1. H_0: High-popularity movies rated **not higher** than low-popularity movies.

   H_1: High-popularity movies rated **higher** than low-popularity movies.

D:

Based on the median number of ratings (197.5), divide movies into two data sets: high-popularity and low-popularity. We have 200 movies in each data sets.

Use Kolmogorov-Smirnov (K-S) test to do the normality check for ratings in both data sets.

Use Welch's t-test to compare the ratings of two data sets, because both two data sets are normal distributed, but the sample variances are different.

Use Mann-Whitney U test for robust and compare the result to Welch's t-test.

Y:

Use the median number of ratings to divide movies into two data set, make sure the comparison between high-popularity movies and low-popularity movies has no bias.

Use K-S test to determine an appropriate statistical test to use when comparing the ratings, because some tests depend on the distribution of the data. (Reason for not using Shapiro-wilk test: large sample size)

Chose Welch's t-test instead of t-test because two data sets are both normal distributed, but the sample variances are different.

Use Mann-Whitney U test to verify the results from Welch's t-test, adding robustness to the result.

F:

High-popularity movies: K-S test p-value = 0.238, sample variance = 0.085

Low-popularity movies: K-S test p-value = 0.571, sample variance = 0.053

These results indicate that both data sets are normally distributed but have different sample variances.

Welch's T-test: p-value = $9.537 * 10^{-52}$, T-test statistic = 17.756

Mann-Whitney U test: p-value = $1.697 * 10^{-40}$, U value = 35404

The Welch's t-test and Mann-Whitney U test results both strongly reject the H_0 and accept H_1.

A:

With the extremely low p-values from both Welch's t-test and the Mann-Whitney U test, we may confidently reject the H_0 and accept the H_1. This leads us to conclude that high-popularity movies rated **higher** than low-popularity movies.

2. H_0: New movies **not rated differently** than old movies.

H_1: New movies **rated differently** than old movies.

D:

Based on the median years of movie published (1999), divide movies into two data sets: new movies and old movies. We have 174 new movies and 226 old movies.

Use histogram and Q-Q plot check the average rating distribution of new movies and old movies. (Figure 1)

Use Kolmogorov-Smirnov (K-S) test to do the normality check for distribution in both data sets.

Randomly pick 174 samples from old movies (Down sampling).

Conduct a t-test and Mann-Whitney U with new movies to compare whether there are differences in ratings.

Do the down sampling 100000 times and compute the mean of 100000 result.

Y:

Use the median year of movie published to divide movies into two data set, make sure the comparison between high-popularity movies and low-popularity movies has no bias.

Use histogram and Q-Q plot to check the distribution can give people an intuition feeling about the data distribution. (Figure 1)

Use K-S test to determine an appropriate statistical test to use when comparing the ratings, because some tests depend on the distribution of the data. (Reason for not using Shapiro-wilk test: large sample size)

Use down sampling is because the sample size is not equal 174≠226.

Use t-test because these two datasets are both normal distributed.

Use Mann-Whitney U test to verify the results from t-test, adding robustness to the result.

F:

New movies: K-S test p-value = 0.347, skewness: 0.134, sample variance: 0.127

Old movies: K-S test p-value = 0.124, skewness: 0.287, sample variance: 0.121

Histogram and Q-Q plot shows these two sets of data appears to have same distribution. (Figure 1)

These results indicate that both data sets appear to have same distribution and has a quasi-normal distribution.

After 100000 times down sampling:

T-test average: p-value = 0.317

Mann-Whitney U average: p-value = 0.316

The t-test and Mann-Whitney U test results both reject the H_1 and accept H_0.

A:

With the very close p-values from both t-test and the Mann-Whitney U test, we may confidently reject the H_1 and accept the H_0. This leads us to conclude that new movies **not rated differently** than old movies.

3. H_0: Male and female viewers do **not rate** Shrek (2001) **differently**.

   H_1: Male and female viewers do **rate** Shrek (2001) **differently**.

D:

Divide sample into two groups: male and female. We have 743 females and 241 males.

Use histogram and Q-Q plot check the rating distribution. (Figure 2)

Use Kolmogorov-Smirnov (K-S) test to do the normality check for distribution in both data sets.

Randomly pick 241 samples from female ratings. (Down sampling)

Conduct a Mann-Whitney U test.

Do the down sampling and Mann-Whitney U test 100000 times and compute the mean of 100000 result.

Y:

Use gender to dived viewer's ratings is the most appropriate control group.

Use histogram and Q-Q plot to check the distribution can give people an intuition feeling about the data distribution. (Figure 2)

Use K-S test to determine an appropriate statistical test to use when comparing the ratings, because some tests depend on the distribution of the data. (Reason for not using Shapiro-wilk test: large sample size)

Use down sampling is because the sample size is not equal $241 \neq 743$.

Use Mann-Whitney U test because both male and female dataset are not normally distributed. Take the mean value of 100000 times result make result more robust.

F:

Female ratings: K-S test p-value = $4.065*10^{-8}$, statistic:0.180

Male ratings: K-S test p-value =$2.461*10^{-07}$, statistic:0.190

Histogram and Q-Q plot shows these two sets of data does not have normal distribution. (Figure 2)

These results indicate that both data sets do not have normal distribution.

After 100000 times down sampling:

Mann-Whitney U average: p-value = 0.172

The t-test and Mann-Whitney U test results both reject the H_1 and accept H_0.

A:

With p-values > 0.005, we may reject the H_1 and accept the H_0. This leads us to conclude that male and female viewers do **not rate** Shrek (2001) **differently**.

4. H0: Male and female viewers **do not** rate movies differently.

   H1: Male and female viewers **do** rate movies differently.

D:

Divide sample into two groups: male ratings and female ratings.

For each movie:

Use Kolmogorov-Smirnov (K-S) test to do the normality check for distribution in both data sets.

Compare male and female sample sizes and down sampling.

If both datasets are normal distributed: Conduct a Welch's t-test.

Else: Conduct a Mann-Whitney U test.

