615 Final Project Report

Wendy Liang

12/13/2020

Introduction

In this project, I will try to gain some insights into the movie industry.

I divide my work into four parts:

- Data Collection
- Data Cleaning And Organization
- Exploration And Findings
- Shiny App

Dataset

Data Description

My dataset contain metadata for all 8,095 movies in the TMDB Top Rated movie Database (before 12/12/2020). The data points include as the following:

- **budget**: The budget of the movie in dollars.
- **genres**: A stringified list of dictionaries that list out all the genres associated with the movie.
- **id**: The ID of the move.
- **original_language**: The language in which the movie was originally shot in.
- **original_title**: The original title of the movie.
- **overview**: A brief blurb of the movie.
- **popularity**: The Popularity Score assigned by TMDB.
- **production_companies**: A stringified list of production companies involved with the making of the movie.
- **production_countries**: A stringified list of countries where the movie was shot/produced in.
- **release_date**: Theatrical Release Date of the movie.
- **revenue**: The total revenue of the movie in dollars.
- **runtime**: The runtime of the movie in minutes.
- **spoken_languages**: A stringified list of spoken languages in the film.
- **title**: The Official Title of the movie.
- **vote_average**: The average rating of the movie.
- **vote_count**: The number of votes by users, as counted by TMDB.

Data Collection

I gain these data from the TMDB API. I use `rjson`, `RCUrl` and `httr` package to scrap in R. The first step is to scrap the top-rate movie dataset, getting the tmdb_id list. Then, use the tmdb_id list as the parameter to re-scrap more information on the TMDB Movie Detail Database. The data collection R file is in github TMDB API.R.

Data Cleaning

Variables `genres`,`production_companies`,`production_contries`,`spoken_languages` are in JSON format. It's hard to use directly in R. So I use both Excel and R to transform them.

In R, I use `str_extract`, `str_c`, `str_split` functions in `stringr` package to extract values with regular expressions.

So the final format is:

genres	production_countries	production_companies	spoken_language
	United States of America	PassionFlix,,,,,,,,,,,,,,,,,,,,,,,,	English,,,,,,,,,
Romance		PassionFlix,,,,,,,,,,,,,,,,,,,,,,,,	English,,,,,,,,,
Romance	United States of America	,,,,,,,,,,,,,,,,,,,,,,,,	English,,,,,,,,,
Drama,Crime	United States of America	Castle Rock Entertainment,,,,,,,,,,,,,,,,,,,,,,,,	English,,,,,,,,,
Drama,Crime	United States of America	Paramount,Alfran Productions,,,,,,,,,,,,,,,,,,,,,,,,	English,Italian,Latin,,,,,,

Exploration And Findings

Figure 1: Movie Production Country Map

Country

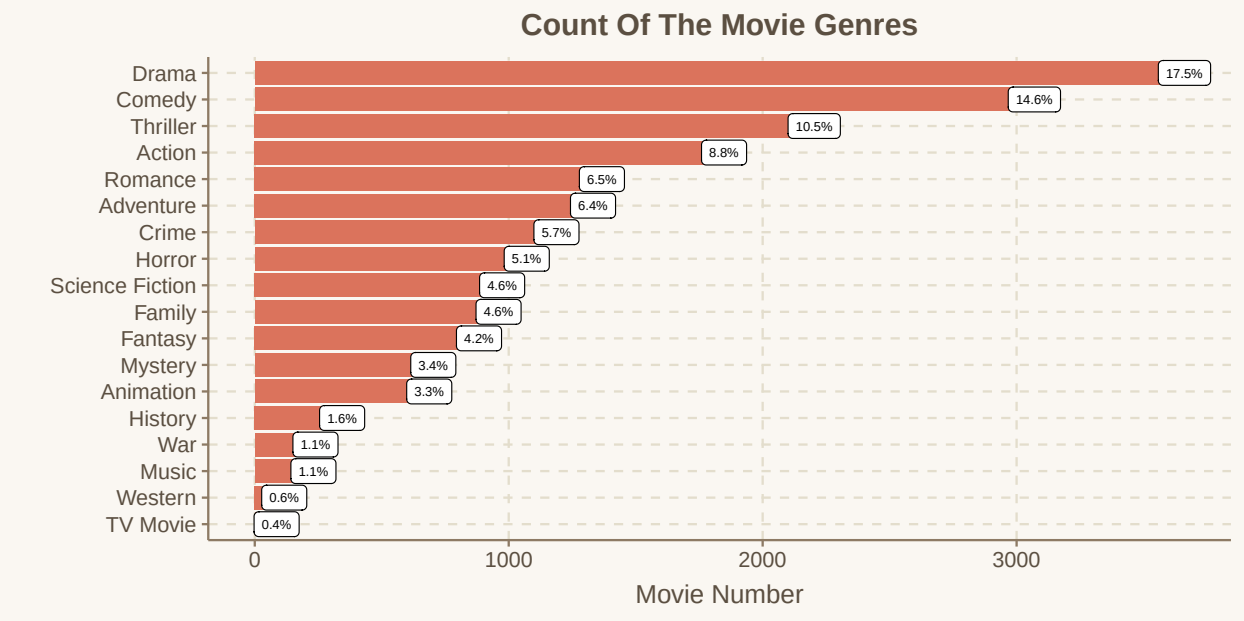
- As we all known, the **U.S** is the biggest film-making country in the world. In the TMDB Top Rate Movie Dataset, there are 5884 top-rated movies from the U.S.
- The film industry in **Europe** has also well developed. In particular, **U.K.** and **France** are the two most popular film-making countries in the world, just smaller than the U.S..

