DATA422 Group Project Report

Play Around with Movie Data

Author

Ahn Joo-Hyun (84701204)

Moon Changmin (72452116)

Komlev Viacheslav (98419435)

Jia Zhiying (54801975)

Introduction

The purpose of the project is about the wrangling and insights of movies data. After decades of growth, the film industry has developed into a huge market. The global film industry shows healthy projections for the coming years, as the global box office revenue is forecast to increase from about 38 billion U.S. dollars in 2016 to nearly 50 billion U.S. dollars in 2020 (statista.com, 2018). It is commonly believed that big companies, famous directors and big-budget productions will produce a successful movie at the box office. However, whether these ideas are supported by scientific evidence and what other factors could contribute to the success of a movie. In this project, we collated and analysed a large amount of movie data and found some interesting statistical rules. These findings are not only the results of the project but also have particular reference significance for investors in the film industry.

The data sources we used

- We used IMDb data as the movie basic information library. Each dataset is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file includes headers that describe what is in each column. A '\N' is used to denote that a particular field is missing or null for that title/name (Figure 1).

  Source Link: https://datasets.IMDbws.com/

  The available datasets are as follows:

  title_basics - Contains the following information for titles (IMDb.com, 2019):

  tconst (string) - unique alphanumeric identifier of the title

  titleType (string) - the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)

  primaryTitle (string) - the more popular title / the title used by the filmmakers on promotional materials at the point of release

  originalTitle (string) - original title, in the original language

  isAdult (boolean) - 0: non-adult title; 1: adult title

  startYear (YYYY) - represents the release year of a title. In the case of TV Series, it is the series start year

  endYear (YYYY) - TV Series end year. '\N' for all other title types

  runtimeMinutes - primary runtime of the title, in minutes

  genres (string array) - includes up to three genres associated with the title

title_ratings - Contains the IMDb rating and votes information for titles

tconst (string) - alphanumeric unique identifier of the title

averageRating - weighted average of all the individual user ratings

numVotes - number of votes the title has received

attributes (array) - Additional terms to describe this alternative title, not enumerated

isOriginalTitle (boolean) - 0: not original title; 1: original title

title_principals - Contains the principal cast/crew for titles

tconst (string) - alphanumeric unique identifier of the title

ordering (integer) - a number to uniquely identify rows for a given titleId

nconst (string) - alphanumeric unique identifier of the name/person

category (string) - the category of job that person was in

job (string) - the specific job title if applicable, else '\N'

characters (string) - the name of the character played if applicable, else '\N'

name_basics - Contains the following information for names:

nconst (string) - alphanumeric unique identifier of the name/person

primaryName (string) - name by which the person is most often credited

birthYear - in YYYY format

deathYear - in YYYY format if applicable, else '\N'

primaryProfession (array of strings) - the top-3 professions of the person

knownForTitles (array of tconsts) - titles the person is known for

title_akas - Contains the following information for titles:

titleId (string) - a tconst, an alphanumeric unique identifier of the title

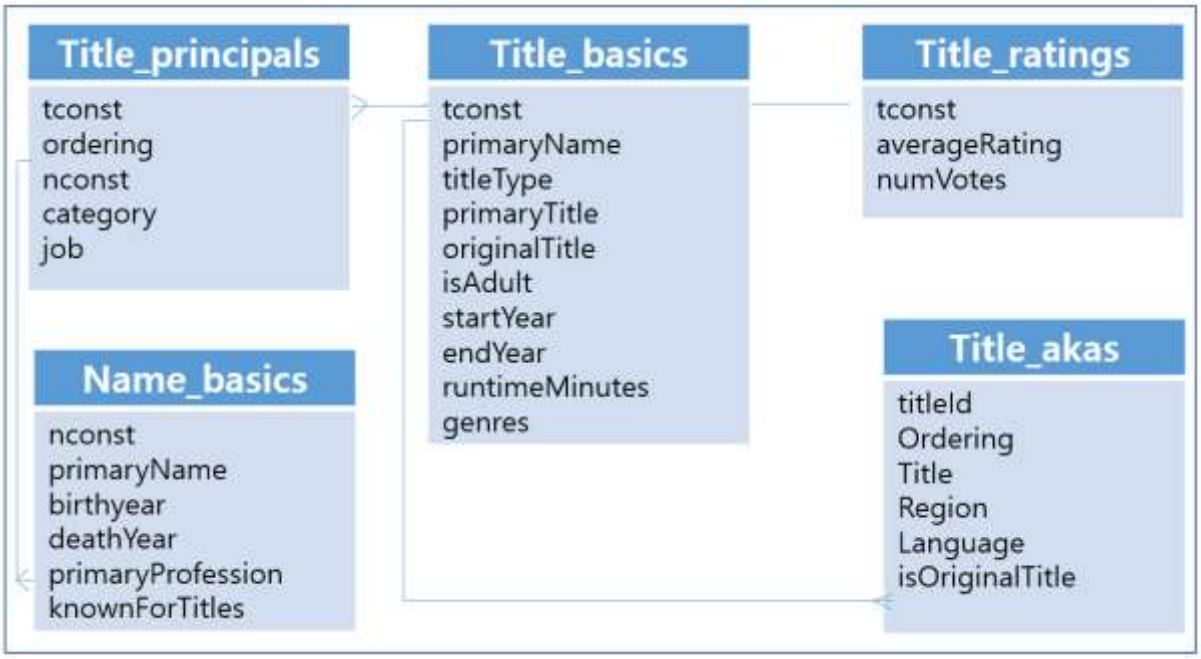
ordering (integer) - a number to uniquely identify rows for a given titleId

title (string) - the localised title

region (string) - the region for this version of the title

language (string) - the language of the title

Figure 1: Initial Data Structure of IMDb



- We extracted the box office dataset from the NUMBERS, including budget, domestic gross and worldwide gross (Figure 2).

  Source Link: https://www.the-numbers.com/movie/budgets/all

Figure 2: Initial Data Structure of The Numbers

	Release Date	Movie	Production Budget	Domestic Gross	Worldwide Gross
1	Dec 17, 2009	Avatar	$425,000,000	$760,507,625	$2,789,705,275
2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$410,600,000	$241,063,875	$1,045,663,875
3	Apr 23, 2019	Avengers: Endgame	$400,000,000	$858,373,000	$2,795,473,000
4	Apr 22, 2015	Avengers: Age of Ultron	$330,600,000	$459,005,868	$1,403,013,963
5	Dec 13, 2017	Star Wars Ep. VIII: The Last Jedi	$317,000,000	$620,181,382	$1,316,721,747
6	Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	$306,000,000	$936,662,225	$2,053,311,220
7	Apr 25, 2018	Avengers: Infinity War	$300,000,000	$678,815,482	$2,048,134,200
8	May 24, 2007	Pirates of the Caribbean: At World's End	$300,000,000	$309,420,425	$963,420,425
9	Nov 13, 2017	Justice League	$300,000,000	$229,024,295	$655,945,209
10	Oct 6, 2015	Spectre	$300,000,000	$200,074,175	$879,620,923
11	Jul 19, 2012	The Dark Knight Rises	$275,000,000	$448,139,099	$1,084,439,099
12	May 23, 2018	Solo: A Star Wars Story	$275,000,000	$213,767,512	$393,151,347

The reasons for choosing the datasets

When selecting data sources, we mainly took two aspects as the selection criteria. Firstly, the data sources should be conducive to us to practice and demonstrate the skills we have learned in this course, such as web scraping, data cleaning. For data source from IMDb, the primary information we needed for analysis is scattered among different data sets. To establish a reasonable data model suitable for our project, we needed to wrangle these data sets. In terms

of box office data, we deliberately avoided the data resources that can be downloaded directly in order to practice web scraping technology and chose to scrape from the website instead.

In the beginning, we selected the website ‘Box Office Mojo' which was a subsidiary of IMDb, as the source of box office data and completed the extraction of the data. However, the global box office records of Box Office Mojo were only 790, which we thought was not large enough for good statistical analysis. We then decided to use the global box office data from ‘The NUMBERS', which provides box office data for 5000 films.

