Q1. Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular?

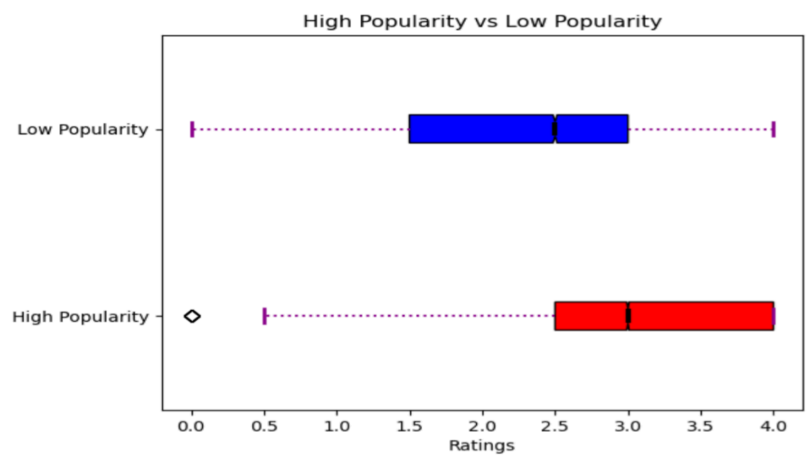
D: Categorized all movies as high popularity or low popularity with respect to the median of each column. Ranked the data and added the ranks to find R1 and R2. Used Mann-Whitney Test to find u statistic and p value.

Y: The data does not follow any typical distribution. The question asks us to compare two independent groups, so we used the Mann-Whitney test. Assumed Null Hypothesis as “Movies that are more popular are not rated higher than movies that are less popular”

F: The medians of both high and low popularity movies are 2.5 and 3 respectively. The u statistic is 800742093. The p value obtained is 0.

A: As the p value is lower than 0.005, it is considered that there is a significant difference in ratings of both groups. So, we reject the null hypothesis. Movies that are more popular are rated higher than movies that are less popular.

```
Number of movies - high popularity: 211
Number of movies - low popularity: 189
Number of Ratings for high popularity movies: 83895
Number of Ratings for low popularity movies: 28319
rank of high:  5094307872.0
rank of low:  1201739133.0
Mannn-Whitney 800742093.0
Mann-Whitney U Statistic: 1575080412.0
0.0000000000 p-value
There is a statistically significant difference in ratings between low and high popularity movies
Median of Highly Popular Movies: 3.0
Median of Lesser Popular Movies: 2.5
```



Q2. Are movies that are newer rated differently than movies that are older?

D: Categorized all movies as old movies or new movies with respect to the median. Used Mann-Whitney Test to find u statistic and p value, used cliff’s delta to learn about the effect size.

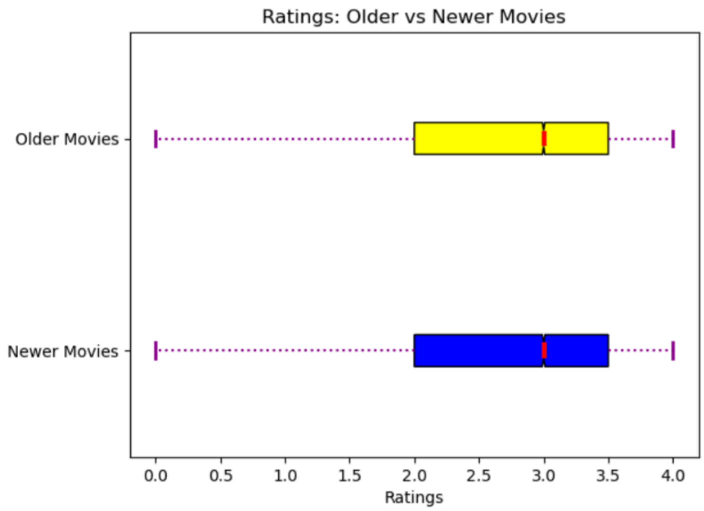
Y: The data does not follow any typical distribution. The question asks us to compare two independent groups, so we used the Mann-Whitney test. Assumed Null Hypothesis as “Movies that are older are not rated differently from newer movies”

F: The median ratings of both older and newer movies are 3 and 3 respectively. The u statistic is 1553577699. The p value obtained is 0.0031.

