May the Data Be with You: A Star Wars Data Analysis

Neal E. Gragg

ITCS-5156 Spring 2023 – Desai

May 4th, 2023

Abstract

This report is presented as a survey of a previous work [Hic14]. Any assertions made within are subjective and do not represent those of the original author.

This paper presents a data analysis of a Star Wars survey, which explores the opinions and attitudes of respondents towards various aspects of the Star Wars franchise. The survey included questions about the characters, the movies, and the Expanded Universe, as well as questions about respondents' demographics and their opinions about the Star Trek franchise. The data was analyzed using various statistical techniques, including descriptive statistics, data visualization, and hypothesis testing. The results of the analysis provide insights into the opinions and attitudes of Star Wars fans towards different aspects of the franchise and highlight some interesting trends and patterns. Overall, this paper contributes to the understanding of the Star Wars fan community and provides useful information for future research and marketing efforts related to the franchise.

1 Introduction

In 2014, Walt Hickey conducted a Star Wars survey with 1186 respondents and published an article titled "America's Favorite Star Wars Movies (and Least Favorite Characters)" [Hic14]. Hickey's survey aimed to uncover the viewership patterns, movie rankings, character preferences, and viewers' responses to the controversial "who shot first?" question. This paper aims to replicate Hickey's original analysis but also delve deeper into the data to answer additional questions. Specifically, this paper will examine the Expanded Universe's popularity among Star Wars fans and the crossover between Star Wars and Star Trek fandoms. To achieve these objectives, this paper will employ exploratory data analysis techniques, including visualization and statistical analysis. The paper's findings will provide valuable insights into the Star Wars fandom and its relationship with the franchise.

1.1 Problem and Motivation

The Star Wars franchise is one of the most popular and beloved movie franchises in cinematic history, with millions of fans across the globe. The franchise has produced numerous films, TV shows, books, and other media, making it a cultural phenomenon that has impacted generations.

Walt Hickey's 2014 article, "America's Favorite Star Wars Movies (and Least Favorite Characters)," is a seminal work in the field of Star Wars data analysis. However, there is a need to revisit Hickey's work and expand upon it, particularly in light of new data analysis methods and related works. Therefore, the motivation behind duplicating and improving upon Hickey's analysis is to provide an updated and more comprehensive understanding of fans' opinions about the Star Wars franchise. By analyzing the data with new methods, this paper aims to explore the fans' opinions and preferences regarding the franchise's movies, characters, and expanded universe, providing an updated and insightful analysis of the Star Wars fandom.

2 Related Works

2.1 StatCrunch – Star Wars Fan Survey

The related work titled "Star Wars Fan Survey" [Sta19] by a group of students from Georgia College explored the relationship between the age at which someone first watched Star Wars and their level of fandom. This study, which used a six-part ranking system to determine an individual's level of fandom, surveyed students at Georgia College and used age, gender, major, favorite movie, favorite character, and science fiction reading/watching habits as demographic variables. While this work does not directly address the same research questions as Hickey's article, it is related in that it investigates factors that may influence a person's level of Star Wars fandom. The Georgia College study focuses on the influence of age, while Hickey's article explores the popularity of Star Wars movies and characters among a broader population. Therefore, these two studies complement each other and provide a more comprehensive understanding of the factors that influence Star Wars fandom.

In Figure 1, StatCrunch provides a pie chart showing the results of their survey that focused on the respondents' favorite Star Wars movie. As opposed to the variation in Hickey's article, this chart shows a more even distribution in favorite movies.

