The Analyzation of Movies and the MovieLens Data

Introduction

In this project, I analyzed the data on movies known as the MovieLens data set that is originally collected by GroupLens Research. The question is: Would we be able to predict the ratings of the movies based upon the user preference or the age of the movie?

The MovieLens data set that I examined contains 10000054 rows, 10677 movies, 797 genres, and 69878 users.

The steps I performed for the analysis of the data:

-Created an age for the movie column

-Made a graphic display of movie, users and ratings in order to find a pattern of the data

-Looked into the determination coefficient of R-squared

-Graphically examined the linear correlation and r-value

-Calculate the RMSE based upon Movie ID, the age of the movie, and user ID

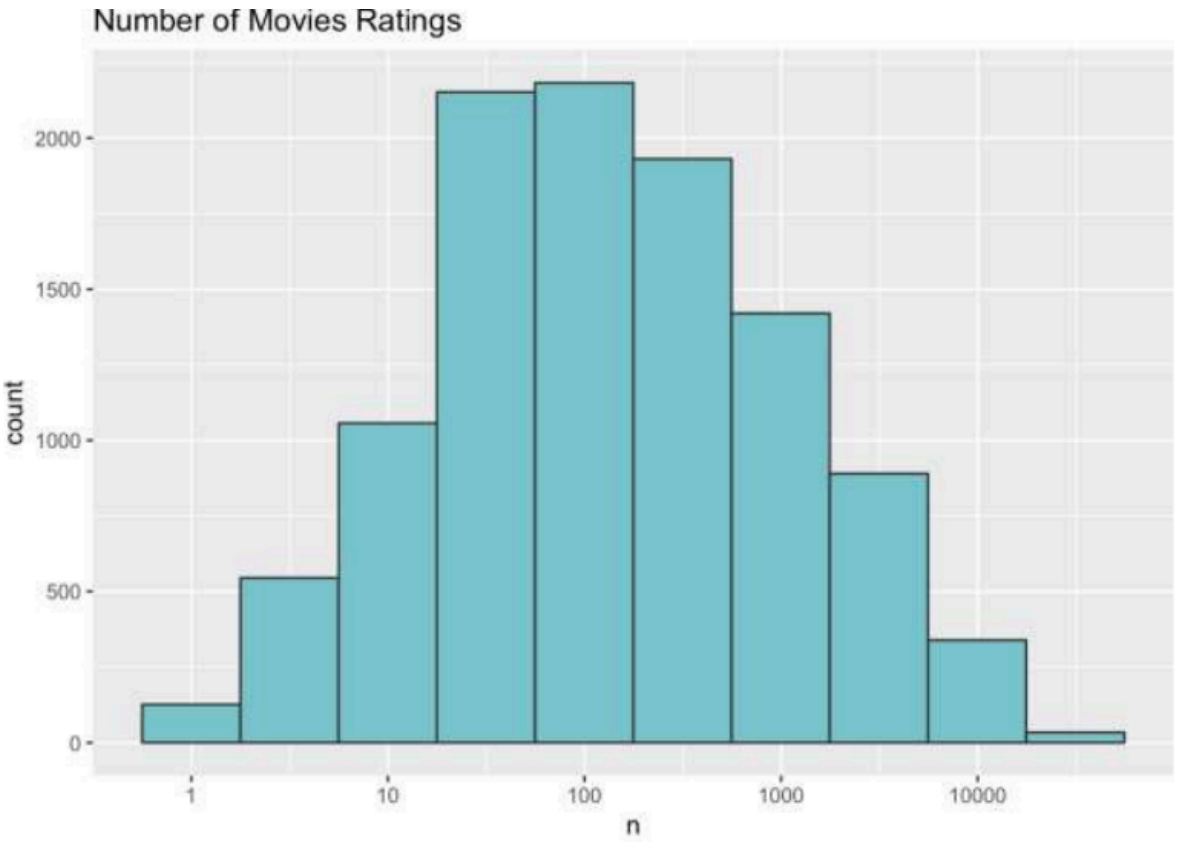
After evaluating the movies through their graphical representations and calculating the RMSE, I concluded that the best predictor for the movie ratings was the Movie ID, User ID. The age of the movie had no change or effect on RMSE.

After calculating the RMSE and observing the graphical representations, we can see that the official RMSE is 0.8252. We can also come to the conclusion that the age of the movie has no change amongst the RMSE and that the movieid, userid is the best predictor for determining the ratings of the movies.

In order for me to prove that the age of the movie was even a factor for predicting the rating of the movies, I calculated the age of the movie and used the date of the movie premier to assist me with this. I also observed the outcome of the user ratings along with all the genres.

