Data Analysis Project 1 - Group #21

General assumptions:

The first assumption we make is that a user has watched a movie if and only if they have rated it. Additionally, we assume that all movie ratings are independent from each other and all the users are independent from each other. This allows us to assume independence among groups.
Given the nature of movie ratings, ordinal data, which has an order but not necessarily equal intervals, using the median as a comparison metrics is the most appropriate approach, given its robustness to skewness, often observed in rating distributions.

1) Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular? [Hint: You can do a median-split of popularity to determine high vs. low popularity movies]

D: We did a median-split of popularity, to split the movie ratings into two groups, one for high popularity movies and one for low popularity ones. We want to perform hypothesis testing:

H0: Movies that are more popular are rated less than or equal to movies that are less popular

H1: Movies that are more popular are rated higher than movies that are less popular

Therefore we conducted a one-tailed Mann-Whitney U test to test our hypotheses after dropping the NaN values.
Y: When testing differences among two groups, we perform Mann-Whitney U test, which is a non parametric test. This test does not rely on assumption of normality and does not require constant interval between data points, which aligns with the nature of ordinal data and the psychological variability of ratings.
F: The p-value is almost **0.0**. The test statistic is U = **741899855.5**
A: With a p-value approaching 0.0, indicating a strong statistical difference, we drop the null hypothesis, concluding that more popular movies have significantly higher ratings given our data and alpha.

2) Are movies that are newer rated differently than movies that are older? [Hint: Do a median split of year of release to contrast movies in terms of whether they are old or new]

D: To determine old and new movies, we split the movies based on the median, which is 1999. We want to perform hypothesis testing:

H0: The ratings of older and newer movies are the same.

H1: The ratings of older and newer movies are different.

We conducted a two-tailed Mann-Whitney U test for the significance testing after dropping the NaN values.
Y: Following the rationale from question one, we chose the Mann-Whitney U test to compare ratings across older and newer movies.
F: The p-value is **0.000128**. The Mann-Whitney U test result is **150258386**
A: Given that the p-value is less than 0.005, we drop the null hypothesis, meaning that there is a significant difference in ratings between old and new movies.

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

D: We collected the ratings for Shrek (2001) distinctive to their gender. Then performed hypothesis testing using a two-tailed Mann-Whitney U test, after dropping the NaN values.

H0: There is no difference in ratings between genders

H1: There is a difference in ratings between genders

Y: Following the same rationale from question 1, we performed a two-tailed Mann-Whitney U test to compare the ratings between the two gender groups.
F: We found a p-value of **0.0505**. The Mann-Whitney U test result is **96830.5.**
A: Ergo, the p-value suggests that we fail to drop the null hypothesis, meaning that our data doesn't provide strong enough evidence to falsify the null hypothesis, meaning we can not conclude that there is a significant difference between the groups.

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

D: We are comparing the ratings across genders for each movie and collecting the number of movies that are statistically different through our hypothesis, using two tailed Mann-Whitney U tests.

H0: There is no difference in ratings between male and female viewers

H1: There is a difference in ratings between male and female viewers.

Y: Again, following the same rationale from question 1 about which test to use, we performed a Mann-Whitney to compare the ratings for each movie across both genders.

F: The proportion of movies rated differently across genders is **12.5%**.

A: Our results show that 50 movies were rated differently, meaning that the proportion of movies that rated differently between male and female is 12.5%**.** However, given the large number of tests conducted, we must consider the potential for alpha inflation, which increases the likelihood of finding a significant result purely by chance, because of the increase of false positives.

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

D: Based on user responses, ratings were collected separately for only children and those with siblings. We did hypothesis testing, using a one-tailed Mann-Whitney U test after dropping the NaN values:

H0: People who are only children enjoy 'The Lion King (1994)' less than people with siblings

H1: People who are only children enjoy 'The Lion King (1994)' equal or more than people with siblings

Y: Following the same rationale from question 1, we chose the Mann-Whitney U test to compare ratings for "The Lion King (1994)" movie between people with siblings and those who are only children.

F: The total number of ratings from only children and siblings groups are 151 and 776 respectively. The p-value is **0.97842**, The Mann-Whitney U test result is **52929**

A: Since the p-value is higher than 0.005, we fail to drop the null hypothesis, indicating that there is no evidence that 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 different by viewers with siblings vs. those without?

D: The ratings for each movie in the dataset were collected separately for only children and those with siblings.We perform hypothesis testing, via conducting a two-tailed Mann-Whitney U test:

H0: People who don't have siblings rated the movie differently compared to those who have siblings.

H1: People who don't have siblings rated the movie equally compared to those who have siblings.

Y: Consistent with question 1 rationale, we chose Mann-Whitney U test to compare ratings between viewers with siblings and without.

F: Out of the 400 movies, 7 movies show a significant only child effect, so the proportion is **1.75%**.

A: The proportion of movies with "only child effect" is 1.75%. However, given the large number of tests conducted, we must consider the potential for alpha inflation, which increases the likelihood of finding a significant result purely by chance, because of the increase of false positives.

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?

D: We split the movie ratings, based on the responses of users to the question of whether they prefer to watch movies socially or alone. We want to do hypothesis testing:

H0: Social viewers enjoy watching 'The Wolf of Wall Street (2013)' less or equal to viewers who prefer to watch movies alone

H1: Social viewers enjoy watching 'The Wolf of Wall Street (2013)' more than viewers who prefer to watch movies alone

We drop the NaN values and perform a one-tailed Mann - Whitney U test.

Y: We chose the Mann Whitney U test to compare the ratings between individuals who prefer watching "The Wolf of Wall Street (2013)" socially and those who prefer to watch it alone. This choice is based on the same rationale as question 1 for using this test.

