Developing Machine learning Model to Predict Movie Ratings from User and Movie Features

Kelvinmg

Executive Summary

The goal of this analysis is to develop the most accurate machine learning model to predict movie ratings based on various features such as movieid and userid interaction. This type of models are reffered to as recommendation systems. There are various techniques of building recommedation systems but we are going to only use machine learning models.

The dataset used in this analysis is obtained from grouplens website and contains data on user movie ratings from movielens. The dataset includes user behavior, movie attributes and ratings.

It contains approximately 10 million ratings with the following key attributes:

- userId: unique identifier for each user.
- movieId: unique identifier for each movie.
- rating: numeric rating (between 0.5 and 5) given by the user.
- timestamp: time when the rating was given (used primarily for time-based analysis).
- title: The title of the movie.
- genres: The genre(s) associated with the movie.

The key steps will be; *simple analysis*, *feature engineering,data exploratory analysis*, *model building* and *model evaluation*.

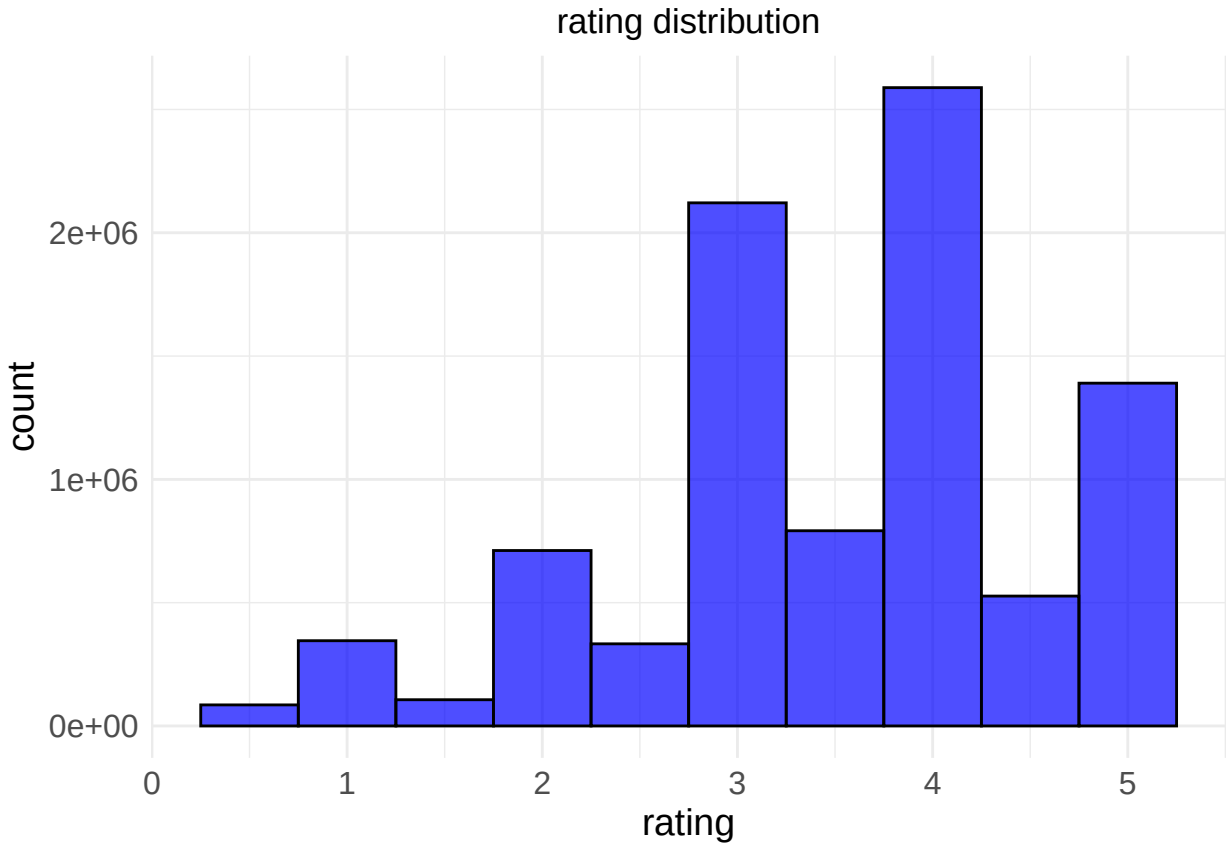
Data Cleaning

```
## 'data.frame':    9000055 obs. of  6 variables:
##  $ userId   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ movieId  : int  122 185 292 316 329 355 356 362 364 370 ...
##  $ rating   : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ timestamp: int  838985046 838983525 838983421 838983392 838983392 838984474 838983653 838984885 83
##  $ title    : chr  "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" ...
##  $ genres   : chr  "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Ac
```

Notice the release years are written in the title in parentheses for each title. The first step will be to extract them and store them in a new variable called release_year. Then we will store the dataset to a new data frame which we will call edx_1

The next thing you notice is that there is a timestamp that can be converted to easy to interpret time. We will extract the year and store it in a new variable called year_rated. Then we will save this to a new dataframe which we will call edx_2.