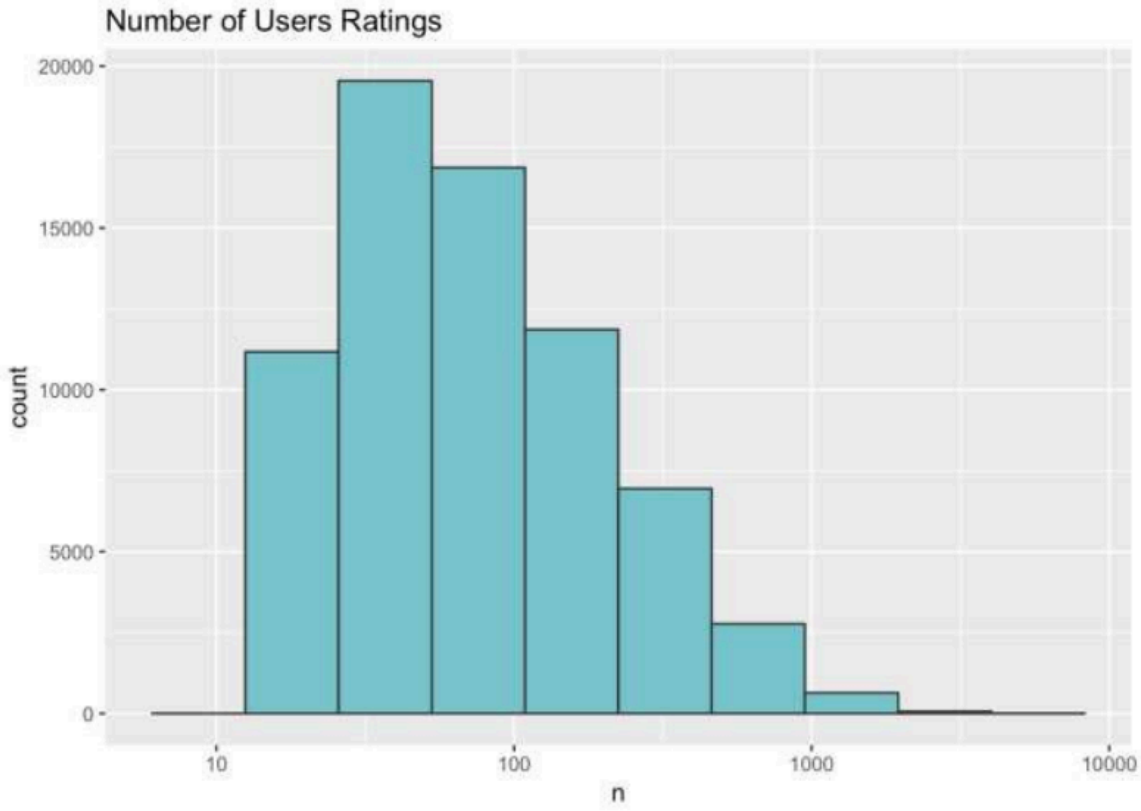
Graphical Representations

Distribution of the data



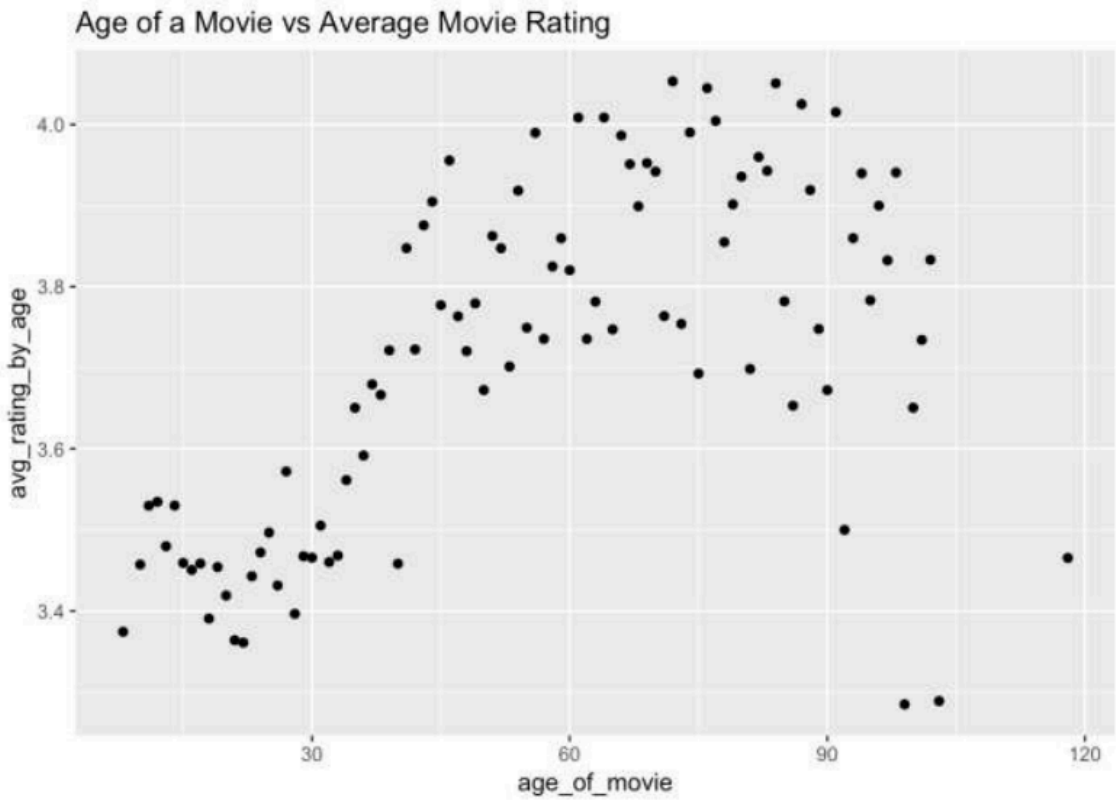
As we can observe from the graph, the distribution of the ratings of the movies varies. There were some movies that we rated more than 10,000 times while some were only rated once.

Distribution of the User Rating



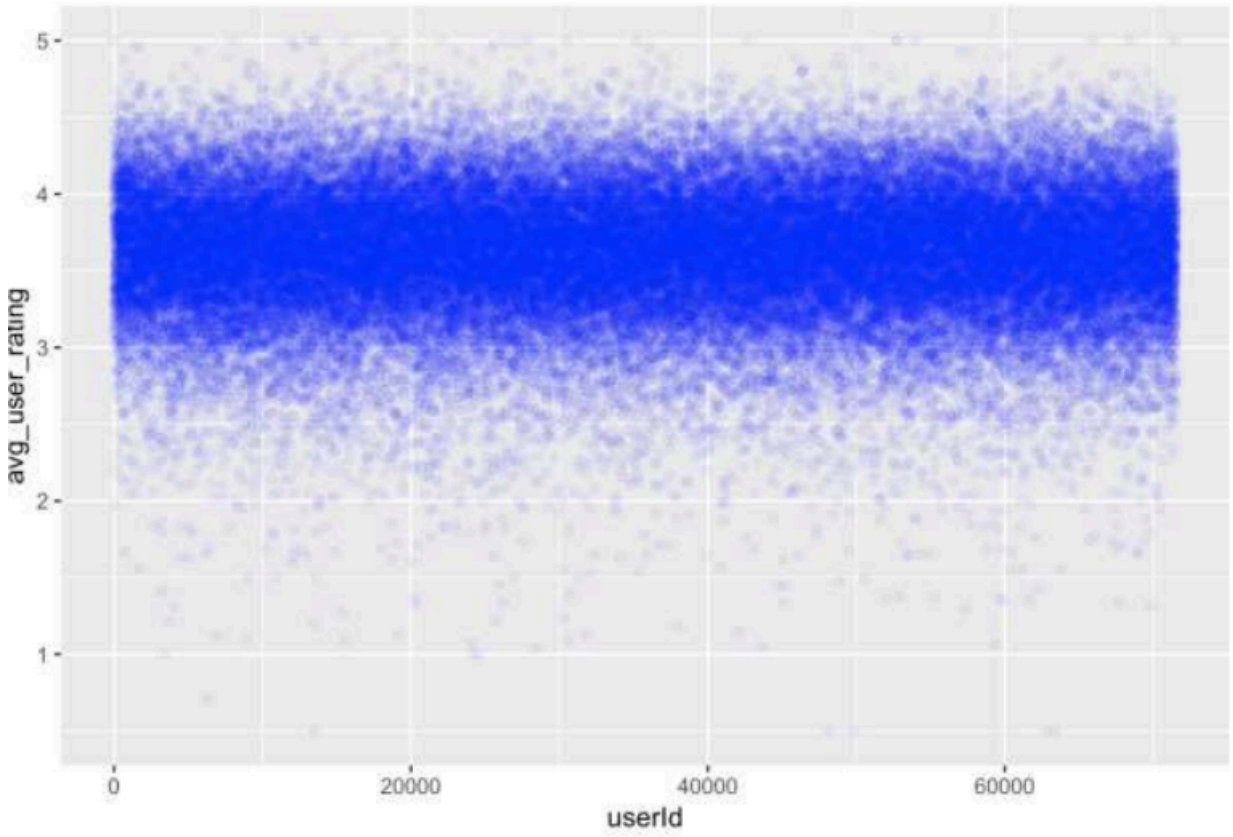
As we can observe from the graph, the data from the user ratings shows that it is, not only skewed to the right, but also 75% of the users rated about 150 movies or even fewer.

Relationship between the Age of the Movie vs. Average Rating of the Movies



The plot above shows more variability the more that the movies age. The newer that the movie is, the lower its rating. The average ratings demonstrate that they increase from the increment of 30 years to 90 years before the ratings of the movies decrease. This could be from less ratings or other factors.

UserID vs. the Average User Rating



As we can observe, the average user ratings by the UserID remain consistent between the increments of 2.5 and 4.5.

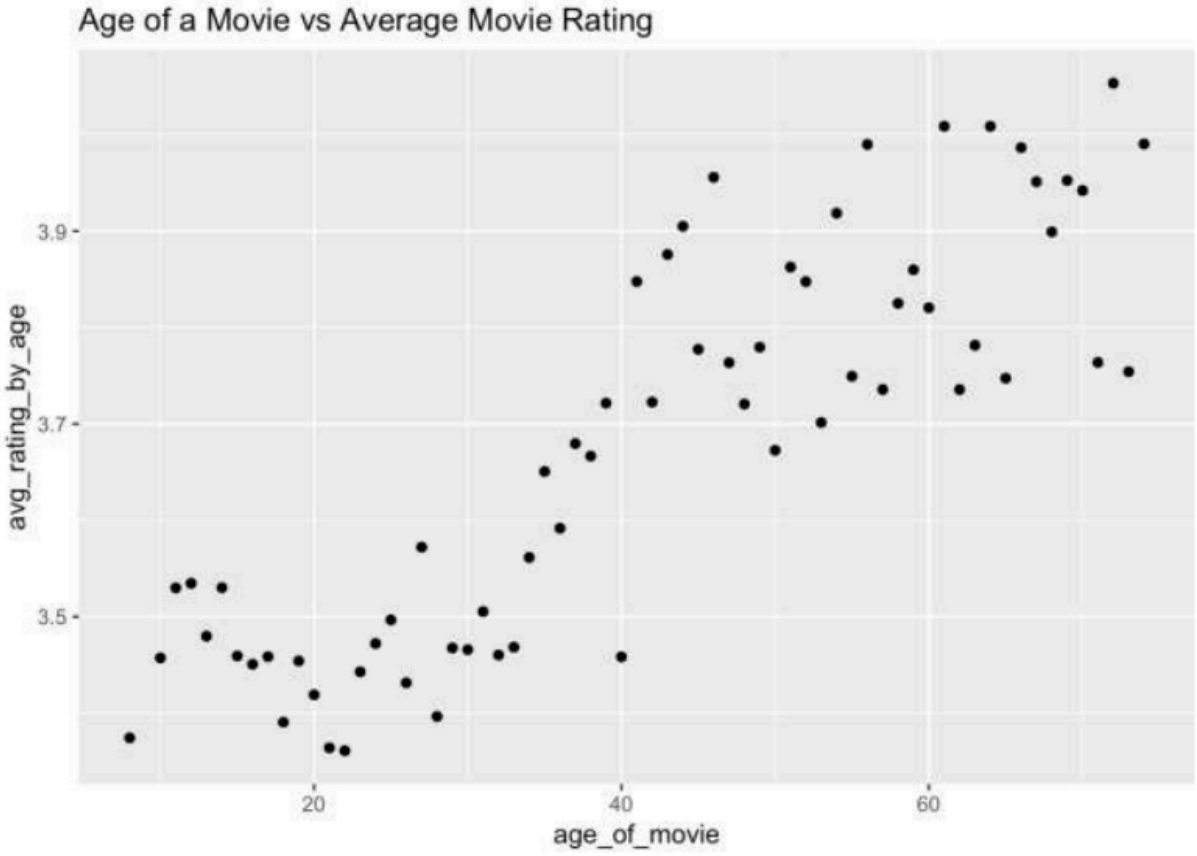
The linear graphical representation between the age of the movie versus average movie rating demonstrates some correlation, but not a strong correlation. It also shows us that there is an R-squared value of 0.30.

