Capstone Project - MovieLens

Recommendation System

Reza Hashemi

August 11, 2019

Abstract

This report is part of the final project capstone to obtain the 'Professional Certificate in Master of Data Science' emited by Harvard University Harvard, through edx platform for education and learning.. 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: $\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$

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 23369607 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':    23369607 obs. of  10 variables:
##  $ userId            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ movieId           : int  122 122 185 185 185 231 292 292 292 292 ...
##  $ rating            : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ timestamp         : int  838985046 838985046 838983525 838983525 838983525 838983392 838983421 838983421 838983421 838983421 ...
##  $ title             : Factor w/ 10676 levels "'burbs, The (1989)",..: 1309 1309 6758 6758 6758 2871 7179 7179 7179 7179 ...
##  $ genres            : Factor w/ 20 levels "(no genres listed)",..: 6 16 2 7 18 6 2 9 17 18 ...
##  $ dates             : Factor w/ 6520453 levels "1995-01-09 06:46:49",..: 218509 218509 218465 218465 218465 218452 218454 218454 218454 218454 ...
##  $ 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	122	5	838985046	Boomerang (1992)	Comedy	1996-08-02 07:24:06	1996	1996-08	1996-08-02
1	122	5	838985046	Boomerang (1992)	Romance	1996-08-02 07:24:06	1996	1996-08	1996-08-02
1	185	5	838983525	Net, The (1995)	Action	1996-08-02 06:58:45	1996	1996-08	1996-08-02
1	185	5	838983525	Net, The (1995)	Crime	1996-08-02 06:58:45	1996	1996-08	1996-08-02
1	185	5	838983525	Net, The (1995)	Thriller	1996-08-02 06:58:45	1996	1996-08	1996-08-02
1	231	5	838983392	Dumb & Dumber (1994)	Comedy	1996-08-02 06:56:32	1996	1996-08	1996-08-02
1	292	5	838983421	Outbreak (1995)	Action	1996-08-02 06:57:01	1996	1996-08	1996-08-02
1	292	5	838983421	Outbreak (1995)	Drama	1996-08-02 06:57:01	1996	1996-08	1996-08-02
1	292	5	838983421	Outbreak (1995)	Sci-Fi	1996-08-02 06:57:01	1996	1996-08	1996-08-02
1	292	5	838983421	Outbreak (1995)	Thriller	1996-08-02 06:57:01	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 :35790   Median : 1748   Median :4.000   Median :1.042e+09
##  Mean   :35887   Mean   : 4276   Mean   :3.527   Mean   :1.035e+09
##  3rd Qu.:53635   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) :  124304   Drama    :3909401
##  Toy Story (1995)    :  119130   Comedy   :3541284
##  Jurassic Park (1993):  117164   Action   :2560649
##  True Lies (1994)    :  113930   Thriller :2325349
##  Aladdin (1992)      :  106070   Adventure:1908692
##  Batman (1989)       :   97372   Romance  :1712232
##  (Other)             :22691637   (Other)  :7412000
##                   dates            date.year    date.year.month
##  1996-02-29 19:00:00:     873   Min.   :1995   1999-12:  684022
##  2005-07-26 15:24:47:     127   1st Qu.:2000   2000-11:  616704
##  1996-03-29 12:04:19:     102   Median :2003   1999-10:  528315
##  1996-04-07 06:40:52:     100   Mean   :2002   2005-03:  526641
##  1996-04-15 06:23:54:     100   3rd Qu.:2005   1996-06:  389438
##  1996-04-01 07:03:49:      99   Max.   :2009   1999-11:  363899
##  (Other)            :23368206                  (Other):20260588
##  date.year.month.day
##  2000-11-20:  142846
##  2005-03-22:  116917
##  1999-12-11:  107247
##  2008-10-29:   93590
```