A: As the p value is lower than 0.005, it is considered that there is a significant difference in ratings of both groups. So we reject the null hypothesis. Since the medians of both groups are 3, cliff’s delta was used to find the effect size. Cliff’s delta gave us a value of 0.016 which said there is a negligible gap between the two groups. Even though statistically there is a difference in how older and newer movies are rated, practically there is not much difference.

```
Median(new movies): 3.0
Median(old movies)): 3.0
0.01668558320588261 negligible
There is slight overlap

MannwhitneyuResult(statistic=1553577699.0, pvalue=1.2849216001533934e-06)
```



Q3. Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?

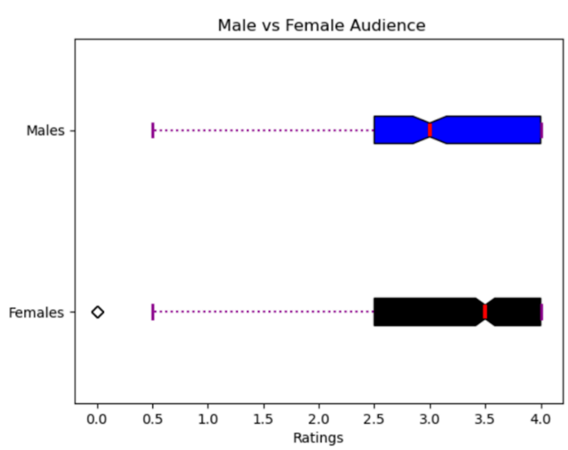
D: Separated based on genders, make two lists - female and male. Used Kolmogorov-Smirnov (KS) Test to find p value.

Y: The data does not follow any typical distribution. Since we can't base this solely off medians, we use a KS test which will compare the CDFs of both groups. Assumed Null Hypothesis as "male and female viewers don't rate it differently".

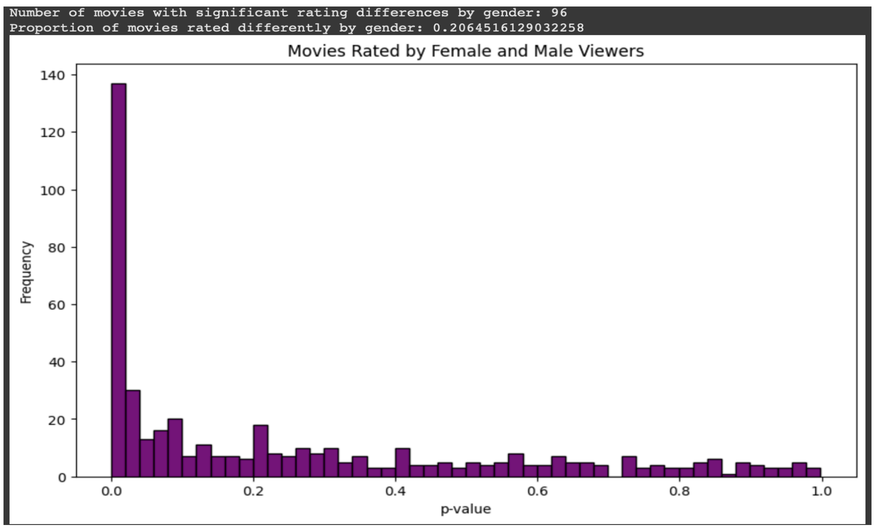
F: The statistic is 0.09796552051512596 and p value is 0.05608204072286342. The median for the female and male list is 3.5 and 3 respectively.

