Wrangling Final Project

Ziyan Wang

April 22, 2019

Idea Introduction

The series movies of Star Wars will be analyzed through different resources with web scrapping, data cleaning, text analysis and visualization skills.Starting from the brief introduction with the Wikipedia page, New York Times API and Character dataset within dplyr are included for detecting more details, such as the review of each movie and the relationship networks among characters.

Github Link

Analysis in this report is done with R(version 3.7). This PDF is generated by the Rmarkdone file. To check futher details, vist the link below: https://github.com/BecaWangRU/StarWars_analysis

Data resources

two websites:

Wikipedia page of Star Wars:

(url: https://en.wikipedia.org/wiki/List_of_Star_Wars_films_and_television_series)

IMDB page of Star Wars:

(url: https://www.imdb.com/list/ls025989264/)

API of New York Times:

Movie Reviews API & Article Search API * need to sign up with email to obtain the individual API key.

Characters Dataset "Starwars" within the dplyr package

Basic Info of the series

Star Wars is an American epic space-opera media franchise created by George Lucas. The franchise began with the eponymous 1977 film and quickly became a worldwide pop-culture phenomenon. Till now there are 11 movies in the whole series, the basic information below is scraped from wikipedia.

date	name
May 25, 1977 (1977-05-25)	Episode IVA New Hope
May 21, 1980 (1980-05-21)	Episode VThe Empire Strikes Back
May 25, 1983 (1983-05-25)	Episode VIReturn of the Jedi
May 19, 1999 (1999-05-19)	Episode IThe Phantom Menace
May 16, 2002 (2002-05-16)	Episode IIAttack of the Clones
May 19, 2005 (2005-05-19)	Episode IIIRevenge of the Sith
December 18, 2015 (2015-12-18)	Episode VIIThe Force Awakens
December 15, 2017 (2017-12-15)	Episode VIIIThe Last Jedi
December 20, 2019 (2019-12-20)	Episode IXThe Rise of Skywalker
December 16, 2016 (2016-12-16)	Rogue One

date	name
May 25, 2018 (2018-05-25)	Solo

Usually web scraping from the wikipedia page is easy, but the tables on the Star Wars page is not standard relational table, making it impossible to retrieve the clean table with node “table”. So in this part, I first scrape the elements of the tables using node “th,td” with the help of SelectorGadget, and then target the date to extract the information needed. It is convenient since, after observing the information scraped, half of the movies are published in May (movies before 2005), and the other half are published in December. And I am wondering if it is related to the Star Wars day, May 04th, indicating “May the force be with you”.

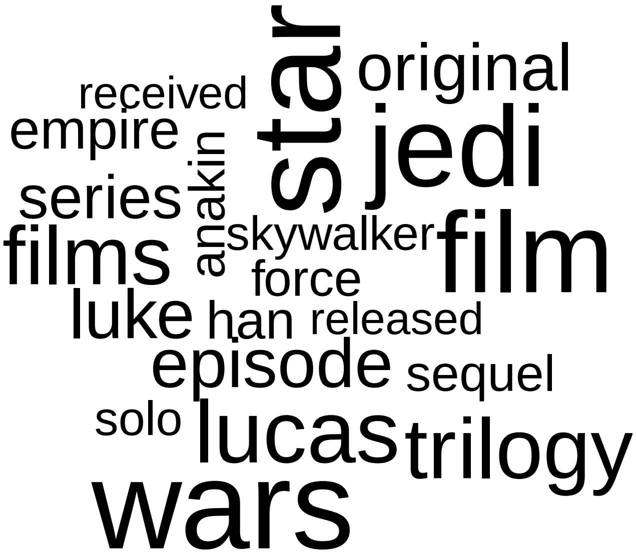
Summary of the films

After knowing the basic names and publishing dates, it is important to know the storyline of the series. By web scraping on the Wiki page again, finding the summary of each story could tell the main characters and the most frequent used words across the whole stories.

