Movielens Capstone Project HarvardX

MovieLens Introduction

The MovieLens data set was collected by GroupLens Research. Can we predict movie ratings based on user preferance, age of a movie? Using the MovieLens data set and penalized least squares, the following R script calculates the RMSE based on user ratings, movieId and the age of the movie.

The MovieLens data set contains 10000055 rows, 10677 movies, 797 genres and 69878 users.

The steps performed for analysis of the data –
- Created an age of movie column
- Graphic displays of movie, users and ratings in order to find a pattern or insight to the behavior of the data.
- Explored Genres to determine if ratings could be predicted by genre.
- Explored the Coefficient of Determination R-Squared
- Graphically explored the linear correlation coefficient, r-value
- Calculate RMSE based on movieId, userId, and age of the movie.

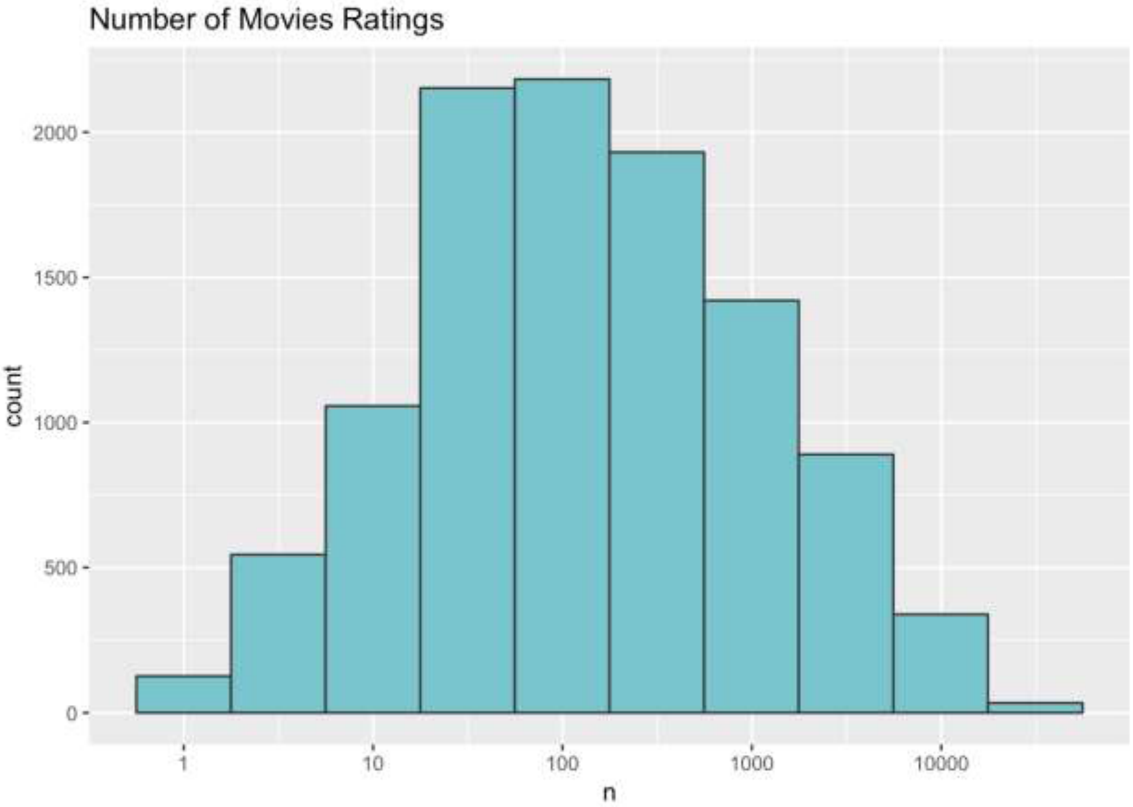
After exploring the movies through graphical representations and calculating RMSE, I found the best predictor for ratings was movieId, userId. The age of the movie didn't change the rmse.

The final RMSE is 0.8252

In order to determine if age of the movie is a factor for predicting rating, I extracted the premier date of the movie, and then calculated the age of the movie. I also looked at individual genres for genre effect, as well as, effects of user ratings.

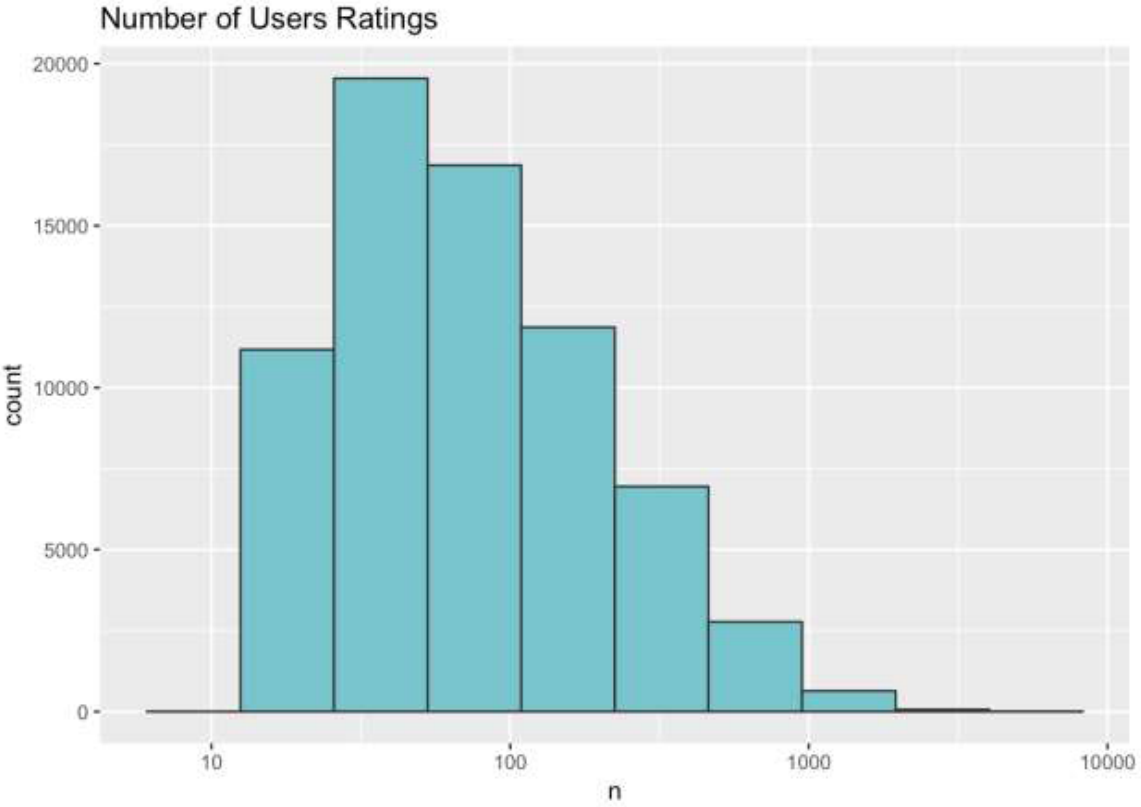
Graph the data

What is the distribution of Movie Ratings?



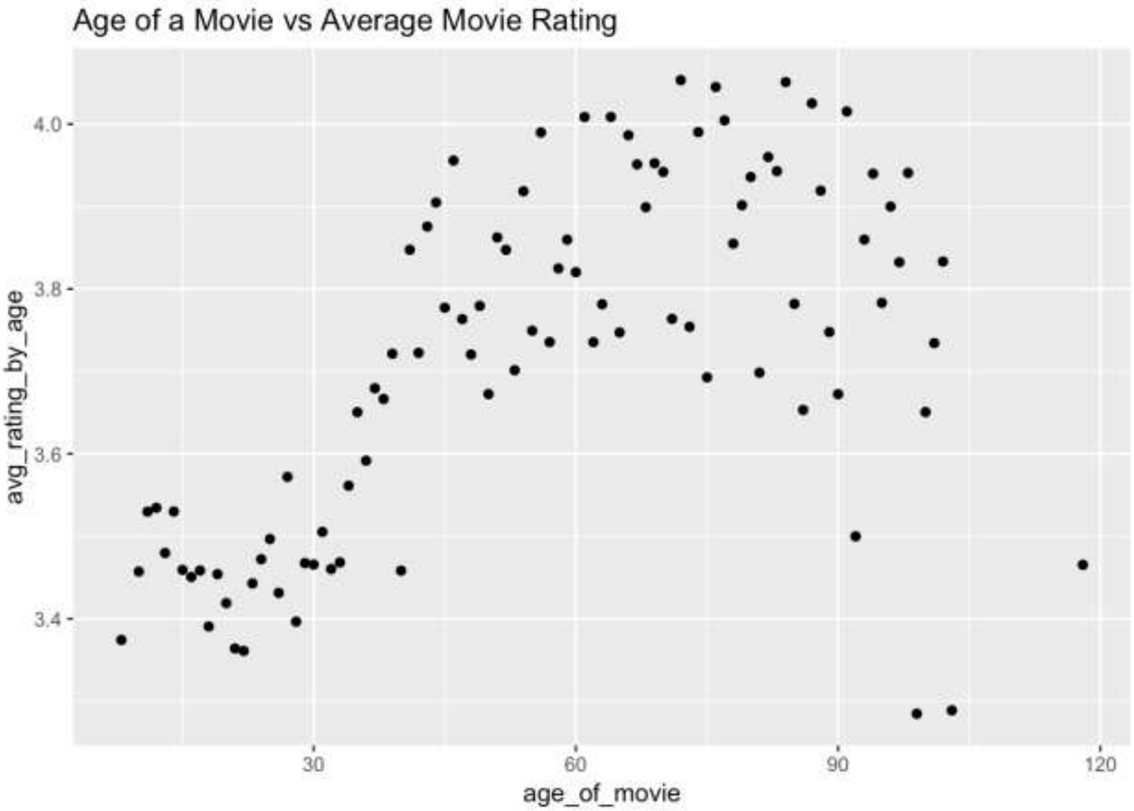
As you can see from the graph, some movies were rated once while other movies were rated more than 10,000 times.

