Capstone 2 Project MileStone

MOVIE PREDICT AND RECOMMENDATION

1. **Statement and problem**

Movies have been around us for long. First they come out with black and white movies, for many years they can make a color and sound effect like today. For today, we have so many film industries that want to bring their idea, and their time to help us entertainment but some movies have revenue less than others or review rating was less than they are expecting. So to help the company come up with ideas and plan for making a good movie. In this project I want to build a machine that can help predict the movie's success or fail with our data we have. And there are many people who are looking for movies to watch everyday, so it's hard for people to look for something they don't know, so I build the recommendation machine, so it can recomandata the movie to the same type as the user who watched it.

2. **Collection data**

The data was in the kaggle compositer, but data just has to the day it was posted as author collected from the API at the time he/she worked on it. So I want to have more data up to today, so I just generated a metadata movie for myself. First I got the last movie ID on the website of TMDb,The Movie Database (TMDb) is a community built movie and TV database. Every piece of data has been added by our amazing community dating back to 2008 , I generate over one by one movie ID. It may take a bit of time, but that can get me an idea how I can pull data from API.

3. **Cleaning and transform data**

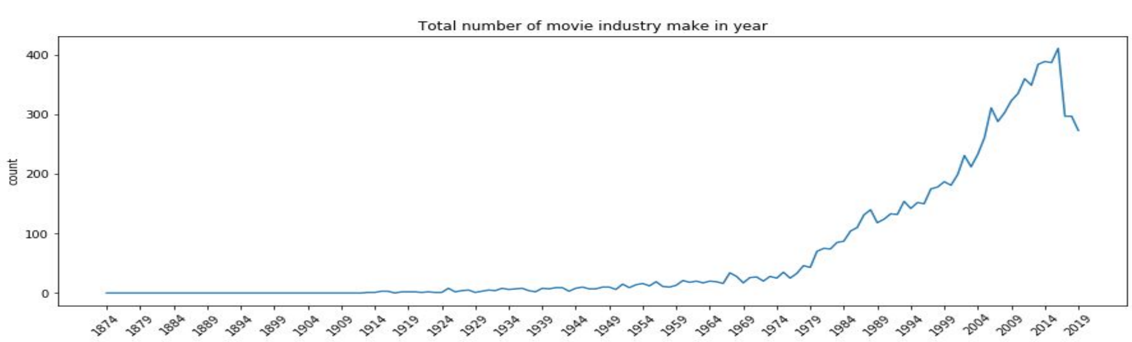
The data format was not usable yet, so we need to apply some methods to make usable data.

First, we needed to drop some columns that were not useful for us.

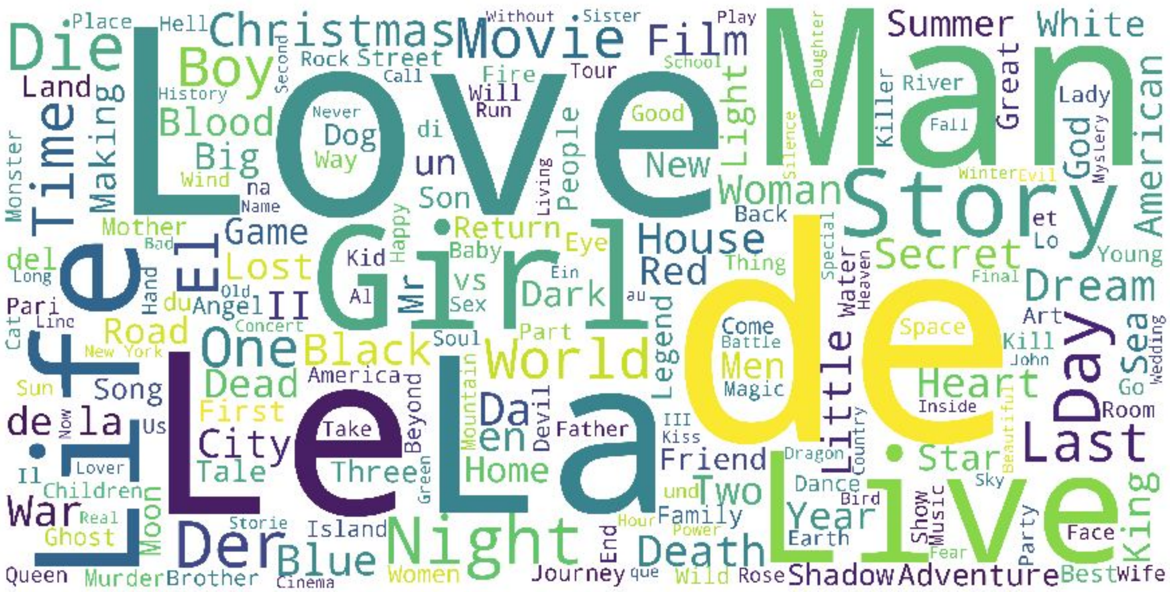
Second, you will see the data after the request from the API gives you a string of dictionaries. The information has the ID and name for those specified columns, but we just need the name only. So I just convert the string of dictionaries to dictionaries and get the list name from the key name. You can find the code I did in this github.

4. **Data Visualization**

a. Total the number of movies was increasing by year, as the graph shows us that from 1874 is the first movie was made, and they are likely to make the same number of movies each year to 1924, from 1924 to 1979 have more movies made, the line began increasing but not much. And after 1979 the number of movies was increasing a lot. As data we have, let's look at top 10 year revenue. This is just an idea to overlook the industry, because our data had a lot of missing revenue data.



b. I used the world cloud to generate the word chart to show the title movie, you can see the chart was show word LOVE, MAN, Live and Life appear so often.



c. So we know the number of movie was make each year was increasing, so the revenue are increasing as well, so you can see the chart below that Total_revenue of each year was increasing to, these number just get us know idea that industry film was growth and growth, but it not actual tell that how much was each year was make, cause we have a lot missing value and we that we don't have information.

year	Average_revenue	Total_revenue	count
2016	7.627827e+07	3.135037e+10	411
2015	7.414786e+07	2.869522e+10	387
2014	7.113407e+07	2.767116e+10	389
2013	7.039734e+07	2.703258e+10	384
2012	7.569661e+07	2.641812e+10	349
2011	6.939981e+07	2.498393e+10	360
2010	7.205477e+07	2.413835e+10	335
2009	7.440009e+07	2.403123e+10	323
2008	6.991971e+07	2.118567e+10	303
2007	7.076921e+07	2.038153e+10	288

Movie was made in 1974 and you can see it is only 1 minute long. It made sense because back there technology still new and they recorded by film so it was cost a lot movie to make a long movie was today.

