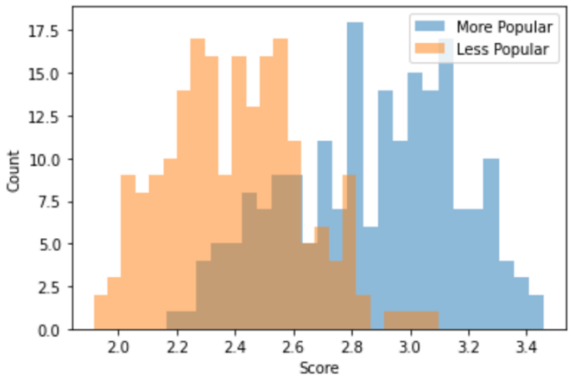
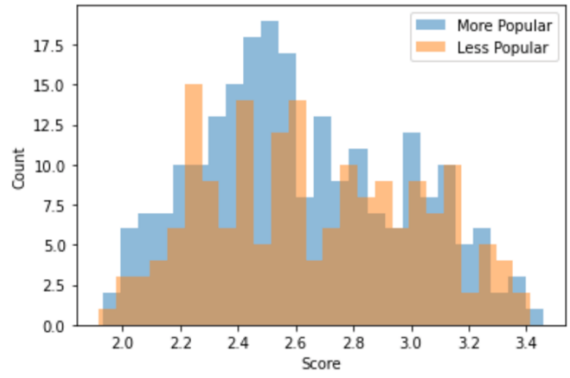
DS-GA 1001
Seonhye Yang (sy3420)
DS-GA 1001 project 1

1. In order to determine if popular movies are rated higher than movies that are less popular movies, we need to perform a median-split. First we sum the number of the ratings of different movies and sort them in ascending order to split them into two columns: low ratings and high ratings. Then with the two ratings, we can perform a Mann Whitney U test. The NaN values were not considered and used np.isnan to take care of Nan values. Then we averaged less popular movies and averaged more popular movies. We use Mann Whitney U test because it is a nonparametric test of the null hypothesis for randomly selected values such as low ratings and high ratings. We get a Mann Whitney U (average ratings of less popular movies and average rating of more popular movies) result of 4596 and pvalue 1.697*10^-40. We have an extremely low pvalue which is smaller than the alpha level 0.005. Since the pvalue is less than 0.005, there is reason to believe that there is a difference between ratings of movies that are more or less popular.

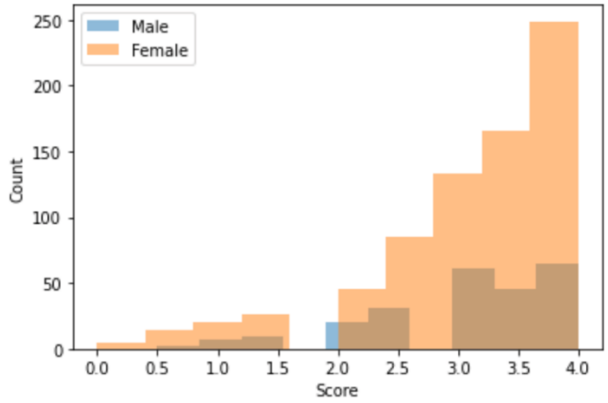


2. In order to determine if newer movies are rated differently than older movies, we need to do a median split on the date (years) then do a Mann Whitney U test on the two rating columns based on years. First, we need to sort the years in ascending order to split it into two columns: older movies and new movies. Since the median of years is 1999, movies released in 2000 would be considered a newer movie. And movies released in 1998 would be considered an older movie. The NaN values were not considered and used np.isnan to take care of Nan values. Then we can perform a Mann Whitney U test on average ratings of new movies and average ratings of older movies. We get a MannWhitneyU statistic of 18330 and a pvalue of 0.275. Our pvalue is greater than alpha = 0.005, we cannot reject the null hypothesis. Therefore, we fail to reject no difference between newer rated movies and older rated movies. There is reason to believe that newer movies are not rated differently than older movies.

