# ASSIGNMENT 1: DATA CLEANING AND SUMMARISING

Han Chien Leow (s3778722)

# 1.0 Data Preparation

## 1.1 Check data types

The data types were checked using the ".dtypes" function, which displayed the data types for all of the columns present in the Star Wars data frame. In addition, it helped to check if the data is loaded properly and matched the source (.csv) file. Therefore, I ensured that the loaded data had appropriate data types assigned and decided not to change any of the original data types because the data types were sensible and able to get the work done.

## 1.2 Typos

The typos in the data were checked using the ".value_counts()" function, which helped to check for data entry errors and outliers with a generated frequency table. It shows all the unique values along with the count. The data entry errors and outliers were manually checked with the generated frequency table for each of the columns in the Star Wars data frame. The typos were quickly fixed by using the ".replace()" function on the columns affected. For example, "Yess" was replaced to "Yes", "Noo" was replaced to "No" and "F" was replaced by "Female", since "F" is the only possible abbreviation for "Female" in the gender column. Thus, all typos were eliminated from the data frame.

## 1.3 Extra-whitespaces

Along the checking with ".value_counts()" function on each of the columns in the data frame, the extra-whitespaces were detected when the letter of the string matches but is not contained inside the same frequency in terms of the unique value, such as "Yes" which contained a frequency of 935, where "Yes " contained a frequency of 1. The extra-whitespaces were fixed using "str.strip()", where it successfully removed all both leading and trailing characters of extra-whitespaces present in the string. The process was repeated throughout the columns which detected similar occurrences.

## 1.4 Upper/Lower-case

Similarly, along the checking with ".value_counts()" function on each of the columns in the data frame, outliers on the capitalization of the letters were detected with some of the strings containing all lower-case letters, such as "yes", "no", "female" and "male". Therefore, the "str.capitalize()" function was used to convert the very first letter of the given series in capital. The function was executed with a for loop in each of the Star Wars columns along the checking if the data types were equal to object, then capitalizing all text data to upper case on the first letter of the string.

## 1.5 Sanity checks

Sanity checks were performed on each of the columns in the data frame using the ".value_counts()" to check for the presence of impossible values. The method of using conditional statements or comparison operators to run sanity checks on the columns was unable to be executed since the data types were mostly categorical and not numerical. For example, we can't use conditional statements or comparison operators to run sanity checks on the "Age" column since the value was displayed in age groups rather than the actual age with them being in object data type. Effectively, we can just utilize the generated frequency table from ".value_counts()" to run the sanity checks for checking the presence of impossible values. During the sanity check, an impossible value of the age being 500 was detected. The index/row where the age was 500, was located and the value of 500 was quickly changed to NaN (as missing value) because the actual value which the respondent intended can't be randomly assumed. So, it was safer and wise to change it to NaN rather than replacing it to another logical value which is guessed.

### 1.6 Missing values

The missing values were checked using ".isna().sum()" to get the total missing values count on each of the columns in the Star Wars data frame. The missing values for each of the Star Wars episodes ranking were filled with the column's mean, so the original episodes ranking of the mean will not be affected after the missing values were filled with the mean as a way to minimize the guessing error. Also, it helped to get more accurate information during data exploration, especially task 2.1. Furthermore, all other missing values were filled with a negative constant value (-1) to mark the fact that they are different from others. The missing values were rechecked using ".isna().sum()" and "starwars.isnull().any().any()" to confirm that there were no any remaining missing values left in the data frame.

## 2.0 Data Exploration

### 2.1 Explore a survey question

The process of exploring the question "Please rank the Star Wars films in order of preference with 1 being your favourite film in the franchise and 6 being your least favourite film.(Star Wars: Episode I The Phantom Menace; Star Wars: Episode II Attack of the Clones; Star Wars: Episode III Revenge of the Sith; Star Wars: Episode IV A New Hope; Star Wars: Episode V The Empire Strikes Back; Star Wars: Episode VI Return of the Jedi)" started by getting the mean from each of the Star Wars episodes rank using the ".mean()" function and assigning them to a variable called "ranking_data" using a for loop. Hence, the "ranking_data" will be containing all the mean of each rank on the Star Wars episodes. The ".mean()" function was used to get the average rank for each of the Star Wars episodes. Next, the bar graph of the average rank by Star Wars episodes were made, which contained a y-axis of rank and x-axis of Star Wars episodes. A bar graph was determined to explore this question because it is the most appropriate visualization to analyse how people rate Star Wars movies as it helps to visualize each of the average rank on the Star Wars episodes compellingly.

