# MovieLens Project

*WONG MEI YING (Catherine_831)*

*Mar 24, 2019*

## Abstract

In this report, a movie recommendation system will be built based on selected training sets of the MovieLens data set. 10M version of the MovieLens data set are extracted from GroupLens which is a research lab in the Department of Computer Science and Engineering at the University of Minnesota. Four algorithms will be applied: *Simple Model without any affects*, *Movie Effects*, *Movie & User Effects* and *Regularization with Movie & User Effects*. The result will be compared and analysed by the performance of Residual Mean Squared Error (RMSE).

## 1. Introduction

Watching movie is a kind of entertainment. There are variety choice of movies. Different persons have their own preference on movie. Recommendation system is a system that predict a rating of preference a user would give to an item. In this project, a movie recommendation system with 5-star scale is to be built to predict how a person to rate a movie. One star represents a bad movie, whereas five stars represents an excellent movie.

10M version of the MovieLens dataset from GroupLens websit (https://grouplens.org/datasets/movielens/10m/ (https://grouplens.org/datasets/movielens/10m/)) will be used. 90% of MovieLens dataset is set as edx set and 10% of MoviLens dataset is set as Validation set. Four algorithms are developed in edx set and predict movie ratings in validation set.

In edx set, 80% of data is set as training data to build the movie recommendation system and the other 20% of data is to evaluate the model by measuring RMSE.

The goal of this project is to develop a machine learning algorithm using the inputs in edx set to predict movie ratings in the validation set. The lower the RMSE, the better the performance of the algorithm.

# 2. Method

## 2.1 Data Cleaning and Data Exploration

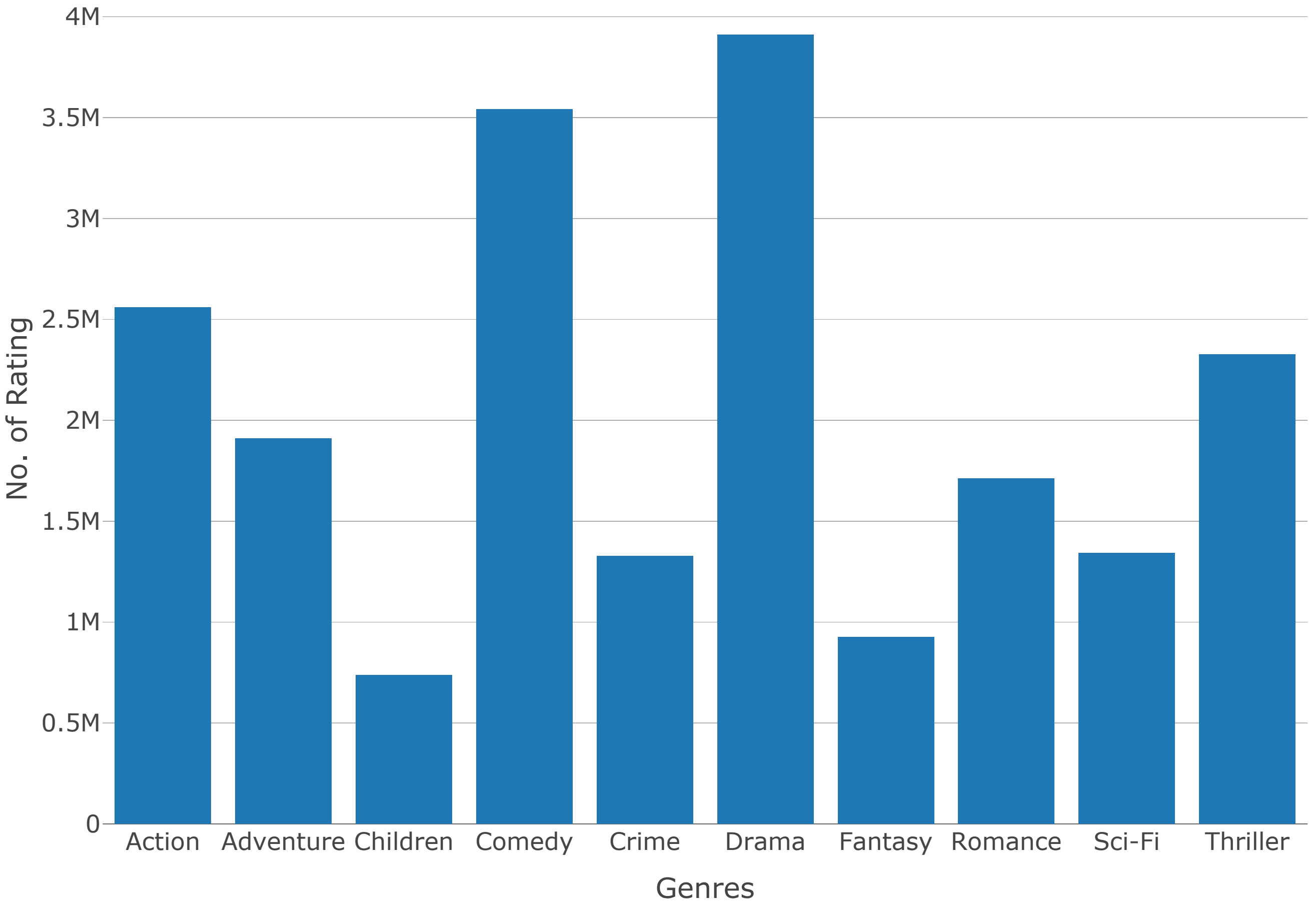
The extracted dataset records different rating of movies by different users. Each row represents a rating given by one user to one movie. The edx data set consists of:

- *9000055* total ratings (1 - 5) from *69878* users for *10677* movies
- movies are categorized into different genres

The structure of the data shown in below charts:

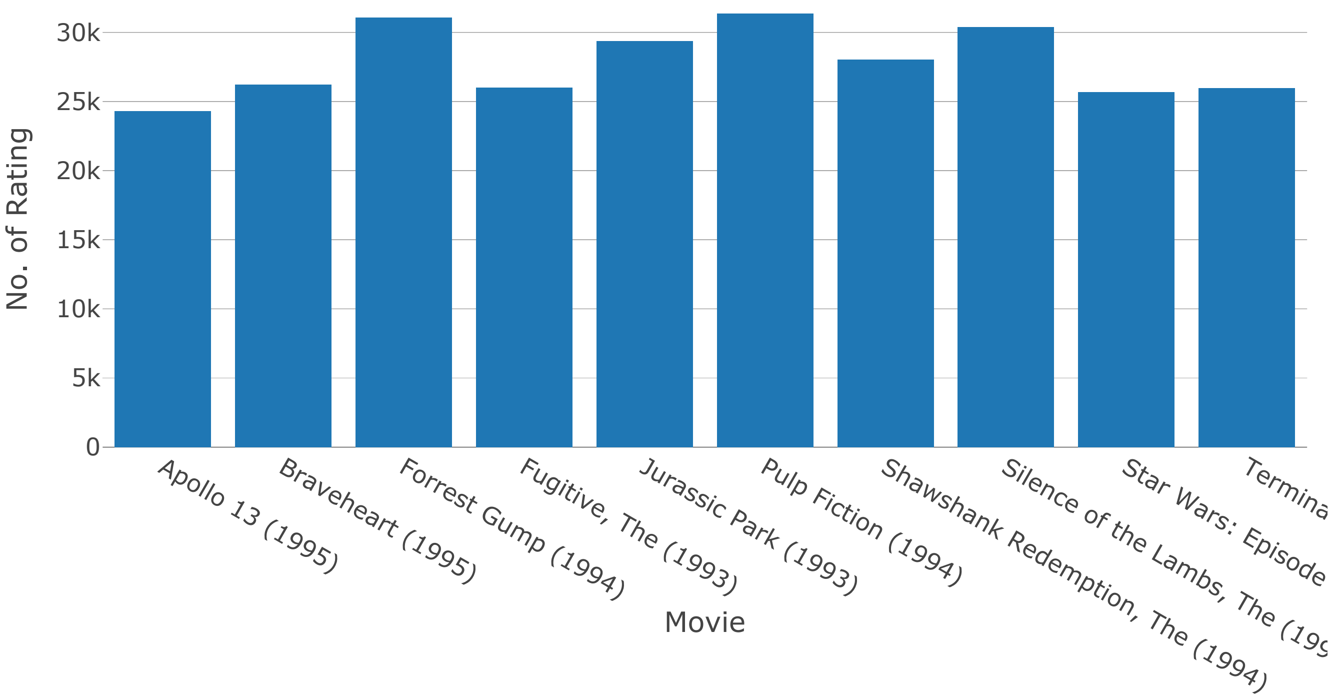
### 2.1.1 Top 10 Rating by Genres

- Below chat shows top 5 genres: **Drama**, **Comedy**, **Action**, **Thriller** and **Adventure.**



