# MovieLens Recommendation

Wilsven Leong

July 10, 2021

## Introduction

In this report, the **MovieLens 10M dataset** was used to create a **movie recommendation system algorithm** that can be used to predict how a certain user will rate a certain movie.

The **MovieLens 10M dataset** consists of 10,000,000 ratings of 10,000 movies by 72,000 users on a five-star scale.

The data was pulled directly from the MovieLens website (https://grouplens.org/datasets/movielens/10m/).

The raw dataset was wrangled into a data frame, then split into the *edx* training dataset and the *validation* testing dataset.

The datasets were cleaned up, wrangled, and coerced into a more usable format.

The *edx* dataset was explored and analyzed by plotting the data through the lenses of different potential effects.

An equation for the root mean squared error (RMSE) was defined as the target parameter.

Several models were trained using the *edx* dataset and evaluated on the *validation* dataset, including naive mean, effects, and regularization. The most effective models were then combined.

Using this method, a **movie recommendation system algorithm** with an **RMSE** of **0.863** was developed.

## Data Analysis and Model Development

### Create the Datasets

The raw datasets were pulled directly from the MovieLens website and saved to a temporary file. From the temporary file, the data was pulled in and coerced into two data frames, the *ratings* data frame, with columns userId, movieId, rating, and timestamp, and the *movies* data frame, with columns movieId, title, and genres. The two data frames were joined together by movieId, creating a new *movielens* data frame with six columns, userId, movieId, rating, timestamp, title, and genres.

***movielens* Dataset** Let's display first six rows of the *movielens* dataset.

```
##    userId movieId rating timestamp                         title
## 1:      1     122      5 838985046              Boomerang (1992)
## 2:      1     185      5 838983525               Net, The (1995)
## 3:      1     231      5 838983392          Dumb & Dumber (1994)
```

```
## 4:      1     292      5 838983421               Outbreak (1995)
## 5:      1     316      5 838983392               Stargate (1994)
## 6:      1     329      5 838983392 Star Trek: Generations (1994)
##                            genres
## 1:                Comedy|Romance
## 2:         Action|Crime|Thriller
## 3:                         Comedy
## 4:  Action|Drama|Sci-Fi|Thriller
## 5:       Action|Adventure|Sci-Fi
## 6: Action|Adventure|Drama|Sci-Fi
```

The *movielens* dataset was then split into two datasets, the *edx* training dataset consisting of 90% of the data and the *temp* dataset consisting of the remaining 10% of the data. Movies that only appear in the *temp* dataset were removed, creating the *validation* testing dataset. Those removed movies were then added to the *edx* dataset.

***edx* Dataset** Let's display first six rows of the *edx* dataset.

```
##    userId movieId rating timestamp                          title
## 1:      1     122      5 838985046              Boomerang (1992)
## 2:      1     185      5 838983525               Net, The (1995)
## 3:      1     292      5 838983421               Outbreak (1995)
## 4:      1     316      5 838983392               Stargate (1994)
## 5:      1     329      5 838983392 Star Trek: Generations (1994)
## 6:      1     355      5 838984474       Flintstones, The (1994)
##                            genres
## 1:                Comedy|Romance
## 2:         Action|Crime|Thriller
## 3:  Action|Drama|Sci-Fi|Thriller
## 4:       Action|Adventure|Sci-Fi
## 5: Action|Adventure|Drama|Sci-Fi
## 6:       Children|Comedy|Fantasy
```

Let's display summary statistics of the *edx* dataset.

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08
##  1st Qu.:18124   1st Qu.:  648   1st Qu.:3.000   1st Qu.:9.468e+08
##  Median :35738   Median : 1834   Median :4.000   Median :1.035e+09
##  Mean   :35870   Mean   : 4122   Mean   :3.512   Mean   :1.033e+09
##  3rd Qu.:53607   3rd Qu.: 3626   3rd Qu.:4.000   3rd Qu.:1.127e+09
##  Max.   :71567   Max.   :65133   Max.   :5.000   Max.   :1.231e+09
##     title              genres
##  Length:9000055     Length:9000055
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

## Clean the Datasets

Looking at the *edx* dataset again, there is some data cleaning that can be done to make the data easier to visualize and analyze.

The timestamp column is the time the review was submitted, formatted as the number of seconds since January 1, 1970. It can be converted to a date_time data type.

The movie release year is included in title column. It can be extracted, added as the new column year, and converted to a numeric data type.

The columns timestamp and year can be used to calculate the number of years between the movie's release year and the year the movie was reviewed and create a new column yearsbetween.

Some movies fall into more than one genre in the genres column. Reviews of movies with more than one genre can be separated out by genre into multiple duplicate reviews with one genre per review.

**Cleaned *edx* Dataset** Let's take a look at the cleaned *edx* dataset. Display first six rows of the cleaned *edx* dataset.

```
## # A tibble: 6 x 8
##   userId movieId rating timestamp           title      genres  year yearsbetween
##    <int>   <dbl>  <dbl> <dttm>              <chr>      <chr>  <dbl>        <dbl>
## 1      1     122      5 1996-08-02 11:24:06 Boomerang~ Comedy  1992            4
## 2      1     122      5 1996-08-02 11:24:06 Boomerang~ Roman~  1992            4
## 3      1     185      5 1996-08-02 10:58:45 Net, The ~ Action  1995            1
## 4      1     185      5 1996-08-02 10:58:45 Net, The ~ Crime   1995            1
## 5      1     185      5 1996-08-02 10:58:45 Net, The ~ Thril~  1995            1
## 6      1     292      5 1996-08-02 10:57:01 Outbreak ~ Action  1995            1
```