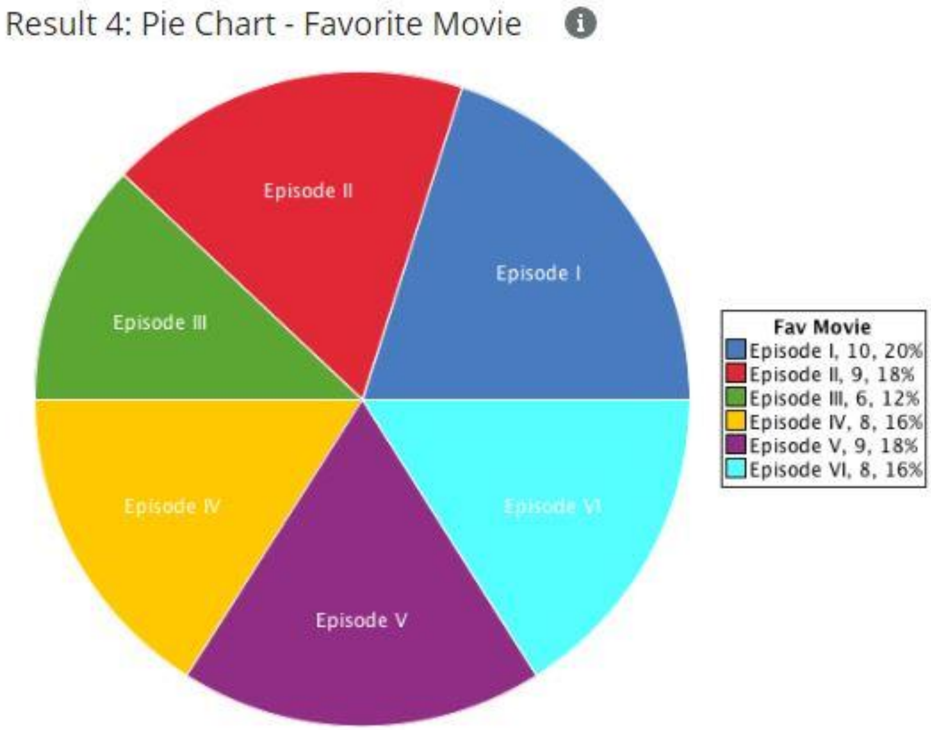

Figure 1: This image shows a pie chart of favorite Star Wars movies [Sta19]

In the next figure, Figure 2, we can see the results of StatCrunch's survey on favorite Star Wars characters. In comparison to Hickey's article, the results are more uniform for most of the characters and certain characters, like Jar Jar Binks and R2-D2, have a much lower favorability rating than the other characters.

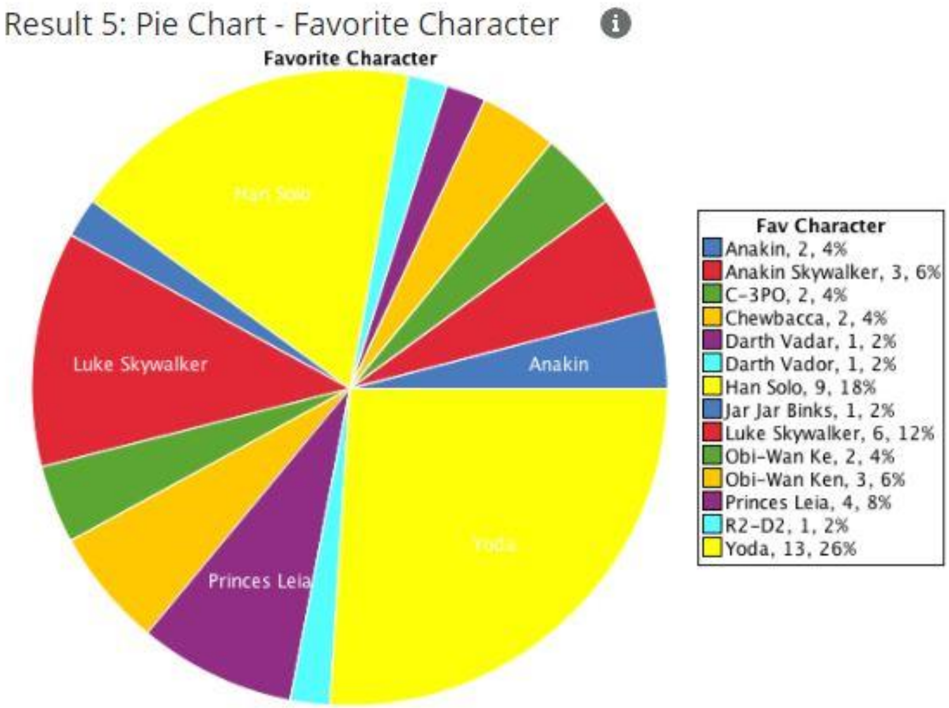

Figure 2: This image shows a pie chart of Star Wars character favorability [Sta19]

2.2 Jedi Temple Archives – Another Star Wars Fan Survey

The article from the Jedi Temple Archives titled " The Results of the Star Wars Fan Survey Are In!" [Tho19] is related to Hickey's article as both articles analyze Star Wars fan responses from surveys. However, the approach and focus of the surveys differ. The survey in the Jedi Temple Archives article focused on analyzing the general make-up and demographics of Star Wars fandom, while Hickey's survey focused on analyzing the respondents' preferences for Star Wars movies, characters, and certain scenes. Additionally, the Jedi Temple Archives article utilized snowball sampling, which may result in a biased sample of Star Wars fans, whereas Hickey's survey was not explicitly described to have used this method. Despite the differences in approach, both articles provide insights into the opinions and behaviors of Star Wars fans.

2.3 In Relation

Both the StatCrunch article and the Jedi Temple Archives article are related to Hickey's article as they are all based on Star Wars surveys. While the StatCrunch article focused on the correlation between age and fandom level, and the Jedi Temple Archives article aimed to reveal insights about the Star Wars fandom and their views on various topics, Hickey's article analyzed the respondents' preferences on the Star Wars movies, characters, and their stance on certain debates. This data analysis project was inspired by Hickey's article, as well as their articles, as we aimed to expand on his analysis and explore new insights from the same data set. This paper seeks to answer new questions and uncover new trends that were not examined in Hickey's article.

