MovieLens Project

Tagaro, Jerome Vincent L.

2024-11-06

1 Introduction

1.1 Movielens Capstone Project

The goal is to create a movie recommendation system using the full-length movielens dataset (available and obtained at: https://grouplens.org/datasets/movielens/latest/ and http://files.grouplens.org/datasets/movielens/ml-10m.zip).

The MovieLens data is downloaded and the provided code for generating the datasets is run. Then, using the generated data set, train a machine learning algorithm using the inputs in one subset of the data as *Input Set* to predict the movie ratings in a *Evaluation Set*. This prediction will be the basis for the recommendation system, i.e. a higher predicted rating implies that a user will like the specified movie. In this report, the *Input Set* will be the `edx` data and the *Evaluation Set* will be the `final_holdout_test` data.
The *Evaluation Set* should only be used for the final model evaluation. Thus, the *Input Set* has to be divided into a smaller *train set* and a *test set*.

The model/s will use linear regression to predict `ratings`. Thus, the models will assume or be similar to the form :
$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon$
Where,
$i = 1, \ldots, n$
$x_{i,n}$ are the corresponding biases/preferences, and,
β_n are the constants/coefficients.

With linear regression, building the model starts with identifying the biases or the $x_{i,n}$ variables. Then, derive corresponding coefficients or β_n variables.

All the models in the project will be evaluated using the *Root Mean Square Error* or *RMSE* between the predicted and actual ratings. *RMSE* can be obtained using the formula:
$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(predicted_i - actual_i)}$
The target *RMSE* of the project is $RMSE < 0.86490$.

This report will use some of the codes in the *previous courses* as starting points. Mainly, the code will use the previous *User Bias* and *Movie Bias* as starting point or for comparison.

1.2 Data set

The entire data set is made up of the *Input Set* `edx` and the *Evaluation set* `final_holdout_test`.

Looking at a snippet of the *Input Set* `edx`:

	userId	movieId	rating	timestamp	title	genres
1	1	122	5	838985046	Boomerang (1992)	Comedy|Romance
2	1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
4	1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
5	1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi

A basic summary of the data sets:

- The *Input Set* has 9000055 rows.
- The *Evaluation Set* has 999999 rows.
- Both Data sets will have 6 columns.
- The column_names are userId, movieId, rating, timestamp, title, genres.

From the summary, the 6 columns can be divided into *inputs* and an *output*.

Input columns:
- The `userId` identifies the user.
- The `movieId` identifies the movie.
- The `timestamp` should be the time the rating is made.
- The `title` is the title associated with the corresponding `movieId`.
- The `genre` column lists out the associated genres for the movie in a string.

These 5 columns are the *inputs* of the model that will be used to train the model for predicting the *output*.

Output column:
- The `rating` column is the rating made by the user on a movie.

The `rating` column is the *output* of the model, and will be the basis of the recommendation system.

The `genres` column show the movies genres in one string. Using *regex* the individual genres can be separated and extracted.

There are 20 unique genres in the entire *Input Set*. These genres are:

Comedy, Romance, Action, Crime, Thriller, Drama, Sci-Fi, Adventure, Children, Fantasy, War, Animation, Musical, Western, Mystery, Film-Noir, Horror, Documentary, IMAX, no genres listed

2 Methods / Analysis

2.1 Split `edx`

The *Evaluation Set* `final_holdout_test` should only be used for the evaluation of the final model. It is thus necessary to further split the *Input Set* `edx` into a *train set* and a *test set* that can be used for testing initial or partial models. The *train set* and *test set* will be *90%* a *10%* of the *Input Set*, respectively.

2.2 Train Set Summary

The *train set* has 8100050 rows and the *test set* has 900005 rows.

Getting the overall average or μ of the *train set* alongside the minimum and maximum values

Average	Minimum	Maximum
3.512484	0.5	5

2.3 Initial Model based on Average

The average of all the ratings can be used as a simple predictor of ratings. It can also serve as the basis for subsequent models.

This model can be represented by the equation
$Y = \mu$
Where Y is the predicted rating and
μ is the overall average

Using the previously obtained $\mu = 3.5124842$, the model evaluated using the *test set* resulted to an *RMSE* is 1.0604198.

After each model, showing all constructed models in a table will be useful for comparison and choosing the final model.

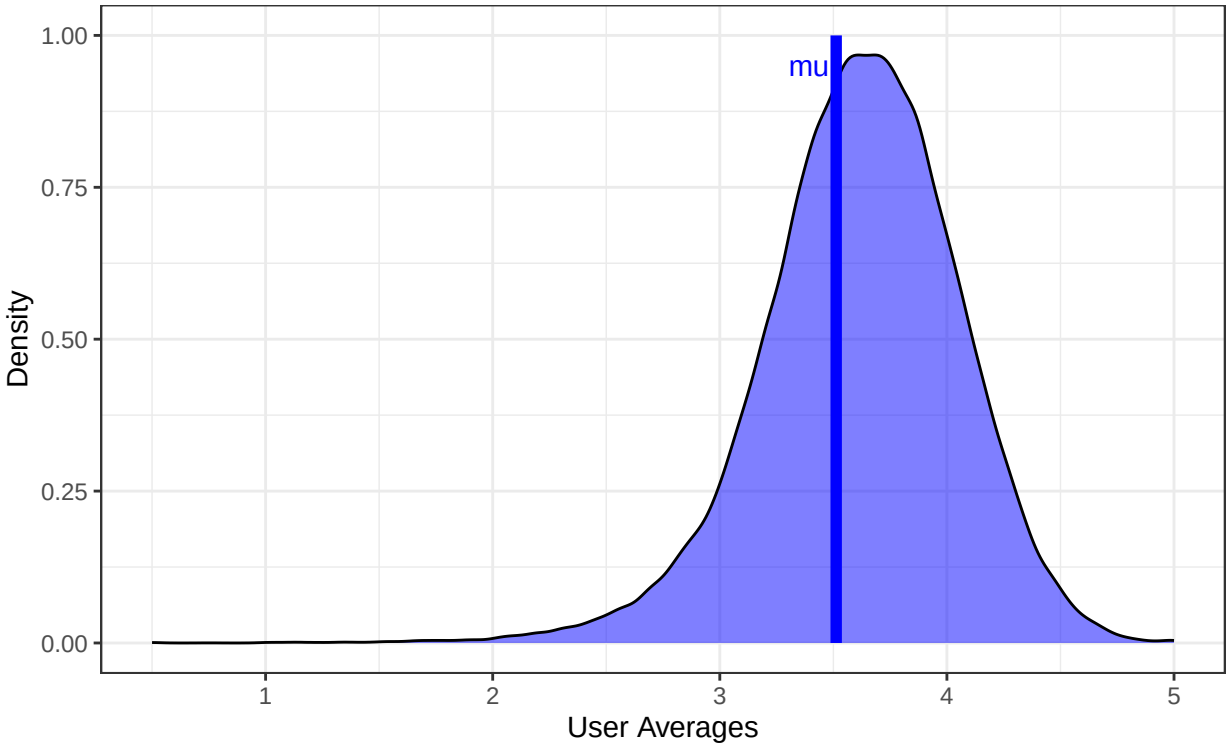
Model	rmse
Overall Average Model	1.06042

2.4 User Bias Model

Each user likely has some biases and tendencies for rating a movie. This bias can be called the *User Bias*. A model can be constructed by taking into account these biases.

2.4.1 User Average

Looking at the density plot of the user averages:



It can be seen that the user averages are either above or below μ. This means that the user average μ_u can be represented as:
$\mu_u = \mu + r_u$
Where r_u is distance of the user average μ_u from μ or *residual* .

2.4.2 User Residual

We can get the residual r_u by subtracting each rating by μ before getting the average.

$r_u = mean(rating - \mu)$

Plotting a density plot of residual r_u:

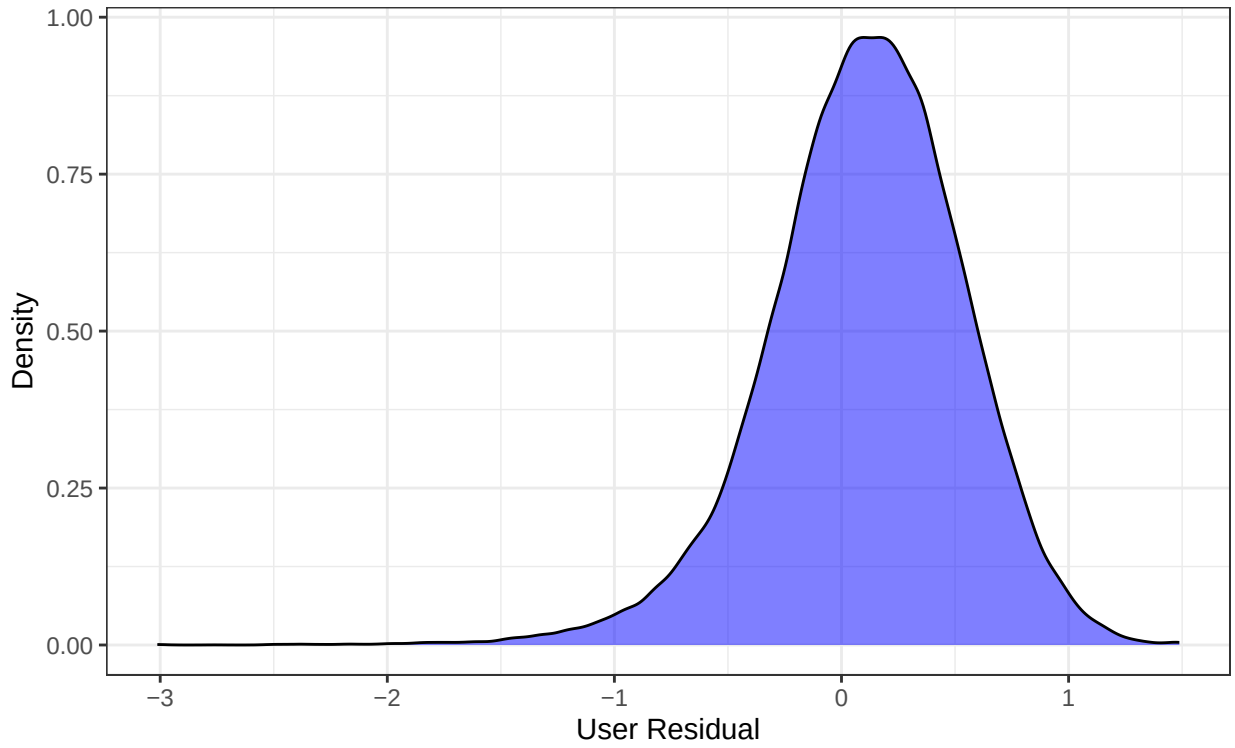


Figure 1: Density Plot of user bias

Looking at the plot, the residual contains 0 and goes from negative to positive value. This makes the residual a better representation of bias. Negative residual value represents negative bias. While, Positive residual value represents positive bias.

2.4.3 Initial Test of the User Bias Model

The residual r_u can now be used to form the next model. The new model based on user bias can be represented by:
$Y = \mu + b_u$
Where b_u is the User Bias.

Plotting the Model results' *predicted ratings* vs *ratings*

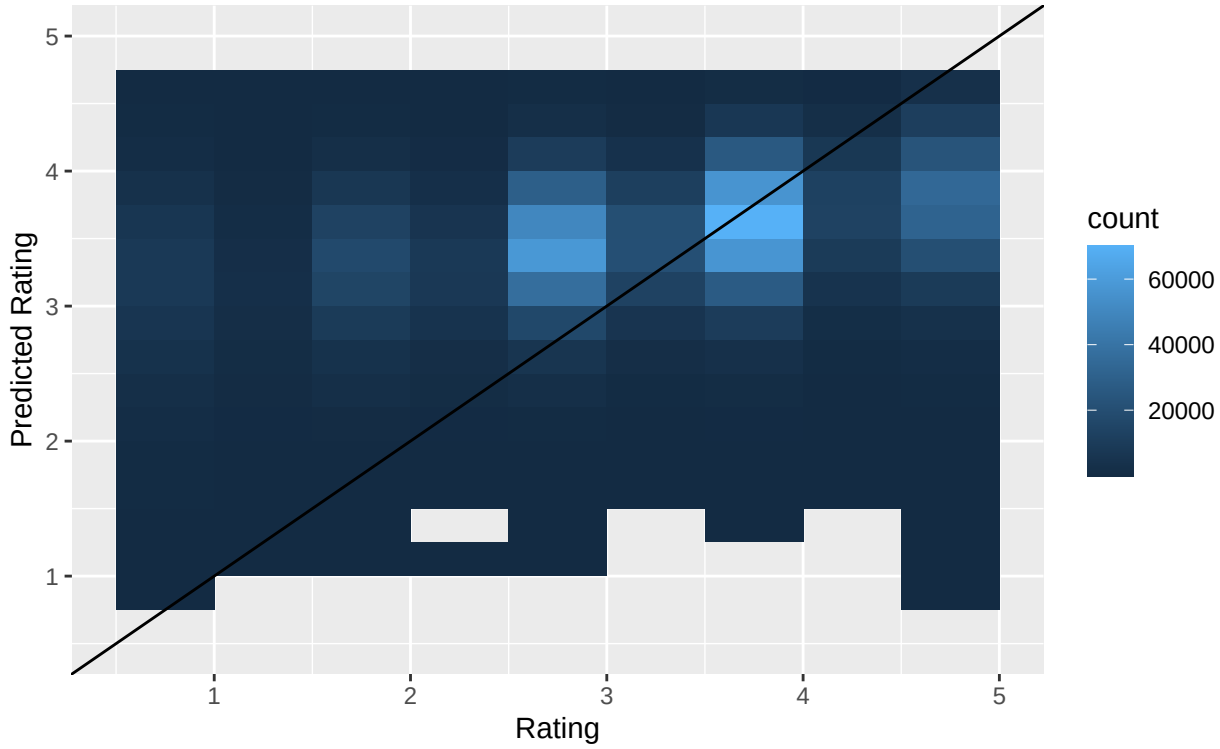


Figure 2: Relationship bet. Predicted and Actual Ratings of the User Bias Model

The *User Bias* model resulted in an RMSE of 0.9782389. Also, the results plot shows that a large part of the results hover around the identity line. This means that, alongside the *RMSE*, the *User Bias* model is accurate enough to predict ratings.

Comparing the User Bias model with the initial model:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389

2.4.4 User Bias Regularization

The results from the *User Bias* Model can be further explored to look for ways of reducing *RMSE*. To approximate the effect on *RMSE*, use the absolute difference between *predicted* and *actual* ratings:
$difference = |predicted - actual|$

Looking at the results:

Table 5: Results with the highest difference

userId	movieId	user_residual	rating	predicted_rating	difference	total
49082	778	1.3473288	0.5	4.859813	4.359813	107
25041	19	1.2375158	0.5	4.750000	4.250000	28
23703	3911	1.1958491	0.5	4.708333	4.208333	96
61700	40851	1.1405008	0.5	4.652985	4.152985	134
15922	1090	1.0360006	0.5	4.548485	4.048485	165
69556	59615	1.0228099	0.5	4.535294	4.035294	85
13438	597	-2.5124842	5.0	1.000000	4.000000	19
17965	57368	0.9875158	0.5	4.500000	4.000000	30
61093	181	1.4875158	1.0	5.000000	4.000000	16
61093	3101	1.4875158	1.0	5.000000	4.000000	16

Most of the observations with highest difference from the *predicted* rating and the *actual* rating have user's total ratings less than 100.

