Movie Analysis Project

Jinran Jin, Kevin Wang, Rundong Guo

Github repository: https://github.com/jinjinran2001/movie_analysis

Question 1: To see if popular movies get better ratings, we made the assumption that the median rating of high-popularity movies is less than or equal to the median rating of low-popular movies. We started by counting how many ratings each movie received as a measure of its popularity. Then, we split the movies into "high popularity" and "low popularity" groups based on whether they had more or fewer ratings than the median number. After that we went with the Mann-Whitney U because it works well for comparing two groups with median, especially because it is non-parametric and everyone's rating to a movie is independent, which can happen with movie ratings. The test gave me a p-value of about **$9.93 \times 10{-}35$**, which is way below our significance cutoff of 0.005. This super low p-value means there's a statistically significant difference in ratings between popular and less popular movies, showing that the trend isn't just random chance. We dropped the null hypothesis that there is no statistically significant difference between popular movies and non-popular counterparts. This points to a positive link between popularity and audience approval, although other factors like release year might also play a role.

Question2: To explore if newer movies are rated differently from older ones, we started by extracting the release year from each movie title and performed a median split on the release years to classify the movies as either "old" or "new." This split allowed me to compare two balanced groups: movies released before the median year (older) and those released after (newer). Then we used the Mann-Whitney U test to compare the ratings of these two groups. We chose the Mann-Whitney U test because it's ideal for comparing two independent groups with potentially non-parametric distributed data, such as movie ratings. Using the median year to split the movies into "old" and "new" groups provides a straightforward way to examine if a movie's release date influences how audiences rate it.The Mann-Whitney U test returned a p-value of approximately **0.1986,** which is above the significance threshold of 0.005. This high p-value indicates that there isn't a statistically significant difference in ratings between older and newer movies, suggesting that the ratings for these two groups are relatively similar, suggesting that we should keep the null hypothesis. Based on this result, we conclude that there's no significant difference in ratings between newer and older movies. This suggests that, on average, whether a movie is old or new doesn't seem to impact how it's rated by audiences in this dataset, implying that factors other than release year may be more important in influencing ratings.

Question 3: To check if people of different genders rate *Shrek (2001)* differently, we pulled the ratings for this movie and grouped them by gender: female and male. Since we are comparing two groups, we used the Mann-Whitney U test. We chose the Mann-Whitney U test because it's well-suited for comparing two independent groups (in this case, male and female viewers) when we want to test if there's a difference in their central tendencies, such as the median rating. The Mann-Whitney U test doesn't require the data to be normally distributed. The test gave a p-value of approximately **0.05053**, which is much higher than our significance threshold of 0.005. This means there isn't a statistically significant difference in how male and female viewers rated *Shrek (2001)* at the 0.005 level. Based on these results, we conclude that enjoyment of *Shrek (2001)* doesn't significantly vary by gender in this dataset, **thus** we should keep the null hypothesis. The observed p-value suggests that any difference in ratings between genders is not large enough to be considered statistically significant under the 0.005 threshold, indicating *Shrek (2001)* has fairly consistent appeal across genders.

Question 4: To see how often movies are rated differently by males and females, We ran a Mann-Whitney U test on each of the 400 movies. For each movie, we split the ratings by gender (female and male) and tested if there was a statistically significant difference in the ratings between these two groups. Then, we counted how many movies showed a significant difference and got the proportion of movies rated differently by male and female viewers. We used the Mann-Whitney U test because it's suitable for comparing two independent groups (female and male) without requiring the ratings to follow a parametric distribution. Testing each movie individually with this method allowed me to determine how often gender actually influences movie ratings. After running the test across all movies, we found that about **12.5%** of the movies showed a statistically significant difference in ratings between males and females at the 0.5% significance level. This means that for these movies, there was a difference in opinions between male and female viewers. In the end, it appears that gender makes a difference in ratings for approximately 12.5% of the movies. This suggests that while the majority of movies are rated similarly across genders, there's a subset of movies where male and female viewers have differing opinions.

Question 5: To find out if only children like *The Lion King (1994)* more than people with siblings, We split up the ratings for this movie based on whether or not the viewer is an only child. Then, we used a **one-sided** Mann-Whitney U test to compare the two groups and see if there's a real difference in how much they enjoy the movie. We chose the Mann-Whitney U test because it's designed to compare two independent groups (female and male) by focusing on the difference in their median ratings. This test is particularly useful here because it assesses whether one group tends to rate movies higher than the other, without relying on specific assumptions about the overall shape of the rating distribution. The test gives a p-value of **0.978**, which is much higher than our cutoff of 0.005. That means there's no significant difference in how only children and those with siblings rate *The Lion King (1994)*. In short, being an only child doesn't seem to make people enjoy *The Lion King (1994)* any more than those with siblings do. Looks like both groups feel pretty much the same about this Disney classic.

