Dataset: IMDB

General description data

The dataset used is the IMDB dataset.

Id	Unique id
Imdb_id	Unique id
Popularity	How popular the movie is
Budget	Budget spent in dollars
Revenue	Revenue made in dollars
Original_title	Original movie title
Cast	Main cast of the movie
Homepage	url of the website of the movie if available
Director	Director of the movie
Tagline	A short sentence regarding the movie
Keywords	Multiple words which describe the movie
Overview	Short description of the movie
Runtime	Duration of the movie in minutes
Genre	Genre(s) that fit the movie
Production_companies	The businesses that produced the movie
Release_date	Date that the movie is shown
Vote_count	Amount of votes for this movie
Vote_average	Average vote for this movie
Release_year	Year in which the movie was released
Budget_adj	Budget adjusted to 2015
Revenue_adj	Revenue adjusted to 2015

	id	popularity	budget	revenue	runtime	vote_count	vote_average	release_year	budget_adj	revenue_adj
count	10844.000000	10844.000000	1.084400e+04	1.084400e+04	10844.000000	10844.000000	10844.000000	10844.000000	1.084400e+04	1.084400e+04
mean	65891.714958	0.647223	1.463269e+07	3.989250e+07	101.292973	217.736997	5.972824	2001.317595	1.756178e+07	5.145429e+07
std	92037.269216	1.000966	3.091919e+07	1.171062e+08	25.248327	576.142810	0.934143	12.819824	3.431520e+07	1.447581e+08
min	5.000000	0.000188	0.000000e+00	0.000000e+00	0.000000	10.000000	1.500000	1960.000000	0.000000e+00	0.000000e+00
25%	10588.750000	0.208132	0.000000e+00	0.000000e+00	90.000000	17.000000	5.400000	1995.000000	0.000000e+00	0.000000e+00
50%	20604.500000	0.384145	0.000000e+00	0.000000e+00	99.000000	38.000000	6.000000	2006.000000	0.000000e+00	0.000000e+00
75%	75171.000000	0.714716	1.500000e+07	2.409062e+07	111.000000	146.000000	6.600000	2011.000000	2.085887e+07	3.386406e+07
max	417859.000000	32.985763	4.250000e+08	2.781506e+09	300.000000	9767.000000	8.900000	2015.000000	4.250000e+08	2.827124e+09

- Popularity differs a lot for all movies. 75% of the movies have a popularity of less then 0.714 while the highest is 32.98. There are a few movies which are a lot more popular than the rest such as Jurassic World(33), Mad Max Fury Road (28), Interstellar(25) and Guardians of the Galaxy (14).
- Atleast 50% of the movies have budget of 0. The same with revenue.
- Some movies have a runtime of 0 minutes. Which seems implausible, but it could be short movies of less than 29seconds which are rounded down to 0 minutes.
- The vote count has a standard deviation 2,5x the mean. Which means that there is a lot of differences between the vote amounts.

```
id                      0
imdb_id                10
popularity              0
budget                  0
revenue                 0
original_title          0
cast                   76
homepage             7930
director               44
tagline              2824
keywords             1493
overview                4
runtime                 0
genres                 23
production_companies 1030
release_date            0
vote_count              0
vote_average            0
release_year            0
budget_adj              0
revenue_adj             0
dtype: int64
```

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.
- The rows with null values in cast, keywords, and genres are omitted in the Questions regarding those columns.
- The columns imdb_id, homepage, director, tagline, overview, Production_companies are not used in the questions. Therefore, the Null values are neglected.
- The homepage column should be dropped.
- In future wranglings year will be binned in the wrangling stage instead Of during the question.

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

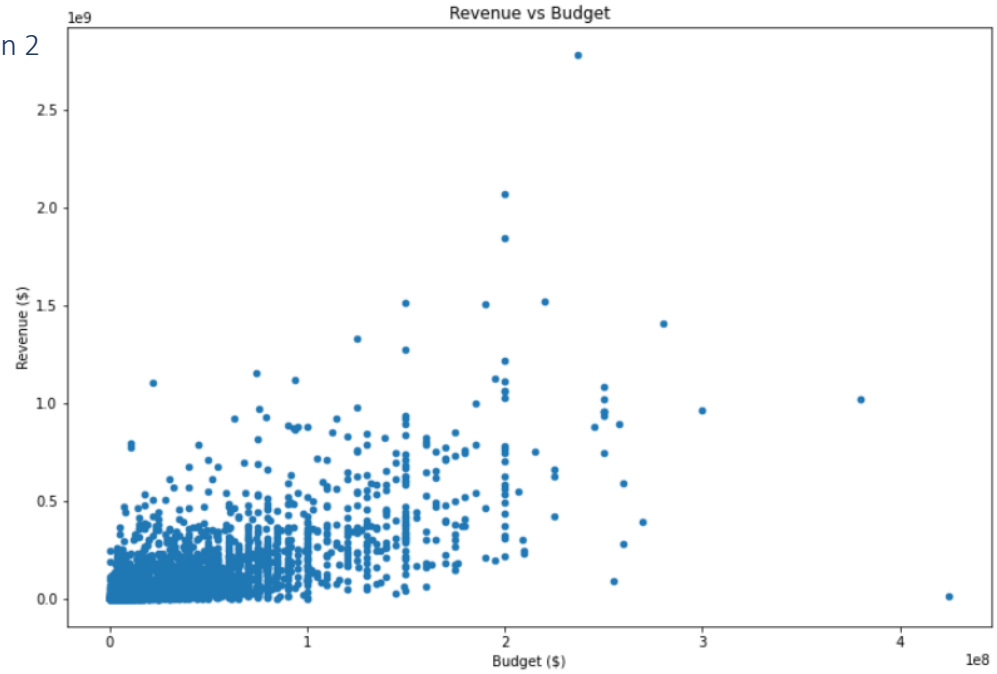
1. I created the product of actors list and genres. All actors that play in one movie with certain genres are a product of all the actors and genres of that movie. This will be done for all movies. Then a groupby is done per actor and genre such that the unique combinations are kept. Then I reset the 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.