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61684 -0.10389  0.00276  0.12759  0.28508
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.4809443  0.0409983  84.905  < 2e-16 ***
## age_of_movie 0.0041241  0.0006489   6.356 7.38e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1781 on 94 degrees of freedom
```

```
## Multiple R-squared:  0.3006, Adjusted R-squared:  0.2931
## F-statistic:  40.4 on 1 and 94 DF,  p-value: 7.377e-09
```

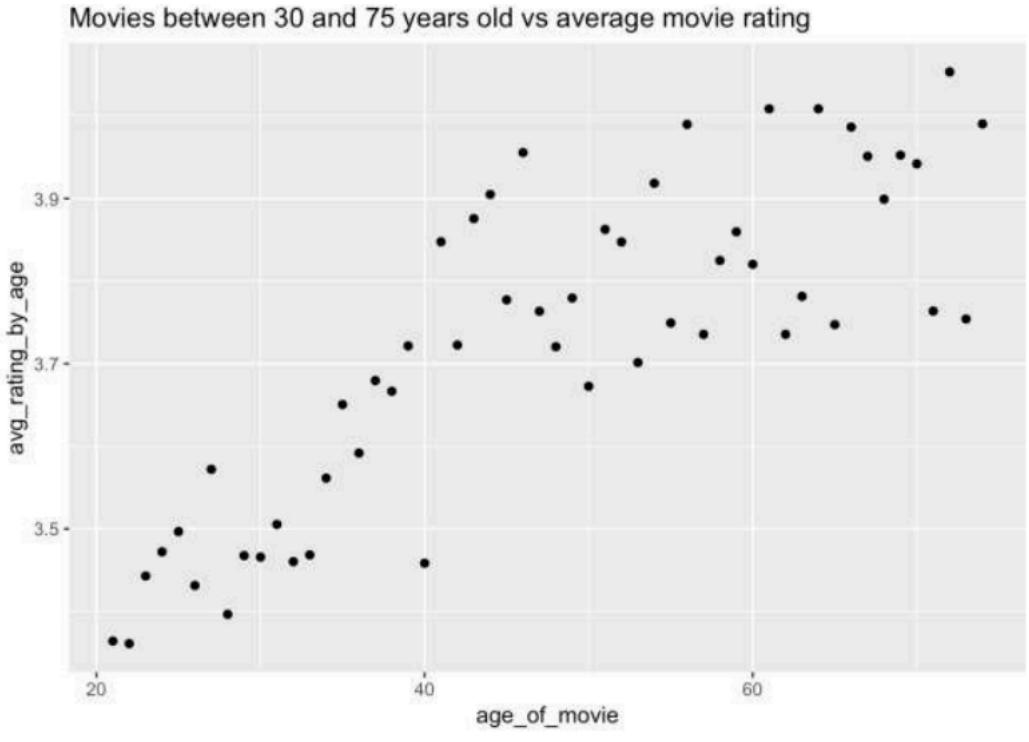
As we can observe from the graph above, the older movies are more inconsistent with the residuals.

Does age effect the R-squared?

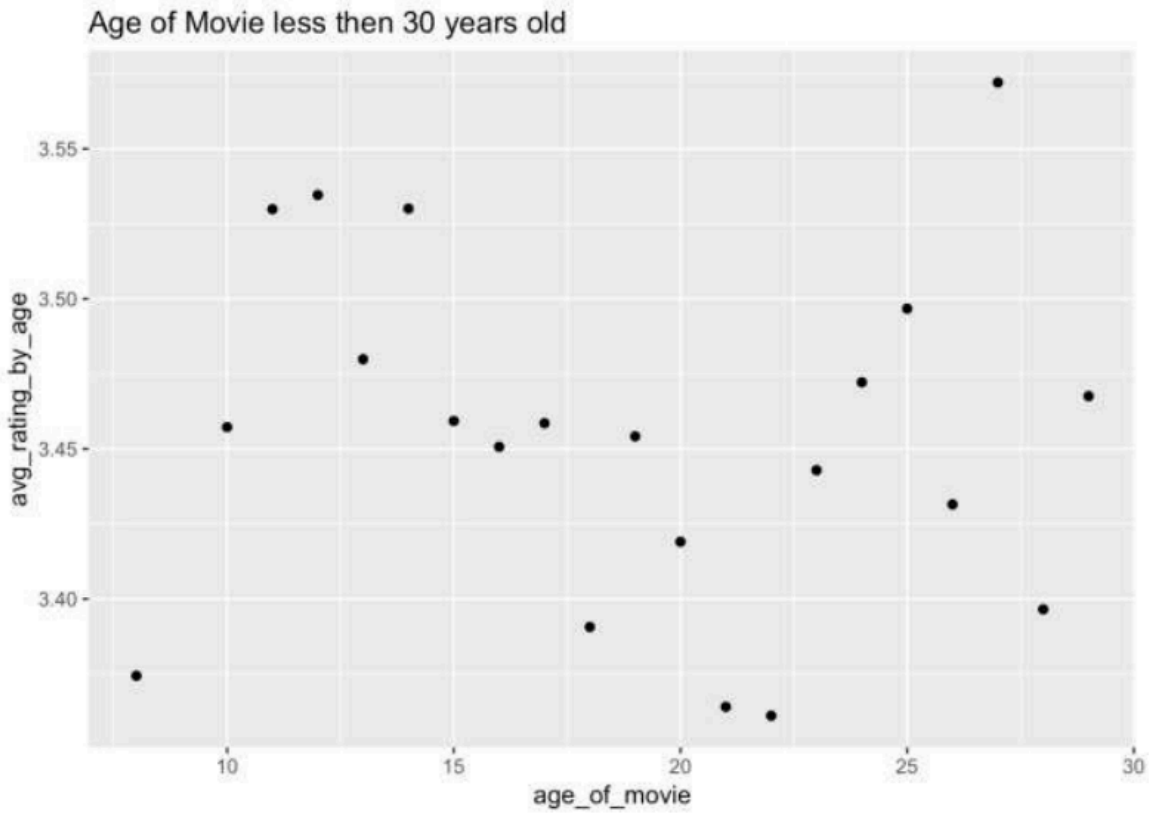


