STAT 6430 – Final Project – R

Tyler Lewris: tal3fj

Rohan Bapat:

Isabelle:

Final Project Report

**Introduction:**

The main objective of our project is to understand and manipulate several related data sets to discover an association related to high movie ratings and determine if certain groups of individuals are more likely to provide a higher movie rating. Our approach was structured and methodical: physically look at each data set before reading into R, read in the data, answer specified questions, generate hypothesis, perform comparisons and analysis on the data, visualize our findings.

**Main Objectives:**

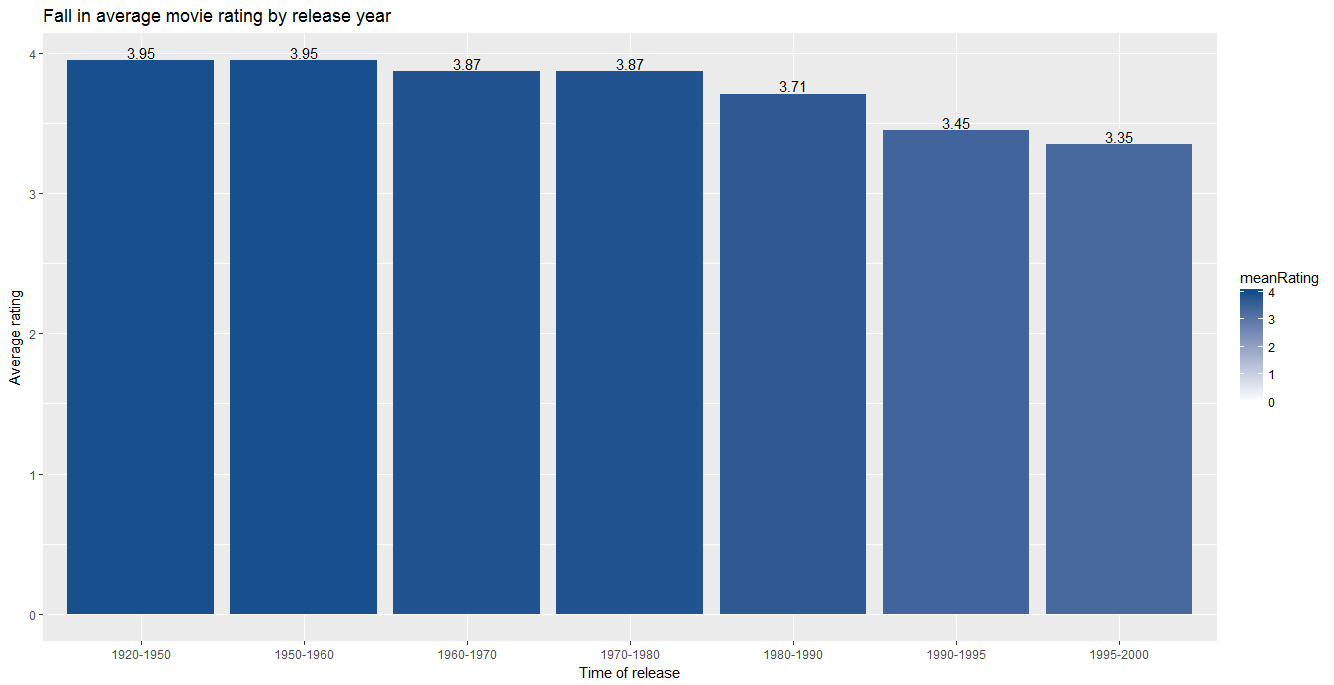
What seems to be associated with a high rating?

The first connection we wanted to define was whether or not certain states provided higher ratings than others. At first glance, it is clear that California provided the greatest number of reviews in our data set.

What groups are most likely to provide higher ratings?

