MovieLens Project

Biljana Novkovic

3/14/2023

INTRODUCTION

Movies are an important part of our modern-day lives. They entertain us, help us escape reality, take us on adventures and make us think. Movies can even change the way we view the world. We used to watch movies in the cinemas, then on our TVs, and nowadays we watch them mainly on various streaming services. This means that we can watch them on demand, and we can pick from thousands of movies at any point in time. But how do we choose what to watch next?

Websites such as IMDB and Rotten Tomatoes aggregate critic and user scores. However, they do not account for the fact that different people have different tastes. This is where machine learning algorithms come into play.

The goal of this project was to create a predictive algorithm that can recommend movies to users based on the MovieLens data set. This dataset has over 9 million ratings by over 69,000 users for over 10,000 movies. In this project, we will split this dataset into a working dataset and the dataset we will use for final evaluation. Using our working dataset, we will look for variables that may help us predict user ratings more accurately. Then we will build and evaluate models based on these variables. Finally, we will pick the most predictive model and evaluate it on our final evaluation dataset.

ANALYSIS

Exploring the Dataset

The MovieLens datasets and initial code were provided by the course. We will use the edx dataset provided for data exploration and model training and testing. We will use the final holdout test set to test our best model as the final step in this project.

	userId	movieId	rating	timestamp	title	genres
1	1	122	5	838985046	Boomerang (1992)	Comedy|Romance
2	1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
4	1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
5	1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi
6	1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi
7	1	355	5	838984474	Flintstones, The (1994)	Children|Comedy|Fantasy

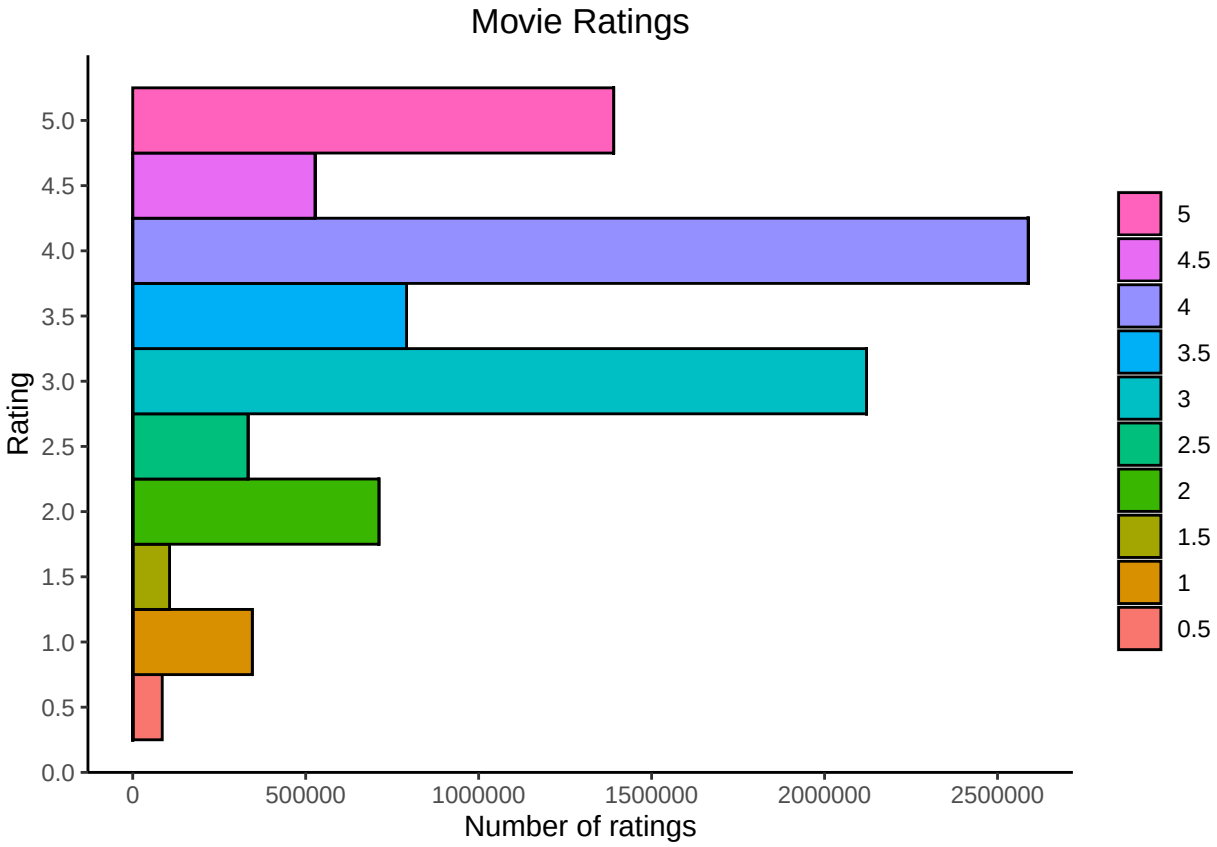
users	movies	nrows
69878	10677	9000055

Our dataset is a table with the following columns: (1) userId, (2) movieId, (3) rating, (4) timestamp, (5) title and (6) genres. It has 9,000,055 rows. From the first two columns, we can see that this dataset has 69,878 distinct users and 10,677 distinct movies.

