MoviesLens.org movie recommendation model

Matthew Hale

Feb-2024

1. Introduction

1a) Project Objective

MovieLens.org is a website that asks users to rate movies they've watched from 0.5 to 5 stars, and based on that information will provide recommendations of other movies that user might like to watch

To provide these recommendations a Machine Learning algorithm has been trained on date the site collects to predict the star rating a user will give to each movie, and return the movies with the highest predicted star ratings as recommendations.

The goal of this project is to train a new Machine Learning algorithm on a large dataset of movie ratings collected from a wide range of MovieLens.org users to predict the star rating each user might give a new film, and do so as accurately as possible.

Applying such a recommender system in a commercial setting would be expected to have significant benefits in terms of customer satisfaction, retention, and therefore commercial performance.

1b) Initial Datasets

The MovieLens data has been sourced from the grouplens.org site here

This data contains over 10 million movie ratings of 10,000 movies by 72,000 users and was released in 2009. Per the ReadMe file at the link above, users were selected at random for inclusion amongst those who had rated at least 20 movies.

1c) Processed/Cleansed Datasets

Initial processing/cleansing was done to join the datasets together into a single flat file, reformat columns where required, and randomly split the data into a training set called **edx** (90% of ratings on which to train the new ML algorithm) and a test set called **final_holdout_test** (10% of ratings which won't be involved in model development and will be used to test the performance of the final model produced).

Please see MovieLensProjectCode.R for the initial cleansing/processing script.

The rowcounts for the two sets are below, showing the ~90/10 split.

```
Number of rows in training set edx = 9,000,055
```

```
Number of rows in test set final_holdout_test = 999,999
```

The 6 columns in the training and test datasets are per the below, listing the data type in each column and some sample entries:

Table 1: Variables in edx training dataset

Variable	Type	Examples
userId	int	50805 42011 45985 19660 46073 13268 6910 30265 65174 18227
movieId	int	1095 6303 593 4210 186 485 596 2402 913 6238
rating	num	4 3.5 4 4 4 2 2.5 2 3.5 3
timestamp	int	952366935 1121891132 854490846 1207988321 1123689113 923662409 1190501210 974970828 1159816589
title	chr	"Glengarry Glen Ross (1992)" "Andromeda Strain, The (1971)" "Silence of the Lambs, The (1991)"
genres	chr	"Drama" "Mystery|Sci-Fi" "Crime|Horror|Thriller" "Action|Crime|Drama|Horror|Thriller" ...

During the course of the exploratory data analysis further processing was done on the data to aid analysis and modelling. This included:

- Splitting the title variable into two: one variable with the title excluding film year, one with the film year (called filmYear)
- Creating the filmDecade variable using filmYear (1980s, 1990s etc)
- Creating the ratingDate variable using the timestamp on the review (and removing the timestamp variable)
- Creating Genres_Cnt variable, counting the number of genres associated with each movie per the genres variable
- Adding a binary variable for each of the 19 genres in the dataset. So 'Comedy|Romance' would be marked as Comedy= TRUE, Romance = TRUE, and all other genres as false (e.g. Action=FALSE, Adventure=FALSE etc).

1d) Modelling Methodology

After initial exploratory data analysis, the predictive model was developed using the edx set, with a number of linear models evaluated using that dataset. K-fold cross validation was conducted on all models to try and ensure the model produced would be 'generalisable' to new data.

The final, selected model was then applied to the final_holdout_test set, and evaluated using root mean squared error (RMSE) - a measure of the deviation between the ML algorithm's predictions and the actual responses of users in the test set. Achieving an RMSE as low as possible in the test set, and getting a closer alignment between predicted and actual ratings, is the ultimate goal of this project. For more see wikipedia here

Whilst other metrics could be used to assess how good a model is, RMSE has a history in movie recommender systems - being the metric of choice for the Netflix challenge in 2006, for example.

2. Data Analysis

The exploratory data analysis conducted is documented below. This primarily involved reviewing the distributions of the variables in the dataset and the movie ratings patterns across those variables, given the modelling task of predicting movie ratings. As part of this process, new variables were created for analysis and modelling, as described in section 1 above.

2.1 Data sparseness

The first step in the process was to review the coverage in the dataset - i.e. how many user/movie combinations we have ratings for and how many gaps there are.

The edx dataset contains the following number of unique users and unique movies:

```
   users movies
1 69,878 10,677
```

If every user had rated every movie in the data, we would have a total ratings volume of:

```
  Combinations
1  746,087,406
```

Given the number of ratings in the edx dataset is just over 9 million, the percentage coverage is therefore:

```
[1] "1.21%"
```

This unsurprising result, that each user has only watched a small fraction of the movies out there, shows we have a sparse matrix with a lot of gaps.

This can be visualised as follows by looking at 100 randomly selected users, and 100 randomly selected movies watched by this user group. The colour of each square indicates the star rating given by the given user for the given film, and no colour shows that the user has not rated the film.

