Dataset: IMDB

Data Wrangling

- I looked at the different values in columns with value counts.
- Then I split the values in the genres, production_companies, cast and keywords to lists in the same field.
- Question 3 null values and 0 values are omitted

Questions

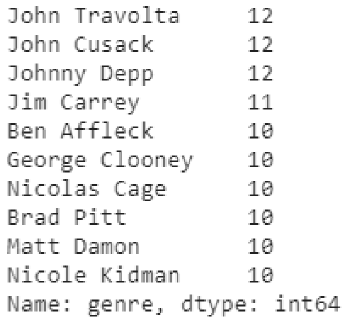
1. Which actor plays in the most different film genres?
2. Is there a connection between budget and revenue? And does it change over time?
3. Does runtime affect the rating?
4. What are the most popular keywords? And do they change over time?

Description of steps which I took to answer the questions

1. I created the product of actors list and genres. All actors that play in one movie with certain genres are a product of all the actors and genres of that movie. This will be done for all movies. Then a groupby is done per actor and genre such that the unique combinations are kept. Then I reset the index to get actor and genre back in the dataframe instead of the index and I did a value counts over the genres. Then I took the actors which play in 10 movies with 10 different genres or more.
2. I selected all movies with a budget and revenue higher than 0. And plotted them using a scatterplot. When this didn’t yield any result, I looked at all movies per year and decided on using bins. Years < 1990 are binned per 10 years and above per 5 years. Then I scatterplotted all the budget vs Revenues of the years. For clearance I added a regression line to indicate the trend.
3. I started with only picking the ‘normal movies’ which I deem to have a length of max 5 hours (300 minutes). Then I grouped by runtime to get the average popularity and vote average per runtime. Then I scatterplotted these. Visually identified outliers, removed them and scatterplotted them again.
4. I generated a dummy list of all the keywords and joined the general dataframe year information on the index. Then I binned them as in question 3 and barchart, summed all the keywords per bin and plotted the 20 most occurring keywords.

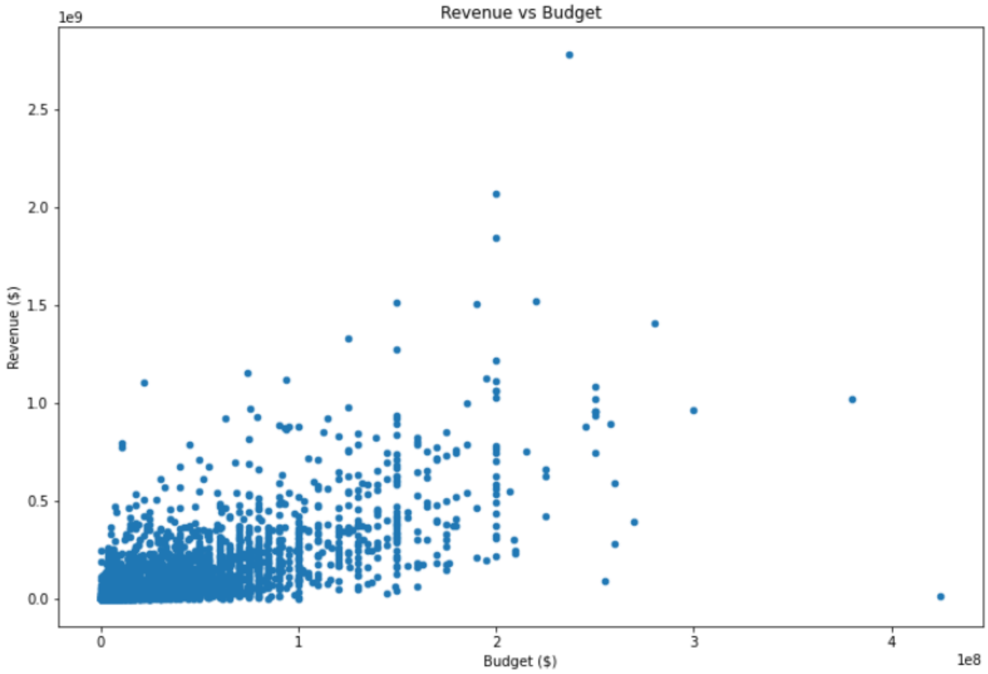
Summary statistics and plots communicating my findings