```
##  2000-11-21:   82659
##  1999-12-12:   79062
##  (Other)   :22747286
```

5 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:

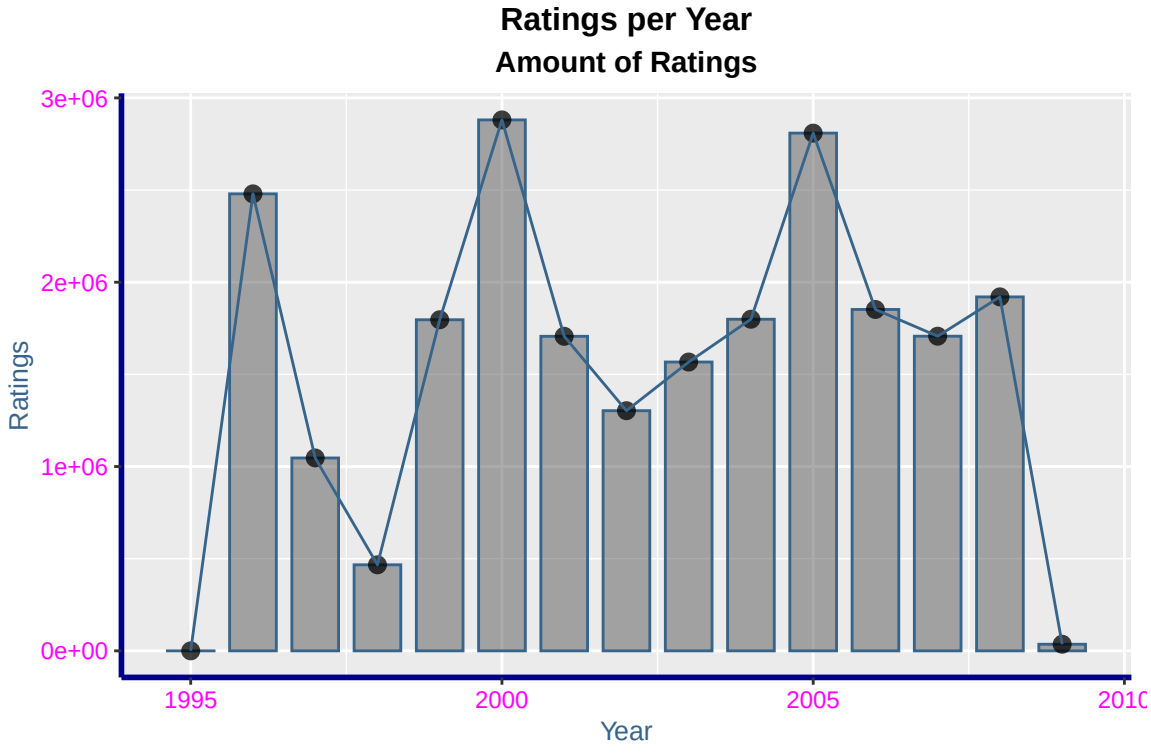
Table 5: Total Amount of each Field

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:

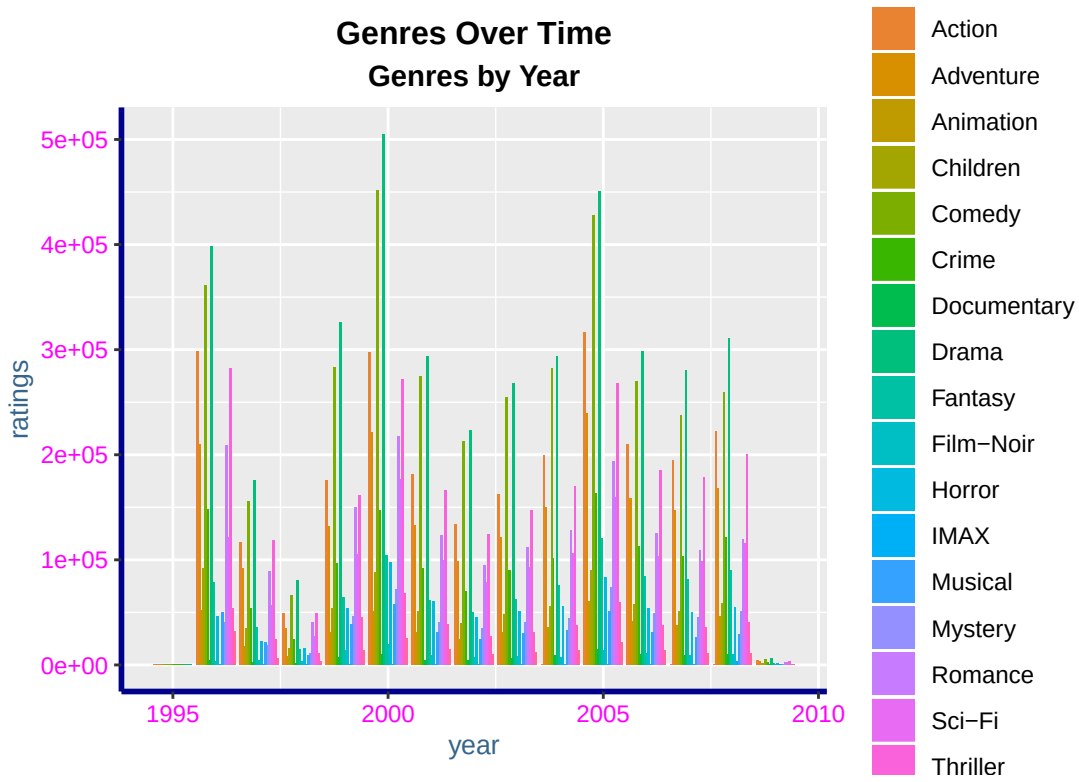
```
## Bar Graph
```



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

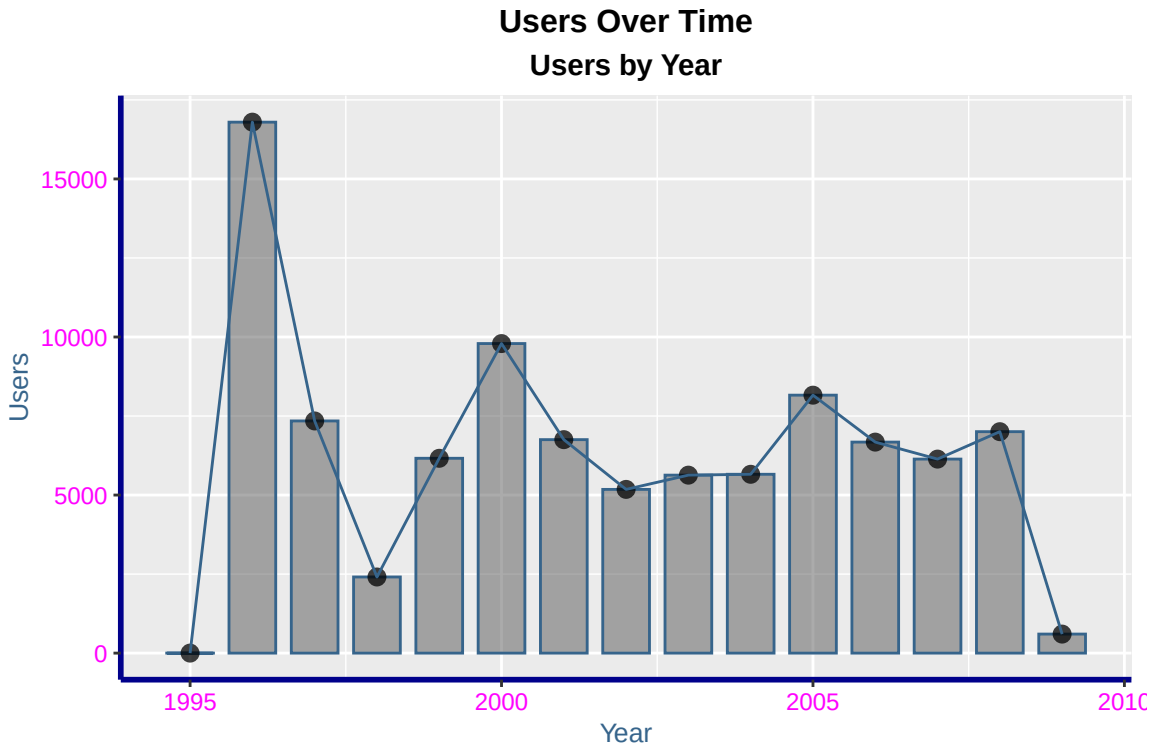
And, an evaluation of genres rating over the years:

```
## 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	3909401
Comedy	3541284
Action	2560649
Thriller	2325349
Adventure	1908692
Romance	1712232
Sci-Fi	1341750
Crime	1326917
Fantasy	925624
Children	737851

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

```
## Descendent order
```

