Capstone Project - MovieLens

Recommendation System

Oscar Reyes

March 22, 2019

Abstract

This report is part of the final project capstone to obtain the 'Professional Certificate in Data Science' emited by Harvard University (HarvadX), through edx platform. The main objective is to create a recommendatin system using the MovieLens dataset, and it must be done training a machine learning algorithm using the inputs in one subset to predict movie ratings in the validation set.

Contents

1 Executive Summary

The main purpose of this project is to develop a machine learning algorithm for a movie recommendation system using the MovieLens dataset, in order of predict movie ratings. The entire dataframe can be found at here, but has been used the 10M version of the MovieLens dataset to make the computation a little easier.

The recommendation system will be created using all the tools learned throughout the courses in this series. I applied different dimensionality reduction algorithms: Matrix Factorization and Neighborhood Approach. It can be used to predit the rating of a user baed on an unrated movie. **RMSE** (Root-Mean-Squared-Error) has been applied as the evaluating criteria to analize the algorithm's performance. The principle used for this project is based on this definition of "recommender system":

> *A recommender system or a recommendation system (sometimes replacing "system" with a synonym such as platform or engine) is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item.* Recommender System Definition.

This project could be the base to develop something simmilar to Amazon or Netflix recommendation systems, because a solution like this take users rating and use this information to predict a customer's rating, in order to anticipate the needs of a customer.

2 Introduction

The 10M version of the MovieLens dataset has been used to make the computation a little easier.

2.1 Selected Data

This dataset contains different users' ratings for different movies (rating score between 1 and 5).

Table 1: Amount of Users and Movies

Users	Movies
69878	10677

3 RMSE

The RMSE (Root Mean Squared Errors) will be used to measure que algorithms quality, and the algorithm qualification will be assigned accordign to the next table:

Table 2: RMSE Valoration

Points	RMSE
0	No RMSE reported
5	RMSE >= 0.90000
10	0.88000 <= RMSE <= 0.89999
15	0.87917 <= RMSE <= 0.87999
20	0.87751 <= RMSE <= 0.87916
25	RMSE <= 0.87750

> The goal of this project is to obtain the lowest possible RMSE, because a RMSE is a measurement of error, and the smaller the error, the better.

And, the function used to calculate the RMSE is:

```
# The RMSE function that will be used in this project is:
RMSE <- function(true_ratings = NULL, predicted_ratings = NULL) {
    sqrt(mean((true_ratings - predicted_ratings)^2))
}
```

The RMSE formula is: **RMSE** $= \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$

Table 3: RMSE Formula Values Definition

Variable	Definition
N	Number of Samples
Predicted	Forecasts
Actual	Observed Values

4 Data Preparation and Preprocessing

4.1 Data Exploration

The MovieLens 10M dataset, contains 23371341 rows and 10 columns, with column names: userId, movieId, rating, timestamp, title, genres, dates, date.year, date.year.month, date.year.month.day, and the dataset structure is:

```
## 'data.frame':    23371341 obs. of  10 variables:
##  $ userId             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ movieId            : int  185 185 185 231 316 316 316 329 329 329 ...
##  $ rating             : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ timestamp          : int  838983525 838983525 838983525 838983392 838983392 838983392 838983392 838983392 838983392 838983392 ...
##  $ title              : Factor w/ 10676 levels "...All the Marbles (a.k.a. The California Dolls) (1981)",..: 6758 6758 6758 2871 8997 8997 8997 8980 8980 8980 ...
##  $ genres             : Factor w/ 20 levels "(no genres listed)",..: 2 7 18 6 2 3 17 2 3 9 ...
##  $ dates              : Factor w/ 6519685 levels "1995-01-09 06:46:49",..: 218367 218367 218367 218355 218355 218355 218355 218355 218355 218355 ...
##  $ date.year          : int  1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 ...
##  $ date.year.month    : Factor w/ 157 levels "1995-01","1996-01",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ date.year.month.day: Factor w/ 4640 levels "1995-01-09","1996-01-29",..: 170 170 170 170 170 170 170 170 170 170 ...
```

4.2 DataLens Data Analysis

The '10 first rows' of 'DataLens dataset' are:

Table 4: First 10 Rows

userId	movieId	rating	timestamp	title	genres	dates	date.year	date.year.month	date.year.month.day
1	185	5	838983525	Net, The (1995)	Action	1996-08-02 05:58:45	1996	1996-08	1996-08-02
1	185	5	838983525	Net, The (1995)	Crime	1996-08-02 05:58:45	1996	1996-08	1996-08-02
1	185	5	838983525	Net, The (1995)	Thriller	1996-08-02 05:58:45	1996	1996-08	1996-08-02
1	231	5	838983392	Dumb & Dumber (1994)	Comedy	1996-08-02 05:56:32	1996	1996-08	1996-08-02
1	316	5	838983392	Stargate (1994)	Action	1996-08-02 05:56:32	1996	1996-08	1996-08-02
1	316	5	838983392	Stargate (1994)	Adventure	1996-08-02 05:56:32	1996	1996-08	1996-08-02
1	316	5	838983392	Stargate (1994)	Sci-Fi	1996-08-02 05:56:32	1996	1996-08	1996-08-02
1	329	5	838983392	Star Trek: Generations (1994)	Action	1996-08-02 05:56:32	1996	1996-08	1996-08-02
1	329	5	838983392	Star Trek: Generations (1994)	Adventure	1996-08-02 05:56:32	1996	1996-08	1996-08-02
1	329	5	838983392	Star Trek: Generations (1994)	Drama	1996-08-02 05:56:32	1996	1996-08	1996-08-02

And, a more detailed information of 'DataLens Dataset' is:

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08
##  1st Qu.:18141   1st Qu.:  616   1st Qu.:3.000   1st Qu.:9.472e+08
##  Median :35785   Median : 1748   Median :4.000   Median :1.042e+09
##  Mean   :35886   Mean   : 4277   Mean   :3.527   Mean   :1.035e+09
##  3rd Qu.:53638   3rd Qu.: 3635   3rd Qu.:4.000   3rd Qu.:1.131e+09
##  Max.   :71567   Max.   :65133   Max.   :5.000   Max.   :1.231e+09
##
##                     title                genres
##  Forrest Gump (1994) :  124252   Drama    :3909983
##  Toy Story (1995)    :  118925   Comedy   :3541027
##  Jurassic Park (1993):  117480   Action   :2560458
##  True Lies (1994)    :  114055   Thriller :2325791
##  Aladdin (1992)      :  105785   Adventure:1908934
##  Batman (1989)       :   97172   Romance  :1711761
##  (Other)             :22693672   (Other)  :7413387
##                    dates           date.year     date.year.month
##  1996-02-29 19:00:00:     871   Min.   :1995   1999-12:  684067
##  2005-07-26 14:24:47:     155   1st Qu.:2000   2000-11:  616617
##  1996-04-15 05:23:54:     109   Median :2003   1999-10:  528296
##  2001-09-04 00:19:04:     104   Mean   :2002   2005-03:  527224
##  1996-03-29 12:04:19:      99   3rd Qu.:2005   1996-06:  389737
##  1996-03-28 17:58:30:      98   Max.   :2009   1999-11:  364225
##  (Other)            :23369905                  (Other):20261175
##  date.year.month.day
##  2000-11-20:  142740
##  2005-03-22:  116954
##  1999-12-11:  107106
```