3 Methods of the Original Analysis

3.1 Analysis Introduction – Films Seen and Ranked

Based on the available information about Walt Hickey's methods in analyzing the Star Wars survey data, it is likely that he used various algorithms, statistical methods, and visualization techniques to derive insights from the survey results. Hickey's article mentions that he analyzed which films the respondents had seen, how they ranked them, how they favored characters, and their response to "who shot first?". It is possible that he used machine learning algorithms such as clustering or decision trees to identify patterns in the responses and determine which factors were most strongly correlated with certain preferences or opinions. Below is how Hickey graphed his results for the data analysis regarding films seen and how they are ranked.

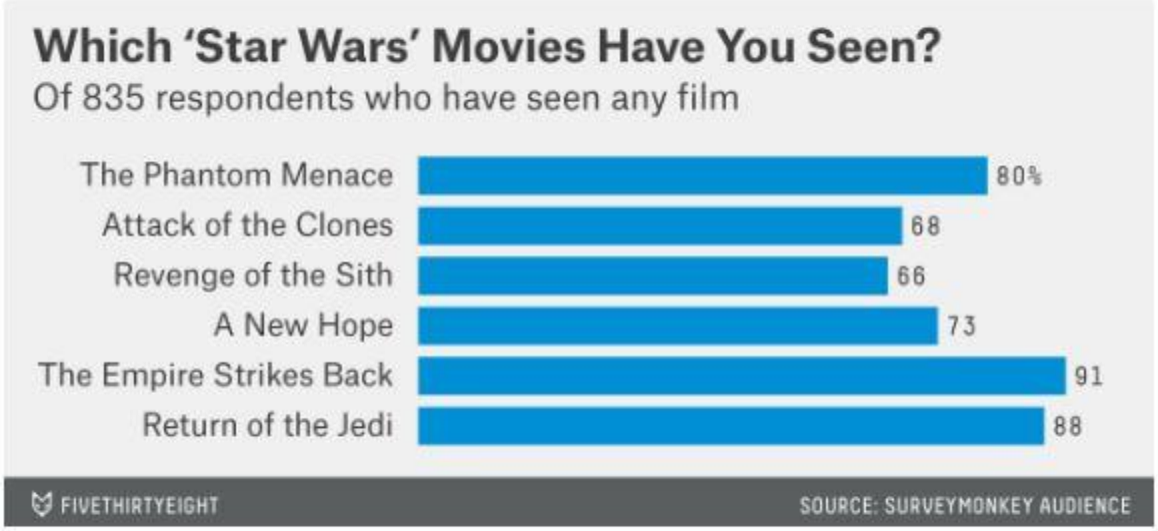

Figure 3: Graph showing the percentage of respondents who saw each film [Hic14]

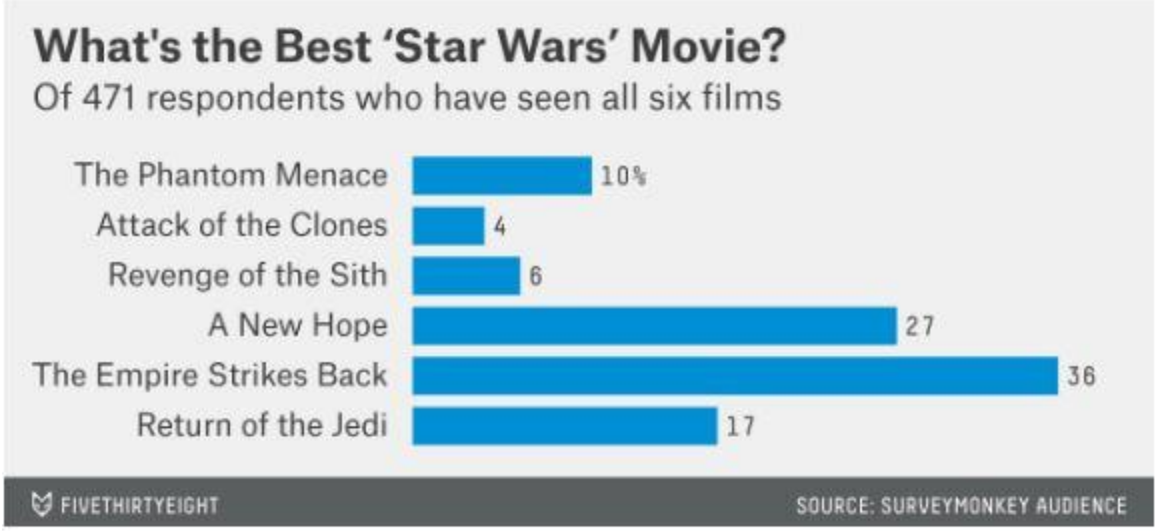

Figure 4: Graph showing how respondents who saw all 6 films ranked them [Hic14]