A: As the p value is higher than 0.005, it is considered that there isn't a significant difference in ratings of both groups. So, we accept the null hypothesis. male and female viewers don't rate the movie 'Shrek (2001)' differently.

```
KstestResult(statistic=0.09796552051512596, pvalue=0.05608204072286342, statistic_location=3.0, statistic_sign=-1)
```

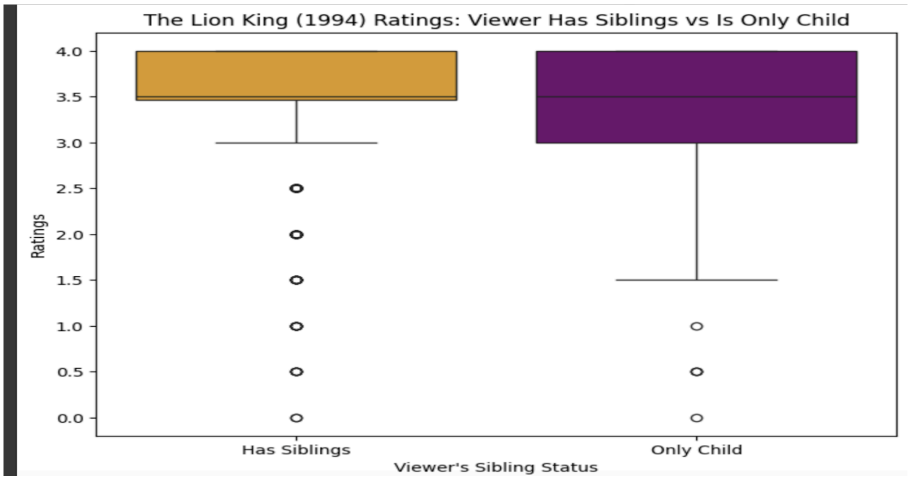


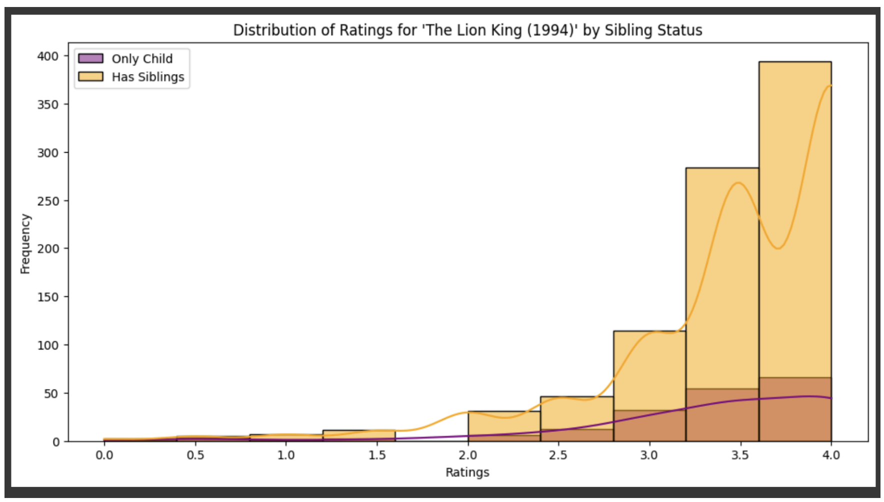
Q4. What proportion of movies are rated differently by male and female viewers?
D: Categorizing the movies rated by female and male viewers into two different groups. Have assumed that the ratings are independent, and the data is not normally distributed.
Y: Using the Mann-Whitney U test as the data is not normally distributed and we have to compare between two independent groups.
F: The proportion of movies rated differently by male and female viewers is obtained through a p value which is 0.2064516129032258. Plotted a histogram to show significance level across all tested movies. Also, there are 96 movies that have been rated differently by male and female viewers.
A: As the analysis found that 96 movies have been rated differently based on gender but since our p value is greater than alpha, we can conclude that it is not statistically significant, and we say that gender does not influence the movie ratings. (A different level of significance can change our results).



