STAR WARS SURVEY DATA ANALYSIS

A PRESENTATION FROM A GALAXY FAR, FAR AWAY

SURVEY DATA BACKGROUND

- The data was taken from a survey conducted via SurveyMonkey on FiveThirtyEight.com
- There were 1186 total respondents
- The data includes questions about people's opinion and attitude about Star Wars films:
  - Whether or not respondents have seen any of the original six movies
  - Whether or not respondents consider themselves to be fans of the franchise
  - Which of the original six movies they had seen
  - How they ranked each of the six movies
- The data also includes how favorably they ranked the major characters
- Each respondent also provides demographic data about themselves, such as their age, gender, location, household income and education

DATA PREPARATION

DATA PREPARATION

- Changed the column names, and had separate columns for the sub-columns (check boxes)
- Filled all the columns of whether respondents had seen each of the six movies with ‘Yes’ or ‘No’ if they had checked the box for that movie

DATA PREPARATION

Typo correction

- Looped through each column and printed out the unique values for that column.

```
expanded_universe_familiarity: ['Yes' nan 'No']

expanded_universe_fan: ['No' nan 'Yes' 'Yess']

star_trek_fan: ['No' 'Yes' nan 'Noo' 'yes' 'no ']

gender: ['Male' nan 'Female' 'F' 'female' 'male']
```

- Made necessary corrections to these typos:

```
Unique value counts in the column star_trek_fan

No      639
Yes     426
yes       1
Noo       1
no        1
```

Before

```
Unique value counts in the column star_trek_fan

No      641
Yes     427
```

After

DATA PREPARATION

conversion to uppercase and removal of extra whitespace

- Removed extra whitespaces, so our exploration functions don't treat a response like **'no '** and 'no' differently since they mean the same thing
- Converted all text values to uppercase for parity

```
star_trek_fan: ['No' 'Yes' nan 'Noo' 'yes' 'no ']
gender: ['Male' nan 'Female' 'F' 'female' 'male']
```

Before

```
star_trek_fan: ['NO' 'YES' nan]
gender: ['MALE' nan 'FEMALE']
```

After

DATA PREPARATION

Sanity Checks

- Looped through each column and printed out the unique values for that column.

```
age: ['18-29' nan '500' '30-44' '> 60' '45-60']

household_income: [nan '$0 - $24,999' '$100,000 - $149,999' '$25,000 - $49,999'
 '$50,000 - $99,999' '$150,000+']

education: ['High school degree' 'Bachelor degree' 'Some college or Associate degree'
 nan 'Graduate degree' 'Less than high school degree']
```

- Made necessary corrections:

```
Unique value counts in the column 'age'

45-60    291
> 60     269
30-44    268
18-29    217
500        1
```

Before

```
Unique value counts in the column 'age'

45-60    291
> 60     269
30-44    268
18-29    217
-1         1
```

After

DATA PREPARATION

MISSING VALUES

- Missing values in the columns were dealt with by setting them all to (-1)
- Putting a -1 allows us to easily identify occurrences of missing values in each column via indexing in the code
- We did not take the column-wise mean because it doesn't work for most columns, as most columns aren't numeric, and the ones that are happened to represent ordinal data
- Taking the column-wise median is a possibility, however, it may not accurately reflect the truth since some columns are interdependent.

DATA PREPARATION

jupyter assignment1_presentation Last Checkpoint: Yesterday at 12:43 AM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Markdown

Out[30]:

	respondent_ID	seen_any_movies	consider_themselves_fan	seen_episode_I	seen_episode_II	seen_episode_III	seen_episode_IV	seen_episode_V	se
0	3292879998	YES	YES	YES	YES	YES	YES	YES	
1	3292879538	NO	-1	NO	NO	NO	NO	NO	
2	3292765271	YES	NO	YES	YES	YES	NO	NO	
3	3292763116	YES	YES	YES	YES	YES	YES	YES	
4	3292731220	YES	YES	YES	YES	YES	YES	YES	
...	...	...	...	...	...	...	...	...	
1181	3288388730	YES	YES	YES	YES	YES	YES	YES	
1182	3288378779	YES	YES	YES	YES	YES	YES	YES	
1183	3288375286	NO	-1	NO	NO	NO	NO	NO	
1184	3288373068	YES	YES	YES	YES	YES	YES	YES	
1185	3288372923	YES	NO	YES	YES	NO	NO	YES	

1186 rows × 38 columns

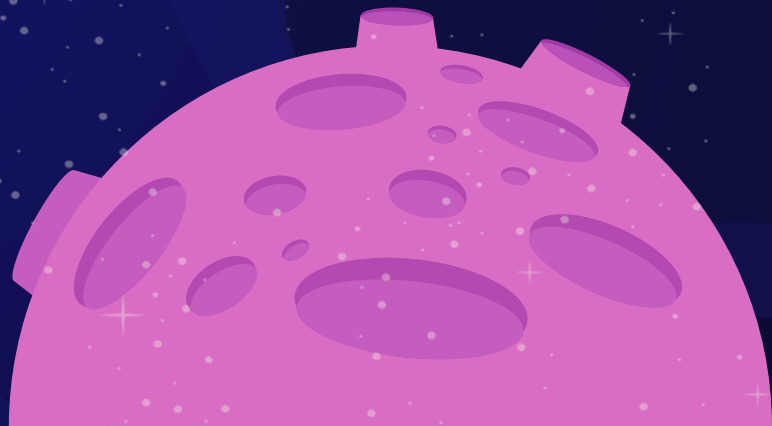
HOW DO PEOPLE RANK STAR WARS MOVIES?

DATA EXPLORATION

First checked if people had ranked the episodes, even if they hadn't seen them.