What is the distribution of ratings by user?

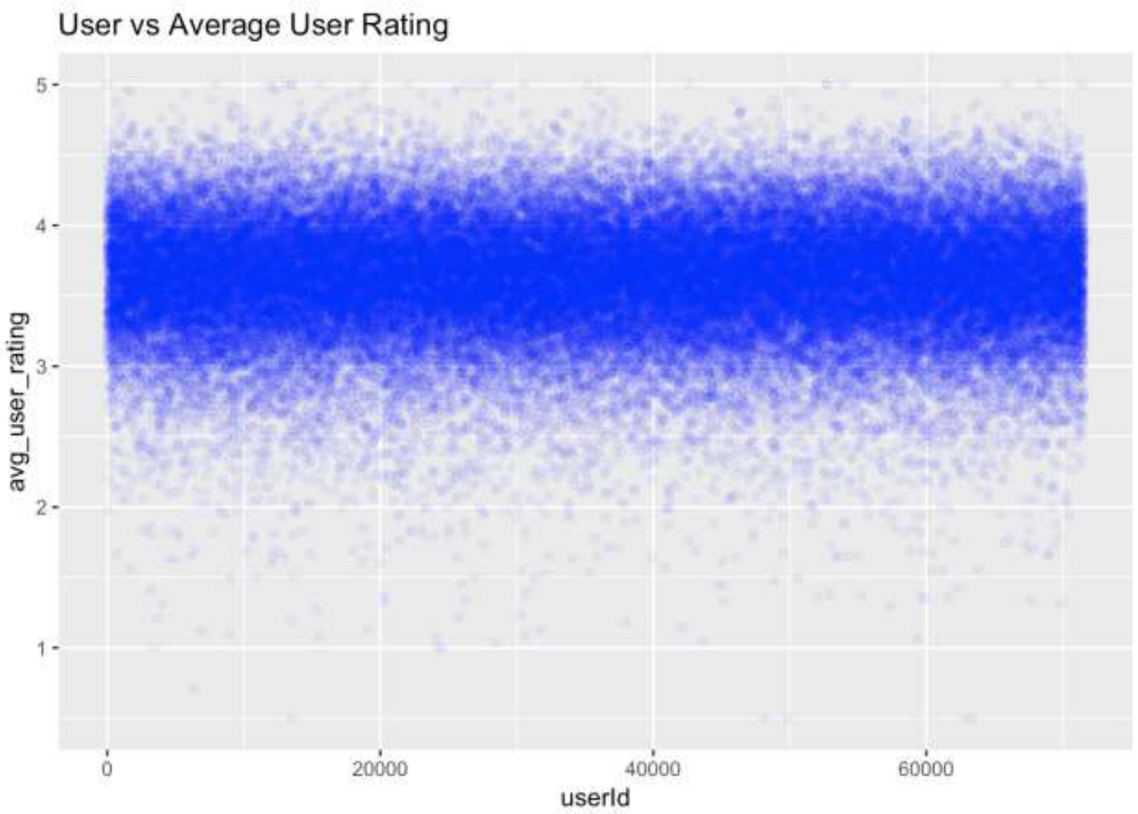


Number of ratings per users is slightly skewed right. 75% of users rated 141 movies or less.

Is there a relationship to the age of a movie and the movies average rating?



The newer the movie, the lower its average rating. The above plot, also, shows more variability as movies age. Average Ratings increase from 30 years old up to approximately older 90 years old, then the ratings drop. Probably due to fewer ratings.

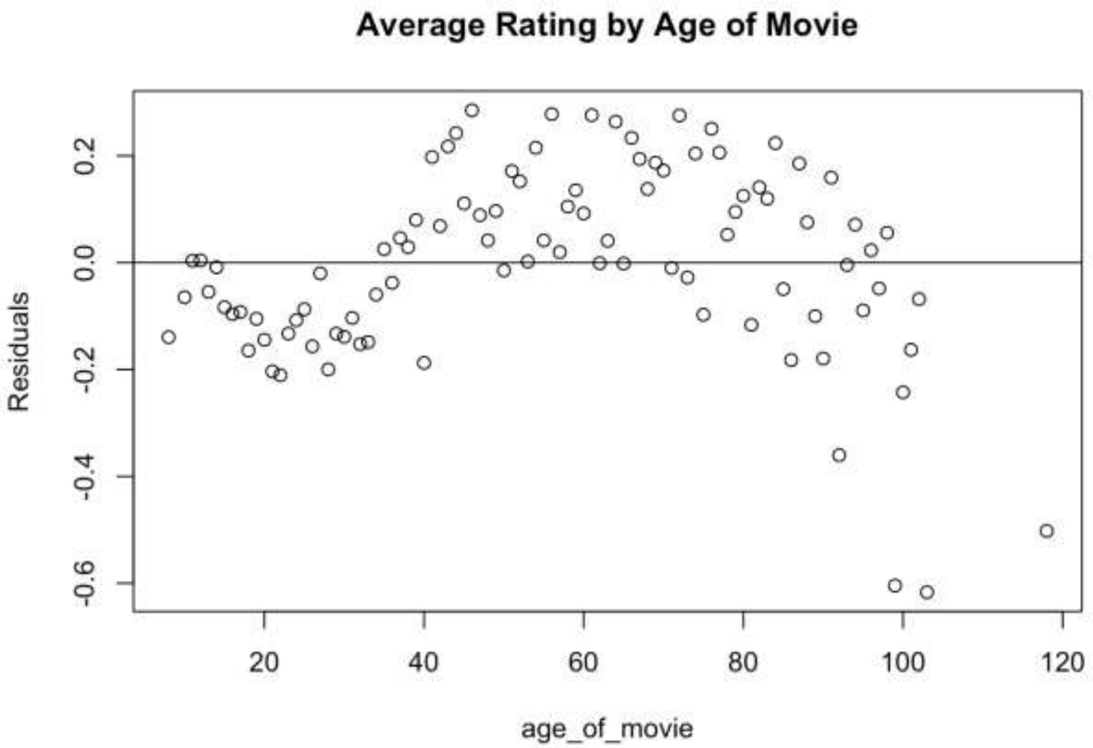


We can see average ratings by user are pretty consistent between 2.5 and 4.5

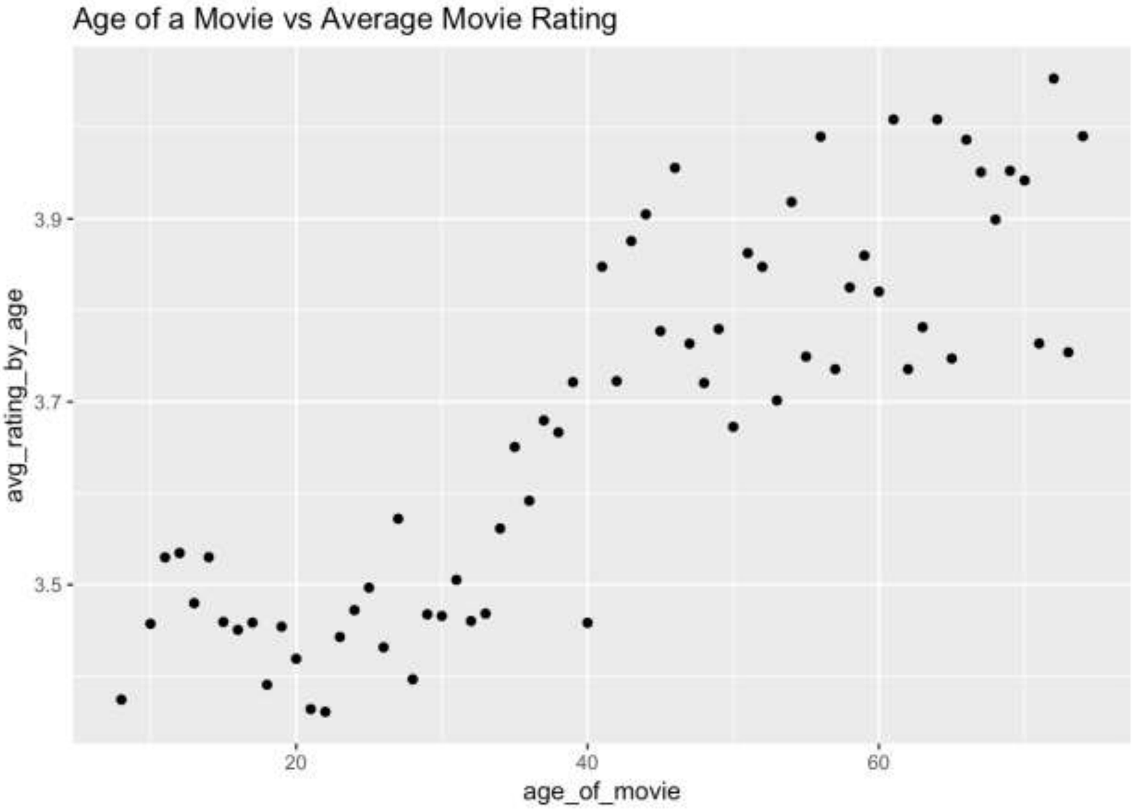
Linear model of the age of a movie vs average rating shows an R-Squared value of 0.30 meaning there is some correlation but its not a strong correlation

```
## 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
```

The residual plot is consistent around zero except for the older movies.