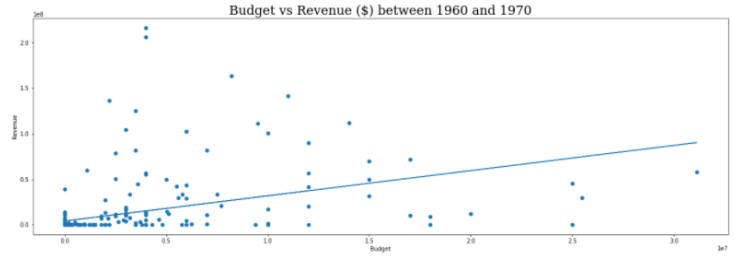
Summary statistics and plots communicating my findings

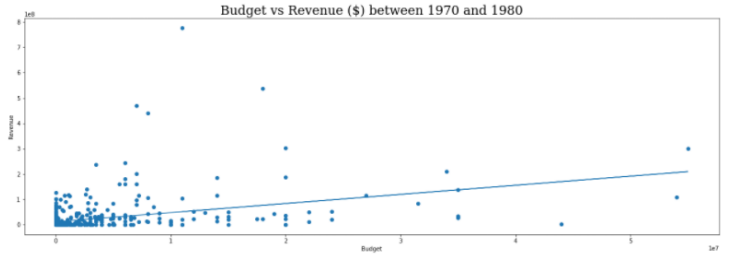
Question 1

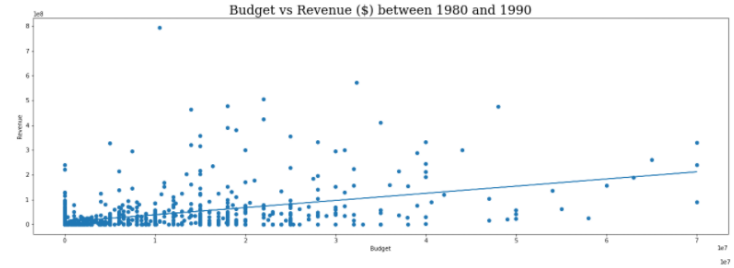
```
John Travolta      12
John Cusack        12
Johnny Depp        12
Jim Carrey         11
Ben Affleck        10
George Clooney     10
Nicolas Cage       10
Brad Pitt          10
Matt Damon         10
Nicole Kidman      10
Name: genre, dtype: int64
```

