Investigation of the Movie Ratings

Cody Stancil, Lulu Ge, Seth Green, Yizhe Ge

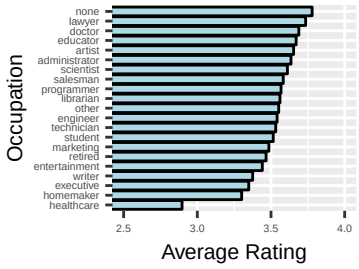
August 2, 2016

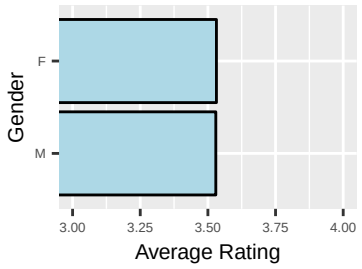
Data Merging

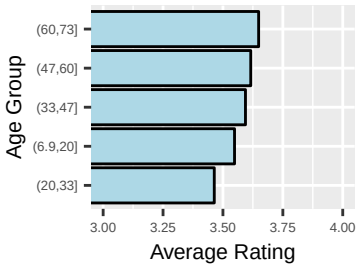
The following data sets were files separated by different delimiters; therefore, we used the read.csv as a method to read in each dataset assigning appropriate variable names. From there, we were able to merge each data set by the `reviewer id` and then the `movie id`. The last data to integrate into the new master data set was the zip code information. We were able to use the zip code information to assign each reviewer to a state in the United States or Canada. After merging all of these data sets the new master data set contained a plethora of information about both the reviewer and the movie that was used to guide our analysis.

Occupation, Gender, Age versus Ratings in General

We first inspected the relationship between different variables and ratings. Among all occupations, lawyer, doctor, and educator have the highest average ratings. For genders, no clear difference was observed between men and women. We assumed that gender does not affect ratings in general. For age group, we saw that older reviewers tend to have higher average ratings. We concluded that older people tend to be more generous on rating the movies.







Further Inspection on Gender versus Ratings

In order to inspect the gender differences, we zoomed into a more granular level. We computed the average ratings for male and female based on genres. The following two datasets show the top ten rated genres for male and female, separately. They show the clear differences: comparing to female, male likes Film-Noir, Documentary, Mystery and Crime more. Western and Sci-Fi appear only in the top-ten list for male, and Musical and Adventure appear only in the top-ten list for female.

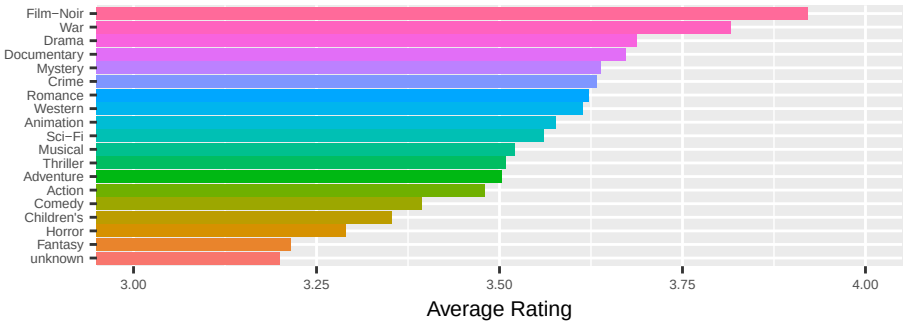
	gender	genre	avgRating
1	M	Film–Noir	3.973294
2	M	War	3.826328
3	M	Drama	3.696957
4	M	Documentary	3.691769
5	M	Mystery	3.664208
6	M	Crime	3.654049
7	M	Western	3.637896
8	M	Romance	3.607072
9	M	Sci–Fi	3.577072
10	M	Animation	3.557471

	gender	genre	avgRating
1	F	War	3.781179
2	F	Film–Noir	3.740260
3	F	Drama	3.662246
4	F	Romance	3.655685
5	F	Musical	3.640083
6	F	Animation	3.627136
7	F	Documentary	3.614973
8	F	Mystery	3.560122
9	F	Crime	3.556299
10	F	Adventure	3.517988