Secondly, the data source should be convenient for us to analyse successful films. However, how to evaluate the success of a movie can be considered from many aspects, such as artistic connotation, technical level, social influence and commercial value. To simplify the problem, the project defines whether a movie is successful from the size of commercial value.

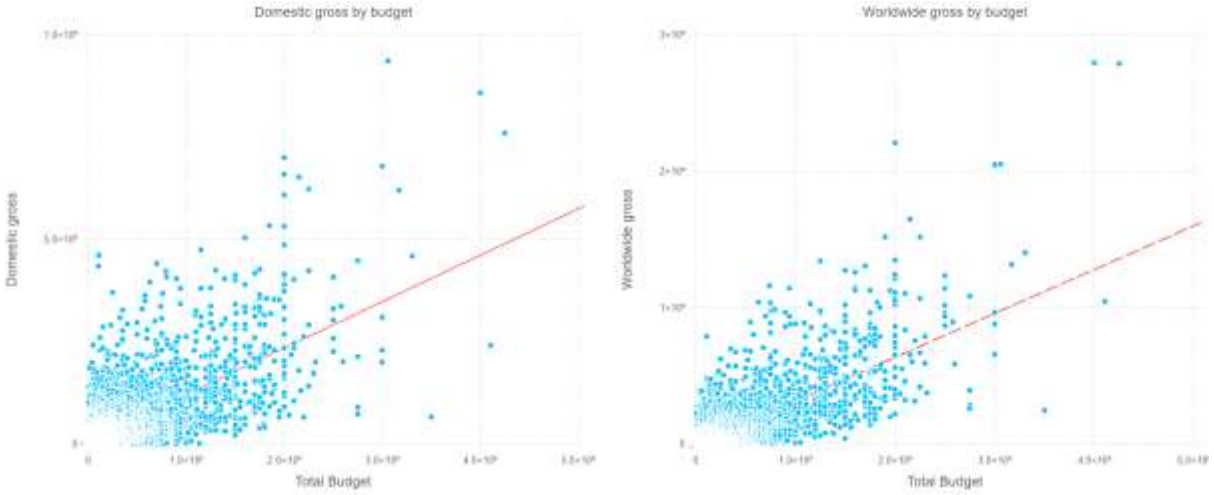
As we know, IMDb is the world's most popular and authoritative source for movie, TV and celebrity content, including ratings and reviews for the newest movie and TV shows, which could be considered as a relatively reliable data source. Besides, The NUMBERS can provide enough box office data to meet our analysis needs. Therefore, we decided to combine the data from these two sources for the next analysis.

The targets we chose

One of our goals is to discover and visualise some interesting patterns in the movie data in terms of the box office. We explored some factors like budget, gross, director, genre, release year, duration and IMDb rating score. Film investors may utilise these findings to produce a successful movie.

Is a high-budget movie sure to bring in good revenue? As can be seen from the linear regression trend line of the scatter plots (Figure 3), there is a positive correlation between budget and box office, which also explains why the investment in commercial films is getting higher and higher in recent years, apparently for better revenue.