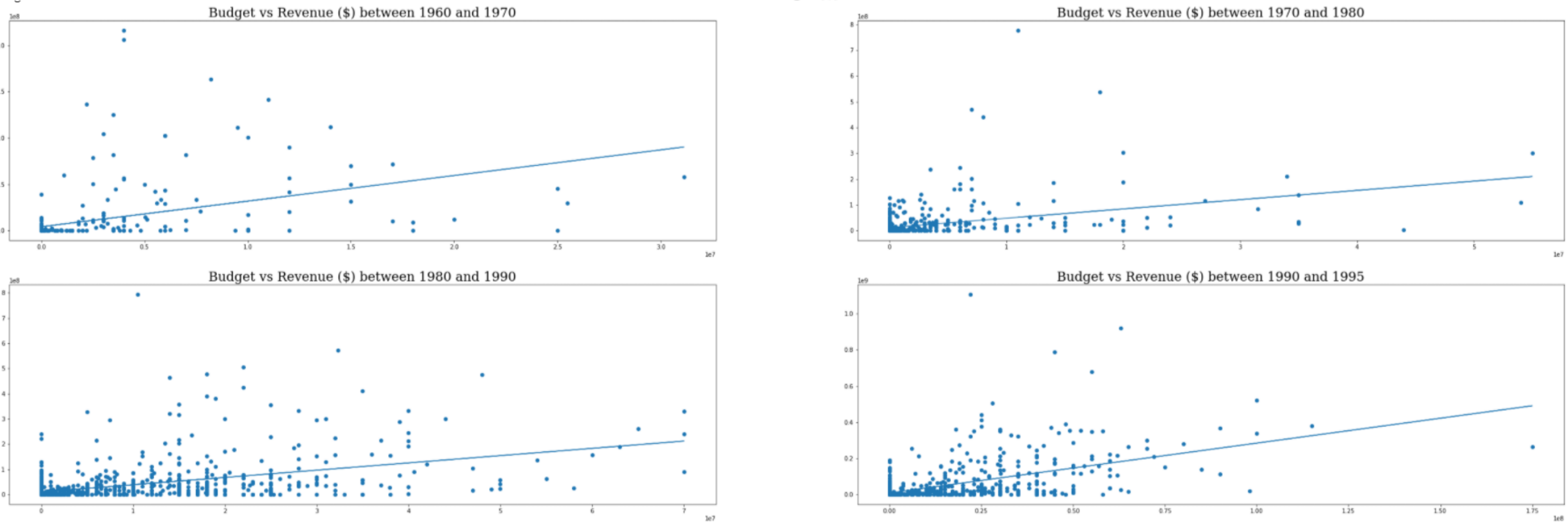
1.

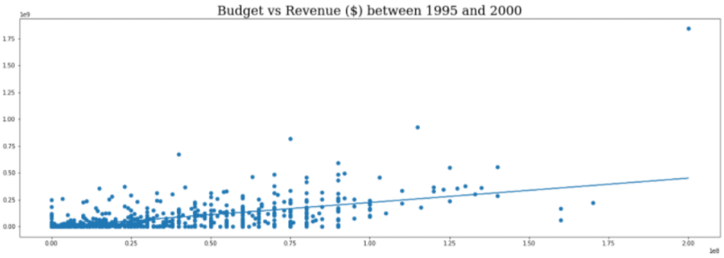


```
John Travolta      12
John Cusack        12
Johnny Depp        12
Jim Carrey         11
Ben Affleck        10
George Clooney     10
Nicolas Cage       10
Brad Pitt          10
Matt Damon         10
Nicole Kidman      10
Name: genre, dtype: int64
```

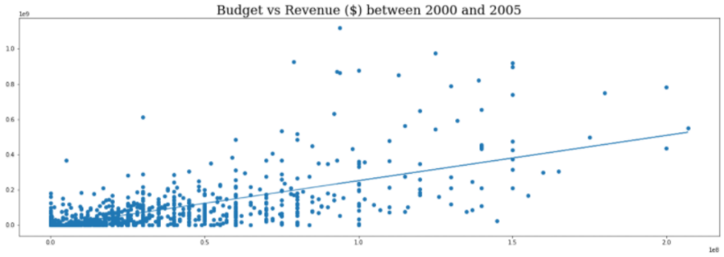
2.