- all the words segmentations in this report have exclude the stopwords, which does not have specific meaning in the context.
- besides, make sure to do the web scraping with both the node “p” (refer to paragraph) and the node “h2” (refer to the heading) so that each movie can be separated using the headings.

```
## [1] "Television[edit]"       "Television specials\n"
```

```
## Joining, by = "word"
```



```
## NULL
```

The wordcloud plot gives the top 20 most frequent words in the desciptions. The words that appears a lot in the summaries include force, jedi, awake etc. The result seems quite reasonable, but only two chracters show up in the most frequent words: luke skywalker and han solo. It should be admitted that these two characters are really important through the movies, but bigrams should be taken into consideration since the names mostly come with first name and last name.

word1	word2	n
star	wars	43

word1	word2	n
empire	strikes	11
han	solo	10
original	trilogy	10
death	star	9
force	awakens	9
obi	wan	9
clone	wars	8
sequel	trilogy	8
wars	156	7
phantom	menace	6
principal	photography	6
special	achievement	6
156	won	5
awakens	162	5
bb	8	5
darth	vader	5
episode	vii	5
jedi	158	5
lando	calrissian	5
lawrence	kasdan	5
luke	skywalker	5
achievement	award	4
jedi	164	4
kylo	ren	4
luke’s	father	4
original	film	4
prequel	trilogy	4
rebel	alliance	4
tv	series	4

In the result above, we can see the names of several other characters than Han Solo and Luke Skywalker: Obi Wan, bb 8, Darth Vader and Lawrence Kasdan, Kylo Ren. However, the summary on the wiki page is not long and specific enough to do further text analysis, so that in the following part resources other than wikipedia will be introduced.

Characters

Previous discussion has already shown several important characters of the movie series. In this part, I will show the analysis of characters in Star Wars with the R dataframe “starwars” from the package “dplyr”, to figure out the popular characters to see if it conforms to the results above.

```
## Warning: Unknown or uninitialised column: 'num_films'.
```

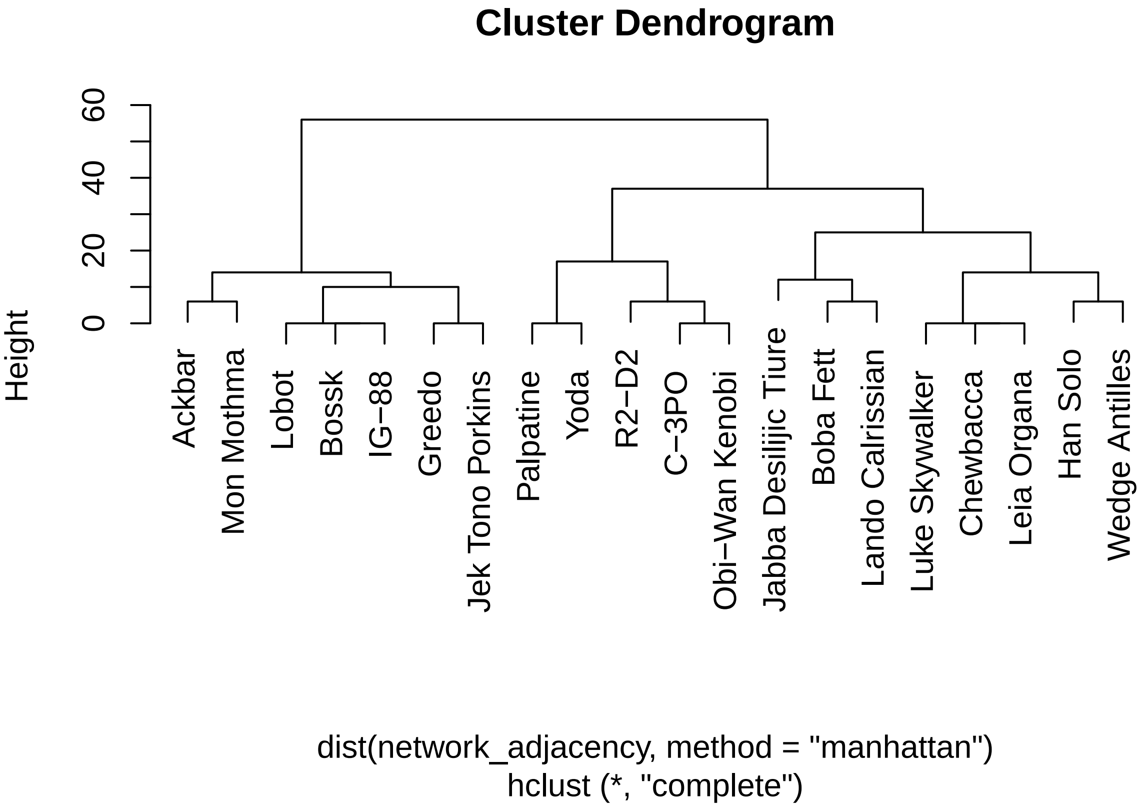
name	species	num_films	films
R2-D2	Droid	7	c(“Attack of the Clones”, “The Phantom Menace”, “Revenge of the S
C-3PO	Droid	6	c(“Attack of the Clones”, “The Phantom Menace”, “Revenge of the S
Obi-Wan Kenobi	Human	6	c(“Attack of the Clones”, “The Phantom Menace”, “Revenge of the S
Luke Skywalker	Human	5	c(“Revenge of the Sith”, “Return of the Jedi”, “The Empire Strikes B
Leia Organa	Human	5	c(“Revenge of the Sith”, “Return of the Jedi”, “The Empire Strikes B
Chewbacca	Wookiee	5	c(“Revenge of the Sith”, “Return of the Jedi”, “The Empire Strikes B
Han Solo	Human	5	c(“Return of the Jedi”, “The Empire Strikes Back”, “A New Hope”, “

