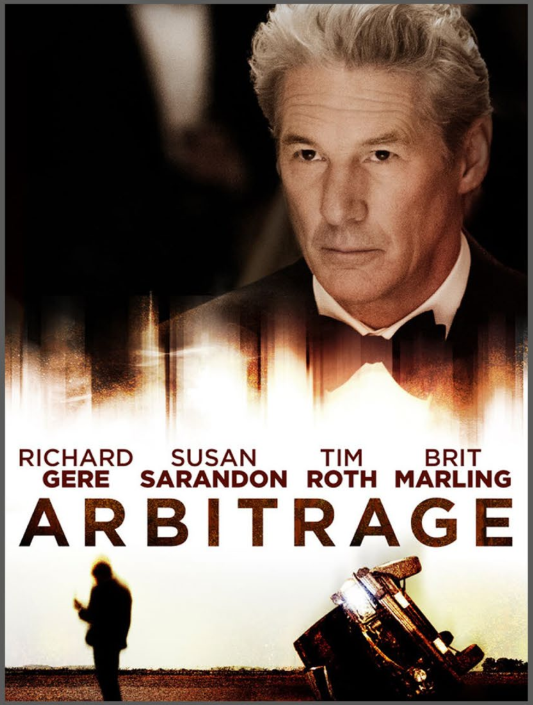
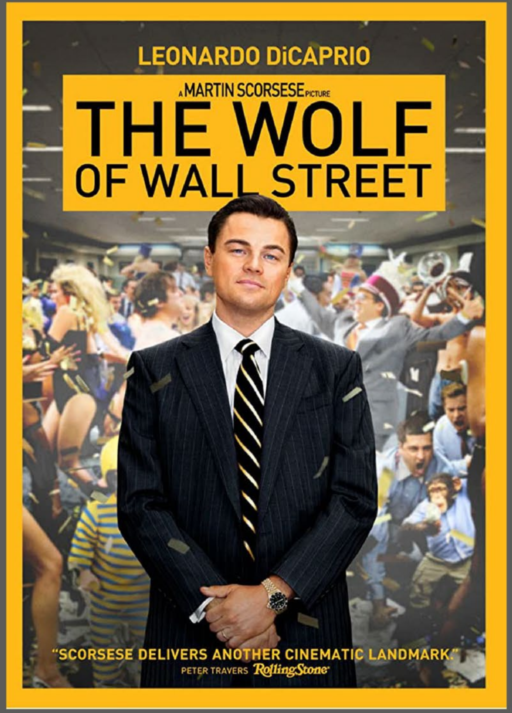
What's in a Hollywood Wall Street Movie?

A DATA ANALYSIS INVESTIGATION





Joram Mutenge

outline

1. Why this project?
2. Tools used
3. Building the dataset
4. Cleaning methods
5. Some findings
6. Takeaways

why this project?

Of Dollars And Data

Act Smarter. Live Richer.

Home | Book | Newsletter | E-Book | Invest with Nick | About

The 13 Best Wall Street Movies [That You Can't Miss]

Posted January 24, 2023 by Nick Maggiulli

@dollarsanddata

NICK MAGGIULLI

Of Dollars And Data focuses on personal finance using data analysis. Nick Maggiulli is the Chief Operating Officer for Ritholtz Wealth Management LLC. For disclosure information

why this project?

- I dream of writing a movie some day.
- Use data analysis to discover some characteristics of a good Wall Street movie.
- Leverage Nick Maggiulli's list of movies.

tools used

LANGUAGES

- Python 3.10
- Regular expressions

LIBRARIES

- Pandas (my favorite data munging tool)
- Numpy
- Matplotlib
- Gensim
- NLTK
- TextBlob

building the dataset

1. Scraped the data from Nick's blog.
2. Added rating and duration data from IMDb.
3. Downloaded subtitles (.srt files) for each movie and converted to markdown (.md files)
4. Cleaned the text with regular expressions.
5. Merged two datasets into one on movie title columns.

building the dataset

- Dataset when scraped from the blog.

	title	description
0	1. Margin Call (2011)	Set in the early stages of the 2008 financial ...
1	2. Wall Street (1987)	This is the classic film that started it all. ...
2	3. The Big Short (2015)	Based on the book by Michael Lewis, The Big Sh...
3	4. Trading Places (1983)	Being the only pure comedy on this list, Tradi...
4	5. The Wolf of Wall Street (2013)	Directed by Martin Scorsese, The Wolf of Wall ...

building the dataset

- Dataset after merging.

