Analysis of a Star Wars survey

Kyrollos Iskandar
S3508880@student.rmit.edu.au

Abstract

In this work, it was set out to analyse how respondents of a Star Wars survey rated the Star Wars films, how their demographics affected how they viewed particular Star Wars characters, and if there were any other relationships between their answers to the survey. After analysing the cleaned data, it was found that respondents tended to rank films that they did not see lower than those they did see. Star Wars fans also had more knowledge, familiarity and understanding of Star Wars, and even Star Trek, than respondents who were not fans of Star Wars. There were also some correlations between respondents' demographics and how they viewed particular Star Wars characters.

1. Introduction

This work is an analysis of a survey done on Star Wars films in which 1,186 respondents were surveyed. It is focused on three questions. These questions were, 'How do people rank/rate Star Wars movies?', 'Are there any relationships between the data?', and 'What are people's attitudes to Star Wars characters based on their demographics?' The preparation of the data for analysis and its exploration were governed by these questions.

2. Data Preparation

The three questions specified in the Introduction meant that entries of respondents who did not see a Star Wars movie were useless. Thus, these entries were set aside.

2.1 Missing values

Approximately 150 people did not specify their household income. However, only less than around 20 people did not specify their gender, age, education or location. This discrepancy between the number of missing values in the household income column and the other four demographics columns was treated as significant. Therefore, the missing household incomes were assigned the value "Not Specified" in an attempt to see whether or not there was any special characteristic in that group of respondents who did not specify their household income. There were two rows with too many missing values in the five demographics columns. They were dropped because information on demographics was required for the aim of this work. There were two rows in which one film was not ranked. I found these rows by extracting the six rankings columns and checking which ones did not add up to 21. When I found them, I filled in the missing rank. A few respondents did not specify their view towards some characters. Their responses could not be inferred. Therefore, the rows with the missing responses were not included in the analyses in which their views on characters was required. There were 606 missing responses about whether or not respondents considered themselves as fans of the Expanded Universe. However, these missing values were expected from respondents who were not familiar with the Expanded Universe in the first place, which were 606 respondents. The number of respondents who were not familiar with the Expanded Universe equalled the number of respondents who did not answer whether or not they were fans of it. These missing values were replaced with "Unfamiliar (N/A)". There were two respondents who did not specify their location. So their locations were assumed to be the modal location of the dataset, which was "Pacific". This action of replacing missing values with the relevant modal value was done also for the "Education" column by age group. For the data analysis, dealt with the columns about whether or not respondents saw any particular film by replacing the names of the episodes & missing values by "Yes" & "No", respectively [1].

2.2 Whitespaces

There were a few entries that had extra whitespaces. These extra whitespaces were stripped off [1].

2.3 Typos

I replaced them with the correct values. For example, I replaced “Yess” with “Yes” & “Noo” with “No”.
There were also multiple entries that meant the same thing, such as “F” and “female” and “Female”. I changed all of these values to “Female” for consistency.

2.4 Impossible values

I dealt with them by checking what values existed in each column [1]. If I found an impossible value, I replaced it with a possible value [2]. Someone specified their age to be 500 years. Making this error is not possible if the respondents answered this question in multiple choice format. They must have typed it. Thus, it was most probable that they were trying to specify an age of 50 years but hit the “0” key a second time by accident. They would not have specified an age of 5 years, otherwise they would have been excluded from the survey.

2.5 Further cleaning of the data

For cleaning the data further, I split the population according to whether or not they saw at least one Star Wars film. There were 100 respondents who did not complete the survey. They answered only that they have seen at least one of the Star Wars films. They were excluded from the analysis because the purpose of this work required that respondents complete the survey. For the process of cleaning the data, these 100 rows were removed because they did not inform the aim of this work. Of the 1,186 responses to the survey, 818 of them were valid for analysis.

3. Data Exploration

3.1 How do people rank/rate Star Wars movies?

Respondents tended to rank films higher than films they did not see (Figure 1). It may have been that they were forced to rank films whether or not they saw them. However, some ranks were missing, suggesting that respondents were not actually forced to rank all films. It is possible, then, that respondents who did not see the films may have at least heard about them from those who did see them.