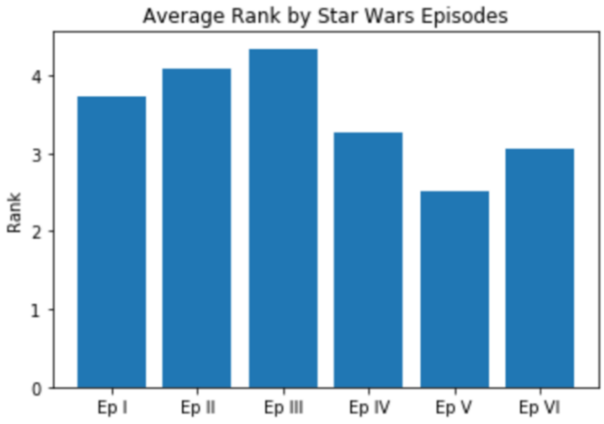

Figure 1: Bar Graph for Average Rank by Star Wars Episodes

Episodes with lower rank indicates better performance and favourite, while episodes with higher rank indicates bad performance and less favourite since they were rated with a scale from 1-6 (in order of preference with 1 being your favourite film in the franchise and 6 being your least favourite film). From the graph, we can see that "Star Wars: Episode V The Empire Strikes Back" performs the best (most favourite), while "Star Wars: Episode III Revenge of the Sith" performs the worst (least favourite). Ep V comes with a mean of 2.51, followed by Ep VI (mean: 3.05), Ep IV (mean: 3.27), Ep I (mean: 3.73), Ep II (mean: 4.09) and Ep III (mean: 4.34). According to Wikipedia, Ep IV, Ep V and Ep VI are the original trilogy of the Star Wars film franchise, while Ep I, Ep II and Ep III are the prequel trilogy of the Star Wars film franchise. From the information retrieved, we found that people tend to prefer the original trilogy compared to the prequel trilogy. Also, it is interesting to note that Ep IV, Ep V and Ep VI are older films, while Ep I, Ep II and Ep III are newer films. From the analysis, we found that people prefer the older Star Wars films compared to the newer Star Wars films. Moreover, the prequel trilogy shows a declining trend on the more recent Star Wars films (Ep I, Ep II & Ep III). In my opinion, the audience might prefer older films because they found the Star Wars film fresh during that period when it was first debuted and love the unique plot as well as the interesting settings in space, which never had been seen before.

## 2.2 Relationships between columns

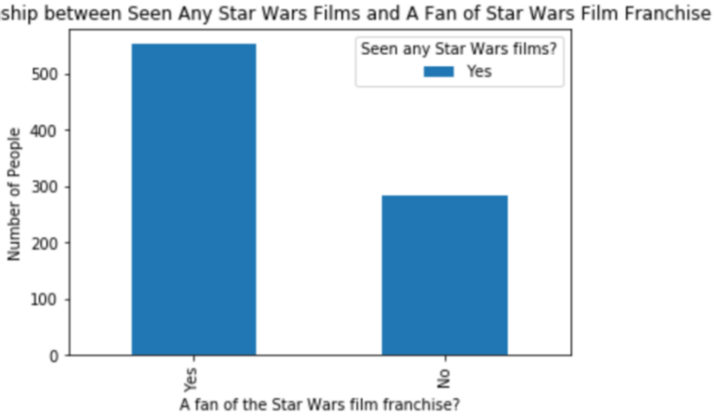

Figure 2: Side-by-side Bar Graph for Relationship between Seen Any Star Wars Films and A Fan of Star Wars Film Franchise

