Data Analysis Project 1
Introduction to Data Science, Section 001

Fall Semester, 2023
Group: DS 34
Members: Tanvi Bansal, Natasha Recoder, Tyler Perez

Introduction

For the movie dataset it was not reasonable to reduce the data to sample means as we could not assume that the difference between a rating of a 1 and a rating of a 2 is the same as the difference between a rating of a 3 and a rating of a 4. Additionally, we cannot assume these ratings are standardized across participants. One person's 3 may be another person's 2. The data was not categorical. We thus ran the Mann Whitney U test and the KS test when comparing two groups or the Kruskal-Wallis test when comparing more than two groups. When running the KS and Mann Whitney U test we believed that not only knowing the medians, but considering the ratings distributions of each group was important. The distributions could give us additional information, such as 'Was the movie polarizing?', thus we prioritized the results of the KS test. However, the median can give us a quick idea of how well-liked the movie was overall, thus we ran both tests.

In order to handle incomplete data we removed NaNs and missing data element wise, with the exception of Q7, Q8, Q10 that required special treatment. This helped maintain statistical power as when answering certain questions we did not have much available data. In order to run these analyses we had to make the assumption that element wise removal of NaNs does not introduce bias, which may not always be true given viewers self-selected into watching certain movies, however we must make the assumption it does not introduce bias in order to run our statistical analyses. We do not have to worry about there being unequal n between groups as the selected tests do not require equal sample sizes.

Q1

Are movies that are more popular rated higher than movies that are less popular?

To determine whether more popular movies are rated higher than less popular movies, we first did a median split on total ratings for each movie to categorize movie popularity, then performed a Mann Whitney U test to compare medians between the two groups of data..

As there were 600 ratings in each group our degrees of freedom was 300.25. We first calculated the median values for each set in order to support findings from our test. We thought more popular movies would yield higher ratings, as users are more likely to see movies that other people dub as good.

The data corroborated that to an extent but not a statistically significant manner given our alpha level, as the median rating for popular movies was 1.0 and 0.5 for not popular ones. Our Whitney Test statistic was U = 169068.0 with an associated p value of 0.03077. Our alpha level was 0.005, therefore we do not reject the null hypothesis. We can conclude that there is no difference in ratings between more popular movies and less popular movies; any observed differences are due to chance. The experiment resulted with an effect size of 0.0607 using Cohen's U, and a power of 0.1830. We have weak power therefore a strong likelihood of a false negative. This is the reality we are given as we did not have ample amount of data.

Q2

Are movies that are newer rated differently than movies that are older?

In this question we were tasked to determine whether more recently released movies are rated differently than movies that are older. To answer this question with the utmost efficacy we employed the KS test to compare the distributions between groups, and the Mann Whitney U test to detect differences in the ratings themselves.

We needed to do some housekeeping in regards to our data prior to calculating any significant values. To begin we calculated the median release year from all of the movies we had ratings for resulting in m = 1999. All movies released before or in 1999 were put into the old movie dataframe, while ones released after 1999 were put into the new movie data frame.

The medians of both groups were equal at 3.0 However, the test statistic results were surprising. The KS statistic was D = 0.009 with associated p value = 0.023. The Mann Whitney U test statistic yielded a U value = 1556898873.0 and p value = 0.001! These two tests were used in unison to first test whether the two sections of our data come from a population with the same sample median, and then secondly to see whether they come from the same distribution. Given our data we rejected the null and concluded that there was a statistically significant difference between the ratings of old and new movies. The KS test results are supported by the histogram of both groups as the groups appear visually similar. However, the Mann Whitney U test results are contradictory to the previous findings that the median values are equivalent between groups, and both groups originate from the same distribution. We have a very large sample size here (n_old=54,948, n_new=57,266) which yields a great amount of power therefore we are likely picking up on smaller effects.. Code used to produce these results can be seen in Appendix III. We could not conclude that the differences in distributions are too great to simply happen by chance, and therefore could not reject the null hypothesis.

Q3

Is enjoyment of 'Shrek (2001)' gendered?

To answer the question "is the enjoyment of Shrek gendered?" We used a KS test to compare the distribution of ratings between male and female viewers, and a Mann Whitney U test to compare the medians. We used ratings as a proxy for enjoyment.