name	species	num_films	films
Greedo	Rodian	5	A New Hope
Jabba Desilijic Tiure	Hutt	5	c(“The Phantom Menace”, “Return of the Jedi”, “A New Hope”)
Wedge Antilles	Human	5	c(“Return of the Jedi”, “The Empire Strikes Back”, “A New Hope”)
Jek Tono Porkins	Human	5	A New Hope
Yoda	Yoda’s species	5	c(“Attack of the Clones”, “The Phantom Menace”, “Revenge of the S
Palpatine	Human	5	c(“Attack of the Clones”, “The Phantom Menace”, “Revenge of the S
Boba Fett	Human	5	c(“Attack of the Clones”, “Return of the Jedi”, “The Empire Strikes B
IG-88	Droid	5	The Empire Strikes Back
Bossk	Trandoshan	5	The Empire Strikes Back
Lando Calrissian	Human	5	c(“Return of the Jedi”, “The Empire Strikes Back”)
Lobot	Human	5	The Empire Strikes Back
Ackbar	Mon Calamari	5	c(“Return of the Jedi”, “The Force Awakens”)
Mon Mothma	Human	5	Return of the Jedi

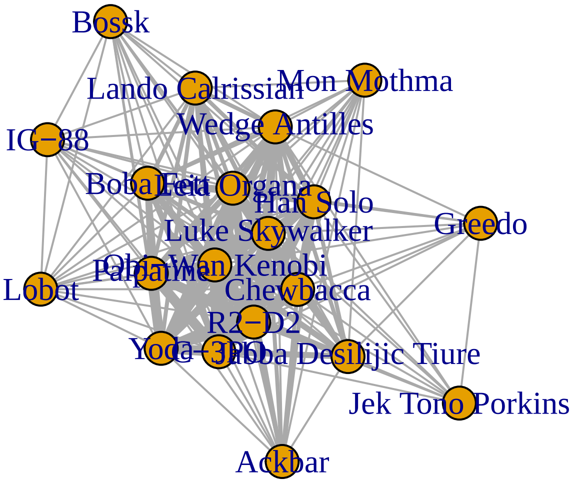
A snippet of my dataframe arranged by the number of films presented has been shown above. We can see the previous names we have are the popular names from the table. But there is only one exception: “Kylo Ren”. This character is the villain in the recent movies, chosen name of Ben Solo. This result imply another drawback of the Starars dataset: the data set is little bit out of date, containing only the first seven films of the whole series.

