Analysis for respondent opinions on Star Wars series

Data here used is the survey conducted on the random sample population with 1186 respondents out of which 963 respondents watched the movie. Here, I will clean the data where it is necessary and the explore it. By cleaning it becomes easy for the preparation of data for further analysis and then by exploring and visually plotting I we can come across some facts or insights from the data.

Task 3.1 --- Data Preparation :

1. **Checking data types :**
   Initially, data types are not properly and I am going to change them in the further tasks leaving this task
   here. I checked it using **dtypes** attribute on the dataframe. In the further steps I changed the data using **astype()** function.
   Syntax
   1. **Dataframe.dtypes # gives data types of all columns.**
   2. **Dataframe["Column name"].astype(type) # converts the column data into the specified data type.**
2. **Typos - Missing values :** Here I checked for typo errors or missing values and replaced accordingly. Initially, I used **unique()** function to check what values are present in each column. From that I got the typos and missing values, I changed them using the function **replace()**.
   Syntax
   1. **Dataframe.["Column name"].unique() # gives values present in the columns.**
   2. **Dataframe["Column name"].replace(np.nan, "Assumed Value", inplace = True)# for missing values.**
   3. **Dataframe["Column name"].replace("Typo", "Actual Value", inplace = True)for typos.**

   Here I Assumed some values in the place of missing values in order to ensure the favourable conditions while exploring the data. i.e., in order to attain smooth flow of analysis. The assumptions that I made are,
   1. For missed ranking values for the columns 9 to 14 i.e., Episode rankings I replaced by 0. As unranked indirectly implies 0 ranking.
   2. Similarly, for columns 15 to 28 i.e., rating for Characters also I replaced it with 0. (But I converted all ratings into integers, which comes in the later steps.)
   3. For all other missing values, I replaced with Not Answered.
   4. There is on special case of typo error in Age Category which is displayed as 500. No Age category will be of that value. So observing other values of that particular columns I came to know that "< 18" is missing. So I renamed 500 as < 18.

   Also I removed the values that are missing or invalid in RespondentId as RespondentID is the identification fur the observation or particular row. It is mandatory that RespondentID is present. I done this using **pd.notnull().**
   Syntax : dataframe = dataframe[pd.notnull(dataframe["Column Name"])]# Removes the row with invalid respondent.
   Now renaming columns which are missed or unnamed.
   1. Now changing column names which hold the Star War movies that the respondents have seen among the entire series as they are not properly named using rename() function as follows.

   "Which of the following Star Wars films have you seen? Please select all that apply." : 'The Phantom Menance',
   "Unnamed: 4" : 'Attack of the Clones',
   "Unnamed: 5" : 'Revenge of the Sith',
   "Unnamed: 6" : 'A New Hope',
   "Unnamed: 7" : 'The Empire Strikes Back',
   "Unnamed: 8" : 'Return of the Jedi'
   And then the data values as,
   "Star Wars: Episode I The Phantom Menace" : "YES",
   "Star Wars: Episode II Attack of the Clones" : "YES",

"Star Wars: Episode III Revenge of the Sith" : "YES",
"Star Wars: Episode IV A New Hope" : "YES",
"Star Wars: Episode V The Empire Strikes Back" : "YES",
"Star Wars: Episode VI Return of the Jedi" : "YES",
np.nan : "NO" # Changed this because I am going to convert it into Boolean values later.

I then mapped these values into the respective columns.

2. Now renaming the columns holding the rankings given by the respondents for the movies among the entire series i.e., columns from 9 to 14 and converting them into float type. Columns are renamed as follows.

"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." : "The Phantom Menance - ranking",
"Unnamed: 10" : "Attack of the Clones - ranking",
"Unnamed: 11" : "Revenge of the Sith - ranking",
"Unnamed: 12" : "A New Hope - ranking",
"Unnamed: 13" : "The Empire Strikes Back - ranking",
"Unnamed: 14" : "Return of the Jedi - ranking",
"Do you consider yourself to be a fan of the Expanded Universe?ÂŒÃ¡" : "Do you consider yourself to be a fan of the Expanded Universe?" # This column is renamed as there is a typo present in the column name.