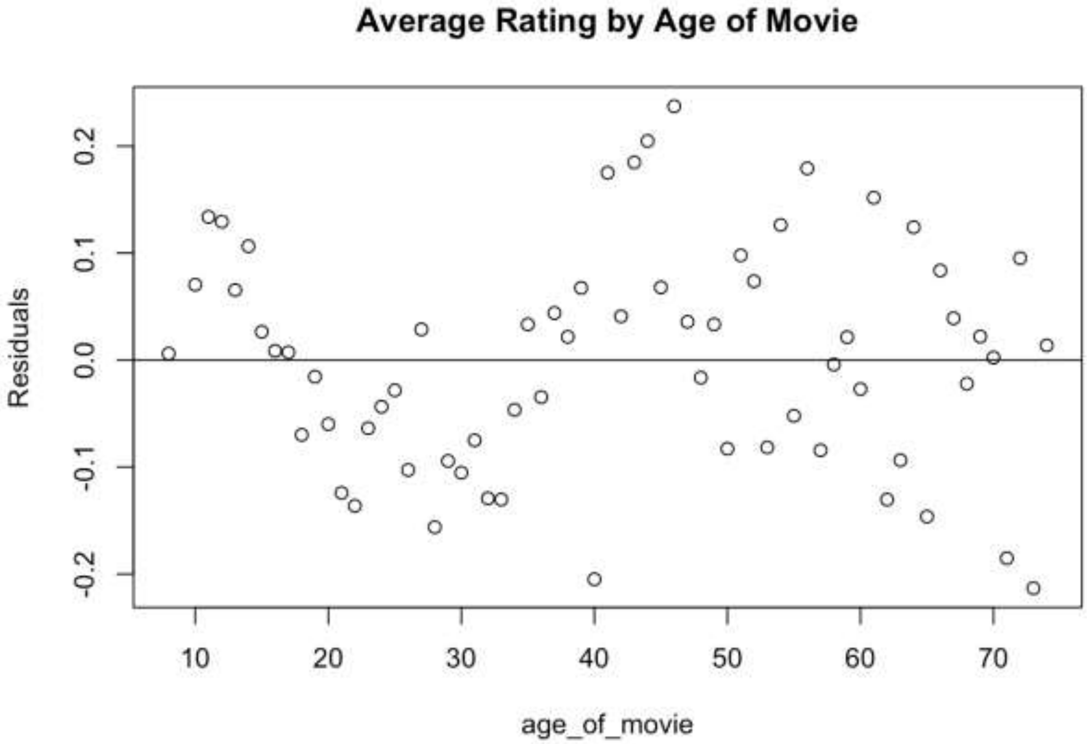


Does the R-Square improve if older movies are removed?

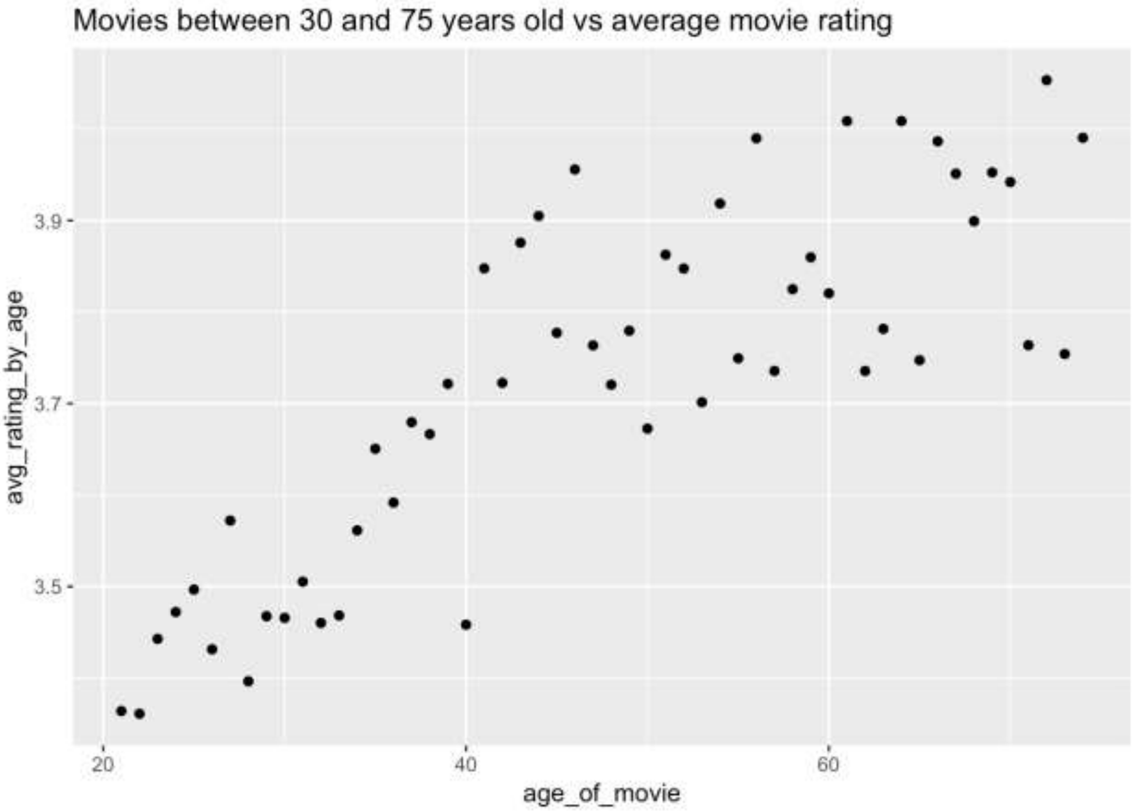


After removing movies greater than 75 years old, the R-Square improves and is now 0.745

```
age_lessthan75_rating.lm <- lm(avg_rating_by_age ~ age_of_movie, data =
age_of_movie_less_than75)
summary(age_lessthan75_rating.lm)
## Call:
## lm(formula = avg_rating_by_age ~ age_of_movie, data =
age_of_movie_less_than75)
## Residuals:
## Min 1Q Median 3Q Max
## -0.21323 -0.07992 0.00663 0.06785 0.23721
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.2946266 0.0307644 107.09 <2e-16 ***
## age_of_movie 0.0092153 0.0006738 13.68 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1044 on 64 degrees of freedom
## Multiple R-squared: 0.7451, Adjusted R-squared: 0.7411
## F-statistic: 187.1 on 1 and 64 DF, p-value: < 2.2e-16
The residual plot now ranges from -0.2 to 0.2.
```



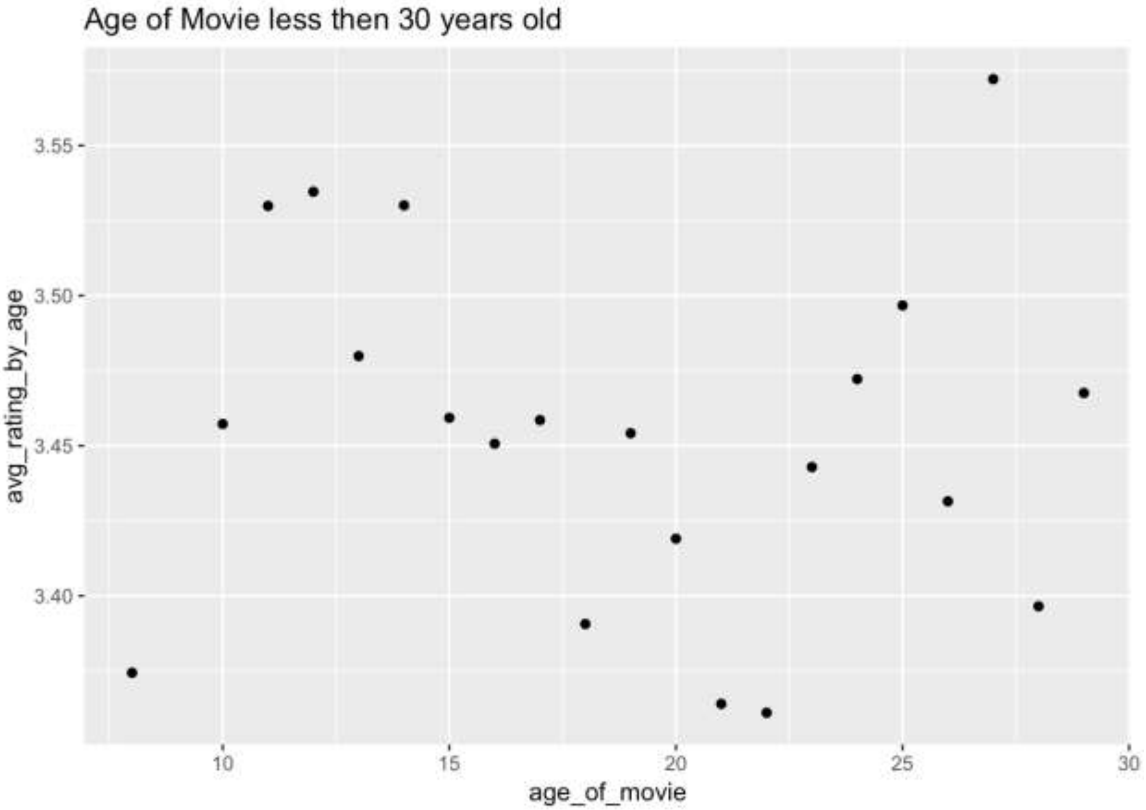
Since movies less than 20 years old appeared to have a negative linear trend, let’s look at movies between 20 and 75 years old as the graph looks more linear in that time frame.