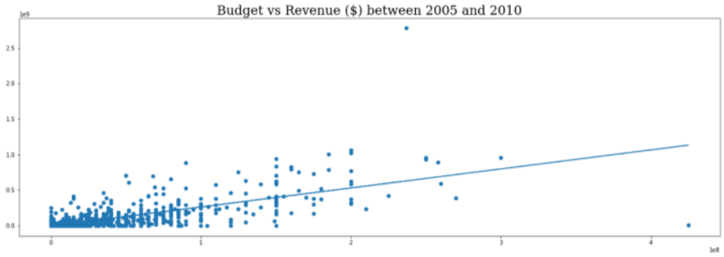




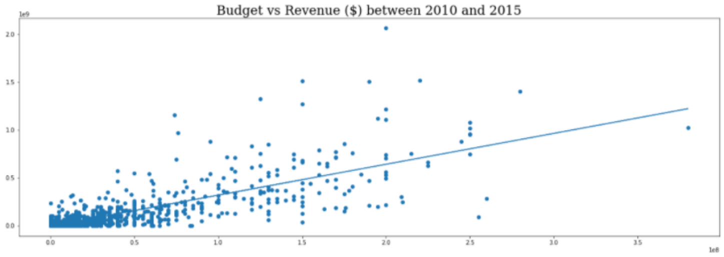
Budget vs Revenue ($) between 1995 and 2000