Q5. Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?
D: Categorizing the ratings of the movie 'The Lion King (1994)' into two different groups. Have assumed that the ratings are independent, and the data is not normally distributed.
Y: Using the Mann-Whitney U test as we have ordinal data and not a normal distribution. It is suitable for such datasets where the data does not meet the requirements of performing parametric tests.
F: Obtained a U-statistic: 71171.0 and a p value that shows the proportion of the people who enjoy movie whether they are only child or with siblings is 0.02667069073543531 showing no significant difference
A: We can conclude that no significant difference in enjoyment between only children and those with siblings (p >= 0.005).

```
Mann-Whitney U Test for 'The Lion King (1994)' Ratings by Only Children vs. Those with Siblings
U-statistic: 71171.0
p-value: 0.02667069073543531
```





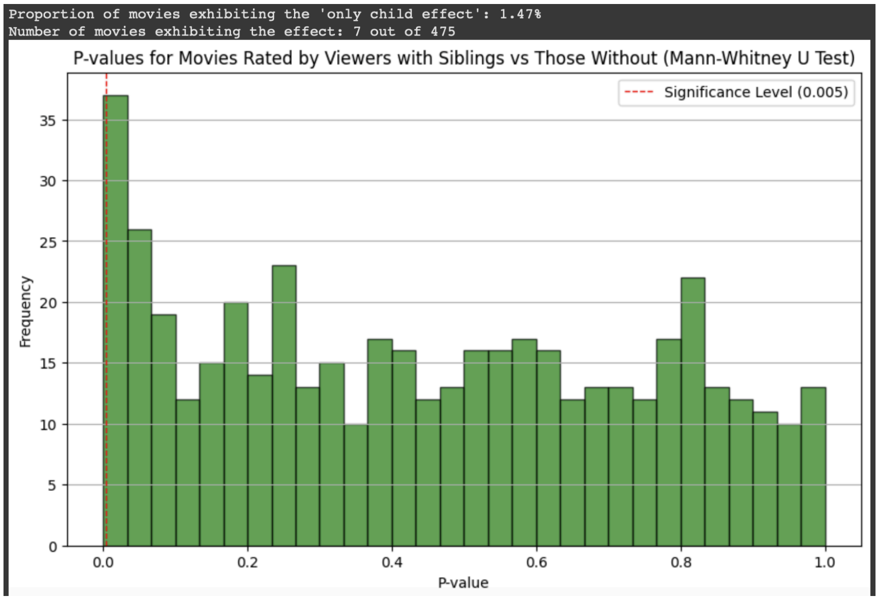
Q6. What proportion of movies exhibit an "only child effect", i.e. are rated different by viewers with siblings vs. those without?

D: Categorizing the data to analyze the only child effect on movie ratings between two independent groups 'with siblings' and 'without'

Y: Using the Mann-Whitney U test as we have ordinal data and not a normal distribution.

F: We got Proportion of movies exhibiting the 'only child effect': 1.47% and Number of movies exhibiting the effect: 7 out of 475

A: We can conclude that 7 movies have shown a difference in ratings between the two groups 'with siblings' and 'without' meaning that this is statistically significant for 7 movies whose p value <= 0.005.



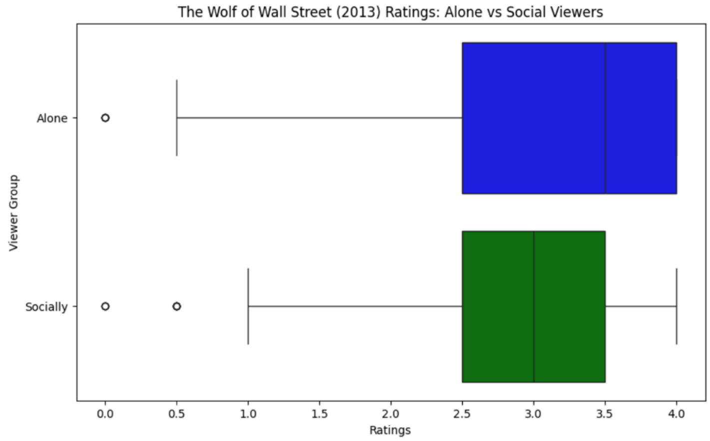
Q7. Do people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone?

D: We conducted a Mann-Whitney U test on ratings of "The Wolf of Wall Street (2013)" to compare groups who prefer watching movies alone versus socially, assuming independence between the two groups, ordinal data for ratings, and non-normality in the distribution.

Y: We chose the Mann-Whitney U test because it is suitable for comparing two independent groups with ordinal or non-normally distributed data, which best matches our dataset. The two-sided test was chosen to identify any difference in ratings without assuming a directional preference.