As we can observe, the R-squared improved and is now 0.745 once we removed movies that were greater than 70 years old. This will change the residual plot now from -0.2 to 0.2 as shown below.

Next, I observed the movies that were between the ages of 30 and 75 years old and compared it to their averages in ratings. I learned that the movies that were younger than this had a negative linear trend as where the movies between these ages appeared to be more linear.

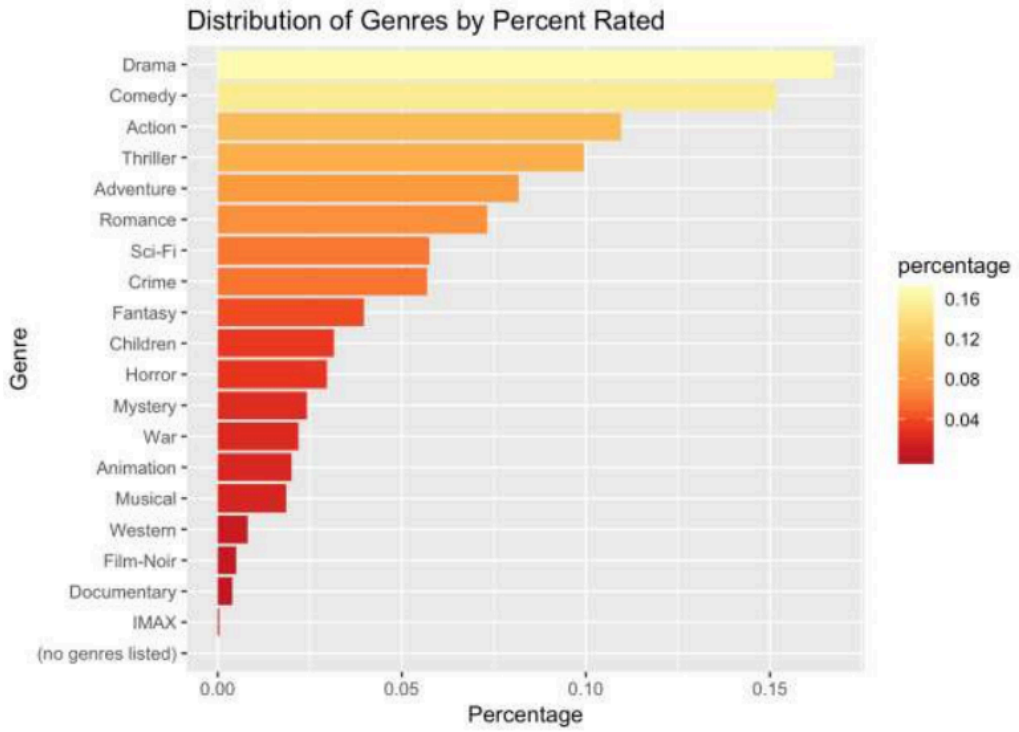


This also impacted the r-square as r-squared appeared to decrease to 0.69. When I observed the movies that were younger than this, the r-squared had gone to nearly zero and the same graph has no correlation as shown below.