- In **Asia**, movies produced from **Japan** and **China** are the two biggest good film-making countries, followed by **India**.

Genre

1. Which are the most commonly occurring genres?

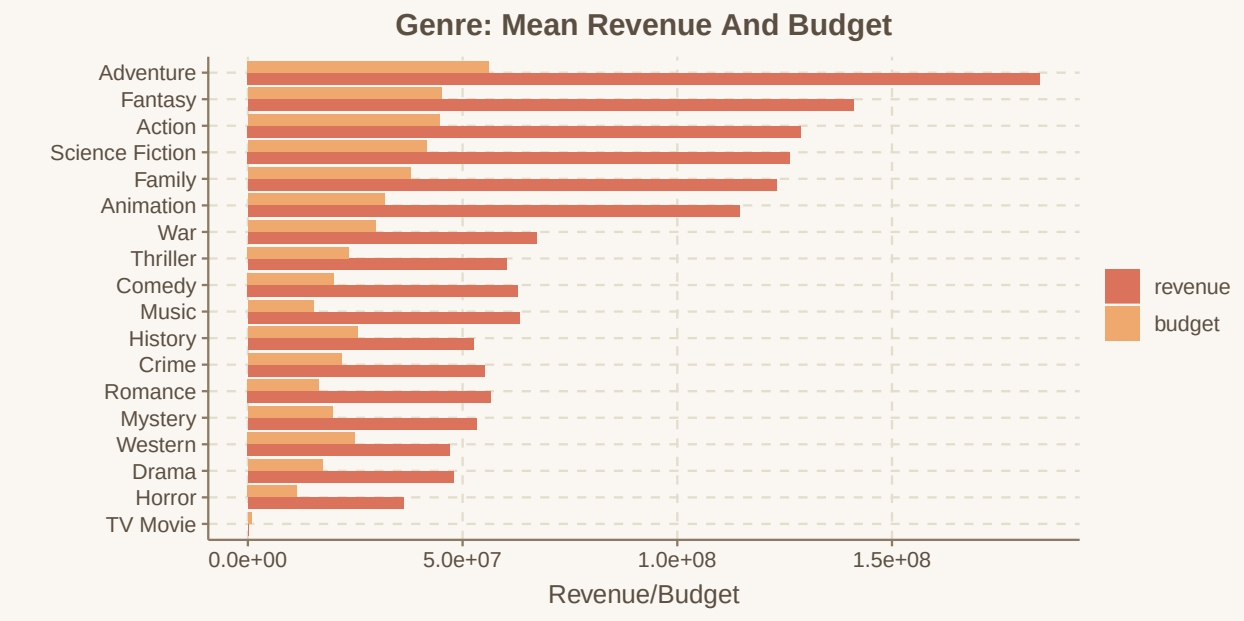
- **Drama** is the most genre with 17.5% proportion among all the top-rated movies.
- **Comedy** is the second most genre with 14.6% proportion and **Thriller** is the third most genre with 10.5% proportion.
- The top 10 genres are Drama, Comedy, Thriller,Romance, Action, Horror, Crime, Adventure, Science Fiction and Family.