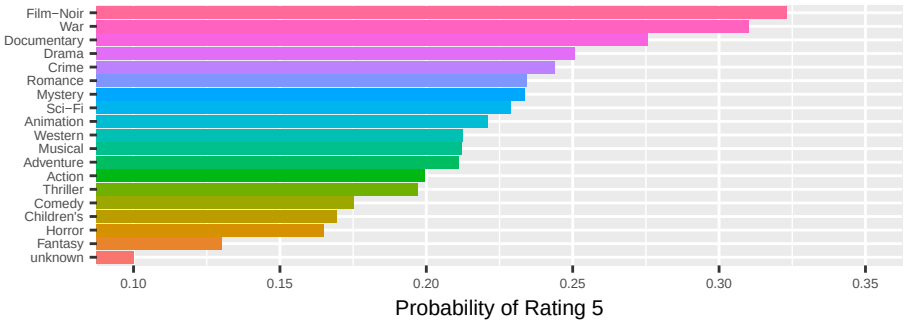
Genres versus Ratings

What seems associated with a high rating? - Genre

If we group the movies by genres (action, animation, comedy, etc), we find out that the top four genres of movies that have highest average ratings are Film Noir, War, Drama, and Documentary.



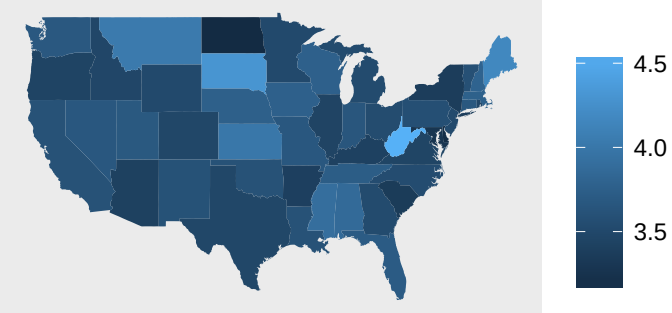
If we compare percentage of top ratings (rating = 5) of each genre, we can see that the first top four genres of movies that have highest percent of high ratings are Film Noir, War, Documentary, and Drama. Although the sorting order is slightly different from the order in average rating, the genre list stays the same. So we may assume that movies with styles Film Noir, War, Drama, and Documentary are more likely to receive high ratings.



Geography: Comparing States

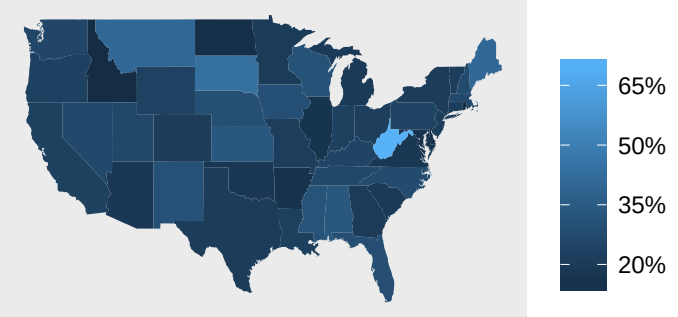
We decided to look at the geographical location of the person giving the review, to see if that had any correlation to the score. Since almost 95% of the data was from the continental United States, we limited our analysis to only those reviewers.

First we calculated an average review score for each state and plotted those scores on a map.



While there is no discernable geographic trend, certain states stick out as having slightly higher averages, notably West Virginia, South Dakota, and Maine. However, the mean rating for our whole sample was 3.53, with a standard deviation of 1.126. When you consider that, no state is actually more than one standard deviation above (or below) the total average.

Next we decided to look at the percentage of reviews from a given state that were 5-star, hoping that this would show a little more variability.

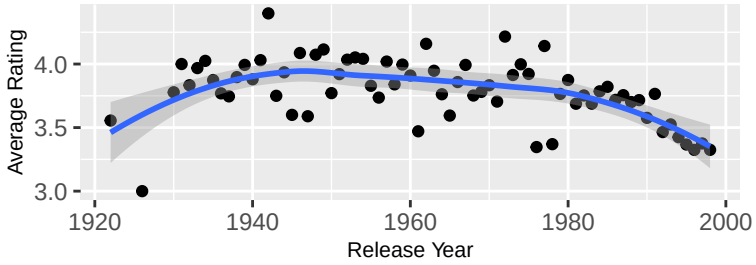


There is indeed slightly more variability, but most notable is the outlier of West Virginia. About 70% of movie reviews coming from West Virginia are 5-star! Any jokes about mountain living aside, this may indicate a problem with our sample. Indeed when we take a quick look at how many unique reviewers there are in some of these states, we get a clue.

