MovieLens Project - Machine Learning Submission

HarvardX Data Science Capstone - PH125.9x

Simon Gibson

2023-12-02

Introduction

For the 9th Course in the HarvardX Data Science course we have been asked to build a movie recommendation system using the MovieLens dataset. This report will cover the initial creation of the data set, exploration of the data, creation and refinement of the algorithm.

This movie recommendation system is similar to systems used by many companies such as Amazon and Netflix to recommend movies, books, and music to customers.

The Movielens data package can be found at the MovieLens homepage.

MovieLens is a project run by GroupLens - a research lab run at the University of Minnesota in North America. MovieLens is a non-commercial collection of movie data and the main set of data contains over 20 million ratings for over 27,000 movies. In this project we are using the 10M dataset.

In order to test the results of the recommendation system we are using the root-mean-square error (RMSE) to measure the difference between the values predicted by the model and the observed values. For this project a RMSE score of less than 0.86490 is the goal.

Methods

The data is divided into 2 sets. The first set is used to train the algorithm and the second set is used to validate the algorithm. By dividing the data the problem of over-training and thus producing skewed results can be avoided.

The creation of the 2 sets involves the following steps. Initially required packages are installed if not installed and then loaded. Next the data is downloaded if the zip files are not found. Column names are set and the data is converted into forms more easily processed. Then the data is joined. Finally the joined data is split into 2 sets - the edx set used to train the algorithm and the final_holdout_test set that will be used to validate the algorithm and calculate the final RMSE score.

Data Exploration

To start with we use the head command to view the first 10 rows of data.

Looking at the first 5 rows of the data in the edX data set we can see the columns we have to work with - userId, movieId, rating, time stamp, title and genre.

Some initial areas of interest here are the time stamp and genres columns. As time passes do movies get higher ratings?

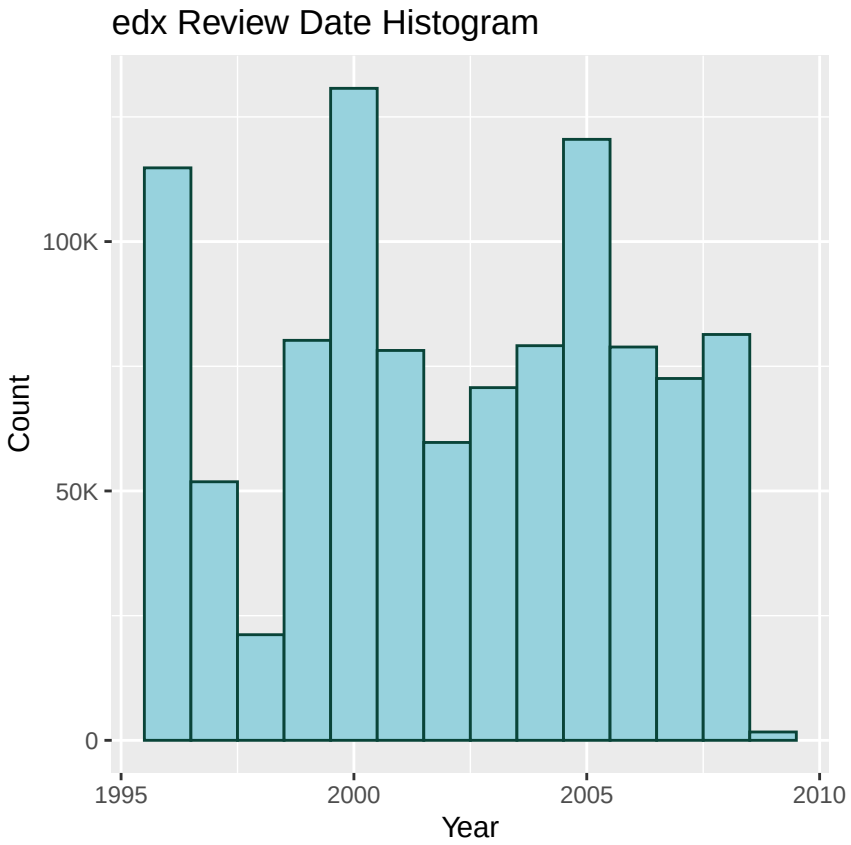
If we take the example of literature, works such as those by the likes of Homer and Shakespeare survive while over time lesser works are weeded out. Possibly there is some survivability bias that means that movies that continue being reviewed are ones that people have enjoyed and have been recommended, for example through word of mouth or via similar recommendation engines.

