Capstone Project: MovieLens

Amit V Singh

6/15/2019

Problem Defnition

This is capstone project on 'Movie recommendation system' to predict the movie rating by a user based on users past rating of movies. The dataset used for this purpose can be found in the following links

- [MovieLens 10M dataset] https://grouplens.org/datasets/movielens/10m/
- [MovieLens 10M dataset - zip file] http://files.grouplens.org/datasets/movielens/ml-10m.zip

The challenge is not so easy given that there are many different type of biases present in the movie reviews. It can be different social, psychological, demographic variations that changes the taste of every single users for a given particular movies. However the problem can still be designed to tacle major biases which can be expressed via mathematical equations relativly easily. The idea here is to develop a model which can effectivly predict movie recommendations for a given user without our judgement being impaired due to different biases. In the algorithm, the prevalences can be suppressed using some clever mathematical tricks. This will become clear as we follow this document.

Data Ingestion

The code is provided in the edx capstone project module https://courses.edx.org/courses/course-v1:HarvardX+PH125.9x+2T2018/courseware/dd9a048b16ca477a8f0aaf1d888f0734/e8800e37aa444297a3a2f35bf84ce452/?child=first

```
#Create test and validation sets
# Create edx set, validation set, and submission file
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
# MovieLens 10M dataset:
# https://grouplens.org/datasets/movielens/10m/
# http://files.grouplens.org/datasets/movielens/ml-10m.zip
dl <- tempfile()
download.file("http://files.grouplens.org/datasets/movielens/ml-10m.zip", dl)
ratings <- read.table(text = gsub("::", "\t", readLines(unzip(dl, "ml-10M100K/ratings.dat"))),
                      col.names = c("userId", "movieId", "rating", "timestamp"))
movies <- str_split_fixed(readLines(unzip(dl, "ml-10M100K/movies.dat")), "\\::", 3)
colnames(movies) <- c("movieId", "title", "genres")
movies <- as.data.frame(movies) %>% mutate(movieId = as.numeric(levels(movieId))[movieId],
                                           title = as.character(title),
                                           genres = as.character(genres))
movielens <- left_join(ratings, movies, by = "movieId")
# Validation set will be 10% of MovieLens data
set.seed(1)
test_index <- createDataPartition(y = movielens$rating, times = 1, p = 0.1, list = FALSE)
edx <- movielens[-test_index,]
temp <- movielens[test_index,]

# Make sure userId and movieId in validation set are also in edx set
validation <- temp %>%
```

```
  semi_join(edx, by = "movieId") %>%
  semi_join(edx, by = "userId")
# Add rows removed from validation set back into edx set
removed <- anti_join(temp, validation)
edx <- rbind(edx, removed)
rm(dl, ratings, movies, test_index, temp, movielens, removed)
```

The above chunk of code gives a partioned dataset for training and testing our dataset. It also removes the unnecessary files from the working directory, which is always a good coding practice ('always clean after you cook').

```
# Validation dataset can be further modified by removing rating column
validation_CM <- validation
validation <- validation %>% select(-rating)
```

```
# extra libraries that might be usefull in analysis and visulizations
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

Once a clean dataset is availble, one must inquire the dataset features and calculate the basic summary statistics

```
## the dataset and its basic summary statistics

# intial 7 rows with header
head(edx)
```

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

```
# basic summary statistics
summary(edx)
```

```
##      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
##
##
##
```

```
# total number of observations
tot_observation <- length(edx$rating) + length(validation$rating)
```

We can see the dataset is in the tidy fromat and ready for exploration and analysis. ## Data pre-processing

```
#  Since RMSE (root mean squre error) is used frequency so lets define a function

RMSE <- function(true_ratings, predicted_ratings){
  sqrt(mean((true_ratings-predicted_ratings)^2,na.rm=T))
}

# lets modify the columns to suitable formats that can be further used for analysis

# Modify the year as a column in the edx & validation datasets
edx <- edx %>% mutate(year = as.numeric(str_sub(title,-5,-2)))
validation <- validation %>% mutate(year = as.numeric(str_sub(title,-5,-2)))
validation_CM <- validation_CM %>% mutate(year = as.numeric(str_sub(title,-5,-2)))

# Modify the genres variable in the edx & validation dataset (column separated)
split_edx  <- edx  %>% separate_rows(genres, sep = "\\|")
split_valid <- validation   %>% separate_rows(genres, sep = "\\|")
split_valid_CM <- validation_CM  %>% separate_rows(genres, sep = "\\|")
```

Data Exploration and general statistics

```
# The 1st rows of the edx & split_edx datasets are presented below:
head(edx)
```

```
##   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 year
## 1                Comedy|Romance 1992
## 2         Action|Crime|Thriller 1995
## 3  Action|Drama|Sci-Fi|Thriller 1995
## 4       Action|Adventure|Sci-Fi 1994
## 5 Action|Adventure|Drama|Sci-Fi 1994
## 6       Children|Comedy|Fantasy 1994
```

```r
head(split_edx)
```

```
##   userId movieId rating timestamp            title   genres year
## 1      1     122      5 838985046 Boomerang (1992)   Comedy 1992
## 2      1     122      5 838985046 Boomerang (1992)  Romance 1992
## 3      1     185      5 838983525  Net, The (1995)   Action 1995
## 4      1     185      5 838983525  Net, The (1995)    Crime 1995
## 5      1     185      5 838983525  Net, The (1995) Thriller 1995
## 6      1     292      5 838983421  Outbreak (1995)   Action 1995
```

```r
# edx Summary Statistics
summary(edx)
```

```
##      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               year
##  Length:9000055     Length:9000055     Min.   :1915
##  Class :character   Class :character   1st Qu.:1987
##  Mode  :character   Mode  :character   Median :1994
##                                        Mean   :1990
##                                        3rd Qu.:1998
##                                        Max.   :2008
```

```r
# Number of unique movies and users in the edx dataset

edx %>% summarize(n_users = n_distinct(userId), n_movies = n_distinct(movieId))
```