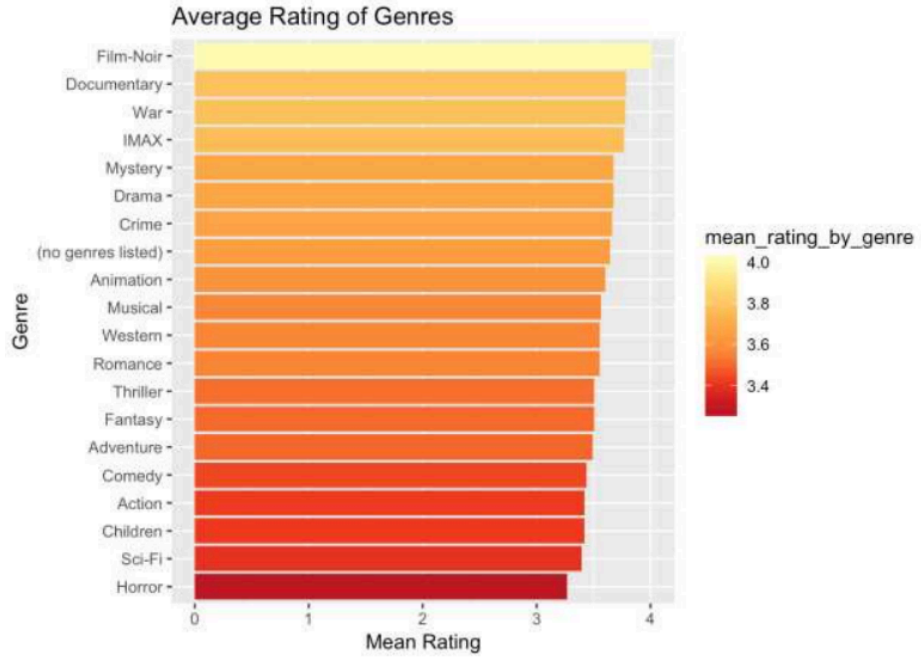


Does the genres of movies have any effect on the ratings?

In order to make this determination, I generated code that allowed for me to view all of the movie genres and the number of movies in each of the genres in order to graphically display what I needed to know.

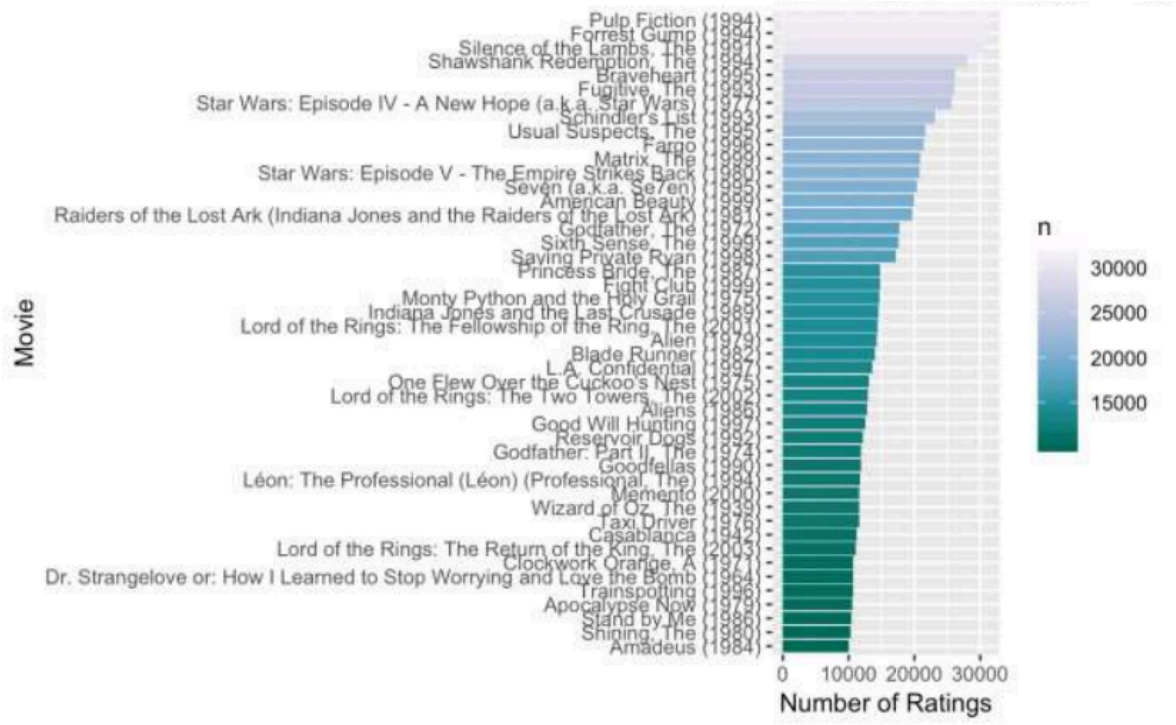


According to the graph above, the movie genre known as “Drama” has the highest percentage of movie ratings while the movie genres known as “IMAX” and “no genres listed” have the smallest. Next, I examined the average movie rating of each genre and found that horror movies contained the smallest average rating while film noir contained the largest as shown below.

