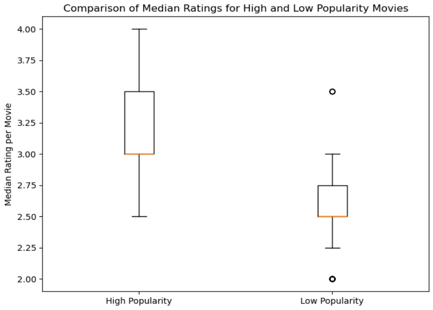
An Analysis of Movie Ratings

Introduction:

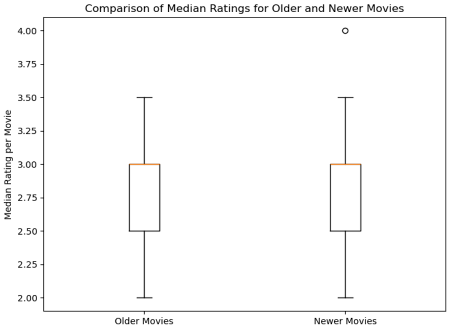
In order to optimize operations, I was tasked with analyzing a dataset consisting of user ratings for 400 different movies. I will use various statistical tests to answer specific questions that will help our company answer ten important questions. I will be using an α level of 0.005 in my research. Each question will be answered in a DYFA format (Do/Why/Find/Answer).

Questions:

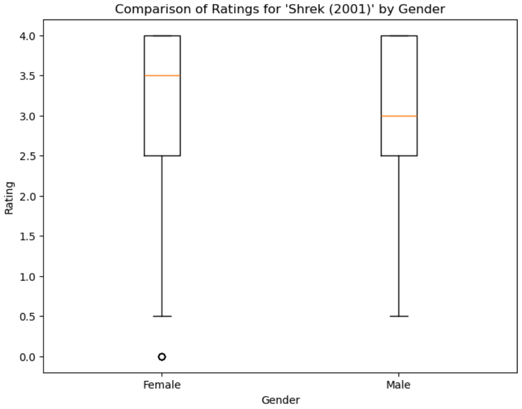
1: Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular?
D: I divided the movies into a high popularity and low popularity group while dropping nulls. I then reduced each individual movie to its median rating in both groups. I ran a Mann Whitney U test to calculate the p-value and then calculated a Cliff's Delta effect size to determine how practical the p-value score was. I assumed each rating was independent and that the data was ordinal in nature. Also, I assumed the data was non-normally distributed.
Y: Due to the nature of the question, we were going to have a high class imbalance. As a result, I reduced each row to its median to normalize both classes. I chose the Mann Whitney U test since I was examining the difference in central tendency between each class. I also used a Cliff Delta score to quantify how practically significant the p-value was.
F: I found a p-value of 1.9858517703414465e-34 and a Cliff's Delta effect score of 0.671375.
A: My findings indicate that my results are statistically significant and have a large effect size. The boxplot above strongly aligns with my statistical findings as the distributions between each group is very different. By reducing each movie to its median, I removed much of the data from my test. My worry was that the class imbalance was so large it would skew my results. Specifically, when I ran the same significance test without normalizing, my p-value was exactly zero.



2: Are movies that are newer rated differently than movies that are older?
D: I divided the data into two groups which were split at the median of the release years. I checked to make sure that each group had similar null proportions, then I dropped the null values. I then ran Mann Whitney U tests on non-normalized and normalized data. I also calculated a Cliff's Delta effect score. I assumed each rating was independent, the data was ordinal, and the distribution was non-normal.
Y: I used the Mann Whitney U test since I was comparing the central tendency of each group again. I ran a normalized test since the results from the non normalized test weren't consistent with the distribution of the data. I calculated a Cliff's Delta score to add an effect size score to the analysis.
F: For my non normalized test, my p-value was 0.00211 and my effect size was -0.01044. For the normalized test, my p-value was 0.19865 and the effect size was 0.07115. To the right is the distribution of for each group. Their distributions are very similar which aligns with the normalized test results.
A: For my normalized distribution, given my p-value of 0.00211, I conclude that it's statistically significant. While it's Cliff's Delta score of -0.01044 means the effect is negligible. For the normalized test, given the p-value is 0.19865, we can conclude that the test was not statistically significant, and the Cliff's Delta score of .07115 indicates that the effect is negligible. My main concern is that I did not need to reduce each movie to its median rating as the Mann Whitney test can handle non-normal distributions. However, since the ratings distribution of new vs old movies is so similar, I feel confident that the normalized test is better suited for the data.