Table 7: Top 10 Rated Movies

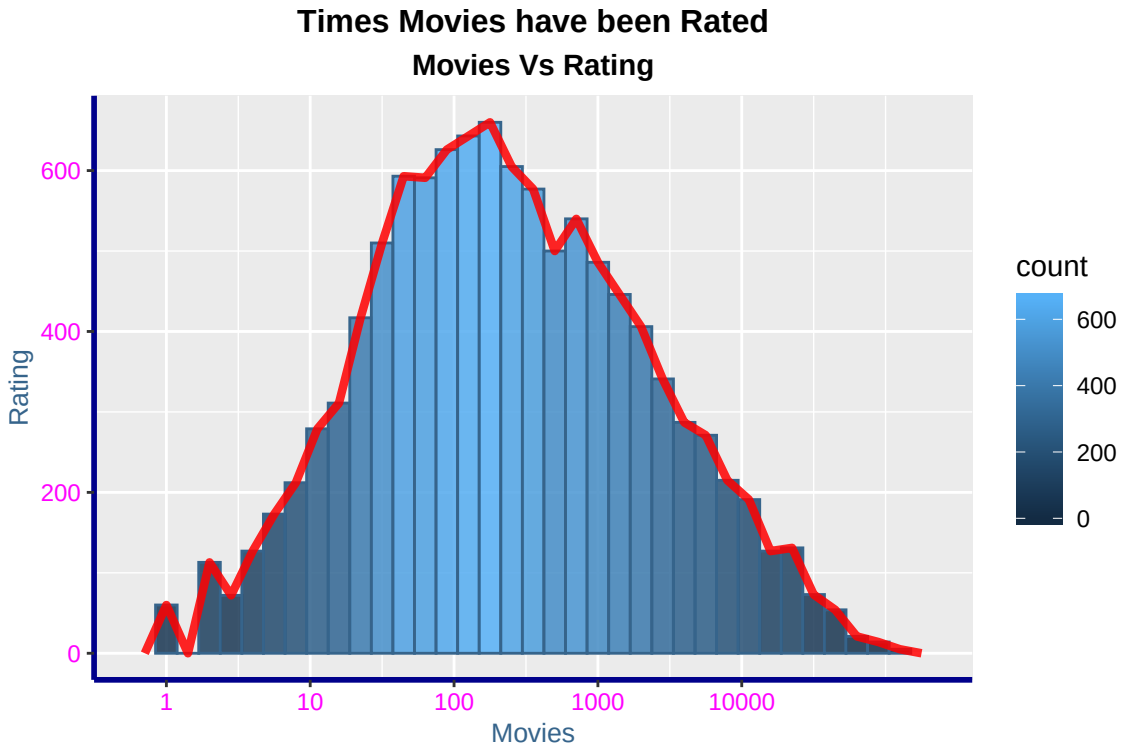
genres	title	count
Comedy	Pulp Fiction (1994)	31336
Crime	Pulp Fiction (1994)	31336
Drama	Pulp Fiction (1994)	31336
Comedy	Forrest Gump (1994)	31076
Drama	Forrest Gump (1994)	31076
Romance	Forrest Gump (1994)	31076
War	Forrest Gump (1994)	31076
Crime	Silence of the Lambs, The (1991)	30280
Horror	Silence of the Lambs, The (1991)	30280
Thriller	Silence of the Lambs, The (1991)	30280

The amount of movies per rating:

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

rating	movies
3.0	10216
4.0	9954
3.5	9810
2.0	9458
2.5	9416
5.0	8618
4.5	8299
1.0	8278
0.5	7197
1.5	7069

Graph of Number of Movies Vs Number of Ratings:



5.4 Analysis by Rating & Year

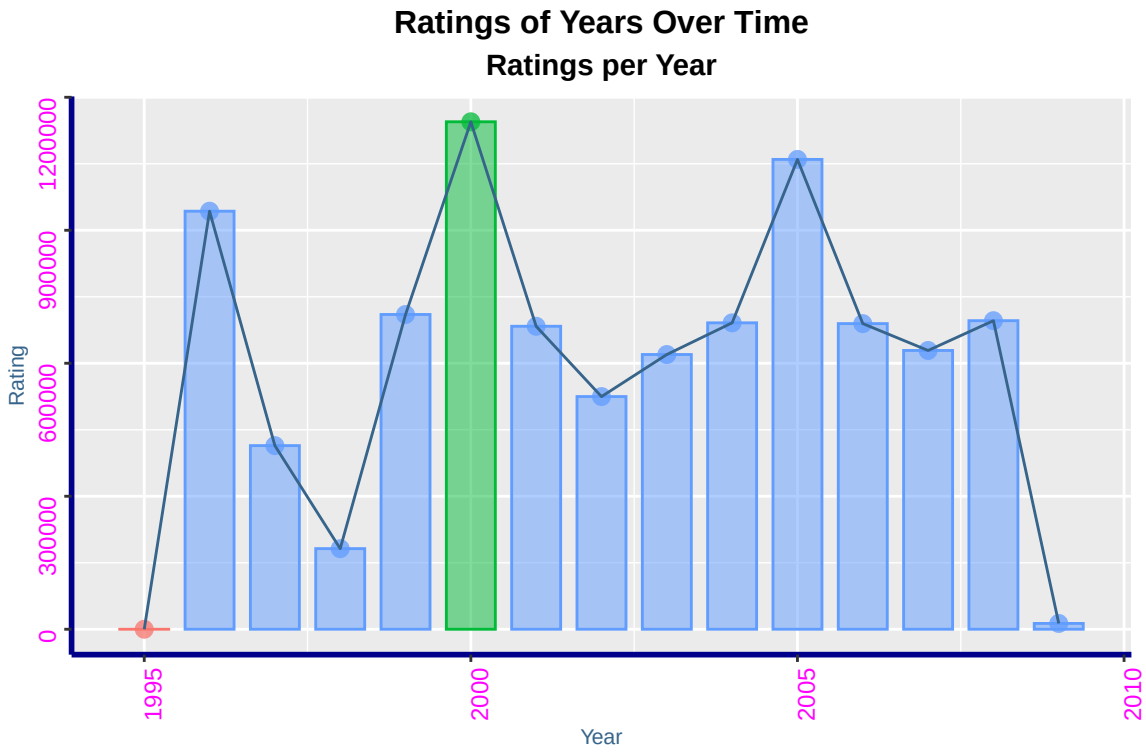
Most rated year: 2000, 1144666
Less rated year: 1995, 2

Table 9: Rating Per Year

date.year	ratings
1995	2
1996	942976
1997	414218
1998	181845
1999	709978
2000	1144666
2001	683412
2002	524826
2003	619707
2004	691191
2005	1059807
2006	689447
2007	628845
2008	696027
2009	13114