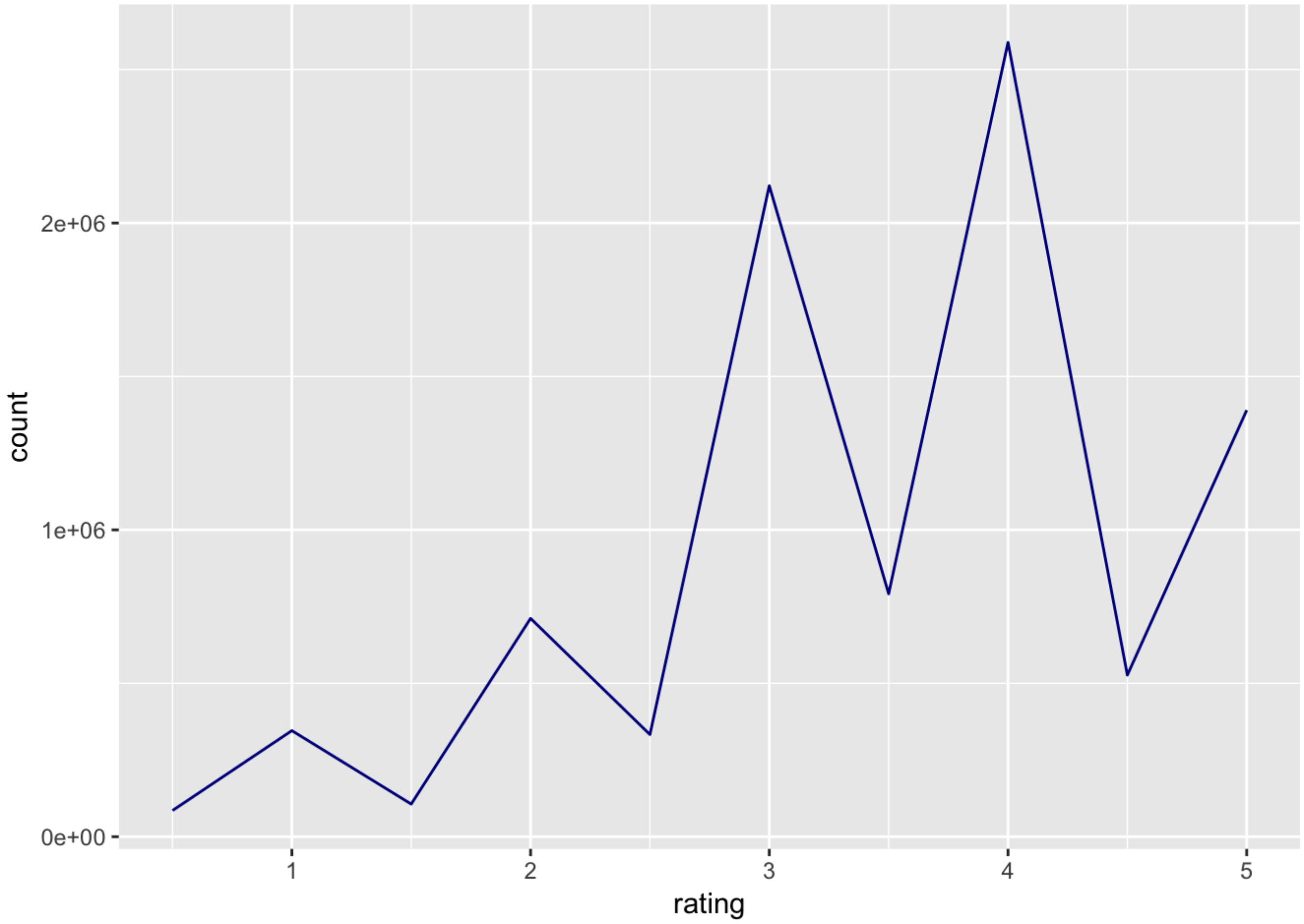
### 2.1.2 Top 10 Rating by Movie

- Below chart shows top five movies: **Pulp Fiction**, **Forrest Gump**, **Silene of the Lambs, The (1991)**, **Jurassic Park (1993)** and **Shawshank Redemption, The (1994)**.



