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 optimizing its future operations, thereby streamlining production focus and advertising strategies. The dataset contains ratings (on a scale of 0-4) for 400 movies from 1,097 research participants. It includes various participant-related variables to help understand 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 $\alpha = 0.005$ in order to cut down on false positives as per the test (Habibzadeh, 2025)

Contents:

Limitations of the data:
The data used in this project is survey-based, observational, and not experimental. Therefore, not all statistical assumptions can be fully upheld 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. Finally, our main dependent variable (movie rating) is ordinal (0-4), which violates the normality assumptions and therefore limits the methods we can apply. Non-parametric tests, such as the KS and Mann-Whitney U tests, are more appropriate but still rely on assumptions that may not hold perfectly. Therefore, this analysis should be interpreted as an observational association rather than a causal effect.

Analysis:

1. **F**ollowing 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 a p-value of $0.000 < 0.005$, indicating that high-popularity movies tend to have higher average standardized ranks than low-popularity movies. However, as the Mann-Whitney U test assumes similar shape distributions, the interpretation should focus on the median shift due to the slight difference in distribution shape seen in *Figure 1*. Independence may not be maintained 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. **T**he 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 a p-value of 0.5416; this large p-value indicates that the 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 *in Figure 2*, which shows 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. **T**o assess whether Shrek (2001) exhibited male and female viewing discrepancies, we first plotted the gender of survey participants, observing that 807 females rated Shrek, compared 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 differences in distribution shape. For the Mann-Whitney U test, we observed a p-value of 0.0505 and a KS test p-value of 0.0561, both of which are greater than 0.005; therefore, we have strong evidence to support the conclusion that there is no statistically significant difference in the gender rating of Shrek (2001).
4. **T**he 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 among male and female viewers, but still exhibit distributions that differ in terms of variability or skewness. Because the question asks for the proportion of movies with any significant difference, we separated the movie ratings of female and male viewers and used the Kolmogorov-Smirnov (KS) test for each movie. The KS test evaluates whether two samples come from different underlying distributions; therefore, it is appropriate for detecting 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. **F**irst, we plotted the status of being an only child among survey participants, identifying 177 only children and 894 participants with siblings. Plotting the ratings distributions for both groups, we observe that visually, children with siblings appear to rate The Lion King higher; however, 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 visualizations in Figure 4 that the opposite relation may be true and warrants further exploration. 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. **B**ecause 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; therefore, it is appropriate for detecting 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, which falls within the rejection region, and showed a significant difference in the population distributions of viewers' movie ratings without and with siblings.
7. **F**irst, we identified our samples, consisting of 270 social watchers and 393 solo watchers. We used a one-sided Mann-Whitney U test to test whether social watchers tended to rate the movie 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 showed a p-value of ~0.9437, which is greater than 0.005; therefore, we don't have significant evidence to conclude that The Wolf of Wall Street (2013) receives more positive ratings from social watchers. Our histograms in *Figure 6* provide evidence that both groups have left-skewed distributions with high ratings across the board.
8. **W**e 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, making it appropriate for identifying movies that were rated higher by social watchers than by solo watchers by 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 among viewers who enjoyed watching movies socially were significantly greater than the median rating of viewers who prefer watching movies alone.

9. **W**e 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 × 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. **T**o assess the consistency of movie franchise quality, 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 two population samples simultaneously, we decided that the Kruskal-Wallis test is the appropriate test, as it evaluates differences in medians between two or more groups. We found that the Kruskal-Wallis test results for all franchises are below 0.005, except for the Harry Potter franchise (*Figure 9*). Thus, we conclude that only the Harry Potter franchise consistently maintains a high quality, while all other franchises exhibit inconsistent quality, as experienced by viewers.
11. ***Extra Credit:*** We aimed to go beyond p-values and analyze the exact effect of gender, social watching, and absence of siblings on movie enjoyment. To accomplish this task, we compute effect sizes. We are unable to use Cohen's d and Hedges' g, as these assumptions require normality and equal variances of the data. We needed to find a non-parametric alternative suitable for ordinal data, such as 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. From a business perspective, we computed Cliff's delta only for movies identified by Mann-Whitney U tests in previous questions, since these tests isolate differences in typical enjoyment between groups, making the resulting effect sizes directly interpretable and actionable for decision-making. We also excluded results from movie titles whose effect sizes were considered negligible ($|\delta| < 0.15$).