The graph of ratings by year is:

```
## Bar Graph Color
```



5.5 Analysis by Rating & Movie

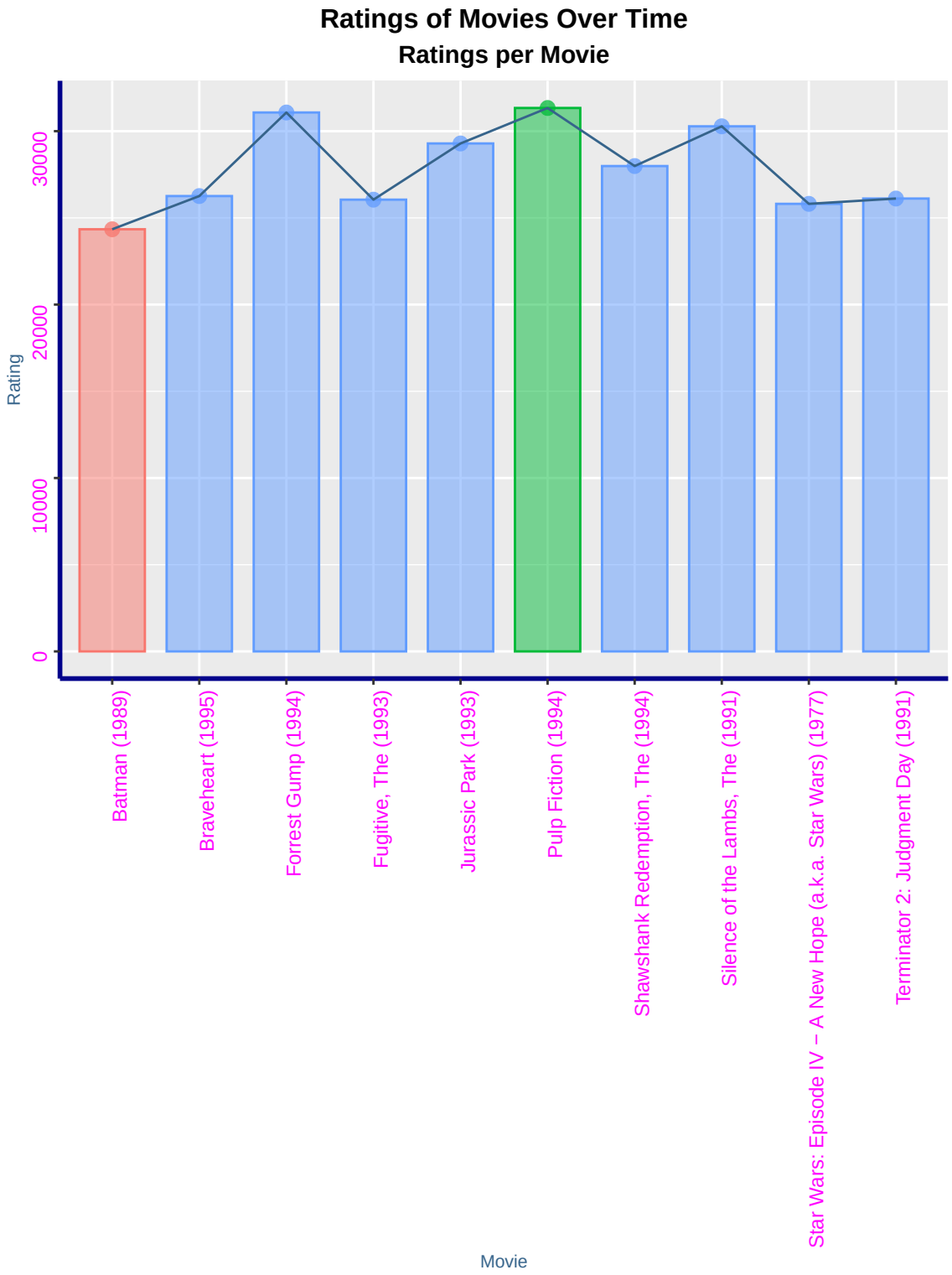
The most rated movie is: 7671, 31336
The less rated movie is: 20, 1

Table 10: Ratings per Movie

title	ratings
Pulp Fiction (1994)	31336
Forrest Gump (1994)	31076
Silence of the Lambs, The (1991)	30280
Jurassic Park (1993)	29291
Shawshank Redemption, The (1994)	27988
Braveheart (1995)	26258
Terminator 2: Judgment Day (1991)	26115
Fugitive, The (1993)	26050
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	25809
Batman (1989)	24343

The graph of ratings by movie is:

```
## Bar Graph Color
```



5.6 Analysis by Rating & Genre

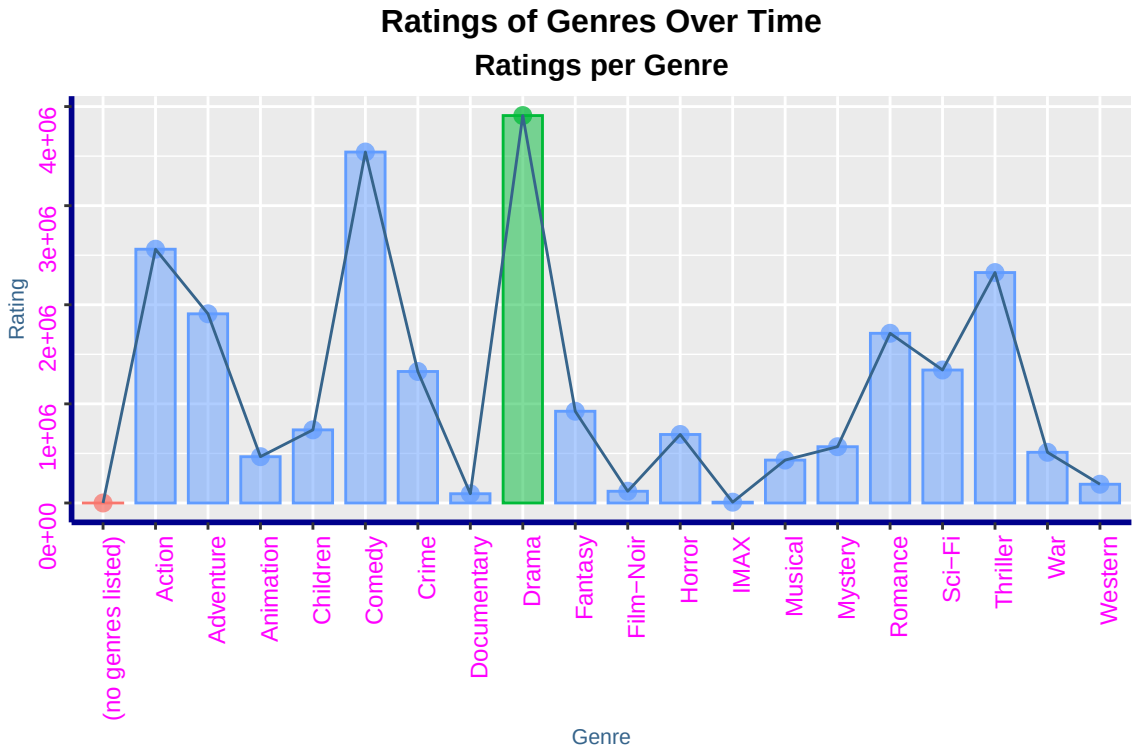
The most rated genre: 9, 3909401
The less rated genre: 1, 6