3. Now let us rename the Character rating column names. i.e., from 15 to 28.

"Please state whether you view the following characters favorably, unfavorably, or are unfamiliar with him/her." : "Han Solo",
"Unnamed: 16" : "Luke Skywalker",
"Unnamed: 17" : "Princess Leia Organa",
"Unnamed: 18" : "Anakin Skywalker",
"Unnamed: 19" : "Obi Wan Kenobi",
"Unnamed: 20" : "Emperor Palpatine",
"Unnamed: 21" : "Darth Vader",
"Unnamed: 22" : "Lando Calrissian",
"Unnamed: 23" : "Boba Fett",
"Unnamed: 24" : "C-3P0",
"Unnamed: 25" : "R2 D2",
"Unnamed: 26" : "Jar Jar Binks",
"Unnamed: 27" : "Padme Amidala",
"Unnamed: 28" : "Yoda"

Renaming the Character rating values for statistical analysis, and mapped to their respective columns.

"Very unfavorably" : -2,
"Somewhat unfavorably" : -1,
"Neither favorably nor unfavorably (neutral)" : 0,
"Somewhat favorably" : 1,
"Very favorably" : 2,
"Unfamiliar (N/A)" : 0,
np.nan : 0 # Missed values changed to 0 as mentioned above.

Now I checked for null values. After checking I came to came to no that there are no null values present i.e., no missing values and the data is perfect.

Syntax : dataframe.isnull().sum()

3. Checking for white spaces :

Now I checked for white spaces using str.isspace() function. Luckily there are no white spaces too. I checked only all columns except integers and floats as they won't have white spaces.

Syntax : dataframe["column name"].str.isspace().sum()

4. Upper Case conversion :

Data values are now converted into Upper case using upper() function. All columns other than numeric are converted here.

As the data is now cleaned I prepared it for my analysis in task 2. I am changing "YES, NO" to "True, False" i.e., into boolean values for easier interpretation and statistical calculations as follows,

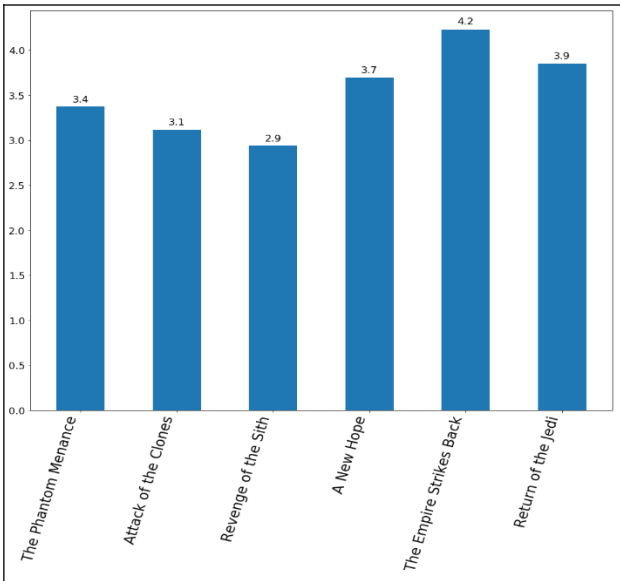
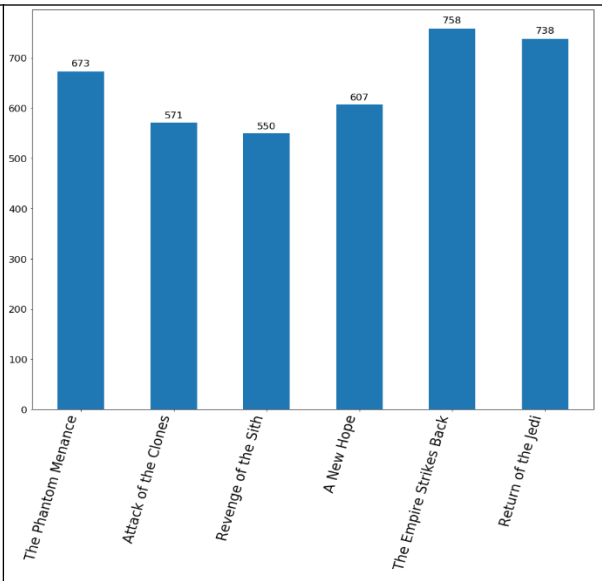
"YES" : True,
"NO" : False,
"NOT ANSWERED" : "NOT ANSWERED" (For seen/not-seen and fan/not-fan columns.)
np.nan : False # Assuming Boolean variable for better results(For columns episode ranking columns.)

Finally, the data is good and data types are fine. Thus, cleaning came to an end. Now let us explore the data and continue the analysis.

Task 3.2 – Data Exploration.

