Group 52: Runhan Chen, Jinrui Fang
DS-GA 1001: Data Analysis Project 1

Data Analysis Report

1. **Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular?**
   - D: We determined the popularity of each movie based on the number of received ratings and categorized them into "High" and "Low" popularity groups using a median split. Subsequently, we conducted an independent sample t-test to compare the average ratings between these two groups.
   - Y: A median split was chosen for its simplicity in categorizing movies, providing a clear dichotomy between more and less popular films. The t-test was employed to objectively ascertain whether a significant difference in average ratings exists between high and low popularity movies, offering a rigorous method to test our hypothesis.
   - F: The t-test result indicated high and low popularity movies have significantly different average ratings(p-value = $2.27{\times}10^{-52}$). And higher popularity movies are rated higher 0.467 than lower popularity movies on average. By visualizing the average ratings between the high and low popularity movies, we found vivid differences from the box plot as well.

   Comparison of Average Ratings between High and Low Popularity Movies
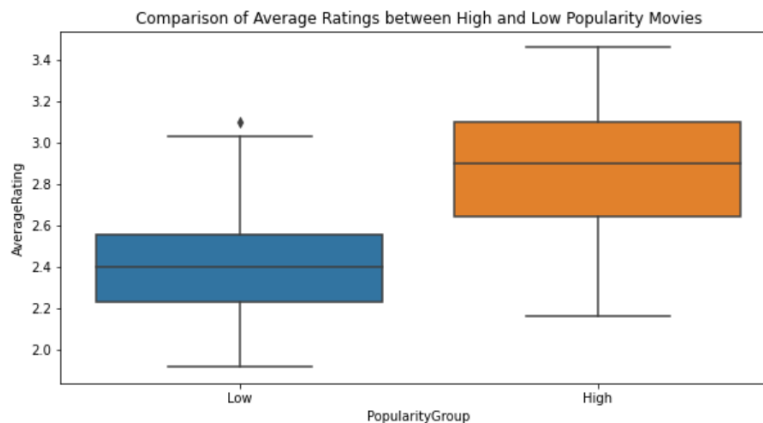
   - A: Given the p-value and visual differences, we concluded that movies that are more popular rated higher than movies that are less popular.

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: We determined the new and old movies based on the median of year among the movies and categorized them into "New" and "Old" movies groups using a median split. A Welch's t-test was then performed to compare the average ratings of the two groups.
   - Y: The median split was used to divide movies into new and old groups. Using Welch's test allows us to determine if there is a statistically significant difference in average ratings between new and old movies. Also, because of the difference size of two groups, we use the Welch's test to do the test.
   - F: The Welch's t-test result indicated new and old movies do not have significantly different average ratings(p-value = 0.2633). By visualizing the average ratings between old and new movies, we found there is no significant difference from the box plot.

Comparison of Average Ratings between Old and New Movies
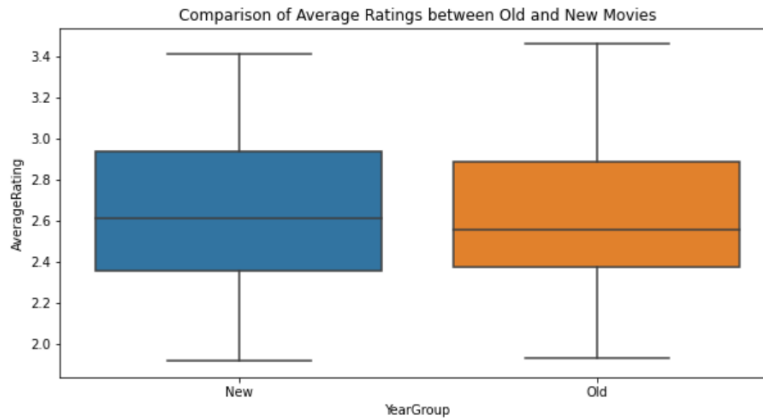
○
- A: Given the p-value and significance level of 0.005, we do not have sufficient evidence to conclude that there is a difference in average ratings between new and old movies.

3. **Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?**
   - We segregated the ratings for 'Shrek (2001)' based on the gender of the respondents, resulting in two groups: female (Gender = 1) and male (Gender = 2). A Welch's t-test was then applied to compare the average ratings of the movie between these two gender groups.
   - Y: Welch's t-test was selected for its suitability in comparing means from female and male groups with unequal variances and sample sizes. The third gender group, self-described (Gender = 3), was excluded due to its insufficient sample size, which could introduce bias and instability in the results.
   - F: The Welch's t-test yielded a p-value of 0.2483, indicating that there is no statistically significant difference in the average ratings of 'Shrek (2001)' between female and male viewers at the 0.005 level of significance.
   - A: Given the p-value, we cannot conclude a statistically significant gender difference in the enjoyment of 'Shrek (2001)'.It is important to acknowledge that by excluding the self-described gender category to avoid possible bias and instability brought by insufficient sample size, we have focused our analysis on female and male categories.