Distribution of Ratings

Next, let's explore the general distribution of ratings in this dataset.

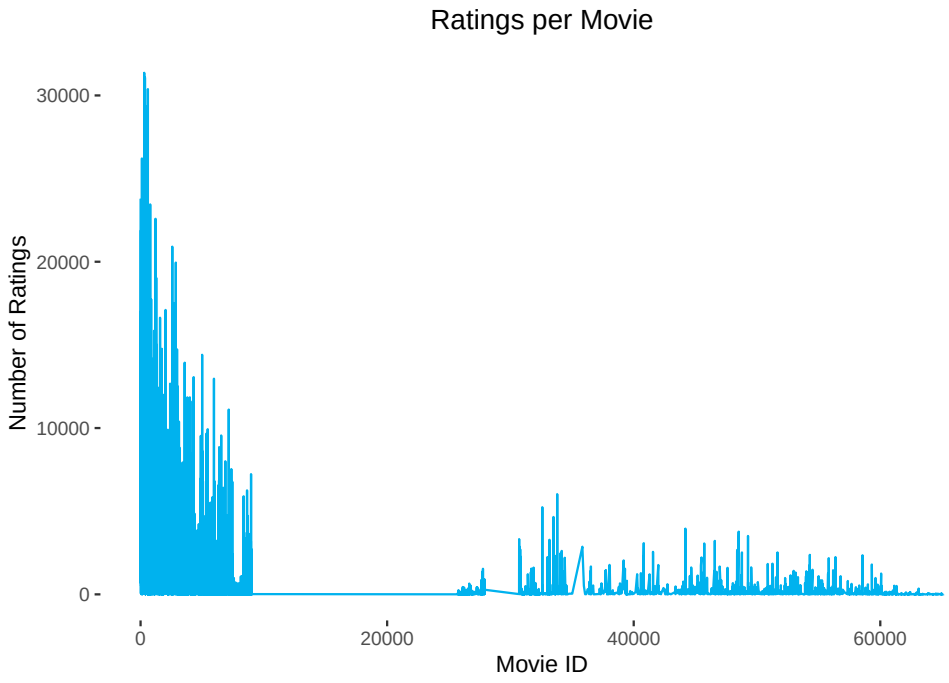
average_rating
3.512465



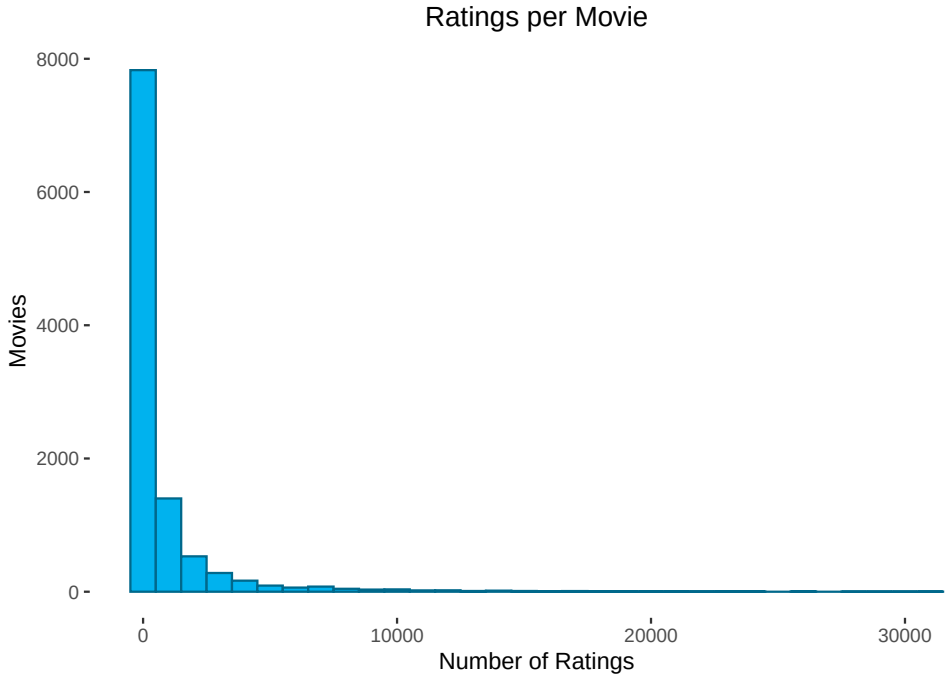
The average rating is around 3.5 stars. From the figure above, we can see that 4 stars is the most common rating given to a movie, with over 2.5 million ratings, followed by 3 and then 5 stars. Relatively fewer movies are rated with 0.5, 1 and 1.5 stars. Because there is no rating of 0 stars, we can assume that 0 stars was not an option available to users in this database.

Movies

Let's look at the number of ratings for each movie. Are they more or less evenly distributed?



We can see that there are a number of movies that receive a lot of ratings, some over 30,000 which is about half of the users in this dataset. The majority of movies, however, have less than 5,000 reviews, and many have far fewer than that.