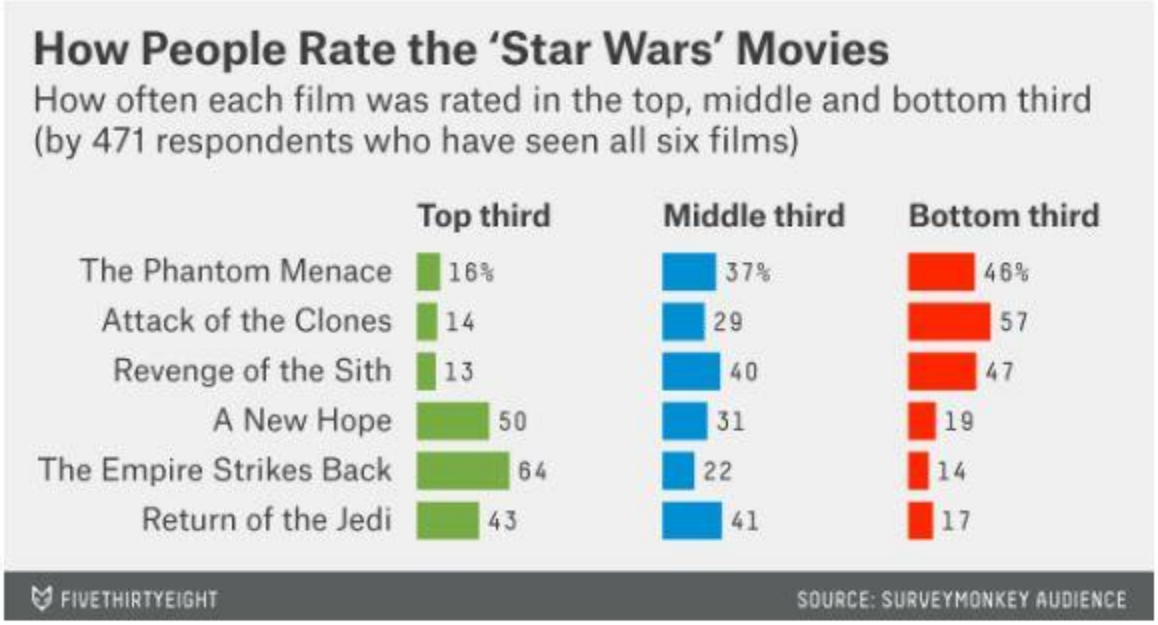

Figure 5: Graph showing how often each film was rated in top, middle, or bottom third [Hic14]

3.2 Analysis of Character Favorability Ratings

Next, we can see below how Hickey graphed the results for his analysis of the data regarding character favorability ratings. Note how low the character Jar Jar Binks rates in favorability.

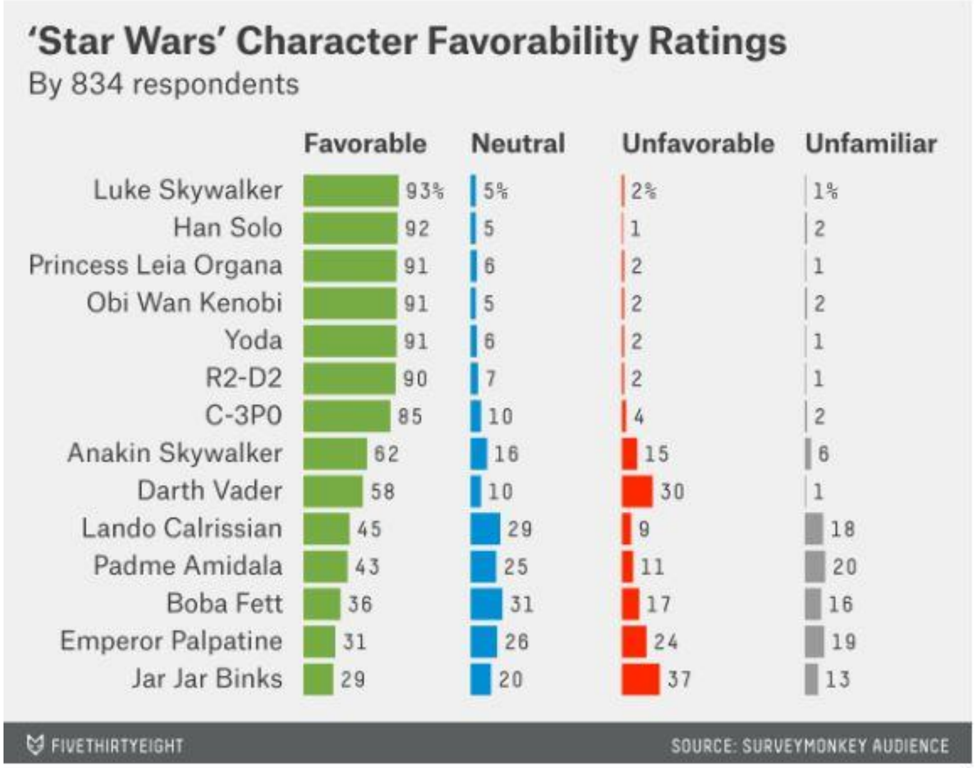

Figure 6: Graph showing Star Wars character favorability ratings [Hic14]

3.3 Analysis of Responses to “Who Shot First?”

Lastly, we can view below how Hickey graphs the responses to the survey question of “who shot first?”.