Display summary statistics of the cleaned *edx* dataset.

```
##      userId         movieId          rating        timestamp
##  Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08
##  1st Qu.:18124   1st Qu.:  648   1st Qu.:3.000   1st Qu.:9.468e+08
##  Median :35738   Median : 1834   Median :4.000   Median :1.035e+09
##  Mean   :35870   Mean   : 4122   Mean   :3.512   Mean   :1.033e+09
##  3rd Qu.:53607   3rd Qu.: 3626   3rd Qu.:4.000   3rd Qu.:1.127e+09
##  Max.   :71567   Max.   :65133   Max.   :5.000   Max.   :1.231e+09
##     title              genres
##  Length:9000055     Length:9000055
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

The same steps were carried out on the *validation* dataset.

## Cursory Data Visualizations and Analysis

All visualizations and analyses were performed with the *edx* training dataset.

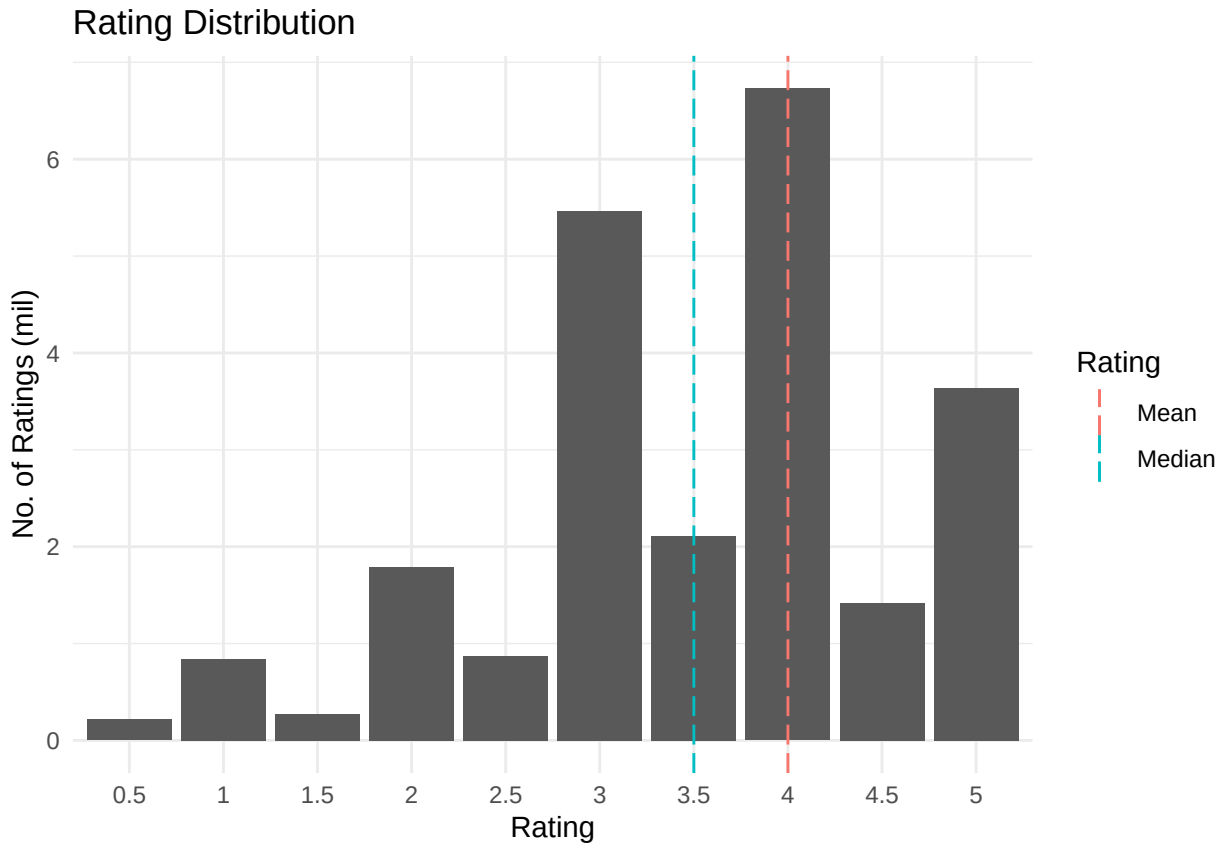
There are 69,878 unique users and 10,677 unique movies in the *edx* training dataset.

```
## # A tibble: 1 x 2
##   uniqueUsers uniqueMovies
##         <int>        <int>
## 1       69878        10677
```

The average rating is 3.5 stars (3.53 stars to be exact). The 4.0 stars is the median rating.

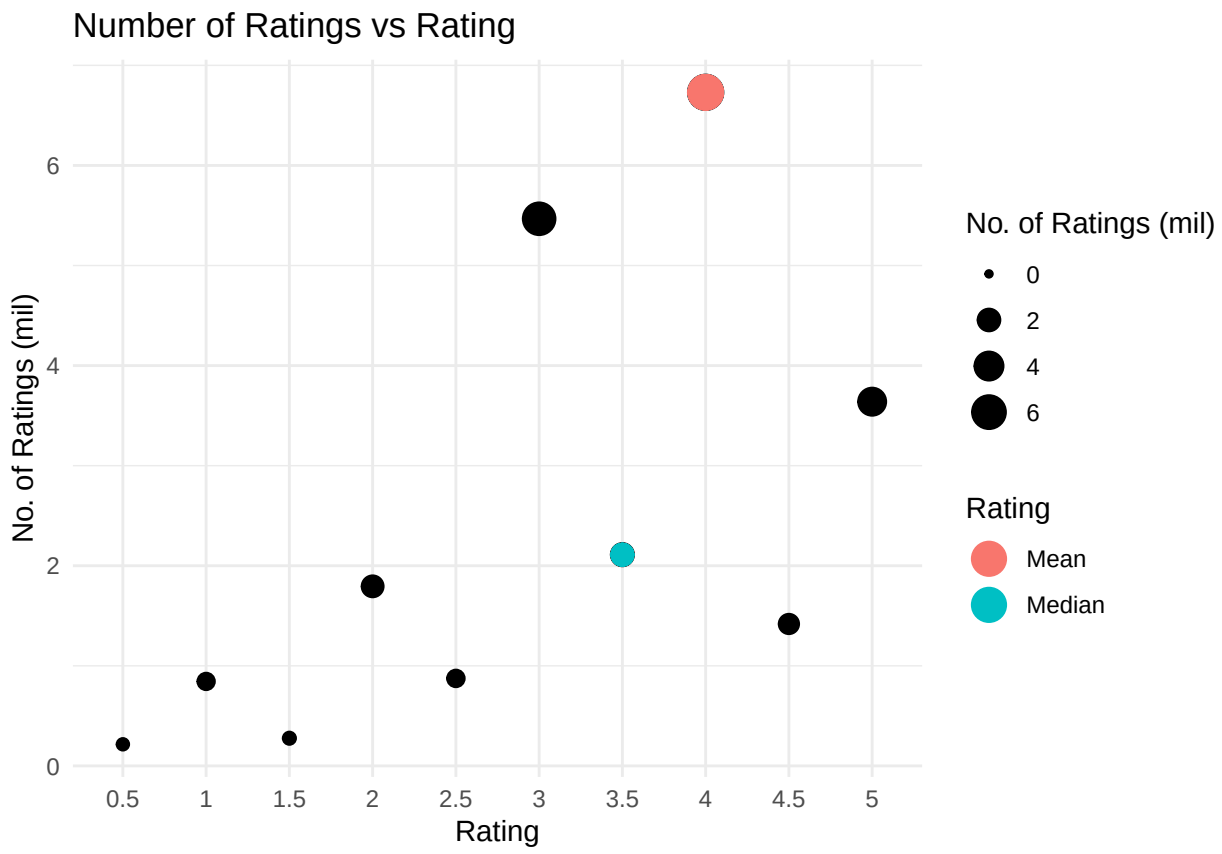
**Ratings** Grouping the data by rating shows that four stars is the most common rating and that full (i.e. 5.0, 4.0, 3.0 etc.) star ratings are given more often than half star ratings (i.e. 4.5, 3.5, 2.5 etc.).