```
##   n_users n_movies
## 1   69878    10677
```

Total movie ratings per genre

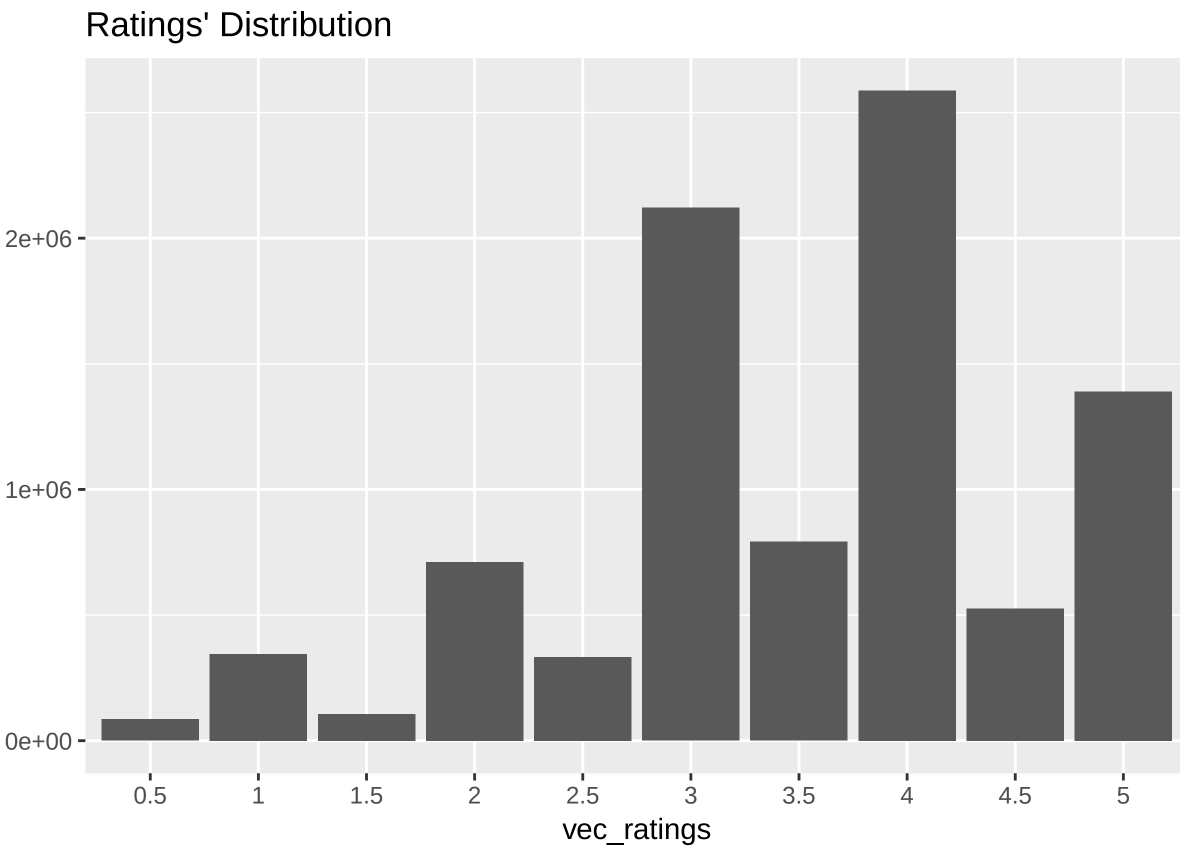
```r
genre_rating <- split_edx%>%
  group_by(genres) %>%
  summarize(count = n()) %>%
  arrange(desc(count))
```

Ratings distribution

