ANLY 501 Project 3 Writeup

Ju Huang, Tianyi Yang, Xinyue Liu, Zikai Zhu

I. Network Analysis of Collaboration Between Movie Studios

(EXTRA CREDIT)

Data Science Questions:

How is the collaboration between studios from 2010-2017? Specifically:

1. What is the network of studio collaboration each year?
2. How has the collaboration changed over years from 2010 to 2017?
3. Do studios with different size (defined by the level of their box office) collaborate with each other? Or do studios only collaborate with similar size?

Data:

1. **Nodes**: A studio is a node. We collected the studio data from OMDBapi.com and the specific column is 'Production'. In total, the network has 79 nodes (i.e. 79 studios have at least one collaboration with another studio in the past 7 years).
2. **Edges:** We collected the studio collaboration data from OMDBapi.com. For every movie, OMDB lists all studios that participate in the movie, and we deem studios listed under one same movie as one instance of collaboration.
3. **Attribute of Nodes**: We code the studios by their amount of box office (in 10 to the power of 5,6...11 dollars) to answer the third question. This boxoffice data is collected from BoxOfficeMojo.com and the specific column is 'Total Gross'.

Method:

We used the Networkx package in Python and we create a simple, undirected network for each year and for 7 year as a whole. For the first question, we calculate local network metrics for each node: betweenness, degree, and clustering coefficient. To answer the second question, we compute the global metric for network for each year: density, triangles, and averages for the centrality metrics. To answer the the third question, we code the studio by their level of box office, and plot the network of all studio collaboration.

Findings:

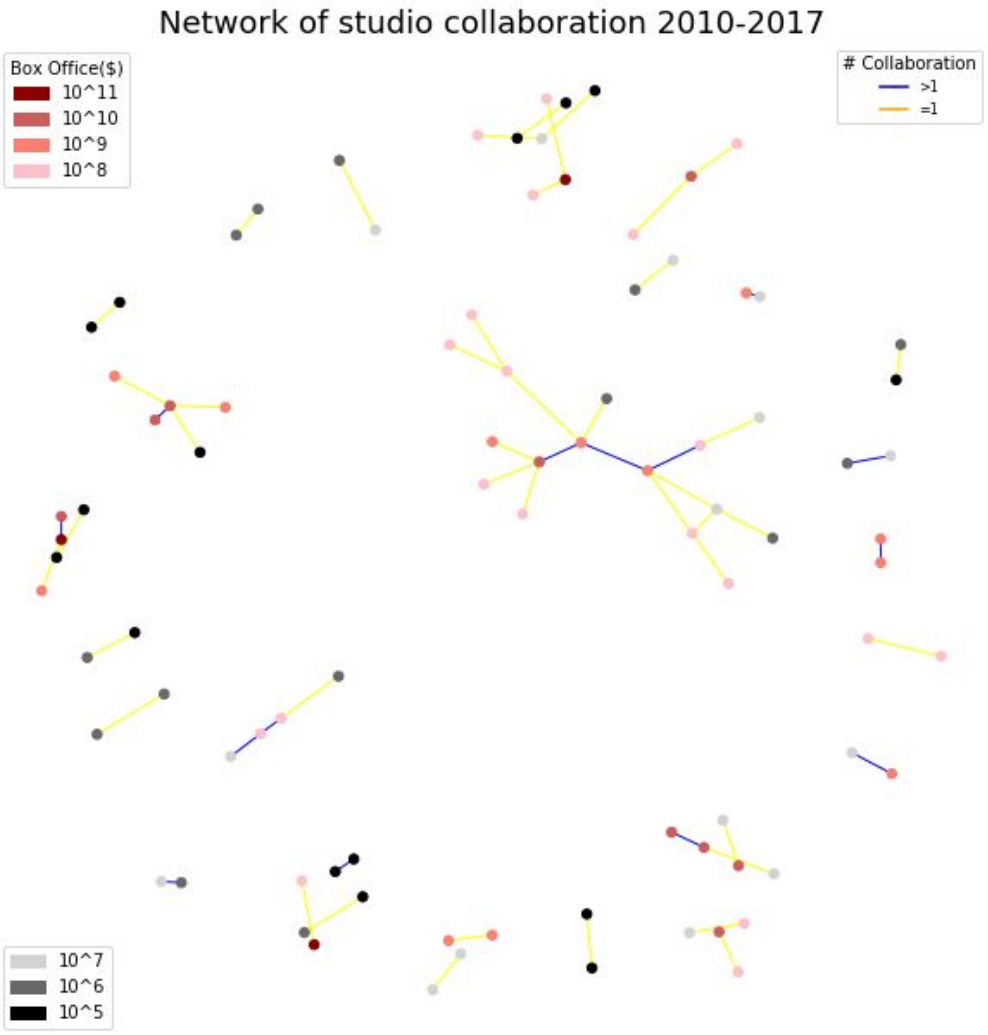
1. There are **very few collaboration** and the network is very **disconnected**

First of all, there are very few nodes each year (as we can see from the printed output with the code). On average, there are only 16 nodes every year. Thus, we can see that very few studios collaborate every year. Second, each year, there are many clusters but very few nodes in each cluster. For example, in 2010, there are 22 nodes and 8 clusters, each cluster having only an average of 2-3 nodes. Thus, there's no one connected network between studios. The network is rather consist of many small clusters (i.e. small-scale collaboration).

2. Across the years, there's **no trend in increasing collaboration**

Looking at the table across 7 years (the printed output ‘Network of studios 2010-2017’ from the code), we can see that there are fewer and fewer edges each year from 2010 to 2016. However, in 2017, there are all of a sudden an increase both in nodes and edges.

