Assignment 1
Practical Data Science with Python

Name : Varun Ramesh
Student ID : s3793675

Table of Contents

Abstract:

The Star Wars data was a survey conducted in the United States. This survey has a detail flow of questions whether they are fans of the series to a detailed character analysis from each movie. Here in this assignment the idea is to clean the data and make appropriate conclusions from graphs and data. The social aspects such as demography of an individual is also taken into account with the type of answers the respondents have answered.

1. Data Preparation:

The dataset Starwars.csv contains 38 different columns of survey data. The data has several columns related to respondent like – Respondent ID, Gender, Age, Household Income, Education and Location.

1.1 Data Retrieving:

Using the function pd.read_csv the Data is being imported into Jupyter Notebook. As the head of the data is looked at, it is noticed that headings of the columns are multi-Indexed. This format proves it to be very hard to work with the data.

So to make it more easy for preparation, the column headings are changed and unnecessary headings are removed. The data frame is renamed to 'starwars'.

1.2 Check Data Types:

Now checking the data types of the data frame, all the columns are of type object except the first column (Respondent ID) which is float64 type.

1.3 Upper/Lower-case:

The whole data frame is converted into upper case to improve the uniformity in the data set. This can prove to be useful when the respondents have filled the survey with different capitalizations (Ex : Yes, YES, yes).

1.4 Extra-Whitespaces:

Whitespaces in the data set can be very hard to find. This will treat the whole String as a new one.

Eg : One of the columns had value counts of YES, NO, YES. Although it looked like there was no difference between the YES strings, there was a blank space present in the String. *Strip* function was used here to remove the white spaces.

1.5 Typos:

Typos were present in the data set and a for loop was made run through all the *Value Counts* of each of the columns. This was in a single output all the typos could be recognize.

After the value counts were obtained, it can be noticed that the Yes and No's were typed differently. So, *replace* function was used to fix these typos.

1.6 Sanity Checks:

As found from the value counts generated, it was found that there was one value of age which was 500. So, assuming this is a typo for 50, replace function was used fit the value in the range '45-60'.

1.7 Missing Values:

- Making True for watched movies and False for null values (movies not watched)
- Now removing SOME null values (renamed to False) from columns Watched|Movies based on the following condition:
    - Have you seen one any of the films == YES
    - If there is a missing value (False/not watched movie)

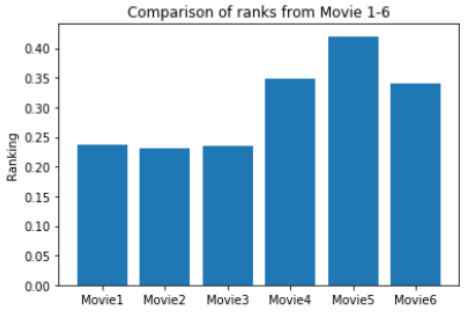
The reason for doing this is, although respondent has not watched all the movies, a ranking is still given. This makes it unfair because he has not even watched that movie. Hence such respondents need to be removed from our analysis.

- Removing all the NaN values from ‘do you consider yourself to be a fan of the Star Wars film franchise?'. This is because the answer can either be an YES or NO, there can be no missing values here.
- The other missing values cannot be removed because that will cause a lot of data to be lost. To make the visualizations more realistic, there must be enough data to get accurate results.
- Dealing missing values for task 3: For task 3, just the familiarity of the characters and the demographics are needed. So nan values are removed here.

2. Data Exploration:

2.1 Explore the survey question: Which movie is rated the best?

Considering only columns where the analysis must be done, the mean of all these columns is taken. This gives us a rank which movie is rated the best.

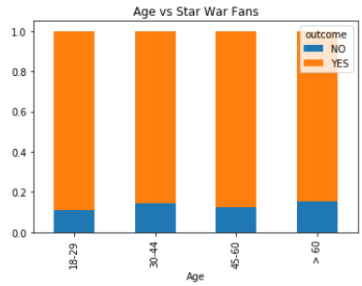


As we can see from the graph, movie 5 (Star Wars: Episode V The Empire Strikes Back) is rated the best while movie2 (Star Wars: Episode II Attack of the Clones) is rated the least.