Also :

Table 6: Results with the highest difference and total < 50

userId	movieId	user_residual	rating	predicted_rating	difference	total
25041	19	1.2375158	0.5	4.750000	4.250000	28
13438	597	-2.5124842	5.0	1.000000	4.000000	19
17965	57368	0.9875158	0.5	4.500000	4.000000	30
61093	181	1.4875158	1.0	5.000000	4.000000	16
61093	3101	1.4875158	1.0	5.000000	4.000000	16
47753	1673	0.9682850	0.5	4.480769	3.980769	26
35677	1342	0.9647885	0.5	4.477273	3.977273	22
52882	44397	0.9236860	0.5	4.436170	3.936170	47
70667	2395	0.9160872	0.5	4.428571	3.928571	14
45308	2858	0.9143450	0.5	4.426829	3.926829	41

Using `difference`, visualize and plot the relationship between total number of ratings, user bias and the difference:

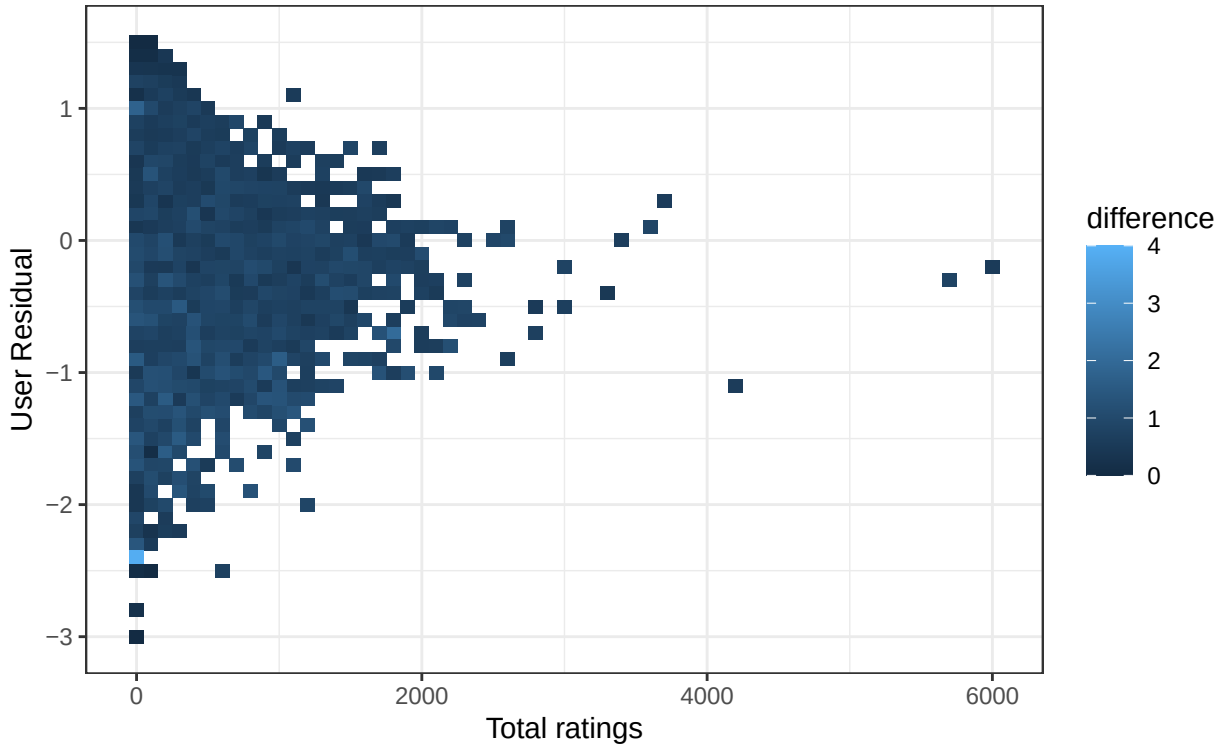


Figure 3: Plot showing the relationship between total ratings, user residual, and difference

This plot shows that the residual value from users with low number of total ratings has a large range of values that might introduce larger errors. **Regularization** can be introduced to lower this range.

To regularize the residual a constant `k` could be added to the divisor when getting the average of the residuals:
From
$r_u = \sum_{i=1}^{n} \frac{rating-\mu}{n}$,
to the
$r_u = \sum_{i=1}^{n} \frac{rating-\mu}{n+k}$

Where `n` is the total number of ratings by the user
And `k` is an adjustable parameter for regularization.

This formula minimizes r_u if the total ratings **n** is low. For larger **n** the effect of k on r_u is low.

Using an initial value $k = 5$, the RMSE of the model is 0.9777804

Plotting the regularized user bias:

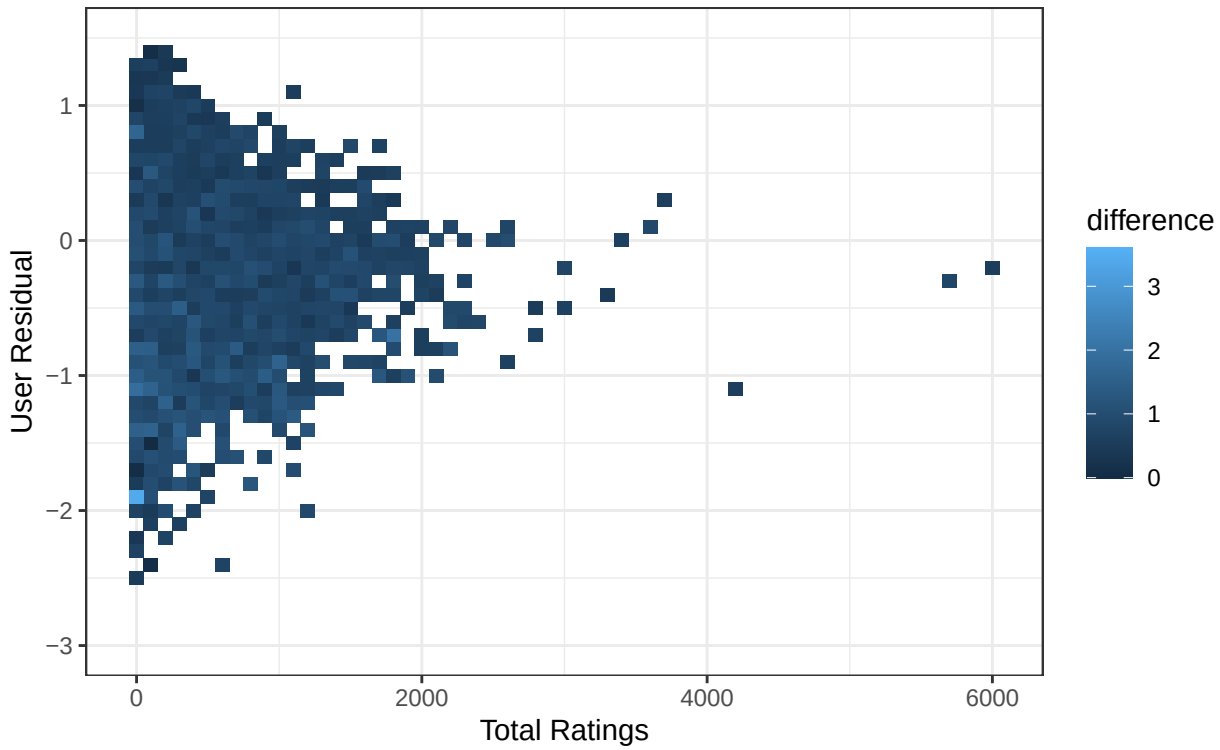


Figure 4: Plot showing the relationship between total ratings, user residual, and difference with Regularization

Comparing the plots of User Bias with and without Regularization :

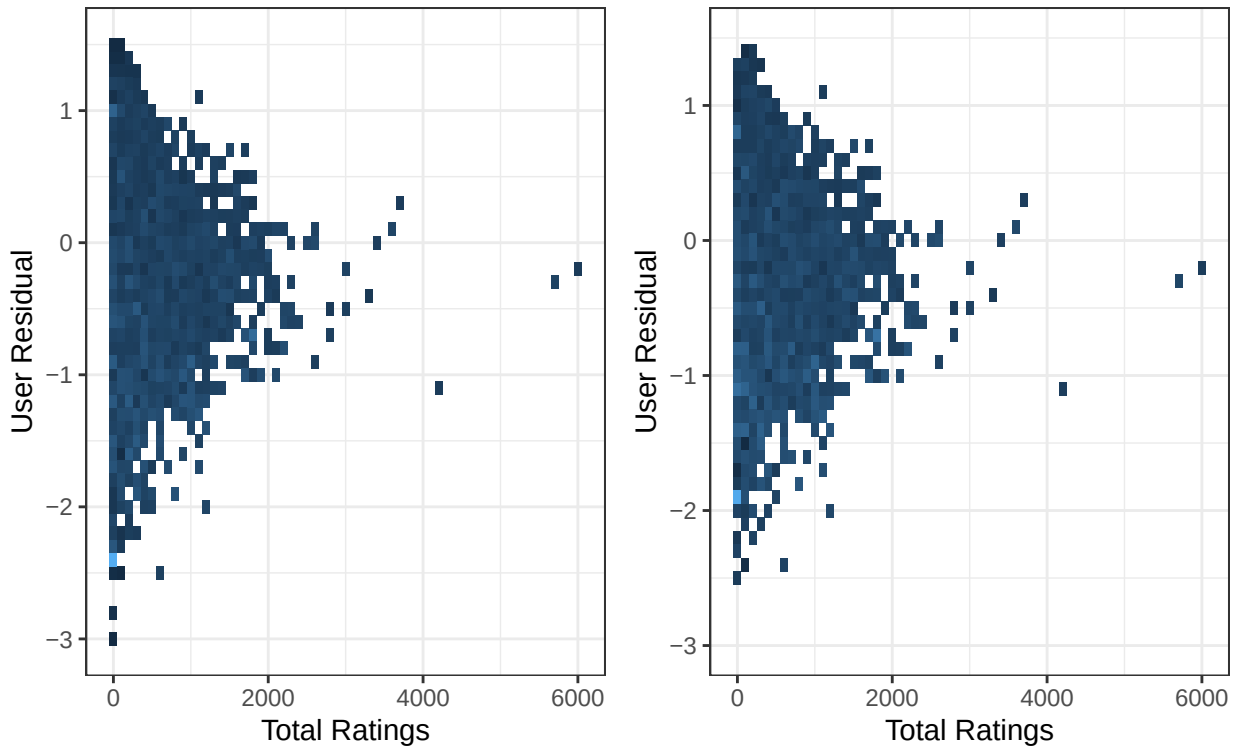


Figure 5: Combined Plot without Regularization(Left) and with Regularization(Right)

The 2 plots above shows that Regularization can minimize the residuals of users with low total number of ratings.

The parameter `k` may still be improved to get the minimum RMSE possible for User Bias:

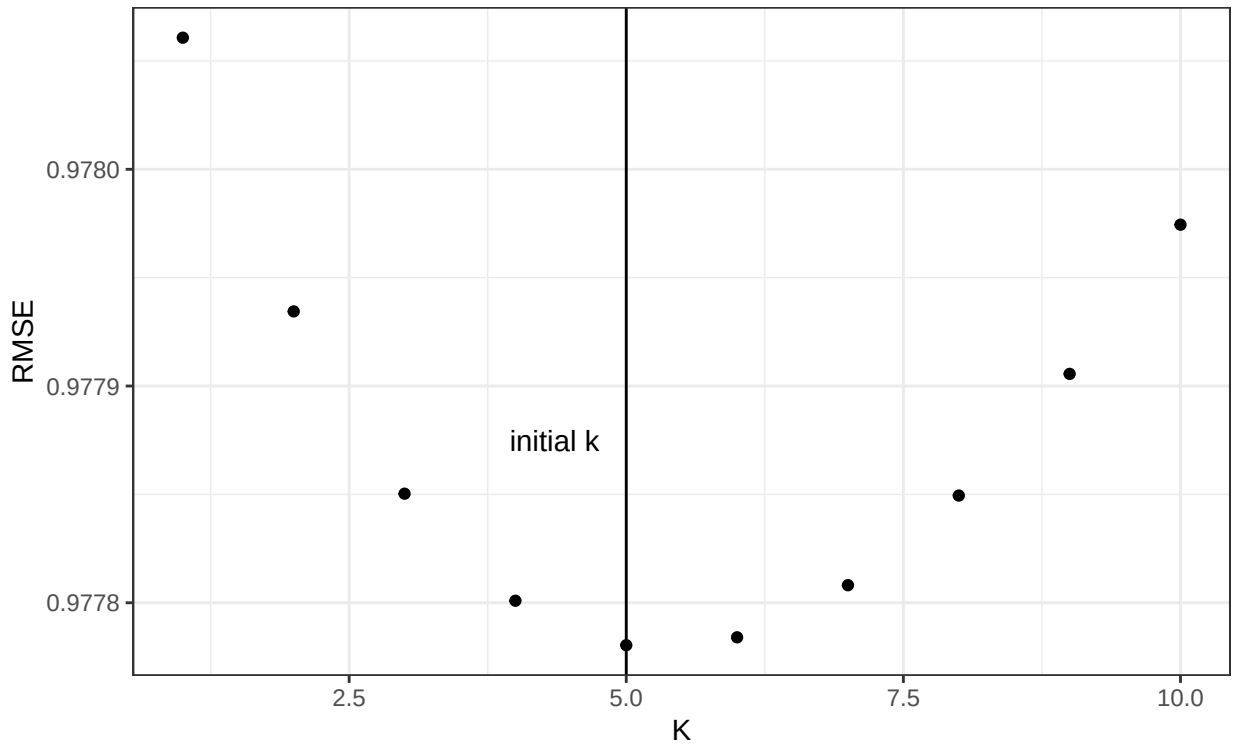


Figure 6: Plot of the resulting RMSE for different parameter k

The results show that the lowest RMSE is 0.9777804 at k of 5.

Plotting the results:

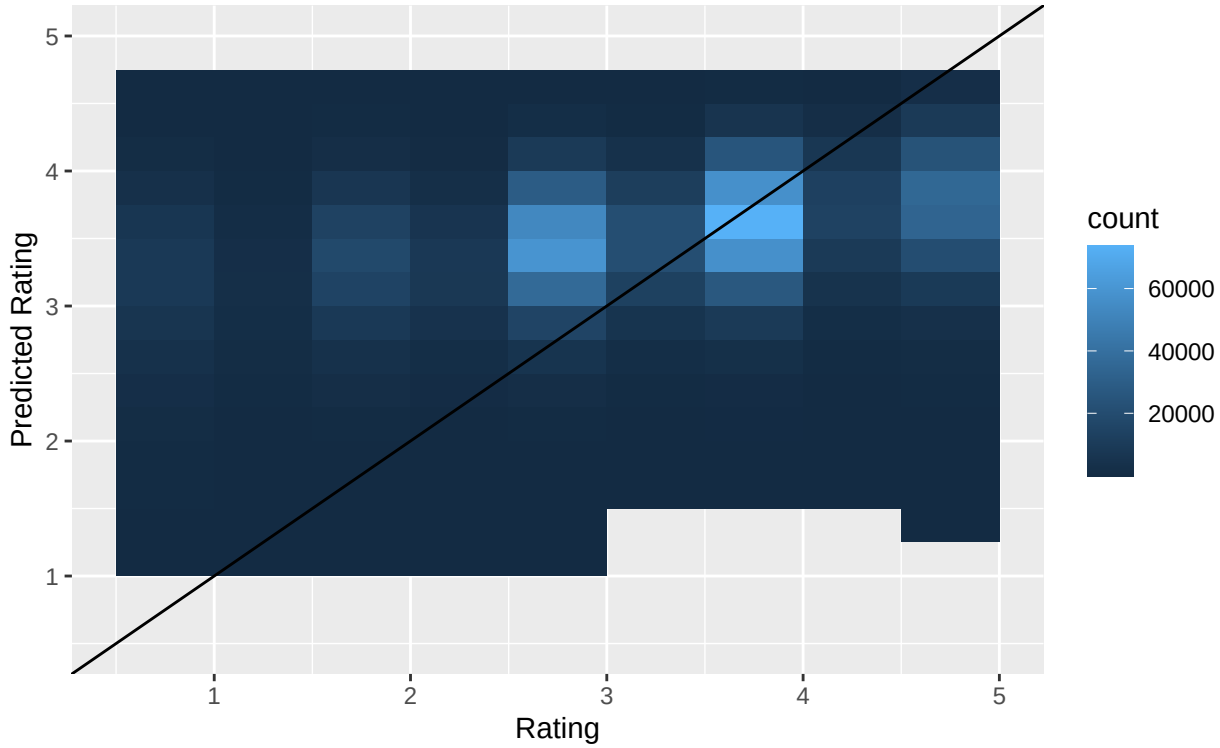


Figure 7: Relationship bet. Predicted and Actual Ratings of the User Bias Model with Regularization

Showing the current table of the current models:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804

The results show that the *User Bias* Model has a lower *RMSE* than the base model. And the regularization resulted to a reduction in *RMSE*

2.5 Movie Bias Model

Movies are generally have either a positive or a negative reception to the audiences. This results in either a below average or above average movie rating. *Good* movies will have ratings closer to the maximum rating, and *Bad* movies will have ratings closer to the minimum rating. This factor can be called the *Movie Bias*, and it can also be used to predict the ratings.

2.5.1 Movie Average

The Movie Bias can be represented by the average rating each movie received.