Compute the mean of 1000 result.

Y:

Use gender to dived viewer's ratings is the most appropriate control group.

Use K-S test to determine an appropriate statistical test to use when comparing the ratings, because some tests depend on the distribution of the data. (Reason for not using Shapiro-wilk test: large sample size)

Use down sampling is because sometimes the sample size is not equal.

Use Mann-Whitney U test for not normally distributed data. Take the mean value of 10000(It takes about 25 mins. Insufficient computing power, otherwise, we would compute 100000 times) times result make result more robust.

Use Welch's t-test for normally distributed data. Take the mean value of 10000 times result make result more robust.

F:

After 10000 times down sampling:

33 of movies was rated different by male and females.

367 of movies was rated no different by male and females.

This would reject the H_0 and accept H_1.

A:

With those samples which has p-values > 0.005, we may reject the H_1 and accept the H_0. This leads us to conclude that Male and female viewers **do** rate movies differently. There are only 9% (33/367) of movies are rated different by male and female viewers.

5. **D**: First, I filtered out the missing data by deleting the rows with N/A and -1 (which means no response). Then, I tested if the data is normally distributed using a hypothesis test and plotted a histogram to verify the result. Next, I used RCT to generate 100 samples in each group with 50 people who enjoy watching movies alone and 50 who do not. Finally, I calculated the average p-value from the 10000 times RCT test using the U test to compare with the significance level.

   Y: Filtering data is an important step in analyzing data. Whether we use a normal distribution or not depends on which test we should use for hypothesis testing. Generally, the t-test is preferred as it has more statistical power. However, if the data is not normally distributed, the U-test should be used instead. The reason we are using RCT test is to eliminate the influence of other confounding factors. By using RCT test, we can obtain a more precise conclusion. We repeat the process 10000 times to obtain a more accurate p-value.

F: K-S test of only children: p value is 3.412348324234531e^(-7)

K-S test of not only children: p value is 2.358702819665285e^(-51)

These values suggest that the data from both groups do not follow a normal distribution, which is also supported by the Figure 3. The average p value within RCT method looping 10000 times by u test is 0.7991836133052952 which means there is no statistically significant difference in ratings between only children and those with siblings.

A: We have a 0.799 probability to conclude that There is no statistically significant difference in ratings between only children and those with siblings. (fail to reject H0)

The limitation is that we can't eliminate all the confounding elements such as gender identity due to insufficient data and limited computation.

6.

D: First, I filtered out the missing data by deleting the rows with N/A and -1 (which means no response). Then, I used RCT to generate 400 samples in each group with 200 people who enjoy watching movies alone and 200 who do not. The samples are generated by sampling with replacement. Finally, I calculated the average p-value from the 1000 times RCT test. For normally distributed data, I used t-test to calculate the average p-value, and for non-normally distributed data, I used U-test instead. Finally, I compared the average p value with the significance level and calculated the proportion of movies exhibit an "only child effect".

Y: Filtering data is an important step in analyzing data. Whether we use a normal distribution or not depends on which test we should use for hypothesis testing. Generally, the t-test is preferred as it has more statistical power. However, if the data is not normally distributed, the U-test should be used instead. The reason we are using RCT test is to eliminate the influence of other confounding factors. By using RCT test, we can obtain a more precise conclusion. We sample with replacement to generate 400 samples per group to mimic sampling from the true underlying distribution. We repeat the process 1000 times to arrive at a more accurate p-value to get a more accurate proportion.

F: Proportion of movies showing the 'only child effect': 0.005

This means only 2 movies have an "only child effect." They are "Leon (1994)" and "The Land That Time Forgot (1974)".

We also tried using 0.8*min(size of group1, size of group2) as sample size for RCT experiments. But such formulation results in each group have limited number of samples. Given that we are using a significance level of 0.005, all the movies do not have significant difference. So instead we use 400 samples per group.

A: Proportion of movies showing the 'only child effect': 0.005. There is only 2 movies average p-value could reject H0. In other words, there is only 2 movies have an “only child effect.”

The limitation is that we can't eliminate all the confounding element such as gender identity due to insufficient data and limited computation.

7.

D: First, I filtered out the missing data by deleting the rows with N/A and -1 (which means no response). Then, I tested if the data is normally distributed using a hypothesis test and plotted a histogram to verify the result. Next, I used RCT to generate 100 samples in each group with 50 people who are only children and 50 with siblings. Finally, I calculated the average p-value from the 10000 times RCT test using the U test to compare with the significance level.

Y: Filtering data is an important step in analyzing data. Whether we use a normal distribution or not depends on which test we should use for hypothesis testing. Generally, the t-test is preferred as it has more statistical power. However, if the data is not normally distributed, the U-test should be used instead. The reason we are using RCT is to eliminate the influence of other confounding factors. By using RCT, we can obtain a more precise conclusion. We repeat the process 10000 times to arrive at a more accurate p-value.

F: K-S test of watching movies socially: p value is 2.216819709934105e^(-14)

K-S test of watching movies lonely: p value is 2.4501961699727183e^(-10)

These values suggest that the data from both groups do not follow a normal distribution, which is also supported by the Figure 4 and 5. The average p value within RCT method looping 10000 times by u test is 0.543898453046458 which means there is no statistically significant difference in ratings for The Wolf of Wall Street (2013) between those who enjoy watching movies alone and those who enjoy them socially.

A: We have a 0.544 probability to conclude that There is no statistically significant difference those who enjoy watching movies alone and those who enjoy them socially (fail to reject H0)

8.

D: First, I filtered out the missing data by deleting the rows with N/A and -1 (which means no response). Then, I used RCT to generate 400 samples in each group with 200 people who are only children and 200 who do not. The samples are generated by sampling with replacement. Finally, I calculated the average p-value from the 1000 times RCT test. For normally distributed data, I used t-test to calculate the average p-value, and for non-normally distributed data, I used U-test instead. Finally, I compared the average p value with the significance level and calculated the proportion of movies exhibit an “social watching effect”.