	movie	year	description	rating	minutes	script
0	Margin Call	2011	Set in the early stages of the 2008 financial ...	7.1	107	Is that them? Jesus Christ. Are they going to ...
1	Wall Street	1987	This is the classic film that started it all. ...	7.3	126	Easy! Excuse me! Good morning. Jackson Steinem...
2	The Big Short	2015	Based on the book by Michael Lewis, The Big Sh...	7.8	130	Frank. How are the wife and kids? You know, fo...
3	Trading Places	1983	Being the only pure comedy on this list, Tradi...	7.5	118	Your breakfast, sir. Pork bellies! I have a hu...
4	The Wolf of Wall Street	2013	Directed by Martin Scorsese, The Wolf of Wall ...	8.2	180	The world of investing can be a jungle. Bulls....

cleaning methods

- Extensively used regular expressions. Here are three sample expressions written:

```
def first_cleaning(dd):
    return (dd
     .assign(script=lambda dd_: dd_.script.str.replace('\d<br>\d{2}:\d{2}:\d{2},\d{3}\ -->\ \d{2}:\d{2}:\d{2},\d{3}<br>|ï»¿', '', regex=True),
             script1=lambda dd_: dd_.script.str.replace('(<br>)+\d|<br><br>|<br>|â™ªâ™ªâ™ª|\d-', ' ', regex=True),
             script2=lambda dd_: dd_.script1.str.replace('\d+\[\ \]|[â™ªâ™ªâ™ª]\ |</i>|<i>|â€¦|\ -\ ', '', regex=True),
             script3=lambda dd_: dd_.script2.str.replace('(<br>)+\d+|<br>-\ |<br>', ' ', regex=True),
             script4=lambda dd_: dd_.script3.str.replace('\[â™ªâ™ªâ™ª\]|(\d+)?\[((([A-Z])+ ?)+)+\]|\d+\*\*\ |\d+\.\.|\d+-\ ', '', regex=True),
             script5=lambda dd_: dd_.script4.str.replace('â€™', "'", regex=True),
             script6=lambda dd_: dd_.script5.str.replace(' [ ] ', '', regex=False),
             script7=lambda dd_: dd_.script6.str.replace('(\d+)?([A-Z])+:\ |<font\ color="#', '', regex=True),
            )
     .pipe(remove_numbers, 'script7')
     .pipe(replace_first_two_digits, 'script7')
     .pipe(add_space_after_punctuation, 'script7')
     .drop(columns=['script','script1','script2','script3','script4','script5','script6'])
     .rename(columns={'script7':'script'})
    )
```

cleaning methods

```
def second_cleaning(df):
    return (dd
     .assign(script=lambda df_: df_.script.str.replace('([A-Z])+\ \d(\ )?:\ ', '', regex=True),
             script1=lambda df_: df_.script.str.replace('</font>', '', regex=False),
             script2=lambda df_: df_.script1.str.replace('e020">', '', regex=False),
             script3=lambda df_: df_.script2.str.replace('(â™)?(\d+)?â™', '', regex=True),
             script4=lambda df_: df_.script3.str.replace('\*(\ )?\*\ ', '', regex=True),
             script5=lambda df_: df_.script4.str.replace('(\d+)?(\*)?\ (\d+\*)?(\ )?', ' ', regex=True),
             script6=lambda df_: df_.script5.str.replace('(\d+)?\[([A-Za-z])+(\])?\ (([A-Za-z])+\ ([A-Za-
             script7=lambda df_: df_.script6.str.replace('\d+\"', '"', regex=True),
             script8=lambda df_: df_.script7.str.replace('\([a-zA-Z]+((\ [a-zA-Z]+)+)?\)|\)|\.{3}|--|\d+\
             script9=lambda df_: df_.script8.str.replace('Subtitles downloaded from www? OpenSubtitles? o
             script10=lambda df_: df_.script9.str.replace(' #', '', regex=False),
             script11=lambda df_: df_.script10.str.replace('Sync for "Wall? Street.1987.BluRay? P? DTS? x
            )
     .drop(columns=['script','script1','script2','script3','script4','script5','script6','script7','scrip
     .rename(columns={'script11':'script'})
    )
```