    Gender-based effects (female effect size $\delta > 0$) Figure 10

    The results identify a subset of films with medium Cliff's delta effect sizes, indicating statistically and practically meaningful differences in enjoyment between female and male viewers. These titles receive higher ratings from female viewers, suggesting strong gender-linked preference patterns. The movies with medium effects are primarily relationship-driven, emotional, or coming-of-age narratives, consistent with established genre preference trends.

Films with small but statistically significant effects also receive slightly higher ratings from female viewers, though the magnitude of these differences is modest. These titles provide weaker preference signals and are better suited for soft recommendation weighting rather than explicit gender-based targeting. Overall, the findings support cautious and selective use of gender information, prioritizing medium-effect titles for actionable personalization while treating small-effect titles as secondary signals.

(male effect $\delta < 0$) Figure 11

All movies shown exhibit negative Cliff's delta values, indicating higher ratings from male viewers. The effects are uniformly small, meaning the differences in enjoyment are statistically significant but modest. These titles are primarily action, war, crime, and science-fiction films, suggesting genre-aligned preference patterns rather than strong gender segmentation. From a business perspective, these results represent weak preference signals and are best suited for soft recommendation weighting or tie-breaking, rather than for direct gender-based targeting.

Only-child effect Figure 12

All movies shown have negative Cliff's delta values, indicating that these titles receive higher ratings from viewers with siblings than from only children. Most effects are small, as seen in Captain America: Civil War (2016), American Pie (1999), and Star Wars: Episode VI, suggesting modest differences in typical enjoyment. Three movies show medium effect sizes: Billy Madison (1995), Happy Gilmore (1996), and The Blue Lagoon (1980), indicating more pronounced rating differences between the two groups. One movie, Feardotcom (2002), exhibits a large effect, reflecting a substantial difference in ratings; however, this result should be interpreted cautiously due to the very small number of only-child ratings for that title. Overall, the only-child effect is movie-specific rather than general, with practical relevance concentrated in a small number of medium and large-effect cases.

Social watching preference effect Figure 13

The table shows seven movies with statistically significant differences in ratings between viewers who prefer watching movies socially and those who prefer watching alone. One movie, North (1994), exhibits a medium positive Cliff's delta, indicating meaningfully higher ratings among social watchers. Two additional movies, The Transporter (2002) and Captain America: Civil War (2016), show small positive effects, suggesting modestly higher enjoyment among social watchers.

The remaining four movies exhibit small negative effects, indicating higher ratings among viewers who prefer to watch alone. These include Inglorious Bastards (2009), The Silence of the Lambs (1991), Apocalypse Now (1979), and Donnie Darko (2001). All negative effects are small in magnitude, indicating limited practical differences despite statistical significance. Overall, the social watching effect is movie-specific and asymmetric, with practical relevance concentrated in the single medium-effect case and weaker signals elsewhere.

