Introduction

The “StarWars” dataset is from a SurveyMonkey poll gathering 1,186 respondents. It contains survey questions and its respective answers about the different Star Wars movies from Episode I to Episode IX. There are also questions about some of its characters such as Luke Skywalker, Yoda and Greedo to name a few. The objective of this study is to explore and generate insights from the following questions (1) What is the most preferred StarWars movie? (2) Comparison of different fan types and (3) Different character preferences based on diverse respondents’ demographics.

I. Data Preparation

The dataset has 1,186 rows and 38 columns. These 38 variables, each with its appropriate datatype from int64, float64 and object, hence, data type conversion is not deemed necessary. Upon checking the *unique* values of each of the variables, several have unusual and missing values which were treated specifically in preparation for further analysis.

- **Typos**: There were typographical and acronyms errors in specific variables namely ‘Fan_SW_Franchise’, ‘Fan_ExpandedUniverse’ and ‘Gender’. A *replace()* function was used to deal with these errors and was successfully transformed from ‘Yess to Yes’, ‘Noo to No’ and ‘F to Female’ respectively.
- **Redundant White Spaces:** Extra white spaces were removed using a *str.strip()* function.
- **Capital Letter mismatches**: Certain binary variables from ‘Fan_Startrek’ and ‘Gender’ contain capital letter mismatches and *str.capitalize()* function was applied to transform ‘yes to Yes’, ‘no to No’ and ‘female to Female’, ‘male to Male’.
- **Sanity Checks:** The *describe()* and *unique()* function were used to check on the demographic variables and an unusual value of ‘500’ in the Age variable was discovered. As 500 seem so unrealistic for an age, it is considered as an outlier and the best approach was to use the r*eplace()* function and change the value with a NaN missing value. This is to be further treated with the rest of the NaN values in the dataset.
- **Missing Values**: In order to detect the missing values, *isnull()* function was applied separately by Categorical and Numerical variables.
    a. Drop rows - Due to a large amount of null values, further inspection led to discovering that 110 rows contain only the first two variables namely ‘Respondent ID’ and ‘Watched_SW’ and the remaining 36 variables have NaN values; Thus, these said rows were dropped as it was deemed not useful to the study and it barely provided any information.
    b. Categorical variables – Null values from Categorical variables were replaced with the word ‘No Response’.
    c. Numerical variables – The same treatment was used to the null values of Numerical variables as there were too many missing values and computing for the mean and mode seems inefficient and will only create an inaccurate assumption.
    d. A final run through of the null values was applied to double check if all null values were treated.

II. Data Exploration

In the data exploration part of the study, it is to be noted that all of the string values with ‘No Response’ were not used in the analysis since it was deemed as a futile response.

Also, all of the movie titles responses from the survey question ‘Which of the following Star Wars films have you seen?’ were replaced with the word ‘Watched’ to properly filter out those respondents who watched ALL 6 movies. This is for an accurate and unbiased representation of each movie in Task 2.1 Figures 1 and 2. Please note that this filtering was only applied to Figures 1 and 2 as it involves ranking of six variables.

Task 2.1 Rank the Star Wars films in order of preference with 1 being your favorite film in the franchise and 6 being your least favorite film.

For an accurate representation of the question, the respondents who watched and ranked **ALL** Star Wars movies for Ep I to Ep VI were **filtered**. A proper ranking of each movie entails that a respondent should have watched and ranked all the 6 movies. It assures an accurate and unbiased vote for each movie. Thus, 471 out of 1,186 respondents have watched and ranked all the 6 movies.

Based on the pie chart in Figure 1 it can be deduced that 35.9% of the respondents who watched all the 6 Star Wars movies ranked Episode V: The Empire Strikes Back as the best movie amongst the choices. It can also be observed that 80.5% of the respondents, the majority prefer Episodes IV, V and VI compared to only 19.5% of respondent who prefers the latter movies of Episode I, II and III.

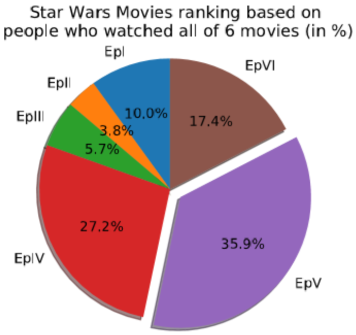

Figure 1: Percetage of people who watch all Star Wars movies and how they ranked each.

How do people rate Star Wars Movies?