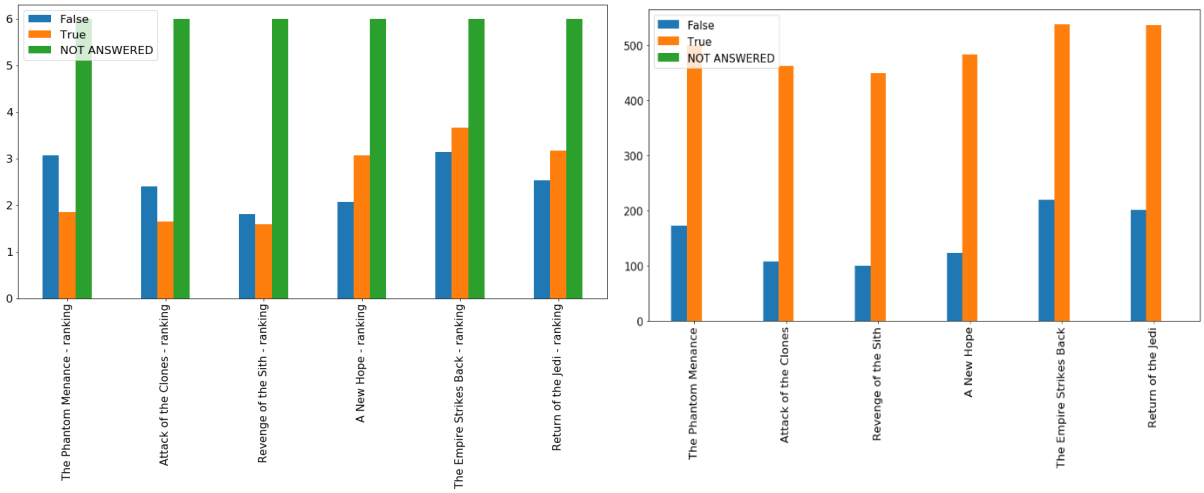
3.2.2 - Exploring the cleaned data - Data Exploration 1 :

Now let us analyse how people rated the star war series. To do this 1st we have to plot a bargraph with ratings (integer rankings as float ratings) of star war movies. Here **The rankings are subtracted with 6 in order to make the highest like movie or episode have the highest rating out of 6. 0***. Also let us check with the number of recipients voted are already seen those movies. That means **most watched movie**. Out of 1186 recipients there are 963 recipients who watched at least 1 movie and among them how many people watched each episode. For this consider a bar plot with Episodes on x – axis and ratings on y – axis.

<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>The ratings of the episodes are as follows :</td>
<td>Number of people who watched each episodes are as follows :</td>
</tr>
<tr>
<td>Episode 1 : 2.3 / 6.0
Episode 2 : 1.9 / 6.0
Episode 3 : 1.7 / 6.0
Episode 4 : 2.7 / 6.0
Episode 5 : 3.5 / 6.0
Episode 6 : 3.0 / 6.0</td>
<td>Episode 1 : 673 / 963
Episode 2 : 571 / 963
Episode 3 : 550 / 963
Episode 4 : 607 / 963
Episode 5 : 758 / 963
Episode 6 : 738 / 963</td>
</tr>
</table>

From this we can say that Episode 5 i.e., **"The Empire Strikes Back"** is the **highest** rated movie with 3.5 rating which means this episode is liked by most of the people who have responded. Followed by "Return of Jedi - Episode 6" with rating 3.0. Coming to the least liked movie which is **"Episode 3 Revenge of the Sith"** with **1.7** rating. We cannot conclude these as the original ratings from this as this survey is taken from a random sample of people. There might be chances that the sample might be Episode 5 favourable sample and more over this may or may not be the most liked episode if the sample changes. But for this Sample survey this is the report.

From this we can say that **Episode 5 i.e., "The Empire Strikes Back"** is the highest watched movie with, followed by "Return of Jedi – Episode 6". Coming to the least watched movie which is **"Episode 3 Revenge of the Sith"** with only 550 viewers. It is more likely to predict that the movie with highest number of viewers is more liked because they saw it. There may be a chance of liking Revenge of Sith too if this is seen by the majority. This can also be implied as the most popular episodes were watched by most of the respondents. From this we can say that if the movie is less attractive or least liked if the number of people watched the movie is less. "Seen by people" is directly proportional to "How good the movie is."
Now let's check the viewers count based on their gender. I just wanted to know whether there was a significant difference in movie rankings based on gender. This graph is removed to adjust the space.
It is clear that viewers are mostly males for the last 3 movies and for the 1st 2 females are the most viewed. Both Males and Females decreased their interest towards the series and majority stopped viewing the movies, this might be one of the reason for the less viewed count and the least rating of episode 3 compared to other movies. Both Episodes 3 and 5 are on the same general agreement according to male and female where they both proportionately watched them. Coming to Not Answered category, they rated favourable to all episodes. This might be a chance of having their disinterest towards the survey. Which makes me to find alternatives in analysing the data.
Now let us see these rankings based on **Fan/Non-Fan** category.