2.2 Relationship between columns:

2.2.1 Comparing Fans of Star wars and their Age:

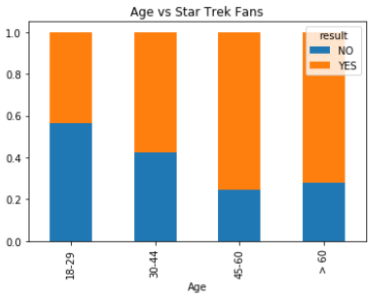
Considering only 2 columns, if there are fans of star wars and their age. Now using *group by* function to group by ‘Age’, this way we can know what the age groups of fans are. This is represented by stacked bar chart.



From this order from highest to lowest fans of age group are from 18-29, 45-60 , 30-44 and finally >60.

2.2.2 Comparing Fans of Star Trek and their Age:

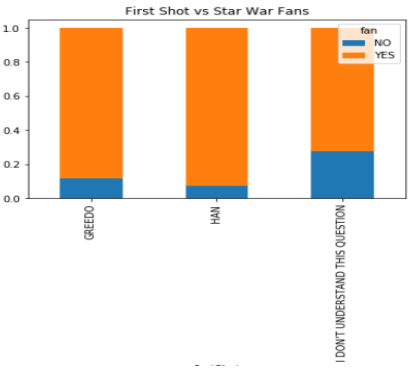
Just like the previous one, similar workflow is used as before and is compared with StarTrek fans



We can notice that the fans of Star Trek are much older compared to Star Wars Fans. This is because Star Trek was released in 1966 but Star Wars was released in 1977.

2.2.3 Fans who got the right answer to 'Which character shot first?

Taking fans of Star Wars column and the column ‘which character was shot first’, in a similar way as before a comparison was made between both of them and then was identified.

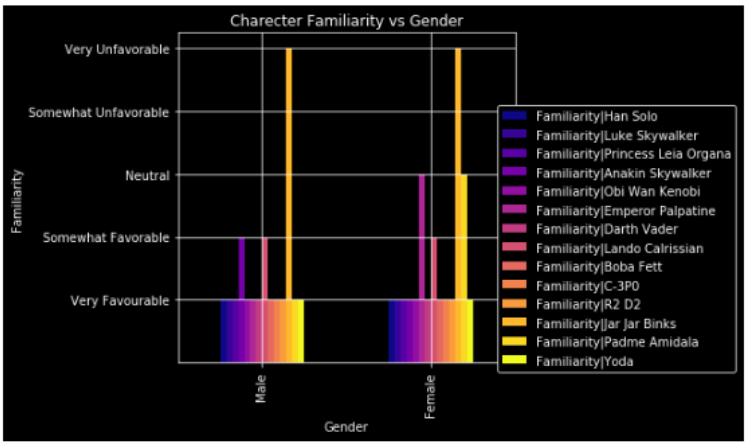


From the above observation, we can see that Fans of StarWar have mostly chosen Han Solo as the right answer. After all Han Solo is the right answer. However, most of the fans have gotten confused with Greedo as after all Greedo was shot at. Finally, the greatest number of non-fans have answered this question as 'I don't understand this question'

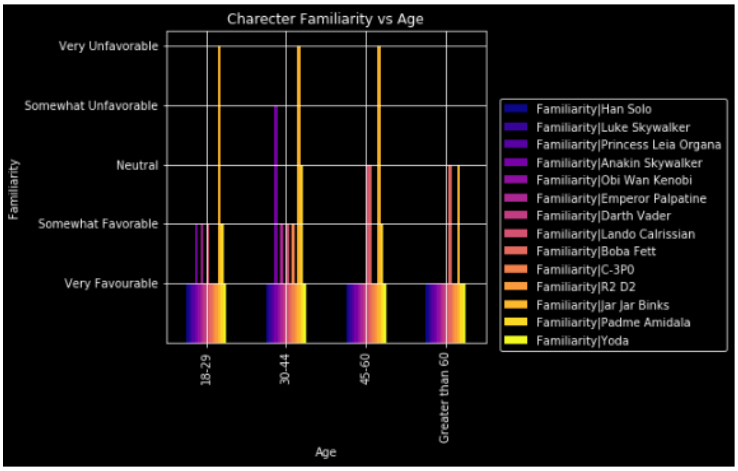
3. Familiarity of Characters vs Demographics