```
## # A tibble: 6 x 2
##   rating numRatings
##   <fct>       <int>
## 1 4         6730401
## 2 3         5467061
## 3 5         3639511
## 4 3.5       2110690
## 5 2         1794243
## 6 4.5       1418248
```




The above rating distribution shows that the users have a general tendency to rate movies between 3 and 4. This is a very general conclusion. We should also further explore the effect of different features to make a good predictive model.

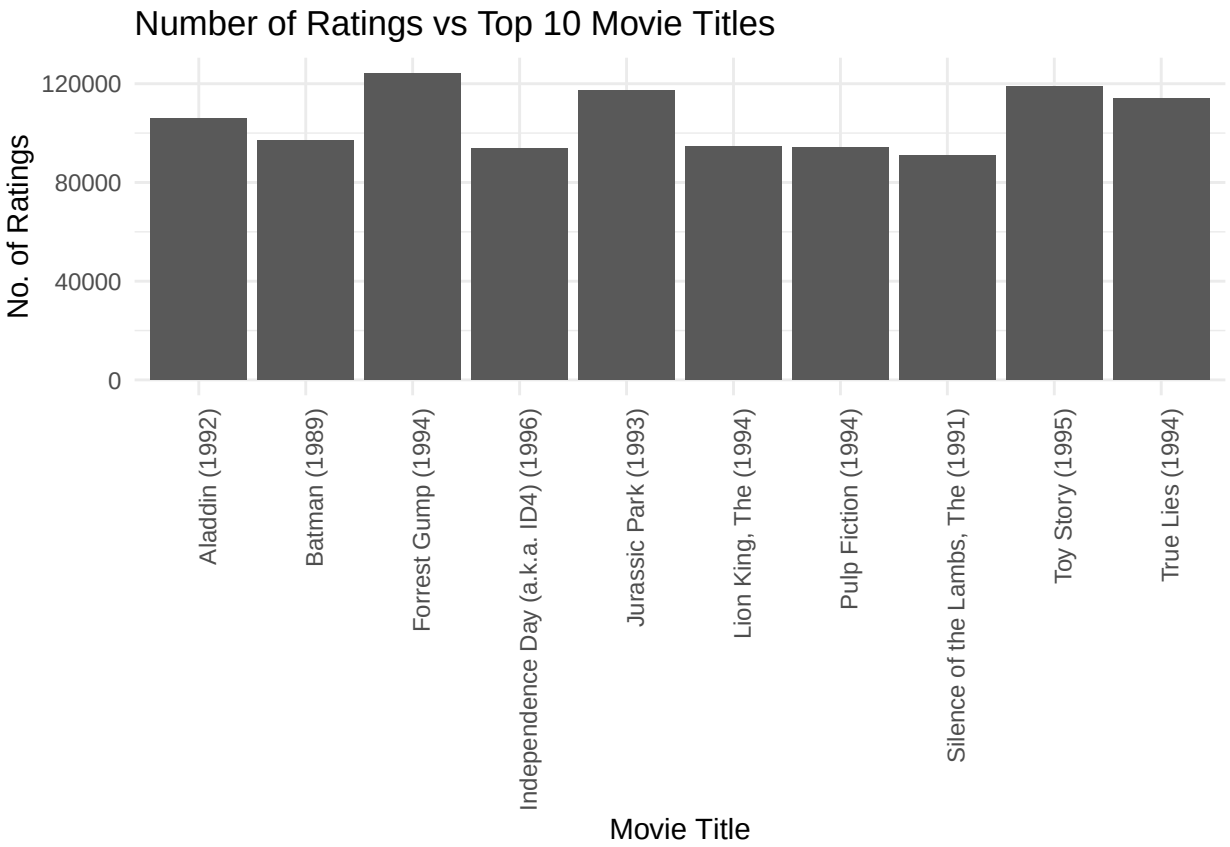
Here's another plot for the frequency of various ratings to help further visualize the most common star ratings.



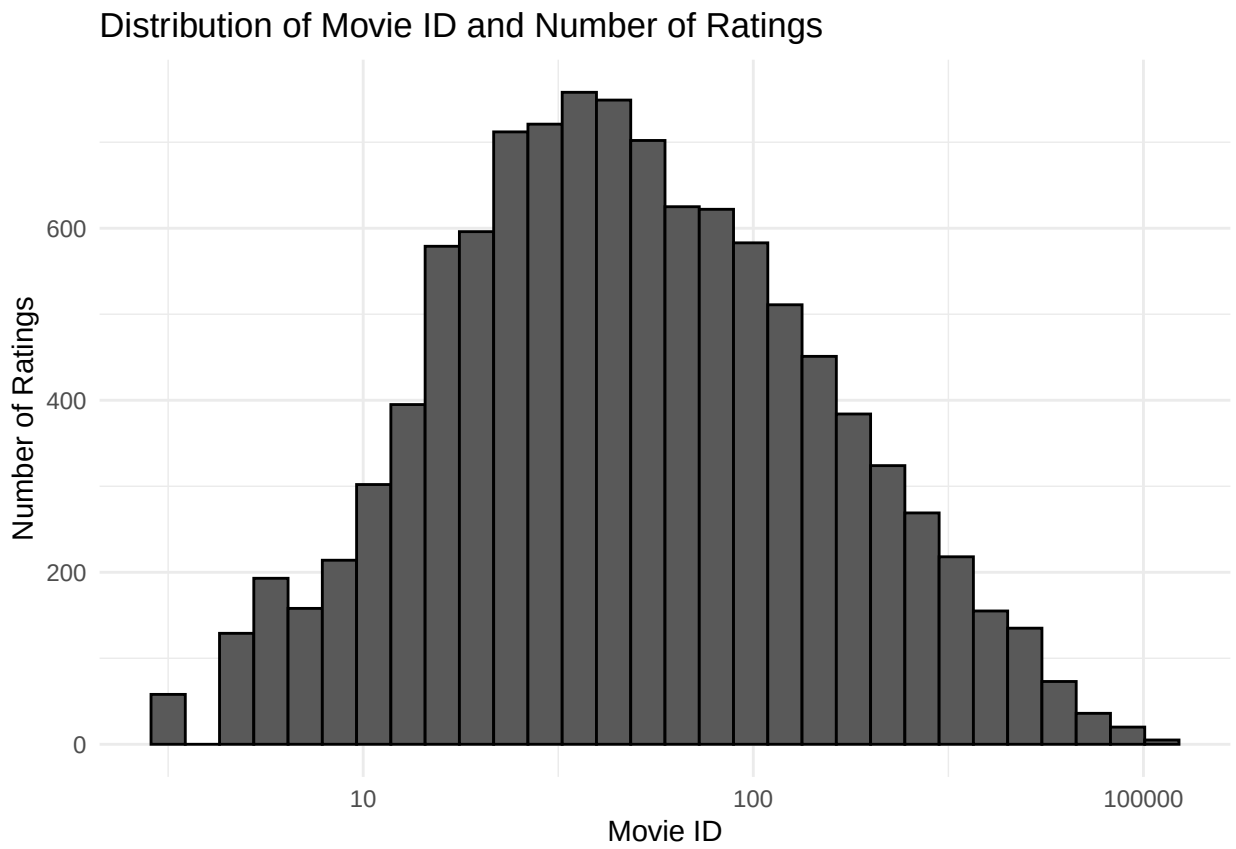

**Movies** Grouping the data by movie shows that in general, movies that are reviewed often have higher average ratings and that there is more variation in average ratings for movies that have few reviews.