Breaking down the respondents between Star Wars fan and non-fans. The graph in Figure 2 below shows that most Star Wars franchise fans with 82.7% of them prefer Episodes IV, V and VI. The same can be said to Non-fans of the Starwars franchise with 65% of them prefer the abovementioned movies. However, there is a bigger percentage of Non-fans who prefer Episode I, II and III compared to Star Wars fans. This suggests that there is a difference in the movie preference between Star Wars fans and non-fans.

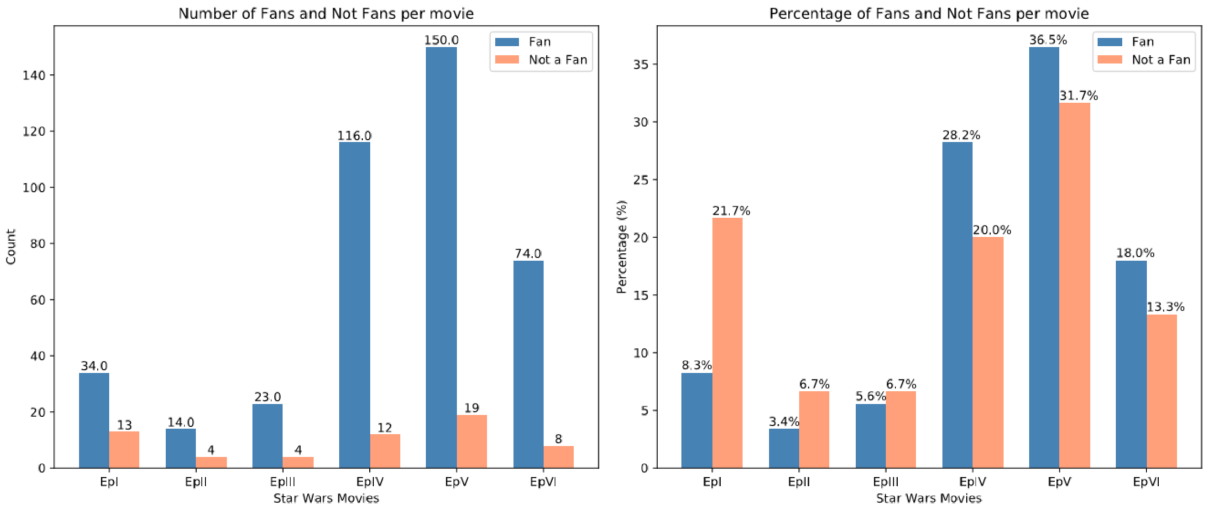

Figure 2: (Left) Number of Fan and Not fans per movie and (Right) Percentage equivalent.

Task 2.2 Relationships between columns

Plot 1: Are you a fan of Star Wars or Star Trek?

The space battle between two major franchises namely Star Wars and Star Trek has been a consistent conversation between space galaxy movie fans. Figure 3 below shows the majority of the respondents are both fans of the Star Wars and Star Trek Franchises with 34%. Although 32.1% of the respondents said they are neither fans of both the Star Wars and Star Trek Franchises. This represents that Star Wars and Star Trek movie franchises are in an equal playing field between space galaxy movie enthusiasts.

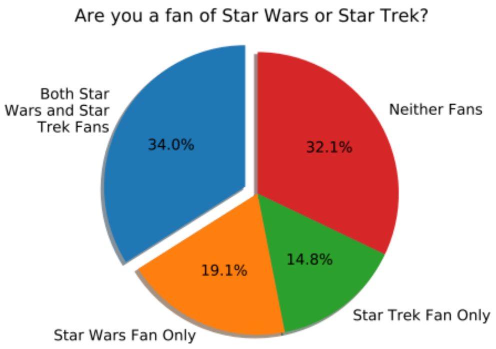

Figure 3: Percentage of Fans and Non-fans of Expanded Universe and Star Trek.

Plot 2: Who shot first (Han or Greedo) based on Fan Type?

This has been a long debate in the Star Wars community. A representation between fans and non-fans shows that 76.9% of the Star Wars fans believe that Han shot first while the Non-fans, who can be assumed that have only watched a few movies are unaware of this debate as 63.6% do not understand this question. It can be inferred that the Star Wars fans from the respondents are avid fans who are well aware of one of the biggest debates in the Star Wars community and the non-fans can be easily distinguished as they are confused with the question.