Below is the graph of different demographics and their familiarity with the characters. Using group by functions with respect to each of the demographics and familiarity of the characters, we can get multiple graphs as showcased below:

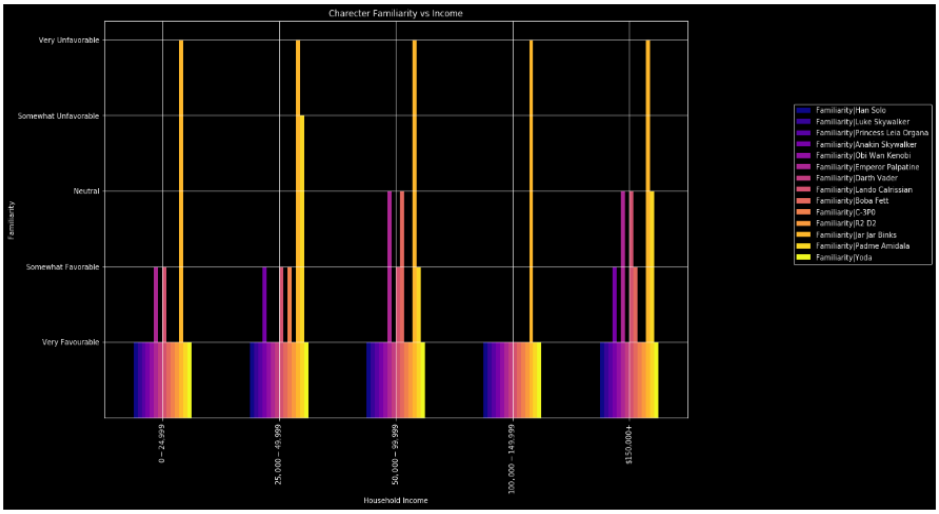
1) Gender: The graph shows us that most of Male population voted 'Very Favourable' compared to female. Females also were close behind when it came to specific character. From the graph the most deviation from both male and female equally was between 'Jar Jar Binks' and 'Lando Calrrissan'



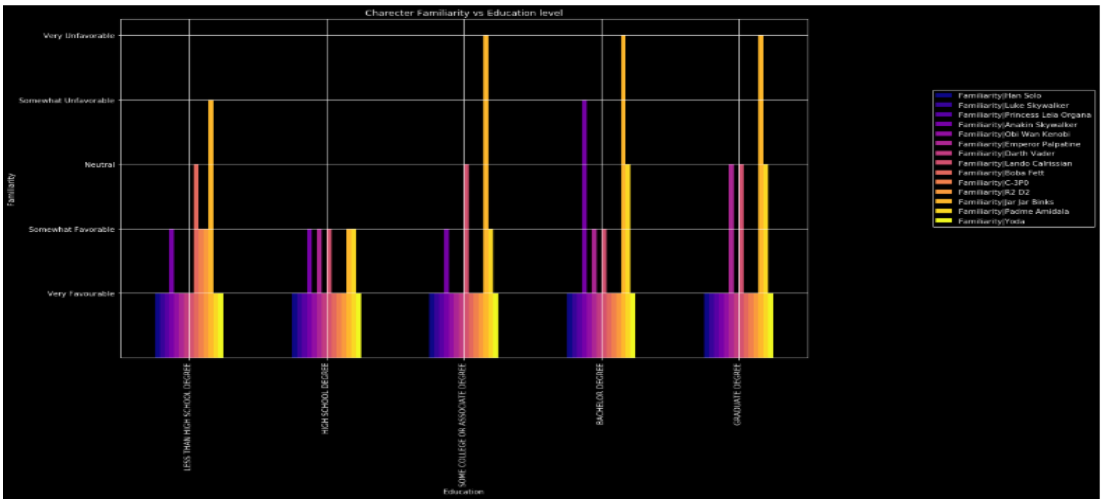
2) Age: People with an age group of > 60 found Jar Jar Binks very favourable while others age groups rated the character very unfavourable.