In the entire data set, there are only three unique reviewers from West Virginia. Presumably one, or maybe two, of them are very enthusiastic in their ratings. With such a small sample, this enthusiasm does not get evened out by other observations. In fact, it looks like a lot of states are sparsely represented. South Dakota, for instance, is represented by only a single reviewer. This casts some doubt on this entire line of analysis, since our sample may be entirely too small to gain any real geographical insights.

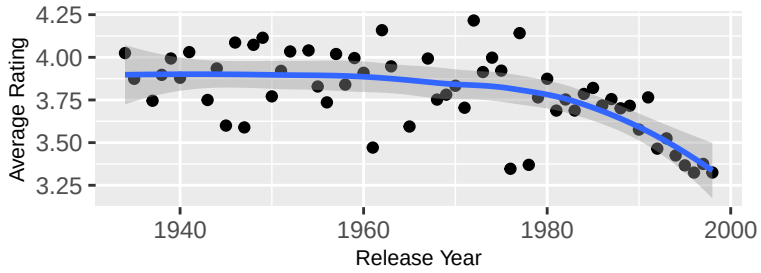
History: Looking at Movies by Release Date

For this section, we grouped the movies by the year that they were released (the `video release date` variable). We then plotted average ratings for the group of movies released in a given year.



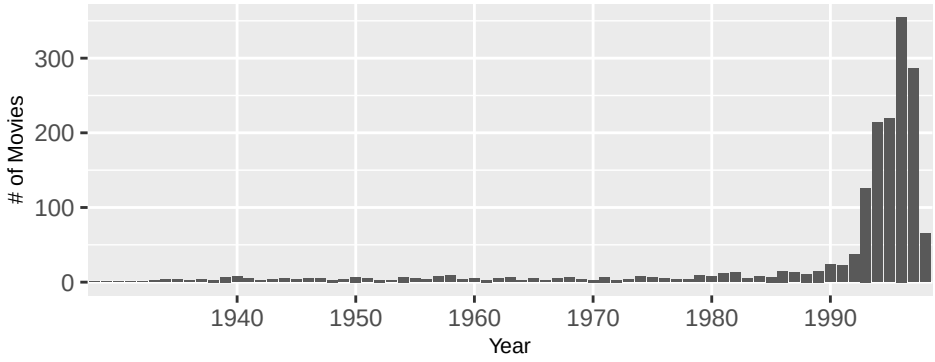
At first glance, it appears that older movies are held in higher estimation by our reviewers. The smoothing line indicates that reviewers were especially hard on movies from the past decade or so (the data was collected in the late 1990's).

We figured there were probably some years that had very few movies and were perhaps obscuring a real trend in the data, so we removed any years with only one or two movies in them and then re-plotted the data.



Removing the sparse years does indeed make the trend much more pronounced. Movies from the 1940's through 1960's are fairly steadily highly rated. This begins to to slip just a little in the 1970's and then takes a real nose dive in the 1980's and 1990's. Are movies from the early days of Hollywood really that much better? Do we just look back with rose-coloured glasses?

We had a hunch, based on our earlier exclusion of sparse years, that maybe there was a discrepency in how many movies were rated from back in the "Golden Days" vs. more recent films. To confirm this, we plotted the number of movies (in our data set) released each year.



The conclusion is overwhelming. Starting in the 1993, the number of movies per year in our data set shoots up by a factor of more than 10. In some cases, there are roughly 100 times as many movies from a given year in the 90's than as some earlier years.

This undermines a persistent falacy that movies (and music and fashion, etc.) used to be better in the good old days. It's really just that we only remember the good stuff. People are still watching 4- and 5-star movies from 50 years ago, but they're not watching (or reviewing) anything bad. Contrast that with the most recent several years, when people are watching a *ton* of different movies: good ones, bad ones, and everything in between. Also, the fact that the jump in count jumps so dramatically at 1993 indicates that maybe this is not a purely random sample of data.