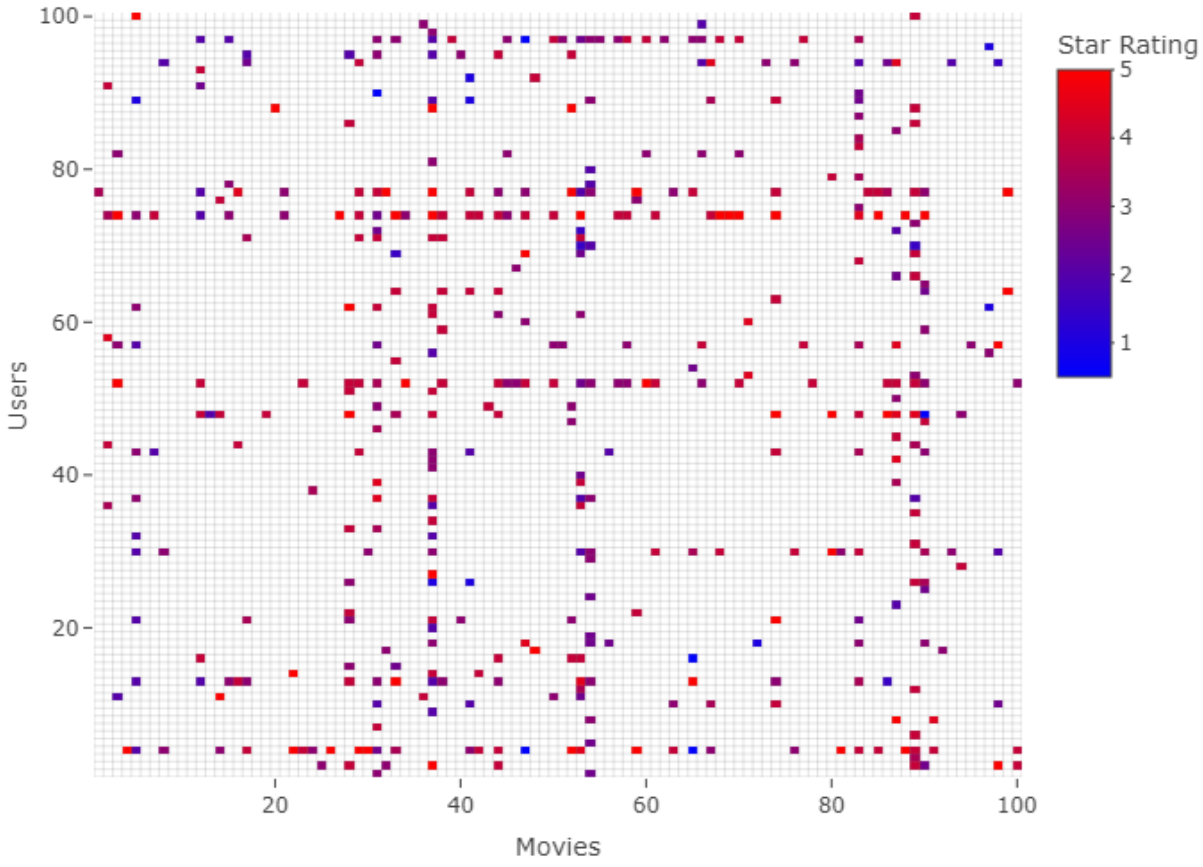

Figure 1: 100 Random Users' Ratings for 100 movies

Here we can see the large % of gaps where we don't have a rating for a movie from a given user.

You can also start to see a number of other insights in the plot, which again are aligned with expectation:

1. from the more populated horizontal lines, you can see that there are some 'super users' in the dataset who have rated a lot more movies than others;
2. some users seem to rate generally higher than other (a tendency for reds in the horizontal lines)
3. from the more populated vertical lines you see that some films have a lot more reviews than others;
4. some films look to be generally rated higher than others.

2.2 Movie Analysis

2.2.1 Most/Least Frequently Rated Movies We can confirm the view that some of the 10,677 movies in the dataset have a lot more reviews than others by plotting a histogram of the distribution of movie ratings.

Understanding the differences in volume of ratings per movie is important, as we might expect more accurate predictions from our models where those predictions are backed by good volumes of data. And vice-versa, poorer predictions where we have a paucity of data.

The histogram below clearly shows a much higher number of films with few ratings, and very few films with up to 30,000+ ratings. Note, the LHS chart is the full distribution, the RHS chart shows the pattern holding for films with up to 200 ratings, with most films having a lower volume of ratings:

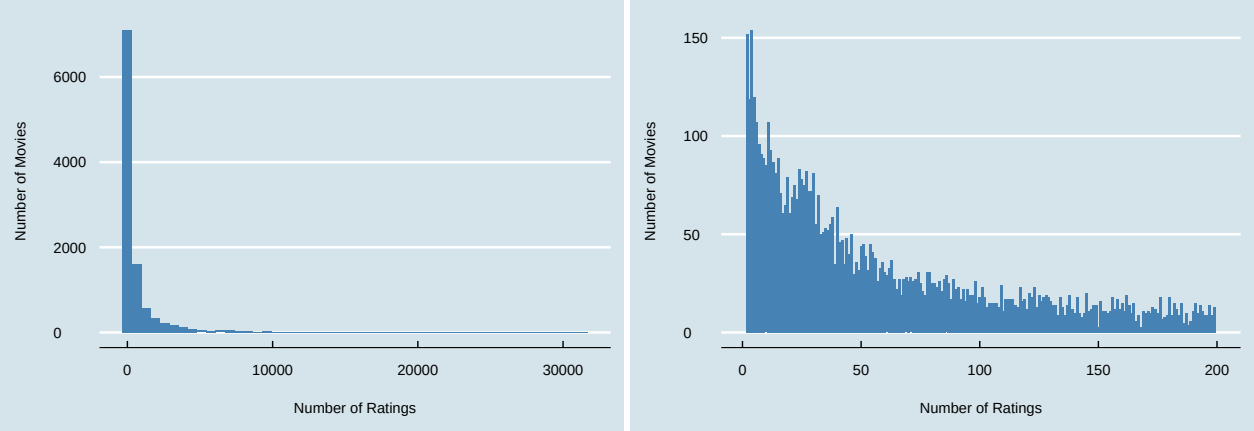

Figure 2: Distribution of Ratings Volume per Movie

The tables below show the top 10 most and bottom 10 movies in terms of ratings volume:

Table 2: Movies with most ratings

title	ratingsVol
Pulp Fiction	31362
Forrest Gump	31079
Silence of the Lambs, The	30382
Jurassic Park	29360
Shawshank Redemption, The	28015
Braveheart	26212
Fugitive, The	26020
Terminator 2: Judgment Day	25984
Star Wars: Episode IV - A New Hope	25672
Batman	24585

Table 3: Movies with least ratings

title	ratingsVol
1, 2, 3, Sun (Un, deuz, trois, sole	1
100 Feet	1
4	1
Accused (Anklaget)	1
Ace of Hearts	1
Ace of Hearts, The	1
Adios, Sabata (Indio Black, sai che	1
Africa addio	1
Aleksandra	1
Bad Blood (Mauvais sang)	1

As you'd expect, the most reviewed films are Hollywood blockbusters, whereas the least reviewed films (a subset of the films with 1 rating) are probably not ones you're heard of:

And the mean and median volume of ratings per movie is as below, with the large difference between mean and median indicating the skewing of the distribution. i.e. if you line up all the movies from least to most ratings, the mid-point movie has 122 ratings, but the overall mean is 842 ratings per movie, which really shows the outsize impact of the handful of blockbuster movies.

```
  Mean_RatingsPerMovie Median_RatingsPerMovie
1                842.9                     122
```

2.2.2 Highest/Lowest Rated Movies In the data we can also start to see the unsurprising result that there is variance in the average rating given to each movie, with some movies generally seen to be better than others.