3. Studios **generally collaborate with studios of the same size**, but not studios that are of different sizes (in terms of box office)



Looking at the visualization of the network across 2010 to 2017 (the printed output ‘Network for studio collaboration 2010-2017’), we can see that in general, warmer-colored nodes collaborate with warmer color nodes, and cooler-colored nodes collaborate cooler-colored nodes (where the warmer the color of the node, the higher the box office). This phenomena is especially true for blue edges (# collaboration > 1). Thus, it is not hard to see that studios with higher box office collaborate with their large-box-office counterparts.

Limitations:

The only information we have about collaboration is from OMDB (IMDB) and we treat that as complete information. This is second hand information, but first hand, so it might not be 100% accurate.

II. Topic Modeling of Movie Plots

Data Science Question:

1. What are the main topics of movie plots?
2. What is the distribution of topics for each plot?
3. What movie plot is similar to other movie plot based on the topic?
4. Which topics are the most popular ones? What about the trend?

Data:

Plots of all movies produced from 2010-2017.
There are 751025 words in the combination of all plots.

Method:

For the first question, we use the gensim package in Python to generate 8 main topics for movie plots. For the second question, we create a table to store the distribution of topics and record the most relevant topic for each plot. For the third question, we look into the table to find out the similar plot based on the topic. For the fourth question, we count for the frequency for each topic to figure out the most popular ones in total and in each year.

Findings:

1. **The topics , frequency ranking and the representative movies**

   Topic: 0 An Adventure Game TOP 6
   The Last Airbender
   The A-Team
   Words: 0.016*"world" + 0.012*"american" + 0.011*"real" + 0.010*"group" + 0.009*"team" + 0.009*"game" + 0.009*"young" + 0.008*"fight" + 0.008*"woman" + 0.008*"citi"

   Topic: 1 Families and Friends TOP 1
   Grown Ups
   Spy Kids All the Time in the World
   Words: 0.019*"friend" + 0.012*"father" + 0.012*"famili" + 0.010*"face" + 0.009*"town" + 0.009*"mother" + 0.008*"love" + 0.008*"forc" + 0.008*"citi" + 0.008*"student"

   Topic: 2 Women's Struggling Lives TOP 4
   Black Swan

The Back-Up Plan
Words: 0.021*"woman" + 0.013*"struggl" + 0.012*"group" + 0.012*"young" + 0.010*"becom" + 0.010*"school" + 0.008*"leav" + 0.008*"high" + 0.007*"friend" + 0.007*"lead"

Topic: 3 Young People's Family Lives and Fights TOP 3
How to Train Your Dragon
The Book of Eli
Words: 0.015*"girl" + 0.015*"young" + 0.010*"famili" + 0.010*"meet" + 0.009*"search" + 0.008*"begin" + 0.008*"becom" + 0.008*"forc" + 0.007*"look" + 0.007*"fight"

Topic: 4 A Documentary of the World of American TOP 5
Shutter Island
The Last Song
Words: 0.013*"world" + 0.013*"documentari" + 0.012*"wife" + 0.009*"young" + 0.009*"chang" + 0.009*"make" + 0.008*"father" + 0.008*"american" + 0.008*"explor" + 0.007*"follow"

Topic: 5 A Journey of Love TOP 2
Despicable Me
Shrek Forever After
Words: 0.024*"world" + 0.020*"love" + 0.014*"young" + 0.013*"famili" + 0.011*"becom" + 0.010*"time" + 0.010*"start" + 0.008*"woman" + 0.008*"face" + 0.008*"journey"

Topic: 6 Dreams of American Young People TOP 8
The Twilight Saga: Eclipse
The Expendables
Words: 0.026*"young" + 0.010*"dream" + 0.009*"becom" + 0.008*"america" + 0.008*"woman" + 0.007*"documentari" + 0.007*"know" + 0.007*"group" + 0.007*"work" + 0.007*"mysteri"

Topic: 7 Crime TOP 7
Inception
Percy Jackson & The Olympians: The Lightning Thief
Words: 0.017*"young" + 0.011*"world" + 0.011*"look" + 0.009*"discov" + 0.008*"famili" + 0.008*"woman" + 0.008*"kill" + 0.008*"battl" + 0.007*"mysteri" + 0.007*"work"

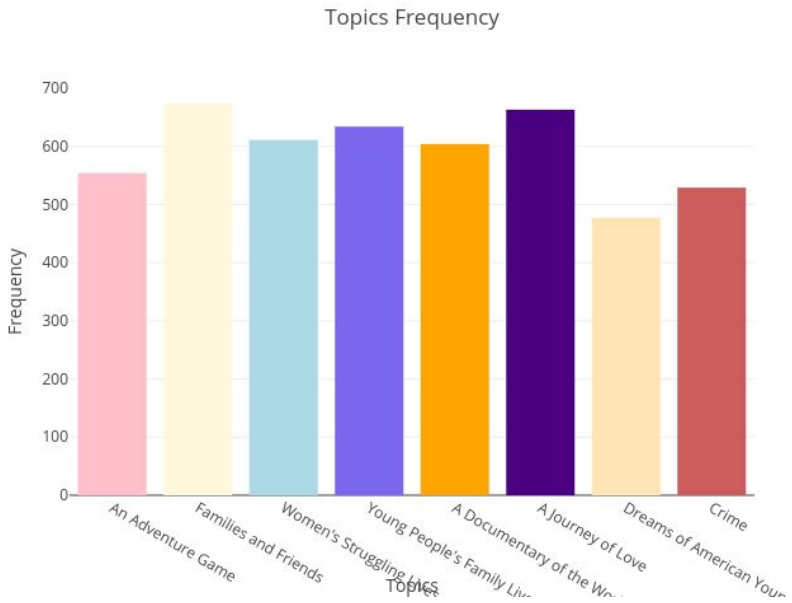
2. **The distribution of topics of each plot**

The ‘score_table’ generated by Python script *topicModeling_v2.py* stores the distribution of topics of each plot and the most relevant one. For example, the *Inception* contains all of the topics, each with different percentage. The topic with the highest % is “Crime”, so it is the best topic for this movie.
Movies with the same topic are similar to each other.