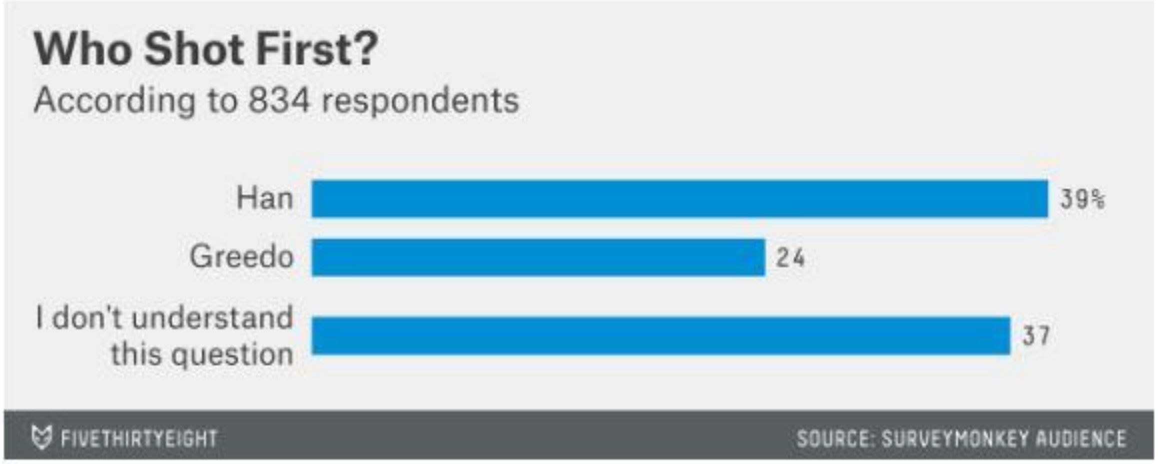

Figure 7: Graph showing results from the question “Who Shot First?” [Hic14]

As for my own analysis technique, I plan to use similar methods such as data analysis and visualization techniques to further investigate the relationships between the different variables in the Star Wars survey. Ultimately, our goal is to build on Hickey's analysis and uncover new insights about the Star Wars fandom and the factors that influence fans' preferences and opinions.

4 My Implementation

4.1 Data Preprocessing

For the data preprocessing of the Star Wars survey, I started by cleaning the data and removing any unnecessary columns or rows, primarily ones that contained numerous missing data values. Then, I transformed the Yes/No responses and film titles/blanks to 1s and 0s for easier analysis. NaN values were left as is and considered as missing data. I also created subsets of the data based on specific criteria, such as only including responses from Star Wars fans or only including responses from those familiar with the Expanded Universe. These subsets were then used for further analysis.

4.2 Data Analysis Cloning

To calculate and graph the percentage of people who have seen each film out of the people who have seen at least one film, I first removed any respondents who have not seen any Star Wars film. Then, I calculated the number of respondents who have seen each film by counting the number of "Yes" responses for each film. Next, I calculated the total number of respondents who have seen at least one film by summing the number of respondents who have seen each film. Finally, I divided the number of respondents who have seen each film by the total number of respondents who have seen at least one film to get the percentage of people who have seen each film.

To graph this data, I created a bar graph with the percentage of people who have seen each film on the x-axis and the film titles on the y-axis. This graph, shown below, allows us to easily compare the viewership of each film among those who have seen at least one Star Wars film. The results shown are identical to those seen in Hickey's article.