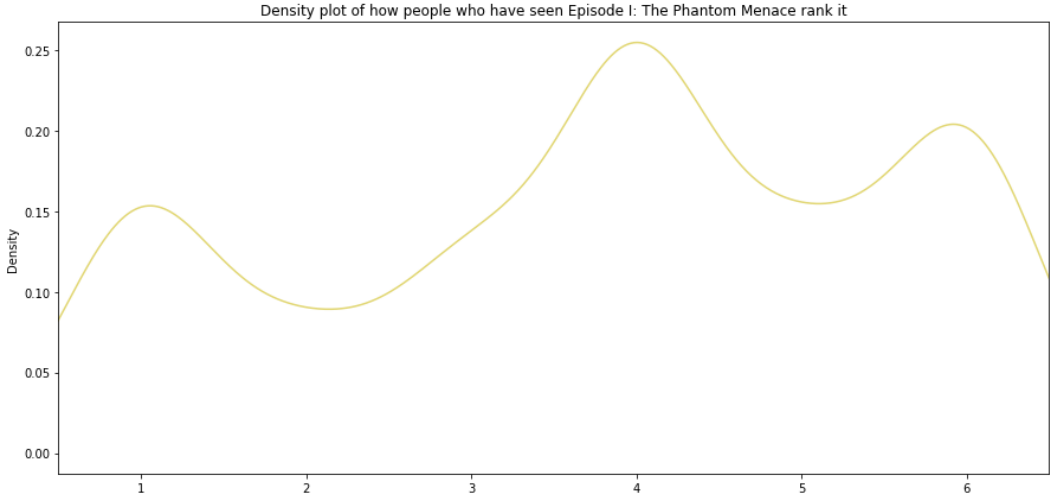
As an example, for Episode I, around 162 respondents marked that they had not seen this episode yet ranked it.

Thus we do not consider these rows when doing our analysis of the rankings as people who haven't seen the movie shouldn't be allowed to rate it.