Plotting the density plot of the Movie Averages:

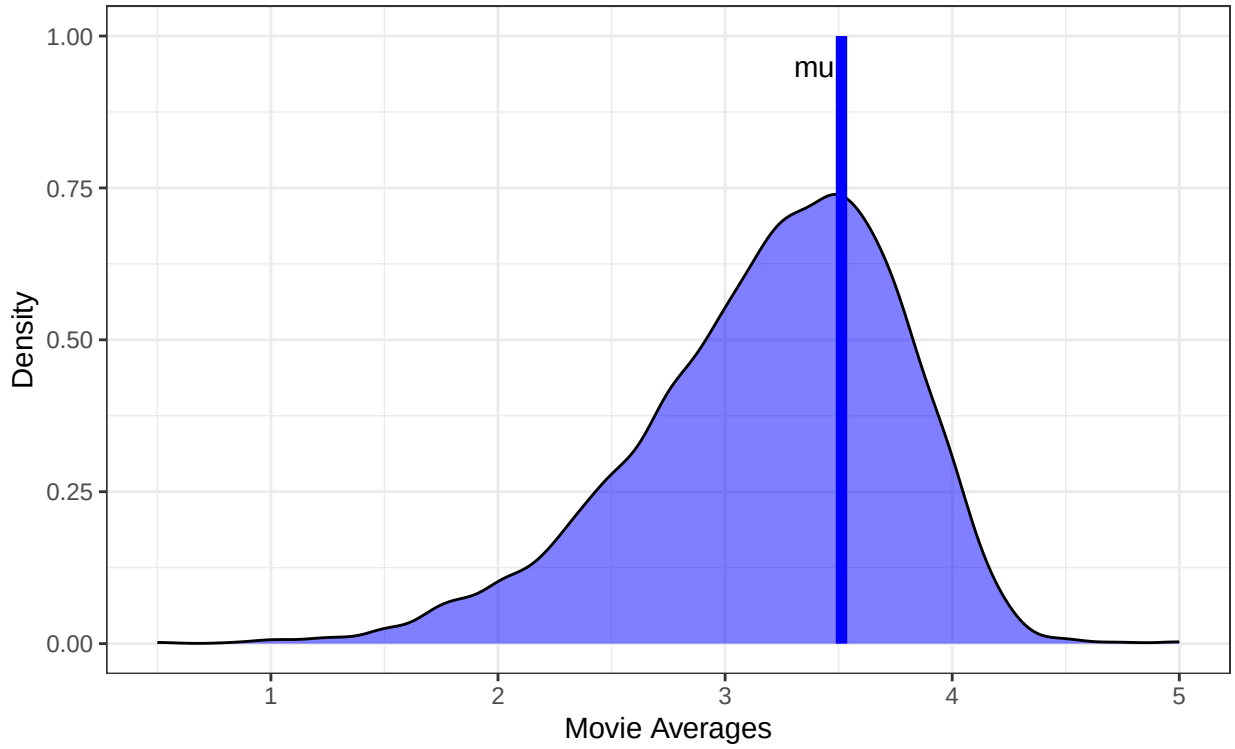


Figure 8: Density Plot of Movie Average

2.5.2 Movie Residual

As with user bias, the movie average μ_m can also be presented as:
$\mu_m = \mu + r_m$
Where r_m is the residual.

Like the User residual, the movie residual r_m can be obtained by using:
$r_m = mean(rating - \mu)$

Plotting the density plot of the movie residual r_m:

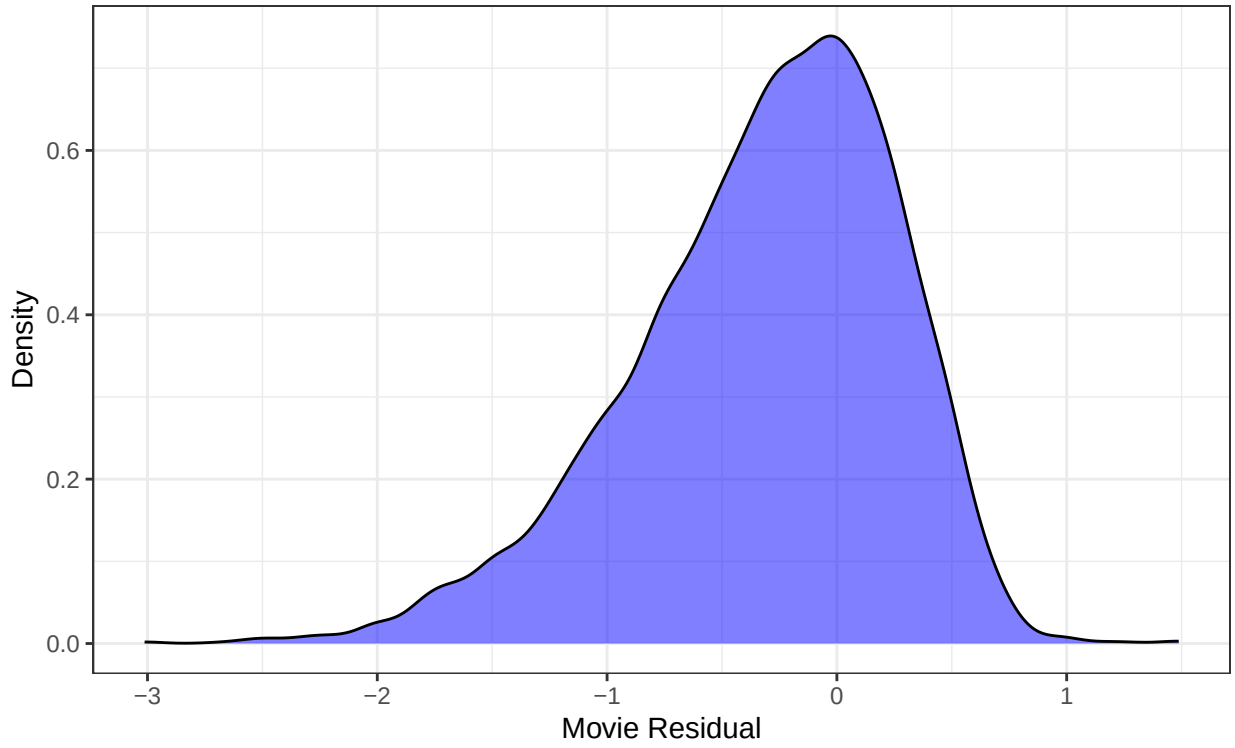


Figure 9: Density Plot of User Residual

The plot shows that the residual contains 0 and goes from negative to positive values. Like in the *User Bias*, the movie residual is a better representation of the *Movie Bias*. Positive residual value means positive bias. While, Negative residual value means negative bias.

2.5.3 Movie Bias Model

The residual r_m can now be used to form the next model. The new model based on movie bias can be represented by:
$Y = \mu + b_m$
Where b_m is the Movie Bias.

Using the Movie Bias model:

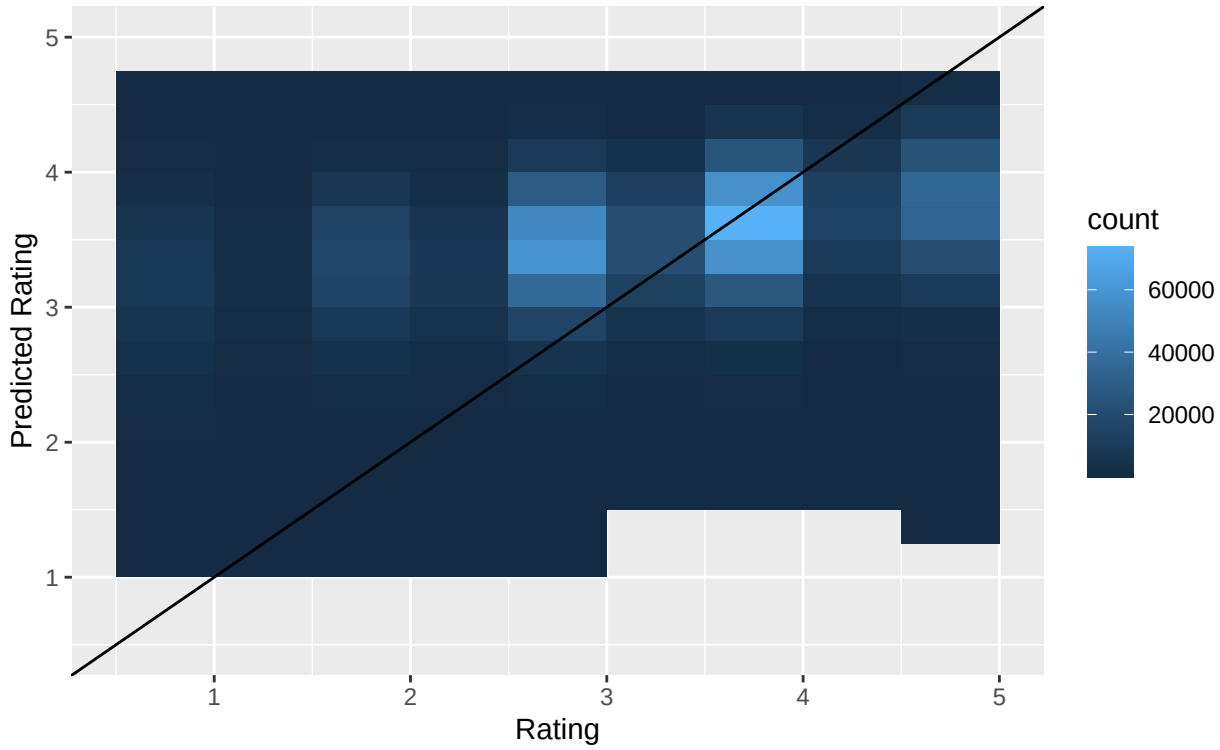


Figure 10: Relationship bet. Predicted and Actual Ratings of the Movie Bias Model

The Resulting RMSE of the *Movie Bias* model is 0.9439072.

Showing the Models table:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804
Movie Bias	0.9439072

2.5.4 Movie Bias Regularization

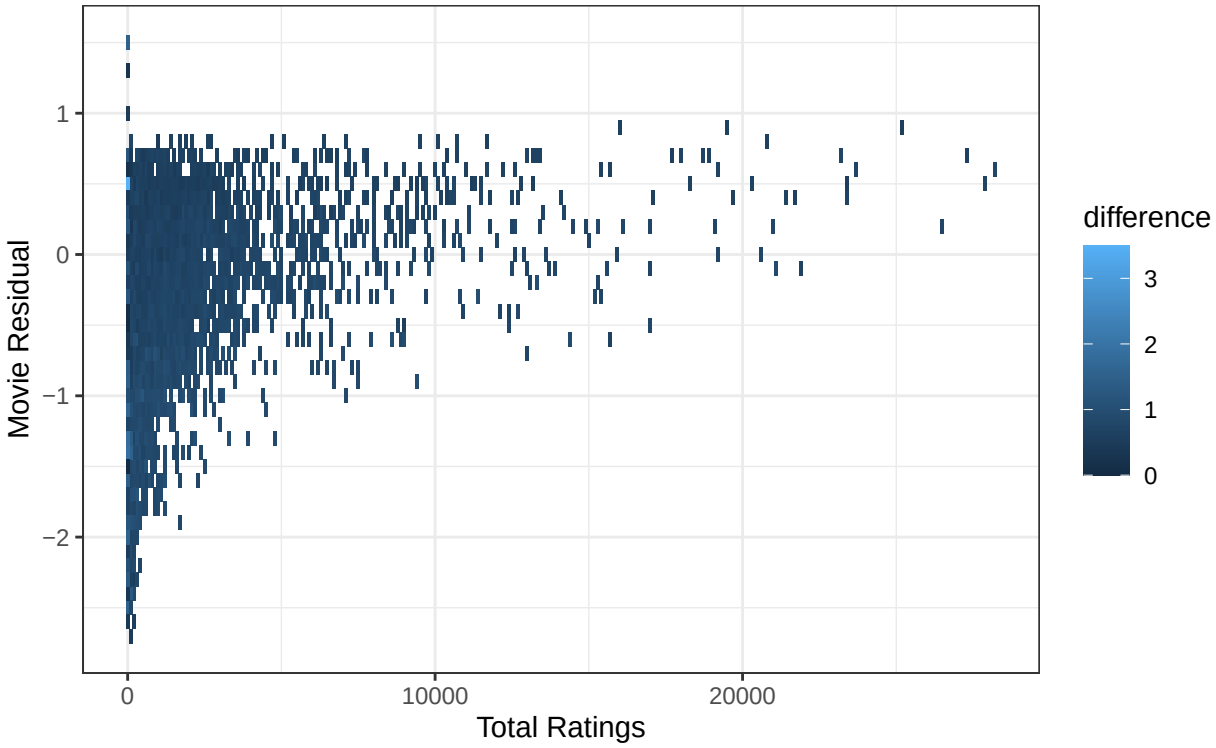
Like the User Bias Model, the results from the Movie Bias Model can be further explored to look for ways of reducing *RMSE*. To approximate the effect on *RMSE*, use the absolute difference between predicted and actual ratings:
$difference = |predicted - actual|$

Looking at the results:

Table 9: Results with the highest difference

userId	movieId	movie_residual	rating	difference	total
5279	6371	-2.4921590	5.0	3.979675	123
9568	318	0.9405396	0.5	3.953024	25183
10680	318	0.9405396	0.5	3.953024	25183
25239	318	0.9405396	0.5	3.953024	25183
26260	318	0.9405396	0.5	3.953024	25183
68400	318	0.9405396	0.5	3.953024	25183
17471	858	0.9015343	0.5	3.914019	15986
19609	858	0.9015343	0.5	3.914019	15986
7510	50	0.8559840	0.5	3.868468	19501
50052	50	0.8559840	0.5	3.868468	19501
62006	50	0.8559840	0.5	3.868468	19501

Using `difference`, Plot the relationship between total number of ratings, movie bias and the difference:



The results and the plot shows that the residual values from movies with lower total number of ratings have a large range of values that might introduce larger errors. Regularization can be introduced to lower this range

To regularize the residual a constant `k` could be added to the divisor when getting the average of the residuals:
From
$r_m = \sum_{i=1}^{n} \frac{rating-\mu}{n}$,
to the
$r_m = \sum_{i=1}^{n} \frac{rating-\mu}{n+k}$

Where `n` is the total number of ratings of the movie.
And `k` is an adjustable parameter for regularization.

This formula minimizes r_m if the total ratings **n** is low. For larger **n** the effect of k on r_m is low.

Using an initial value $k = 5$, the RMSE of the model is 0.9438815.

Plotting the relationship between total number of ratings, movie bias and the difference with Regularization:

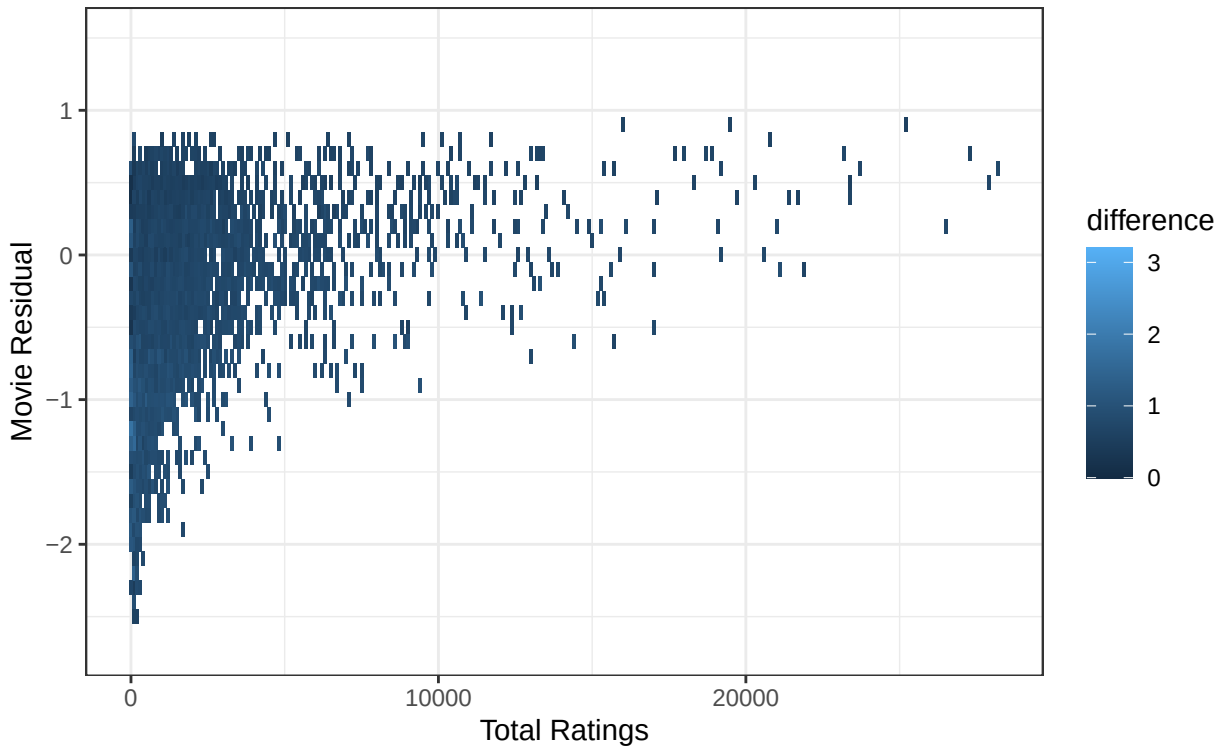


Figure 11: Plot showing the relationship between total ratings, movie residual and difference with Regularization

Comparing the plot of Movie Bias with and without Regularization :

