Case Study 2 - Analyzing Data from MovieLens with R

Trevor Shaw

Introduction

Data is a very important tool used by movie companies in order to evaluate the performance of a movie in terms of how well it was received by the public. A rating system is common method used to evaluate movie performance. In this particular case study, we are going to look at a rating system utilizing the integers 1-5. We will have to take an assumption about this rating system because we are not explicitly told what these rating mean. Let's assume 1 = Very Bad, 2 = Bad, 3 = Okay, 4 = Good, and 5 = Very Good. It is important to understand that the statistics derived from the data are objective by nature but, there are many subjective elements to a rating system which we will explore in the following sections.

Section 1: Statistics and Conjectures

First let's download our data from the internet at the follow link and import into R as a data frame: [link] (https://raw.githubusercontent.com/dnchari/DS501_MovieLens/master/Results/unifiedMLDataMulti.csv)

```
movielens = 'https://raw.githubusercontent.com/dnchari/DS501_MovieLens/master/Results/unifiedMLDataMulti
mlData = read.csv(movielens)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Here are some brief summary statistics about the data:

- 15 movies have an average rating greater than or equal to 4.5 overall
- 22 movies have an average rating greater than or equal to 4.5 among men
- 20 moves have and average rating greater than or equal to 4.5 among women
- 59 movies have a median rating greater than or equal to 4.5 among men over age 30
- 91 movies have a median rating greater than or equal to 4.5 among women over age 30
- There are 1,662 movies in the data set

Here are the ten most popular movies by mean rating: 1. A Close Shave (4.49) 2. Schindler's List (4.47) 3. The Wrong Trousers (4.47) 4.Casablanca (4.46) 5. Shawshank Redemption (4.45) 6. Rear Window (4.39) 7. The Usual Suspects (4.39) 8. Star Wars (4.36) 9. 12 Angry Men (4.34) 10. The Third Man (4.33)

This Top 10 list was established by taking the mean rating of those movies with at least 100 ratings. This list intentionally does not consider any other factors such as gender, age or occupation. I chose only movies with at least 100 ratings because mean rating becomes more trustworthy with larger sample sizes. We will explore this in more detail later in the report.

Potential biases might exist in the data. We can see this by looking at which groups might be most easy to please. One may be tempted to make the conjecture that children (10 and under) would have the highest average rating. This is not true. Rating actually positively correlates with an increase in age. The average rating among children is actually 3.57 whereas the average rating among mature adults (50+) is 3.66. Although, the correlation is there it is slight and it is fair to say age is not a factor strongly affecting popularity.

The difference between the average rating among men and women is 0.01 So, gender is not a factor strongly affecting popularity. We will more deeply explore the gender variable among other variables further in the report.

When I was first exploring the top 10 movies, I suspected a strong male bias. To no surprise of my own, I discovered 75% of the +200,000 ratings in the data set are male. This could potentially be attributed or that all of the movies in the data set were released prior to 1999 where the percentage of male movie directors was even higher.

Section 2: Using Histograms to Learn about the Data

R has many effective visualization tools. Many of the advanced tools can be imported into R from a variety of sources. For this report, I chose to stick with the R base package. Since R is free, any one can download the software and create a report like this without installing any extra visualization packages. Neat!

Next, you will see 4 figures. Figure 1 tells us that most ratings were 3, 4, or 5. Figure 2 tells us that the majority of movies had very few ratings. Figures 3 and 4 give us an idea of the distribution of ratings and how the distribution changes when rating count is considered.

Figure 1: Ratings of all movies.

