Case Study 2 - Analyzing data from MovieLens

Tyler Nardone DS 501 - Data Science with R

Introduction

Desired outcome of the case study. In this case study we will look at the movies data set from MovieLens. It contains data about users and how they rate movies. The idea is to analyze the data set, make conjectures, support or refute those conjectures with data, and tell a story about the data!

Problem 1: Importing the MovieLens data set and merging it into a single data frame

https://raw.githubusercontent.com/dnchari/DS501_MovieLens/master/Results/unifiedMLDataMulti.csv

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
## [1] 212257      8
```

```
##     user_id      movie_title           genre               rating     
##  Min.   :  1.0   Length:212257      Length:212257      Min.   :1.000  
##  1st Qu.:255.0   Class :character   Class :character   1st Qu.:3.000  
##  Median :450.0   Mode  :character   Mode  :character   Median :4.000  
##  Mean   :464.6                                         Mean   :3.551  
##  3rd Qu.:689.0                                         3rd Qu.:4.000  
##  Max.   :943.0                                         Max.   :5.000  
##  release_date            age           gender           occupation       
##  Length:212257      Min.   : 7.00   Length:212257      Length:212257     
##  Class :character   1st Qu.:24.00   Class :character   Class :character  
##  Mode  :character   Median :30.00   Mode  :character   Mode  :character  
##                     Mean   :32.77                                        
##                     3rd Qu.:40.00                                        
##                     Max.   :73.00                                        
```

```
## [1] "user_id"      "movie_title"  "genre"        "rating"       "release_date"
## [6] "age"          "gender"       "occupation"  
```

```
##   user_id           movie_title     genre rating release_date age gender
## 1       1 101 Dalmatians (1996) Childrens      2   1996-11-27  24      M
## 2       1 101 Dalmatians (1996)    Comedy      2   1996-11-27  24      M
## 3     101 101 Dalmatians (1996)    Comedy      3   1996-11-27  15      M
## 4     101 101 Dalmatians (1996) Childrens      3   1996-11-27  15      M
## 5      13 101 Dalmatians (1996) Childrens      2   1996-11-27  47      M
##   occupation
## 1 technician
## 2 technician
## 3    student
## 4    student
## 5   educator
```

Report some basic details of the data you collected. For example:

- How many movies have an average rating over 4.5 overall?

It can be seen from the head of the data frame that the same movie, user ID, and rating can show up on multiple rows if the movie falls into more than one genre. Therefore, while assessing movie ratings it will be appropriate to only consider one rating per user ID per movie in order to avoid inflating a movie's average rating if it falls into multiple genres. This is done by creating a new data frame that does not contain the genre column and therefore only contains unique reviews, and we will perform our analysis of ratings based off of this new data frame.

```
## [1] 11
```

- How many movies have an average rating over 4.5 among men? How about women?

```
## [1] 18
```

```
## [1] 16
```

- How many movies have an median rating over 4.5 among men over age 30? How about women over age 30?

```
## [1] 47
```

```
## [1] 70
```

- What are the ten most popular movies?
  - Choose what you consider to be a reasonable definition of "popular".

Popularity is a relatively subjective term that can carry multiple interpretations. For example one measure of popularity could be the number of reviews, with a higher number of reviews indicating that more people watched a given movie. On the other hand, popularity could also depend on the average rating as well, since many people would only consider a movie to be popular if it is highly rated. Therefore it will be appropriate to blend these two distinct measures by creating a "popularity score" that will be the multiplicative product of average movie rating and the relative number number of reviews that a given movie has.

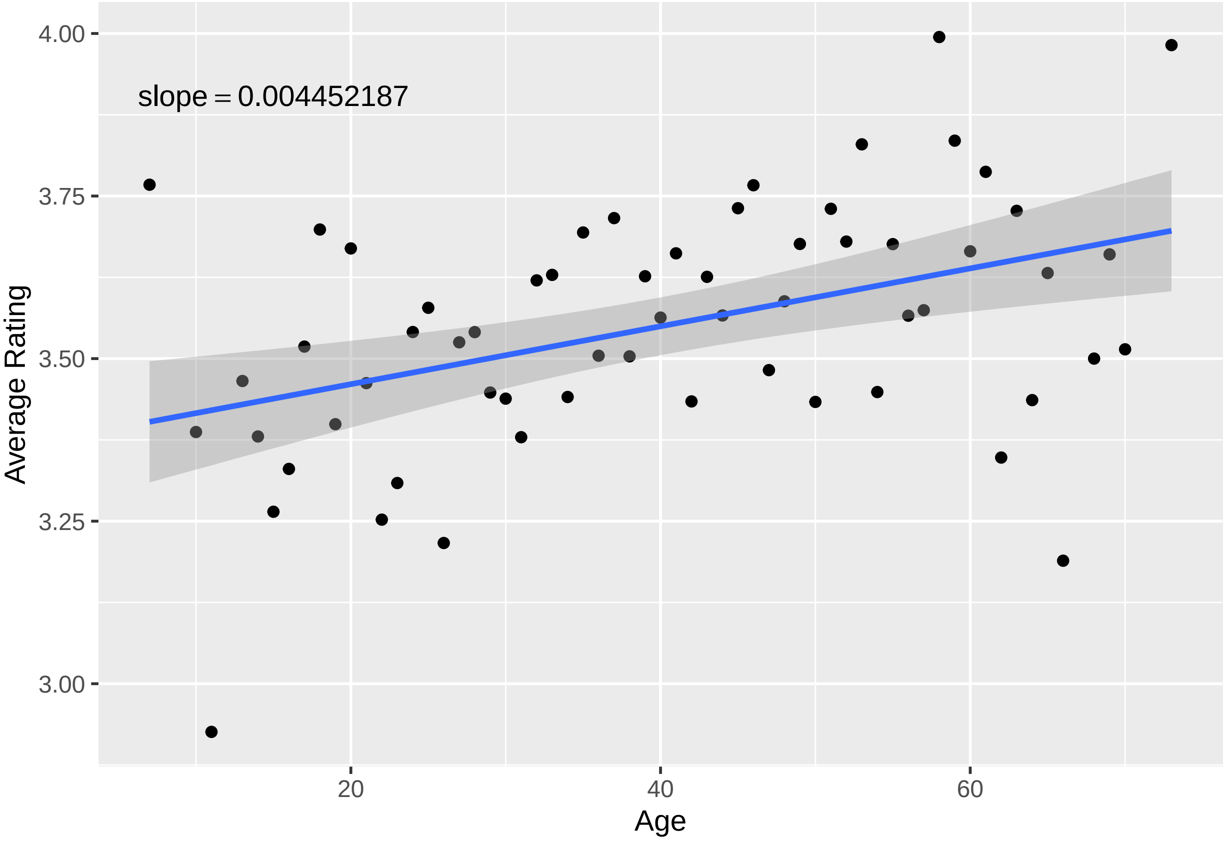
Be prepared to defend this choice.

The column "rel_freq" (below) is the number of rows (i.e. number of reviews) for a given movie, divided by the total number of reviews (i.e. the number of rows in ratingsByUserID). Popularity score is then the product of average rating and relative frequency of reviews. This is appropriate because if two movies share the same average rating, the one with the higher relative frequency will have the higher popularity score, and vice versa. This measure therefore elevates movies that both are highly rated, and have a high number of reviews relative to the rest of the data frame.

