# CMPT 353 Project Report

## ***Wikidata, Movies, and Success***

Summer, 2019

Presented by:<br>
DeLong Li 301297103<br>
George He 301315033

**Q1. Do the various criteria for success (critic reviews, audience reviews, profit/loss) correlate with each other? Is there something you can say about better or worse kinds of “success”?**

**Refine the original question:**

1. Study the correlations between the following 6 variables: audience_average, audience_percent, audience_ratings, critic_average, critic_percent and made_profit.
2. Based on the result, what are the possible definition of “success”?

**Data gathering and cleaning:**

We started with the provided wikidata-movies.json.gz and discovered that the cost and revenue of each movie are missing, so we modified build_wikidata_movies.py to include ‘nbox’ and ‘ncost’ columns, then we uploaded build_wikidata_movies.py and transform_download.py to cluster and executed them to get newwiki-movies.json.gz.

After selecting relevant columns and dropping rows with NaN, we ended up with only 817 movies with valid revenue and cost, which size is very small compared to data obtained in rotten-tomatoes.json.gz (12764). Therefore we decided to analyze the correlation between all other 5 variables first, then inner join two dataframes by rotten_tomatoes_id and check if profit is correlated with any of the other 5 variables.

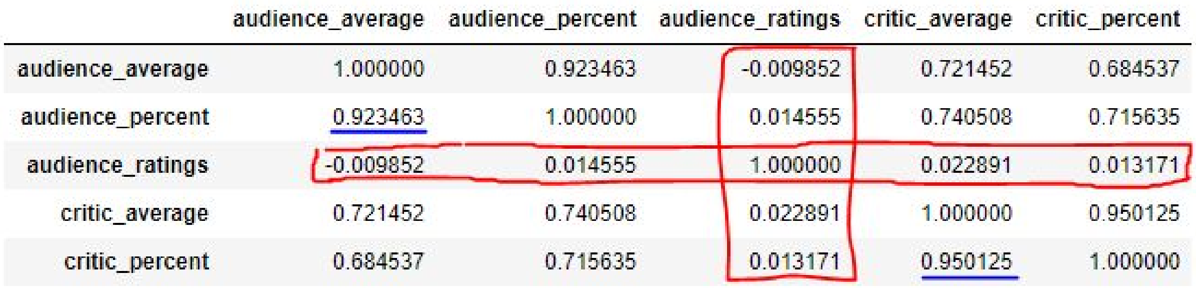
We loaded rotten-tomatoes.json.gz as a pandas dataframe (roto), filtered out movies that have audience_ratings < 1000 because we think movies that have less than 1000 reviews are not credible thus cannot be included into analysis.

We decided to include the audience_ratings column into analysis, considering the possibility of it relating to the success of movies and having correlation with other variables.

**Analyzing:**

We started by scatter plotting every pair of variables. By examining all plots, we figured that every plot without audience_rating showed some weak positive linear trend, with two pairs audience_averange, audience_percent and critic_averange, critic_percent being more significant, and others with a lot of variance.

We then used .corr() function on dataframe roto to produce the following r-values for pairwise correlation of columns (variables). The result is shown below.



| | audience_average | audience_percent | audience_ratings | critic_average | critic_percent |
|---|---|---|---|---|---|
| audience_average | 1.000000 | 0.923463 | -0.009852 | 0.721452 | 0.684537 |
| audience_percent | 0.923463 | 1.000000 | 0.014555 | 0.740508 | 0.715635 |
| audience_ratings | -0.009852 | 0.014555 | 1.000000 | 0.022891 | 0.013171 |
| critic_average | 0.721452 | 0.740508 | 0.022891 | 1.000000 | 0.950125 |
| critic_percent | 0.684537 | 0.715635 | 0.013171 | 0.950125 | 1.000000 |

We interpreted the results:

1. From r values underlined in blue,  there exists a strong positive linear relationship between audience_average and audience_percent (r = 0.92), as well as between critic_average and critic_percent (r = 0.95). The implication is that audience/critic who likes the movie is almost certain to give a higher rating and vice versa.