Name	**Best Topic**	Topic 0: An Adventure Game	Topic 1: Families and Friends	Topic 2: Women's Struggling Lives	Topic3 :Young People's Family Lives and Fights	Topic4: A Documentary of the World of American	Topic 5: A Journey of Love	Topic6 : Dreams of American Young People	Topic 7 :Crime	Year
Inception	7	0.011	0.110	0.011	0.011	0.011	0.011	0.286	0.547	2010

3. **The most popular topics and the trend.**

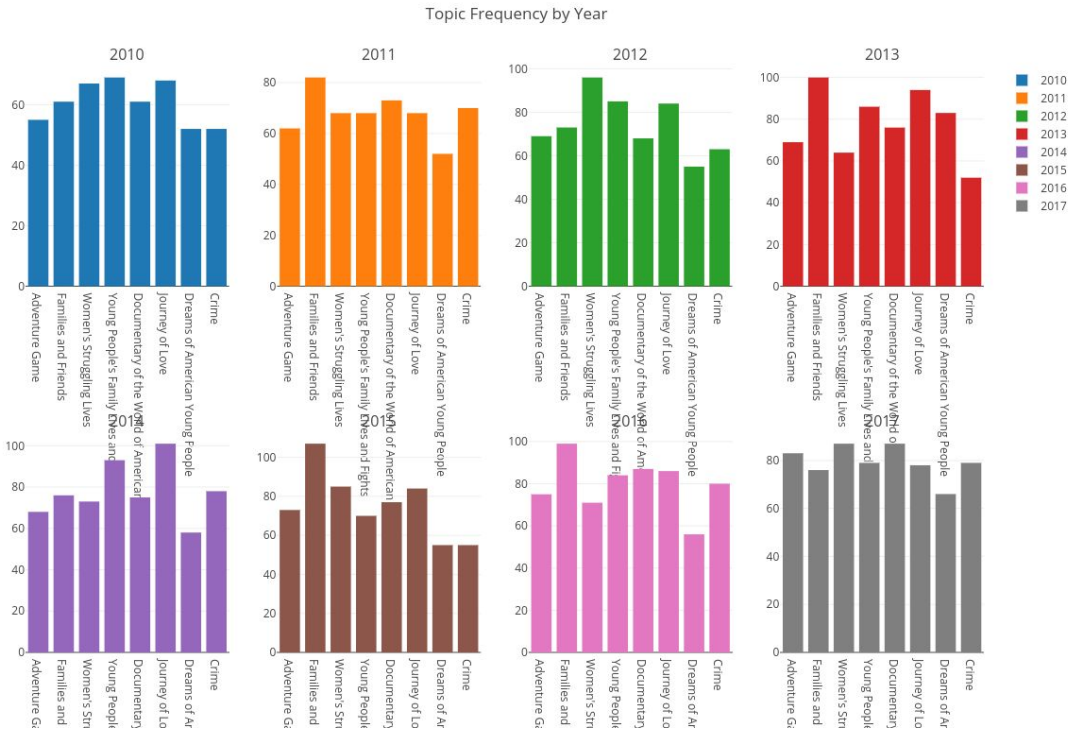
The most popular topic is ‘Families and Friends’. Though with some little difference, in general, topics are evenly distributed.



(This is a screenshot of the plotly graph, please see link below for the interactive graph)
https://plot.ly/~juhuang/18

Though topic ‘Families and Friends’ came the first in some years like 2011, 2013, 2015, and 2016, the topic distribution is still relevantly flat. Moreover, few topic took the first place for two consecutive years, which means the most popular topic last year seems have nothing to do with

the most topic the next year. In general, the trend of the popular topics did not change a lot during the 8 years.



(This is a screenshot of the plotly graph, please see link below for the interactive graph)
https://plot.ly/~juhuang/20

Wordcloud:

As can be seen from the word cloud, the most frequently used words are ‘world’, ‘family’, ‘friend’, ‘love’, ‘young’ ‘man’ and ‘woman’. With these words, we can almost imagine a story ourselves: a young man and a young woman fall in love with each other and explore the world together. During their adventure, something happens to their families and friends. How do they

solve the challenges? This is what the movie is about!

Limitations:

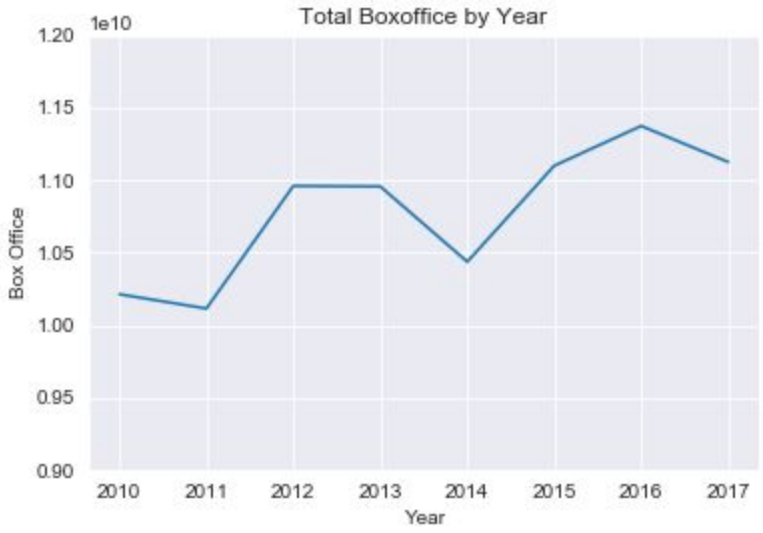
1. There are some overlap between each topic. (eg. Topic “Young People's Family Lives and Fights” have similar content with “An Adventure Game” and “Families and Friends”.)
2. The representative percentage of words for each topic is relevantly low. The highest one is only 0.026.
   (Topic: 6 Dreams of American Young People
   Words: **0.026***"young" + 0.010*"dream" + 0.009*"becom" + 0.008*"america" + 0.008*"woman" + 0.007*"documentari" + 0.007*"know" + 0.007*"group" + 0.007*"work" + 0.007*"mysteri")
3. The topic names are manually decided, and may not perfectly match every movie under the topic.(eg. *The Social Network* is categorized as topic ‘Families and Friends’)
4. The number of topics are manually chosen. Choosing fewer topics leads to a even lower representative percentage of words for each topic, however, more topics lead to a more vague categorization of topics.