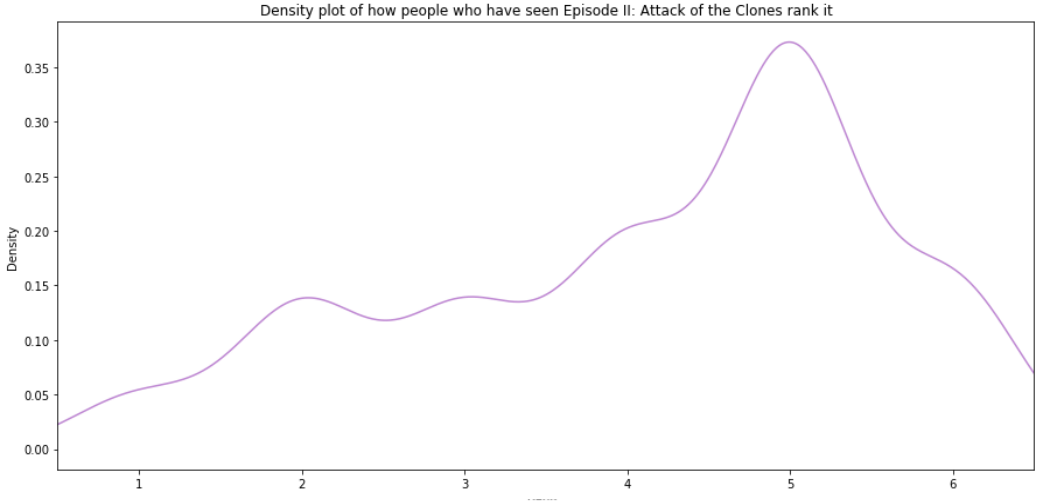
How people rank

Episode I: The Phantom Menace



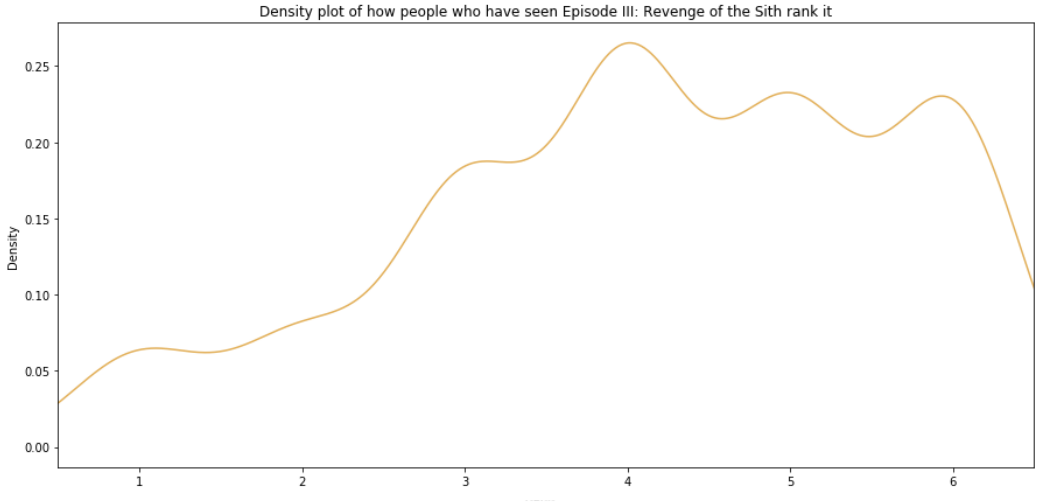
How people rank

Episode II: Attack of the Clones



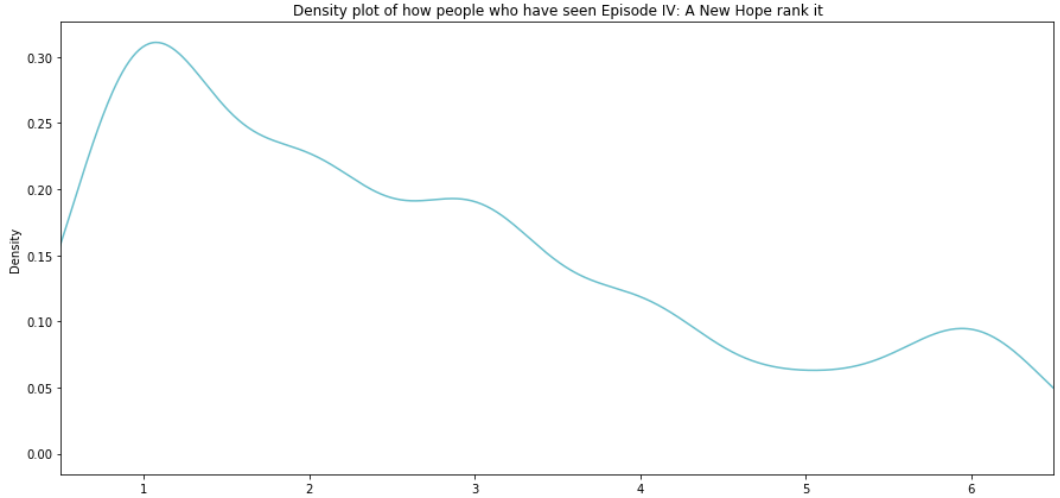
How People Rank

Episode III: Revenge of the Sith



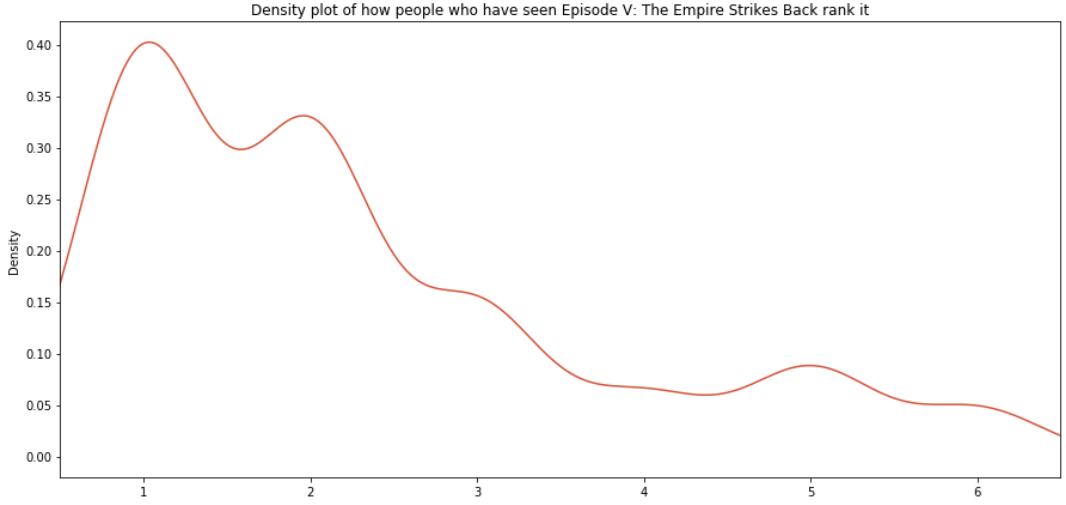
How people rank

Episode IV: A New Hope



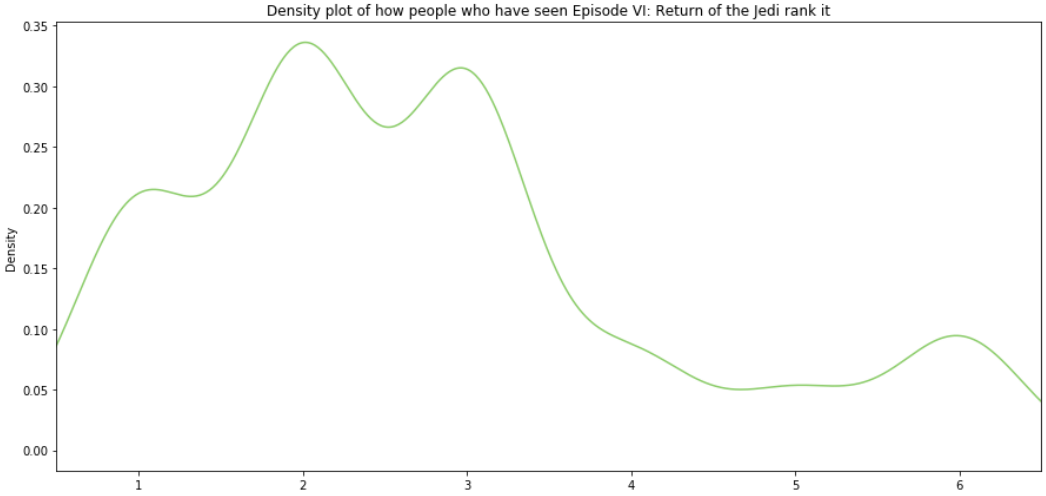
How people rank

Episode V: The Empire Strikes Back