```
##   movieId                     title numRatings avgRating
## 1     356       Forrest Gump (1994)     124316      4.01
## 2       1          Toy Story (1995)     118950      3.93
## 3     480      Jurassic Park (1993)     117440      3.66
## 4     380          True Lies (1994)     114115       3.5
## 5     ...                      <NA>        ...       ...
## 6   64611 Forgotten One, The (1990)          1       3.5
## 7   64897             Mr. Wu (1927)          1         3
## 8   64944 Face of a Fugitive (1959)          1         3
## 9   64976              Hexed (1993)          1       1.5
```

Let's visualize the top 10 movies with the most number of ratings.



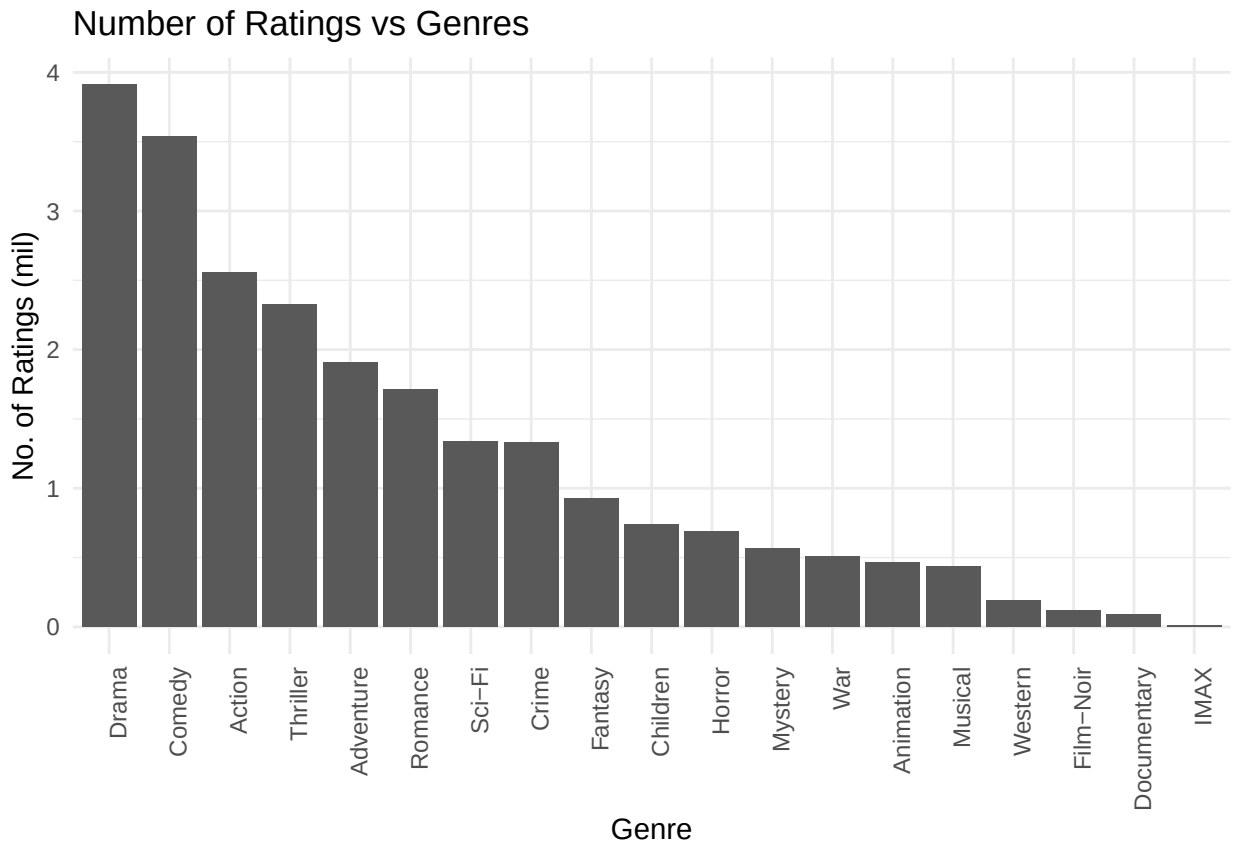

