Movie Lens Ratings through our lens

Nitish Bahadur, Mohmad El-Rifai, Farha Mohsin, Satishraju Rajendran, Li Wenjing

WPI

March 2016

Abstract

We study two datasets 1MM ratings from movie lens and 22MM ratings from movie lens. Exploratory analysis is performed with 1MM data set. Using the 22MM ratings, we empirically study the ratings time series after the release date of the movie. Moreover, we study if there is any relationship between the ratings and the profitability of the movie.

Keywords: Movie Lens, TMDB, Data Mining;

Introduction:

Entertainment is big business in the United States. It is expected to generate over 679 billion US dollars in value over the next four years proving its worth in domestic markets and as a major U.S. export. The film industry is considered a cornerstone of the industry. Technological advances, globalization, and improving internet penetration is changing film production life cycle.

Our primary motivation is to empirically study the 1MM movie lens ratings data set to evaluate how men and women rate movies; is the rating similar; what genres of movies are popular.

Ambitiously, we extend our work by studying 22MM movie lens rating. We specifically studied how the median ratings change after 90, 180, 365, and greater than 365 days. Moreover, we compared these analytics to profitability of a movie, as reported in TMDB database.

The remainder of the paper is structured as follows. The next section describes the data that we have used in empirical testing. Analysis and testing methodologies report each question in the case study.

Data

Movie Lens 1MM[1] dataset (aka dataset 1), Movie Lens 22MM dataset (aka dataset 2) and TMDB[2] dataset was used.

Movie Lens: (1MM)

The movie lens dataset comprises 3 files:

i. Ratings.dat – the file contains 1,000,210 instances. The file contains range of user ids between 1 and 6040, movie identifiers between 1 and 3952, ratings on a 5 point scale, and a timestamp in epoch format.

ii. Users.dat – the file contains 6040 instances. Along with the user identifier, the other attributes are gender, age, occupation and zip code. Gender, Age, and Occupation are nominal attributes. While there are 7 age brackets, there are 21 occupation types.

iii. Movies.dat – The 3952 movie instances has movie title and a pipe delimited genres. The dataset uses 20 genres.

The Movie Database (TMDB):

To enhance our movie attributes we use the links.csv file from 22MM dataset to get the TMBD identifiers for each movie lens identifier in our 1MM dataset. Using TMDB_ID, we wrote a REST client to download the following attributes to a file:

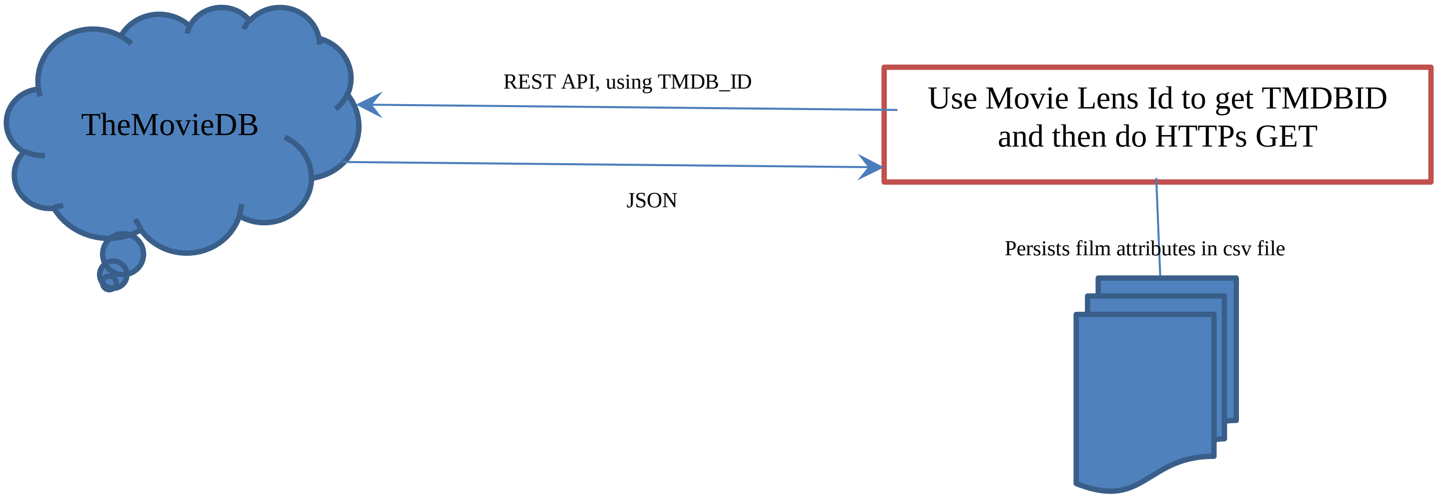
- Original Language
- Original Title
- Release Date

[1] Larger data sets such as 22MM are more current but the user profile is sanitized.

[2] https://www.themoviedb.org/?language=en

- Budget
- Revenue
- Runtime
- Vote Average
- Vote Count
- Popularity
- Genres
- Production Companies
- Production Countries

.


TheMovieDB
REST API, using TMDB_ID
Use Movie Lens Id to get TMDBID and then do HTTPs GET
JSON
Persists film attributes in csv file

Movie Lens: (22MM)

Unlike 1MM dataset, this was comma separated. While the dataset contained user identifier, it did not contain any user information. It contained 22MM ratings.

Analysis / Testing Methodologies / Empirical Results

This section corresponds to the questions asked in the case study.

1.1 Data loading

Data is loaded and merged using pandas. To load the data, we use the split function to separate the columns in each row and then build a list dynamically. The list is finally appended to a pandas data frame. The same paradigm is followed for movies, users and ratings dataset. Subsequently, we merge the merged data set using pandas HDFStore method. For details, please refer to the Case Study 2 IPython notebook.

1.2 Basic Summary Statistics

The section below presents some of the summary statistics.

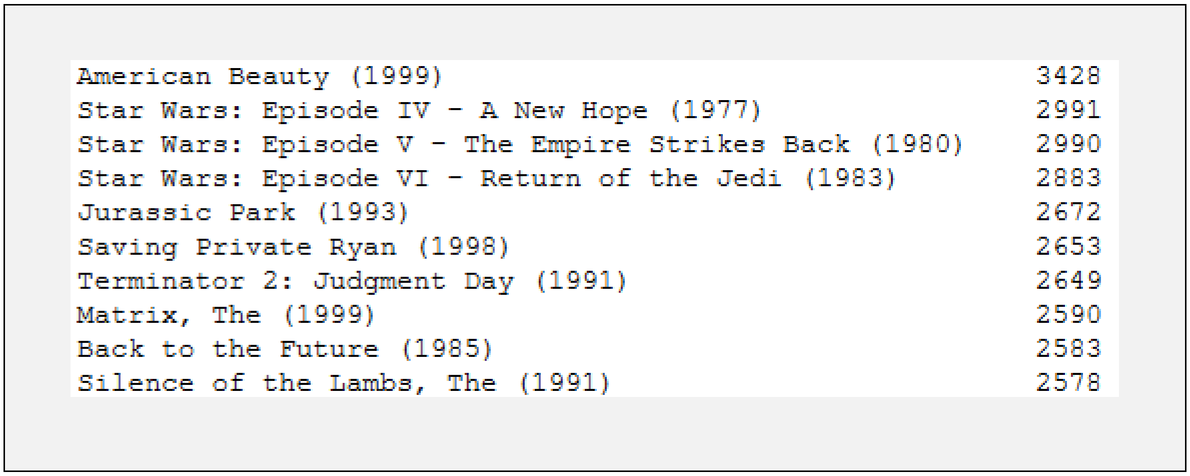
# of movies with average rating of 4.5	29
# of movies with average rating of 4.5 by men	29
# of movies with average rating of 4.5 by women	70

As illustrated above, women have highly rated more movies than men. Because mean is sensitive to outliers we further investigate the number of movies which has a median of 4.5 or higher; additionally, we dissect (profile) this number by segregating men and women with age more than 30 years.