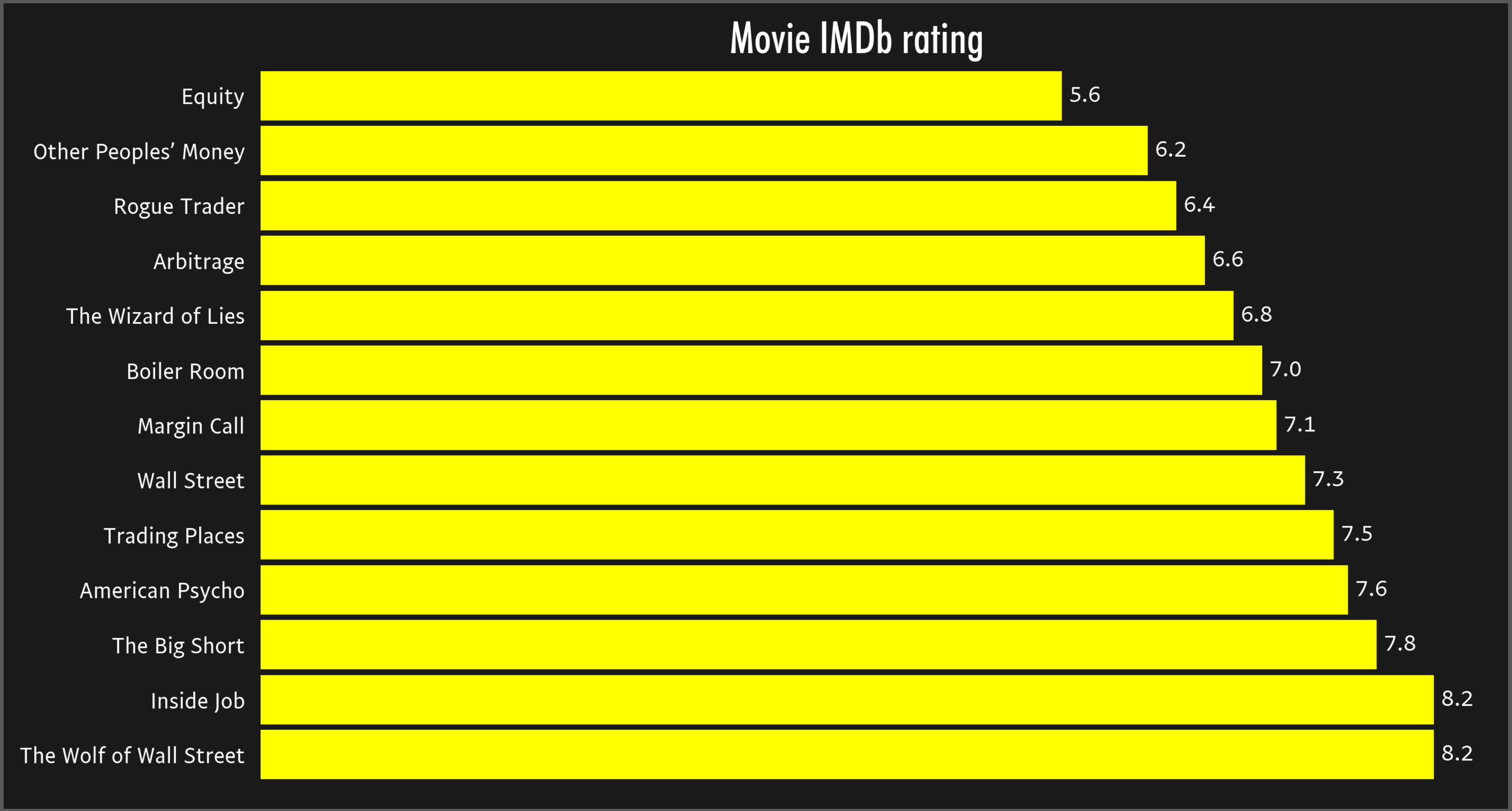
cleaning methods

```
def third_cleaning(df):
    return (dd
     .assign(script=lambda df_: df_.script.str.replace('Best watched using Open Subtitles MKV Player', '', regex=False),
             script1=lambda df_: df_.script.str.replace('clanking]', '', regex=False),
             script2=lambda df_: df_.script1.str.replace(' Visiontext subtitles: Paul Sofer', '', regex=False),
             script3=lambda df_: df_.script2.str.replace('Cleaned, corrected and OCR issues fixed by Tronar Hiya,', '', r
             script4=lambda df_: df_.script3.str.replace('Sync for "Wall? Street.1987.BluRay? P.DTS? x CHD" ::nlsinh@gmai
            )
     .drop(columns=['script','script1','script2','script3'])
     .rename(columns={'script4':'script'})
    )
```