The KS test resulted in test statistic D of 0.153, corresponding to a p value of 0.056. Given the p value is greater than our α level of 0.005 we cannot reject the null hypothesis that the distribution of enjoyment of Shrek is not gendered. We therefore maintain our null hypothesis that the enjoyment of Shrek is not gendered. Our findings were supported visually **(Figure 3a, 3b)**. These distributions did not look identical, nor dramatically different, which was supported by our statistical analysis which confirmed that we could not assume that these distributions were different by an effect other than chance.

To learn more about our data we also ran the Mann Whitney U test. The median of the male ratings was 3.0 while the median female rating was 3.5. The test resulted in a U statistic of 82232.5 and a p value of 0.025; the p value was greater than our α level of 0.005, therefore we could not reject our null hypothesis that the median rating of Shrek by men does not differ from the median rating of Shrek by women. This supported our finding from the KS test that the distribution of the ratings of Shrek between men and women is not significantly different.

We therefore conclude that the enjoyment of Shrek is not gendered, with the limitation that we had to assume element-wise removal of NaNs would not bias the data. We were further limited by a data set that is not large leading to lowered statistical power, particularly in the male ratings of Shrek.

Q4

What proportion of movies are rated differently by male and female viewers?

To determine what proportion of movies are rated differently by male and female viewers, for each movie we ran the KS test to determine if the distribution of male vs female ratings differed for each movie and the Mann Whitney U test to determine if the median ratings between groups differed significantly, then calculated the proportion that had statistically significant differences.

The KS test resulted in 25 movies having a p value smaller than our α level of 0.005; this finding indicated that there was a statistically significant difference in ratings distribution between males and females in 6.25% of movies. The Mann Whitney U test resulted in a statistically significant difference in between the median rating of a particular movie by male vs female viewers 15.25% of the time.

We believe that the overall distribution of ratings of a movie is more informative than one median. This is because how the ratings are distributed can tell us a lot about how the movie was perceived. Thus we conclude that 6.25% of movies are rated differently by male and female viewers.This must be put into the context that our sample sizes were relatively small for each movie, thus we did not have much statistical power, and we had to make the assumption that element-wise removal and NaNs and self-selection into watching certain movies would not bias our data. Another limitation is potential alpha-inflation. We made many comparisons (400, one per movie) and thus the more comparisons we make, the more likely we are to find statistical significance

Q5

Do only children enjoy 'The Lion King (1994)' more than people with siblings?

In this question we were tasked with determining whether participants who were an only child enjoyed "The Lion King (1994)" more than people with siblings. We performed a Mann Whitney U test to determine whether there was a significant median difference in ratings between these two groups.

We ran the Mann Whitney U test resulting in a test statistic U of 52929.0 with an associated p value of 0.0216. Given that the p value is not lower than our α level of 0.005, we fail to reject the null hypothesis and conclude that the difference in ratings for The Lion King between only children and non-only children could indeed happen by random chance. However the observed non statistically significant difference was not in the direction in which the question prompted, as it was actually the non only childs who gave a higher median rating to the movie. This however seems to be a product of chance more than anything else.. The calculated median rating value for only children and non-only children was 3.5 and 4.0, respectively, which would indicate children with siblings rating the movie higher. Through some analysis and postulation, this can make some sort of sense. As the Lion King was a childrens movie, one would be more inclined to watch the movie if they had someone close to them to watch it with. Having a sibling to watch it with may lead to more enjoyment. Yet with that being said we cannot say with great confidence that this categorical variable truly is a great identifier regarding whether someone will enjoy the Lion King more. Our effect size was 0.0966 and our power was 0.0425. This is catalyzed by our minimal data available, and leads to be more prone in making a Type 2 error.

Q6

What proportion of movies exhibit an "only child effect"?

We were tasked to determine what proportion of movies are rated differently by viewers with siblings and viewers without siblings. Previously we saw there was a strong only child effect for the Lion King, it would be captivating to see if many other movies also possess this effect.