Besides, the character dataframe can also generate the network graph, helping to visualize the relationship between characters. If the characters shows up in same movies, we count one to build up the adjacency matrix and then do the cluster.

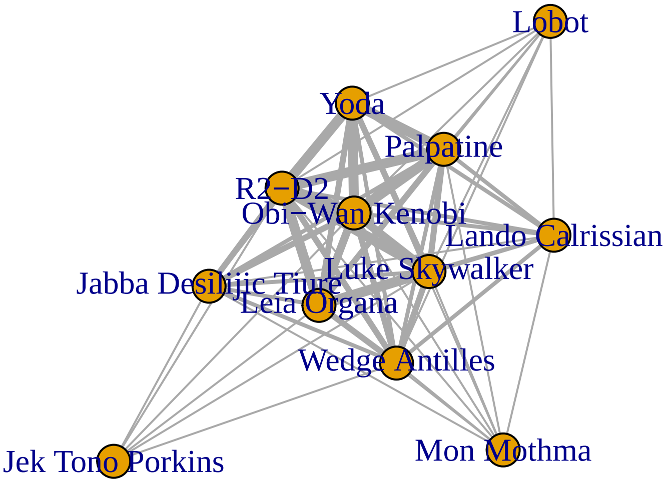
- Only the important characters are selected in this graph analysis, which are the top 20 characters that appears in most movies.



In this clustering plot, we can see the result kind of make sense based on the story. R2-D2 is really close to C-3PO. Both of them are roberts. Luke and Leia are twins and are really close in this cluster. Another kind of visualization of the character relationships is the following networks:



The network seems a little bit messy, so I tried to provide another plot with only top 11 characters, making the core relationships more clear:



The first messy network indicates that generating the networks has another draw back: counting one for the mutual appearance in the same movie makes the values in the adjacency matrix small, making the cluster result less precise. In fact, I think about using the reviews of each film to generate another network graph by counting the numbers that characters mutually appear in one review, making the values in adjacency matrix bigger. But it should be really hard. It is not easy to recognize the names of characters from the reviews. If do the words segmentation and compare it to the Starwars dataset, the new character in the last four films cannot be match. Moreover, the names might start with Capital Letters, but after the unnest_tokens function, words are automatically transformed into smaller letter. So this might be a problem that I left here.

Reviews of the movies

Getting the reviews of each movie helps to know the response of the movie. The New York Times Movie API gives the access to these reviews. By signing in and creating application, the metadata of movie reviews can be obtained, which does not contain the whole reviews, requiring the further web scaping with the urls provided in the metadata. Another quick tip is that New York Times seems change the organization of each article after 2010, indicating that different node and format before and after these time point.

After obtaining the nine reviews, we could do text analysis again, finding out the most frequent words or characters in the reviews. Besides, we can use sentiment analysis to try finding out the positive or negative feelings showing towards the movies.

- but the New York Times API just provide 9 reviews of the movie series, meaning that two more movies are not included in the reviews. One is the Episode V The Empire Strikes Back, published in 1980-05-21, another is the Episode III Revenge of the Sith, published in 2005-05-19.

the whole reviews analysis

First put the reviews together, to detect the most popular words in the text.

```
## Joining, by = "word"
```

top10	freq	top20	freq	top30	freq
wars	64	lucas	17	skywalker	13
star	58	it's	15	story	13
movie	29	menace	15	darth	12
force	22	leia	14	lucas's	12
film	21	luke	14	vader	12
jedi	21	phantom	14	called	11
time	21	solo	14	empire	11
han	19	space	14	episode	11
series	19	abrams	13	princess	11
awakens	17	johnson	13	creatures	10