4. **What proportion of movies are rated differently by male and female viewers?**
   - D: We conducted independent Welch's t-tests for each of the 400 movies in the dataset to compare the mean ratings between male (Gender = 2) and female (Gender = 1) viewers.
   - Y: Welch's t-test was chosen due to its ability to handle unequal variances and sample sizes between the two groups, making it a suitable option for this comparison. By identifying the number of movies with significant rating differences between genders, we can calculate the proportion of such movies.
   - F: There are 45 movies that have different average ratings between female and male groups.
   - A: The proportion of movies rated differently by male and female viewers is 11.25%.

5. **Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?**
   - D: We classified the respondents into two categories based on their sibling status: only children (Only Child = 1) and individuals with siblings (Only Child = 0). Following this, we applied Welch's t-test to compare the average ratings given to 'The Lion King (1994)' between these two groups.
   - Y: Welch's t-test was selected as it is well-suited for comparisons between two groups with potentially unequal variances and sample sizes. This method allows for a robust comparison of average ratings, helping us understand if being an only child influences the enjoyment of 'The Lion King (1994)'.
   - F: The Welch's test result(p-value = 0.061) indicated that the average ratings between the two groups were not found to be significantly different.

● A: Based on p-value and the significance level (0.005), there is insufficient evidence to conclude that there is a significant difference in the enjoyment of 'The Lion King (1994)' between only children and individuals 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: For each movie in our dataset, we segregated the respondents into two groups: individuals who are only children (Only Child = 1) and individuals with siblings (Only Child = 0). We proceeded to conduct Welch's t-tests to compare the average ratings for each movie between these two groups.
   ● Y: Welch's t-test was selected as it is well-suited for comparisons between two groups with potentially unequal variances and sample sizes. This approach helps us to investigate whether being an only child or having siblings influences how a viewer rates a movie, helping us identify any potential "only child effect."
   ● F: There are 6 movies which are statistically significantly different between the two groups.
   ● A: The proportion of movies exhibiting an "only child effect" is around 1.5%.

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 categorized respondents based on their answers in Column 477, creating two groups: those who believe movies are best enjoyed alone (Movies are best enjoyed alone = 1), and those who do not (Movies are best enjoyed alone = 0). We then applied Welch's t-test to compare the average ratings for 'The Wolf of Wall Street (2013)' between these two groups.
   ● Y:Welch's t-test is robust to variations in variance and sample size between the two groups, making it a suitable choice for this analysis. This test enables us to evaluate whether the preference for watching movies alone has an impact on the enjoyment of 'The Wolf of Wall Street (2013)'. Any responses marked as -1 (no response) were excluded to maintain accuracy in our analysis.
   ● F: The Welch's test result(p-value = 0.121) indicated there is no statistically significant difference in the average ratings of 'The Wolf of Wall Street (2013)' between the two groups.
   ● A: Based on the data and significance level (0.005), there is insufficient evidence to conclude that there is significant difference between the two groups.

8. **What proportion of movies exhibit such a "social watching" effect?**
   ● D: We split every movie into two groups, one is watching movies socially, the other is watching them alone. Then, we conducted Welch's test to compare the average rating for each movie between the two groups.
   ● Y:Using Welch's test allows us to determine whether there is a statistically significant difference in average ratings between the two groups for each movie.
   ● F: There are 6 movies which are statistically significantly different between the two groups.
   ● A: The proportion of movies exhibiting an "social watching" effect is around 1.5%.

9. **Is the ratings distribution of 'Home Alone (1990)' different from that of 'Finding Nemo (2003)'?**
   ● D: We performed a two-sample Kolmogorov-Smirnov test to compare the ratings distributions of 'Home Alone (1990)' and 'Finding Nemo (2003)'. Additionally, we created comparative visualizations (histograms) to visually assess the distributions.
   ● Y: The Kolmogorov-Smirnov test was chosen for its utility in comparing two samples' distributions without making assumptions about their specific distributions. The visual inspection through histograms provides an intuitive and immediate grasp of the distributions' shapes, centers, spreads, and potential outliers.

- F: The Kolmogorov-Smirnov test resulted in a test statistic of 0.1526 and a p-value of 6.38×10^{-10}. The histogram illustrates that 'Finding Nemo (2003)' generally received higher ratings, predominantly in the 3.5 to 4.0 range. In contrast, 'Home Alone (1990)' has a more varied distribution with a spike in the 0.5 to 1.5 range and another in the 2.0 to 3.0 range. While both movies have ratings clustering around the higher end, the frequency and distribution vary considerably.
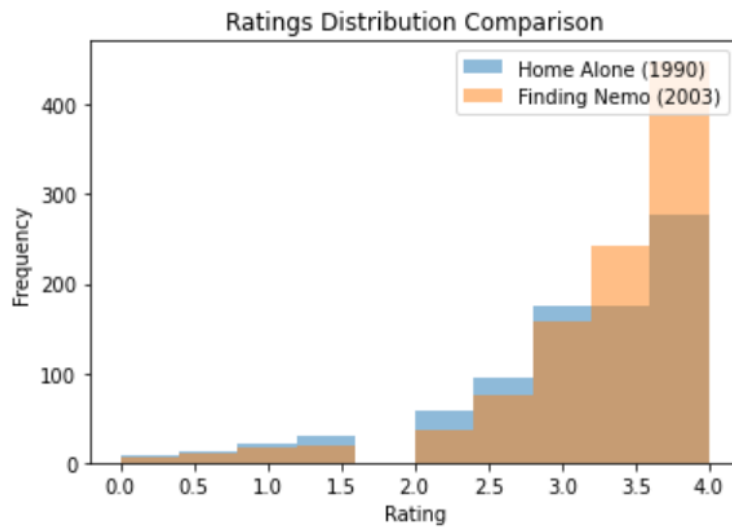
- A: Given the p-value, which is below the 0.005 significance level, we have evidence to conclude that the ratings distributions of 'Home Alone (1990)' and 'Finding Nemo (2003)' are significantly different. The graph confirms our conclusion based on the KS test that the ratings distributions of the two movies are significantly different.