When we look at the distribution of reviews per movie, we can see that almost 8,000 movies out of our total 10,677 have less than a thousand reviews.

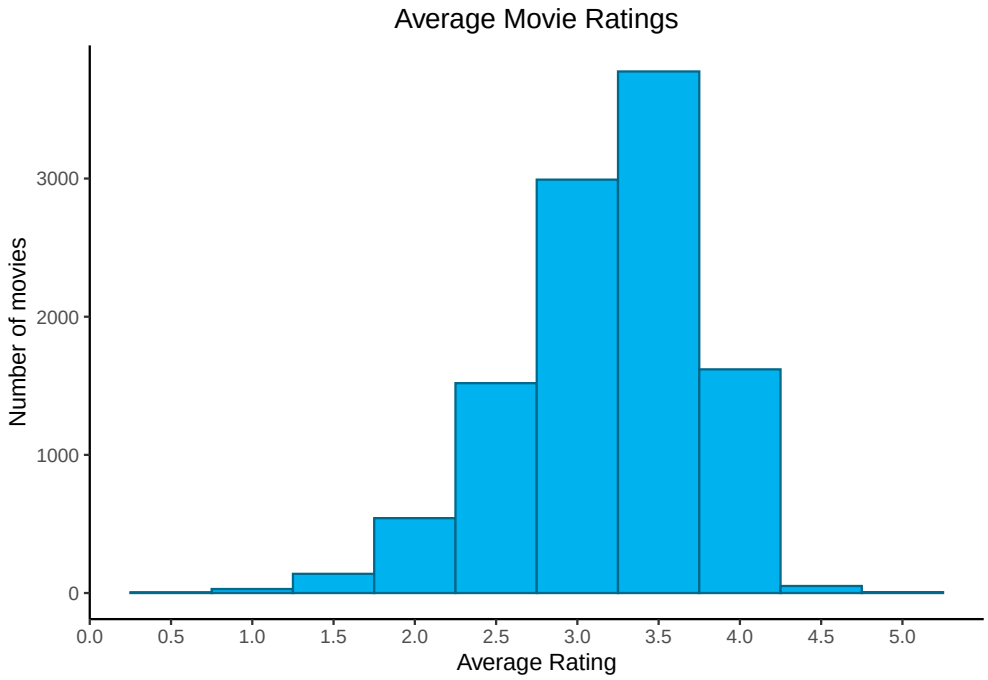
Let's look at the most and least reviewed movies.

movieId	title	reviews
296	Pulp Fiction (1994)	31362
356	Forrest Gump (1994)	31079
593	Silence of the Lambs, The (1991)	30382
480	Jurassic Park (1993)	29360
318	Shawshank Redemption, The (1994)	28015
110	Braveheart (1995)	26212
457	Fugitive, The (1993)	25998
589	Terminator 2: Judgment Day (1991)	25984
260	Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	25672
150	Apollo 13 (1995)	24284

movieId	title	reviews
3191	Quarry, The (1998)	1
3226	Hellhounds on My Trail (1999)	1
3234	Train Ride to Hollywood (1978)	1
3356	Condo Painting (2000)	1
3383	Big Fella (1937)	1
3561	Stacy's Knights (1982)	1
3583	Black Tights (1-2-3-4 ou Les Collants noirs) (1960)	1
4071	Dog Run (1996)	1
4075	Monkey's Tale, A (Les Château des singes) (1999)	1
4820	Won't Anybody Listen? (2000)	1

We can see that the most reviewed movies are large blockbusters such as: Pulp Fiction, Forrest Gump and The Silence of the Lambs. Each of these 3 movies has over 30,000 reviews. The least reviewed movies have only a single review and include some pretty obscure entries.

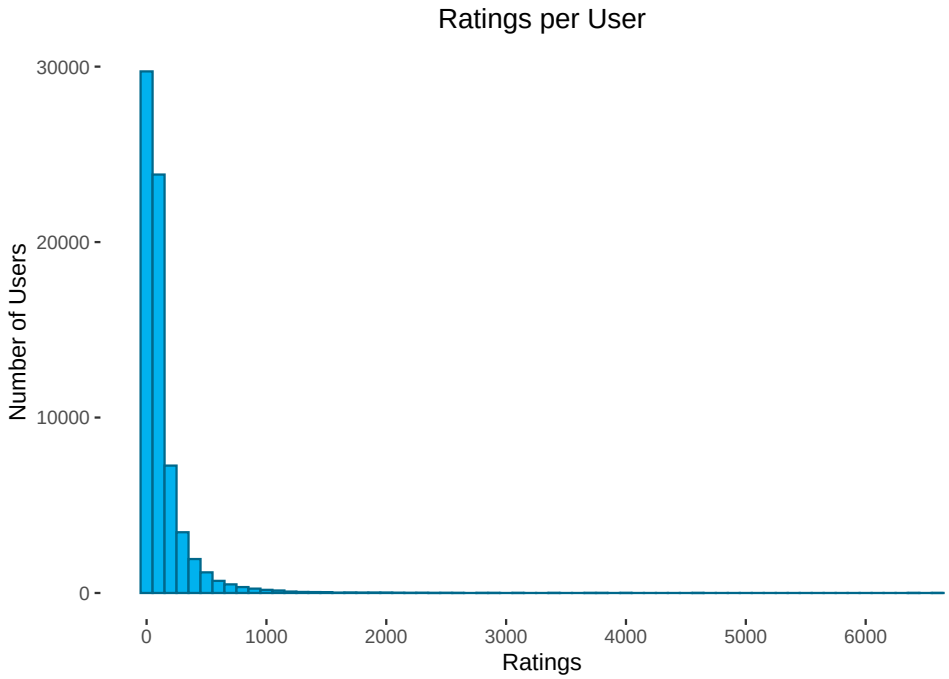
What is the distribution of reviews between movies?