```
## 2008-10-29:   93329
## 2000-11-21:   82538
## 1999-12-12:   79147
## (Other)   :22749527
```

5 Methods & Analysis - Visualize the Importance of Variables

5.1 All Data

Each variable and its amount in the data set is:
<Dates are grouped by month>

In the table we can see the total amount of each field in the dataset:

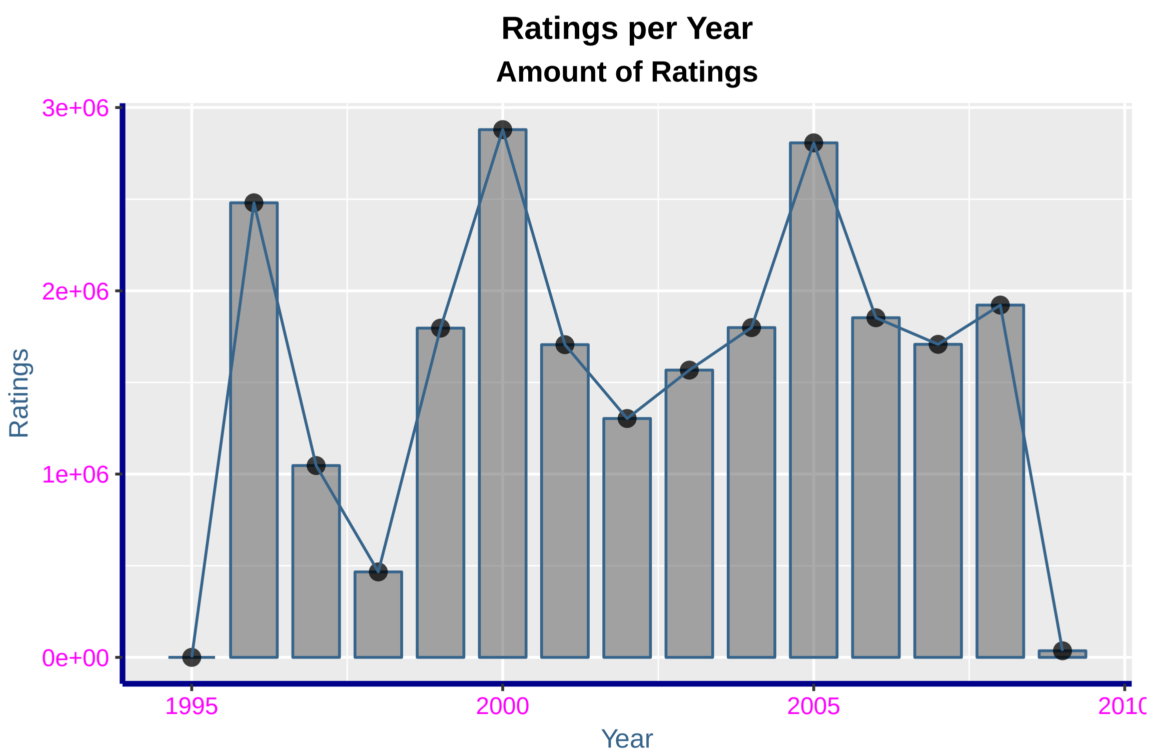
Table 5: Total Amount of each Field

Field	Amount
Dates - Year	15
Dates - Month	157
Genres	20
Ratings	10
Titles	10676
Users	69878

5.2 Analysis by Date (timestamp)

The dataset contains information of 15 years, since: 1995 to 2010. And, we can see the behavior of ratings over the years:

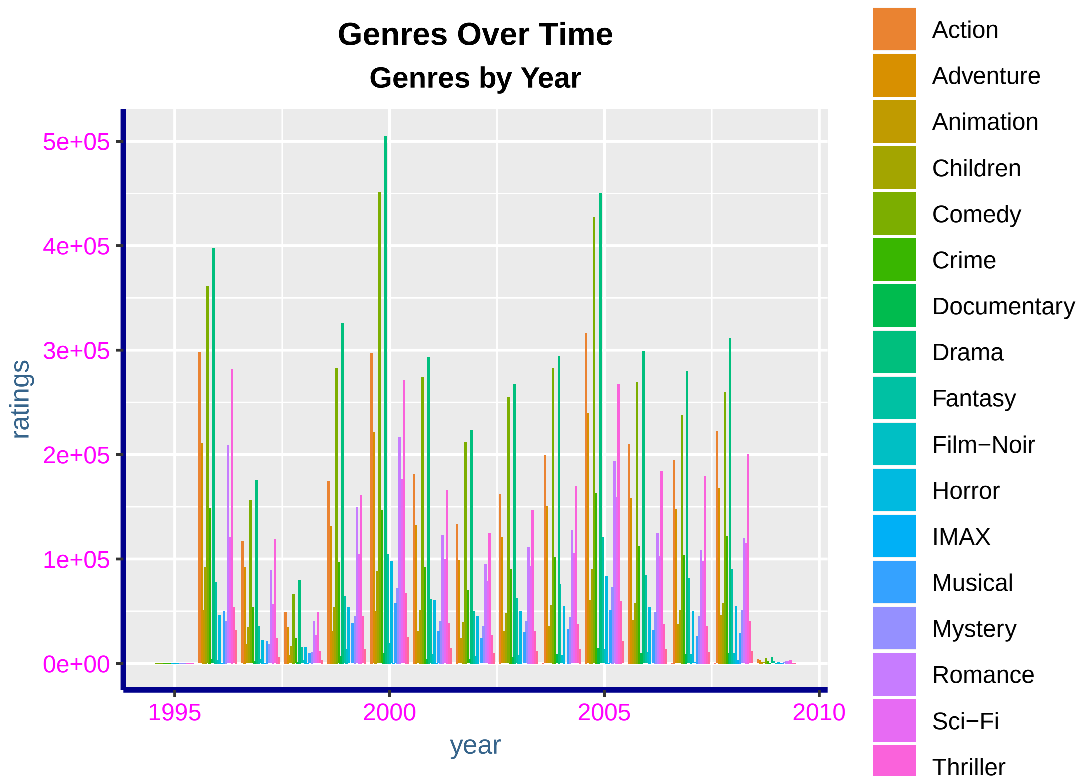
```
## Bar Graph
```



An evaluation of ratings per year won't let us to identify the year with most ratings amount, because the behavior was irregular.

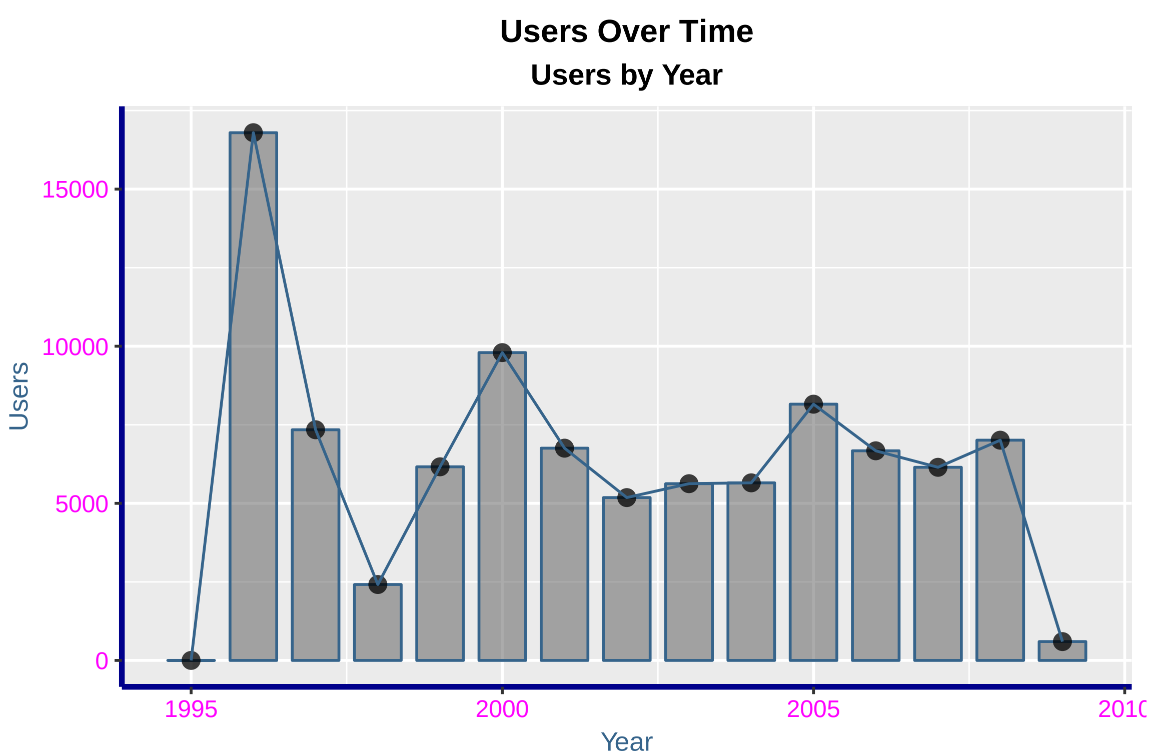
And, an evaluation of genres rating over the years:

```
## Col Graph
```



Users by year:

```
## Bar Graph
```



It won't be useful to add date into overall prediction, as result of the analysis of previous graphics, in which we can see that the year does not represent an evident influence over the ratings, but nevertheless, if we make an evaluation of successful movies on each year, it could be a point of analysis. But, this is not the case.

5.3 Analysis by Genres

After separating all genres in the Data, we have obtained a total of 20 different genres, the following table shows the genres list and the amount of times that each one appear on data:

Amount of movies per genres:

```
## Descendent order
```

Table 6: Top 10 Genres

genres	count
Drama	3909983
Comedy	3541027
Action	2560458
Thriller	2325791
Adventure	1908934
Romance	1711761
Sci-Fi	1341297
Crime	1327780
Fantasy	925654
Children	738267

Drama, Comedy, Action, and Thriller are the most likely rated, which movies are the most rated?