```
Rating=mlData$rating
hist(Rating, breaks=50, main="Figure 1: All Ratings")
```

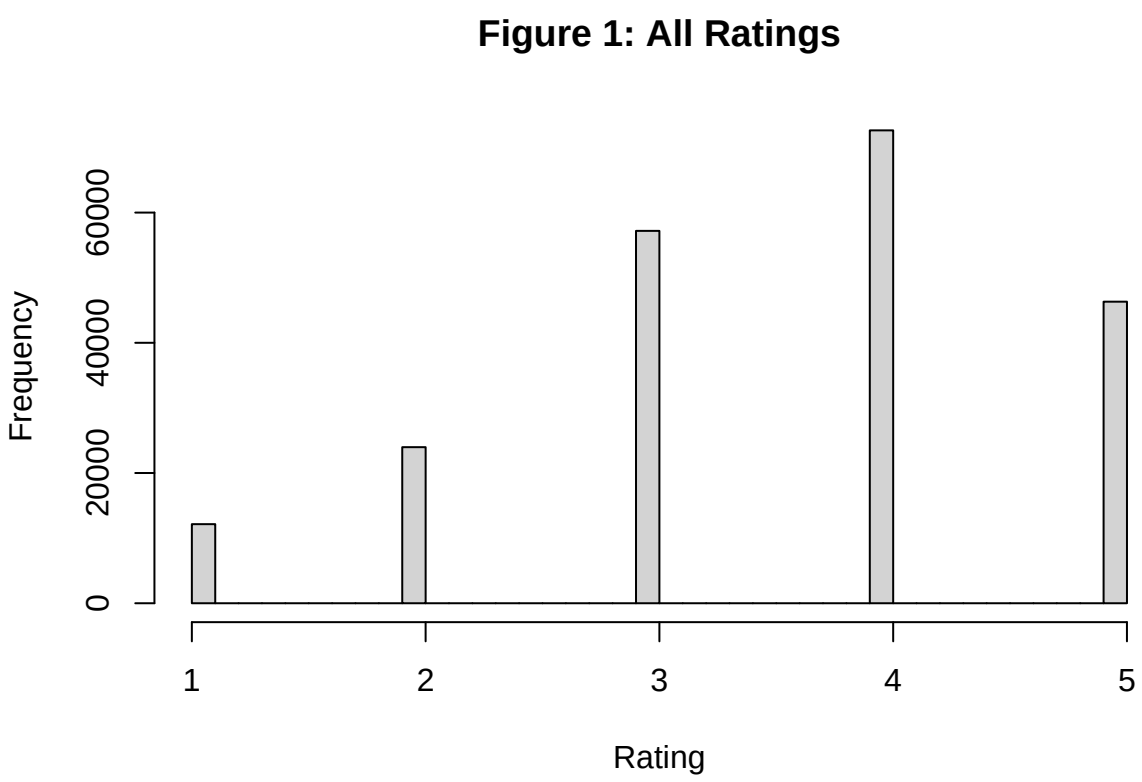


Figure 2: Number of Ratings by Movie

```
library(dplyr)
Number_Ratings_per_Movie = table(mlData$movie_title)
hist(Number_Ratings_per_Movie, breaks=50, xlab="Number of Times Rated", ylab="Number of Movies", main="
```

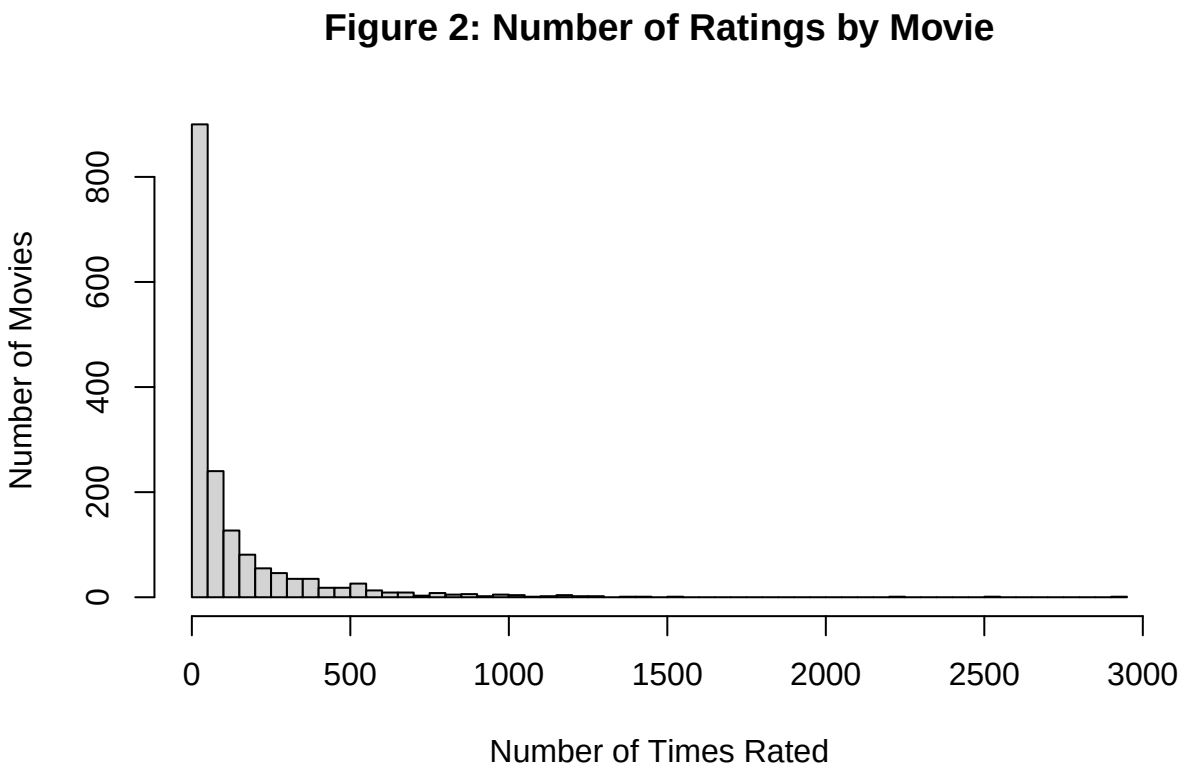


Figure 3: Average Rating by Movie