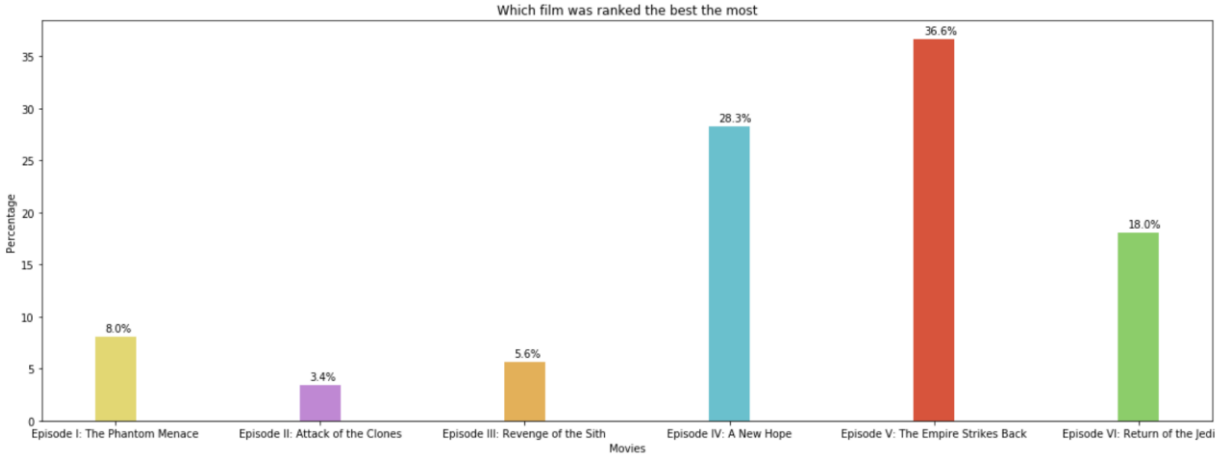
How people rank

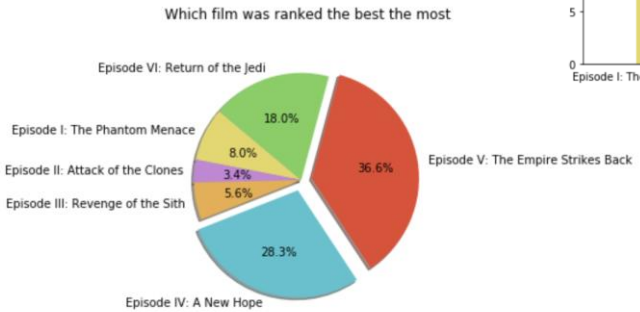
Episode VI: Return of the Jedi



BEST STAR WARS MOVIE?

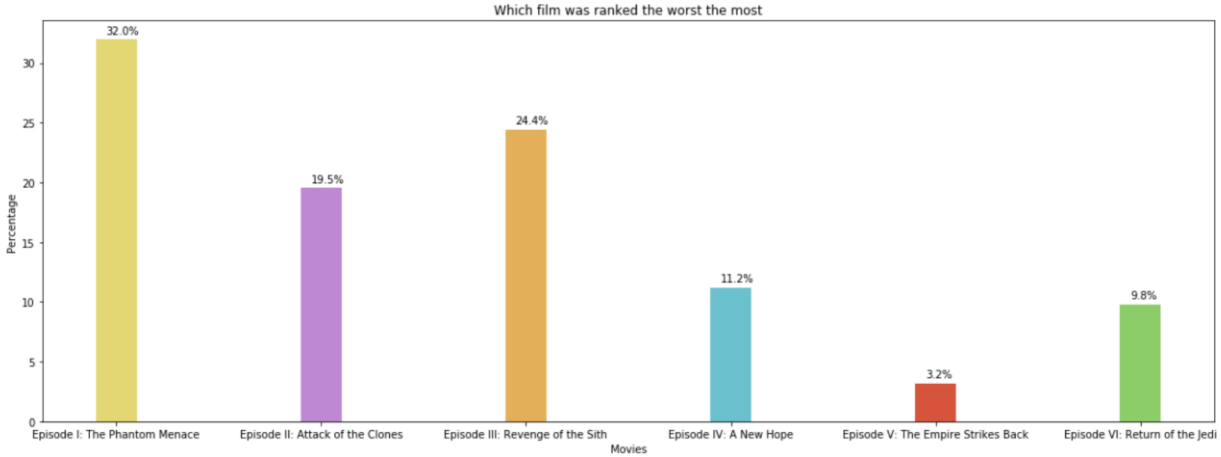
Which film was ranked the best the most

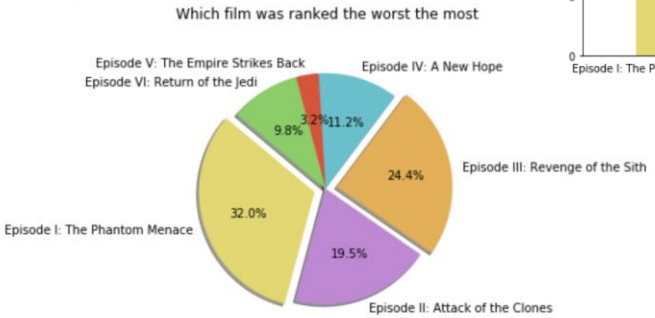




WORST STAR WARS MOVIE?

Which film was ranked the worst the most





RELATIONSHIP BETWEEN COLUMNS

BACKGROUND

Episode I was very political, with many extended sequences of the movie dedicated to diplomacy

