Project 1 : Data Analysis

Q1 : Determining whether or not movies that are more popular are rated higher than movies that are less popular

- **Null Hypothesis (H0):** More popular movies do not have higher ratings than less popular movies (when popularity is operationalized as having more ratings).
- **Alternative Hypothesis (H1):** More popular movies have higher ratings than less popular movies (when popularity is operationalized as having more ratings).

I counted the total number of ratings that each movie in the data set received and did a median split over these results, then classified the movies into two categories accordingly: high popularity (those above or equal to the median) and low popularity (those less than the median). Since ratings are ordinal data, and since this question examines rank differences between two mutually exclusive groups (high/low popularity), I performed a one-tailed Mann-Whitney U test to determine whether or not more popular movies are rated higher than lower popularity movies. To do this, I made the assumptions that the data is similar in shape across the two groups., i.e. that the underlying distributions of the two groups are similar, and that there is independence within the groups, i.e. that there is no relationship between the observations in each group.

The Mann-Whitney U test yielded a p-value of approximately 0. At an alpha level of 0.005, since the p-value < 0.005, the test indicates that the positive difference in ratings of more popular and less popular movies is highly significant. Therefore, the null hypothesis is false; ratings of more popular movies are significantly higher than those of less popular movies.However, it is important to note that the assumptions made when performing this test in reality could be untrue. Confounding factors such as group overlap (same individuals rating both newer and older movies), media attention, or other external factors could influence results. Additionally, it is important to not the limitations to the scope of this result, as for this test, popularity of a movie has been operationalized as its rating count, when in reality, a movie could be popular but not have many ratings.

Q2 : Determining whether or not movies that are newer are rated differently than movies that are older

- **Null Hypothesis (H0):** There is no significant difference in the ratings of newer movies and the ratings of older movies.
- **Alternate Hypothesis (H1):** There is a significant difference in the ratings of newer movies and the ratings of older movies.

Extracting the year of each movie from its title, I implemented a median split over these years and classified the movies into two categories accordingly: newer movies (those above or equal to the median) and older movies (those less than the median). Since ratings are ordinal data, and since this question examines rank differences between two mutually exclusive groups (new/old movies), I performed a two-tailed Mann-Whitney U test to determine whether or not the newer movies are rated differently than the older movies. In doing this, I made the assumptions that the data is similar in shape across the two groups., i.e. that the underlying distributions of the two groups are similar, and that there is independence within the groups, i.e. that there is no relationship between the observations in each group.

The results of the Mann-Whitney U test yielded a p-value of approximately 0. At an alpha level of 0.005, since the p-value < 0.005, the test indicates that the difference in ratings of newer and older movies is highly significant. Therefore, the null hypothesis is false; there is a significant difference in the ratings of newer and older movies. However, it is important to note that the assumptions made when performing this test in reality could be untrue. Confounding factors such as group overlap (same individuals rating both newer and older movies), media attention, or other external factors could influence results.

Q3 : Determining whether or not enjoyment of ‘Shrek (2001)’ is gendered

- **Null Hypothesis (H0):** There is no significant difference in the enjoyment of ‘Shrek’ (2001) for female viewers and male viewers.
- **Alternate Hypothesis (H1):** There is a significant difference in the enjoyment of ‘Shrek’ (2001) for female viewers and male viewers.

Project 1 : Data Analysis

To preserve the sample size, I handled the missing data using element-wise removal of null values and performed a two-tailed Mann Whitney U test over the ratings of male viewers and female viewers for ‘Shrek’ (2001) to determine if a significant difference in ratings exists among male and female viewers of this movie. I chose this test because it is appropriate for determining differences in the rankings of two independent, mutually exclusive samples. For this test, I reasonably assumed that the data is similar in shape across the two groups., i.e. that the underlying distributions of the two groups are similar, and that there is independence within the groups, i.e. that there is no relationship between the observations in each group.
The results of the Mann-Whitney U test yielded a p-value of approximately 0.0505. At an alpha level of 0.005, since the p-value > 0.005, the test indicates that the difference in ratings between male and female viewers of ‘Shrek’ (2001) is not statistically significant. Therefore, the null hypothesis is true; there is no significant difference in the enjoyment of ‘Shrek’ (2001) for female viewers and male viewers. I thus conclude that enjoyment of ‘Shrek’ (2001) is not gendered.

Q4 : Determining the proportion of movies that are rated differently by male and female viewers

- **Null Hypothesis (H0):** There is no difference in the ratings of movies by male and female viewers.
- **Alternative Hypothesis (H1):** There is a difference in the ratings of movies by male and female viewers.

I handled the missing data using element-wise removal of null values and performed a two-tailed Mann Whitney U test over the ratings of viewers in each group (male vs female viewers). For this test, I reasonably assumed that the data is similar in shape across the two groups., i.e. that the underlying distributions of the two groups are similar, and that there is independence within the groups, i.e. that there is no relationship between the observations in each group. From the results, I computed the proportion of movies showing a significant difference in ratings when the two groups were compared (from Mann Whitney U test) and divided this number by the total number of movies evaluated. I chose to handle the NaN values element-wise in order to preserve the sample size, and chose the Mann Whitney U test because this data is ordinal and we are interested in the difference in ratings between two mutually exclusive groups.
Out of the 400 movies evaluated, 50 of these movies had significant p-values at an alpha level of 0.005 (i.e. p-value < 0.005), meaning that male viewers and female viewers rated 50 out of 400 of these movies significantly differently. From these results, I determine that the null hypothesis is false and accept the alternative hypothesis, stating that there is a difference in the ratings of movies by male and female viewers. 12.5% of movies are rated differently by male and female viewers according to this test.

Q5 : Determining whether or not people who are only children enjoy ‘The Lion King (1994)’ more than people with siblings

- **Null Hypothesis (H0):** People who are only children do not enjoy ‘The Lion King’ (1994) more than those who are not only children.
- **Alternative Hypothesis (H1):** People who are only children enjoy ‘The Lion King’ (1994) more than those who are not only children.

To preserve the sample size, I handled the missing data by removing NaN values in the rows of ratings for each group (alone and social viewers). I then performed a one-tailed Mann-Whitney U test to compare the ratings between the two groups and determine if only children rated the movie higher. I chose this test because it is appropriate for determining if a positive difference exists in the rankings of two independent, mutually exclusive samples, particularly when the data does not follow a normal distribution. In doing so, I reasonably assumed that the data is similar in shape across the two groups., i.e. that the underlying distributions of the two groups are similar, and that there is independence within the groups, i.e. that there is no relationship between the observations in each group.
The results of the Mann-Whitney U test yielded a p-value of approximately 0.9784. At an alpha level of 0.005, since the p-value > 0.005, the test indicates that the difference in ratings between only children and not only children is not statistically significant. Therefore, the null hypothesis is true; there is no significant difference in enjoyment of

‘The Lion King’ (1994) between viewers who are only children and viewers who have siblings. I thus conclude that only children do not enjoy ‘The Lion King’(1994) more than viewers with siblings.

Q6 : Determining the proportion of movies that exhibit an “only child effect”

- **Null Hypothesis (H0):** There is no difference in the ratings of movies for viewers who are an only child and viewers who are not an only child.
- **Alternative Hypothesis (H1):** There is a difference in the ratings of movies for viewers who are an only child and viewers who are not an only child.

I handled the missing data using element-wise removal of null values and performed a two-tailed Mann Whitney U test over the ratings of viewers in each group (only child vs not only child). In doing so, I reasonably assumed that the data is similar in shape across the two groups., i.e. that the underlying distributions of the two groups are similar, and that there is independence within the groups, i.e. that there is no relationship between the observations in each group.

From the results, I computed the proportion of movies exhibiting an ‘only child effect’, taking the number of movies whose ratings were significantly different when the two groups were compared (from Mann Whitney U test) and dividing this number by the total number of movies evaluated. I chose to handle the NaN values element-wise in order to preserve the sample size, and chose the Mann Whitney U test because this data is ordinal and we are interested in the difference in ratings between two mutually exclusive groups.

Out of the 400 movies evaluated, of 7 of these movies had significant p-values at an alpha level of 0.005 (i.e. p-value < 0.005), meaning that 7 out of 400 of these movies were shown to have significant differences in the ratings of viewers who are an only child and viewers who are not an only child. From these results, I determine that the null hypothesis is false and accept the alternative hypothesis for the ‘only child effect’. 1.75% of movies exhibit this ‘only child effect’ according to these results.

Q7 : Determining whether or not people who like to watch movies socially enjoy ‘The Wolf of Wall Street (2013)’ more than those who prefer to watch them alone

- **Null Hypothesis (H0):** People who prefer to watch movies socially do not enjoy ‘The Wolf of Wall Street’ (2013) more than those who prefer to watch alone.
- **Alternative Hypothesis (H1):** People who prefer to watch movies socially enjoy ‘The Wolf of Wall Street’ (2013) more than those who prefer to watch alone.

To preserve the sample size, I handled the missing data by removing NaN values in the rows of ratings for each group (alone and social viewers). I then performed a one-tailed Mann-Whitney U test to compare the ratings between the two groups and determine if social viewers rated the movie higher. I chose this test because it is appropriate for determining if a positive difference exists in the rankings of two independent, mutually exclusive samples, particularly when the data does not follow a normal distribution. For this test, I reasonably assumed that the data is similar in shape across the two groups., i.e. that the underlying distributions of the two groups are similar, and that there is independence within the groups, i.e. that there is no relationship between the observations in each group.

The results of the Mann-Whitney U test yielded a p-value of approximately 0.9437. At an alpha level of 0.005, since the p-value > 0.005, the test indicates that the difference in ratings between social viewers and non-social viewers is not statistically significant. Therefore, the null hypothesis is true: there is no significant difference in enjoyment of ‘The Wolf of Wall Street’ (2013) among viewers who prefer social watching and those who prefer to watch alone. I thus conclude that those who prefer to watch movies socially do not enjoy ‘The Wolf of Wall Street’ (2013) more than those who prefer to watch alone.

Q8 : Determining the proportion of movies that exhibit a “social watching” effect

- **Null Hypothesis (H0):** There is no difference in the ratings of movies for viewers watching socially and viewers watching alone.

- **Alternative Hypothesis (H1):** There is a difference in the ratings of movies for viewers watching socially and viewers watching alone.

I handled the missing data using element-wise removal of null values and performed a two-tailed Mann Whitney U test over the ratings of viewers in each group (alone vs social watching preference). For this test, I reasonably assumed that the data is similar in shape across the two groups., i.e. that the underlying distributions of the two groups are similar, and that there is independence within the groups, i.e. that there is no relationship between the observations in each group. From the results, I computed the proportion of movies exhibiting a 'social watching' effect, taking the number of movies whose ratings were significantly different (from Mann Whitney U test) between the two groups (alone vs socially), and dividing by the total number of movies evaluated. I chose to handle the NaN values element-wise in order to preserve the sample size, and the Mann Whitney U test since this data is ordinal and because we are interested in the difference in ratings between two mutually exclusive groups.
Out of the 400 movies evaluated, 10 of these movies had significant p-values at an alpha level of 0.005 (i.e. p-value < 0.005), meaning that 10 out of 400 of these movies were shown to have significant differences in the ratings of viewers who watch alone and viewers who watch socially. From these results, I determine that the null hypothesis is false and accept the alternative hypothesis for the 'social watching effect'. 2.5% of movies exhibit this 'social watching' effect, according to these results.

Q9 : Determining whether or not the ratings distribution of 'Home Alone (1990)' is different from that of 'Finding Nemo (2003)'.

- **Null Hypothesis (H0):** The ratings distribution of 'Home Alone' (1990) is not different from the ratings distribution of 'Finding Nemo (2003)'.
- **Alternative Hypothesis (H1):** The ratings distribution of 'Home Alone' (1990) is different from the ratings distribution of 'Finding Nemo (2003)'.