```
mean_rating_table=mlData %>% group_by(movie_title) %>%
summarize(mean_rating = mean(rating))
hist(mean_rating_table$mean_rating, breaks=50, main="Figure 3: Average Rating by Movie", xlab="Mean Rati
```

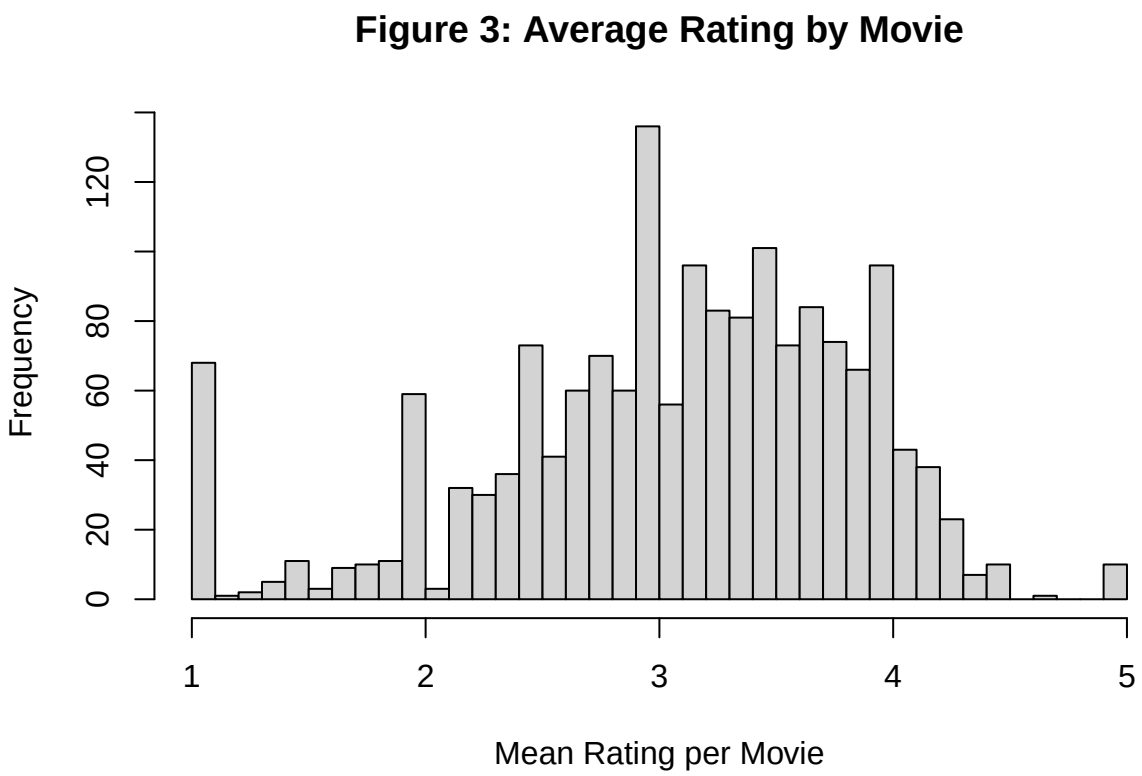
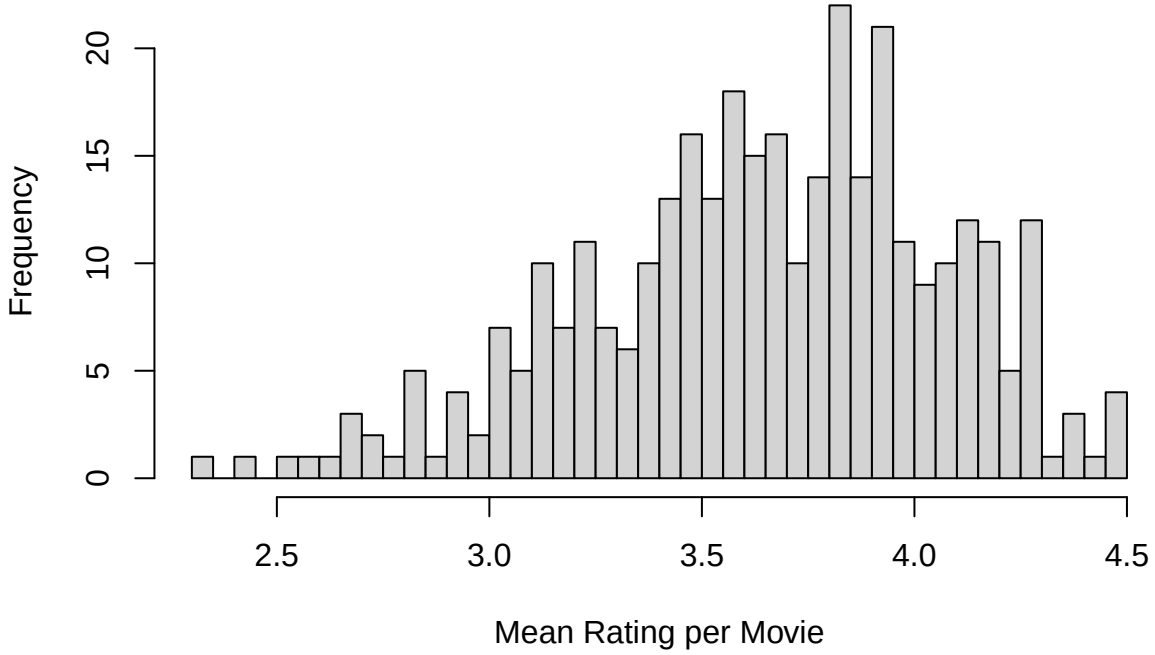


Figure 4: Average Rating for Movies (At Least 100 Ratings)

