FA-25 DS-GA 1001
Interview in Data Science

Data Analysis Project 1
Hypothesis Testing of Movie Rating Data

Group: Luke Ducker, Joseph Tadros, Beibarys Nyussupov

Abstract: This report aims to answer the following ten questions to assist Fictitious Movie Studio Ltd in optimising its future operations in order to streamline production focus and advertising strategies. The dataset contains ratings (0-4) of 400 movies from 1097 research participants. It includes various participant-related variables to assist in understanding participant profiles. Ratings are only given for movies that participants have watched. To answer the set of questions, relevant statistical tests were used, dependent on various assumptions; justifications for these tests are provided throughout the report. Additional plots are included for visual representations and to support our conclusions. Throughout the report, we are using $\boldsymbol{\alpha} = 0.005$ in order to cut down on false positives as per the test (Habibzadeh, 2025)

Contents:

1. Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular?
2. Are movies that are newer rated differently than movies that are older?
3. Is enjoyment of ‘Shrek (2001)’ gendered, i.e. do male and female viewers rate it differently?
4. What proportion of movies are rated differently by male and female viewers?
5. Do people who are only children enjoy ‘The Lion King (1994)’ more than people with siblings?
6. What proportion of movies exhibit an “only child effect”, i.e. are rated differently by viewers with siblings vs. those without?
7. Do people who like to watch movies socially enjoy ‘The Wolf of Wall Street (2013)’ more than those who prefer to watch them alone?
8. What proportion of movies exhibit such a “social watching” effect?
9. Is the ratings distribution of ‘Home Alone (1990)’ different from that of ‘Finding Nemo (2003)’?
10. There are ratings on movies from several franchises ([‘Star Wars’, ‘Harry Potter’, ‘The Matrix’, ‘Indiana Jones’, ‘Jurassic Park’, ‘Pirates of the Caribbean’, ‘Toy Story’, ‘Batman’]) in this dataset. How many of these are of inconsistent quality, as experienced by viewers?

Limitations of the data:

The data used in this project is survey-based, observational, and not experimental. Because of that, not all statistical assumptions can fully hold in this context. First, independence of observations is not guaranteed - the same participants rated many movies, and participants may share social, cultural, or demographic similarities that influence the ratings in a correlated manner. Second, the data is not randomized and not sampled from the actual population. It is a voluntary survey sample, meaning that results might not generalize to the full population of movie watchers. Finally, our main dependent variable (movie rating) is ordinal (0-4), which violates normality assumptions and therefore limits how much we can operate on parametric methods. Non-parametric tests such as KS and Mann-Whitney U are more appropriate, but still rely on assumptions that may not hold perfectly. Therefore, this analysis should be interpreted as observational associations rather than causal effects.

Analysis:

1. Following a median split on the number of ratings per movie to classify high and low popularity, we used the standardised average user rank scores and the Mann-Whitney U test to explore the question. We attempted to remove ‘psychological differences’ in rating values by our rank transform to standardise differences in individuals' rating scales. Our approach also assumes popularity can be reasonably captured by the volume of ratings. The Mann-Whitney U test was used as it directly tests for median rank differences. The test yields p-value = 0.000 < 0.005, indicating that high popularity movies tend to have higher average standardised ranks than low popularity movies. However, as the Mann-Whitney U assumes similar shape distributions, interpretation should focus on median shift due to our slight difference in distribution shape seen in *Figure 1*. Independence may not hold due to the nature of rating systems, and while the observed significant difference between high and low popularity movies is robust, the effect size may be influenced by systematic user patterns.
2. The Kolmogorov-Smirnov (KS) test was used to compare the distribution of standardised movie ratings for new and old movies, where the age groups were defined using a median split of release year. We used the same standardisation to eliminate individual rating biases. The KS test was chosen as it is non-parametric, making no assumptions about the underlying distributions of the data. The KS test reported p-value = 0.5416; this large p-value indicates that differences in the rating distributions of old and new movies are not statistically significant. We can see that this conclusion is supported by the histogram *Figure 2* showing the two similar overlapping distributions. However, we acknowledge here that independence between ratings is violated as users rate multiple movies. Furthermore, the median split may not be the most effective categorisation of what is viewed as ‘Old’ and ‘New’ by the raters or general public, as this is open to personal interpretation.
3. To assess whether Shrek (2001) exhibited male and female viewing discrepancies, we first plotted the gender of survey participants, observing that 807 females rated Shrek as opposed to 260 males. We then used two non-parametric tests: the Mann-Whitney U test to compare medians and the Kolmogorov-Smirnov test to compare empirical distributions. We assumed the ratings for each gender were independent and randomly sampled. The Mann-Whitney U test was chosen as the movie ratings are ordinal and may not be normally distributed. The KS test complements this by checking for distribution shape differences. For the Mann-Whitney U test, we observed p-value = 0.0505 and KS test p-value = 0.0561, which are both greater than 0.005; therefore, we have strong evidence to support that there isn't a statistically significant difference in the gender rating of Shrek (2001).

4. The difference in ratings for a particular movie can come from multiple aspects of the distribution (central tendency, spread, or shape), not necessarily just differences in median. For example, two movies might have equal median ratings between male and female viewers, but still have distributions that differ in variability or skewness. Because the question asks for the proportion of movies with any significant difference, we separated female and male viewers' movie ratings and used the Kolmogorov-Smirnov (KS) test for each movie. The KS test evaluates whether two samples come from different underlying distributions, and therefore is appropriate to detect differences beyond just central shifts. All test results (including p-values) for each movie are in *Figure 3*. We find that only 6.25% (25 movies) had p-values lower than 0.005, showing a significant difference in distributions between female and male viewers.

5. First, we plotted the only child status of survey participants, identifying 177 only children and 894 participants with siblings. Plotting the ratings distributions for both groups, we see that visually, children with siblings seem to rate The Lion King higher, although here we acknowledge that the sample size difference is a limiting factor to this conclusion. To assess whether only children enjoy The Lion King more, we did a one-sided Mann-Whitney U test and observed a p-value = 0.978 > 0.005, so we conclude there is no significant evidence to support that only children enjoy The Lion King more than people with siblings. However, one can observe from the visualisations in *Figure 4* that the opposite relation may be true and may want to be explored further. The large difference in our sample size and their distributions is a limiting factor to the utility of the Mann-Whitney U test in this case.

6. Because the question asks for the proportion of movies with any significant difference, we separated movie ratings of viewers with siblings and without, and used the Kolmogorov-Smirnov (KS) test for each movie. The KS test evaluates whether two samples come from different underlying distributions, and therefore is appropriate to detect differences beyond just central shifts. All test results (including p-values) for each movie are in *Figure 5*. We find that only 0.75% - 3 movies had p-values lower than 0.005 rejection region and showed a significant difference in population distributions of viewers' movie ratings without and with siblings.

7. First, we identified our samples of 270 social watchers and 393 alone watchers. We used a one-sided Mann-Whitney U test to test whether social watchers tended to rate the movies higher than solo watchers. The non-parametric test was chosen as ratings are ordinal and non-normal. The Mann-Whitney U test is appropriate in this case to compare median differences without assuming equal variances. A one-sided test can be adopted due to the directionality of the test question given. Our results evidenced a p-value = ~0.9437 > 0.005; therefore, we don't have significant evidence to conclude that The Wolf of Wall Street (2013) exhibits more positive ratings for social watchers. Our histograms in *Figure 6* provide evidence that both groups have left-skewed distributions with high ratings across the board.

8. We separated movie ratings of viewers who prefer to watch movies socially and alone, and then conducted a one-sided Mann-Whitney U test for each movie to identify the proportion of movies that exhibit a "social watching". The Mann-Whitney U test evaluates the median difference between two samples, so it is appropriate for identifying movies that were rated higher by social watchers than by solo watchers, comparing median ratings. All results (including p-values) are presented in *Figure 7*. We find that only 1.5% - 6 movies had p-values lower than 0.005 and showed a "social watching effect" where the median ratings for these movies of viewers who enjoyed watching movies socially were significantly greater than the median rating of viewers who prefer watching movies alone.

9. We performed a Kolmogorov-Smirnov (KS) test on the ratings distributions for Home Alone (1990) and Finding Nemo (2003). We find the KS test to be appropriate because we want to test if our two samples are generated from the same underlying distributions (i.e., if the two samples came from the same population). Our test yields a p-value of 6.38 x 10^{-10}, which is less than our alpha-value of 0.005. Thus, we find that there is a significant difference in rating distributions between Home Alone (1990) and Finding Nemo (2003) that is extremely unlikely to have occurred by random chance (*Figure 8*). In practical terms, this indicates that the two movies come from different distributions.

10. In order to assess the consistency of the quality of movie franchises, we performed a Kruskal-Wallis test on the ratings data for each franchise, where each movie's ratings represented a single population sample. Given the ordinal, non-normal nature of the movie ratings data and the need to compare more than 2 population samples at a time, we decided that the Kruskal-Wallis test is the appropriate test, as it evaluates differences in median between 2+ groups. We found that the Kruskal-Wallis test results for all of the franchises are below 0.005, except for that of the Harry Potter franchise (*Figure 9*). Thus, we conclude that only the Harry Potter franchise is of consistent quality, while all other franchises are of inconsistent quality, as experienced by the viewers.

11. ***Extra Credit:*** We hypothesize that movies with longer runtimes tend to be rated higher than movies with shorter runtimes. Using a third-party IMDb movie api (https://imdbapi.dev/), we were able to find the runtime, in seconds, for 368 movies of the 400 movies, or 92% of the movies. As in questions 1 and 2, we removed "psychological differences" in ratings by performing a rank transform across each user's reviews and standardized the ranks to be between 0 and 1. We then averaged these standardized ranks for each movie. Two sample populations were created by splitting the average standardized ranks for each movie based on the median movie runtime of 6960 seconds (*Figure 10*). A one-sided Mann-Whitney U test was used to test if the median rank of the sample population of longer runtime movies is greater than that of shorter runtime movies. Our test yields a p-value of 1.40 x 10^{-6} and a U statistic of 21594. Thus, we confirm our hypothesis that longer runtimes tend to be rated higher than shorter runtimes.

12. ***Extra Credit:*** We aimed at going beyond p-values and analyzing the exact effect of gender, social watching, and absence of siblings on movie enjoyment. To accomplish this task, we compute effect sizes. We are not able to use Cohen's D and Hedges' G as these assume normality and equal variances of the data. We had to find a non-parametric alternative suitable for ordinal data, like movie ratings.

    To go beyond statistical significance and quantify practical importance, we computed effect sizes using Cliff's Delta (δ). Cliff's delta is a non-parametric effect size measure that quantifies the degree of distributional non-overlap between two groups on ordinal or non-normal data (Meissel & Yao, 2024). It compares all pairwise values between two groups and ranges from -1 to +1, where 0 indicates complete overlap and ±1 indicates no overlap. Positive values indicate a shift toward higher ratings in Group 1, while negative values indicate that Group 2 tends to rate higher. As recommended by the paper, δ magnitudes can be interpreted using conventional thresholds: $|\delta| < 0.15$ = negligible, 0.15-0.33 = small, 0.33-0.47 = medium, and ≥ 0.47 = large. Because Cliff's delta does not assume normality, equal variance, or metric scale measurement, it is more appropriate than traditional standardized mean difference measures (e.g., Cohen's d) for ordinal movie ratings. To avoid p-hacking, we computed Cliff's delta values for movies that were identified by KS/Mann-Whitney U tests in previous questions. We also cleaned results from movie titles whose effect sizes were considered to be negligible ($|\delta| < 0.15$ = negligible).

Gender-based effects (female effect size > male) Figure 11

Medium effect sizes indicate commercially meaningful segmentation potential: these films are substantially more appealing to female viewers, and therefore could be used for female-targeted personalization and catalog curation (romance, emotional drama, coming-of-age). Small effect films still lean female, but are weaker signals - they could be used for softer, non-aggressive recommendation filtering rather than direct audience targeting.

(male effect size > female) Figure 12

These films lean toward action, combat, intensity, and "adrenaline-driven" themes - males consistently prefer them more. Even though the effects are small in magnitude, the direction is stable enough to be useful for recommendation ranking.

Only-child effect Figure 13

The only-child effect resulted in only two movies with medium effect sizes. Both effect sizes reveal that these movies are consistently being rated higher by viewers with siblings than by viewers without siblings. Both movies represent the genre of chaotic comedy, which suggests that shared-sibling household environments may be more familiar with chaotic, competitive, and playful humor - and therefore find these movies more enjoyable. This also means that there does not appear to be a movie that viewers with no siblings prefer more than viewers with siblings. This association is a low priority and much less consistent in the number of movies than the gender effect, and should not drive global personalization of movie recommendations, but can be used as a weak micro-signal for humour sub-genre targeting.

Social watching preference effect Figure 14

To properly analyze both social and alone watching effects, we had to conduct another two-tailed KS test to identify movies that have any significant difference between viewers' rating distribution who enjoy watching movies socially and viewers who enjoy watching movies alone. Hypotheses and results of this test can be seen in *Figure 14*. This action can not be considered as p-hacking, since we are not interested in finding more significant movies, but we are exploring both social and alone watching effects instead of focusing solely on social watching.

Only one movie with a small effect size was found - Donnie Darko (2001). We can treat it as a weak cue. Since the effect size is negative (~-0.28), it indicates that alone watchers tend to rate this movie higher than social watchers. Movies that were rated higher by social watchers than by individual watchers were not identified using effect size. While Donnie Darko shows a small tendency toward higher enjoyment when watched alone, this is not strong enough to build reliable recommendation logic. In practice, this suggests that social watching preference might not be a universal segment driver in movie enjoyment within this sample, unlike gender, which presented clearer and systematic data patterns.

In this project, Cliff's delta quantifies associational separation between group ratings. We use medium-to-large magnitude deltas as cues for segment-aware curation and gentle recommendation ranking. While movie titles with small deltas are considered "weak", they can also be used as optional signals. All statements provided are non-causal and purely observational. Additionally, because the KS test operates on distributional differences and our data is ordinal and discrete, effect size interpretation should always be read in the context of which significance test was used.

Bibliography

Free IMDB API (no date) *Free IMDb API*. Available at: https://imdbapi.dev/ (Accessed: 03 November 2025).

Habibzadeh, F. (2025) *On the effect of flexible adjustment of the P value significance threshold on the reproducibility of randomized clinical trials*, *PloS one*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC12165351/ (Accessed: 28 Oct)

Meissel, K. and Yao, E.S. (2024) *Using Cliff's delta as a non-parametric effect size measure: An accessible web app and R tutorial*, *Practical Assessment, Research, and Evaluation*. Available at: https://openpublishing.library.umass.edu/pare/article/id/1977/ (Accessed: 03 November 2025).

Appendix

Figure 1

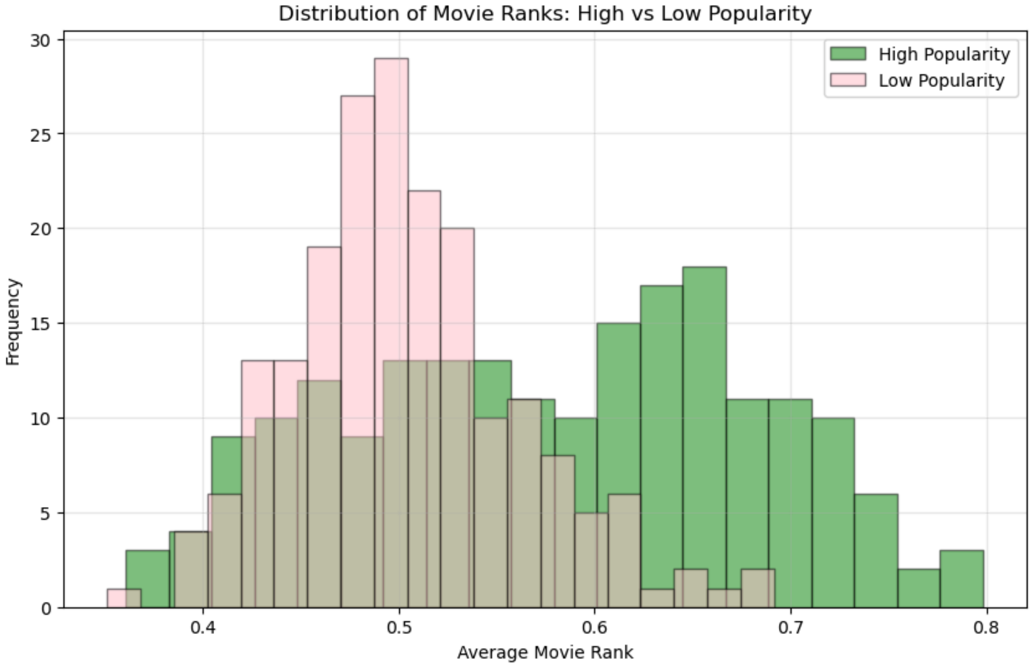


Figure 2

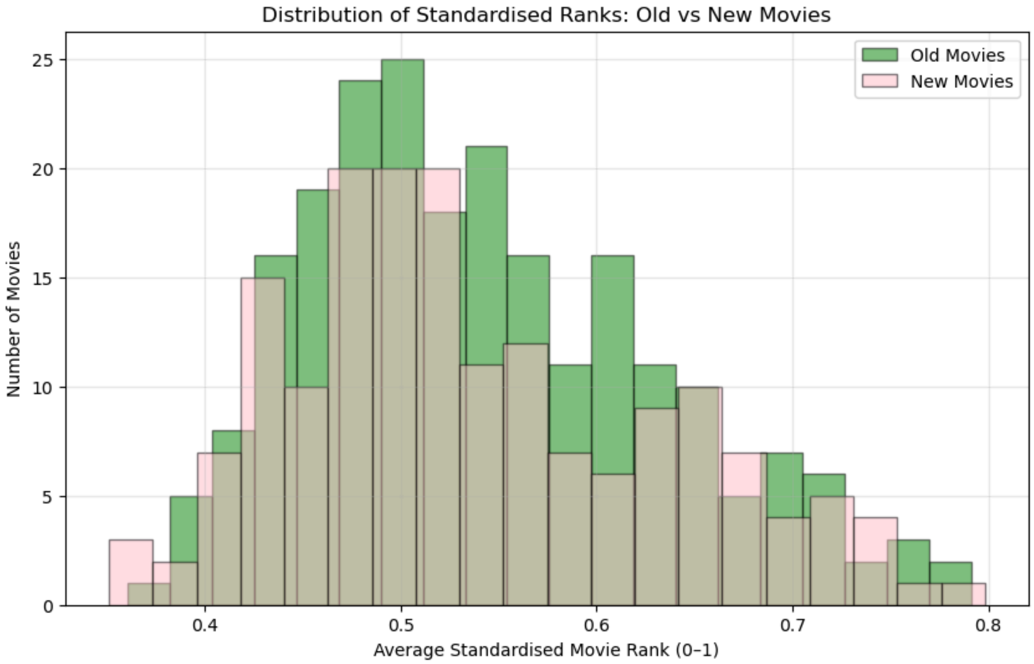


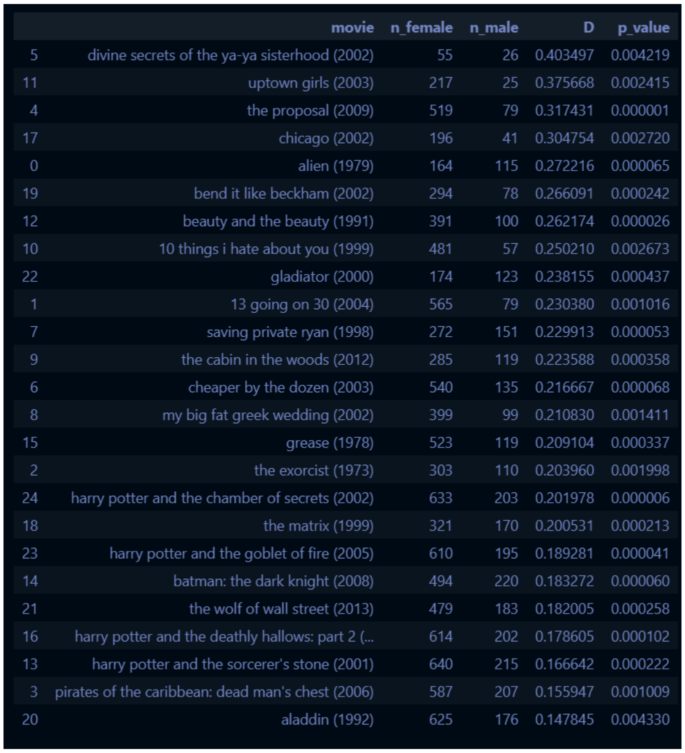
Figure 3

Null hypothesis (H_0): Female and male viewers' movie ratings come from the same population distribution.

Alternative **(***H_a***)**: Female and male viewers' movie ratings come from different population distributions.

Rejection region (significance level): 0.005

If the p-value is smaller than 0.005, it means such a result would be very unlikely by chance under the null hypothesis, so we reject the H0 and conclude that female and male viewers' ratings come from different population distributions. If the p-value is larger than 0.005, then the observed difference could easily occur by random variation, so we fail to reject the H0.



	movie	n_female	n_male	D	p_value
5	divine secrets of the ya-ya sisterhood (2002)	55	26	0.403497	0.004219
11	uptown girls (2003)	217	25	0.375668	0.002415
4	the proposal (2009)	519	79	0.317431	0.000001
17	chicago (2002)	196	41	0.304754	0.002720
0	alien (1979)	164	115	0.272216	0.000065
19	bend it like beckham (2002)	294	78	0.266091	0.000242
12	beauty and the beauty (1991)	391	100	0.262174	0.000026
10	10 things i hate about you (1999)	481	57	0.250210	0.002673
22	gladiator (2000)	174	123	0.238155	0.000437
1	13 going on 30 (2004)	565	79	0.230380	0.001016
7	saving private ryan (1998)	272	151	0.229913	0.000053
9	the cabin in the woods (2012)	285	119	0.223588	0.000358
6	cheaper by the dozen (2003)	540	135	0.216667	0.000068
8	my big fat greek wedding (2002)	399	99	0.210830	0.001411
15	grease (1978)	523	119	0.209104	0.000337
2	the exorcist (1973)	303	110	0.203960	0.001998
24	harry potter and the chamber of secrets (2002)	633	203	0.201978	0.000006
18	the matrix (1999)	321	170	0.200531	0.000213
23	harry potter and the goblet of fire (2005)	610	195	0.189281	0.000041
14	batman: the dark knight (2008)	494	220	0.183272	0.000060
21	the wolf of wall street (2013)	479	183	0.182005	0.000258
16	harry potter and the deathly hallows: part 2 (...	614	202	0.178605	0.000102
13	harry potter and the sorcerer's stone (2001)	640	215	0.166642	0.000222
3	pirates of the caribbean: dead man's chest (2006)	587	207	0.155947	0.001009
20	aladdin (1992)	625	176	0.147845	0.004330

Figure 4

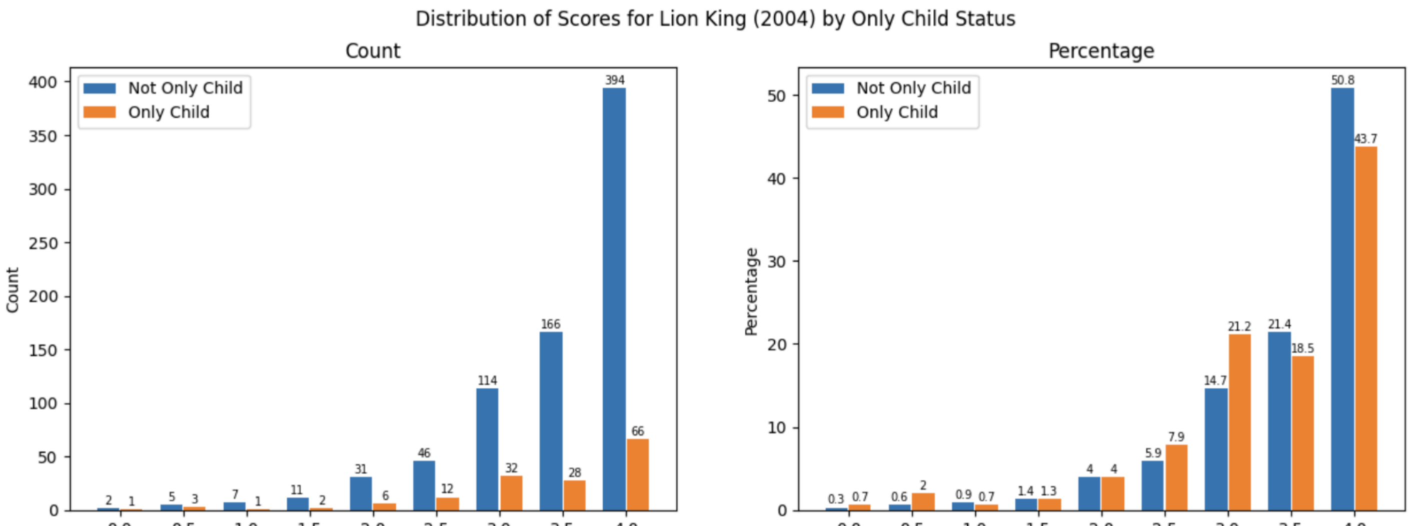


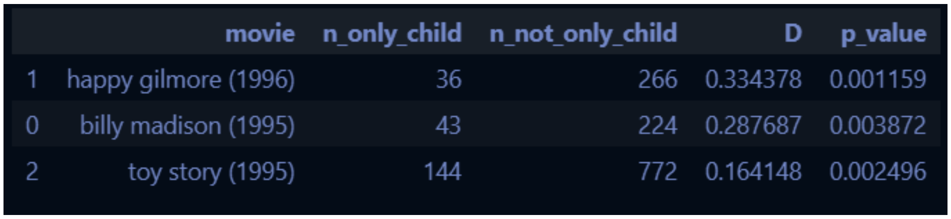
Figure 5

Null hypothesis (H_0): Movie ratings of viewers with no siblings and with siblings come from the same population distribution.

Alternative **(***H_a***)**: Movie ratings of viewers with no siblings and with siblings come from different population distributions

Rejection region (significance level): 0.005

If the p-value is smaller than 0.005, it means such a result would be very unlikely by chance under the null hypothesis, so we reject the H0 and conclude that the population distributions differ. If the p-value is larger than 0.005, then the observed difference could easily occur by random variation, so we fail to reject the H0.



	movie	n_only_child	n_not_only_child	D	p_value
1	happy gilmore (1996)	36	266	0.334378	0.001159
0	billy madison (1995)	43	224	0.287687	0.003872
2	toy story (1995)	144	772	0.164148	0.002496

Figure 6

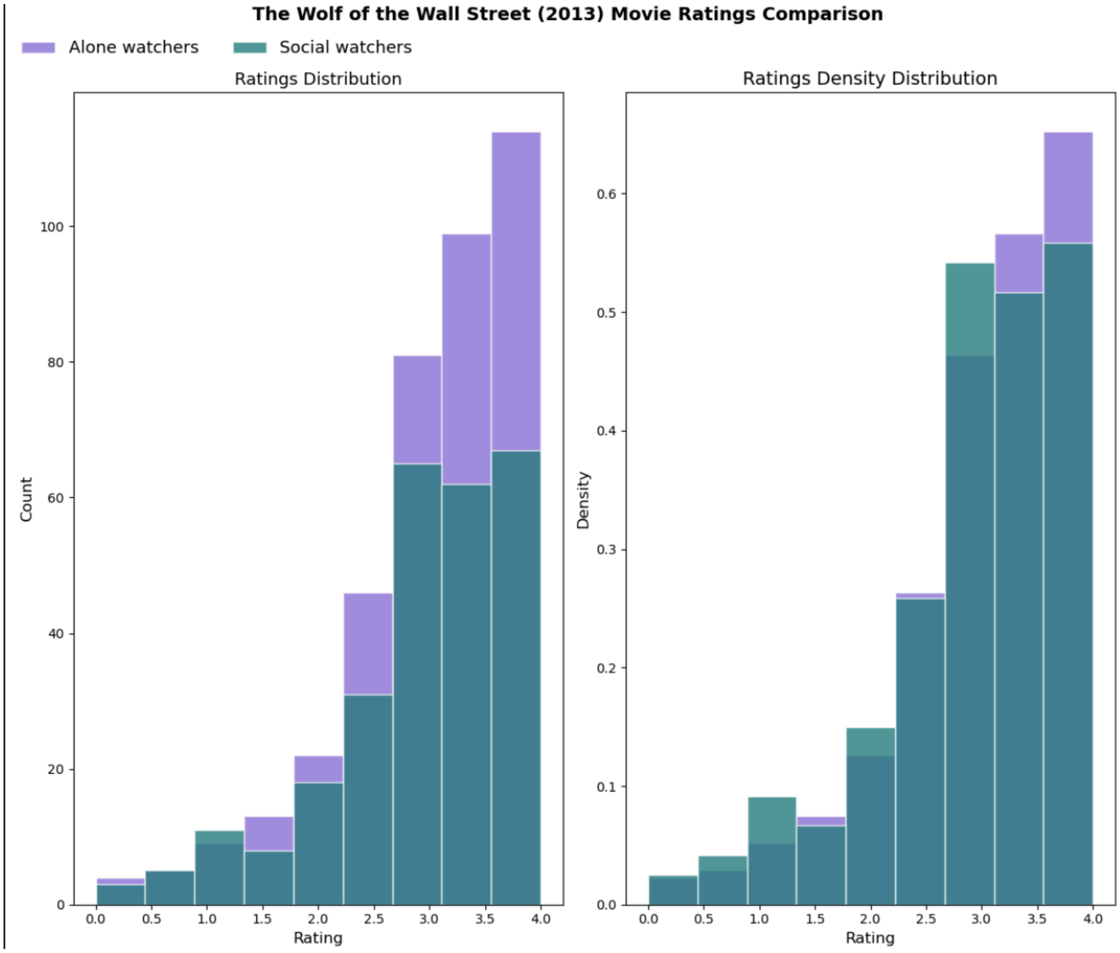


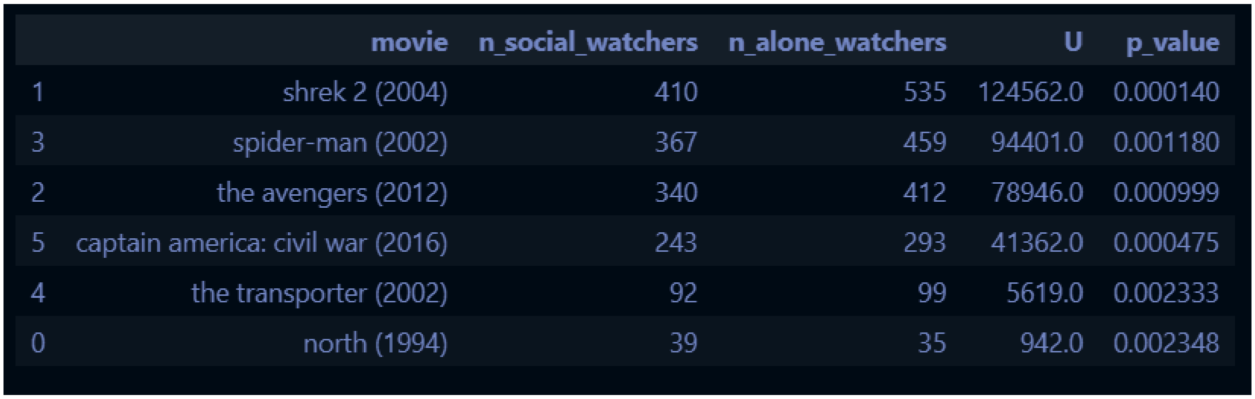
Figure 7

Null hypothesis (H_0): The median rating of the movie is not significantly different between viewers who enjoy watching movies socially and alone.

Alternative hypothesis (Ha): The median rating of viewers who enjoy watching movies socially is significantly greater than the median rating of viewers who enjoy watching movies alone.

Rejection region (significance level): 0.005

If the p-value is smaller than 0.005, it means such a result would be very unlikely under the null hypothesis, so we reject the H0 and conclude that the median movie rating of viewers who enjoy watching movies socially is significantly greater than the movie rating of viewers who enjoy watching movies alone. If the p-value is larger than 0.005, then the observed difference could easily occur by random variation, so we fail to reject the H0.



	movie	n_social_watchers	n_alone_watchers	U	p_value
1	shrek 2 (2004)	410	535	124562.0	0.000140
3	spider-man (2002)	367	459	94401.0	0.001180
2	the avengers (2012)	340	412	78946.0	0.000999
5	captain america: civil war (2016)	243	293	41362.0	0.000475
4	the transporter (2002)	92	99	5619.0	0.002333
0	north (1994)	39	35	942.0	0.002348

Figure 8

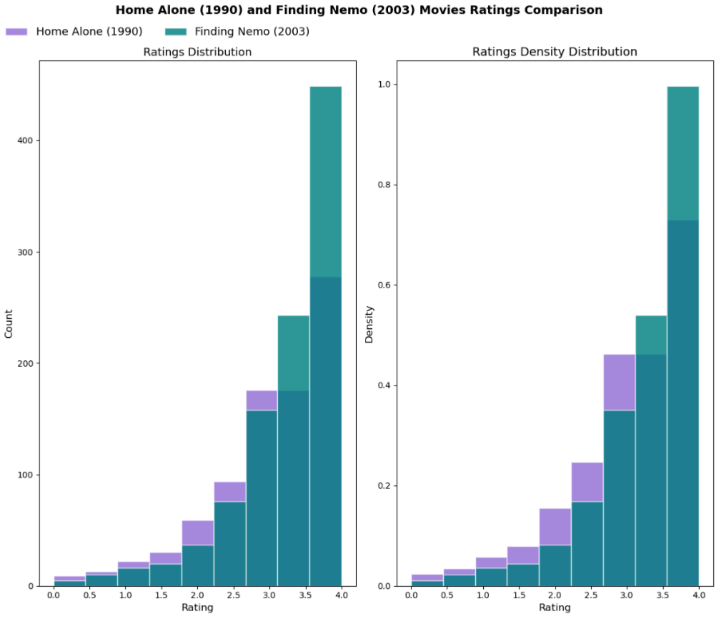
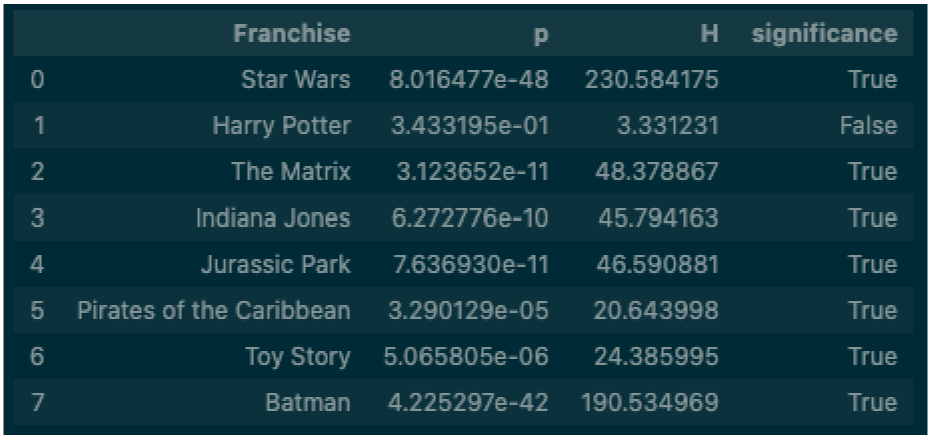


Figure 9



	Franchise	p	H	significance
0	Star Wars	8.016477e-48	230.584175	True
1	Harry Potter	3.433195e-01	3.331231	False
2	The Matrix	3.123652e-11	48.378867	True
3	Indiana Jones	6.272776e-10	45.794163	True
4	Jurassic Park	7.636930e-11	46.590881	True
5	Pirates of the Caribbean	3.290129e-05	20.643998	True
6	Toy Story	5.065805e-06	24.385995	True
7	Batman	4.225297e-42	190.534969	True

Figure 10

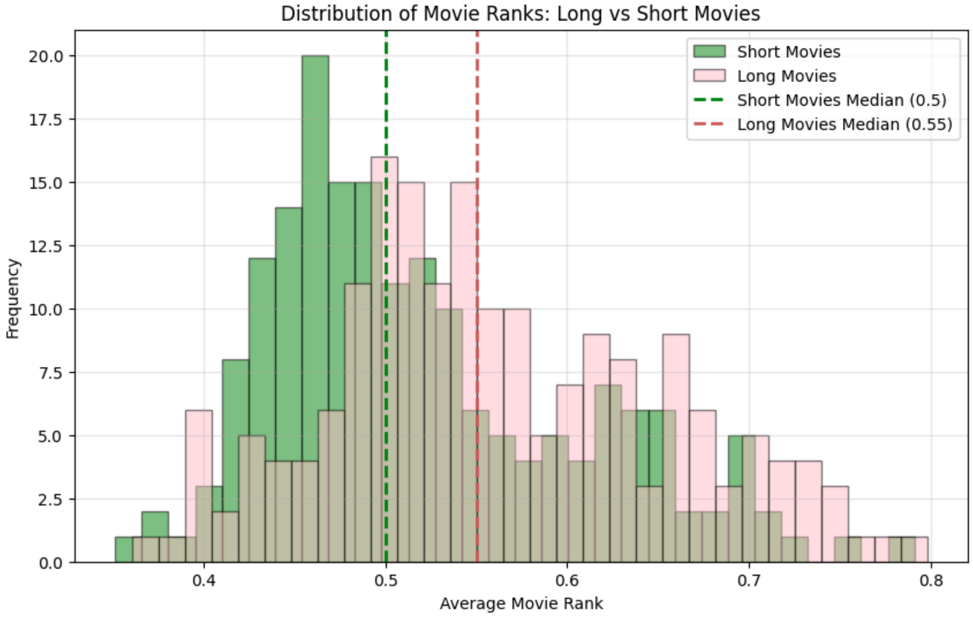
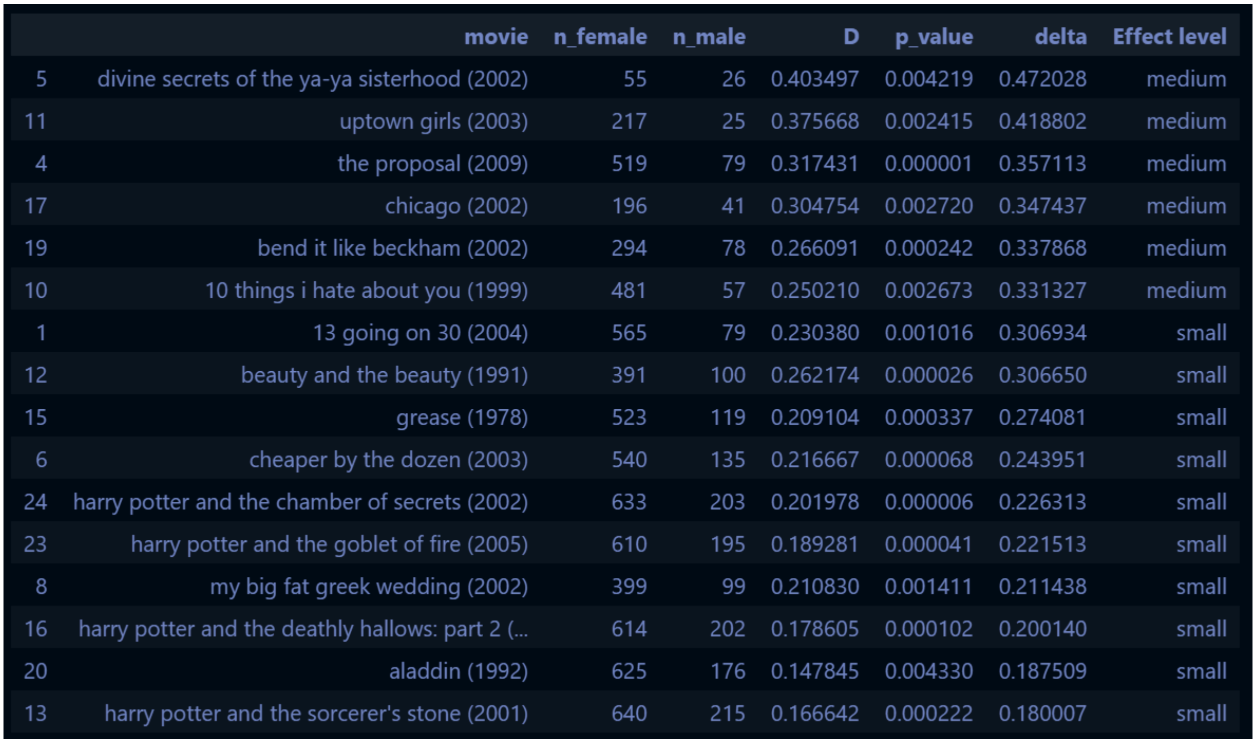
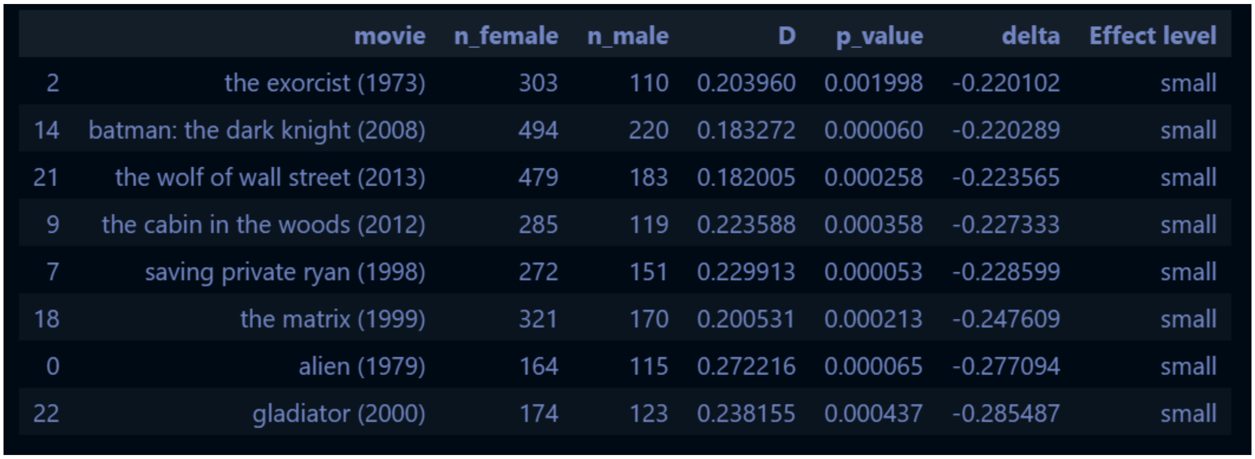


Figure 11



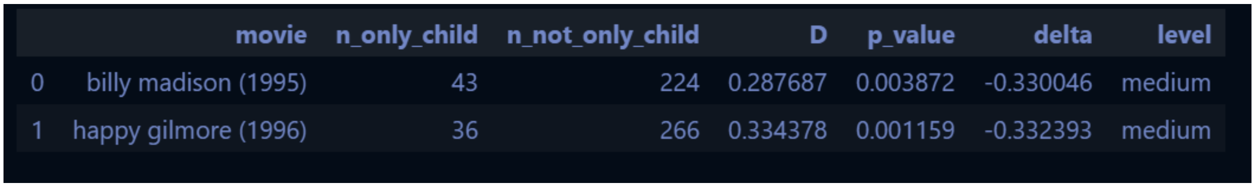
	movie	n_female	n_male	D	p_value	delta	Effect level
5	divine secrets of the ya-ya sisterhood (2002)	55	26	0.403497	0.004219	0.472028	medium
11	uptown girls (2003)	217	25	0.375668	0.002415	0.418802	medium
4	the proposal (2009)	519	79	0.317431	0.000001	0.357113	medium
17	chicago (2002)	196	41	0.304754	0.002720	0.347437	medium
19	bend it like beckham (2002)	294	78	0.266091	0.000242	0.337868	medium
10	10 things i hate about you (1999)	481	57	0.250210	0.002673	0.331327	medium
1	13 going on 30 (2004)	565	79	0.230380	0.001016	0.306934	small
12	beauty and the beauty (1991)	391	100	0.262174	0.000026	0.306650	small
15	grease (1978)	523	119	0.209104	0.000337	0.274081	small
6	cheaper by the dozen (2003)	540	135	0.216667	0.000068	0.243951	small
24	harry potter and the chamber of secrets (2002)	633	203	0.201978	0.000006	0.226313	small
23	harry potter and the goblet of fire (2005)	610	195	0.189281	0.000041	0.221513	small
8	my big fat greek wedding (2002)	399	99	0.210830	0.001411	0.211438	small
16	harry potter and the deathly hallows: part 2 (...	614	202	0.178605	0.000102	0.200140	small
20	aladdin (1992)	625	176	0.147845	0.004330	0.187509	small
13	harry potter and the sorcerer's stone (2001)	640	215	0.166642	0.000222	0.180007	small

Figure 12



	movie	n_female	n_male	D	p_value	delta	Effect level
2	the exorcist (1973)	303	110	0.203960	0.001998	-0.220102	small
14	batman: the dark knight (2008)	494	220	0.183272	0.000060	-0.220289	small
21	the wolf of wall street (2013)	479	183	0.182005	0.000258	-0.223565	small
9	the cabin in the woods (2012)	285	119	0.223588	0.000358	-0.227333	small
7	saving private ryan (1998)	272	151	0.229913	0.000053	-0.228599	small
18	the matrix (1999)	321	170	0.200531	0.000213	-0.247609	small
0	alien (1979)	164	115	0.272216	0.000065	-0.277094	small
22	gladiator (2000)	174	123	0.238155	0.000437	-0.285487	small

Figure 13

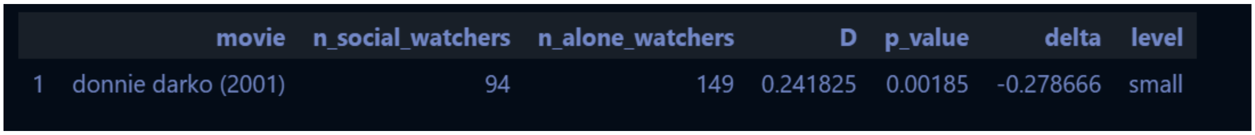


	movie	n_only_child	n_not_only_child	D	p_value	delta	level
0	billy madison (1995)	43	224	0.287687	0.003872	-0.330046	medium
1	happy gilmore (1996)	36	266	0.334378	0.001159	-0.332393	medium

Figure 14

Null hypothesis (H0): Movie ratings of viewers who enjoy social watching and alone watching come from the same population distribution.
Alternative hypothesis (Ha): Movie ratings of viewers who enjoy social watching and alone watching come from different distributions.
*Significance level:*0.005
Interpretation of p-value: Probability of observing the data (or more extreme data) by chance if the null hypothesis were true.

If the p-value is smaller than 0.005, it means such a result would be very unlikely by chance under the null hypothesis, so we reject the H0 and conclude that the population distributions differ. If the p-value is larger than 0.005, then the observed difference could easily occur by random variation, so we fail to reject the H0.



	movie	n_social_watchers	n_alone_watchers	D	p_value	delta	level
1	donnie darko (2001)	94	149	0.241825	0.00185	-0.278666	small