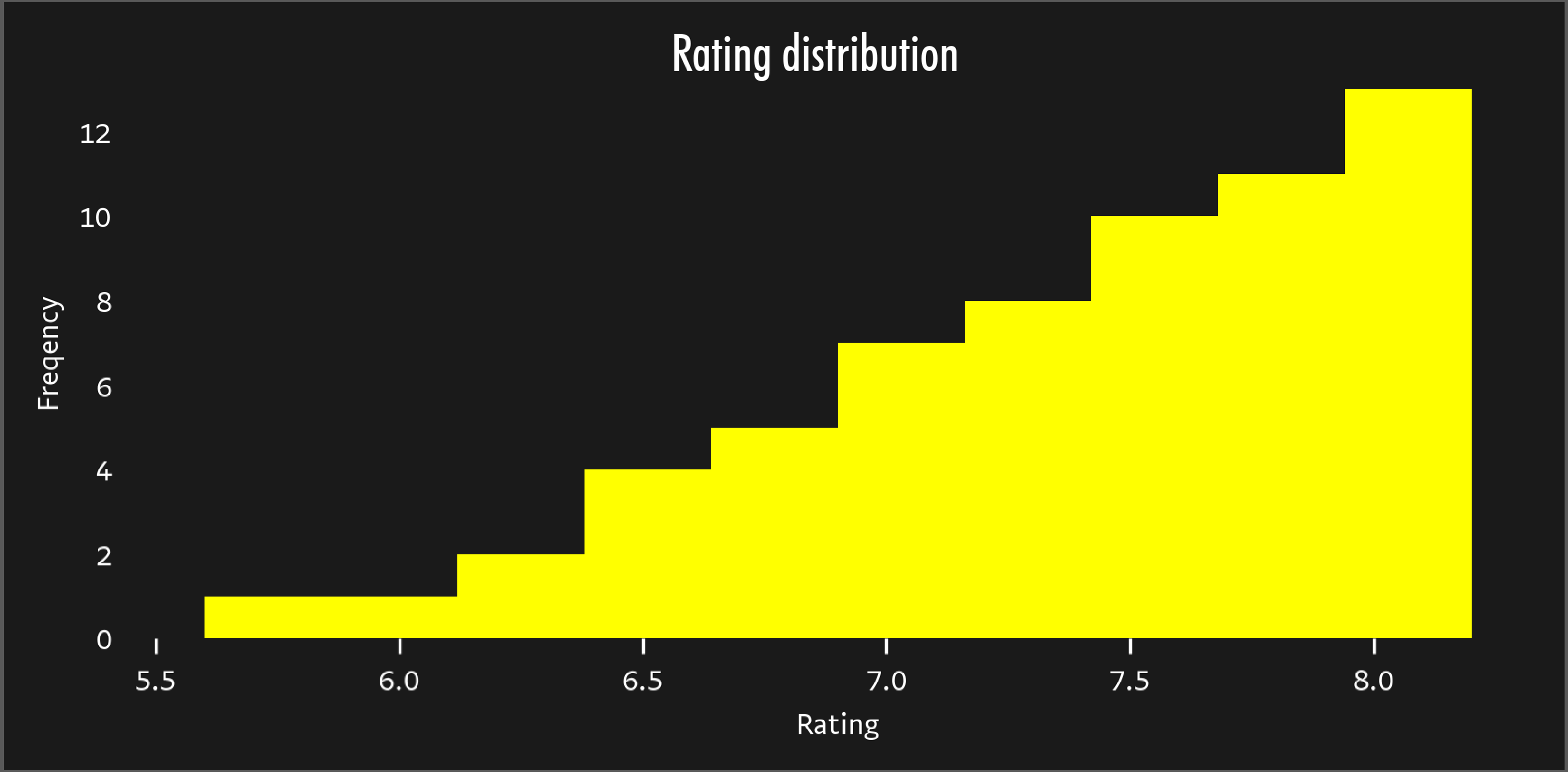
some findings

- How are these 13 Wall Street movies rated?
- Which movie on the list has the highest rating?
- What does the cumulative rating distribution look like?
- How long are these movies?
- Which decade had the most movies?
- Is there a relationship between rating and movie duration?