```
rating_freq_table=mlData %>% group_by(movie_title) %>%
summarize(rating_frequency = n_distinct(user_id))

mean_freq_table=merge(mean_rating_table, rating_freq_table, by.x= "movie_title")
mean_freq_table=mean_freq_table[order(mean_freq_table$rating_frequency, decreasing=TRUE),]
mean_freq_table=mean_freq_table[1:337,]
hist(mean_freq_table$mean_rating, breaks=50, main="Figure 4: Average Rating by Movie", xlab="Mean Ratin
```

Figure 4: Average Rating by Movie



The tails of the histogram in Figure 3 stretch quite far. The top tail stretches all the way to 5 which means that those movies received a 5 every single time it was rated. In figure 4, when we only consider movies with at least 100 ratings, we can see that the tails have less of a spread. Figures 3 and 4 tells us the more ratings we have in a movie, the more confidence we can have in the accuracy in the movie rating. It would be hard to trust if a movie was actually good if it was rated only a few times and received all high ratings.

Let's look at the extreme ends of our rating scale (ratings of 1's and 5's). 27.5% of all ratings were a 1 or a 5. One might be inclined to suggest that children would have the highest percentage extreme ratings. 32% of rating by children 10 and under gave ratings of 1 or 5 while 27% of all other ratings were a 1 or 5. Children do appear slightly more likely to give a an extreme rating but, the difference between from rest of the population is small. However, there are only 200 child ratings to consider. It would be interesting to see if these statistics changed with a larger sample of children.

31% of female ratings were extreme and 27% of male ratings were extreme. Again, we see a slight difference. However, 75% of the ratings are male and it would be interesting if the female percentage dropped with more female ratings.

One interesting observation is that genre does seem to have a moderate effect on rating. Film-Noir and War have the highest percentage of extreme ratings at 42% while Fantasy (23%) and Children (28%) have the lowest percentage of extreme ratings. Film-Noir is also the most highly rated genre at 3.92 and the difference between that and the lowest rated genre Fantasy is 0.71.

Problem 3: Correlation: Men versus Women

The correlation coefficient between the average rating of males and females by movie is 0.52, suggesting a moderate positive correlation between the ratings. What this tells us is that, generally speaking, there is not a huge difference between the average male and female rating of most movies. Further supporting this

notion, the difference in male and female mean ratings is less than 0.62 for 90% of movies with at least 50 ratings. This number jumps to 97% when you consider movies with at least 200 ratings. The correlation gets stronger when the sample size increases (see figures 5 and 6).

##Figure 5: Men vs. Women Mean Rating by Movie