Bibliography

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

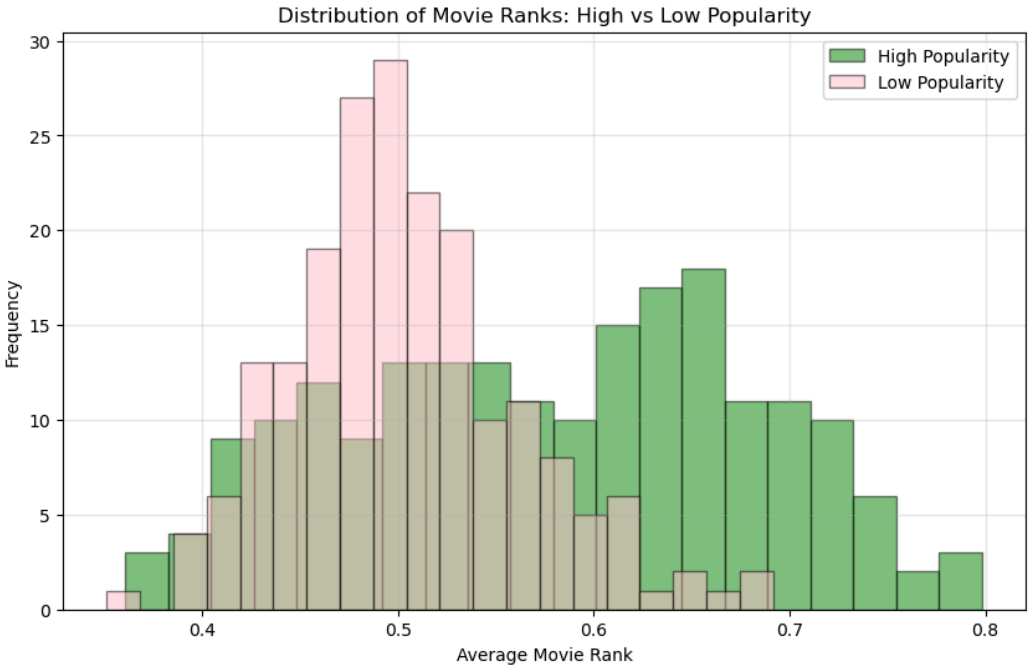

Figure 2

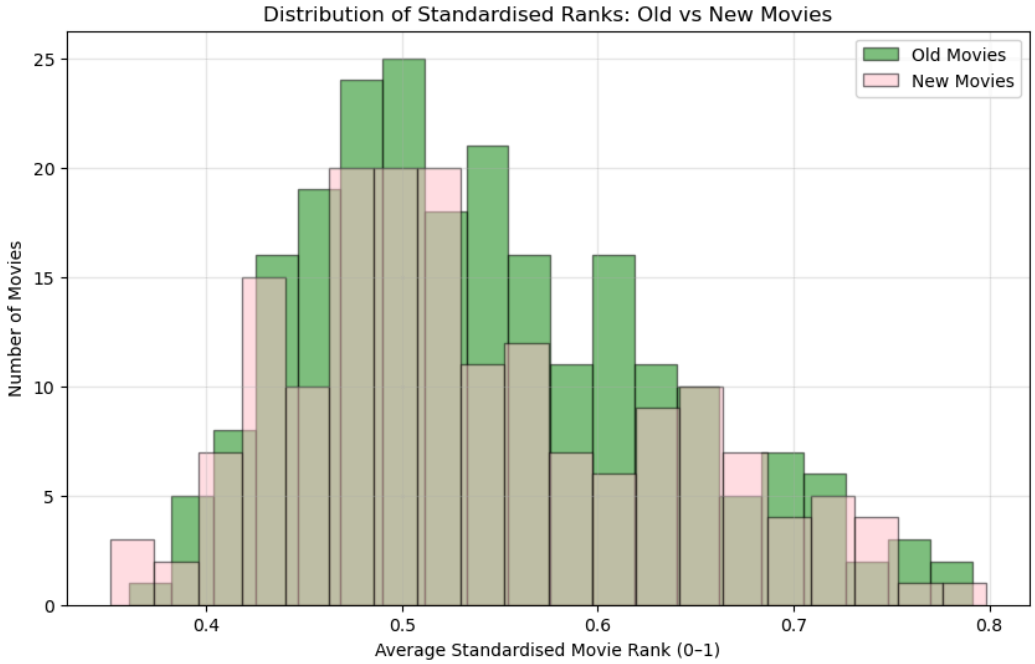

Figure 3

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

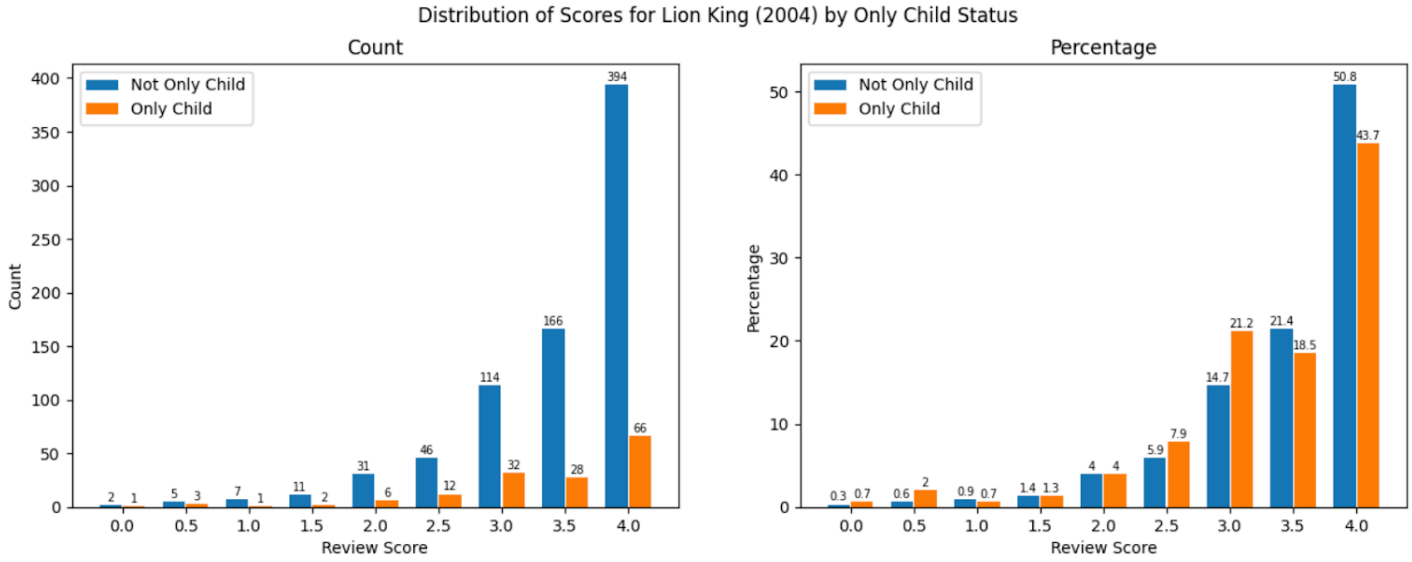

Figure 5

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

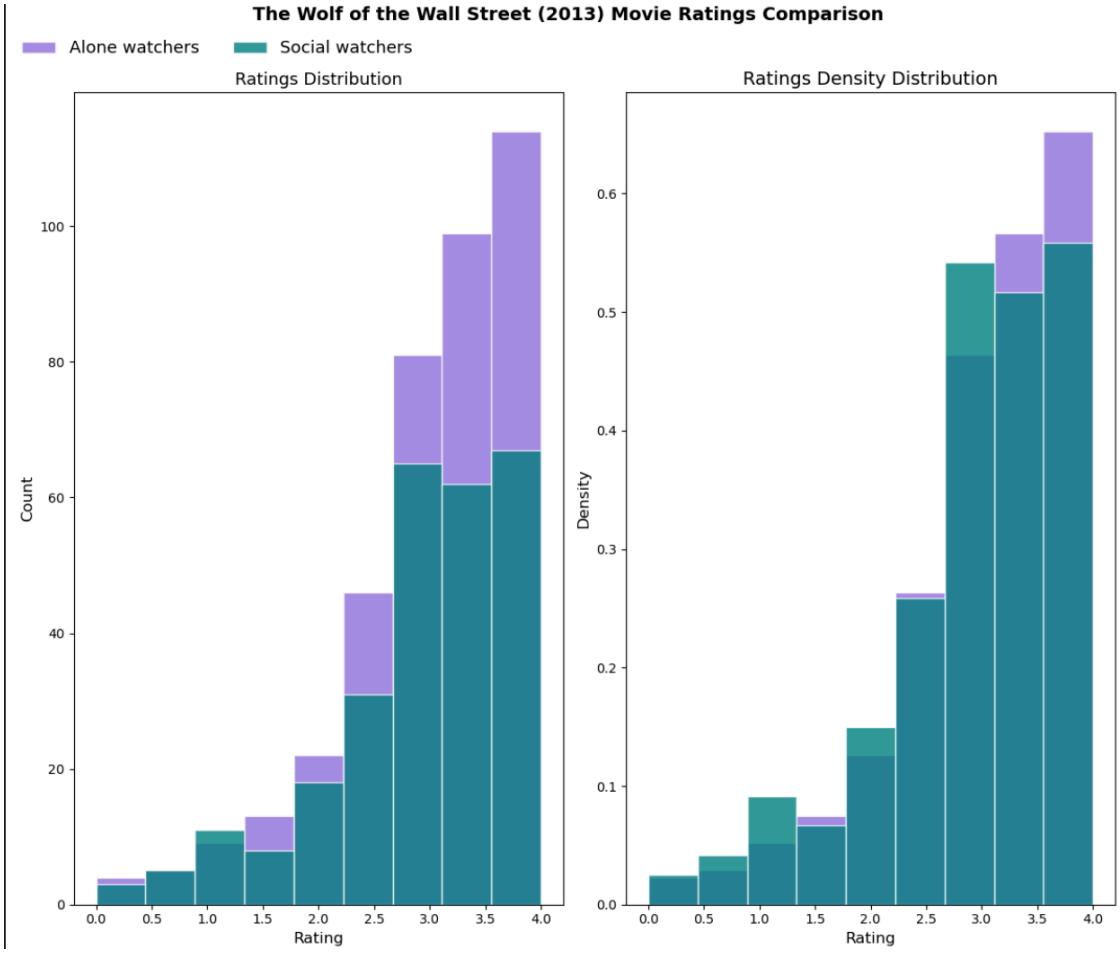

Figure 7

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

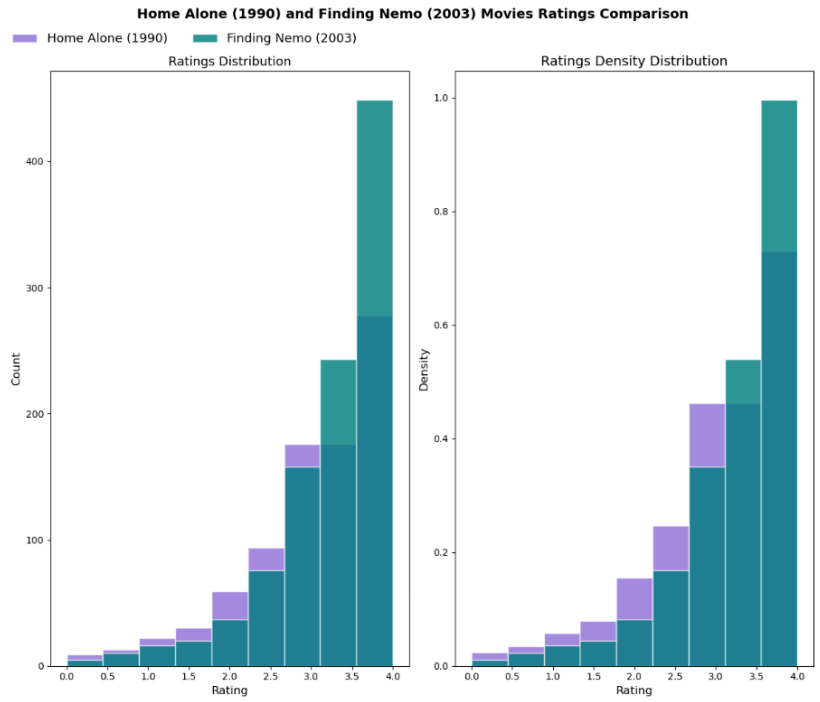

Figure 9

	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

	movie	n_female	n_male	U	p_value	delta	Effect level
17	divine secrets of the ya-ya sisterhood (2002)	55	26	1052.5	5.835228e-04	0.472028	medium
14	andaz apna apna (1994)	50	23	844.0	1.264299e-03	0.467826	medium
26	the firm (1993)	48	26	909.0	1.132727e-03	0.456731	medium
29	uptown girls (2003)	217	25	3848.5	4.978236e-04	0.418802	medium