```
## # A tibble: 10 x 4
##    movie_title                      average_rating rel_freq popularity_score
##    <chr>                                     <dbl>    <dbl>            <dbl>
##  1 Star Wars (1977)                           4.36  0.00585           0.0255
##  2 Fargo (1996)                               4.16  0.00509           0.0212
##  3 Return of the Jedi (1983)                  4.01  0.00508           0.0204
##  4 Contact (1997)                             3.80  0.00510           0.0194
##  5 Raiders of the Lost Ark (1981)             4.25  0.00421           0.0179
##  6 Godfather, The (1972)                        4.28  0.00414           0.0177
##  7 English Patient, The (1996)                  3.66  0.00482           0.0176
##  8 Toy Story (1995)                             3.88  0.00453           0.0176
##  9 Silence of the Lambs, The (1991)             4.29  0.00391           0.0168
## 10 Scream (1996)                                3.44  0.00479           0.0165
```

- Make some conjectures about how easy various groups are to please? Support your answers with data!
  - For example, one might conjecture that people between the ages of 1 and 10 are the easiest to please since they are all young children. This conjecture may or may not be true, but how would you support or disprove either conclusion with with data?

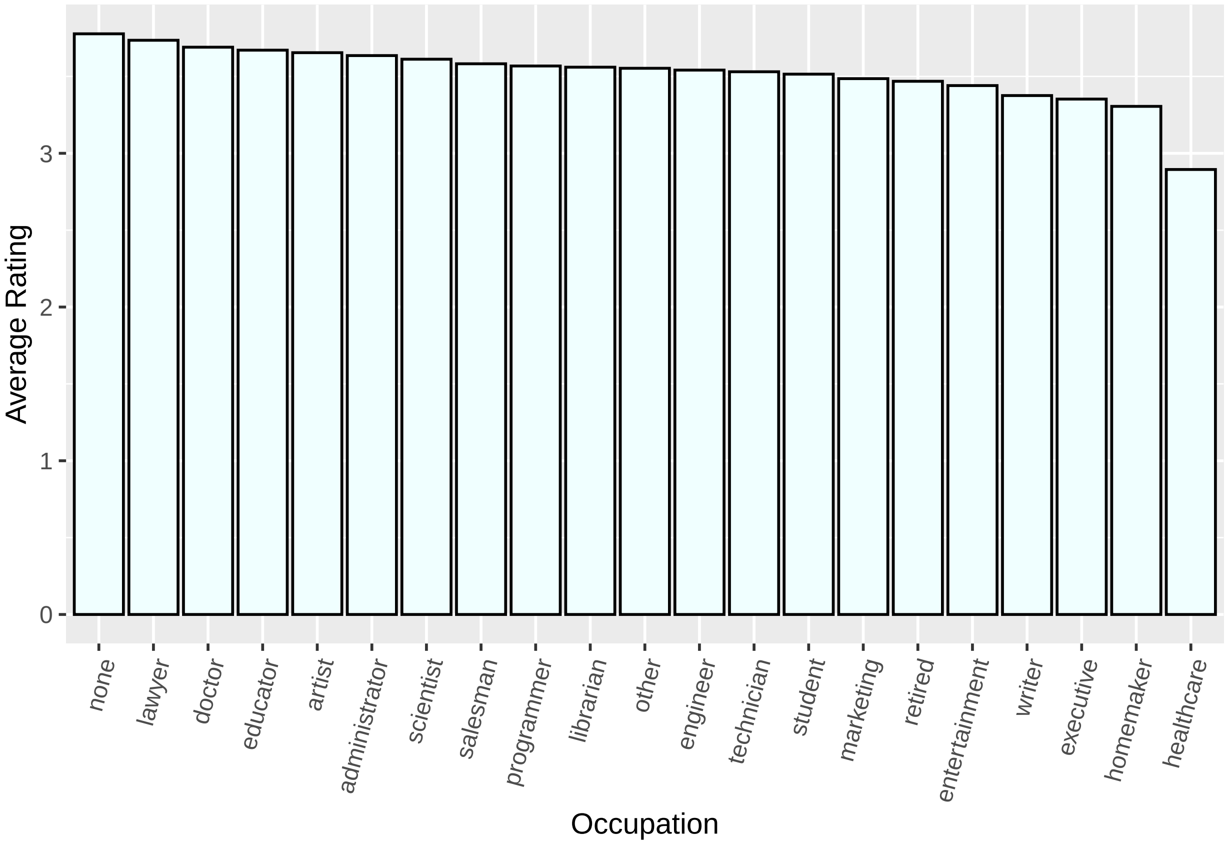