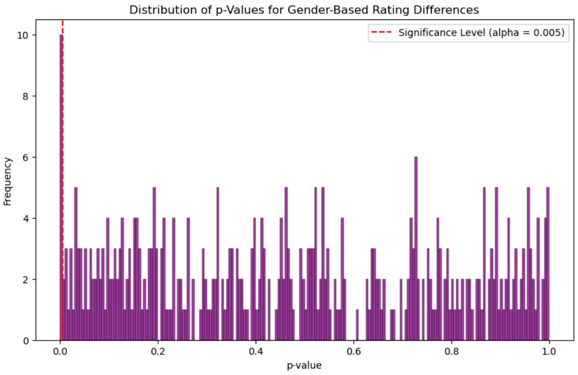


3: Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?
D: I broke up the data into two groups that compare Shrek ratings by gender. I dropped the nulls in an element based fashion where I assumed that each rating was independent given the gender. I then ran a Mann Whitney test for the p-value and calculated a Cliff's Delta score for the effect size. I assumed the data was non-normal and ordinal.
Y: I used a Mann Whitney U test because I was concerned with comparing the central tendencies of each group. Also, the Mann Whitney U test can handle the high class imbalance. Even with the class imbalance, each group had similar null values so I felt comfortable with dropping the nulls. The



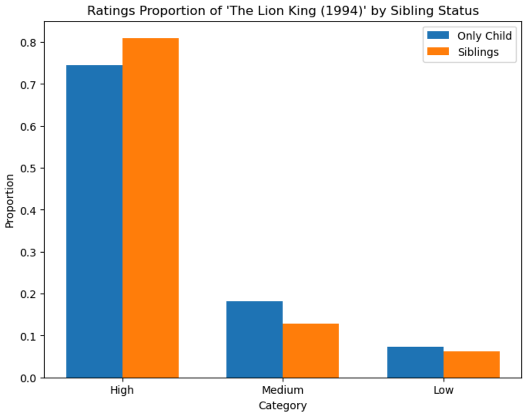
Cliff’s Delta effect score would allow me to add context to any p-value my significance test calculates.
F: My p-value was 0.0505 and my Cliff’s Delta score was 0.0815. These values are consistent with the boxplot above. The data distributions have some differences, but overall they aren’t different enough for statistical significance.
A: Given that my p-value was 0.0505, we can conclude that the test results are not statistically significant. Meaning, the different genders don’t rate the movie differently. This makes sense when we consider how the effect size is negligible as it is 0.0815. One limitation in the test is that the sample size of the female viewers is considerably larger than the sample size of the male viewers.

4: What proportion of movies are rated differently by male and female viewers?
D: I also used a Mann Whitney U test for this question. I iterated through all 400 movie columns. If the proportion of nulls for each gender were within 10% of each other, I dropped the nulls from the column. If the proportion of nulls between genders were greater than 10%, I skipped the movie in general. I then calculated the proportion of movies that had statistical differences. My approach assumed that each movie was independent from each other. Also, the ratings for each movie were independent of the other ratings. I treated the data as ordinal and non-normal.
Y: I followed a very similar process as I did for question 3. I used the Mann Whitney U test because I wanted to compare the central tendency for each movie. I knew the test could handle potential class imbalances. I had to set a null threshold because it was impractical to analyze each movie’s nulls individually.
F: After filtering, I was left with 243 movies and 20 of them had p-values that were statistically significant. Specifically, 8.23% of the movies had statistically significant differences in ratings between genders. The chart to the right shows the distribution of all the p-values for the movies. We can see the 20 movies with statistically significant values.