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?**

- D: We segmented the dataset by franchise, calculated the mean and standard deviation of movie ratings for each, and visualized the results using a bar chart. A threshold was established at the average standard deviation across all franchises, and this was represented with a line on the bar chart to identify franchises with inconsistent quality.
- Y: Standard deviation was chosen as it quantifies variability, providing a direct measure of inconsistency in movie ratings; franchises with greater standard deviation exhibit more varied viewer experiences, implying inconsistent quality. The bar chart and threshold line visually demonstrate these variances and enable an easy identification of franchises that fall above the average, signaling inconsistency.
- F: The standard deviations were: 'Star Wars' (0.28), 'Harry Potter' (0.022), 'The Matrix' (0.19), 'Indiana Jones' (0.16), 'Jurassic Park' (0.15), 'Pirates of the Caribbean' (0.09), 'Toy Story' (0.064), and 'Batman' (0.37). 'Batman' showed the highest inconsistency in quality, followed by 'Star Wars', while 'Harry Potter' was the most consistent. The threshold was calculated to be 0.165, placing 'Star Wars', 'Batman', and 'The Matrix' above it, indicating these franchises have inconsistent quality.

Histogram of Standard Deviation of Franchises
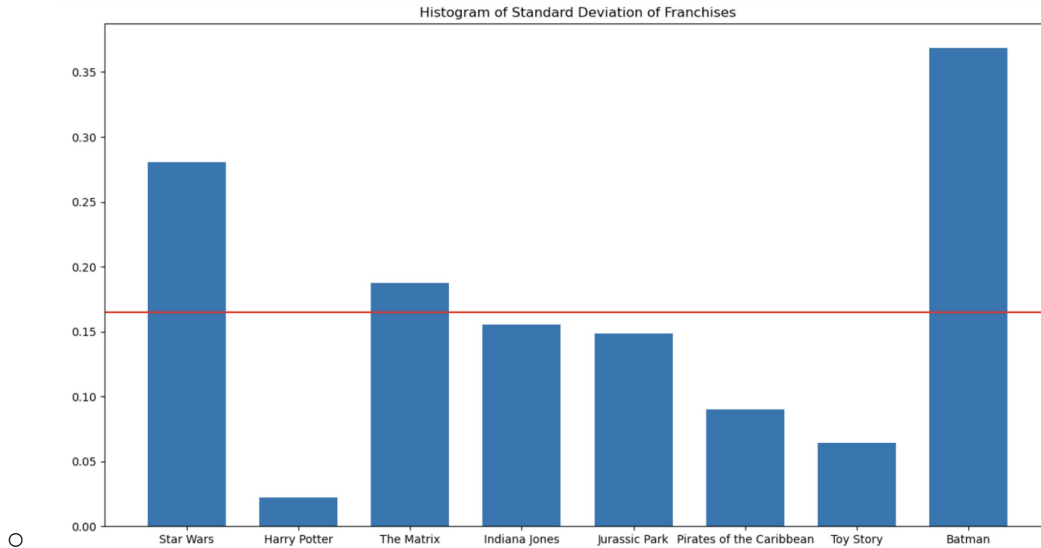
- A: Based on our analysis, 'Star Wars', 'Batman', and 'The Matrix' are identified as **three** franchises with inconsistent quality according to viewer ratings. It's important to note that this approach assumes a normal distribution of ratings within each franchise and that the calculated standard deviation is an accurate representation of inconsistency. Any deviations from these assumptions could affect the validity of our conclusions. Furthermore, setting the threshold at the average standard deviation is a subjective choice, and different thresholds could yield different results.

11. **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**

    Q: What portion of movies that if the audience feels like the things on the screen are happening to me, the feeling for more and less, there is a difference in average ratings between the two groups?

- D: We split every movie into two groups using the median-split, one is watching movies that feel like the things on the screen are happening to me, the other is feeling less. Then, we conducted Welch's test to compare the average rating for each movie between the two groups.
- Y: Using Welch's test allows us to determine whether there is a statistically significant difference in average ratings between the two groups for each movie.
- F: There are 36 movies which are statistically significantly different between the two groups.
- A: The proportion of movies that have this effect is around 9%.