From the result above, it is easy to observe that some of the words come from the same phrases, especially some names are divided into first name and last name. So the following procedure is to split the reviews into bigrams (phrases contain two words). The table below shows the top 30 most popular bigrams in the whole reviews of Star Wars.

top 10_word1	top 10_word2	freq	top 20_word1	top 20_word2	freq	top 30_word1	top 30_word2	freq
star	wars	43	phantom	menace	6	lawrence	kasdan	5
empire	strikes	11	principal	photography	6	luke	skywalker	5
han	solo	10	special	achievement	6	achievement	award	4
original	trilogy	10	156	won	5	jedi	164	4
death	star	9	awakens	162	5	kylo	ren	4
force	awakens	9	bb	8	5	luke's	father	4
obi	wan	9	darth	vader	5	original	film	4
clone	wars	8	episode	vii	5	prequel	trilogy	4
sequel	trilogy	8	jedi	158	5	rebel	alliance	4
wars	156	7	lando	calrissian	5	tv	series	4

From the two tables above, we can see the bigrams mainly come from the names of charcters, and names of the movies. And some of the bigrams contain numbers that just appear out of expect.

reviews analysis respectively

There are 9 reviews provided by the New York Times, so we can get 9 separate results from each of them. We can count the word frequency and do some sentiment analysis to help decide whether the review is positive or negative. The feature “direction” is generated by using the number of positive words frequencies minus the number negative words frequencies for each movie.

movie	direction	web_url
Return of the Jedi	1	http://www.nytimes.com/1983/05/25/movies/lucas-returns-wit
Rogue One: A Star Wars Story	-7	http://www.nytimes.com/2016/12/13/movies/star-wars-rogue-
Solo: A Star Wars Story	-15	http://www.nytimes.com/2018/05/15/movies/solo-a-star-wars-
Star Wars	8	http://www.nytimes.com/1977/05/26/archives/star-wars-a-trip-
Star Wars: Episode I - The Phantom Menace	-5	http://www.nytimes.com/1999/05/19/movies/film-review-in-the
Star Wars: Episode II Attack of the Clones	-22	http://www.nytimes.com/2002/05/10/movies/film-festival-revie
Star Wars: Episode VII - The Force Awakens	8	http://www.nytimes.com/2015/12/18/movies/star-wars-the-for
Star Wars: The Clone Wars	-3	http://www.nytimes.com/2008/08/15/movies/15clon.html
Star Wars: The Last Jedi	-19	http://www.nytimes.com/2017/12/12/movies/star-wars-the-last

As the review indicates, almost all the reviews turns to the negative side. Especially three of them are obvious negative while the other six are around the zero. So how do the sentiment analysis performs? I introduce the ratings comes from IMDB to help evaluate by web scraping of the IMDB page for Star Wars.

