Practical Data Science with Python COSC 2670/2738

Syed Haider Saleem

s3796258

16, April 2020

Data Preparation

We have been provided with a StarWars.csv file and the first step to get on with it is definitely to load the data using appropriate functions.

```
# Loading Data
import pandas as pd
StarWarsDF = pd.read_csv("StarWars.csv", encoding = "unicode-escape", header=[0], skiprows=[1])
```

The header for our data is composed of two rows, but for the purpose of our own ease and understanding I have only loaded the first row as a header and skipped the second row but to preserve that information I have renamed the data using previous information.

Task 1.2: Check data types

Once our data is loaded into the dataframe it's time to ensure whether our data is loaded in the right format, check its data types, columns , and most importantly its shape to make sure all the columns and rows are loaded correctly.

Our data had 1186 rows and 37 features (excluding ResponderID which was not required) having the right data types of object and floats where required. We were required to change the names of our columns for the purpose of ease and understanding as mentioned above.

Task 1.3: Typos

Typos are one of the most common issues in datasets where free text is provided to users in which case users do sometimes enter those terms which are not grammatically incorrect hence we need to check for that as well. We identified those using “value_counts()” and looping through each column and some typos that we ran into were “Yess” instead of “Yes” and “Noo” instead of “No” etc.

For resolving it we created a mask from which the data frame was passed through and typos were removed.

Task 1.4,1.5 : Extra-whitespaces and Upper/Lower case

Extra-whitespace issue was resolved by using the “strip()” function which strips the unnecessary whitespace from both left and right side of the String value. For the purpose of converting all values to upper case we used the “upper()” function and the reason why it was required to convert all the values to upper was because few values were in small cases and few in upper which interns make our dataframe treat it like different values which we do not want.

We applied those functions by looping through each column of the dataframe and we also needed to make sure to apply these functions to those columns who can be converted to string like the object type and not on float types.

Task 1.6: Sanity checks

Sanity checks are applied to validate our data using simple rules and checks so that the particular data could follow some specific rules. In our case we applied sanity checks by looping through each column and printing value_counts() for each column and correcting it afterwards manually, reason being that most of our columns life for example age is in object type and it has ranges instead of float values hence we cannot apply conditional checks. So what we did is identify values that looked problematic and removed them manually.

Task 1.7: Missing values

One of the major data preprocessing tasks includes dealing with missing values as a real life dataset do have missing values and you need to treat them accordingly.

In our case we primarily have two data types which are object and float and hence we need to deal with them separately. In case of float values we replaced the missing values with the mean of that column because we wanted to maintain the behaviour of data in that column even though that column has discrete values of 1 to 6 and we could have simply used median as well but that could severely affect the behaviour of our data when we were taking the mean of rankings of movies in our task 2.1 also there are no outliers in those particular columns, hence we sticked with mean.

For the object columns we simply replaced the missing values with “No Response” hence were able to save the data while also keeping a note of those places where values were missing.

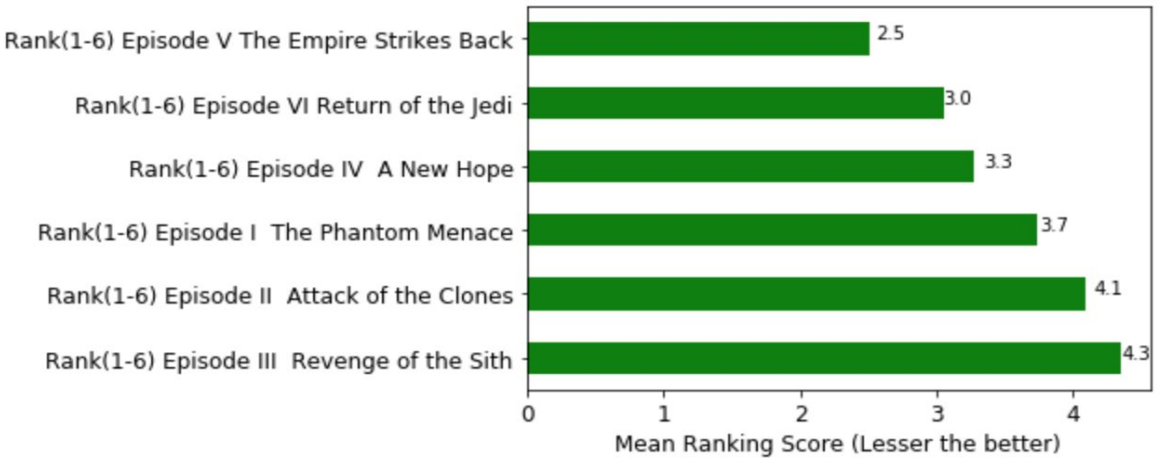
The reason why we did not go for removing the missing values at all was because we were worried that we could lose some vital information along with quite a percentage of our valuable data.

Data Exploration:

Now once we have preprocessed our data now it is ready for exploration using the power of visualisation to make sense of your data and to find useful trends that we might not be able to find otherwise.

Task 2.1: Explore a survey question

Now the survey question that we needed to answer was **“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”**. For this purpose what we have done is taken the mean of each column with each representing the ranking of each movie, what this does is gives us the mean (average) ranking for each movie which could be used to compare and then rank each movie column in that order. For the visualisation part I have used horizontal barchart as it is one of the most simple and powerful visualisation type.



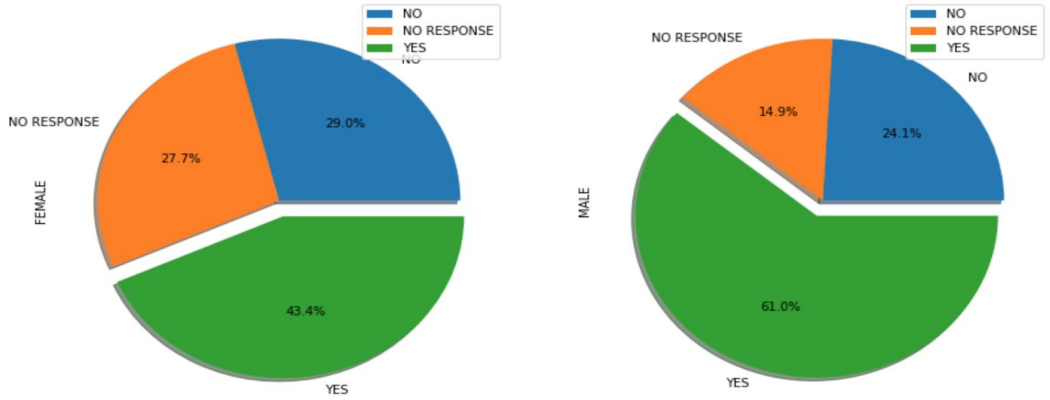
From this barchart we can clearly see that **“Episode V The Empire Strikes Back”** is the most prefered movie among the viewers in our data with the lowest mean value of 2.5.

Task 2.2: Relationships between columns

Relationship1:

Hypothesis : Star Wars has bigger male fanship as compared to female’s.

For making an analysis of this hypothesis we need **'Gender**’ and **'Do you consider yourself to be a fan of the Star Wars film franchise?'** columns, we groped the data using pandas groupby function and then plotted the results using pie chart.



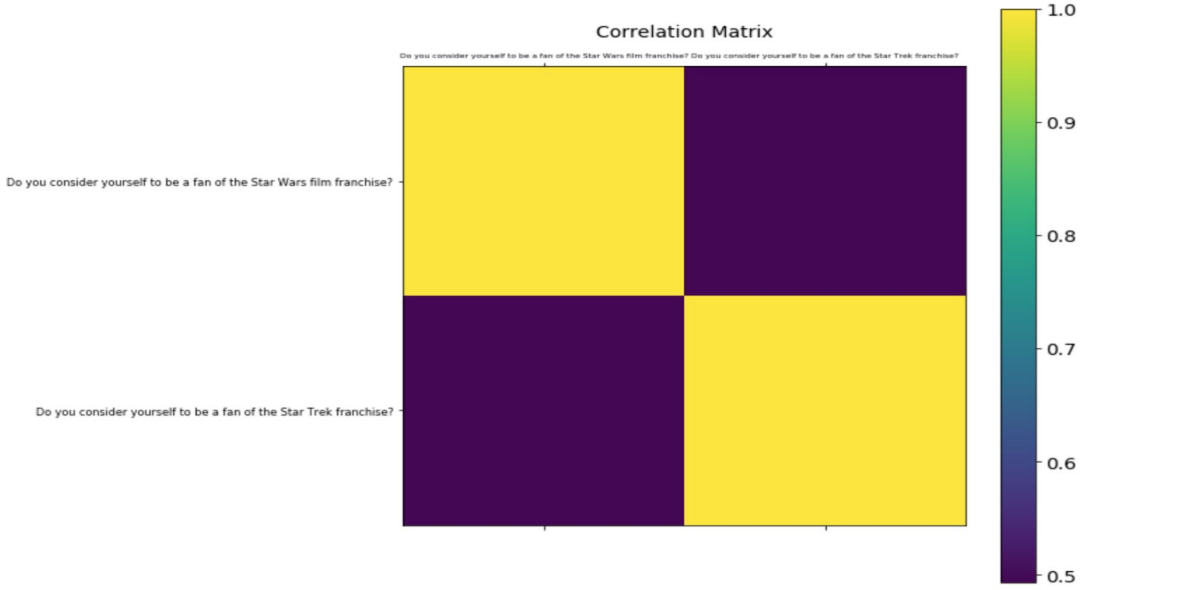
We can clearly see that male's are leading with 61% (yes) as compared to 43.4% (yes) in terms of answer to the question of are you a fan of Star War movies which proves our hypothesis and we can confidently say that Star Wars have a bigger male fanbase according to our survey.

Relationship 2:

Hypothesis: Star Trek fans are also the fans of Star Wars Film franchise.

For this purpose what we are taking **"Do you consider yourself to be a fan of the Star Trek franchise?"** and **"Do you consider yourself to be a fan of the Star Wars film franchise?"** as our target columns and after taking their correlation with each other the value comes out to be

	Do you consider yourself to be a fan of the Star Wars film franchise?	Do you consider yourself to be a fan of the Star Trek franchise?
Do you consider yourself to be a fan of the Star Wars film franchise?	1.000000	0.493177
Do you consider yourself to be a fan of the Star Trek franchise?	0.493177	1.000000



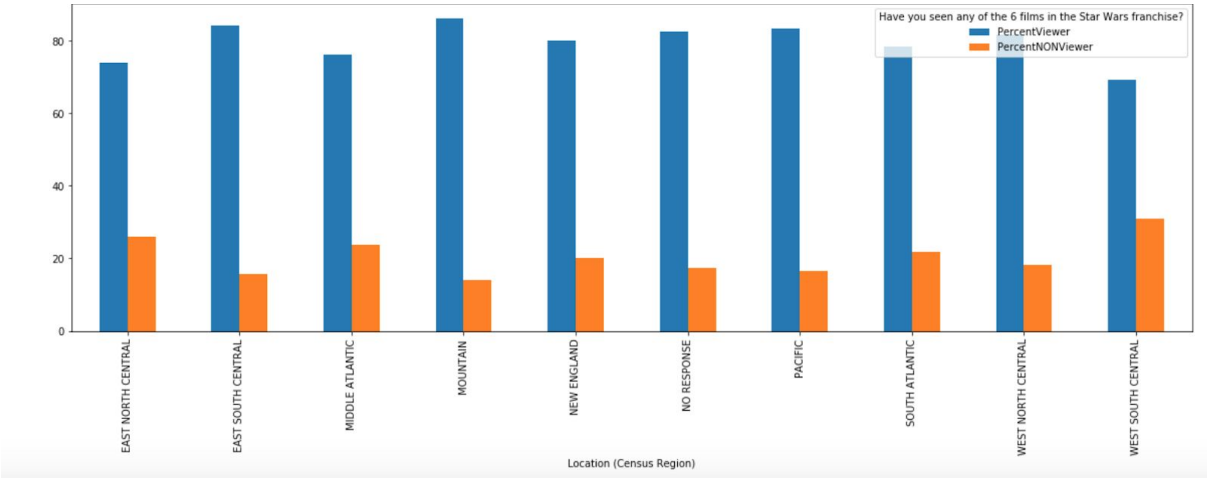
This correlation matrix gives us the value of 0.49 (almost 0.5) which means that there is a moderate positive correlation between the two columns which in other words means that

there is a moderate chance that Star Trek fan will also be a fan of Star War Movie series but we cant say it with a 100% certainty hence our hypothesis that a Star Trek fan will be a definite fan of Star War series is not proved yet with this data.

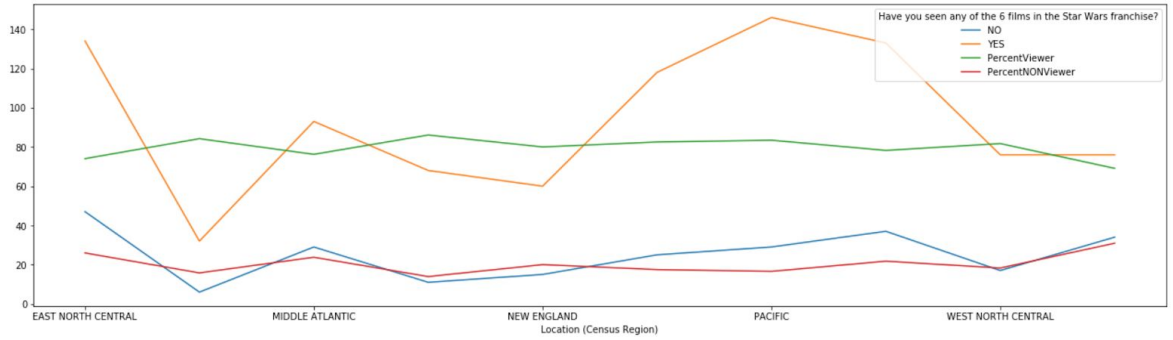
Relationship 3:

Hypothesis: Star Wars is watched mostly in the pacific region

For working on this hypothesis we would be needing **'Location (Census Region)'** and **'Have you seen any of the 6 films in the Star Wars franchise?'** as our target columns. Then we will be using groupby to obtain the resulting data which clearly says that the count in pacific region is the highest. But then when we dig deep we realise that it is only because we have more examples from pacific region in our survey data hence we need to create another column having the percentage of viewers corresponding to the total number of people surveyed in that specific region and this is what we came up with.



We can see that the viewership is almost evenly distributed among different regions (in terms of percentage of viewers saying yes and total viewers asked) and to get a more clear picture of the steady nature of the graph let's have a look at the line plot.



The green line shows the percent of viewers saying yes to the question that they have watched and we can see that the graph is steady hence rejecting our hypothesis that it is mostly watched in pacific region and telling us that the viewership of StarWars is just about evenly distributed throughout the world.

Task 3.2.3: Data Exploration

In this subsection we were required to explore the relationship between people's demographics and their attitude towards the Star War characters.

Method of Exploration

For this purpose we created multiple horizontal barcharts each corresponding to each demographic for example “Age” and “Household Income” and plot them against each of the Star War characters. We are plotting on the basis of count of that particular value which might be biased based on the number of observations taken from that particular area or set hence as a result we are provided with multiple graphs for each character.

Findings

What we find is that almost in all cases the attitude of people with different demographics about a character is the same in general with minor changes.

Han Solo, Luke Skywalker, Princess Leia, Obi Wan, C-3P0, R2 D2, Yoda are the characters which are generally liked by people of all demographics with almost the same intensity while demographic of people of age bracket 45-60 having salary of around 50k and in their bachelors are a bit more biased towards these characters which again could be because of the sample size.

While characters like **Anakin Skywalker, Lando, Boba Fett, Padme Amidala** are viewed as neutral characters having mixed opinions about them in general.

Darth Vader, Jar Jar Binks and Emperor Palpatine are generally disliked characters as they mostly lie in somewhat to very unfavorable range.