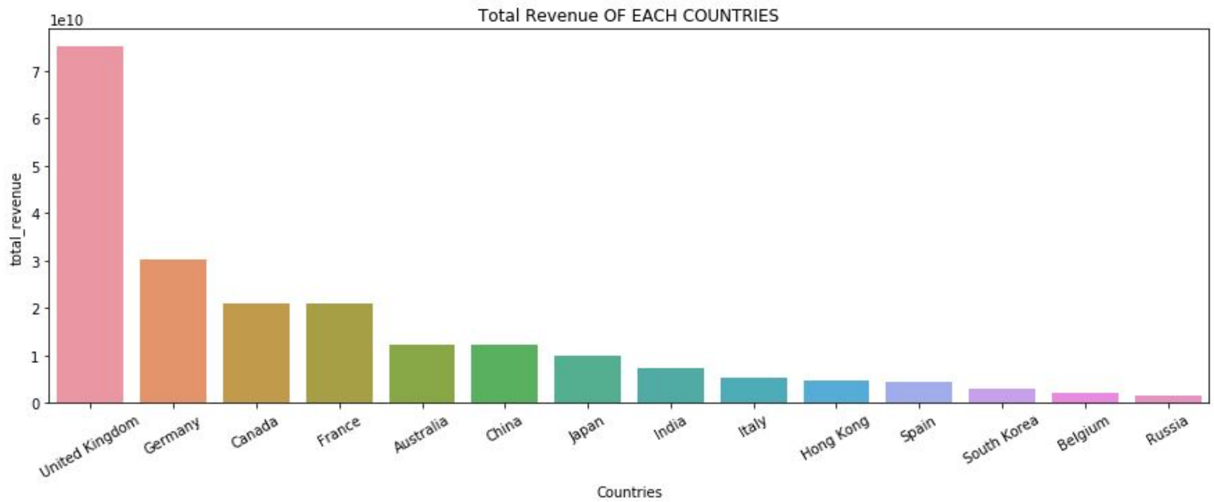
	year	title	runtime
34693	1874	Passage of Venus	1.0
34690	1878	Sallie Gardner at a Gallop	1.0
70946	1881	Athlete Swinging a Pick	1.0
41270	1883	Buffalo Running	1.0
135279	1885	L'homme machine	1.0

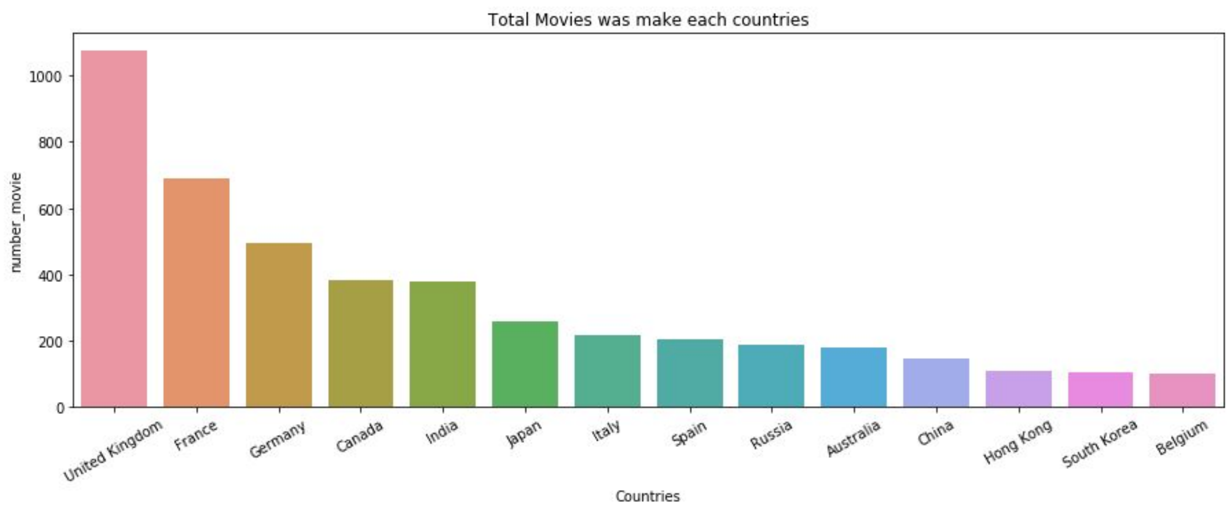
Recent movie had a time run longer, as you know standard most movie was runaround 90 to 120 min long.

	year	title	runtime
174446	2019	Sunday	13.0
174441	2019	Queen + Béjart - Ballet For Life	58.0
126731	2019	The Ocean Washed Open Your Grave	3.0
174467	2019	Entropia	28.0
210551	2019	Jorge	20.0

d. Next, let us divide countries and see where they make more and have best success in the film industry on a data set. The chart below, you can see that the USA was the top 1 make movie, then after that was the UK and France, and Germany.

	average_revenue	total_revenue	number_movie
Countries			
United States of America	7.714978e+07	4.976932e+11	6451
United Kingdom	6.990096e+07	7.514353e+10	1075
France	3.021624e+07	2.084921e+10	690
Germany	6.078097e+07	3.014736e+10	496
Canada	5.485054e+07	2.089806e+10	381
India	1.911960e+07	7.246328e+09	379
Japan	3.894279e+07	9.969354e+09	256
Italy	2.382187e+07	5.169346e+09	217
Spain	2.157035e+07	4.378782e+09	203
Russia	8.155989e+06	1.517014e+09	186

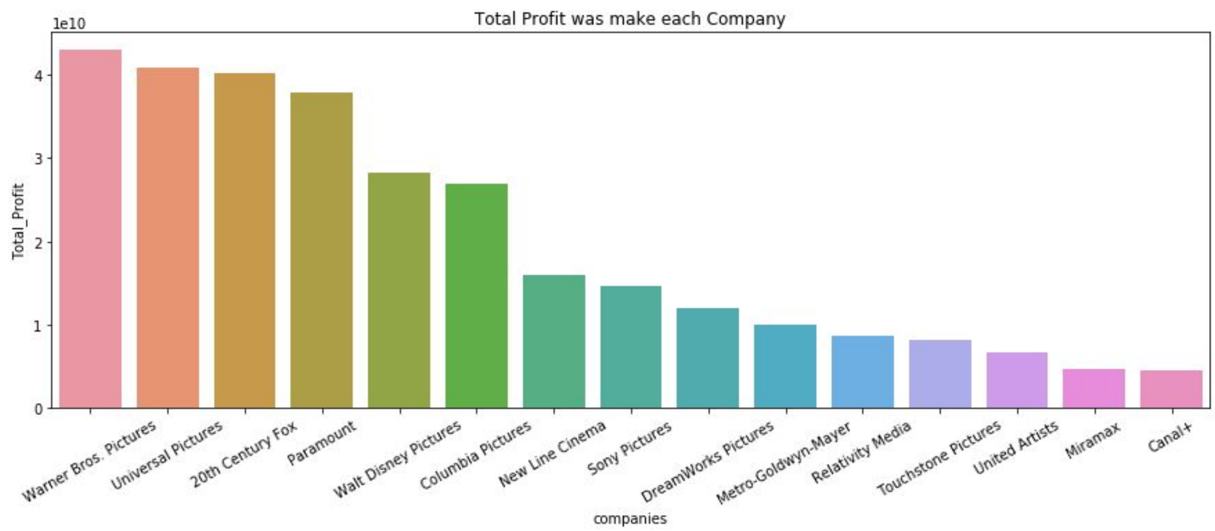


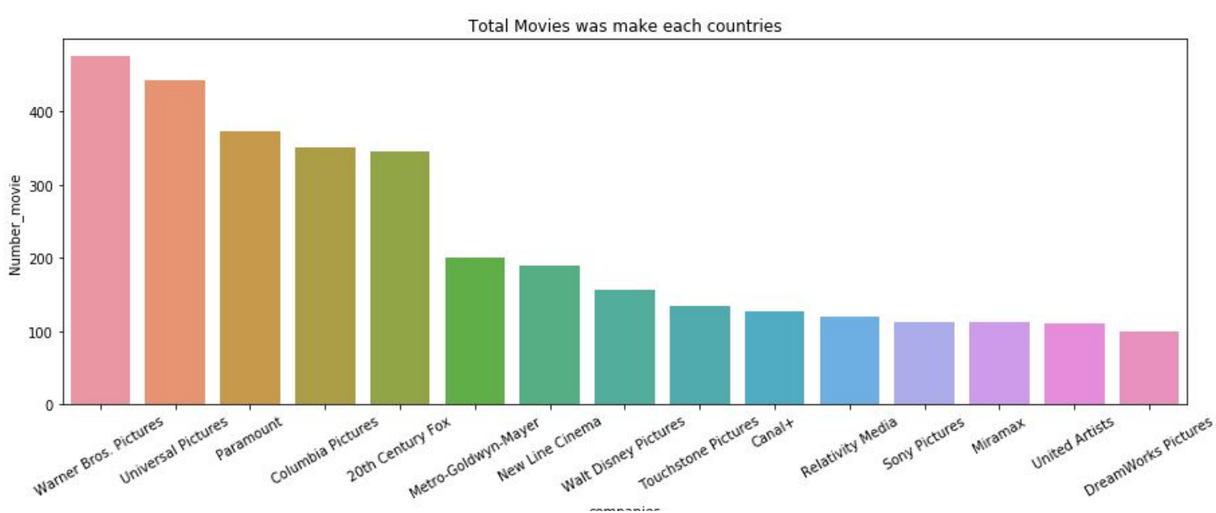


France is rank 3 product movie, but their revenue was at rank 5. China is kind of the opposite, they were rank 12 of countries' product movies, but their revenue was rank 7.

e. Now we will get more detail about each company and which was best, so the chart below that you can see Warners Bros. Pictures, Universal Pictures, Paramount, Columbia Pictures, and 20th Century Fox are top 5 companies that make most profit and number movie they are make overall.

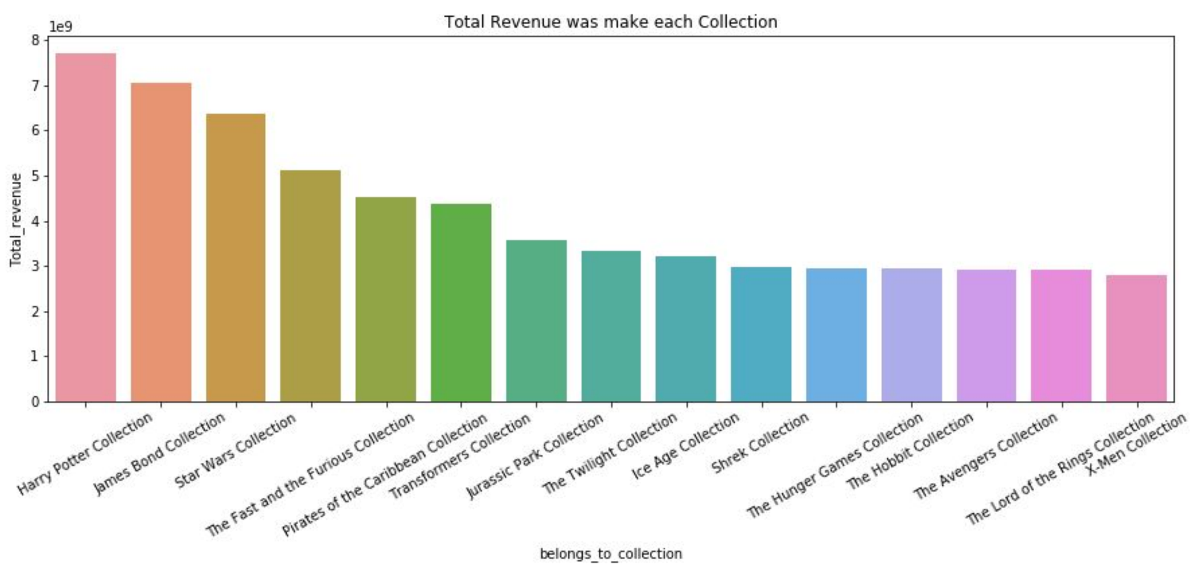
companies	Avergae_profit	Total_Profit	Number_movie
Warner Bros. Pictures	9.032809e+07	4.299617e+10	476
Universal Pictures	9.229253e+07	4.088559e+10	443
Paramount	1.014296e+08	3.793466e+10	374
Columbia Pictures	7.684656e+07	2.697314e+10	351
20th Century Fox	1.164543e+08	4.017672e+10	345
Metro-Goldwyn-Mayer	5.015126e+07	1.003025e+10	200
New Line Cinema	8.415754e+07	1.598993e+10	190
Walt Disney Pictures	1.797291e+08	2.821747e+10	157
Touchstone Pictures	6.081461e+07	8.209972e+09	135
Canal+	3.521476e+07	4.437060e+09	126



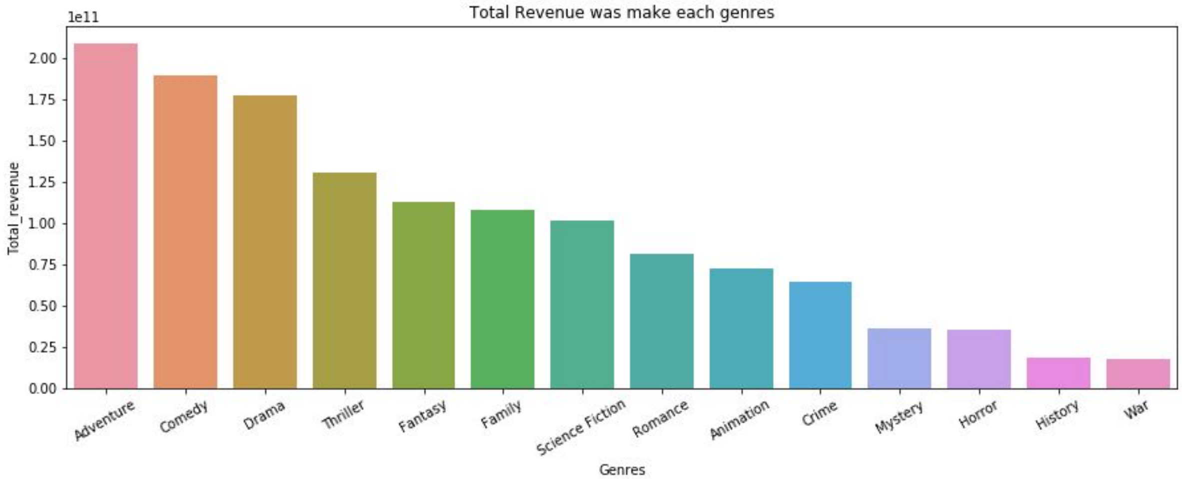


f. Some movies are so good, so they keep making a new series after another. So you can see below that the top 1 have revenue is the Harry Potter collection. And Jame Bond was number 2, after that Star War, the Fast and Furious. These movies you can see are so popular, so it made sense that it was in top 4 revenue of all the collection movies.

belongs_to_collection	Average_revenue	Total_revenue	Total_movie
Harry Potter Collection	9.633598e+08	7.706879e+09	8
James Bond Collection	2.827486e+08	7.068715e+09	25
Star Wars Collection	9.112054e+08	6.378438e+09	7
The Fast and the Furious Collection	6.406373e+08	5.125099e+09	8
Pirates of the Caribbean Collection	9.043154e+08	4.521577e+09	5
Transformers Collection	8.758574e+08	4.379287e+09	5
Jurassic Park Collection	8.948083e+08	3.579233e+09	4
The Twilight Collection	6.686215e+08	3.343107e+09	5
Ice Age Collection	6.433533e+08	3.216767e+09	5
Shrek Collection	7.411823e+08	2.964729e+09	4

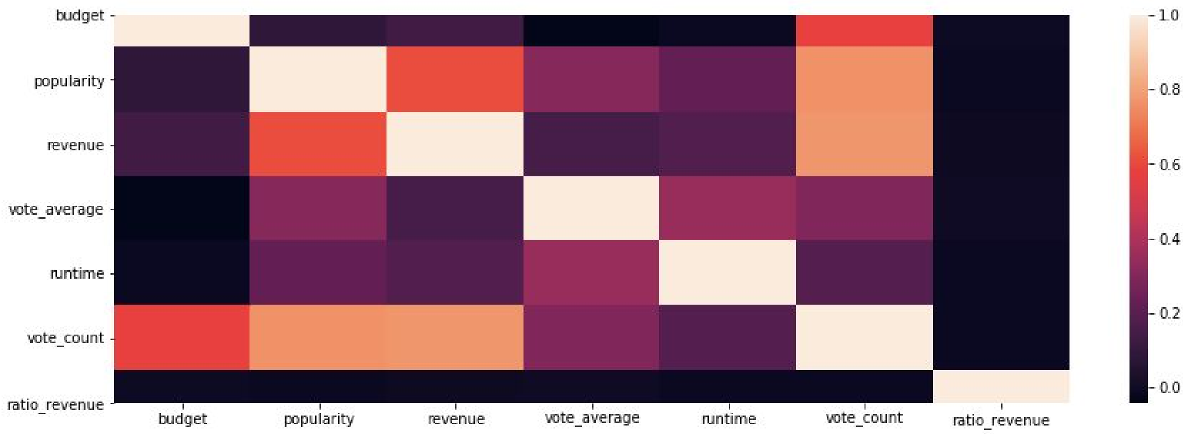


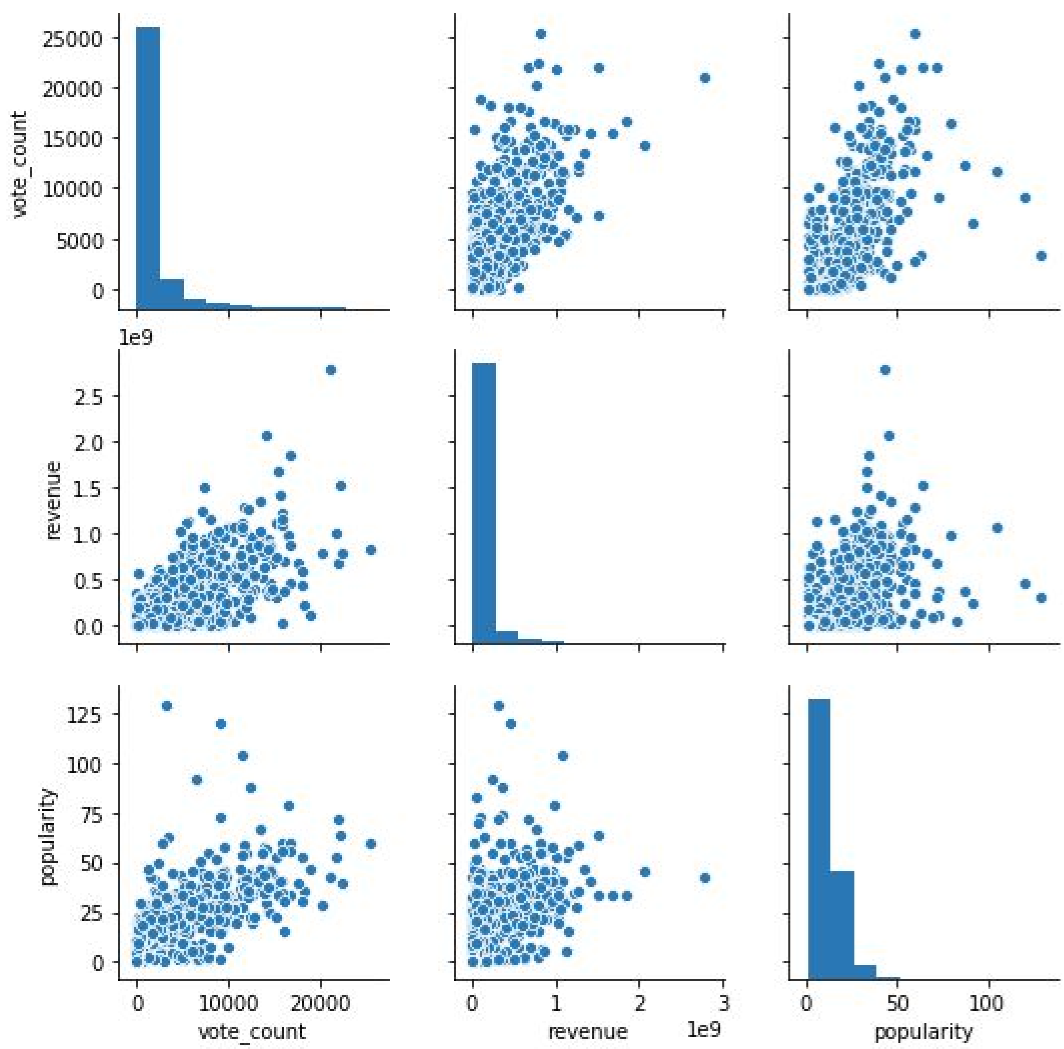
g. Genres of action, adventures, and comedy are the top 3 movies that have most revenue. It can be said people are more likely interested in these kinds of movies.



Genres	Avergae_revenue	Total_revenue	Total_movie
Action	1.025900e+08	2.134897e+11	2081
Adventure	1.640720e+08	2.086996e+11	1272
Comedy	5.798689e+07	1.889213e+11	3258
Drama	3.835340e+07	1.773078e+11	4623
Thriller	6.120269e+07	1.302393e+11	2128
Fantasy	1.463286e+08	1.122341e+11	767
Family	1.299991e+08	1.080293e+11	831
Science Fiction	1.181741e+08	1.016297e+11	860
Romance	4.568787e+07	8.073046e+10	1767
Animation	1.467283e+08	7.189686e+10	490

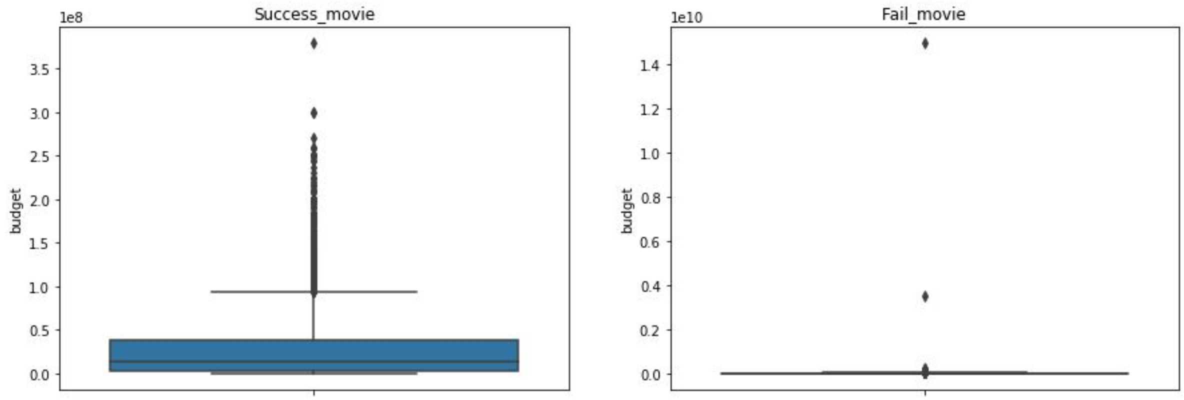
h. Now, we can see the relationship between these numeric columns. As the heatmap plot below show us that they are not have much relation, but some of them are have most is around 0.6 or 0.7 correlation.



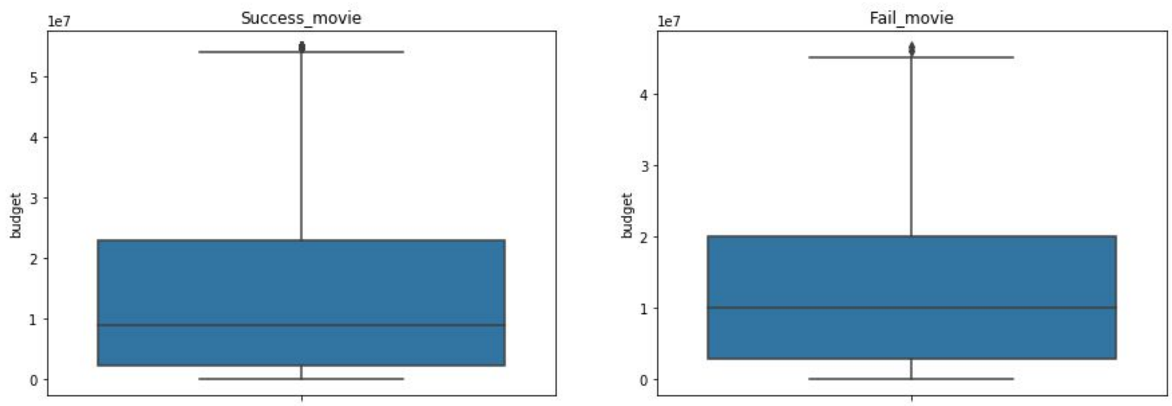


i. Inference Statistic
   - We will take a looking what different the average budget of success movie and fail movie, but data is have so outlier. That we need to remove it.