```
## Descendent order
```

Table 7: Top 10 Rated Movies

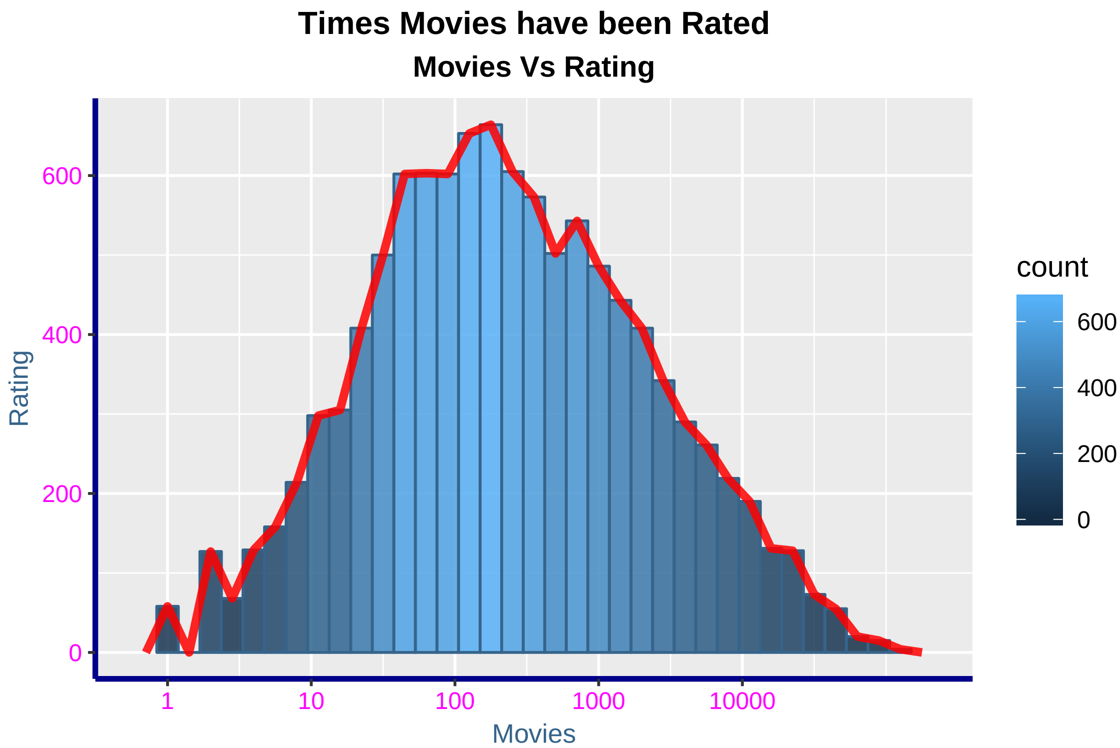
genres	title	count
Comedy	Pulp Fiction (1994)	31388
Crime	Pulp Fiction (1994)	31388
Drama	Pulp Fiction (1994)	31388
Comedy	Forrest Gump (1994)	31063
Drama	Forrest Gump (1994)	31063
Romance	Forrest Gump (1994)	31063
War	Forrest Gump (1994)	31063
Crime	Silence of the Lambs, The (1991)	30327
Horror	Silence of the Lambs, The (1991)	30327
Thriller	Silence of the Lambs, The (1991)	30327

The amount of movies per rating:

Table 8: Amount of Movies per Rating, with Different ID

rating	movies
3.0	10209
4.0	9960
3.5	9798
2.0	9479
2.5	9386
5.0	8575
4.5	8275
1.0	8263
0.5	7195
1.5	7103

Graph of Number of Movies Vs Number of Ratings:



5.4 Analysis by Rating & Year

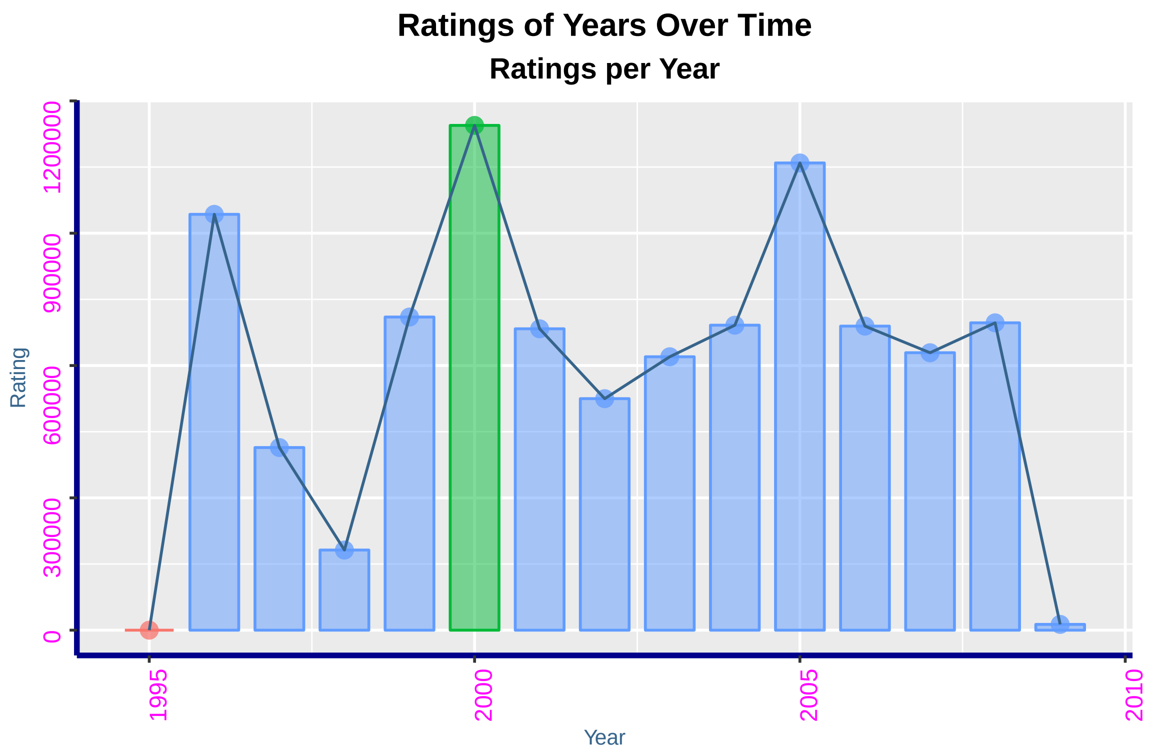
Most rated year: 2000, 1144387
Less rated year: 1995, 2

Table 9: Rating Per Year

date.year	ratings
1995	2
1996	942799
1997	414075
1998	181684
1999	709981
2000	1144387
2001	683261
2002	524918
2003	619900
2004	691430
2005	1059302
2006	689322
2007	629058
2008	696813
2009	13122

The graph of ratings by year is:

```
## Bar Graph Color
```



5.5 Analysis by Rating & Movie

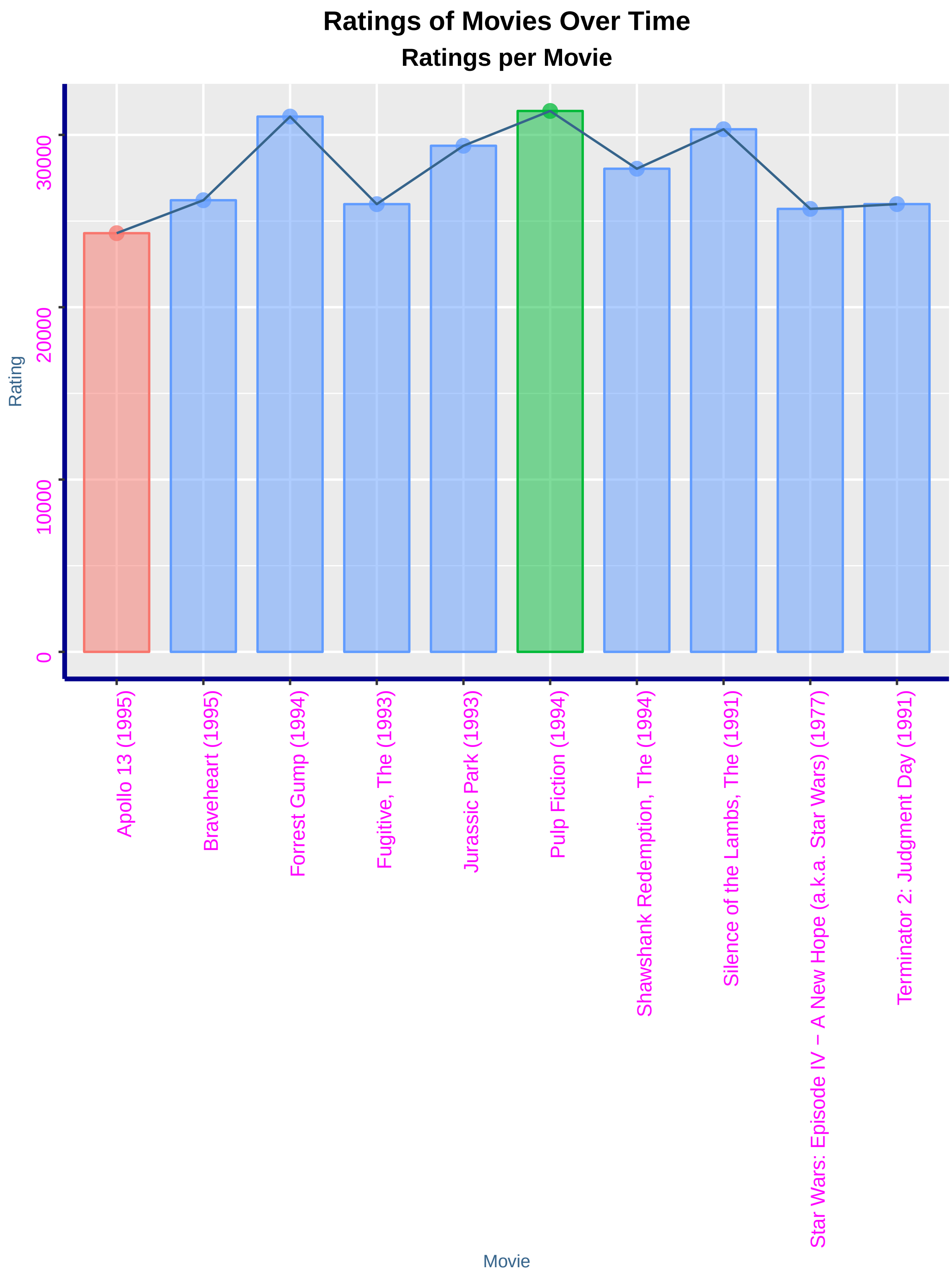
The most rated movie is: Pulp Fiction (1994), 31388
The less rated movie is: 1, 2, 3, Sun (Un, deuz, trois, soleil) (1993), 1

Table 10: Ratings per Movie

title	ratings
Pulp Fiction (1994)	31388
Forrest Gump (1994)	31063
Silence of the Lambs, The (1991)	30327
Jurassic Park (1993)	29370
Shawshank Redemption, The (1994)	28037
Braveheart (1995)	26209
Terminator 2: Judgment Day (1991)	25984
Fugitive, The (1993)	25982
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	25707
Apollo 13 (1995)	24297

The graph of ratings by movie is:

```
## Bar Graph Color
```



5.6 Analysis by Rating & Genre

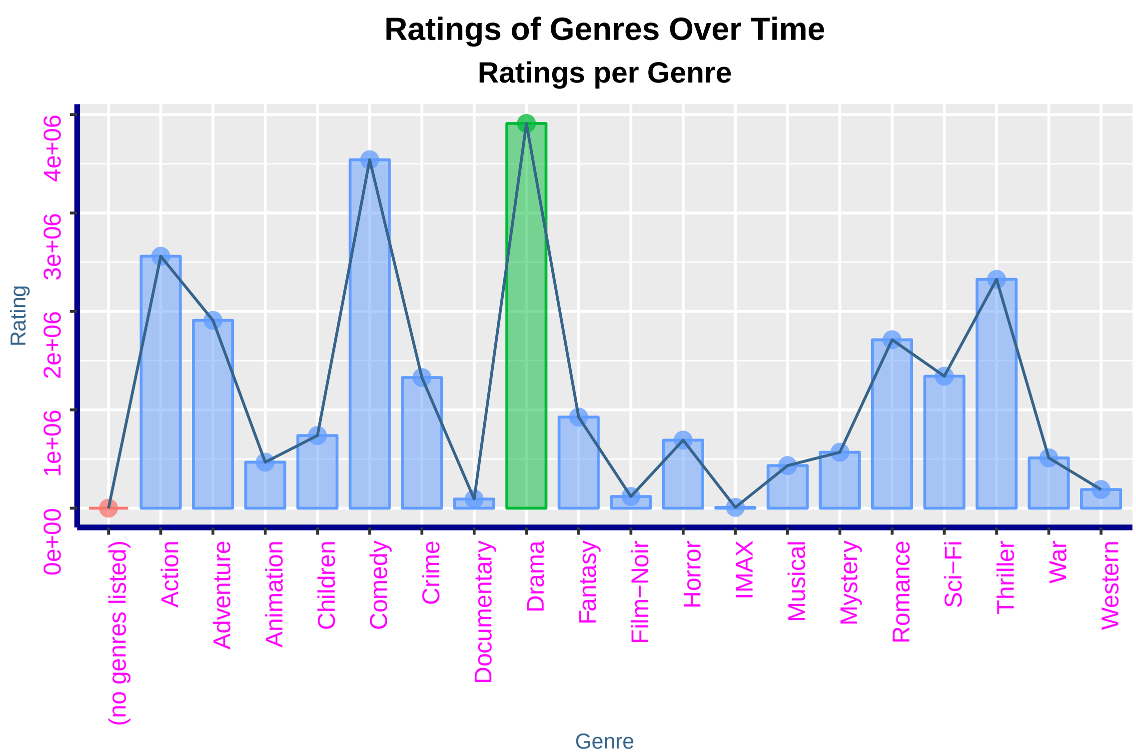
The most rated genre: Drama, 3909983
The less rated genre: (no genres listed), 7

The graph of ratings by genre is:

```
## Bar Graph Color
```

Table 11: Ratings per Genre

genres	ratings
Drama	3909983
Comedy	3541027
Action	2560458
Thriller	2325791
Adventure	1908934
Romance	1711761
Sci-Fi	1341297
Crime	1327780
Fantasy	925654
Children	738267
Horror	691429
Mystery	568333
War	511057
Animation	467357
Musical	433116
Western	189404
Film-Noir	118510
Documentary	93002
IMAX	8174
(no genres listed)	7



5.7 Analysis of Ratings & User

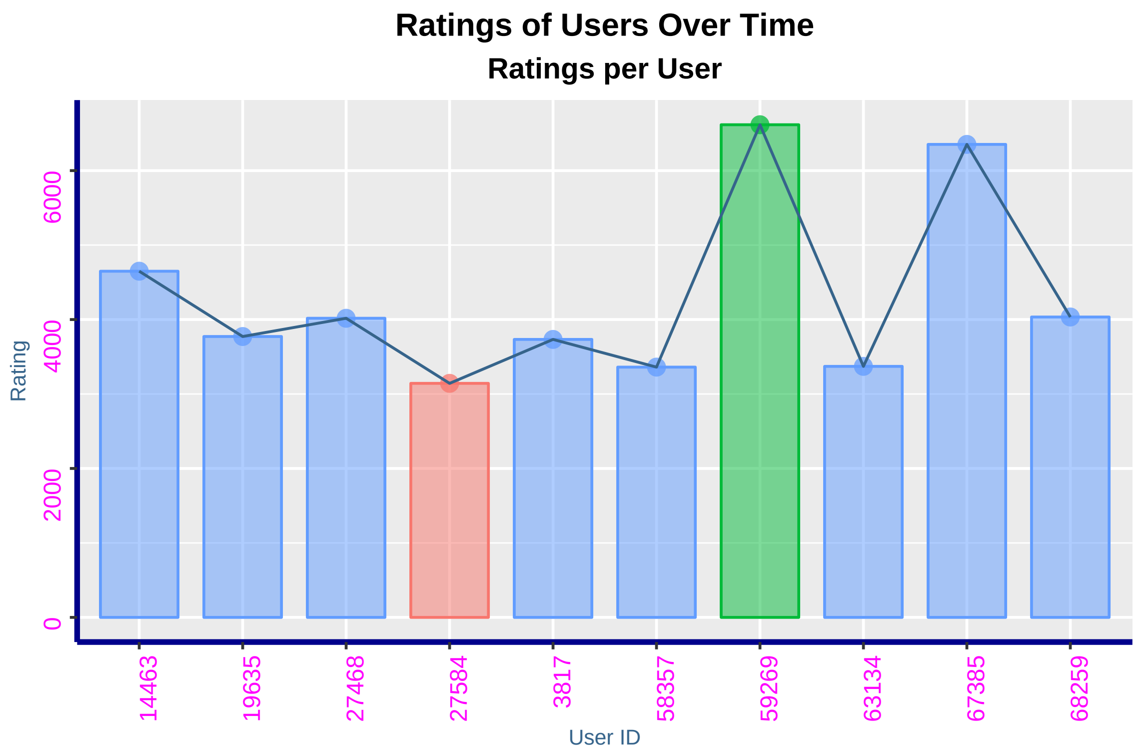
The most user ratings: 59269, 6616
The less user ratings: 62516, 10

The graph of ratings by user is:

Table 12: Ratings per User

userId	ratings
59269	6616
67385	6352
14463	4649
68259	4034
27468	4017
19635	3772
3817	3733
63134	3371
58357	3361
27584	3143