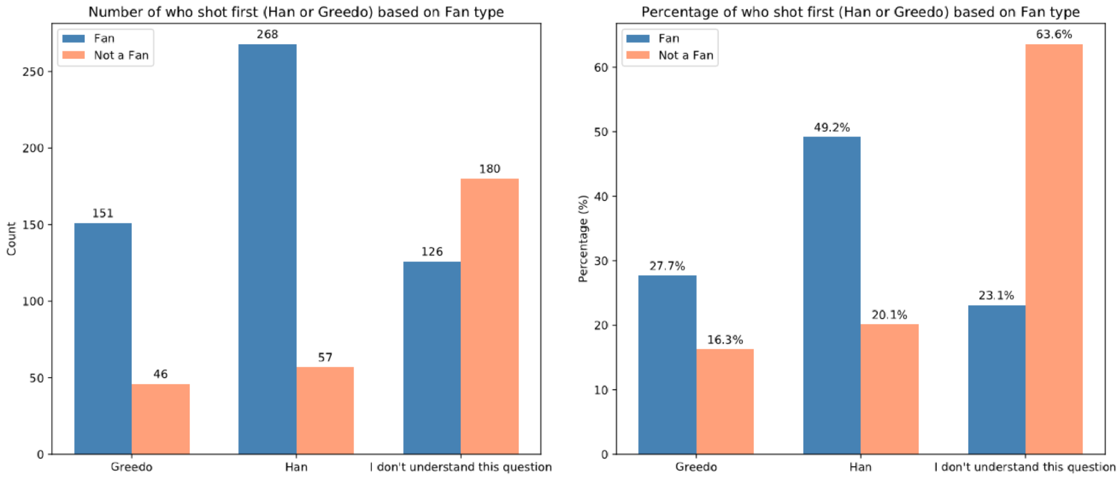

Figure 4: (Left) Number of who shot first (Han or Greedo) between Star Wars fans and non – fans. (Right) Percentage equivalent.

Plot 3: Favorability Ratings Between R2D2 VS C3P0

Two of the most loved robot characters in the Star Wars movies are namely C-3P0 and R2-D2, comparing their favorability likeness from each other it can be inferred that majority of the respondents are highly favorable of these two characters with 58% and 69% answered “Very Favorably” respectively. It can be assumed that C-3P0 and R2-D2 had a positive impact on the viewers of the Star Wars movies as they are portrayed as allies of the protagonists, Han Solo and Luke Skywalker.

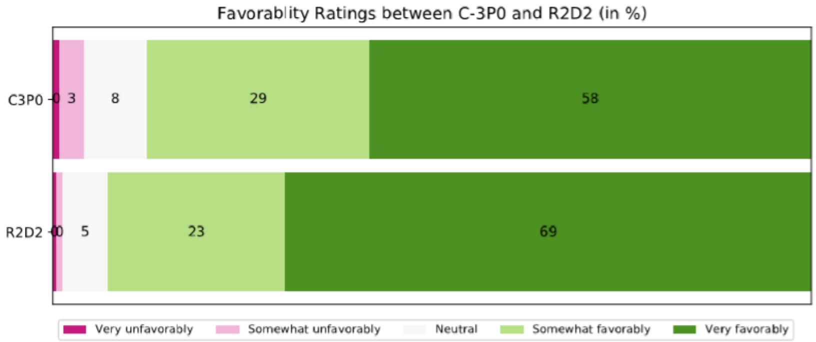

Figure 5: Percentage of favorability ratings between C-3P0 and R2-D2.

Task 2.3: Relationship between Star Wars characters between people’s demographics

A chi-square test of association with alpha = 0.05 level of significance is the most appropriate test to check if there are some relation between two categorical variables such as a specific Star War character and the corresponding demographics from Gender, Age, Income, Education and Location.

1. The **Null** hypothesis H0: There is no association between both variables.
2. The **Alternate** hypothesis H1: There is evidence to suggest there is an association between two variables.

Based on the heat map in Figure 6 below shows that the lightest yellow areas have attained a p-value of less than 0.05. Thus, we reject the null hypothesis and these variables are considered to have an association. Some of these variables are namely Emperor Palpatine and Gender with a p-value= 0.00 , Luke Skywalker and Age with a p-value=0.00, and R2D and Income with a p-value = 0.03, etc. It can also be observed that there is no significant association between any of the Star Wars characters and the Education and Location demographics.