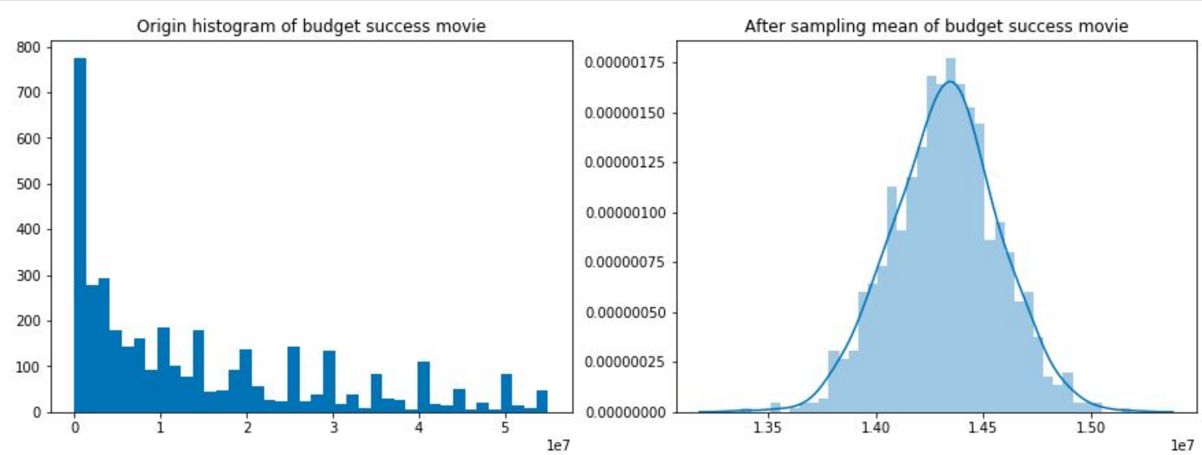
+ Plot box plot before we remove the outlier, so can notice that the failed movie has a largest gap of outliers. It is bad for our conclusion.

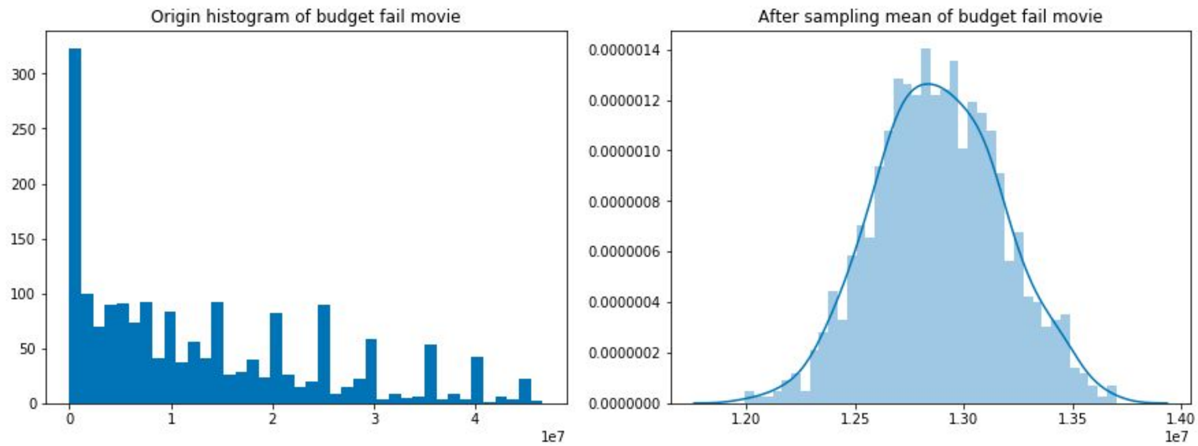


+ These plot below is when we remove the outlier, box plot show is more better

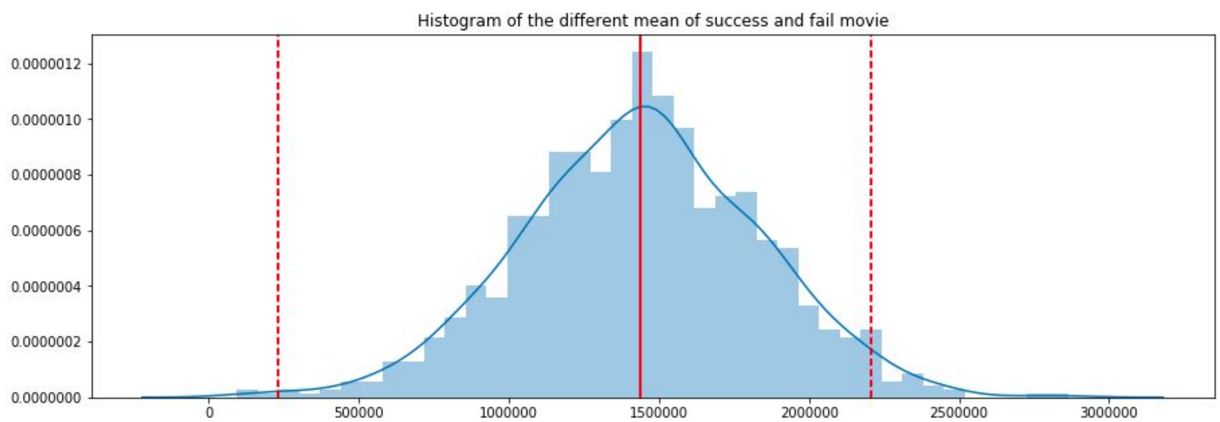


- We used the bootstrap resampling to tell that mean of success is 14M and fail movie 13M





- We perform the hypothesis test to test the different mean average of success and fail movies.
    - We identified the null hypothesis and alternative hypothesis.
        - H0: the different average of success and fail is >= 1438763
        - Ha: the different average is < 1438763
    - With significant is 5% or 0.05
    - After our test and we got the p-value = 0.501
    - Then we can’t reject our null hypothesis.