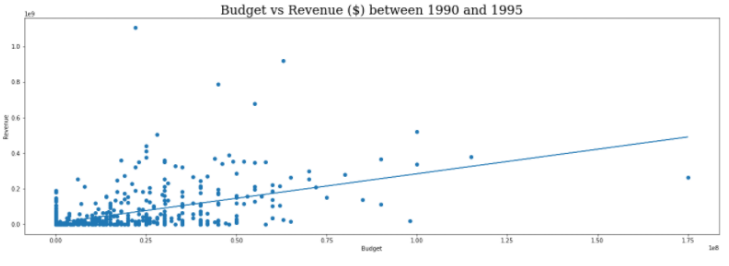
Question 2

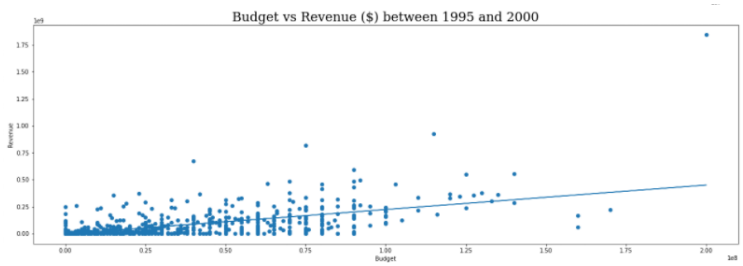


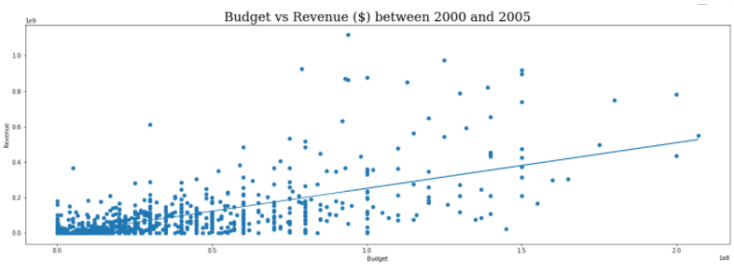


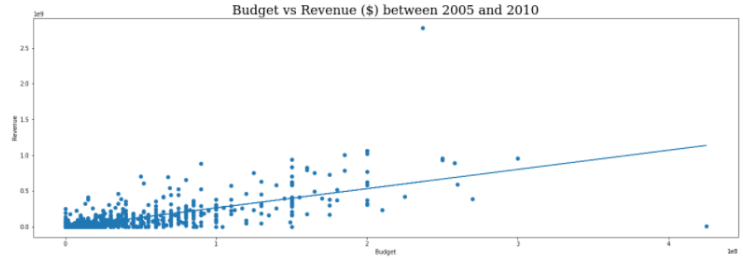


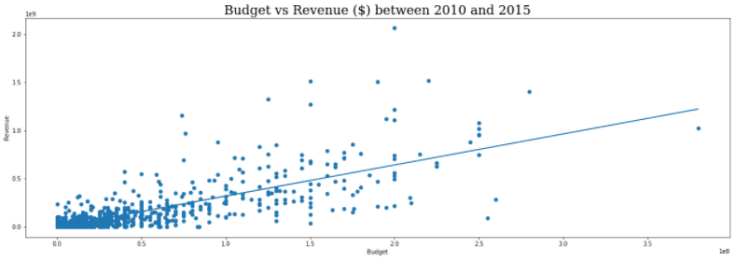




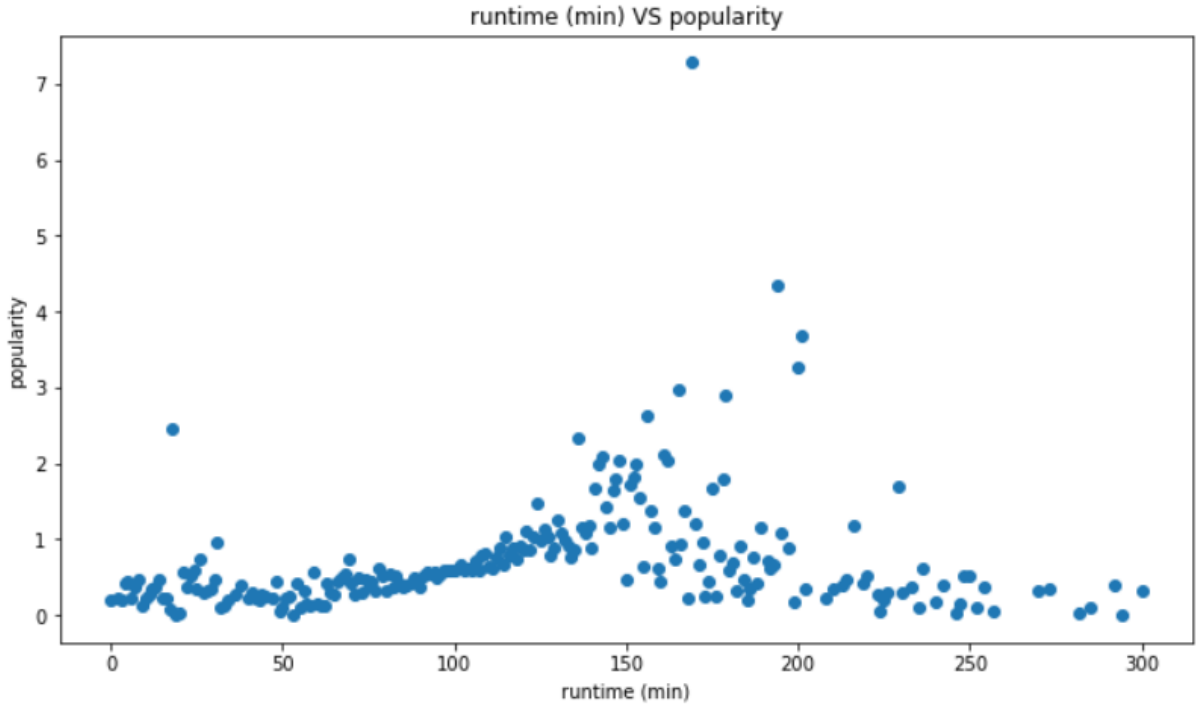

Budget vs Revenue ($) between 1995 and 2000