Graphs:

1. Number of reviews by state
2. Average rating by movie release decade
3. Age by Genre
4. Gender by Genre



**Conclusion:**

**Appendix 1**: Answers to Questions

1. **Which percentage of each rating was given?**

Rating: Percentage:

1. 6.11%
2. 11.37%
3. 27.145%
4. 34.174%
5. 21.201%
6. **Which reviewers were the top-10 in terms of number of movies reviewed? (Provide the reviewer number and the number of movies reviewed. If there is a tie for 10th place, include all that tied.)**

Reviewer ID: Number of Movies Reviewed:

405 737

655 685

13 636

450 540

276 518

416 493

537 490

303 484

234 480

393 448

1. **Find a 95% confidence interval for the average rating among all reviewers, and a 95% confidence interval for the average rating among the top-10 reviewers. Does there appear to be evidence that the two groups differs?**

Upper and lower limits of 95% confidence interval for all reviewers:

3.559737 and 3.616645

Upper and lower limits of 95% confidence interval for top-10 reviewers:

2.754554 and 3.586818

Does there appear to be evidence that the two groups differ?

Although the confidence intervals slightly overlap, top-10 reviewers has a lower limit than that of all reviewers which can differentiate the two groups. However, it is difficult to make conclusions off this data alone, as the sample size of the top-10 reviewers is clearly lower than that of all reviewers. Further research could be done to answer this question in its entirety. For instance, understanding the standard deviation of just the top-10 reviewers and comparing it to the standard deviation of all reviewers without factoring in sample size could provide more insight amongst the two groups.

All: 0.4452335

Top10 : 0.5817123

95% confidence that the mean of that sample will lie within the upper end and lower range

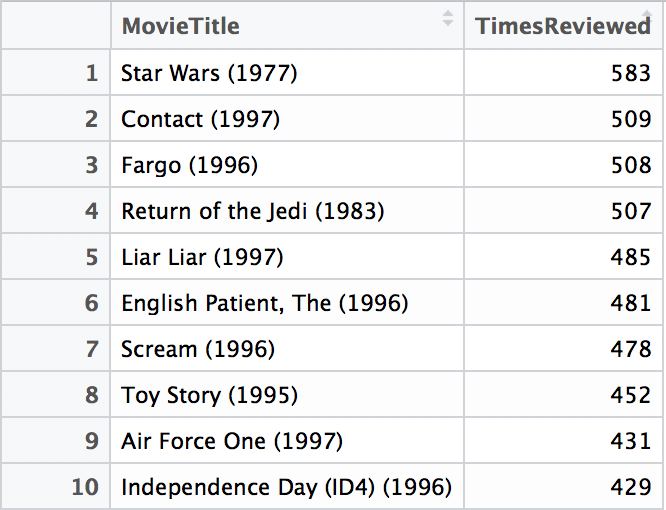
mean 3.588191for all review

mean 3.170686 for top ten

samples of data set over and over again- calculated corresponding interval – 95% of that data would contain the true mean

1. **Which movies were the top-10 based on of number of times reviewed? (Provide the movie title and the number of times reviewed. If there is a tie for 10th place, include all that tied.)**

Top-10 movies based on number of times reviewed:

****

1. **Which genre occurred most often, based on the number of reviews. Which was least often? (Don’t include “unknown” as a genre for this question.)**

Genre that occurred most often based on number of reviews:

Drama : 39895 reviews

Genre that occurred least often based on number of reviews:

Documentary : 758 reviews

1. **What percentage of reviews involved movies classified in at least two genres?**

69.938% of all reviews involved movies classified in at least two genres.

1. **Give a 95% confidence interval for the average rating for male reviewers, and do the same for female reviewers.**

95% confidence interval for the average rating for male reviewers:

3.521309 and 3.537269

95% confidence interval for the average rating for female reviewers:

3.517202 and 3.545813

1. **Which state/territory/Canada/unknown produced the top-5 most reviews?**

State/Territory/Canada/Unknown that produced the top-5 most reviews:

CA, MN, NY, IL, TX

1. **What percentage of movies have exactly 1 review? 2 reviews? 3 reviews? Continue to 20 reviews.**



1. **Which genre had the highest average review, and which had the lowest average review?**

Genre with highest average review: FilmNoir

Genre with lowest average review: Unkown

1. **Repeat the previous question, for reviewers age 30 and under and then for reviewers over 30.**

Highest average review for reviewers age 30 and under: FilmNoir

Lowest average review for reviewers age 30 and under: Fantasy

Highest average review for reviewers over 30: FilmNoir

Lowest average review for reviewers over 30: Unknown