```r
vec_ratings <- as.vector(edx$rating)
unique(vec_ratings)
```

```
##  [1] 5.0 3.0 2.0 4.0 4.5 3.5 1.0 1.5 2.5 0.5
```

```r
vec_ratings <- vec_ratings[vec_ratings != 0]
vec_ratings <- factor(vec_ratings)
qplot(vec_ratings) +
  ggtitle("Ratings' Distribution")
```



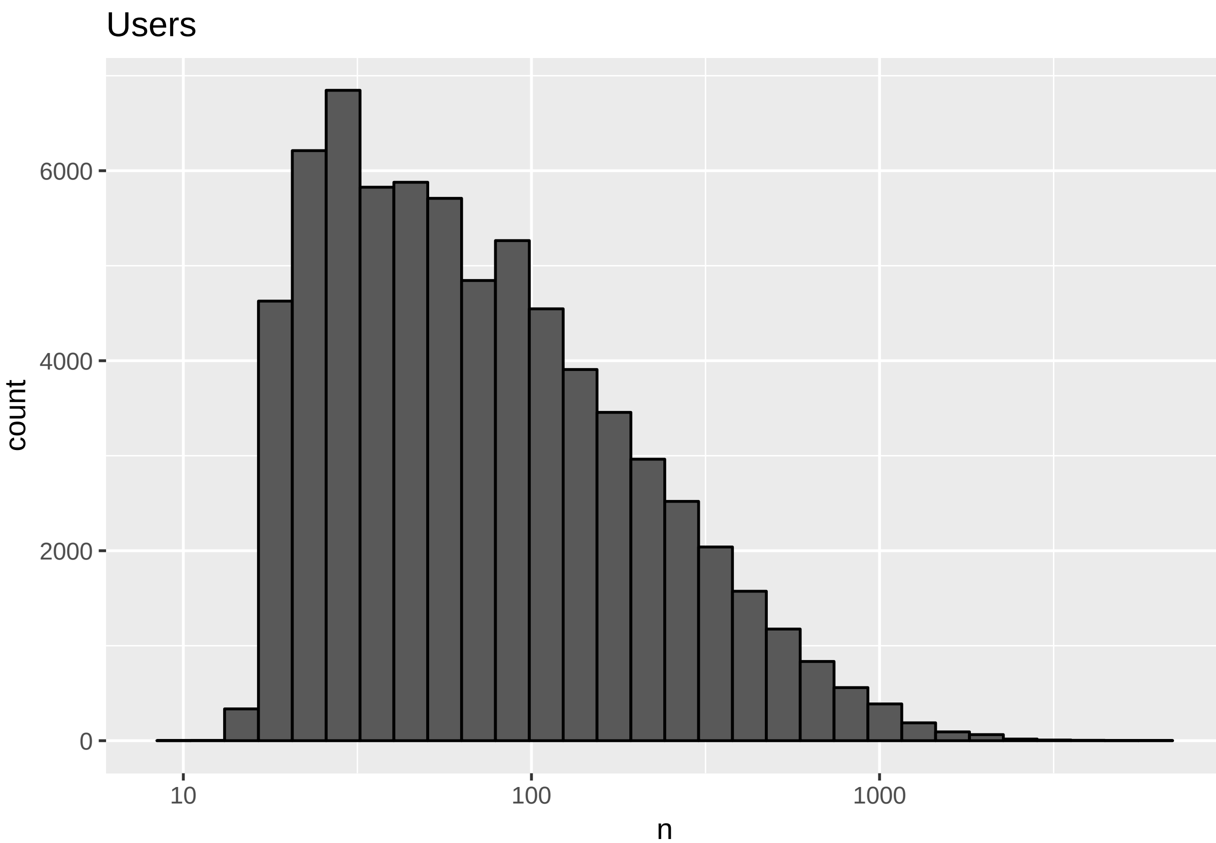
The above rating distribution shows that the users have a general tendency to rate movies between 3 and 4. This is a very general conclusion. We should further explore the effect of different features to make a good predictive model.

Data Analysis Strategies

- Some movies are rated more often than others (e.g. blockbusters are rated higher). How to incorporate this in our model: find movie bias.
- Some users are positive and some have negative reviews beacuase of their own personal liking/disliking regardless of movie. How to address this characteristics: find users bias.
- The popularity of the movie genre depends strongly on the contemporary issues. So we should also explore the time dependent analysis. How to approach this idea: find the genre popularity over the years
- Do the users mindset also evolve over time? This can also effect the average rating od movies over the years. How do visulize such effect: plot rating vs release year

The distribution of each user's ratings for movies. This shows the users bias

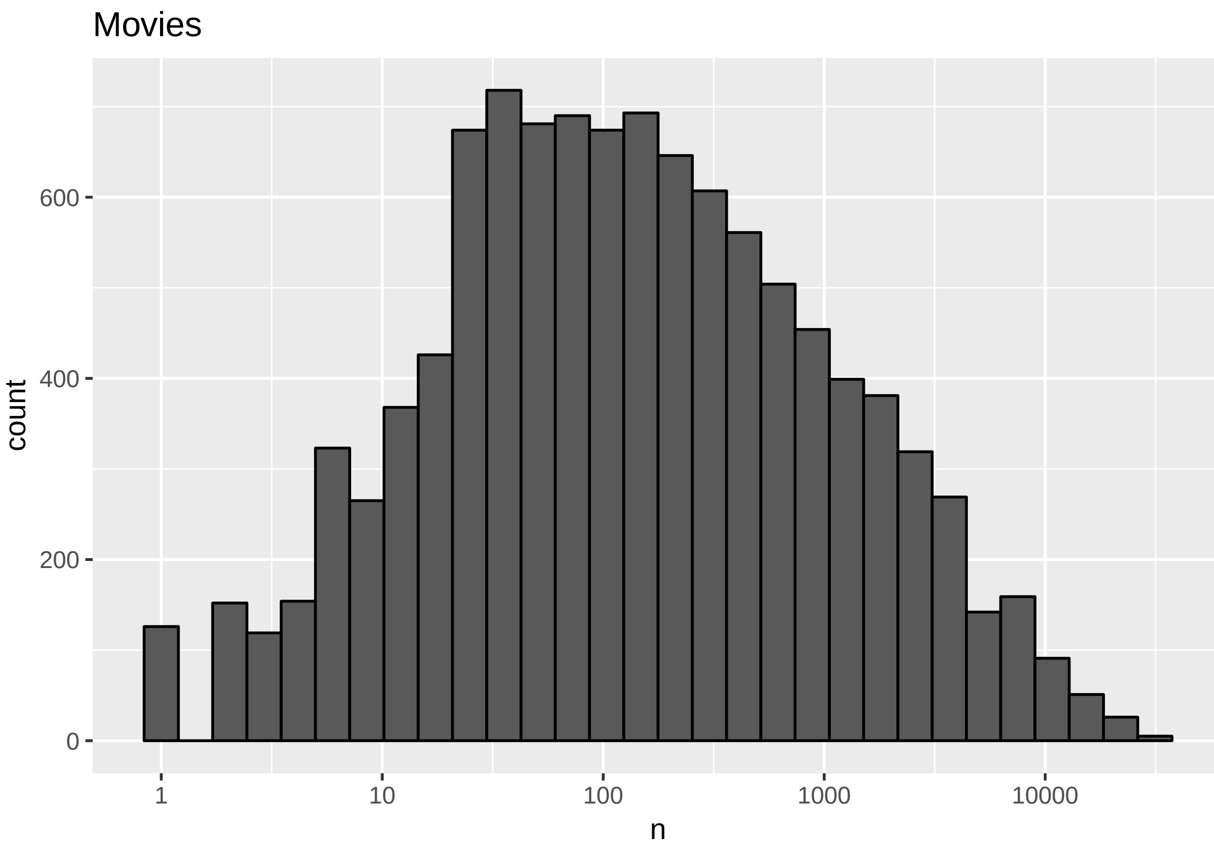
```
edx %>% count(userId) %>%
  ggplot(aes(n)) +
  geom_histogram(bins = 30, color = "black") +
  scale_x_log10() +
  ggtitle("Users")
```



Above plot shows that not every user is equaly active. Some users have rated very few movie and their oponion may bias the prediction results.

Some movies are rated more often than others.Below is their distribution. This explores movie biases.

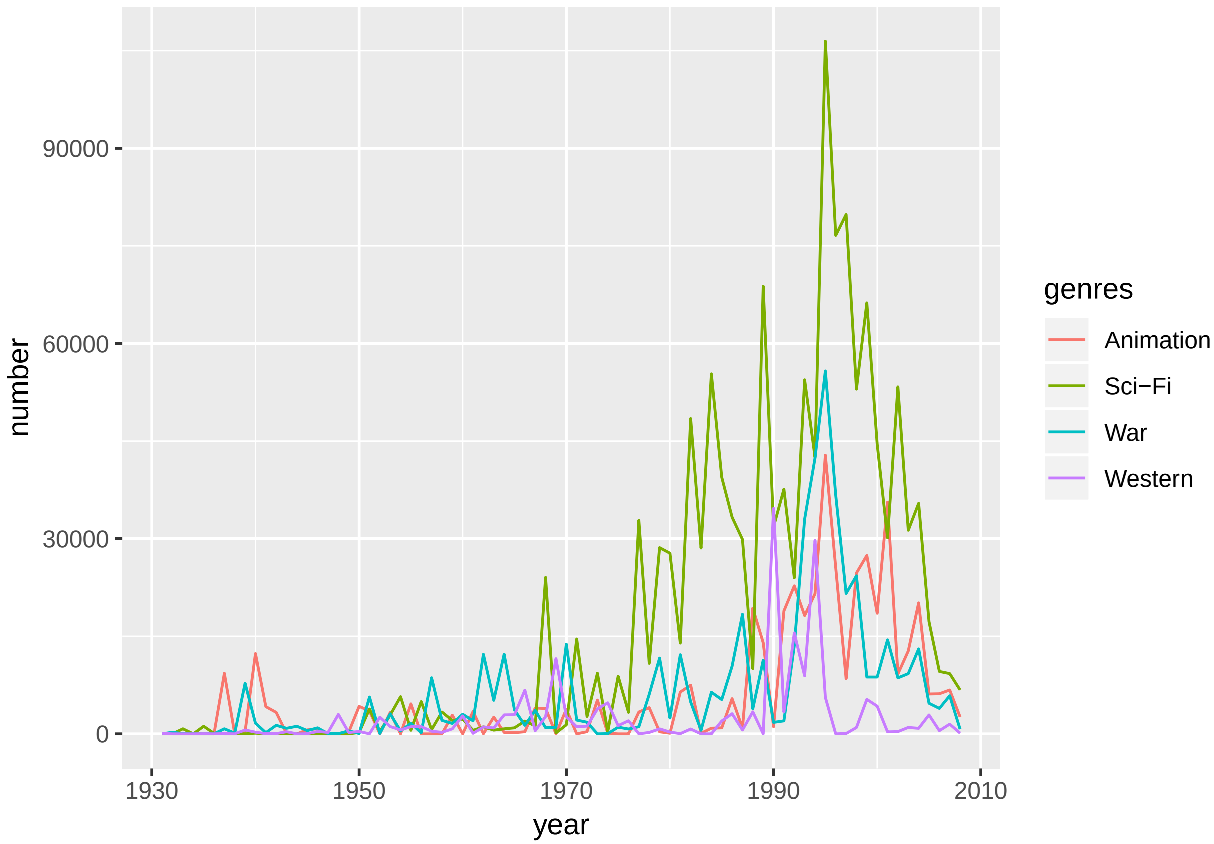
```
edx %>%
  count(movieId) %>%
  ggplot(aes(n)) +
  geom_histogram(bins = 30, color = "black") +
  scale_x_log10() +
  ggtitle("Movies")
```



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

Genres popularity per year. Here we tackle the issue of temporal evolution of users taste over different popular genre.