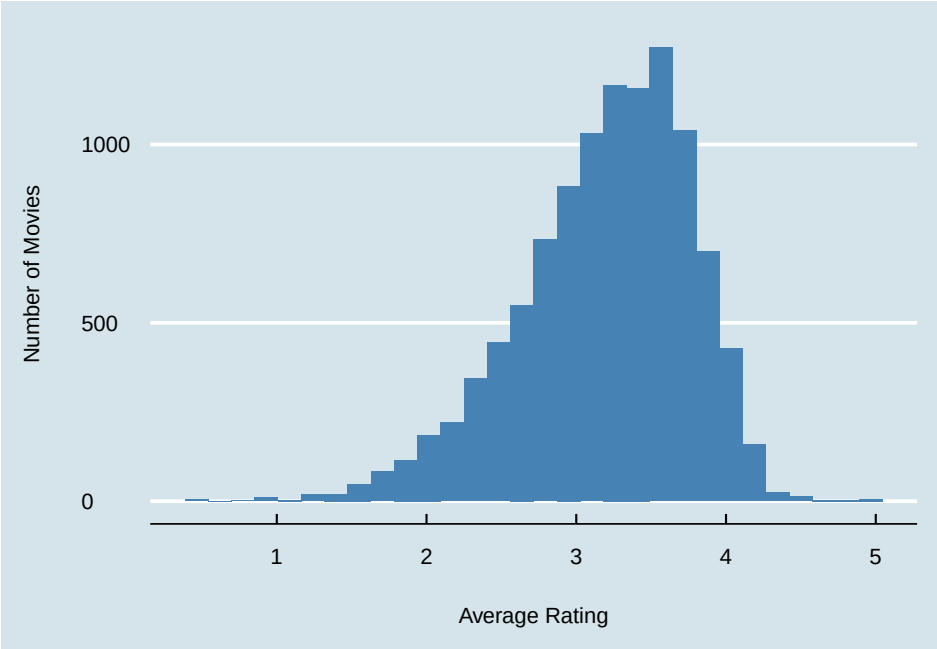

Figure 3: Distribution of Average Ratings per Movie

The tables below give the best rated and worst rated movies that have at least 100 reviews:

Table 4: Movies with highest ratings

title	Avgrating
Shawshank Redemption, The	4.46
Godfather, The	4.42
Usual Suspects, The	4.37
Schindler's List	4.36
Casablanca	4.32
Rear Window	4.32
Sunset Blvd. (a.k.a. Sunset Bouleva	4.32
Double Indemnity	4.31
Paths of Glory	4.31
Seven Samurai (Shichinin no samurai	4.31

Table 5: Movies with lowest ratings

title	Avgrating
From Justin to Kelly	0.90
Pokémon Heroes	1.03
Glitter	1.18
Pokemon 4 Ever (a.k.a. Pokémon 4: T	1.18
Barney's Great Adventure	1.19
Gigli	1.19
Son of the Mask	1.30
House of the Dead, The	1.35
Turbo: A Power Rangers Movie	1.36
Pokémon 3: The Movie	1.38

This is all looks aligned with expectation, with some films that always rank highly in best ever film lists featured in the top 10 ranking, and some less highly respected films in the bottom 10 - the Pokemon films in particular seem to have fared badly, taking 3 spots of the bottom 10.

2.2.3 Most/Least Rated Movie Genres, and Highest/Lowest Rated We now come up a level and look at movies grouped into genres. Per the tables below, we can see that each movie can have multiple genres associated with it in the data, and we can start to see the most rated genre combinations and the highest/lowest rated.

The total volume of genre combinations in the data is:

```
  genres
1    797
```

The top and bottom ten genre combinations (with at least 1000 ratings) are as follows:

Table 6: Genre Combinations with highest ratings

genres	AvgRating
Drama|Film-Noir|Romance	4.30
Action|Crime|Drama|IMAX	4.30
Animation|Children|Comedy|Crim	4.28
Film-Noir|Mystery	4.24
Crime|Film-Noir|Mystery	4.22
Film-Noir|Romance|Thriller	4.22
Crime|Film-Noir|Thriller	4.21
Crime|Mystery|Thriller	4.20
Action|Adventure|Comedy|Fantas	4.20
Crime|Thriller|War	4.17

Table 7: Genre Combinations with lowest ratings

genres	AvgRating
Action|Children	2.04
Action|Adventure|Children|Come	2.06
Crime|Sci-Fi|Thriller	2.18
Action|Adventure|Fantasy|Thril	2.21
Action|Adventure|Comedy|Fantas	2.26
Action|Children|Fantasy	2.28
Children|Comedy|Sci-Fi	2.28
Action|Comedy|Musical	2.43
Action|Crime|Fantasy	2.49
Fantasy|Horror|Thriller	2.59

So we can start to see a preference in the set for Crime, Thrillers, Dramas. And Sci-Fi, Children's Films and Fantasy falling lower down the list.

Taking things a step further and breaking out the genres field into binary variables, then looking at the ratings associated with each genre, we can get a clearer view.

Note, given each movie can be assigned to more than one genre, the total volume of ratings in the chart below sums to more than the total number of ratings in the edx dataset.

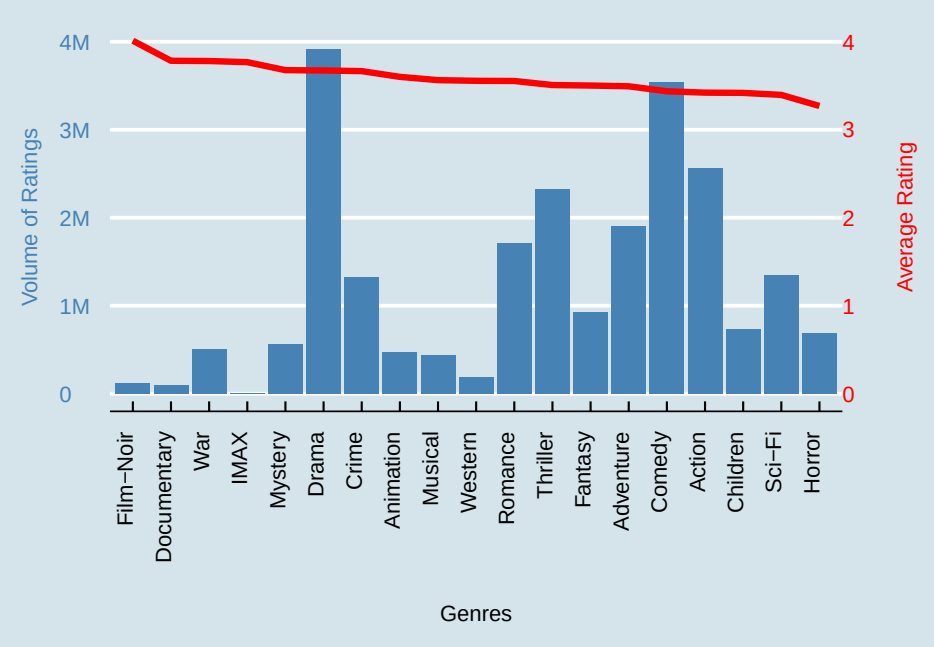

Figure 4: Average Ratings and Ratings Volume per Genre

From this new view using the binary variables, we can see that Drama, Comedy, Action and Thriller movies are the most rated, and the highest rated genres on average are Film-Noir and Documentary, although they have few ratings. Of genres with 1m ratings +, it's Drama and Crime that are the highest rated.

The worst rated on average are Horror and Sci-Fi movies.

2.2.4 Number of genres associated with a movie Another avenue of investigation around genres was whether the volume of genres associated with a given movie would have any bearing on ratings for that movie - with an initial speculative hypothesis being that more genres, and an identity crisis for the movie, might lead to lower ratings.