Figure 3: Distribution of Gross by Budget

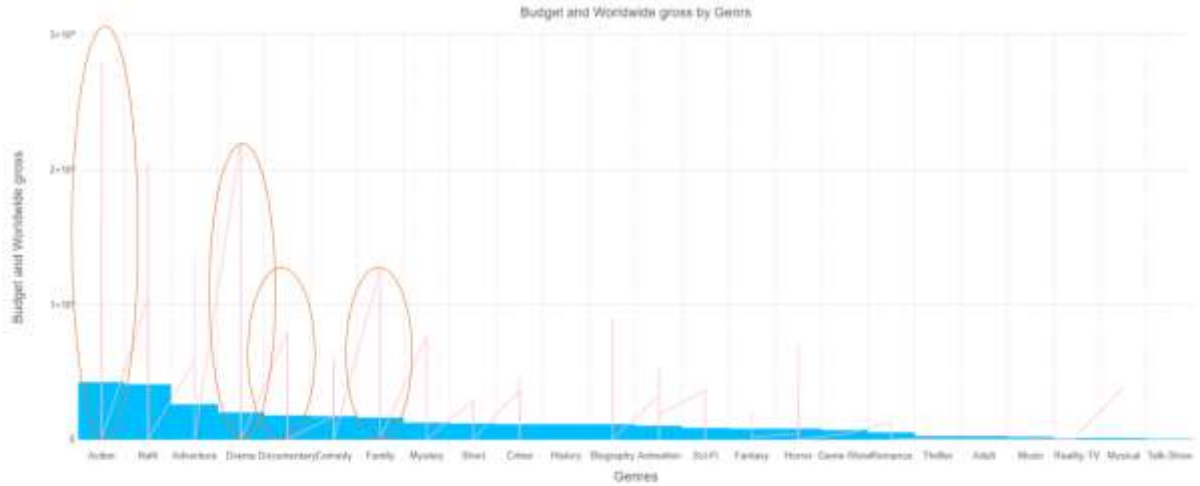


Since a higher budget can bring in the higher box office, what factors can attract a higher budget? As can be seen from the scatter plots (Figure 4), the running-time lies between 100 min and 200 min. Also, the average rating of movies lies on between 5.5 and 7.5 scores.

Figure 4: Distribution of Budget by Duration and Rating Score

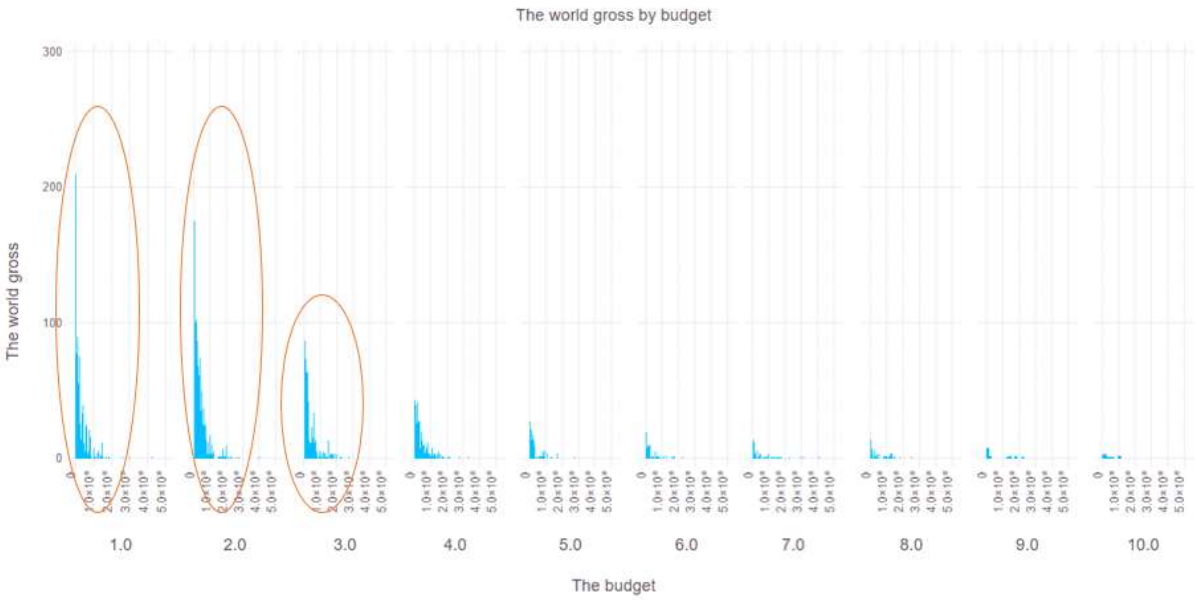
In addition, we suspected that there is a significant difference between the marketing budget and worldwide revenue by the genre to the successful movies. As a result, the action, dramas and family type of films are enjoying relatively good success in preparation for marketing budgets, followed by documentaries (Figure 5).