EXPLORATION PAIR 1

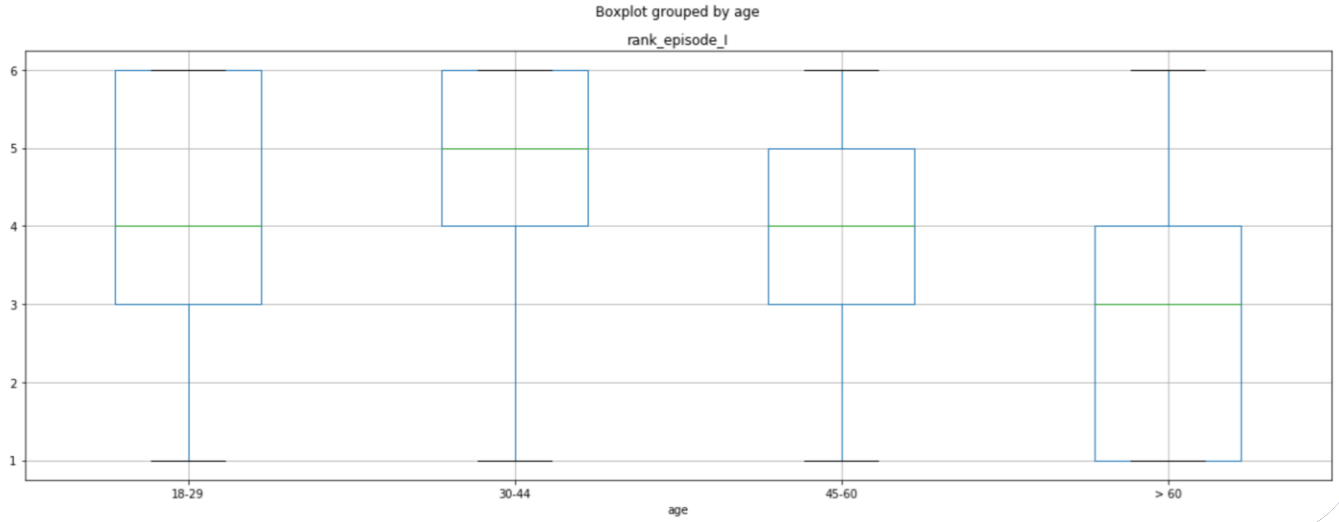
Relationship between respondent's age and how they rank **Episode I: The Phantom Menace**

HYPOTHESIS

Older generations would've understood the political aspect better and enjoyed this movie better than younger audiences, thus ranking it as the best over all others

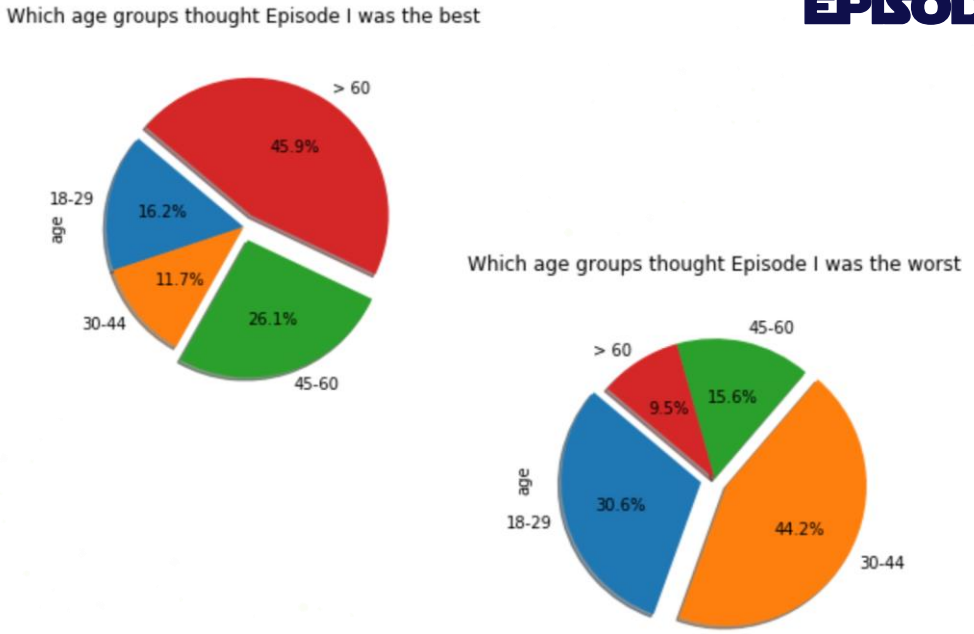
Boxplot of the relationship between respondent's age and how they rank

Episode I: The Phantom Menace



Pie charts of the relationship between respondent's age and how they rank

Episode I: The Phantom Menace



72%

of respondents who thought Episode I was the best are over the age of 45

74.8%

of respondents who thought Episode I was the worst are aged 18-44

EXPLORATION PAIR 2

Relationship between respondents who consider themselves fans, and their subsequent answer to the notoriously debated question

'WHO SHOT FIRST?'

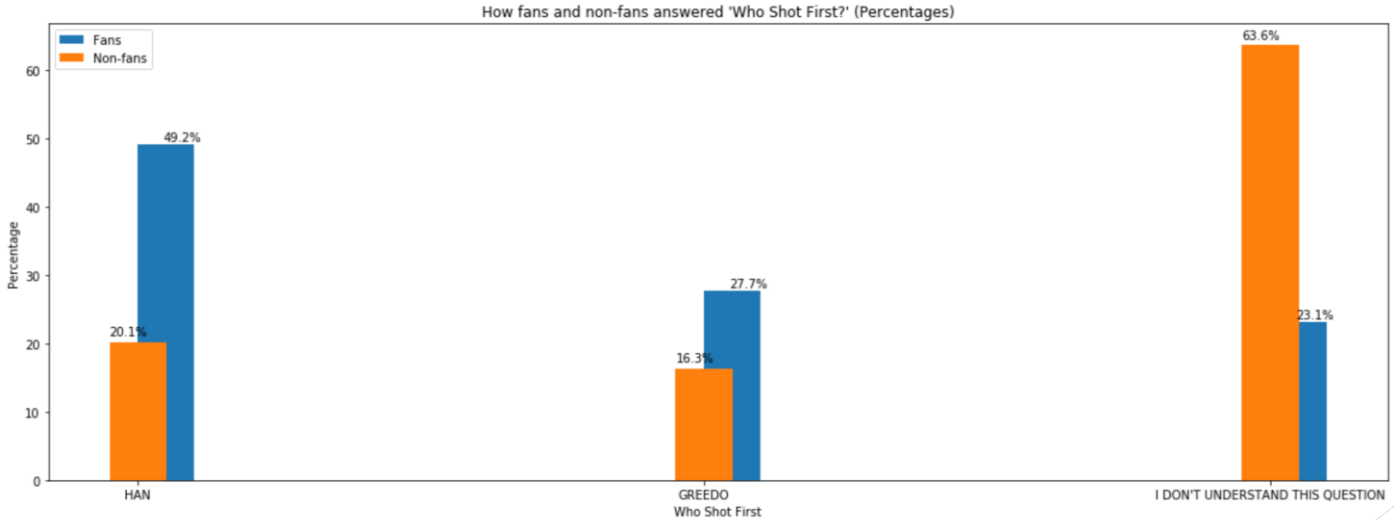
BACKGROUND

The original scene in Episode IV had Han Solo shoot Greedo first, but this was later changed in subsequent versions of the movie