We can see that a lot of movies get 3 and 3.5 stars on average. Movies with less than 2 and more than 4 stars are relatively uncommon.

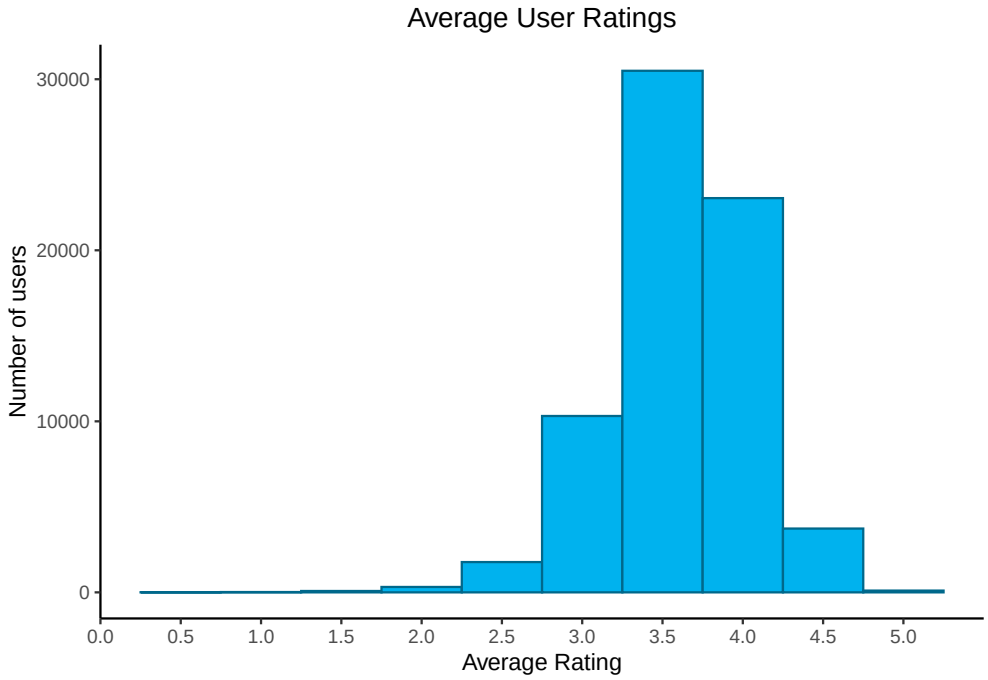
Users

Now let's look at the users. Are some users rating way more movies than others?



The majority of users have rated less than 200 movies. In fact, almost half of the users have rated a 100 movies or less.

How about user bias? Are some users more likely to review movies favorably than others?



We can see that many users have a review average of about 3.5. Users that rate movies very favorably (average of 4.5 or 5) or unfavorably (average of 2.5 or below) are less common, but they do exist.

Genres

Next, let's look at the genres. There are 18 genres in the edx dataset. One movie doesn't have a genre, and several are classified as IMAX, which we will ignore for this analysis.

```
## # A tibble: 20 x 2
##    genres                n
##    <chr>             <int>
##  1 Drama              5336
##  2 Comedy             3703
##  3 Thriller           1705
##  4 Romance            1685
##  5 Action             1473
##  6 Crime              1117
##  7 Adventure          1025
##  8 Horror             1013
##  9 Sci-Fi              754
## 10 Fantasy             543
## 11 Children            528
## 12 War                 510
## 13 Mystery             509
## 14 Documentary         481
## 15 Musical             436
## 16 Animation           286
## 17 Western             275
## 18 Film-Noir           148
```

```
## 19 IMAX                   29
## 20 (no genres listed)      1
```