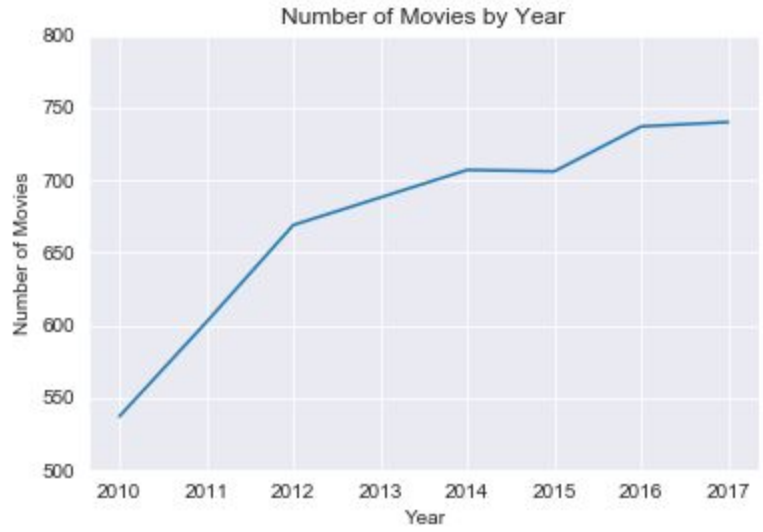
III. Data Visualization

1.Overall Trend of Box Office

In order to analyze the trend of the movie industry from 2010 to 2017, whether the industry is growing, declining or maintain stable over these years, we simply pick the most intuitive and straightforward features from both sides. On the production side, we explore the trend of

number of movies produced from 2010 to 2017; and on the market side, we explore the trend of box office of all movies year over year. It turns out that although there are some fluctuations, the overall trend of number of movies and total box office receipt is growing, and thus we say movie industry is growing.



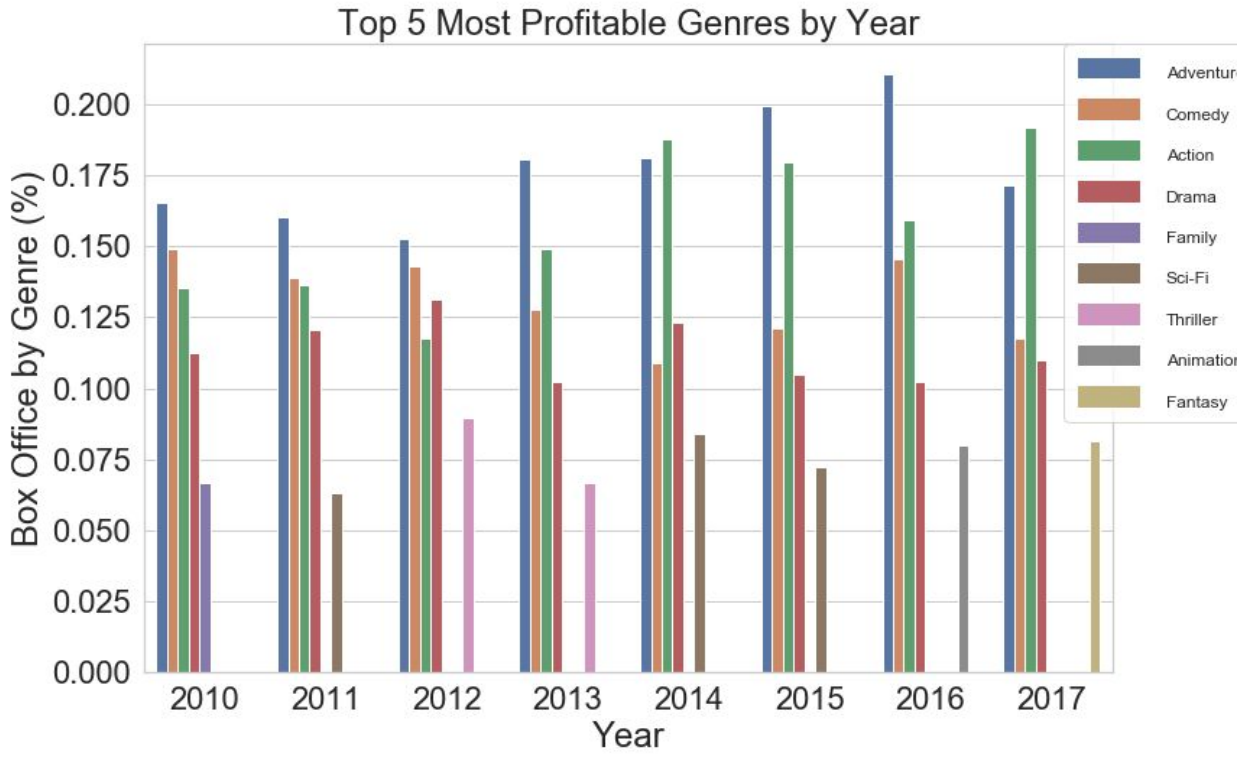


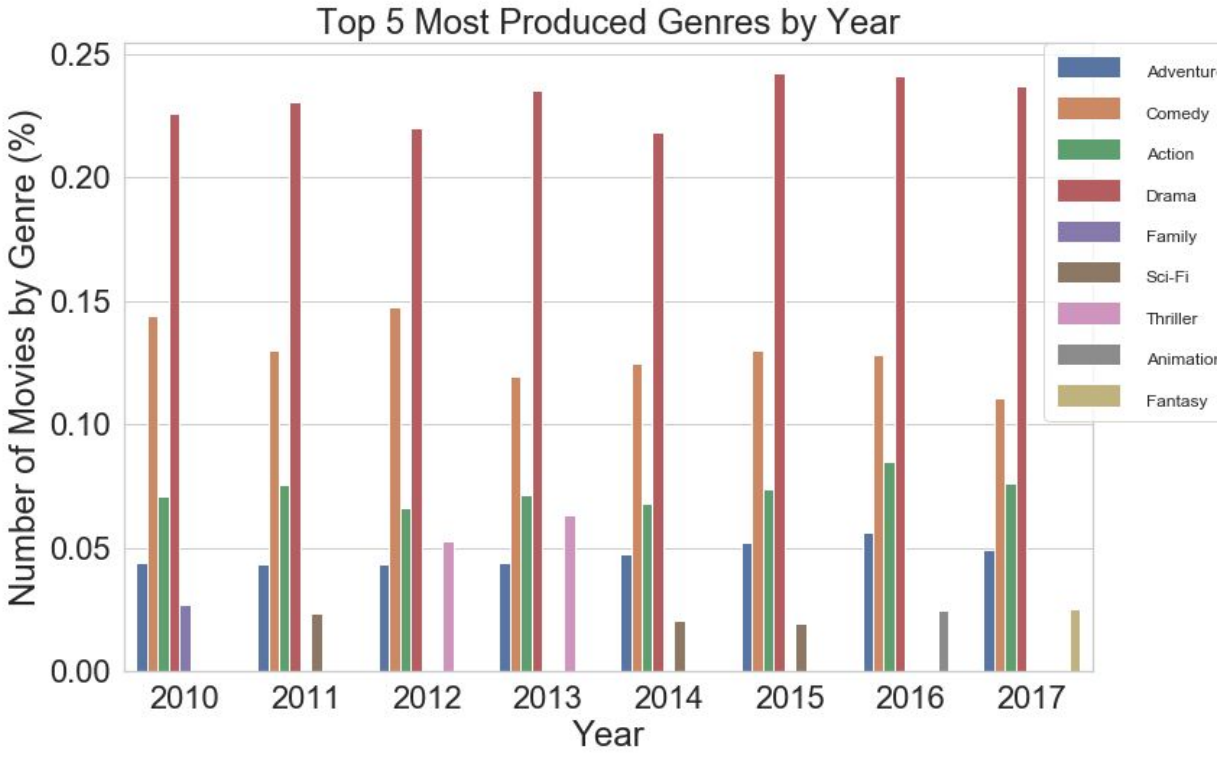
2. Genre Analysis

One of the factors we want to explore is movie genre. We believe there might be some changes in movie genres and could be related to the growth. So we collected the main genres of each movie and try to explore if movies of some certain genres produced less or more often throughout the years. However, as we can see from this bar graph, although there are some

fluctuations, the proportion of genres produced(the x-axis), especially the top genres, does not change a lot.

On the market side, we analyzed the total box office of movies in each genre (the y-axis) to see if audience favor different genres over the years, but it also turns out the most popular genres remain the same throughout the years. The top 3 popular genres are always Adventure, Action and Comedy. The blue, orange and green bars in each year group. However, there is one obvious thing we can observe from the graph, that is the popular genres become more and more popular. That means theaters are more willingly to screen popular genres, or audience's taste become more concentrate, or a combination. The top genres contribute more to the total box office than before.





On one side, we are happy to see that although the top genres make more and more money, the production side still produce relevantly same proportion of different genres. This help protect the diversity of movie production.

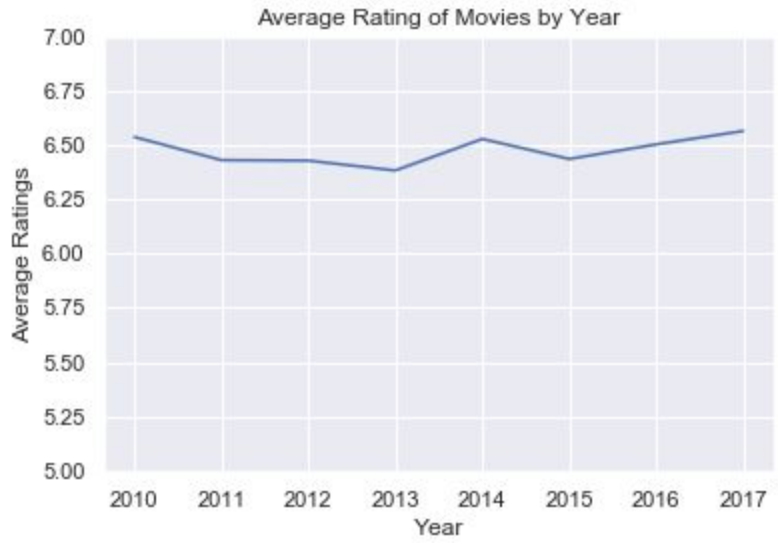
However, the market is strongly driven by box office. Popular genres are more popular year over year. This is sad because we believe that the diversity of market demand would contribute to the diversity of movie production and help the movie industry maintain a virtuous cycle. We hope the niche movie market could obtain more attentions.

Limitations:

The data we gathered from IMDB only record 3 main genres for each movie, but many movies actually contain far more than 3 genres.

3. Line plot of average rating from 2010 to 2017

Since box office is growing year over year, we make a hypothesis that the quality of movies are getting better over time. One variable that shows quality of movie is the average rating (average of Rotten Tomato, IMDB, and Metasocre). Thus we plotted the rating over year.



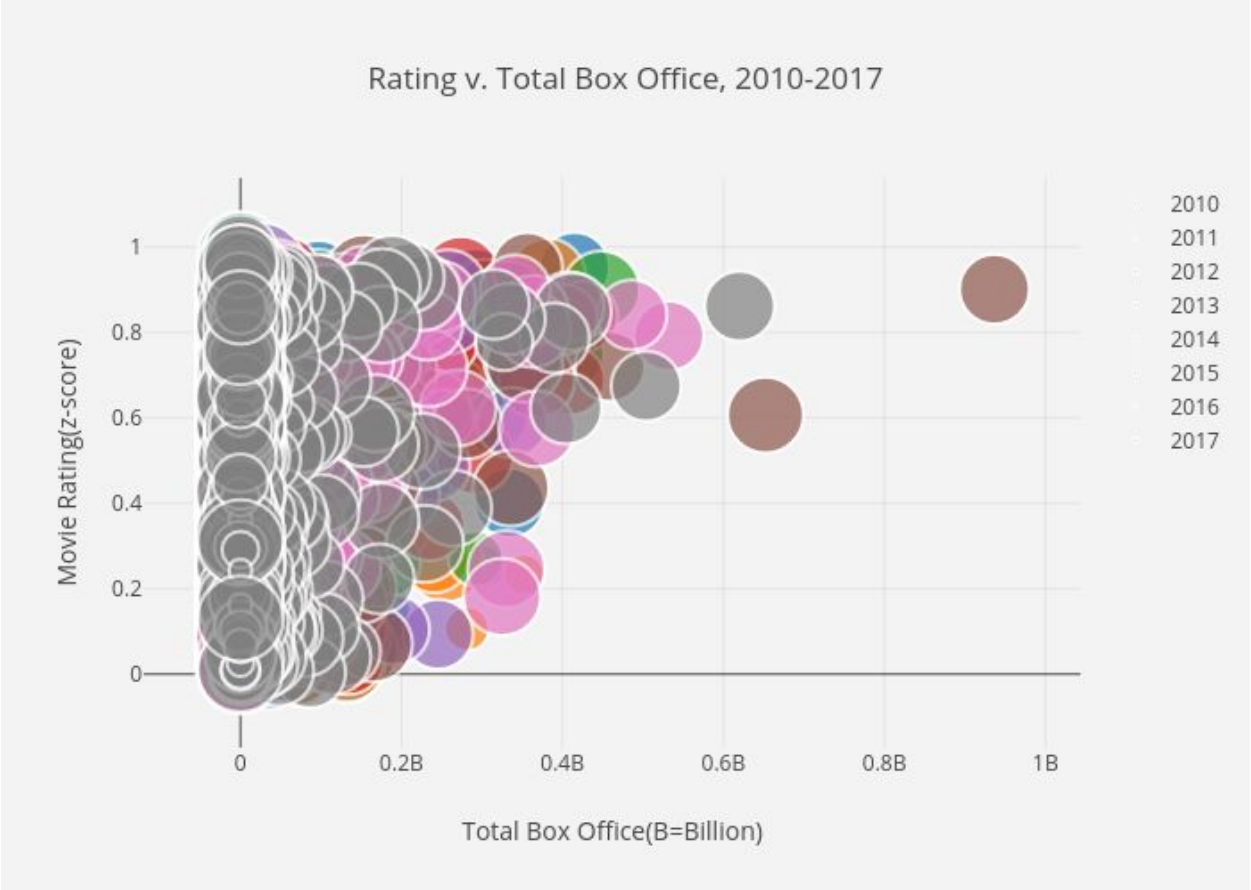
Findings:

The above graph shows the rating trend during these years. From the graph, we can see that the average rating for each year does not change a lot.t fluctuates around 6.5. Thus, the hypothesis that movie qualities are improving is not proved true.

Limitations:

Although the rating is an important indicator of movie quality, there's also a possibility that each year, audience give out 'relative ratings', rather than 'absolute ratings', i.e. audience judge a movie through comparing it with movies of that year, rather than of all time. Thus, it is possible that ratings will never move up over years. However, rating is the best proxy we have so far to prove our hypothesis.

4. Bubble Graph of Rating vs Box Office (with Year and Studio Size Information)



(This is a screenshot of the plotly graph, please see link below for the interactive graph)
https://plot.ly/~cassiezzz/8/rating-v-total-box-office-2010-2017/#/

For the bubble graph, we have ratings on the y-axis, total gross on the x-axis, year for the bubble colors and studio size for the bubble size (the more movies the studio produces, the larger the size of the bubble).

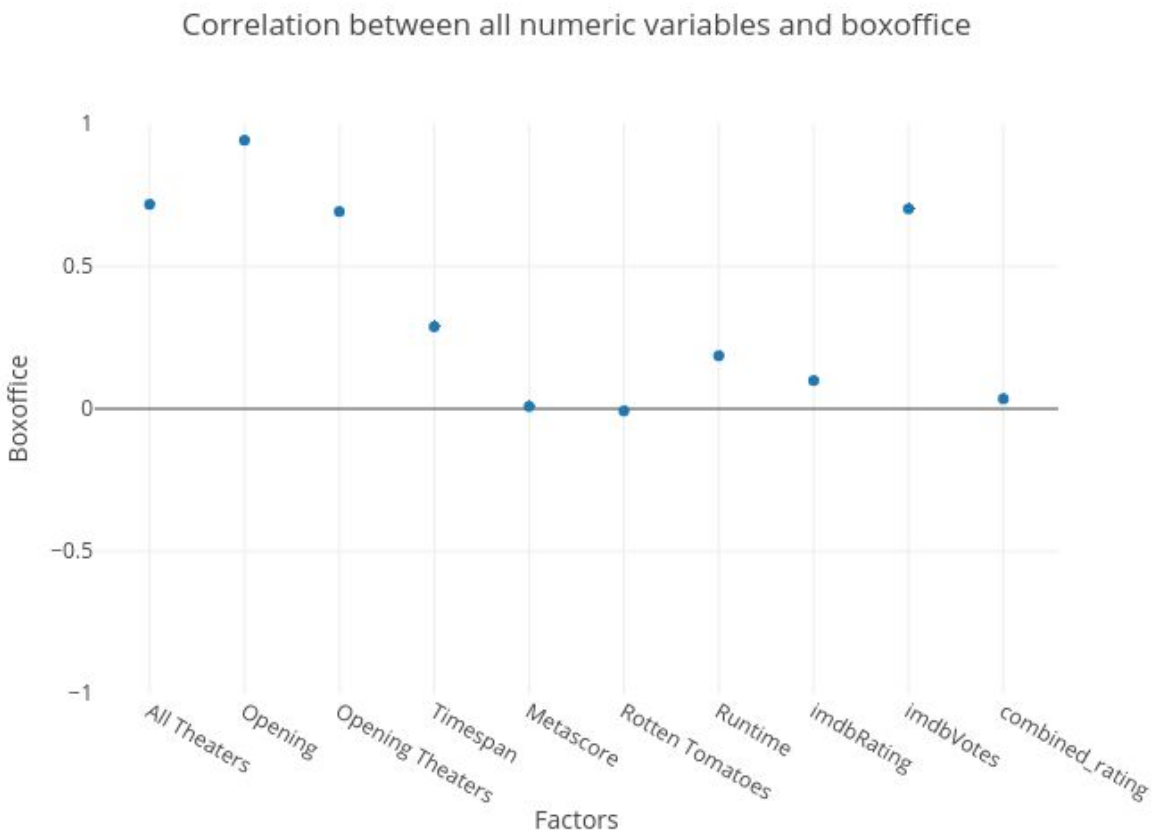
Findings:

1. Ratings and the box office is not strictly correlated. There are a lot of movies with high rating having very small box office. However, if a movie does not have a relatively high rating, then it probably does not have a large box office. If a movie has a large box office, its rating is probably also relatively high. In short, high rating is a necessary but not sufficient to have high box office.
2. Bubble size depends on the studio size, which is the number of movies that the studio produces. It shows that large studio might also produce a lot of movies with low ratings or small box office. However, most movies with large box office are produced by large studio such as Buena Vista (known as Walt Disney Company) and Universal Pictures.
3. Distribution of bubbles for every year is similar. It depicts that the movie market does not have great change every year.

Limitations:

1. Some movies don't have any ratings so we delete them from the dataset (about 500 movies).
2. Some movie lost only one or two rating data, we fill them as the mean of ratings we have.

5. Correlations Between Numeric Variables and Box Office



(This is a screenshot of the plotly graph, please see link below for the interactive graph)
https://plot.ly/~juhuang/14

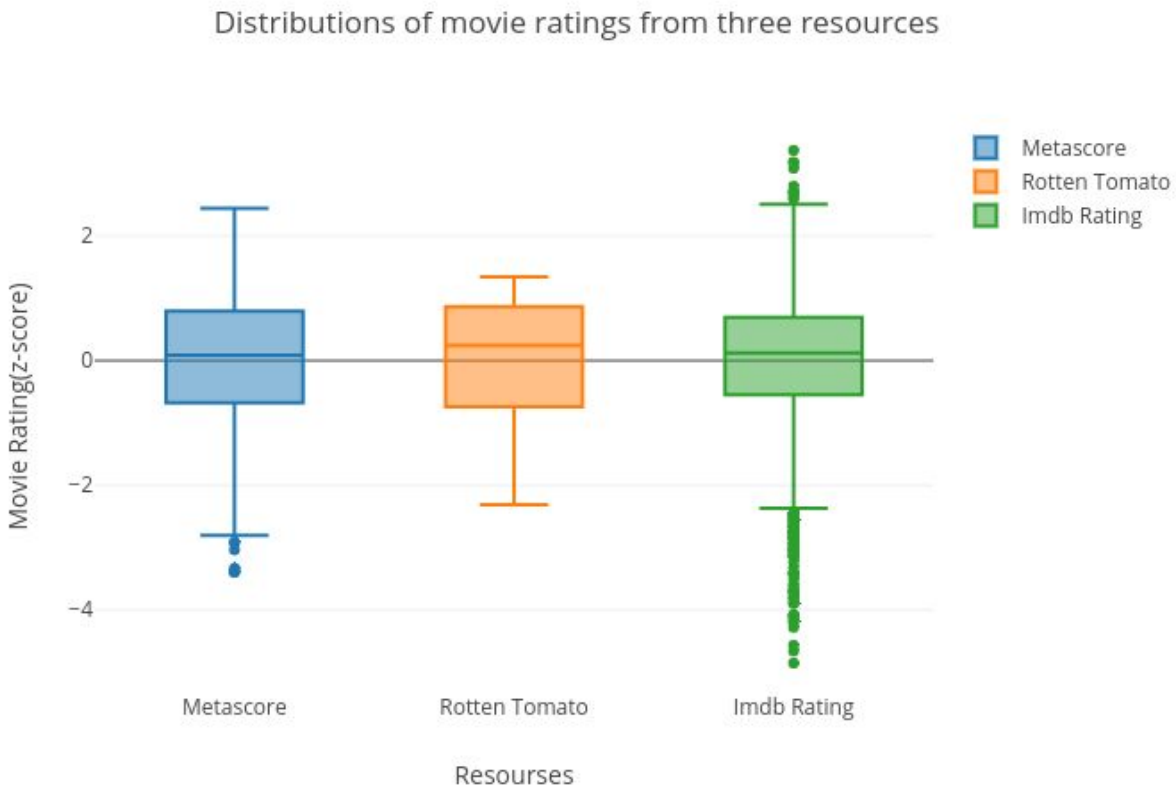
For the scatter plot, we have correlation with boxoffice on the y-axis, numeric variable names on the x-axis.