### 2.1.3 List of Given Ratings in order from Most to Least

| rating | count |
|---:|---:|
| 4.0 | 2588430 |
| 3.0 | 2121240 |
| 5.0 | 1390114 |
| 3.5 | 791624 |
| 2.0 | 711422 |
| 4.5 | 526736 |
| 1.0 | 345679 |
| 2.5 | 333010 |
| 1.5 | 106426 |
| 0.5 | 85374 |



In general, half star ratings are less common than whole star ratings (e.g., there are fewer ratings of 3.5 than there are ratings of 3 or 4, etc.). The above table shows that there is no rating of zero.

## 2.2 RMSE Definitation

Evaluation of prediction is based on Residual Mean Squared Error (RMSE). RMSE is the typical error made when predicting a movie rating. The lower the RMSE, the better the performance of the predication. RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{u,i} \left( \hat{y}_{u,i} - y_{u,i} \right)^2}$$

where:

- $y_{u,i}$ is the rating of movie $i$ by user $u$,
- $\hat{y}_{u,i}$ is the prediction,
- $N$ is the number of user/movie combinations and the sum occurring over all these combinations

## 2.3 Models

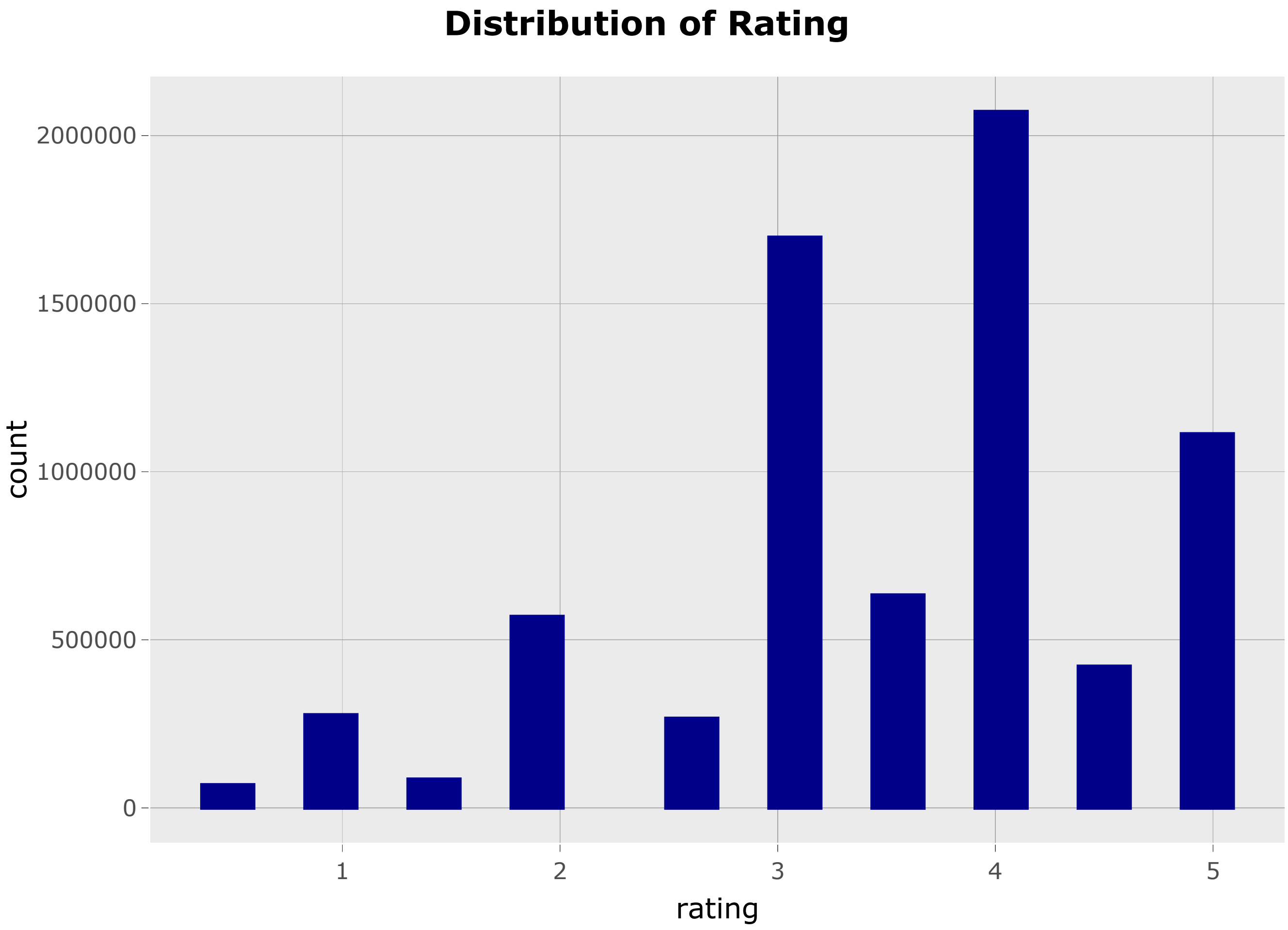
### 2.3.1 Model: Simplest model without any effects