To investigate the conjecture above, let us group the data by age to see if younger reviewers tend to leave higher reviews. Will younger reviewers leave higher reviews on average than older reviewers?



The above chart does in fact suggest that the average rating varies with age, however we can see that average rating actually increases very slightly with age. This would suggest that on average, younger reviewers are more difficult to please since on average they leave lower reviews. However we can see that the slope of the regression line is very low, and in fact might be too low to suggest that there is a measurable difference among age at all.

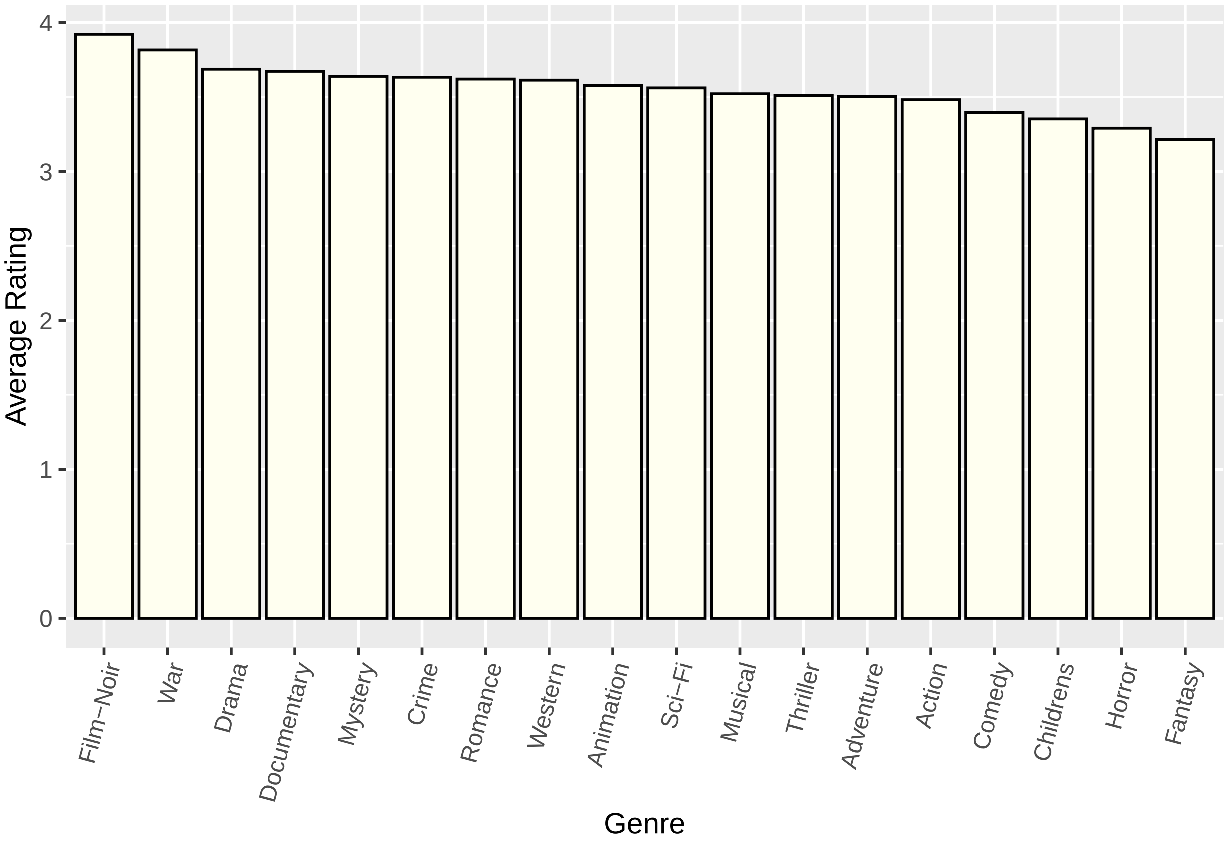
Be sure to come up with your own conjectures and support them with data!

How does average rating tend to vary with occupation? Do some occupations tend to leave higher reviews on average than others?



The above chart seems to suggest there is moderate but not much variation in ratings due to occupation alone, although on average healthcare workers do tend to leave lower reviews relative to other occupations in the data frame.

Do certain genres tend to be favored by reviewers more than others?

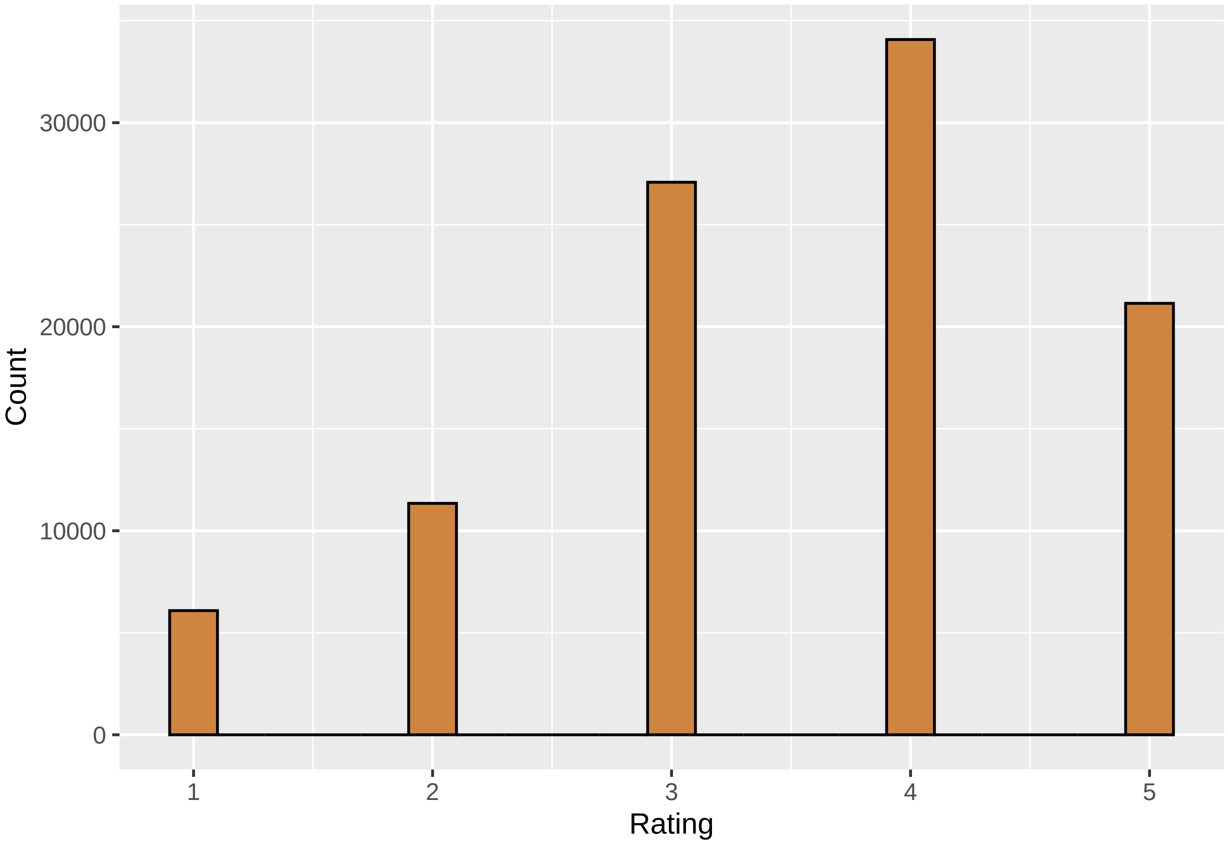


Again, there does not appear to be significant variation in average rating due to genre alone, with only about 1/2 point difference between the highest rated genre and the lowest rated genre.