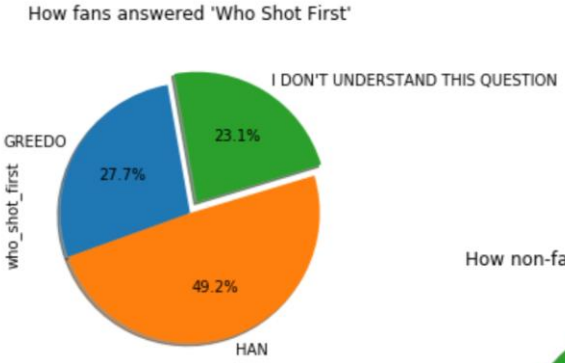
HYPOTHESIS

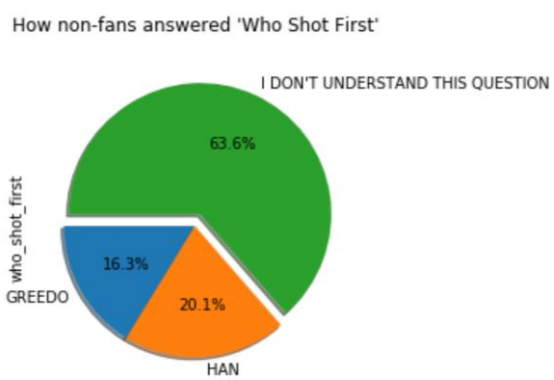
Real fans should know that Han shot first, or at the very least, understand the question.

Bar Graph of How Respondents Answered the Question 'Who Shot First?'



Pie Charts of How Fans and Non-Fans Answered the Question 'Who Shot First?'





49.2%

of respondents who are fans answered 'Han'

76.9%

of respondents who are fans understood the question

63.6%

of respondents who are not fans did not understand the question

EXPLORATION PAIR 3

Relationship between respondents who are Star Wars fans and Star Trek fans

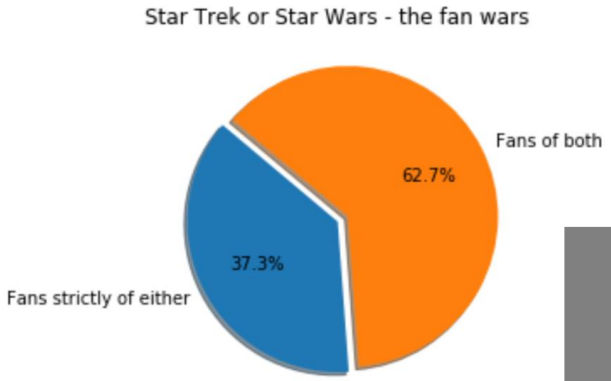
BACKGROUND

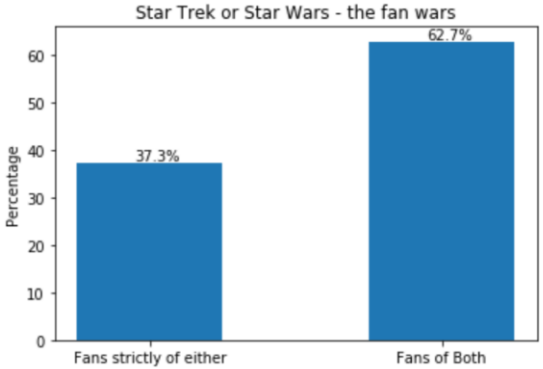
People's love for both these properties is so much, that there is a bit of a rivalry between the two fandoms.

HYPOTHESIS

Due to this intense rivalry, it is likely that we may find in the data that people only consider themselves fans of one of the franchises, but not both.

Bar Graph of How Respondents Answered the Question 'Who Shot First?'