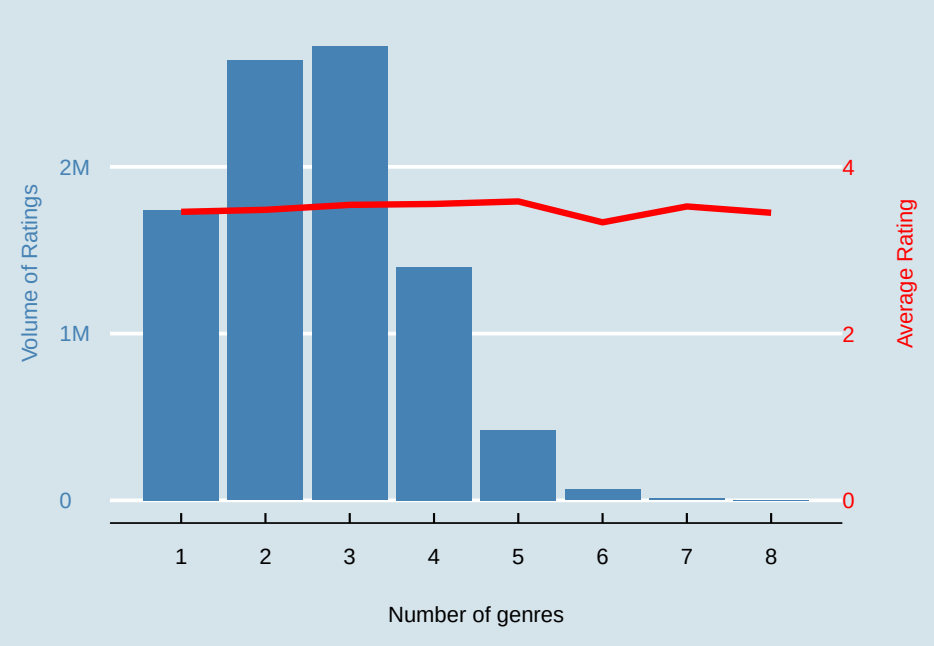

Figure 5: Average Rating and Ratings Volume by number of Genres associated with a Movie

The visual above doesn't support the view that the number of genres linked to a film bears much relation to the rating given, but the genre preferences in the dataset certainly look like they could potentially be exploited in the modelling stage.

2.2.4 Distribution and Ratings of Movies by Movie Release Year One other piece of information we have about each movie is the year in which it was released. The chart below shows the distribution of movies by year, and the average ratings associated with each.

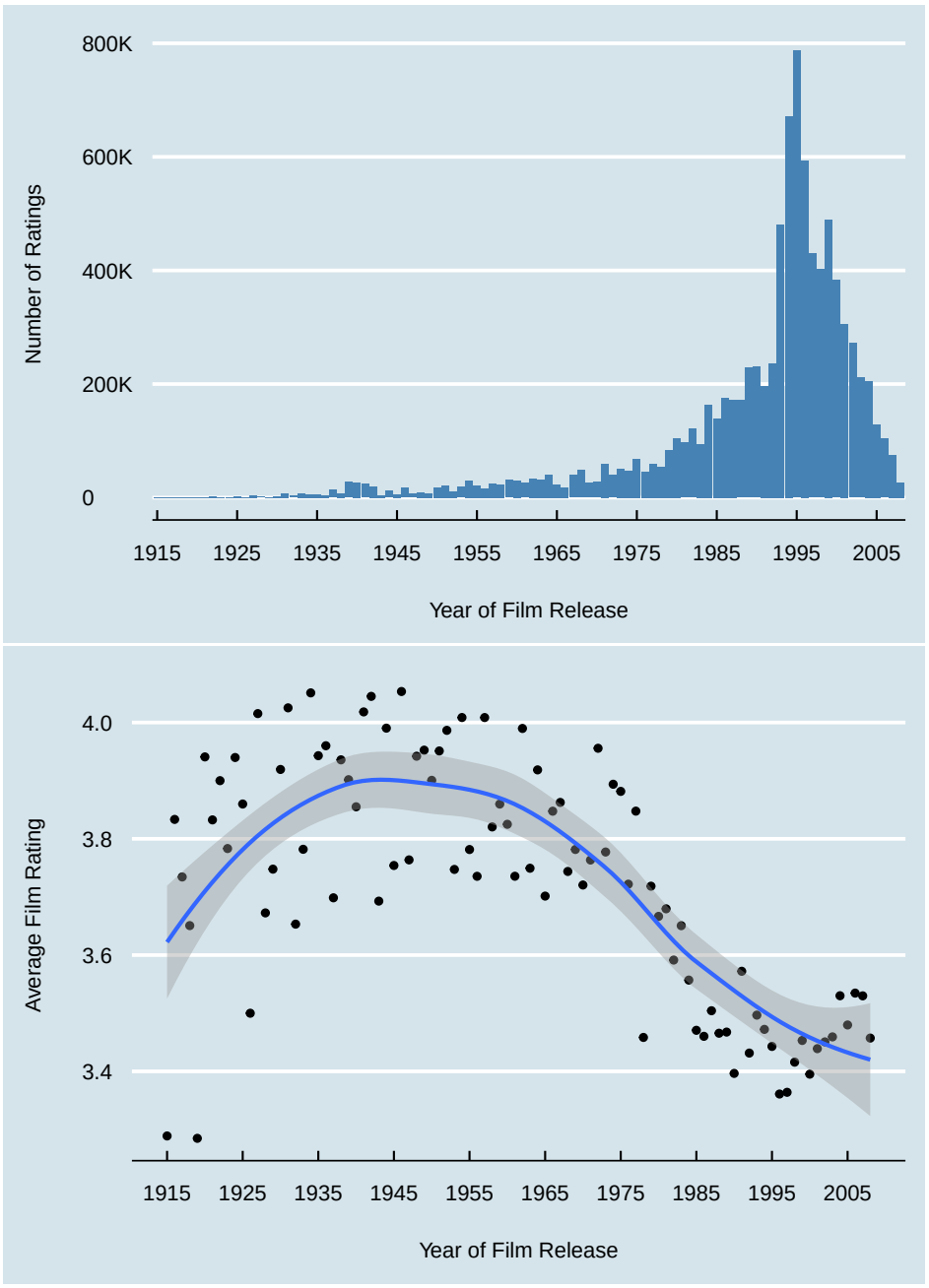

Figure 6: Distribution of Ratings and Average Rating by Movie Year

The chart shows a skewing of the ratings volume towards the 1980s, 1990s, and 2000s, with the 1990s in particular dominating.

In terms of ratings, it's years in the 'Golden Age' of Hollywood that have higher movie ratings on average, dropping off from the 1960s onwards.