Summary

In conclusion, while it is difficult to actually predict a movie's rating based on this data, we certainly identified some trends. For instance, certain genres (Musicals, Westerns, Sci-Fi) were rated differently by men and women, and certain occupations (doctors, lawyers, the unemployed) rate movies slightly higher. The strongest trend we observed was a higher average rating for older movies, as opposed to movies from the 1980's and 1990's. We discussed a possible explanation for that trend above. However, we identified a serious problem with sample size and skewed sampling in many of the variables. For the year analysis, roughly 75% of our data is for movies released between 1993 and 1998. For the occupation analysis, our highest raters (unemployed) and lowest (healthcare) are only represented by 10 and 16 reviewers respectively. Before drawing any real conclusions, some more research into the source data, and some selective subsetting of the data, could prove fruitful.

Appendix

Question 1:
Which reviewer reviewed the most movies?
Answer:

```
## [1] "Reviewer ID: 405"
```

Question 2:
Which state/territory/Canada produced the top-5 most reviews?
Answer:

```
##    Var1  Freq
## 5    CA 13842
## 25   MN  7635
## 36   NY  6882
## 16   IL  5740
## 46   TX  5042
```

Question 3:
What percentage of reviews involved movies classified in at least two genres?
Answer:

```
## [1] "69.94%"
```

