

1.DATA PREPERATION

1.1 STRUCTUR THE COLUMNS

RespondentID	Have you seen any of the 6 films in the Star Wars franchise?	Do you consider yourself to be a fan of the Star Wars film franchise?	Which of the following Star Wars films have you seen? Please select all that apply.	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	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.	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unn	
0	NaN	Response	Response	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	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
1	3.292880e+09	Yes	Yes	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	3	2	1	4	

The structure of initial data does not make much sense. There are few issues that catches attention on the surface:

1. The column headings do not make much sense.
2. The first row of the dataframe is not an observation rather is the options for the provided questions.

Initial browsing through the data reveals a pattern that a column containing a question is followed by their available responses as Unnamed, but the actual options for that question is present in the 1st row of dataset. Also, using the question and options provided as it is will not be appropriate for analysis due to their large size.

Solution: To tackle the above stated problem, I renamed all the columns in a concise way. To control any information loss due to our imputation and to facilitate use of original information later in the analysis I prepared a dictionary.

To structure the columns properly I removed the 1st row with index [0] and preserved the options in a dictionary called **options**.

RespondentID	Q1	Q2	Q3/O1	Q3/O2	Q3/O3	Q3/O4	Q3/O5	Q3/O6	Q4/O1	Q4/O2	Q4/O3	Q4/O4	Q4/O5	Q4/O6	Q5/O1	Q5/O2	Q	
1	3.292880e+09	Yes	Yes	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	3	2	1	4	5	6	Very favorably	Very favorably	fav

Result

1.2 UNNECESSARY OBSERVATIONS

Going a little bit further in my analysis I noticed a trend that any respondent who have not answered the Q2 have not answered rest of the questions as well, so all the columns contains NaN.

RespondentID	Q1	Q2	Q3/O1	Q3/O2	Q3/O3	Q3/O4	Q3/O5	Q3/O6	Q4/O1	Q4/O2	Q4/O3	Q4/O4	Q4/O5	Q4/O6	Q5/O1	Q5/O2	Q5/O3	Q5/O4	Q
11	3.292638e+09	Yes	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Solution: A straight up elimination of all these observations were done as they can't provide any meaningful information in our analysis. But I am going to keep respondents who have answered Q1(Have you seen any of 6 films...) as No because doing so makes sense in two ways:

1. Respondents who have not watched any movies are going to skip all answers about movies.

2. They can provide valuable demographic information and can provide interesting insights.

Before going any further in our analysis, I took a step and made a subset of responses containing Q1 as Yes and named new dataframe as **df_yes**. As stated above responses with Q2 as NaN does not contain any other information except demographics. Rest of the responses is still kept in original dataframe to be used in further analysis.

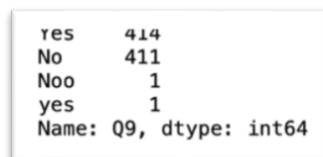
1.3 WHITESPACES

I took a strategy of scanning the data column wise to look for potential errors and problems. Only column Q1 came out as containing whitespaces in some of its observation. With values 'Yes' and 'Yes ' both present in the same column.

Solution: A simple **str.strip()** took out all the unnecessary whitespaces from the column.

1.4 TYPOS

Q2, Q8 ,Q9 and Gender columns had typographical errors which was creating an issue of



```
yes    414
No     411
Noo     1
yes     1
Name: Q9, dtype: int64
```

representing the same information in a different way.
The figure on the left shows such a case in column Q9.

Solution: I used a simple method in all of these cases taking example of the above stated problem in column Q9. I defined a dictionary with key as typos and values as the desired outcome of typos and used it with a **.replce()** for example : **df_yes['Q9'].replace({'yes':'Yes','Noo':'No'})**

1.5 CHANGING DATATYPES

- a) Q3/O1-Q3/O6 contains responses for Q3(which addresses whether or not respondent have watched the particular movie column is representing) contains name of the movie if the respondent answered yes and a NaN otherwise.
- b) Q4/O1-Q4/O6 contains ratings for different movies of the franchise but datatype of each column comes up as object which is needed to be changed to int.

Solution: To address 1st problem I and to make data more presentable I wrote a function **impute_val** to impute 1 in place of a yes response and 0 otherwise. To address the 2nd issue **.astype(int)** is used.

1.6 SANITY CHECK

While analyzing Age column an observation came up which contained an observation of Age as 500

Solution: Since, from all the possible solutions to treat this extreme value like imputing it with a mode(as column is categoric) removing it altogether as it provides misleading information was a logical choice.