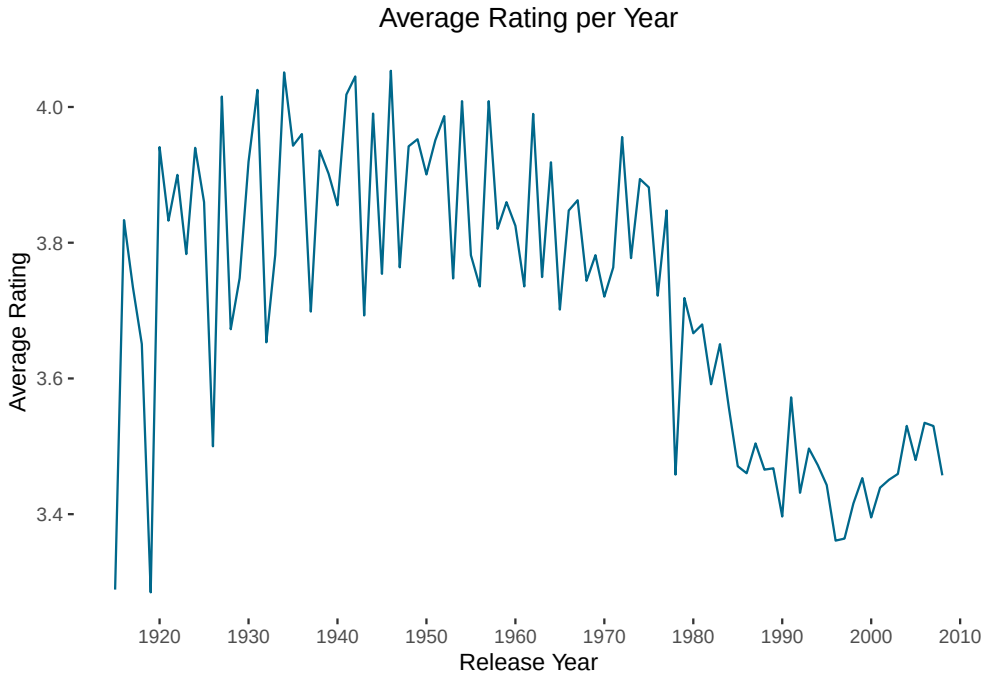
Dramas are the most common. Over 5,000 movies, about half in this database, have been classified as dramas. Comedy is the next common genre, followed by thriller, romance and action. Film-noir is the least represented genre, followed by western and animation.

```
## # A tibble: 20 x 2
##    genres              average
##    <chr>                 <dbl>
##  1 Film-Noir              4.01
##  2 Documentary            3.78
##  3 War                    3.78
##  4 IMAX                   3.77
##  5 Mystery                3.68
##  6 Drama                  3.67
##  7 Crime                  3.67
##  8 (no genres listed)     3.64
##  9 Animation              3.60
## 10 Musical                3.56
## 11 Western                3.56
## 12 Romance                3.55
## 13 Thriller               3.51
## 14 Fantasy                3.50
## 15 Adventure              3.49
## 16 Comedy                 3.44
## 17 Action                 3.42
## 18 Children               3.42
## 19 Sci-Fi                 3.40
## 20 Horror                 3.27
```

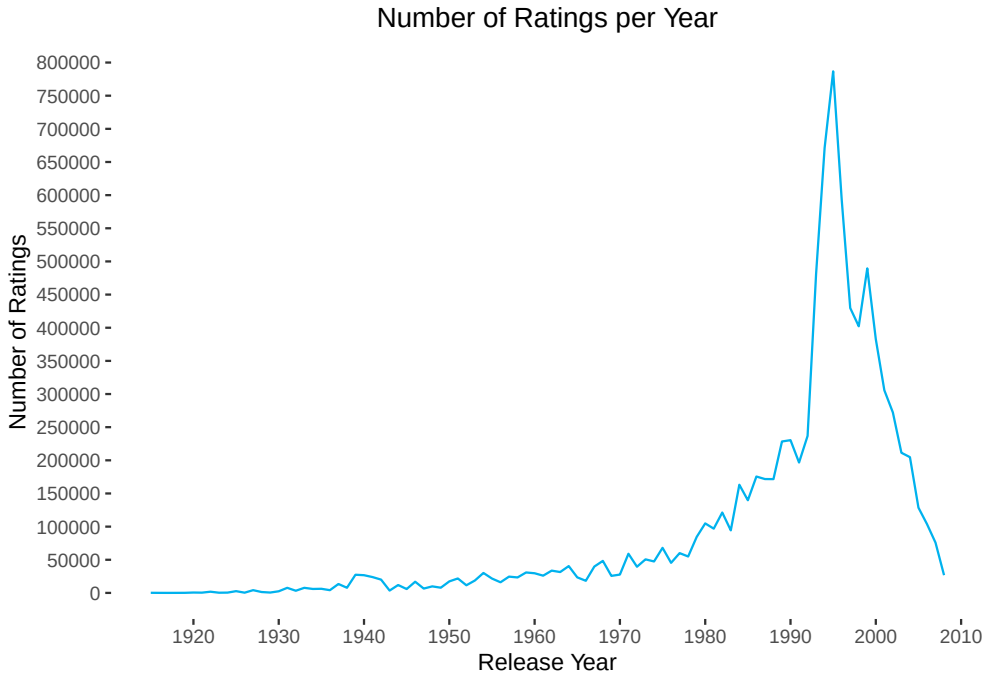
Different genres have different rating averages. For example, film-noir, documentaries and war movies have the highest average ratings (4.01, 3.78 and 3.78), while childrens' movies, sci-fi and horror movies have the lower ratings (3.42, 3.4 and 3.27).

Year of Release

Finally, let's look at the year of release. Do people tend to prefer newer or older movies?