The plot above appears to be a linear trend; however, the r-square decreased to 0.69.

```
## Residuals:
## Min 1Q Median 3Q Max
## -0.235567 -0.077940 -0.009169 0.068137 0.246532
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.2313562 0.0473472 68.25 < 2e-16 ***
## age_of_movie 0.0103880 0.0009471 10.97 3.88e-15 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1085 on 52 degrees of freedom
## Multiple R-squared: 0.6982, Adjusted R-squared: 0.6924
## F-statistic: 120.3 on 1 and 52 DF, p-value: 3.882e-15
```

What is the distribution of movies that are less than 30 years old?



For movies less than 30 years old there appears to be quite a bit of variation. We can see from the linear model that r-squared is nearly zero.

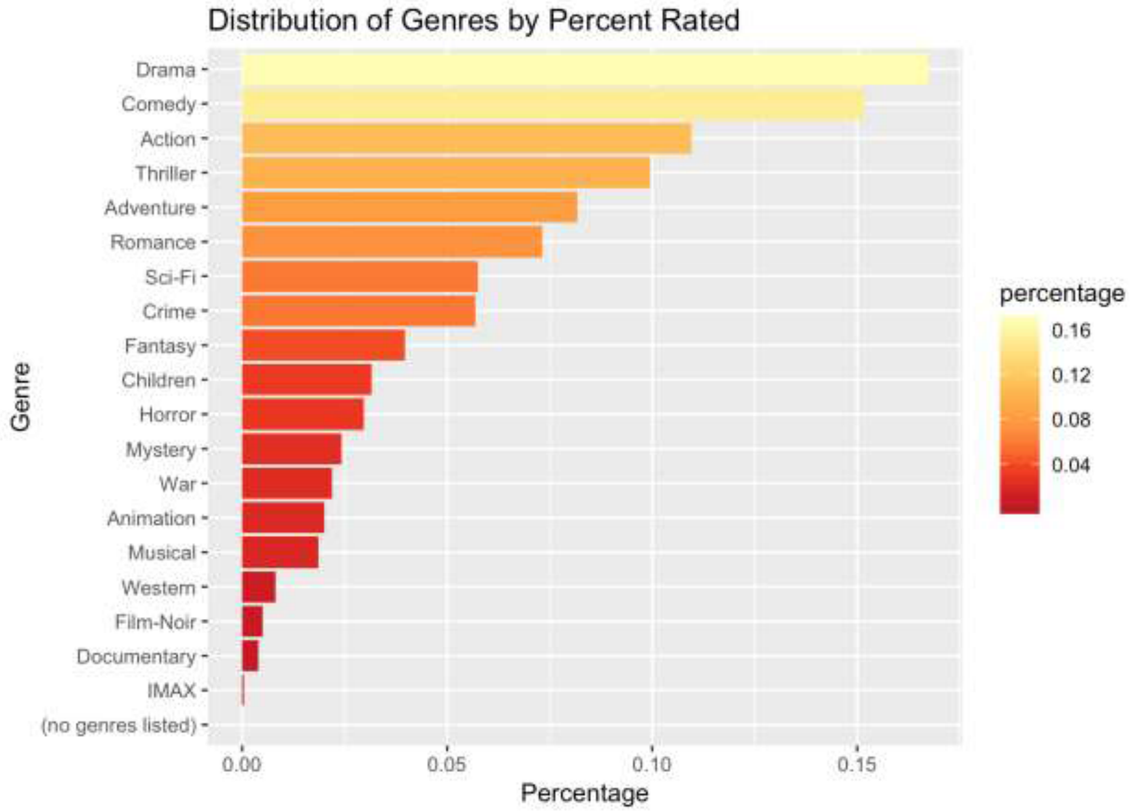
```
## Residuals:
## Min 1Q Median 3Q Max
## -0.091058 -0.034589 -0.000233 0.021613 0.123826
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.4688469 0.0420239 82.545 <2e-16 ***
## age_of_movie -0.0007611 0.0021095 -0.361 0.722
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05933 on 19 degrees of freedom
## Multiple R-squared: 0.006805, Adjusted R-squared: -0.04547
## F-statistic: 0.1302 on 1 and 19 DF, p-value: 0.7222
```

The age of a movie did seem to affect the outcome of the average rating. This is possibly due to a higher number of ratings for older movies.

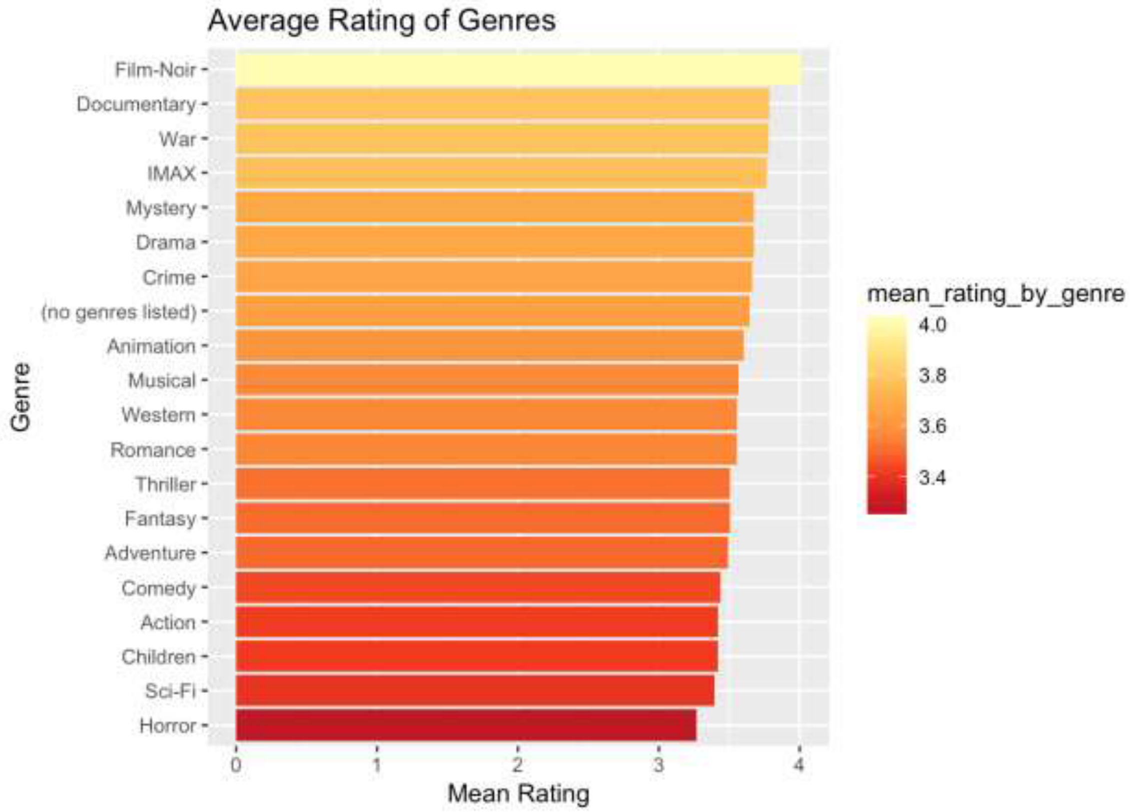
Do Genres have an effect on ratings?

I extracted the genres from the data with the idea to do an analysis on each genre and counted the number of movies in each genre

```
## genres n
## <chr> <int>
## 1 (no genres listed) 7
## 2 Action 2560545
## 3 Adventure 1908892
## 4 Animation 467168
## 5 Children 737994
## 6 Comedy 3540930
## 7 Crime 1327715
## 8 Documentary 93066
## 9 Drama 3910127
## 10 Fantasy 925637
## 11 Film-Noir 118541
## 12 Horror 691485
## 13 IMAX 8181
## 14 Musical 433080
## 15 Mystery 568332
## 16 Romance 1712100
## 17 Sci-Fi 1341183
## 18 Thriller 2325899
## 19 War 511147
## 20 Western 189394
```



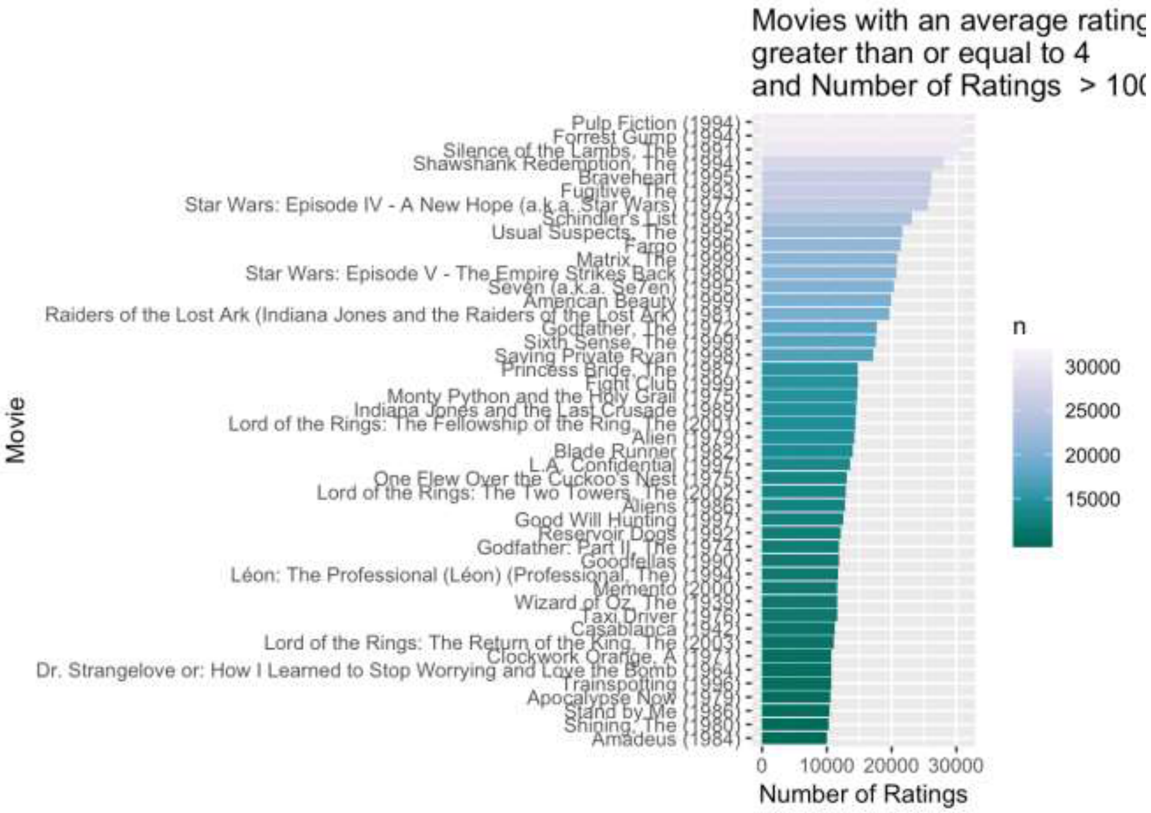
Drama had the highest percentage of ratings, with Documentary IMAX, and (no genres listed) having the smallest percentage of ratings.



