**Predicting Box Office Success through Indicator analysis**

20190196 김태형

20190361 안선호

20190573 전재완

## Abstract

Using the R program, we are planning to process movie metadata from kaggle to analyze the relations between several indicators. Our data wrangling consists of three parts. First part is data importing, which is done by using read.csv(). Second part is ‘tidy-ing’, which is transforming themassive metadata into a simple tibble form. Third part is transforming, which is aimed to remove ‘Not Available’ data in the tibble and create a new variable named ‘Earnings Rate’. After the wrangling, we are going to check three relations: genre - earnings rate, budget - revenue, budget - ratings. Our current expectation is that there will be a linear relationship between these factors, but after the first analysis we are planning to process data such as removing outliers and changing the expectation model to find the most accurate model. Through this project we can anticipate what kind of movies will be popular, and predict several indices of movies(e.g. ratings, revenue) to be released.

[**Abstract**](#_m89r6cddhnac) **1**

[**1. Introduction**](#_cfq0iwo3mjt5) **2**

[**2. Data Description**](#_v6eo9ocylord) **2**

[**3. Data Wrangling**](#_18qfkswpbqm) **3**

[**4. Data Analysis**](#_gx94mh7ad607) **3**

[**5. Expectation**](#_2sx9t61himsw) **4**

## 1. Introduction

By the 21st century, the film industry had developed significantly. Despite the decline in movie theater sales due to COVID-19, people watch various movies using movie platforms such as ‘Netflix’ and ‘Watcha’. The film industry, which is receiving great attention, continues to change and develop. Genres that were not previously spotlighted suddenly become popular (e.g. Avengers of Marvel Studio), and new genres such as SF and fantasy are made. Also, film companies are willing to produce good films at a cost of enormous budget (Avengers: Endgame's production cost is a whopping $1.56 billion).

It is important to know which movies are popular in this fashionable situation. Therefore, we would like to look at various indicators related to movies to find out what kind of movies are popular. It will be analyzed using the movie metadata in Kaggle, and the main direction of the study is finding factors that affect the ratings of movies, and finding out the factors of each movie which affect the movie’s revenue through data analysis using R.

## 2. Data Description

Before analyzing the data, We would like to introduce what items are in the dataset.

* **movies\_metadata.csv:** The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.
* **keywords.csv:** Contains the movie plot keywords for our MovieLens movies. Available in the form of a stringified JSON Object.
* **credits.csv:** Consists of Cast and Crew Information for all our movies. Available in the form of a stringified JSON Object.
* **links.csv:** The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset.
* **links\_small.csv:** Contains the TMDB and IMDB IDs of a small subset of 9,000 movies of the Full Dataset.
* **ratings\_small.csv:** The subset of 100,000 ratings from 700 users on 9,000 movies.

Among them, we will mainly use 'movies\_metadata' and 'ratings\_small'.

## 3. Data Wrangling

Data Wrangling consists of three parts. Import, Tidy, and Transform.

* Import

Using the read.csv() function, we are going to read movies\_metadata.csv into R with the option *header = TRUE*.

* Tidy

We are going to use the “tidyr” package to create tidy data. From the data we imported above. Then we’ll transform it into a tibble using the as\_tibble() function so that we can use the “dplyr” package when we transform the data.

* Transform

movies\_matadata consists of 24 columns and 45455 rows of data. We are going to use the “dplyr” package to transform data.

The column that we need is budget, genres, revenue, runtime, vote\_average and vote\_count. So using the select() function, we can pick only budget, genres, revenue, runtime, vote\_average and vote\_count.

Then we need to remove data that we can’t use. Using the filter() function, we are planning to remove movies that include NA as value. Then using the mutate() function, produce a new variable that represents “earnings rate” which can be calculated by revenue/budget.

Also we need to make a new table consisting of genre and earnings rate. One movie can have multiple genres. For example, genres data for movie “Toy Story” is [{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': 10751, 'name': 'Family'}]. So “Toy Story” should be calculated for each Animation, Comedy, and Family. By running a loop for the table we made above, make a new table that would have three rows(Animation, Comedy, Family) for Toy Story.

## 4. Data Analysis

We are planning to create models for each genre - earnings rate, budget - revenue, budget - running time. Currently before making the data, we are expecting a linear relationship between each factor, but this may be wrong. From the first data we extracted, we are planning to remove the outliers and change the model from linear to other to get the correct model for these relations.

## 5. Expectation

Before the analysis, we expected the result to establish some hypothesis such as the following table.

|  |  |
| --- | --- |
| Relation | Expectation(Hypothesis) |
| genre - earnings rate | The genre which appears with high frequency will have a high earning ratio. |
| budget - revenue | The movie which has more budget will earn more. |
| budget - ratings | The movie which has more budget will receive good ratings |

*Table 1. Hypothesis of Project*

Explanations of these three expectations are like below:

1. **genre-earnings ratio**

Movie makers will make high earnings ratio movies so that they can earn more money.

1. **budget - revenue**

It is an easy expectation that movie makers will pay more money just to earn more.

1. **budget - ratings**

To get good ratings, movie makers will pay more money.

In these three relations: genre - earnings ratio, budget - revenue, budget - ratings, we’ll prove or modify our hypothesis by future analysis.