Budget vs Revenue ($) between 2000 and 2005

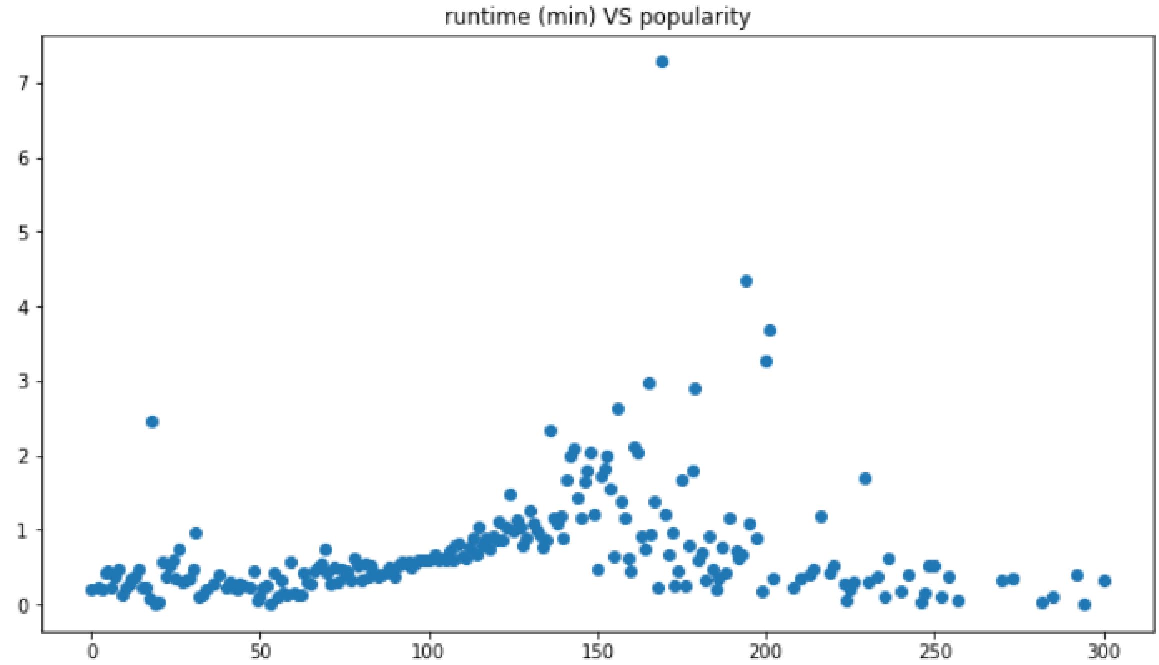


Budget vs Revenue ($) between 2005 and 2010



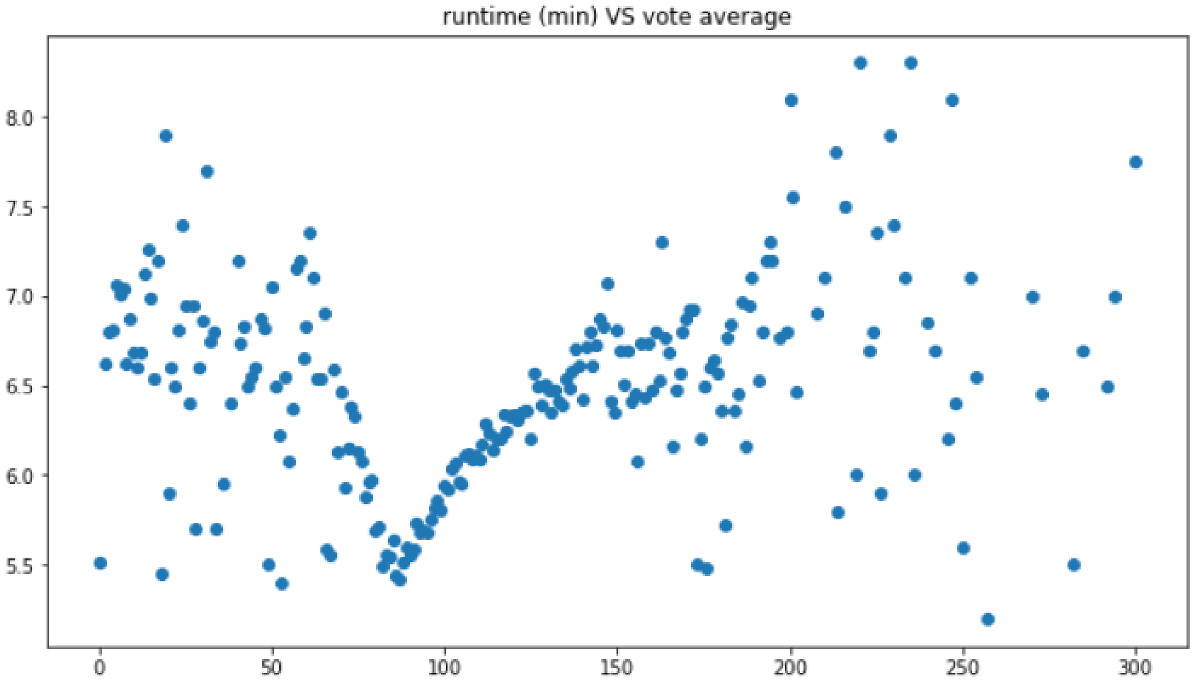
Budget vs Revenue ($) between 2010 and 2015

3.