5. **Prediction machine model**

    a. Preparation the data

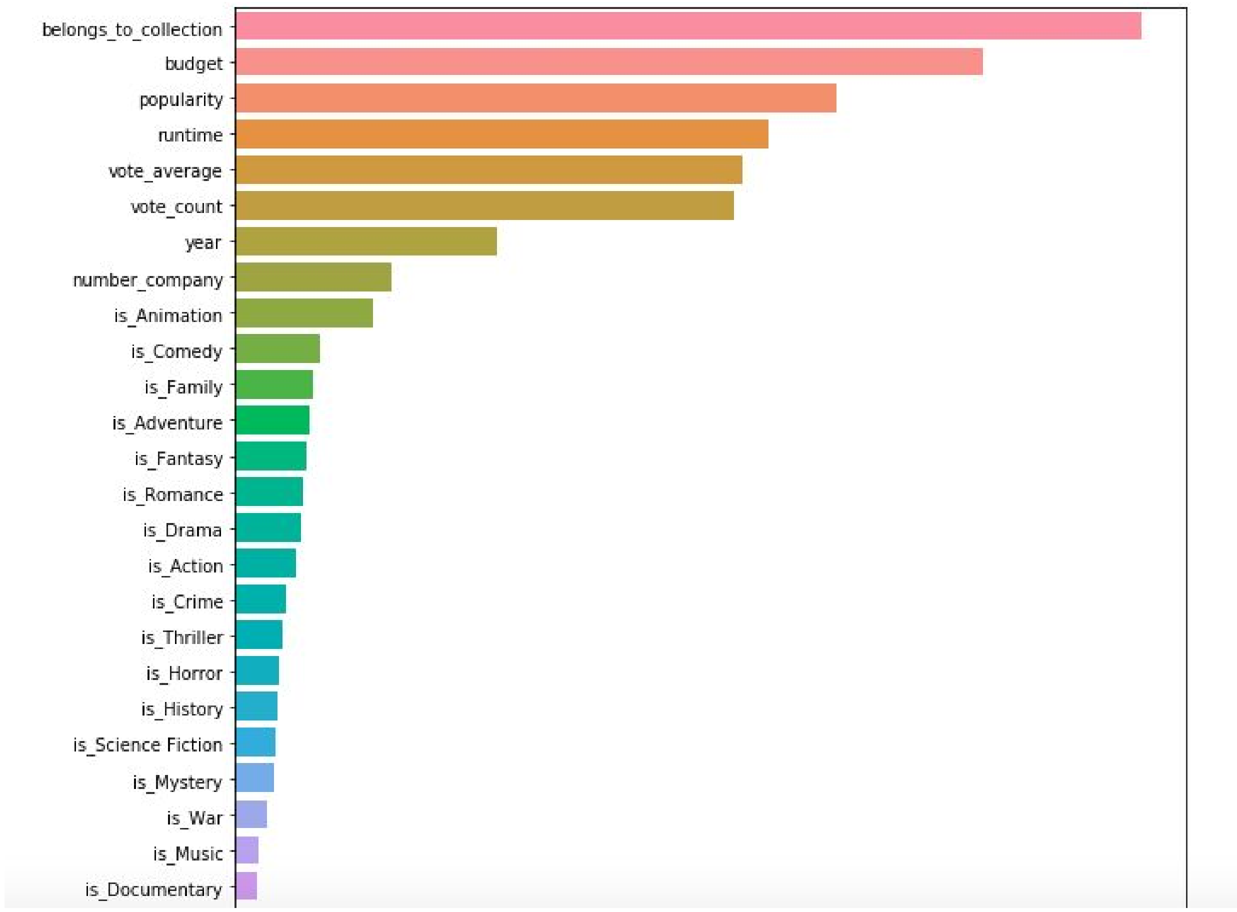
    Our data have category and numeric data, so first we go through our train model. We need to transform these category columns to numeric. Also these are new feature for our model

    b. Random Forest Classification

    After training our data with random forest classification, we have our model score is 0.78. As we have a classification report show bellow.

	precision	recall	f1-score	support
0	0.67	0.54	0.60	785
1	0.82	0.89	0.85	1860
accuracy			0.78	2645
macro avg	0.75	0.71	0.73	2645
weighted avg	0.78	0.78	0.78	2645

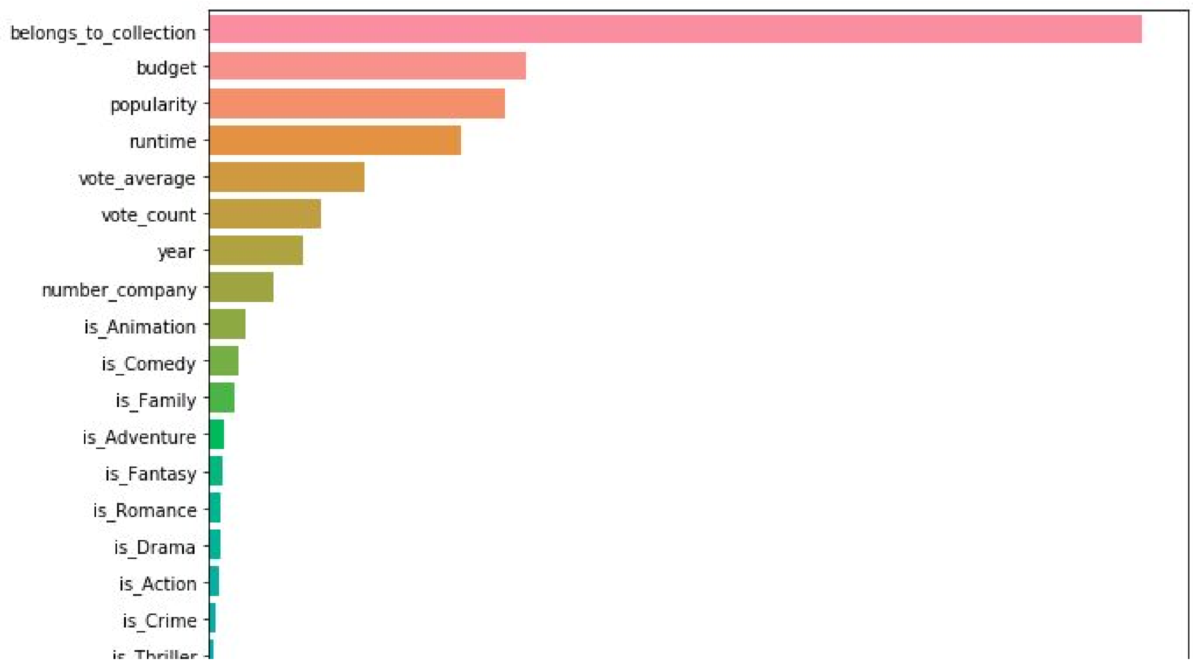
You can see the importance of our model as shown below. It shows that belong_to_collect and budget have the most score when we are predicted success or fail of movie



c. Gradient Boosting Classification
Let see how our Gradient Boosting models work with our data set. It showed the same score as random forest is 0.78. The classification report shows that these 2 models are the same accuracy, precision and recall.

	precision	recall	f1-score	support
0	0.67	0.54	0.60	785
1	0.82	0.88	0.85	1860
accuracy			0.78	2645
macro avg	0.74	0.71	0.73	2645
weighted avg	0.78	0.78	0.78	2645

But feature importance is a little different, that this model uses less features than random forest. It can point out that most features are belong_to_collection. So we can reduce our features, it can make our model train faster.



d. Logistic regression Model
   Let's use the basic classification model, logistic regression. This model seems to perform worse than these 2 above. So we are not using this model.

	precision	recall	f1-score	support
0	1.00	0.00	0.01	785
1	0.70	1.00	0.83	1860
accuracy			0.70	2645
macro avg	0.85	0.50	0.42	2645
weighted avg	0.79	0.70	0.58	2645

e. Conclusion

As the result above, we can choose a random forest or gradient boosting model for our prediction model.