In this model, it is assumed same rating for all users and the differences are the random variation. The equation is as follows:

$$Y_{u,i} = \mu + \varepsilon_{u,i}$$

where:

- $\varepsilon_{u,i}$ is independent errors sampled from the same distribution centered at 0,
- $\mu$ is the “true” rating for all movies



The average rating is *3.5124821* which is to be calculated for RMSE.

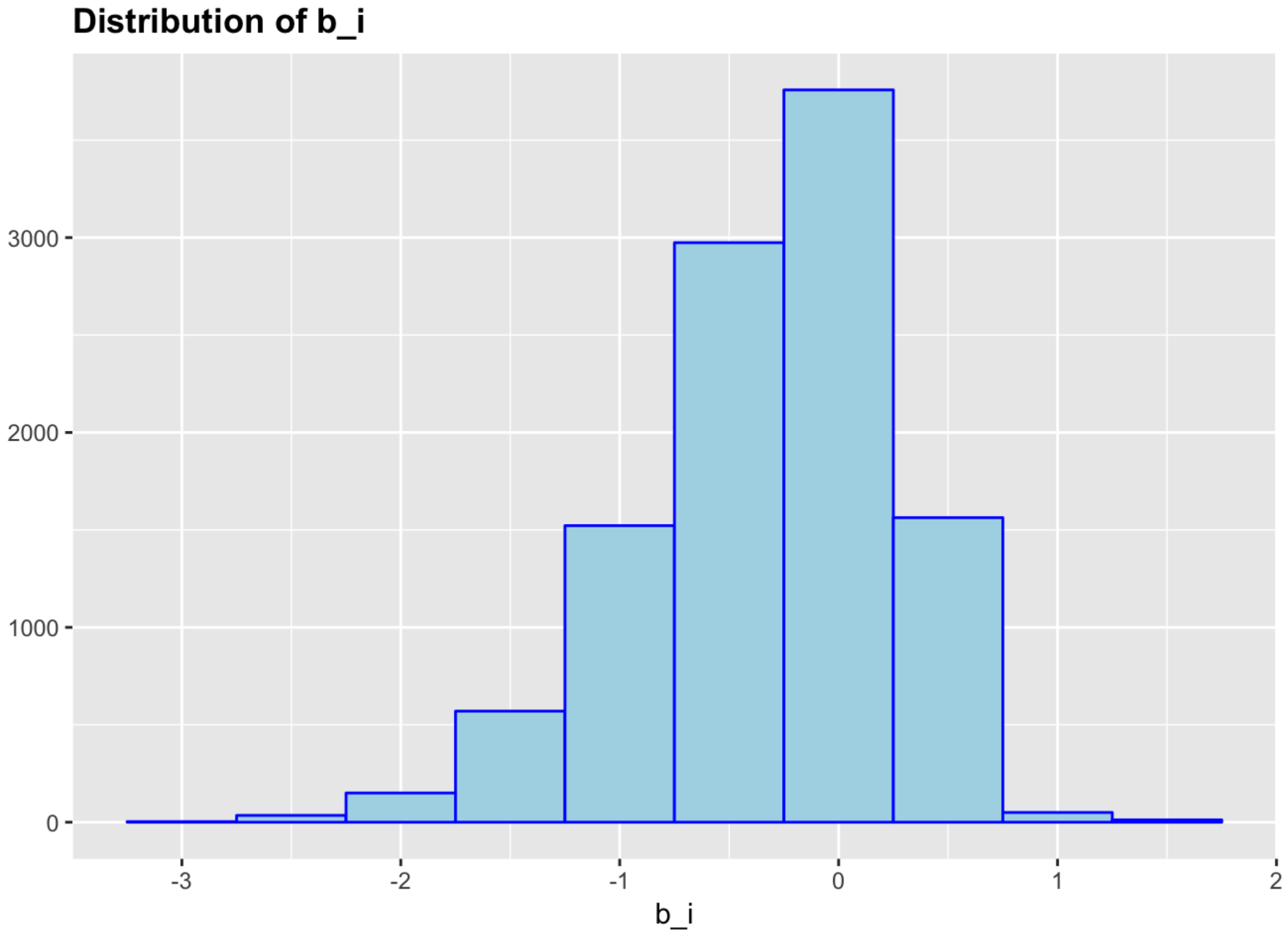
### 2.3.2 Model: Movie Effects

In reality, movies are rated differently. Some movies are rated higher than other movies. In this model, different movies are rated with different rating. The equation is as follows:

$$Y_{u,i} = \mu + b_i + \varepsilon_{u,i}$$

where:

- $b_i$ is the bias on movies,
- $\varepsilon_{u,i}$ is independent errors sampled from the same distribution centered at 0,
- $\mu$ is the “true” rating for all movies



From the above chart, the variation of *b_i* varies much ranging from approximate -3.25 to approximate 1.75. With refer to the first model, the average rating is around 3.5. For $b_i$ is 1.5, the rating will be a 5-star rating. $b_i$ is included to calculate the RMSE.

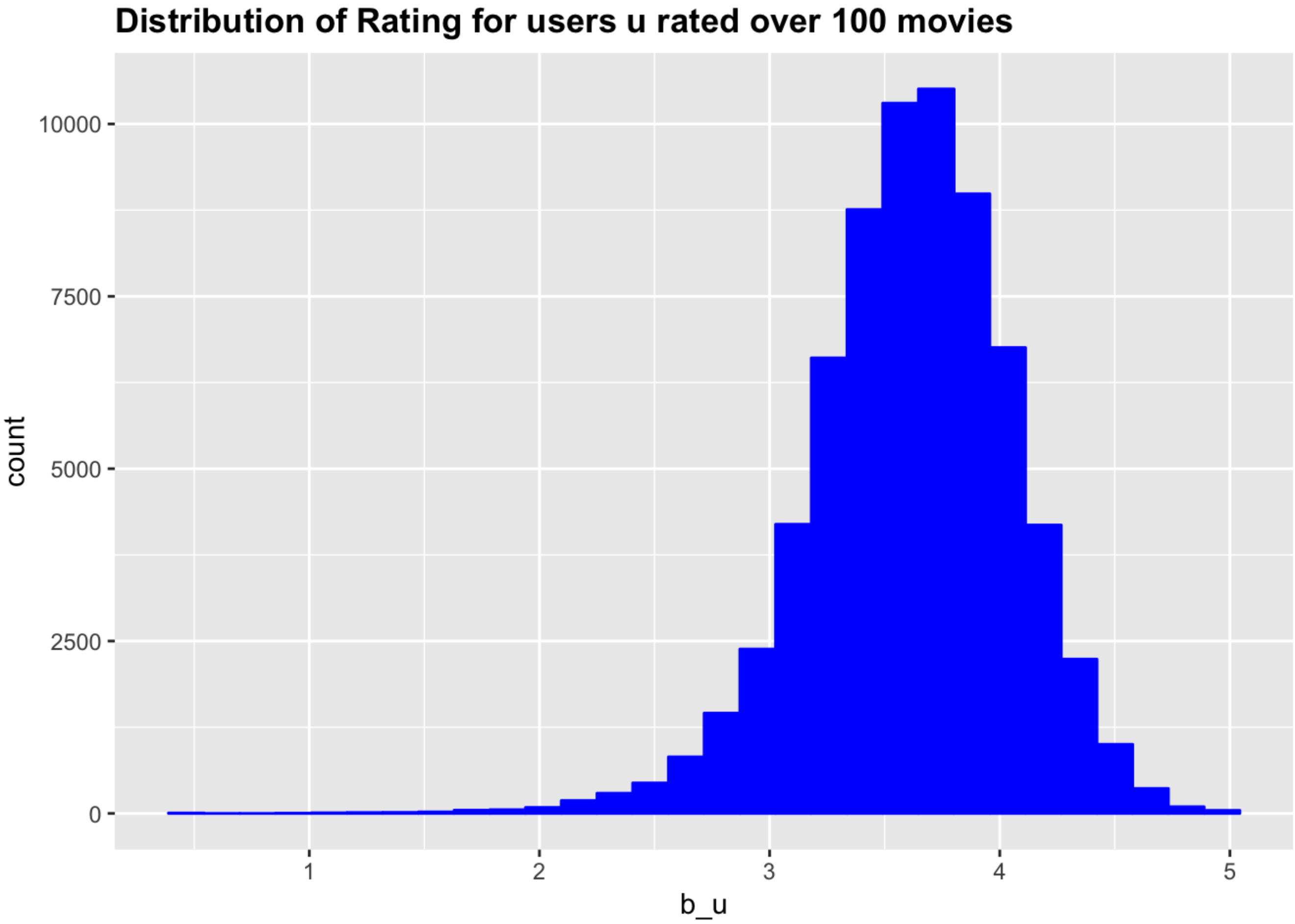
### 2.3.3 Model: Movie Effects and User Effects