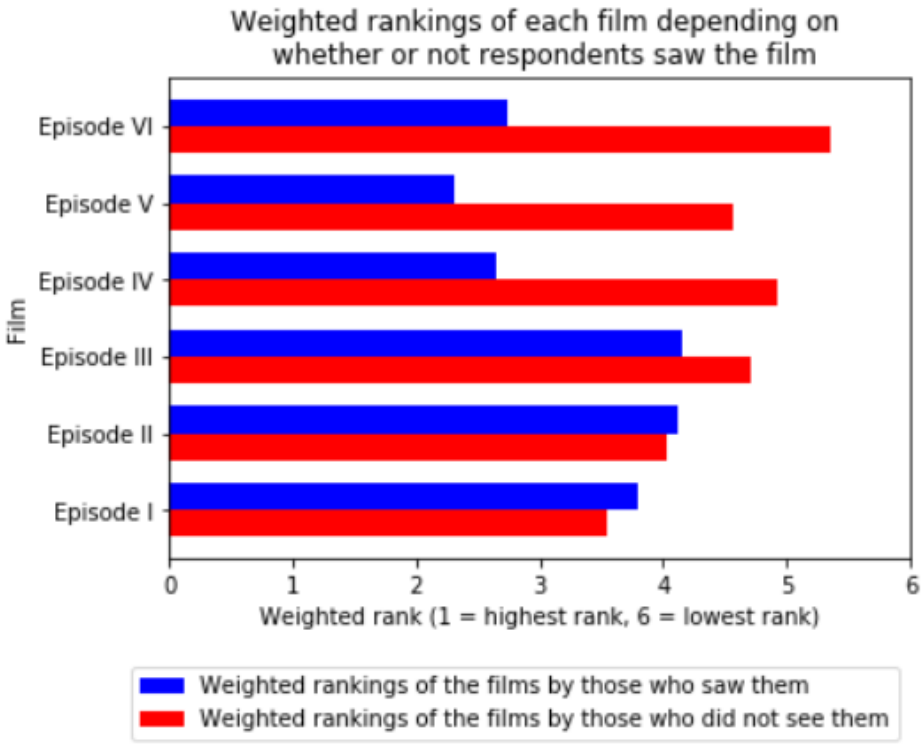

Figure 1: Average/weighted rankings of the six Star Wars films by fans and not fans of the Star Wars film franchise.

3.2 Are there any relationships between the data?

Fans of the Star Wars film franchise tended also to be fans of the Star Trek franchise (Figure 2). Fans of the Star Wars film franchise also tended to be more familiar with the Expanded Universe compared to respondents who were not fans of the film franchise (Figure 3). Fans of the Star Wars film franchise also tended to answer the question about who shot first more successfully than respondents who were not fans of the film franchise (Figure 4). Characters usually did not appear in all six films. Therefore, how respondents viewed particular characters is relevant only to the films in which these characters appeared in [3].

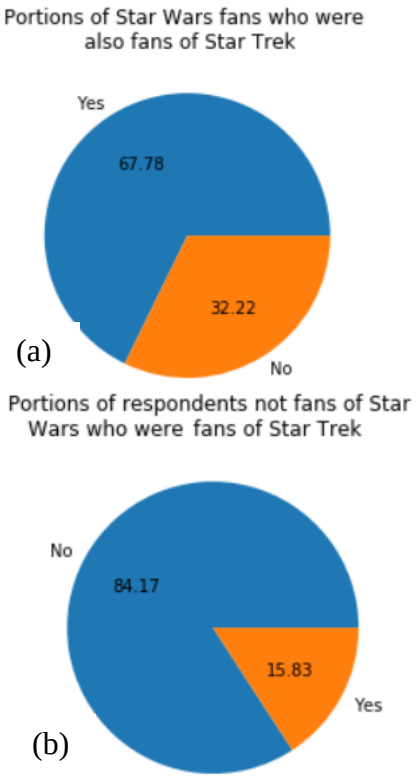

Figure 2: Relationship between whether (a) or not (b) respondents were fans of the Star Wars film franchise and whether or not they were also fans of the Star Trek franchise.

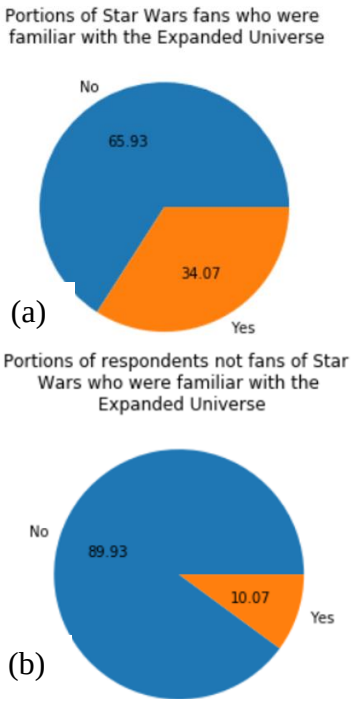

Figure 3: Relationship between whether (a) or not (b) respondents were fans of the Star Wars film franchise and whether or not they were familiar with the Expanded Universe.

3.3 What are people's attitudes to Star Wars characters based on their demographics?

For answering this question quantitatively, views such as the string entries of how respondents viewed the characters were converted to numerical views. In a copy of the original Pandas DataFrame, "*Very favorably*", "*Somewhat favorably*", "*Somewhat unfavorably*" and "*Very unfavorably*" were replaced with "2", "1", "-1" and "2", respectively. Whether respondents viewed characters neutrally or were unfamiliar with them were interpreted as "0". It was found that Han Solo, Lukewalker, Princess Leia Organa, Obi Wan Kenobi, C-3P0, R2 D2 and Yoda were the most popular characters while Jar Jar Binks and Emperor Palpatine were the least popular. The other characters' popularities were somewhere in between. Jar Jar Binks was especially unpopular among males (Figure 5), as well as respondents of age 18 – 44 (Figure 6), and respondents with Bachelor and Graduate degrees (Figure not shown). Females favoured C-3P0 and R2 D2 more than males did, while males favoured Darth Vader more than females did.