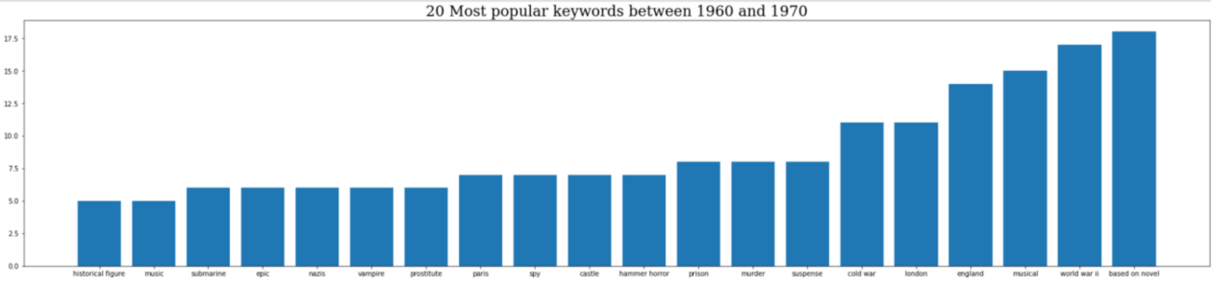
runtime (min) VS popularity

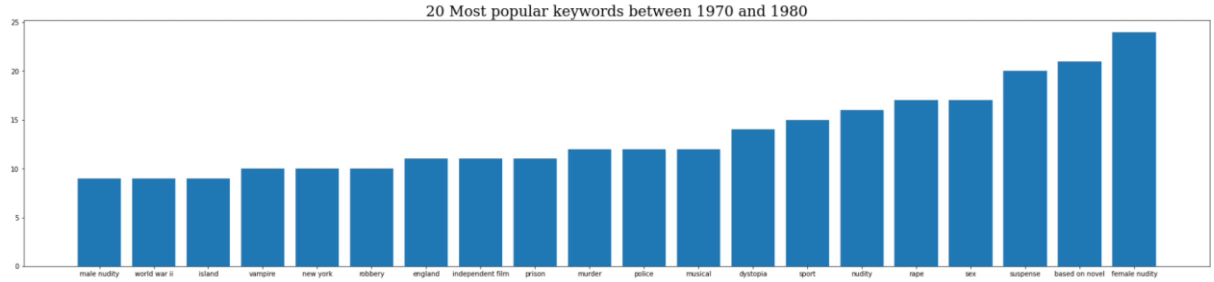
As is seen there is a an outlier in the runtime VS popularity scatterplot. Therefore, I deleted this outlier (It was Jurassic world).

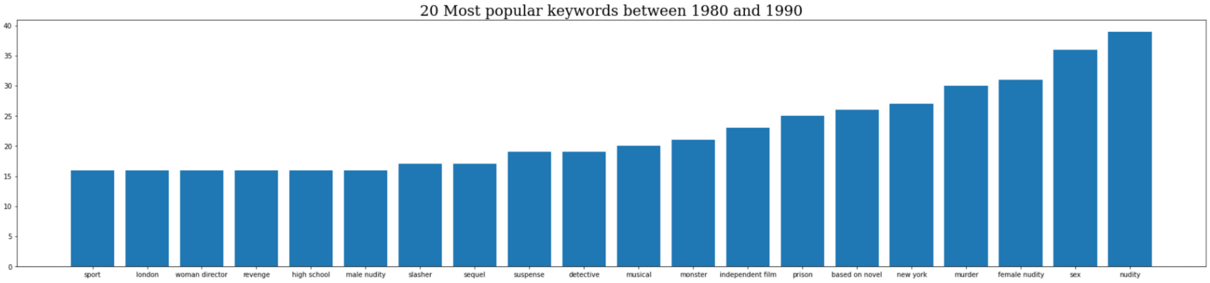


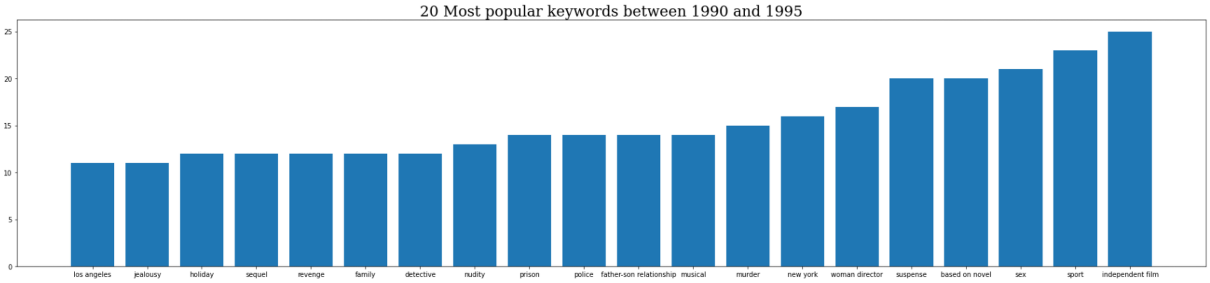
runtime (min) VS vote average

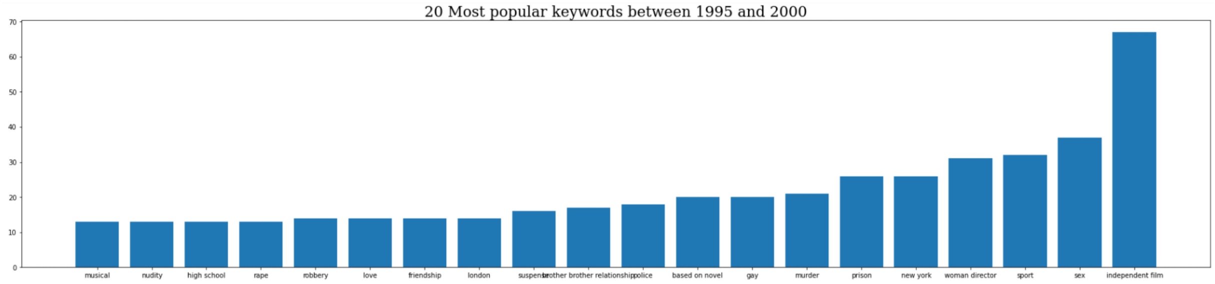
4.