# of movies with median rating of 4.5 or more	92
# of movies with median rating of 4.5 or more among men over 30	105
# of movies with median rating of 4.5 or more among women over 30	187

1.3 Popular Movies

We believe ratings and popularity are two different concepts. Just because a movie is highly rated, it does not make it popular. For example, several art movies in Cannes festival are critically acclaimed and highly rated. Unfortunately, this does not translate into making it a block buster, both in terms of popularity and box office collection. If a movie is rated by large number of users, it is popular; hence we count the number of ratings per movie and present the top 10 popular movies.



```
American Beauty (1999)                                    3428
Star Wars: Episode IV - A New Hope (1977)                 2991
Star Wars: Episode V - The Empire Strikes Back (1980)     2990
Star Wars: Episode VI - Return of the Jedi (1983)         2883
Jurassic Park (1993)                                      2672
Saving Private Ryan (1998)                                2653
Terminator 2: Judgment Day (1991)                         2649
Matrix, The (1999)                                        2590
Back to the Future (1985)                                 2583
Silence of the Lambs, The (1991)                          2578
```

As evident from the picture above, long running movies typically get multiple chances over long horizon to get rated. Hence we see a lot of movies from 1990's. Additionally, it could be a symptom of the dataset too.

1.3 Ease to Please

We evaluated 2 criteria to determine if a particular age bracket was easier to please than another:

Median: Our hypothesis was if we were to find higher median values across age brackets then we could hypothesize that the particular age bracket was easier to please. Surprisingly, the median across all age bracket is 4. We abandoned this technique.

Omega: Because of the sensitivity of mean, we did not use mean but instead used Omega, which is a summary statistics that takes into account all 4 moments, mean, variance, skewness, and kurtosis. Since Omega is the ratio of upper partial moment to lower partial moment and we use excess above the median we will consider the maximum omega as our decision making criteria.

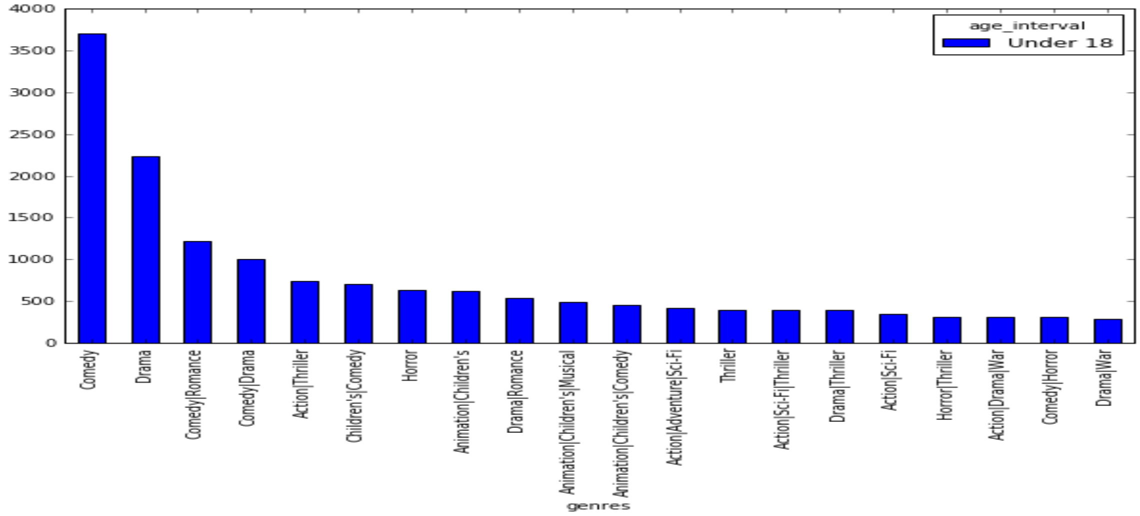
age	omega
18-24	0.631889
25-34	0.651326
35-44	0.681956
45-49	0.691269
50-55	0.697442
56+	0.692048
Under 18	0.608657

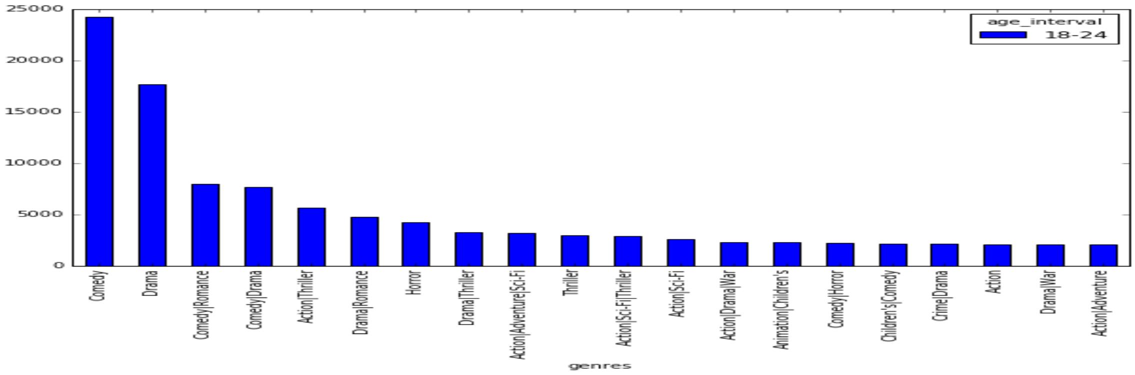
Even these numbers are closely clustered in the 0.6 – 0.7 range. Hence we cannot guess which group is easiest to please.

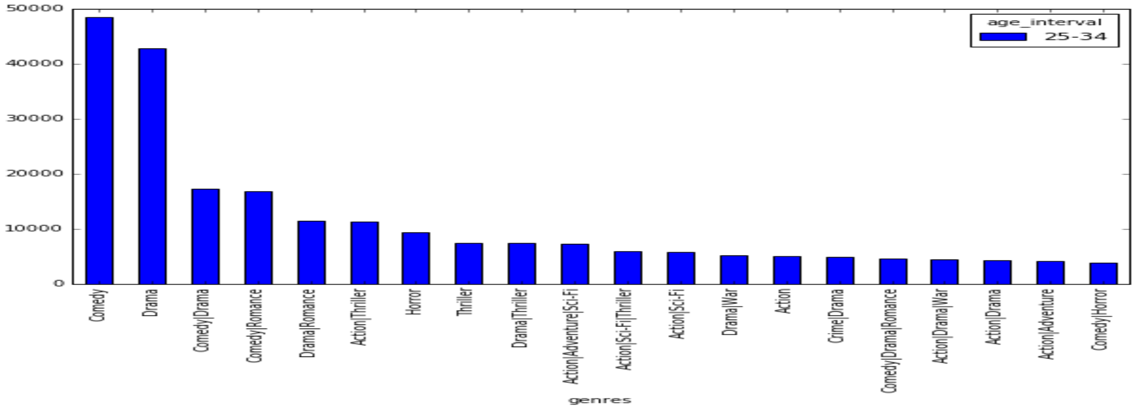
Another Prediction:

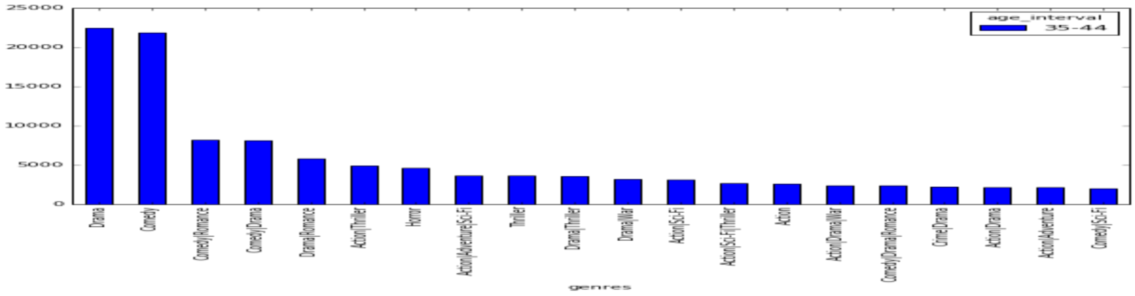
Which genre of movie is most watched by different age group.