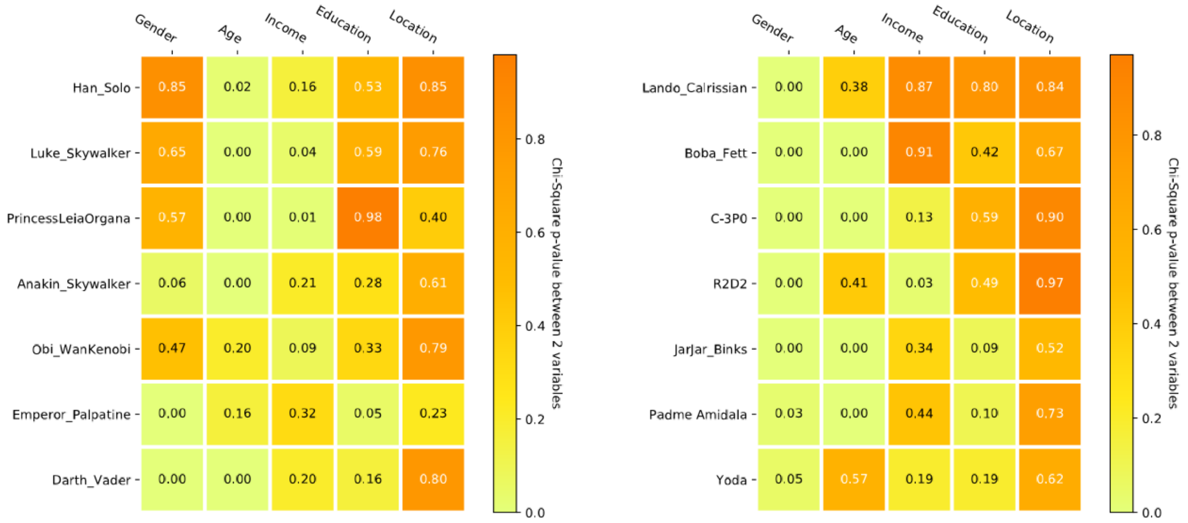

Figure 6: A heatmap of chi-square correlation between Star Wars characters and the respondents’ demographics.

Zooming on the associated variables from the chi-square test, in Figure 7 below between Darth Vader and Age with a p-value = 0.00 shows that the male respondents are highly favorable of Darth Vader as opposed to a certain number female gender who are very unfavorable of Darth Vader. On the right side of Figure 7, the association between Boba Fett and the respondents' Age with a p-value=0.00 shows that both male and female respondents are neither favorable of the said character.

These are a few examples of the Star Wars characters that are associated with the Gender demographics and it can be inferred that the male and female genders have different preferences and favorability on certain characters especially the antagonist characters like Darth Vader and unfamiliar and niche characters such as Boba Fett.

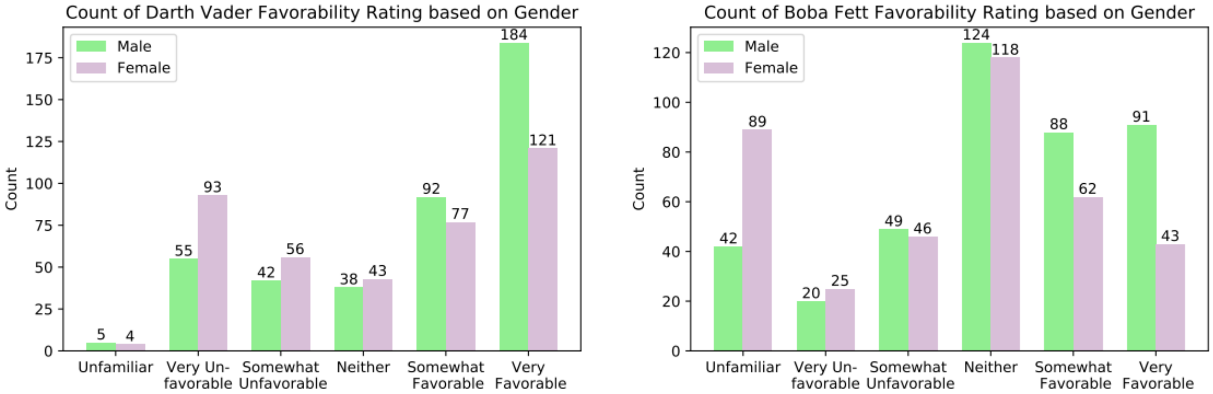

Figure 7: (Left) Count of Darth Vader favorability ratings based on Gender.
(Right) Count of Boba Fett favorability ratings based on Gender.

Meanwhile, concentrating on the significant association with the Age variable in Figure 8 below between Luke Skywalker and Age with a p-value = 0.00 shows that the different age range had a unanimous answer that most of them are highly favorable of Luke Skywalker. The same can be said between Anakin Skywalker and Age with p-value=0.00 shows that most of its respondents with different age range showed a unanimous response of positive favorability towards Anakin Skywalker. It can be inferred that depending on the perception a Star Wars characters regardless of age, respondents highly regard the main protagonists of the movie.

