Movies Popularity Prediction

Data Analytics
Prof. Uday C. Menon
Fall B, 2020

Group: Data Decipherment

Group Members:
Yilin Liu (yl4585), Savannah Wang (rw2840),
Dongyang Yin (dy2416), Zongqian Wu (zw2679)

Introduction

The dataset is an ensemble of data collected from TMDB and GroupLens. The files contain metadata and credits for all 45,000 movies listed in the Full MovieLens Datasets, including cast, release dates, genres, etc.

This project includes three steps. First, we cleaned the dataset by deleting missing values, converting some attributes to binary variables, etc. Second, we analyzed and visualized the dataset so that the data is shown directly through plots. Finally, we created several models and chose the most accurate model to represent our results, and made further conclusions.

Data Preprocessing & Cleaning

The dataset on Kaggle has a large size with very detailed information about each movie. To predict movie popularity efficiently, we browsed the whole dataset and decided to focus on analyzing movies_metadata.csv and credits.csv, which contain information that is highly related to popularity. We first worked on each dataset separately.

Movies Metadata

We extracted 11 columns for later data processing (see [Appendix 1a] for information of columns). We dropped data with duplicate values and wrong inputs. When exploring our dataset, we detected that around 25% of the movies don't have information about genres. Considering that genres play an important role in helping us predict movies' popularity, we removed those movies without genres. The dataset now contains about 33,000 movies after removing useless data. Another primary problem that we had to deal with was that the inputs under some columns could not be used directly. These columns are "genres", "production_companies", "production_countries", and "spoken_languages".

For the input under genres, we first checked the unique genres in the dataset (20 genres in total) and then made each genre a new column. If a movie belongs to some genres, one is labelled to them otherwise 0. We did the similar procedure for the column "status". For production companies, since there are more than 20,000 companies, we collected a list of top 500 production companies by web-scraping the link and added a new column called "top500_company_count" to count the number of top 500 production companies of each movie. The similar strategy was implemented on "production_countries" and "spoken_languages". We collected the top 20 frequent production countries and languages from using our dataset and created two columns "top20_freq_countries_count" and "top20_freq_langs_count" accordingly. We also transferred the content under "release_data" to three new columns which are "release_year", "release_month", and "release_day". For the column of "adult", "True" was replaced with 1 and "False" was replaced with 0.

We also established a new column called "LDA_closest_topic". We used LDA to generate three topics based on movie overviews after trying topic numbers of 5, 4, and 3 (see [Appendix 1b]). Then the most approximate topic was assigned to each movie. Although we already have genres of each movie, we did this topic modeling to obtain more detailed information of what a movie is about.

We also found that there exist outliers in our dataset. For example, from looking at the boxplot of "popularity" (see [Appendix 1c]), we can see that most movies have pretty low popularity, and some have extremely large values. When considering if we should keep them, we looked up the names of those outlier movies, which are "Minions", "Wonder Women", "Big Hero 6", and etc. We found that those movies are very popular around the world. It seems that it is reasonable for them to have such high popularity. Thus, we decided to keep those outliers in

our analysis. In the end, we deleted the original columns “genres”, “production_companies”, “production_countries”, “release_date”, “spoken_languages”, and “status”.

Credits

The dataset, ‘Credits’ had three columns originally, id, Cast, and Crew, and 45019 rows. ‘Cast’ and ‘Crew’ columns came in the JSON format and included detailed information about the casting team and crew team of each movie. We dropped the unnecessary data and wrong data to get an informatic dataset that we can directly work with: Dropping all data with Na values for Cast and Crew and all duplicates values in movie id.

Cast majority gender: For 'Cast', the name and gender of each actor/actress are extracted from JSON format using eval(), then we constructed a function named ‘get_cast_list’ which returns a list of names and majority gender. If males played a majority role in the cast, return an ‘M’; if more females, then return an ‘F’, and a ‘B’ stand for balanced gender. 65% of the casting team are the male dominated cast; 20% are balanced, and only 14% are female dominated.