Table 1: EDX Dataset Overview - First 10 Rows

	userId	movieId	rating	timestamp	title	genres
2	1	185	5.0	838983525	Net, The (1995)	Action|Crime|Thriller
20	1	589	5.0	838983778	Terminator 2: Judgment Day (1991)	Action|Sci-Fi
23	2	110	5.0	868245777	Braveheart (1995)	Action|Drama|War
34	2	786	3.0	868244562	Eraser (1996)	Action|Drama|Thriller
49	3	1252	4.0	1133571071	Chinatown (1974)	Crime|Film-Noir|Mystery|Thriller
55	3	1597	4.5	1133571226	Conspiracy Theory (1997)	Drama|Mystery|Romance|Thriller
78	4	39	3.0	844417037	Clueless (1995)	Comedy|Romance
83	4	165	5.0	844416699	Die Hard: With a Vengeance (1995)	Action|Crime|Thriller
87	4	266	5.0	844417070	Legends of the Fall (1994)	Drama|Romance|War|Western
91	4	329	5.0	844416796	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi

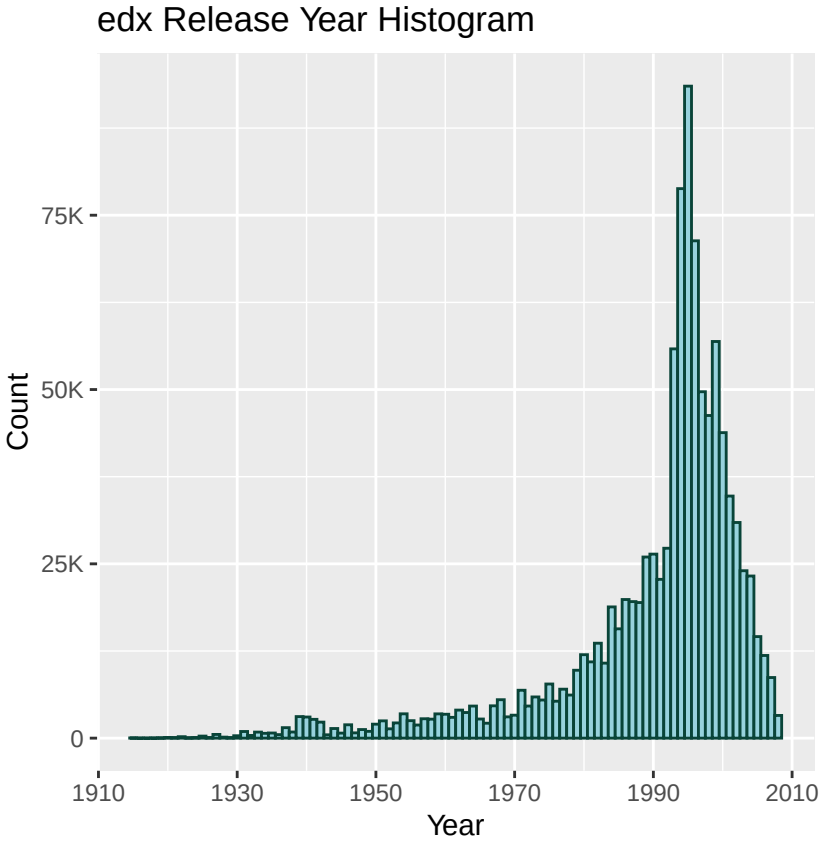
The genre column also shows collections of genre keywords, rather than single genres. These collections could also prove to be useful.

If we convert the time stamps, we can see that the oldest review is dated 1996-01-29 13:00:00 and the most recent time stamp is 2009-01-05 17:51:47.