The feature we will analyse first is the rating distribution: \



This reveals the mode rating which is 4 followed by 3, it does not convey much about the type of distribution. We can confirm this by checking the number of people who used each rating.

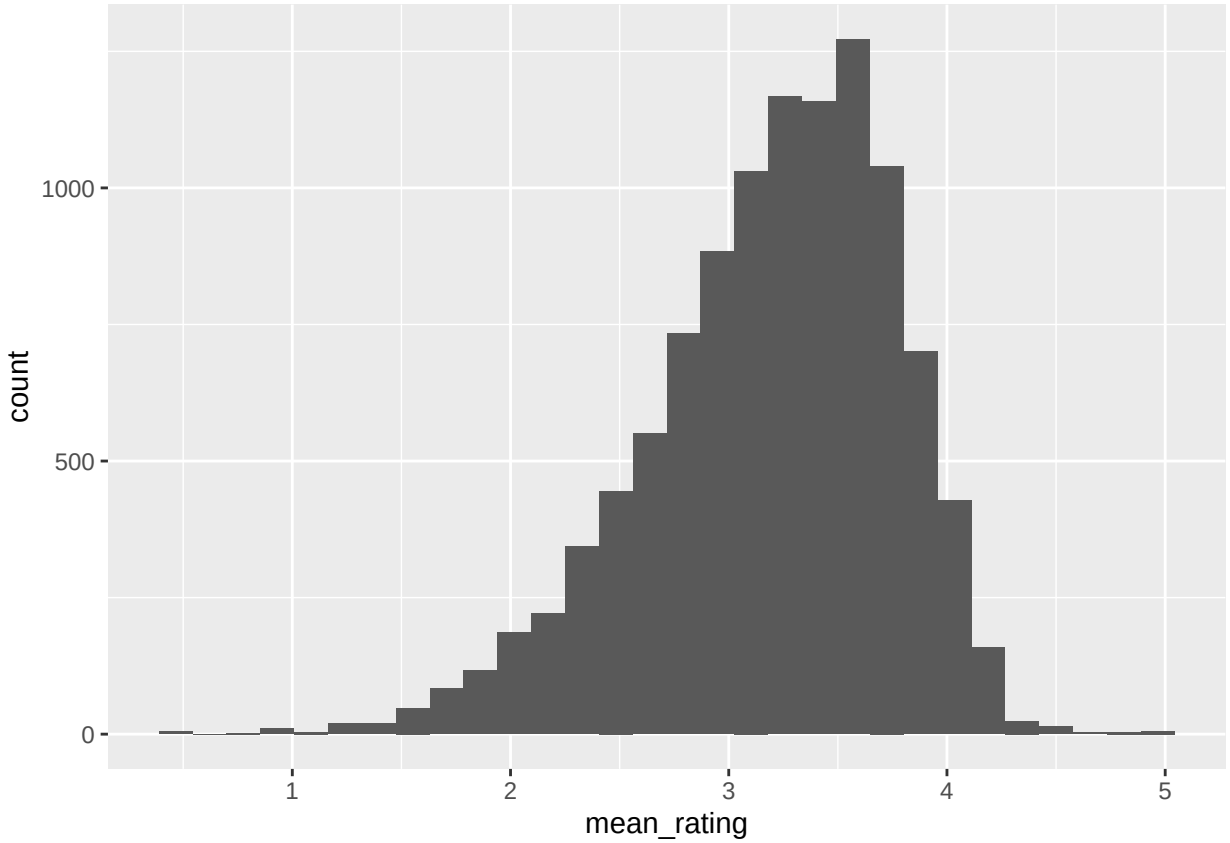
The next process is performing data exploratory analysis on the engineered dataset edx_2 to get an idea of how the different features interact.

first analysis will be to analyse the features summarized by each movie to understand if there is a movie bias across the various features.