F: The p-value is almost **0.9436**. Test statistic was U = **49303.5**

A: Given that our p-value is 0.9436, we fail to drop the null hypothesis, meaning that we can not make a case for social viewers rating the movie higher than those who prefer to watch it alone.

8) What proportion of movies exhibit such a "social watching" effect?

D: For each movie we separated the ratings from social viewers and from those who prefer to watch them alone. After that we dropped the NaN values, we then performed a one-tailed Mann - Whitney U test for each movie.

Y: We used the one-tailed Mann Whitney U test to identify the proportion of movies where ratings differ between people watching movies socially and those who prefer watching alone. This logic for using this test aligns with our rationale from question 1.
F: The proportion of movies that exhibit such a "social watching" effect is **1.5%**
A: A statistically significant higher rating by social viewers was observed in 7 out of 400 movies (1.5% of the movies). However, given the large number of tests conducted, we must consider the potential for alpha inflation, which increases the likelihood of finding a significant result purely by chance, because of the increase of false positives.

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

D: We collected the ratings for both movies to see if their distribution differs from each other. After dropping the NaN values, we conduct a two-tailed Kolmogorov-Smirnov test to do hypothesis testing.

H0: There is no difference between the ratings distribution of Home Alone (2001) and Finding Nemo (2003)

H1: There distribution from Home Alone (2001) and Finding Nemo (2003) are different

Y: We performed a two-tailed Kolmogorov-Smirnov test to compare distributions of ratings for "Home Alone" and "Finding Nemo". This test was chosen because it is suitable for comparing overall distributions of independent samples.
F: We found a p-value of **6.379e-10** and a test statistic D of **0.15269080020897632**.
A: Ergo, the p-value suggests we drop our null hypothesis, implying that there is a difference between the distribution of the two movies.

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? [Hint: You can use the keywords in quotation marks featured in this question to identify the movies that are part of each franchise]

D: We aggregated movies by franchises based on standard franchise names. We then performed hypothesis testing using a Kruskal-Wallis test.

H0: The rating distribution across movies within each franchise is the same.

H1: The rating distribution across movies within each franchise is different.

Y: We used a Kruskal-Wallis test to compare viewer ratings across multiple movies within each franchise. This test was chosen because it is appropriate for comparing more than two independent groups without assuming normality.
F: We found the following p-values:

Franchise Name	***P-Value***	***Kruskal-Wallis H statistic***
Star-Wars	*8.016e-48*	*230.584*
Harry-Potter	*0.343*	*3.331*
The Matrix	*3.123e-11*	*48.378*
Indiana Jones	*6.272e-10*	*45.794*
Jurassic Park	*7.636e-11*	*46.590*
Pirates of the Caribbean	*3.290e-05*	*20.643*
Toy Story	*5.065e-06*	*24.385*
Batman	*4.225e-42*	*190.534*

A: Based on the p-value, we can drop the null hypothesis for all franchises except **Harry Potter,** meaning significant inconsistencies in ratings for the other 7 franchises.

Something Interesting:

EC) Do people with high sensation-seeking scores rate movies differently than people with low sensation-seeking scores?

D: We split the people into two groups, high-sensation and low-sensation seekers. To do that we took the sum of all their answers to all the sensation seeking questions and then did a median split on that. Whenever an answer to such a question was NaN we imputed it with zero, because we didn't want to exclude users who may have partial experience with sensation seeking activities. After that, we got the movie ratings of the two groups and we performed hypothesis testing using a two-tailed Mann-Whitney U test after dropping the NaN values.

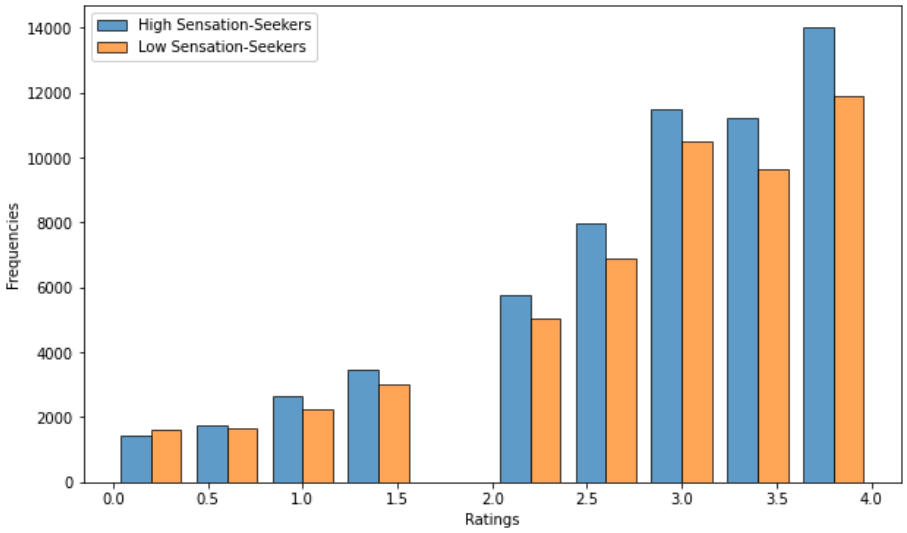
H0: High sensation-seekers rate movies equally as low sensation-seekers

H1: High sensation-seekers rate movies differently than low sensation-seekers

Y: Following the rationale from question 1, we chose the Mann-Whitney U to compare ratings between high sensation seekers and low sensation seekers.

F: The p-value is **0.00006** and the test statistic U=**1588943638**

A: Given that the p-value is less than 0.005, we drop the null hypothesis, which means that high-sensation seekers rate movies significantly differently than low sensation-seekers.



Project Repository:

Github: https://github.com/NickGreen99/movie-database-analysis