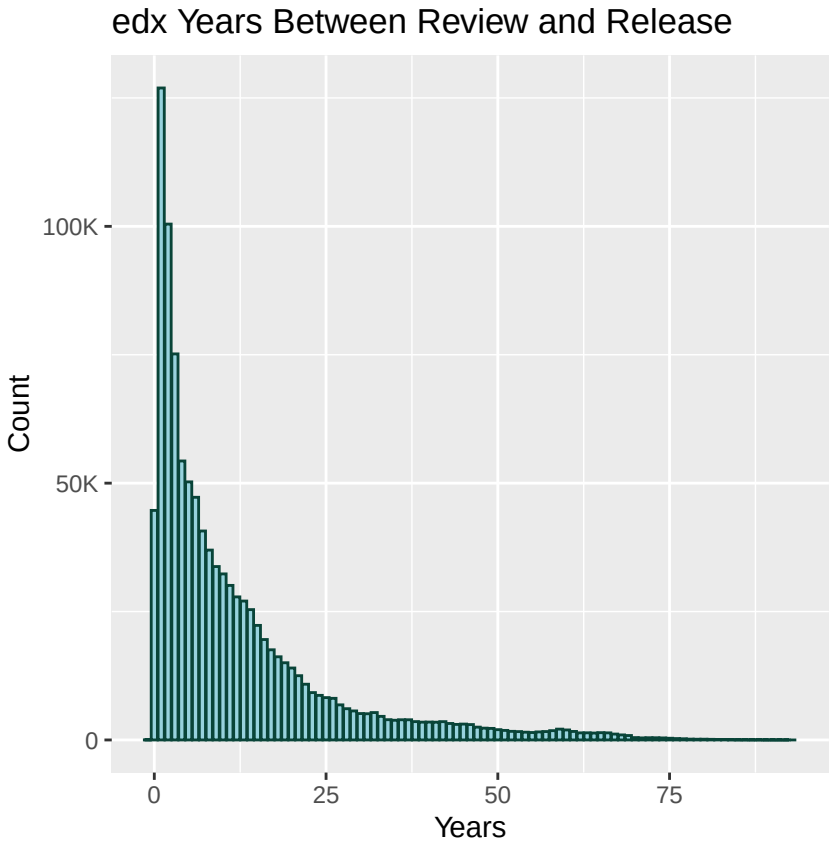


If we extract the release year from the title column we find that the earliest movie reviewed was released in 1915 and the most recently reviewed movie was released in 2008. This makes sense given the final review in the data set was received at 2009-01-05 17:51:47. During the time the dataset was collated there were 2 ways people were exposed to movies - at theaters and at home on video cassette. Theaters predominantly showed new releases with occasional film festivals and late night showings of classic films such as Clockwork Orange and The Rocky Horror Picture Show. Video Cassettes were rented from local video stores and were a mix of