Figure 5: Distribution of Budget by Genre



Is there any relationship between the marketing budget and world gross by words of a successful movie title? The plot (Figure 6) has shown that the top 3 of the successful movies have below 3 words in the title.

Figure 6: Distribution of Gross by Title Words

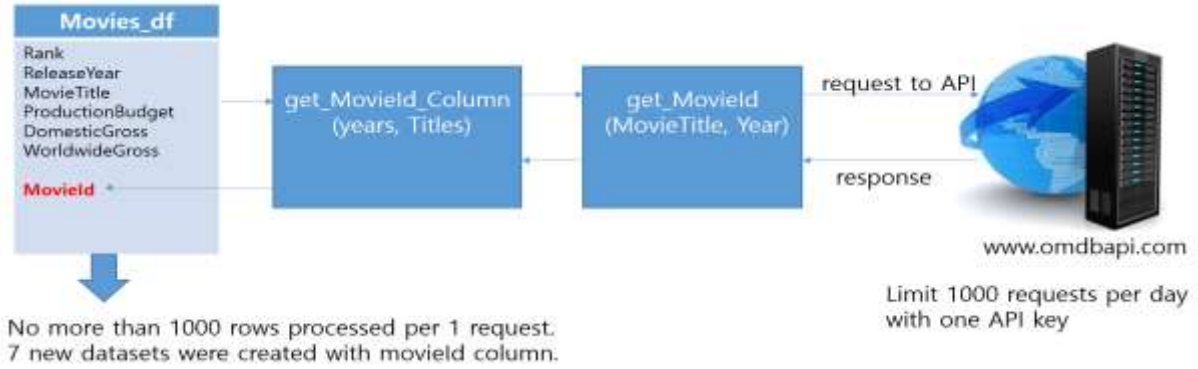


The difficulties we met

Although we have successfully obtained data from two sources, unfortunately, the dataset from the Numbers has no column with movie Id. Additionally, the column MovieTitle contains movie titles which differ from movie title in IMDb dataset. Therefore, it was impossible to join these different datasets by movie title.

In order to join two different datasets, an API function from www.omdbapi.com was used to get moviId by title and year. MovieId was retrieved from movies_df dataset by sending request to API with movie title and year when movie was issued. The function get_MovieId_Column spreads moviId value in movies_df dataset for each row by sending request via get_MovieID function. To use API service, it is necessary to get API key. While it is restricted to send no more than 1000 requests to the server API per day. To tackle this problem several API keys was activated and then used. As a result, several tables with about 1000 rows with movie ID were formed from movies_df dataset. Then, they were combined together to get one dataset with financial information (Figure 7).

Figure 7: Movie Id Retrieval



The techniques we used

In our project, the following data wrangling techniques was implemented:

- Data were checked by using R functions glimpse, problems, kable
- Missing data and NA values were excluded so that it could not affect the main idea
- Data were scrapped by using httr library and its functions. Also API function was implemented
- Converted date data in characters into ‘yyyy-mm-dd’ format by using ludridate package
- Data were partly filtered by years
- Reshaping was implemented (gbind was used to unite several dataframes)
- Mutate was used for creating new columns
- Columns with several values were separated (separate) and gathered(gather) into new tables. For instance, column Genres.

Figure 8: Movie Data Wrangling

tconst	1	2	3
<chr>	<chr>	<chr>	<chr>
tt0015724	Drama	Mystery	Romance
tt0023331	Documentary	NA	NA
tt0035423	Comedy	Fantasy	Romance

tconst	genreId	genre
<chr>	<chr>	<chr>
tt0015724	1	Drama
tt0015724	2	Mystery
tt0015724	3	Romance