Different users have preference for different movies. User effects will also be taken into this model. The equation is as follows:

$$Y_{u,i} = \mu + b_i + b_u + \varepsilon_{u,i}$$

where:

- $b_i$ is the bias on movies,
- $b_u$ is the user-specific effects,
- $\varepsilon_{u,i}$ is independent errors sampled from the same distribution centered at 0,
- $\mu$ is the “true” rating for all movies



It shows the rating variation of users is substantial. Both $b_i$ and $b_u$ is to be calculated in RMSE.

### 2.3.4 Model: Regularization with Movie & User Effects

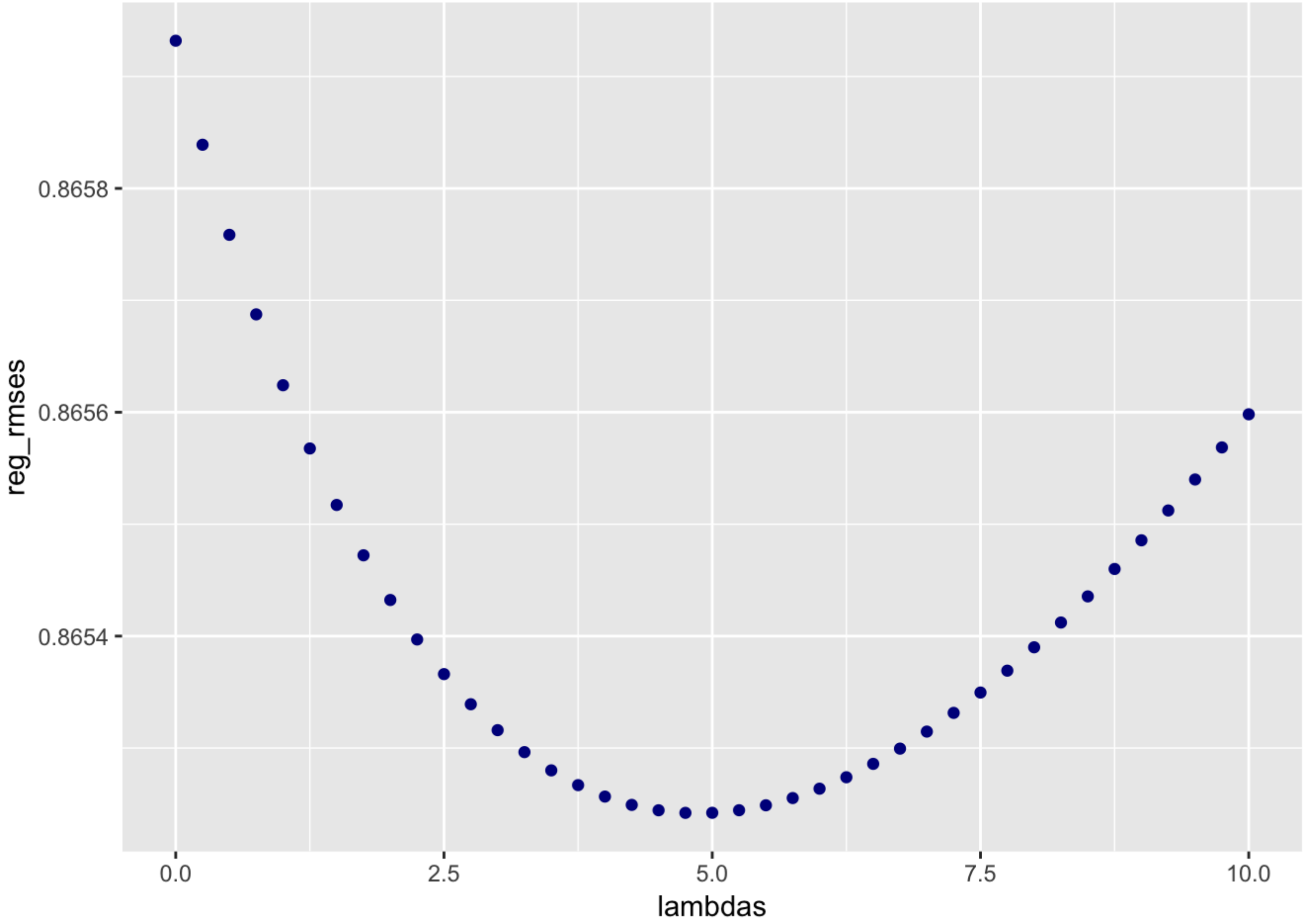
Regularization adds a penalty on different parameters of the model to reduce the noise of the training data and improve the generalization abilities of the model. It will penalize large estimates by small sample sizes. The equation is as follows:

$$\frac{1}{N}\sum_{u,i}(Y_{u,i} - u - b_i - b_u)^2 + \lambda(\sum_i b_i^2 + \sum_u b_u^2)$$

where:

- $\lambda$ is a penalty,
- $b_i$ is the bias on movies,
- $b_u$ is the user-specific effects,
- $\varepsilon_{u,i}$ is independent errors sampled from the same distribution centered at 0,
- $\mu$ is the “true” rating for all movies

To choose the penalty terms, cross-validation is used:



The lambdas for best RMSE is *4.75* which is to be calculated for RMSE.

# 3. Results

## 3.1 Result of Four Models

***3.1.1 Model: Simplest model without any effects***

RMWE is *1.0599043*