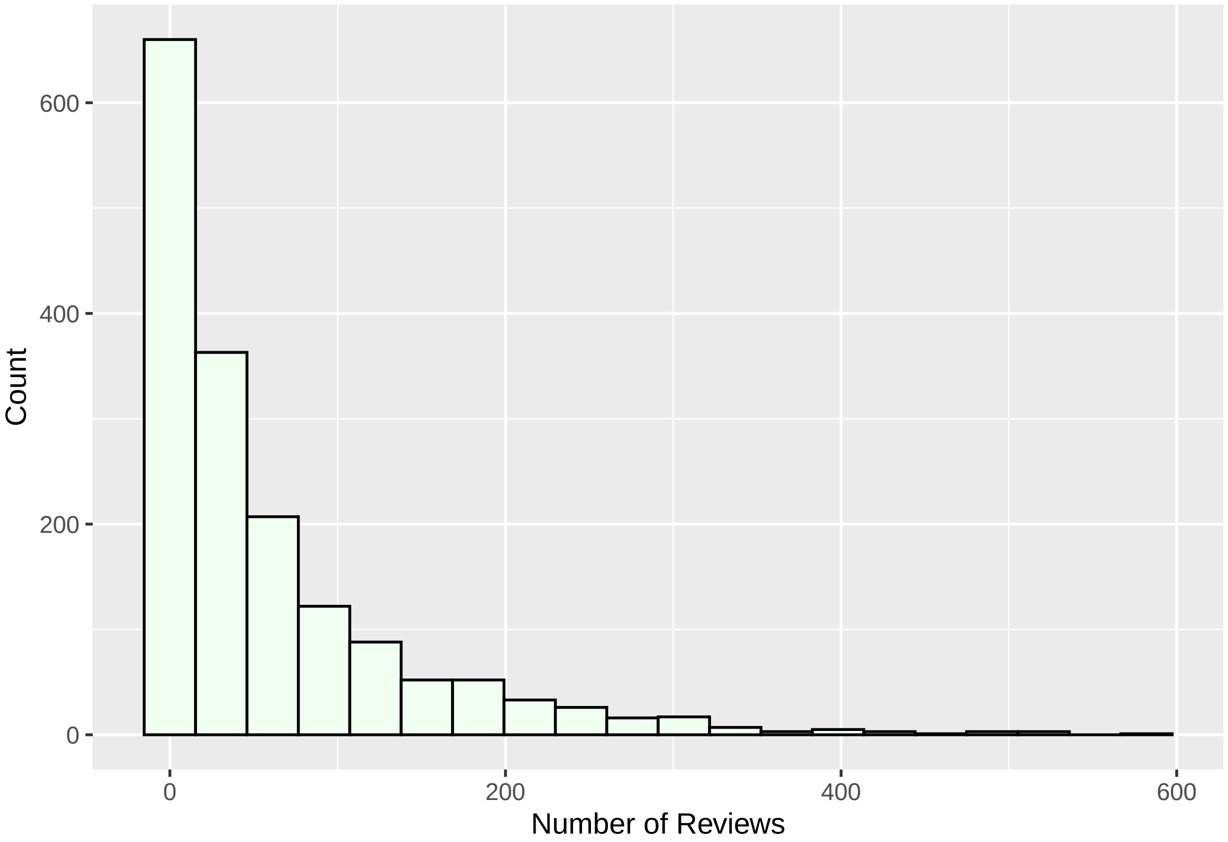
Problem 2: Expand our investigation to histograms

An obvious issue with any inferences drawn from Problem 1 is that we did not consider how many times a movie was rated.

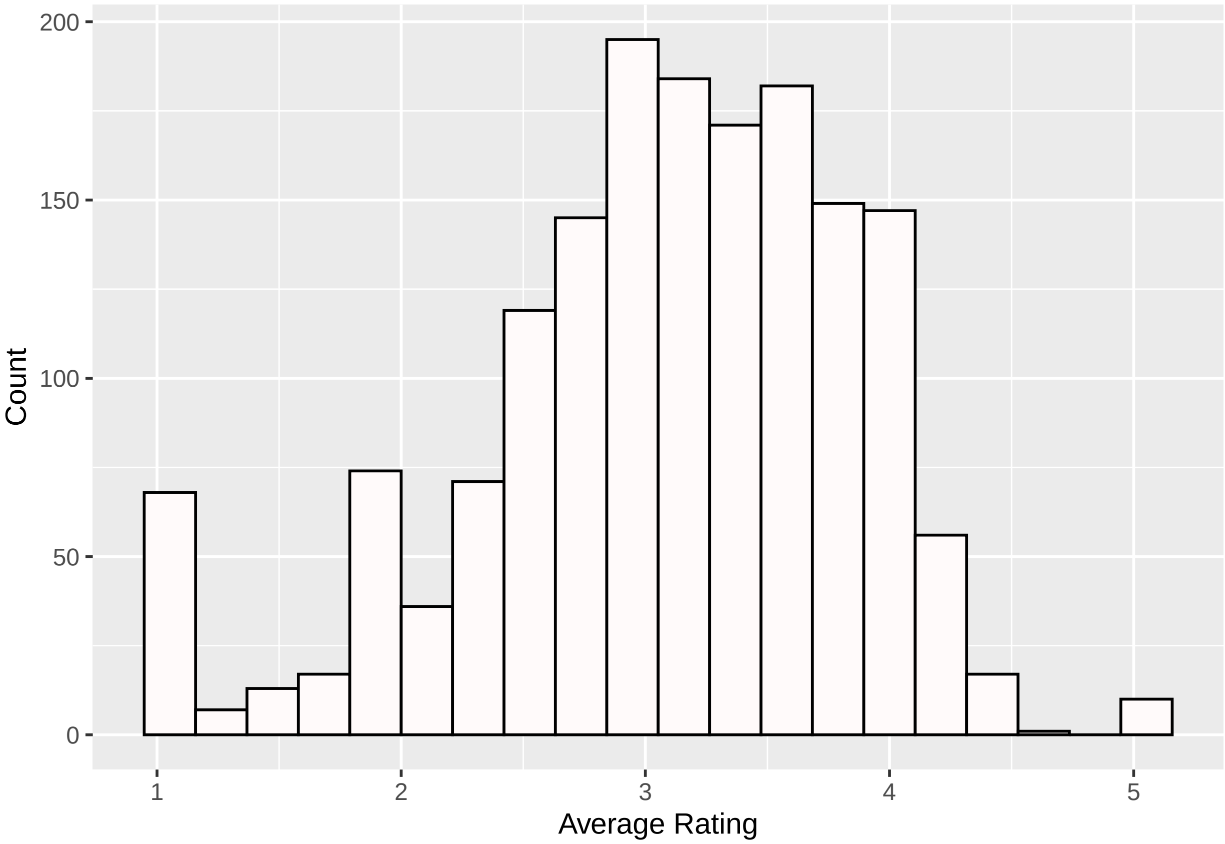
- Plot a histogram of the ratings of all movies.



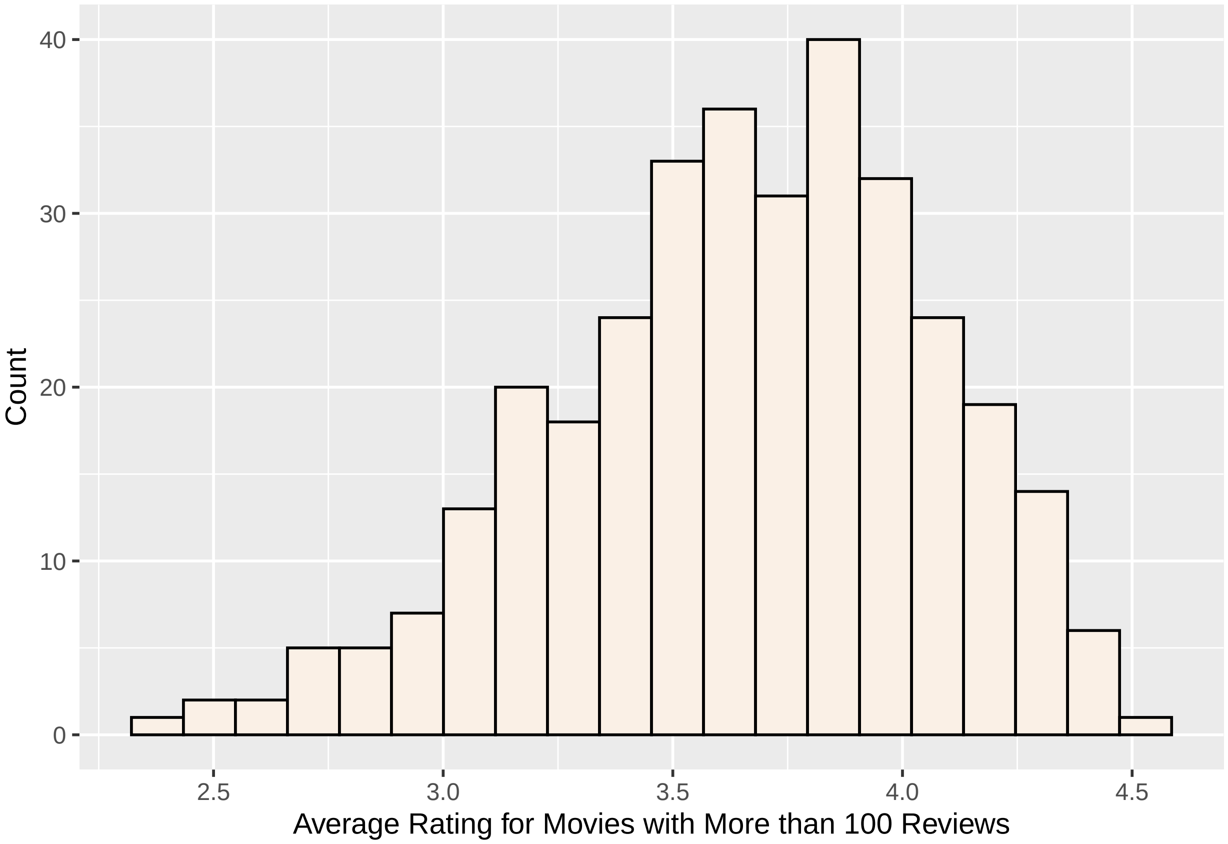
- Plot a histogram of the number of ratings each movie received.



- Plot a histogram of the average rating for each movie.



- Plot a histogram of the average rating for movies which are rated more than 100 times.



What do you observe about the tails of the histogram where you use all the movies versus the one where you only use movies rated more than 100 times?

Movies with more than 100 reviews tend to have a skew towards higher average ratings compared to all movies considered at once.

Which highly rated movies would you trust are actually good? Those rated more than 100 times or those rated less than 100 times?

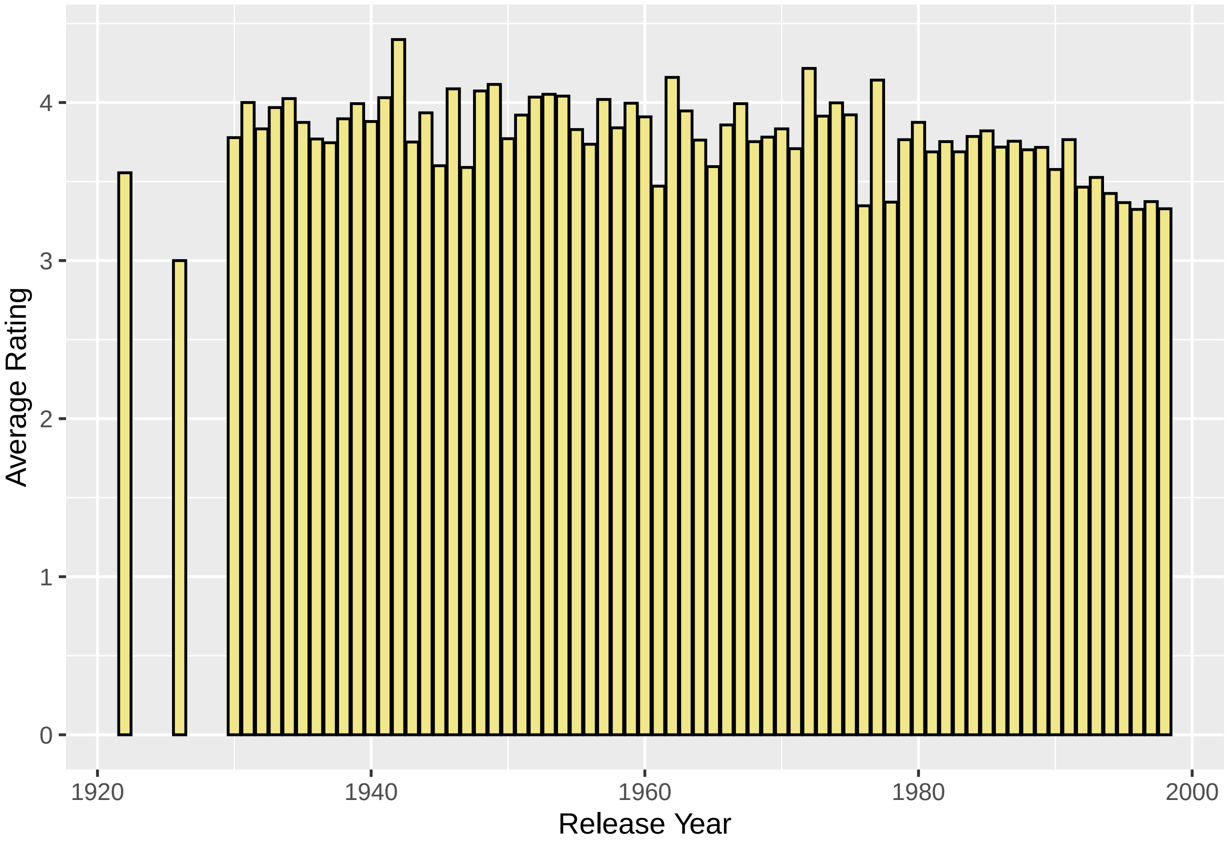
For the same reason that a popularity score was formed, movies rated more than 100 times can be better trusted as actually good when paired with a high average rating. This condition suggests a large amount of independent agreement of the higher average rating.