Some movies are rated more often than others. This is because some movies are blockbusters and are highly anticipated movies while other movies are less well known. Below is their distribution. This explores movie biases.

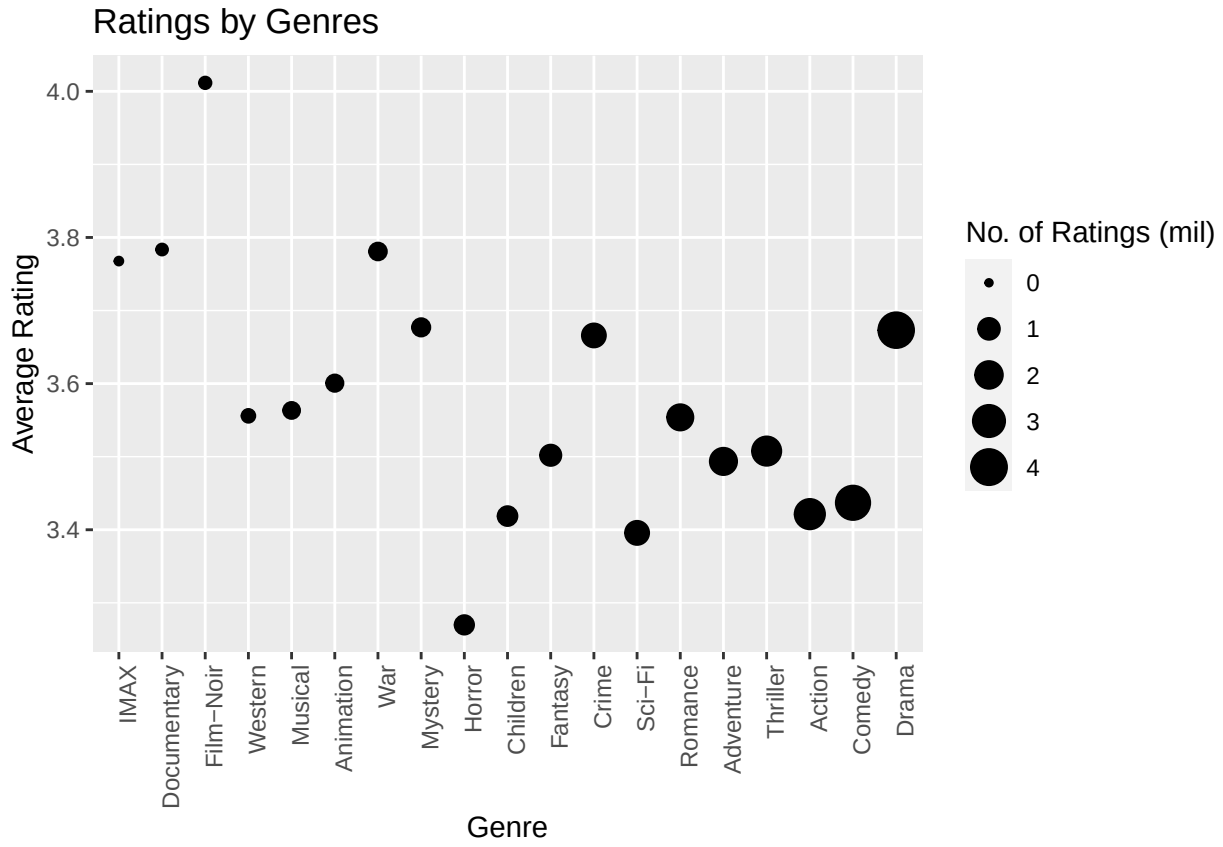



The histogram shows some movies have been rated very few number of times. So they should be given lower importance in movie prediction.

**Genres** Let's also visualize the genres and respective number of ratings to see which genres are the more popular ones. Do note that most movies have multiple genres.

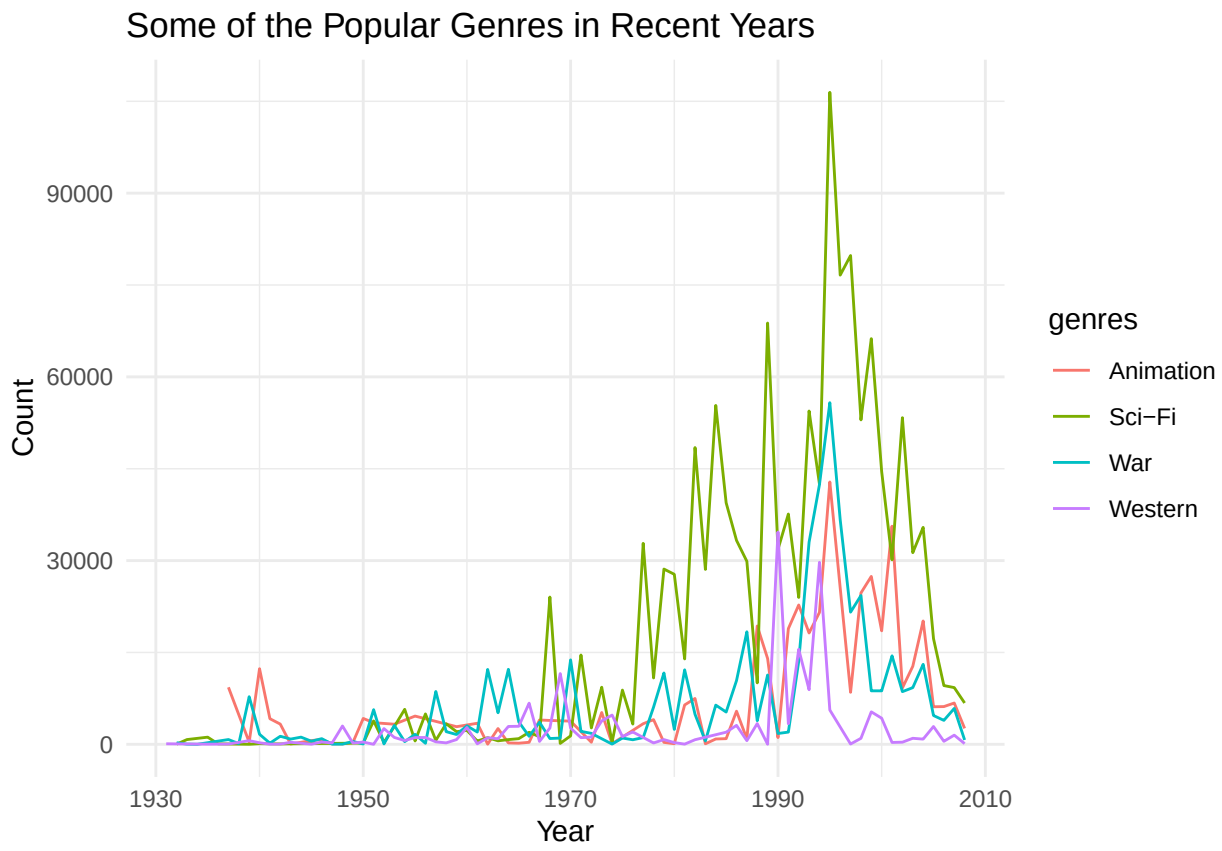
```
##               genres numRatings avgRating
## 1              Drama    3910127      3.67
## 2             Comedy    3540930      3.44
## 3             Action    2560545      3.42
## 4           Thriller    2325899      3.51
## 5               <NA>        ...       ...
## 6          Film-Noir     118541      4.01
## 7        Documentary      93066      3.78
## 8               IMAX       8181      3.77
## 9 (no genres listed)          7      3.64
```


Number of Ratings vs Genres
No. of Ratings (mil)
4
3
2
1
0
Drama
Comedy
Action
Thriller
Adventure
Romance
Sci−Fi
Crime
Fantasy
Children
Horror
Mystery
War
Animation
Musical
Western
Film−Noir
Documentary
IMAX
Genre

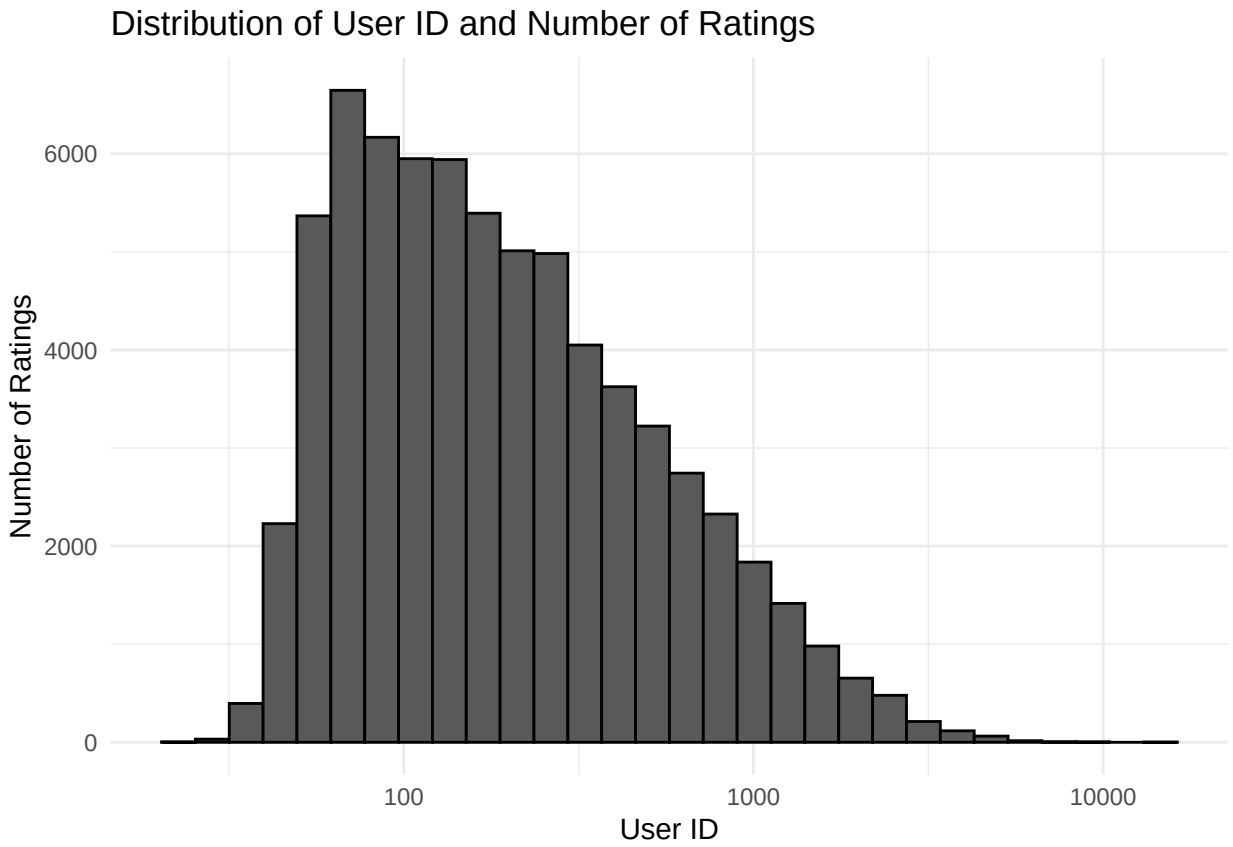



Grouping the data by genre shows that the most common genres are Drama, Comedy, and Action. Best rated genres like Film-Noir, War, and Documentary have fewer movies and ratings.