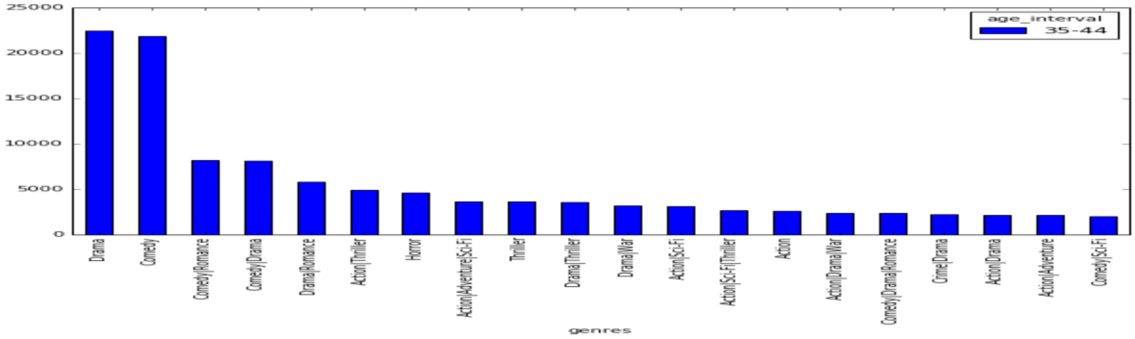
Below is the plot for most watched genre of movie for every age group.




25000
20000
15000
10000
5000
0
age_interval
18-24
Comedy
Drama
Comedy|Romance
Comedy|Drama
Action|Thriller
Drama|Romance
Horror
Drama|Thriller
Action|Adventure|Sci-Fi
Thriller
Action|Sci-Fi|Thriller
Action|Sci-Fi
Action|Drama|War
Animation|Children's
Comedy|Horror
Children's|Comedy
Crime|Drama
Action
Drama|War
Action|Adventure
genres


50000
40000
30000
20000
10000
0
age_interval
25-34
Comedy
Drama
Comedy|Drama
Comedy|Romance
Drama|Romance
Action|Thriller
Horror
Thriller
Drama|Thriller
Action|Adventure|Sci-Fi
Action|Sci-Fi|Thriller
Action|Sci-Fi
Drama|War
Action
Crime|Drama
Comedy|Drama|Romance
Action|Drama|War
Action|Drama
Action|Adventure
Comedy|Horror
genres

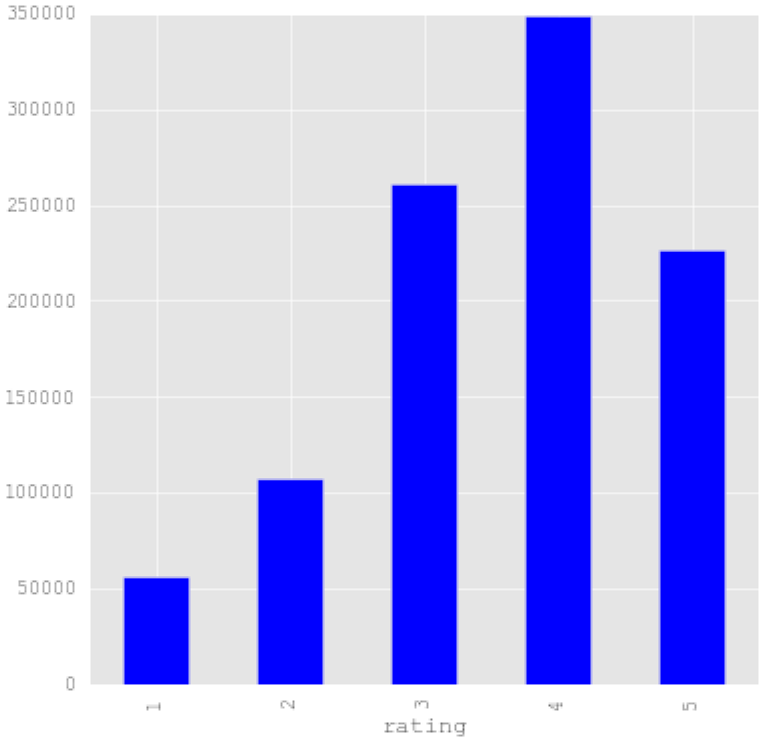

25000
20000
15000
10000
5000
0
age_interval
35-44
Drama
Comedy
Comedy|Romance
Comedy|Drama
Drama|Romance
Action|Thriller
Horror
Action|Adventure|Sci-Fi
Thriller
Drama|Thriller
Drama|War
Action|Sci-Fi
Action|Sci-Fi|Thriller
Action
Action|Drama|War
Comedy|Drama|Romance
Crime|Drama
Action|Drama
Action|Adventure
Comedy|Sci-Fi
genres


25000
20000
15000
10000
5000
0
age_interval
35-44
Drama
Comedy
Comedy|Romance
Comedy|Drama
Drama|Romance
Action|Thriller
Horror
Action|Adventure|Sci-Fi
Thriller
Drama|Thriller
Drama|War
Action|Sci-Fi
Action|Sci-Fi|Thriller
Action
Action|Drama|War
Comedy|Drama|Romance
Crime|Drama
Action|Drama
Action|Adventure
Comedy|Sci-Fi
genres

Although, the rating of each group varies for different genres, ***Drama and Comedy*** *are the most watched genres across all the age groups. It's easy to please all the age groups with drama and comedy movies to watch.*

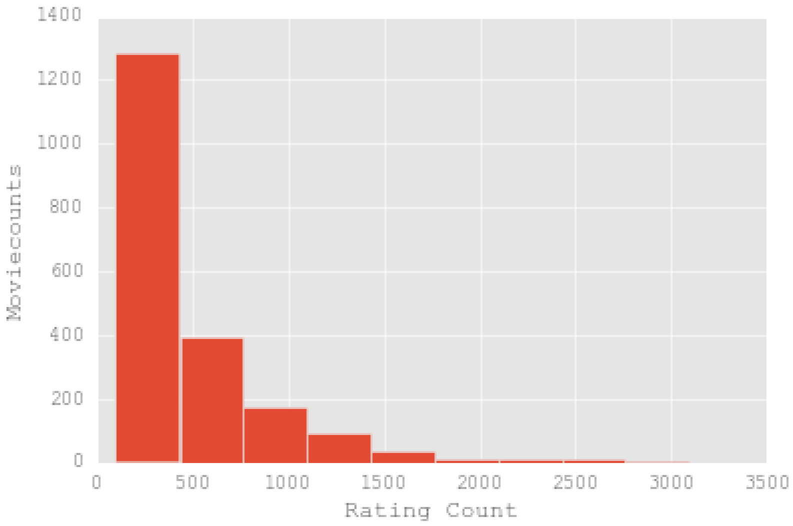
2.1 Ratings of All Movies

The histogram below shows rating of all movies in our dataset.