- Make some conjectures about the distribution of ratings? Support your answers with data!
  - For example, what age range do you think has more extreme ratings? Do you think children are more or less likely to rate a movie 1 or 5?

As shown in the chart above, there appears to be very little variance in average ratings due to age alone.

Be sure to come up with your own conjectures and support them with data!

How have ratings changed over time? For example, if we extract the year from release_date for every movie, what will the average ratings for all movies released in a given year look like, and will that change over time?

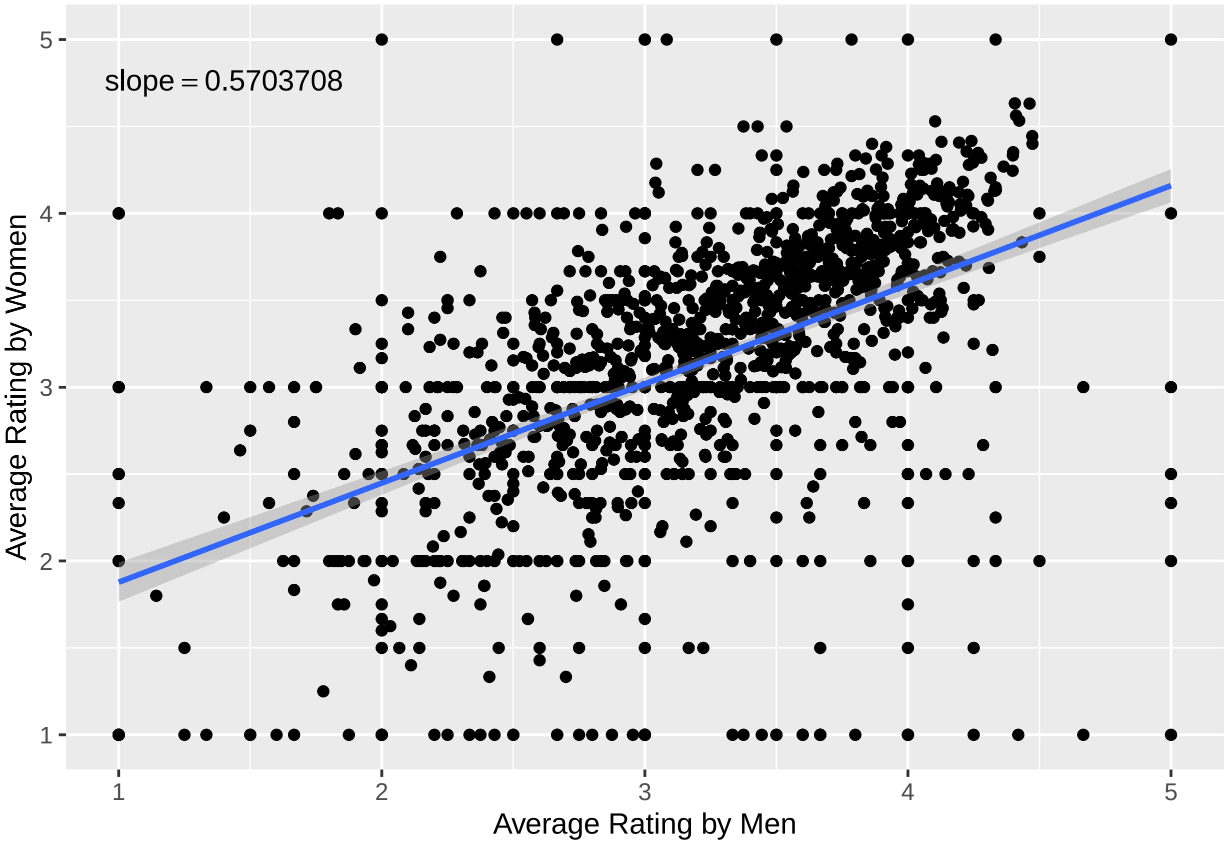


Above we can see that when all movies released within a given year are considered together, there does not appear to be much variance year to year in average ratings. From about 1985 onward, there does appear to be a slight but minor downwards trend.

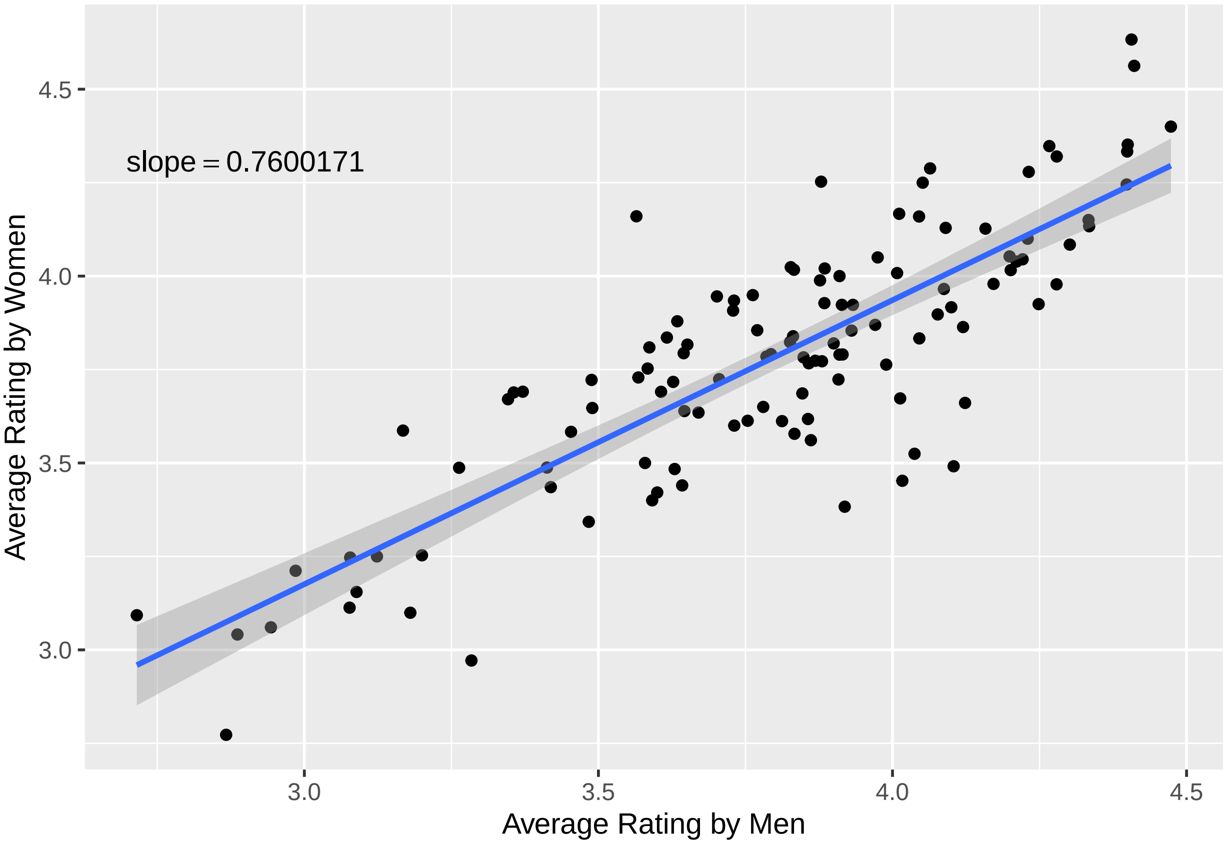
Problem 3: Correlation: Men versus women

Let us look more closely at the relationship between the pieces of data we have.

- Make a scatter plot of men versus women and their mean rating for every movie.



- Make a scatter plot of men versus women and their mean rating for movies rated more than 200 times.



- Compute the correlation coefficient between the ratings of men and women.
    - What do you observe?