The KS test was run on the ratings between the only children group and non-only children group for each movie. The results indicated that only 3 movies ("Happy Gilmore", "Billy Madison", and "Toy Story") had a p value lower than our α level of 0.005, allowing us to reject the null hypothesis that there is no difference in ratings between groups for 3/400 movies. We conclude that 0.75% of movies possessed a strong enough only child effect in which the observed data was too unlikely to happen just by chance. There are some key takeaways from this number. One being that the existence of an only child effect is observed to be superficial and not very existent in the slightest given our experiment. Whether one grows up as an only child or not does not seem to be a dominant predictive or differential factor in movie preference. This comes with caveats: we cannot state these conclusions with great confidence. This lack of confidence comes from the low amount of Power that comes with KS Tests.

Given we are comparing each movie to itself in the two dataframes, the power and the probability of making a Type II error varies greatly from one movie to another due to differing amounts of data for each movie.

Q7

Do social movie watchers enjoy 'Wolf of Wall Street (2013)' more than those who prefer to watch movies alone?

In order to compare the enjoyability of The Wolf of Wall Street between social movie watchers and alone movie watchers, the KS test was used to compare the ratings distributions and the Mann Whitney U test to approximate median differences between the two groups. We row-wise removed missing and "no response" entries to limit bias in the assignment of groups. A two-way KS test is run resulting in a test statistic value of D = 0.064 and a p value of 0.524. Given an $\alpha = 0.005$, we cannot reject the null hypothesis that there is no difference in the distributions of the two groups. Therefore we proceed with the conclusion that the distributions of ratings between groups are the same. Although the distributions appear similar in Figure 7 (Appendix I), the results imply that any observed differences are likely due to chance alone.

A two-way Mann Whitney U test is run resulting in a test statistic value of U = 49034, and p value = 0.113. We estimate the effect size as ES = 0.465, indicating mediocre power. Given $\alpha = 0.005$ and p value = 0.113, we see p value > α and therefore cannot reject the null hypothesis that there is no median location difference between groups. We conclude that there is no difference in the enjoyability of The Wolf of Wall Street between groups. Although there is evidence that we have some risk of committing a β-error, we are limited by the sample size of available data to pursue further risk mitigation efforts.

Q8

What proportion of movies have a "social watching" effect?

To determine which proportion of movies have a social watching effect, the KS test and Mann Whitney U test were used to determine distribution and median differences. The movies that exhibit significantly different *positive* median location differences for social watchers (ratings median shifted higher than alone watchers) are counted and computed as a proportion of all movies. Row-wise removal of missing and "no response" data was chosen to limit bias in the assignment of groups.

A KS test is run resulting in a test statistic value D = 0.009 and p value = 0.026. With an $\alpha = 0.005$, we cannot reject the null hypothesis that there is no difference between the distribution of ratings between groups across all movies. We proceed with the conclusion that the distribution of ratings between social movie watchers and alone movie watchers are the same, further supported by the histograms of each group in Figure 8 (Appendix I). A Mann Whitney U test is run between group 1 and group 2 for each movie resulting in 10 out of 400 movies with a p-value < $\alpha = 0.005$. Of these 10 movies, we find from a one-way Mann Whitney U test that 6 movies have a *positive* location shift of ratings. For these 6 movies, an effect size is estimated as ranging from 0.560 - 0.690 by calculating the proportion of the obtained U value to the maximum U value. This implies mediocre power for this finding. Given the p < $\alpha = 0.005$ for 6 movies, we conclude that 6/400, or 1.5% of movies have a social watching effect. Although there is some evidence of β-error risk, we are limited by the sample size of available data and the cleaning techniques required to eliminate bias or confounding effects.

Q9

Is the ratings distribution of 'Home Alone' different than that of 'Finding Nemo'?

To answer the question, 'is the ratings distribution of 'Home Alone' different than that of 'Finding Nemo' we ran the KS test to compare the ratings distributions of the two movies. We did not run the Mann Whitney U test as we are comparing the ratings distributions, not simply the medians.

Running the KS test yielded a test statistic of $D = 0.153$ with a corresponding p-value of 6.378×10^{-10}. This is below our α of 0.005. Thus, we rejected our null hypothesis that the distributions were identical and concluded that the ratings distribution of Home Alone is significantly different from the ratings distribution of Finding Nemo given the differences were extremely unlikely to have arisen from chance. Inspecting visually, we see in figures 9a, 9b below that there is a higher concentration of 4 ratings for Finding Nemo than Home Alone, while Home Alone has a higher proportion of 3 ratings. We must take this conclusion into context of our limitations, including element-wise removal of NaNs and self-selection of participants into watching any given movie.