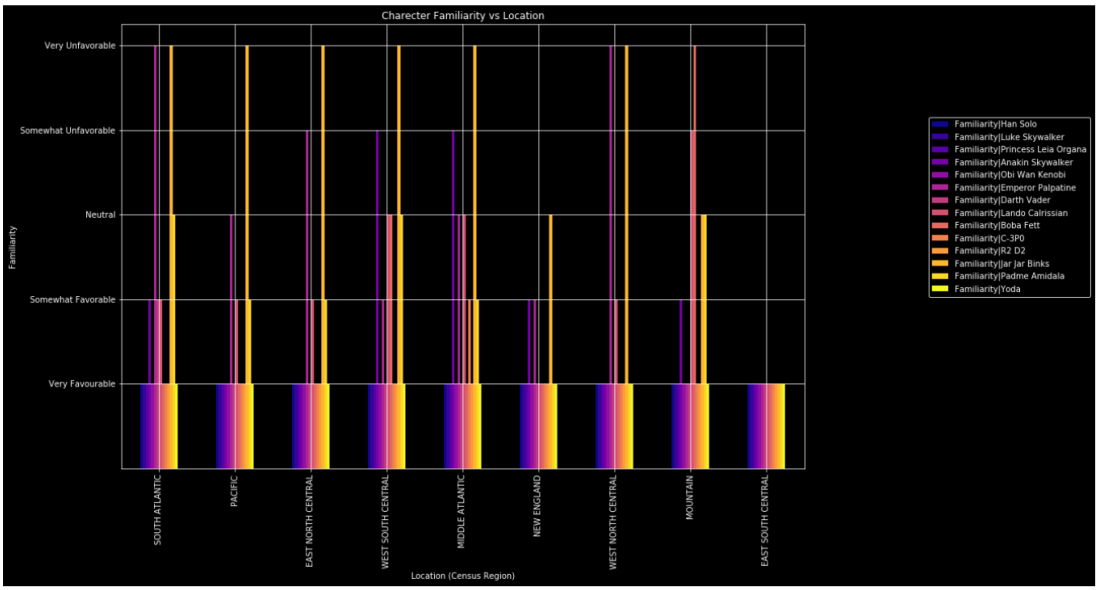
3) Income: As we can see from the graph, there is similarity for most characters on how people have voted even when their income groups were different. People who had an income of more than USD 150,000 were a little more unsure of the characters than the ones who had lesser income. We can also notice that the character familiarity was similar between USD 50K-99K income group and USD 150K income group. Income group ranging from $100K-150K were very sure of the characters of the movies.



4) Education: From the graph we can see that people with less than high school degree are very favourable to most of the characters compared to all other education level. While other education levels chose very favourable for 'R2D2','C-3PD', 'Boba Fett'. Most people having High School degree chose either very favourable or somewhat favourable which indicates a more understanding of the characters.



5) Location: People from South Atlantic and West North Central, consider Emperor Palpatine Very Unfavourable while others regions either consider him Neutral to Very Favourable. Jar Jar Binks is considered Very Unfavourable by all regions except New England and East South Central. Most other characters to a large extent are considered Very Favourable.



Conclusion

- Star War fans are younger than Star Trek fans
- Most fans got the answer to who shot first as Han Solo but some of them got confused with Greedo.
- Most demographics with diverse opinions were: Female, age of 30-44, annual salary 50K – 99K USD, education of Lesser than High School, and from West South Central. And people with a strong opinion with very favourable: Male, age of 18-29, annual salary 100K – 150K USD, education of High School degree, and from East South Central.