Information from IMDB

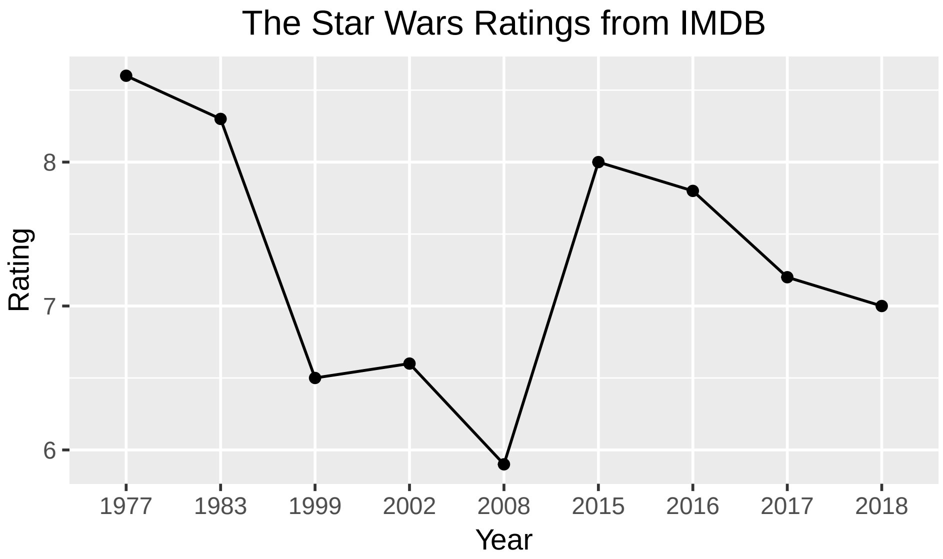
Rating for each movie

Scraping from this url requires writing the regular expressions to get the information needed. And using the publishing year to match the movies is more convenient than matching with the name of the movies since the publishing year for each movie is unique. And the table below shows the Ratings and the directions.

movie	year	Rate	direction
Return of the Jedi	1983	8.30	1
Rogue One: A Star Wars Story	2016	7.80	-7
Solo: A Star Wars Story	2018	7.0	-15
Star Wars	1977	8.60	8
Star Wars: Episode I - The Phantom Menace	1999	6.50	-5
Star Wars: Episode II Attack of the Clones	2002	6.60	-22
Star Wars: Episode VII - The Force Awakens	2015	8.0	8
Star Wars: The Clone Wars	2008	5.90	-3
Star Wars: The Last Jedi	2017	7.20	-19

As the table shows, the three movies: Return of the Jed(1983), Episode IV A New Hope(1977), and Star Wars: Episode VII - The Force Awakens(2015), are highly rated. All of the three movies get ratings over 8.0. The reviews must be kind of observative and critic, so that all the directions been caculated do not reveal strong positive feelings.

Besides, we can have a quick look at all the Ratings of Star Wars movies via the line plot:



The rating plot shows that the first two movies publishing around the 1980s are highly rated, but the following three movies are not that recommended. And the movies after 2015 have decreasing ratings as well. Only the four movies: Return of the Jed(1983), Episode IV A New Hope(1977), Star Wars: Episode VII - The Force Awakens(2015), and Rogue One: A Star Wars Story (2016) have relative high ratings.

Votes ang Gross from IMDB page

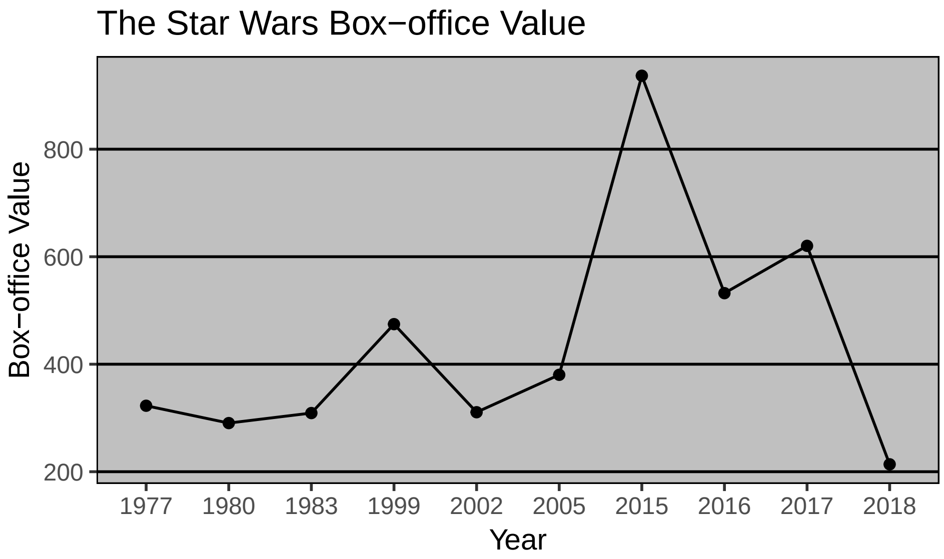
Here is a sample of the tidy description of Star Wars movies from IMDB:

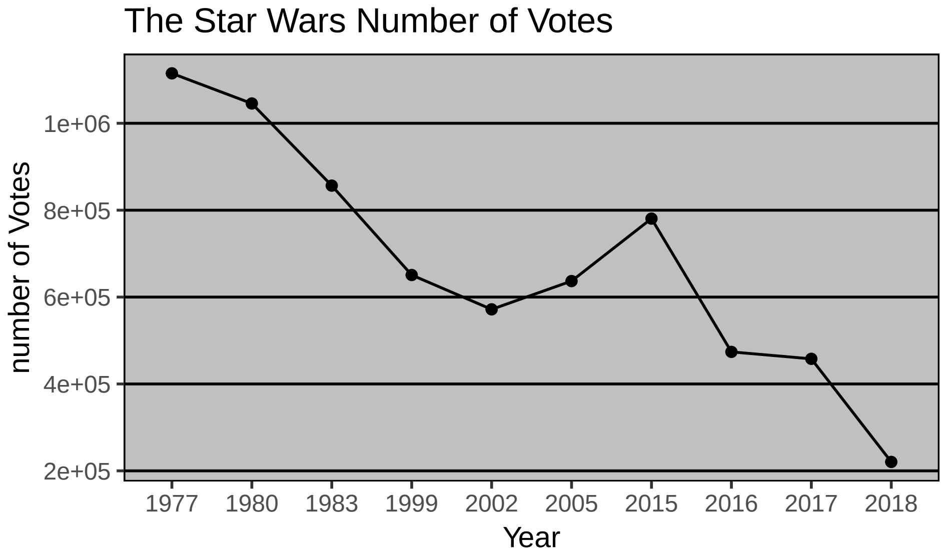
[1] "1.Star Wars: Episode V - The Empire Strikes Back(1980)PG | 124 min | Action Adventure Fantasy8.70Rate1Rate2Rate3Rate4Rate5Rate6Rate7Rate8Rate9Rate10Rate0Error: please try again.82MetascoreAfter the Rebels are brutally overpowered by the Empire on the ice planet Hoth Luke Skywalker begins Jedi training with Yoda while his friends are pursued by Darth Vader.Director:Irvin Kershner | Stars:Mark Hamill Harrison Ford Carrie Fisher Billy Dee WilliamsVotes:1045110|Gross:$290.48M"

The description gives extra information of the number of votes and the box-office value. These information also help to understand the popularity for the whole series.

name	Year	Box	Rate	Votes
Episode VThe Empire Strikes Back	1980	290.48	8.70	1045577
Episode IVA New Hope	1977	322.74	8.60	1114942
Rogue One	2016	532.18	7.80	473819
Episode VIReturn of the Jedi	1983	309.13	8.30	856531
Episode VIIIThe Last Jedi	2017	620.18	7.20	457839
Episode VIIThe Force Awakens	2015	936.66	8.0	780609
Solo	2018	213.77	7.0	220555
Episode IIIRevenge of the Sith	2005	380.26	7.60	636872
Episode IIAttack of the Clones	2002	310.68	6.60	571709
Episode IThe Phantom Menace	1999	474.54	6.50	650759

SO is there any relation among the Rate, Votes and Box? Here are several plot help visualize the relationships.