1.7 MISSING VALUE

Two different approaches were implemented to treat missing values at two different situation.

1. **Manual imputation:** In Q4/O3 and Q4/O1 columns 1 missing value was found in each of them.

Q4/O1	Q4/O2	Q4/O3	Q4/O4	Q4/O5	Q4/O6	Q5/O1	Q5/O2	Q5/O3	Q5/O4
1	2	NaN	4	5	3	Very favorably	Somewhat favorably	Neither favorably nor unfavorably (neutral)	1

From close observation it comes out that the missing value should be remaining 1 rating which in this case is 6.

2. **Predicting missing value:** In pursuit to free our dataset from all null values when I checked null values in all the columns Household Income has a significant amount of NaN about 161 values. Replacing these values with mode or simply removing these observations can either lead to a misrepresentation of observation or loss in information from our dataset. So, a logical decision of using a classification prediction model was made to predict these absent observations.

This whole process was done in some simple steps:

1. Subsetting data in 2 separate dataset one containing observation where household income was NaN and one where all the data is present which will be used in building model.
2. Substituting all the nominal data with dummy variable and ordinal data with integer encoding.
3. Fitting a decision tree with our processed data and use cross validation to tune hyperparameters for the model.
4. Predicting missing data and substituting it back to main dataframe.

After imputing all the predicted value back to dataframe I am going to drop all the remaining observations containing null in any column.

Conclusion:

After doing all our pre-processing the original dataset reduced by a 16.76% and what's left is a clean dataset.

3. DATA EXPLORATION

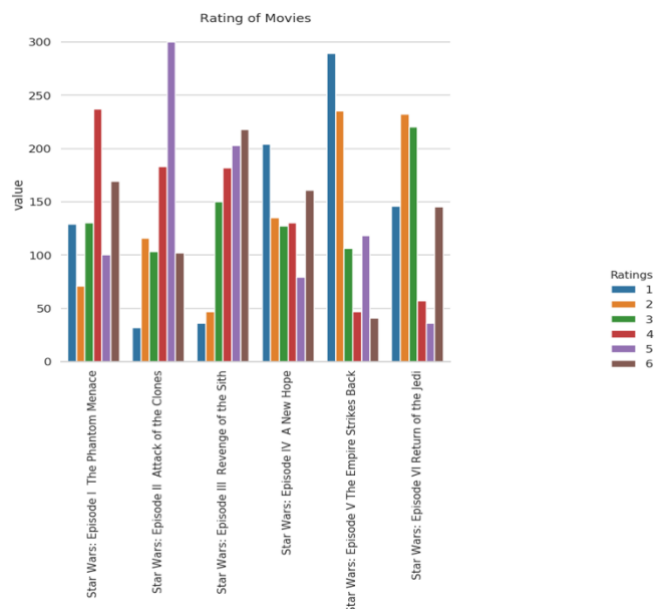
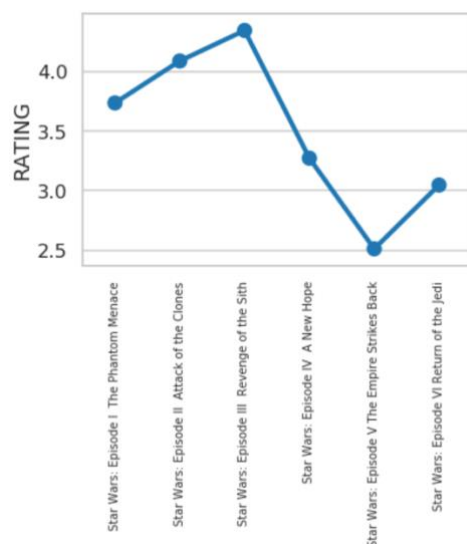


Figure1:average rating of movie

Figure 2: different rating of all movies

2.1 EXPLORING RATING OF MOVIES

Fig 1 shows average rating of movies in the franchise and an increasing trend can be clearly seen till the 3rd movie of the franchise after which ratings started improving drastically. Fig 2 corroborate the first conclusion as it can be clearly seen that the 5th movie in the franchise received most number of top rating.