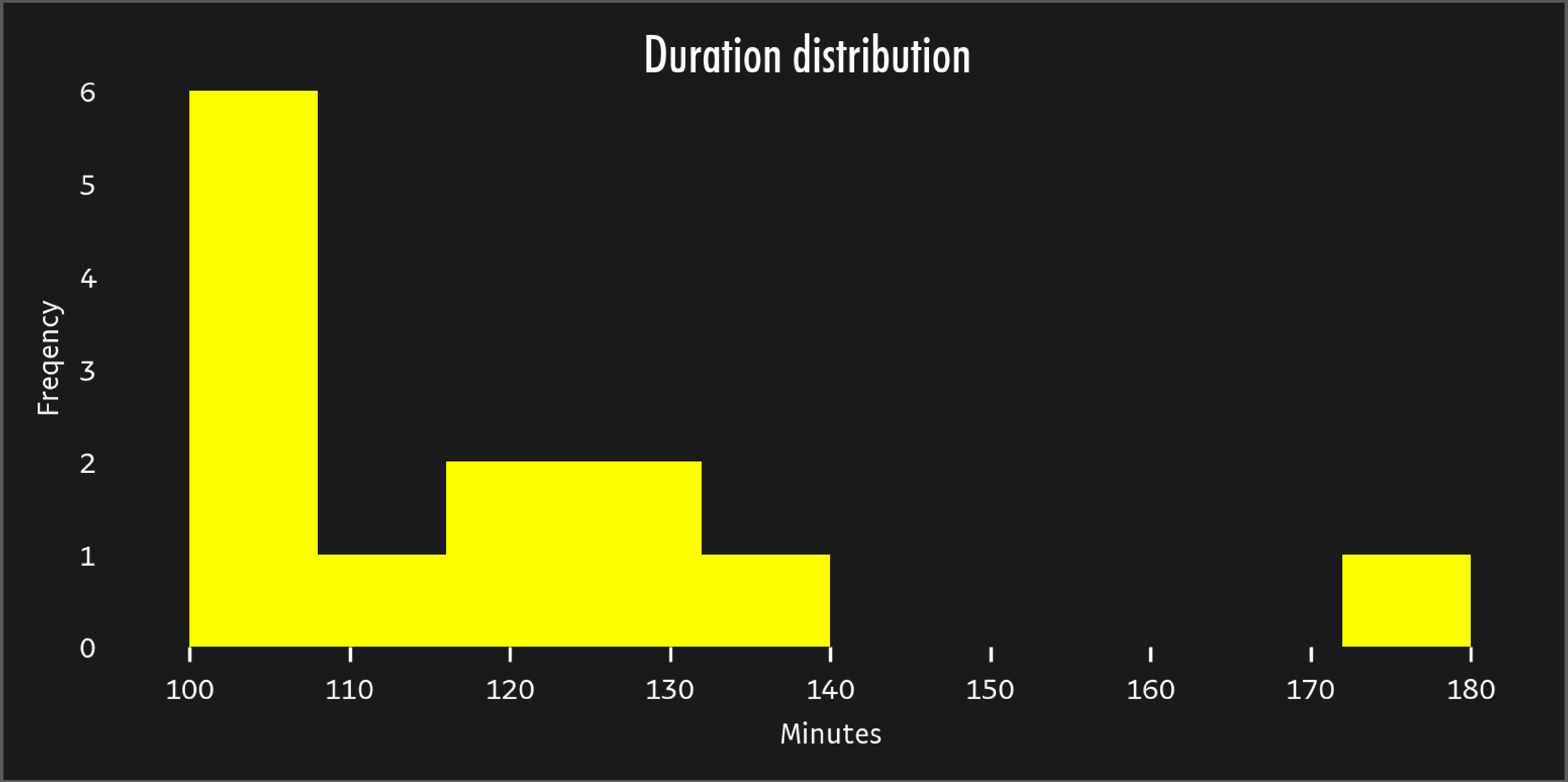
some findings



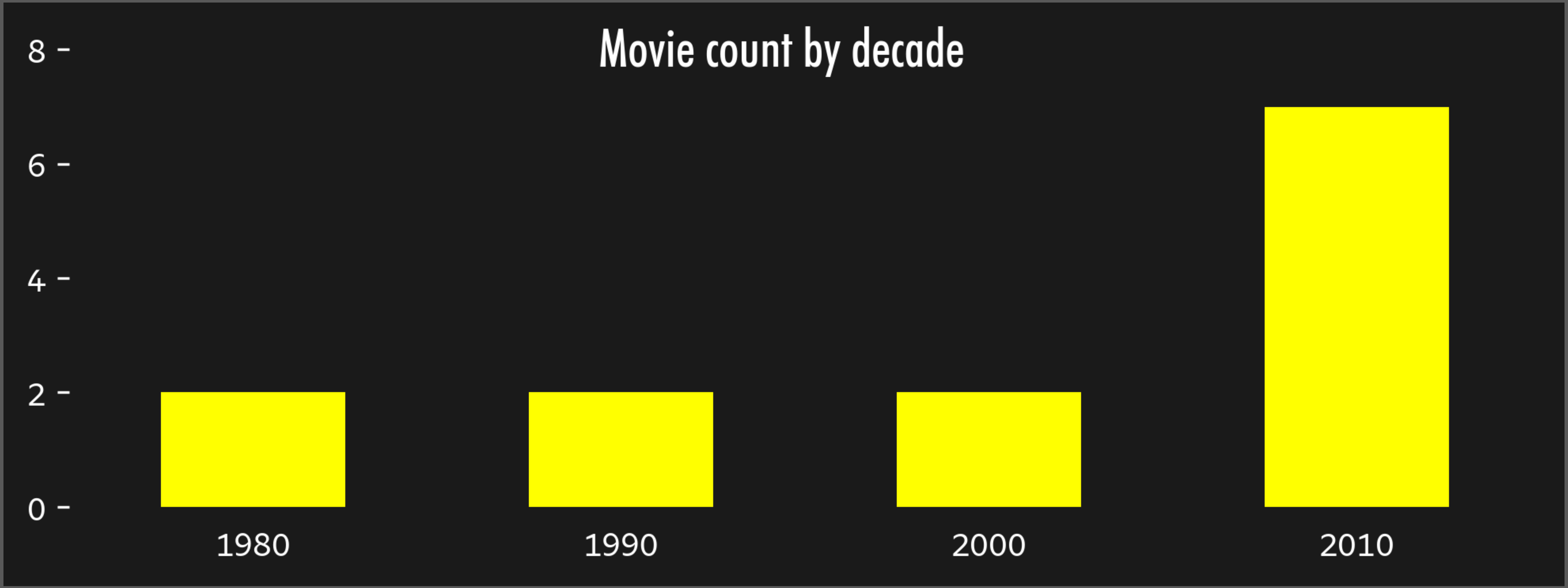
some findings



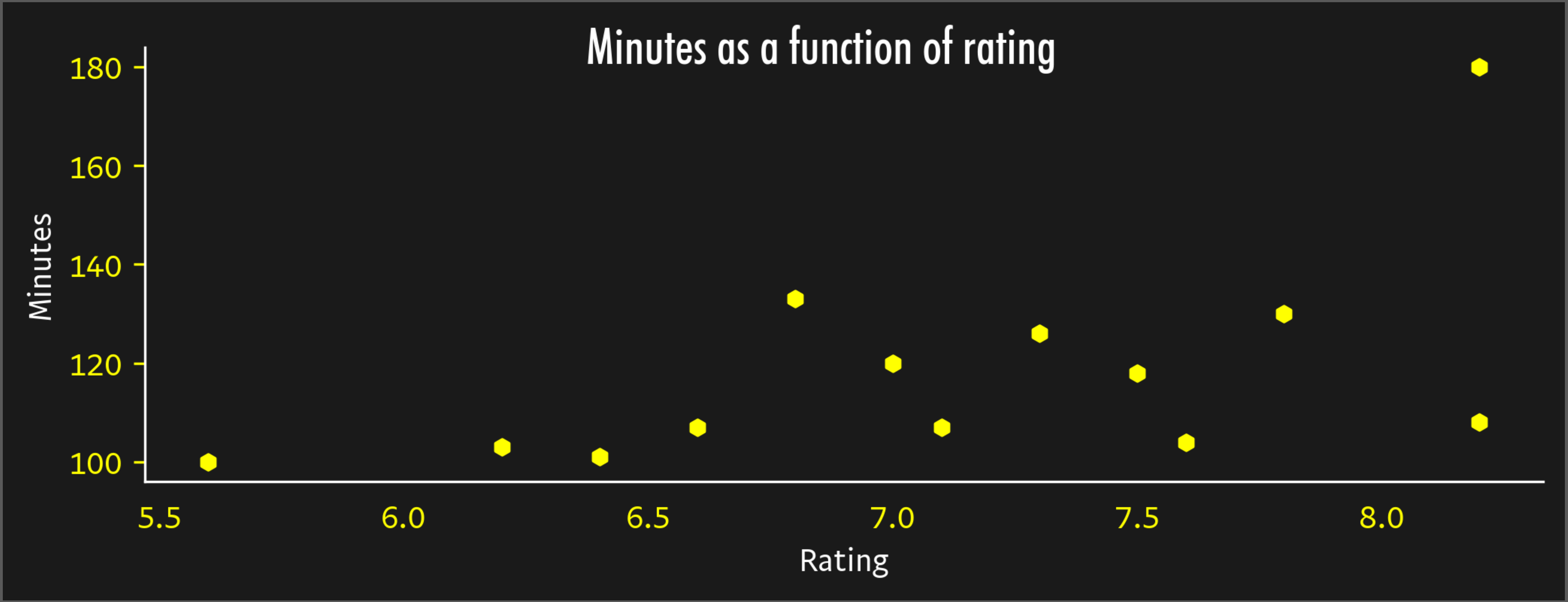
some findings



some findings



some findings



some findings

- What topics are contained in the movies?
- Are there some common topics?
- Sentiment vs polarity in the movies?