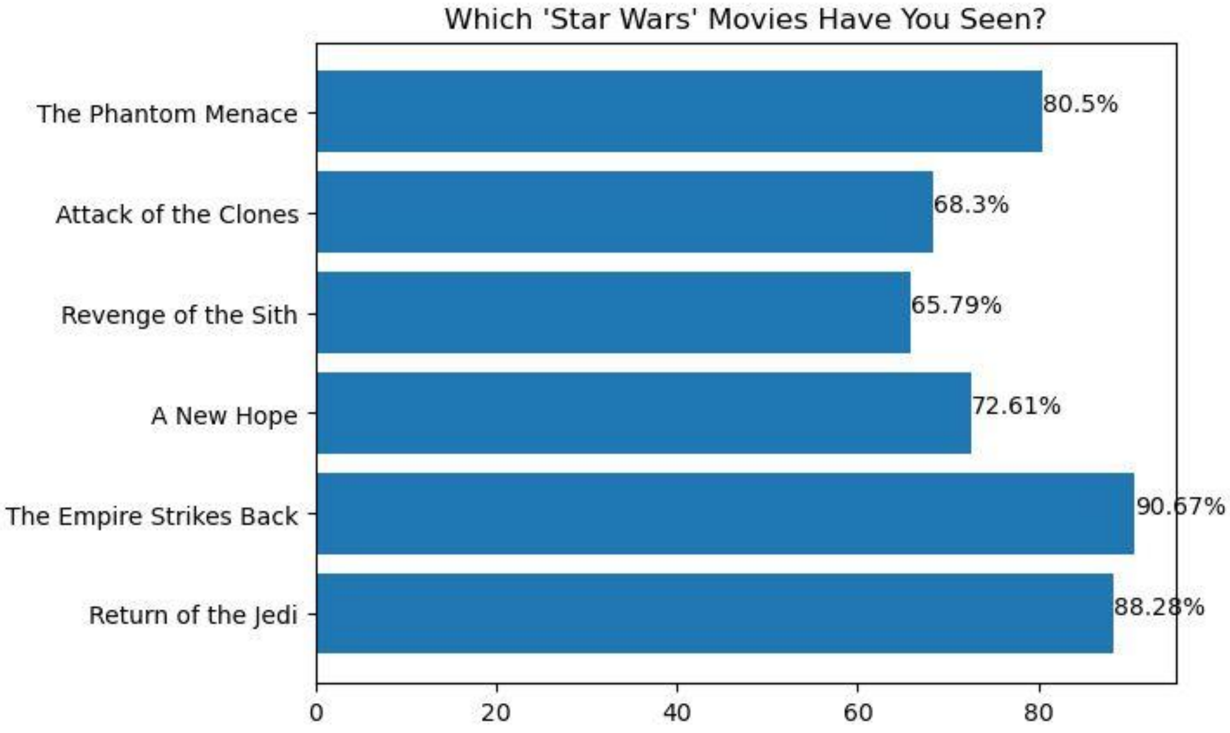

Figure 8: My graph displaying the percentages of respondents who have seen each film

To narrow down the data to those who have seen all six films, I filtered the dataset to only include respondents who have seen all six films. Then, I created a new dataset that consisted of the rankings for each of the six films for these respondents. I then calculated the number of respondents who gave each rank to each film and plotted this information in a stacked bar graph.

The y-axis of the graph represents the ranking given to each film, with 1 being the highest rank and 6 being the lowest rank. The x-axis represents the number of respondents who gave each rank to each film. Each film is represented by a colored bar, and the length of each bar corresponds to the total number of respondents who ranked that film at each particular rank. The graph shown below allows us to see how the respondents who have seen all six films ranked each one, and which films were the most and least popular among this group. While the graph differs from Hickey's, the data it displays is consistent with his findings where the original trilogy dominates the upper ranks, and the prequel trilogy falls into the lower ranks.

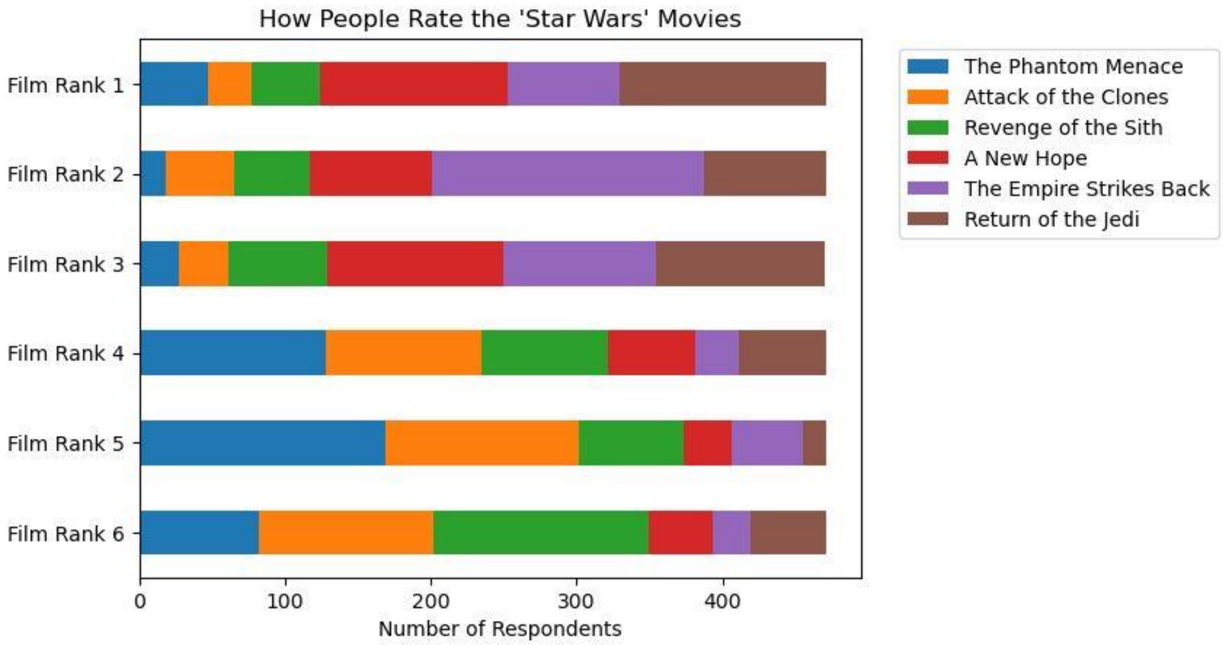

Figure 9: My graph showing how people rated each Star Wars film

To calculate the character favorability ratings, I first filtered the data to include only respondents who had seen at least one film. Then, for each character, I counted the number of respondents who rated them as Favorable, Neutral, Unfavorable, or Unfamiliar. Next, I calculated the percentage of respondents in each rating category for each character. Finally, I used this data to create a stacked bar graph with character names on the y-axis and the percentage of respondents on the x-axis. Each character's bar was divided into four segments, corresponding to their Favorable, Neutral, Unfavorable, and Unfamiliar ratings, respectively. The resulting graph shown below allowed us to easily compare the favorability ratings of each character across all respondents who had seen at least one film.