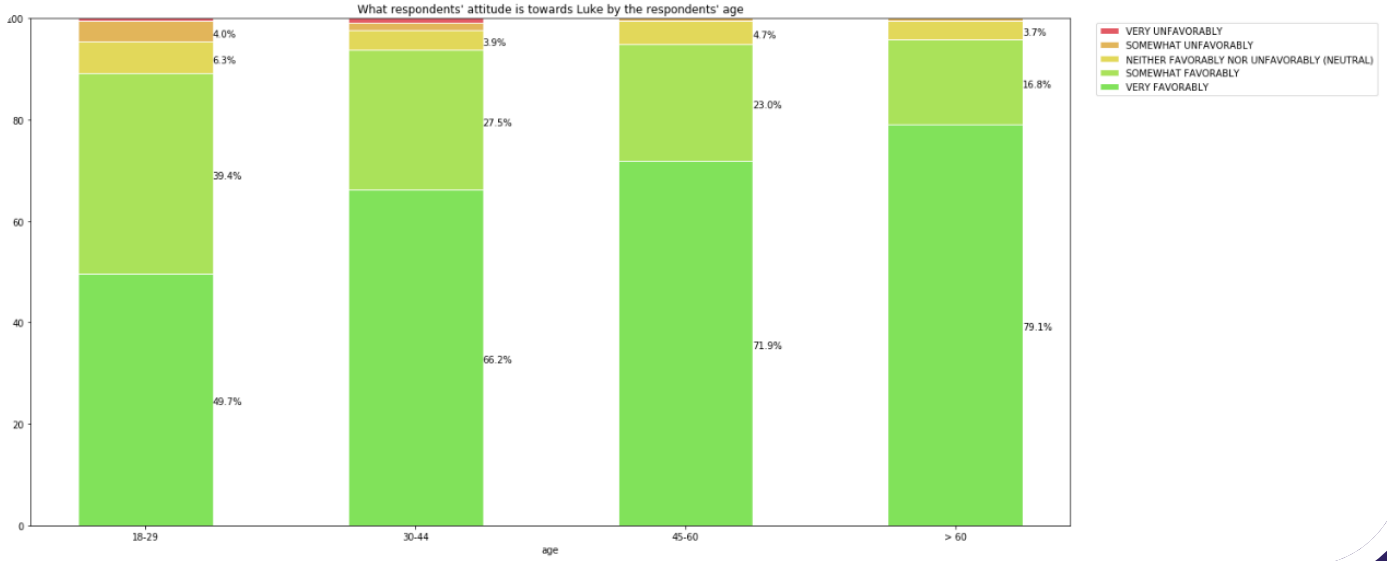


62.7%
of respondents are fans of both franchises

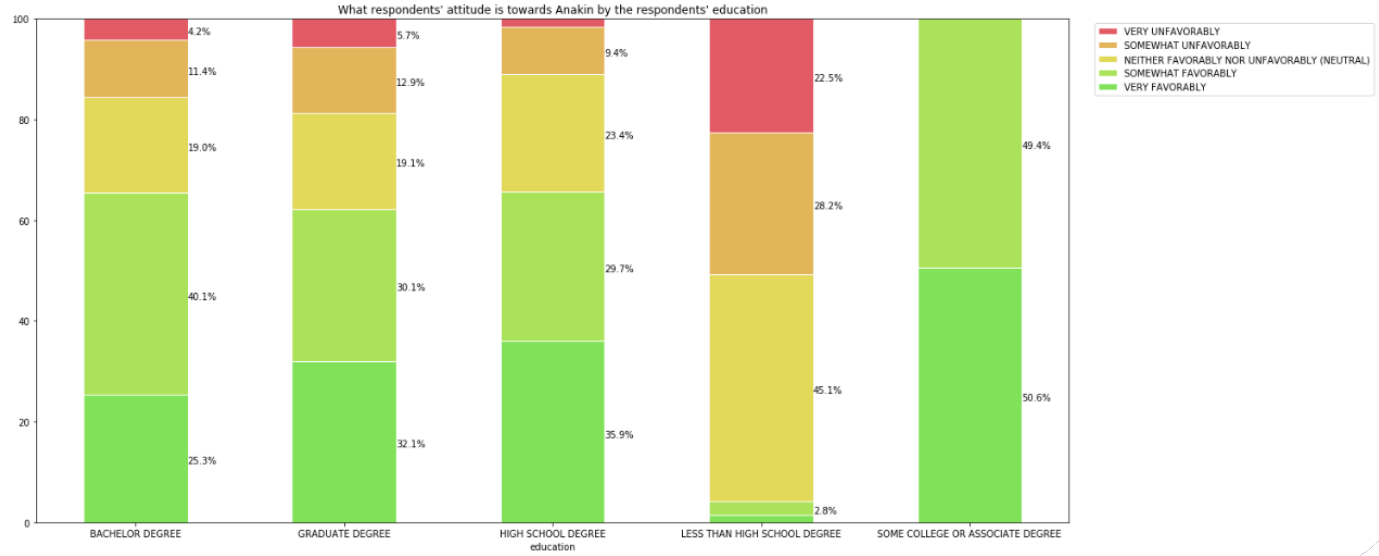
37.3%
of respondents are fans strictly of either one

CHARACTER AFFINITIES BY DEMOGRAPHICS

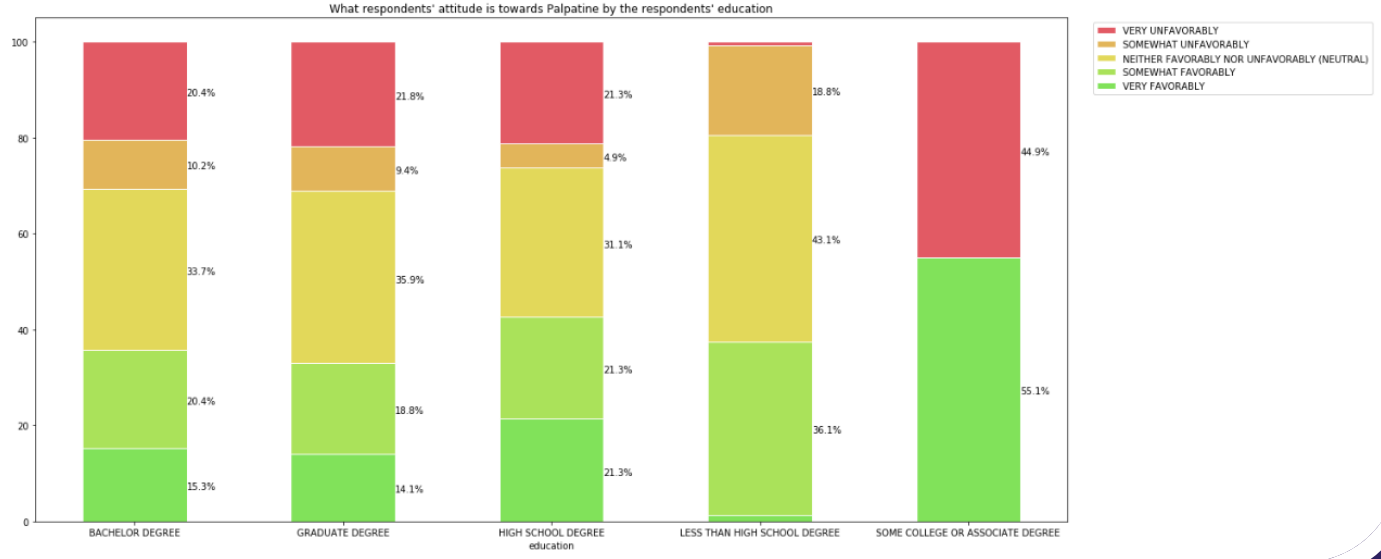
Respondents' attitude towards Luke Skywalker by age