Y: Filtering data is an important step in analyzing data. Whether we use a normal distribution or not depends on which test we should use for hypothesis testing. Generally, the t-test is preferred as it has more statistical power. However, if the data is not normally distributed, the U-test should be used instead. The reason we are using RCT test is to eliminate the influence of other confounding factors. By using RCT test, we can obtain a more precise conclusion. We repeat the process 1000 times to arrive at a more accurate p-value to get a more accurate proportion.

F: Proportion of movies showing the ' social watching effect ':
This means only 7 movie has an “social watching effect.”
They are The Evil Dead (1981)
Donnie Darko (2001)
Mulholland Dr. (2001)
Apocalypse Now (1979)
American History X (1998)
Midnight Cowboy (1969)
Fatal Attraction (1987)

We also tried using 0.8*min(size of group1, size of group2) as sample size for RCT experiments. But such formulation results in each group have limited number of samples. Given that we are using a significance level of 0.005, all the movies do not have significant difference. Hence we generate 400 samples with replacement per group.

A: Proportion of movies showing the 'only child effect': 0.0175. There is only 7 movies’ average p-value could reject H0. In other words, there is only 7 movies have an “social watching effect.”

The limitation is that we can't eliminate all the confounding element such as gender identity due to insufficient data and limited computation.

9.

D: To compare the distribution of two sets of data, Chi square test is a suitable choice. And I want to graph the data separately and calculate the key values. Also, to conduct Chi-square test the groups need to be independent and the expected frequency in contingency table cell need to have least 5 in 80% of the cells.

Y: Having a graphical representation of the two sets of data and the average and standard deviation also gives me a basic idea of the two sets of data. And the movie rating can be considered as ordinal data, so Chi square test might be more suitable compared to K-S test.

F: Home Alone mean: 3.13 standard deviation: 0.9

Finding Nemo mean: 3.39 standard deviation: 0.788

p-value: 1.72*10^-10

Degrees of Freedom: 3

And we get the bar charts for these two movies. (Figure6)

A: The p-value is less than 0.005 so we reject null hypothesis concluded that there is a difference in the distribution of these two data sets. To ensure the expected frequency in contingency table, I generated a binned contingency table that combined the ratings into broader categories.

10.

D: I would like to start by visualizing the ratings of the movies in. each of these series and calculate the standard deviation for each and see which of those series the standard deviation is compared to the mean standard deviation. Finally introduce the ANOVA test.

Y: The visual chart gives me a basic subjective judgment. Selecting movies by singling out deviations from the average standard deviation lists which series have mixed ratings. The ANOVA test verifies the consistency of the variance in a set of data.

F: Visual graphs were drawn and the results for each standard deviation value and ANOVA test were calculated.

A: By using the ANOVA test, the p-value for Star War is 1.53*10^-45, which smaller than 0.005, indicating inconsistent quality. The p-value for Harry Potter is 0.509, which bigger than 0.005, indicating consistent quality. The p-value for The Matrix is 2.137*10^-11, which smaller than 0.005, indicating inconsistent quality. The p-value for Indiana Jones is 2.26*10^-09, which smaller than 0.005, indicating inconsistent quality. The p-value for Jurassic Park is 1.84*10^-10, indicating inconsistent quality. The p-value for Pirates of the Caribbean is 6.58*10^-5, indicating inconsistent quality. The p-value for Toy Story is 0.00048, indicating inconsistent quality. The p-value for Batman is 1.54*10^-44, indicating inconsistent quality.

The bar charts for these franchises are shown below. (Figure 7-14)

People with higher empathy for movies rate movies higher. Also compared to all 400 movies listed, franchises movies have a lower correlation of ratings.

11.

D: Calculate the correlation between the rating of each movie and the rating of a single metric. Select the movies that are part of a series and spend the correlation values of the two parts in a scatter plot. And introduce independent samples t-test.

Y: Using a scatterplot helps me visualize the difference between the correlation of these two groups. Calculating the correlation helps us to see if there is a relationship between the rating of each movie and people having stronger empathy for the movie. T-test can compare the means of two independent groups which are correlation coefficients of empathy and movie ratings for franchise movies and all movies.

F: Calculated the correlation between the overall 400 movies and people on empathy and the series of movies. Also made scatter plots to both sets of data. (Figure 15-16) And a t-test was made. The average correlation between overall movie ratings and empathy is 0.1, which is higher than the average correlation between series and empathy which is 0.06. The P-value is 0.000843.

A: The p-value is much smaller than 0.005. We can reject the null hypothesis and concluded that we can argue that although people with stronger empathy for movies have a weak tendency to give movies higher scores. However, there is a much weaker tendency to rate movies in series higher. This may be because movie series contain more fictional worldviews, special effects and exciting scenes than heartfelt and thought-provoking stories.