A: Based on the findings, only 8.23% of the movies are rated differently by gender. Based on my approach where I drop nulls in movies that have large differences in nulls by gender, I end up excluding a large number of movies from the analysis. This is a limitation in my ability to give a complete representation of all the movies. However, dropping nulls from every movie regardless of null proportion could lead to my data being skewed.

5: Do people who are only children enjoy ‘The Lion King (1994)’ more than people with siblings?
D: I used a Chi-Square test to answer this question. I grouped user ratings into low, medium and high groups. I then broke the data out into if the viewer was an only child or not. I calculated a Cramér’s V to determine the effect size as well.
Y: I treated the data as categorical in this question. As a result, using a Chi-Square test was the appropriate statistical test to use in this question. Chi-Square tests analyze the differences in frequencies between categories. The Cramér’s V score will add an effect size score to add more context.
F: My p-value was 0.229 and my Cramér’s V score is 0.062. Referencing the chart to the right, we can see that the proportions of ratings by category are very similar across sibling status. My resulting contingency table was 3 rows by 2 columns and the degrees of freedom for my contingency table was 2.

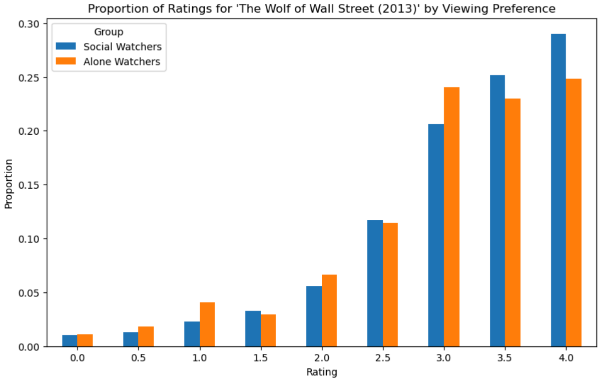


A: Given that my p-value is 0.229, we can conclude that the test results are not statistically significant. Also, the effect size being .062 indicates that there is a very weak association between each sibling status group. A potential limitation to my result is that the Chi-Square test isn’t as well suited for imbalanced data as Mann Whitney U tests. The approach I’m using might not be as effective as using a Mann Whitney test, but since my counts are high enough, I feel confident in my approach. People who are only children don’t rate “The Lion King” differently than people 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: I used a Chi-Square test to analyze statistical significance in this question. I grouped the ratings into low, medium and high categories. I checked if null proportions in each movie are within 0.1 of each other based on sibling status. If they were, I dropped the nulls from the column. If the null proportions were past that threshold, I skipped over those movies from the analysis. I also made sure that each cell in the contingency table had an expected count higher than 5. Also, I allowed for one cell to be between 1 and 5 values per table. I then calculated the overall proportion of significant differences in ratings based on sibling status. Also, I treat each movie as independent in the analysis. I assumed the data is non-normal and categorical.
Y: Like in question 5, I treated the data as categorical, so Chi-Square tests were appropriate. I dropped nulls in order to shield the analysis from highly skewed movies. Also, for Chi-Square tests to be most effective, each cell in the contingency table should

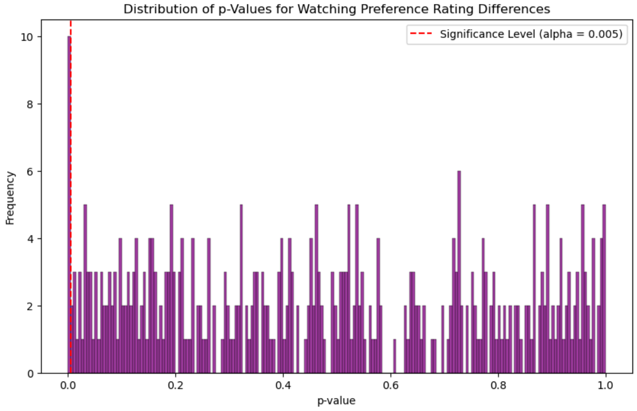
have over 5 values per cell. Also, research has indicated that if df > 1, it is permissible for 20% of cells to have a value between 1 and 5 (Agresti, 177). My contingency tables had a df of 2.
F: After filtering, I had 363 movies with 6 movies showing statistically significant differences for a percentage of 1.65%
A: My assumption to use a Chi-Square test instead of the Mann Whitney U test could have led to suboptimal results. Some classes had data imbalances, and the Mann Whitney U test is better suited for that. I wanted to approach these last two problems differently than the questions I've already tackled. In conclusion, most movies do not have an "only child effect".