Film Noir had the highest average rating, while Horror had the lowest average rating.

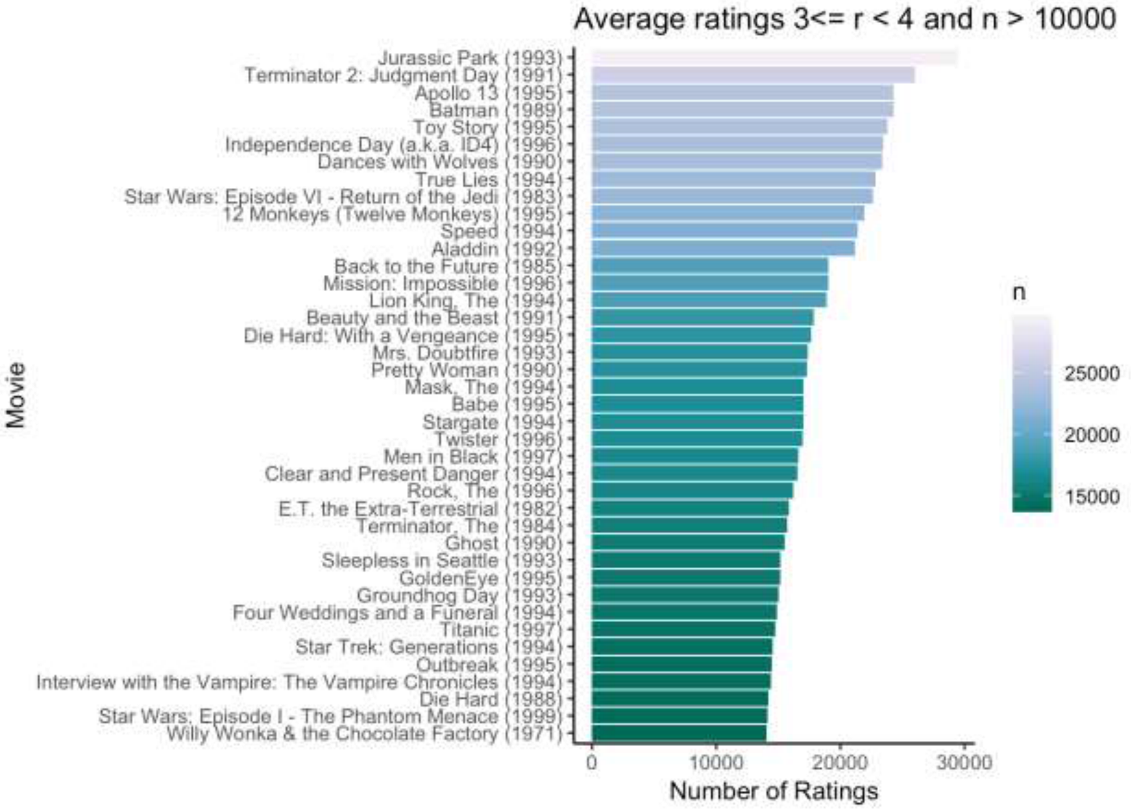
Explore movie ratings based on number of ratings and value of the rating.

Graph of movies with more than 10000 ratings and a mean rating greater than 4.

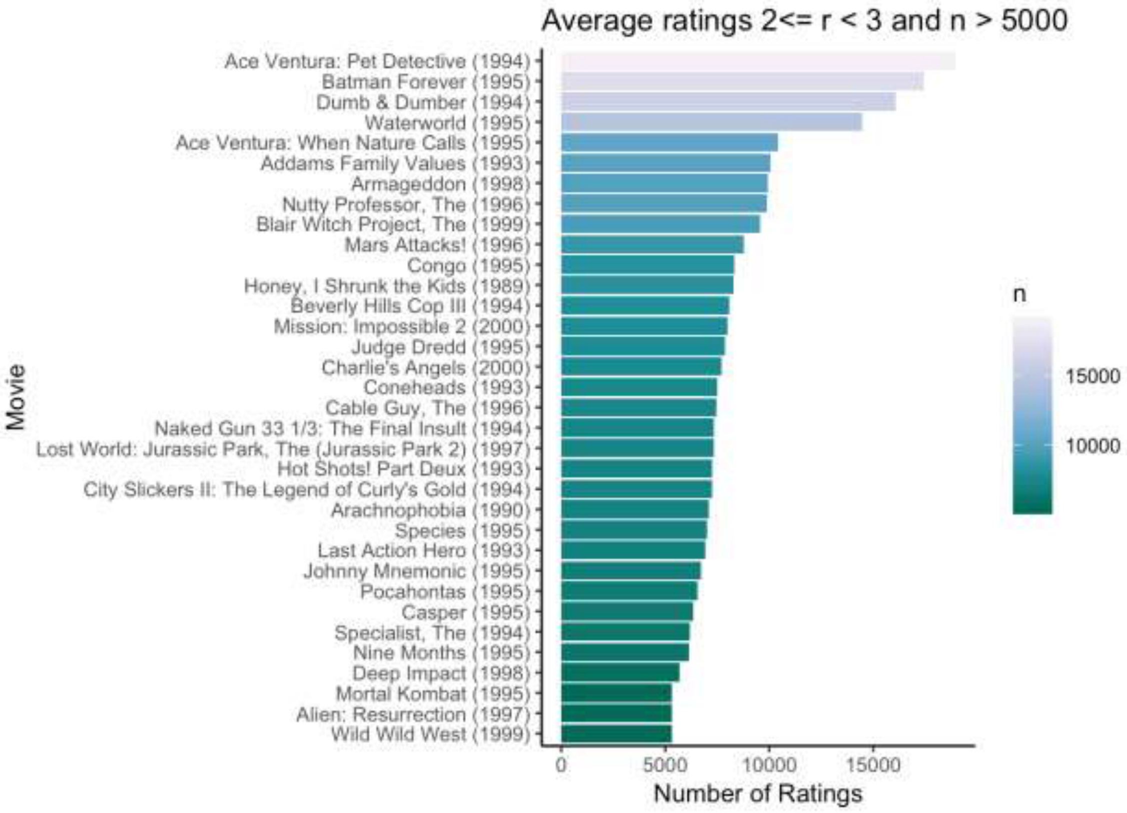


Pulp Fiction had the highest average rating for movies who were rated more than 10000 time.

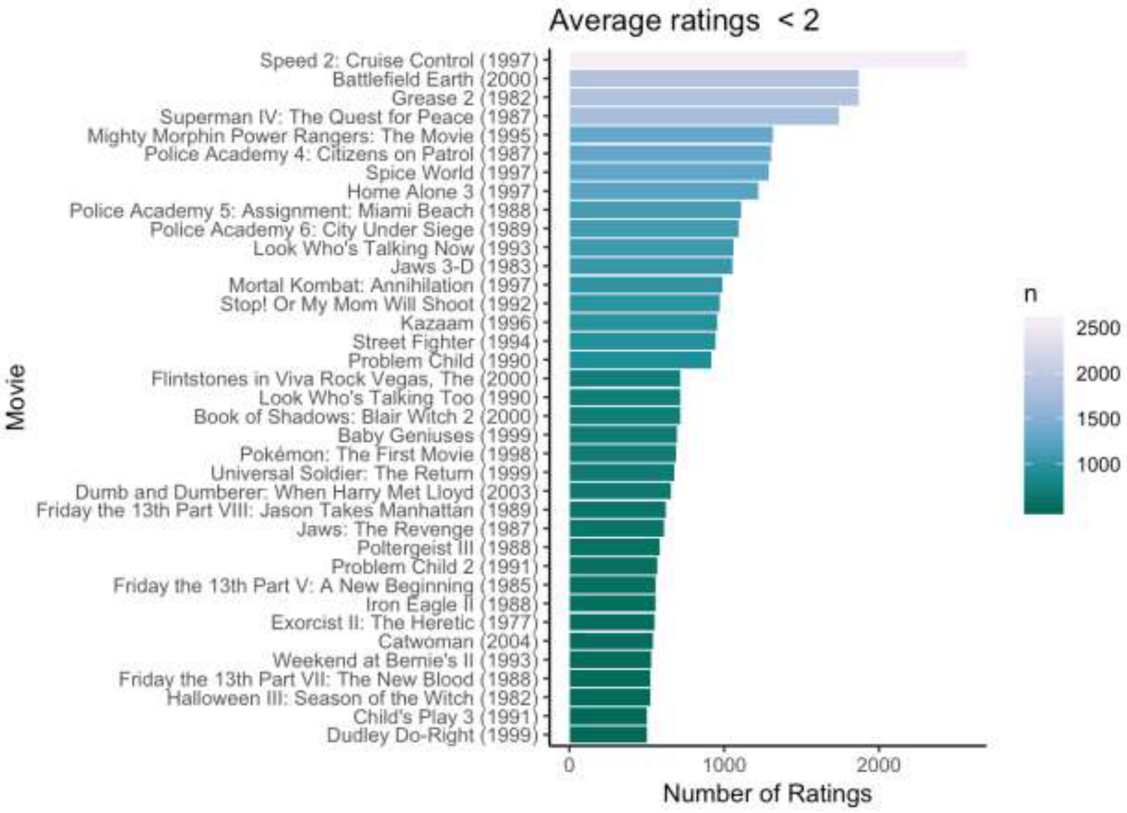
Examine Movies with ratings between 3 and 4 and more than 10000 ratings



Movies with an average rating between 2 and 3. Since movies with lower rating also have fewer ratings, let’s look at number of ratings greater than 5000.