Appendix：

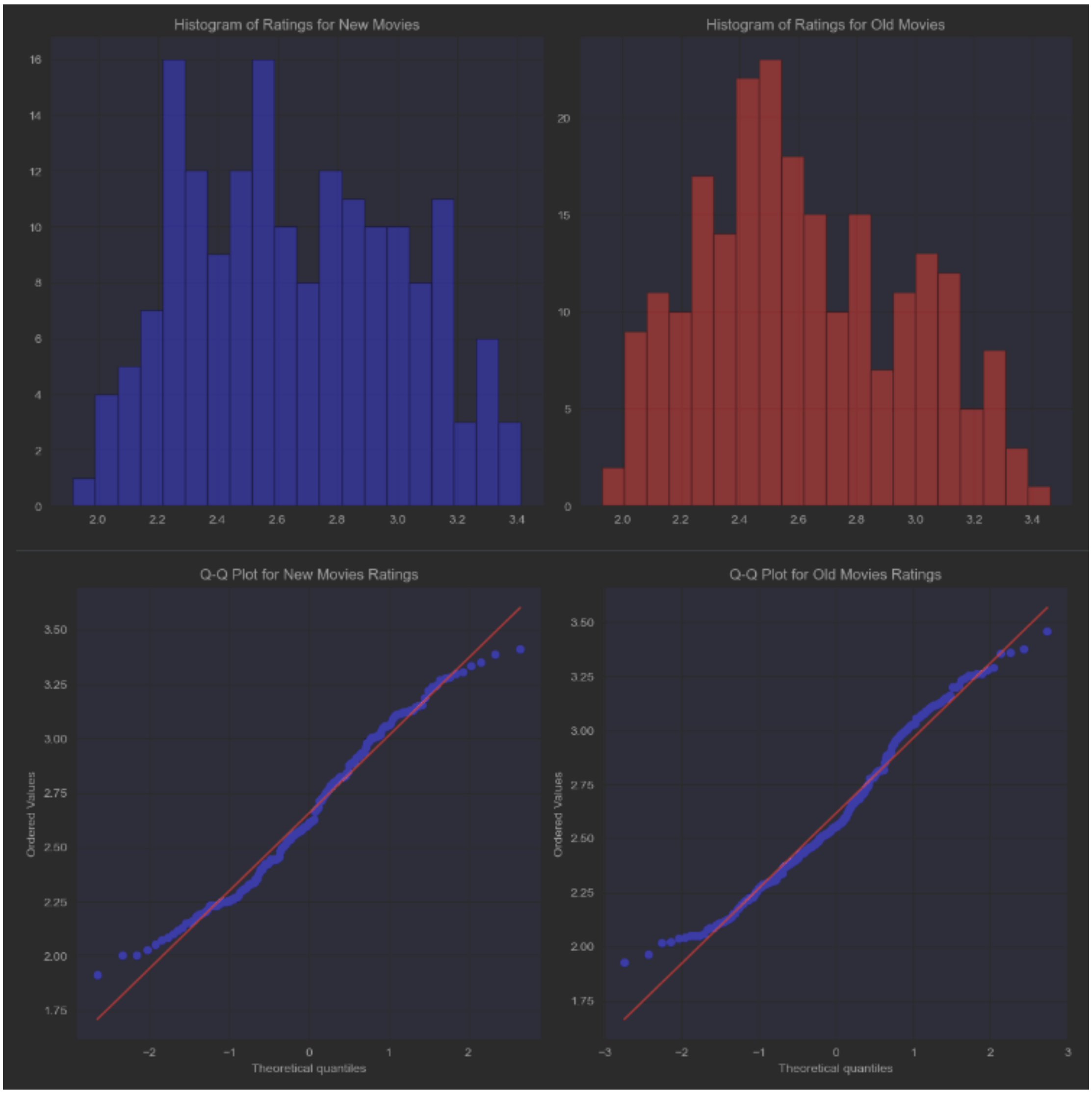

Figure1

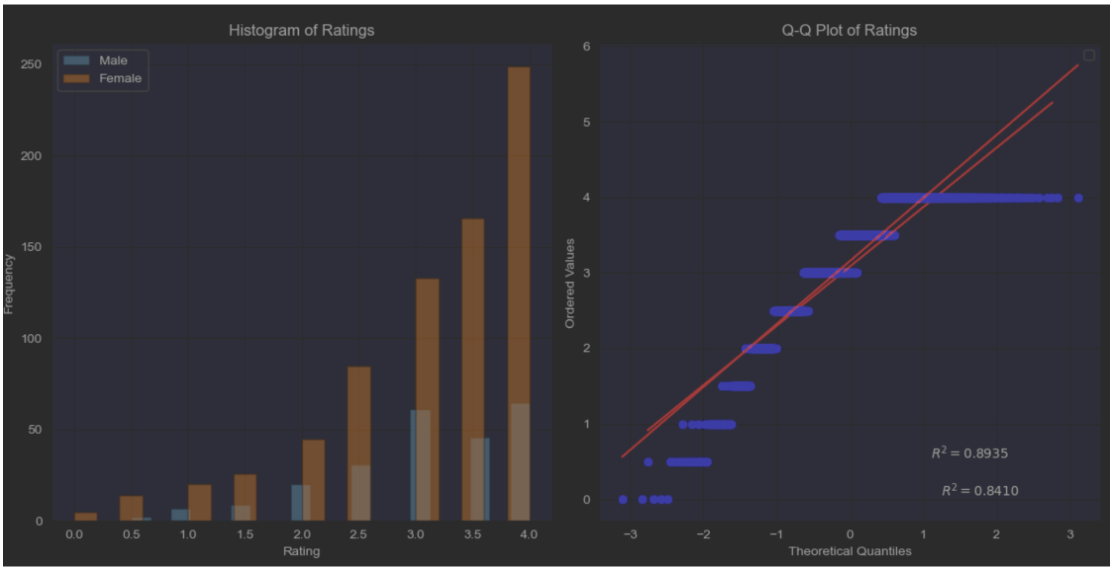

Figure2

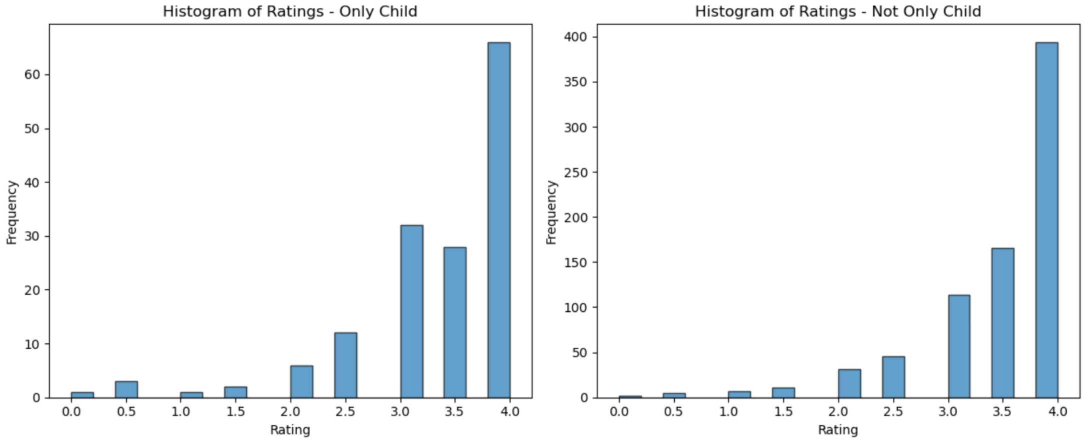

Figure3