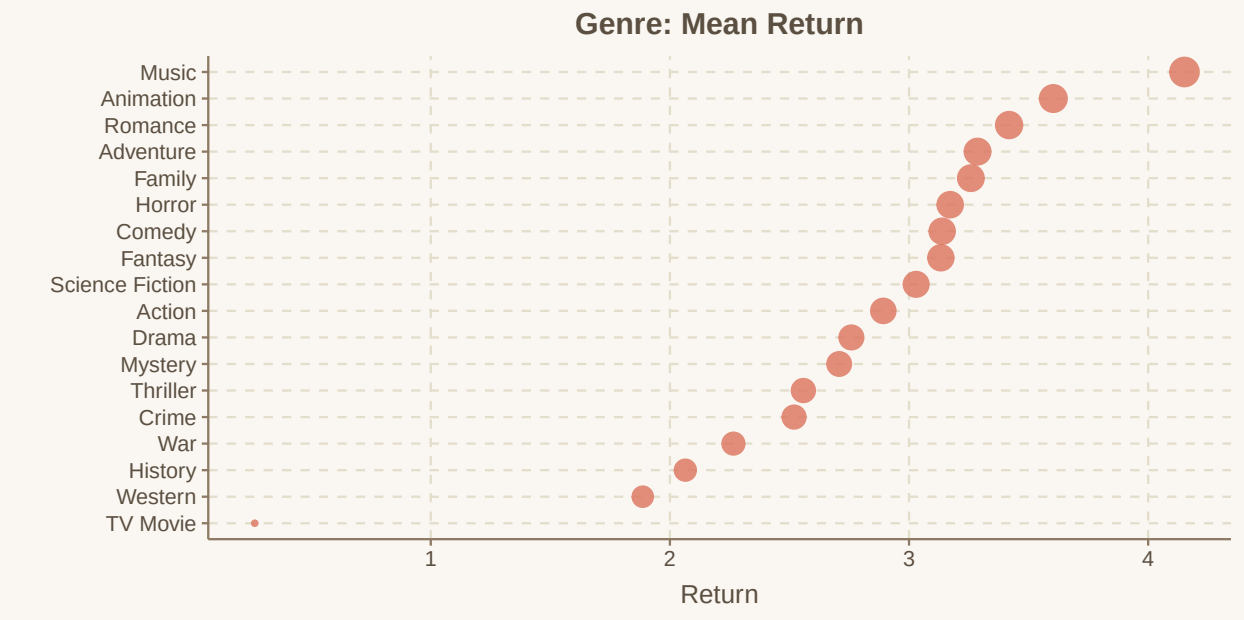


2. Which genres have the highest revenue? From this bar plot, it seems like large difference of revenue and budget between all the genres.

- **Adventure** and **Fantasy** movies have the highest revenue and budget.
- **Documentary** and **Foreign** movies have the lowest revenue and budget.
- Return(revenue/budget) indicates which kind of movies are profitable. Among the top-rated movies, **Music** is the most profitable genre, whose return is lager than 4. **Animation** and **Romance** rank second and third profitable genres. **Western** has the least return, followed by **History** and **War**. I think these kinds of movies need many grandeur that take too many money.

P.S. Since our dataset describes the top-rated movies, I guess the returns are lager than the average value of the industry.