```
genres_popularity <- split_edx %>%
  na.omit() %>% # omit missing values
  select(movieId, year, genres) %>% # select columns we are interested in
  mutate(genres = as.factor(genres)) %>% # turn genres in factors
  group_by(year, genres) %>% # group data by year and genre
  summarise(number = n()) %>% # count
  complete(year = full_seq(year, 1), genres, fill = list(number = 0)) # add missing years/genres

# Genres vs year; 4 genres are chosen for readability: animation, sci-fi, war and western movies.

genres_popularity %>%
  filter(year > 1930) %>%
  filter(genres %in% c("War", "Sci-Fi", "Animation", "Western")) %>%
  ggplot(aes(x = year, y = number)) +
  geom_line(aes(color=genres)) +
  scale_fill_brewer(palette = "Paired")
```

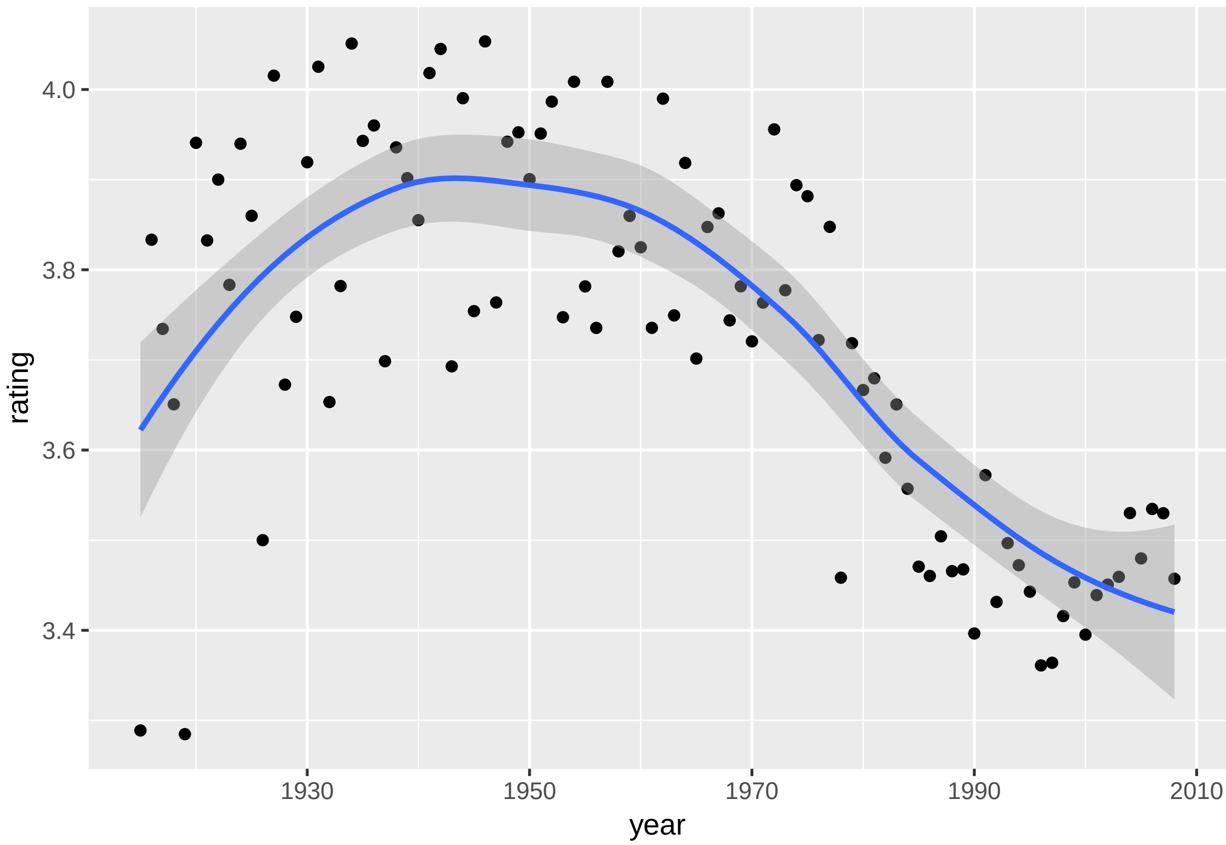


This plots depicts some genre become more popular over others for differnt period of time.

Rating vs release year. Here a general trent of movie viewers and thier rating habits can be explored.