Q10

How many franchises contain movies with inconsistent quality?

To determine which franchises contain movies with inconsistent quality we used a Kruskal-Wallis test to detect median differences between movies. For each franchise we row-wise removed NaNs and missing data entries to eliminate any systematic differences in users who have not watched all movies. The Kruskal-Wallis test was run on the user ratings between movie groups, repeated for all 8 franchises (data table in Appendix II). The test resulted in two out of eight franchises - Harry Potter (p value = 0.118), Pirates of the Caribbean (p value = 0.036) - having larger p values, and the remaining six having very low p values (p < 10e-5). We estimate effect sizes for Harry Potter and Pirates of the Caribbean franchises as 0.002, 0.004 respectively. This indicates very low power.

Given a significance level $\alpha = 0.005$, we reject the null hypothesis for all but two franchises and conclude that six movies have shifted rating locations relative to one another. Although there is evidence of committing a β-error, we are limited by the effort to eliminate systematic differences between users. It may be more insightful to accept that ratings from users that have not watched all movies in a franchise may introduce some bias but help to achieve a higher power. We accept the conclusion that 6 movie franchises have inconsistent movie quality.

Extra Credit

Does a person's distractibility affect how much they enjoy movies?

We investigated whether how much someone gets distracted as a whole impacts how they rate movies. To answer this question we took people's self-reported level of distractibility (1-5) and compared how people in each of these distractibility levels rate movies overall. We use a Kruskal-Wallis test given the reasons stated in the introduction. We compared all the ratings of any movie from each distractibility level group compared to another.

Running the Kruskal-Wallis test resulted in a Kruskal Wallis H statistic of 515.8 which, with 4 degrees of freedom, corresponds to a p-value of 2.536×10^{-110}. Given this is lower than our α of 0.005 we can therefore confidently reject the null hypothesis that the population median of all groups are equal. Thus, we know that how much someone gets distracted as a whole impacts how they rate movies. We must take our limitations into consideration however when coming to this conclusion. People self-rated their own levels of distractibility, and thus this cannot be a fully reliable metric of their true levels of distraction. We also must consider that people self-selected into watching certain movies, and perhaps there is a confounding factor that causes one distractibility group to watch a certain group of movies more than another group, which could result in different ratings.

Appendix

I. Figures

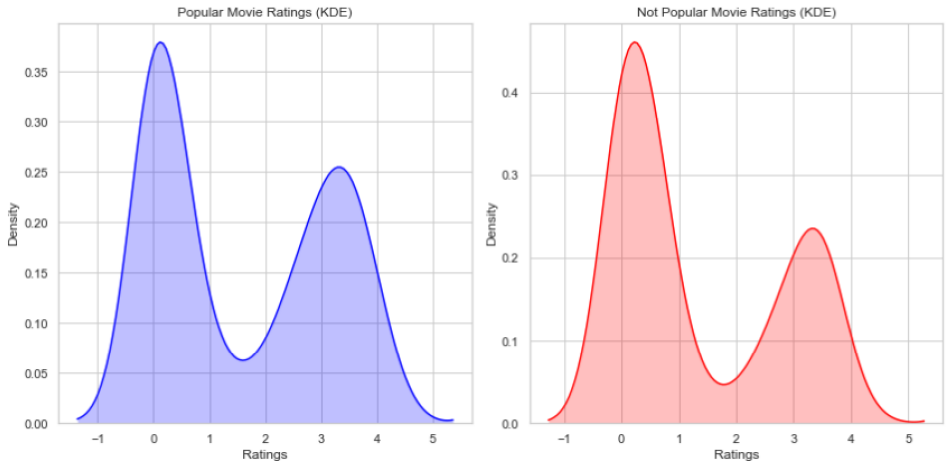

Figure 1. KDE of Popular Movies (Left) vs. Non Popular Movies (Right)

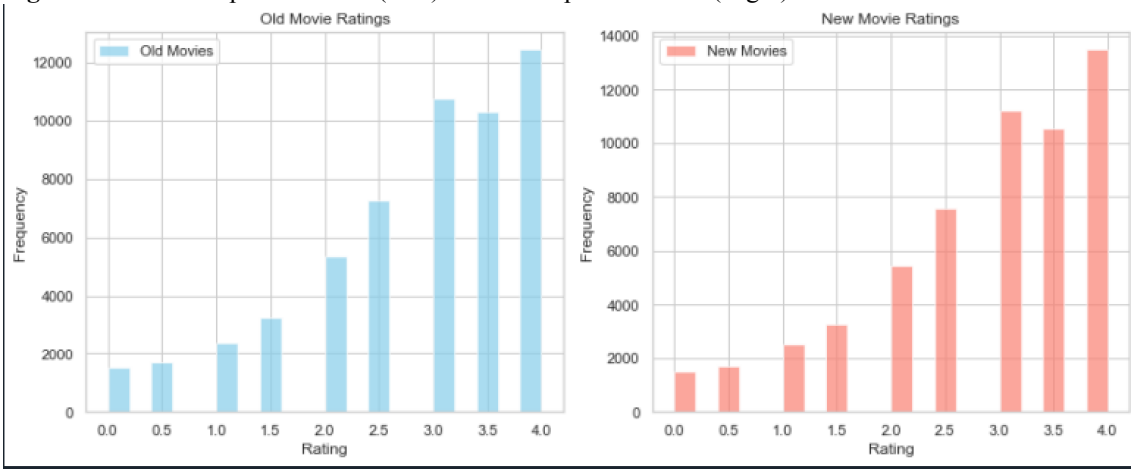

Figure 2. Old Movie Histogram (Left) vs. New Movie Histogram (Right)

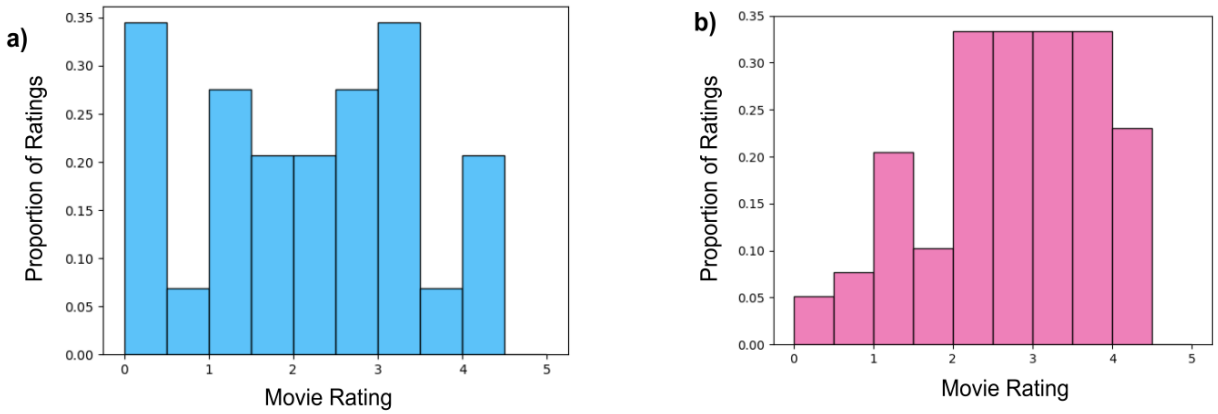

Figure 3. Ratings distributions of 'Shrek (2001)' by gender. Rating options include 0 through 4, inclusive, in 0.5 unit increments. Histogram is left inclusive. **(a)** Ratings by male viewers. **(b)** Ratings by female viewers

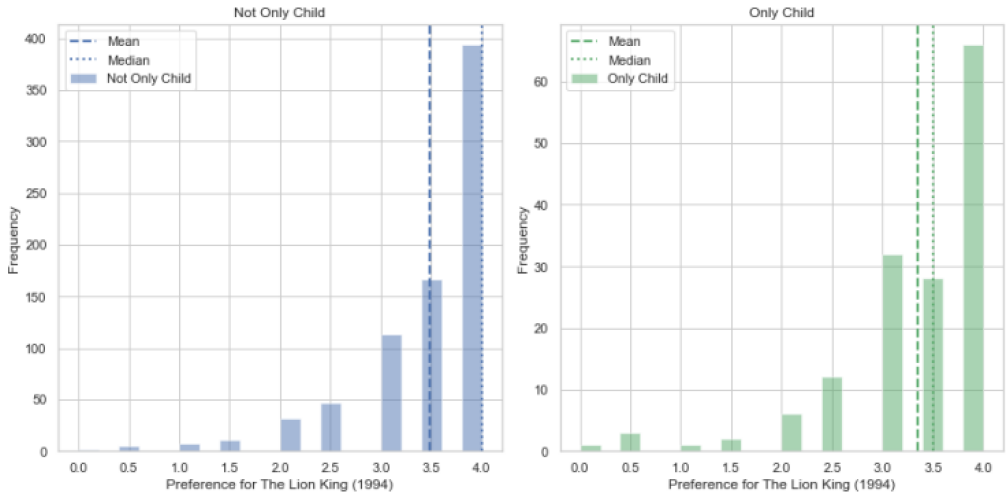

Figure 5. Histogram of ratings of non-only child ratings for "The Lion King (1994)" (Left) versus those of only children (Right)

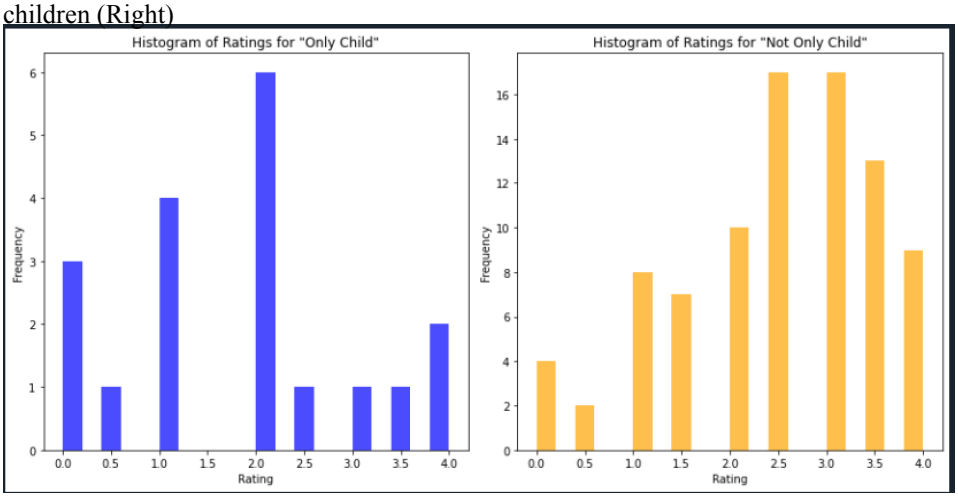

Figure 6. Histogram of ratings from only children (Left) with a histogram of the ratings given by non-only children (Right).

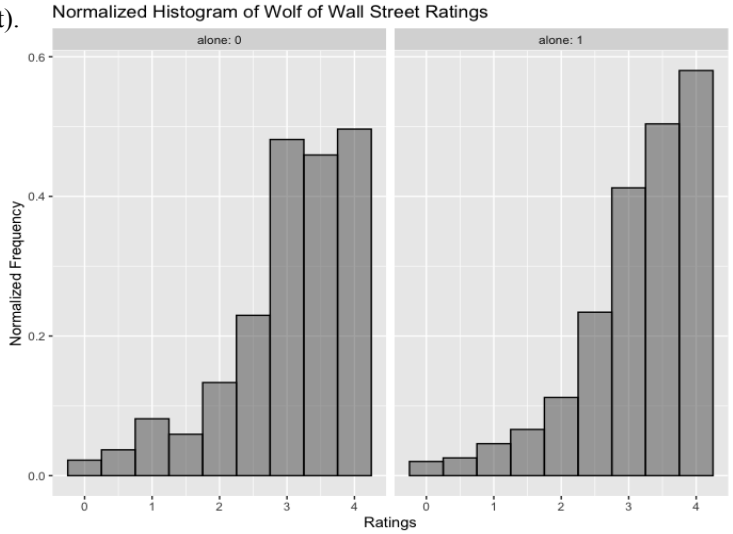

Figure 7. Ratings histogram for Wolf of Wall Street between social movie watchers (Left) and alone movie watchers (Right)

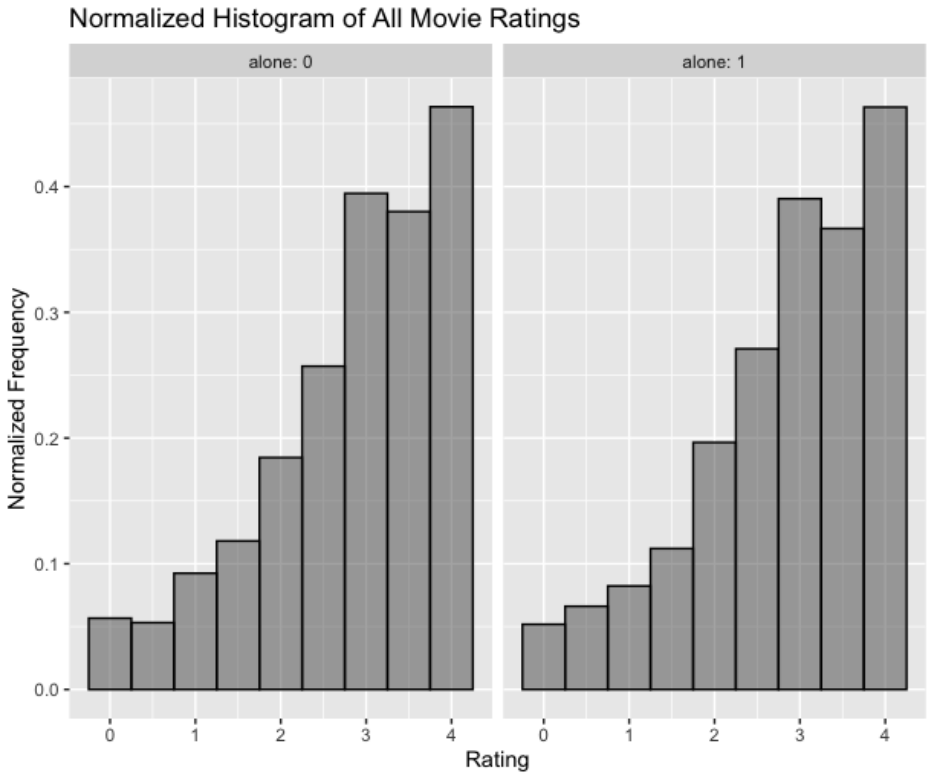

Figure 8. Distribution of ratings for all movies between social movie watchers (left) and alone movie watchers (right)

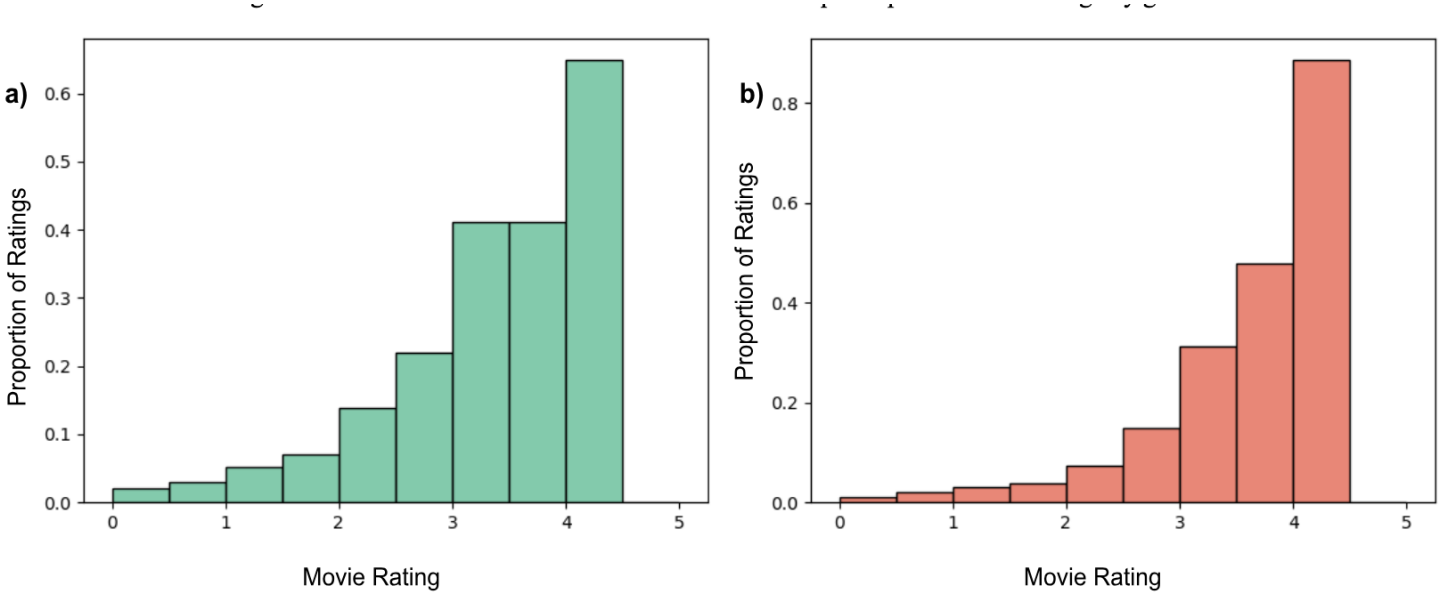

Figure 9. Ratings distributions of 'Home Alone (1990)' and 'Finding Nemo (2003)'. Rating options include 0 through 4, inclusive, in 0.5 unit increments. Histogram is left inclusive. **(a)** Ratings of 'Home Alone (1990)'. **(b)** Ratings of 'Finding Nemo (2003)'

II. Data table result from Kruskal-Wallis test performed on ratings between movies within a franchise, compiled over all franchises (referenced in Q10)

Franchise	Degrees of Freedom	Test Statistic	P-value
Star Wars	5	193.510	6.940e-40
Harry Potter	3	5.874	0.118
The Matrix	2	40.323	1.753e-9
Indiana Jones	3	54.194	1.020e-11
Jurassic Park	2	49.427	1.849e-11
Pirates of the Caribbean	2	6.660	0.036
Toy Story	2	23.497	7.902e-6
Batman	2	84.658	4.138e-19

III. Code for Q3

```
import pandas as pd
from scipy.stats import mannwhitneyu
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import ks_2samp
import numpy as np

# Read the CSV file into a DataFrame
file_path = '/Users/tylerperez/Downloads/movieReplicationSet.csv'
movies = pd.read_csv(file_path)  # Skip the first row containing movie titles
movies = movies.iloc[:, 0:400]

# Extract years from the headers and store them in a list
years = [int(name.split("(")[-1].split(")")[0]) for name in movies.columns]

# Calculate the median release year
median_year = int(pd.Series(years).median())

# Create two DataFrames based on the median year
old_movies = movies.loc[:, [year <= median_year for year in years]]
new_movies = movies.loc[:, [year > median_year for year in years]]


# Extract ratings into separate DataFrames for both old and new movies
old_ratings = old_movies.stack().reset_index(drop=True)
```

```
new_ratings = new_movies.stack().reset_index(drop=True)

# Remove null values from old_ratings and new_ratings
old_ratings = old_ratings.dropna()
new_ratings = new_ratings.dropna()


# Remove rows with blank values (empty strings)
old_ratings = old_ratings[old_ratings != ""]
new_ratings = new_ratings[new_ratings != ""]


# Perform a Mann-Whitney U test to compare old and new ratings
statistic, p_value = mannwhitneyu(old_ratings, new_ratings)
# Perform a two-sample KS test to compare old and new ratings
statistic_ks, p_value_ks = ks_2samp(old_ratings, new_ratings)

# Create two histograms using Matplotlib
plt.figure(figsize=(10, 6))
plt.hist(old_ratings, bins=20, alpha=0.5, color="blue", label="Old Movies")
plt.hist(new_ratings, bins=20, alpha=0.5, color="orange", label="New Movies")
plt.xlabel("Ratings")
plt.ylabel("Frequency")
plt.title("Distribution of Ratings for Old vs. New Movies")
plt.legend()
plt.show()
```