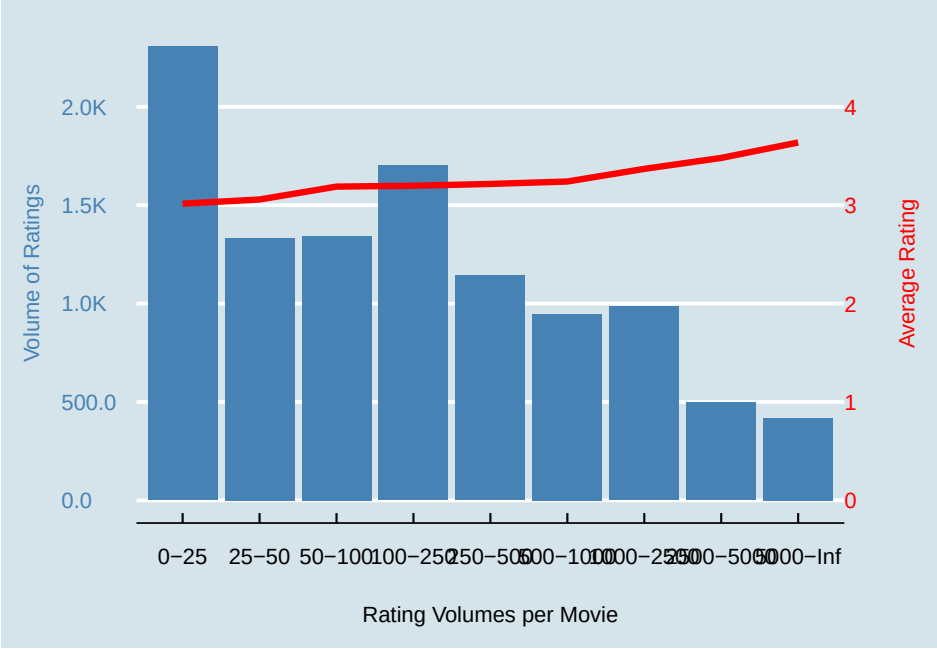

Figure 7: Average Ratings by Number of Ratings a Movie has had

So interestingly there does appear to be a correlation - movies with a higher number of reviews tend to be better rated on the average than films with fewer reviews.

2.3 User Related

2.3.1 Distribution of Volume of Ratings per User Moving onto a focus on the 69,878 different users in the dataset, we can start by looking at the distributions of the number of ratings per user.

Again, you'd expect to be able to make better predictions for users who have given you more information about their movie preferences through their ratings.

The histograms below (LHS being the full distribution, RHS zooming in to those with under 200 ratings) shows most users having reviewed smaller volumes of films, with a handful of enthusiasts having rated thousands of movies.

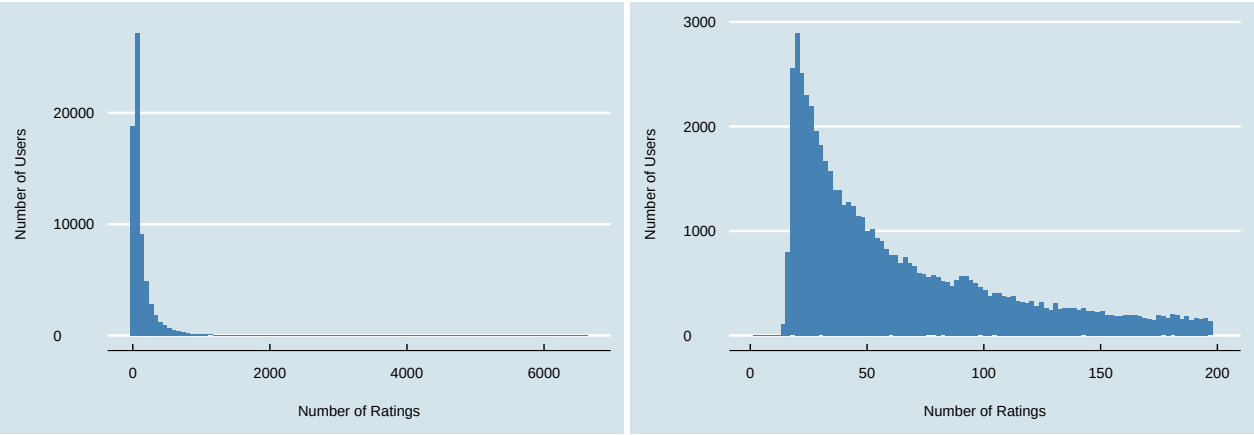

Figure 8: Distribution of Ratings Volume per User

This is clear from the tables below, which show the top and bottom ten users in terms of review volumes.

Table 8: Movies with most ratings

userId	ratingsVol
59269	6616
67385	6360
14463	4648
68259	4036
27468	4023
19635	3771
3817	3733
63134	3371
58357	3361
27584	3142

Table 9: Movies with least ratings

userId	ratingsVol
62516	10
22170	12
15719	13
50608	13
901	14
1833	14
2476	14
5214	14
9689	14
10364	14

And the mean and median volume of ratings per user is as below, with the difference between mean and median also indicating the skewing of the distribution (albeit to a lesser extent than the skewing of the ratings per movie distribution).

```
  Mean_RatingsPerUser Median_RatingsPerUser
1               128.8                    62
```

2.3.2 Distribution of Average Rating per User In terms of ratings, we can plot that distribution to see the unsurprising result that some users on average will give out higher ratings than others.

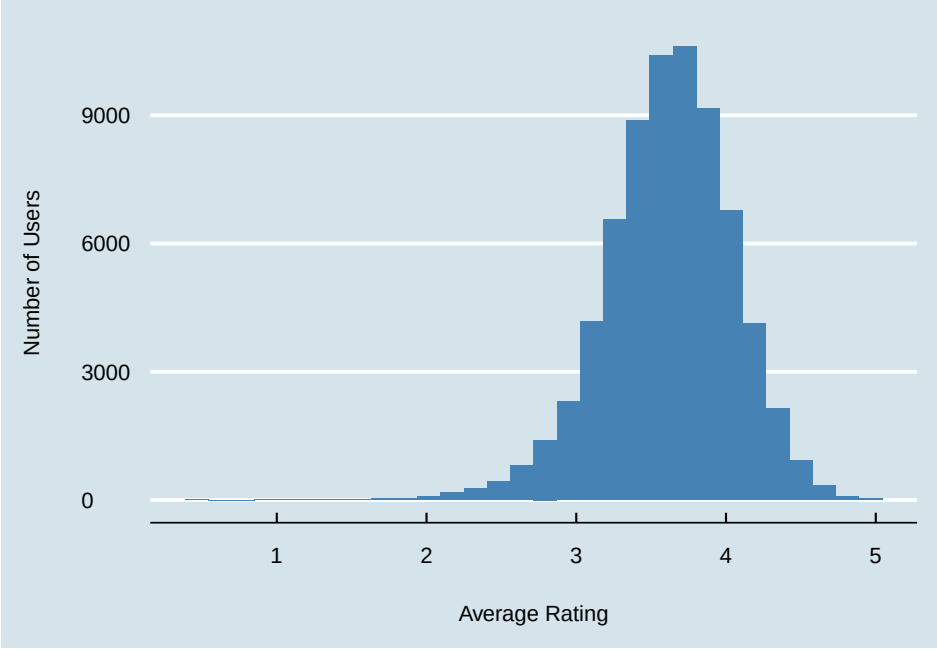

Figure 9: Distribution of Average Ratings per User

2.3.3 Date of rating vs Average Rating We have no information about the users in the data beyond their ratings for each movie they rated (no demographic information etc). We do, however, have date of

their reviews, which we can review to see how it may relate to movie ratings.