F: The Mann-Whitney U statistic was 56806.5, with a p-value of 0.1128, which indicates no statistically significant difference between the groups. Additionally, the Cliff's Delta effect size was 0.0707, indicating a small and non-meaningful difference in ratings between those who watch alone and those who watch socially.

A: Given the p-value of 0.1128, we conclude there is no statistically significant difference in ratings for "The Wolf of Wall Street (2013)" between individuals who prefer watching alone and those who prefer watching socially. This analysis assumes independent samples and that ratings reflect ordinal, non-normal data. Limitations include potential unaccounted dependencies within the sample or underlying differences in distribution.



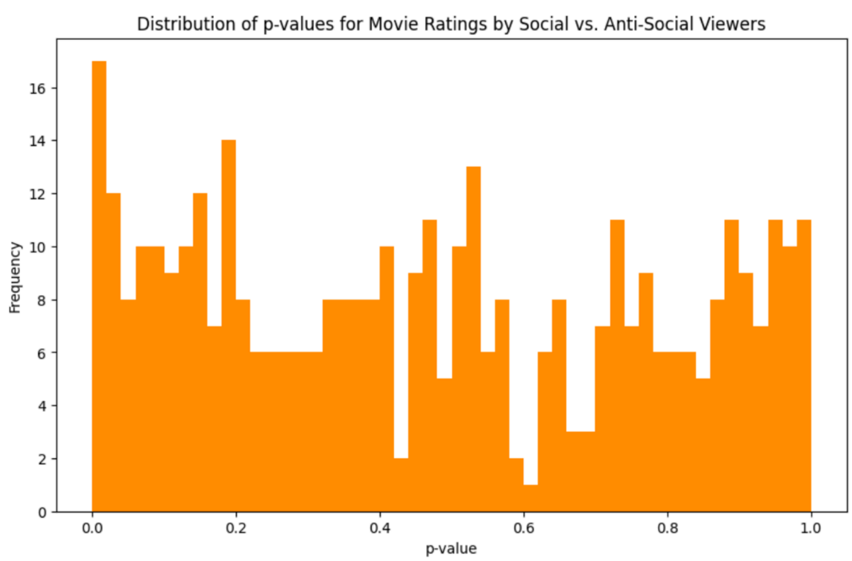
Q8. What proportion of movies exhibit such a "social watching" effect?

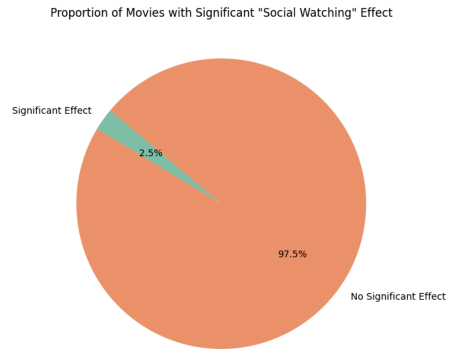
D: We conducted a series of Mann-Whitney U tests to compare ratings for each movie between groups of viewers who prefer watching movies alone versus socially, assuming independence of samples, ordinal rating data, and non-normality in the distribution of ratings.

Y: We used the Mann-Whitney U test for each movie because it is a non-parametric test suitable for comparing two independent groups with ordinal or non-normally distributed data, which fits the characteristics of the dataset. A two-sided test was used to detect any difference in ratings without assuming a specific direction.

F: We found that out of all the movies analyzed, 10 had statistically significant differences in ratings between the two groups ($p \leq 0.005$), representing 2.5% of the total. This result was visualized with a histogram of p-values and a pie chart showing the proportion of significant results.

A: Given that only 2.5% of movies show a statistically significant difference in ratings based on watching preference ($p \leq 0.005$), we conclude that there is generally no strong "social watching" effect across most movies. This analysis assumes independent responses and ordinal data, with limitations that include potential multiple testing issues and the assumption that all viewer ratings are independent.