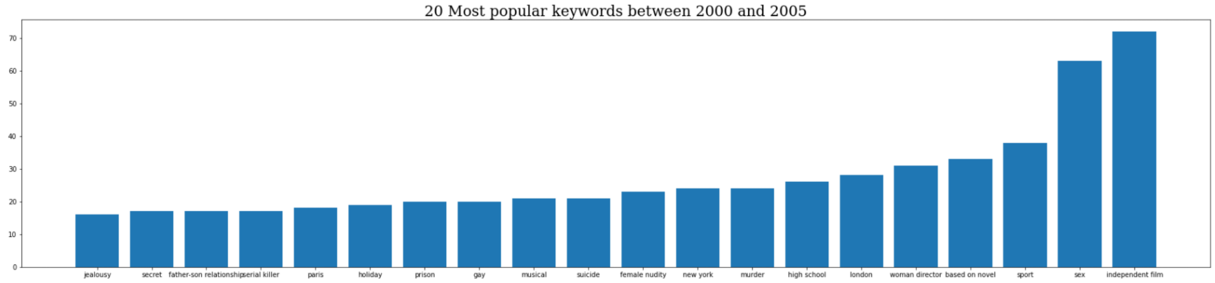

20 Most popular keywords between 1960 and 1970

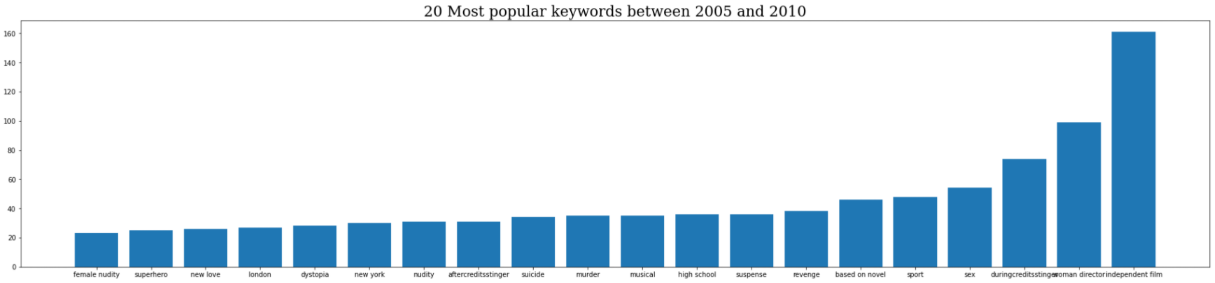

20 Most popular keywords between 1970 and 1980

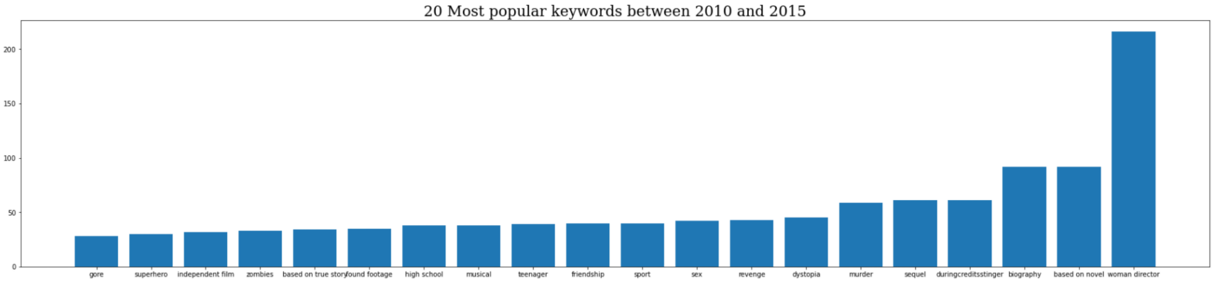

20 Most popular keywords between 1980 and 1990


20 Most popular keywords between 1990 and 1995


20 Most popular keywords between 1995 and 2000
0
10
20
30
40
50
60
70
musical
nudity
high school
rape
robbery
love
friendship
london
suspense
brother brother relationship
police
based on novel
gay
murder
prison
new york
woman director
sport
sex
independent film