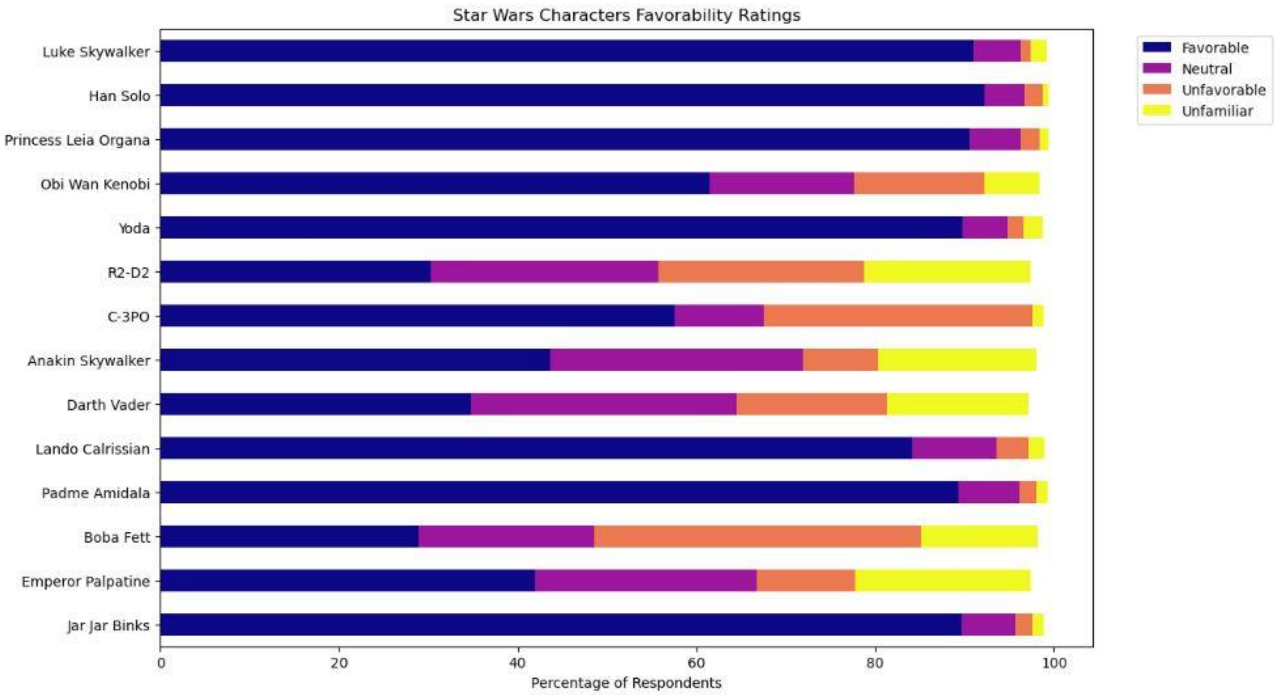

Figure 10: My graph displaying character favorability ratings

However, we can see that my analysis differs greatly from Hickey's analysis. As I had stated before, Hickey's analysis had a notably low favorability rating for the character Jar Jar Binks (as well as Emperor Palpatine), however my data shows a much higher favorability rating. There are a number of factors that could contribute to this difference and to choose just one would be mere conjecture. Factors could include, but not be limited to, bias, data mishandling, or technical errors.

Lastly, to analyze and graph the survey data for the question "who shot first?", I first extracted the data for all respondents who had seen at least one film. Then, I calculated the percentage of respondents who answered each option: Han, Greedo, or 'I don't understand this question.' I used these percentages to create a horizontal bar graph, where the y-axis represented the answer options, and the x-axis represented the percentage of respondents. The bars were color-coded to correspond to each answer option, with blue representing Han, green representing Greedo, and grey representing 'I don't understand this question.' The graph shown below allows us to visualize the distribution of answers and see which option was the most popular among respondents.

