2020

COSC2670 Assignment 1

PRACTICAL DATA SCIENCE WITH PYTHON

NIKHIL SHARMA S3833151

1) Data Preparation:

The first step of the project was to prepare the data. There were various steps involved in data preparation since it is one of the most important steps of the data science process. There were 37 columns in the dataset so going through all the columns and checking it was a very tedious task but it needed to be done because if the data is not prepared well, it can affect the other steps of the data science process.

Below are the assumptions that were made while preparing the data:

- There were 100 records in which the person has answered “Yes” to the survey question ”Have you seen any of the 6 films in the Star Wars franchise?” but did not answer any of the other questions. Also there were 10 records in which a person has said “No” to the survey question ”Have you seen any of the 6 films in the Star Wars franchise?” but did not answer any of the other questions. These 100 records were deleted by considering them bad responses. These 110 records were deleted because they were considered as bad records.
- If a person has not answered the question “Do you consider yourself to be a fan of the Star Wars film franchise?” means that they are not a fan.
- If a person has not answered the question “Which of the following Star Wars films have you seen? Please select all that apply”, then the null value has been replace with “No”, assuming that had a person watched a movie he would have answered.
- If a person has not answered the question “Please state whether you view the following characters favorably, unfavorably, or are unfamiliar with him/her.” means that the person is unfamiliar with them.
- If a person has not answered the question “Are you familiar with the Expanded Universe?” means that they are not familiar with the expanded universe.
- If a person has not answered the question “Do you consider yourself to be a fan of the Expanded Universe?” means that they are not a fan of the expanded universe.
- If a person has not answered the question “Do you consider yourself to be a fan of the Star Trek franchise?” means that they are not a fan.

The following steps were followed in order to prepare the data for the exploration phase.

After this the below structure was followed next to clean the data.

1.1) Data Retrieving:

This was done by the read_csv() function of the pandas library. While importing the data, the names of the columns were renamed because the previous names were confusing and were not good for the analysis phase. To rename the columns the “name” argument of the read_csv function was used. Once the data was imported, it was checked by skimming through it to check if it was imported properly.

1.2) Typos:

There were a few typos in the columns “fan_Star_Wars”, “fan_Expanded_Universe”, “fan_Star_Trek_franchise” and “Gender”. They were corrected using the replace() function.

1.3) Extra-whitespaces:

Extra white spaces were removed from all the non-numeric columns using the strip() function.

1.4) Sanity Checks:

The sanity checks were done for all the columns using the value_counts() function. There was an impossible value of “500” in the “Age” column which was replace with the mode of that column which was the age group ”45-60”. It was corrected using the replace function.

1.5) Missing values:

- The missing values for the “fan_Star_Wars” column was replaced with “No”.
- The missing values for the six columns for the “Which of the following Star Wars films have you seen? Please select all that apply.” question was replaced with “No”.
- The missing values for the six columns for the “Please 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.” question were replaced with 0 indicating that the person did not rate the movie.
- The missing values for the fourteen columns for the “Please state whether you view the following characters favorably, unfavorably, or are unfamiliar with him/her.” question were replaced with “Unfamiliar (N/A)” indicating that the person does not know the character.
- The missing values for the “Which character shot first?” column were replaced with two values:
    - “NA”: These were for the records which have answered “No” to the survey question ”Have you seen any of the 6 films in the Star Wars franchise?”
    - With the mode of the column which was “Han” for the records which have answered “Yes” to the survey question ”Have you seen any of the 6 films in the Star Wars franchise?”
- The missing values for the “Are you familiar with the Expanded Universe?” column is replaced with “No”.
- The missing values for the “Do you consider yourself to be a fan of the Expanded Universe?” column is replaced with “No”.
- The missing values for the “Do you consider yourself to be a fan of the Star Trek franchise?” column is replaced with “No”.
- The missing values for the “Gender” and “Age” columns is replaced with the mode of that column.
- The missing values for the “Household Income” column is replaced with “Not Available” indicating that the person did not answer this question.
- The missing values for the “Education” and “Location (Census Region)” columns is replaced with the mode of that column.