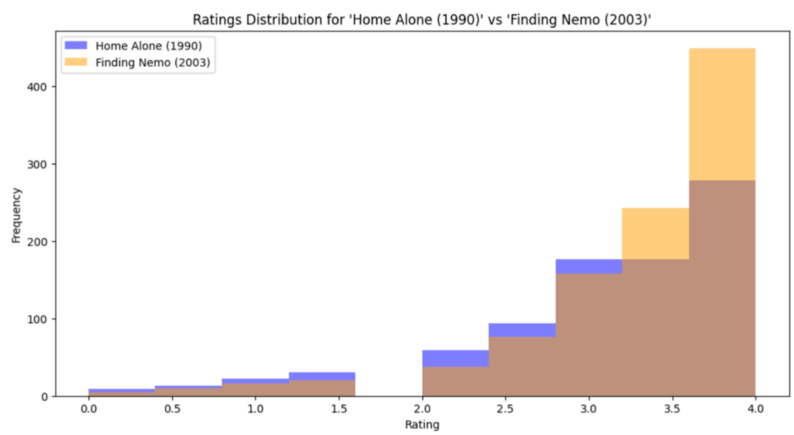
Q9. Is the ratings distribution of 'Home Alone (1990)' different than that of 'Finding Nemo (2003)'?

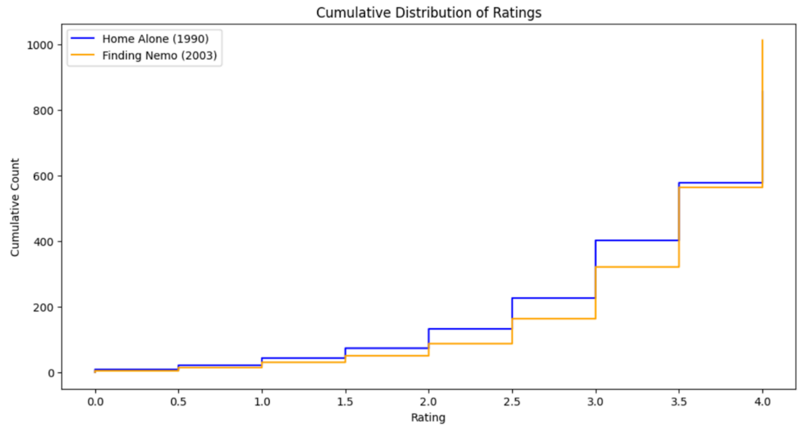
D: We performed a Kolmogorov-Smirnov (KS) test to compare the distributions of ratings for "Home Alone (1990)" and "Finding Nemo (2003)", assuming independence of ratings between the two movies and that the KS test is appropriate for comparing continuous distributions.

Y: We chose the KS test because it is a non-parametric test that assesses whether two independent samples are drawn from the same distribution, which is suitable here for identifying potential distributional differences in ratings without assuming normality.

F: The KS statistic was 0.1527, with a p-value of 6.38×10^ {−10}, indicating a statistically significant difference between the two rating distributions. This result suggests that the ratings for "Home Alone (1990)" and "Finding Nemo (2003)" do not come from the same distribution.

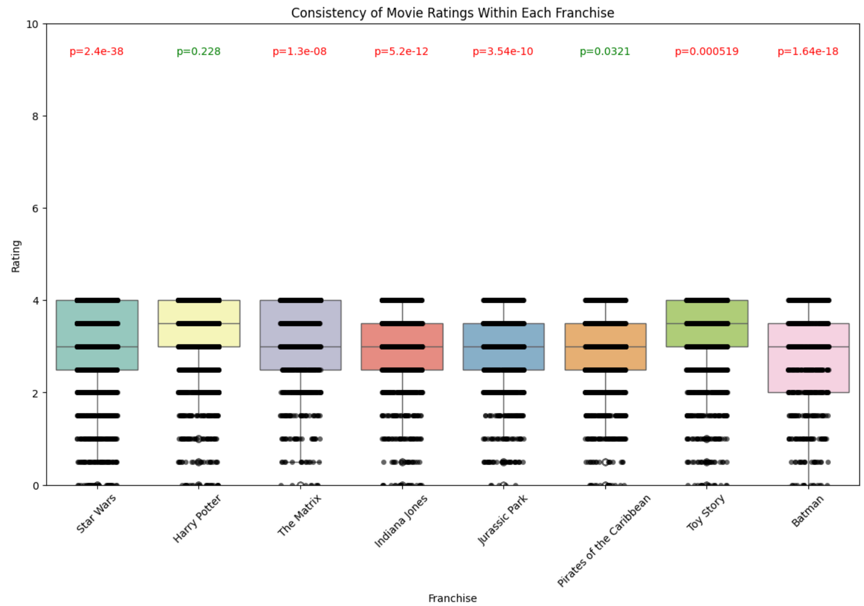
A: Given the p-value of 6.38×10^ {−10}, we conclude that there is a statistically significant difference in the rating distributions of "Home Alone (1990)" and "Finding Nemo (2003)". This conclusion assumes that the KS test is suitable for the data and that the ratings are independent. Limitations include potential sensitivity to large sample sizes, which can detect minor distributional differences as significant.