The hypothesis for the relationship between seen any Star Wars films and a fan of Star Wars film franchise is people who seen any Star Wars films are more likely to be a fan of the Star Wars film franchise. The hypothesis is accepted because the votes containing "Yes" of being a fan of the Star Wars film franchise from people who seen any Star Wars films (552 people) are more compared to the votes containing "No" of not being a fan of the Star Wars film franchise from people who seen any Star Wars films (284 people). 59% of the respondents from a number of 936 people who seen any of the Star Wars film voted "Yes" for being a fan of the Star Wars film franchise, signifies that people who watch a Star Wars film are more likely to be a fan of the Star Wars film franchise. It is interesting to note that there are no respondent who voted of being a fan of the Star Wars film franchise despite not watching any of the films because there might be some people who did not seen any Star Wars film but a fan of the Star Wars film franchise. This error is due to the limitation during the survey from the respondents, where the respondents can't proceed answering the other questions when not seen any of the Star Wars film.

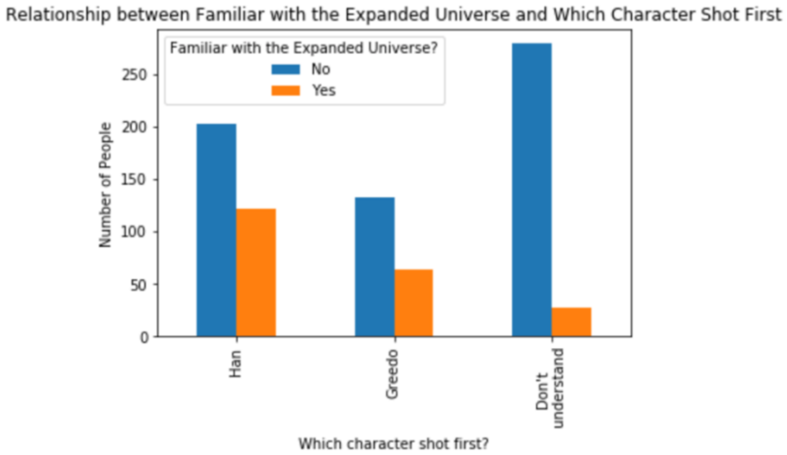

Figure 3: Side-by-side Bar Graph for Relationship between Familiar with the Expanded Universe and Which Character Shot First

The hypothesis for the relationship between familiar with the expanded universe and which character shot first is people who are familiar with the expanded universe are more likely to know that Han Solo is the character shot first. The hypothesis is accepted because the majority from the people who are familiar with the expanded universe picked "Han", while the majority from the people who are not familiar with the expanded universe picked "Don't Understand". 122 number of people who are familiar with the expanded universe picked "Han" as the character shot first, followed by "Greedo" with 64 people and "Don't understand" with 27 people, while 279 number of people who are not familiar with the expanded universe picked "Don't understand", followed by "Han" with 203 people and "Greedo" with 133 people. From the graph, we can see the significant difference on "Don't understand" between the people who are familiar and not familiar with the expanded universe. In addition, it is interesting to know that most people who are not familiar with the expanded universe know "Han" as the character shot first while the count from "Don't understand" is ignored.

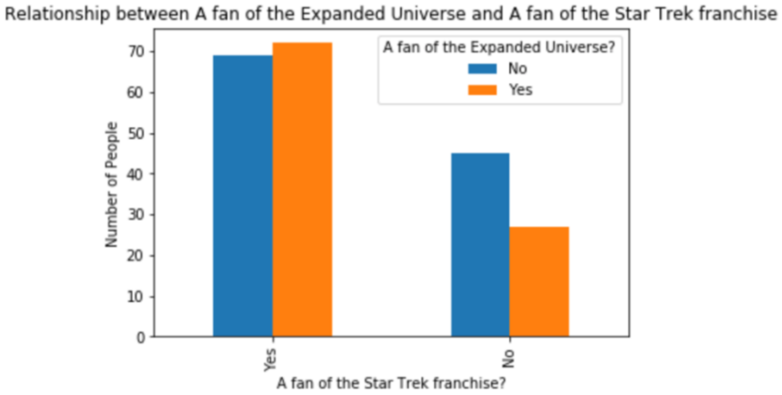

Figure 4: Side-by-side Bar Graph for Relationship between A fan of the Expanded Universe and A fan of the Star Trek franchise