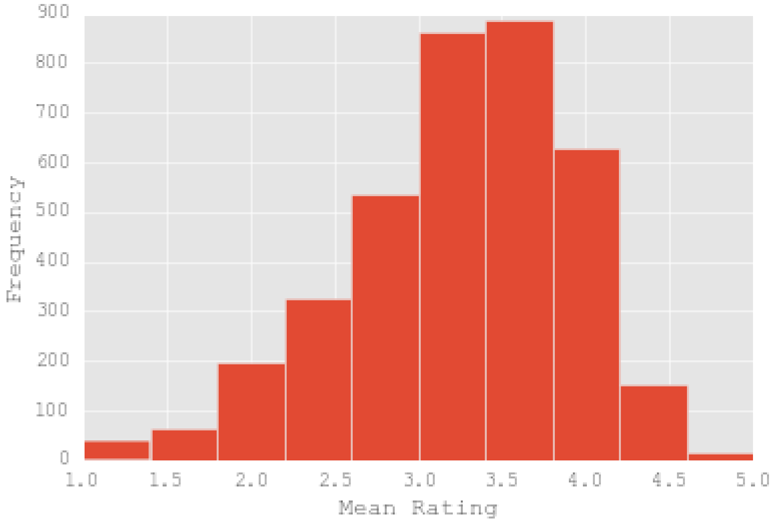
2.2 Number of Ratings each movie received

The histogram below shows the number of ratings each movie received.



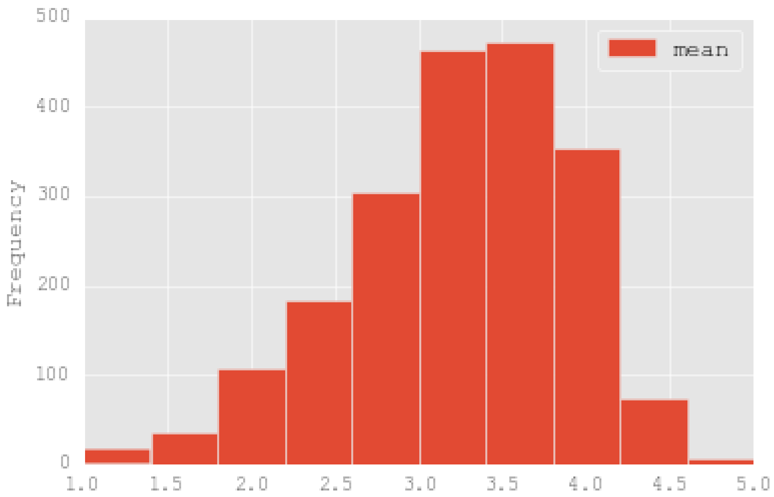
2.3 Average Ratings by movie

The histogram below shows the number of ratings each movie received.



2.4 Average Ratings by movie which are rated more than 100 times

The histogram below shows the number of ratings each movie that received at least 100 ratings. This is done to reduce sparsity and provide consistency in our rating histogram.



2.4.1 Tail Distribution Analysis

The tail is much fatter in case of "all movies" compared to movies with greater than 100 count. which means more data is distributed at the extreams in case of "all movies" and less data is distributed around the average.

2.4.2 Confidence in Rating

With the given dataset, movies rated more than 100 times can be trusted over ones that are rated less than 100 times, as we can see the tail is much fatter with the latter one.

3.2.4.3 Conjecture about the distribution if ratings

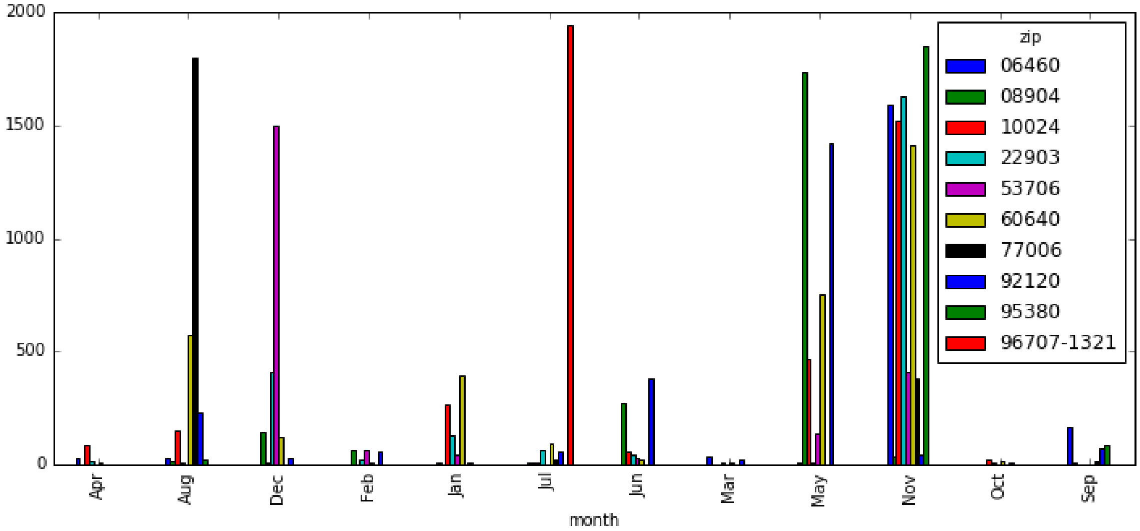
To analyse the rating distribution for different zip codes in the dataset throughout the year.

Top 10 zip codes where maximum participation in rating is determined based on the rating count. here is the top 10 zip codes:

1. 96707-1321 - Hawaii
2. 95380 - CA
3. 77006 - TX

4. 08904 - NJ
5. 22903 - VA (virginia)
6. 06460 - CT
7. 10024 - NY
8. 53706 - WI
9. 92120 - CA (SD)
10. 60640 - IL

plot a graph to show the distribution of the ratings throughout the year for every month.



As we can see Nov is where we have lot of ratting traffic followed by May.

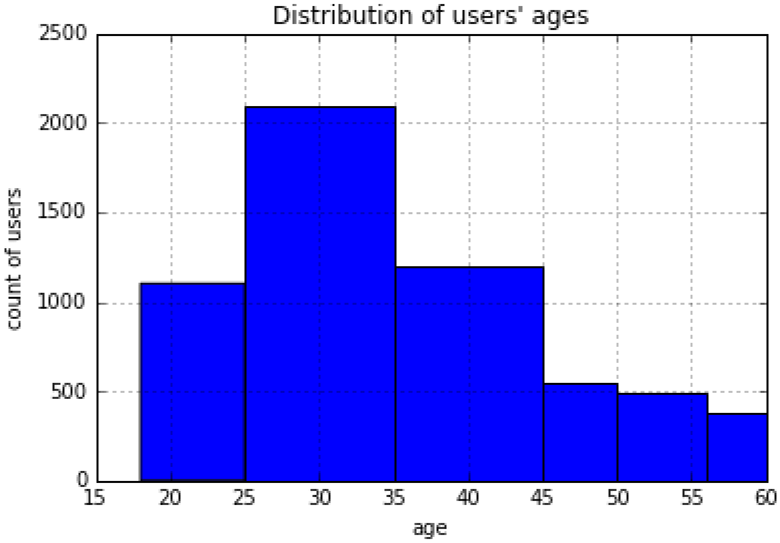
TOP 3 zip codes with highest rating in NOV is from CT,NJ,VA .

Which means, for the given dataset - people in East coast tend to watch more movies in November, due to "WINTER" or due to holidays "Thanks giving" where families get together as a tradition. Which also means, releasing a movie around NOV would make it more popular.

Analyze rating distribution based on user age group.

