# Final Report on IMDb Data Analysis

Ruixuan Zhao, Ruochen Yin, Chushi Shi, YiHsuan Tsai, Lokeswar Sadasivuni

# Introduction

The goal of the project is to get insights in the IMDb data (https://www.imdb.com/interfaces/). We mainly focus on three questions:
1. What is the trend of ratings and audience interest over time?
2. Is there any difference in audience interest among different kinds of productions?
3. The best production and movie people in the eyes of audience?

To answer these questions, we paralleize our computation in 850 jobs on the CHTC platform, plot related graphs and make a conclusion under statistical results. In general, we find that there is no signficant changes on ratings and audience interest from 1840 to 2019. For the difference among different kinds of production, apart from that movie is always dominated among the audience, specially some modern entertainment and production, like video games and TV series, become more and more fashionable, and they have called more audience attention and higher ratings in recent 30 years. Specifically, we list top10 movie people by role and top5 productions by category.

# Data Description & Cleaning

Our data is refreshed daily on website, so we use 7 tsv.files(size) from IMDb Datasets on 12/06/2019 and their total size is 4 GB. The followed are 4 main tsv.files of 7 with description and selected features.

Basics

| variable | description |
|---|---|
| tconst | title Id, alphanumeric unique identifier of the title |
| titleType | the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc) |
| primaryTitle | the more popular title / the title used by the filmmakers on promotional materials at the point of release |
| startYear | represents the release year of a title. In the case of TV Series, it is the series start year |

Principals

| variable | description |
|---|---|
| tconst | title Id, alphanumeric unique identifier of the title |
| nconst | person Id, alphanumeric unique identifier of the name/person |
| category | the category of job that person was in |

Ratings

| variable | description |
|---|---|
| tconst | title Id, alphanumeric unique identifier of the title |
| averageRating | weighted average of all the individual user ratings |
| numVotes | number of votes the title has received |

Name

| variable | description |
|---|---|
| nconst | person Id, alphanumeric unique identifier of the name/person |
| primaryName | name by which the person is most often credited |

We $wget$ 7 tsv.files from http://datasets.imdbws.com/filename.tsv.gz (http://datasets.imdbws.com/filename.tsv.gz), $gzip$ these seven $tsv.gz$ files, and transfer them to csv type by $tr$. In $title.basics.csv$, we get all the types of production and $awk$ to the $type.txt$ file. Then we merge $title.basics.csv$ and $title.ratings.csv$ to $merge\_br.csv$ by $tconst$, and merge $title.principals.csv$ and $title.ratings.csv$ by $tconst$. In addition, for dealing with different type of casts conveniently, we split the $title.principals.csv$ to five csv files, $actor.csv$, $actress.csv$, $director.csv$, $producer.csv$, $composer.csv$. During the cleaning, we deduplicate related csv files for the error of overlap in database.

