Data Analysis Project 1

Shuyan Liu • Saakshi More • Antonio Vela Gartner

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

We wanted to see if popular movies get higher ratings. First, we looked at how many ratings each movie contained and found the median of all 400 movies. We then split the movies into two groups. Movies with more ratings than the median (high popularity group) and movies with fewer ratings (low popularity group). We conducted a K-S test that showed that the data does not follow a normal distribution. A Mann-Whitney U test instead of an independent sample T-test will give us more statistical power in such cases. The test returned a U-Statistic: 35404.0 P-Value: 1.6971433120157929e-40. Given our alpha level of 0.005, we reject the null hypothesis and conclude that high popularity movies have significantly higher average ratings than movies that are less popular.

2. **Are movies that are newer rated differently than movies that are older?**

We wanted to see if newer movies get rated differently than older movies. First, we extracted the year from each movie from its title. We then found the median and split the movies into two groups: movies made after this year (newer movies) and movies made before (older movies). We did a K-S test to see if the ratings of both groups followed a normal distribution, the test result suggested that our data is not normally distributed. Since we only want to know if there is a difference in our two groups, and there is no specific direction, a two-sided Mann-Whitney U test was conducted. The test returned T-Statistic: 21891.5 P-Value: 0.1010700414217255. Given our alpha level, we failed to reject the null hypothesis and concluded that there is no significant difference between the ratings of newer and older movies.

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

In our analysis to determine whether the enjoyment of the movie 'Shrek (2001)' is gendered, we first isolated the ratings for 'Shrek' and matched them with the gender identities of the respondents, on male and female viewers. To ensure the appropriateness of our subsequent statistical tests, we conducted a Kolmogorov-Smirnov (KS) test on both datasets to assess the normality of their distributions. The results indicated that neither the male nor female ratings followed a normal distribution. Given this, we opted for the Mann-Whitney U test to compare the two groups. The test yielded a U statistic of 96830.5 and a p-value of 0.05054. Since this p-value exceeds our alpha of 0.005, we failed to reject the null hypothesis and conclude that there is no statistically significant difference in the ratings of 'Shrek (2001)' between male and female viewers. We used an element-wise method to remove the null values that resulted in a different sample size for each group, female: 743, male: 241. While Mann-Whitney U test does not require sample size, this could reduce the power of our test.

4. **What proportion of movies are rated differently by male and female viewers?**

We are now interested in knowing what proportion of movies are rated differently by male and female, to achieve this we selected the data for every movie independently and performed a normality KS-test on the ratings given by male and females, if the normality null hypothesis was rejected for either we performed a U-test between the two. In the case that both female and male data were normally distributed, we performed a Levene test to check whether the variances were statistically different to decide between a T-test or a Welsh test. This approach allows us to use the test with the highest statistical power for each movie. At the end we found that 12.5% of the movies were differentially rated by males and females.

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

To analyze the effect of having or not siblings in the ratings of The Lion King (1994) we performed a U-test between two groups because the data was not normally distributed (see Figure 1.). With a p-value of ~0.97 bigger than our threshold, we fail to reject the null hypothesis that the amount of siblings does not affect enjoyment of The Lion King.

6. **What proportion of movies exhibit an "only child effect"?**

To address the question of what proportion of movies are rated differently by only childs we took a similar approach to when we found the proportion of movies rated differently between male and female viewers. In this case, however, each movie's data was divided into two groups (only child or not). As before, we chose the test with the most statistical power for each of the movies by checking for normality and then for similarity of variances in both groups. Here we found that 1.75% of movies are rated differently by only childs.

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?**

To assess this, we first divided people who rated this movie into 2 groups, social and non-social watchers. We did a Levene's test to find if there was a statistically significant difference in the variance of the ratings given by these two groups. The p-value for the test was 0.37, indicating no variance. Then we plotted the distribution of the two groups to check for normality (see Figure 2) because many significance tests assume normal distribution. Since the data was evidently not normally distributed but the two groups had equal variance, we proceeded with a 'greater than' Mann-Whitney test to check if social watchers enjoyed the movie more. The p-value was 0.94, implying that people who like to watch movies socially do not enjoy the movie more than those who like to watch it alone.

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

We took a similar approach to this proportion-related question as well - divided all users into two groups: social and non-social watchers per movie. We chose the test with the most statistical power for each movie by checking for normality and then for similarity of variances in both groups. Here we found that 1.5% of movies exhibit a social watching effect i.e., have people enjoy watching movies socially enjoyed them more than those who prefer to watch them alone.

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

We created 2 groups with the ratings of the 2 movies. We checked for variance using Levene's test where we obtained a p-value of the order of 10e-06, indicating significant variance difference in the 2 groups. Then we plotted the distribution histograms for a normality check (see Figure 3), where we found that the data was not distributed normally. Thereafter, we did a KS test to find if there is a significant difference between the distributions of the two samples. P-value was of the order of 10e-10, indicating significant difference between the groups.