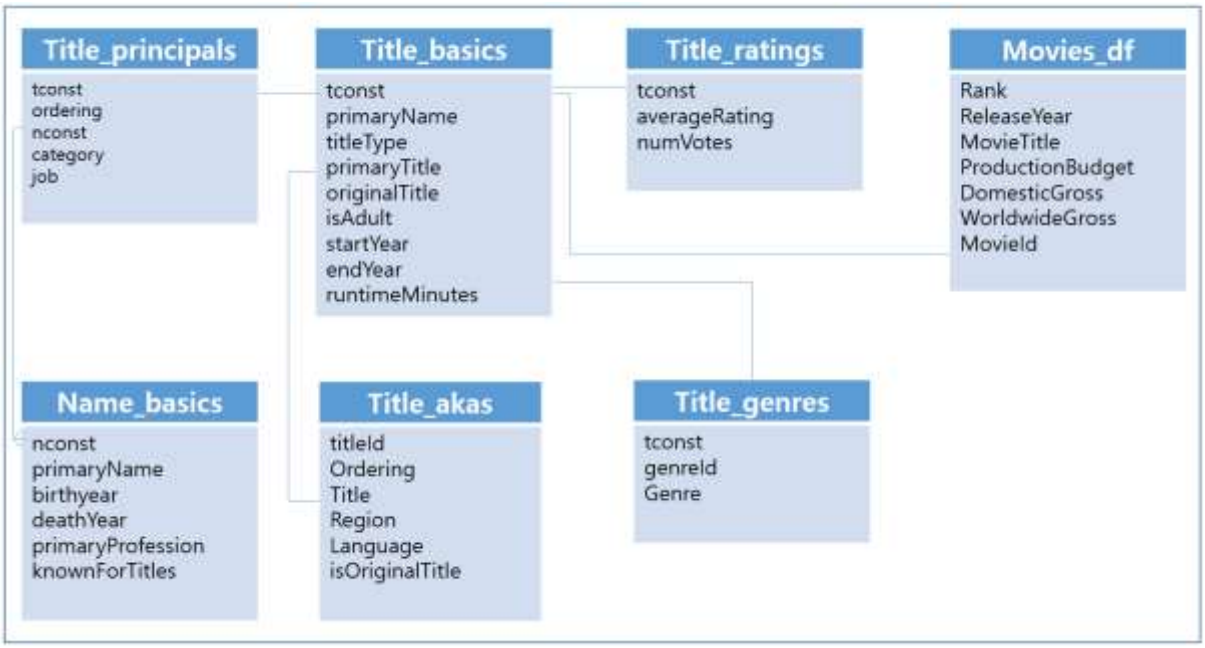
Achievement and limitations

After data scraping, wrangling and analysis, we finally built a reasonable data model (Figure 9). By retrieving MovieId, we can connect different datasets to be analysed. IMDb datasets contain millions of records, and it was a challenge to work with all rows. The dataset was jointed with financial data, which reduced the number of rows, other datasets were obtained by using joint functions. Then, a tidy and joined dataset was used for making plot and analysis.

Furthermore, we found some interesting relationships between movie features, and we could make some investment recommendations for the film market based on these findings, such as investing the blockbuster movies or the movies with no more than three hours duration. Action, adventure, drama, family and documentary are relatively lucrative types of film.

However, the project also has some limitations. Due to time constraints, we only extracted some major movie elements, but a movie will involve many aspects. To better analyse the factors that affect the success of a movie, we should also add the main actors, cultural background and other factors. Also, we would extract the main content of the movie and carry out corpus analysis to determine what kind of content is more attractive to the market.

Figure 9: Data Model



References

Amy Watson. (Dec 19, 2018). Statista. *Film Industry - Statistics & Facts*. Retrieved from https://www.statista.com

IMDB. (2019). IMDB. *IMDb Datasets*. Retrieved from https://www.IMDb.com/interfaces/

DataFrames.jl. (n.d.). Retrieved from https://juliadata.github.io/DataFrames.jl/stable/

Gadfly.jl. (n.d.). (n.d.). Retrieved from http://gadflyjl.org/stable/

PackageCompiler.jl. (n.d.). Retrieved from https://github.com/JuliaLang/PackageCompiler.jl

GLM.jl. (n.d.). Retrieved from https://github.com/JuliaStats/GLM.jl

Package 'lubridate'. (Apr 11, 2018). Retrieved from https://cran.r-project.org/web/packages/lubridate/lubridate.pdf