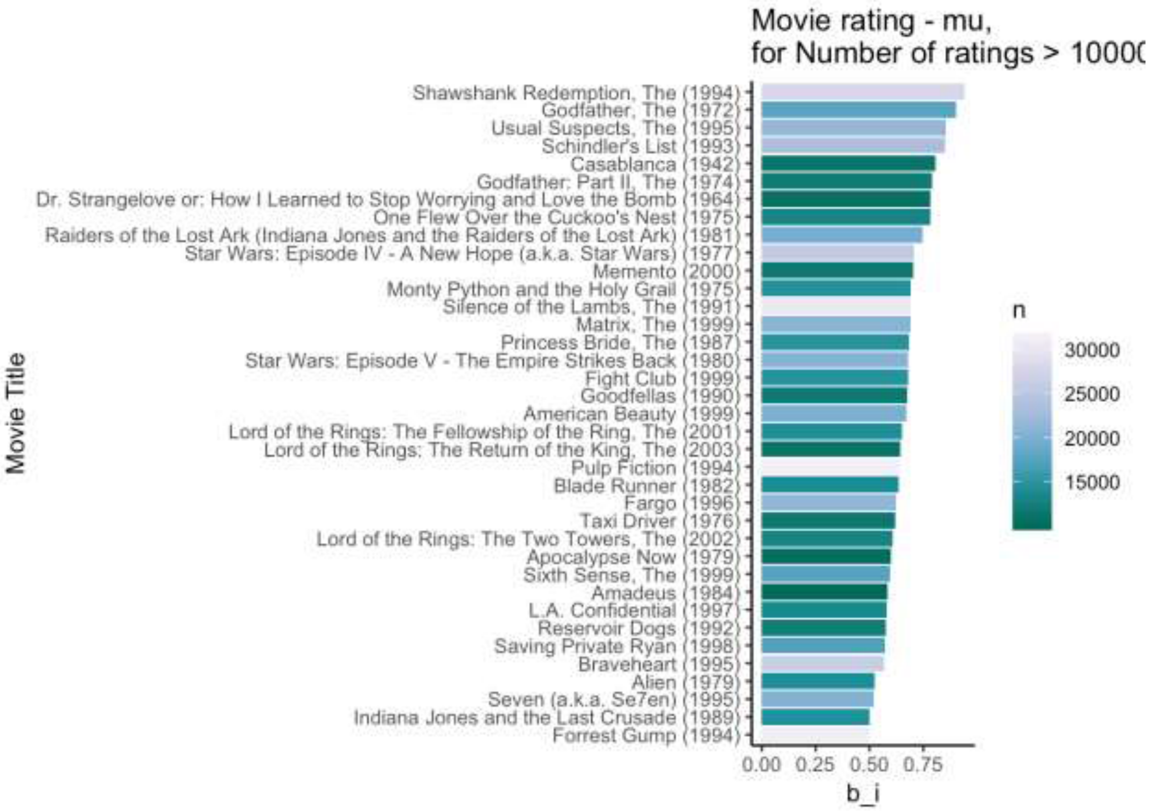


Movies with a rating less than 2 and number of ratings greater than 500.

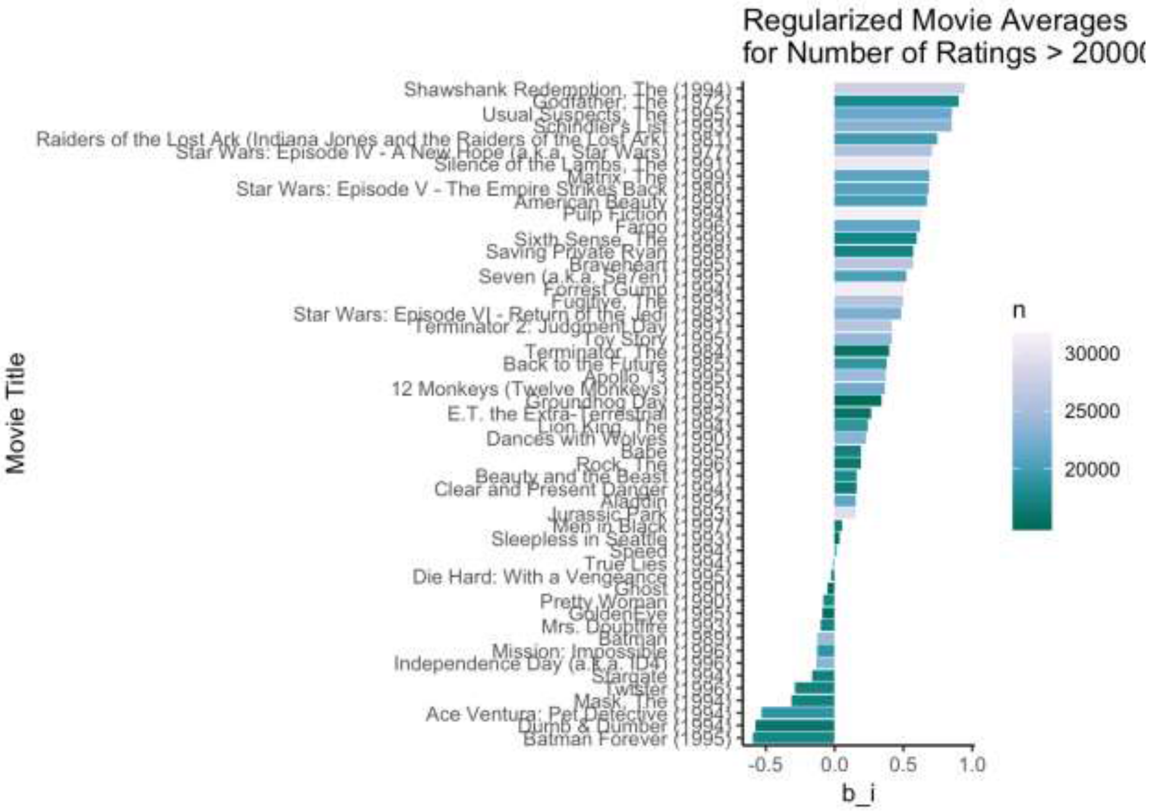


Compute the least squares for movieId

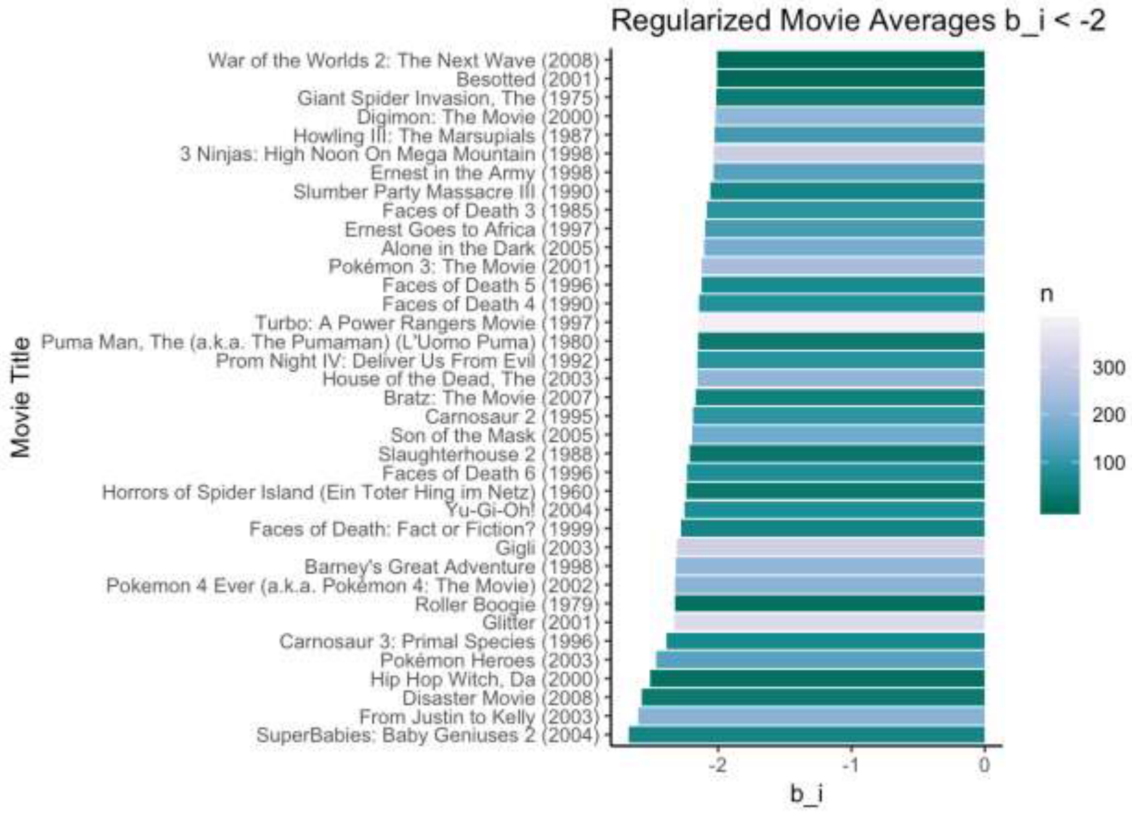
Which movies have many ratings and a rating larger than the average mu.