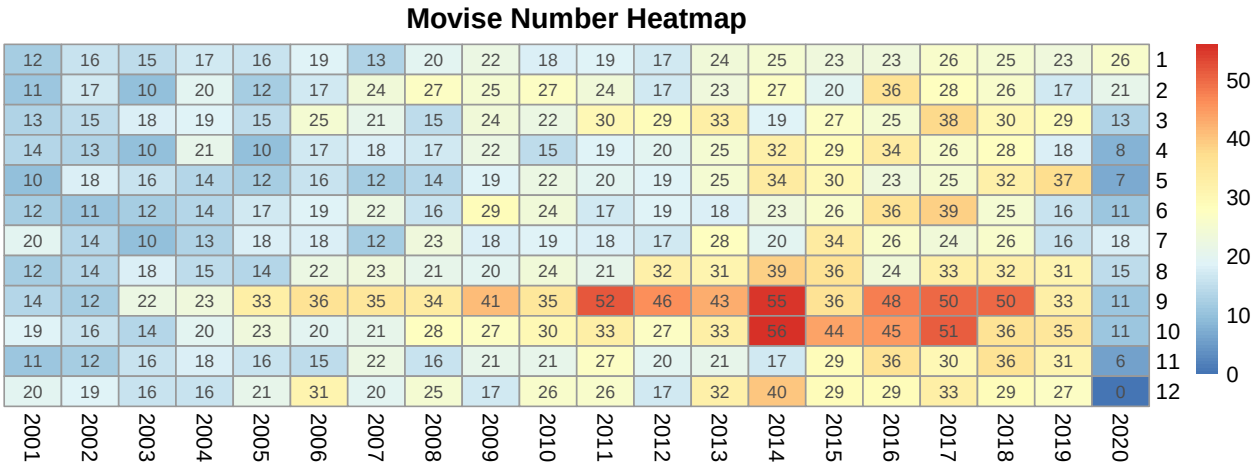




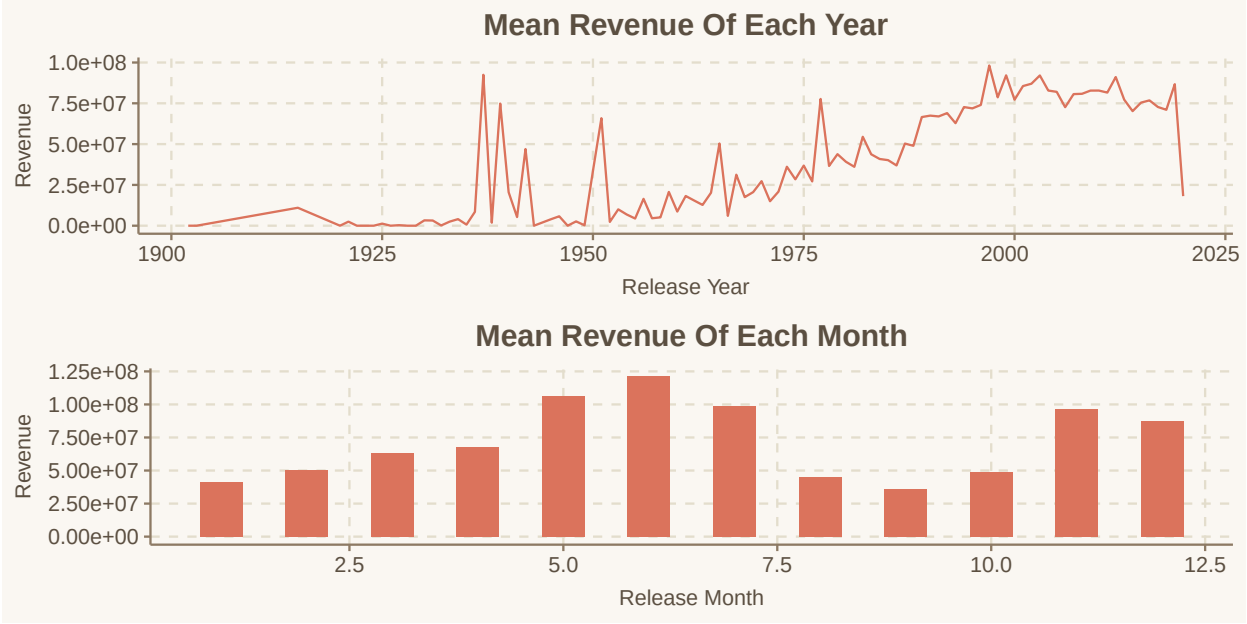
Year & Month

1.Which years and months released most movies?

- The number of released movies are increasing in the recent 20 years, which indicates that the film industry in the world is thriving.
- It seems like summer is a popular movie season since there are many movie released from **August** to **October**.



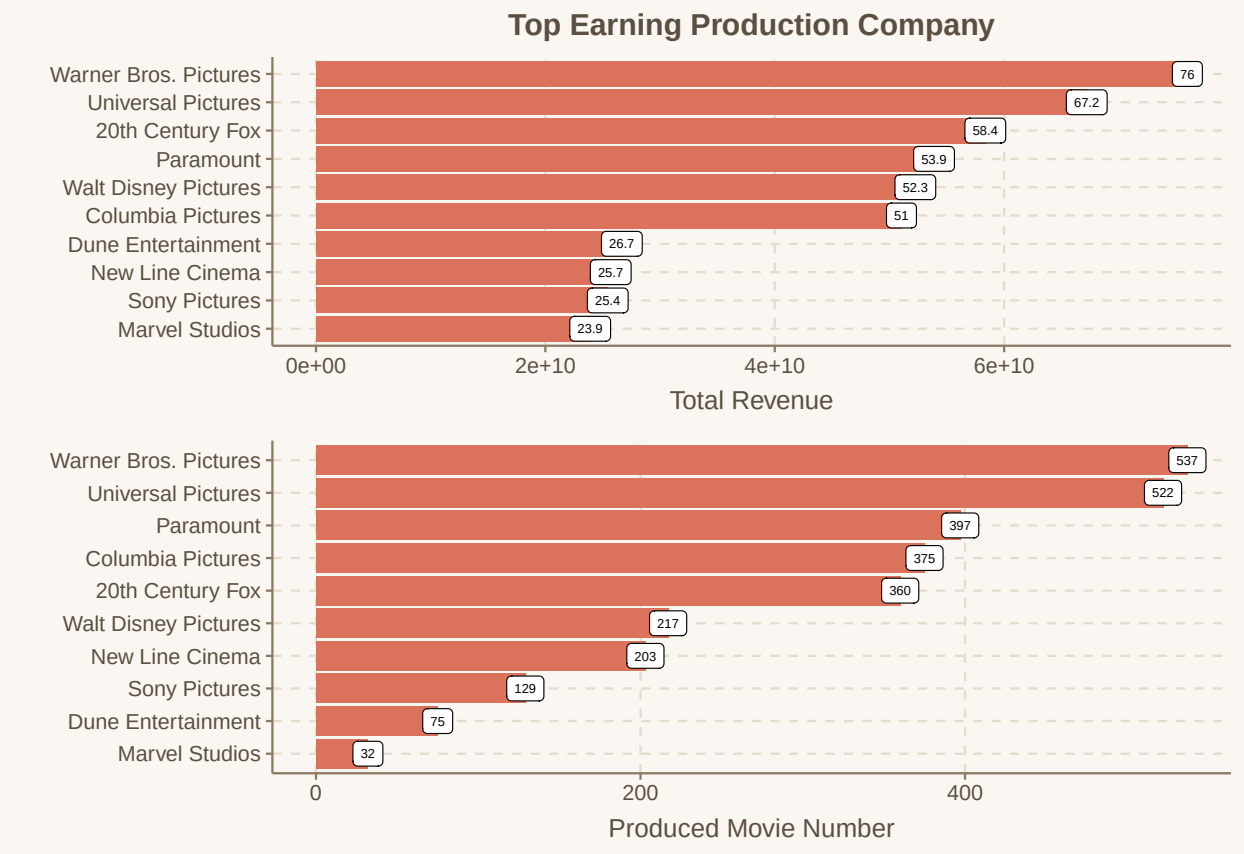
2.Which years and months have the highest revenue?