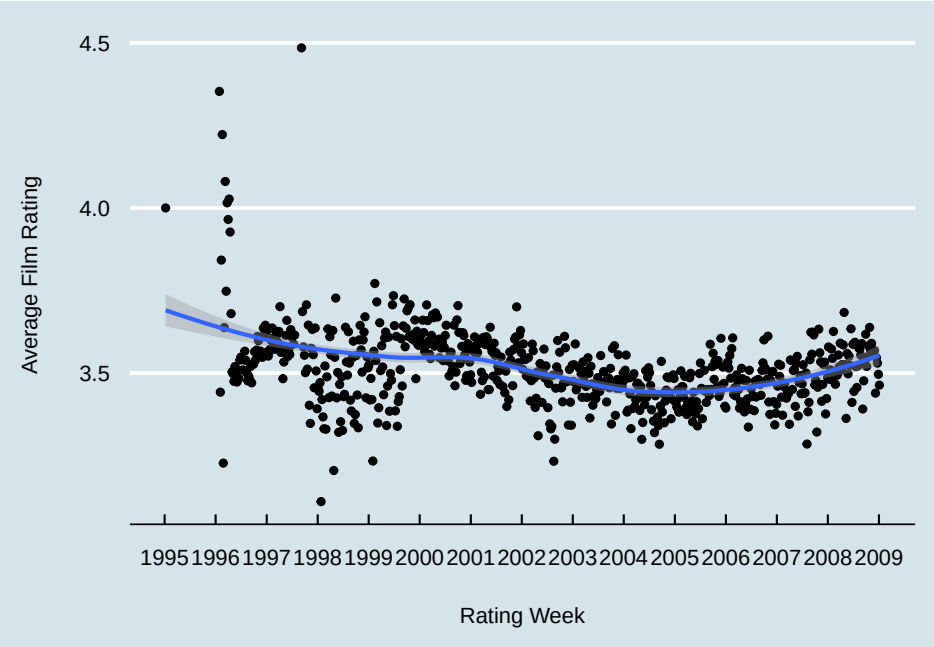

Figure 10: Average Ratings by Rating Week

This shows that the ratings were given in the period from the mid-90s to the late 2000s, and there appears to be a general decline in the ratings from the mid-90s to the mid-00s, followed by a tick up as we move towards 2010.

2.3.4 Hour of rating vs Average Rating One other piece of information we have is the time of each rating. We can therefore plot this to see if this has any impact on ratings.

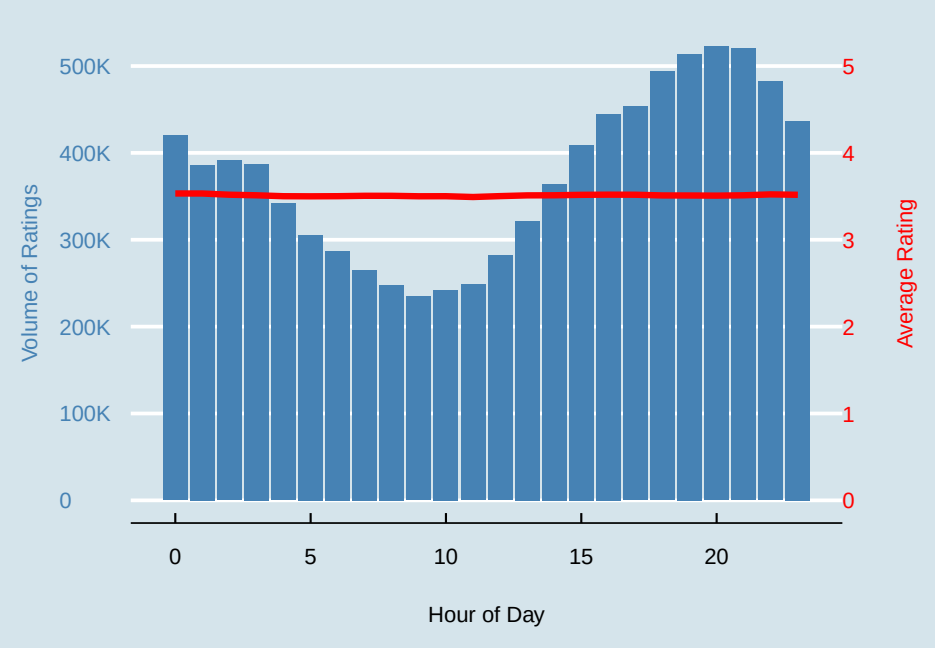

Figure 11: Average Ratings by Rating Hour of Day

Whilst interesting to see the daily pattern of ratings volume (with most ratings being done in the evenings) there appears to be very little impact on the ratings given.

2.3.5 Days between Rating and User's first rating vs Average Rating The chart below now looks at the days between the rating being given and the first time the user rated any movie, to see if users start off more optimistic and become harsher critics over time.

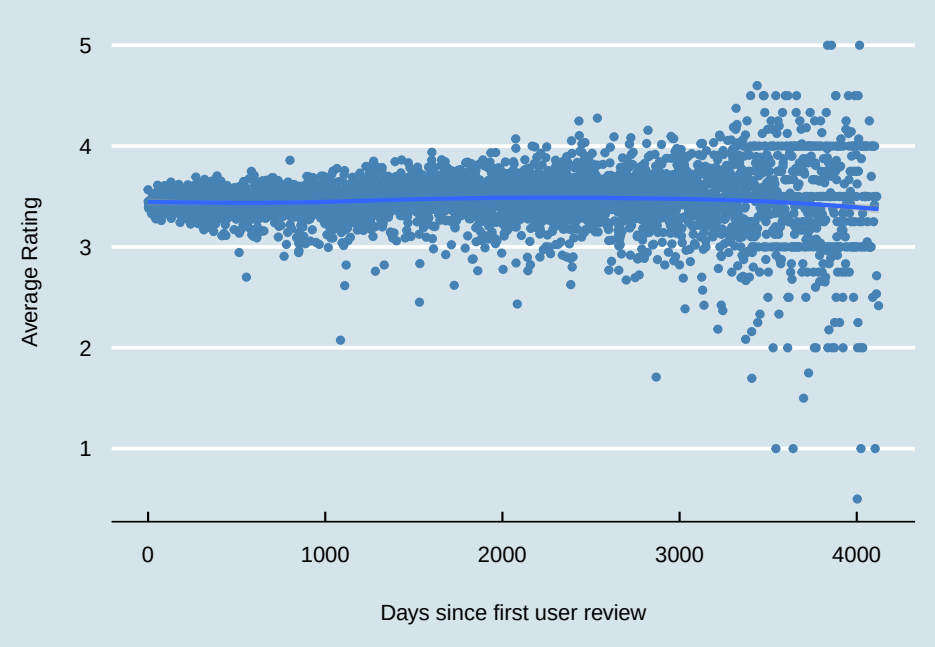

Figure 12: Average Ratings by Days since first user rating

Again, this appears to have little impact on ratings at the aggregate level.