```
edx %>% group_by(year) %>%
  summarize(rating = mean(rating)) %>%
  ggplot(aes(year, rating)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The genral trend shows modern users ralativly rate movies lower.

Data Analysis: Model Preparation

```
#Initiate RMSE results to compare various models
rmse_results <- data_frame()
```

Simplest possible model

Dataset's mean rating is used to predict the same rating for all movies, regardless of the user and movie.

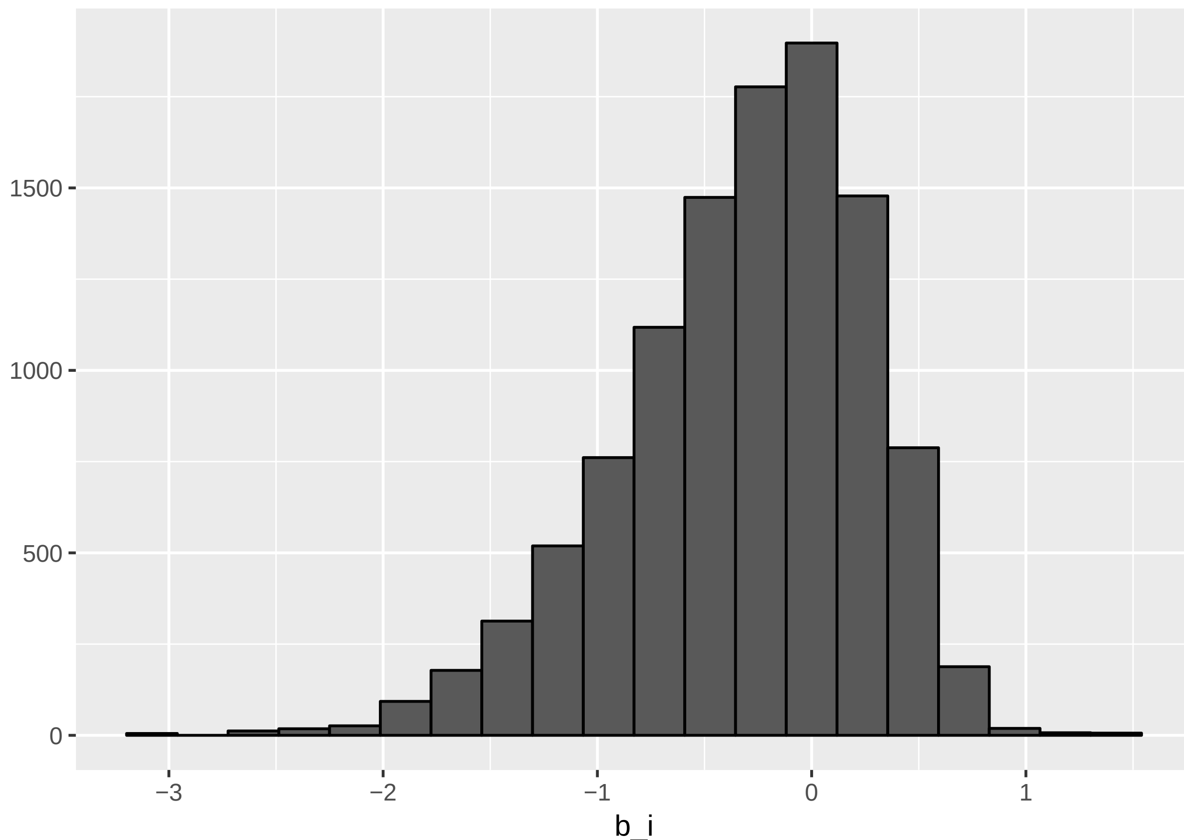
```
mu <- mean(edx$rating)
mu
```

```
## [1] 3.512465
```

Penalty Term (b_i)- Movie Effect

Different movies are rated differently. As shown in the exploration, the histogram is not symmetric and is skewed towards negative rating effect. The movie effect can be taken into accout by taking he difference from mean rating as shown in the following chunk of code.

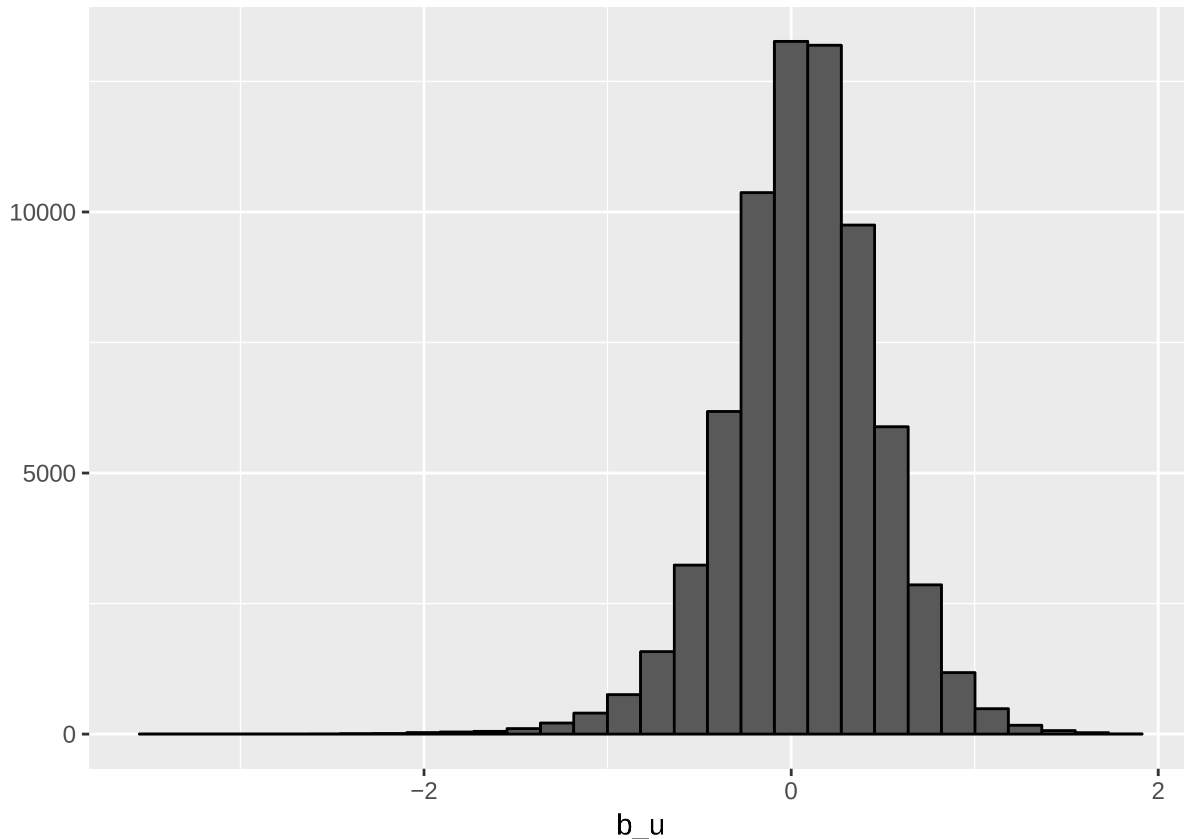
```
movie_avgs_norm <- edx %>%
  group_by(movieId) %>%
  summarize(b_i = mean(rating - mu))
movie_avgs_norm %>% qplot(b_i, geom ="histogram", bins = 20, data = ., color = I("black"))
```