```
gmean_rating_table=mlData %>% group_by(movie_title, gender) %>%
summarize(mean_rating = mean(rating))
```

```
## `summarise()` has grouped output by 'movie_title'. You can override using the
## `.groups` argument.
```

```
grating_freq_table=mlData %>% group_by(movie_title, gender) %>%
summarize(rating_frequency = n_distinct(user_id))
```

```
## `summarise()` has grouped output by 'movie_title'. You can override using the
## `.groups` argument.
```

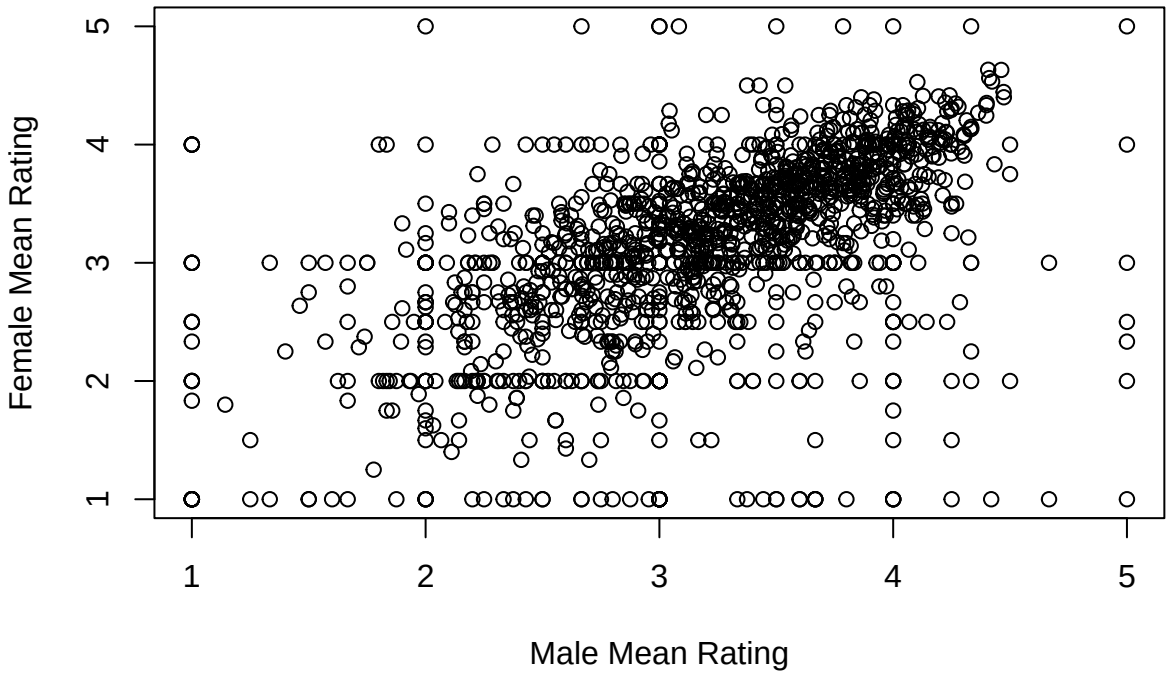
```
gmean_freq_table=merge(gmean_rating_table, grating_freq_table, by= "movie_title")

Males = filter(gmean_freq_table, gender.x == "M")
Males = Males %>% group_by(movie_title) %>%
summarize(mean_rating = mean(mean_rating))
Females = filter(gmean_freq_table, gender.x == "F")
Females = Females %>% group_by(movie_title) %>%
summarize(mean_rating = mean(mean_rating))
gender_summary_merged=merge(Males, Females, by = "movie_title", all = TRUE)

options(dplyr.summarise.inform = FALSE)