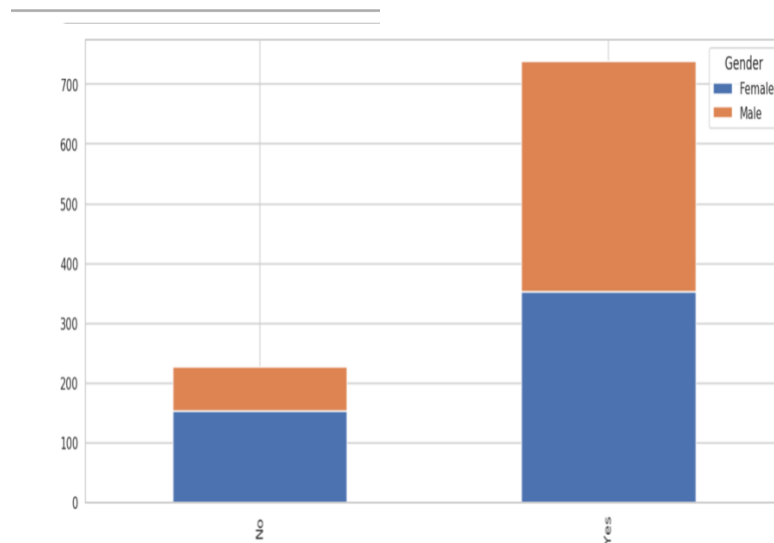
2.2 RELATIONSHIP BETWEEN COLUMN

1. **Ho:** The distribution of respondents who have watch or have not watched movies from Star Wars franchise is same among different genders.

Ha: The distribution of respondents who have watch or have not watched movies from Star Wars franchise is not same among different genders.

Gender	Female	Male
Q1		
No	152	74
Yes	352	386

The data shows that significantly lower number of male respondents have not watched any movie from the franchise and more of male respondents have watched movie from franchise. But, in order to prove the hypothesis, I am going to us **Chi- sq goodness of fit test.**



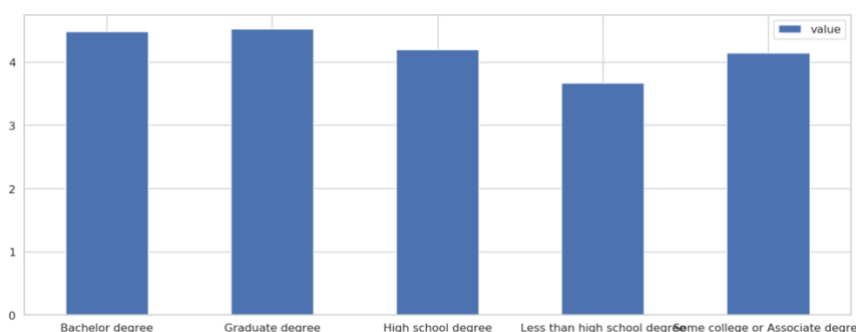
A Chi-square goodness of fit test was used to determine weather a respondent who have watch or have not watched movies from star wars franchise is same among different genders The test was statistically significant, $\chi^2 = 25.75$, $df = 1$, $p < .001$. This suggests that the distribution of respondents who have watch or have not watched movies from star wars franchise is not same among different genders.

2.

In my analysis Star Wars Episode III had the highest average rating so for this hypothesis test I am going build my hypothesis around it.

Ho: Different education levels have a difference in average rating for Star Wars Episode III.

Ha: Different education levels don't have a difference in average rating for Star Wars Episode III.



By looking at average rating for Star Wars Episode II it is hard to draw a conclusion weather averages differ in different education level.

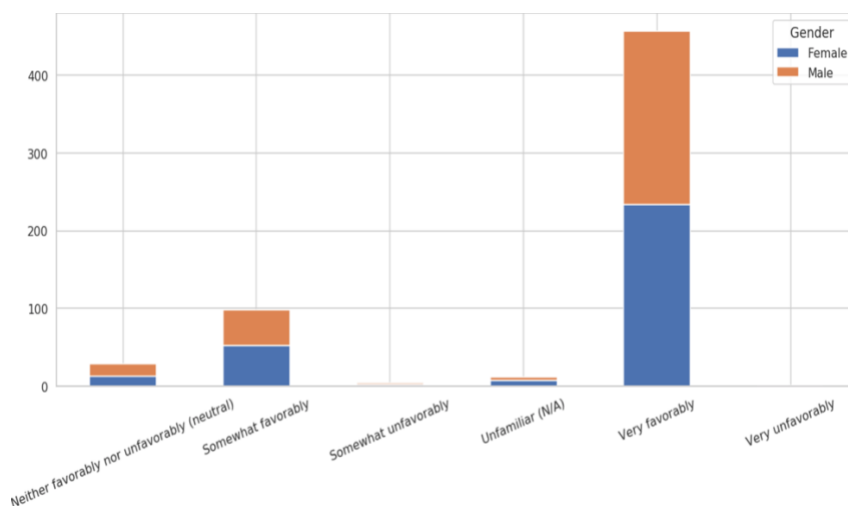
	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	4.0	22.816293	5.704073	2.903812	0.021122
Residual	733.0	1439.861214	1.964340	NaN	NaN

A one-way ANOVA test was used to check the variance of average rating for Star Wars Episode III among different groups. The test was statically significant with $F=2.9$ and $p\text{-value}<0.05$. Hence a conclusion can be drawn that there is a difference in average rating for Star Wars Episode III among different education levels.