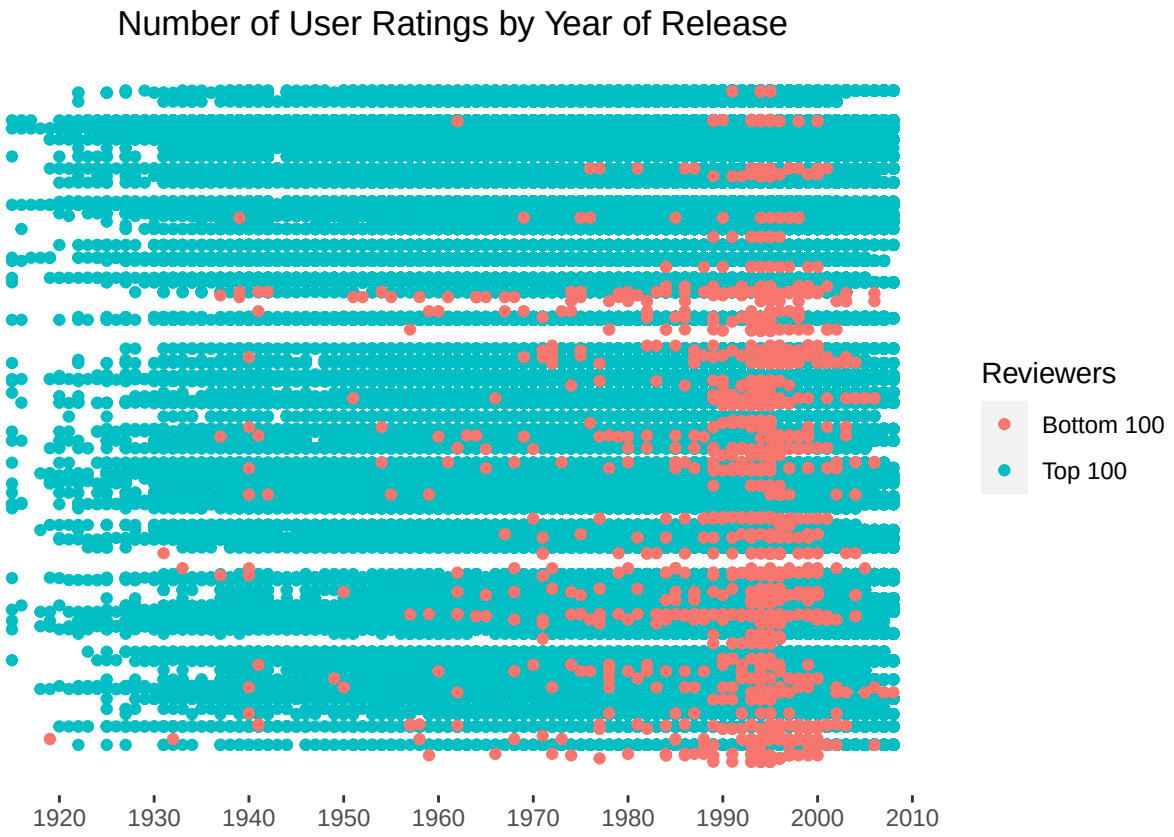


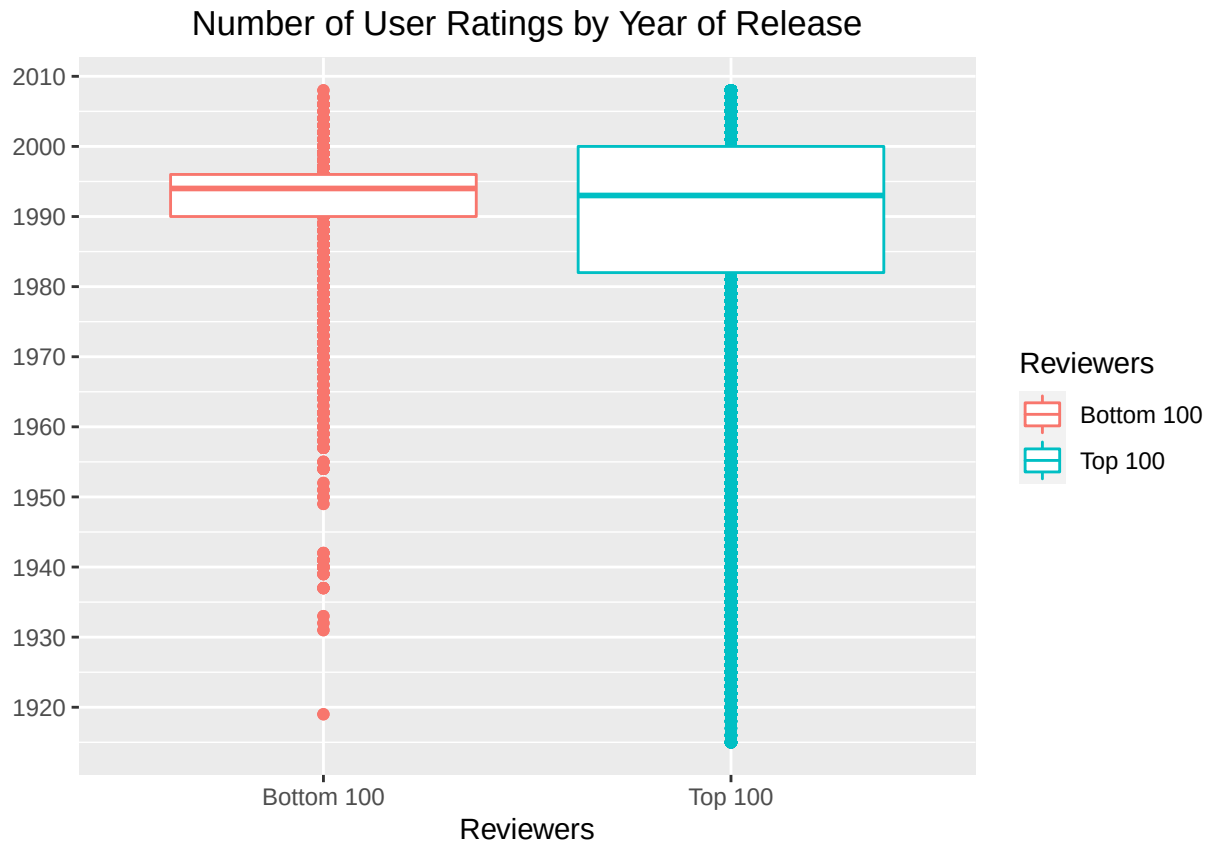
Somewhat surprisingly, newer movies tend to have lower ratings. In fact, movie ratings are higher than average for the majority of the years up to mid 80s, when ratings seem to drop. But is there bias in the data? Likely, a much higher number of users only watch and review newer movies.