Q10: 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?

D: We performed a one-way ANOVA test on the ratings of movies within each franchise to assess if there were significant differences in quality as perceived by viewers. A p-value threshold of 0.005 was used to identify franchises with inconsistent quality.
Y: We used the ANOVA test because it is effective for comparing the mean ratings across multiple movies within each franchise, helping to determine if viewers perceived significant differences in quality. The stricter p-value threshold (0.005) was chosen to ensure that only substantial inconsistencies were identified.
F: The findings indicate that 6 out of the 8 franchises were inconsistent, with p-values below 0.005 (highlighted in red). These franchises showed statistically significant differences in ratings among movies, suggesting that viewers perceived some movies within these franchises as being of different quality than others. Franchises with p-values above 0.005 (in green) did not show significant inconsistencies.
A: Based on the ANOVA results and a 0.005 significance threshold, I conclude that 6 franchises exhibit inconsistent quality in movie ratings, as experienced by viewers. This conclusion assumes that each movie’s ratings are independent and normally distributed within each group. Limitations include potential rating scale limitations and possible sensitivity to minor quality variations among movies.



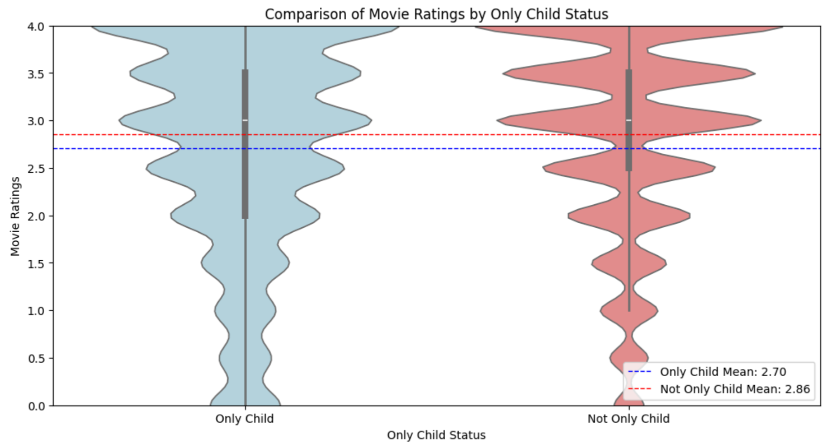
Extra Credit: Tell us something interesting and true (supported by a significance test of some kind) about the movies in this dataset that is not already covered by the questions above.

Is there a significant difference in movie ratings between only children and non-only children?

Null Hypothesis (H0): There is no significant difference in movie ratings between only children and non-only children.

Alternative Hypothesis (H1): There is a significant difference in movie ratings between only children and non-only children.

D: We conducted an independent sample t-test to compare the mean movie ratings of only children and non-only children, assuming that the ratings for each group are independent and approximately normally distributed.
Y: We selected the independent samples t-test as it allows for a direct comparison of means between two groups, testing for a significant difference in average ratings. This method was chosen over non-parametric tests because it provides more power when data can be assumed to be approximately normal.
F: The t-test yielded a t-statistic of -18.31 and a p-value of 8.51e-75, far below the 0.005 threshold. The mean rating for only children was 2.70, compared to 2.86 for non-only children, suggesting a statistically significant but modest difference in movie ratings between the groups.
A: Given the extremely low p-value (8.51e-75), we conclude that there is a statistically significant difference in movie ratings between only children and non-only children, with non-only children giving slightly higher ratings. This analysis assumes that ratings are independent within groups and approximately normally distributed. However, due to the modest difference in means, this significant result may have limited practical impact.