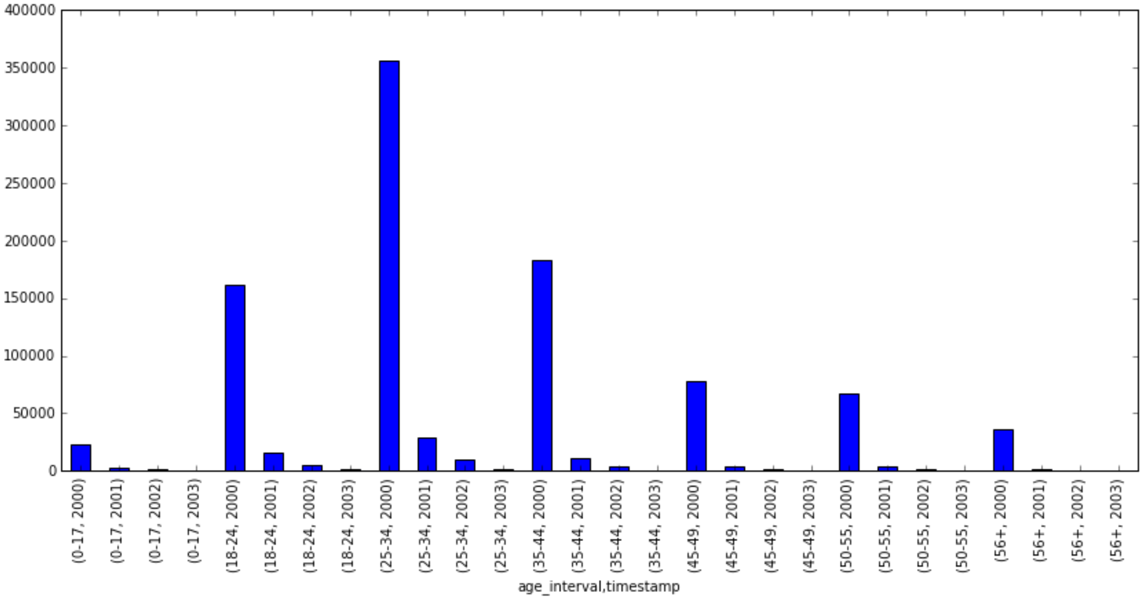
first we would like to see the user distribution in the given dataet.

below is the plot of user distribution by age group.



from the above plot - user group of age 25-35 is relatively much larger. which means, age group “25-35” has maximum participation in the movie ratings and is the age group that watch maximum movies.

Below is the plot on user rating counts for different age group over a period of time from 2000 to 2003.

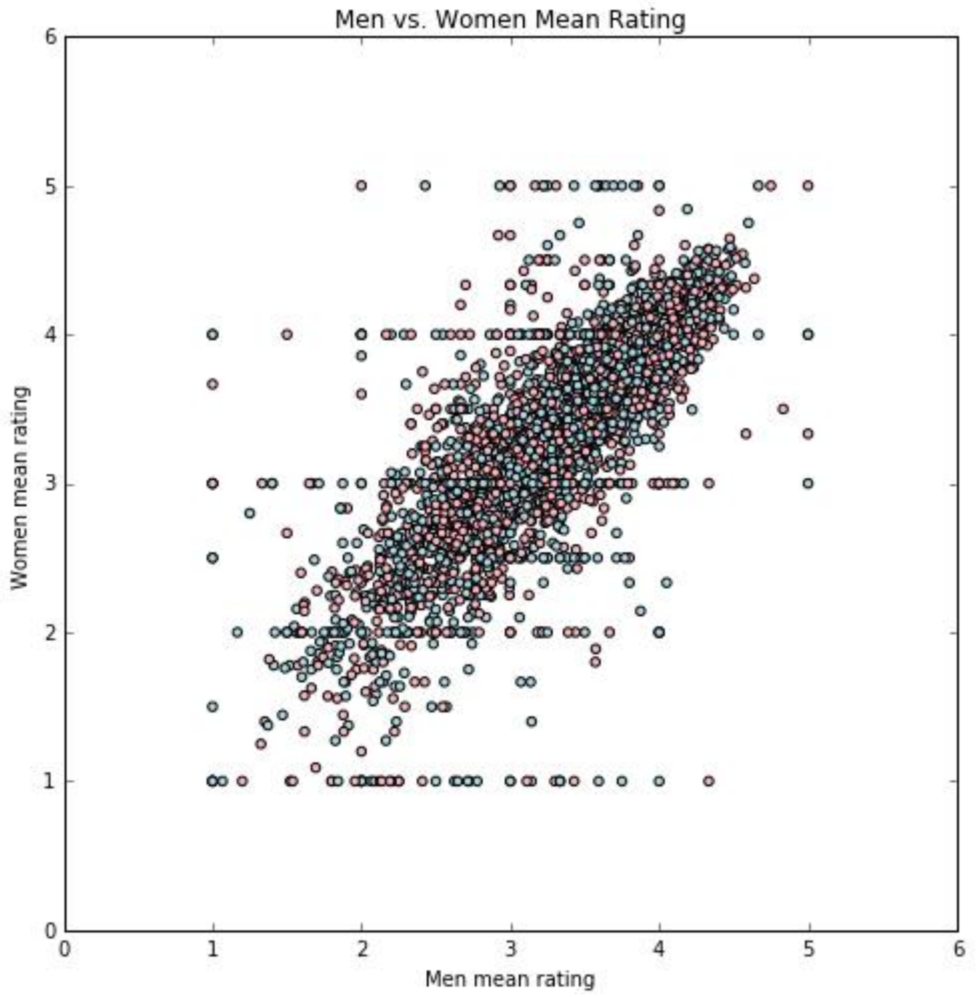


The above plot indicates, most of ratings for the moves in the dataset happened in 2000 and significantly dropped over 2001 - 2003.

We think, Most of the move in our dataset are dated, which means people watched them and rated them in 2000 and new movies were already trending over 2001-2003 so, there were not much rating traffics for dated movies.

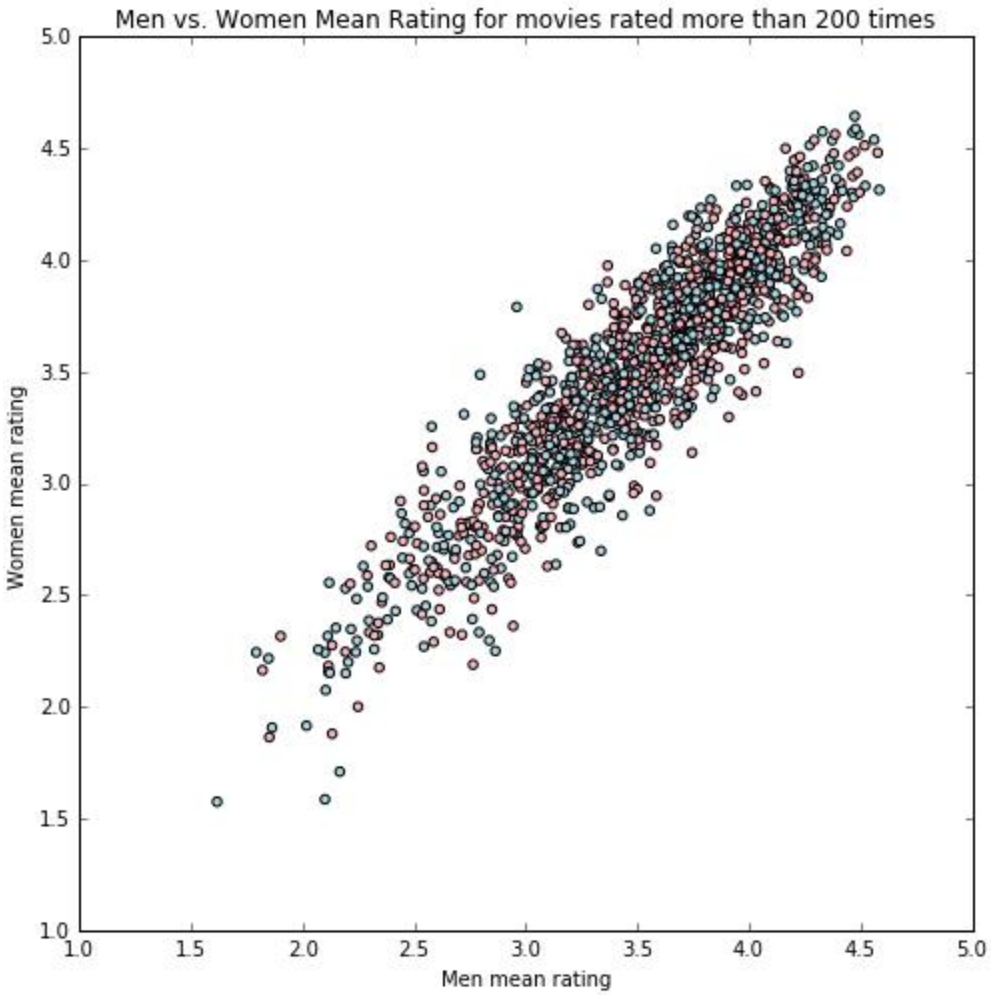
3.1 Men vs. Women Movie Rating Scatter Plot