The graph of ratings by genre is:

```
## Bar Graph Color
```

Table 11: Ratings per Genre

genres	ratings
Drama	3909401
Comedy	3541284
Action	2560649
Thriller	2325349
Adventure	1908692
Romance	1712232
Sci-Fi	1341750
Crime	1326917
Fantasy	925624
Children	737851
Horror	691407
Mystery	567865
War	511330
Animation	467220
Musical	432960
Western	189234
Film-Noir	118394
Documentary	93252
IMAX	8190
(no genres listed)	6



5.7 Analysis of Ratings & User

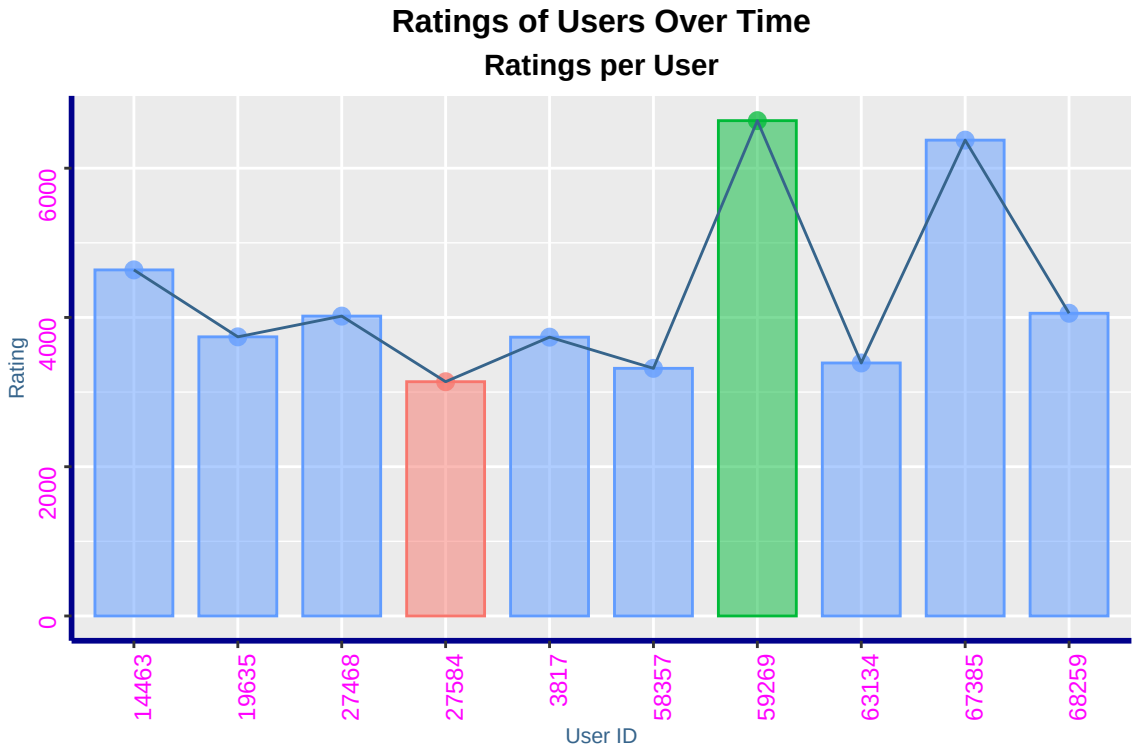
The most user ratings: 59269, 6637
The less user ratings: 22325, 13

The graph of ratings by user is:

Table 12: Ratings per User

userId	ratings
59269	6637
67385	6376
14463	4637
68259	4056
27468	4018
19635	3740
3817	3736
63134	3390
58357	3318
27584	3139

```
## Bar Graph Color
```



5.8 Analysis by Title

The most rated title by year:

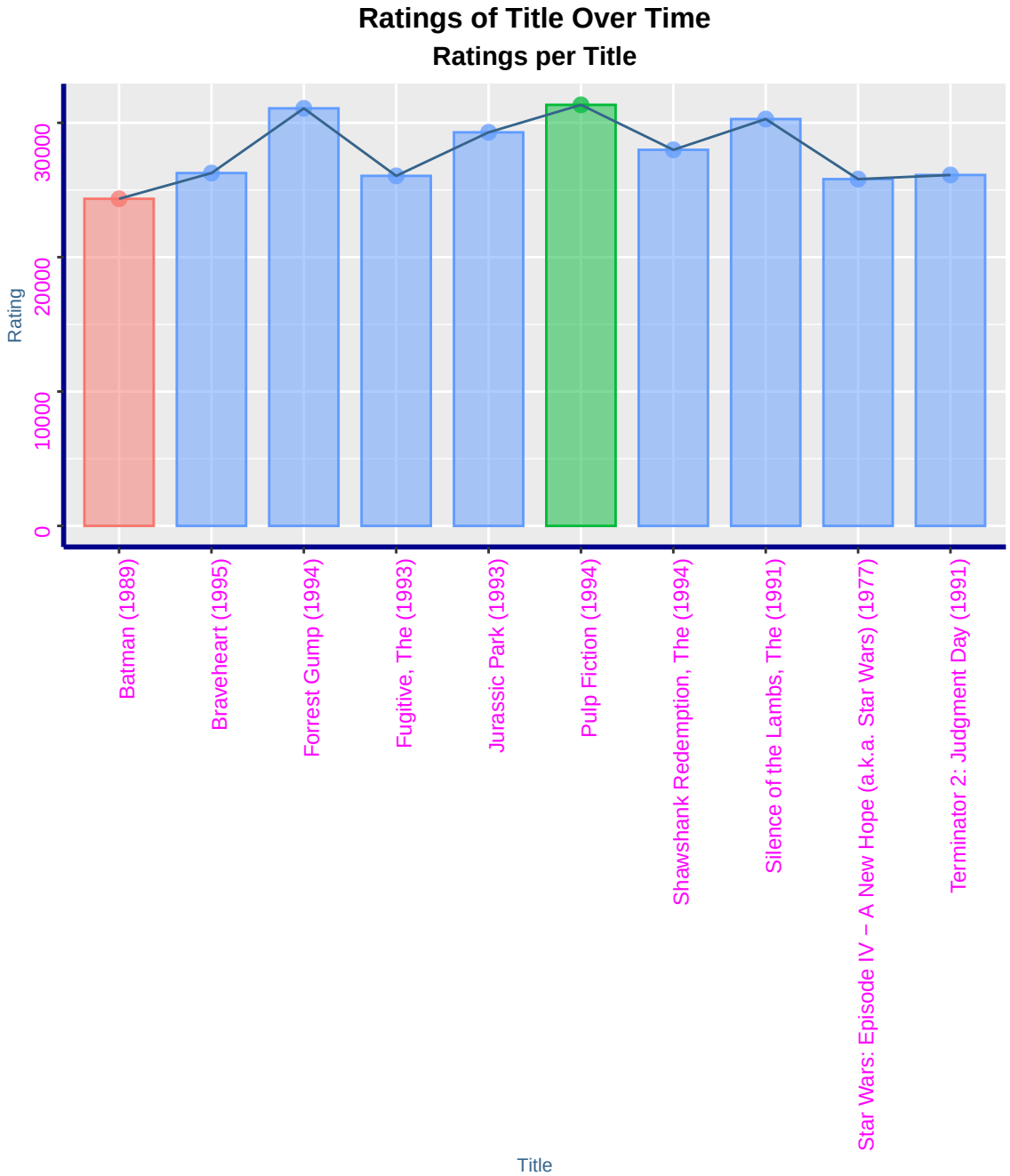
The most rated title: 7671, 31336
The less rated title: 20, 1