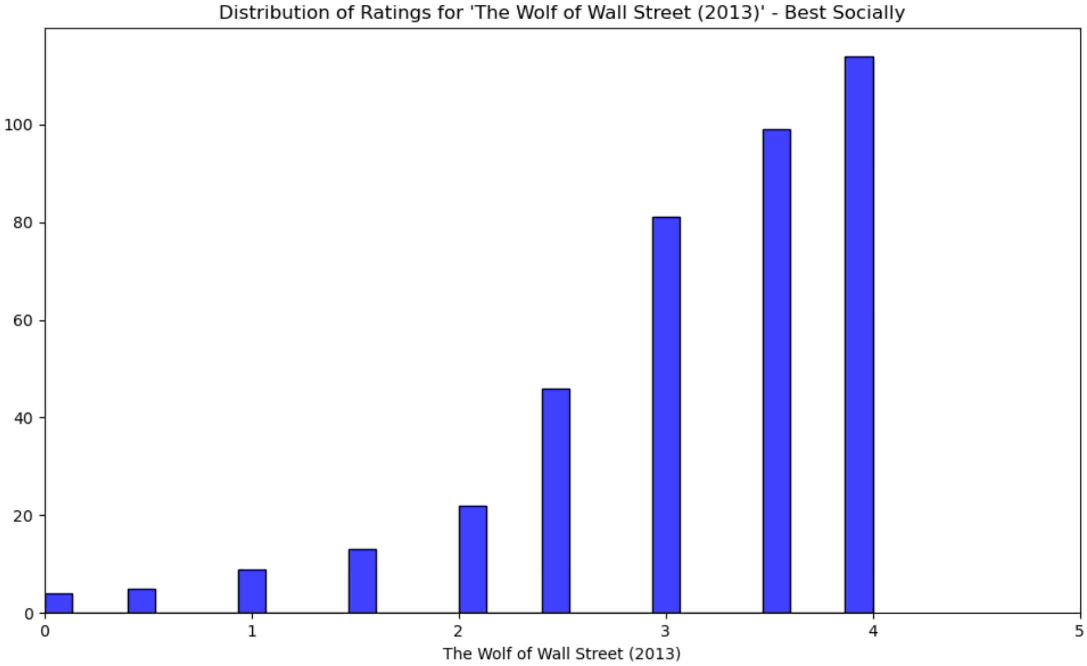

Figure4

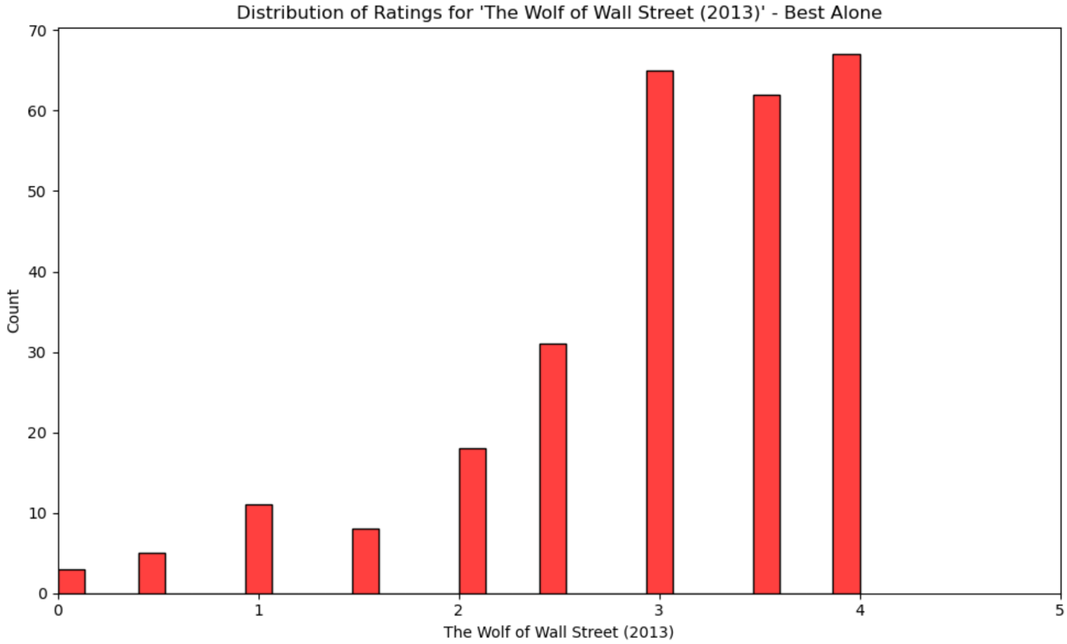

Figure5

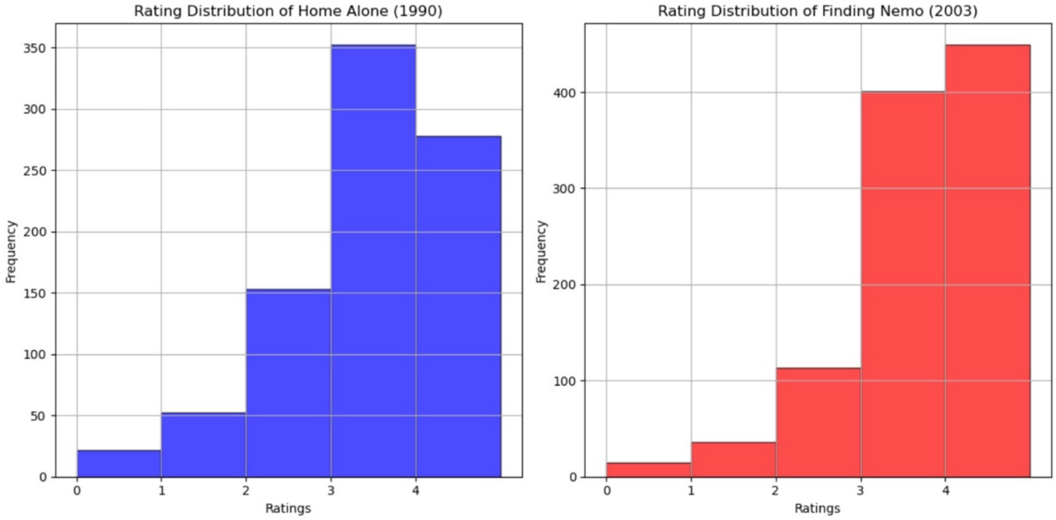

Figure6