The scatter plot below shows the mean rating of men versus the mean rating of women. The plot shows the similarity between the ratings.



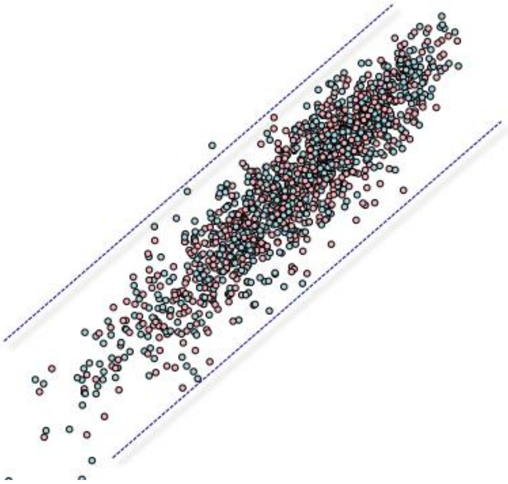
3.2 Scatter Plot for Men/Women mean rating for movies rated > 200 times

As the scatter plot below shows, there is less noise between ratings once we include more number of ratings per movie.



3.3 Correlation Coefficient

The correlation coefficient between men and women is 0.763. This is also evident from the picture above, which is extracted from scatter plot above.



3.4 Are Men ratings similar to Women ratings?

By just looking at the mean and standard deviation between the men and women ratings, we cannot conjecture one gender ratings given the other. To do this we calculate the mean of both men and women across age interval and run a one tailed t-test to determine if the two means are statistically different. We further run a F-test to see if the variance between the 2 groups of men and women across age interval similar. The t-test is presented below.

$$H_0: \mu_x = \mu_y versus\ H_1: \mu_x > \mu_y at\ the\ \alpha\ level\ of\ significance,$$

$$reject\ H_0 if\ t \geq t_{\alpha,n+m,2}$$

$$where, n\ and\ m\ are\ sample\ size\ of\ each\ age\ interval.$$

To calculate t statistics we use the following formula:

$$t = \frac{Mean\ Men\ Age\ Interval_i - Mean\ Women\ Age\ Interval_i}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Age Interval	Mean Men Rating	Mean Women Rating	T-Statistics
Under 18	3.517461	3.616291	-6.320286517
18 - 24	3.525476	3.453145	11.47357814
25-34	3.52678	3.6067	-18.8007493
35-44	3.604434	3.659653	-9.877633632
45-49	3.627942	3.663044	-4.316414113
50-55	3.687098	3.79711	-12.08298342
56+	3.720327	3.915534	-15.43630283

All the values are critical and null hypothesis can be rejected in every case. The mean rating between men and women for each age interval is different.

Using an F-Test, we further test if the variance of men and women across age interval similar. We fail to reject the null hypothesis.

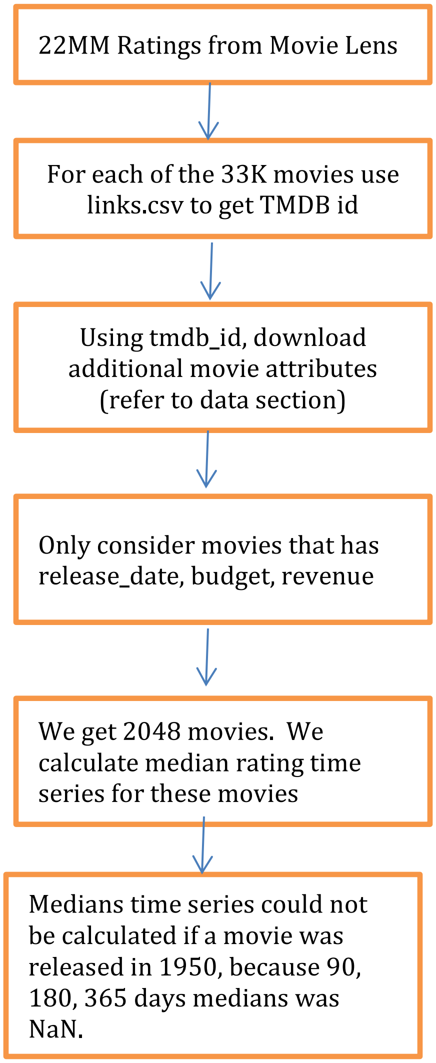
Age Interval	StdDev Men Rating	StdDev Women Rating	F-Statistics	Critical Value
Under 18	1.214797	1.192325	0.963345069	1.030677124
18 - 24	1.16167	1.17729	1.027073118	1.012684624
25-34	1.132786	1.106069	0.953385835	1.008827354
35-44	1.078132	1.076955	0.997817786	1.012159543
45-49	1.062387	1.072365	1.018872326	1.017956132
50-55	1.069039	1.033605	0.934807321	1.020234716
56+	1.066283	1.036587	0.945075589	1.028280745

Based on the 1MM dataset we find it interesting that the mean of men and women are statistically different, however, the standard deviation is statistically not different.

4. Business Intelligence

To study how the median movie ratings move within first 90, 180, 365, and finally greater than 365 days was an interesting and challenging problem to us. We hypothesized that there might be biases in the ratings and positive early ratings often resulted in better profitability.

For this experiment we use 22MM movie ratings from movie lens and for each of the 33043 movies we download additional movie attributes such as release date, production companies, director, revenue, budget, etc. 1MM dataset was limiting in 2 ways: first, the movie attributes were absent so no economic study such as how much did it take to make the movie or what was the revenue collection at the box office could be performed; second, the sample set was small.


22MM Ratings from Movie Lens
For each of the 33K movies use links.csv to get TMDB id
Using tmdb_id, download additional movie attributes (refer to data section)
Only consider movies that has release_date, budget, revenue
We get 2048 movies. We calculate median rating time series for these movies
Medians time series could not be calculated if a movie was released in 1950, because 90, 180, 365 days medians was NaN.

Data quality has been extremely challenging. For some movies the budget was $1, and for some revenue was $100. To provide consistency we only consider movies that have a budget of $100,000.

After removing the NaN, we extensively study the 1569 movies.