Findings:

1. Ratings and the box office is not correlated, which means movies with good reputation may not be movies with high boxoffice.
2. Opening is extremely correlated with boxoffice, which means if a movie performs good on the first day, it is very much likely that it has a high boxoffice at the end.

3. All theaters and opening theaters are highly correlated with boxoffice, which means that the more theaters play the movie, the higher boxoffice it will be.
4. imdbVotes is highly correlated with boxoffice. It makes sense intuitively. The number of audience in theaters is highly correlated with the number of votes on IMDB.

6. Movie Rating Distributions of Three Major Rating Websites



(This is a screenshot of the plotly graph, please see link below for the interactive graph)
https://plot.ly/~juhuang/12

The above graph shows three boxplot. Each boxplot represents the distributions of movie ratings from three major rating websites: Metascore, Rotten Tomato, IMDB.

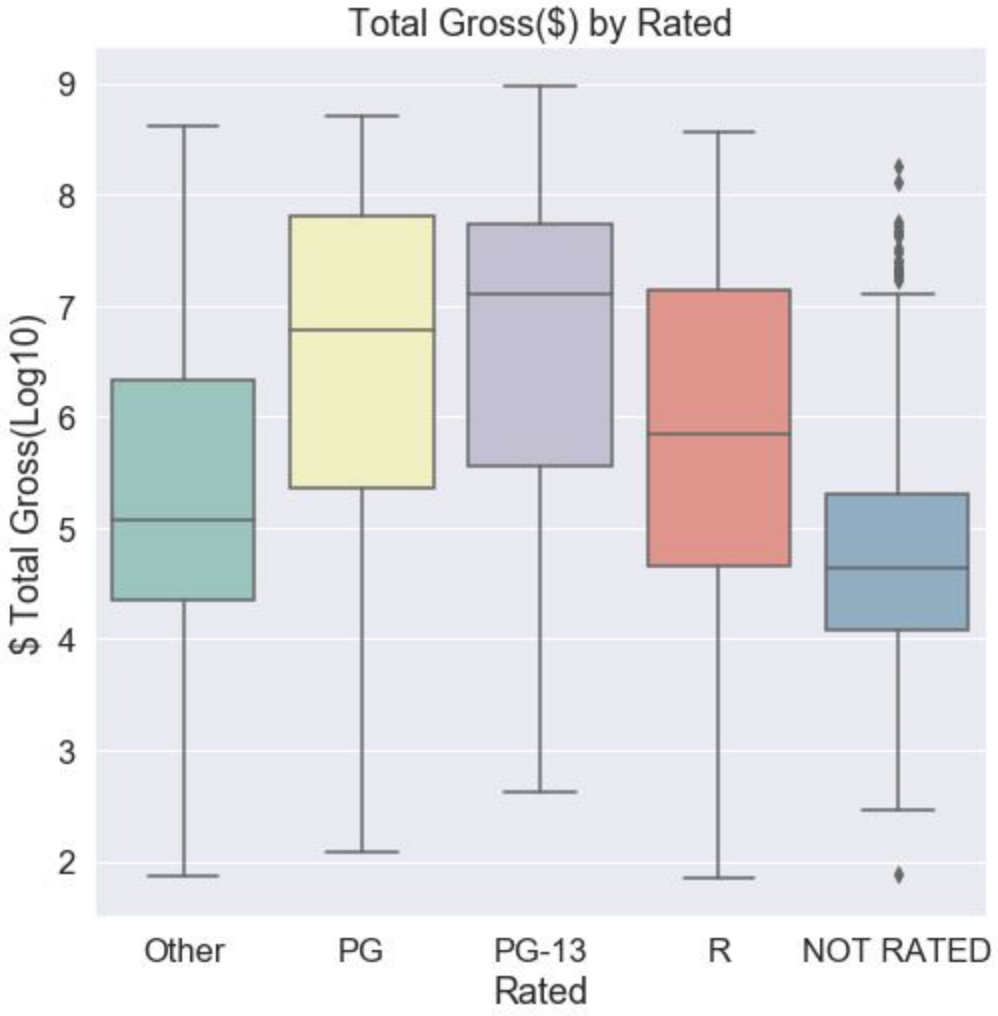
Findings:
Compared with other two websites, people on imdb Rating tend to rate extremely on movies. They give very high scores for good movies and very low scores for poor movies.

Please see Visualization 7 from Topic Modeling Section - Topic Frequency By Year.
Please see Visualization 8 from Topic Modeling Section - Wordcloud of Topics.

Visualization: Impact of Movie Rating (PG, R, etc) on Box Office

(EXTRA CREDIT)

From the linear regression result (in Project II), we found out that how movie is rated is a significant variable that impact the box office. We will visualize how movies with different rating differs in box office.



Findings:

From the above graph, we can see that PG-13 has the highest box office, followed by PG, R, Other & Not Rated. This shows that movies that are not children-friendly (PG-13) can gain the most popularity. However, movies that contains too much children-unfriendly content (R) is not as popular.