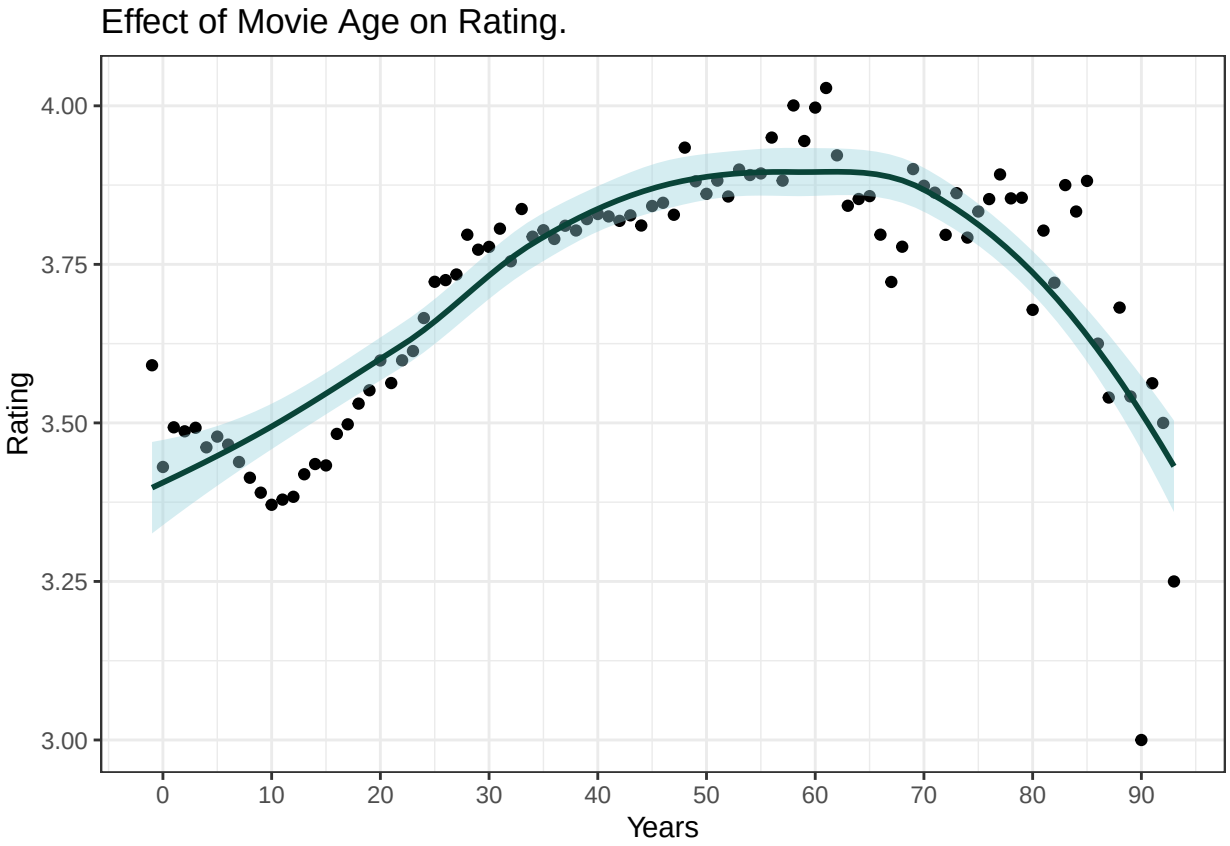
recent releases and classic films.



If we look at the relationship between rating and the length of time between the release of the movie and the review we get the following.



If we combine the age of the movie with the mean rating we get the following graph which shows that older movies have higher average ratings. These movies would predominately be the older movies offered by video stores or exhibited during film festivals. Due to long tail effects these higher ratings are to be expected.



Next we can use the summary command to produce result summaries of the results of various model fitting functions.

Table 2: EDX Dataset Summary

userId	movieId	rating	timestamp	title	genres	date
Min. : 1	Min. : 1	Min. :0.500	Min. :8.229e+08	Length:1041390	Length:1041390	Min. :19
1st Qu.:18059	1st Qu.: 612	1st Qu.:3.000	1st Qu.:9.456e+08	Class :character	Class :character	1st Qu.:1
Median :35682	Median : 1777	Median :4.000	Median :1.033e+09	Mode :character	Mode :character	Median :
Mean :35849	Mean : 4145	Mean :3.515	Mean :1.031e+09	NA	NA	Mean :20
3rd Qu.:53630	3rd Qu.: 3617	3rd Qu.:4.000	3rd Qu.:1.126e+09	NA	NA	3rd Qu.:2
Max. :71567	Max. :65133	Max. :5.000	Max. :1.231e+09	NA	NA	Max. :20

As we can see from the summary, from a statistical perspective in the current form, the most useful column is the rating row. The time stamp row is in Unix epoch time (seconds from the 1st of January 1970) so that will need to be converted to a human readable format if that is found to be useful.

The following table shows the distinct number of User IDs, Movie IDs, Titles, and Genres. The last column is a check for any unset variables. This will return TRUE if present and FALSE if not.

Table 3: Summary of Movielens Data Set

Users	MoviesIds	Titles	Genres	MissingValues
69878	10677	10676	797	FALSE

The number of movies reviewed is higher than the number of reviewers. Also we can see that the number of genres is quite large due to the usage of different arrays of keywords to describe the movies. Also we can see that all there are no "Not Available" or missing values.