I handled missing data using row-wise removal of the NaN values, then I performed a two-tailed KS test for the two samples of ratings. I used row-wise removal to handle the missing data so that the test would consider the ratings for individuals who have seen both movies and avoid survivorship bias. I chose to use the KS test for significance because we are interested in comparing the underlying distribution of ratings for the two movies, which is captured by the KS test. To perform this test, I assumed that a specific individual's ratings for each of the two movies are independent.
From the KS test, I obtained a p-value of approximately 0. At an alpha level of 0.005, my results show that the null hypothesis is false. Given that the p-value < 0.005, I therefore accept the Alternate Hypothesis (H1) and conclude that the ratings distribution of 'Home Alone' (1990) is significantly different from the ratings distribution of 'Finding Nemo (2003)'.

Q10 : Determining how many of the franchises included in the data are of inconsistent quality

- **Null Hypothesis (H0):** Movies within a franchise are of consistent quality.
- **Alternative Hypothesis (H1):** Movies within a franchise are of inconsistent quality.

I first handled the missing data using row-wise removal of NaN values, then I performed a Kruskal Wallis test over the movies for each franchise to determine any significant difference in ratings across the movies of a given set (franchise). I assumed that the samples being compared (i.e. movie ratings within a franchise) are independent to perform this test. I chose the Kruskal Wallis test because determining consistency of quality within a franchise involves comparing ordinal data of more than 2 groups, with groups being the number of movies included in the dataset for each franchise. I handled the missing values using row-wise removal in order to reduce bias that could arise from users who only rated one/some of the movies within a franchise, and avoid skewing the results. Given that this analysis intends to show consistency or lack thereof of ratings for movies within a franchise, removing entire rows having NaN values ensures that the trend of quality as experienced by each individual viewer is accurately captured and preserved.
My results showed that, at an alpha level of 0.005, the null hypothesis held true only for the 'Harry Potter' (p-value of approximately 0.1179) and 'Pirates of the Caribbean' (p-value of approximately 0.0358) franchises. P-values for

all other franchises tested were less than 0.005 (p-values were approximately 0). In context, this result suggests that the median of the movie rating for ‘Harry Potter’ movies and ‘Pirates of the Caribbean’ movies is significantly consistent within the respective franchise, whereas movies with the ‘Star Wars’, ‘The Matrix’ , ‘Indiana Jones’ , ‘Jurassic Park’ , ‘Toy Story’ , and ‘Batman’ franchises show a significantly different median across movies within their respective franchise. Therefore, from these ratings of individuals who have rated every movie (included in the dataset) for a particular franchise, ‘Harry Potter’ movies and ‘Pirates of the Caribbean’ movies are of consistent quality, whereas the respective movies of all other franchises tested are of inconsistent quality. However, because it is probable that a person who enjoys one movie will likely enjoy another movie within that same franchise, it is important to note that the independence assumption made when performing this test in reality is likely untrue.

Extra Credit: Determining if there is a significant difference in ratings of ‘Grease’ (1978) between shy viewers and outgoing viewers

- **Null Hypothesis (H0):** There is no difference in the ratings of ‘Grease’ (1978) by shy viewers and outgoing viewers.
- **Alternate Hypothesis (H1):** There is a difference in the ratings of ‘Grease’ (1978) by shy viewers and outgoing viewers.

I created two new data frames of ratings for ‘Grease’ merged with viewers who identify as shy and those who identify as outgoing. I handled the missing data by performing row-wise removal of the NaN values so that the test would consider the ratings for individuals who have seen the movie. I chose this test because it is appropriate for determining if a positive difference exists in the rankings of two independent, mutually exclusive samples. I reasonably assumed that the data is similar in shape across the two groups and that there is independence within the groups.

The Mann Whitney U test returned a p-value of approximately 1, indicating that the difference in ratings by shy and outgoing individuals for ‘Grease’ is highly insignificant. The null hypothesis holds true; I conclude therefore that there is no significant difference in the ratings of ‘Grease’ (1978) by shy viewers and outgoing viewers.