1.6) Upper/Lower-case:

All the non-numeric columns were converted to upper case. This was done using the upper() function.

1.7) Check data types:

All the columns except for the “RespondentID” column was converted into categorical data since all the other columns were categorical.

2) Data Exploration:

2.1) Analysing how people rate Star Wars Movie

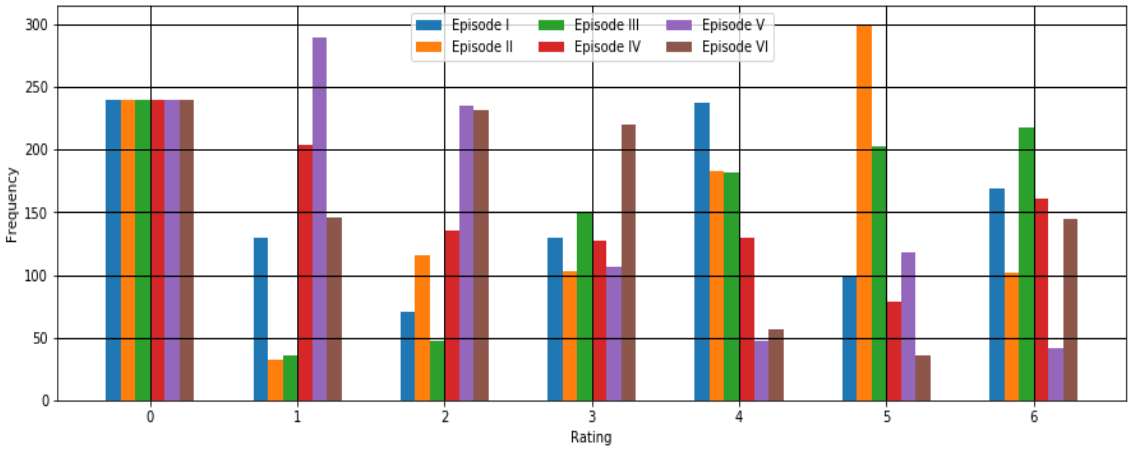

Figure 1

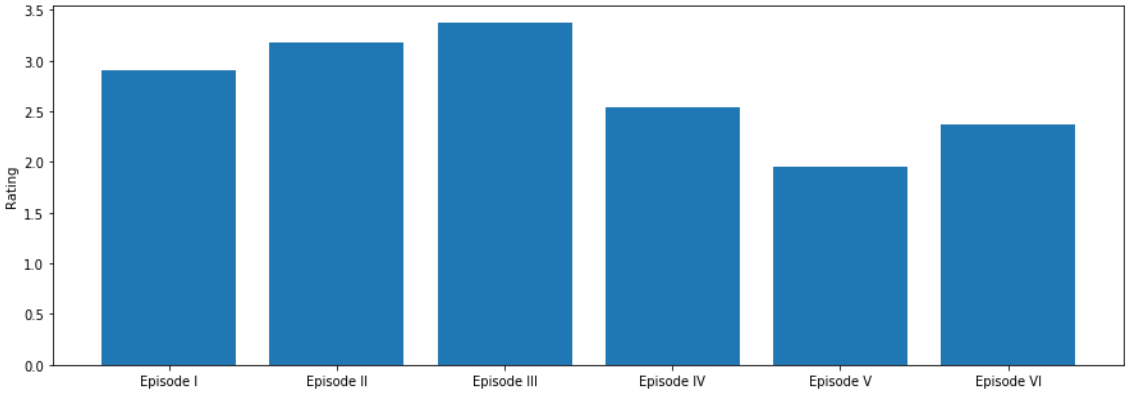

Figure 2

Figure 1 shows that around 250 people did not rate the movie because they did not watch it, so the ranking is only for those people who have watched the movie. Figure 2 shows that Episode V was the most like movie of all time followed by Episode VI, Episode IV, Episode I, Episode II, Episode III. An interesting fact that can be visualized from the plot that people liked the first trilogy (Episode IV, Episode V, Episode VI) more as compared to the second trilogy (Episode I, Episode II, Episode III). Episode III was no least liked of them all. As the movies were released, they