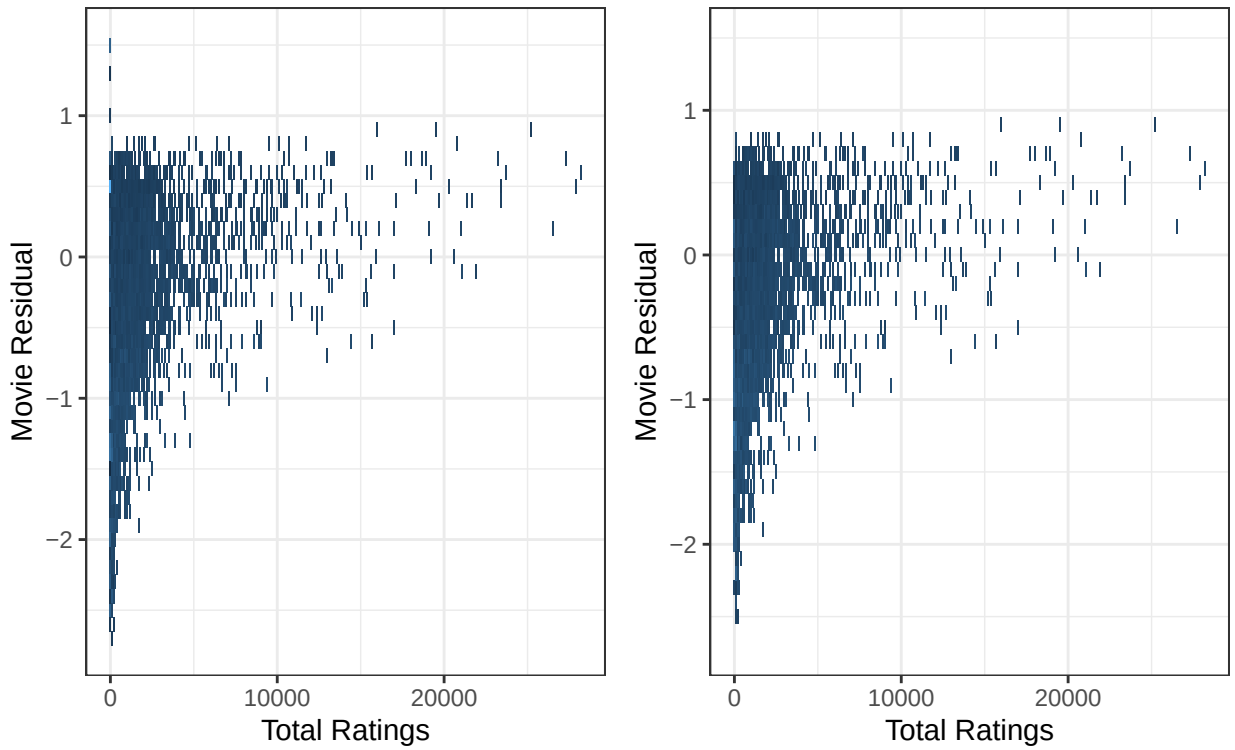
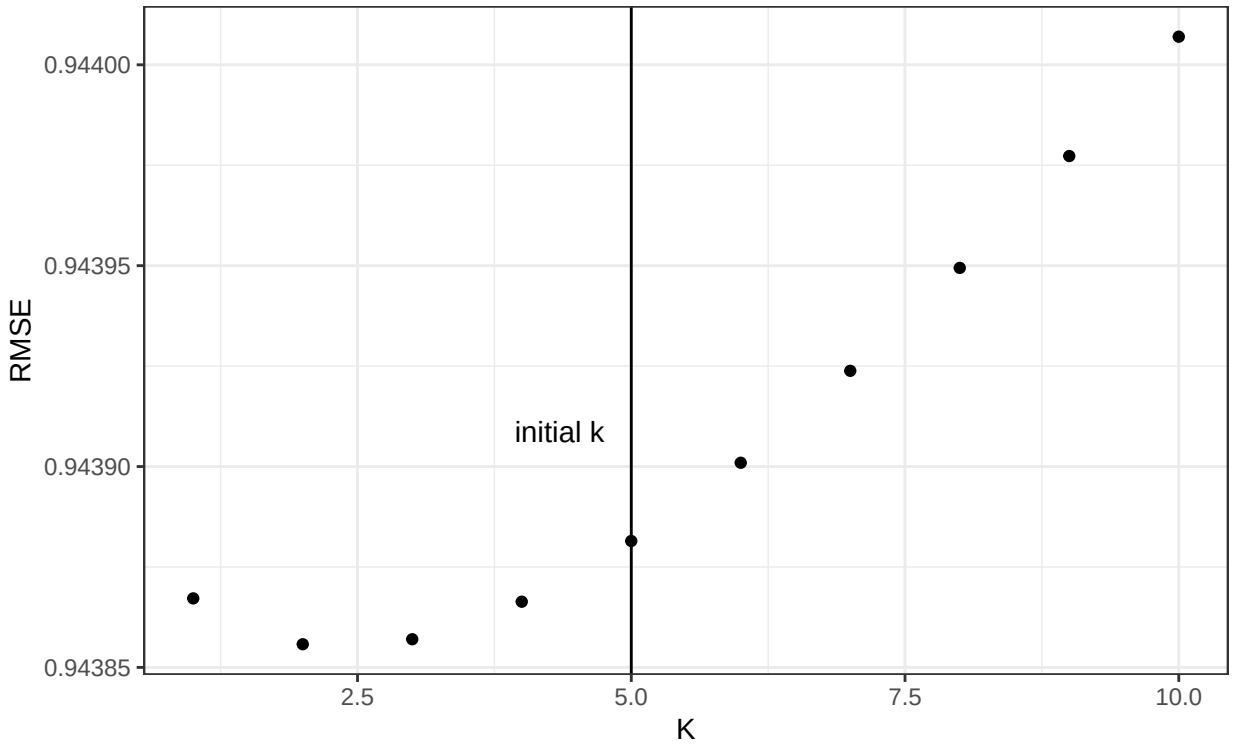


Figure 12: Combined relationship plot without Regularization(Left) and with Regularization(Right)

Regularization minimized the residuals of movies with lower total number of ratings, while keeping the residuals of movies with higher number of ratings almost the same.

As with the *User Bias* the `k` for *Movie Bias* can still be optimized to get better RMSE.



The results show that the lowest RMSE is 0.9438558 at k of 2.

Plotting the results:

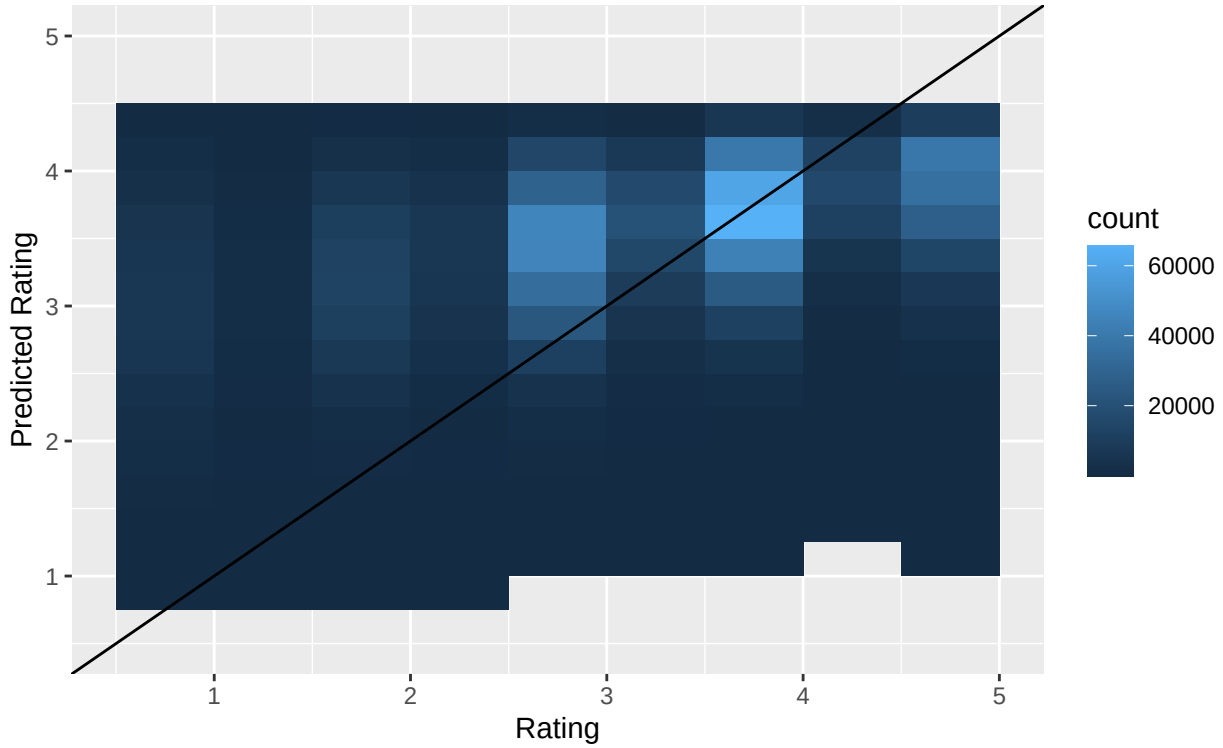


Figure 13: Relationship bet. Predicted and Actual Ratings of the User Bias Model

Showing the current table of models:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804
Movie Bias	0.9439072
Movie Bias w/ Regularization	0.9438557

The results show that both *User Bias* and *Movie Bias* has a lower RMSE than the initial model. Also, Regularization resulted to some reduction in RMSE in both biases.

2.6 Combine User and Movie Model

The User and Movie Bias model can now be joined to look at how it influences the RMSE. Also, it could show the effect of combining Biases into one model.

The two previous models can be combined to get a model represented by:
$Y = \mu + b_{user} + b_{movie}$
Where b_{user} is the user bias
And b_{movie} is the movie bias

Plotting the results of the *Movie + User Bias Model* without Regularization:

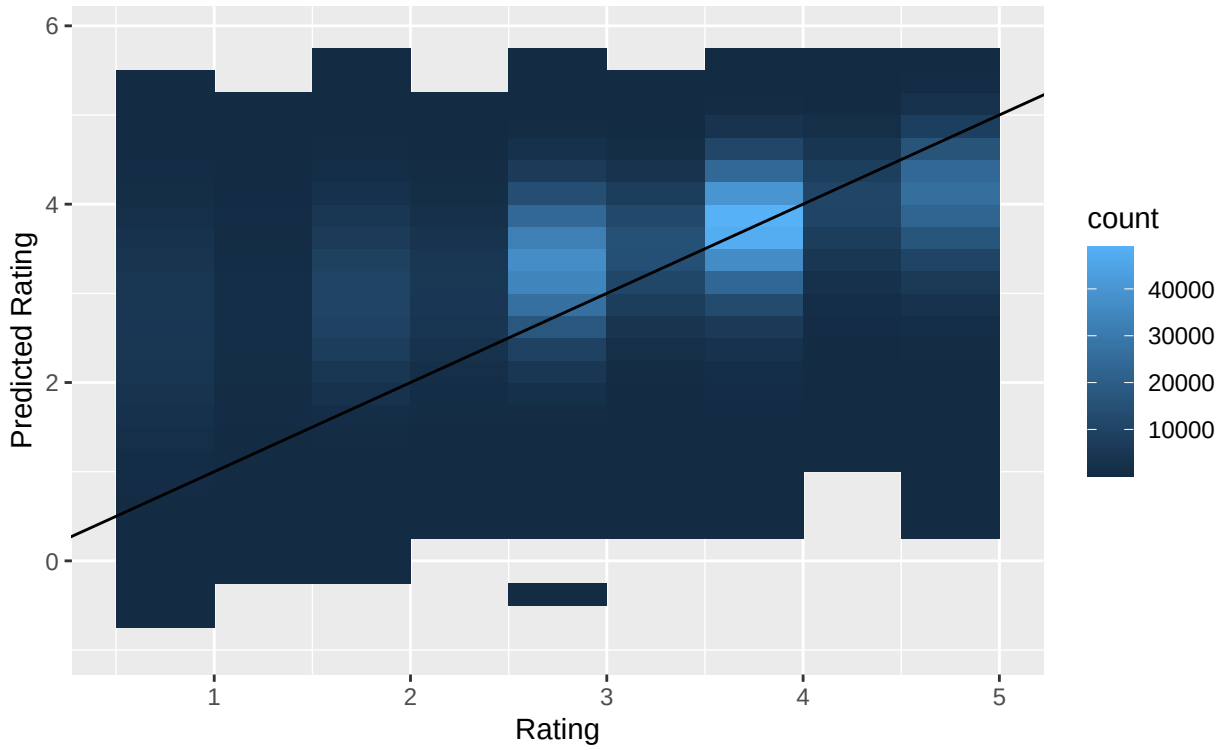


Figure 14: Relationship bet. Predicted and Actual Ratings of the Movie + User Bias Model

Plotting the results of the *Movie + User Bias Model* with Regularization:

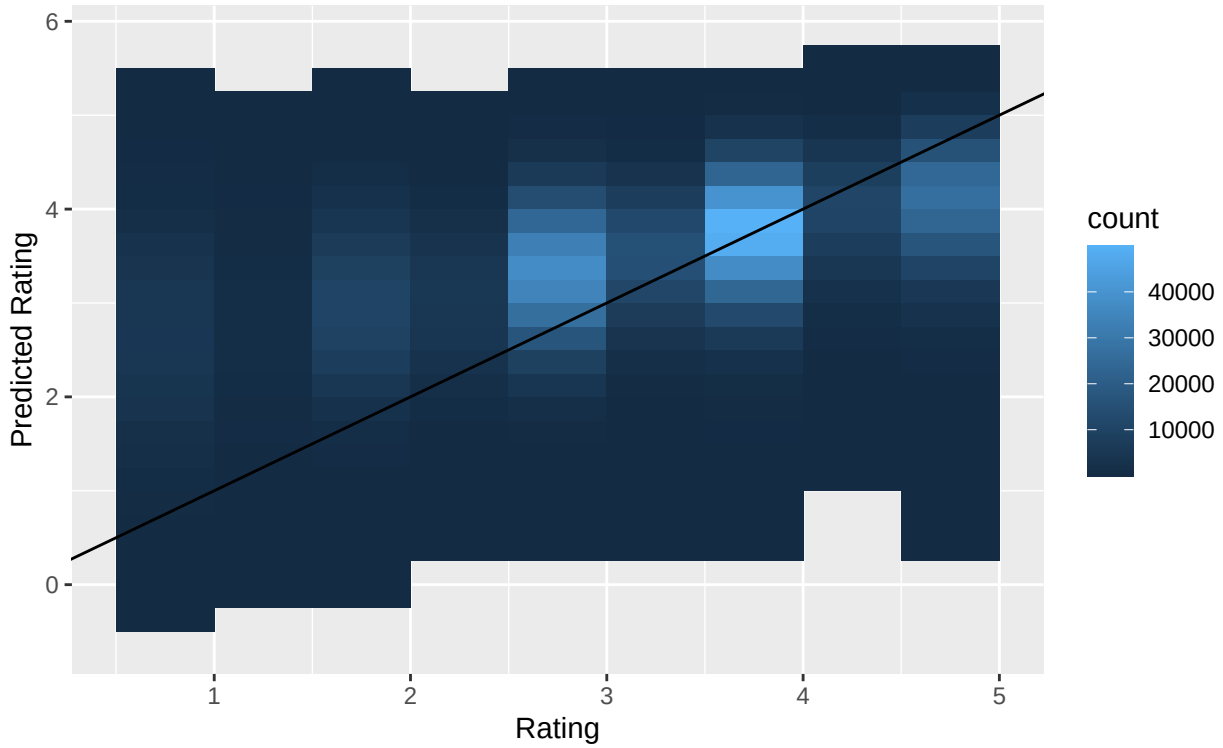


Figure 15: Relationship bet. Predicted and Actual Ratings of the Movie + User Bias Model with Regularization

The RMSE of the combined model without Regularization is 0.8854581. While, the RMSE of the combined model with Regularization is 0.8834019.

Showing the current table of models:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804
Movie Bias	0.9439072
Movie Bias w/ Regularization	0.9438557
Movie Bias + User Bias	0.8854581
Movie Bias + User Bias w/ Regularization	0.8834019

Notice that the combined model of the *User Bias* and *Movie Bias*, resulted in a far lower *RMSE* than the previous models. And the combined model with Regularization still has an *RMSE* lower than the model without Regularization.
This means that combining different biases can reduce *RMSE*, and that combined models are still influenced by Regularization.

2.7 Genre Bias

The genres of movies could have some effect on the ratings. Also, users most likely has a preference for some genres of movies over other genres. It is worth exploring the effect of genre preference on each users rating.

To extract the details for each genre, a *regex* pattern of each genre is used. The pattern will be used to detect if the `genres` column/string has that genre.
For example, the genre pattern used for detecting genre Comedy : [a-zA-Z|]*Comedy*[*a-zA-Z*|].