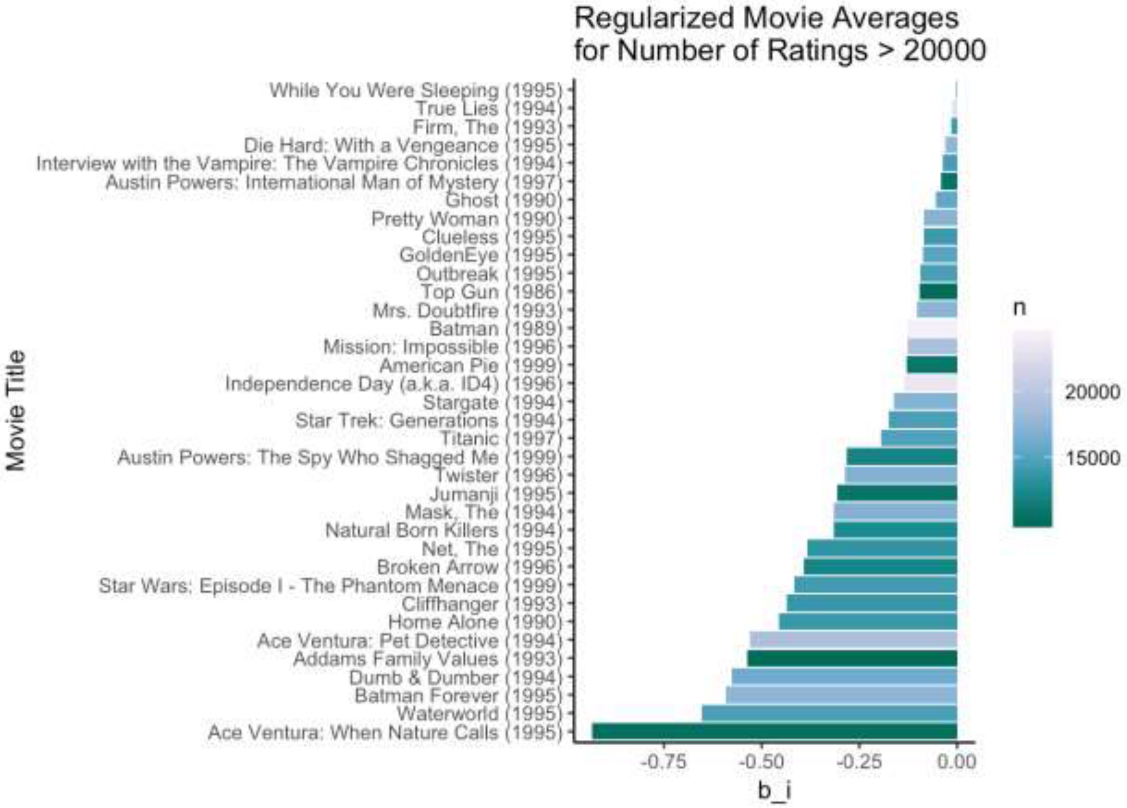
Regularized Movie Averages for movies with more than 20000 ratings.



Regularized Movie Averages for the movies with regularized ratings less than 2

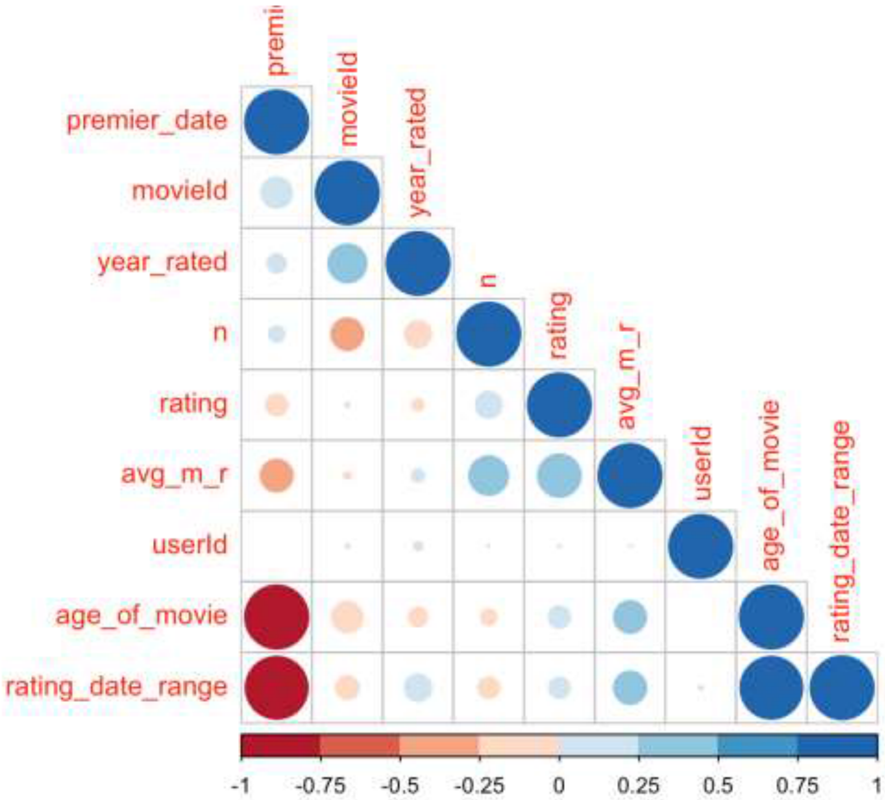

Regularized Movie Averages b_i < -2
Movie Title
War of the Worlds 2: The Next Wave (2008)
Besotted (2001)
Giant Spider Invasion, The (1975)
Digimon: The Movie (2000)
Howling III: The Marsupials (1987)
3 Ninjas: High Noon On Mega Mountain (1998)
Ernest in the Army (1998)
Slumber Party Massacre III (1990)
Faces of Death 3 (1985)
Ernest Goes to Africa (1997)
Alone in the Dark (2005)
Pokémon 3: The Movie (2001)
Faces of Death 5 (1996)
Faces of Death 4 (1990)
Turbo: A Power Rangers Movie (1997)
Puma Man, The (a.k.a. The Pumaman) (L'Uomo Puma) (1980)
Prom Night IV: Deliver Us From Evil (1992)
House of the Dead, The (2003)
Bratz: The Movie (2007)
Carnosaur 2 (1995)
Son of the Mask (2005)
Slaughterhouse 2 (1988)
Faces of Death 6 (1996)
Horrors of Spider Island (Ein Toter Hing im Netz) (1960)
Yu-Gi-Oh! (2004)
Faces of Death: Fact or Fiction? (1999)
Gigli (2003)
Barney's Great Adventure (1998)
Pokemon 4 Ever (a.k.a. Pokémon 4: The Movie) (2002)
Roller Boogie (1979)
Glitter (2001)
Carnosaur 3: Primal Species (1996)
Pokémon Heroes (2003)
Hip Hop Witch, Da (2000)
Disaster Movie (2008)
From Justin to Kelly (2003)
SuperBabies: Baby Geniuses 2 (2004)
-2
-1
0
b_i
n
300
200
100

Movies with number of ratings larger than 1000 and regularized average less than 0. Movies with average ratings less than mu.

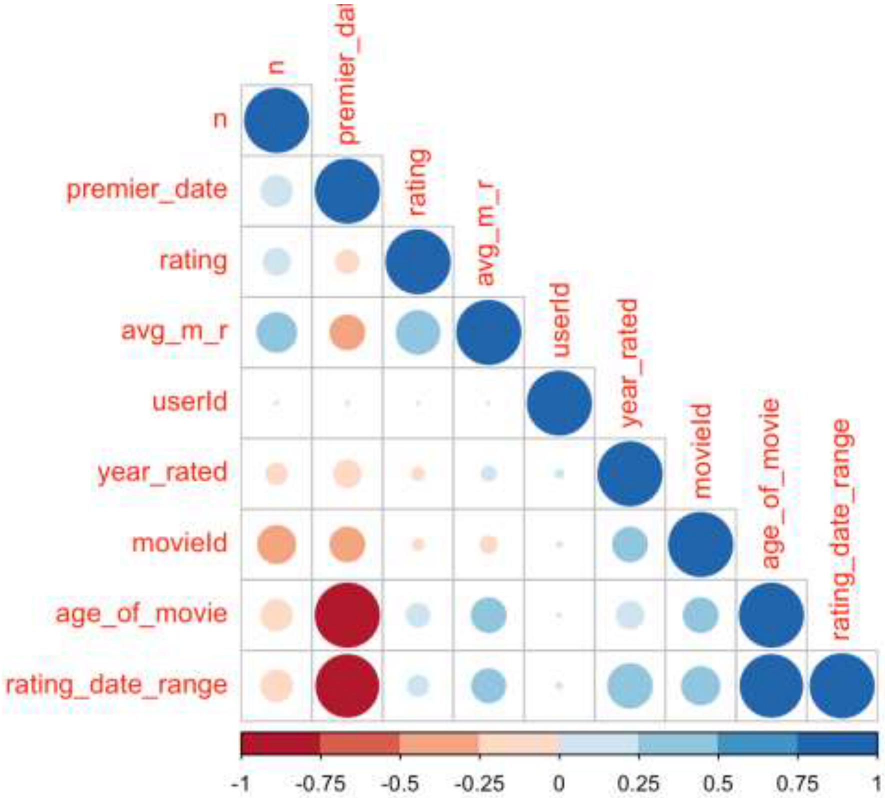


Is there a correlation between ratings, users, movieId, age of movie and number of ratings?