Rating per title:

```
## Bar Graph Color
```

Table 13: Rating per Title

title	ratings
Pulp Fiction (1994)	31336
Forrest Gump (1994)	31076
Silence of the Lambs, The (1991)	30280
Jurassic Park (1993)	29291
Shawshank Redemption, The (1994)	27988
Braveheart (1995)	26258
Terminator 2: Judgment Day (1991)	26115
Fugitive, The (1993)	26050
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	25809
Batman (1989)	24343

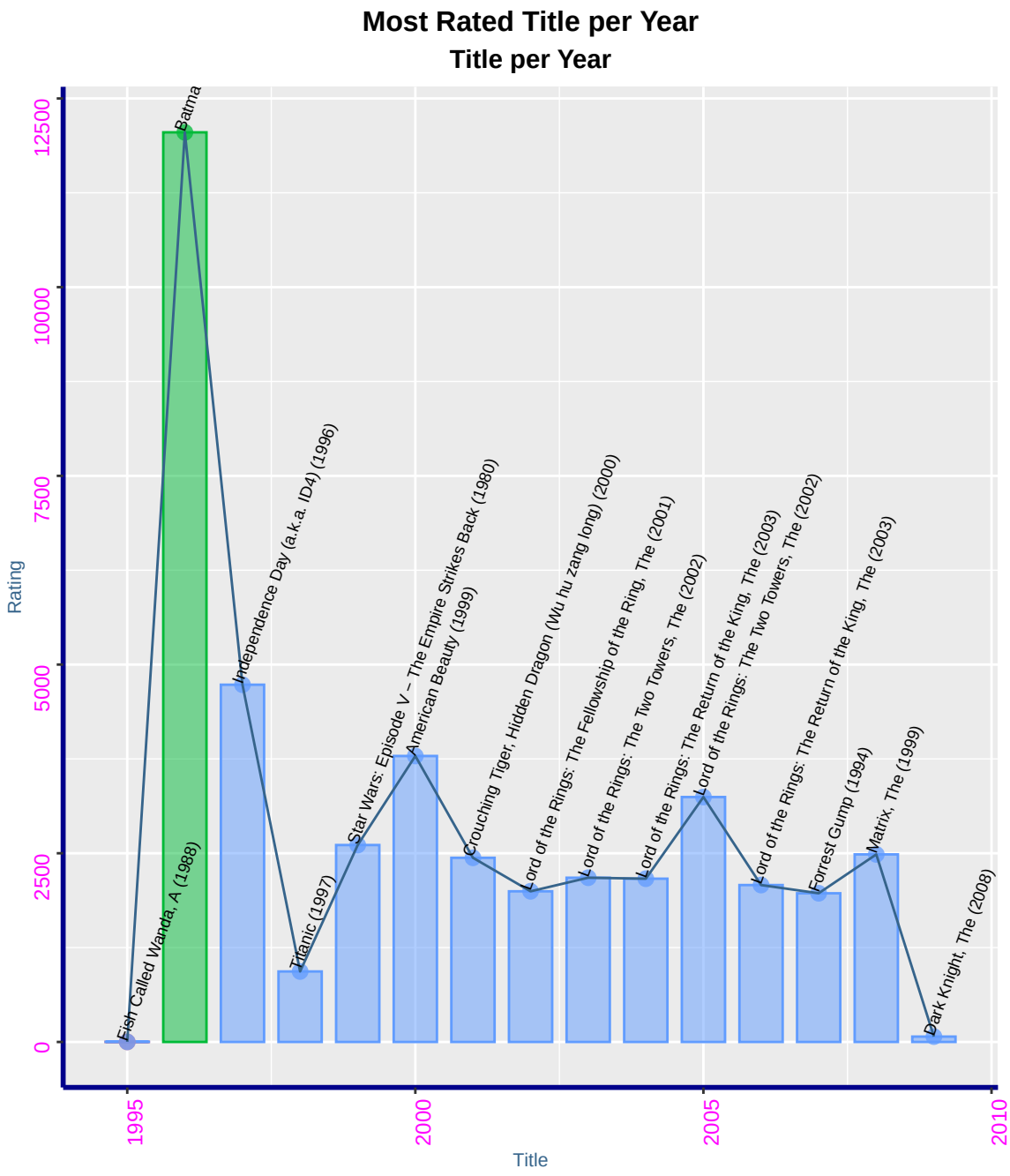


Most rated title per year:

Table 14: Most Rated Title per Year

date.year	title	ratings
1995	Fish Called Wanda, A (1988)	1
1995	Get Shorty (1995)	1
1996	Batman (1989)	12051
1997	Independence Day (a.k.a. ID4) (1996)	4733
1998	Titanic (1997)	934
1999	Star Wars: Episode V - The Empire Strikes Back (1980)	2609
2000	American Beauty (1999)	3789
2001	Crouching Tiger, Hidden Dragon (Wu hu zang long) (2000)	2440
2002	Lord of the Rings: The Fellowship of the Ring, The (2001)	1996
2003	Lord of the Rings: The Two Towers, The (2002)	2176
2004	Lord of the Rings: The Return of the King, The (2003)	2163
2005	Lord of the Rings: The Two Towers, The (2002)	3245
2006	Lord of the Rings: The Return of the King, The (2003)	2079
2007	Forrest Gump (1994)	1970
2008	Matrix, The (1999)	2485
2009	Dark Knight, The (2008)	71

```
## Bar Graph Color
```



5.9 Analysis by Users

A table of user with more ratings:

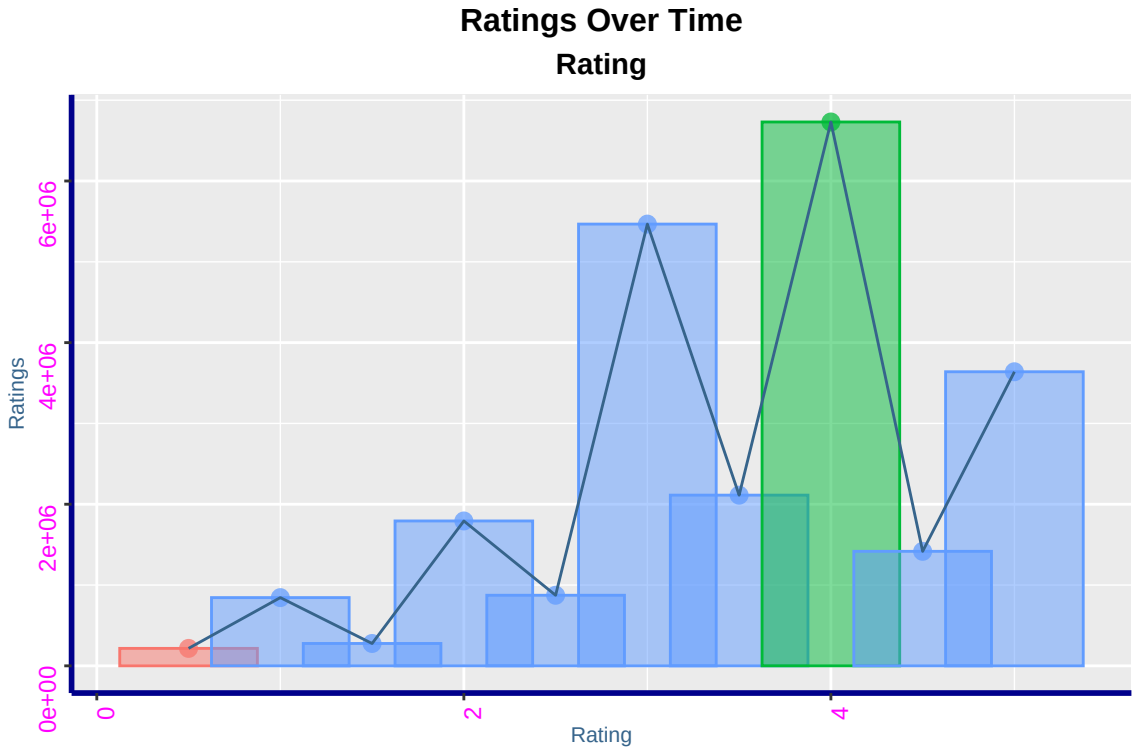
The user with most ratings has the ID: 59269, 6637
The user with less ratings has the ID: 22325, 13

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	6730156	28.7987556
3.0	5466754	23.3925799
5.0	3639299	15.5727865
3.5	2112391	9.0390523
2.0	1792891	7.6718920
4.5	1416963	6.0632727
2.5	873585	3.7381245
1.0	844605	3.6141173
1.5	276775	1.1843374
0.5	216188	0.9250819



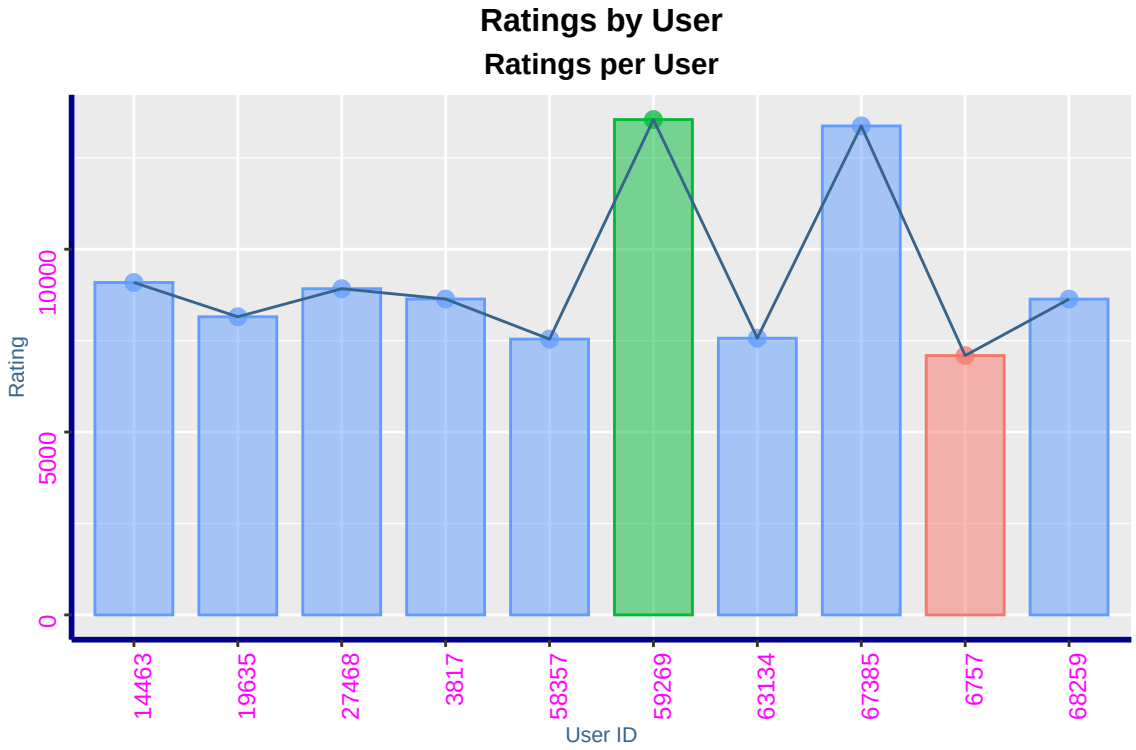
Amount of users per rating:

A table that shows all ratings per user:

Graph of times that a user has rated a movie:

Table 16: Ratings per User

userId	ratings
59269	13545
67385	13371
14463	9090
27468	8920
3817	8638
68259	8636
19635	8153
63134	7567
58357	7542
6757	7092



6 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.