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: I used a Mann Whitney U test to analyze statistical differences for this question. I calculated null proportions between social watchers and individual watchers based on 'The Wolf of Wall Street'. When I saw they had very comparable proportions, I dropped the nulls. I then calculated a Cliff's Delta effect size score. I assumed the data is non-normal and ordinal.
Y: I chose a Man Whitney U test since the data was slightly imbalanced (393 vs 270 values). I dropped the nulls because since both viewer groups had a similar proportion, I felt dropping the nulls didn't introduce additional bias. I wanted to add an effect score to provide additional context to our significance score.
F: My p-value score was 0.113 and my Cliff's Delta score was 0.071. As you can see in the chart to the right, the distribution between each viewer group is slightly different, but mostly consistent. This aligns with the test results we found.
A: Given that my p-value is 0.113, we can conclude that the test is not statistically significant. Also, since the Cliff's Delta effect size value is 0.071, the effect size is negligible. Therefore, there is not a difference in how social vs alone watchers rate "The Wolf of Wall Street".

8: What proportion of movies exhibit such a "social watching" effect?
D: I ran a Mann Whitney U test for this question. I iterated through all 400 rows of movies. I dropped nulls in movies that had null proportions within 0.1 of each other based on their social watching status. I skipped over the movie entirely if the null proportions fell outside that range. I then calculated the proportion of movies that had statistically significant results. I assumed each movie and its ratings are independent from each other in the data. I also assumed that the data is non-normal and ordinal.

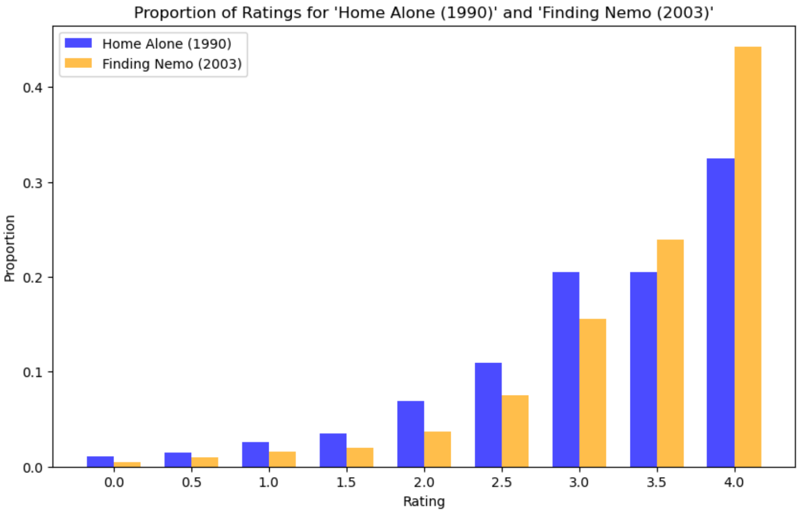


Y: I chose a Mann Whitney U test since each movie had a level of imbalance in their data. I wanted to compare the central tendency of each movie's ratings based on if they were a social watcher or not. I couldn't analyze each movie's nulls individually, so I set the null threshold.
F: After dealing with nulls, I was left with 399 movies. After the analysis, 10 total movies were found to be statistically significant for a percentage of 2.51%. To the right we can see a histogram of the different p-values for each movie. We can see the 10 movies with statistically significant results.
A: Based on the results, only a very small portion of the data is statistically significant. However, according to the chart, the alpha level of < 0.005 is still the largest bar in the histogram. A very small number of movies exhibit a "social watching" effect.