TOPICS
Equity

talk	cachet	fuck	fucking	ipo
work	people	company	call	take
something	money	day	never	time

some findings

TOPICS
Margin Call

fuck	today	fucking	people	firm
guys	last	time	years	jesus
work	day	understand	done	take

TOPICS
Arbitrage

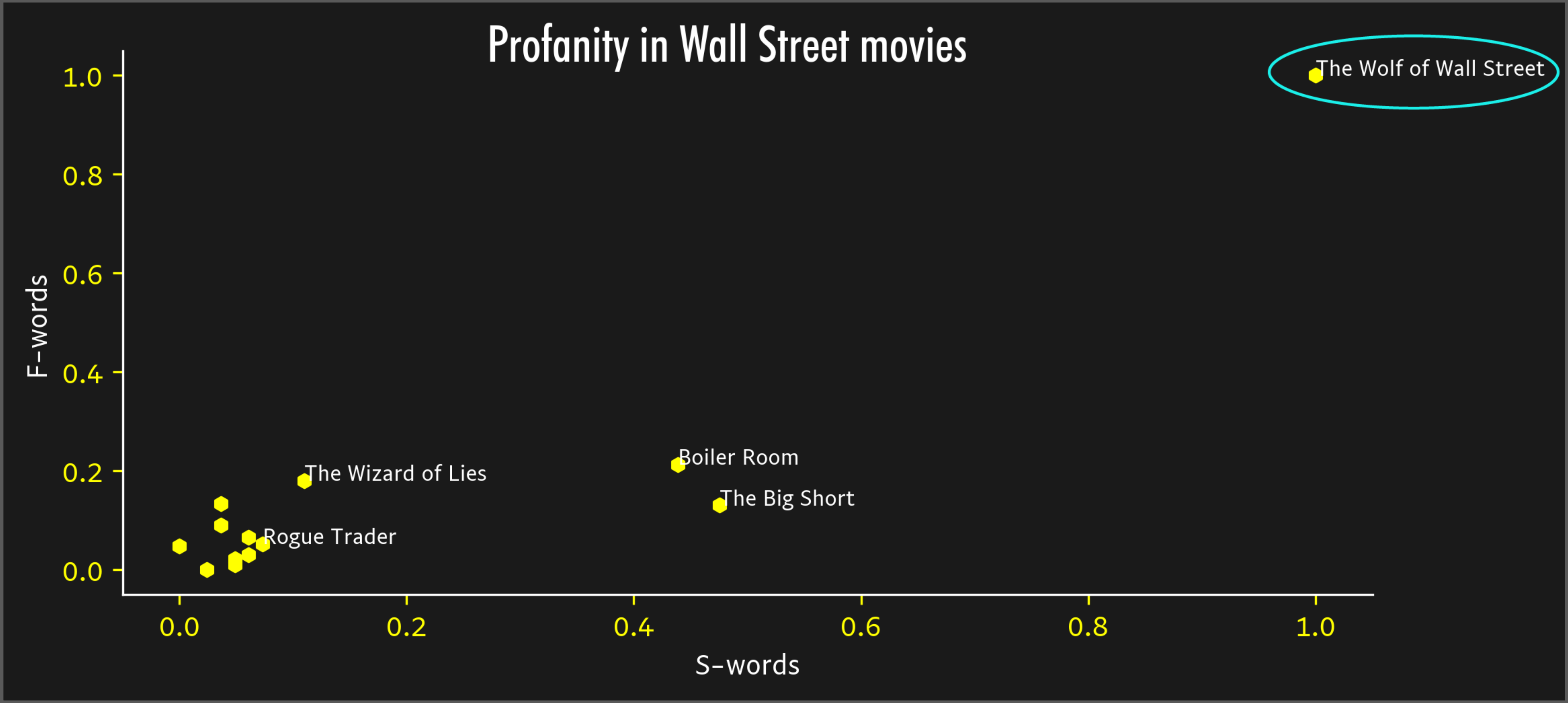
fucking	take	fuck	money	call
talk	something	deal	time	car
wait	detective	people	brooke	phone