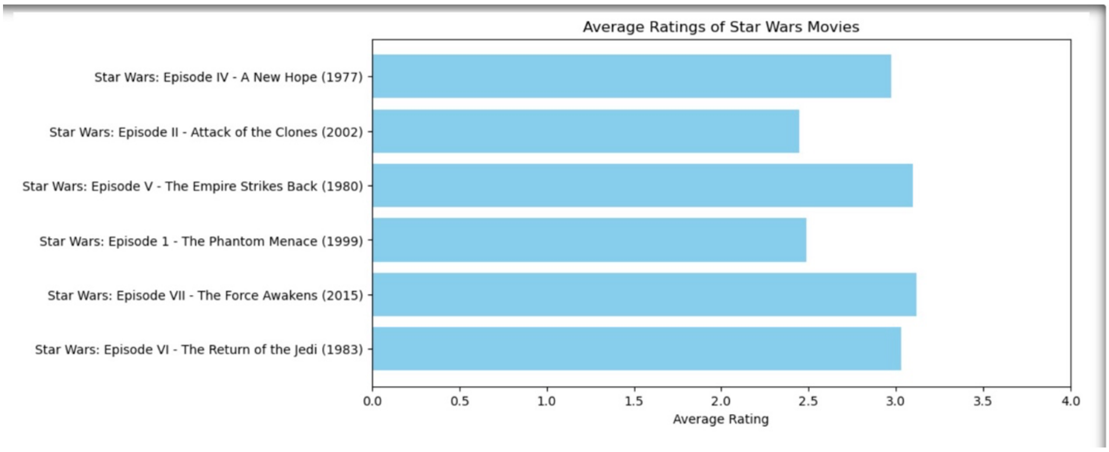

Figure7

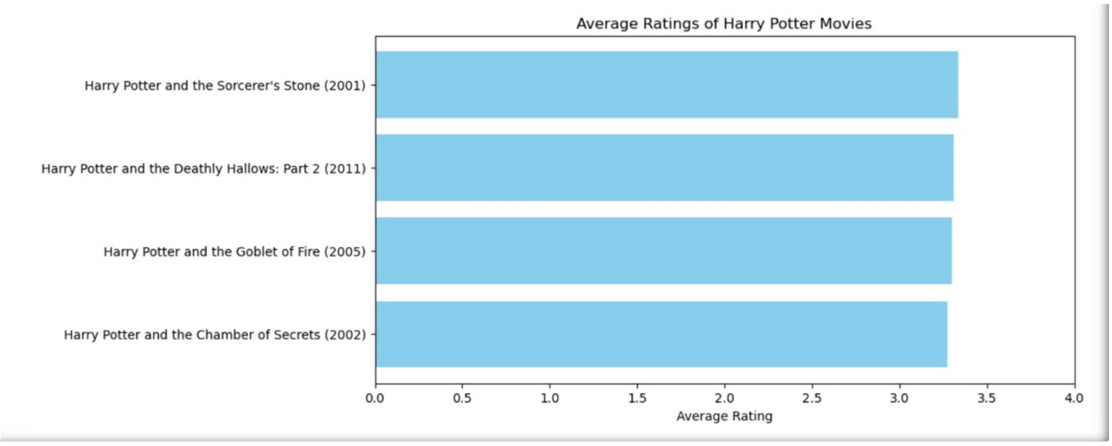

Figure8

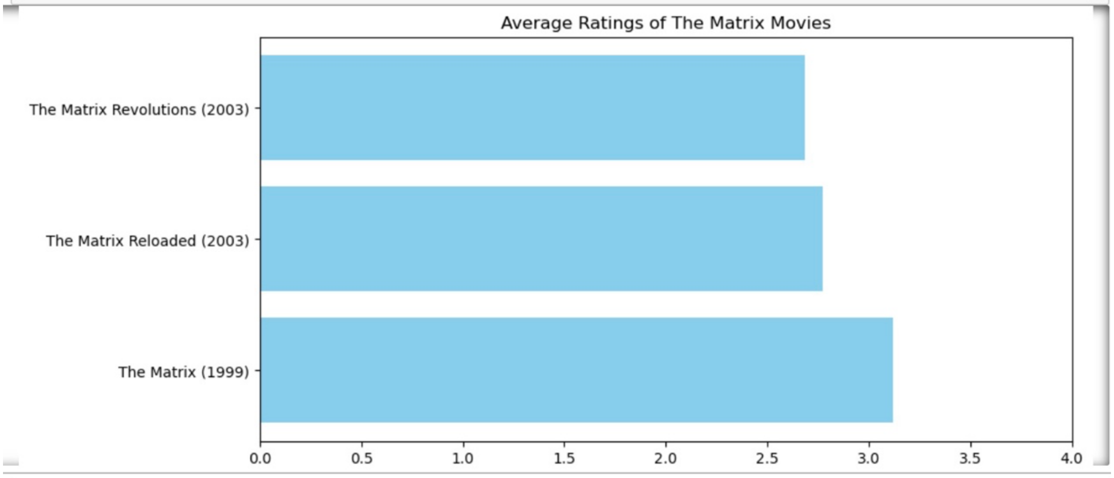

Figure9

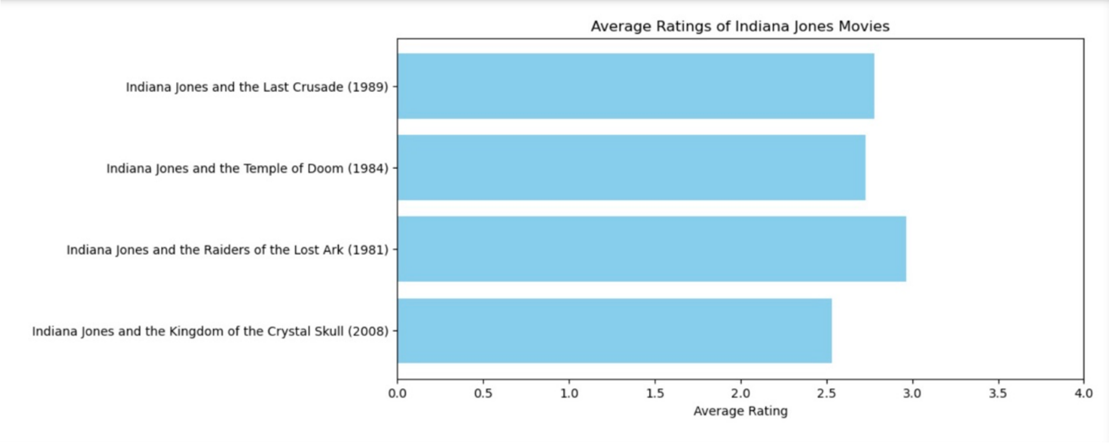