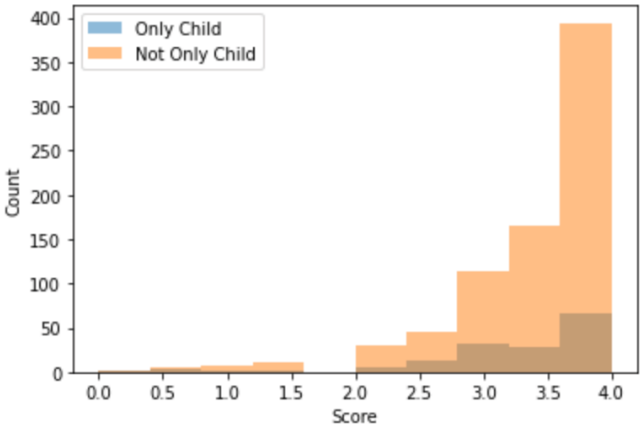


3. To determine the enjoyment of Shrek (2001) between males and females, we need to split Shrek (2001) column according to genders. Then, we will have two columns: ratings of Shrek (2001) for males

and ratings of Shrek (2001) for females. The NaN values were not considered and used np.isnan to take care of Nan values in the Shrek (2001) data. For the hypothesis testing, a Mann Whitney U test was most appropriate for a nonparametric test of the null hypothesis. We get a Mann Whitney U test of 82232.5 and a pvalue of 0.0505. The pvalue is greater than alpha = 0.005 which implies that we cannot reject the null hypothesis. Therefore, we fail to reject that enjoyment of Shrek (2001) is not gendered. We have reason to believe that enjoyment of Shrek (2001) is not gendered.

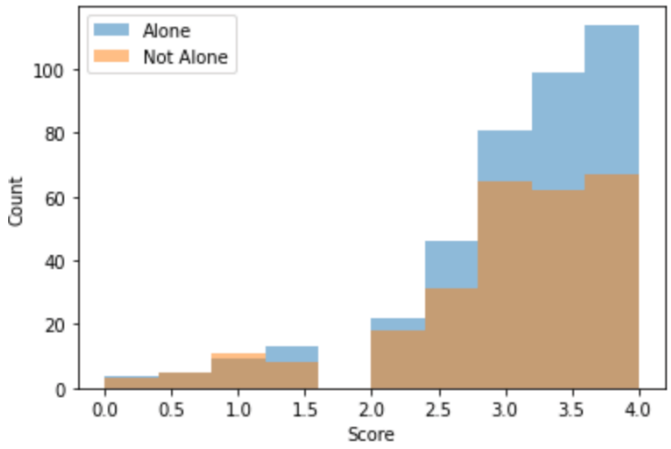


4. To determine the proportion of movies are rated differently by male and female viewers, the Mann Whitney U test was performed for each movie according to gender. Since there are 400 pvalues from the Mann Whitney U test performed for all movies, pvalues less than alpha = 0.005 was divided by total to get the proportion. The result was (all pvalues < 0.005)/400) = 0.125.
5. To determine whether people who are only child enjoy the lion king (1994) more than people with siblings, we need to create two columns from the lion king according to only child and not only child. For instance, values with 1 in “only child” column suggests someone is an only child. Whereas values with 0 in “only child” column suggests someone is not an only child or has siblings. Then performed a Mann Whitney U test of Lion King (1994) for only child and not only child. We get a Mann Whitney U test of 64247 and a pvalue of 0.04319. The pvalue is greater than alpha = 0.005 which implies that we cannot reject the null hypothesis. Therefore, we fail to reject not enjoyed differently by only child and siblings. We have reason to believe that only child does not enjoy the Lion King more than those with siblings.



6. To find the proportion of movies that exhibited “only child effect”, the Mann Whitney U test was performed for each movie according to only child and not only child. Since there are 400 pvalues from the Mann Whitney U test performed for all movies, pvalues less than alpha = 0.005 was divided by total to get the proportion. The result was (all pvalues < 0.005)/400) = 0.0175.

7. To determine whether people who prefer to watch the Wolf of Wall Street (2013) alone more than those who prefer to watch them alone, we need to create two columns from the Wolf of Wall Street (2013) according to alone and not alone. For instance, values with 1 in “Movies are best enjoyed alone” column suggests someone prefers to watch movies alone. Whereas values with 0 in “Movies are best enjoyed alone” column suggests someone prefers to not watch movies alone. Then performed a Mann Whitney U test of the Wolf of Wall Street (2013) for alone and not alone. We get a Mann Whitney U test of 49303.5 and a pvalue of 0.1127. The pvalue is greater than alpha = 0.005 which implies that we cannot reject the null hypothesis. Therefore, we fail to reject not enjoyed differently by watching movies alone or watching not alone. The Wolf of Wall Street has the same enjoyment or rating in a group and when watched alone.



.

8. To find the proportion of movies that exhibited “social watching effect”, the Mann Whitney U test was performed for each movie according to those who prefer watching alone and those who don’t prefer watching alone. Since there are 400 pvalues from the Mann Whitney U test performed for all movies, pvalues less than alpha = 0.005 was divided by total to get the proportion. The result was (all pvalues < 0.005)/400) = 0.025.
9. To determine whether the ratings distribution of “Home Alone (1990)” is different than that of “Finding Nemo (2003)”, we need to perform a KS-test. KS-test is appropriate because it’s a nonparametric test that can be used to compare samples with a reference probability distribution. When performing a KS-test, the NaN values in both “Home Alone (1990)” and Finding Nemo (2003)” were not considered by using np.isnan. The results of KS-test were statistics = 0.1526 and pvalue = 6.379*10^-10. The pvalue is less than alpha = 0.005 which indicates that we can reject the null hypothesis. Therefore, there is reason to believe that ratings distribution of “Home Alone (1993)” is different than that of “Finding Nemo (2003)”.
10. Here we must find if a specific movie in a given franchise is rated inconsistently when compared to the rest of the franchise. To accomplish this, we first gather the movies within each franchise, all six Star Wars movies for example. We then take a specific movie in the franchise and collect the ratings for it. After this, we average the ratings for each person who gave a rating for the other movies in the franchise. Using this data, we can then perform a Mann-Whitney U test against these two pieces of data to get a P-Value which will indicate if a specific movie is rated inconsistently from the rest of the franchise. Our results are as follows:

Star Wars

- Star Wars: Episode IV - A New Hope (1977) is significantly inconsistent from the Star Wars franchise (2.3450502229954436e-08)

- Star Wars: Episode II - Attack of the Clones (2002) is significantly inconsistent from the Star Wars franchise (2.1000042457004973e-12)
- Star Wars: Episode V - The Empire Strikes Back (1980) is significantly inconsistent from the Star Wars franchise (2.0913070644738035e-15)
- Star Wars: Episode 1 - The Phantom Menace (1999) is significantly inconsistent from the Star Wars franchise (1.6842626621129465e-06)
- Star Wars: Episode VII - The Force Awakens (2015) is significantly inconsistent from the Star Wars franchise (1.2122254655003124e-19)
- Star Wars: Episode VI - The Return of the Jedi (1983) is significantly inconsistent from the Star Wars franchise (1.7352516154109993e-10)

We can gather that the Star Wars films are generally inconsistent, this is likely due to the average rating of the films being skewed between the 2 sets of releases.

Harry Potter
- Harry Potter and the Sorcerer's Stone (2001) is significantly inconsistent from the Harry Potter franchise (6.54317976666447e-05)
- Harry Potter and the Deathly Hallows: Part 2 (2011) is not significantly inconsistent from the Harry Potter franchise (0.00939031794691776)
- Harry Potter and the Goblet of Fire (2005) is not significantly inconsistent from the Harry Potter franchise (0.050548556315244624)
- Harry Potter and the Chamber of Secrets (2002) is not significantly inconsistent from the Harry Potter franchise (0.06831980448883875)

Harry Potter produces consistent films, with only the first one being inconsistent.

The Matrix
- The Matrix Revolutions (2003) is significantly inconsistent from the The Matrix franchise (0.00011009709352874405)
- The Matrix Reloaded (2003) is not significantly inconsistent from the The Matrix franchise (0.0845918279549569)
- The Matrix (1999) is significantly inconsistent from the The Matrix franchise (1.1973713955148384e-13)

Indiana Jones
- Indiana Jones and the Last Crusade (1989) is not significantly inconsistent from the Indiana Jones franchise (0.01369744358753365)
- Indiana Jones and the Temple of Doom (1984) is not significantly inconsistent from the Indiana Jones franchise (0.2955137949555886)
- Indiana Jones and the Raiders of the Lost Ark (1981) is significantly inconsistent from the Indiana Jones franchise (4.8138786843341715e-12)
- Indiana Jones and the Kingdom of the Crystal Skull (2008) is significantly inconsistent from the Indiana Jones franchise (0.001487123008825187)

It would seem that the most critically acclaimed film in the series and the worst rated one in the series significantly differ, but otherwise Indiana Joes is consistent in ranking.

Jurassic Park

- The Lost World: Jurassic Park (1997) is not significantly inconsistent from the Jurassic Park franchise (0.2257846937054424)
- Jurassic Park III (2001) is significantly inconsistent from the Jurassic Park franchise (6.07414364628455e-06)
- Jurassic Park (1993) is significantly inconsistent from the Jurassic Park franchise (6.345104910188874e-11)

The films in the original series are rated inconsistently, likely higher or lower while the new set of films would be considered “average”.

Pirates of the Caribbean
- Pirates of the Caribbean: Dead Man's Chest (2006) is not significantly inconsistent from the Pirates of the Caribbean franchise (0.05276909093173878)
- Pirates of the Caribbean: At World's End (2007) is not significantly inconsistent from the Pirates of the Caribbean franchise (0.3364364890779179)
- Pirates of the Caribbean: The Curse of the Black Pearl (2003) is significantly inconsistent from the Pirates of the Caribbean franchise (3.633538040026717e-08)

Pirates of the Caribbean is a consistent franchise, with only the first film being inconsistent.

Toy Story
- Toy Story 2 (1999) is significantly inconsistent from the Toy Story franchise (0.0030214363595904643)
- Toy Story 3 (2010) is significantly inconsistent from the Toy Story franchise (0.0010162148764891253)
- Toy Story (1995) is significantly inconsistent from the Toy Story franchise (1.083656038429831e-05)

The Toy Story franchise is generally inconsistent.

Batman
- Batman & Robin (1997) is significantly inconsistent from the Batman franchise (6.0799921498227264e-30)
- Batman (1989) is significantly inconsistent from the Batman franchise (0.0011564209421138915)
- Batman: The Dark Knight (2008) is significantly inconsistent from the Batman franchise (2.720230526187444e-35)

The Batman franchise is generally inconsistent, likely due to different directors over the course of time.

Extra Credit:
In this extra credit problem, we wanted to observe the difference between males and females when they are alone and in groups. This is an interesting because it may reveal how people react differently to movies not just because of gender, but if differently gendered people act differently in a movie social situation. Our results are as follows: 0.004884091588092327 Alone males rate movies significantly differently from alone females 0.4436414283231227 Males in groups do not rate movies significantly differently females in groups 0.054261363758888896 Males who are alone do not rate movies significantly differently males in groups 0.3098783531995011 Females who are alone do not rate movies significantly differently females in groups Before we begin analysis, it is useful to note that previously, we found that 1/8th of movies are gendered, or a majority of movie ratings are not gendered. Looking at our first result, we can see that there is a statistically significant difference in the way alone males and alone females rate movies. This is surprising when compared to the initial result since we initially found that most movies don't have a gendered rating difference. However, when we look at our other results, we can see that when we add a groups factor, the significance disappears, which does line up with our findings from before which indicates that 2.5% of movies exhibit social

differences. In all, this indicates that movie ratings only show a gendered difference when watching alone, and that group watching of any kind removes any gender related differences that might be present in ratings.