1e9
Revenue
Budget
1e8

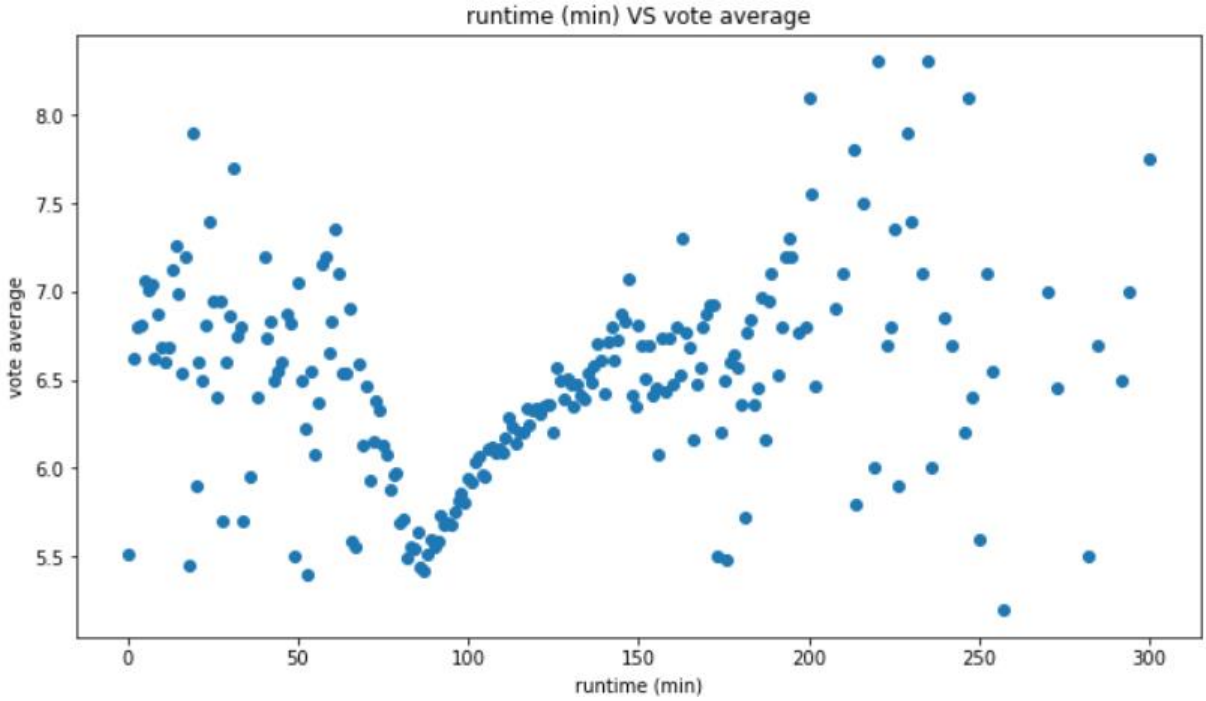

Budget vs Revenue ($) between 2000 and 2005
1e9
Revenue
Budget
1e8


Budget vs Revenue ($) between 2005 and 2010
1e9
Revenue
Budget
1e8

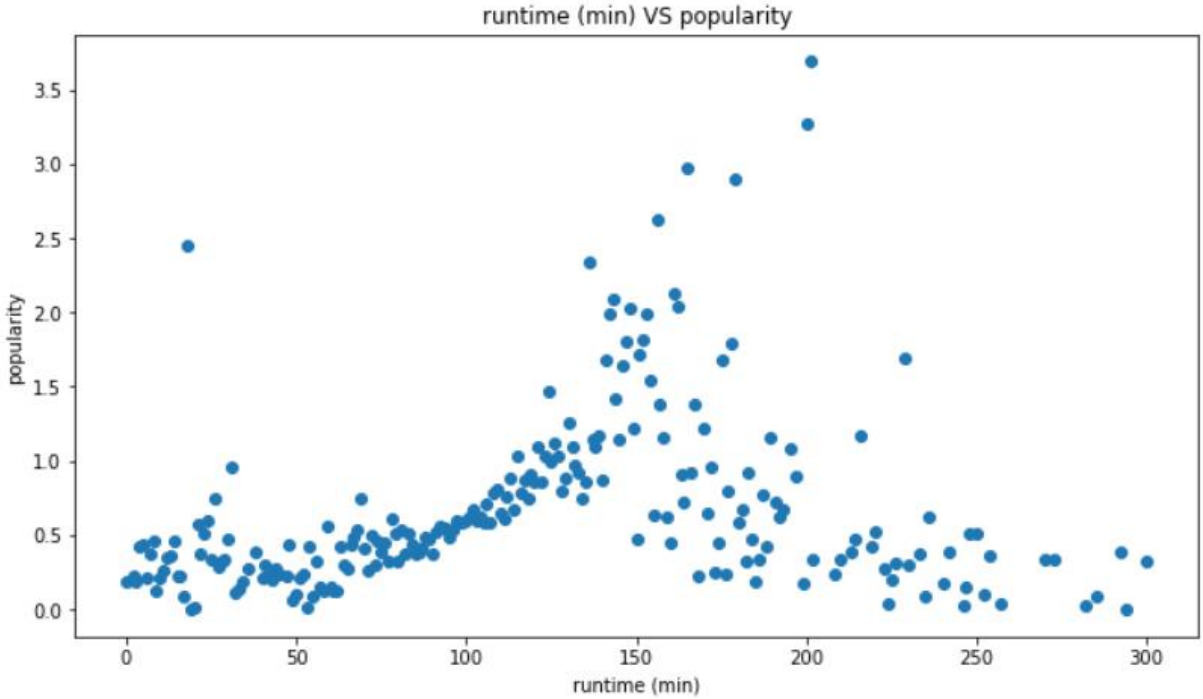

Budget vs Revenue ($) between 2010 and 2015
1e9
Revenue
Budget
1e8

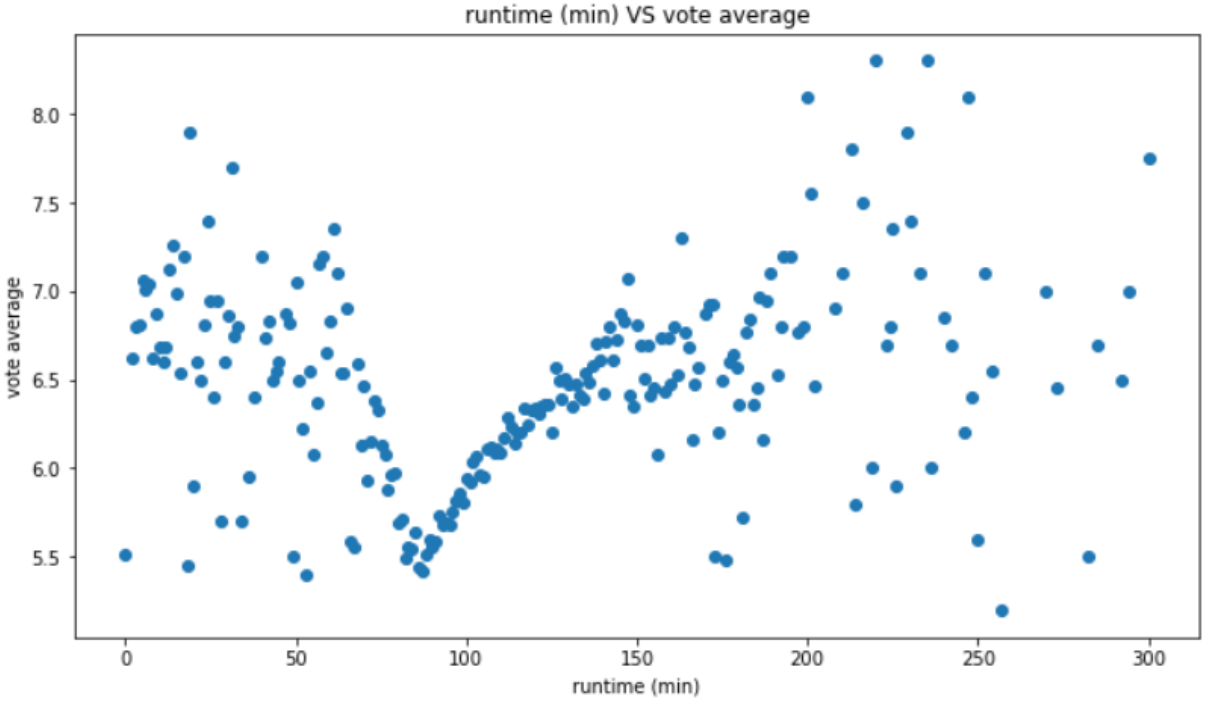
Question 3



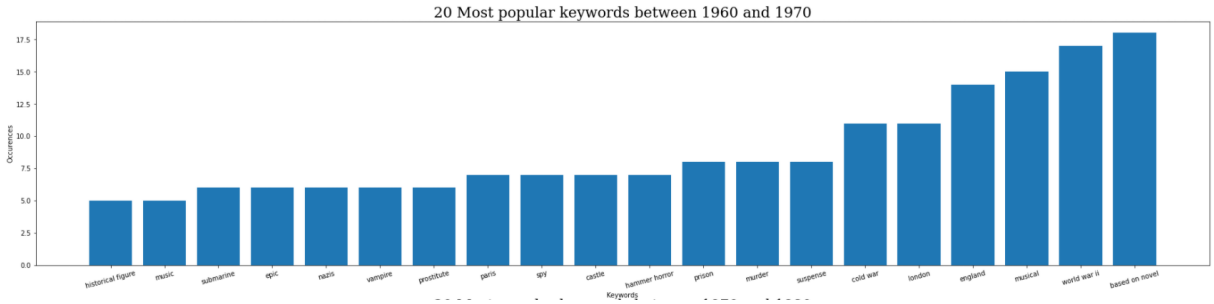


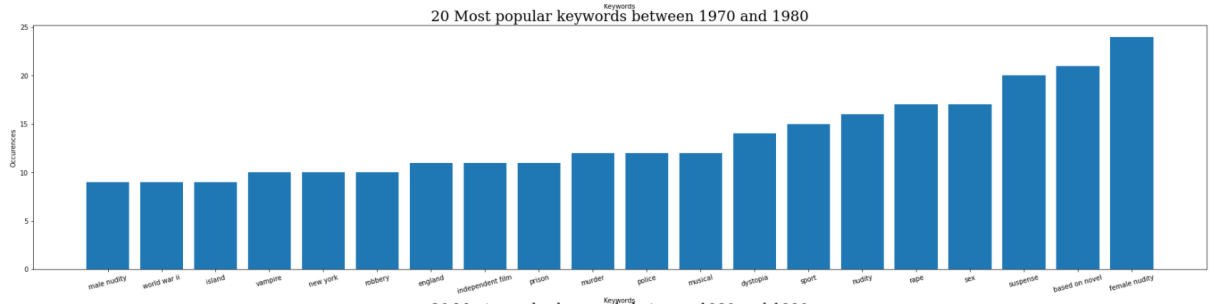
As is seen there is a an outlier in the runtime VS popularity scatterplot. Therefore, I deleted this outlier (It was Jurassic world).

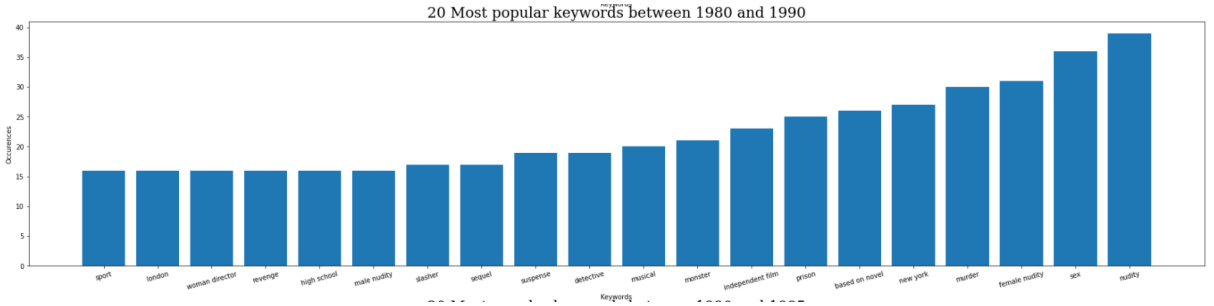

runtime (min) VS popularity
popularity
runtime (min)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
50
100
150
200
250
300

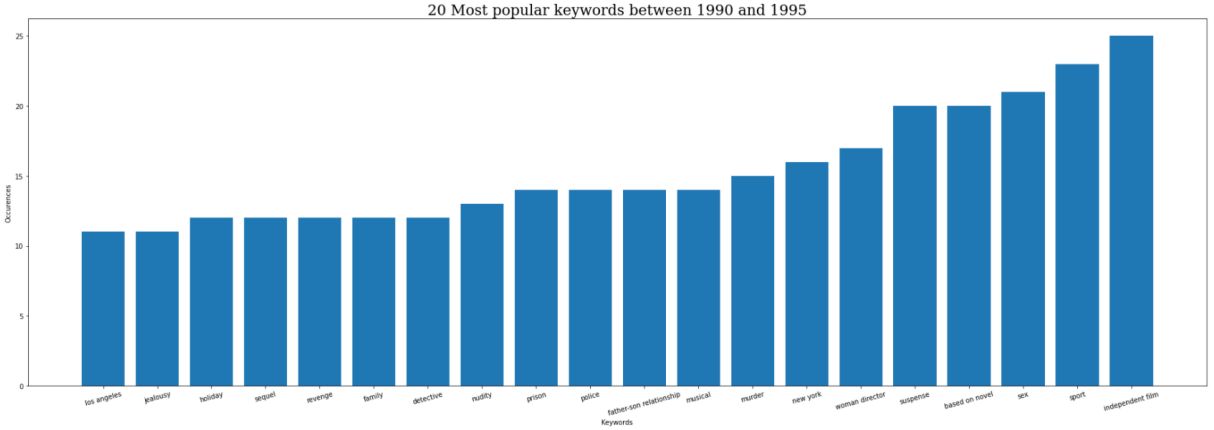

runtime (min) VS vote average
vote average
runtime (min)
5.5
6.0
6.5
7.0
7.5
8.0
0
50
100
150
200
250
300