Figure10

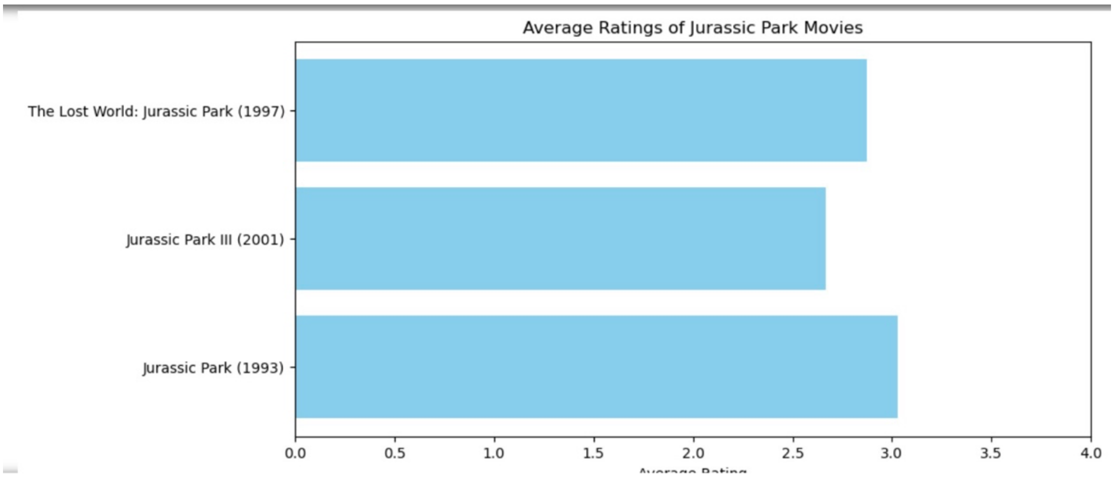

Figure11

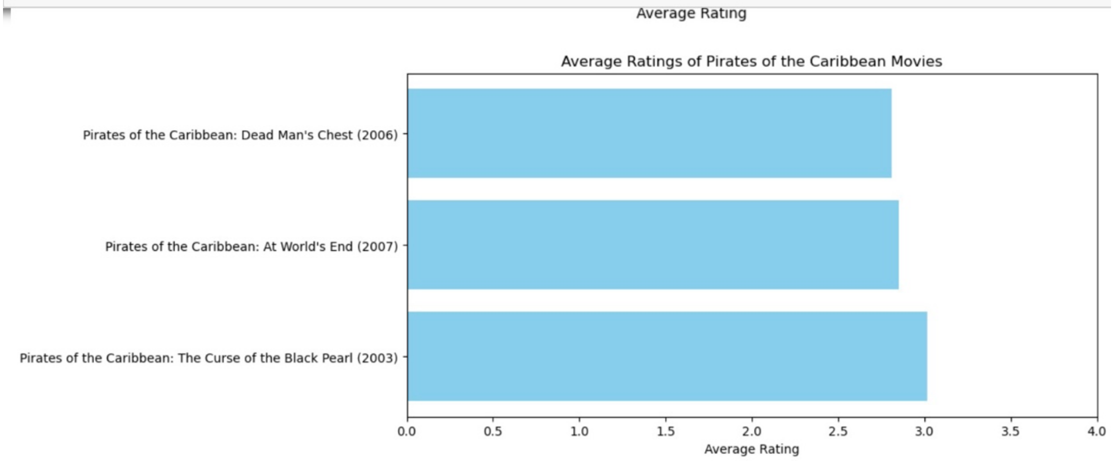

Figure12

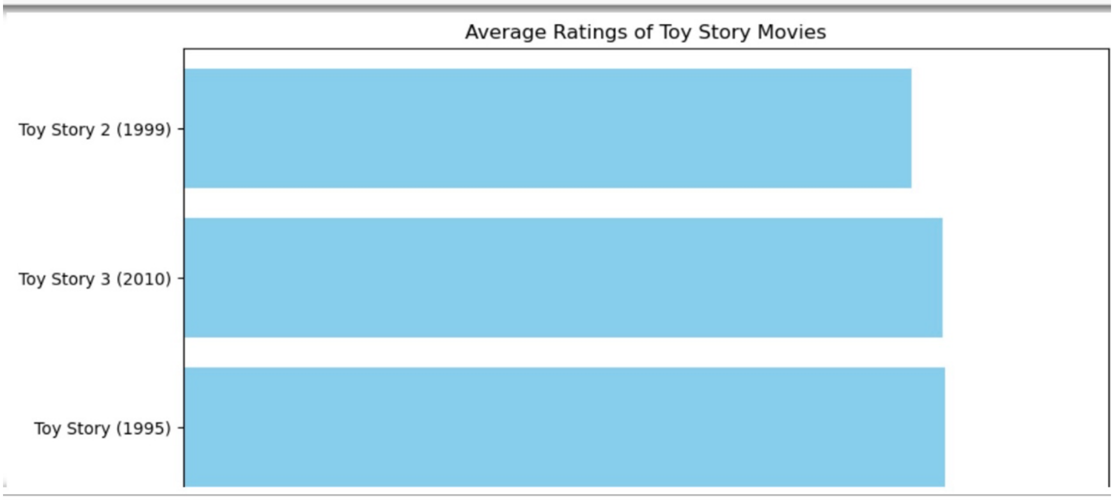