```
## [1] 0.5149489
```

When all movies are considered in the analysis, a correlation coefficient of 0.515 suggests a moderate amount of correlation between the average ratings by men and the average ratings by women. However this paired with the visual scatter plot suggests that there are many instances that do not follow this trend and it would be difficult to argue that there is a concrete and associative relationship between these two variables.

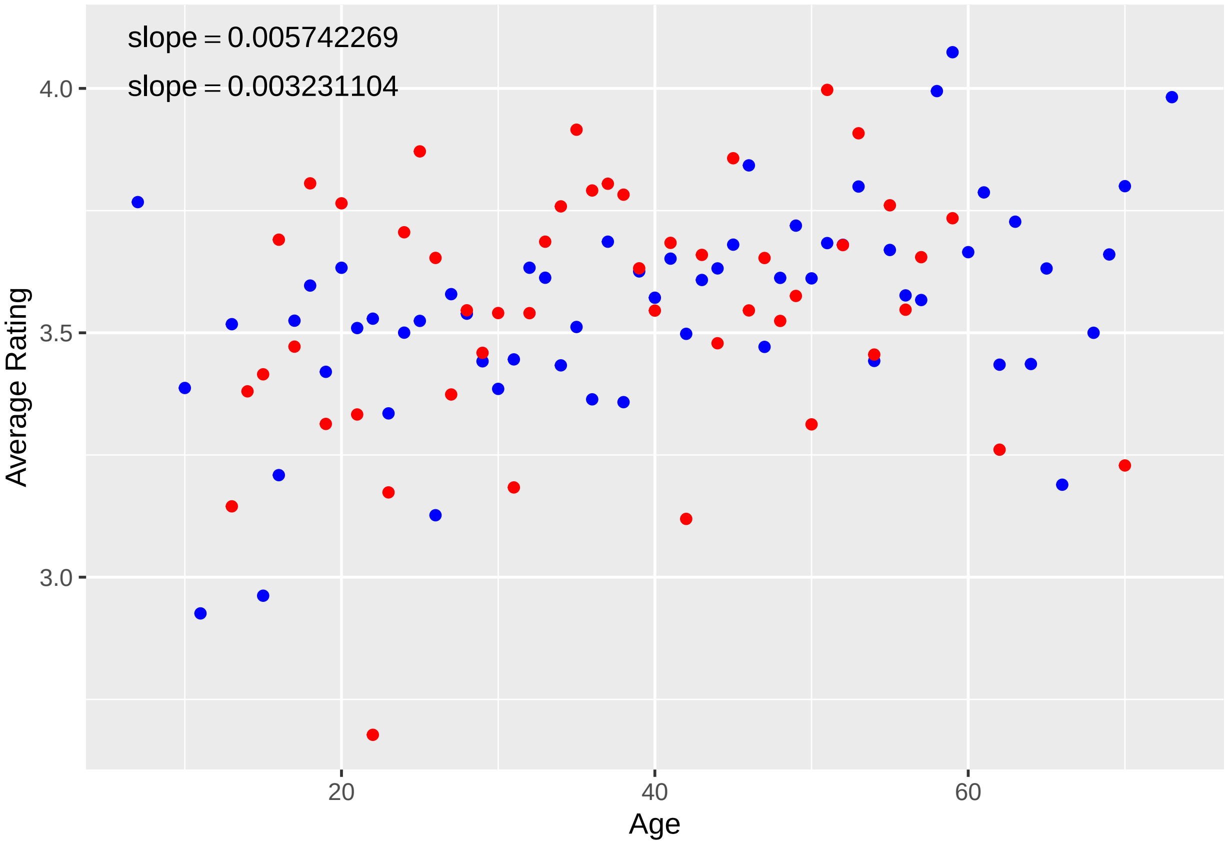
```
## [1] 0.8311729
```

Are the ratings similar or not? Support your answer with data!

A correlation coefficient of 0.831 suggests a much higher correlation when only considering movies with more than 200 reviews. This seems to indicate that there is in fact a positive correlation between these two variables, and a slope of the regression line of 0.76 in this context indicates that on average, men tend to leave higher reviews than women.

- Conjecture under what circumstances the rating given by one gender can be used to predict the rating given by the other gender.
    - For example, are men and women more similar when they are younger or older?

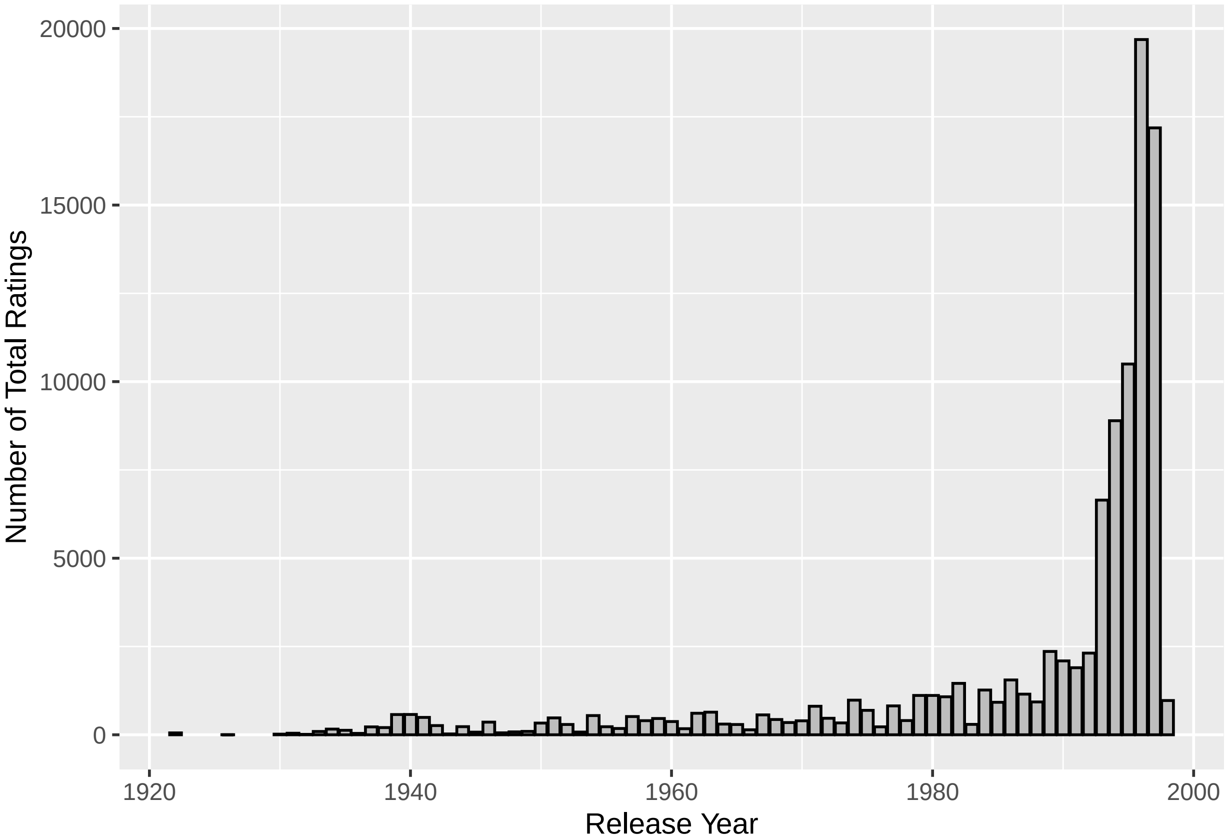
To assess the above conjecture, the average rating vs. age chart will be recreated, with two different columns plotted – one for men and one for women. It will be then be seen if these two plots diverge or if they fall in relatively the same place.



Both men and women tend to show very little difference in the average ratings by age, which agrees with the aggregate result that was found previously in the analysis. This seems to indicate that for both genders, there is very little variation in average ratings due to age alone.

Be sure to come up with your own conjectures and support them with data!