- During the very early period , the revenues are in very low level. It's easy to understand, Surprisingly, there are several climax: 1937,1939,1942,1951 before 1950s. **Why?** I guess some outstanding movies were released in these years which made the average revenue higher. So, it cannot represent the entire movie revenue case at that time.
- After 1950s, the average revenue has an increasing trend among years. After the 20th century, the annual average revenue has stayed stable.
- Now, we discuss the monthly average revenue. From the second plot, it seems like **May** to **July** have the highest average revenue. This can be attributed to the fact that **blockbuster movies** are usually released in the summer when the kids are out of school and the parents are on vacation and therefore, the audience is more likely to spend their disposable income on entertainment. Moreover, these event movies are always high rated by audiences.

Production Company

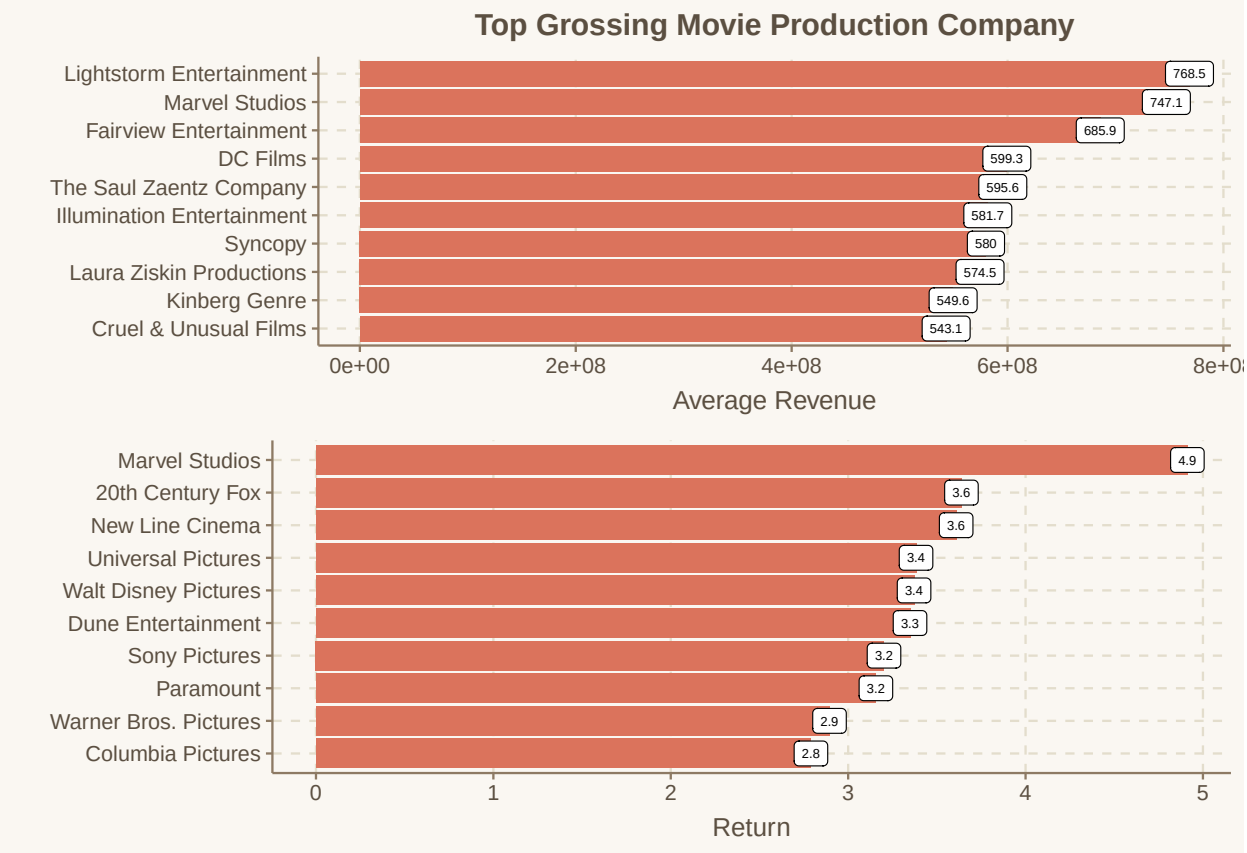
1. Which companies earn the highest total revenue?



- Undoubtedly, **Warner Bros** is the highest revenue earning movie production company among the most 50 successful companies. It has earned $ 76 billion from 537 movies.
- **Universal Pictures** and **20th Century Fox** win the silver and bronze medals respectively, with $ 67.2 billion from 522 movies and with $ 58.4 billion from 360 movies.