Question 6: We chose to use the **Mann-Whitney U-test** due to the nature of the dataset. This test is appropriate in this case because it does not require the data to be continuous and does not assume a normal distribution. To determine the proportion of people who exhibit an "only child effect," we needed to conduct a two-tailed test to identify any significant differences. In our code, we created a for loop (for i in range(400)) to capture all 400 movies in the dataset. For each movie, we created a DataFrame using the variables "rating" and "only child." We then separated the ratings into only_child_rating and not_only_child_rating to account for the fact that some participants are only children while others have siblings. We used a two-tailed test because we are interested in any significant differences, whether higher or lower; specifically, we aim to determine if only children rate some movies significantly higher or lower. The calculated p-value is **0.0175**, indicating that approximately 1.75% of the movies exhibit an "only child effect" at the 0.005 significance level.

Question 7: We use a Mann-Whitney U-test here because of the nature of the data and because we are comparing two groups: those who like to watch movies socially and those who prefer to watch them alone. Mann-Whitney is a non-parametric test that compares two groups which means it is suitable for this situation. We created a DataFrame q7_data with two columns called 'rating' and 'social'. To account for the two groups, we use q7_alone_rating and q7_not_alone_rating to separate the two groups into those who are socially and those who prefer alone. We then run a one-tailed stats.manwhitneyu to see if those who prefer socially have a greater rating on the movie than the other groups. The p-value we found is **0.056**, which is higher than the 0.005 significance level. This means we should **keep** the null-hypothesis that people who are more socially enjoy the movie more than those who prefer alone.

Question 8: This question is similar to question 6, meaning a two-tailed Mann-Whitney U test is appropriate in this situation. We require a two-tailed test because we want to check if there is a significant difference in ratings between socially watching and those alone, whether it is higher or lower. We created a for_loop to go over ratings of each of the 400 movies in this dataset. Using alone_rating and not_alone_rating, we are able to separate those people who prefer socially and those who prefer alone. We use stats.mannwhitneyu(...) to run Mann-Whitney U test. alternative='two-sided' to ensure this is a two-tailed test. The result is displayed in sum(np.array(q8_p_values) < 0.005) / 400 to show the result. The output is **0.025**, which means about 2.5 percent of the movies exhibit significant 'social watching' effect at the 0.005 significance level.

Question 9: We are not using a Mann-Whitney U test this time because we are trying to find the rating distribution of two movies rather than comparing the ranks of the data. Given that the data set is not continuous, a KS test is appropriate to answer this question given that it is a non-parametric test used to test ratings distribution between two groups.We are able to retrieve the ratings for Home alone and Finding Nemo using home_alone_rating = data['Home Alone (1990)'] and finding_nemo_rating = data['Finding Nemo (2003)']. We performed the Kolmogorov-Smirnov Test using ks_statistic, ks_p_value = stats.ks_2samp(...). We are able to find the p-value using ks_p_value. Given that the P-value is **6.38e-10**, which is well below significant level 0.005, we can conclude that there are statistically significant differences in the ratings distribution between Home alone and Finding Nemo, thus dropping the null hypothesis.

Question 10: We didn't use the Kolmogorov-Smirnov Test for this question because we want to find the distribution of ratings across multiple groups. Hence, we decided to use the non-parametric Kruskal-Wallis test for this question.We created a dictionary called franchise_films to store the ratings of each franchise. Using a for loop to iterate each movie and franchise_films[franchise] = franchise_movies, we are able to store each franchise's movie's ratings. We performed the test using kw_statistic, p_value = stats.kruskal(*franchise_films[franchise]) to find the p-values. If the p-value is below 0.005, then we are able to conclude that there is a significant difference in ratings across movies, suggesting that there exists inconsistent quality inside the franchise. The p-values indicate that franchises like *Star Wars*, *The Matrix*, *Indiana Jones*, *Jurassic Park*, *Pirates of the Caribbean*, *Toy Story*, and *Batman*, exhibit significant differences in ratings across their movies, which suggests inconsistent qualities. Only the *Harry Potter* franchise shows no statistically significant difference in ratings, indicating relatively consistent quality.

Extra: We tried to answer the question: Do people who enjoy driving fast (>=4) tend to rate The Wolf of Wall Street (2013) differently? To answer this question, we used Mann-Whitney U test to test the hypothesis. Mann-Whitney U test is appropriate here because we are comparing two groups, those who enjoy driving fast and those do not. The test is two-tailed because we want to test both statistical significance that is both higher and lower. To solve this problem, we divided the two groups through enjoy_driving_fast = data[data['I enjoy driving fast'] >= 4] and do_not_enjoy_driving_fast = data[data['I enjoy driving fast'] < 4]. Then we tried to find their ratings through fast_rating = enjoy_driving_fast['The Wolf of Wall Street (2013)'].dropna(),andnot_fast_rating = do_not_enjoy_driving_fast['The Wolf of Wall Street (2013)'].dropna(). We performed the two-tailed test using alternative='two-sided' and found out that the p-value is **6.72e-5.** Because the p-value is so small and below the significance level of 0.005, we concluded that there is statistically significant difference in ratings between people who enjoy driving fast and those who do not, thus dropping the null hypothesis.