year	n
1995	786762
1994	671376
1996	593518
1999	489537
1993	481184
1997	429751
1998	402187
2000	382763
2001	305705
2002	272180

Indeed, we can observe an increase in reviews throughout the 70s and 80s, with a sharp increase and peak in the mid 90s. The top 7 most reviewed movies are from the 90s. People who seek out older movies may be more likely to be cinefiles, and may tend to rate those movies higher. Let's check if that is true by taking the 100 users with the most reviews and the 100 users with the least reviews, and by checking the year of release of the movies they tend to rate.





We can definitely see that the users with least reviews tend to review more recent movies, peaking in the 90s, while the most prolific reviewers in our dataset tend to review movies across various decades.

RESULTS

First, we need to partition the dataset into a train and test set, and make sure they both look at the same movies and have the same user pool. We assign 80% of the data to the train and 20% to the test set. We also define the residual means squared error (RMSE), that we will use to evaluate our models, as RMSE = sqrt(mean((true_ratings - predicted_ratings)^2).

The Simplest Model

Let's start with the simplest of all models, where we always predict the average rating.

method	RMSE
Average Rating	1.059998

We can see that this model has an RMSE of about 1.06. We can do better than that.

The Movie Effect

Next, we will take into account that different movies have different average ratings. Not all movies are great. Good movies are more likely to have higher reviews and bad movies are more likely to have lower reviews.

method	RMSE
Average Rating	1.0599983
Movie Effect Model	0.9434448

Factoring in the "movie effect" improves our prediction from an RMSE of about 1.06 to 0.943. How about the genre of the movie? We've seen that some genres tend to get more favorable ratings than others.

The Movie + Genre Effect

To evaluate if adding the information about genre can help improve our prediction, let's add it to our previous model. For computational efficiency, we will use the combinations of genres available in the genre column.

method	RMSE
Average Rating	1.0599983
Movie Effect Model	0.9434448
Movie + Genre Effects Model	0.9434448

Information about movie genres does not seem to improve our prediction compared to the previous model that only included movie information.

The Movie + User Effect

Let's check what happens when we adjust for different user biases in addition to movie biases.

method	RMSE
Average Rating	1.0599983
Movie Effect Model	0.9434448
Movie + Genre Effects Model	0.9434448
Movie + User Effects Model	0.8659837

Adding user information improves our prediction from an RMSE of about 0.943 to 0.866.

The Movie + User + Year Effect

Next, let's check if adding the year of release to the previous model improves our prediction.

method	RMSE
Average Rating	1.0599983
Movie Effect Model	0.9434448
Movie + Genre Effects Model	0.9434448
Movie + User Effects Model	0.8659837
Movie + User + Year Effects Model	0.8656643

Information about the release year of the movie does not seem to further improve our model.