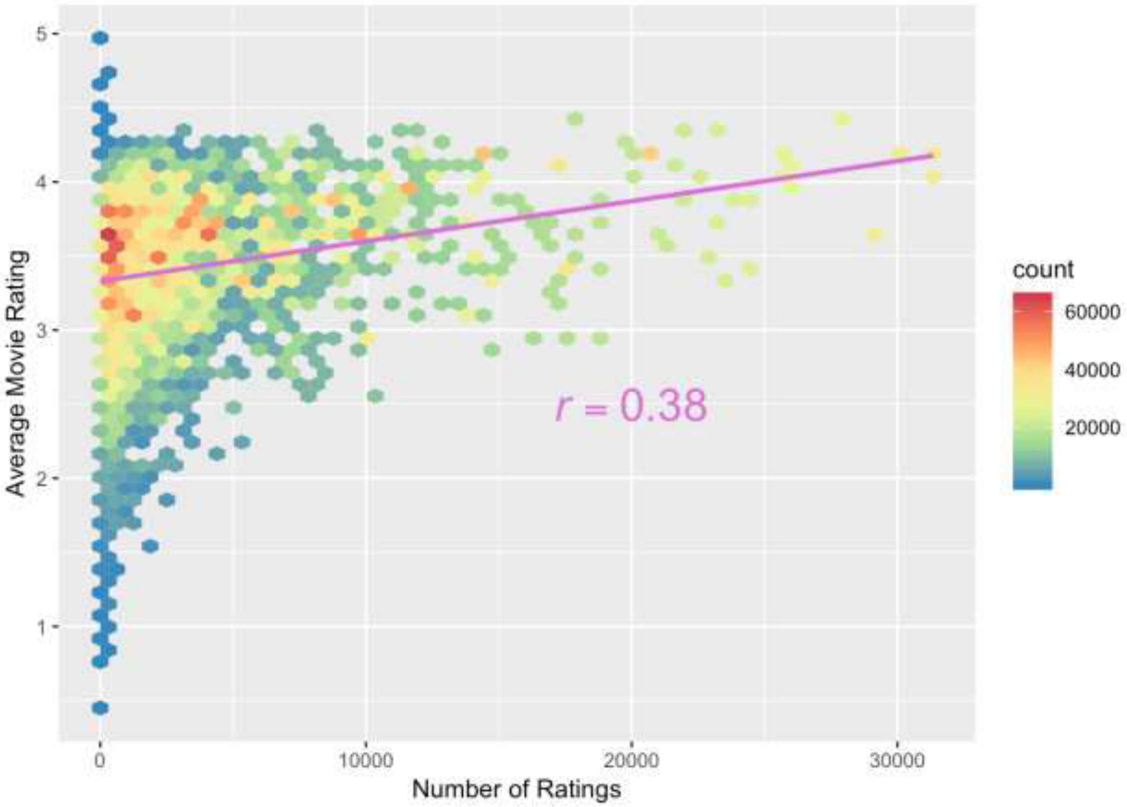
Graph the correlation


premi
movieId
year_rated
n
rating
avg_m_r
userId
age_of_movie
rating_date_range
premier_date
movieId
year_rated
n
rating
avg_m_r
userId
age_of_movie
rating_date_range
-1
-0.75
-0.5
-0.25
0
0.25
0.5
0.75
1

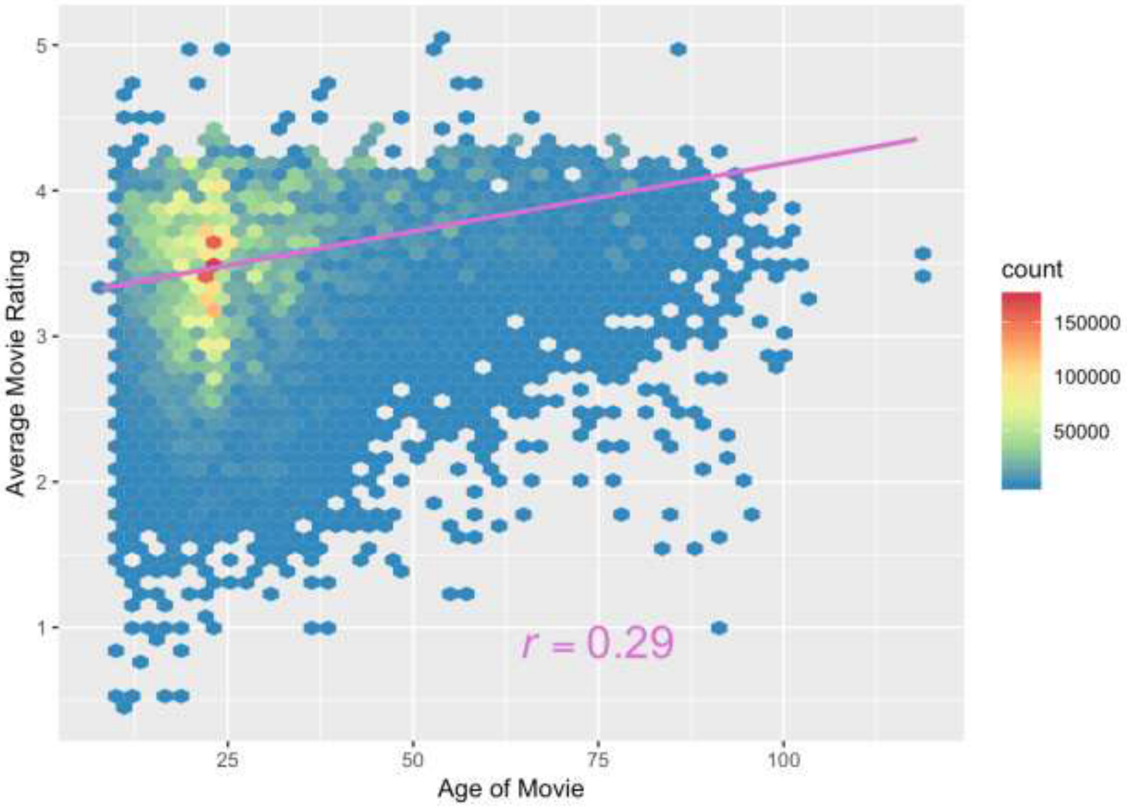
What is the effect of the age of the movie?



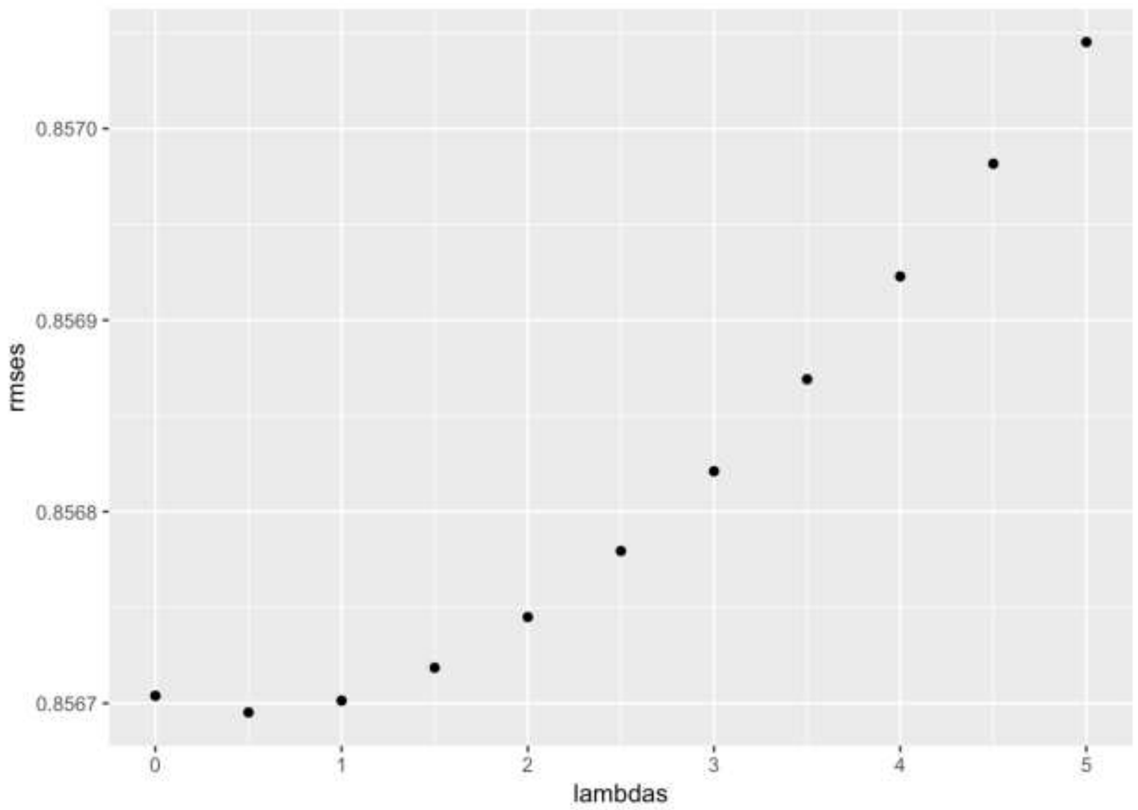
Is there a relationship between number of ratings and the average rating?



Is there an Age Effect on Movie Ratings?



Calculate the RMSE



Using the model on the Validation data

```
RMSE(predicted_ratings, validation$rating)
## [1] 0.8252108
```

I originally calculated a b_a for the age of a movie but found it didn’t lower my RMSE so took it out and didn’t include it in this script.

#I used movieId and userId to calculate the RMSE and was able to achieve an
RMSE = 0.8252