Figure13

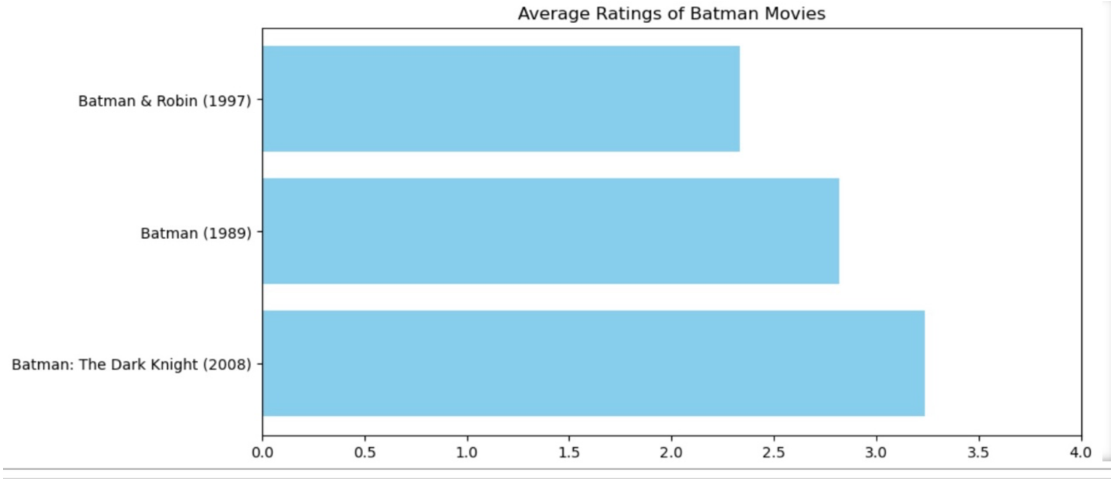

Figure14

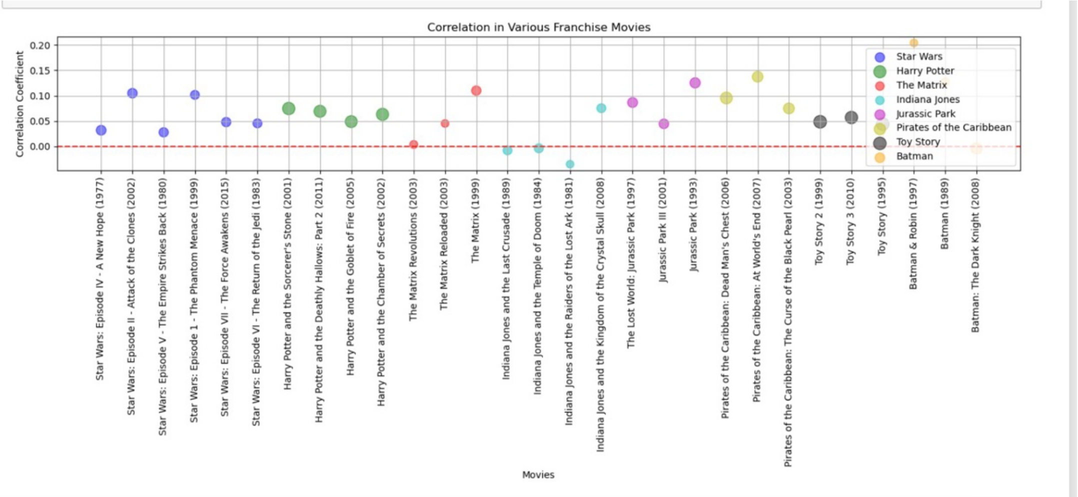

Figure15

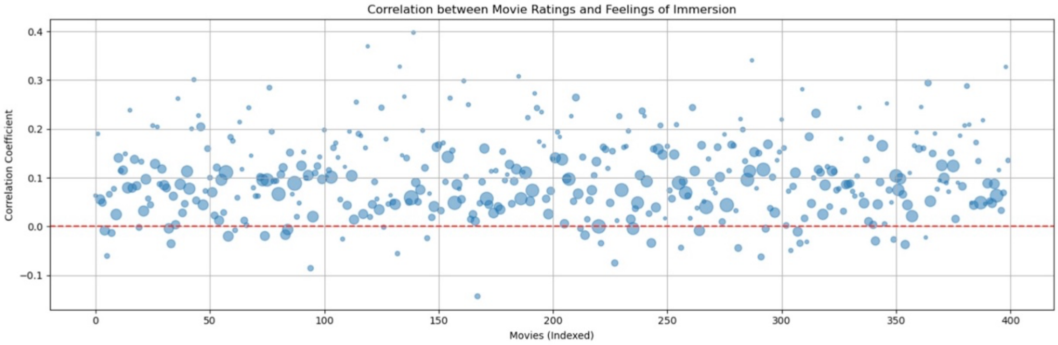

Figure16