Penalty Term (b_u)- User Effect

Different users are different in terms of how they rate movies. Some cranky users may rate a good movie lower or some very generous users just don't care for accessment. We have already seen this pattern in our data exploration plot (user bias). We can calculate it using this code.

```
user_avgs_norm <- edx %>%
  left_join(movie_avgs_norm, by='movieId') %>%
  group_by(userId) %>%
  summarize(b_u = mean(rating - mu - b_i))
user_avgs_norm %>% qplot(b_u, geom ="histogram", bins = 30, data = ., color = I("black"))
```



Model Creation

The quality of the model will be assessed by the RMSE (the lower the better).

Baseline Model

Its simply a model which ignores all the feathers and simply calculates mean rating. This model acts as a baseline model and we will try to improve RMSE relative to this baseline standard model.

```
# baseline Model: just the mean
baseline_rmse <- RMSE(validation_CM$rating,mu)
## Test results based on simple prediction
baseline_rmse
```

```
## [1] 1.061202
```

```
## Check results
rmse_results <- data_frame(method = "Using mean only", RMSE = baseline_rmse)
rmse_results
```

```
## # A tibble: 1 x 2
##   method           RMSE
##   <chr>           <dbl>
## 1 Using mean only  1.06
```

Movie Effect Model

An improvement in the RMSE is achieved by adding the movie effect.

```r
# Movie effects only
predicted_ratings_movie_norm <- validation %>%
  left_join(movie_avgs_norm, by='movieId') %>%
  mutate(pred = mu + b_i)
model_1_rmse <- RMSE(validation_CM$rating,predicted_ratings_movie_norm$pred)
rmse_results <- bind_rows(rmse_results,
                          data_frame(method="Movie Effect Model",
                                     RMSE = model_1_rmse ))
rmse_results %>% knitr::kable()
```

method	RMSE
Using mean only	1.0612018
Movie Effect Model	0.9439087

```r
rmse_results
```

```
## # A tibble: 2 x 2
##   method              RMSE
##   <chr>              <dbl>
## 1 Using mean only    1.06
## 2 Movie Effect Model 0.944
```

The error has drop by 5% and motivates us to move on this path further.

Movie and User Effect Model

Given that movie and users biases both obscure the prediction of movie rating, a further improvement in the RMSE is achieved by adding the user effect.

```r
# Use test set,join movie averages & user averages
# Prediction equals the mean with user effect b_u & movie effect b_i
predicted_ratings_user_norm <- validation %>%
  left_join(movie_avgs_norm, by='movieId') %>%
  left_join(user_avgs_norm, by='userId') %>%
  mutate(pred = mu + b_i + b_u)
# test and save rmse results
model_2_rmse <- RMSE(validation_CM$rating,predicted_ratings_user_norm$pred)
rmse_results <- bind_rows(rmse_results,
                          data_frame(method="Movie and User Effect Model",
                                     RMSE = model_2_rmse ))
rmse_results %>% knitr::kable()
```

method	RMSE
Using mean only	1.0612018
Movie Effect Model	0.9439087
Movie and User Effect Model	0.8653488

```r
rmse_results
```

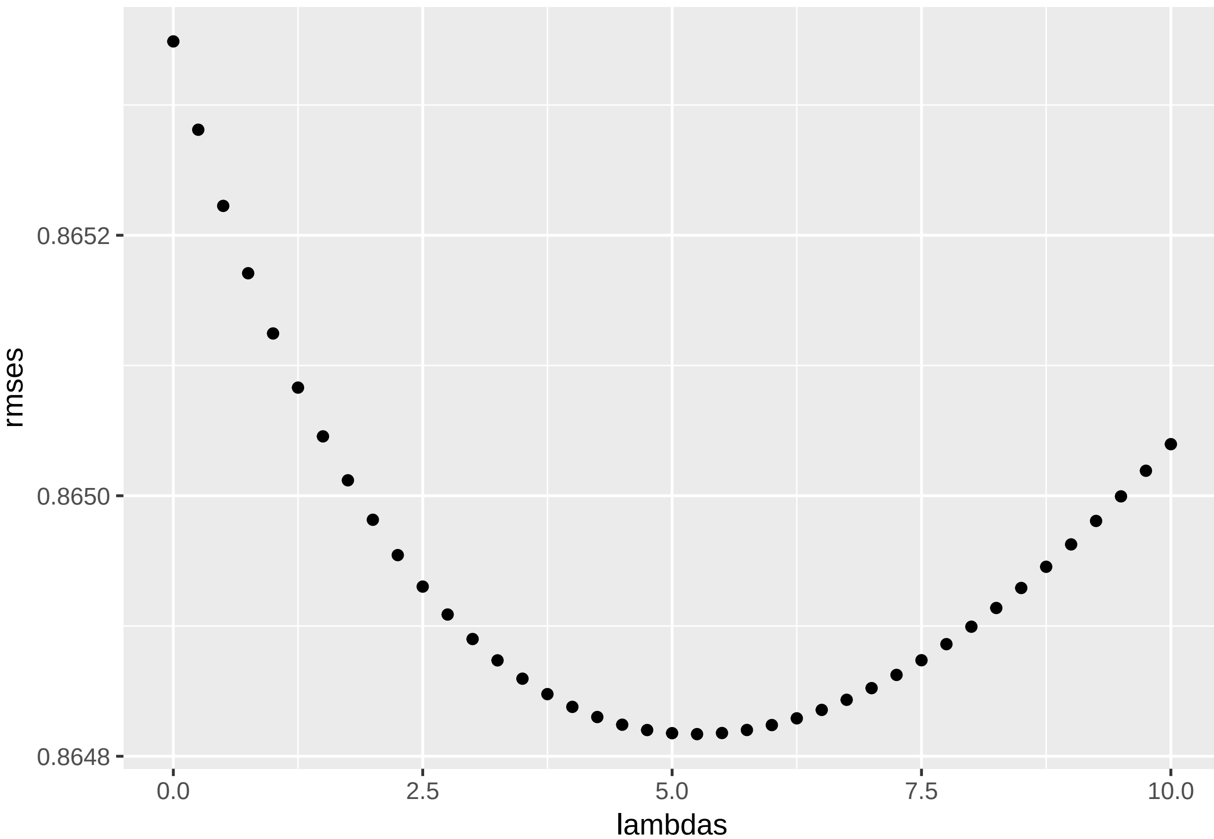
```
## # A tibble: 3 x 2
##   method                       RMSE
##   <chr>                       <dbl>
## 1 Using mean only             1.06
## 2 Movie Effect Model          0.944
## 3 Movie and User Effect Model 0.865
```