20	can't hardly wait (1998)	69	26	1258.5	2.335726e-03	0.403010	medium
11	funny girl (1968)	106	29	2091.5	2.622892e-03	0.360768	medium
16	girl interrupted (1999)	195	28	3710.5	1.648810e-03	0.359158	medium
15	the proposal (2009)	519	79	27821.5	1.602214e-07	0.357113	medium
38	chicago (2002)	196	41	5414.0	3.694106e-04	0.347437	medium
41	bend it like beckham (2002)	294	78	15340.0	3.207870e-06	0.337868	medium
28	10 things i hate about you (1999)	481	57	18250.5	2.171767e-05	0.331327	medium
32	the holiday (2006)	229	36	5417.0	2.102445e-03	0.314168	small
6	13 going on 30 (2004)	565	79	29167.5	6.916496e-06	0.306934	small
30	beauty and the beauty (1991)	391	100	25545.0	8.920452e-07	0.306650	small
18	ghostbusters (2016)	289	99	18333.0	2.364462e-05	0.281535	small
31	knight and day (2010)	140	54	4838.0	2.200775e-03	0.279894	small
35	grease (1978)	523	119	39647.5	1.986581e-06	0.274081	small
19	cheaper by the dozen (2003)	540	135	45342.0	8.660131e-06	0.243951	small
22	room (2015)	239	56	8245.5	4.408124e-03	0.232143	small
8	clueless (1995)	558	70	24036.0	1.254298e-03	0.230722	small
49	harry potter and the chamber of secrets (2002)	633	203	78790.0	3.930701e-07	0.226313	small
47	harry potter and the goblet of fire (2005)	610	195	72649.5	1.123946e-06	0.221513	small
24	my big fat greek wedding (2002)	399	99	23926.5	8.908822e-04	0.211438	small
37	harry potter and the deathly hallows: part 2 (...	614	202	74425.5	7.366151e-06	0.200140	small
42	aladdin (1992)	625	176	65313.0	8.361613e-05	0.187509	small
33	harry potter and the sorcerer's stone (2001)	640	215	81184.5	2.899731e-05	0.180007	small
21	transformers: age of extinction (2014)	335	161	31600.5	1.772995e-03	0.171799	small
44	iron man 3 (2013)	391	184	41353.5	3.172237e-03	0.149602	small