To show how the *regex* patterns can be used, get a summary data of each genre :

Table 12: Summary of Genres with average rating and total number of ratings

genres	average	total
Comedy	3.437398	3187490
Romance	3.554146	1541755
Action	3.421245	2304615
Crime	3.665500	1195646
Thriller	3.507234	2093845
Drama	3.673142	3519264
Sci-Fi	3.395243	1207351
Adventure	3.493778	1717900
Children	3.418868	664205
Fantasy	3.502035	833594
War	3.780294	459944
Animation	3.601325	420561
Musical	3.564145	389789
Western	3.556208	170537
Mystery	3.676628	511672
Film-Noir	4.010691	106636
Horror	3.269543	621937
Documentary	3.784079	83681
IMAX	3.770532	7330
no genres listed	3.642857	7

From the summary, an overview of the average ratings of each genre can be plotted:

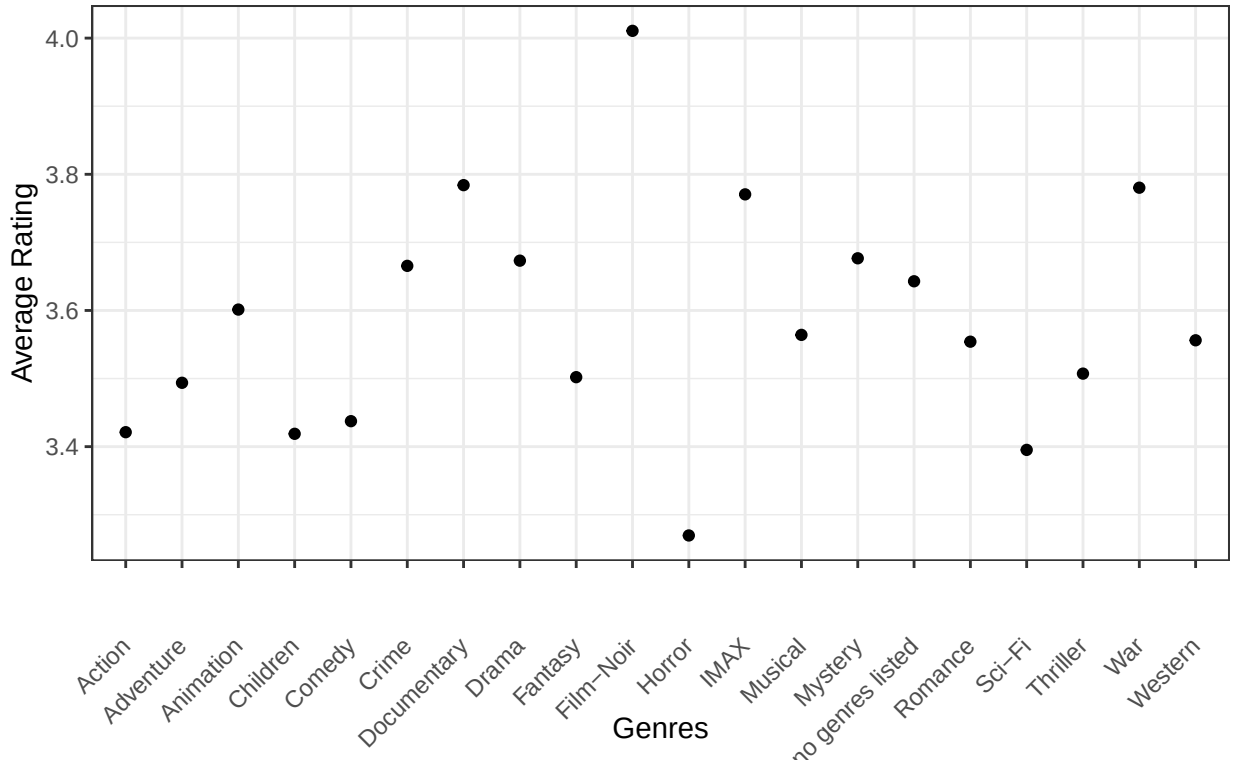


Figure 16: Plot of the average ratings per genre

2.8 User-Genre Bias

Looking at the plot of Genre averages, it can be seen that there are differences in the average ratings of each genre. This shows that *genre* could impact the users' ratings.
Using the formula:
$r_{g,m} = mean(rating - \mu)$
Where $r_{g,m}$ is the genre residual of the genre m.
The genre preference of each user can be extracted.

Showing part of the extracted preference:

Table 13: User preference summary for two genres

userId	Comedy_total	Comedy_residual	Romance_total	Romance_residual
1	9	1.4875158	5	1.4875158
2	4	-0.5124842	3	-0.8458176
3	9	0.1541824	10	0.6375158
4	17	0.1345746	7	0.3446586

Each user will have different residuals representing the difference between preferences.

To see the how the User-Genre Bias will look like, show the entire genre preference of a user:

Table 14: A User's Genre Bias

userId	residual_type	residual_value
2	Comedy	-0.5124842
2	Romance	-0.8458176
2	Action	-0.2902620
2	Crime	0.0000000
2	Thriller	-0.7124842
2	Drama	-0.2624842
2	Sci-Fi	0.4875158
2	Adventure	0.0589443
2	Children	-0.5124842
2	Fantasy	-0.5124842
2	War	0.0000000
2	Animation	0.0000000
2	Musical	-0.5124842
2	Western	1.4875158
2	Mystery	-1.5124842
2	Film-Noir	0.0000000
2	Horror	0.0000000
2	Documentary	0.0000000
2	IMAX	0.0000000
2	no genres listed	0.0000000

Looking at the values of residuals, each user can have a positive or negative bias on a genre. This models the users' preferences. The genres where the user did not have any ratings for are set to a residual value of

0. This residual values can be called the *User-Genre Bias*.

2.8.1 User-Genre Bias Model

The genre preference of each user can be used to make a model the same way the *User Bias* and *Movie Bias* are used to construct the *User Bias* Model and *Movie Bias* Model. However the previously obtained *User-Genre Bias* is a list of users' preference per genre. For each movie the *User-Genre Bias* will vary depending on the movie's genre.

To get the *User-Genre Bias* b_{ug} of a movie m, use the formula:

$b_{g,u,m} = \sum i = 1^n g_{m,i} b_{g,u,i}$
Where **n** is the total number of genres,
$g_{m,i}$ is **1** if movie m is in the genre i, **0** otherwise.
$b_{g,u,i}$ is the genre preference/residual of user u for genre i.

Using the obtained *User-Genre Bias*, a model can be made using the formula:
$Y = \mu + b_{g,u,m}$
Where $b_{g,u,m}$ is the user-genre bias for the movie m

Putting the two formula together, the User-Genre Formula will be:
$Y_m = \mu + \sum_{i=1}^{n} (g_{m,i} b_{g,u,i})$
Plotting the results:

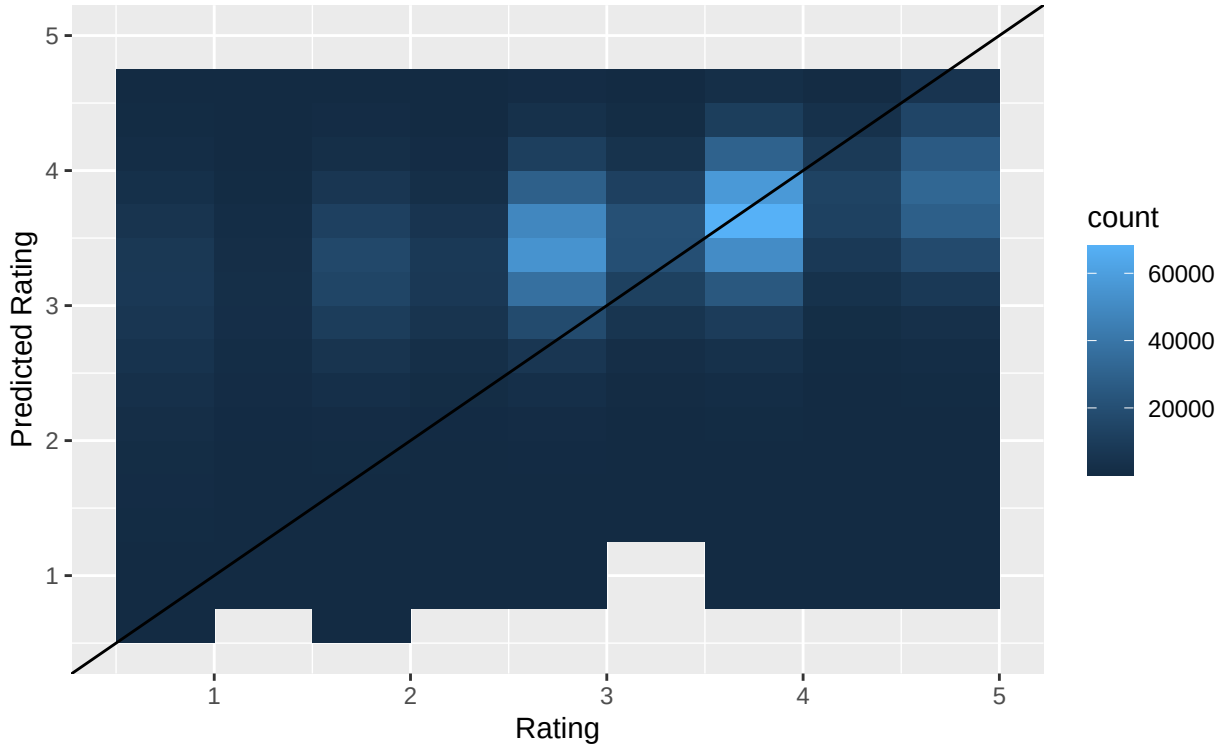


Figure 17: Relationship bet. Predicted and Actual Ratings of User-Genre Bias Model

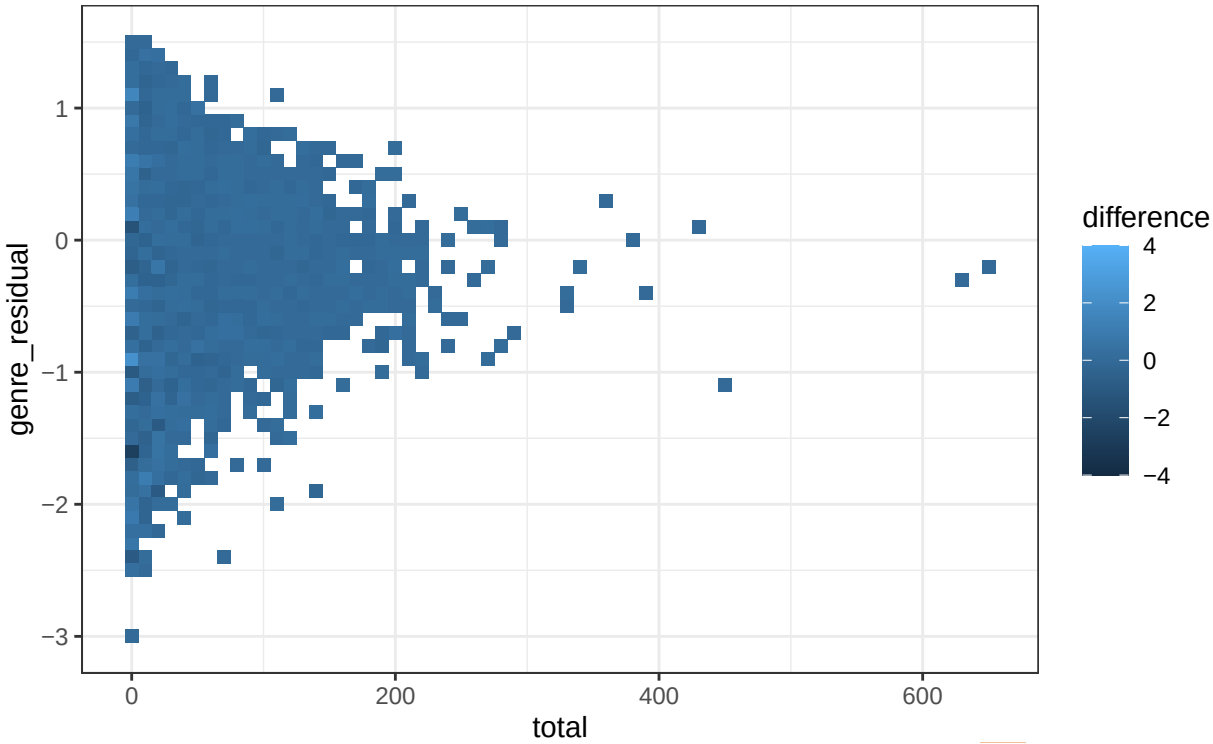
The model made using the *User-Genre Bias* has an *RMSE* of 0.9532481.

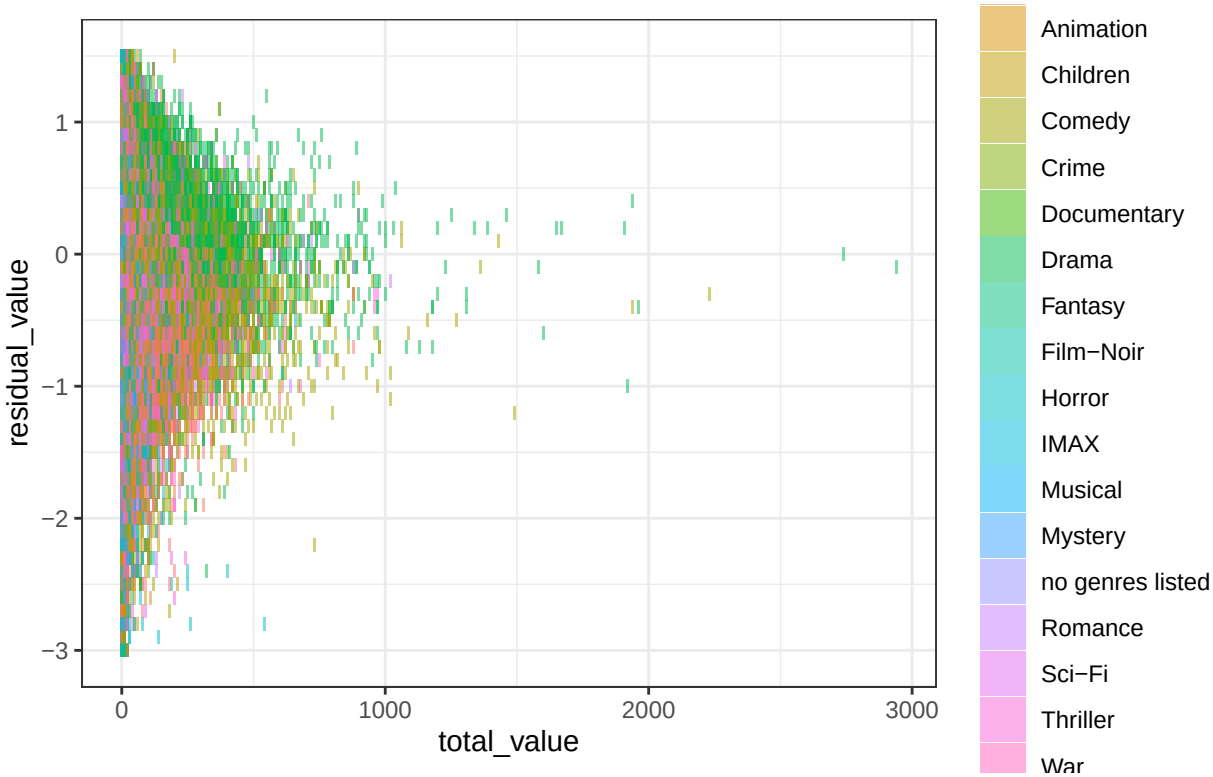
Showing the Model List for comparison:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804
Movie Bias	0.9439072
Movie Bias w/ Regularization	0.9438557
Movie Bias + User Bias	0.8854581
Movie Bias + User Bias w/ Regularization	0.8834019
User-Genre Bias Model	0.9532481

2.8.2 User-Genre Bias Regularization

From the previous biases, Regularization helped reduce RMSE by reducing error at lower *total ratings*. To check if regularization can help reduce RMSE for *User-Genre Bias*, plot the user-genre bias against total ratings:





The two plot looks the same as with user and movie bias, where low total ratings have a large range of values that may introduce larger errors.
Using the same concept of regularization, the formula for the genre residual r_g of each genre i will be :
$r_{gi} = \sum_1^n \frac{rating - \mu}{n+k}$

Using an initial `k = 5`, the resulting RMSE is 0.9547064.