# Statistical Computation

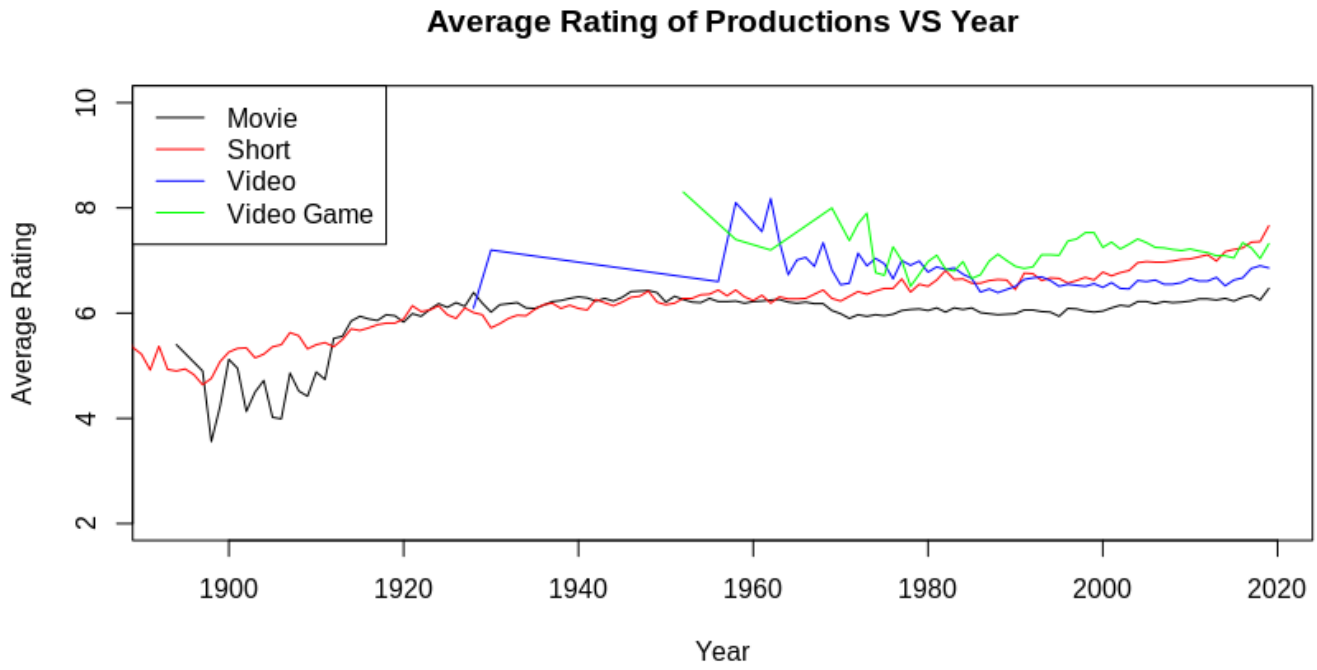
To answer question 1&2, we select $averageRating$ and $numVotes$ as criterion of ratings and audience interests, and parallize to calculate average ratings and voting quantity of each production(movie, short, tvShort, etc) in each year on the CHTC. Here, for a more accurate result in statistics, firstly we split category to 2 big parts. The one is traditional categories(Movie, Short, Video and Video Game) and the other is TV categories(TV Movie, TV Episode, TV Mini Series, TV Short and TV Special). Then, we plot the trend of average ratings and voting over time by category seperately.

To answer question 3, we process $merge\_br.csv$ file, and parallize to sort and get 5 production with the highest ratings in each category and output to $best.csv$. For more accruate and significance, we filter out the production with less than or equal to 800 reviews. To get best movie people, we select the data in movie category, and prarllize 5 jobs to process $actor.csv$, $actress.csv$, $director.csv$, $producer.csv$, $composer.csv$. Specifically, we calculate average ratings of all works for each movie people, and sort the Top 10 person by role. And we also get all works of 3 one of these top 10 person by role.