```
## Bar Graph Color
```



5.8 Analysis by Title

The most rated title by year:

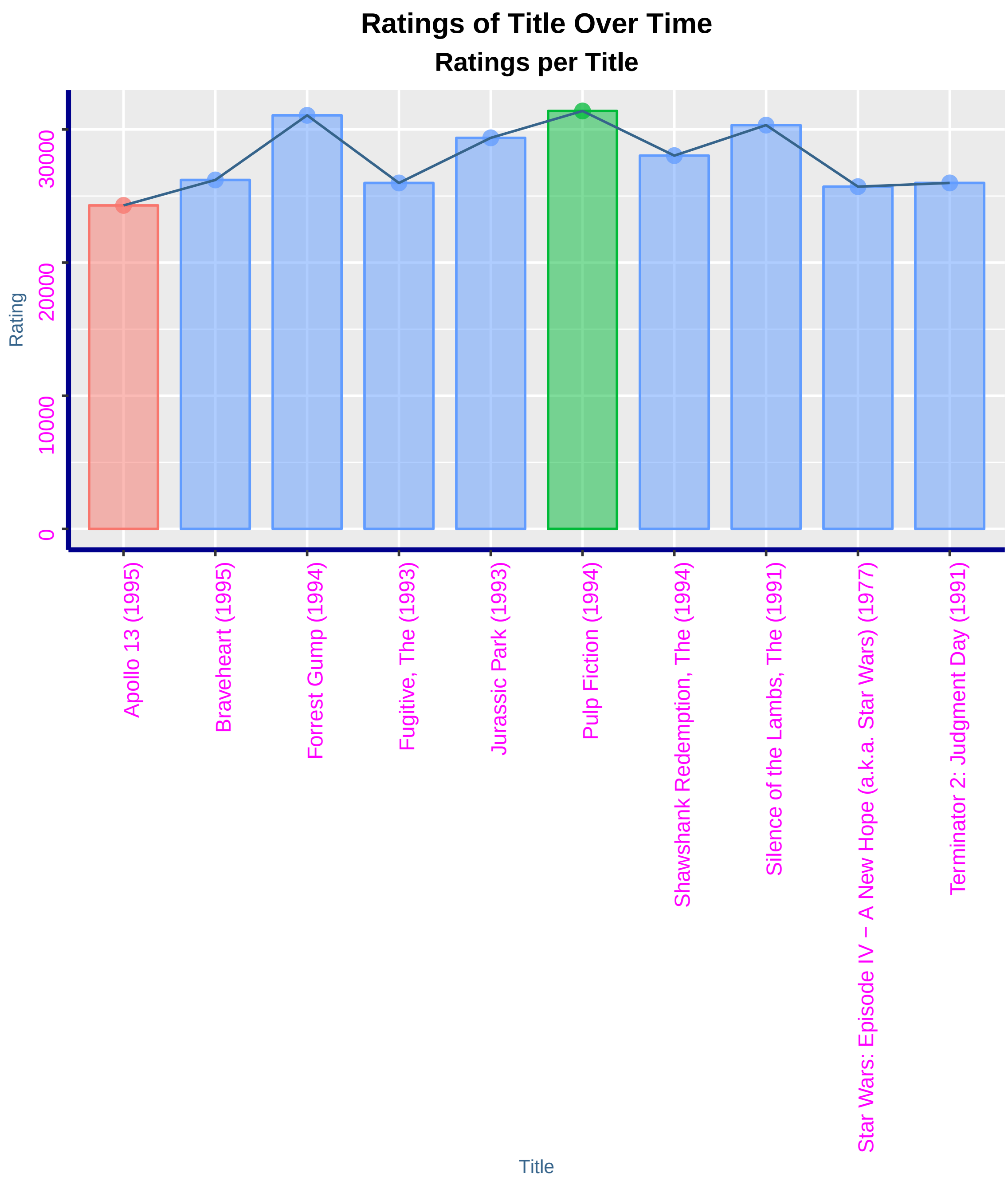
The most rated title: Pulp Fiction (1994), 31388
The less rated title: 1, 2, 3, Sun (Un, deuz, trois, soleil) (1993), 1

Rating per title:

```
## Bar Graph Color
```

Table 13: Rating per Title

title	ratings
Pulp Fiction (1994)	31388
Forrest Gump (1994)	31063
Silence of the Lambs, The (1991)	30327
Jurassic Park (1993)	29370
Shawshank Redemption, The (1994)	28037
Braveheart (1995)	26209
Terminator 2: Judgment Day (1991)	25984
Fugitive, The (1993)	25982
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	25707
Apollo 13 (1995)	24297

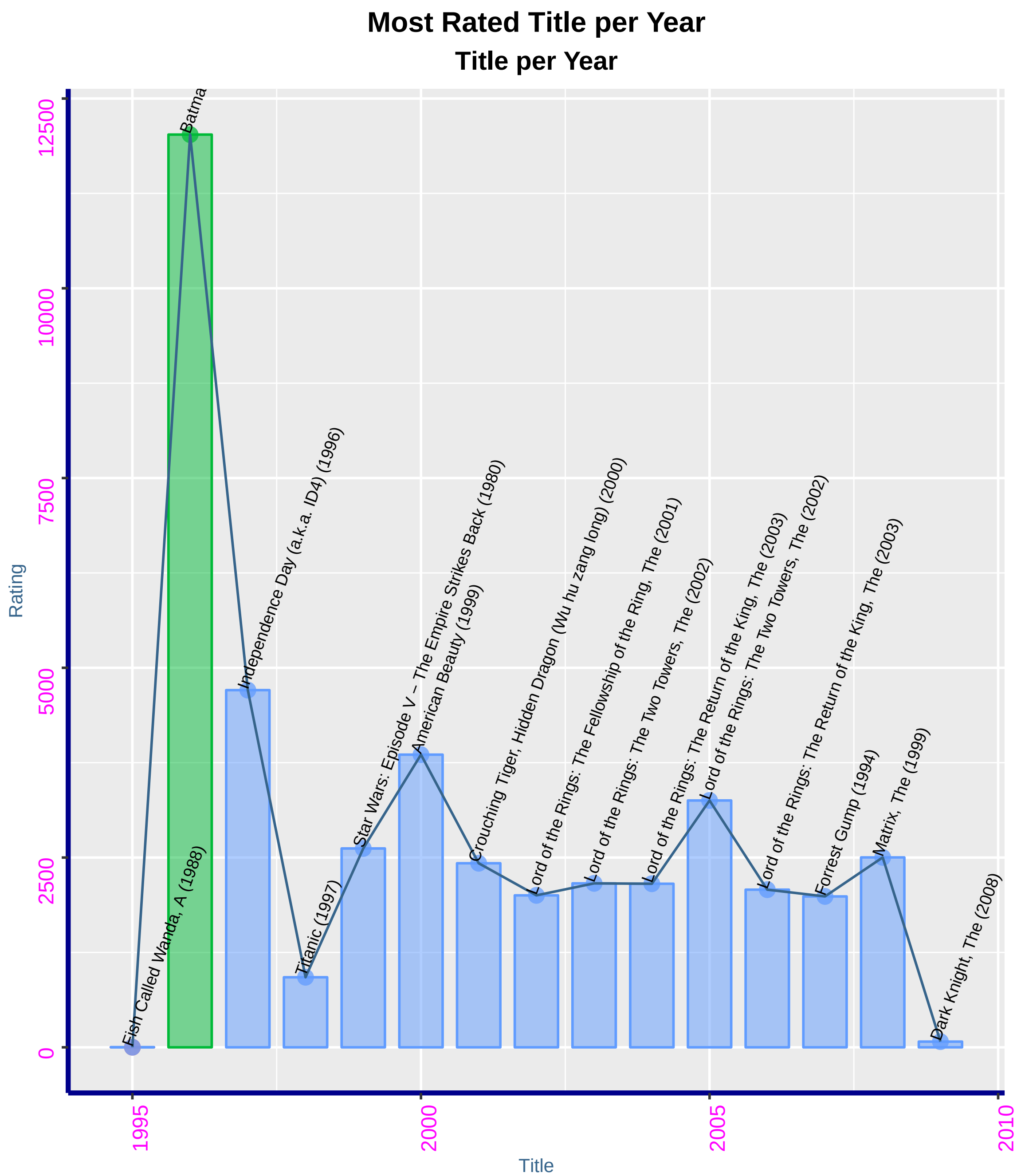


Most rated title per year:

Table 14: Most Rated Title per Year

date.year	title	ratings
1995	Fish Called Wanda, A (1988)	1
1995	Seven (a.k.a. Se7en) (1995)	1
1996	Batman (1989)	12025
1997	Independence Day (a.k.a. ID4) (1996)	4706
1998	Titanic (1997)	923
1999	Star Wars: Episode V - The Empire Strikes Back (1980)	2620
2000	American Beauty (1999)	3856
2001	Crouching Tiger, Hidden Dragon (Wu hu zang long) (2000)	2425
2002	Lord of the Rings: The Fellowship of the Ring, The (2001)	2002
2003	Lord of the Rings: The Two Towers, The (2002)	2160
2004	Lord of the Rings: The Return of the King, The (2003)	2155
2005	Lord of the Rings: The Two Towers, The (2002)	3252
2006	Lord of the Rings: The Return of the King, The (2003)	2077
2007	Forrest Gump (1994)	1988
2008	Matrix, The (1999)	2502
2009	Dark Knight, The (2008)	75

