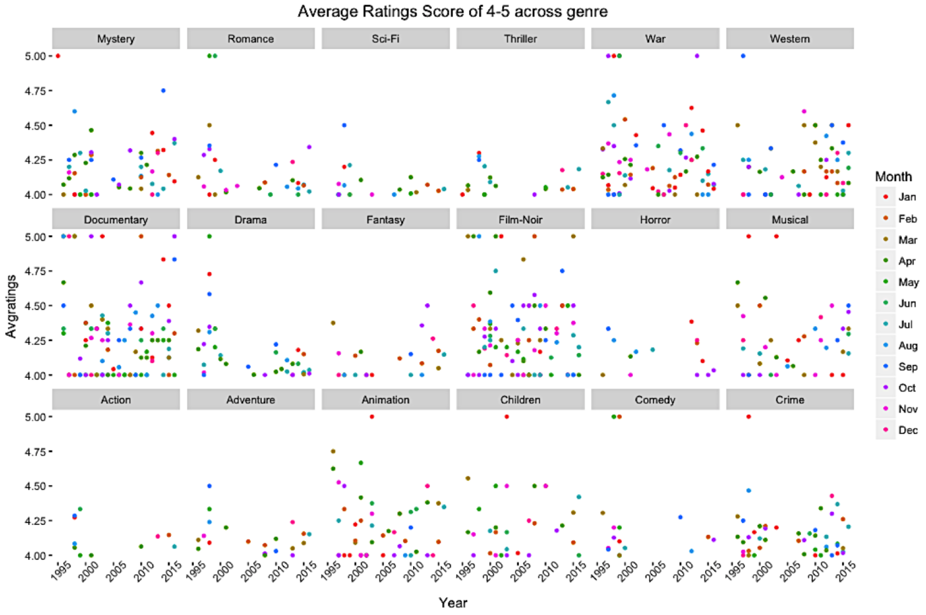


Figure 12: Average rating score (4-5) by genre over time.

6 Conclusion

The study was primarily divided into three main sections to answer below questions related to MovieLens dataset.

(i) **Distribution of ratings, movies and users over time** - The distribution of movie, users and ratings over time shows stable trend with stable growth. We also observed, older the movie in the database higher the ratings received because of the exposure over the years. Comparison of growth-rate of ratings with the growth-rate of movies and growth-rate of ratings show high correlation of 0.94 and 0.95 respectively. Therefore ratings over-time follows a trend which is a combination of trends for movie and ratings. We also looked at the impact of user behaviour on weekend vs. weekdays on ratings which shows the users are more active (70%) in the mid-week and slows down (30%) over the weekend.

(ii) **Impact of mood of the user-base on rating score** - For this analysis we looked at the average rating score per user across different time frames like year, month, day, hour. Data did not show any distinct trend or pattern. Concluding, time-frame has limited impact on rating scores. A notable observation is the shift in the rating pattern for the last batch(current batch, 2011-2015). In this batch approximately 50% of the rating scores are average and the 25% each divided between poor and high rating scores in comparison to other batches where it was divided as 80-20 between average and high/poor rating scores.

(iii) **How each genre is rated over-time** - Here we looked at genres by dividing them into poor, average and high rating scores. Genres like war, film Noir and documentary received more than 75 rating scores in high rating score category. Approximately 17.5% of the times genres were classified under this category. The break-up for average ratings score had the highest number of ratings with

70% of the ratings scores. The most popular genres in this slot were action, adventure, comedy and thriller which received more than 200 rating scores. Finally in the below 3 rating score category, not many movies received poor ratings except horror which received 58 poor rating scores. Sci-fi, children, action musical, comedy and western received 25-30 poor rating scores. Overall 9.4% times genres were classified under this category.

One of the limitations of the study is limited variables to understand and study the rating pattern, user mood and behaviour. More information on user demographics like age, occupation, can help bring in more insights on the trend breaks and level changes observed for e.g. there is high level change in the year 2000. Moreover the dataset is cleaned for users less than 20 ratings affecting in-depth analysis on user behaviour. The growth analysis will be more robust once we have 24 periods to calculate moving average growths. Currently we have 22 periods, of which first 4 periods cannot be part of analysis due to data dumps in the inception phase.

Further analysis can be carried out on causality of changes observed in span of ratings, impact of new users joining the dataset, new ratings input or new movies added.

References

[1] Paper on temporal analysis of rating Datasets (from UCL)

[2] MovieLens Dataset Exploratory Analysis on R studio (Internet)