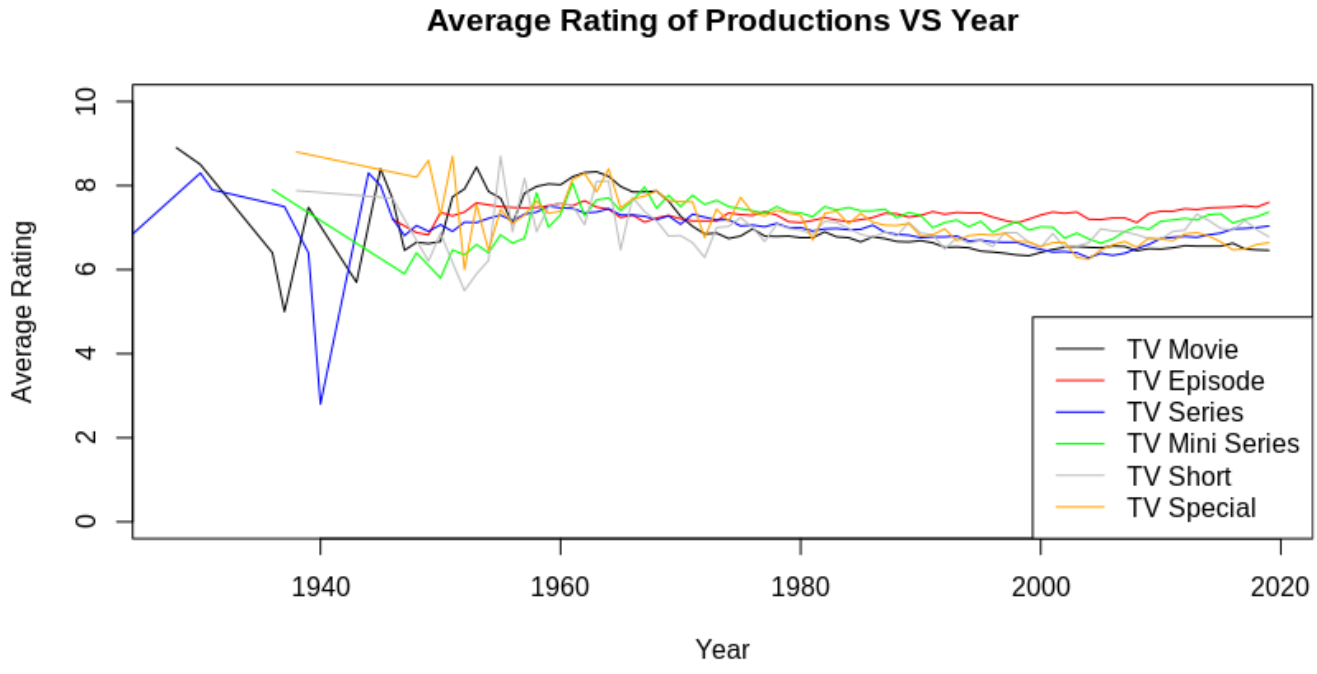
For above whole process, we run three part of works on CHTC which contains 835, 10 and 5 parallel jobs, and for each part of work, it takes no more than 20 min to get the results. Morever, our data is refreshed daily, so the code can get real-time results.

# Result

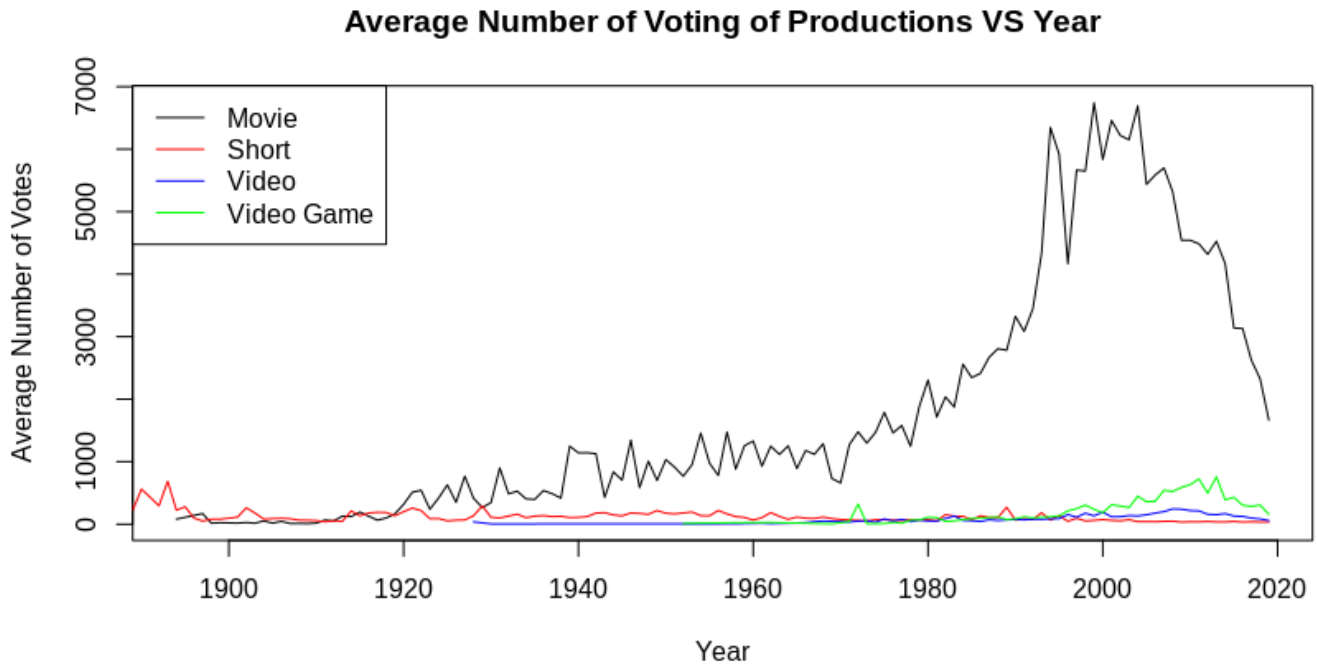
For 2 category parts, traditional categories and TV categories. We get 4 graphs about ratings and the numbers of voting.