2.3.5 Days between Rating and Movie's first rating vs Average Rating The chart below now looks at the days between the rating being given by a specific user and the first time the movie was ever rated by any user.

Figure 13: Average Ratings by Days since first movie rating

Overall beyond the spike on the RHS, which appears to be associated with limited volume given the increase in the spread of the data, there is limited evidence of a trend. You might expect higher ratings initially, if it's generally a movie's fans who are watching the movie first post release, and then a reversion to the mean. This looks like it may be happening in the first year to a small extent, but beyond that the pattern feels hard to explain.

It should be remembered that the above ratings plots are only 2-dimensional though, and there may well be partial correlations once other factors are accounted for (in this case, things like the year of film release, given a good number of movies were released prior to the start of the rating window)

2.3.6 Volume of User's Reviews vs the Average Rating As well as when they rated, we also know how many films a user rated. The below groups together user with similar volumnes of reviews to understand if this impacts ratings given.

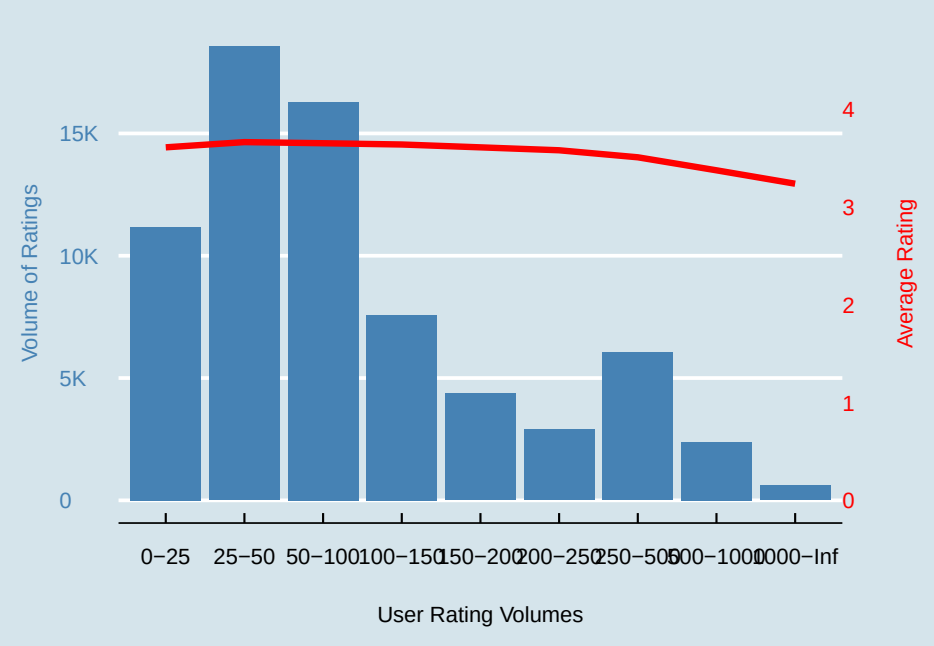

Figure 14: Average ratings by number of movies a user has rated

Interestingly this does look like it may impact to an extent, with the super users who have rated several hundred movies or more tending to be harsher critics than those with fewer reviews. Or an alternative reading could that users with more ratings will have watched a broader range of movies, and have gone beyond the popular into the realms of the less popular, which (although not always the case) may be less popular for a reason.

2.4 User and Movie Related: User's movie type preferences

We can also look at specific users, and start to see their genre preferences and movie year preferences emerge. For example, if we select a specific user at random with over 500 ratings given, we can plot the ratings per genre and pull out their favourite films. We can also see if they appear to have a preference for movies from a certain era.

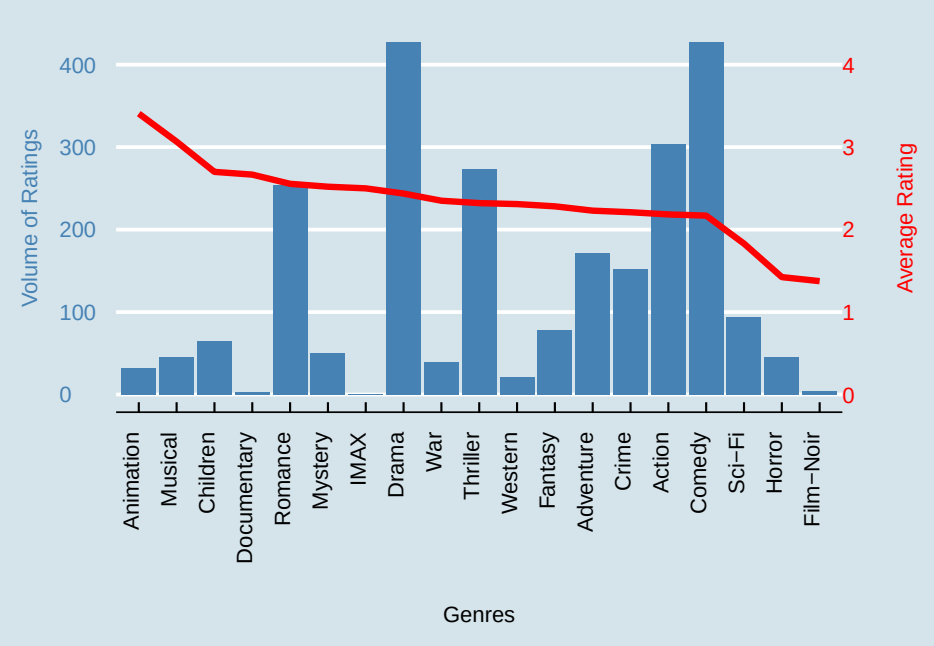

Figure 15: Genre ratings for chosen super user

So we can start to see the specific users preferences begin to appear from the chart above. They seem to be a low rater overall vs the whole user group first of all, but clearly have a distinct preference for Animation, Musicals and Children's movies - a pattern that looks quite different to what we saw for the whole user group, particularly for the Children's movie genre which ranked poorly overall.