9: Is the ratings distribution of 'Home Alone (1990)' different than that of 'Finding Nemo (2003)?
D: I used a Kolmogorov-Smirnov test to analyze this question. I calculated the null proportions of the ratings for "Home Alone" and "Finding Nemo". The percentages were considerably different, so I made sure to analyze the behavioral columns to ensure there was no bias introduced by dropping them. I ended up finding no pattern, so I dropped the nulls.
Y: I chose the Kolmogorov-Smirnov test to analyze this question because the question was asking about distribution differences. I dropped the nulls after finding out there were no behavioral columns that produced extra bias. I calculated the proportions of missing values for each behavioral column based on each movie. Once I determined that the missingness was consistent in pattern for each movie, I felt comfortable dropping the columns.
F: My p-value was 0.00000000064 and my Cliff's Delta value was -0.9674. As you can see in the chart, the proportions of values are considerably different. Home Alone has consistently lower ratings. The histogram matches my findings very well. My

effect size score indicates that Home Alone's ratings are usually smaller than Finding Nemo scores, and that's backed up by the histogram.

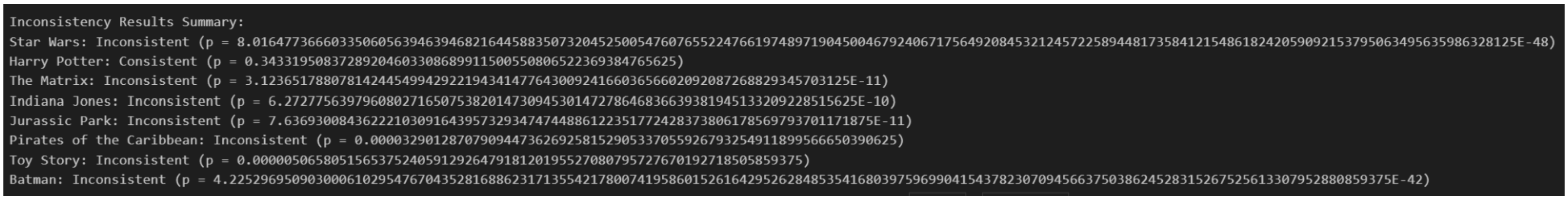
A: Given the p-value is < 0.005, we can conclude that the test results are statistically significant. Also, the effect size is very strong according to the Cliff's Delta value. I didn't want to impute values since the data in question are user ratings. My assumption on dropping nulls as long as they have similar proportions of missingness based on the behavioral columns could be problematic. My idea might run into problems if there are dependencies between the behavioral columns.



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: I used a Kruskal-Wallis test to analyze this question. I split the data up by movie franchise. I ensured each movie had enough ratings for the test to run, and then I ran the test. I assume that the ratings for each franchise are independent of the other franchises. I also assume the data is non-normal, and ordinal. I also added an epsilon squared effect size score.

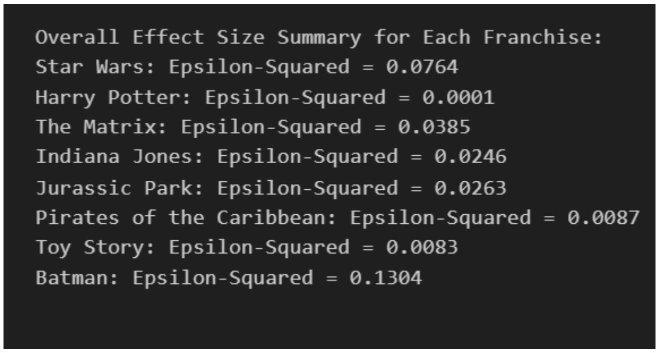
Y: I chose to use a Kruskal-Wallis test because the data is ordinal and each movie franchise has at least three individual movies. I ensured that each movie had enough data to run the test so that I didn't add unnecessary bias. I wanted to add an effect size score in order to add an extra layer of understanding.



```
Inconsistency Results Summary:
Star Wars: Inconsistent (p = 8.01647736660335060563946394682164458835073204525005476076552247661974897190450046792406717564920845321245722589448173584121548618242059092153795063495635986328125E-48)
Harry Potter: Consistent (p = 0.34331950837289204603308689911500550806522369384765625)
The Matrix: Inconsistent (p = 3.1236517880781424454994292219434147764300924166036566020920872688293457031 25E-11)
Indiana Jones: Inconsistent (p = 6.2727756397960802716507538201473094530147278646836639381945133209228515625E-10)
Jurassic Park: Inconsistent (p = 7.636930084362221030916439573293474744886122351772428373806178569793701171875E-11)
Pirates of the Caribbean: Inconsistent (p = 0.0000329012870790944736269258152905337055926793254911899566650390625)
Toy Story: Inconsistent (p = 0.0000050658051565375240591292647918120195527080795727670192718505859375)
Batman: Inconsistent (p = 4.2252969509030006102954767043528168862317135542178007419586015261642952628485354168039759699041543782307094566375038624528315267525613307952880859375E-42)
```