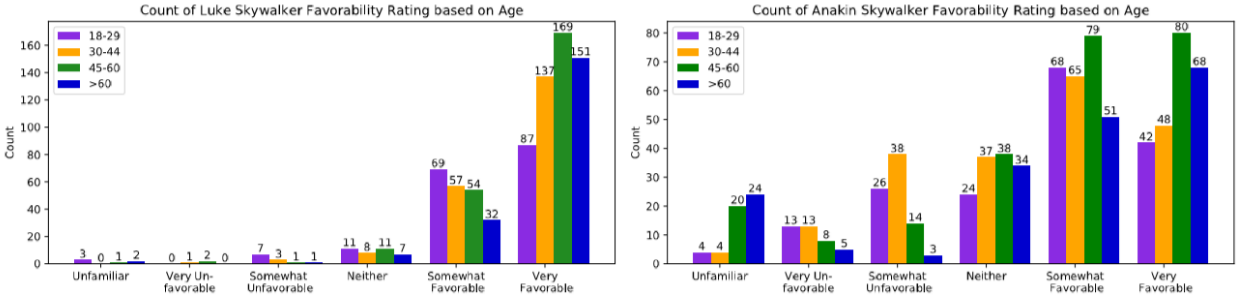

Figure 8: (Left) Count of Luke Skywalker favorability ratings based on Age.
(Right) Count of Anakin Skywalker favorability ratings based on Age.

Lastly, some examples of the associated variables with the Income variable in Figure 9 are Princess Leia Organa and Income with p-value = 0.01 and R2D2 and Income with p-value=0.03. The same observation with Figure 8 can be said, since two characters are deemed positive allies of the protagonists, regardless of the different range of income most respondents answered a high favorability with Princess Leia Organa and R2D2.

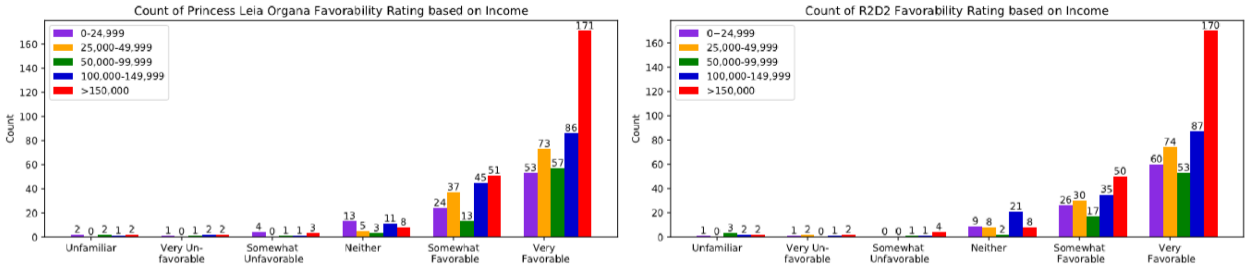

Figure 9: (Left) Count of Princess Leia Organa favorability ratings based on Income.
(Right) Count of R2D2 favorability ratings based on Income.

Conclusion

Based on the generated insights, these findings suggest that the respondents had differences and similarities with their perspective and response towards the Star Wars survey questions. There were different preferences of Star Wars movies between fans and non-fans, different answers on who shot first between fans and non-fans. However, regardless of these differences, there were also some noticeable similarities in the responses. Some respondents showed a similar perspective regardless of the different demographics. For instance, the favorability ratings of protagonists like Luke Skywalker and some of its allies like Princess Leia and R2D2 were deemed positively by most respondents despite differences in Age and Income. The results of the chi-square test of association also confirmed that there is a significant association between specific Star Wars characters and the demographics

References:

FiveThirtyEight (2014). America's Favorite 'Star Wars' Movies (And Least Favorite Characters). Retrieved from https://fivethirtyeight.com/features/americas-favorite-star-wars-movies-and-least-favorite-characters/

Towards data science (2020). Chi-Squared Test for Feature Selection with implementation in Python. Retrieved from https://towardsdatascience.com/chi-squared-test-for-feature-selection-with-implementation-in-python-65b4ae7696db

Matplotlib. Creating annotated heatmaps. Retrieved from https://matplotlib.org/3.1.1/gallery/images_contours_and_fields/image_annotated_heatmap.html#sphx-glr-gallery-images-contours-and-fields-image-annotated-heatmap-py