```
## Bar Graph Color
```



5.9 Analysis by Users

A table of user with more ratings:

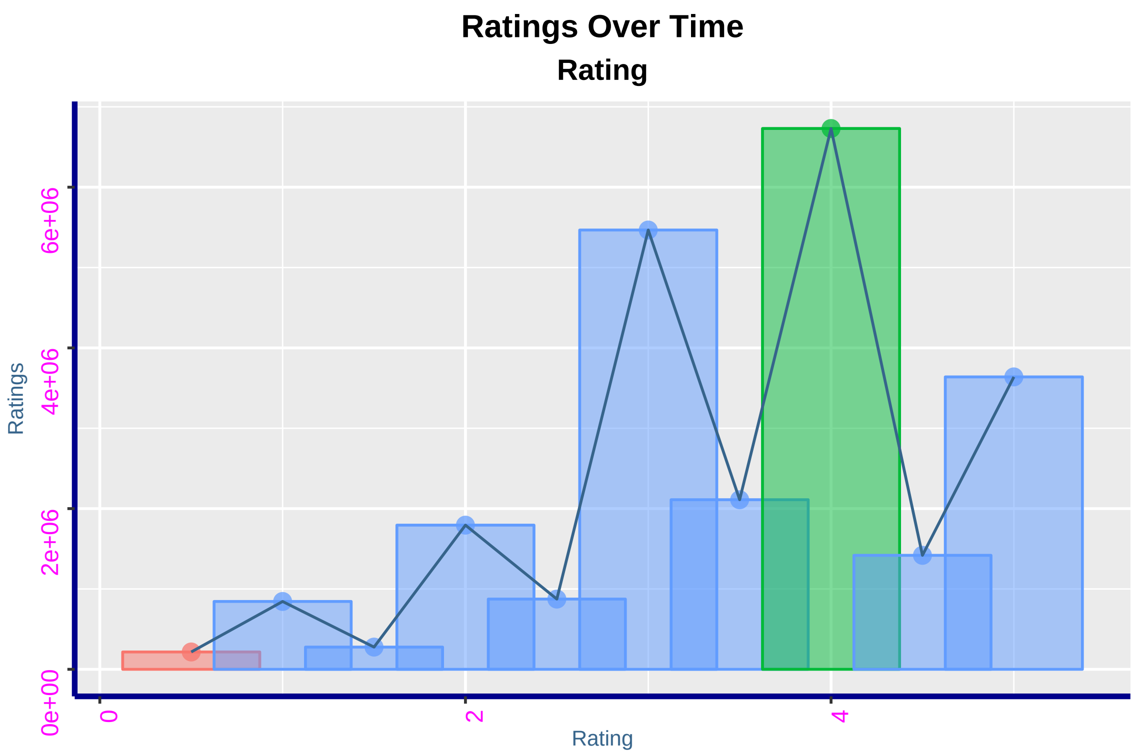
The user with most ratings has the ID: 59269, 6616
The user with less ratings has the ID: 62516, 10

Users rated movies with 4.0 over 28%, more than quarter of time

Graph of user's ratings:

Table 15: Ratings per Rating Value

rating	ratings	percent
4.0	6730401	28.7976672
3.0	5467061	23.3921579
5.0	3639055	15.5705871
3.5	2110690	9.0311035
2.0	1794243	7.6771076
4.5	1418622	6.0699213
2.5	874290	3.7408637
1.0	844336	3.6126981
1.5	276711	1.1839757
0.5	215932	0.9239179



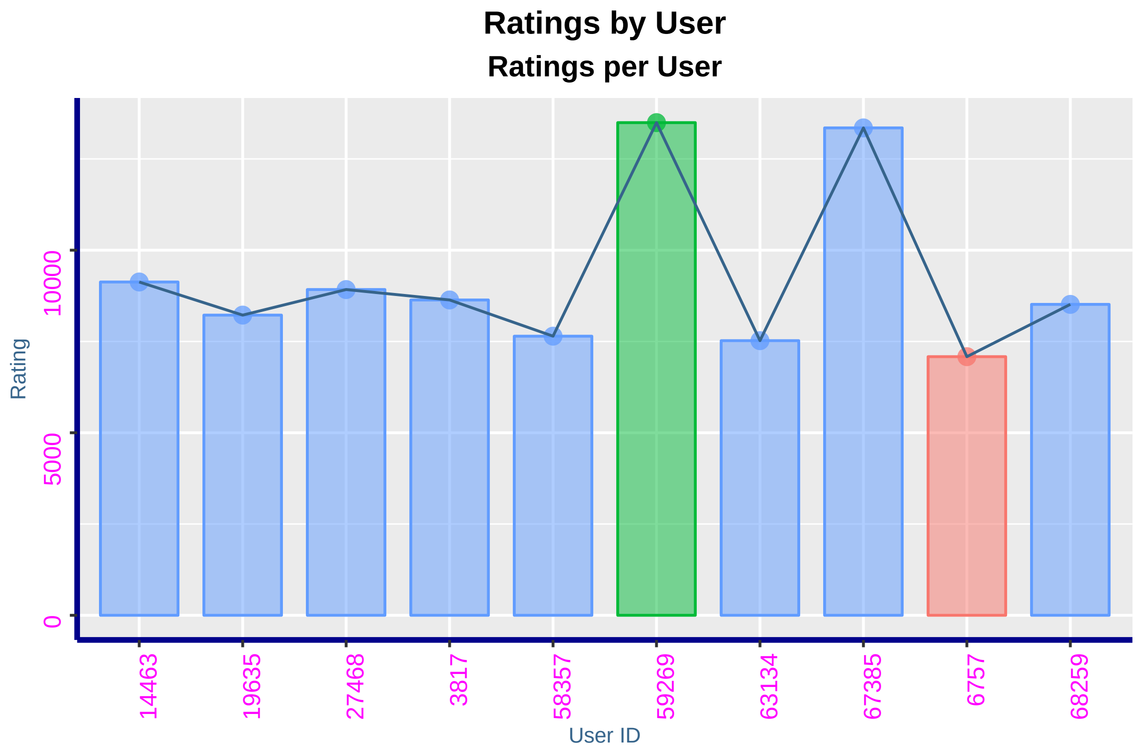
Amount of users per rating:

A table that shows all ratings per user:

Table 16: Ratings per User

userId	ratings
59269	13494
67385	13350
14463	9129
27468	8922
3817	8637
68259	8516
19635	8222
58357	7645
63134	7522
6757	7084

Graph of times that a user has rated a movie:



6 Results

6.1 Model Building & Training

The model used for developing the prediction algorithm follows: the mean rating μ is modified by one or more bias terms b with a residual error ϵ expected.
$Yu, i = \mu + b_i + b_u + b_g + \epsilon_{i,u,g}$

Let's start writing a loss-function that computes the RMSE (Residual Mean Squared Error), as accuracy meassure.

6.2 Baseline Model

Let's start with a baseline model, the most basic recommendation system. This baseline includes the average of all users accross all movies and use the average to predict all ratings:
$Y_{u,i} = \mu + \epsilon_{u,i}$

No is time to predict a new rating to be the average tating of all movies in the training dataset, and it will be the 'Baseline RMSE'.
mu = 3.5270058 and baseline **RMSE** = 1.0522507

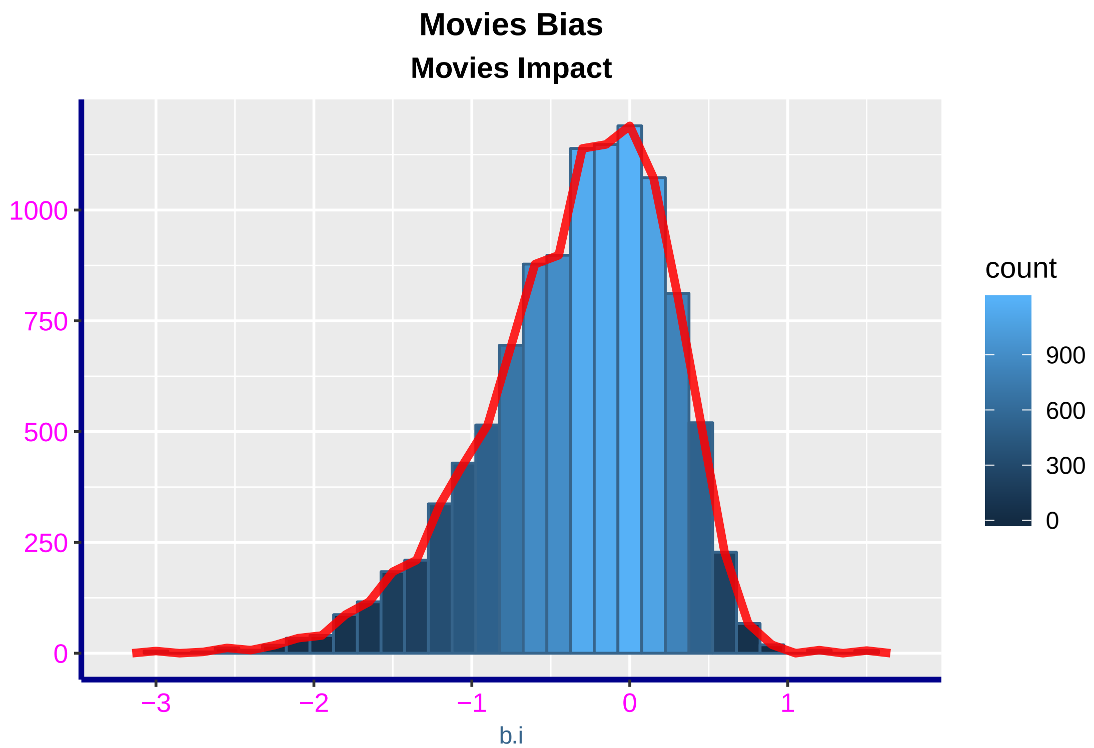
Table 17: RMSEs Comparisson

method	RMSE
Baseline	1.052251

6.3 Movies Bias

In order of improve the model, we will analyze the movies bias effect.

In the next graph we can make a visual evaluation of Movies Bias



An `lm` evaluation is not possible because the dataset is too big, and the computer could crash by memory. The formula is: $Y_{u,i} = \mu + b_i + \epsilon_{u,i}$

To solve the previous restriction, we can estimate the movie bias as $\hat{b_i} = y_{u,i} - \mu$ for each `i` movie. The the equation to use is: $\hat{y_{u,i}} = \hat{\mu} + \hat{b_i}$

In this table we can see the RMSE produced by Movies Bias

Table 18: RMSEs Comparisson

method	RMSE
Baseline	1.0522507
Movies Bias	0.9405627

We can see an improvement of Movies Bias over Baseline.

6.4 Users Bias

Is time for testing the users bias, and evaluate the impact over the model.

Now, is time to see the impact of User Bias over the model.

Table 19: RMSEs Comparisson

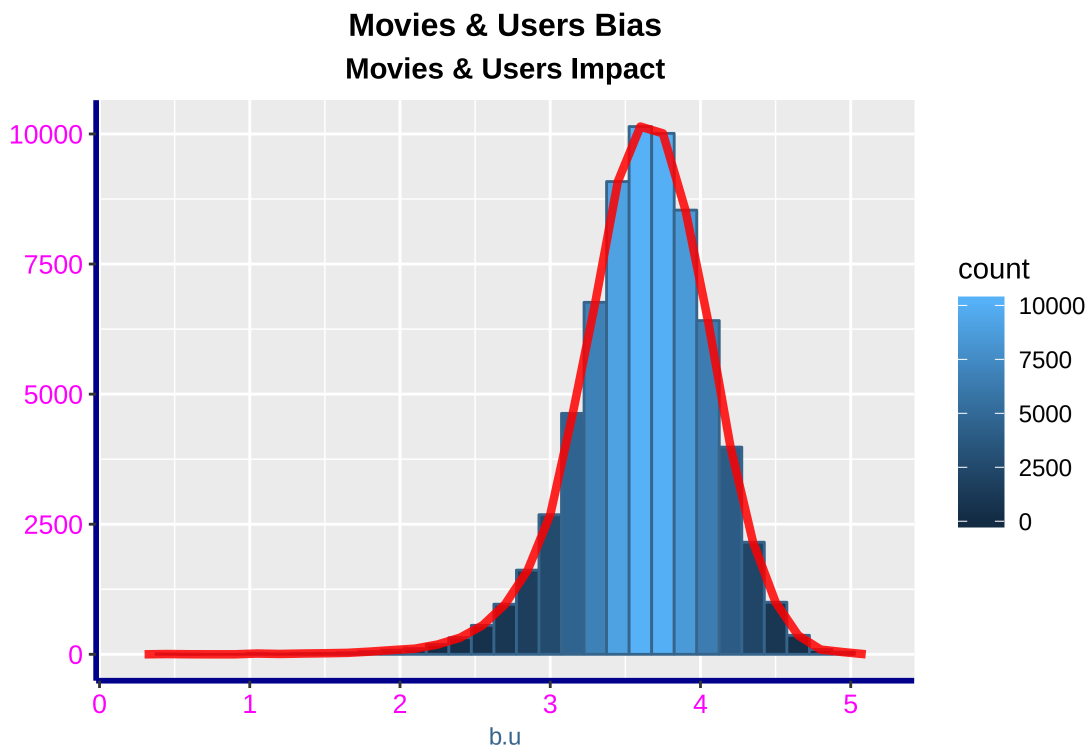
method	RMSE
Baseline	1.0522507
Movies Bias	0.9405627
Users Bias	0.9790568

6.5 Movies & Users Bias

The next evaluation will include the Movies and Users bias.