tend to get worse over time according to the survey ranking. Another interesting fact is that even though time was passing, and the technology was advancing the movies did not get better. Hence, we can say that evolution technology did not have an impact on the Star Wars movie. The above data also shows that around 250 people did not rate the movie because they did not watch it, so the ranking is only for those people who have watched the movie.

2.2) Relationship between columns

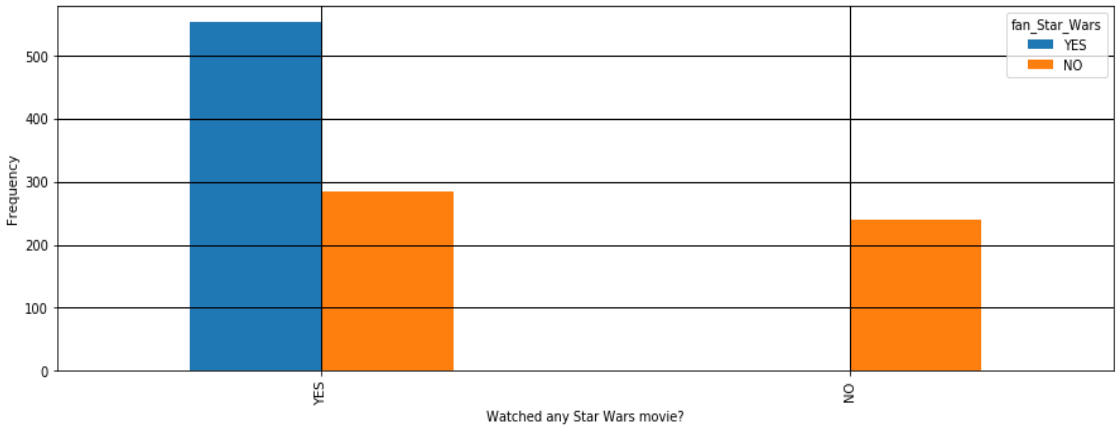

Figure 3

In figure 3 the comparison is done between the columns “Have you seen any of the 6 films in the Star Wars franchise?” and “Do you consider yourself to be a fan of the Star Wars film franchise?”. The plot shows that if a person has seen any of the Star Wars movie then there is 2/3 chance that they are a fan of the movie. Around 875 people have watched any of the Star Wars movie out of which around 575 people are Star Wars fan and a little less than 300 people are not a fan of the Star Wars movie.

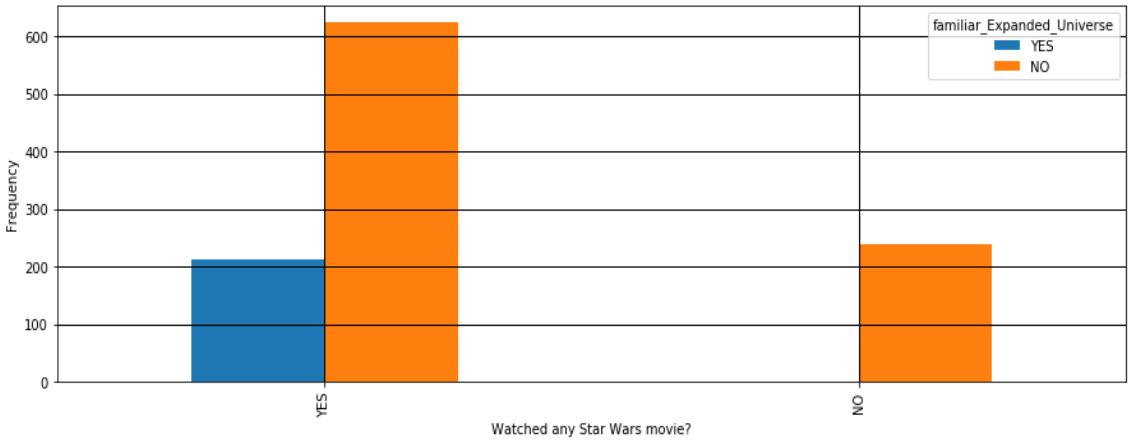

Figure 4

In figure 4 the comparison is done between the columns “Have you seen any of the 6 films in the Star Wars franchise?” and “Are you familiar with the Expanded Universe?”. The plot shows that if a person has seen any of the Star Wars movie then there is 1/4 chance that they are familiar with the expanded universe. Around 875 people have watched any of the Star Wars movie out of which a little over 200 people are familiar with the expanded universe and around 650 people are not a familiar with the expanded universe.

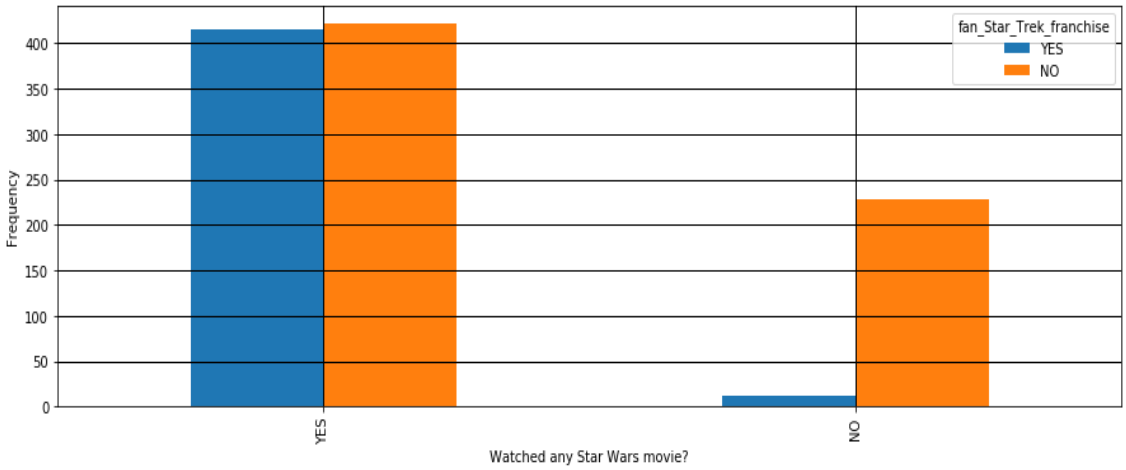

Figure 5

In figure 5 the comparison is done between the columns “Have you seen any of the 6 films in the Star Wars franchise?” and “Do you consider yourself to be a fan of the Star Trek franchise?”. The plot shows that if a person has seen any of the Star Wars movie then there approximately is a 50-50 chance that they are a fan of the Star Trek franchise as well. Around 875 people have watched any of the Star Wars movie out of which around 420 people are a fan of the Star Trek franchise and around 650 people are not a familiar with the expanded universe. Another interesting observation is that out of the 250 people who have not watched any of the Star Wars movie around 225 people are not a fan of the Star Trek franchise.

2.3) Relationship between people’s demographics and their attitude to Star Wars characters

Peoples demographic have a very little impact on how they view a character because the result of all the visualization is very similar in all the cases for all the demographics. Luke Skywalker, Han Solo, Princess Leia Organa, Obi Wan Kenobi, Yoda, R2-D2 and C-3PO were the most favourable characters amongst all the characters and Jar Jar Binks, Emperor Palpatine and Boba Fett were some of the least favourable characters amongst all. But some interesting conclusions about the demographics can be made that Star Wars is most popular in people who have income in the range $50,000 - $99,999, also it is most popular in people who at least have some college or associate degree and lastly it is most popular in people who are from the region East-North Central, South Atlantic and Pacific.