Top cast: To build a more direct connection between cast and movie popularity, we need to set a binary variable, ‘Top cast’, to show if the cast is popular or not. Since movies in our dataset are in a worldwide range and the range of release time is over 140 years, we constructed a list including names of top 100 actors’ and actresses' names, 200 names in total by using web scraping from IMDB: worldwide top 100 actors (actresses) of all the time. If any of the names in the casting team of a movie is in this top list, we consider the movie is performed by a popular casting team and marked ‘1’ for ‘Top cast’. It turns out that only 7% of movies have a top cast team.

Top director: For Crew, we did the same procedure and got the name of the director. If the director is in the IMDB list of worldwide top 100 directors of all the time, we consider the movie is directed by a top director and marked ‘1’ for a binary variable, ‘Top director’. It turns out that only 1% of movies are directed by a top director.

Lastly, we dropped two original columns: ‘Crew’ and ‘Cast’, because they will significantly delay the efficiency of our work due to the large size.

Data Merging

Finally, by merging movies_metadata_clean.csv and credits_clean.csv with movie id we got our Movie_Data_Update.csv with 32142 entries and 43 columns in total, which means that we kept around 97% of the useful dataset.

Explore Data Analysis & Variables Selection

All fields are divided into four categories: Released Date Features, Genres Features, Overview Features and Producer Features, and we demonstrate the relationship between them and Popularity by data visualization.

Released Date Features

We think release dates have an impact on the popularity. The relationship between movies and release date can be seen in [Appendix 2a] & [Appendix 2b]. The movie releases and popular movies percentage almost increased every year. We can observe that most movies release during Fall and January, but movies released in Summer tend to be popular. While most movies released on the first day of a month, only a small portion of them could be popular.

Genres Features

In total, there are 20 types of genres labels. The top 3 most popular genres are Adventure, Fantasy and Animation, and the top 3 most unpopular genres are Foreign, Documentary and TV

Movie. The popularity difference of Adventure and Foreign is over 5. The popularity between different genres are significantly different, which can be easily concluded in [Appendix 3].

Overview Features

As discussed, movies are divided into 3 groups based on overviews using LDA models. The Popularity of Overview topics graph is in [Appendix 4]. The number of popular movies shows highest in the second topic, and popular movies percentage shows highest in the first group. This means the number of popular movies and popular movies percentage varies in different topics. Therefore, we will include this feature into our model.

Producer Features

Better producers are more likely to lead to higher popularity. We define many variables as Producer features, like Top Cast, Top Director, Number of Top 500 Company, Number of Top 20 Country etc. We take Top Cast ([Appendix 5a]) and Number of Top Company ([Appendix 5b]) as examples. The average popularity is higher when there is any top cast in the producer team, and it is similar for the movies produced by one or more top 500 movie companies. The result actually corresponds to our prior expectation.

Model Training & Results Comparison

Machine Learning Models

i). Linear & Logistic Regression

For this dataset, we first used regression analysis to obtain the fitting values for movies' popularity. Since the initial data for popularity is continuous and hard to measure, we split the popularity into 1 and 0. 1 standards for the popularity of that movie is greater than 5, and 0 otherwise. Another reason for adding this threshold is that it is ambiguous to tell how popular the movie is by looking at the numerical value. With the threshold, our models will be more straightforward in predicting whether a movie is popular or not. We selected the threshold of 5 based on our observation to movies and the popularity distribution. Also, setting the threshold to 5 won't greatly affect the relationship between popularity and other variables. For instance, the order of the mean popularity of genres doesn't have much change after using the threshold to transfer popularity to a binary variable (see [Appendix 6]). This project aims to create a model to determine whether a movie is popular or not. Thus, each movie would be assigned a probability between 0 and 1, with a sum up to 1. To confirm we should use logistic regression instead of linear regression, we need to test the results of using a linear regression model that did not show as well as the logistic model.