This is a good improvement from our last model.

Regularization based approach (motivated by Netflix challenge)

We have noticed in our data exploraton, some users are more activly participated in movie reviewing. There are also users who have rated very few movies (less than 30 movies). On the other hand, some movies are rated very few times (say 1 or 2). These are basically noisy estimates that we should not trust, especially when it comes to prediction. Additionally, RMSE are sensative to large errors. Large errors can increase our residual mean squared error. So we must put a penalty term to give less importance to such effect.

```r
# lambda is a tuning parameter
# Use cross-validation to choose it.
lambdas <- seq(0, 10, 0.25)
# For each lambda,find b_i & b_u, followed by rating prediction & testing
# note:the below code could take some time
rmses <- sapply(lambdas, function(l){

  mu <- mean(edx$rating)

  b_i <- edx %>%
    group_by(movieId) %>%
    summarize(b_i = sum(rating - mu)/(n()+l))

  b_u <- edx %>%
    left_join(b_i, by="movieId") %>%
    group_by(userId) %>%
    summarize(b_u = sum(rating - b_i - mu)/(n()+l))

  predicted_ratings <- validation %>%
    left_join(b_i, by = "movieId") %>%
    left_join(b_u, by = "userId") %>%
    mutate(pred = mu + b_i + b_u) %>%
    .$pred

  return(RMSE(validation_CM$rating,predicted_ratings))
})


# Plot rmses vs lambdas to select the optimal lambda
qplot(lambdas, rmses)
```




```
lambda <- lambdas[which.min(rmses)]
lambda
```

```
## [1] 5.25
```

```
# Compute regularized estimates of b_i using lambda
movie_avgs_reg <- edx %>%
  group_by(movieId) %>%
  summarize(b_i = sum(rating - mu)/(n()+lambda), n_i = n())
# Compute regularized estimates of b_u using lambda
user_avgs_reg <- edx %>%
  left_join(movie_avgs_reg, by='movieId') %>%
  group_by(userId) %>%
  summarize(b_u = sum(rating - mu - b_i)/(n()+lambda), n_u = n())
# Predict ratings
predicted_ratings_reg <- validation %>%
  left_join(movie_avgs_reg, by='movieId') %>%
  left_join(user_avgs_reg, by='userId') %>%
  mutate(pred = mu + b_i + b_u) %>%
  .$pred
# Test and save results
model_3_rmse <- RMSE(validation_CM$rating,predicted_ratings_reg)
rmse_results <- bind_rows(rmse_results,
                          data_frame(method="Regularized Movie and User Effect Model",
                                     RMSE = model_3_rmse ))
rmse_results %>% knitr::kable()
```

method	RMSE
Using mean only	1.0612018
Movie Effect Model	0.9439087
Movie and User Effect Model	0.8653488
Regularized Movie and User Effect Model	0.8648170

```
rmse_results
```

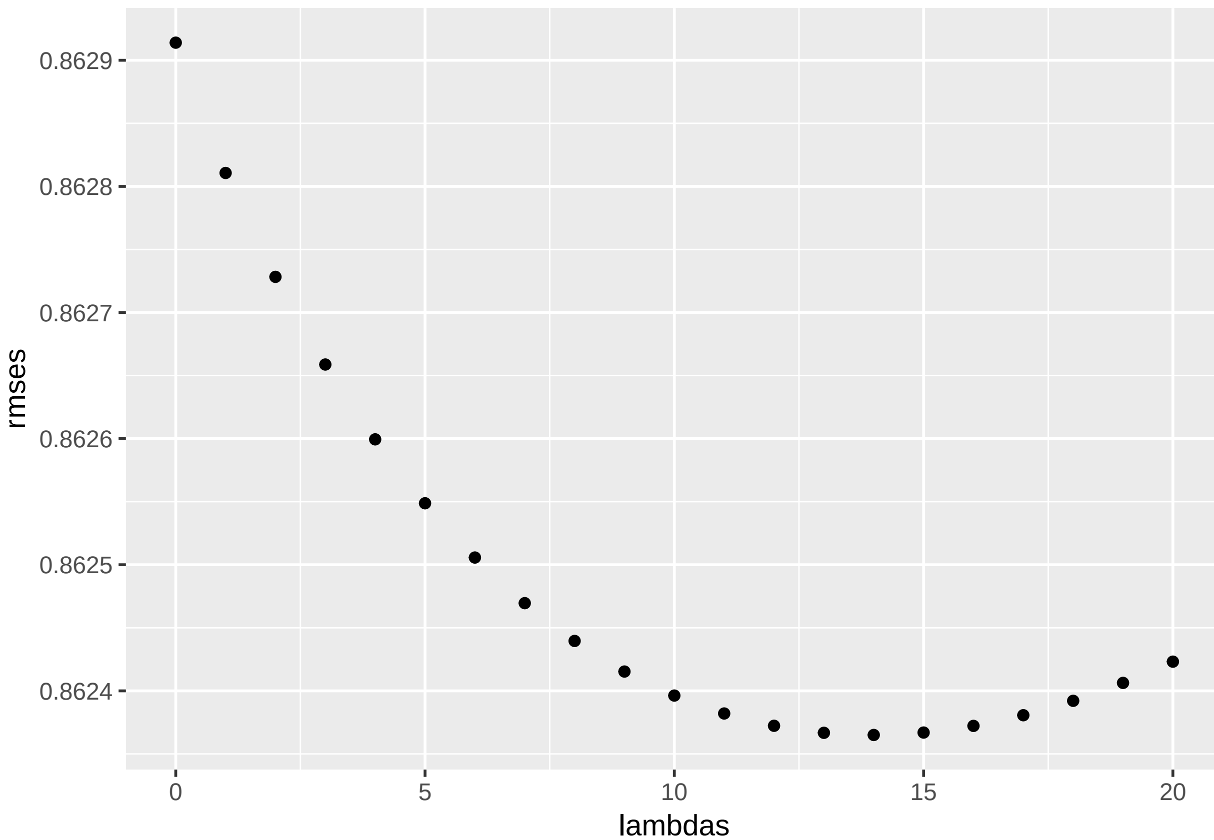
```
## # A tibble: 4 x 2
##   method                                   RMSE
##   <chr>                                   <dbl>
## 1 Using mean only                         1.06
## 2 Movie Effect Model                      0.944
## 3 Movie and User Effect Model             0.865
## 4 Regularized Movie and User Effect Model 0.865
```