The hypothesis for the relationship between a fan of the expanded universe and a fan of the Star Trek franchise is people who are a fan of the expanded universe are more likely to be a fan of the Star Trek franchise. The hypothesis is accepted when the fans of the expanded universe who voted "Yes" as a fan of the Star Trek franchise are more compared to the people who are not a fan of the expanded universe and voted "Yes" as a fan of the Star Trek franchise.

Also, the observation supports our hypothesis when the fans of the expanded universe who voted “No” as not a fan of the Star Trek franchise are less compared to the people who are not a fan of the expanded universe and voted “No” as not a fan of the Star Trek franchise. From the observation, we can also see that despite being a fan or not a fan of the expanded universe, people are more likely a fan of the Star Trek film franchise rather than not a fan of the Star Trek film franchise.

## 2.3 Explore a specific relationship

The relationship between people's demographics (Gender, Age, Household Income, Education, Location) and their attitude to Star Wars characters were explored using a side-by-side bar graph because it is the most significant and easily understandable visualization. Well-known Star Wars characters who contained the most favourability (Han Solo), average favourability (Darth Vader) and the most unfavourability (Jar Jar Binks) were chosen to explore the relationship between people's demographics (Gender, Age, Household Income, Education, Location) and their attitude to Star Wars characters. Other Star Wars characters were ignored as they were not as popular as Han Solo, Darth Vader, Jar Jar Binks and will add redundancy to the information retrieved.

From the graph for the relationship between gender and their attitude to Star Wars characters, the data shows that different gender has different attitude to Star Wars characters. For example, Han Solo is more favourably for male, Darth Vader is more favourably for male, while more unfavourably for female and Jar Jar Binks is more favourably for female, while more unfavourably for male. Thus, it proves that gender affects their attitude to Star Wars characters.

From the graph for the relationship between age and their attitude to Star Wars characters, the data shows that different age has different attitude to Star Wars characters. For example, Han Solo is more favourably in older adult group with age from 45 to 60, Darth Vader is more favourably in younger adult group with age from 39 to 44, while more unfavourably in older adult group with age from 45 to 60 and Jar Jar Binks is more favourably in older group with age from 45 to 60, while more unfavourably on younger adult group with age from 30 to 44. Thus, it proves that age affects their attitude to Star Wars characters.

From the graph for the relationship between household income and their attitude to Star Wars characters, the data shows that most favourability (from very favourably to very unfavourably) is picked by household income of 50,000 to 99,999 on all the 3 characters, Han Solo, Darth Vader and Jar Jar Binks. This might due to the fact that more respondents with high household income watched Star Wars compared to the respondents with low household income, resulting most of the data being picked by household income of 50,000 to 99,999.

From the graph for the relationship between education and their attitude to Star Wars characters, the data shows that different education has different attitude to Star Wars characters. For example, people with bachelor degree are more favourably to Han Solo, people with some college or associate degree are more favourably to Darth Vader, while people with bachelor degree are more unfavourably to Darth Vader and people with some college or associate degree are more favourably to Jar Jar Binks, while people with bachelor degree are more unfavourably to Jar Jar Binks. This shows that people with certain education are more favourably/unfavourably towards certain Star Wars characters. Thus, it proves that education affects their attitude to Star Wars characters.

From the graph for the relationship between location and their attitude to Star Wars characters, the data shows that people from different location have different attitude to Star Wars characters. For example, people from the Pacific are more favourably towards Han Solo, people from the East North Central are more favourably towards Darth Vader and people from the Pacific are more favourably towards Jar Jar Binks. Similarly, people from Middle Atlantic are more unfavourably towards Han Solo, people from East North Central are more unfavourably towards Darth Vader and people from the Pacific are more unfavourably towards Jar Jar Binks. This proves that location affects the attitude to Star Wars characters.