- For my last hypothesis I am made a crossable to look for most popular Star Wars character in which Han Solo came up top with 549 vary favorable votes.

Ho: The distribution of respondents among different genders have same response towards Han Solo.

Ha: The distribution of respondents among different genders have different response towards Han Solo.



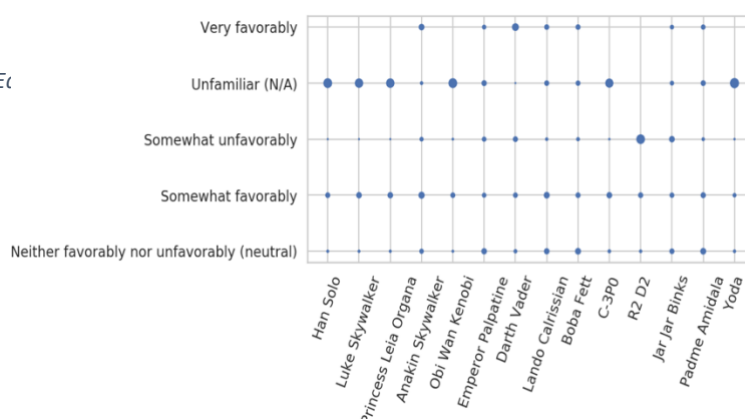
```
(1.902604474386418, 0.8624507905276039, 5, array([[ 14.86189684, 14.13810316],
[ 50.22296173, 47.77703827],
[ 2.04991681, 1.95008319],
[ 6.14975042, 5.85024958],
[234.20299501, 222.79700499],
[ 0.5124792, 0.4875208 ]]))
```

A Chi-square goodness of fit test was used to determine whether the distribution of respondents among different genders have same response towards Han Solo. The test was statistically insignificant, $\chi^2 = 1.9$, $df = 1$, $p > .001$. This suggests that the distribution of respondents among different genders have same response towards Han Solo.

2.3. RELATIONSHIP BETWEEN DEMOGRAPHICS AND CHARACTERS

To explore the relationship between demographics of people and their attitude towards characters in the franchise I had to take a different approach because to do so a large number of comparisons between various variables need to be done.

Figure 1E



Location (Census Region)	value
0 East North Central	560
1 East South Central	146
2 Middle Atlantic	430
3 Mountain	305
4 New England	276
5 Pacific	604
6 South Atlantic	541
7 West North Central	315
8 West South Central	330

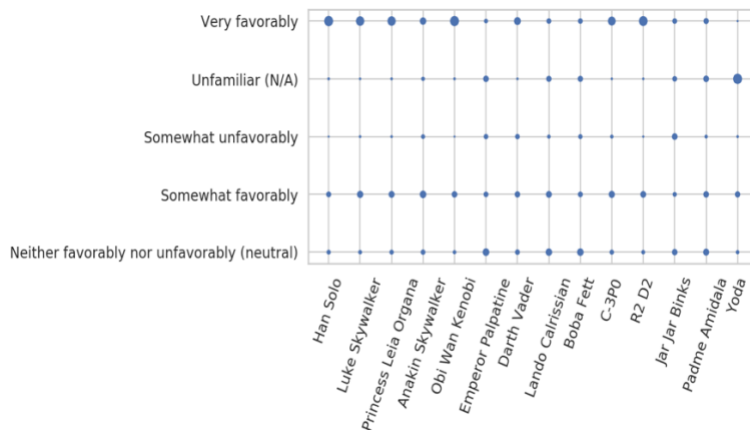


Figure 2pacific

As from the table above Pacific and East north central are the biggest followers of franchise but on further investigation of characters most of the characters were unfamiliar in East north central region. But, its is not true for pacific which have the largest viewership almost all characters are rated very favorable.

In my further analysis I found respondent with education level over a high school degree have a much better chance to find characters very favorable, below 1st image show result for respondents with bachelor's degree followed by respondents with high school degree.



Also, middle income (\$50,000-\$99,000) people found character as very favorably as compare to respondents with any other income level. Which is corroborated by the fact that about 40.8% of franchise followers lie in this income level.