Plotting the results for regularization:

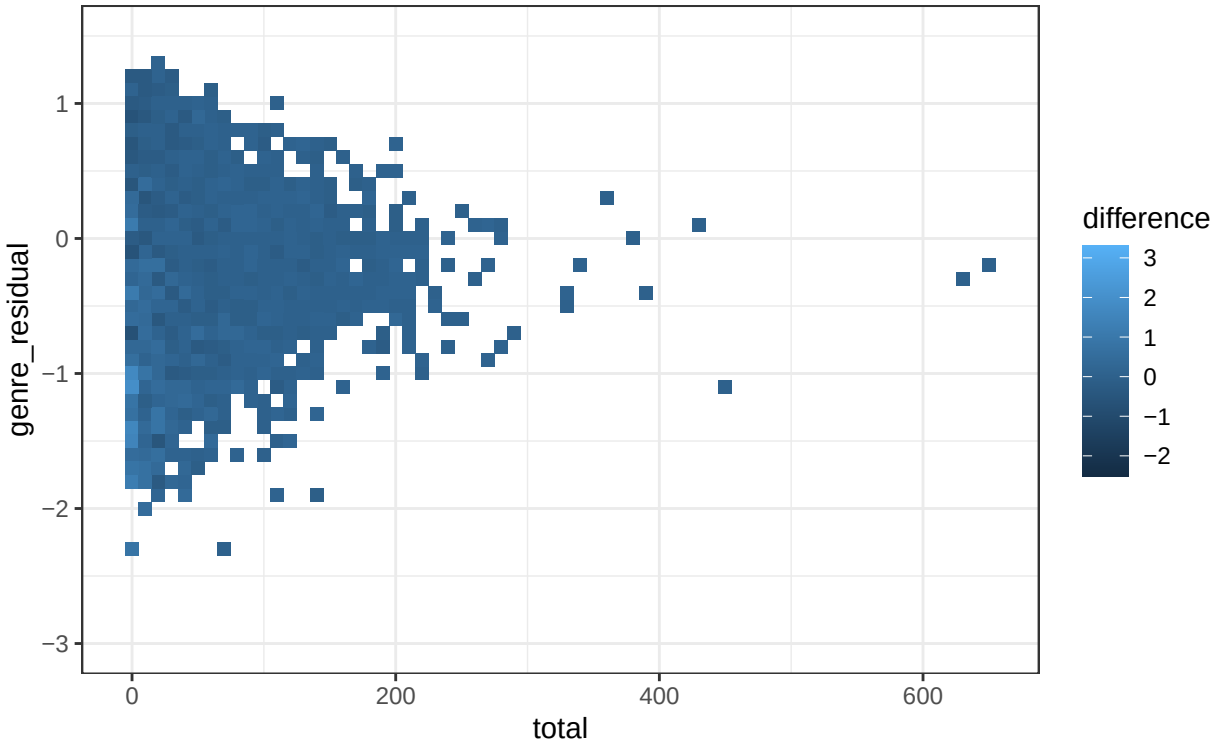


Figure 18: Combined Genre Residual vs. total number of ratings by user

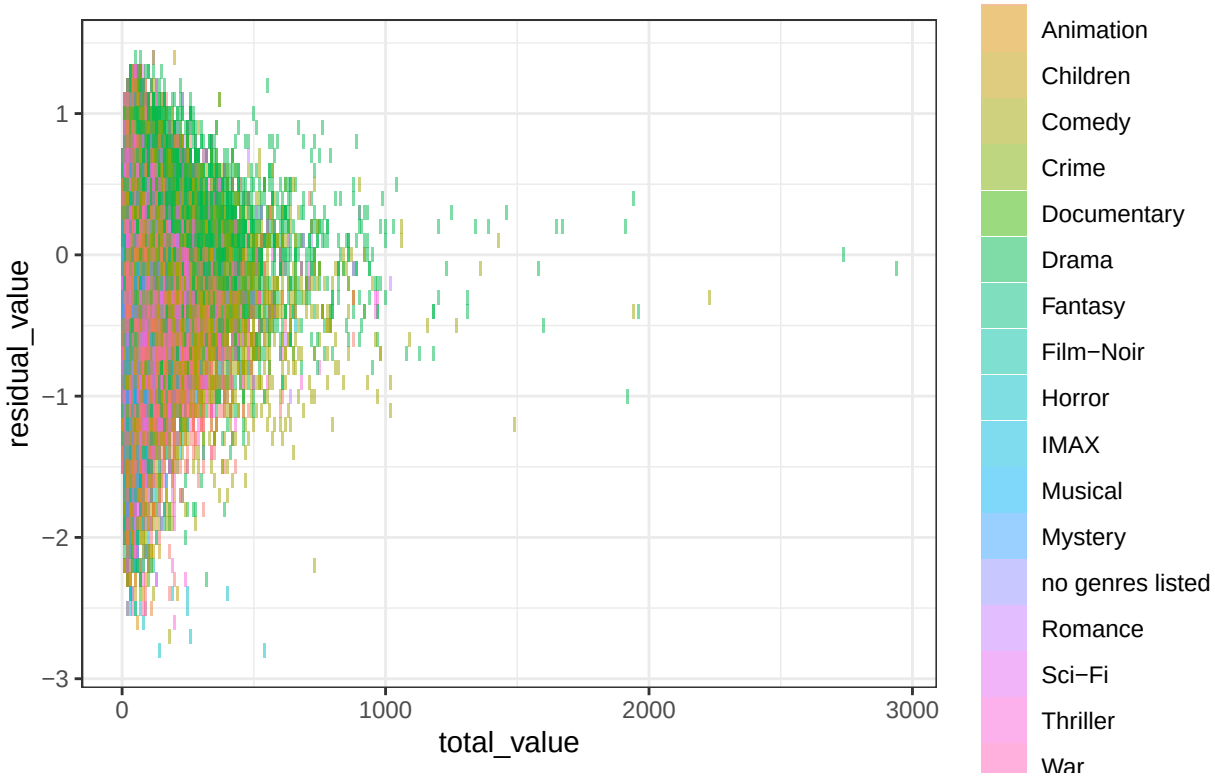


Figure 19: Per Genre Residual vs. total number of ratings by user

Comparing the current results to the results without regularization:

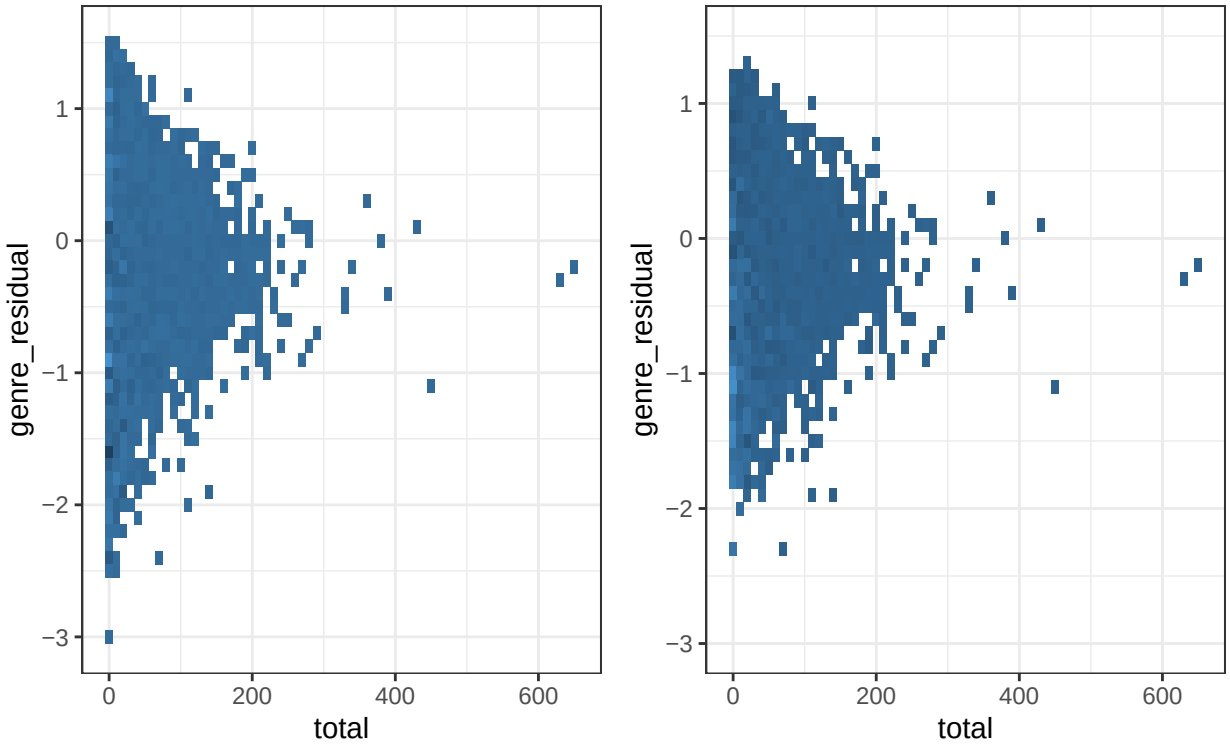
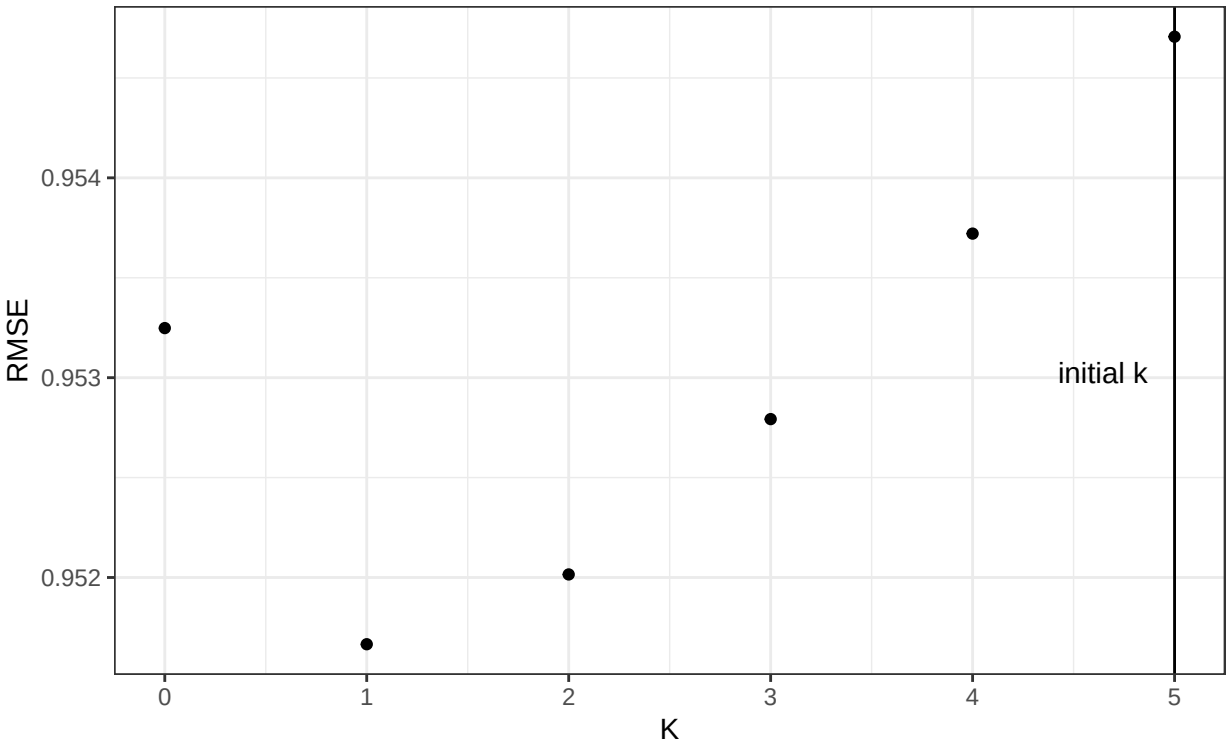


Figure 20: Combined User-Genre Residual without Regularization (Left) and with Regularization (Right)

The regularized residual plot shows a smaller range than without regularization.

However, the k parameter can still be tuned.

Plotting the relationship bet. K and $RMSE$:



The plot shows that the lowest RMSE is 0.951666 at k of 1.

Plotting the results:

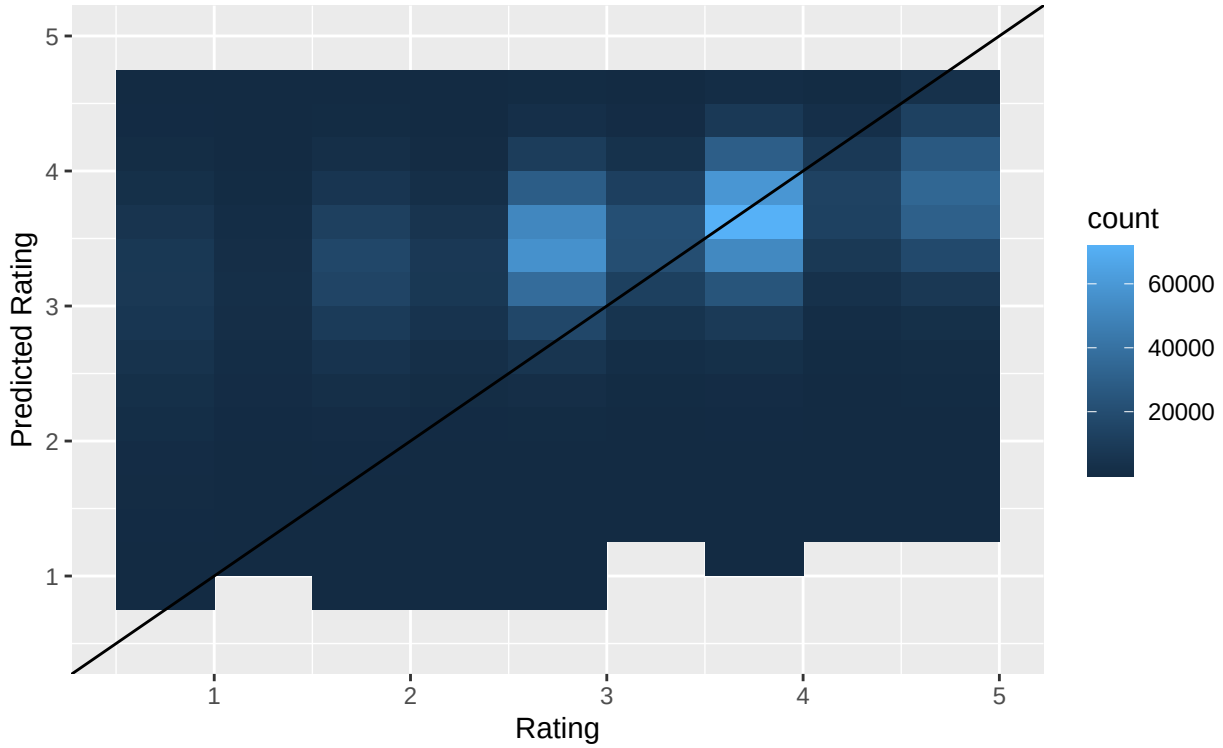


Figure 21: Relationship bet Predicted and Actual Ratings for User-Genre Bias Model with Regularization.s

Putting the Regularized *User-Genre Bias* in the table:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804
Movie Bias	0.9439072
Movie Bias w/ Regularization	0.9438557
Movie Bias + User Bias	0.8854581
Movie Bias + User Bias w/ Regularization	0.8834019
User-Genre Bias Model	0.9532481
User-Genre Bias Model w/ Regularization	0.9516660

2.8.3 Movie-Genre Bias

Using the same idea as the User-Genre Bias Model, movies' ratings may also be explained by its Genres. Try a regularized Movie-Genre bias with the formula:
$Y_m = \mu + \sum_{i=1}^{n}(g_{m,i}b_{g,m,i})$
Where,
$g_{m,i}$ is **1** if movie m is in the genre i, **0** otherwise.
$b_{g,m,i}$ is the effect of genre i on movie m.

Using the initial `k` of 1, the RMSE of initial *Movie Genre Model* is 1.0150828.

While the resulting *RMSE* is a little high, it is still lower than the base model. And like before, the k may still be optimized.
Find a better k for Movie-Genre Bias:

Plotting the results:

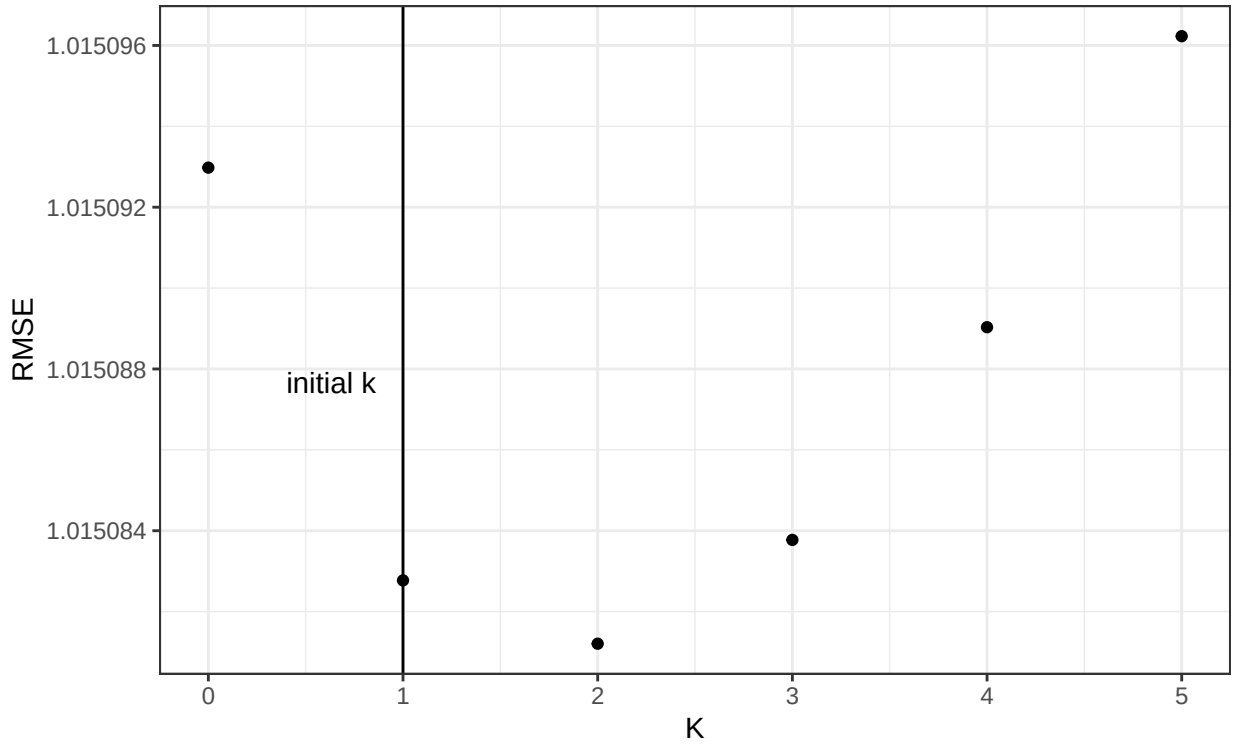


Figure 22: Plot of the RMSE of each parameter k

The lowest RMSE is 1.0150812 at k of 2.