The first features we will analyse will be the mean rating for every movie and amount of reviews the movie had, to do this we will store the data in a dataframe called edx_summary_movieid. then try and get the best 6 movies with more than 100 reviews.

the bottom 6 with more than 100 reviews

an histogram to understand the distribution of mean rating

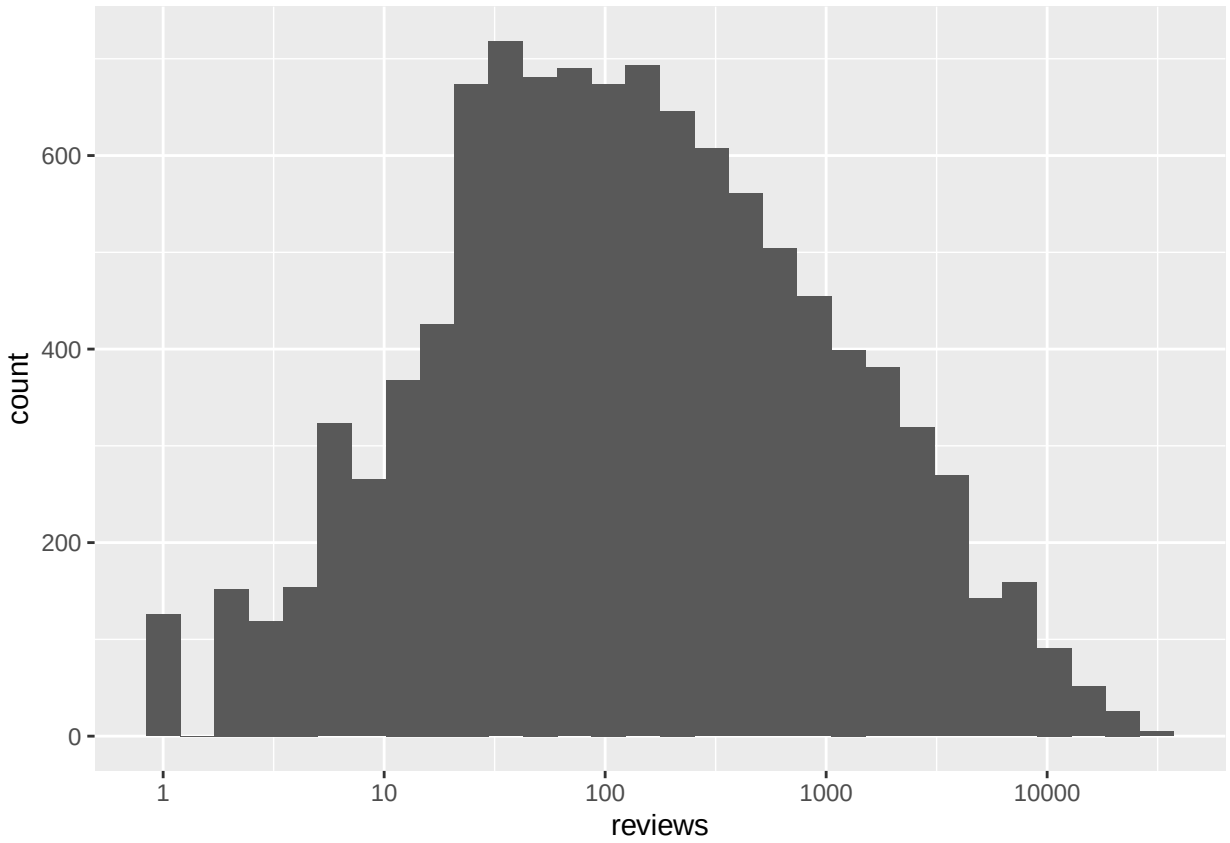


The distribution is skewed to the left meaning there are relatively few mean rating values with most of them being clustered on the higher end, creating a long tail on the left side of the distribution. The histogram shows the most movies had a rating of 3 to about 3.7 near the mean.

The mean is:

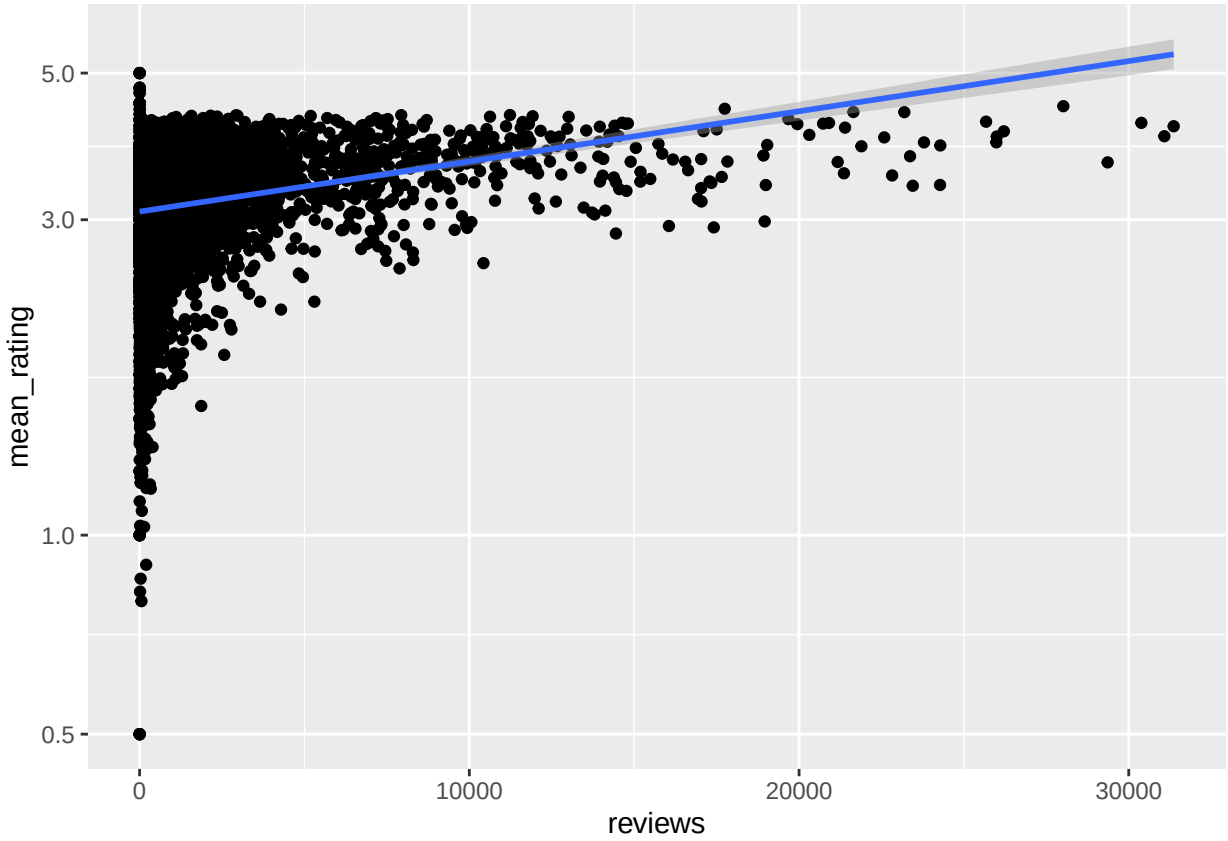
```
## [1] 3.512465
```

Next we will analyse the reviews histogram



The histogram shows that most movies have moderate amount of reviews

next will be to plot the mean rating against the reviews to observe the relationship




```
##
## Call:
## lm(formula = mean_rating ~ reviews, data = edx_summary_movieid)
##
## Coefficients:
## (Intercept)      reviews
##   3.146e+00    5.396e-05
```