10. **Franchises' inconsistent quality?**

For each franchise, we selected the columns that belonged to the franchise. We only kept rows that had no null values since it would be invalid to check for inconsistency in the ratings given by a user if they haven't watched all movies of the franchise. We conducted a one-way f-statistic for each franchise to check if there is any significant difference among at least one pair of movies within the franchise. We found that from the franchises' listed, only Harry Potter and Pirates of the Caribbean had consistent quality (there was no significant difference in the movies under these franchises).

11. **Extra Credit: Are movies with short titles seen more frequently than movies with long titles?**

To test the hypothesis that movies with shorter names are usually more frequently seen, we divided the data set between movies with more or less than or equal to 'x' words and performed a one-sided U test (since data is not normal,

see Figure 4.) on the ratings count for each group. Importantly, we assigned series movies to the correct group based on the series name, not on the movie name. For instance, if the threshold x=2 any "Harry Potter" movie would be in the group "less than or equal" and any "Pirates of The Caribbean" would be in the "more than" group. This was done so that not all franchises would be in the same group. Under this set-up the length of the name did not seem to affect the popularity of the movie (for x ranging from 1 to 7), for instance, for x=2 the p-value was ~0.28.
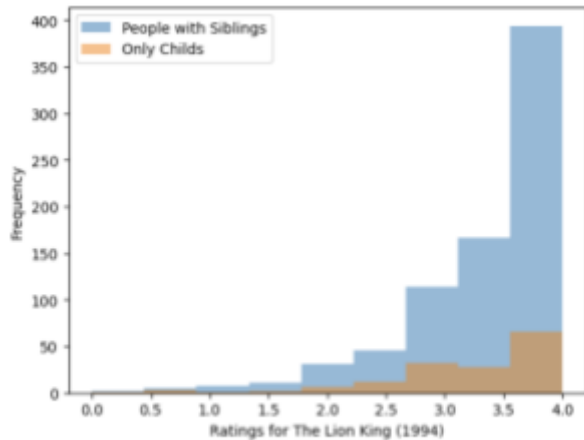
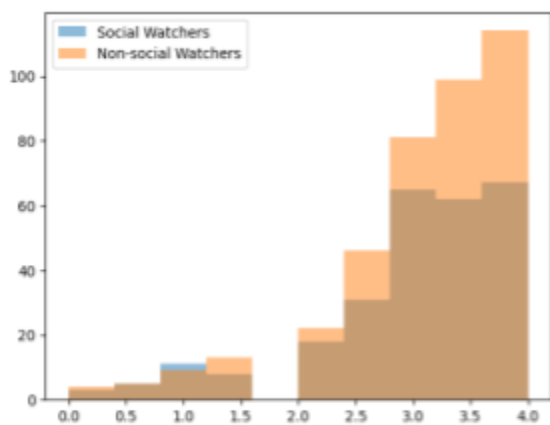Figure 1. Distribution of ratings for People with Siblings v/s Only Childs for 'The Lion King (1994)'

Figure 2. Distribution of ratings for Social v/s Non-Social Watchers for 'Wolf of the Wall Street'
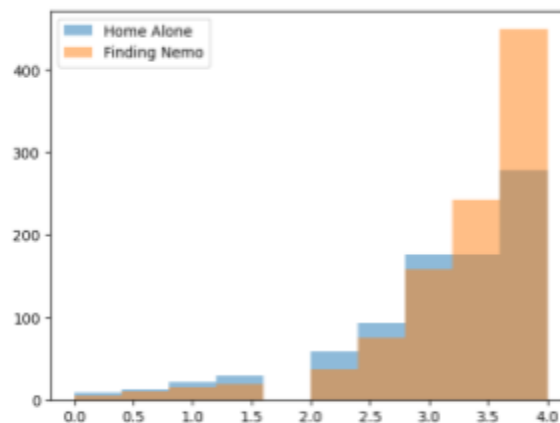

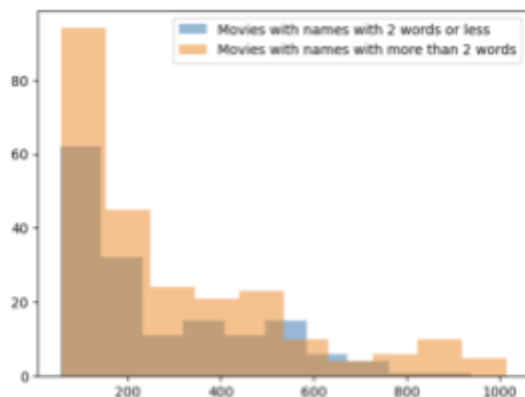Figure 3. Distribution of ratings for 'Home Alone' v/s 'Finding Nemo'

Figure 4. Distribution of number of ratings for movies with names shorter or longer than 2 words