- Out of 1569 movies, 81% were profitable and 19% lost money. Profitability is defined as (revenue – budget) / budget. We conjecture if there is a sampling bias in this rating. Behaviorally, people would like to be positive trend follower, as opposed to be iconoclastic.
- Additionally, the number of people who saw the profitable movie was 10 folds (7,139,640 vs. 708,765) higher, making us wonder is the ratings skewed towards popular movies (defined as number of people rating a movie)
- After considering a budget floor of $100,000 we segregate the movies into 3 categories: Budget of more than 100MM; 10-100MM;1-10MM and less than 1MM. We do this to remove any biases budget creates and we believe movies in the same bucket such as 100MM will have similar marketing and media planning. The table below summarizes what we find.

100MM Category

Name	Profitability	First 90 day Median Rating	% of Total Number of ratings
Avatar (2009)	1074%	4	13%
Harry Potter and the Deathly Hallows: Part 2 (2011)	962%	4	49%
Titanic (1997)	823%	5	47%
Frozen (2013)	749%	4	7%
Star Wars: Episode I - The Phantom Menace (1999)	704%	4	17%

Name	Loss	First 90 day Median Rating	% of Total Number of ratings
Stealth (2005)	-43%	3	12%
Flushed Away (2006)	-57%	3.5	1%
13th Warrior, The (1999)	-61%	3	38%
Australia (2008)	-62%	3.5	18%
Lone Ranger, The (2013)	-65%	3	4%

10MM – 100MM Category

Name	Profitability	First 90 day Median Rating	% of Total Number of ratings
Intouchables (2011)	3181%	4.25	0.44%
King's Speech, The (2010)	2661%	3.5	0.02%
The Fault in Our Stars (2014)	2443%	4	0.48%
Black Swan (2010)	2422%	4	3.11%
Conjuring, The (2013)	2346%	3.5	9.74%

1MM – 10MM Category

Name	Profitability	First 90 day Median Rating	% of Total Number of ratings
My Big Fat Greek Wedding (2002)	7275%	4	3.5%
Full Monty, The (1997)	7267%	4	1.8%
Paranormal Activity 2 (2010)	5817%	3.5	4.4%
Saw II (2005)	3723%	3.5	0.8%
Juno (2007)	2985%	4	1.7%

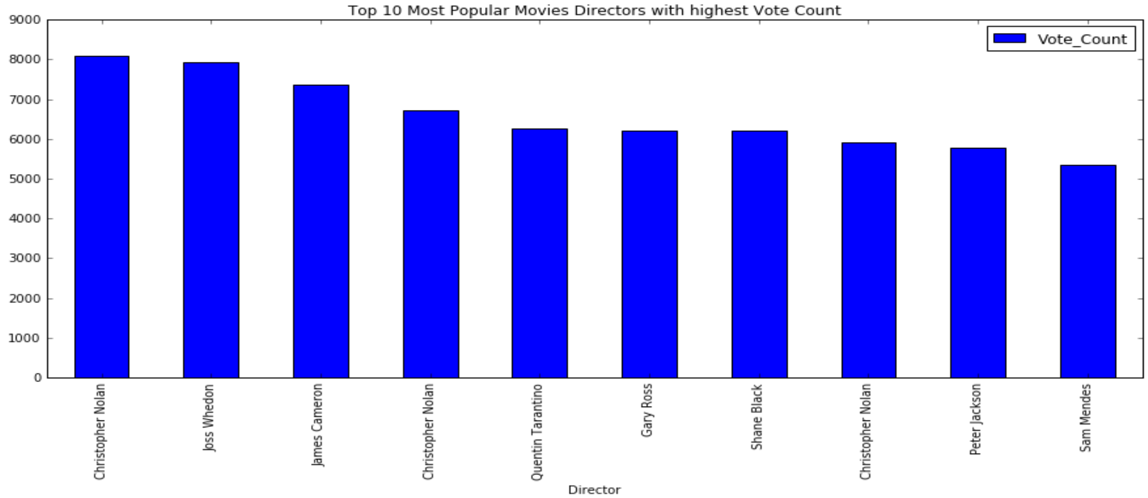
Less than 1MM

Name	Profitability	First 90 day Median Rating	% of Total Number of ratings
Napoleon Dynamite (2004)	11430%	4	1.85%
Swingers (1996)	2153%	4	6.42%
Tadpole (2002)	1828%	3	0.13%
Another Earth (2011)	788%	4	0.01%
Murderball (2005)	483%	4	0.12%

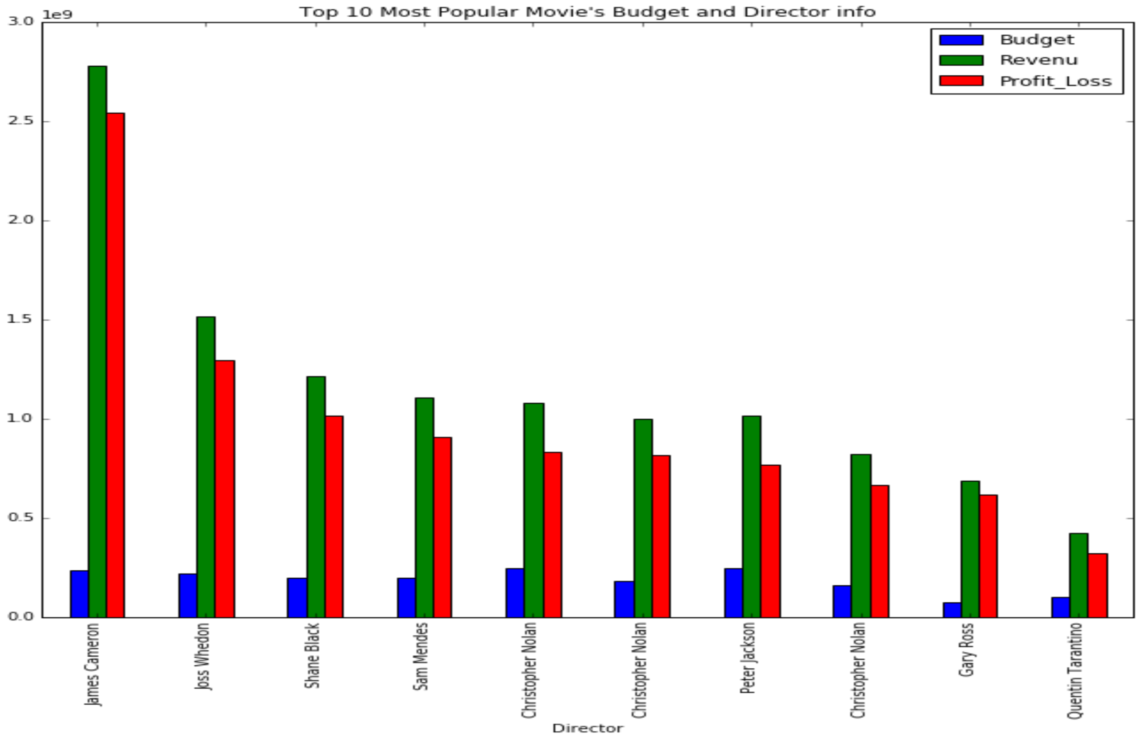
Because of the data quality, we find no loss making movie in the less than 1MM category. We believe this is akin to survivorship bias; why report about a small budget loss making movie?

Business Intelligence observation I:
Based on this dataset we can observe that movies with over 100MM that initially have a high median rating are highly profitable. Moreover, these movies are generally from Action and Sci-Fi genres. This should help producers spend money on social media to create a buzz and work on improving first 90 day median ratings.