Appendix

Q1: The inbuilt Mann-Whitney U statistic is giving the higher of the u values and p value is 0. We have found the U statistic manually and the value is 800742093.

```
import pandas as pd
import numpy as np
data = pd.read_csv("/Users/alkaashok/Downloads/movieReplicationSet.csv")
#print(data)
#Taking the movie columns
movie_data = data.iloc[:,:400]
#Median for each column
movie_medians = movie_data.median()
movie_popularity = {}
#Highly Popular Movies = High , Lesser Popular Movies = Low
for movie in movie_medians.index:
    overall_median = movie_medians.median()
    movie_popularity[movie] = 'high' if movie_medians[movie]>= overall_median else 'low'
#print(len(movie_popularity.keys()))
#print(movie_popularity)
from scipy.stats import mannwhitneyu, rankdata
#Splitting High and Low Popular Movies into two lists
high_popularity_movies = [movie for movie, popularity in movie_popularity.items() if popularity ==
'high']
low_popularity_movies = [movie for movie, popularity in movie_popularity.items() if popularity ==
'low']
# Number of movies in both lists
num_high_popularity_movies = sum(1 for popularity in movie_popularity.values() if popularity == 'high')
num_low_popularity_movies = sum(1 for popularity in movie_popularity.values() if popularity == 'low')
print("Number of movies - high popularity:",num_high_popularity_movies)
print("Number of movies - low popularity:",num_low_popularity_movies)
#Flatten the data values to an array
```

```
high_popularity_ratings = data[high_popularity_movies].values.flatten()
low_popularity_ratings = data[low_popularity_movies].values.flatten()
#Remove the nan values from both lists
high_popularity_ratings = high_popularity_ratings[~pd.isna(high_popularity_ratings)]
low_popularity_ratings = low_popularity_ratings[~pd.isna(low_popularity_ratings)]
#Number of Ratings for both lists
print("Number of Ratings for high popularity movies:",len(high_popularity_ratings))
print("Number of Ratings for low popularity movies:",len(low_popularity_ratings))
#Ranking the data
combined_ratings = list(high_popularity_ratings) + list(low_popularity_ratings)
ranks = rankdata(combined_ratings)
#combined_ratings['rank'] = combined_ratings[''].rank()
#Separating the higher and lower popularity from the combined list
ranks_high_popularity = ranks[:len(high_popularity_ratings)]
ranks_low_popularity = ranks[len(high_popularity_ratings):]
#Computing the Sum of Ranks to get R1 and R2 values to enter into the Mann-Whitney U Equation
sum_ranks_high_popularity = sum(ranks_high_popularity)
sum_ranks_low_popularity = sum(ranks_low_popularity)
print("rank of high: ", sum_ranks_high_popularity )
print("rank of low: ", sum_ranks_low_popularity )
# Manual (Gives the minimum of both U1 and U2)
U1 =  (len(high_popularity_ratings)*len(low_popularity_ratings)) +
(len(high_popularity_ratings)*(len(high_popularity_ratings)+1)/2)-sum_ranks_high_popularity
U2 =  (len(high_popularity_ratings)*len(low_popularity_ratings)) +
(len(low_popularity_ratings)*(len(low_popularity_ratings)+1)/2)-sum_ranks_low_popularity
whitney = min(U1,U2)
print("Mannn-Whitney",whitney)
#Inbuilt Function (Gives the Higher value? why?p value is 0)
u_statistic, p_value = mannwhitneyu(high_popularity_ratings,low_popularity_ratings, alternative =
'two-sided')
print(f'Mann-Whitney U Statistic: {u_statistic}')
print("{0:.10f} p-value".format(p_value))
if p_value < 0.005:
    print("There is a statistically significant difference in ratings between low and high popularity movies")
else:
    print("There is no statistically significant difference in ratings between low and high popularity
movies")
```