However, genre popularity changes every year. Here we tackle the issue of temporal evolution of users taste over different popular genre over the years.




This plots depicts how some genres are more popular over others during different periods of time.

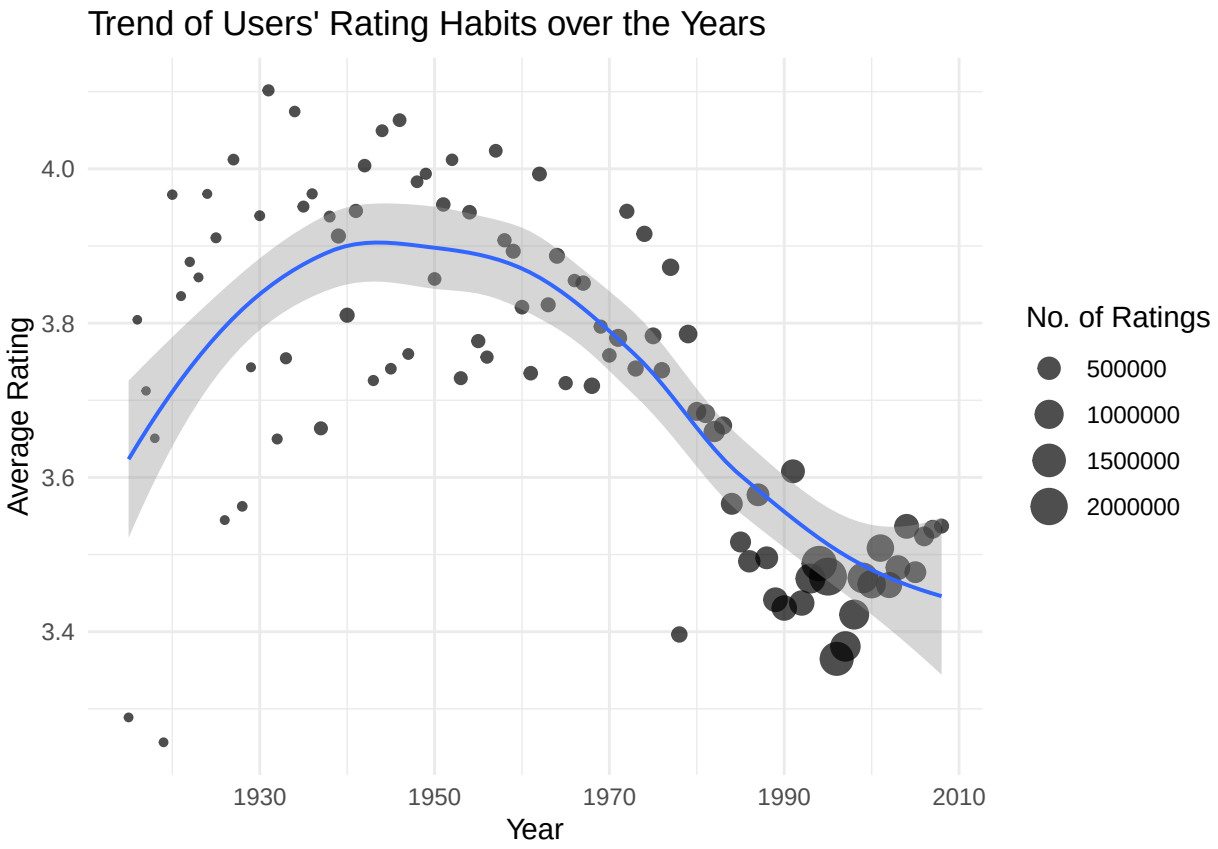



**Users**

The plot above shows that not every user is equally active. Some users have rated very few movies and their opinion may contribute user bias to the prediction results.

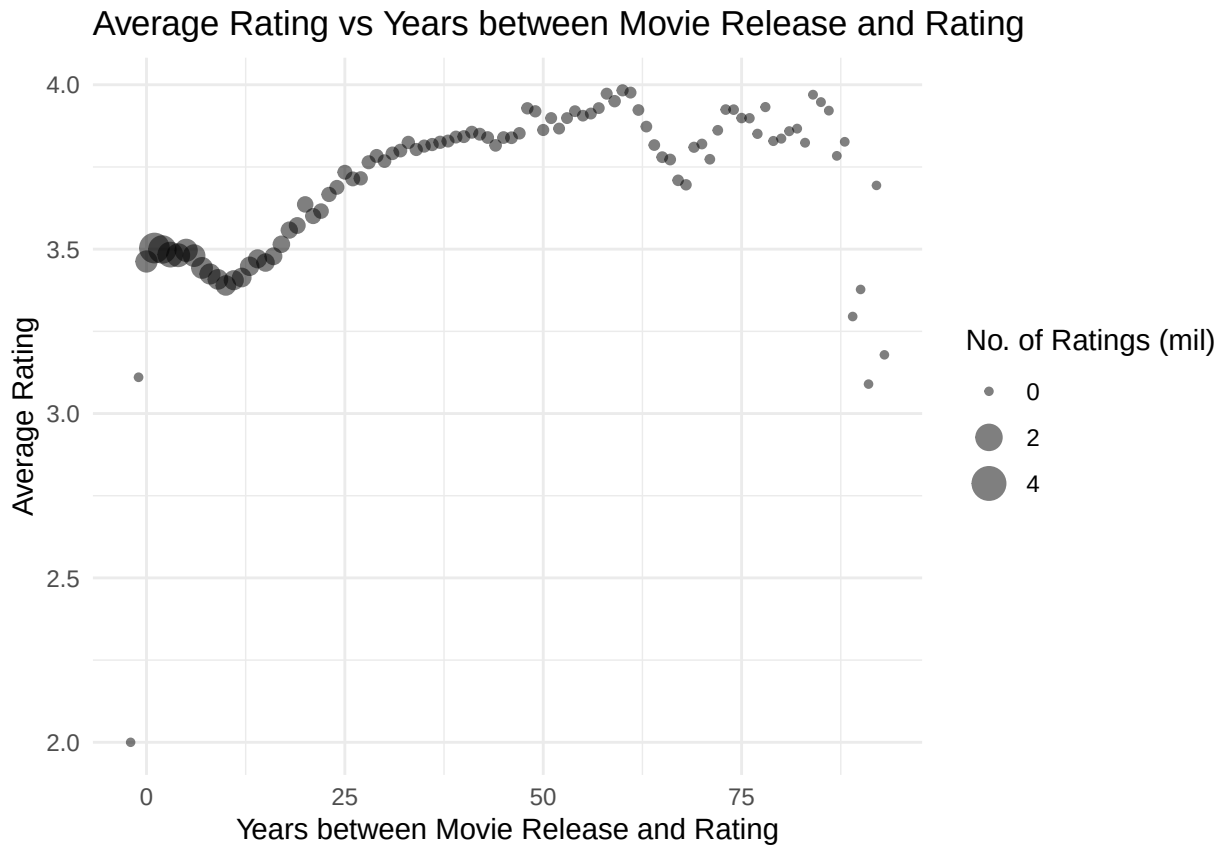
**Release Year** Grouping the data by movie release year shows that movies are better rated in pre-1980 years than post-1980 years and that movies released in recent years have received more ratings. In other words, the general trend shows modern users rate movies on relatively lower rating.