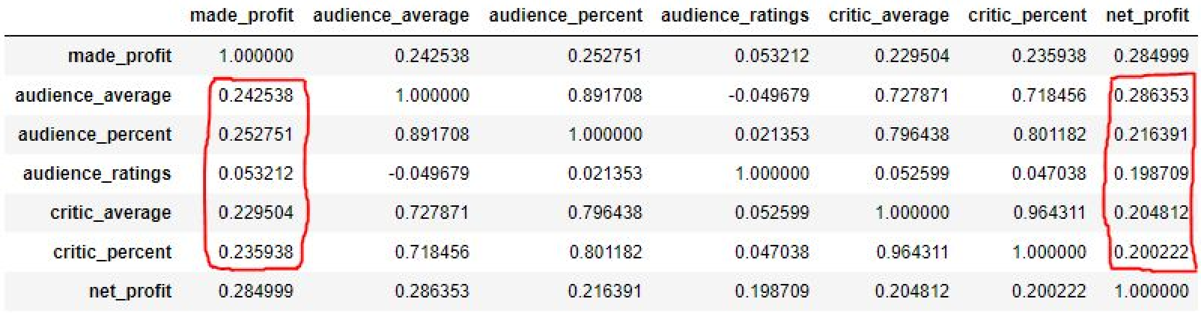
2. From all correlation coefficients circled in red that are close to 0 and corresponding scatter plots, we think it is safe to conclude that audience_ratings is not correlated with any other 4 variables. Meaning there is no relationship between number of audience reviews and score of audience/critic reviews.

3. All other r-values except those on the diagonal are in the range 0.68~0.75. We think there exist relatively strong correlations between audience reviews and critic reviews (both average and percent). Therefore if audience/critic review score is high on a certain movie, the corresponding critic/audience score is likely to be high as well.

After examining the correlation coefficients, we performed 6 linear regression tests on each pair chosen from four correlated variables using linregress(), along with p-value verification and residual normality test. The results yielded significant p-values for all 6 regression line tests, but no pair of variables passed residual normality test, not even the two pairs that have correlation coefficients larger than 0.9. Though the residual plots look reasonably normal enough to us and have n ≥ 40 for all tests.

We now include profit into correlation analysis. First we outer joined two dataframes by 'rotten_tomatoes_id', dropped rows with NaN, then we created an extra column ‘net_profit’. The column is created by calculating nbox - ncost.

At last, we again called corr() on the joined dataframe and result is not exciting:



| | made_profit | audience_average | audience_percent | audience_ratings | critic_average | critic_percent | net_profit |
|---|---|---|---|---|---|---|---|
| made_profit | 1.000000 | 0.242538 | 0.252751 | 0.053212 | 0.229504 | 0.235938 | 0.284999 |
| audience_average | 0.242538 | 1.000000 | 0.891708 | -0.049679 | 0.727871 | 0.718456 | 0.286353 |
| audience_percent | 0.252751 | 0.891708 | 1.000000 | 0.021353 | 0.796438 | 0.801182 | 0.216391 |
| audience_ratings | 0.053212 | -0.049679 | 0.021353 | 1.000000 | 0.052599 | 0.047038 | 0.198709 |
| critic_average | 0.229504 | 0.727871 | 0.796438 | 0.052599 | 1.000000 | 0.964311 | 0.204812 |
| critic_percent | 0.235938 | 0.718456 | 0.801182 | 0.047038 | 0.964311 | 1.000000 | 0.200222 |
| net_profit | 0.284999 | 0.286353 | 0.216391 | 0.198709 | 0.204812 | 0.200222 | 1.000000 |

All values circled in red are close to 0 (with max = 0.28) which means that most likely, profit is not correlated with any of the other 5 variables.