6.1 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 = 3.5270036 and baseline **RMSE** = 1.0522745

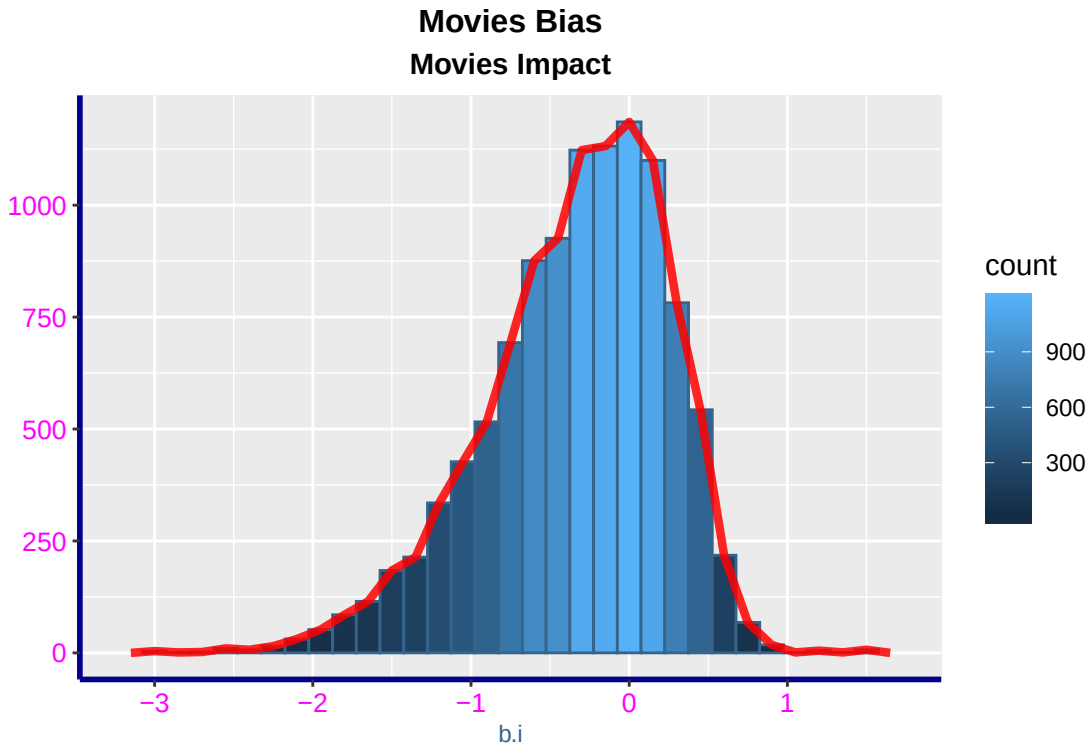
Table 17: RMSEs Comparisson

method	RMSE
Baseline	1.052275

6.2 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	1.0522745
Movies Bias	0.9405772

We can see an improvement of Movies Bias over Baseline.

6.3 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

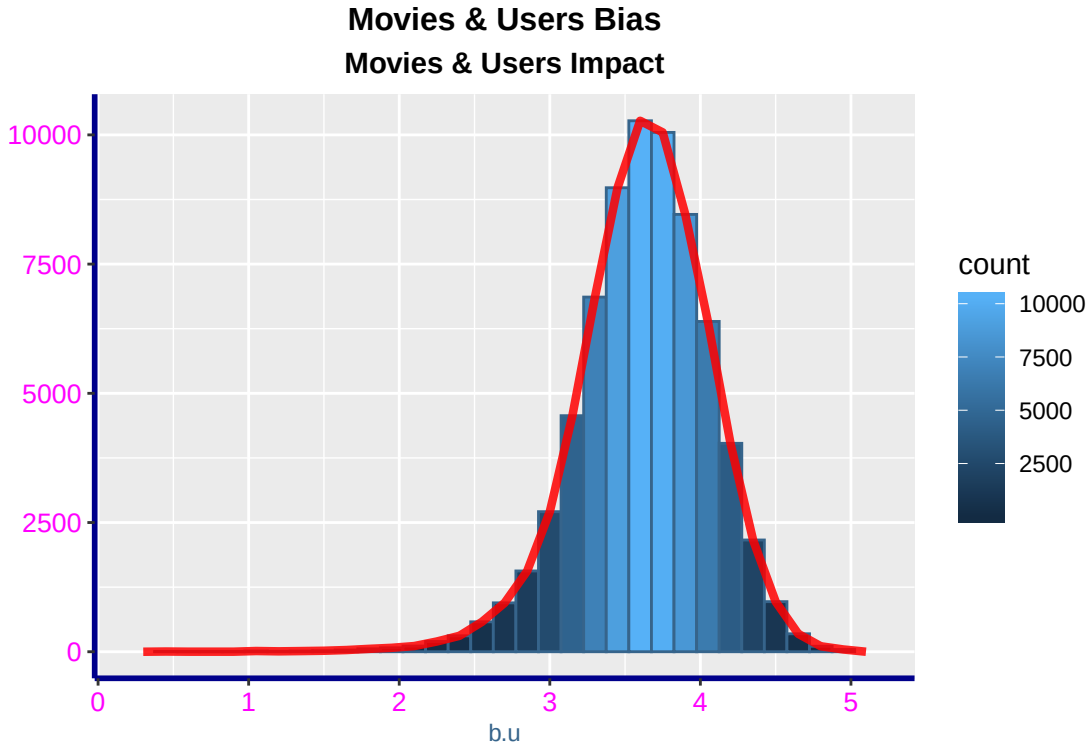
method	RMSE
Baseline	1.0522745
Movies Bias	0.9405772
Users Bias	0.9790470

6.4 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	1.0522745
Movies Bias	0.9405772
Users Bias	0.9790470
Movies & Users Bias	0.8539940

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)	-3.956928
Shawshank Redemption, The (1994)	-3.956928
Shawshank Redemption, The (1994)	-3.956928
Shawshank Redemption, The (1994)	-3.956928
Shawshank Redemption, The (1994)	-3.956928
Shawshank Redemption, The (1994)	-3.956928
Shawshank Redemption, The (1994)	-3.956928
Shawshank Redemption, The (1994)	-3.956928
Shawshank Redemption, The (1994)	-3.956928
Shawshank Redemption, The (1994)	-3.956928

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)	1.472996	1
Satan's Tango (SÃfÂ¡tÃfÂ¡ntangÃfÂ³) (1994)	1.472996	2
Shadows of Forgotten Ancestors (1964)	1.472996	2
Fighting Elegy (Kenka erejii) (1966)	1.472996	2
Sun Alley (Sonnenallee) (1999)	1.472996	2
Blue Light, The (Das Blaue Licht) (1932)	1.472996	3
Constantine's Sword (2007)	1.472996	1
Human Condition II, The (Ningen no joken II) (1959)	1.306330	6
Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva) (1980)	1.222996	4
Human Condition III, The (Ningen no joken III) (1961)	1.222996	8

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

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:

Table 23: 10 Worst Movies Rating

title	b.i	n
Besotted (2001)	-3.027004	2
Hi-Line, The (1999)	-3.027004	1
Grief (1993)	-3.027004	1
Accused (Anklaget) (2005)	-3.027004	1
War of the Worlds 2: The Next Wave (2008)	-2.777004	2
SuperBabies: Baby Geniuses 2 (2004)	-2.713444	59
Hip Hop Witch, Da (2000)	-2.693670	36
From Justin to Kelly (2003)	-2.597711	396
Disaster Movie (2008)	-2.543133	31
Stacy's Knights (1982)	-2.527004	1

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

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: