Introduction

This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue. Certain columns, like ‘cast’ and ‘genres’,

contain multiple values separated by pipe (|) characters.

There are some odd characters in the ‘cast’ column. Don’t worry about cleaning them. You can leave them as is.

The final two columns ending with “_adj” show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.

Questions:

1. Which genres are more popular and profitable
2. Which company make the most profit
3. What keywords are the most popular
4. Which movies are the most popular or profitable
5. The correlation between features

Investigation process

Data Wrangling

1. Remove features with many NaN values
2. Remove features that are not considered('homepage', 'overview', 'release_date', 'budget','revenue', 'imdb_id')
3. Remove rows with NaN value

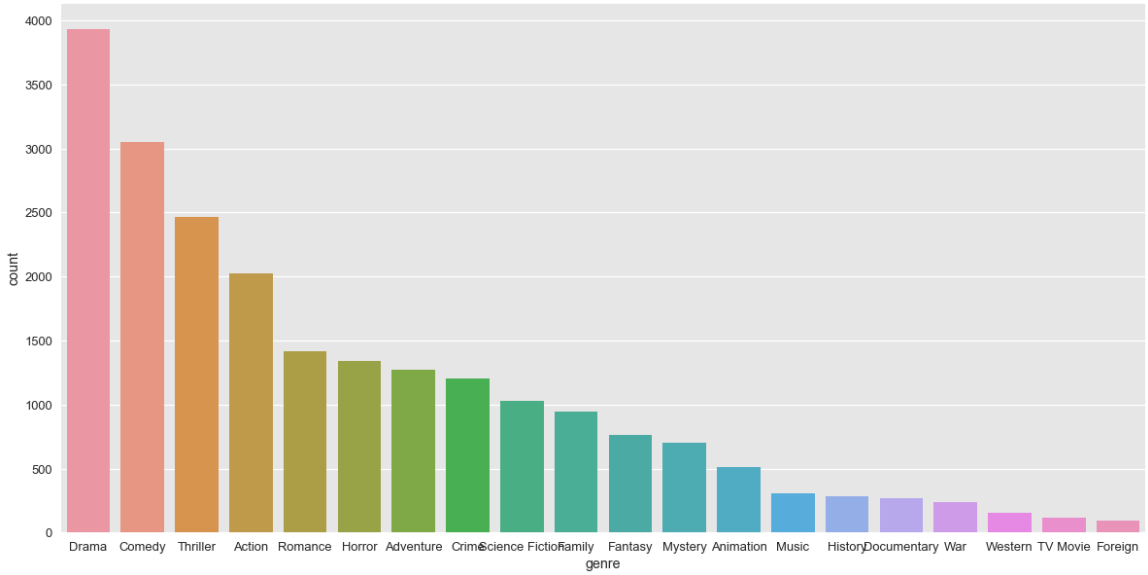
Exploratory Data Analysis

Summary statistics

	id	popularity	runtime	vote_count	vote_average	release_year	budget_adj	revenue_adj
count	8667.000000	8667.000000	8667.000000	8667.000000	8667.000000	8667.000000	8.667000e+03	8.667000e+03
mean	57001.470520	0.741920	103.813776	264.187031	6.003496	2000.130957	2.130849e+07	6.361717e+07
std	86237.736398	1.087596	26.205981	635.161318	0.893552	13.220940	3.710067e+07	1.592796e+08
min	5.000000	0.000188	0.000000	10.000000	1.500000	1960.000000	0.000000e+00	0.000000e+00
25%	9817.500000	0.250007	91.000000	20.000000	5.500000	1993.000000	0.000000e+00	0.000000e+00
50%	16219.000000	0.449397	100.000000	54.000000	6.100000	2004.000000	2.139935e+06	1.905315e+05
75%	54002.000000	0.842818	113.000000	200.000000	6.600000	2011.000000	2.811797e+07	5.521258e+07
max	417859.000000	32.985763	705.000000	9767.000000	8.700000	2015.000000	4.250000e+08	2.827124e+09

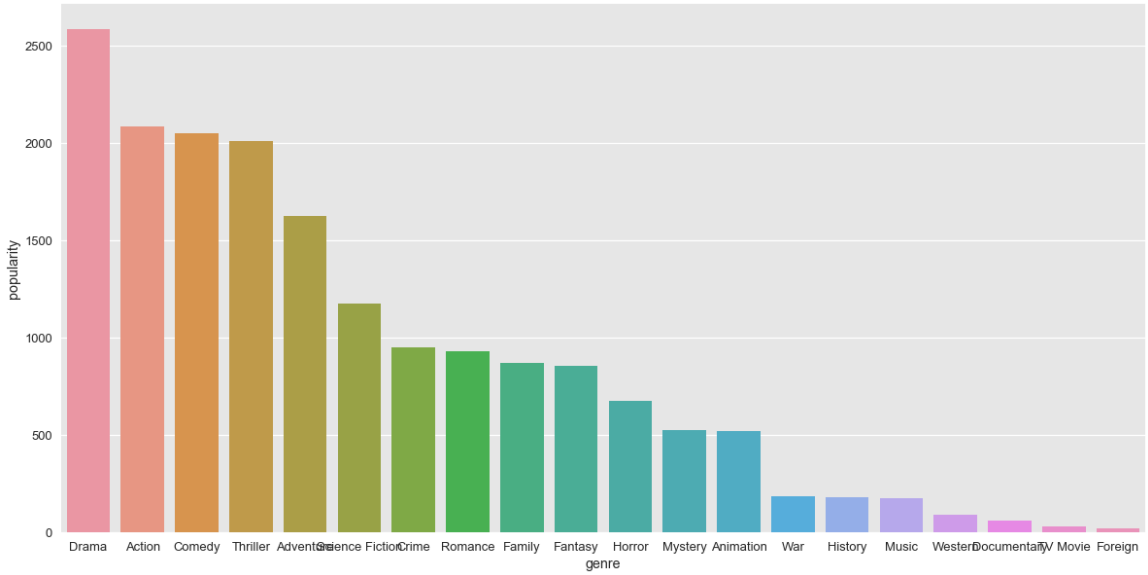
Research Question 1 (Which genres are more popular and profitable)

1. split the genres which a movie belong to from combined value into single genre value
2. count the number of each genre



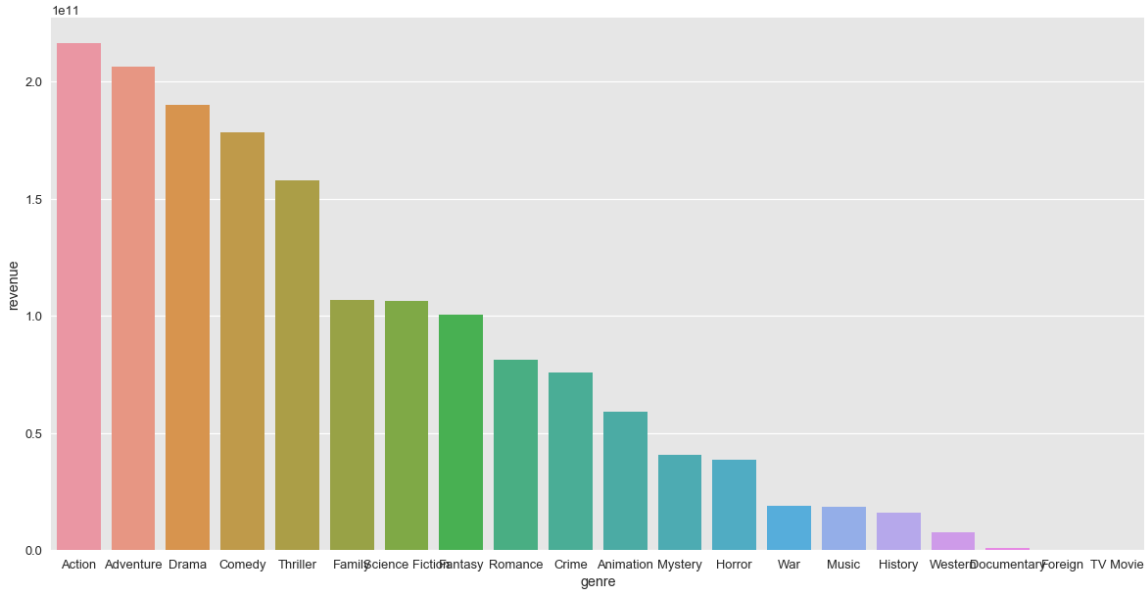
Drame, Comedy, Thriller and Action are four most-made genres. The number of drama is the most.

3. calculate popularity of each genre



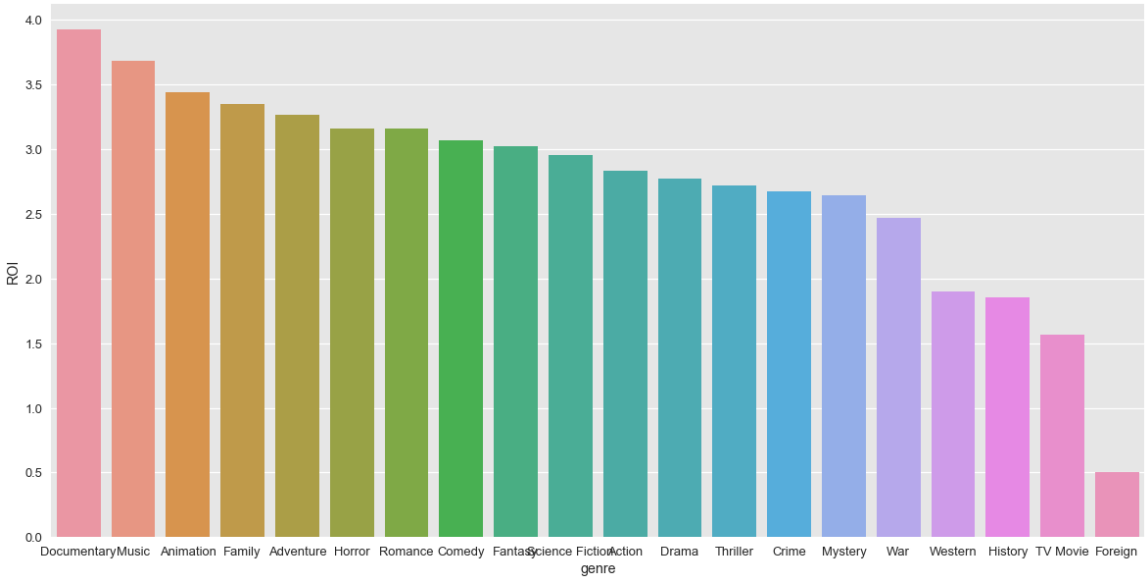
Drama is the most popular genre, following by action, comedy and thriller.

4. calculate revenue for each genre



Action, Adventure, Drama, Comedy and Thriller are the most profitable genres.

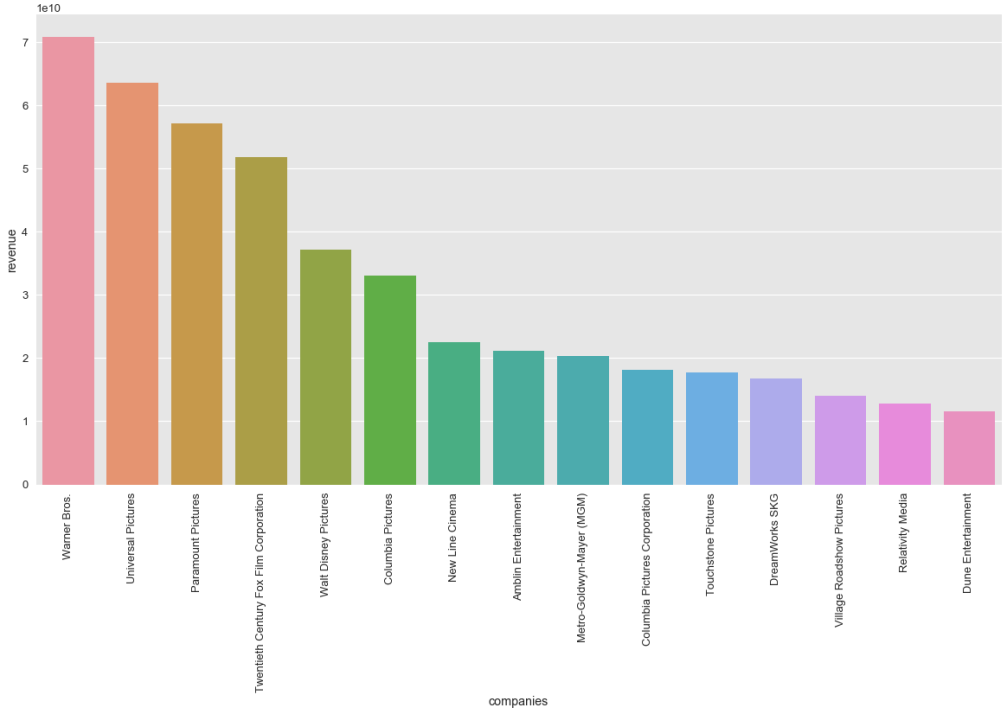
5. calculate ROI for each genre



Documentary and Music have highest ROI with all others having fairly close ROI

Research Question 2 (Which company make the most profit)

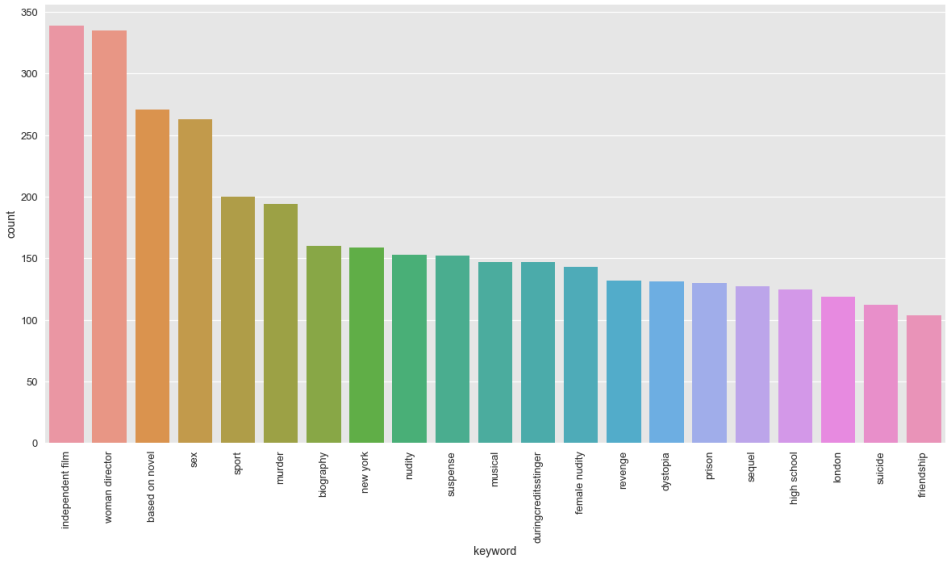
1. split the company from a combined value to single value
2. extract the companies which have over 50 movies
3. calculate the revenue for each c



Warner Bros, Universal Pictures, Paramount Pictures and Twentieth Century Fox Film Coportion have highest revenue with Warner Bros highest

Research Question 3 (What keywords are the most popular)

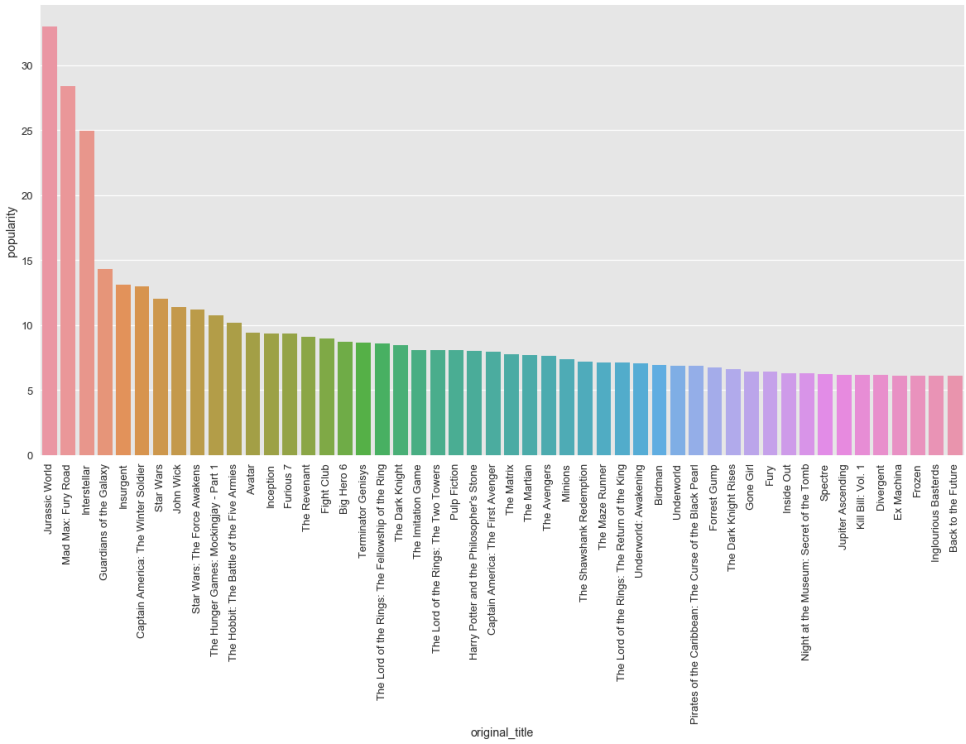
1. split the keywords from a combined value to single keyword value
2. extract keywords which have over 100 frequency



'independent film', 'woman director', 'based on novel' and 'sex' are the most popular key words

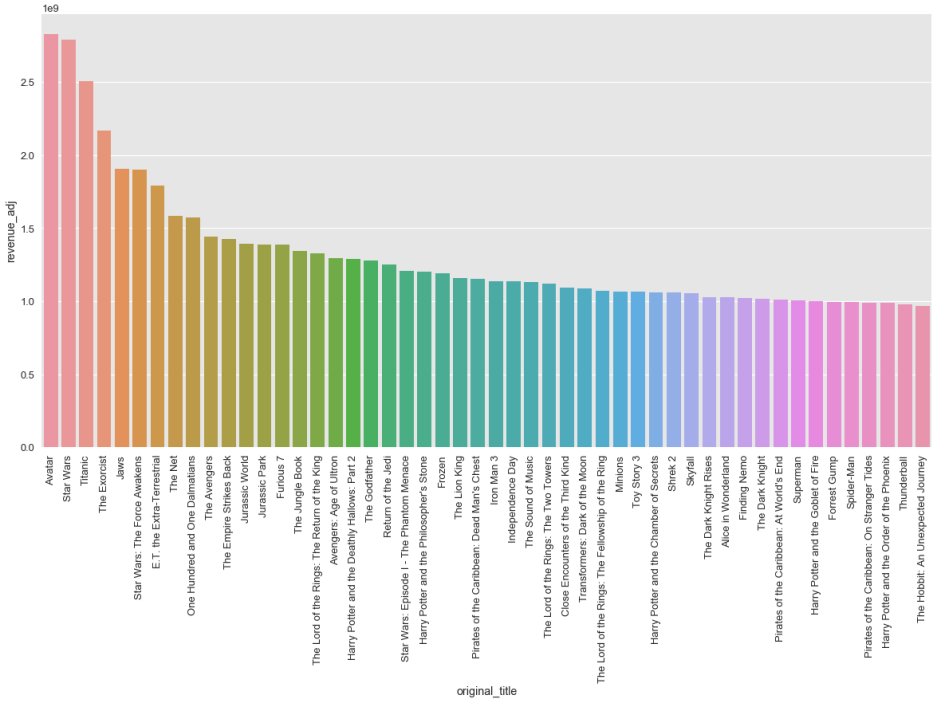
Research Question 4 (Which movies are the most popular or profitable)

1. calculate movie popularity



'Jurassic World', 'Mad Max: Fury Road', 'Interstellar' are way more popular movies than others

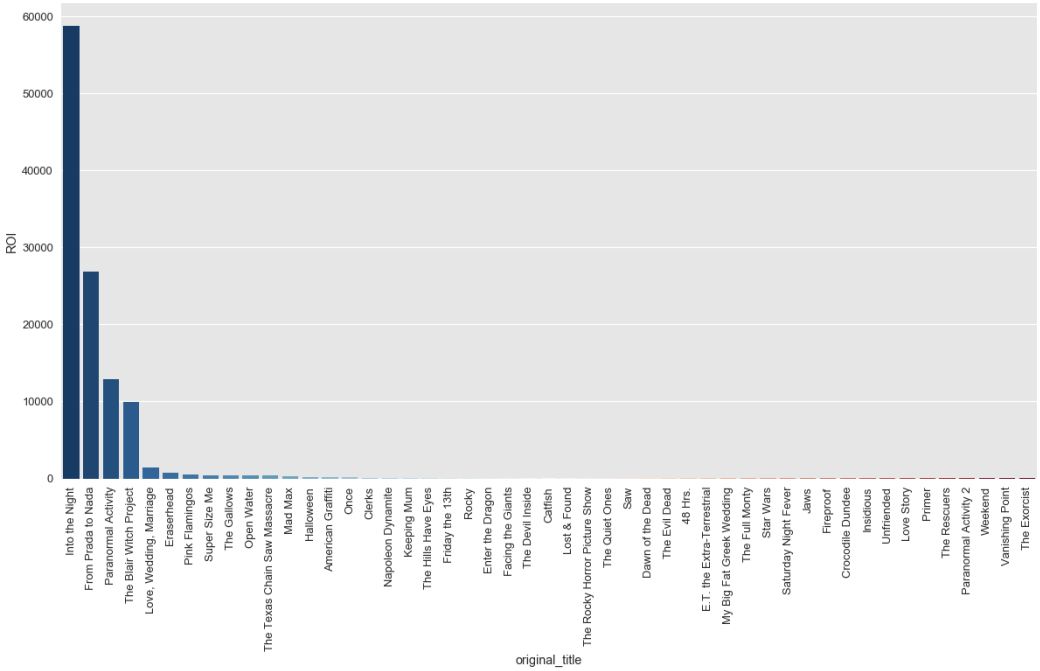
2. calculate movie revenue



'Avatar', 'Star Wars' and 'Titanic' produce the most revenue

3. calculate ROI for each movie

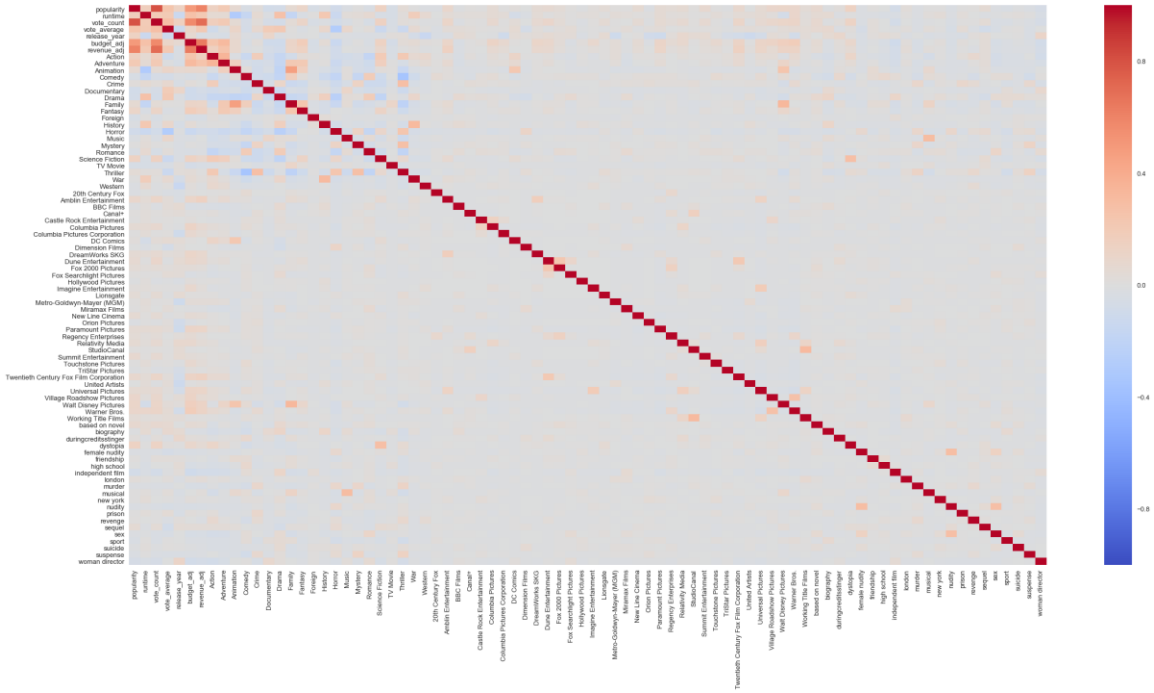
'The Karate Kid, Part II' has so much higher ROI than others which is over 1.1 million



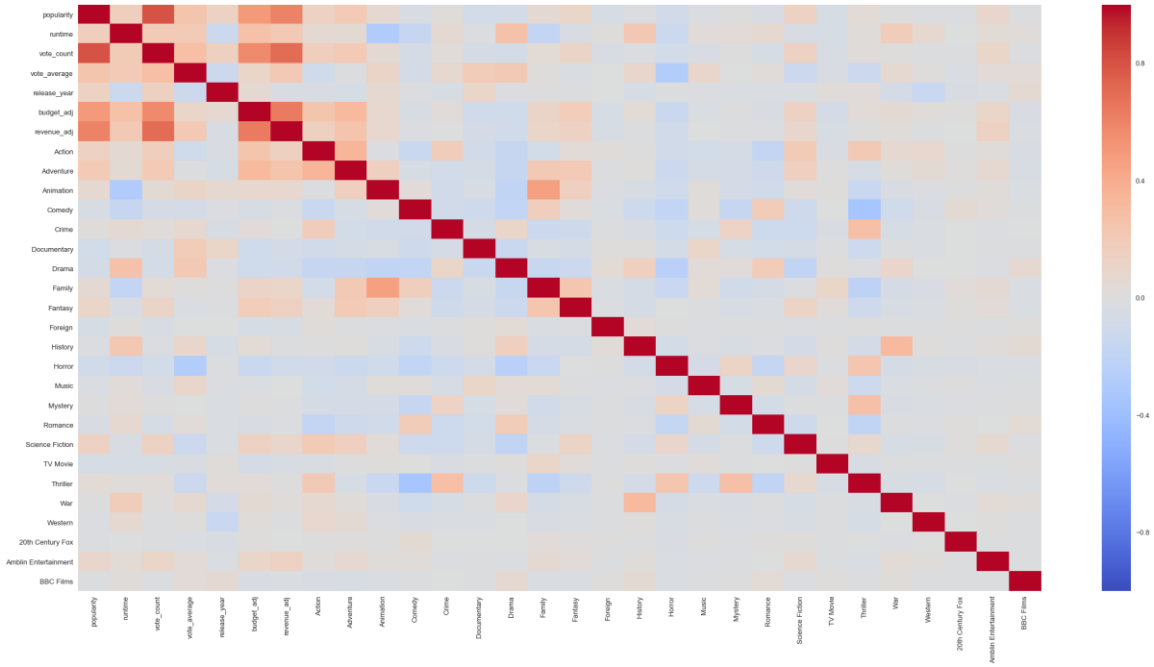
For the rest of movies except 'The Karate Kid, Part II', 'Into the Night', 'From Prada to Nada', 'Paranormal Activity','The Blair Witch Project' have pretty high ROI compared with others.

Research Question 5 (The correlation between features)

1. combine single genre, single company with over 50 frequency and single keyword with over 100 frequency data frames into original data



It is hard to see what is going on, but it seems correlations concentrate on the left corner, so plot the first 30 features



from the correlation plot,

for popularity, vote_count, budget_adj, revenue_adj have positive correlation with it.

for runtime, genre Animation, Comedy and family seem to have negative correlation with it.

for vote_count, revenue_adj and budget_adj have relatively high positive correlation.

for vote_average, genre Horror has negative correlation.

for budget_adj, revenue_adj has high correlation with it.

Conclusions:

- Drame, Comedy, Thriller and Action are four most-made genres. The number of drama is the most.
- Drama is the most popular genre, following by action, comedy and thriller.
- Action, Adventure, Drama, Comedy and Thriller are the most profitable genres.
- Documentary and Music have highest ROI with all others having fairly close ROI
- Warner Bros, Universal Pictures, Paramount Pictures and Twentieth Century Fox Film Coportion have highest revenue with Warner Bros highest
- 'independent film', 'woman director', 'based on novel' and 'sex' are the most popular key words
- 'Jurassic World', 'Mad Max: Fury Road', 'Interstellar' are the most popular movies and way more than others
- 'Avatar', 'Star Wars' and 'Titanic' produce the most revenue
- 'The Karate Kid, Part II' has so much higher ROI than others which is over 1.1 million. For the rest of movies except 'The Karate Kid, Part II', 'Into the Night', 'From Prada to Nada', 'Paranormal Activity','The Blair Witch Project' have pretty high ROI compared with others.

- from the correlation plot,

1. for popularity, vote_count, budget_adj, revenue_adj have positive correlation with it
2. for runtime, genre Animation, Comedy and family seem to have negative correlation with it
3. for vote_count, revenue_adj and budget_adj have relatively high positive correlation
4. for vote_average, genre Horror has negative correlation
5. for budget_adj, revenue_adj has high correlation with it