```
##   year numRatings avgRating
## 1 1915        360      3.29
## 2 1916         92       3.8
## 3 1917         33      3.71
## 4 1918         73      3.65
## 5  ...        ...       ...
## 6 2005     393004      3.48
## 7 2006     277026      3.52
## 8 2007     226367      3.53
## 9 2008      79971      3.54
```




**Years between Release and Review** Grouping the data by the number of years between release and review shows that movies are generally rated higher when there is more time between a movie's release and the time it was reviewed.

```
##   yearsbetween numRatings avgRating
## 1           -2          3         2
## 2           -1        290      3.11
## 3            0    1010761      3.46
## 4            1    2784854       3.5
## 5          ...        ...       ...
## 6           90        102      3.38
## 7           91         67      3.09
## 8           92         49      3.69
## 9           93         28      3.18
```




## Defining RMSE

The goal of this project is to develop an algorithm with the lowest possible residual mean squared error (RMSE). RMSE is defined as the error that the algorithm makes when predicting a rating, or:

$$\sqrt{\frac{1}{N}\sum_{e}(\hat{y}_e - y_e)^2}$$

where $N$ is the total number of user or movie ratings, $\hat{y}_e$ is the predicted rating for a particular review given effects $e$, and $y_e$ is the actual rating for a particular review given effects $e$.

An RMSE of 1 would mean that on average, the rating that the algorithm predicted is one star off the actual rating.

## Modeling Approach

### A Simple Model - Average

The simplest model predicts the same rating for each review, regardless of effects like movie, user, genre, etc. This model can be defined as:

$$Y = \mu + \epsilon$$

where $Y$ is the outcome (predicted rating), $\mu$ is the average rating, and $\epsilon$ is the error.

The **RMSE** of the **Average** model is **1.053**.

### Introducing Effects

Introducing effects allows the model to take variability into account. Looking at the visualizations above, for example, some movies are, on average, rated higher than others and certain genres tend to receive lower average ratings than others. The effects model can be defined as:

$$Y = \mu + e_a + \epsilon$$

where $e_a$ is the effect term of effect $a$.

For modeling purposes, the least square estimate of $e_a$ is the average of $Y_a - \mu$ for each instance of effect $a$.

Based on the above visualizations, movie, user, genre, year released, and years between release and review effects were all introduced to the model.

### Movie Effect

The **Average + Movie Effect** model is defined as

$$Y = \mu + e_m + \epsilon$$

where $e_m$ is the effect term for movie $m$.

The **RMSE** of the **Average + Movie Effect** model is **0.941**.

### User Effect

The **Average + User Effect** model is defined as

$$Y = \mu + e_u + \epsilon$$

where $e_u$ is the effect term for user $u$.

The **RMSE** of the **Average + User Effect** model is **0.973**.

### Genre Effect

The **Average + Genre Effect** model is defined as

$$Y = \mu + e_g + \epsilon$$

where $e_g$ is the effect term for genre $g$.

The **RMSE** of the **Average + Genre Effect** model is **1.046**.

### Year Effect

The **Average + Year Effect** model is defined as

$$Y = \mu + e_y + \epsilon$$

where $e_y$ is the effect term for release year $y$.

The **RMSE** of the **Average + Year Effect** model is **1.042**.

### Years between Effect

The **Average + Years between Effect** model is defined as

$$Y = \mu + e_y b + \epsilon$$

where $e_y b$ is the effect term for years between the movie's release and review $yb$.

The **RMSE** of the **Average + Years between Effect** model is **1.045**.

### Introducing Regularization

Looking at the visualizations above again, there is a lot of variation in the number of ratings that different movies receive, different users give, etc. Regularization will introduce a penalized term that will have a great effect on large predicted ratings stemming from small group sizes while having little effect on predicted ratings stemming from large group sizes.

$$e_a = \frac{\sum_1^{n_a}(Y_a - \mu)}{n_a + \lambda_a}$$

where $n_a$ is the number of ratings for effect $a$, $Y_a$ is the average rating for effect $a$, and $\lambda_a$ is the penalization term for effect $a$.

### Movie Regularization

The **Average + Movie Effect + Regularization** model is defined as

$$Y = \mu + e_m + \epsilon$$

where

$$e_m = \frac{\sum_1^{n_m}(Y_m - \mu)}{n_m + \lambda_m}$$

The **RMSE** of the **Average + Movie Effect + Regularization** model is **0.941**, which is no improvement over the non-regularized model.

## Results - The Best Model

Looking that the models described above, only two of them, **Movie Effect** and **User Effect** made significant improvements to the **Average** model.

```
## # A tibble: 7 x 2
##   model                                rmse
##   <chr>                               <dbl>
## 1 Average                             1.05
## 2 Movie Effect                        0.941
## 3 User Effect                         0.973
## 4 Average + Genre Effect              1.05
## 5 Average + Year Effect               1.04
## 6 Average + Years between Effect      1.04
## 7 Average + Movie Effect + Regularization 0.941
```

By combining these two effects, the model should become more accurate.

The **Average + Movie + User Effects** model is defined as

$$Y = \mu + e_m + e_u + \epsilon$$

**Best Effects Model**

```
## # A tibble: 1 x 2
##   model                                rmse
##   <chr>                               <dbl>
## 1 Average + Movie Effect + User Effect 0.863
```

The **RMSE** of the **Average + Movie + User Effect** model is **0.863**.

# Conclusions

After visually analyzing and examining the data and testing several models, an algorithm to predict movie ratings with an **RMSE** of **0.863** was developed by defining a model that included effects.

$$Y = \mu + e_m + e_u + \epsilon$$