Question 4

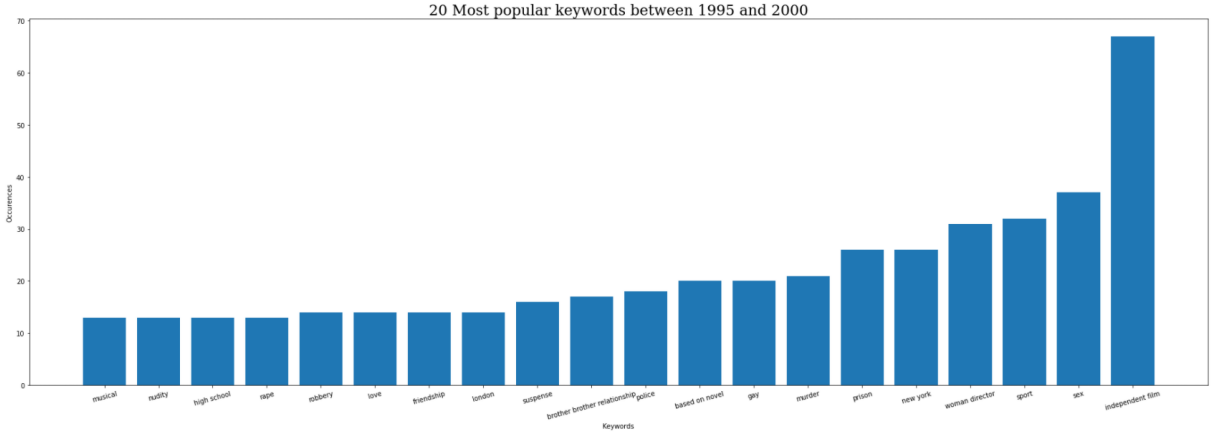

20 Most popular keywords between 1960 and 1970
Occurences
historical figure
music
submarine
epic
nazis
vampire
prostitute
paris
spy
castle
hammer horror
prison
murder
suspense
cold war
london
england
musical
world war ii
based on novel
Keywords


20 Most popular keywords between 1970 and 1980
Occurences
male nudity
world war ii
island
vampire
new york
robbery
england
independent film
prison
murder
police
musical
dystopia
sport
nudity
rape
sex
suspense
based on novel
female nudity
Keywords

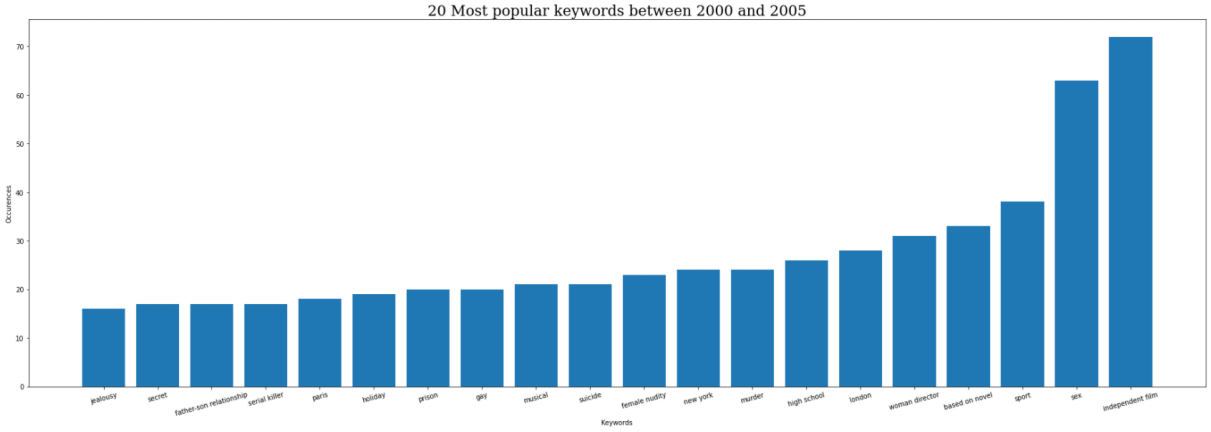

20 Most popular keywords between 1980 and 1990
Occurences
sport
london
woman director
revenge
high school
male nudity
slasher
sequel
suspense
detective
musical
monster
independent film
prison
based on novel
new york
murder
female nudity
sex
nudity
Keywords


20 Most popular keywords between 1990 and 1995
Occurences
los angeles
jealousy
holiday
sequel
revenge
family
detective
nudity
prison
police
father-son relationship
musical
murder
new york
woman director
suspense
based on novel
sex
sport
independent film
Keywords