Question 4:
What percentage of movies have 1, 2, 3, ... reviews? (Need a percentage for each.)
Answer:

```
##  Number of Reviews Percentage
##                  1 8.38287753
##                  2 4.04280618
##                  3 3.56718193
##                  4 3.80499405
##                  5 3.03210464
##                  6 2.31866825
##                  7 2.61593341
##                  8 1.78359096
##                  9 1.96195006
##                 10 1.96195006
##                 11 1.18906064
##                 12 1.66468490
##                 13 1.48632580
##                 14 0.83234245
##                 15 1.30796671
##                 16 1.12960761
##                 17 0.59453032
##                 18 1.42687277
##                 19 1.07015458
##                 20 0.71343639
##                 21 0.83234245
##                 22 1.01070155
```

```
##                23 0.71343639
##                24 0.71343639
##                25 0.83234245
##                26 0.83234245
##                27 1.01070155
##                28 0.71343639
##                29 0.53507729
##                30 0.47562426
##                31 0.89179548
##                32 0.83234245
##                33 0.41617122
##                34 0.71343639
##                35 0.41617122
##                36 0.23781213
##                37 0.47562426
##                38 0.41617122
##                39 0.95124851
##                40 0.71343639
##                41 0.65398335
##                42 0.41617122
##                43 0.71343639
##                44 1.01070155
##                45 0.59453032
##                46 0.71343639
##                47 0.41617122
##                48 0.53507729
##                49 0.47562426
##                50 0.59453032
##                51 0.23781213
##                52 0.29726516
##                53 0.47562426
##                54 0.35671819
##                55 0.23781213
##                56 0.17835910
##                57 0.53507729
##                58 0.47562426
##                59 0.59453032
##                60 0.29726516
##                61 0.11890606
##                62 0.23781213
##                63 0.35671819
##                64 0.71343639
##                65 0.47562426
##                66 0.71343639
##                67 0.53507729
##                68 0.41617122
##                69 0.35671819
##                70 0.23781213
##                71 0.29726516
##                72 0.35671819
##                73 0.35671819
##                74 0.23781213
##                75 0.23781213
##                76 0.17835910
```

```
##                77 0.23781213
##                78 0.11890606
##                79 0.29726516
##                80 0.35671819
##                81 0.41617122
##                82 0.41617122
##                83 0.11890606
##                84 0.11890606
##                85 0.29726516
##                86 0.35671819
##                87 0.29726516
##                88 0.05945303
##                89 0.29726516
##                90 0.29726516
##                91 0.35671819
##                92 0.35671819
##                93 0.35671819
##                95 0.17835910
##                96 0.23781213
##                97 0.23781213
##                98 0.17835910
##                99 0.05945303
##               100 0.23781213
##               101 0.41617122
##               102 0.23781213
##               103 0.05945303
##               104 0.29726516
##               105 0.05945303
##               106 0.23781213
##               107 0.05945303
##               108 0.05945303
##               109 0.05945303
##               110 0.05945303
##               111 0.11890606
##               112 0.23781213
##               113 0.05945303
##               114 0.17835910
##               115 0.23781213
##               116 0.23781213
##               117 0.17835910
##               118 0.05945303
##               119 0.17835910
##               120 0.17835910
##               121 0.35671819
##               122 0.05945303
##               123 0.05945303
##               124 0.41617122
##               125 0.23781213
##               126 0.11890606
##               127 0.35671819
##               128 0.41617122
##               129 0.23781213
##               130 0.11890606
##               131 0.11890606
```

```
##                 132 0.05945303
##                 133 0.05945303
##                 134 0.23781213
##                 136 0.23781213
##                 137 0.41617122
##                 138 0.11890606
##                 142 0.05945303
##                 143 0.17835910
##                 145 0.11890606
##                 146 0.05945303
##                 147 0.11890606
##                 148 0.23781213
##                 149 0.17835910
##                 150 0.11890606
##                 151 0.17835910
##                 152 0.05945303
##                 153 0.11890606
##                 154 0.05945303
##                 155 0.05945303
##                 156 0.05945303
##                 157 0.17835910
##                 158 0.11890606
##                 160 0.17835910
##                 161 0.05945303
##                 162 0.29726516
##                 163 0.05945303
##                 164 0.17835910
##                 165 0.05945303
##                 166 0.05945303
##                 168 0.11890606
##                 169 0.17835910
##                 170 0.23781213
##                 171 0.35671819
##                 172 0.11890606
##                 173 0.05945303
##                 174 0.17835910
##                 175 0.17835910
##                 176 0.11890606
##                 177 0.05945303
##                 178 0.11890606
##                 179 0.17835910
##                 180 0.17835910
##                 182 0.11890606
##                 183 0.05945303
##                 184 0.05945303
##                 185 0.05945303
##                 187 0.05945303
##                 188 0.05945303
##                 189 0.05945303
##                 190 0.05945303
##                 191 0.05945303
##                 192 0.05945303
##                 193 0.05945303
##                 194 0.11890606
```

```
##                 195 0.05945303
##                 197 0.05945303
##                 198 0.11890606
##                 199 0.05945303
##                 200 0.05945303
##                 201 0.11890606
##                 202 0.05945303
##                 206 0.17835910
##                 208 0.11890606
##                 209 0.17835910
##                 211 0.05945303
##                 212 0.05945303
##                 213 0.05945303
##                 215 0.05945303
##                 216 0.05945303
##                 217 0.05945303
##                 218 0.05945303
##                 219 0.23781213
##                 220 0.11890606
##                 221 0.17835910
##                 222 0.05945303
##                 223 0.05945303
##                 226 0.05945303
##                 227 0.11890606
##                 230 0.11890606
##                 231 0.05945303
##                 232 0.05945303
##                 236 0.05945303
##                 239 0.11890606
##                 240 0.11890606
##                 241 0.05945303
##                 243 0.11890606
##                 244 0.11890606
##                 246 0.05945303
##                 247 0.05945303
##                 250 0.05945303
##                 251 0.23781213
##                 254 0.05945303
##                 255 0.05945303
##                 256 0.11890606
##                 259 0.11890606
##                 261 0.05945303
##                 264 0.05945303
##                 267 0.11890606
##                 268 0.05945303
##                 272 0.05945303
##                 275 0.05945303
##                 276 0.11890606
##                 280 0.11890606
##                 283 0.05945303
##                 284 0.05945303
##                 290 0.05945303
##                 291 0.05945303
##                 293 0.17835910
```

```
##                 295 0.11890606
##                 297 0.11890606
##                 298 0.11890606
##                 299 0.05945303
##                 300 0.05945303
##                 301 0.05945303
##                 303 0.05945303
##                 315 0.05945303
##                 316 0.11890606
##                 321 0.05945303
##                 324 0.05945303
##                 326 0.05945303
##                 331 0.05945303
##                 336 0.05945303
##                 344 0.05945303
##                 350 0.11890606
##                 365 0.05945303
##                 367 0.05945303
##                 378 0.05945303
##                 384 0.05945303
##                 390 0.05945303
##                 392 0.05945303
##                 394 0.05945303
##                 413 0.05945303
##                 420 0.05945303
##                 429 0.05945303
##                 431 0.05945303
##                 452 0.05945303
##                 478 0.05945303
##                 481 0.05945303
##                 485 0.05945303
##                 507 0.05945303
##                 508 0.05945303
##                 509 0.05945303
##                 583 0.05945303
```