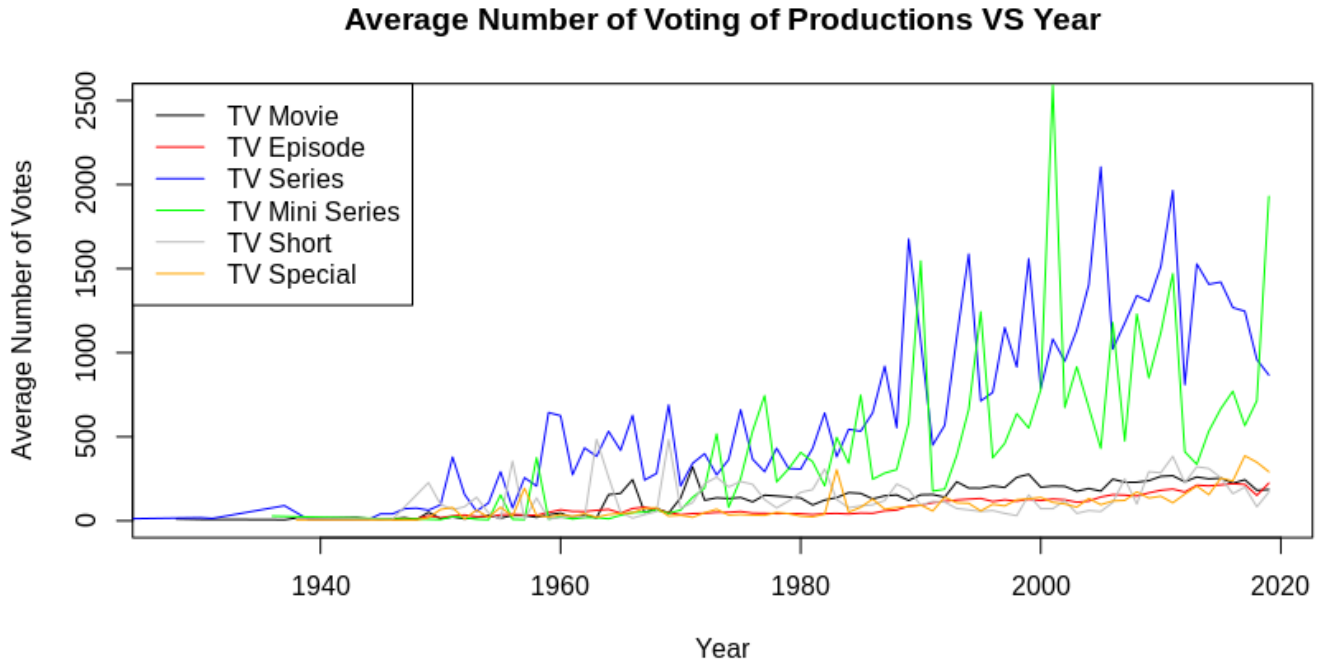

From the plot of average ratings by traditional categories over year, we can find that there is no significant changes for average ratings by each catgory over year, even if video and video games have a noticeable fluctuation in around 1960. For the difference among categories, in general, the average ratings of video and video games are slightly higher than that of movie and short. But in the recent 20 years, the average ratings of short have a significant increment.




On the other hand, in TV categories, there is also no significant changes for average ratings by each catgory over year, even if TV movie and series have a noticeable fluctuation around 1940. In general, TV Episode and Mini Series have higher ratings than others.




For audience interests part, we can find the average numbers of voting in movie category is significantly larger than that in other categories, especially in recent 20 years. From 1980-2019, there is a bell-shaped curve in movie category. Specifically, the average grows to the peak in 2000 and gradually decline. For other categories, there is no noticable changes.




For audience interests by TV categories, TV Series and TV Mini series have a significant fluctuation over year and the amplitude of fluctuations increases over year. Other categories are steady. Among different categories, TV Series and TV Mini Series gain much more attentions than other categories.

Because there are many results for question 3, we just list some result samples for question 3.

## Top 5 productions in movie category in 2018

| tconst | titleType | primaryTitle | startYear | averageRating | numVotes |
|---|---|---|---|---|---|
| tt11351186 | movie | Und vorne hilft der liebe Gott | 2018 | 10.0 | 10 |

| tconst | titleType | primaryTitle | startYear | averageRating | numVotes |
|---|---|---|---|---|---|
| tt5926500 | movie | Lumpinee | 2018 | 9.8 | 5 |
| tt7813466 | movie | Transhumance | 2018 | 9.8 | 5 |
| tt8109740 | movie | 30 KM/H | 2018 | 9.8 | 19 |
| tt9059120 | movie | Live The Stream: The Story of Joe Humphreys | 2018 | 9.8 | 6 |

### Top 10 movie people and their average rating

| nconst | AveRating | primaryName |
|---|---|---|
| nm0015147 | 8.67 | Sitki Ak<U+00E7>atepe |
| nm0464060 | 8.67 | Todor Kolev |
| nm0831262 | 8.66 | Danilo ‘Bata’ Stojkovic |
| nm0135908 | 8.62 | Toma Caragiu |
| nm0130025 | 8.6 | Puiu Calinescu |
| nm0705880 | 8.6 | Dem Radulescu |
| nm0304261 | 8.53 | Gemini Ganesan |
| nm0017988 | 8.5 | Mija Aleksic |
| nm0904303 | 8.48 | Pavle Vuisic |
| nm0218849 | 8.47 | Aleksandr Demyanenko |

# Weakness & Future Work

In our analysis, these followed problems maybe exist.
1. Whether the numbers of voting can represent well audience interests. The numbers of voting represent that the audience has watched the production and rated it, so it is sentive to time and other variables. For example, for a just released movie, people have less chance to access it than the old movie, and perhaps they just go to cinema but not watch online. So it is possible that a moive of large numbers of voting is not because of more interests from the audience but more ways to access it.
2. There are too many latent variables to influence whether a production or a movie people is good, and they also might make an influence on each other. So we need a more data and further research to get conclusion who is best production or movie people.

Next, we will work on these issues in the future.
1. To find data about box office, it will be useful to evaluate cast and production.
2. The data on IMDb is real-time updated and refresh dayly, so we can run the script everyday and the lastest result.
3. To find the trend of changes in TOP cast and production, deep into theirs features and find what features

create a good cast or production.
4. Do the regression and test on related features, and find more significant features to get a more powerful interpretation on our question.

# Conclusion

From our results, we can make these conclusion as followed.
1. For most categories, ratings and audience interests don’t change much over time.
2. Movie is always dominated among the audience but most movies in the market do not have a higher ratings.
3. Modern entertainment and production, like video games and production by TV categories are more and more popular in recent year, and at same time, they also get higher ratings.