Using k of 2, the resulting RMSE has 1.0150828.

Plotting the results:

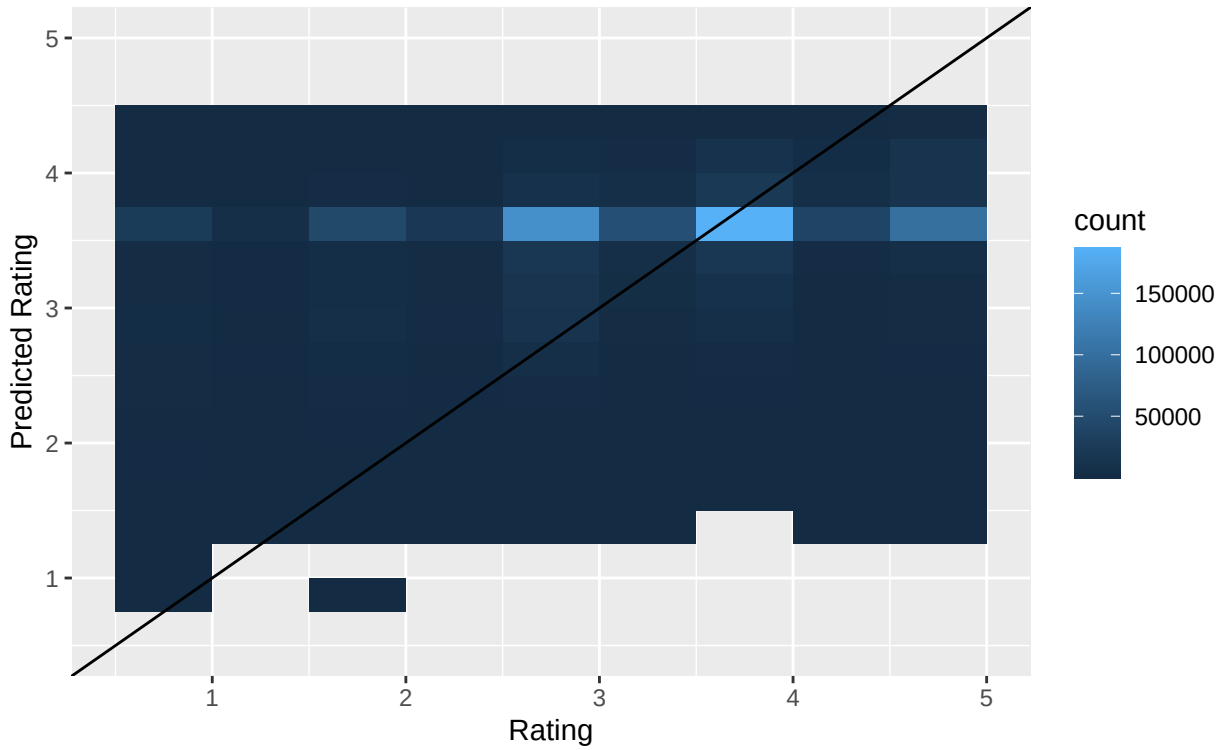


Figure 23: Predicted vs. Actual Rating of the Movie-Genre Bias Model

The results plot looked like the model barely predicted the actual ratings. This is also shown by the High *RMSE*. However, it may still be used in conjuction with other biases.
Showing the Current Model Table:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804
Movie Bias	0.9439072
Movie Bias w/ Regularization	0.9438557
Movie Bias + User Bias	0.8854581
Movie Bias + User Bias w/ Regularization	0.8834019
User-Genre Bias Model	0.9532481
User-Genre Bias Model w/ Regularization	0.9516660
Movie-Genre Bias Model w/ Regularization	1.0150828

2.9 Combine User Bias, Movie Bias, and Genre Bias

Currently, there are 4 different bias obtained:
User Bias, *Movie Bias*, *Movie-Genre Bias*, *User-Genre Bias*
Notice that the 4 biases can be split into
User Biases : *User Bias*, *User-Genre Bias*
Movie Biases : *Movie Bias*, *Movie-Genre Bias*

2.9.1 Combined User Biases

Testing the combination of User Biases using the formula:
$Y = \mu + b_u + b_{ug}$
Where* b_u is the user bias
And b_{ug} is the user-genre bias

Plotting the results for the combined User-Bias Model:

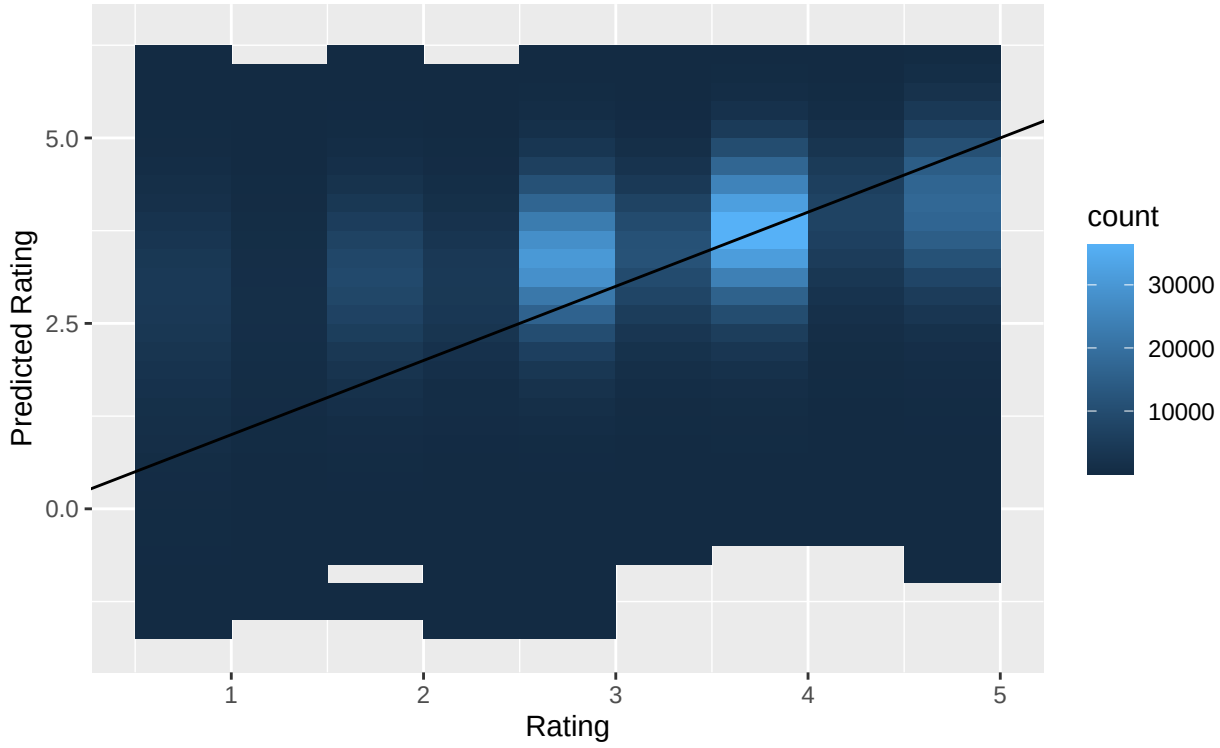


Figure 24: Predicted vs. Actual Ratings of Combined User Bias Model

The combined User Bias Model has an *RMSE* of 1.0425754.

2.9.2 Combined Movie Biases

Testing the combination of User Biases using the formula:
$Y = \mu + b_m + b_{mg}$
Where b_m is the movie bias
And b_{mg} is the movie-genre bias

Plotting the results for the *Combined Movie Bias* Model:

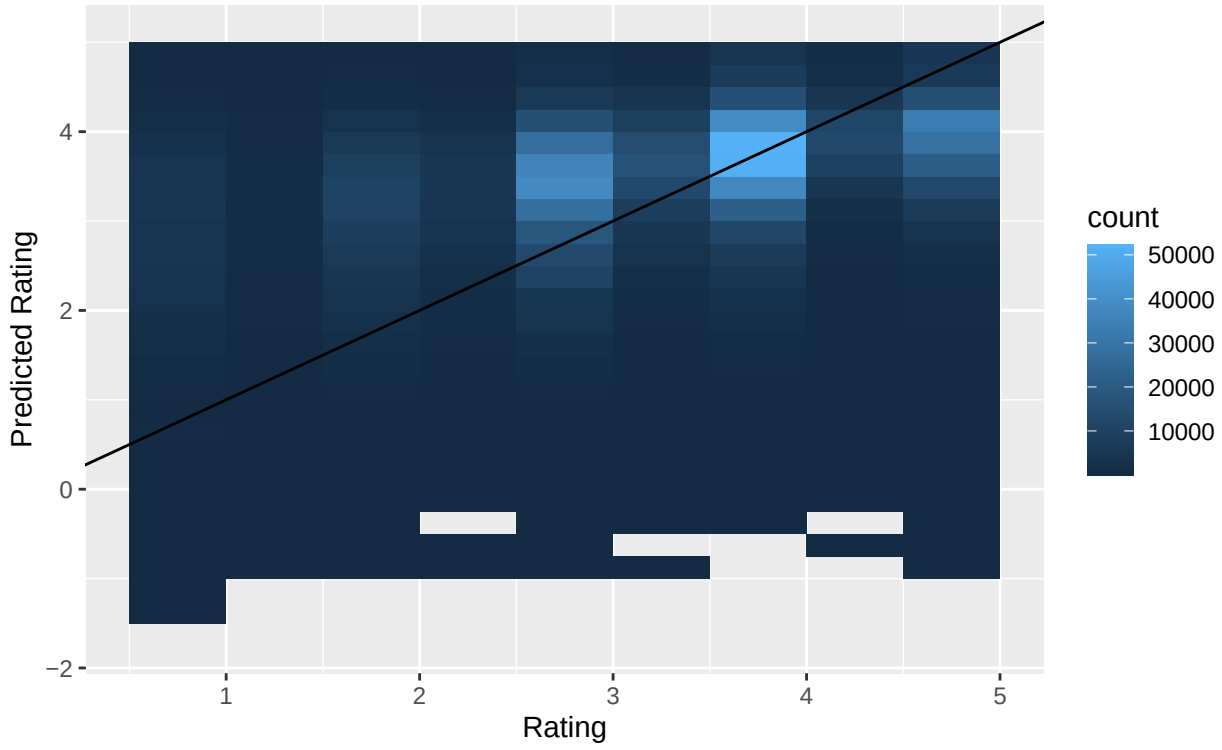


Figure 25: Predicted vs. Actual Ratings of Combined Movie Bias Model.

The combined Movie Bias Model has an *RMSE* of 0.9925297.

The current Model Table:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804
Movie Bias	0.9439072
Movie Bias w/ Regularization	0.9438557
Movie Bias + User Bias	0.8854581
Movie Bias + User Bias w/ Regularization	0.8834019
User-Genre Bias Model	0.9532481
User-Genre Bias Model w/ Regularization	0.9516660
Movie-Genre Bias Model w/ Regularization	1.0150828
Combined User Biases Model	1.0425754
Combined Movie Bias Model	0.9925297

2.9.3 Combined Biases:

The two models can be combined using the formula:
$Y = \mu + [b_u + b_{ug}] + [b_m + b_{mg}]$
Where,
b_u and b_{ug} are the user bias and user-genre bias of user u.
b_m and b_{mg} are the movie bias and movie-genre bias of user m.

Plotting the results for the combined Model:

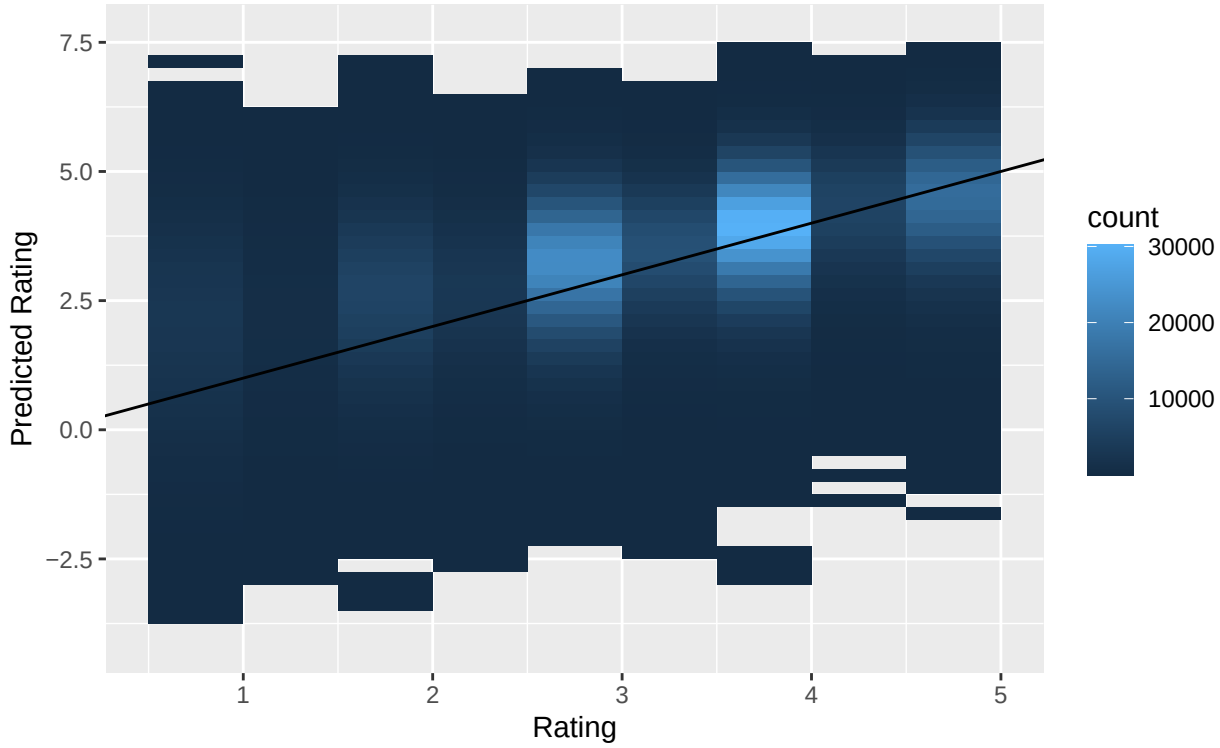


Figure 26: Predicted vs. Actual Ratings of User+Movie with Genre Model

The model resulted to an *RMSE* of 1.0627675.

Showing the models table:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804
Movie Bias	0.9439072
Movie Bias w/ Regularization	0.9438557
Movie Bias + User Bias	0.8854581
Movie Bias + User Bias w/ Regularization	0.8834019
User-Genre Bias Model	0.9532481
User-Genre Bias Model w/ Regularization	0.9516660
Movie-Genre Bias Model w/ Regularization	1.0150828
Combined User Biases Model	1.0425754
Combined Movie Bias Model	0.9925297
User + Movie with Genre Model	1.0627675

2.10 Adjustments

Adjust the last model to minimize RMSE more.