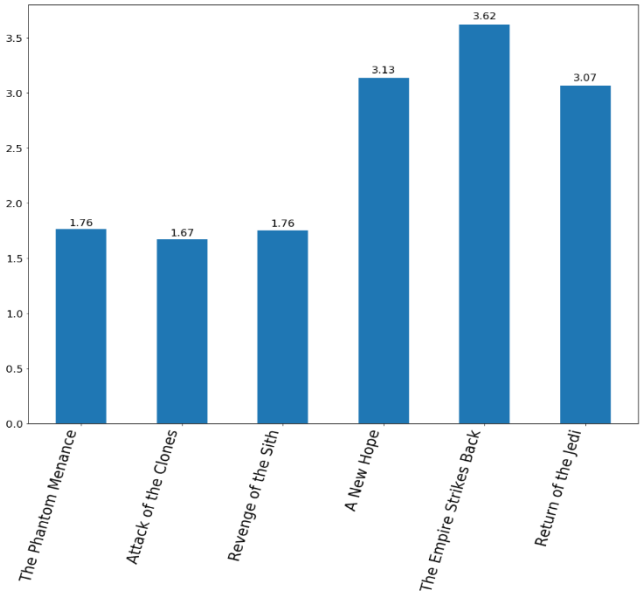


Fans are more likely that they watched the original movies more, also most of them have seen all 6. Coming to non-fans, they mostly preferred original movies than prequels. Now let us consider ratings with this classification.
Fans : Again they showed their love towards original series by rating the originals high.
Non - Fans : They given ratings for episodes 1 and 5 mostly similar and leaving this point to consider for analysis. They might have given some proper ratings as they are not favourable to any of the episodes leaving their opinions genuine (Just an assumption).
Not Answered : Similar to Gender (Coming to Not Answered Category, they rated favourable to all episodes. This might be a chance of having their disinterest towards the survey.)
At this particular point only 15 to 20 % difference is present from the most viewed to the least viewed movies by fans, where-as coming to non-fans the difference is nearly 60%. This can be taken as an observation for analysing the data. From this we can say that non – fans are more unlikely to watch at least one movie. This helps us in leaving the consideration of both non – fans and fans again for the next analysis.

Now let us check number of people seen each movie and sample people who saw all movies for better analysis
From the below analysis we can also predict how people generally rate the movies. Here, the respondents splitted into two groups namely,

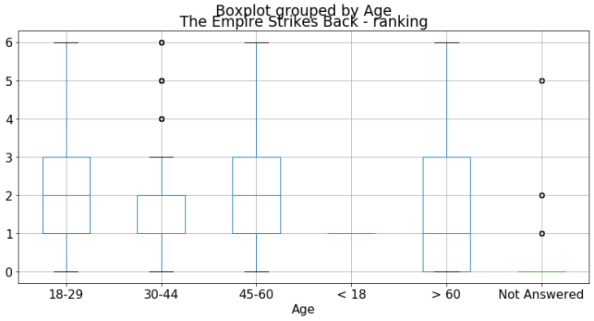
1. Original followers -- Episodes 4, 5, 6
2. Prequel followers -- Episodes 1, 2, 3

Respondents who voted for The Empire Strikes Back : Episode - 5 as their favourite are more likely to rate A New Hope : Episode - 4 and in turn rated Return of the Jedi : Episode - 6 too. Similarly, people who voted for The Phantom Menace : Episode - 1 as the most liked film were more likely to rate Episodes 2 and 3. This is just the point of views of some random people and might vary if the sample changes. There is also a chance that respondents are voting the movies blindly even if they didn't watch them. They may vote based on their own perspectives or even based on the word of mouth from the fellow people. This results in the flaw in the analysis that we attained so far. So the result might be more accurate if we only consider the respondent votings who have **seen all the 6 movies**. As The Empire Strikes Back : Episode - 5 is seen by more people it might get the highest rating as the respondents who have seen it might not have seen the other episode considering all responses not only the responses from respondents seen all 6.

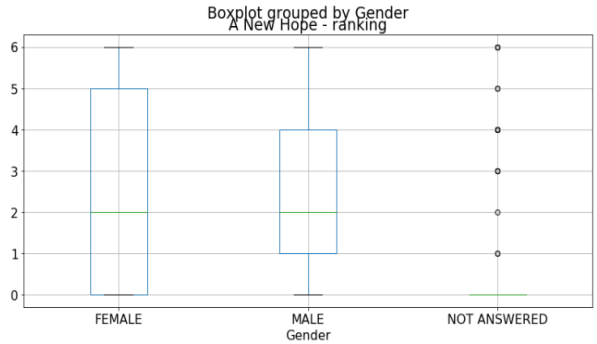


Task 3.2.2
Visualisation based on Plausible hypothesis for 3 pairs of columns - Data Exploration 2

1. **Pair 1 - Visualisation** Consider the highest rated episode with age.



Mostly all voted it prety well, but coming to categories people aged more than 60 has a wide range of voting compared to 18-29 and 45 - 60 which are almost similar. This implies people aged 18 to 29 and 45 to 60 have similar opinions. coming to mid aged that is 30 to 44 have the most similar ratings among them and on average they voted high for this episode. Coming to Not answered as they are disinteredted plot is like that with outliers maximising leaving the plot negligible.

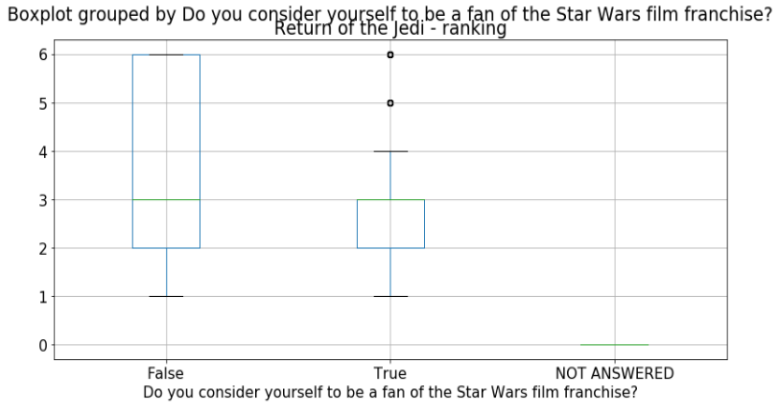


2. **Pair 2 – Visualisation**

Consider the previous episode of the highest rated episode from the same triology with gender.

Male : As usual rated highest.
Female : Had a wide range of ratings.
Not Answered : Disinterested.

3. **Pair 3 – Visualisation**
   Consider the next episode of the highest rated episode from the same triology with Fan/Not-Fan.



Fans : As usual voted high as it is the original hypothesis.
Non - Fans : Had a wide range of voting as the are not favourable with original series.
Not Answered : As usual Disinterested.

Task 3.2.3

Demographic analysis - Data Exploration 3(For visualisation kindly refer python notebook)

Here I checked for the relations between people's demographics and their attitude towards Star Wars Characters. The visualisations in this session are not attached due to spacing issues. Please kindly refer to assignment1.ipybn file

1. **With respect to Gender : Subplots -> Bar Graph**
   Gender wise Character ratings(Popularity/Attitude of Respondents)
2. **With respect to Age : Subplots -> Bar Graph**
   Age wise Character ratings(Popularity/Attitude of Respondents)
3. **With respect to Education : Subplots -> Bar Graph**
   Education wise Character ratings(Popularity/Attitude of Respondents)
4. . **With respect to Location : Subplots -> Bar Graph**
   Location wise Character ratings(Popularity/Attitude of Respondents)

Popularity based rating : Bar Graph.

Now I made a visualisation to see how each Star War characters were rated by respondents based on their popularity ratings that are changed during data cleaning.
From the graph we can say that Han Solo is most liked among all and Jar Jar Binks is the most disliked character among all Star War characters with negative rating. This might be enough to show how popular the characters were. Both more precisely and more visually let us analyse what is the attitude of respondents towards each of these characters by visualising a pie - chart.
For the pie chart I am taking polarity, to show Respondents *Attitude - Possitive/ Negative / Neutral* towards Star Wars characters.

Overall audience attitude based on polarity : "Pie Chart"

Favourable – Possitive **Unfavourable - Negative**

From these pie charts we can say that Protogonists(i.e., main characters around whom the plot runs) like Han Solo, are more liked than the characters with negative roles like Jar Jar Blinks are more hated.