How much has reviewer engagement (i.e. number of reviews in a given year) changed over time? Do movies receive a higher number of reviews now compared to the past?



Clearly, movies released from around the early 1990s onward have a significantly higher number of reviews relative to movies from earlier decades.

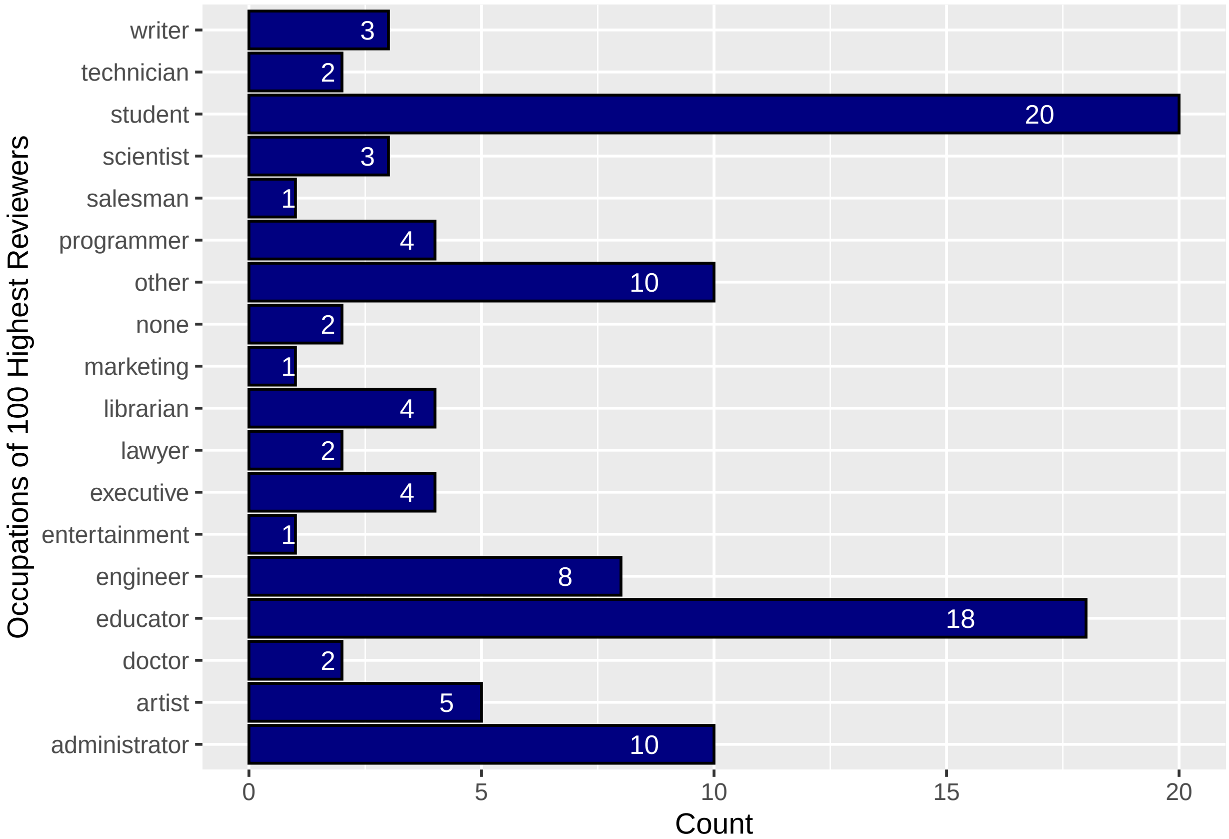
Problem 4: Open Ended Question: Business Intelligence

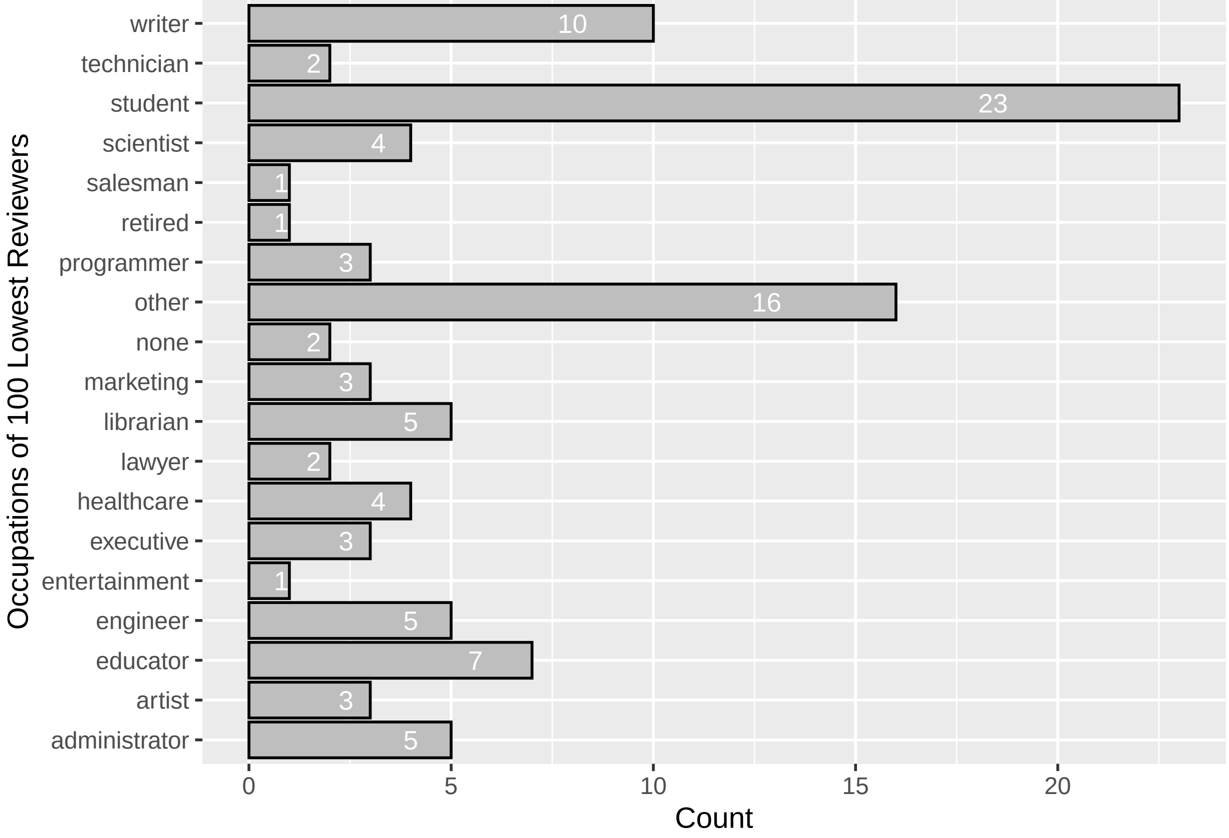
- Do any of your conjectures in Problems 1, 2, and 3 provide insights that a movie company might be interested in?

If a movie company is interested in making a new movie with an optimal chance of being highly rated, they may be interested in knowing which genres historically score high ratings, and also perhaps which professions tend to review movies highly on average. In theory, they could then tailor a movie in a specific genre to appeal somehow to that profession, and perhaps they would have a better chance at the movie receiving a high rating. It may also be good to know that on average, there is not much variation due to the age of the reviewer, and also that on average men tend to leave slightly higher reviews than women.

- Propose a business question that you think this data can answer.
- Suppose you are a Data Scientist at a movie company. Convince your boss that your conjecture is correct!

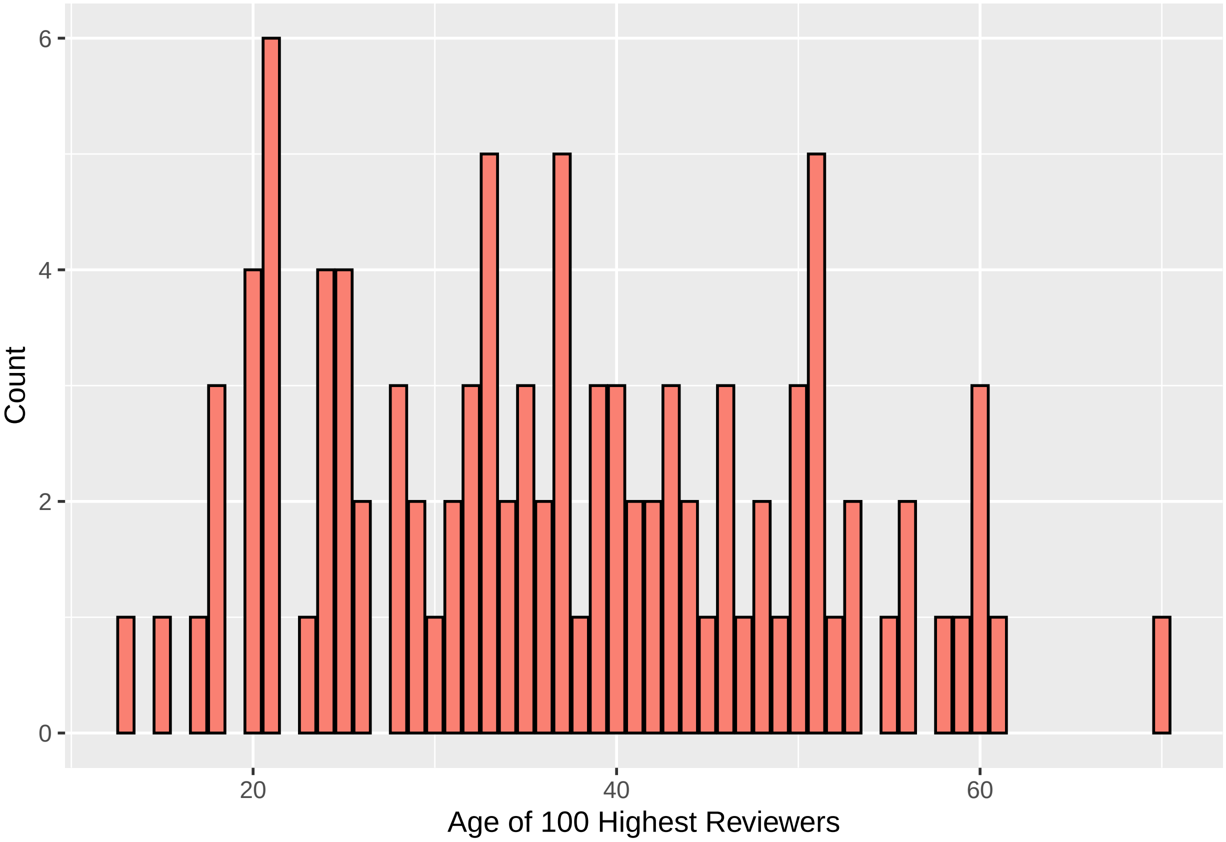
What can we learn about the most extreme reviewers? For example, among people who leave both the 100 highest and 100 lowest reviews in general, do they have any identifying traits?

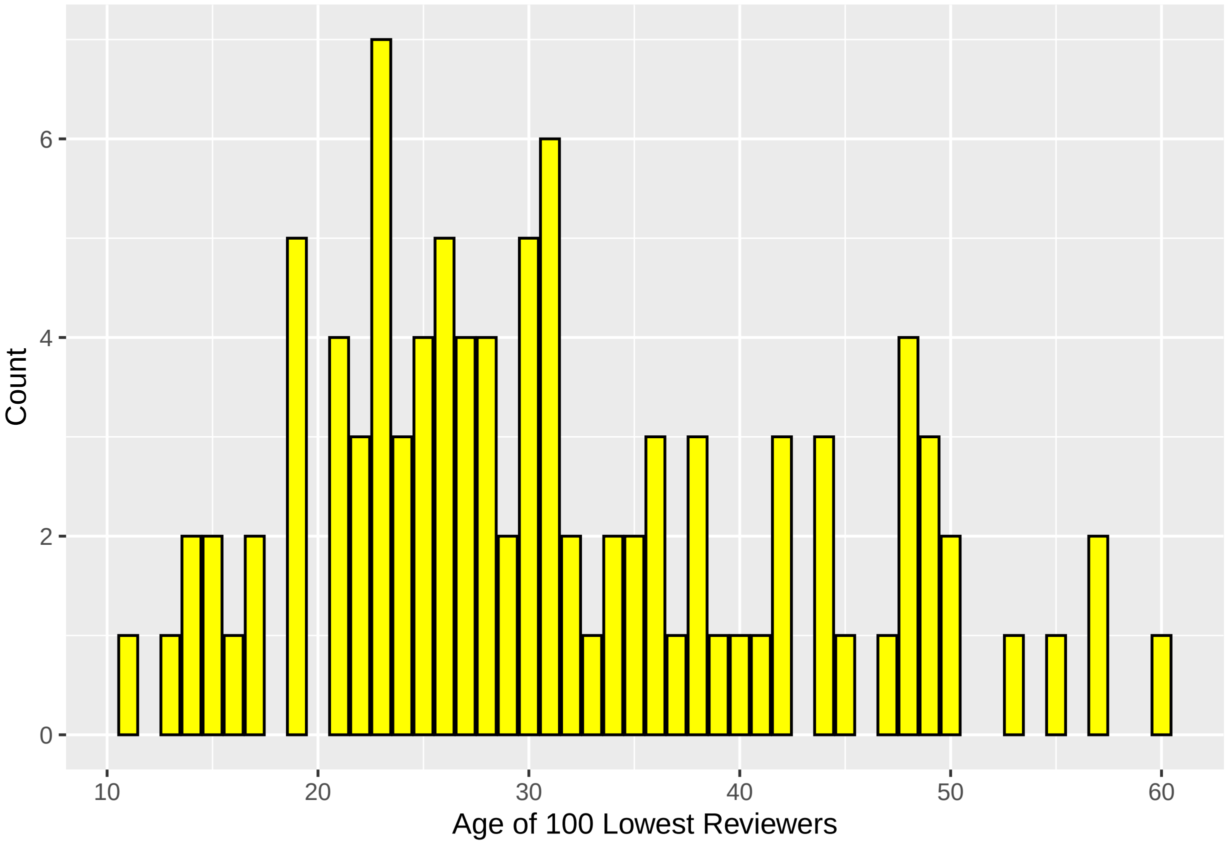




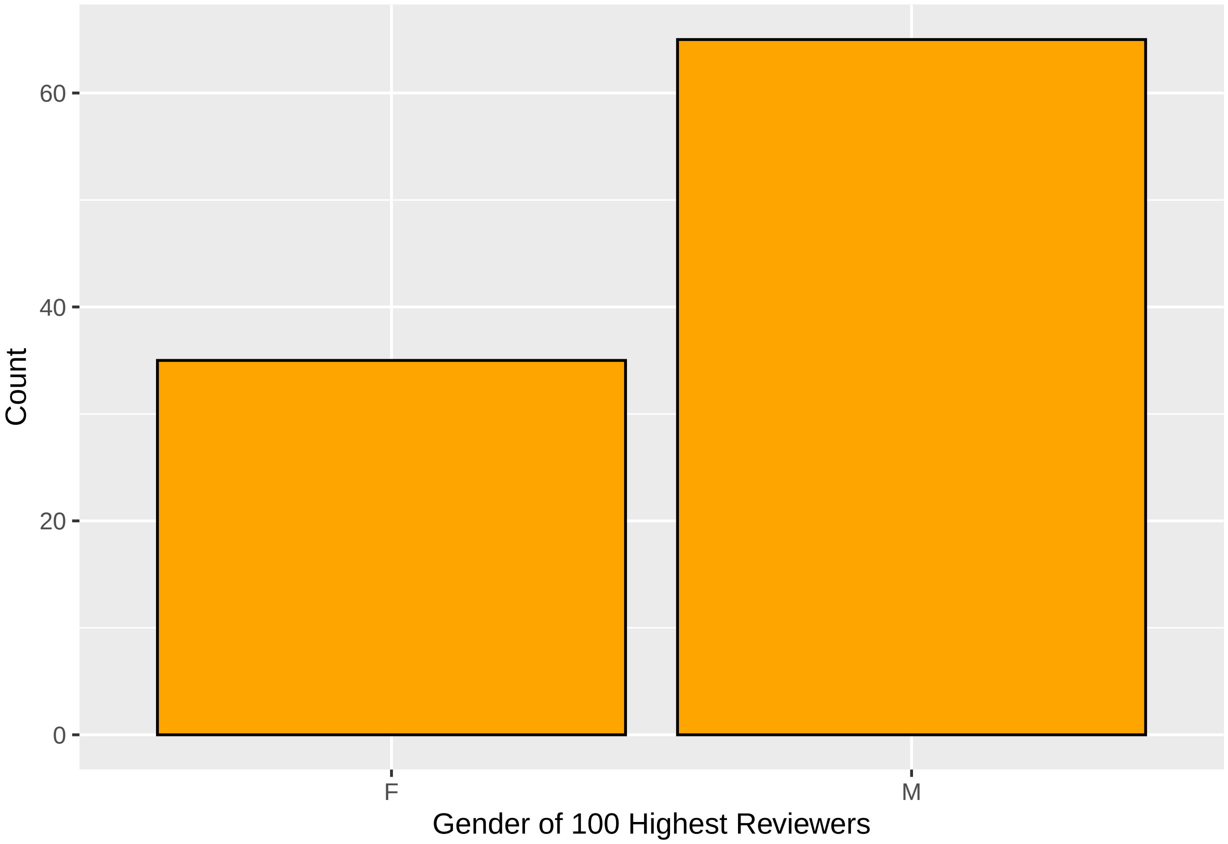
It is interesting to see that both students and “other” make up a significant potion of both the 100 highest and