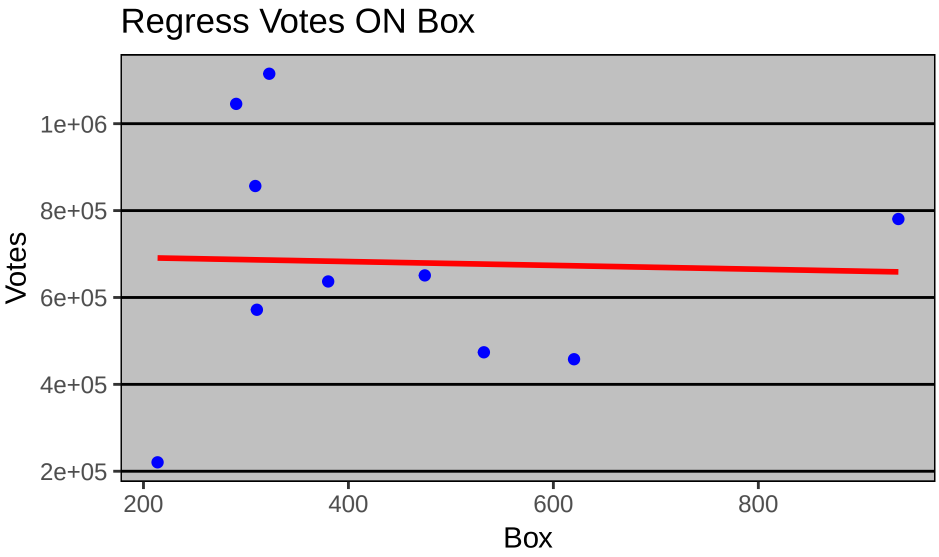




The Box-office value reached the peak at 2015, when published the movie "Episode VIIThe Force Awakens". It seems that the box office of the movie after the 21st century are generally higher, which can be the result of economic development – more people can afford the movies. But surprisingly the Box-office of "Solo: A Star Wars Story (2018)" is also low and true, with only 213mllion dollars.

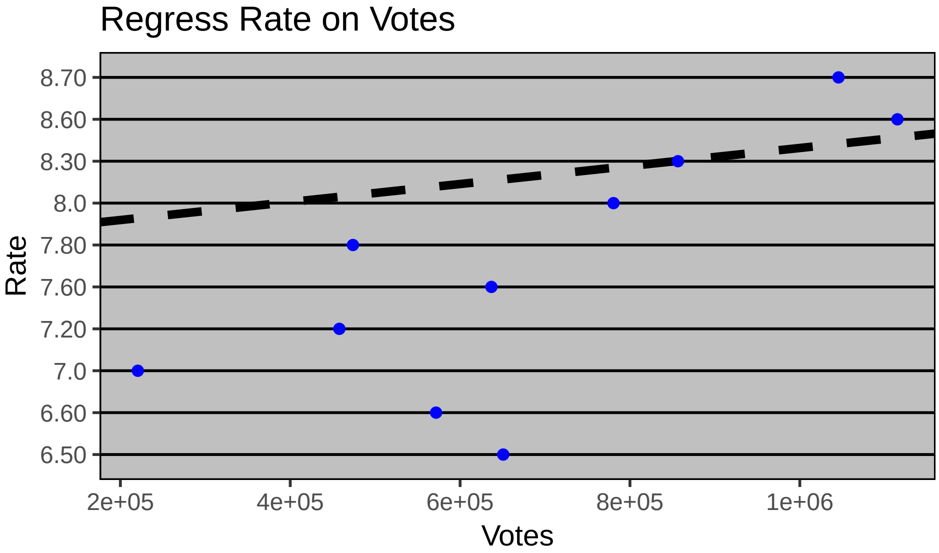
However, the number of votes keeps decreasing. Only the first two movies published at 1990s get the votes over 1,000,000, which is quite surprising. People may not curious about this kind of fiction movies after seeing the first several of them.

Futher, we can try to find out the relationship between Box-office value and the number of votes. Is the movie with higher number of votes, which means higher attention, get higher box-office value?



The regression line above shows little relationship between Box and Votes.

The same question happens between Rates and the number of votes.



This plot shows that generally more number of votes (more attention paying to the movie) lead to higher ratings on the movie.

conclusion

From the analysis above, the most popular movies of the whole series are: (1)Return of the Jed(1983),(2) Episode IV A New Hope(1977), (3)Star Wars: Episode VII - The Force Awakens(2015). The first two movies are the most famous two since theses space-opera fiction movies are rare at 1990s. The following movies extend the story with same background, but having less influence than the first two movies.