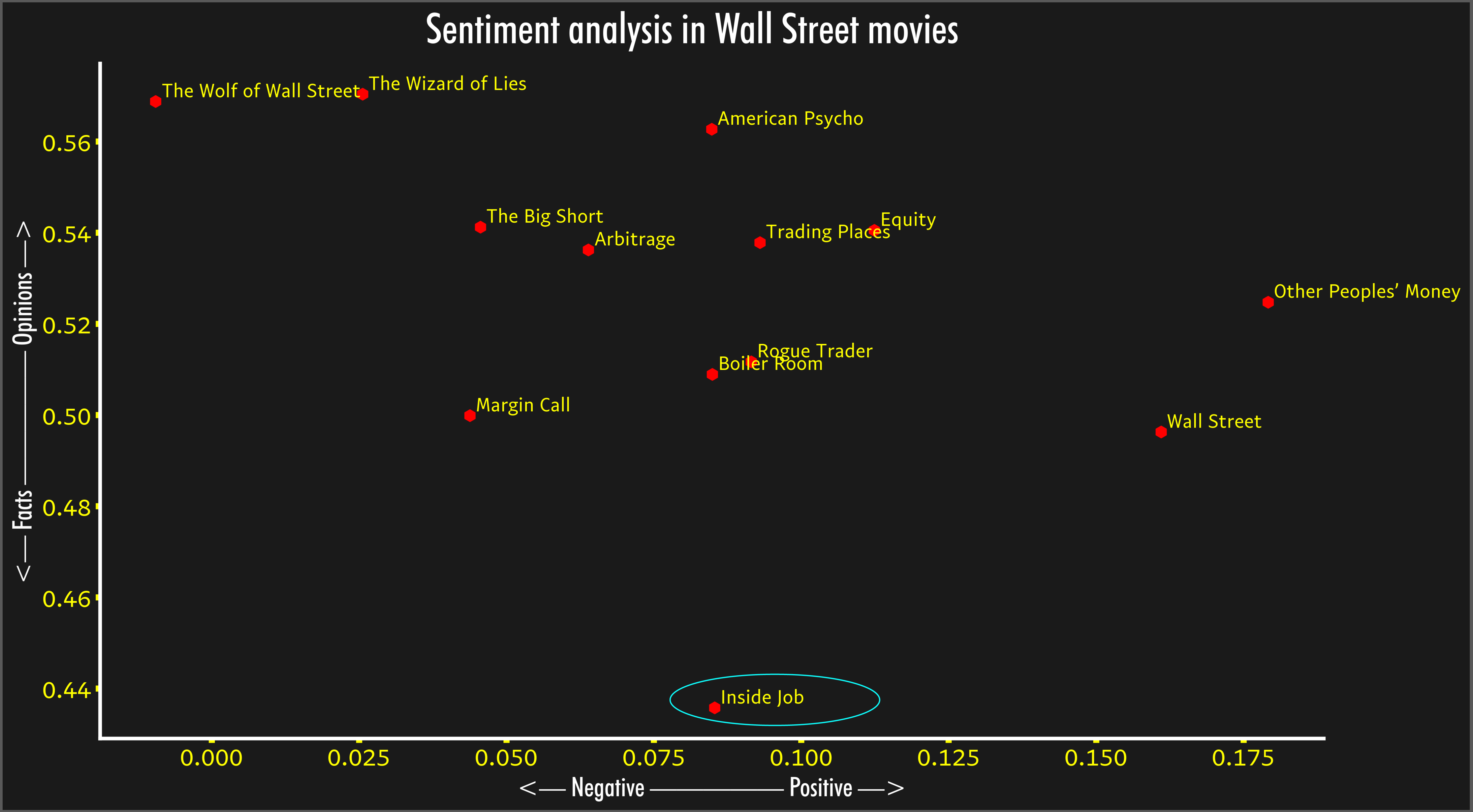
some findings

- This profanity can't be a coincidence. Let's investigate!

	Margin Call	Wall Street	The Big Short	Trading Places	The Wolf of Wall Street	American Psycho	Arbitrage	Equity	Inside Job	Boiler Room	Rogue Trader	The Wizard of Lies	Other Peoples' Money
f_word	45	15	65	11	498	24	67	33	0	106	26	90	5
s_word	6	8	42	7	85	3	6	8	5	39	9	12	7

some findings





Sentiment analysis in Wall Street movies
The Wolf of Wall Street
The Wizard of Lies
American Psycho
The Big Short
Arbitrage
Trading Places
Equity
Other Peoples' Money
Rogue Trader
Boiler Room
Margin Call
Wall Street
Inside Job
0.56
0.54
0.52
0.50
0.48
0.46
0.44
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
<— Facts ——— Opinions —>
<— Negative ——— Positive —>

takeaways

- Wall Street movies tend to be between 1:30 min to 2 hrs long.
- Wall Street movies are replete with profanity.
- They are more fictional than factual.
- They are very negative on the sentiment scale.