To model with regression methods, we used a training and testing dataset and separated out the feature set and target value for the dataset. By using the sklearn's linear_model library, we got our training data to fit the model and decided on a threshold value which is 0.28 that assigns either 0 or 1 to each test case. Using the results of the confusion matrix, we can calculate metrics that help us evaluate the model. We saw that the precision which measures the proportion of cases identified as positive that is actually positive is only half of the percent and 61% tradeoff between precision and recall. The accuracy, in the end, is 0.74. To make a comparison, we also try the logistic regression model. The result of accuracy is nearly 0.80, which is greater than that of linear regression.

Thus, we can conclude that it is more efficient by using the logistic model for our dataset. For a binary dependent variable, logistic regression is the fittest model for predicting whether the movie is popular or not.

ii). Other Models: KNN, Random Forests, and Neural Networks

Next, we used other famous algorithms for classification, such as K-Nearest Neighbors, Random Forests, and Neural Networks. These algorithms have different targets and results for measuring the dataset, so let us introduce the details of each model below:

- KNN: We tried different values of parameter k which is the number of nearest neighbors, and we observed the error that decreases and then increases is at approximately k=45.
- Random Forests: To avoid overfitting and find the best ensemble, we set 50 n_estimators, 15 max_depth, 4 min_samples_leaf, and 8 min_samples_split. Then, we used the feature importance graph to highlight important features.
- Neural Networks: We used one hot encoder to encode the popularity column and started with one hidden layer with 36 nodes and 500 epochs. Then using sklearn to classify the data into multiple layers and got the suitable model.

Model Results Comparison

From the table below, models are compared based on MSE, RMSE, and Accuracy. The results of these four methods are pretty similar, but we can know that the Random Forests model is the best model for predicting the movies' popularity, and the Neural Network does not perform very well. The reasons might be our available data is relatively small, no overfitting data, and the subsets of the features are figured in some trees or the other.

Metric	Logistics Regression	KNN	Random Forests	Neural Network
Mean Squared Error	0.204	0.205	**0.179**	0.213
Root Mean Squared Error	0.452	0.453	**0.423**	0.462
Accuracy	0.796	0.795	**0.821**	0.786

Prediction & Application

Since the Random Forests model is the most suitable one for our data, and by using the feature importance graph, we knew the top 3 significant features are top director, top cats, and cast majority gender. Thus, we can predict the popularity of one movie mainly based on these three features. The application can be used in movie theaters. To make a large profit, the theater can give more time to release the movies which are defined as popular by our model.

Future Improvements

First, we noticed that the gap between well-known movies and other movies is significantly large. For instance, the popularity of Minions is 547, and most of them have popularities around 2. Therefore, since several movies are too popular to define, it is hard for us to choose a threshold to differentiate all the movies. After a thoughtful investigation, we let the threshold to be 5 for modeling our dataset.

Second, since Neural Network is the most popular algorithm for modeling today, we can explore and conduct other methods to let the data fit this algorithm.

Last but not least, to put it into real usage, we still need to improve our model. Here, we only consider covering all the movies. Maybe, it is more accurate to delete some outliers which either have too large or too small popularity values. Also, we can add more variables to achieve better performance in predicting the popularity of movies.

Appendix

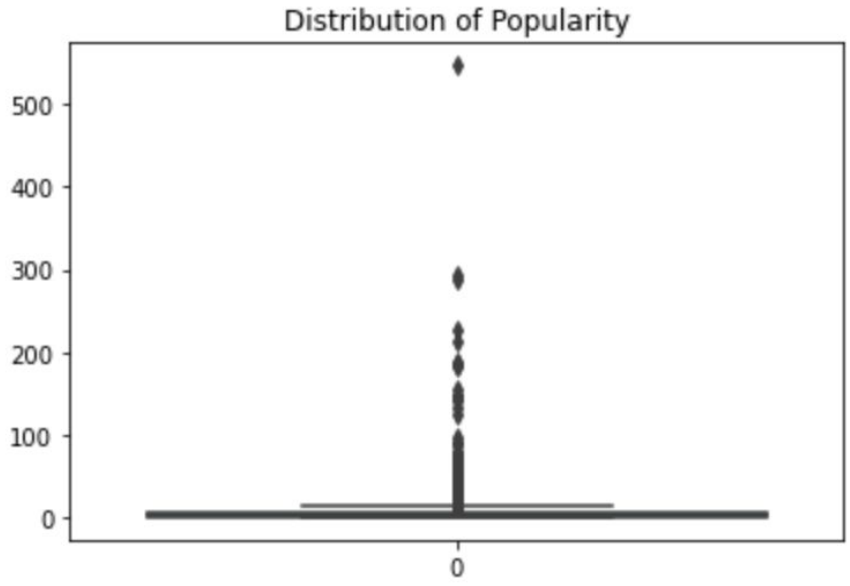
[Appendix 1a] Information about Extracted Columns

Column Name	**Input Example**
id	82702
adult	False
genres	[{'id': 14, 'name': 'Fantasy'}, {'id': 28, 'name': 'Action'}, {'id': 12, 'name': 'Adventure'}, {'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': 10751, 'name': 'Family'}]
overview	The thrilling second chapter of the epic How To Train Your Dragon trilogy brings back the fantastical world of Hiccup and Toothless five years later. While Astrid, Snotlout and the rest of the gang are challenging each other to dragon races (the island's new favorite contact sport), the now inseparable pair journey through the skies, charting unmapped territories and exploring new worlds. When one of their adventures leads to the discovery of a secret ice cave that is home to hundreds of new wild dragons and the mysterious Dragon Rider, the two friends find themselves at the center of a battle to protect the peace.
popularity	12.256689
production_companies	[{'name': 'DreamWorks Animation', 'id': 521}, {'name': 'Mad Hatter Entertainment', 'id': 20154}]
production_countries	[{'iso_3166_1': 'US', 'name': 'United States of America'}]
release_date	6/12/2014
runtime	102
spoken_languages	[{'iso_639_1': 'en', 'name': 'English'}]
status	Released

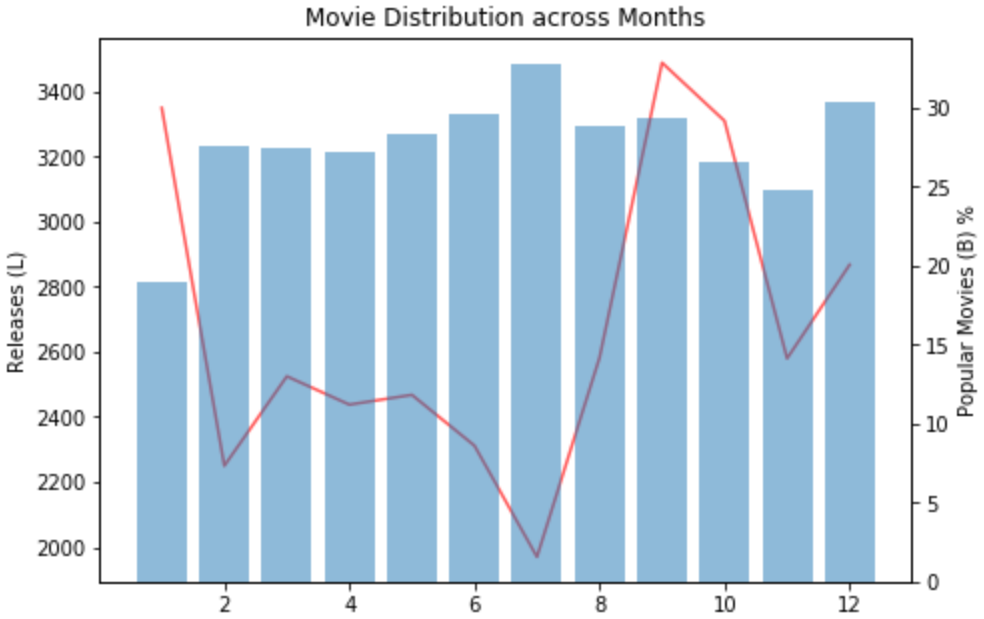
[Appendix 1b] LDA Topics

```
[   (   0,
        '0.005*"police" + 0.004*"world" + 0.004*"young" + 0.004*"group" + '
        '0.003*"town" + 0.003*"gang" + 0.003*"save" + 0.003*"murder"'),
    (   1,
        '0.010*"life" + 0.008*"young" + 0.008*"love" + 0.007*"family" + '
        '0.005*"father" + 0.004*"girl" + 0.004*"mother" + 0.004*"friends"'),
    (   2,
        '0.008*"war" + 0.007*"world" + 0.004*"life" + 0.004*"documentary" + '
        '0.003*"american" + 0.003*"comedy" + 0.003*"history" + 0.002*"people"')]
```