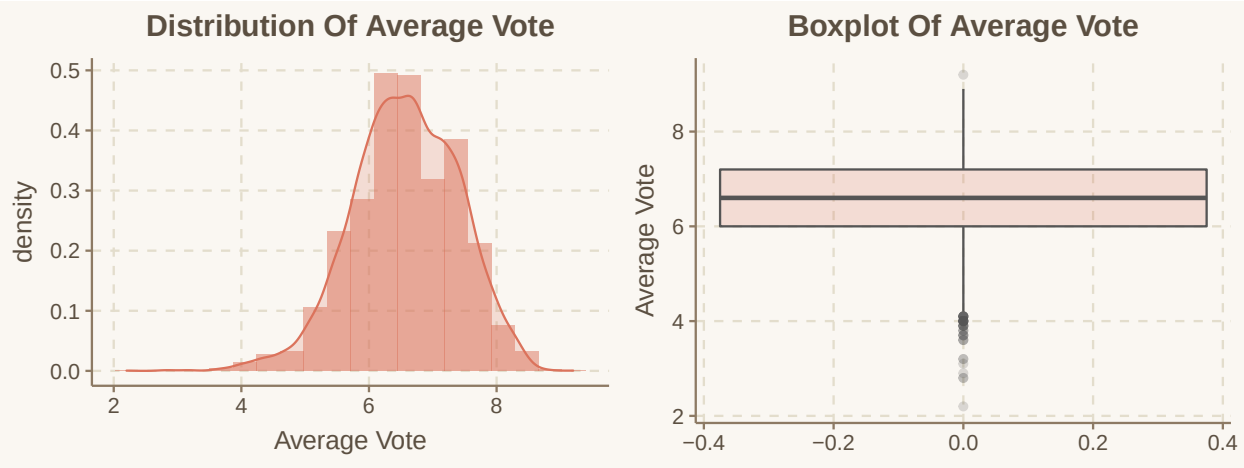
2. Which companies made the highest-revenue movie?

- Movies made by **Lightstorm Entertainment** has the highest revenue – $ 768.5 million. Movies made by **Marvel Studios** has the second highest revenue – $ 747.1 million. We can also find familiar companies like **DC Films** that has good grossing performance.
- Movies made by **Marvel Studios** has the highest return of investment – 490%. Movies made by **20th Century Fox** and **New Line Cinema** has the second highest return of investment – 360%.



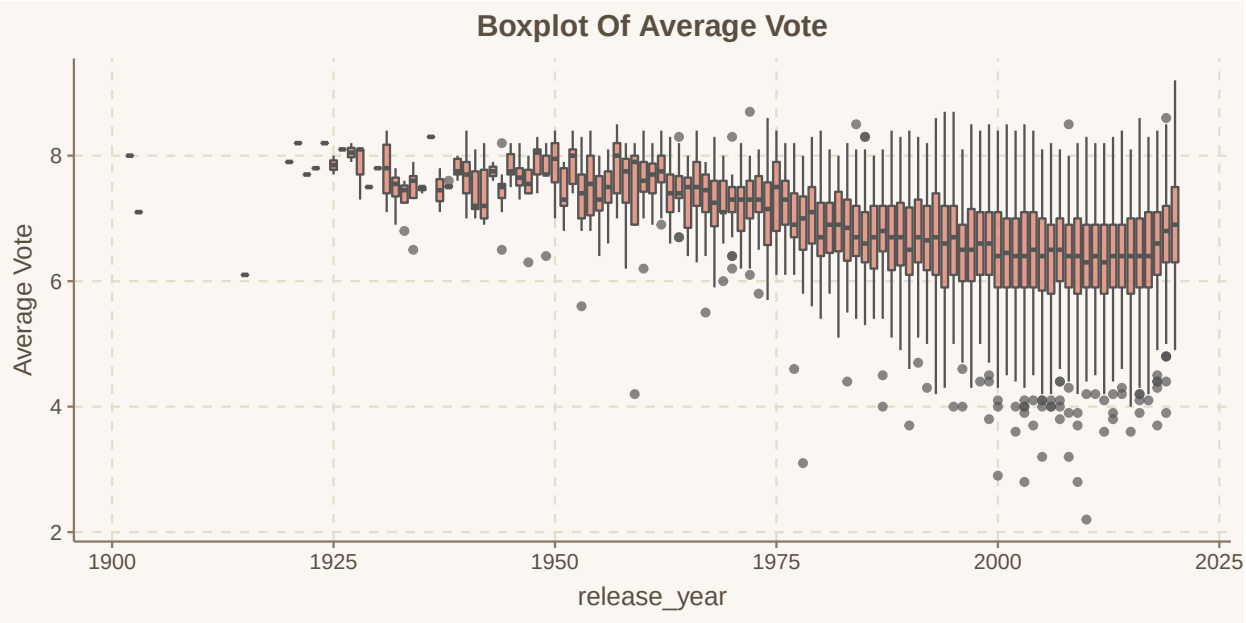
Popularity, Vote Average and Vote Count

1.How the vote scores distribute? This density plot shows the distribution of TMDB users' average vote to the TMDB Top Rate Movie Dataset. It seems like most vote scores concentrates on **6.4** to **7.6**. Few top-rated movies are voted below **4.8** and beyond **8.4** scores.



2.Dose Users get more strict along with time? From this plot, it seems like the overall rating trend is decreasing among the years. The annual average vote has decreased from 8 to 6 scores until 2017 and has increased a little form 2018.