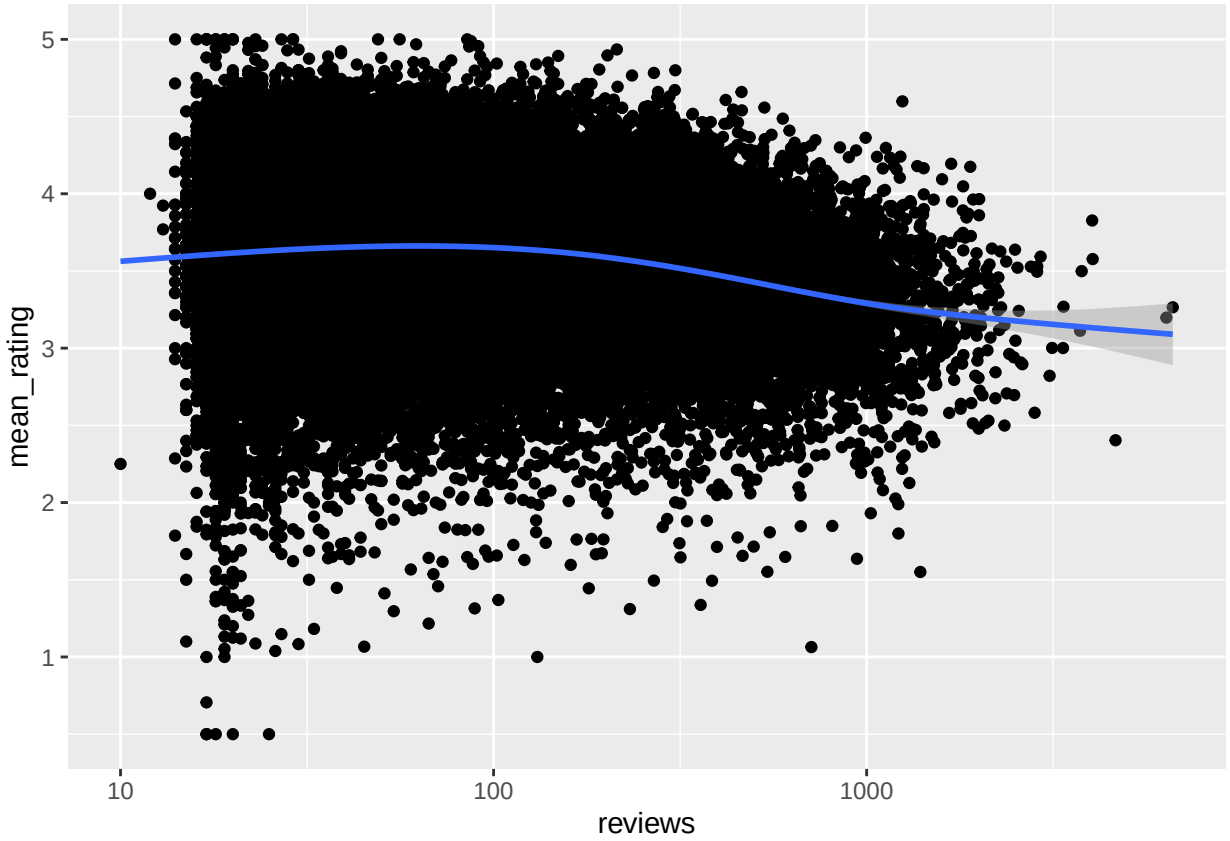
```
## [1] 0.2114161
```

The plot shows that from the rating of 3.14 there is a weak linear relationship with the correlation showing how that the 2 variables are weakly related. This tells us that there is a small movie bias.\

our second analysis will be to analyse the features summarized by each user to understand if there is a user bias across the various features.

Top 6 users who had the highest mean ratings after reviewing more than 100 times in the dataset.

checking the users with the most reviews and their mean rating



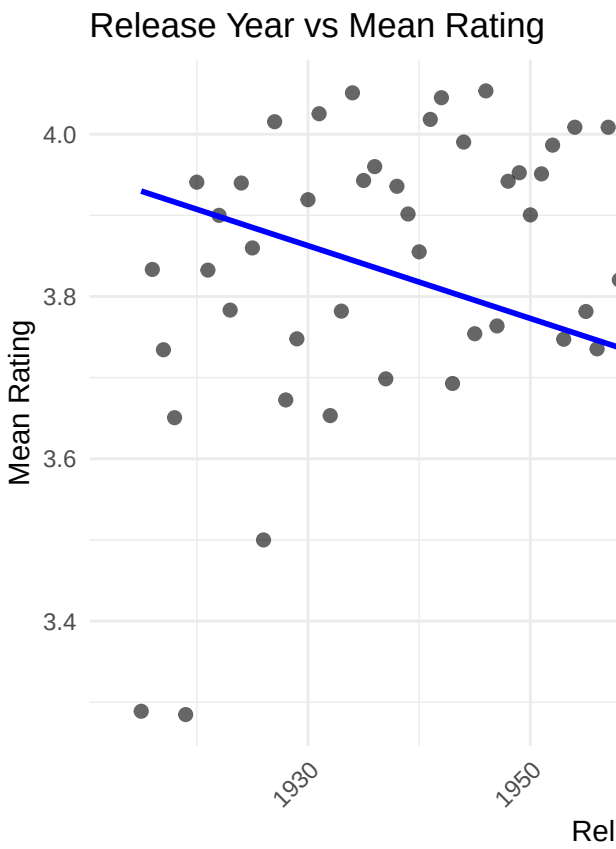

```
##
## Call:
## lm(formula = mean_rating ~ reviews, data = edx_summary_userid)
##
## Coefficients:
## (Intercept)      reviews
##   3.6576963   -0.0003424

## [1] -0.1550551
```

The features are be negatively correlated but weakly. This tells us that there is a small user bias. \