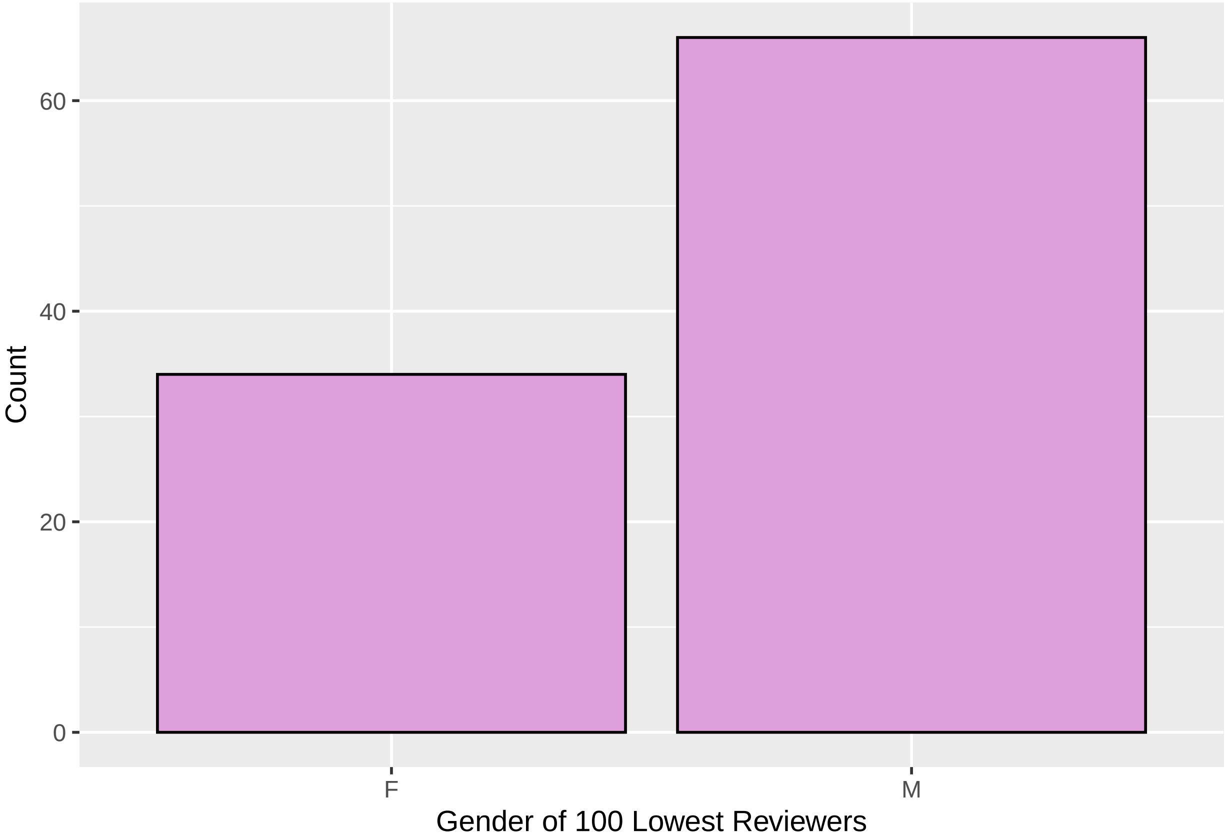
100 lowest reviewers.





For the 100 highest reviewers, there does not appear to be much variance in age; the distribution appears fairly uniform or even slightly normally distributed. For the 100 lowest reviewers on the other hand, the data does appear to be skewed more towards people in their 20s.





In both the 100 highest and the 100 lowest reviewers, men appear to be represented about twice as much as women, which may suggest that on average they have more extreme opinions.

Double check that the split of gender is actually the same for both 100 highest and 100 lowest reviewers:

```
## [1] 65
```

```
## [1] 35
```

```
## [1] 65
```

```
## [1] 35
```

Coincidentally, the split is exactly the same.

Done

All set!

What do you need to submit?

1. Report: please prepare a report based on what you found in the data.

- What data you collected?

The data is all based off of the MovieLens data set that was provided.

- Why this topic is interesting or important to you? (Motivations)

People can have very varying opinions on the movies they watch, and it is interesting to try and form aggregate conclusions based on a large data set of movie reviews.

- How did you analyze the data?

Much of this analysis was based on trying to find aggregate trends based on average reviews that people leave.

- What did you find in the data? (please include figures or tables in the report)

There were many findings as shown above, and in general some factors seems to affect peoples' ratings on average, and some factors do not.

2. R Code with RMarkdown, compile it to PDF

How to submit: - Submit PDF file on Course Webpage on Canvas only. Do not email it to me.