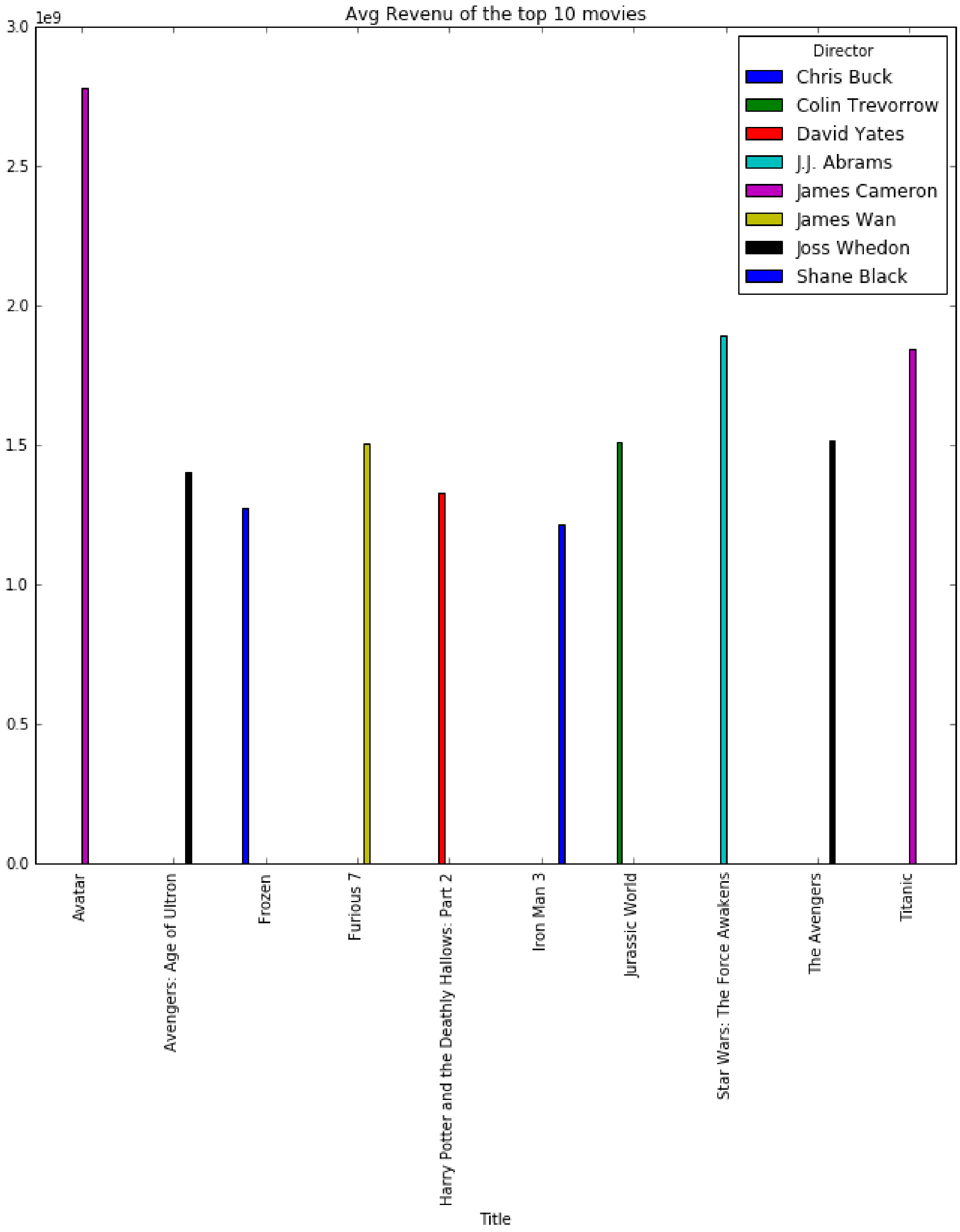
This got us further interested in finding out the directors that were part of these big budget movies. To determine the most profitable Director we leveraged the TMDB database, as mentioned earlier. The plot of TOP 10 movies with maximum vote count (Most popular) mapped to the directors.



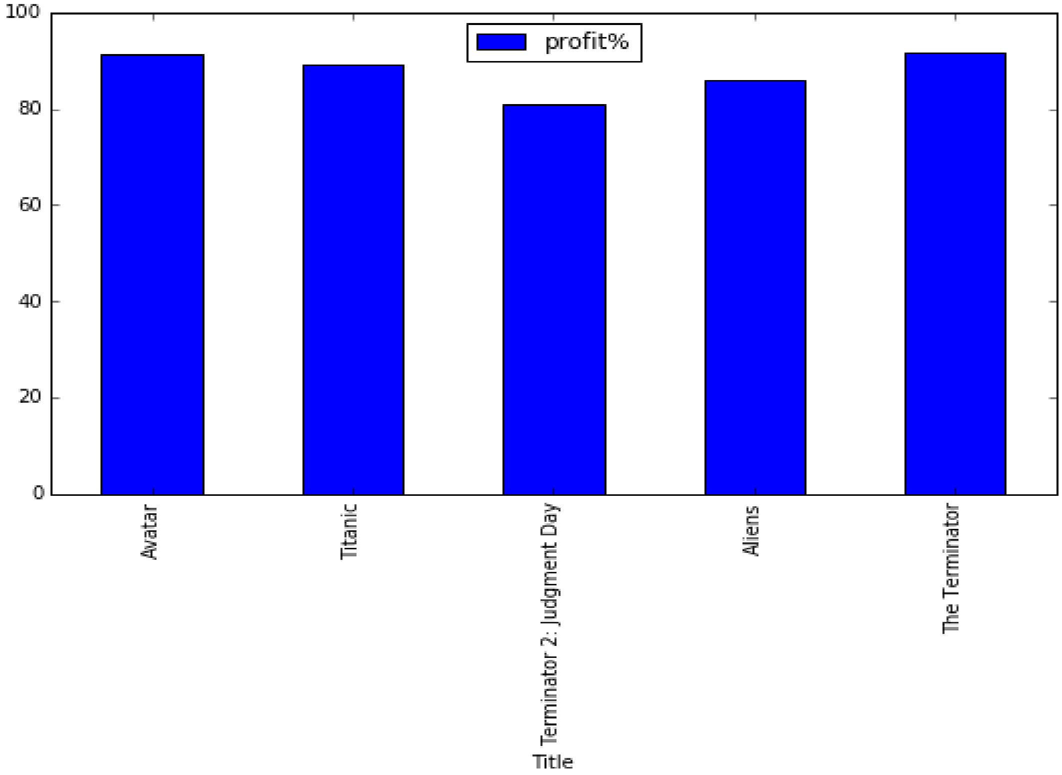
Further analysis on the directors of these top ten movies based on the budgets/revenue:



As the plot above indicates "James Cameron" has the maximum movie with maximum profit. Further, we plotted the top 10 mean revenue of movies.


Avg Revenu of the top 10 movies
1e9
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Director
Chris Buck
Colin Trevorrow
David Yates
J.J. Abrams
James Cameron
James Wan
Joss Whedon
Shane Black
Avatar
Avengers: Age of Ultron
Frozen
Furious 7
Harry Potter and the Deathly Hallows: Part 2
Iron Man 3
Jurassic World
Star Wars: The Force Awakens
The Avengers
Titanic
Title

The plots confirmed that “James Cameron” had the movie with highest profit; both top 2 movies are made by James Cameron. The plot below indicates that generally James Cameron movies are profitable.



Business Intelligence observation II:

If you are making a big budget movie, strongly consider having “James Cameron” as your director. From above analysis, In the given dataset "James Cameron" has made the highest profit movie. Both top two highest profit movies were directed by "James Cameron". The profit margin is at least ~80% for the movies in dataset.