Deeper analysis to understand the engineered features year_rate and release_year

top 6 highest mean rating release years with more than 100 reviews



The relationship between the release year and the mean rating that year

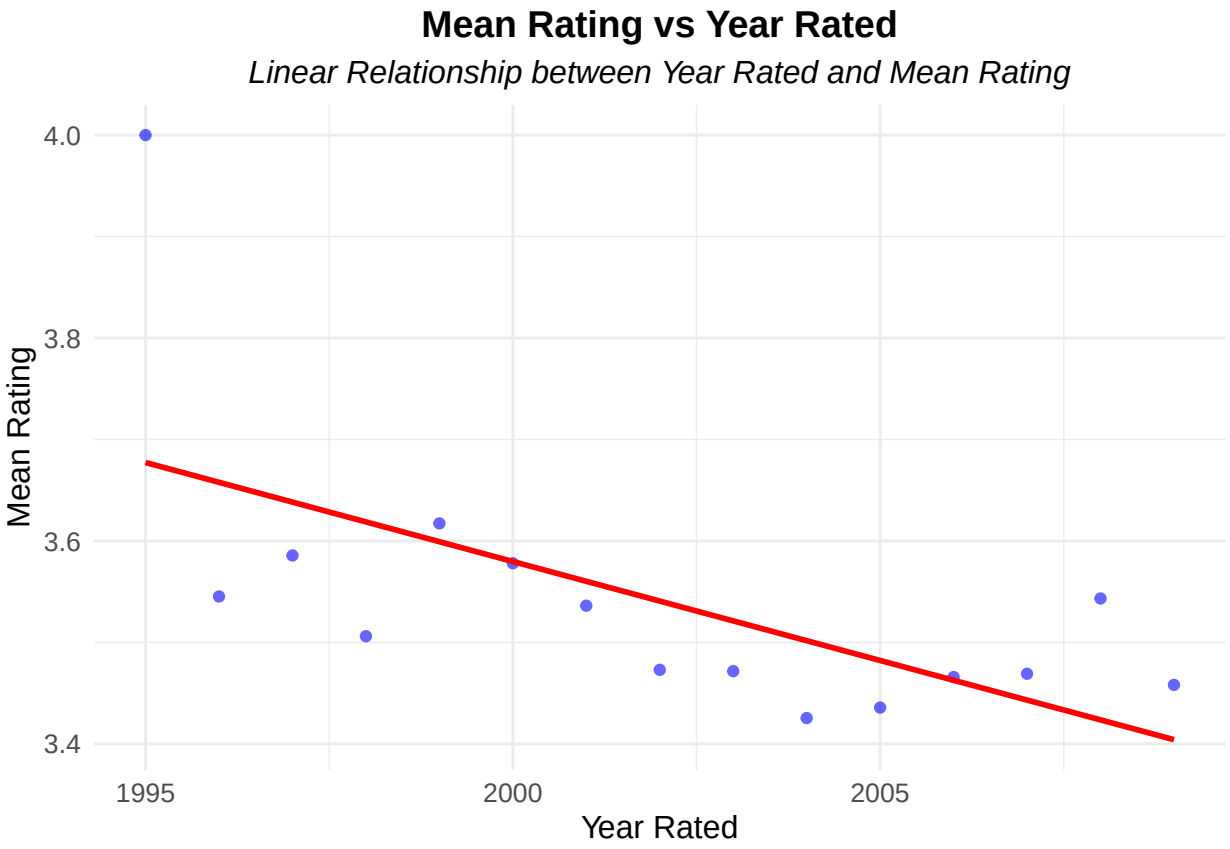
```
## [1] -0.582716
```

The mean rating has an inverse relationship with the release year. The correlation tells us that the features are strongly negatively correlated which tells us that there is a release year bias.

The analysis between year rated and mean rating

Top 6 highest mean ratings for year rated with more than 100 reviews

We will test the relationship between year rate and mean rating




```
## [1] -0.6252134
```

There is a strong negative correlation between the two features year rated and mean rating. This shows there is a strong year rated bias.

Model building

We will test three models; the movie bias, movie and user biases and then regularize the movie and user biases. We will then compare the Rmses of the three models to see which had the lowest meaning highest accuracy.

we will test our models with the RMSE

##Our baseline model is the guessing model/naive model. we use the average to guess all the outcomes as it represents our expected prediction for all the cases. i.e expected mu.