It's hard to say the audiences get more strict to assess a movie. There may be another explanation. There are less movies in the early year, but the percentage of high-rated movies is high. For these old movies, the TMDB users (modern people) can just notice and vote the outstanding ones while the low-quality movies are always ignored. In this case, the much higher average vote of old movies is reasonable.



3.What are the top reputable movies? Firstly, I filter the vote count that more than 2000. If the users who voted for a movie are too little, maybe the rating is not meaningful. Below is the top 20 reputable movies table:

production_countries <chr>	title <chr>	vote_average <dbl>	release_year <dbl>
United States of America	The Shawshank Redemption	8.7	1994
United States of America	The Godfather	8.7	1972
India	Dilwale Dulhania Le Jayenge	8.7	1995
United States of America	Schindler's List	8.6	1993
Japan	Your Name.	8.6	2016
United States of America	The Godfather: Part II	8.6	1974
Japan	Spirited Away	8.5	2001
South Korea	Parasite	8.5	2019
United States of America	The Green Mile	8.5	1999
United States of America	Pulp Fiction	8.5	1994

Not surprisingly **The Shawshank Redemption** and **The Godfather** are the two highest rated movies. In fact, these two famous movies are in all kind of Top Movies Lists in the world.

It seems like the top rated movies are most old movies before 20th century. There are only two movies – **Your Name** and **Parasite** released after 2010 rank the top 20. I believe many people are familiar with these two phenomenal movies from Asia. They are directed separately by the most famous directors in Japan and South Korea – Makoto Niitsu and Bong Joon ho. By the way, **Parasite** won the Oscar last year.

Shiny App

After exploring the dataset, I build an interactive web appby Shiny to present our data and plots more visually.

On the dashboard,there are two menu items: the first one is *Movie Detail Table* with one table box, the second one is *Movie Genre Plot* with two plot boxes.

On the side bar, users can select the year, country and genre.

Here is the link to my app: Final Mapping Shiny Link.

Reference

- R shinydashboard example
- TMDB API Doc
- Kaggle full movie dataset (2017)