***3.1.2 Model: Movie Effects***

RMSE is *0.9437429*

***3.1.3 Model: Movie Effects and User Effects***

RMSE is *0.865932*

***3.1.4 Model: Regularization with Movie & User Effects*** RMSE is *0.8652421*

| method | RMSE |
|---|---:|
| Just the Average | 1.0599043 |
| Movie Effect Model | 0.9437429 |
| Movie + User Effects Model | 0.8659320 |
| Regularized Movie + User Effect Model | 0.8652421 |

The results show that the RMSE is improving with more effects taken into consideration. The **best** model is **Regularization with Movie & User Effects** with **RMSE 0.8652421**

## 3.2 RMSE for Validation Set by Regularized Movie & User Effect Model

In the edx set, **Regularization with Movie & User Effects Model** is the best model, i.e. the lowest RMSE. This model is applied to Validation set. The lambdas for best RMSE is *5.25* which is to be calculated for RMSE. The **RMSE** for **Validation Set** is **0.864817**.

# 4. Conclusions

To conclude, four models, including “Simplest model without any effects”“,”Model with Movie Effects“,”Model with both Movie effects & User effects" and “Regularization with Movie & User effects Model”, are applied and “Regularization with Movie & User effects Model” got the best result, i.e. best RMSE. “Regularization with Movie & User effects Model” is successfully applied on the validation set to calculate the RMSE.