plot(gender_summary_merged$mean_rating.x, gender_summary_merged$mean_rating.y, xlab="Male Mean Rating",
```

Figure 5: Male/Female Correlation



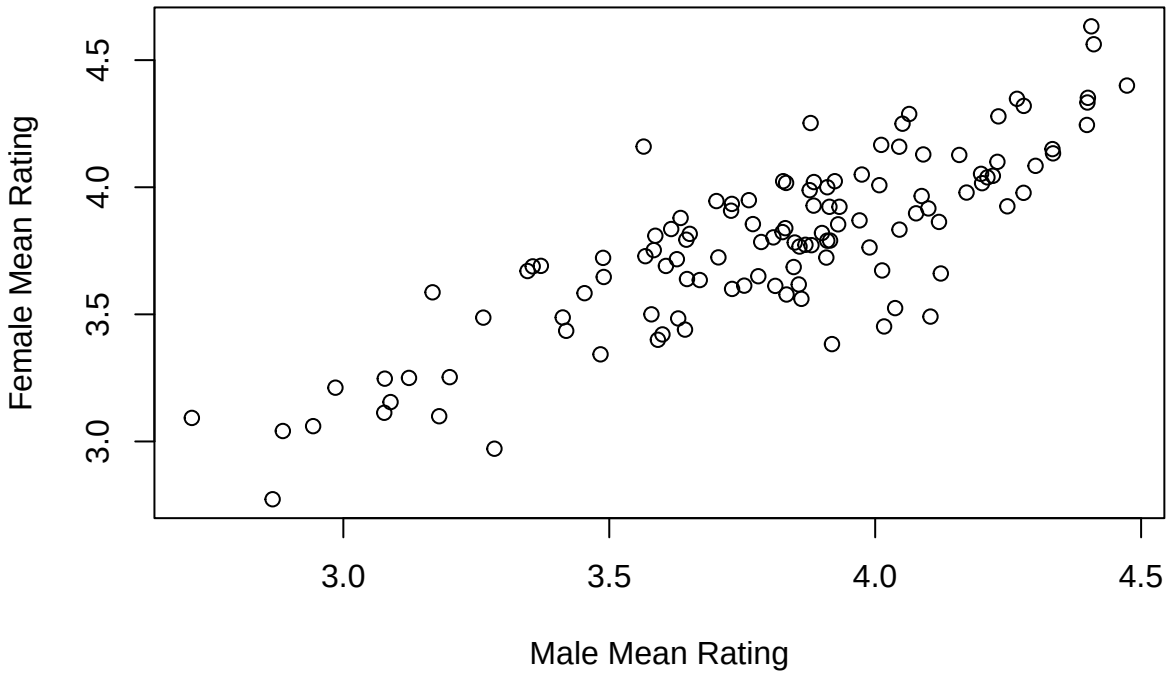
##Figure 6: Men vs. Women Mean Rating by Movie with at Least 200 Ratings

```
gender_summary_merged_frequency=merge(gender_summary_merged, rating_freq_table, by = "movie_title", all
gender_summary_merged_frequency=gender_summary_merged_frequency[order(gender_summary_merged_frequency$ra
gender_summary_merged_frequency=gender_summary_merged_frequency[1:118,]

options(dplyr.summarise.inform = FALSE)

plot(gender_summary_merged_frequency$mean_rating.x, gender_summary_merged_frequency$mean_rating.y, xlab=
```

Figure 6: Male/Female Correlation of Movies with 200+ Ratings



The data is telling us that gender does not have a strong impact on movie rating and, the more ratings you have the smaller the average difference will be between male and female ratings. The more ratings you have on a movie, the more accurately one will be able to predict the rating of one gender when given the other because the will be very close to each other.

Based on the data discussed in section 2, I wanted to make the conjecture that the average ratings between genders would have a greater difference when looking at mean rating by genre. For example, I was thinking that females would rate Romance movies much higher than males. However, my conjecture was wrong. The average difference between male and female ratings for Romance was only 0.05 and the greatest difference by genre was Film-Noir at 0.23. Genre still seems to have a moderate impact on overall mean rating. Among movies with at least 100 ratings, the difference between the highest rated genre (Film-Noir) and the lowest rated genre (Fantasy) 3.92 to 3.28.

Occupation has a moderate impact on rating similar to genre. Healthcare workers give the lowest ratings with a mean rating of 2.96 while lawyers give the highest ratings at 3.74. The highest mean rating actually belonged to occupation: "none" but, I am assuming that mean the occupation was not filled out by the rater and not "unemployed".

Section 4: Conclusion

The data we have discussed to this point has shown us three obvious insights:

1. The trustworthiness of the ratings improves with sample size
2. The variables of gender and age have small impacts on ratings
3. The variables of genre and occupation have moderate impacts on ratings

But, how could a movie company improve the ratings of its movies?

I believe that focusing movies towards specific audiences could improve overall rating. The reason the ratings for healthcare workers are the lowest could be because there is no genre in the data set which appeals specifically to them. Film-Noir and Documentaries are the only two genres that have an average rating above 4 among artists. The nature of these genres likely has a specif appeal to artists.