Model Regularization

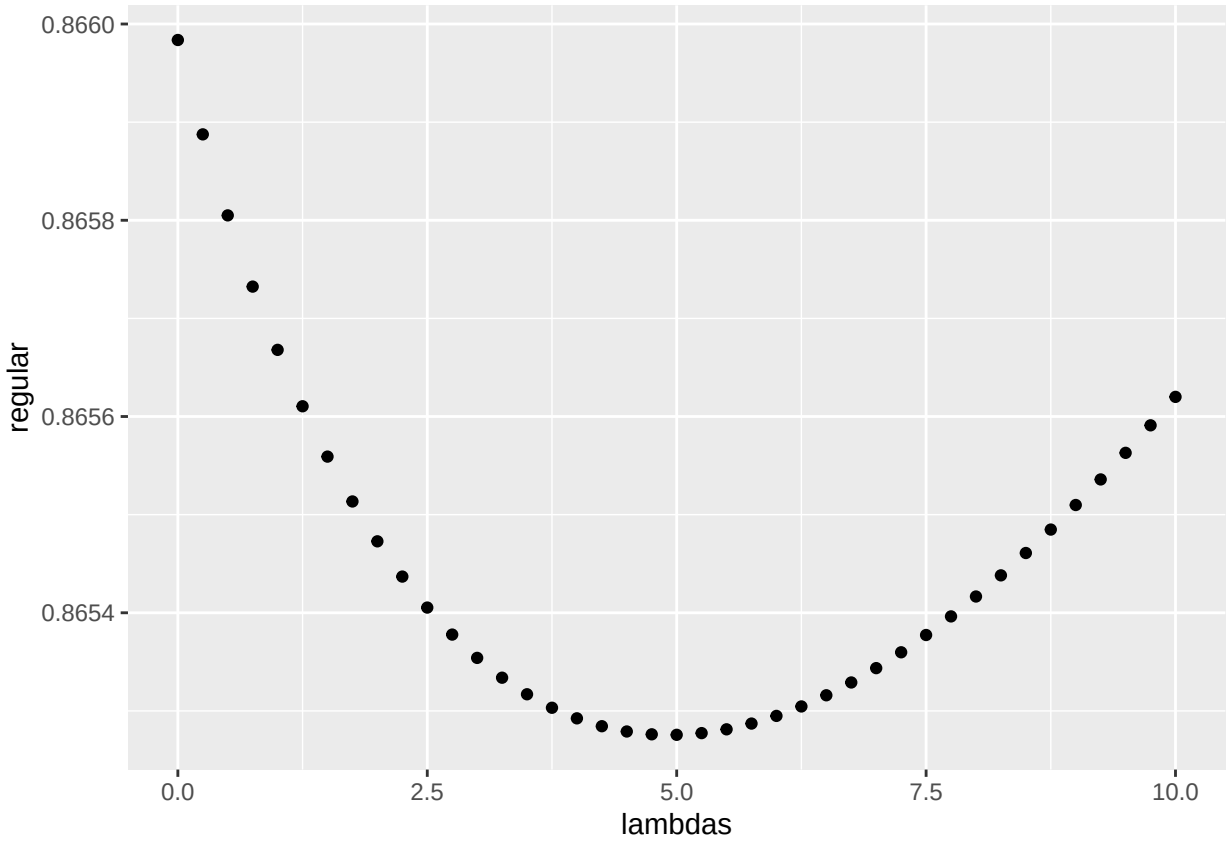
The best model we have trained so far may still suffer from biases. For example, our best and worst reviewed movies seem to be really obscure movies with few ratings.

movieId	title	avg_rating
3226	Hellhounds on My Trail (1999)	5.000000
25789	Shanghai Express (1932)	5.000000
33264	Satan's Tango (Sátántangó) (1994)	5.000000
42783	Shadows of Forgotten Ancestors (1964)	5.000000
51209	Fighting Elegy (Kenka erejii) (1966)	5.000000
53355	Sun Alley (Sonnenallee) (1999)	5.000000
64275	Blue Light, The (Das Blaue Licht) (1932)	5.000000
64280	Hospital (1970)	5.000000
26048	Human Condition II, The (Ningen no joken II) (1959)	4.833333
65001	Constantine's Sword (2007)	4.750000

movieId	title	avg_rating
6606	Purpose (2002)	1.0000000
55324	Relative Strangers (2006)	1.0000000
61348	Disaster Movie (2008)	0.8478261
6483	From Justin to Kelly (2003)	0.8187500
8859	SuperBabies: Baby Geniuses 2 (2004)	0.7843137
7282	Hip Hop Witch, Da (2000)	0.6363636
8394	Hi-Line, The (1999)	0.5000000

movieId	title	avg_rating
59655	Patti Smith: Dream of Life (2008)	0.5000000
61768	Accused (Anklaget) (2005)	0.5000000
64999	War of the Worlds 2: The Next Wave (2008)	0.5000000

We need to account for this bias by constraining the total variability of the effect sizes. We can do that by penalizing larger or smaller estimates that come from small sample sizes. Let's add a tuning parameter lambda, that shrinks the estimates that are outliers.



method	RMSE
Average Rating	1.0599983
Movie Effect Model	0.9434448
Movie + Genre Effects Model	0.9434448
Movie + User Effects Model	0.8659837
Movie + User + Year Effects Model	0.8656643
Regularized Movie + User Effect Model	0.8652756

We can see that lambda = 5 is the best tuning parameter for this model. In addition, regularization improves the results slightly, from RMSE of about 0.866 to 0.865.

Final Test

We will use the final holdout dataset to test our best model, which is the regularized model with movie and user effects.

method	RMSE
Final Regularized Movie + User Effect Model	0.8648177

CONCLUSION

We have explored the MovieLens dataset, looking at the predictive power of different variables, including the movie, user, genre and year of release. The Regularized Movie + User Effect model was the best performing model, with a slight improvement over the second best unregularized Movie + User Effect model. Variables such as genre and year of release did not improve the predictive power of our model, likely because the information carried in these variables was redundant. In the future, this models can be further improved by employing more advanced machine learning models.