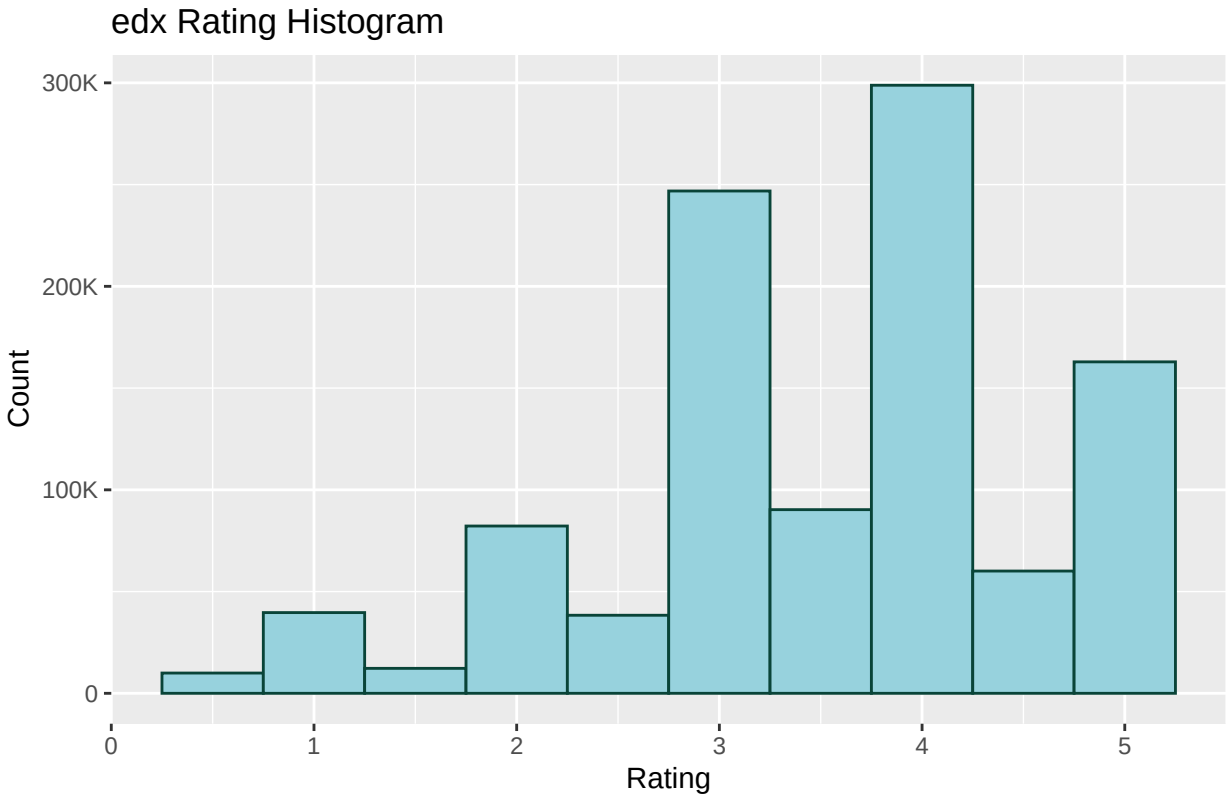
An initial look at the ratings using the summary function gave a mean of 3.515 and a median value of 4.

The following table shows the count of ratings. Whole numbers are much more commonly chosen when rating movies than decimal ratings.

Table 4: Rating Distribution

Var1	Freq
0.5	9961
1	39676
1.5	12271
2	82214
2.5	38355
3	246853
3.5	90259
4	298828
4.5	60085
5	162888

From this we can see that people are more likely to rate movies in whole numbers. If we plot this as a graph it is much more evident.



Whole number ratings

Now we will look at the data to see if rating with whole numbers compared to decimals has any impact.

First whole numbers - the subset of the edx dataset that has ratings 1, 2, 3, 4, 5:

Table 5: Whole Number Ratings

Users	MoviesIds	Titles	Genres
68928	10145	10144	784

Decimal Point Ratings

Then the decimal ratings - the subset of the edx dataset with ratings 0.5, 1.5, 2.5, 3.5 or 4.5:

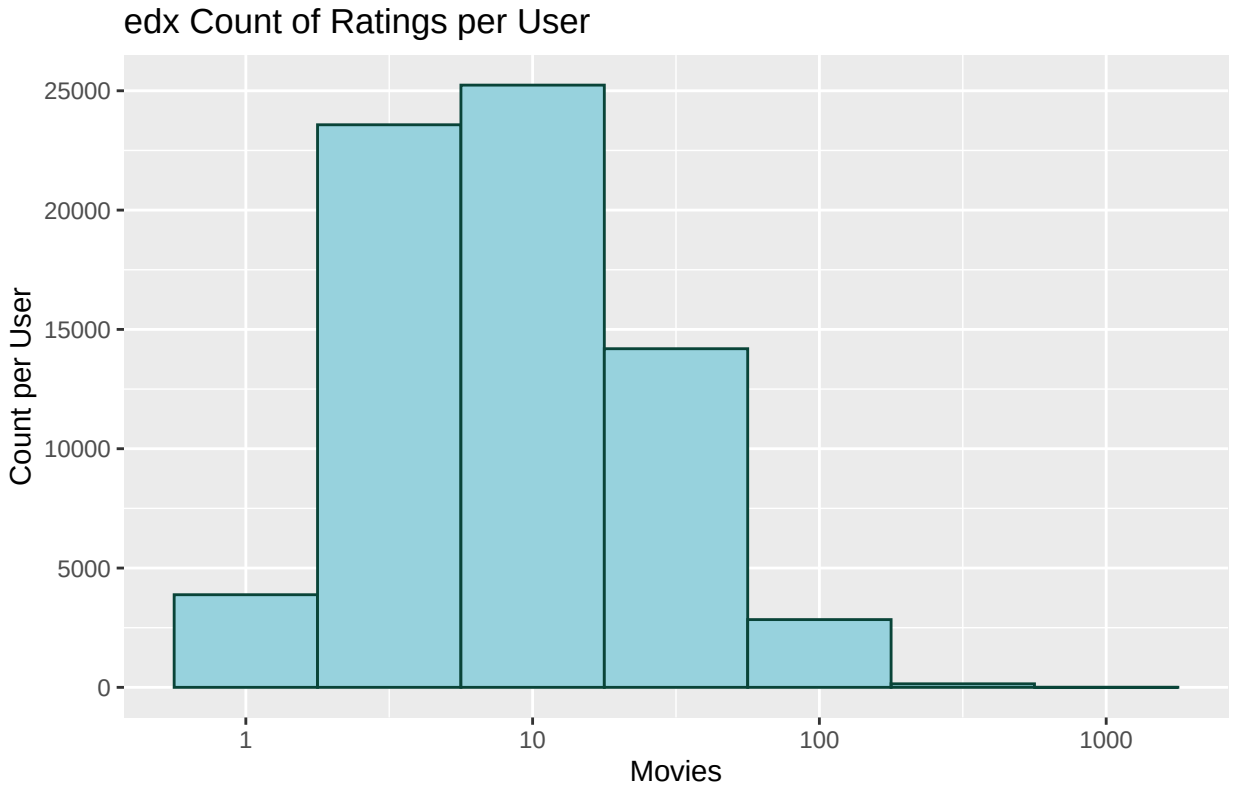
Table 6: Decimal Point Ratings

Users	MoviesIds	Titles	Genres
22674	9119	9118	770

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   2.500   3.500   3.345   4.500   4.500
```

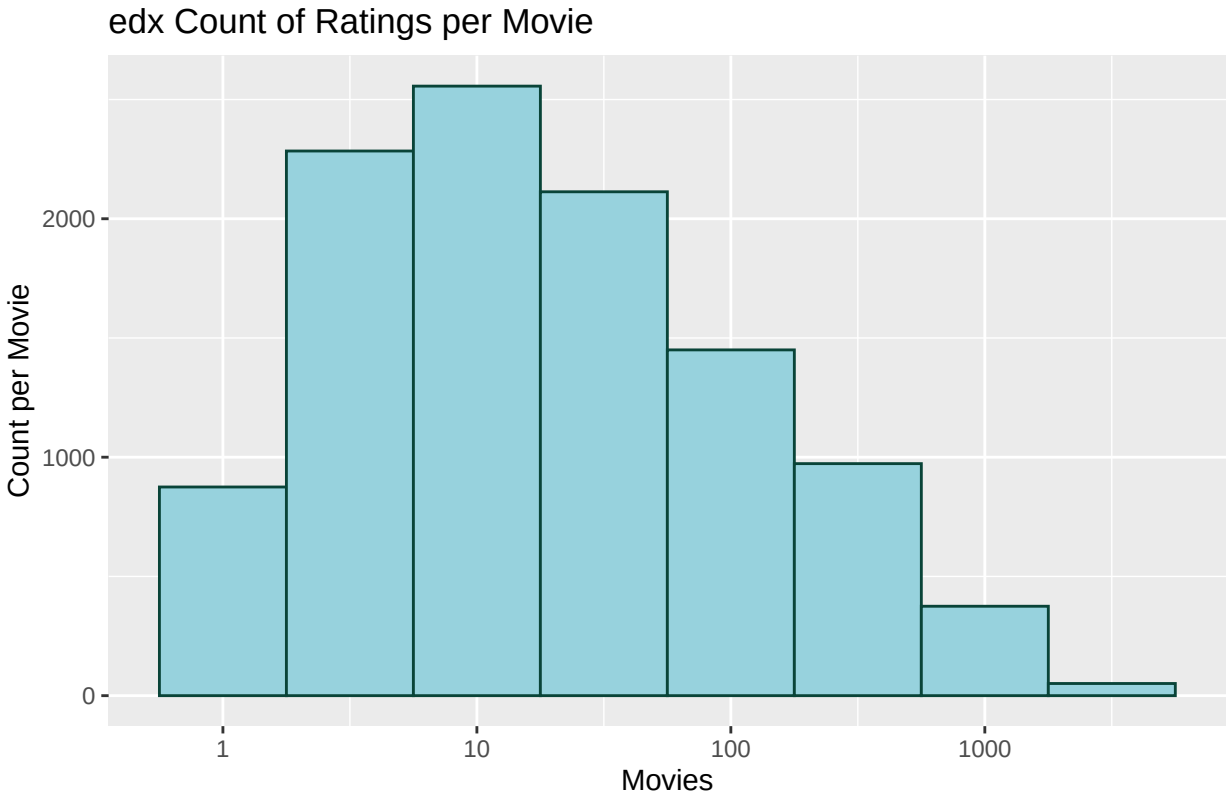