20 Most popular keywords between 2000 and 2005
0
10
20
30
40
50
60
70
jealousy
secret
father-son relationship
serial killer
paris
holiday
prison
gay
musical
suicide
female nudity
new york
murder
high school
london
woman director
based on novel
sport
sex
independent film


20 Most popular keywords between 2005 and 2010
0
20
40
60
80
100
120
140
160
female nudity
superhero
new love
london
dystopia
new york
nudity
aftercreditsstinger
suicide
murder
musical
high school
suspense
revenge
based on novel
sport
sex
duringcreditsstinger
woman director
independent film


20 Most popular keywords between 2010 and 2015
0
50
100
150
200
gore
superhero
independent film
zombies
based on true story
found footage
high school
musical
teenager
friendship
sport
sex
revenge
dystopia
murder
sequel
duringcreditsstinger
biography
based on novel
woman director

Conclusions

1. The 3 actors with that play the most different genres start their name with 'John'. Except "John Cusack" they are all familiar to me in some sense. Even though I'm not interested in actors this indicates that the (most of the) actors are probably more recent actors. This is in line with the value count where it is obvious that there are more films produced more recently. Which could imply actors play more movies but also could imply that there are more diverse movies to play in are available. I haven't investigated the spread of available genres yet.
2. Over the dataset there is no clear conclusion that shows that there is a correlation between budget and revenue. When looking at the binned years there is in earlier years no clear conclusion from the scatter cloud. The later time periods 1995-2000, 2005-2010 and 2010-2015 show a clearer trend. In all time periods the regression line is positive which coincides with my hypothesis that there is at least a positive correlation between films with higher budget and higher revenue.
3. Popularity: Here we can easily see that movies with runtime lower than +/- 120 minutes in general more popular if they have more runtime. Where films that are especially long 230 min+ are less popular. The scatterplot sort of forms an upward wig.

   Vote Average: There seems to be a big spread in in movies with less than 70 minutes runtime and more than 150 minutes runtime. Between 70 and 150 minutes The lowest vote average is around 85 minutes and is increasing after in a more or less straight line up.
4. A lot can be inferred from this. Some conclusions are:
   a. Between 1990 and 2010 Independent films were popular. And that number drastically dropped after 2010
   b. There is an increasing surge in woman director films. Either in the profession or it being mentioned in the keywords.
   c. Between 1970 and 2010 sex or nudity has always been in the top 3 keywords.
   d. In earlier periods there were far less movies. In the category of 1960 and 1970 if 2 more movies mentioned music in the keywords it would shoot from number 20 to number 10 in popular keywords

Limitation

All findings above are about the movie database that is mentioned in the project. This does not necessarily reflect all movies but only the movies available to us.

Things I would like to have tips/help with

- I have some problems with x/y-labels in subplots Q3/Q4
- Also I wanted to rotate the x-labels in Q4. I did this once in Seaborn but could not get this working in this instance.
- In question 3 I wanted to set the axis on a set frame such that the graphs would be better comparable.

Sources

https://matplotlib.org/stable/gallery/text_labels_and_annotations/text_fontdict.html

https://stackoverflow.com/questions/12444716/how-do-i-set-the-figure-title-and-axes-labels-font-size-in-matplotlib

https://stackoverflow.com/questions/25239933/how-to-add-title-to-subplots-in-matplotlib

https://stackoverflow.com/questions/14770735/how-do-i-change-the-figure-size-with-subplots

https://matplotlib.org/stable/gallery/lines_bars_and_markers/scatter_with_legend.html

https://stackoverflow.com/questions/44970881/matplotlib-multiple-scatter-subplots-with-shared-colour-bar

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.subplots.html

https://www.kite.com/python/answers/how-to-plot-a-linear-regression-line-on-a-scatter-plot-in-python

https://stackoverflow.com/questions/29034928/pandas-convert-a-column-of-list-to-dummies