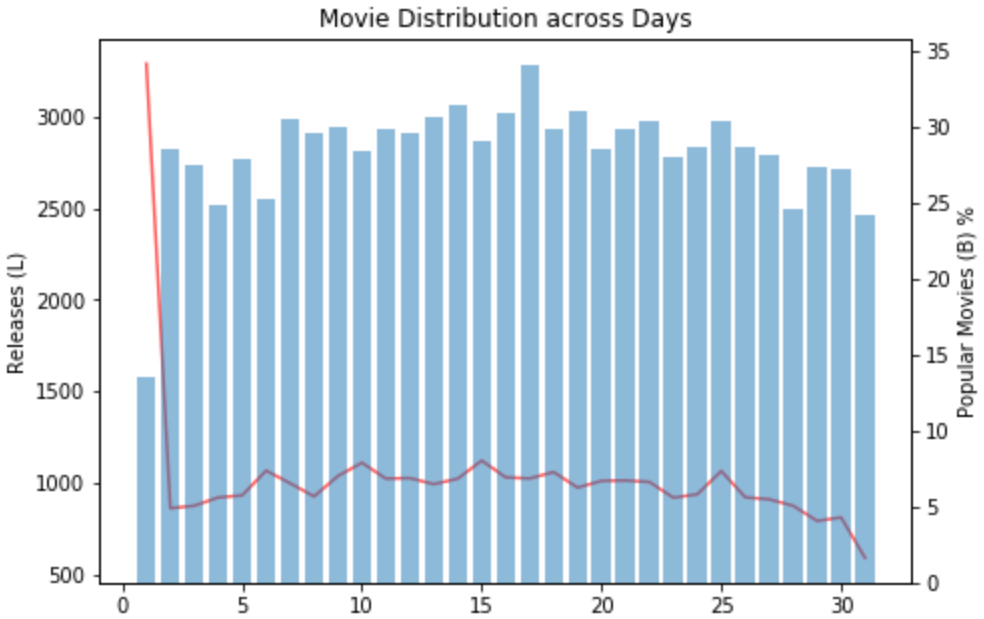
[Appendix 1c] Boxplot of Popularity



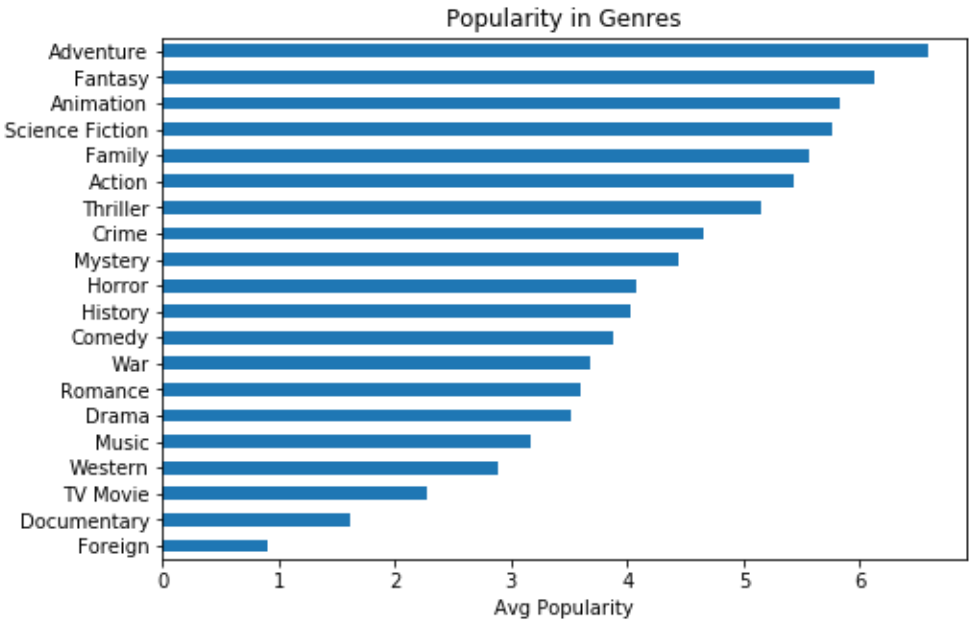