Figure 11

	movie	n_female	n_male	U	p_value	delta	Effect level
5	the lost world: jurassic park (1997)	392	157	26028.5	0.004066	-0.154150	small
45	inception (2010)	513	197	42413.0	0.000561	-0.160646	small
3	star wars: episode iv - a new hope (1977)	343	185	26577.0	0.001666	-0.162336	small
7	inglorious bastards (2009)	266	155	16914.0	0.001623	-0.179529	small
48	jaws (1975)	353	138	19755.5	0.000973	-0.188919	small
27	die hard (1988)	169	119	8025.0	0.003061	-0.201929	small
36	8 mile (2002)	205	119	9723.0	0.001897	-0.202869	small
4	indiana jones and the raiders of the lost ark ...	194	135	10432.5	0.001388	-0.203322	small
0	django unchained (2012)	269	167	17761.0	0.000155	-0.209269	small
9	the exorcist (1973)	303	110	12997.0	0.000554	-0.220102	small
2	fargo (1996)	149	94	5461.5	0.003309	-0.220120	small
34	batman: the dark knight (2008)	494	220	42369.5	0.000001	-0.220289	small
43	the wolf of wall street (2013)	479	183	34030.0	0.000005	-0.223565	small
25	the cabin in the woods (2012)	285	119	13102.5	0.000278	-0.227333	small
23	saving private ryan (1998)	272	151	15841.5	0.000062	-0.228599	small
40	the matrix (1999)	321	170	20529.0	0.000004	-0.247609	small
1	alien (1979)	164	115	6817.0	0.000065	-0.277094	small
46	gladiator (2000)	174	123	7646.0	0.000019	-0.285487	small
39	apocalypse now (1979)	95	72	2363.0	0.000552	-0.309064	small
12	the thing (1982)	84	70	1983.0	0.000465	-0.325510	small

Figure 12

	movie	n_only_child	n_not_only_child	U	p_value	delta	level
6	captain america: civil war (2016)	104	431	18403.5	0.003736	-0.178855	small
3	american pie (1999)	75	367	10903.0	0.004168	-0.207775	small
4	star wars: episode vi - the return of the jedi...	72	378	10705.0	0.003346	-0.213330	small
0	billy madison (1995)	43	224	3226.5	0.000538	-0.330046	medium
2	happy gilmore (1996)	36	266	3196.5	0.001075	-0.332393	medium
1	the blue lagoon (1980)	28	135	1200.5	0.002135	-0.364815	medium
5	feardotcom (2002)	12	44	123.0	0.004604	-0.534091	large

Figure 13

Null hypothesis (H0): The median rating of the movie is equal between the two groups.
Alternative hypothesis (Ha): The median rating of the movie differs between the two groups.
*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, we reject H0 and conclude that the movie exhibits a statistically significant difference in typical enjoyment between the two groups.
If the p-value is larger than 0.005, we fail to reject H0 and conclude that the observed difference is consistent with random variation.

	movie	n_social_watchers	n_alone_watchers	U	p_value	delta	level
1	north (1994)	39	35	942.0	0.004696	0.380220	medium
8	the transporter (2002)	92	99	5619.0	0.004666	0.233860	small
9	captain america: civil war (2016)	243	293	41362.0	0.000950	0.161870	small
2	inglorious bastards (2009)	168	257	18197.0	0.004998	-0.157078	small
0	the silence of the lambs (1991)	203	291	24344.5	0.000554	-0.175783	small
7	apocalypse now (1979)	69	98	2498.5	0.003731	-0.261017	small
6	donnie darko (2001)	94	149	5051.5	0.000198	-0.278666	small