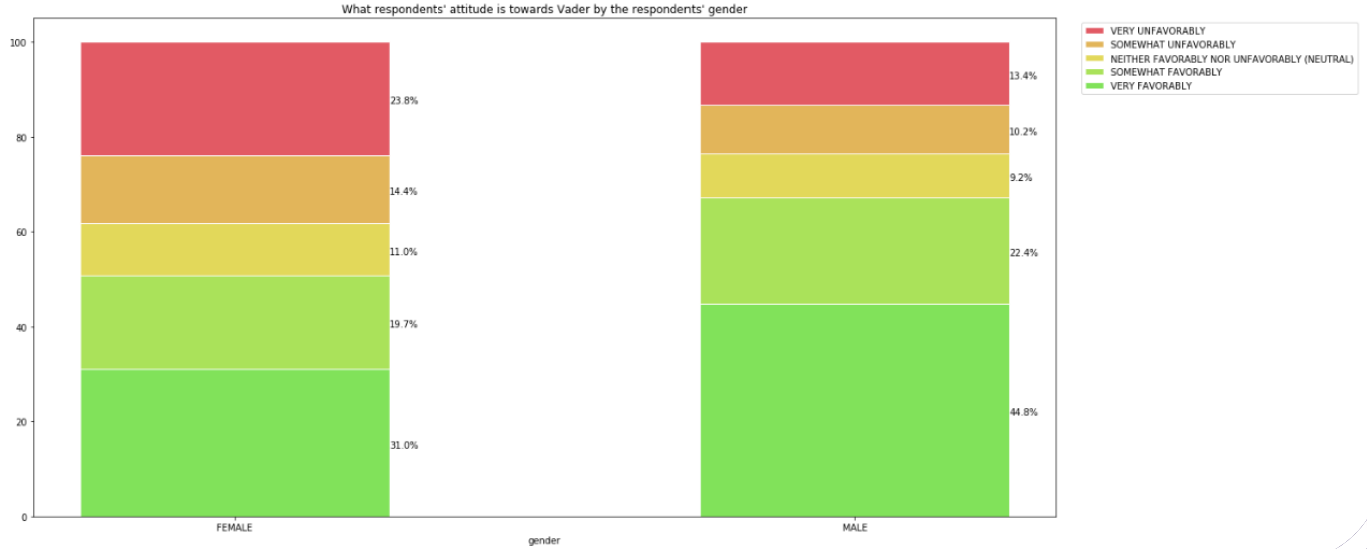
Respondents' attitude towards Anakin by age



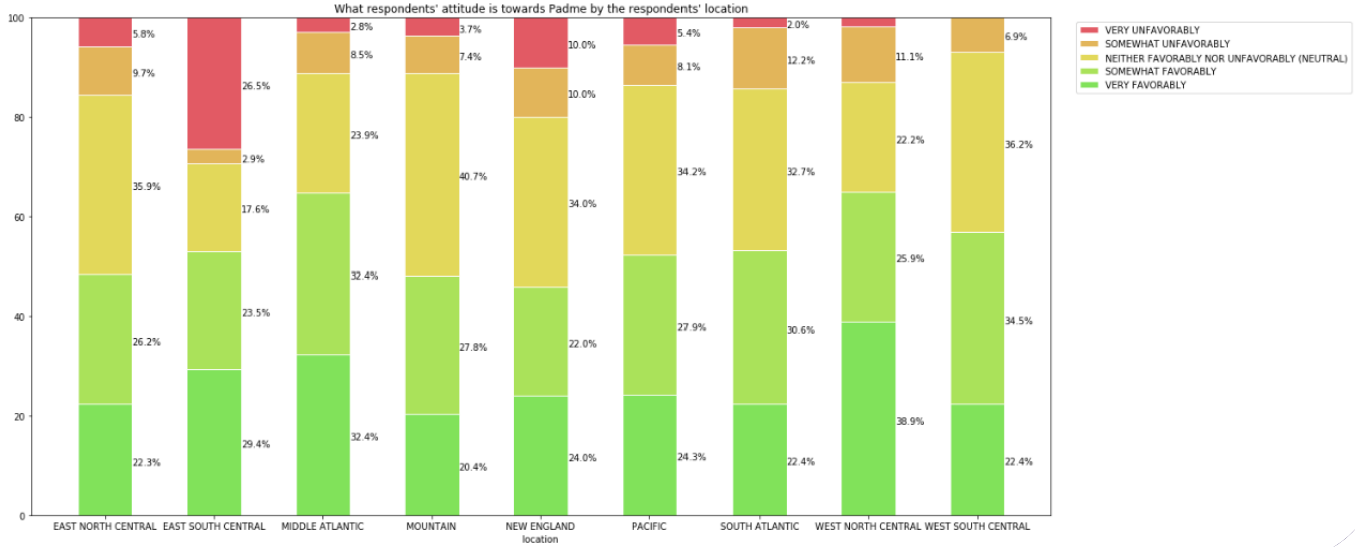
Respondents' attitude towards Palpatine by education



Respondents' attitude towards Darth Vader by gender



Respondents' attitude towards Padme by Location



THANKS!