[Appendix 2a] Movie Distribution across Months



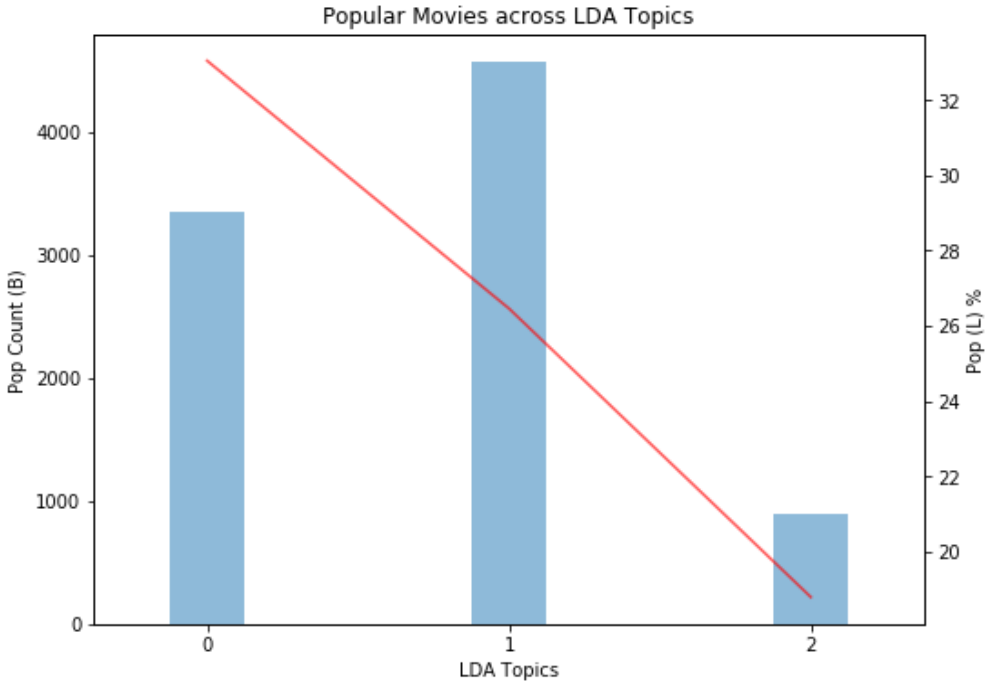
[Appendix 2b] Movie Distribution across Days