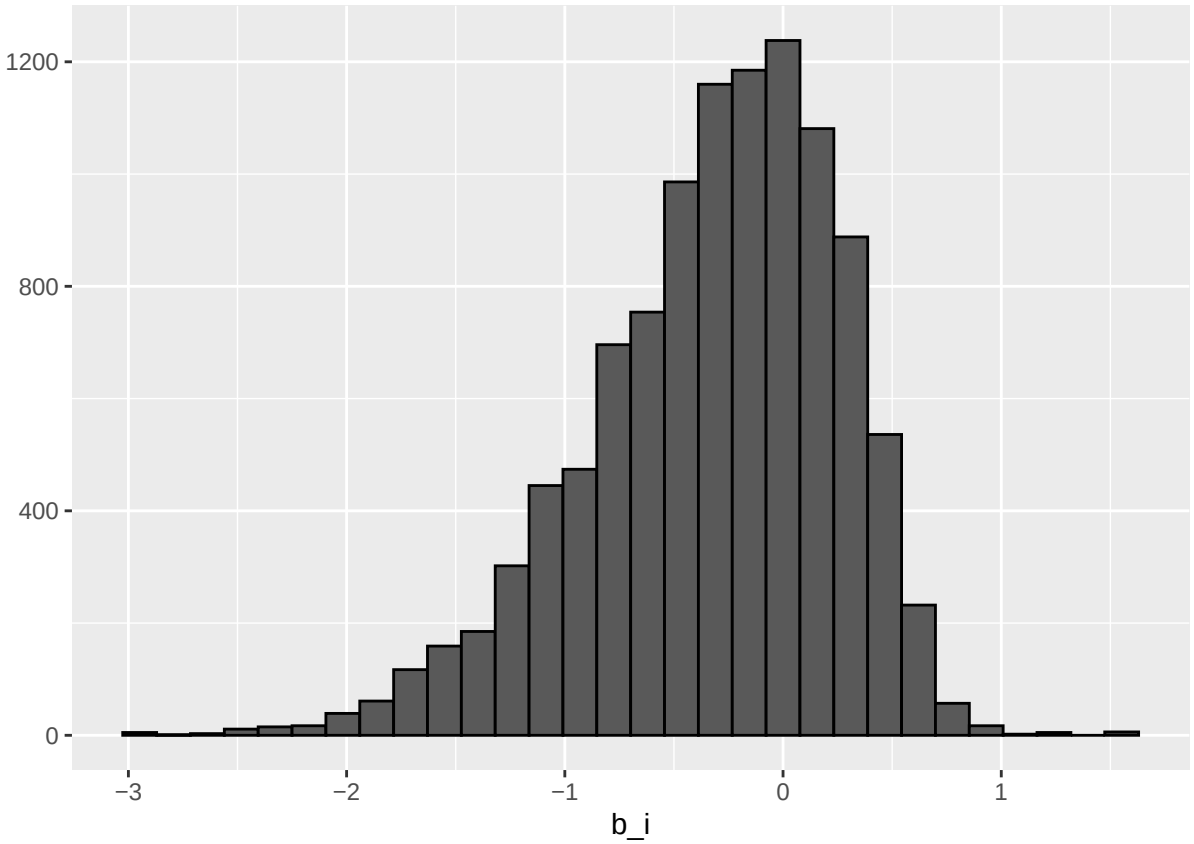
```
## [1] 3.512465
```

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

method	RMSE
baseline model	1.061202

An RMSE of 1.061202 is relatively low, but considering that our data ranges from 0.5 to 5, this value may be acceptable but a movie company would need better accuracy to be more effective. we will work on improving the model and use this as the baseline to evaluate whether the other models have a noticeable impact.