4. Conclusion

It was found that respondents tended to rank films lower if they did not see them compared to films they saw. Also, Star Wars fans tended to also be fans of the Star Trek franchise, be familiar with the Expanded Universe and have a better knowledge of which Star Wars character shot first compared to respondents who were not fans. There was also some correlation between respondents' gender, age and education and their view towards particular Star Wars characters.

References

1. Ren, Y., *Data Curation*, in *Practical Data Science*, Y. Ren, Editor. 2020, RMIT University: 124 La Trobe St, Melbourne VIC 3000.
2. COSC 2670/2738 Practical Data Science with Python, *Practical Data Science Tute/Lab 02*, in *Practical Data Science Tute/Lab*, COSC 2670/2738 Practical Data Science with Python, Editor. 2020, RMIT University: 124 La Trobe St, Melbourne VIC 3000.
3. Lucasfilm Ltd. *Star Wars*. 2020 [cited 2020; Available from: https://www.starwars.com/databank/.

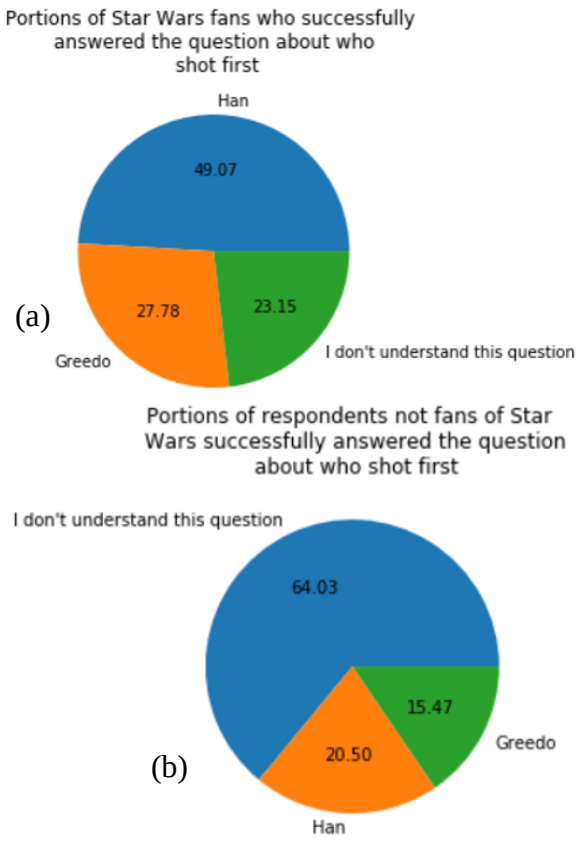

Figure 4: Relationship between whether (a) or not (b) respondents were fans of the Star Wars film franchise and whether or not they successfully answered the question about which Star Wars character shot first.

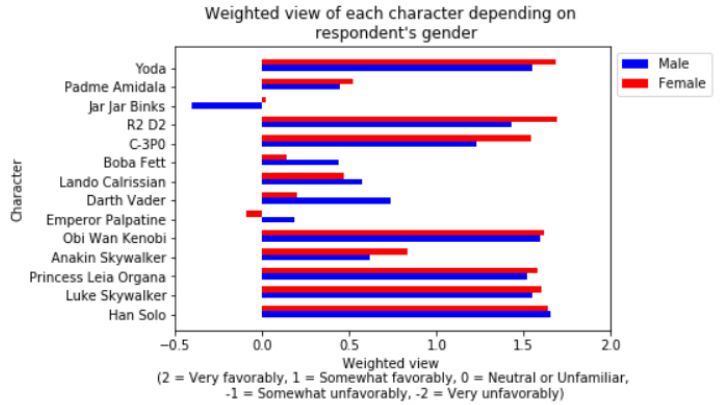

Figure 5: Relationship between respondents' gender and how they viewed particular Star Wars characters.

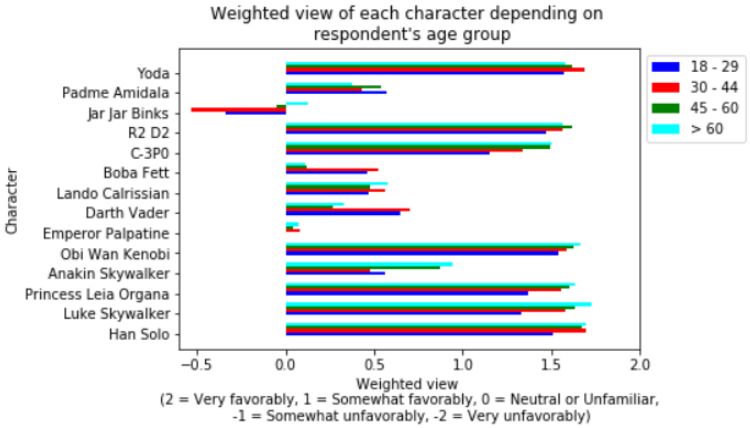

Figure 6: Relationship between respondents' age group and how they viewed particular Star Wars characters.