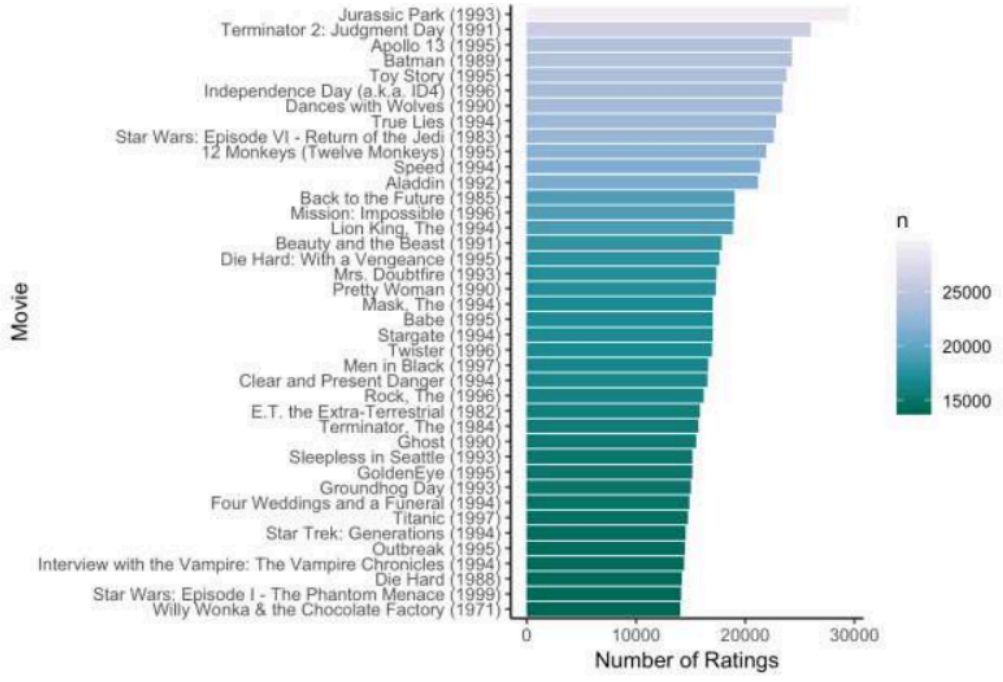


The Ratings of the Movies based on the quality of quantity of Movie Ratings

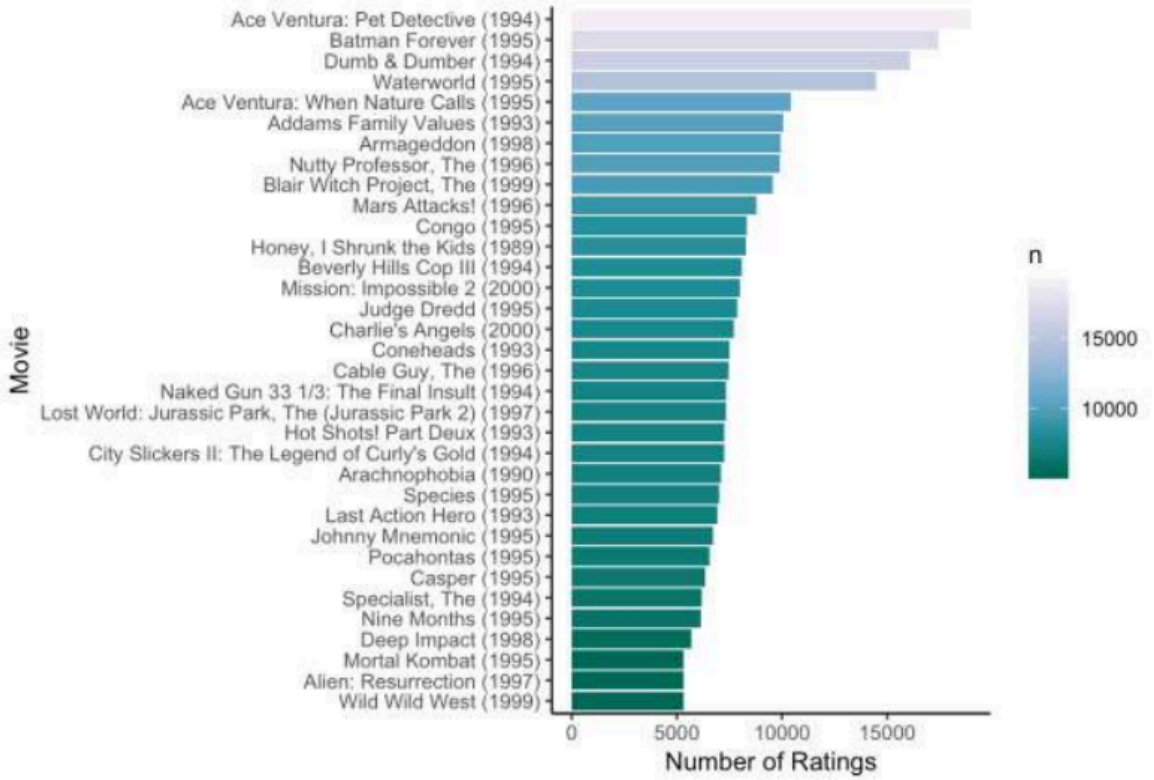
To do this, I graphically displayed the movies that had average ratings of 4 and was known to have over 10000 movie ratings. By doing this, I found that the movie known as “Pulp Fiction” had the highest mean as shown below.



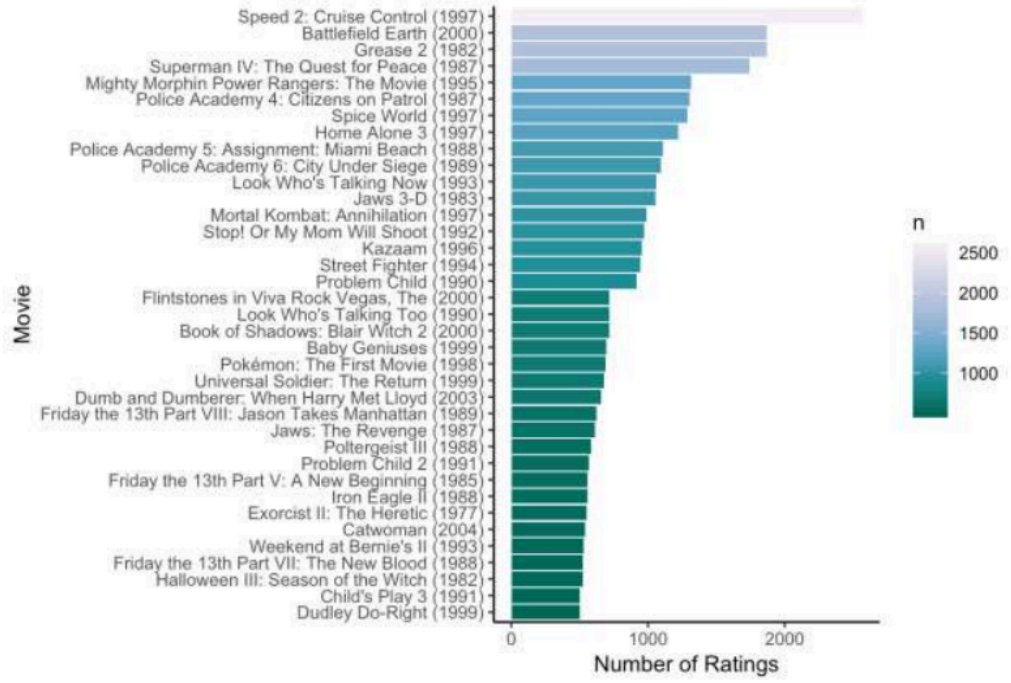
I also cleaned the graphic display up a bit more and observed the movies that contained an average movie rating between 3 and 4 and contained more than 10000 ratings. Here, is that graph:



I observed the movies that had an average movie rating between the numbers of 2 and 3 and contained more than 5000 ratings.

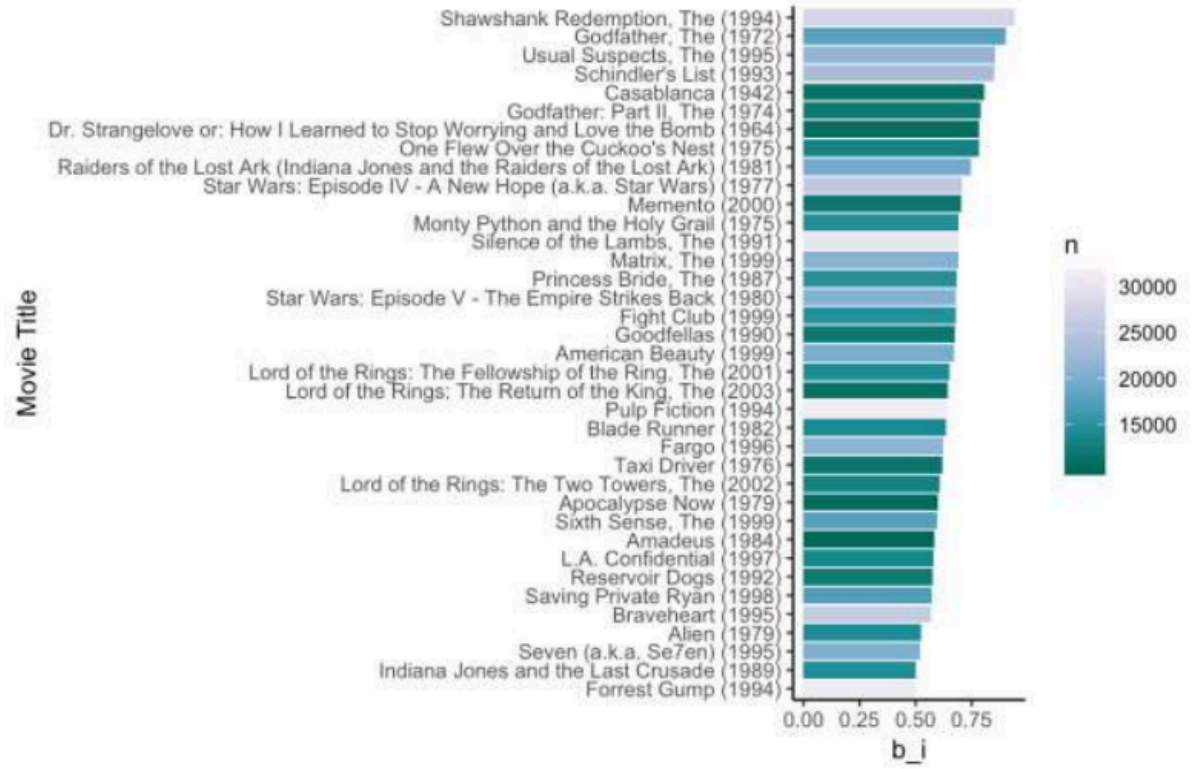


Next, I examined the movies that had a movie rating that was greater than 500, but less than 2 as seen below.

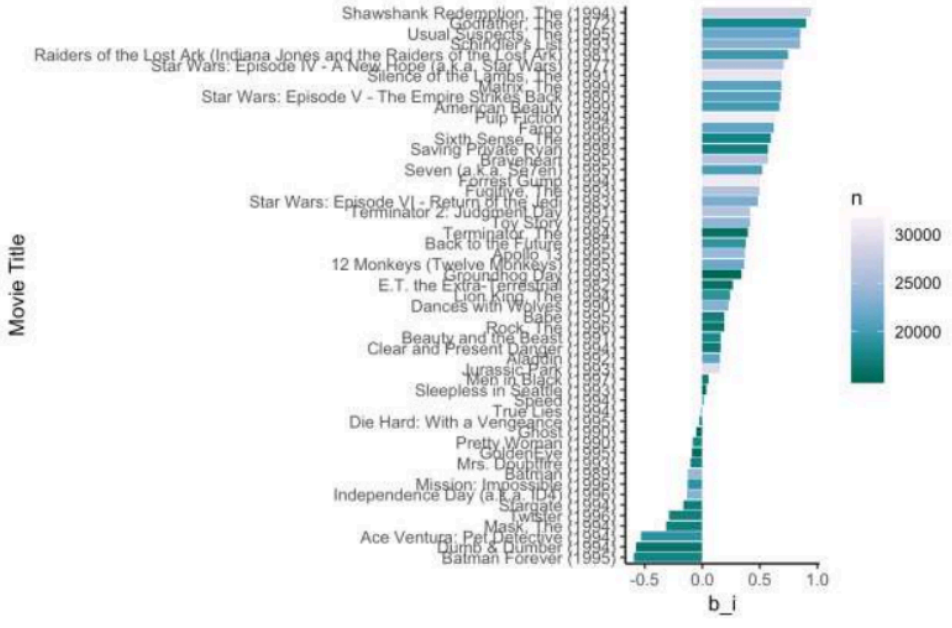


Calculate Least Squares for the Movie ID

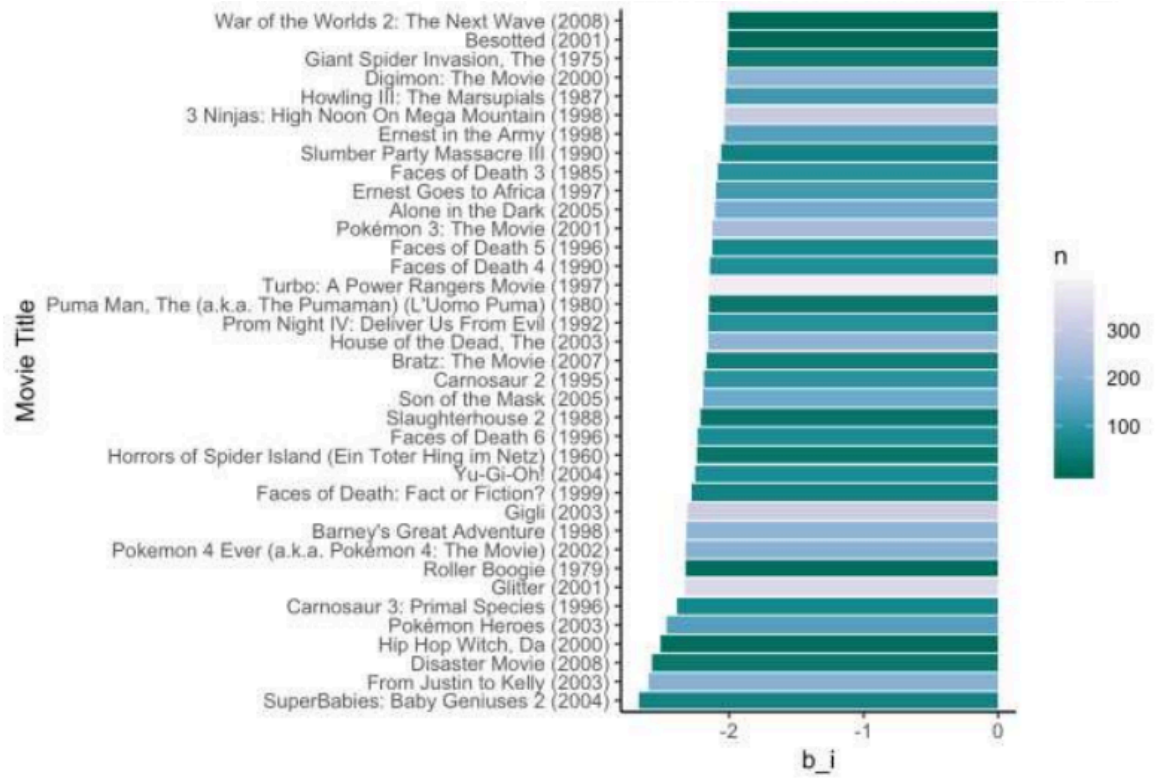
What movies have a large quantity of ratings and the ratings are larger than the average p?



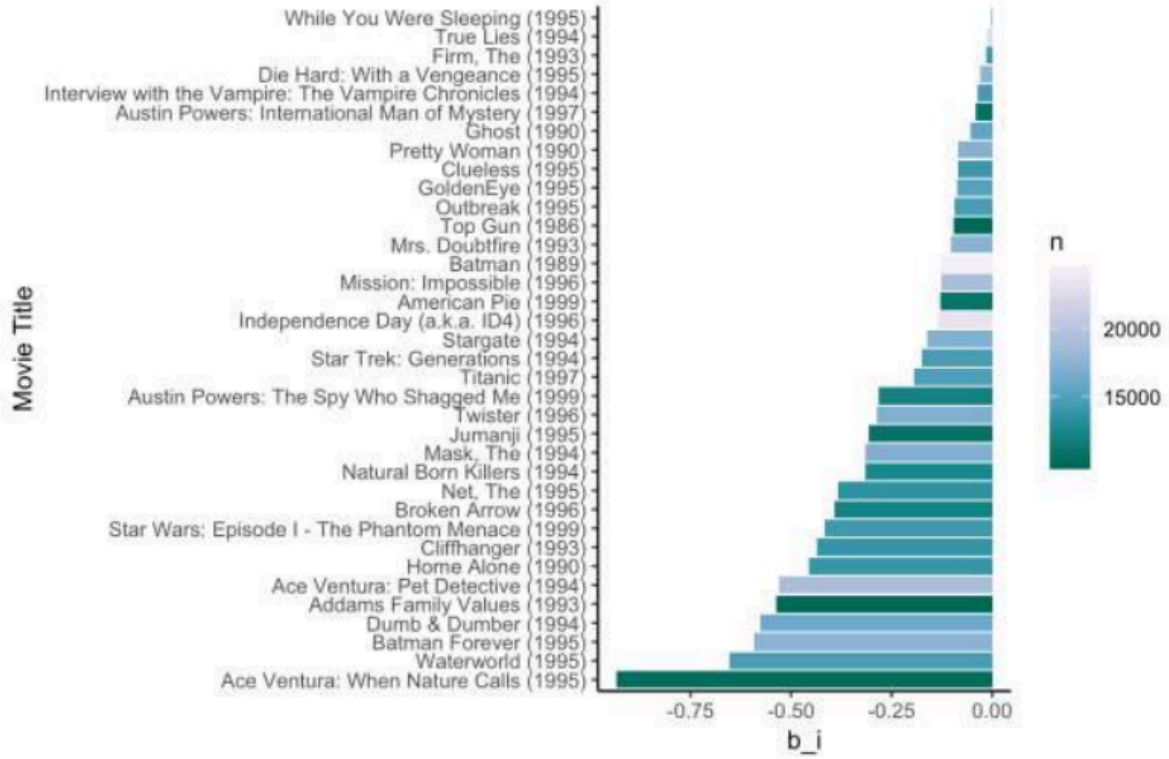
The Regular Averages for the Movies in the data with more than 20000 movie ratings



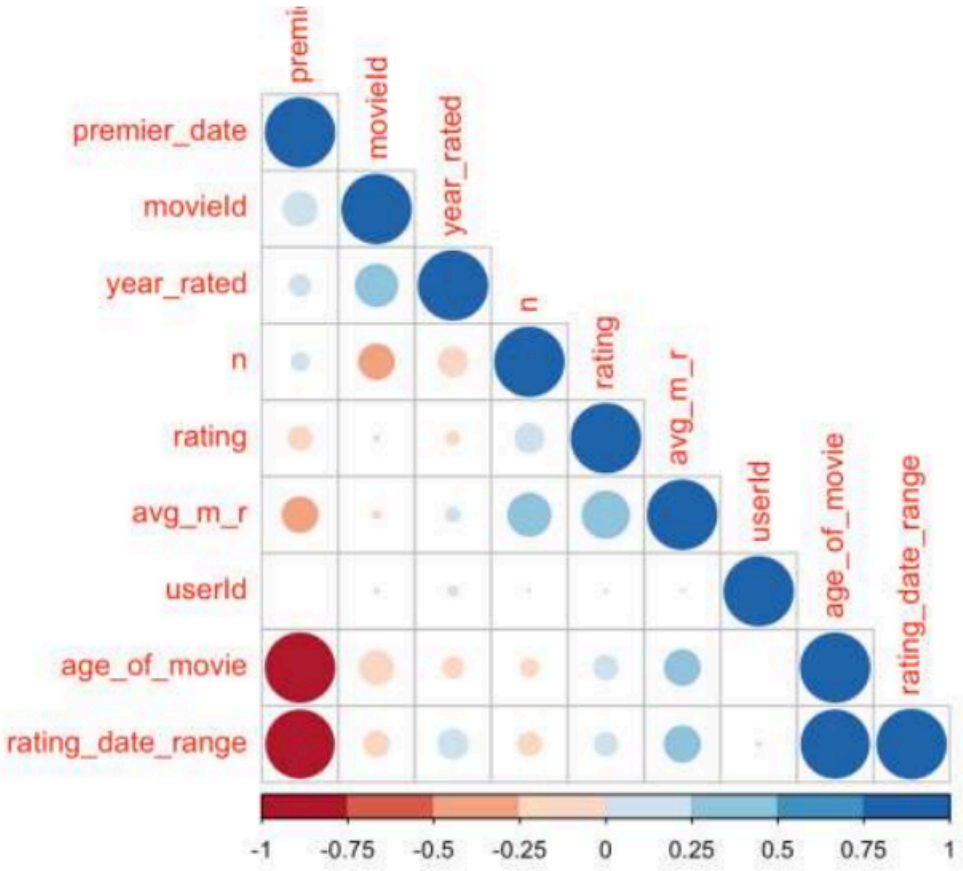