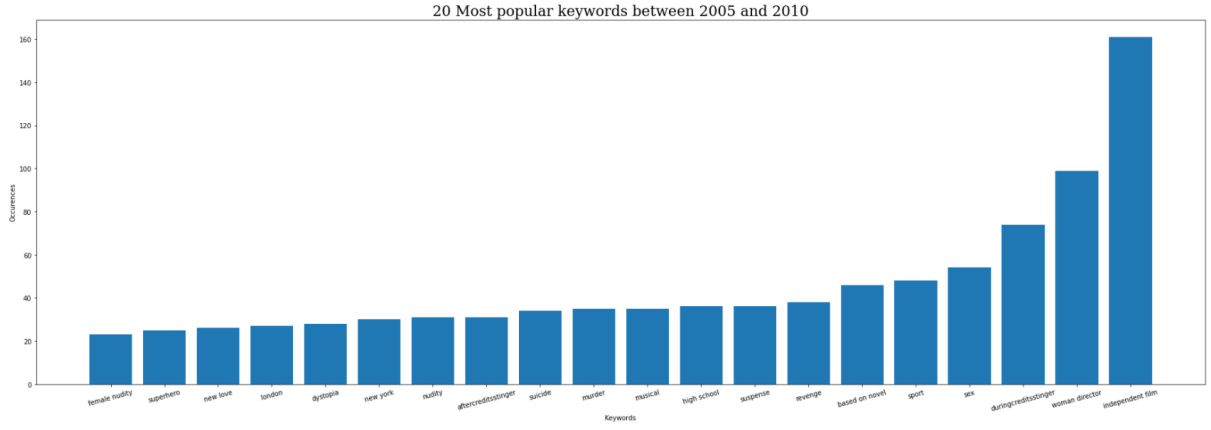
20 Most popular keywords between 1995 and 2000



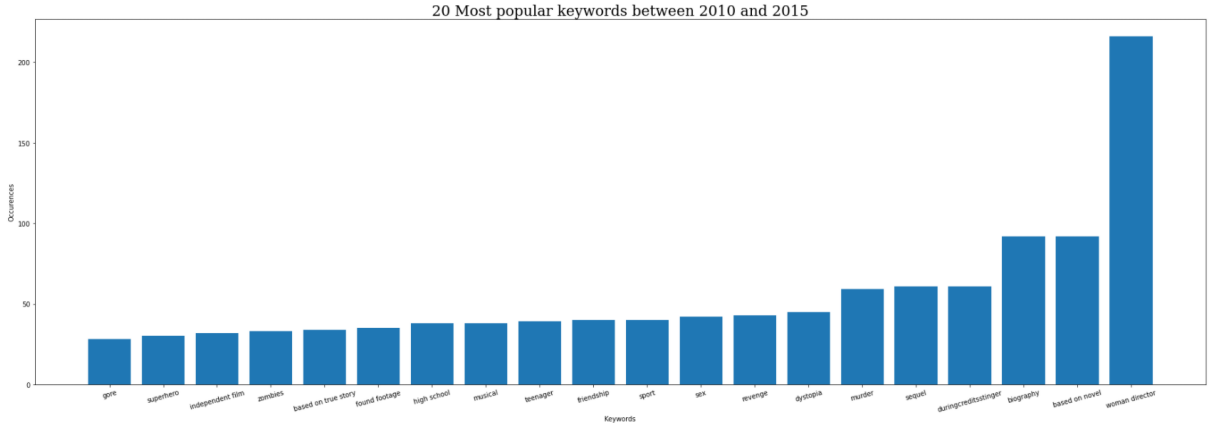
20 Most popular keywords between 2000 and 2005



20 Most popular keywords between 2005 and 2010



20 Most popular keywords between 2010 and 2015



Conclusions

Question 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.

Question 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.

Question 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.

Question 4

A lot can be inferred from this. Some conclusions are:

1. Between 1990 and 2010 Independent films were popular. And that number drastically dropped after 2010
2. There is an increasing surge in woman director films. Either in the profession or it being mentioned in the keywords.
3. Between 1970 and 2010 sex or nudity has always been in the top 3 keywords.
4. 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

Limitations

As a general remark I would like to clarify that all findings are about the movie database that is mentioned in the project. This does not necessarily reflect all movies ever made but only the movies in the dataset. Correlations between variables in this dataset may not be correlations outside of this dataset.

In the first section, the data that is used is could be skewed because there are more different genres nowadays (or less, I didn't look into this). Therefore, 'newer' actors play more different genres. This could be solved by more closely inspecting the genres changing over time.

In the second section, there is no normalization or currency conversion. Therefore, we are limited to the numerical values of budget and revenue. Besides, a lot of movies have budgets and revenues equal to 0 in the dataset. These are omitted from the calculations. There could be a common property as to why those are missing. This could lead to a skew in the results.

In the third section, movies are averaged over their runtime. There aren't an equal amount of movies for every runtime. The lower and higher runtimes therefore are less 'normalized' through a lot of samples and carry a higher variation. (Which is also visible in the graph). Those averages are less reliable. Next to that, some movies could be country specific i.e. movies in Russian language without subtitles. More people from Russia would vote on that movie compared to the average movie. This could lead to a bias in the results. For example, Russian voters tend to give higher/lower votes than Canadian voters.

In the last section, I looked at the keyword occurrences in different time bins. Some keywords didn't exist in the early time periods. Therefore, a comparison would be inappropriate. Besides, there were a lot less movies in the earlier years, thus possible creating a skew and be more susceptible to outliers.

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