At the other end of the scale, they rank Film Noir - a polar opposite to the rating for the user group as a whole, although admittedly with a low volume of movies rated.

If we pull out the top and bottom 40 movies for this user, we can see more information emerge:

Table 10: Movies with best ratings

title	Avgrating
Beauty and the Beast	5.0
Casino Royale	5.0
Dirty Dancing	5.0
Few Good Men, A	5.0
Finding Nemo	5.0
Finding Neverland	5.0
Little Mermaid, The	5.0
Lord of the Rings: The Fellowship o	5.0
Lord of the Rings: The Return of th	5.0
Lord of the Rings: The Two Towers,	5.0
Monsters, Inc.	5.0
Overboard	5.0
Pinocchio	5.0
Pretty Woman	5.0
Robin Hood: Prince of Thieves	5.0
Shrek	5.0
Shrek 2	5.0
St. Elmo's Fire	5.0
Titanic	5.0
Top Gun	5.0
About Last Night...	4.5
Absolute Power	4.5
Adventures in Babysitting	4.5
Aladdin	4.5
Bend It Like Beckham	4.5
Bridget Jones: The Edge of Reason	4.5
Cocktail	4.5
Coyote Ugly	4.5
E.T. the Extra-Terrestrial	4.5
Firm, The	4.5
Four Weddings and a Funeral	4.5
Ghost	4.5
Gorillas in the Mist	4.5
Heaven Can Wait	4.5
How to Lose a Guy in 10 Days	4.5
In the Name of the Father	4.5
Jungle Book, The	4.5
Karate Kid, The	4.5
Lethal Weapon 2	4.5
Lion King, The	4.5

Table 11: Movies with worst ratings

title	Avgrating
2 Fast 2 Furious	0.5
Accidental Tourist, The	0.5
After Life (Wandafuru raifu)	0.5
Amelie (Fabuleux destin d'Amélie Po	0.5
American Psycho	0.5
Anaconda	0.5
Aviator, The	0.5
Big Lebowski, The	0.5
Bill & Ted's Bogus Journey	0.5
Blade Runner	0.5
Blow Dry (a.k.a. Never Better)	0.5
Boomerang	0.5
Bowfinger	0.5
Brewster's Millions	0.5
Butcher's Wife, The	0.5
Croupier	0.5
Crow, The	0.5
Deep Blue Sea	0.5
Divorce, Le	0.5
Do the Right Thing	0.5
Duets	0.5
Evolution	0.5
eXistenZ	0.5
Faculty, The	0.5
Falling Down	0.5
Fear and Loathing in Las Vegas	0.5
Freaky Friday	0.5
From Dusk Till Dawn	0.5
Gods Must Be Crazy II, The	0.5
Gods Must Be Crazy, The	0.5
Grease 2	0.5
Gremlins	0.5
Hamlet	0.5
Haunting, The	0.5
House on Haunted Hill	0.5
Hulk	0.5
Jaws	0.5
Jaws 2	0.5
Jaws 3-D	0.5
Jaws: The Revenge	0.5

So here we can see a good number of animated films given a 5 star rating: Beauty and the Beast, The Little Mermaid, Monsters Inc etc, as well as some other info not captured in the genre breakdowns; for instance there seems to be a soft spot for British films here too with the 5-star ratings for Bend it Like Beckham, Four Weddings, Casino Royale and Bridget Jones. And also the impact of franchises is visible, with the Lord of the Rings trilogy all rated 5-star.

On the other hand, in the lowest rated films we can see a good number of comedies and Sci-Fi films - even

respected films like Blade Runner getting a half star rating. The franchise effect is evident here as well - having rated Jaws, Jaws 2 and Jaws 3-D at half a star apiece, it's not surprising to see Jaws: The Revenge also faring poorly.

3. Modelling Methodology and Results

Having analysed the training dataset to improve understanding and to assess potential model drivers, the modelling process was undertaken to construct a prediction of how each user would rate movies they hadn't previously rated.

3.1) Modelling Methodology

A set of linear models was built to try and achieve an accurate predictive model without overfitting to the training data. Initial work was done using the processed edx training set as a whole, and k-fold cross validation was then done to get RMSE results more likely to be representative of the performance in the final holdout set.

The 5 model variants trialled are:

1. Using the simple mean rating across all ratings in the training set as a baseline
2. Adding an average movie bias to reflect that some movies are better/worse ranked than the average movie
3. Adding an average user effect to reflect some users generally rate higher/lower than average
4. A regularized version of model 3, adding a penalty for users/movies with low rating volumes
5. An amended version of model 3, adding user specific genre-preference biases - only genres rated 5 or more times by a given user alter the prediction, and the average of the user's biases across the genres referenced for a specific movie are used.

3.2) Modelling Results

The average RMSE in the validation sets when performing 5-fold cross validation, for each of the 5 models, are below:

Table 12: Average RMSE from validation sets in 5-Fold Cross Validation

Model	RMSE
1	1.0603301
2	0.9437003
3	0.8661773
4	0.8654820
5	0.8513982

Given the result in the cross validation, model 5 was chosen as the best model of those trialled. Having selected model 5, the final holdout test set was then used for the first time in the project as the acid test of the model's performance. The final RMSE using the test set is given below.

```
[1] 0.8489658
```

4. Conclusion

The model appears to give a good prediction by capturing the impact of the average movie rating, the deviation from the overall average for each movie, the deviation from the average for each user, and user-specific genre preferences.

With more time and resources, the following are potential avenues for further investigation:

1. Regularizing the chosen model. Whilst a minimum volume of 5 genre-specific ratings was used as a threshold for inclusion in the final model, which does add a regularizing effect, this hyperparameter could be selected via k-fold cross validation, or alternatively a penalty term could be added into the objective function when minimising root mean squared error, per model 4
2. The addition of other effects could be assessed such as user-specific movie year preference
3. Other groupings of movies beyond genre could be assessed using matrix factorisation; this may reveal other structures in the data such as user preferences for specific franchises (e.g. Harry Potter, Lord of the Rings etc), or blockbusters, or superhero movies etc.

It would also be preferable to continue investigating with more than 8gb RAM to give more freedom to investigate other model structures or drivers - managing memory issues (e.g. through breaking the data up into chunks etc) was a significant issue in the course of this project.

Finally, whilst trying to utilise the available data to build an effective model, there are numerous other personal attributes which you'd expect would correlate with a user's preference for certain movies. Adding richer demographic data into the modelling set may well also be an avenue to improved model performance.

Overall however, the final model presented here does give a significant enhancement in accuracy versus a naive movie averages approach, and the benefits of improved prediction in a commercial context is likely to be higher customer satisfaction and higher retention, which would be expected to drive commercial performance.