Appendix

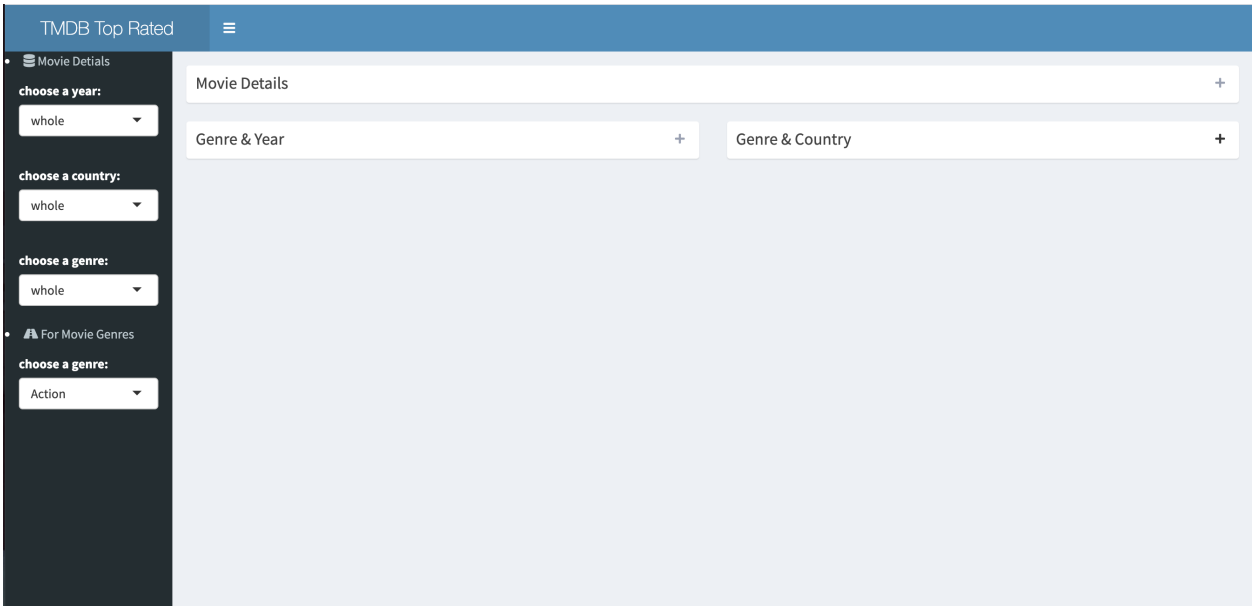


Figure 2: Whole Interface

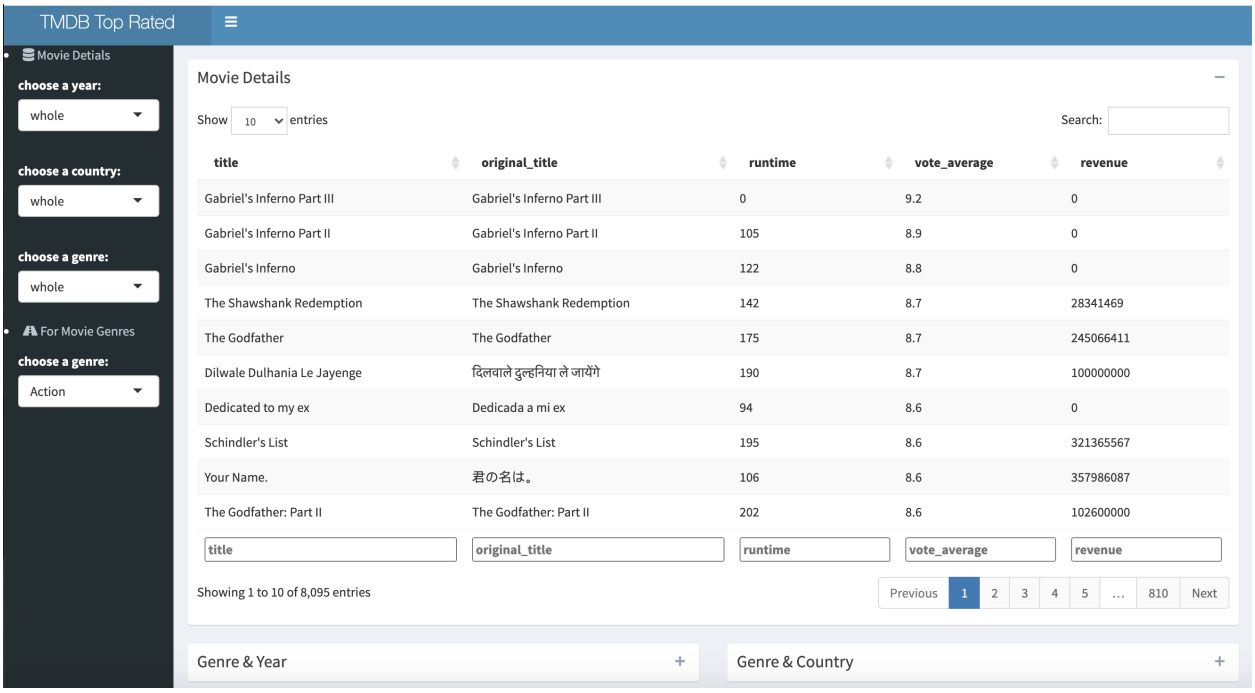


Figure 3: Interface 1

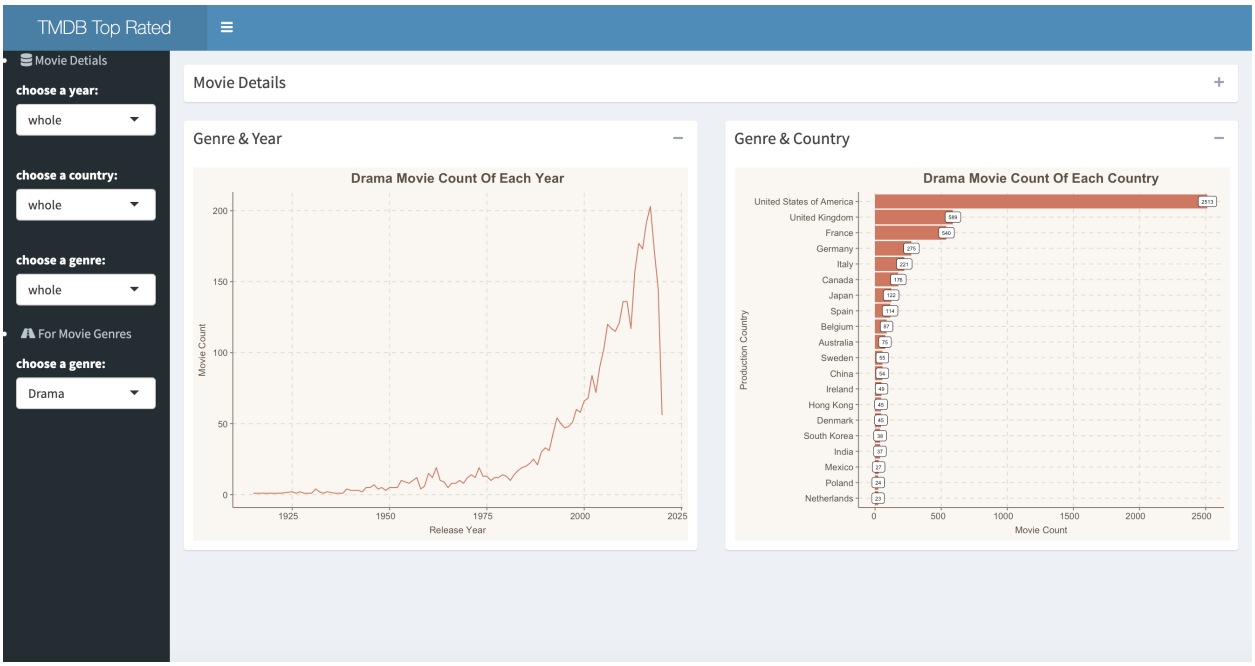


Figure 4: Interface 2