**Conclusion:**

1. audience reviews are positively correlated with critic reviews. Audience/Critic averages are strongly correlated with audience/critic percents. Even though the residuals are not normally distributed for all pair of correlated variables, the requirement for normality can be softened with kind-of-normal data and n ≥ 40. Which means it is feasible to make reliable predictions especially on average and percent of audience/critic reviews. At last, profit and number of reviews are not correlated with any other variables.

2. Based on the results, we think better or worse kind of movies can be interpreted w.r.t three independent variables: profit, audience/critic review score and the number of reviews (popularity).
   If a movie has good profit, very positive audience/critic reviews and high number of reviews, then it is obviously very successful (better success), but if a movie has only considerable amount of reviews and non of the other two, then maybe it is “successful” in some way but obviously not as successful as the former (worse success). Therefore the degree of “success” of a movie can be evaluated with respect to the movie’s performance on all three criterias.

**Limitations:**

1. Too few data points for profit/loss analysis. The conclusion that profit has no correlation with other variables may not be correct considering the fact that we only had 817 movies with valid profit to work on.

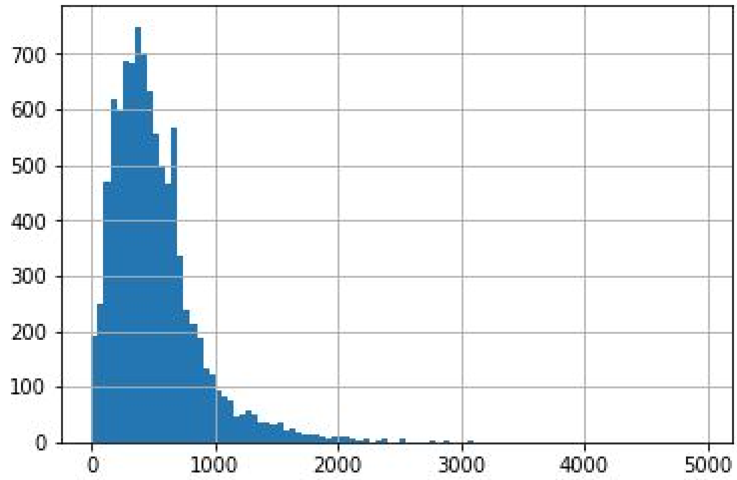
2. With more time, we could have trained a linear model to see if it is actually feasible to make predictions on two pairs of variable that have correlation coefficients > 0.9 and output the accuracy scores.

3. Without knowing how the review scores and ratings are collected, we cannot guarantee no bias in data. For example the rating of a movie can be greatly affected by companies that specialize in click farming.

**Q2. Explore the possibility of predicting genre using the corresponding plot summary.**

**Data gathering and cleaning:**

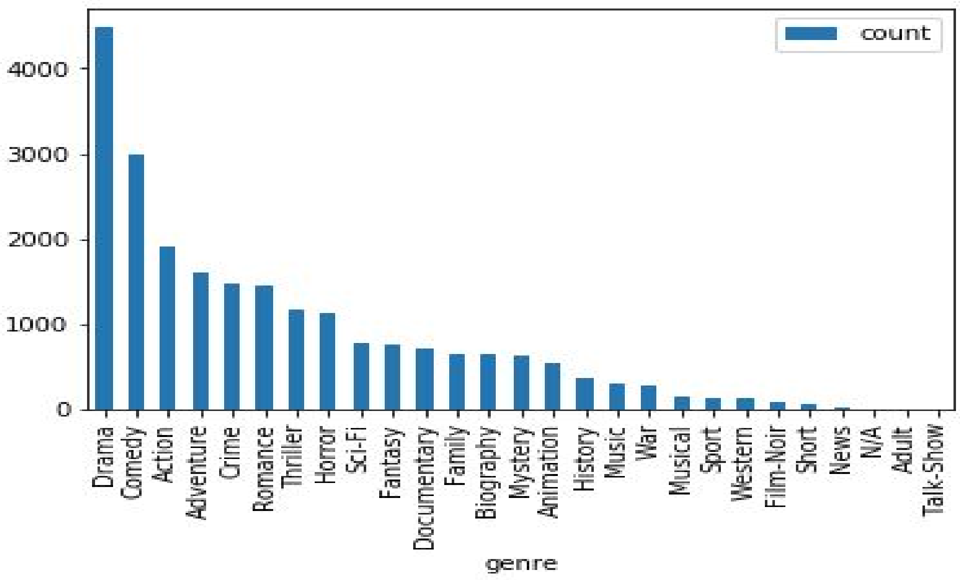
We plotted the length of all plot summary as histogram:



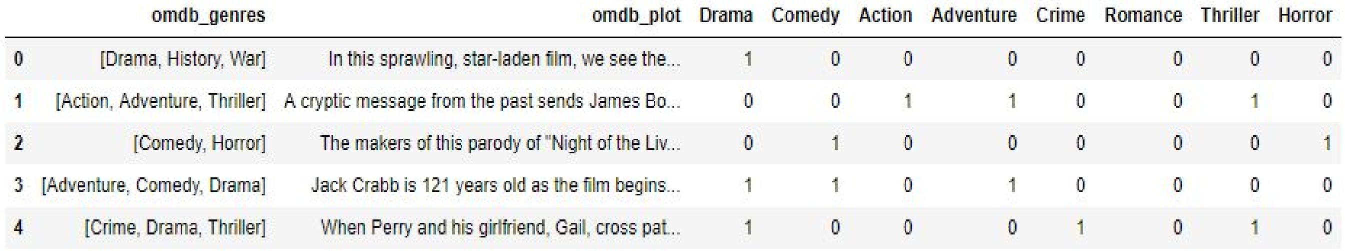
The minimum length is 3 characters and max is 7104. We decided to removed all documents that have < 20 characters and > 1500 character, since they cannot be predicted reasonably due to the unreasonable size. We ended up with 9276 plot summaries.

**Analyzing:**

We then wrote a function to find all genres and their counts, there are 27 genres in total:

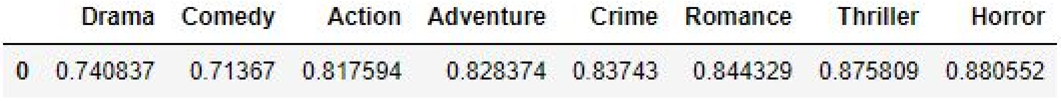


We decided to predict each of the 8 genres that have count > 1000 separately using all plots. We started by creating a column for each of the 8 genres that consists of 0 and 1, 1 means the corresponding movie belongs to that genre and 0 means it does not.



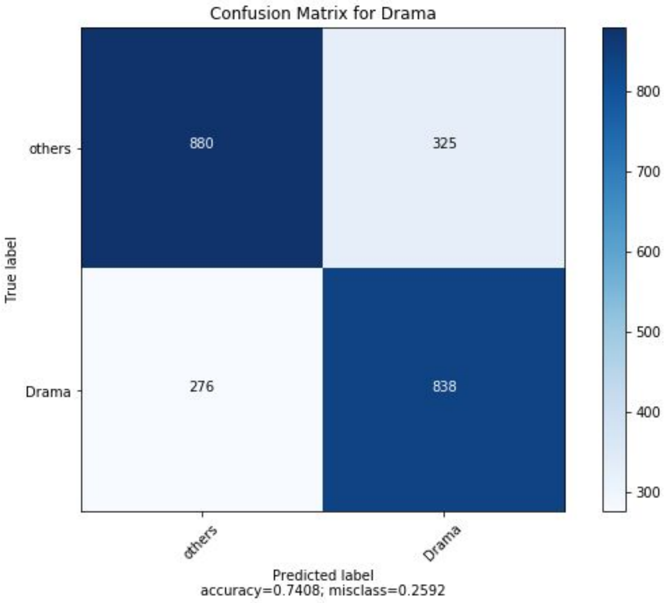
| | omdb_genres | omdb_plot | Drama | Comedy | Action | Adventure | Crime | Romance | Thriller | Horror |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | [Drama, History, War] | In this sprawling, star-laden film, we see the... | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | [Action, Adventure, Thriller] | A cryptic message from the past sends James Bo... | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 2 | [Comedy, Horror] | The makers of this parody of "Night of the Liv... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | [Adventure, Comedy, Drama] | Jack Crabb is 121 years old as the film begins... | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | [Crime, Drama, Thriller] | When Perry and his girlfriend, Gail, cross pat... | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |

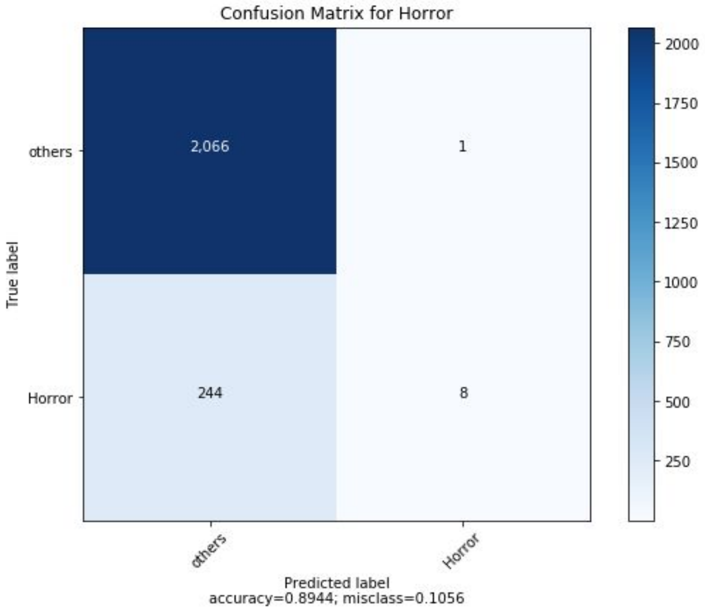
Then for each of the genre, we set up X as summary plots and y as the corresponding genre's column of 1 and 0. Constructed the TfidfVectorizer to fit_transform plot summaries to feature vectors. Then we created MultinomialNB model to fit then predict the genre, as the model is suited for Natural Language Processing. The accuracy score for each genre is shown below:



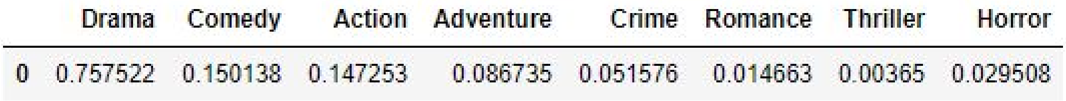
| | Drama | Comedy | Action | Adventure | Crime | Romance | Thriller | Horror |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.740837 | 0.71367 | 0.817594 | 0.828374 | 0.83743 | 0.844329 | 0.875809 | 0.880552 |

Through observation we found that the accuracy score tends to increase as the predicted genre gets more and more uncommon (having less count). We think it is because when the genre's count gets small relative to the overall training set data, the model can achieve very high accuracy score just by making more 0 predictions. A model that always predict 0 for genre Horror will have accuracy score 0.877 = (9276-1139)/9276. Therefore in order to see if the model is actually making correct and reasonable predictions, we plotted the confusion matrix for each genre using a function found online. Take Drama and Horror for example:





Though the first confusion matrix looks promising. The model made 2066 negative predictions with only 8 positive predictions for genre Horror while having accuracy score of 0.8944 as depicted by the second confusion matrix. The impressive accuracy score thus cannot be trusted and we have to rely on the True Positive Rate (when it’s actually yes, how often does the model predict yes) to evaluate the model’s validity. We calculated the True Positive Rate for each model respectively:



| | Drama | Comedy | Action | Adventure | Crime | Romance | Thriller | Horror |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.757522 | 0.150138 | 0.147253 | 0.086735 | 0.051576 | 0.014663 | 0.00365 | 0.029508 |

**Conclusion:**

The True Positive Rate declines rapidly as the model switches from Drama to other genres, the extremely low TPR confirms the fact that all models after Drama are simply making more and more negative prediction to achieve high accuracy score. While it seems the model is completely useless, the True Positive Rate for Drama is 0.75, which reasonably differs from its proportion in overall count 4489/9276 = 0.4839. Therefore we think the model for Drama is the only model with authentic and meaningful accuracy score.