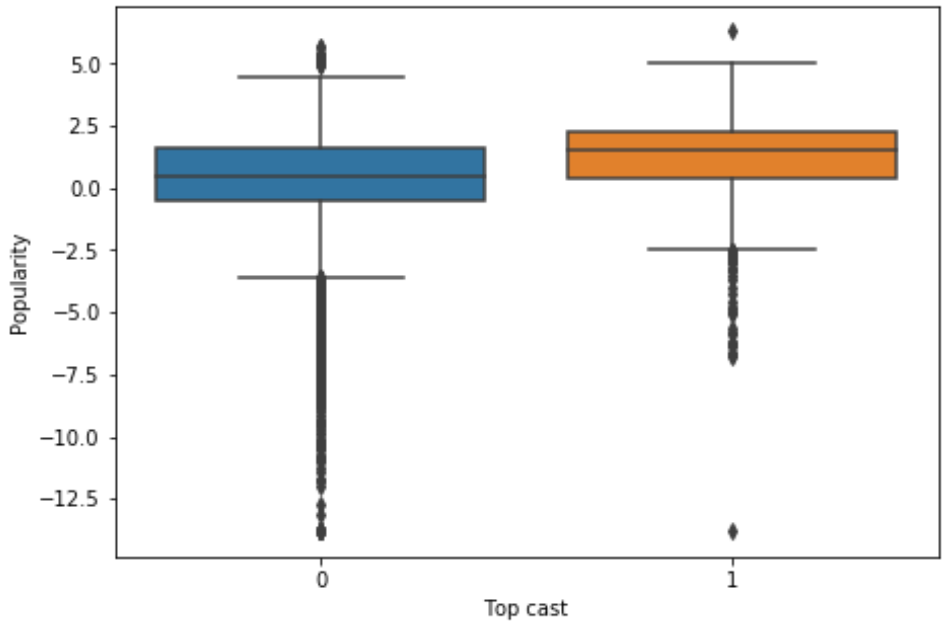
[Appendix 3] Popularity of Movie Genres



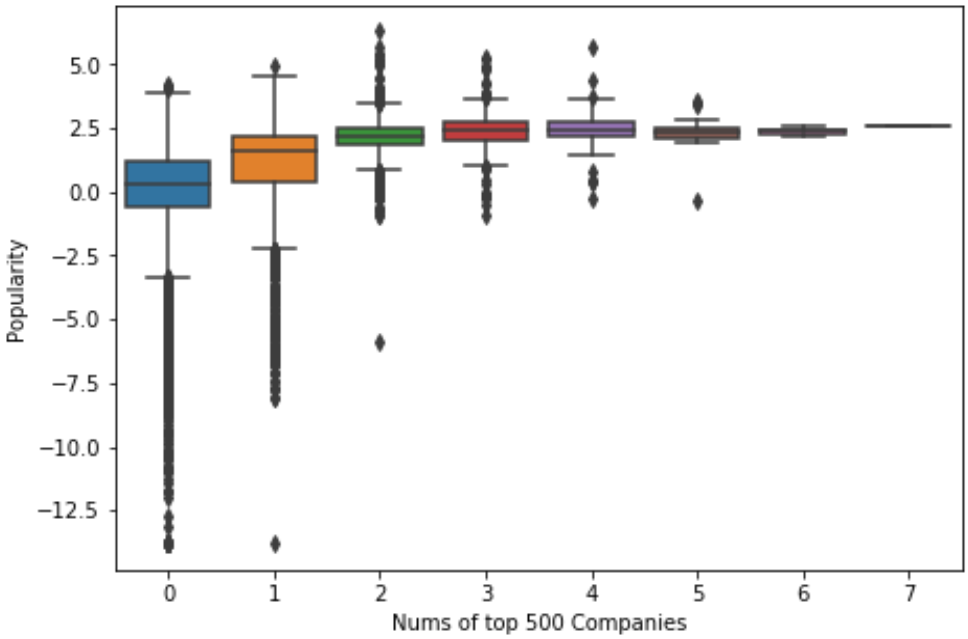
[Appendix 4] Popularity of Overview Topics



[Appendix 5a] Popularity of Top Cast



[Appendix 5b] Popularity of Top Company



[Appendix 6] Comparison of the Order of Mean Popularity of Genres