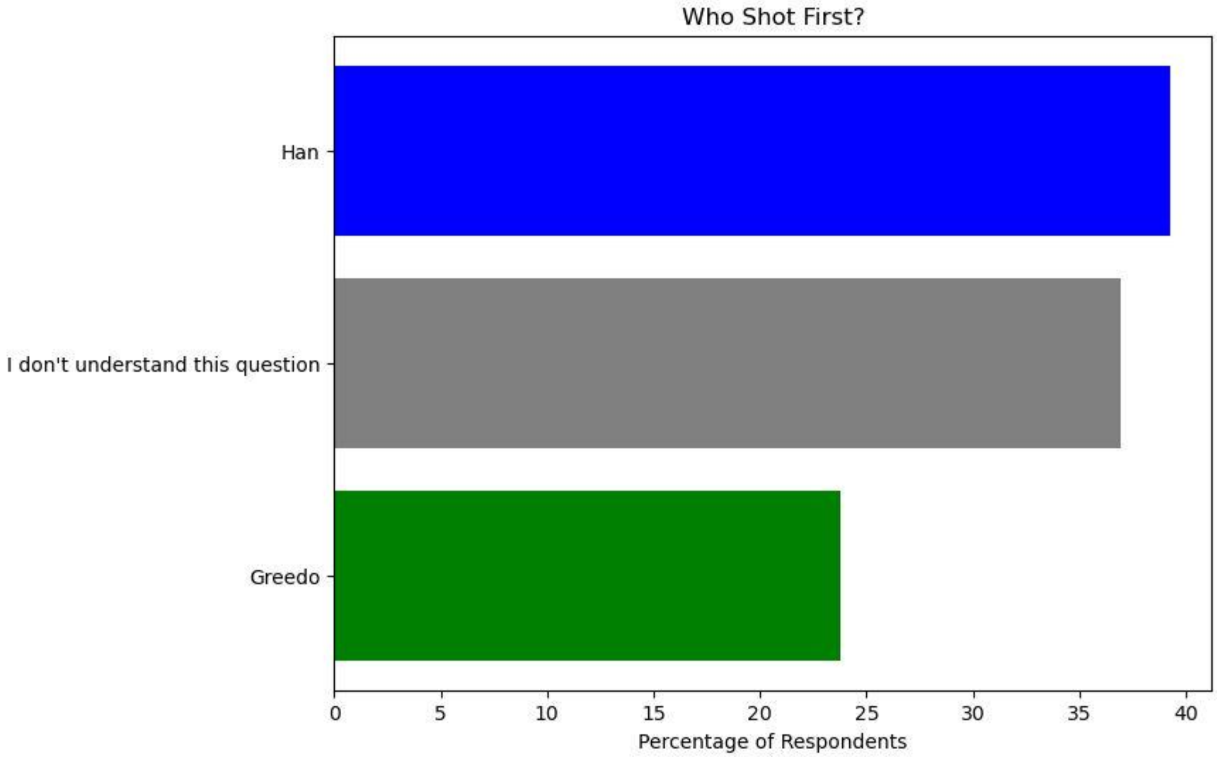

Figure 11: My graph showing the responses to the “Who Shot First?” question

4.3 My Conclusions and Afterthoughts

In conclusion, the analysis of the Star Wars survey data provided interesting insights into the preferences and opinions of Star Wars fans. From the analysis of the films seen by respondents, it is evident that the original trilogy has been the most widely watched by fans, with The Empire Strikes Back being the most popular. The analysis of the ranking of the films by those who had seen all six films indicated that the original trilogy remained the most favored among fans.

The analysis of character favorability revealed that Han Solo was the most popular character, but Jar Jar Binks was actually surprisingly popular, with a shockingly high favorability rating. Analysis of the "who shot first?" question revealed that a majority of respondents believed Han Solo shot first, contrary to George Lucas' later revision to the scene.

Overall, the analysis of the Star Wars survey data provided a fascinating glimpse into the preferences and opinions of Star Wars fans. Through the use of data analysis and visualization techniques, we were able to gain insights into how fans view the Star Wars films and characters, as well as their opinions on certain contentious topics.

To enhance this study in the future, we could gather more responses and include new questions that cover the new films, TV shows, and characters. This would provide a more up-to-date picture of the current state of the Star Wars fandom and its preferences. Additionally, prediction modeling using machine learning could be implemented to predict the most popular films, characters, and storylines in the future. This could be done by analyzing trends in the data and using algorithms to make predictions based on past behavior.

The insights gained from this data could influence future marketing strategies and media developments. For example, by understanding which characters are most popular among fans, media companies can tailor their marketing campaigns and merchandise to better appeal to the fandom. Additionally, understanding the perceptions and preferences of fans can help media companies develop storylines that resonate with fans and create a more positive reception for new films and TV shows. Overall, this data can provide valuable insights to guide future decisions in the development and marketing of the Star Wars franchise.

Thank you, and may the Force be with you!

References

[Hic14] Hickey, Walt. "America's Favorite Star Wars Movies and Least Favorite Characters." FiveThirtyEight, ABC News Internet Ventures, 22 July 2014, https://fivethirtyeight.com/features/americas-favorite-star-wars-movies-and-least-favorite-characters/.

[Sta19] "StatCrunch." StatCrunch, Pearson Education Inc., 2019, https://www.statcrunch.com/reports/view?reportid=27489&tab=preview .

[Tho19] Thomas. "The Results of the Star Wars Fan Survey Are In!" Jedi Temple Archives, 31 May 2019, https://www.jeditemplearchives.com/2019-05-31-the-results-of-the-star-wars-fan-survey-are-in/.