**Limitations:**

1. With more time, we could have achieved higher accuracy score by filtering the summary plots better using other tools as well instead of just relying on TfidfVectorizer.

2. We do not have enough data points for categories after drama, which may be the cause of biased model.

3. We could have used GridSearchCV to optimize the model for predicting genre Drama.

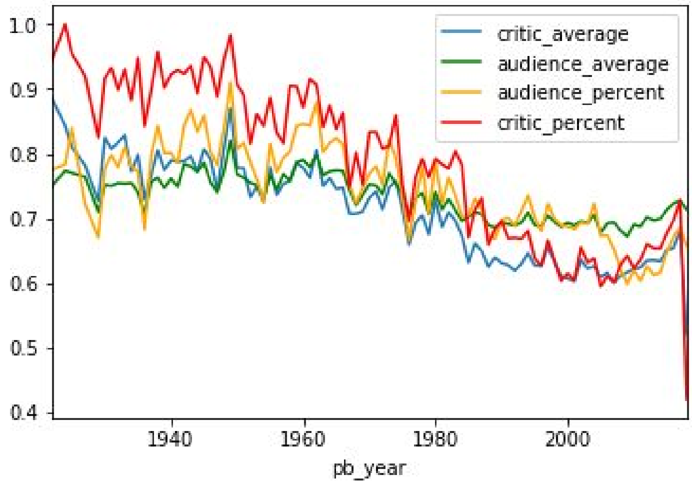
**Q3. Have any of these things changed over time (depending on the movie's release date)?**

**Refine the original question:**

1. Has the reviews (audience_average, audience_percent, audience_ratings, critic_average, critic_percent) for the top three genres changed over time?
2. Has the number of movies produced for top three genres for each year changed over time?

**Data gathering and cleaning:**
We loaded rotten-tomatoes.json.gz and wikidata-movies.json.gz, dropped rows with NaN and joined them after selecting relevant columns and scaled all reviews to the range 0~1. Then we converted the publication date (which is of type string) to date object then to timestamp. By examining all 12760 movies, we found that the maximum number of genres any movie can has is 11. We then decided to take the top three genres into analysis, these are: drama (4797), comedy (2551) and action (1624).

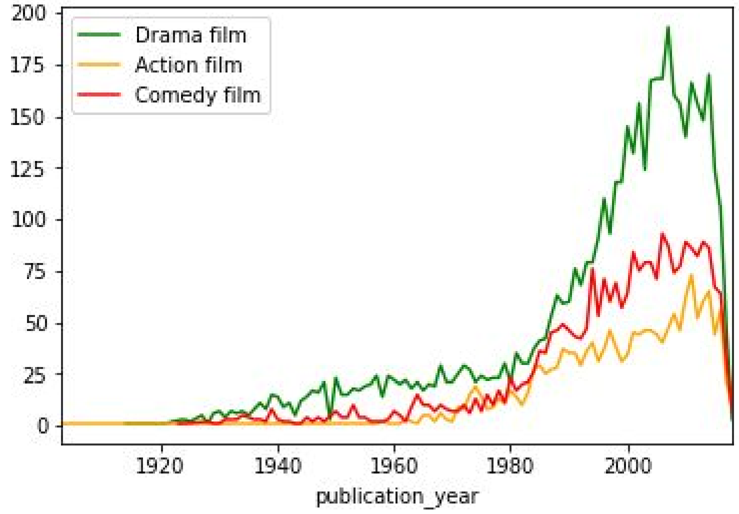
**Implement and Analyze:**
First we want to investigate if the reviews for top three genres have changed over time. Since many movies have multiple genres, we decided to include a movie into a corresponding genre if that movie is of that genre regardless of other genres it belongs to. After classification, we ended up with three dataframes, by grouping them by year and taking the average of reviews, we produced one plot for each genre, take drama for example.
Drama:



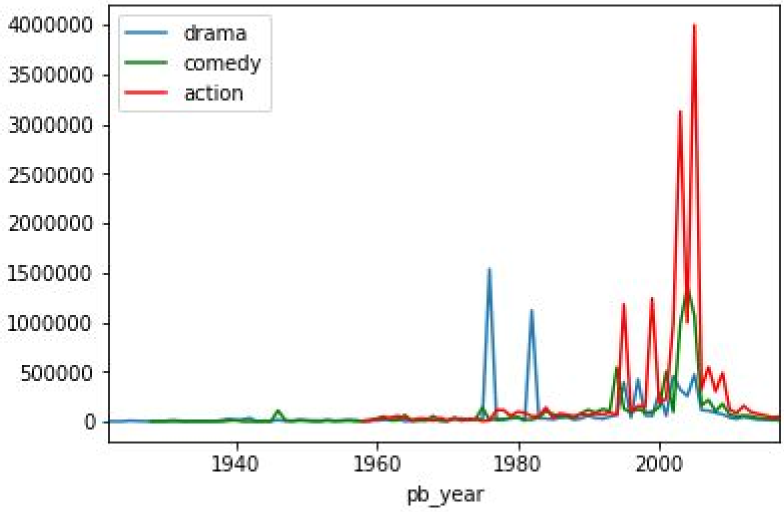
We observed a downward trend starting in all plots since 1920 and gradually changing to upward trend around 2000. All four kinds of reviews (audience and critic, average and percent) tend to move along same directions with more fluctuations in critic’s reviews.

Even though the trend is obvious in each plot, linregress yielded extremely small negative slope for each variable in each plot (-e-10~0). Therefore we cannot conclude that the reviews for the top three genres has changed over the years.

Then we continue to examine the number of movies produced for each genre in each year. This is done by using groupby and count on three dataframes each containing movies of one genre. The plot is shown below:



We compared it with the averaged audience_ratings per year which is shown below:



From the second graph we can see that the most popular genre just after 2000 is action and followed by comedy, and lastly drama. However, the first plot suggested that most movies produced are of genre drama, which is almost twice the number of action movies produced in each year.

**Conclusion:**

1. The reviews of top three genres may be decreasing over the years but we cannot be sure because of the extremely small negative slope
2. The number of movies produced for each of the three genres are increasing exponentially with drama being the most popular genre in movie industry. However the audience's attention is mostly attracted by action movies.

# Project Experience Summary

DeLong Li:

- Initial general data cleaning and manipulating the data for easy analysis
- Collaborating, brainstorming with teammate about how to implement the data science to find answers for questions, as well as coming up with the overall structure of program.
- Utilize pandas learned in class to find out correlation among criteria for movie success
- Utilize pandas to come up with a procedure to separate data into groups based on genre and do analysis on the potentially changing factors.
- Use SFU gitlab for documentation and version control
- Overviewing and putting together work of teammates to make overall analysis

George He:

- Initial general data cleaning and manipulating the data for easy analysis
- Performed ETL processes on several data sets using python tools such as numpy and pandas.
- Brainstormed with another teammate to produce creative questions and solutions.
- Analyzed correlation between multiple variables effectively and came up with meaningful interpretation.
- Applied RandomForestClassifier to do multilabel classification for the features of the movies, and ranked the features by their importance to see which are the most related.
- Applied both Tfidfvectorizer and Word Count Vectorizer to the plot summaries of the movies to do natural language processing and explored if it is possible to predict a movie's success from its plot summary.