6. **Recommendation System**

   a. Basic Recommendation

We will recommend movies to users for the first time. We can recommend top 20 movies with higher ratings to users or users can choose which genre they want to watch and our system will display 15 movies with higher ratings for that genre.

Show top 20 movie:

```
recommend_top_20_movie()
```

	title	genres
0	In a Lonely Place (1950)	Drama|Film-Noir|Mystery|Romance
1	Paperman (2012)	Animation|Comedy|Romance
2	Diabolique (Les diaboliques) (1955)	Horror|Mystery|Thriller
3	Paradise Now (2005)	Crime|Drama|Thriller|War
4	Best Years of Our Lives, The (1946)	Drama|War
5	Drunken Master (Jui kuen) (1978)	Action|Comedy
6	Inherit the Wind (1960)	Drama
7	Tell No One (Ne le dis à personne) (2006)	Crime|Drama|Mystery|Thriller
8	For the Birds (2000)	Animation|Children|Comedy
9	Wind Rises, The (Kaze tachinu) (2013)	Animation|Drama|Romance
10	Court Jester, The (1956)	Adventure|Comedy|Musical
11	Godfather, The (1972)	Crime|Drama
12	Shawshank Redemption, The (1994)	Crime|Drama
13	Tom Jones (1963)	Adventure|Comedy|Romance
14	Gladiator (1992)	Action|Drama
15	On the Waterfront (1954)	Crime|Drama
16	Kid, The (1921)	Comedy|Drama
17	When We Were Kings (1996)	Documentary
18	Carnal Knowledge (1971)	Comedy|Drama

Show top 15 movie have choose by Genres:

```python
get_recommend_genre("Drama")
```

	title	genres
0	Best Years of Our Lives, The (1946)	Drama|War
1	Inherit the Wind (1960)	Drama
2	Godfather, The (1972)	Crime|Drama
3	Shawshank Redemption, The (1994)	Crime|Drama
4	Gladiator (1992)	Action|Drama
5	On the Waterfront (1954)	Crime|Drama
6	All About Eve (1950)	Drama
7	Ran (1985)	Drama|War
8	Mister Roberts (1955)	Comedy|Drama|War
9	Godfather: Part II, The (1974)	Crime|Drama
10	Paths of Glory (1957)	Drama|War
11	Lifeboat (1944)	Drama|War
12	Rush (2013)	Action|Drama
13	Modern Times (1936)	Comedy|Drama|Romance
14	Philadelphia Story, The (1940)	Comedy|Drama|Romance

b. Correlation-base recommendation

Use Pearson's r correlation to recommend a movie that is most similar to the movie that user have early watch

Based on all of the users rating the movie, we can calculate the correlation of movies.

Users can search movies by name, or the system will recommend the next movie based on the user have been watched.

```
get_recommendation_movie_corr('Dangerous Minds (1995)')
```

```
/Users/hungnguyen/miniconda3/lib/python3.7/site-packages/numpy/li
freedom <= 0 for slice
  c = cov(x, y, rowvar)
/Users/hungnguyen/miniconda3/lib/python3.7/site-packages/numpy/li
ero encountered in true_divide
  c *= np.true_divide(1, fact)
```

	title	genres
0	...And Justice for All (1979)	Drama|Thriller
1	127 Hours (2010)	Adventure|Drama|Thriller
2	2 Days in the Valley (1996)	Crime|Film-Noir
3	2 Fast 2 Furious (Fast and the Furious 2, The)...	Action|Crime|Thriller
4	21 (2008)	Crime|Drama|Romance|Thriller
5	300 (2007)	Action|Fantasy|War|IMAX
6	40 Days and 40 Nights (2002)	Comedy|Romance
7	About Last Night... (1986)	Comedy|Drama|Romance
8	Absolute Power (1997)	Mystery|Thriller
9	Addams Family, The (1991)	Children|Comedy|Fantasy
10	Adventureland (2009)	Comedy|Drama
11	Adventures in Babysitting (1987)	Adventure|Comedy
12	Adventures of Baron Munchausen, The (1988)	Adventure|Comedy|Fantasy
13	Aeon Flux (2005)	Action|Sci-Fi

These list movies strongly correlate with the Dangerous Minds. So people like Dangerous Minds may like to watch these movies as well.

c. Model-based Collaborative filtering system recommendation
We recommend movies based on our model. So we don't have a recall back to our data set. We do the same as the correlation recommended above, but this time we use Singular Value Decomposition (SVD) to create the model, and then we can recommend movies based on that model. It will be faster and more accurate.

```python
model_base_recommendation('Dangerous Minds (1995)')
```

	title	genres
0	Dangerous Minds (1995)	Drama
1	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
2	Waterworld (1995)	Action|Adventure|Sci-Fi
3	Legends of the Fall (1994)	Drama|Romance|War|Western
4	Tombstone (1993)	Action|Drama|Western
5	Dances with Wolves (1990)	Adventure|Drama|Western
6	Apollo 13 (1995)	Adventure|Drama|IMAX
7	Cliffhanger (1993)	Action|Adventure|Thriller
8	Ace Ventura: Pet Detective (1994)	Comedy
9	Firm, The (1993)	Drama|Thriller
10	Speed (1994)	Action|Romance|Thriller
11	True Lies (1994)	Action|Adventure|Comedy|Romance|Thriller
12	Die Hard: With a Vengeance (1995)	Action|Crime|Thriller
13	Crimson Tide (1995)	Drama|Thriller|War
14	Net, The (1995)	Action|Crime|Thriller