Ratings Per User

Now we turn to the count of ratings per user.

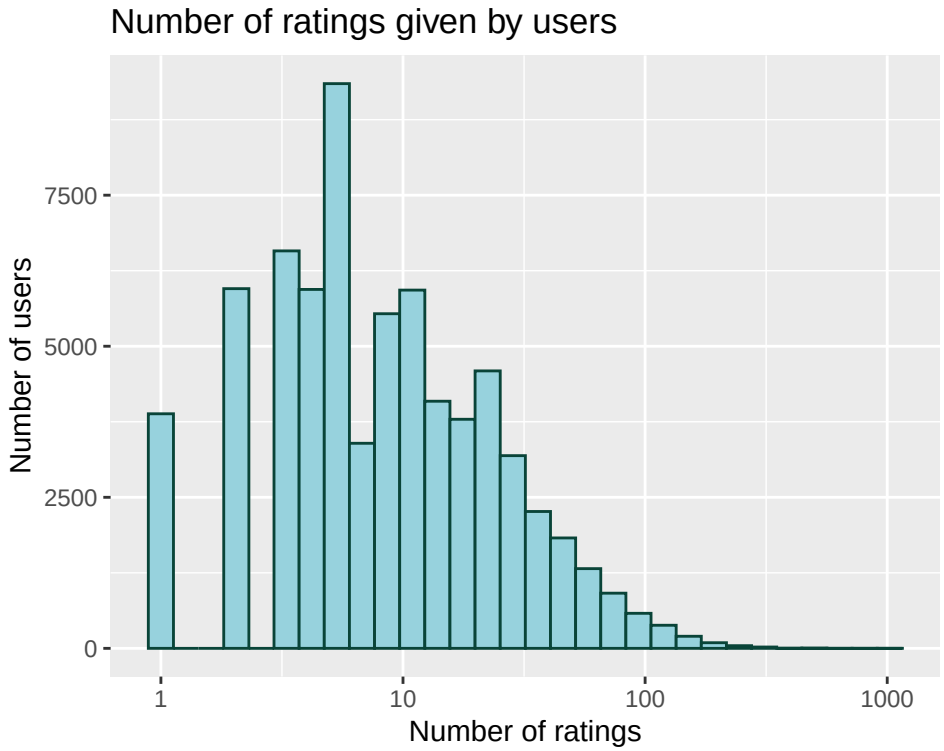


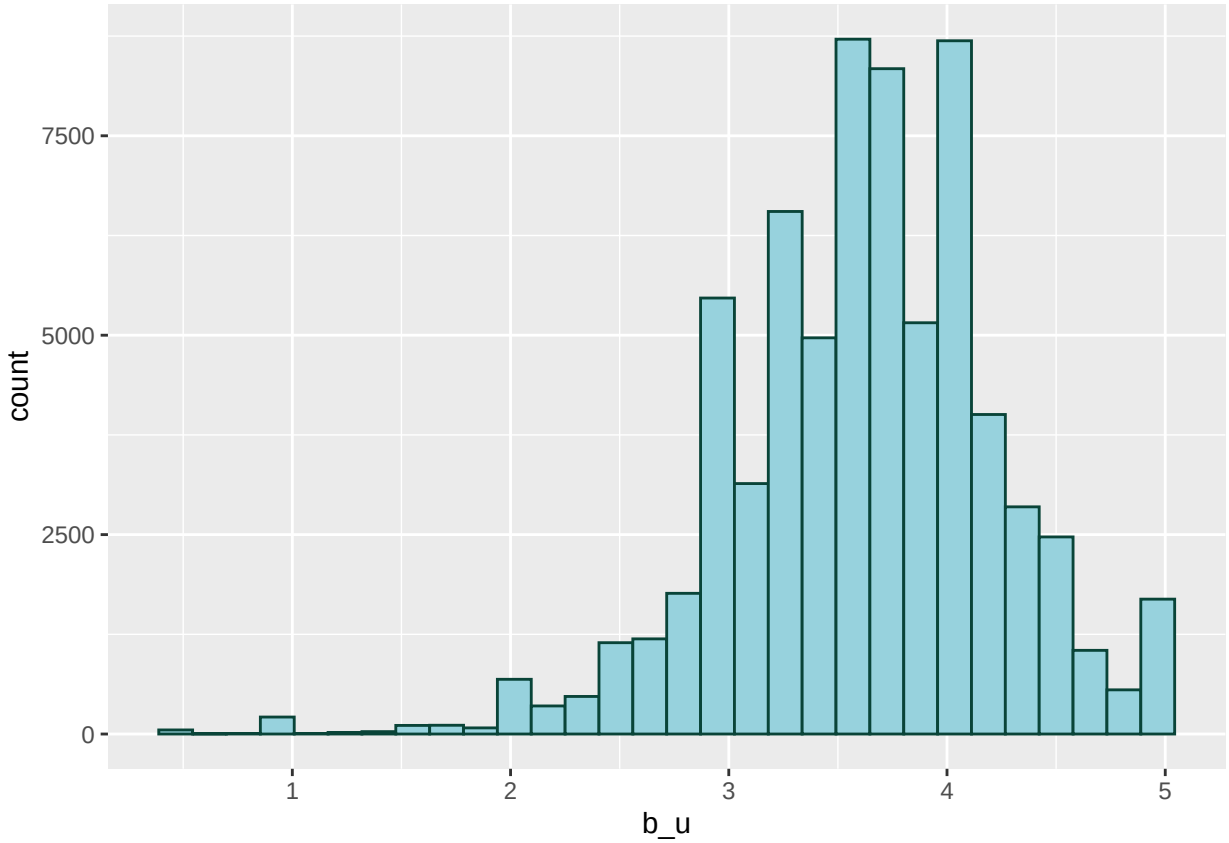
Ratings Per Movie

Again we can see that some movies are more popular than others and therefore have more reviews than less popular films.



If we look at the average number of films reviewed by each reviewer we get the following results.


Number of ratings given by users
Number of users
7500
5000
2500
0
1
10
100
1000
Number of ratings


count
7500
5000
2500
0
1
2
3
4
5
b_u