Regularization using movies, users, years and genres.

The approach utilized in the above model is implemented below with the added genres and release year effects.

```
# b_y and b_g represent the year & genre effects, respectively
lambdas <- seq(0, 20, 1)
# Note: the below code could take some time
rmses <- sapply(lambdas, function(l){

  mu <- mean(edx$rating)

  b_i <- split_edx %>%
    group_by(movieId) %>%
    summarize(b_i = sum(rating - mu)/(n()+l))

  b_u <- split_edx %>%
    left_join(b_i, by="movieId") %>%
    group_by(userId) %>%
    summarize(b_u = sum(rating - b_i - mu)/(n()+l))

  b_y <- split_edx %>%
    left_join(b_i, by='movieId') %>%
    left_join(b_u, by='userId') %>%
    group_by(year) %>%
    summarize(b_y = sum(rating - mu - b_i - b_u)/(n()+lambda), n_y = n())

  b_g <- split_edx %>%
    left_join(b_i, by='movieId') %>%
    left_join(b_u, by='userId') %>%
    left_join(b_y, by = 'year') %>%
    group_by(genres) %>%
    summarize(b_g = sum(rating - mu - b_i - b_u - b_y)/(n()+lambda), n_g = n())

    predicted_ratings <- split_valid %>%
    left_join(b_i, by='movieId') %>%
    left_join(b_u, by='userId') %>%
```

```r
    left_join(b_y, by = 'year') %>%
    left_join(b_g, by = 'genres') %>%
    mutate(pred = mu + b_i + b_u + b_y + b_g) %>%
    .$pred

  return(RMSE(split_valid_CM$rating,predicted_ratings))
})
# Compute new predictions using the optimal lambda
# Test and save results
qplot(lambdas, rmses)
```




```r
lambda_2 <- lambdas[which.min(rmses)]
lambda_2
```

```
## [1] 14
```

```r
movie_reg_avgs_2 <- split_edx %>%
  group_by(movieId) %>%
  summarize(b_i = sum(rating - mu)/(n()+lambda_2), n_i = n())
user_reg_avgs_2 <- split_edx %>%
  left_join(movie_reg_avgs_2, by='movieId') %>%
  group_by(userId) %>%
  summarize(b_u = sum(rating - mu - b_i)/(n()+lambda_2), n_u = n())
year_reg_avgs <- split_edx %>%
  left_join(movie_reg_avgs_2, by='movieId') %>%
  left_join(user_reg_avgs_2, by='userId') %>%
  group_by(year) %>%
  summarize(b_y = sum(rating - mu - b_i - b_u)/(n()+lambda_2), n_y = n())
```

```r
genre_reg_avgs <- split_edx %>%
  left_join(movie_reg_avgs_2, by='movieId') %>%
  left_join(user_reg_avgs_2, by='userId') %>%
  left_join(year_reg_avgs, by = 'year') %>%
  group_by(genres) %>%
  summarize(b_g = sum(rating - mu - b_i - b_u - b_y)/(n()+lambda_2), n_g = n())
predicted_ratings <- split_valid %>%
  left_join(movie_reg_avgs_2, by='movieId') %>%
  left_join(user_reg_avgs_2, by='userId') %>%
  left_join(year_reg_avgs, by = 'year') %>%
  left_join(genre_reg_avgs, by = 'genres') %>%
  mutate(pred = mu + b_i + b_u + b_y + b_g) %>%
  .$pred
model_4_rmse <- RMSE(split_valid_CM$rating,predicted_ratings)
rmse_results <- bind_rows(rmse_results,
                          data_frame(method="Reg Movie, User, Year, and Genre Effect Model",
                                     RMSE = model_4_rmse ))
rmse_results %>% knitr::kable()
```

method	RMSE
Using mean only	1.0612018
Movie Effect Model	0.9439087
Movie and User Effect Model	0.8653488
Regularized Movie and User Effect Model	0.8648170
Reg Movie, User, Year, and Genre Effect Model	0.8623650

3. Results

RMSE overview

The RMSE values for the used models are shown below:

```r
rmse_results %>% knitr::kable()
```


method	RMSE
Using mean only	1.0612018
Movie Effect Model	0.9439087
Movie and User Effect Model	0.8653488
Regularized Movie and User Effect Model	0.8648170
Reg Movie, User, Year, and Genre Effect Model	0.8623650

Concluding Remarks

The RMSE table shows an improvement of the model over different assumptions. The simplest model 'Using mean only' calculates the RMSE more than 1, which means we may miss the rating by one star (not good!!). Then incorporating 'Movie effect' and 'Movie and user effect' on model gives an improvement by 5% and 13.5%. This is substantial improvement given the simplicity of the model. A deeper insight in the data revealed some data point in the feathers have large effect on errors. So a regulization model was used to penalise such data points. The final RMSE is 0.8623 with an improvement over 13.3% with respect to the baseline model. This implies we can trust our prediction for movie rating given by a users.

(**References** 1. https://github.com/johnfelipe/MovieLens-2 2. https://github.com/cmrad/Updated-MovieLens-Rating-Predic