2.10.1 Shift ratings within the range

Looking at this part of the data:/

userId	movieId	predicted_rating	rating
9074	1193	7.638474	5
59987	1247	7.486009	5
9074	1136	7.432250	5
59987	1269	7.430430	5
24125	905	7.415927	5
7605	608	7.411985	5
59987	945	7.389761	5
13898	750	7.380282	5
44413	750	7.364117	5
46291	1193	7.354595	5

The maximum predicted rating goes beyond the maximum rating (5) of the training set.
Also:

userId	movieId	predicted_rating	rating
30272	3268	-4.122185	1.0
30272	2631	-3.616819	1.0
12138	2461	-3.591486	0.5
11994	3695	-3.576681	1.0
23159	810	-3.541335	1.0
24176	2286	-3.366109	1.0
49236	8859	-3.311284	2.0
15471	8859	-3.300011	0.5
24176	2050	-3.287547	1.0
24176	2458	-3.256018	1.0

The minimum predicted rating goes lower than the minimum rating (0.5) of the training set.
Adjusting the result:

Plotting the results of the adjusted model:

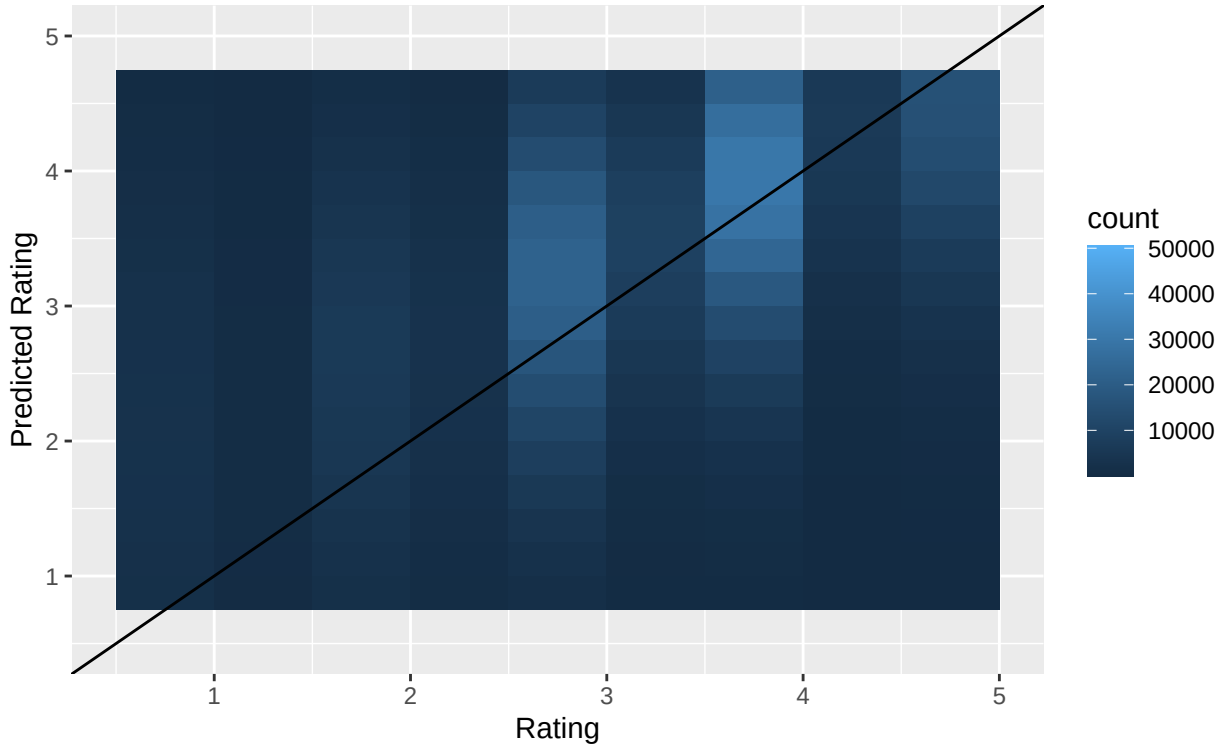


Figure 27: Predicted vs. Actual Ratings of Adjusted Model

After adjustment the RMSE is 1.0169438.

Showing the table of models:

Model	rmse
Overall Average Model	1.0604198
User Residual	0.9782389
User Bias w/ Regularization	0.9777804
Movie Bias	0.9439072
Movie Bias w/ Regularization	0.9438557
Movie Bias + User Bias	0.8854581
Movie Bias + User Bias w/ Regularization	0.8834019
User-Genre Bias Model	0.9532481
User-Genre Bias Model w/ Regularization	0.9516660
Movie-Genre Bias Model w/ Regularization	1.0150828
Combined User Biases Model	1.0425754
Combined Movie Bias Model	0.9925297
User + Movie with Genre Model	1.0627675
Adjusted User + Movie w/ Genre Effect Model	1.0169438

2.10.2 Multipliers

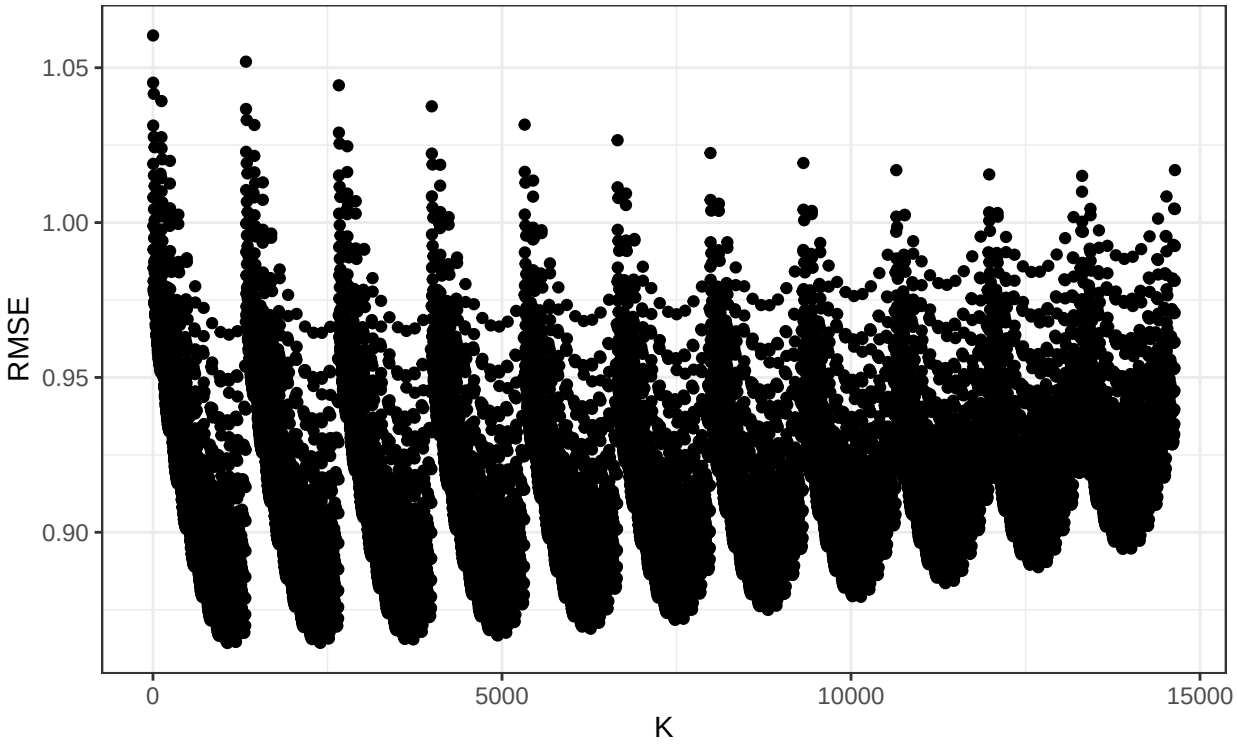
Notice from the previous models, the User Bias and the User-Genre Bias both describes the users' preference and the Movie Bias and the Movie-Genre Bias both describes the movies' effect on ratings. This can be better seen by taking the correlations:

user_correlation	movie_correlation
0.9099028	0.6367303

As the correlations show, the biases are correlated with each other. It is possible that there is some overlap in their effects which resulted in the higher *RMSE*. To try and compensate for this overlap, the model can be changed to:

$Y = \mu + [(x_u)b_u + (x_{ug})b_{ug}] + [(x_m)b_m + (x_{mg})b_{mg}]$
Where x_u, x_{ug}, x_m and x_{mg} are the coefficients of the user and movie biases that can reduce the overlap with the genre bias.
Look for the coefficients by checking the *RMSE* of the model when the coefficients are changed from 0 to 1 with 0.1 intervals.

Plot:



Showing best results:

	row_id	ur	ug	mr	mg	min_rmse
1068	1068	0	0.9	0.8	0	0.864265

The results show for the user biases, only the user-genre bias is needed. While for the movie biases, the movie-genre bias is not needed.

The current best Model to use is thus:
$Y = \mu + [(0)b_u + (0.9)b_{ug}] + [(0.8)b_m + (0)b_{mg}]$
or,
$Y = \mu + (0.9)b_{ug} + (0.8)b_m$

Plotting the results:

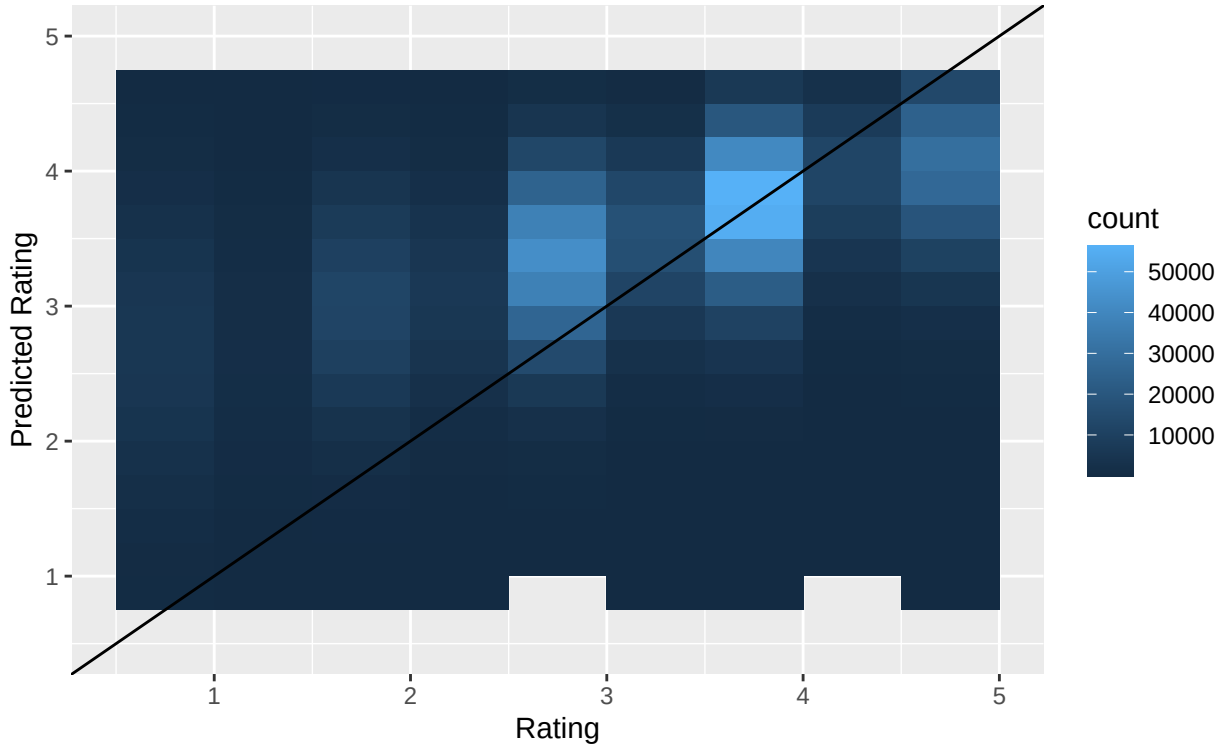


Figure 28: Predicted vs. Actual Ratings of Adjusted Model with Coefficients.

The current model resulted to an $RMSE$ of 0.864265. The $RMSE$ 0.864265 has reached the goal of $RMSE < 0.86490$.

The Full list of models:

Model	rmse	< 0.86490
Overall Average Model	1.0604198	FALSE
User Residual	0.9782389	FALSE
User Bias w/ Regularization	0.9777804	FALSE
Movie Bias	0.9439072	FALSE
Movie Bias w/ Regularization	0.9438557	FALSE
Movie Bias + User Bias	0.8854581	FALSE
Movie Bias + User Bias w/ Regularization	0.8834019	FALSE
User-Genre Bias Model	0.9532481	FALSE
User-Genre Bias Model w/ Regularization	0.9516660	FALSE
Movie-Genre Bias Model w/ Regularization	1.0150828	FALSE
Combined User Biases Model	1.0425754	FALSE
Combined Movie Bias Model	0.9925297	FALSE
User + Movie with Genre Model	1.0627675	FALSE
Adjusted User + Movie w/ Genre Effect Model	1.0169438	FALSE
Adjusted User + Movie w/ Genre Effect & coefficients Model	0.8642650	TRUE

For the final evaluation using the *Validation Set*, the entire *Input Set* `edx` can be used to train the model or obtain biases. This will result to a better model or a better fit, since the *Input Set* will have more data for training.

2.11 Final Model

Previously, the coefficients obtained resulted to a coefficient of `0` for *User Bias* and *Movie-Genre Bias*. This means that for the final model, the two Biases can be ignored.

Using the User-Genre x_{ug}:
$x_{ug} = \sum_{i=1}^{n}(g_{m,i}b_{g,u,i})$
Where,
$g_{m,i}$ is **1** if movie m is in the genre i, **0** otherwise.
$b_{g,u,i}$ is the genre preference/residual of user u for genre i.

The final model formula is:
$Y = \mu + (0.9)x_{ug} + (0.8)x_m$
Where,
μ is the overall average rating,
x_{ug} is the User-Genre Bias.
x_m is the Movie Bias.

For the final model, the entire *Input Set* `edx` was used to train the model.

3 Results

Using the Final Movie and User-Genre Biases, we can evaluate the model on the `final_holdout_test`.

From the start of the Final Model to the end of prediction, the time taken is:

Time elapsed
18.66 mins

Using the Final Model, the RMSE of the Final Evaluation on the *validation set* `final_holdout_test`:

Model	rmse
Final Evaluation of the Final Model	0.8634292

The Final Result is within the target RMSE of `<0.86490`.

Plotting the Results:

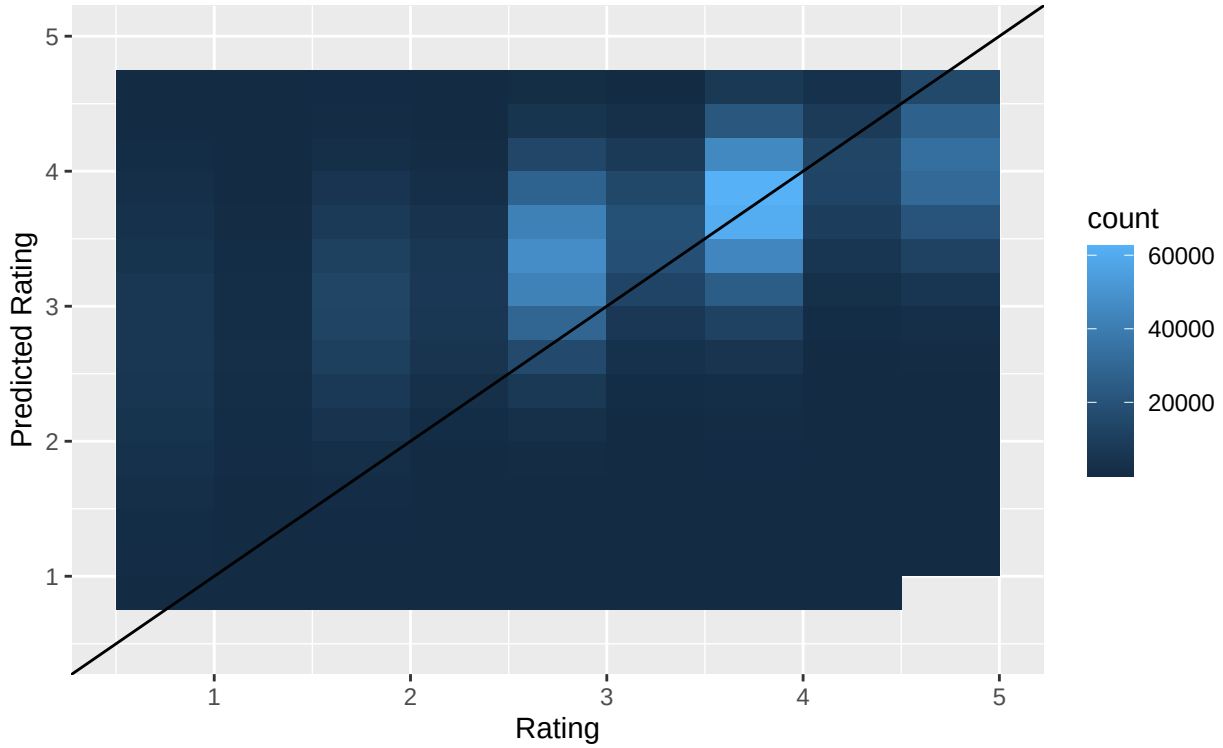


Figure 29: Predicted vs. Actual Ratings for the Final Model

The model has a high accuracy for ratings between *3.5* to *4*, or ratings close to the average 3.5124652. However, the Model still has problems predicting the lower end of ratings (0.5-1.5).

4 Conclusion

The final model was obtained by first obtaining the User and Movie Bias of the training set. Then, based on how the User and Movie Bias are obtained, derive a way to obtain the effect of movie genre on the user's and movie's ratings. This genre effect were the User-Genre Bias and the Movie-Genre Bias. Each of the 4 biases were optimized using regularization and the optimization of the regularization parameter. The model also took into account predicted ratings outside the range of the input ratings. Finally, find the coefficients or multiplier of each of the 4 Bias. The obtained coefficients showed that the User Bias and the Movie-Genre Bias can be removed. By using the *Input Set* `edx`, the final model, evaluated on the *validation set* `final_holdout_test`, resulted to an RMSE less than 0.86490 at 0.8634292.

Since, the final model mainly uses past ratings to predict future ratings, it will likely fail to predict users or movies with no prior ratings made. Also, the final model was trained from a data set with fixed number of rows/data, which is different from practical applications with new data coming in regularly. The model can be adapted to practical applications by updating the initially obtained biases with the incoming data.