F: Every movie franchise except for Harry Potter showed statistically significant differences. Above are the resulting p-values. We can see that every franchise except Harry Potter has an extremely small p-value. Attached at the end of the report are the histograms of all the ratings. We can see how they all have inconsistent ratings except for Harry Potter. Additionally, Each franchise except Harry Potter shows some type of effect with Batman showing the largest effect. According to the histogram, this makes sense as “Batman: The Dark Knight” is largely different in its ratings compared to the other movies in the franchise.

A: In conclusion, only Harry Potter as a movie franchise has consistent ratings across its movies. One downside in the analysis is that not every movie in the franchises are represented in the dataset. Also, the assumption that the ratings for each movie franchise might not actually be independent. Many of these movie franchises likely have overlapping fandoms.



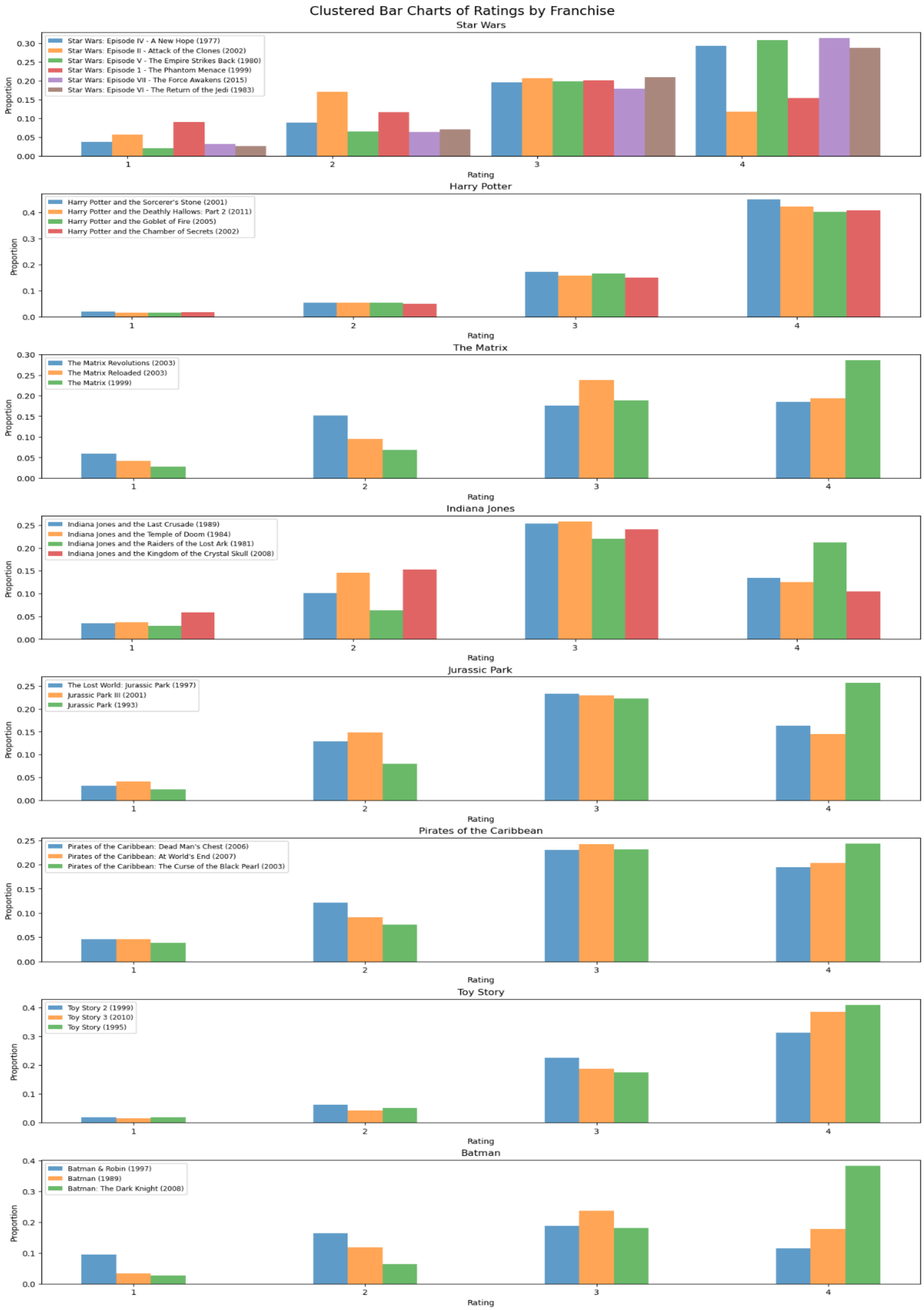
```
Overall Effect Size Summary for Each Franchise:
Star Wars: Epsilon-Squared = 0.0764
Harry Potter: Epsilon-Squared = 0.0001
The Matrix: Epsilon-Squared = 0.0385
Indiana Jones: Epsilon-Squared = 0.0246
Jurassic Park: Epsilon-Squared = 0.0263
Pirates of the Caribbean: Epsilon-Squared = 0.0087
Toy Story: Epsilon-Squared = 0.0083
Batman: Epsilon-Squared = 0.1304
```