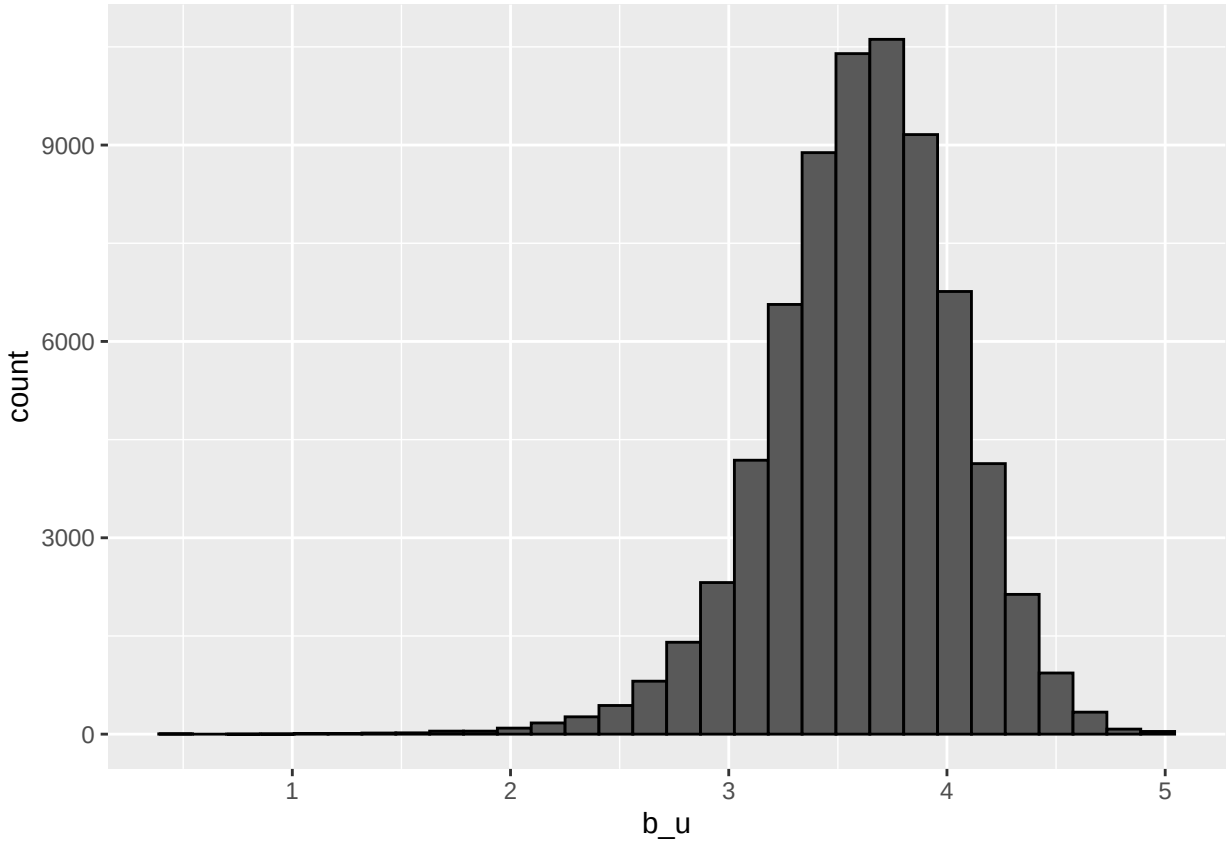
#The first model we are going to try is the movie effect We analyzed this effect in the exploratory data analysis. To reduce the inaccuracy, we will calculate the bias for each movie and incorporate it into our baseline prediction (u). This adjustment should help improve the model's accuracy.



method	RMSE
baseline model	1.0612018
Movie Effect Model	0.9439087

The RMSE is 0.9439087 and is better than the baseline model. more improvements can be made by adding the second feature in our analysis.

#The second model we are going to incorporate the second feature analysed the user bias to the movie bias and baseline m(u)



method	RMSE
baseline model	1.0612018
Movie Effect Model	0.9439087
Movie and User Effects Model	0.8653488

The Rmse of 0.8653488 is decent for our range of data and can be improved again by optimizing the biases which can be done using a prameter that can be tuned lambda giving us a chance to use cross validation.

The third model we will use Regularized movie bias and user bias which can be achieved by tuning the lambda

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

method	RMSE
baseline model	1.0612018
Movie Effect Model	0.9439087
Movie and User Effects Model	0.8653488
Regularized movie and user effect model	0.8648170

The regression of 0.8648170 is the best of all the models tested coming from regularized movie and user effect.

Conclusion

This report examined various models for predicting movie ratings, beginning with a baseline model and progressively adding movie and user effects. The models evaluated include:

Baseline Model: RMSE = 1.0612018 Movie Effect Model: RMSE = 0.9439087 Movie and User Effects Model: RMSE = 0.8653488 Regularized Movie and User Effect Model: RMSE = 0.8648170

The analysis revealed that incorporating the movie effect led to improvement in prediction accuracy, followed by the addition of both movie and user effects. The regularized movie and user effect model provided the best performance, yielding a slight reduction in RMSE. This suggests that factoring in both movie-specific and user-specific biases, along with applying regularization, enhances prediction accuracy.

Limitations

some limitations include assuming the models assume linear relationships between movie and user effects, which might not fully account for more underlying relationships in the data. The other limitation would be utilizing only basic features meaning there still more accuracy that can be gotten for example utilising the years and genres.

Future Work

Incorporating Additional Features: Introducing more features, such as genre preferences, movie release year, or user demographic data, could help capture more underlying factors in the data. Utilizing more advanced models like matrix factorization, deep learning, or collaborative filtering techniques could lead to better accuracy performance.