In this analysis we will include the user effect (b_u).

First, we can see a graph with the users rating average:



We can see that most of the users have an average between 3 and 4.5, and in the table we can see an improvement in the RMSE over the previous calculated RMSEs.

Table 20: RMSEs Comparisson

method	RMSE
Baseline	1.0522507
Movies Bias	0.9405627
Users Bias	0.9790568
Movies & Users Bias	0.8540140

7 Regularization

We can see that in the previous RMSEs, Movies Bias and Users Bias are not the best option, but the Users and Movies Bias has the smallest RMSE. Is time to identify if our previous analysis contains any error, we will start with the Movies Bias. Let's see which is the result obtained with first ten (10) movies, ordered in descendant mode.

Table 21: Largest Errors

title	residual
Shawshank Redemption, The (1994)	-3.95588
Shawshank Redemption, The (1994)	-3.95588
Shawshank Redemption, The (1994)	-3.95588
Shawshank Redemption, The (1994)	-3.95588
Shawshank Redemption, The (1994)	-3.95588
Shawshank Redemption, The (1994)	-3.95588
Shawshank Redemption, The (1994)	-3.95588
Shawshank Redemption, The (1994)	-3.95588
Shawshank Redemption, The (1994)	-3.95588
Shawshank Redemption, The (1994)	-3.95588

We will reduce the repeated movies, to one, in order to identify the mistakes in a better way. And, after joined the titles, the top Best Movies Ratings, are:

Table 22: 10 Best Movies Rating

title	b.i	n
Hellhounds on My Trail (1999)	1.472994	1
Satan's Tango (SÃ¡tÃ¡ntangÃ³) (1994)	1.472994	2
Shadows of Forgotten Ancestors (1964)	1.472994	2
Fighting Elegy (Kenka erejii) (1966)	1.472994	2
Sun Alley (Sonnenallee) (1999)	1.472994	2
Blue Light, The (Das Blaue Licht) (1932)	1.472994	3
More (1998)	1.222994	24
Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva) (1980)	1.222994	4
Human Condition II, The (Ningen no joken II) (1959)	1.222994	8
Human Condition III, The (Ningen no joken III) (1961)	1.222994	8

And, finally, after joined the titles, the top 10 Worst Movies Ratings, are:

Table 23: 10 Worst Movies Rating

title	b.i	n
Besotted (2001)	-3.027006	2
Hi-Line, The (1999)	-3.027006	1
Accused (Anklaget) (2005)	-3.027006	1
Confessions of a Superhero (2007)	-3.027006	1
War of the Worlds 2: The Next Wave (2008)	-3.027006	2
SuperBabies: Baby Geniuses 2 (2004)	-2.732363	56
Hip Hop Witch, Da (2000)	-2.705577	42
Disaster Movie (2008)	-2.667631	32
From Justin to Kelly (2003)	-2.624996	398
Criminals (1996)	-2.527006	2

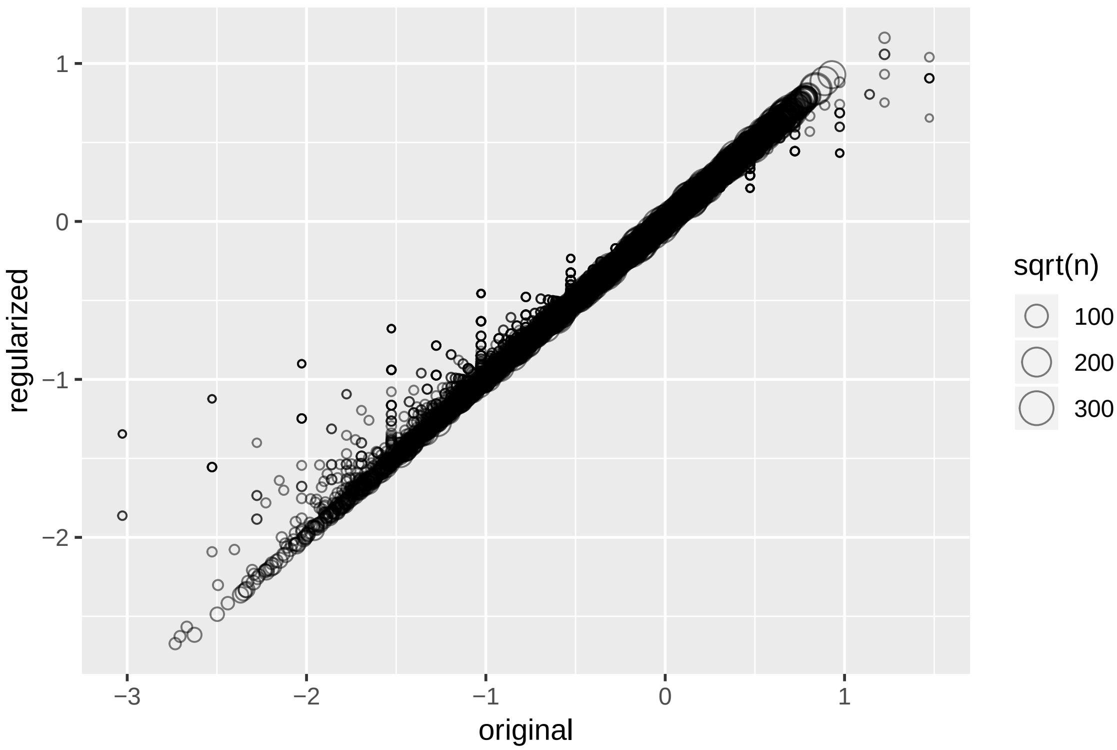
Most of the movies rated as Best Rated and Worst Rated are not popular, in recent years, and these movies do not have to much ratings, so is required a better analysis. In order of optimize b_i we use the follwing equation:

$$\frac{1}{N}\sum_{u,i}(y_{u,i}-\mu-b_i)^2+\lambda\sum_i b_i^2$$

And, the same reduced equation is:

$$\hat{b_i}(\lambda)=\frac{1}{\lambda+n_i}\sum_{u=1}^{n_i}(Y_{u,i}-\hat{\mu})$$