Appendix:

Citation:

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). John Wiley & Sons.

Question 10 Histogram:


Clustered Bar Charts of Ratings by Franchise
Star Wars
Star Wars: Episode IV - A New Hope (1977)
Star Wars: Episode II - Attack of the Clones (2002)
Star Wars: Episode V - The Empire Strikes Back (1980)
Star Wars: Episode 1 - The Phantom Menace (1999)
Star Wars: Episode VII - The Force Awakens (2015)
Star Wars: Episode VI - The Return of the Jedi (1983)
Proportion
Rating
Harry Potter
Harry Potter and the Sorcerer's Stone (2001)
Harry Potter and the Deathly Hallows: Part 2 (2011)
Harry Potter and the Goblet of Fire (2005)
Harry Potter and the Chamber of Secrets (2002)
Proportion
Rating
The Matrix
The Matrix Revolutions (2003)
The Matrix Reloaded (2003)
The Matrix (1999)
Proportion
Rating
Indiana Jones
Indiana Jones and the Last Crusade (1989)
Indiana Jones and the Temple of Doom (1984)
Indiana Jones and the Raiders of the Lost Ark (1981)
Indiana Jones and the Kingdom of the Crystal Skull (2008)
Proportion
Rating
Jurassic Park
The Lost World: Jurassic Park (1997)
Jurassic Park III (2001)
Jurassic Park (1993)
Proportion
Rating
Pirates of the Caribbean
Pirates of the Caribbean: Dead Man's Chest (2006)
Pirates of the Caribbean: At World's End (2007)
Pirates of the Caribbean: The Curse of the Black Pearl (2003)
Proportion
Rating
Toy Story
Toy Story 2 (1999)
Toy Story 3 (2010)
Toy Story (1995)
Proportion
Rating
Batman
Batman & Robin (1997)
Batman (1989)
Batman: The Dark Knight (2008)
Proportion
Rating