The Regular Averages for the Movies that Contain a movie rating of less than 2



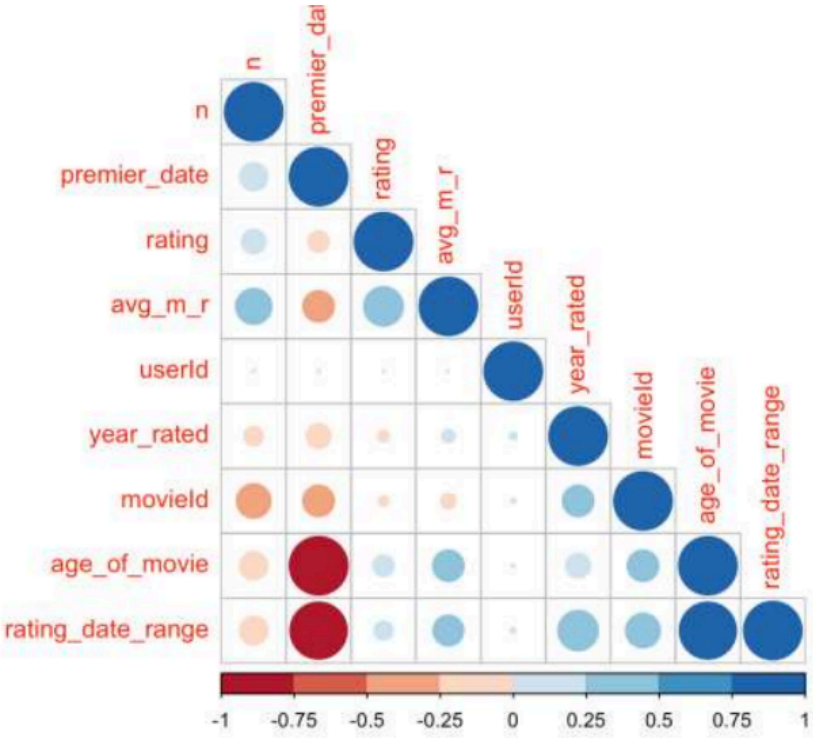
Movies that have a mean rating of less than p



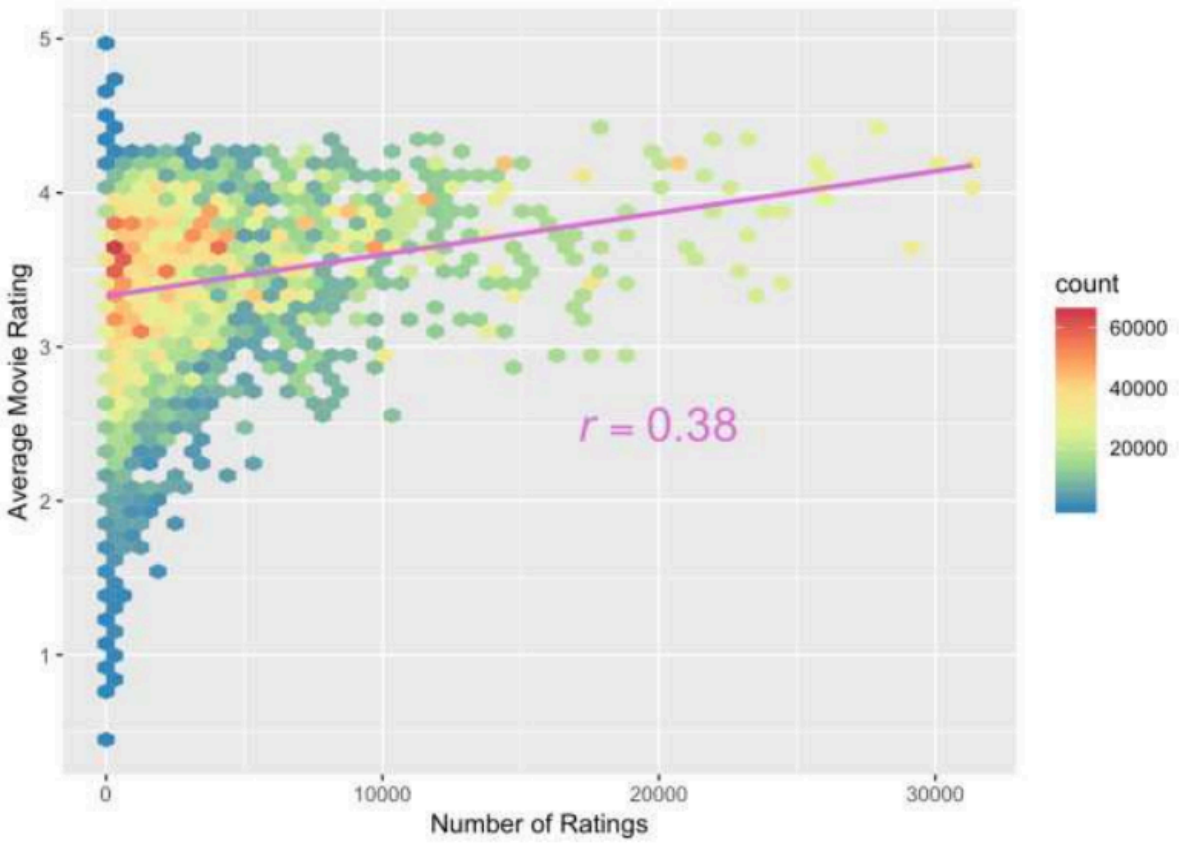
Is there any pattern of correlation between movieid, userid, the age of the movie, the ratings of the movies, and how many ratings the movie has?



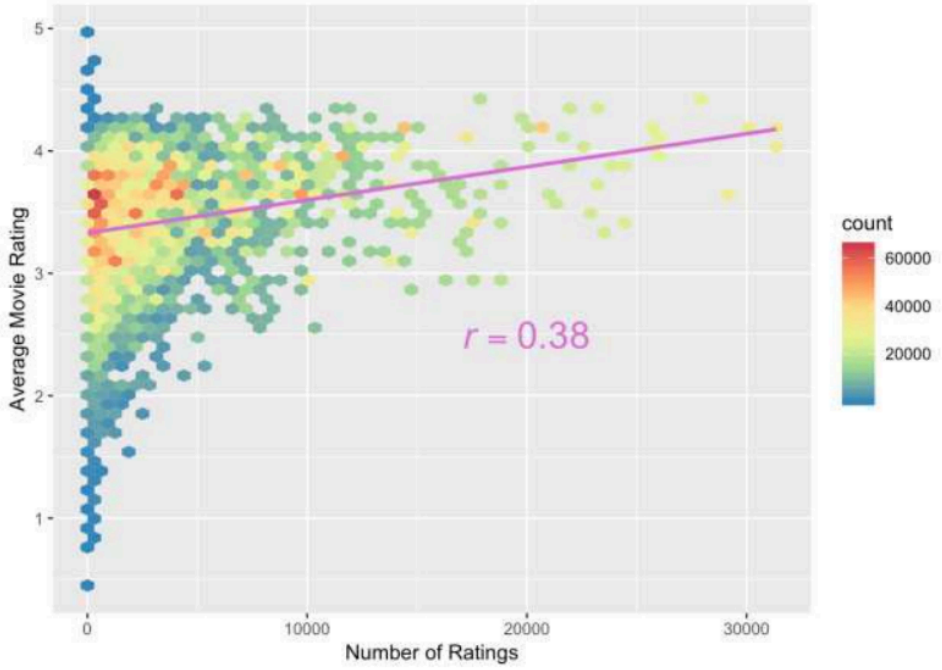
Age of the Movie and its Effect



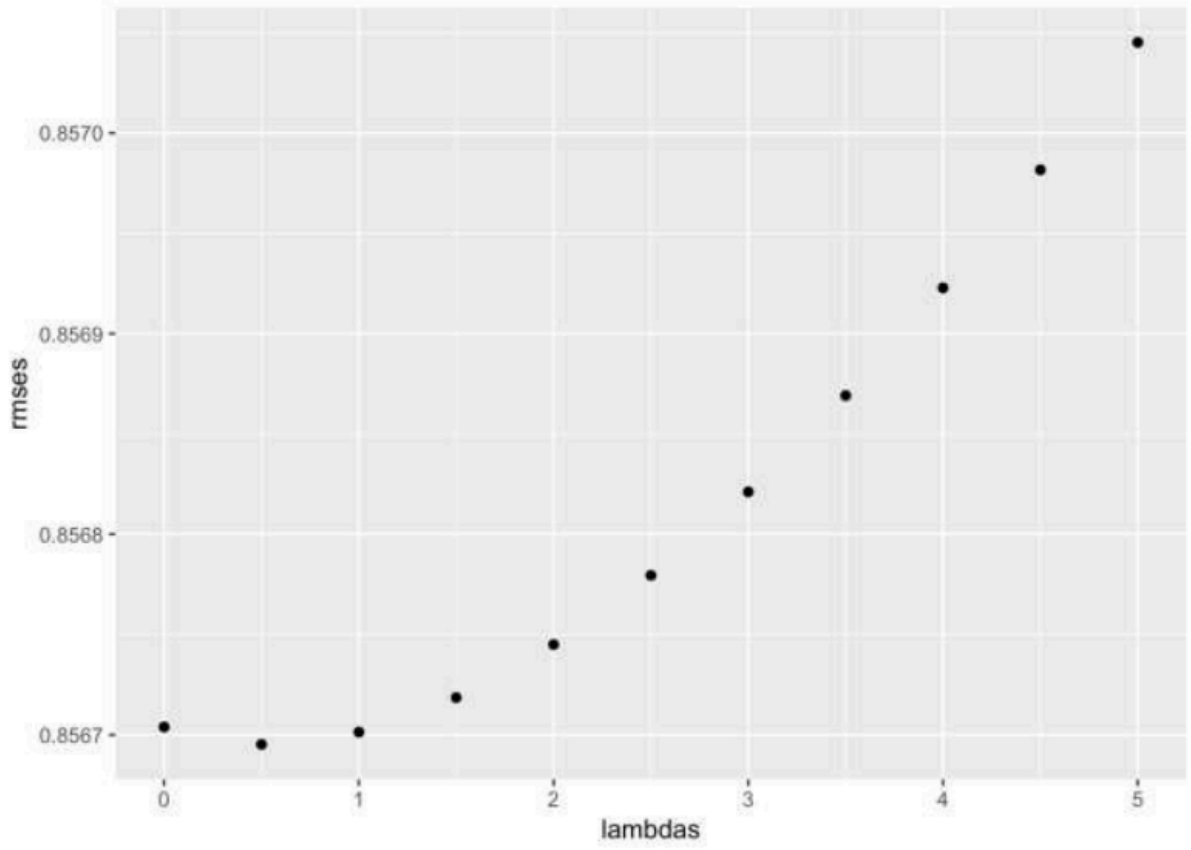
Is there any pattern and/or relationship between the average movie ratings and the number of movie ratings?



Average Movie Rating vs. The Age of the Movie



Computing and graphically displaying RMSE



By this, I achieved an RMSE score of 0.8252.