The regularization method allows us to add a lambd to penalizes movies with large estimates from a small sample size. In this graph, we can see the estimates shrink with penalty:



After regularization procedure, we should have a different information, let's see it:

10 best rated movies after regularization:

Table 24: Top 10 of Best Regularized Movies

title	b.i	n
More (1998)	1.1624499	24
Human Condition II, The (Ningen no joken II) (1959)	1.0577247	8
Human Condition III, The (Ningen no joken III) (1961)	1.0577247	8
Blue Light, The (Das Blaue Licht) (1932)	1.0397606	3
Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva) (1980)	0.9318051	4
Shawshank Redemption, The (1994)	0.9288325	28037
Satan's Tango (SÃ¡tÃ¡ntangÃ³) (1994)	0.9064580	2
Shadows of Forgotten Ancestors (1964)	0.9064580	2
Fighting Elegy (Kenka erejii) (1966)	0.9064580	2
Sun Alley (Sonnenallee) (1999)	0.9064580	2

10 worst rated movies after regularization:

Table 25: Top 10 of Worst Regularized Movies

title	b.i	n
SuperBabies: Baby Geniuses 2 (2004)	-2.672704	56
Hip Hop Witch, Da (2000)	-2.627381	42
From Justin to Kelly (2003)	-2.616777	398
Disaster Movie (2008)	-2.567344	32
PokÃ©mon Heroes (2003)	-2.486465	274
Carnosaur 3: Primal Species (1996)	-2.416559	136
Pokemon 4 Ever (a.k.a. PokÃ©mon 4: The Movie) (2002)	-2.364126	804
Glitter (2001)	-2.348603	1017
Barney's Great Adventure (1998)	-2.332497	416
Gigli (2003)	-2.330613	939

Is time to validate if regularization represents some better performance. And, the RMSEs table comparisson:

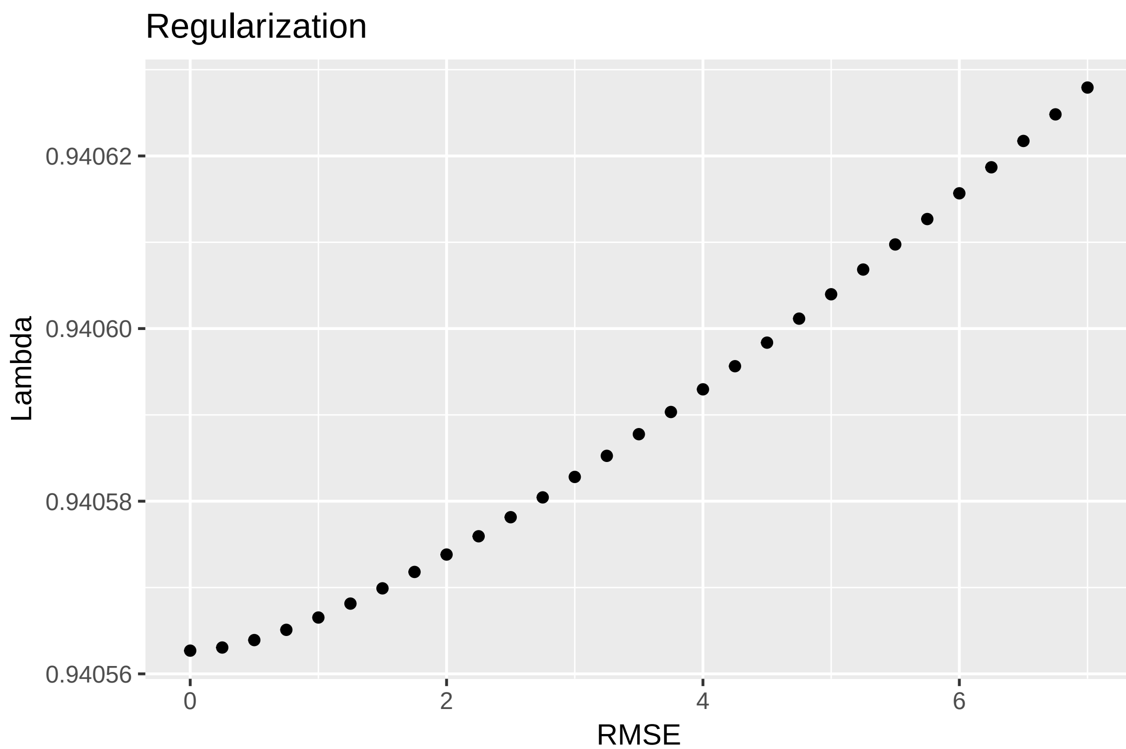
Table 26: RMSEs Comparisson

method	RMSE
Baseline	1.0522507
Movies Bias	0.9405627
Users Bias	0.9790568
Movies & Users Bias	0.8540140
Regularized - Movies Bias	0.9405681

We can see a better perfomance betwee Movies Bias and Regularization Movies Bias. And, we can validate that in the Regularization Movies Bias top 10 best and top 10 worst movies have a more logical ranking, according to the historical information.

7.1 Identify Lambda

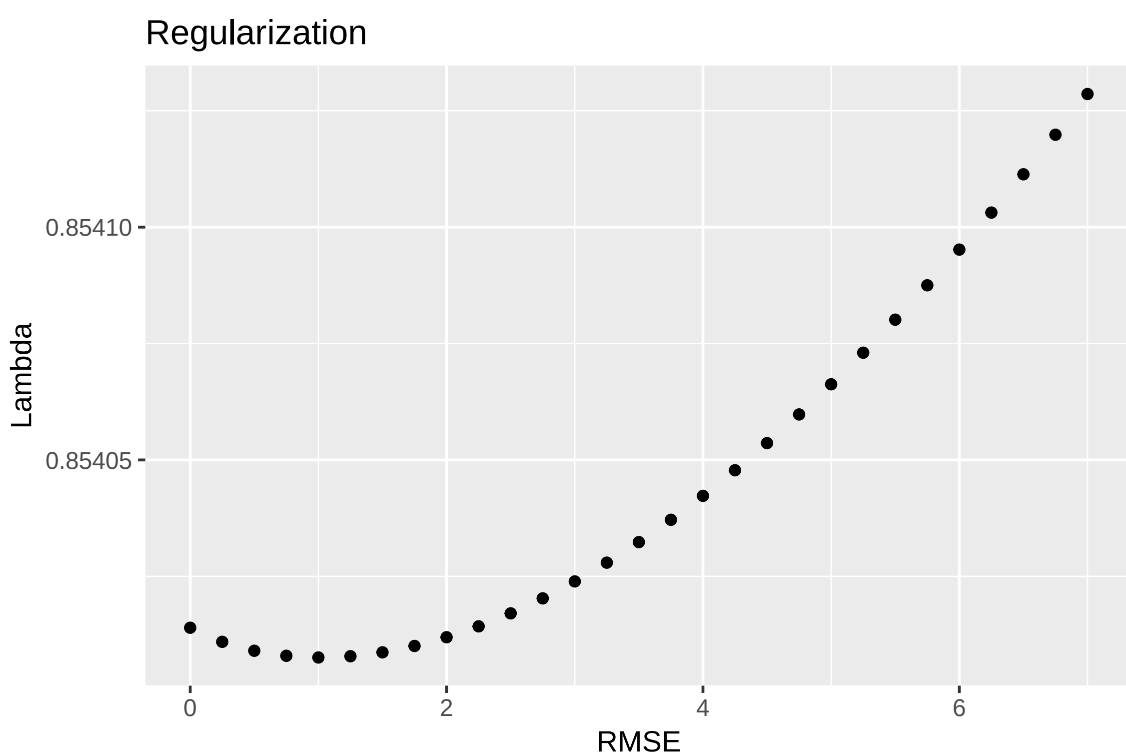
Now we must identify a `lambda` value to find the lowest RMSE, to identify it, we will use a function with a sequence of number from 0 to 7, applied to movies bias. And, we will se in a graph the λ which produces the lowest RMSE.




```
## [1] 0
```

7.2 Regularization Users & Movies Bias

We will find the λ which provides the smallest RMSE:



The best λ is: 1.

Table 27: RMSEs Comparisson

method	RMSE
Baseline	1.0522507
Movies Bias	0.9405627
Users Bias	0.9790568
Movies & Users Bias	0.8540140
Regularized - Movies Bias	0.9405681
Regularized Moves & User	0.8540076

8 Conclusion

Machine Learning has been used to create a `Movie Recommendation System`, and deduce that a join between `Users & Movies Bias`, produce the best RMSE improvement for our System. The best and lowest obtained RMSE = 0.8540076}