**CISC7201 Introduction to Data Science Programming**

**Course Project 2019/2020**

Group Members: MB955473 Tou Ka Him

MB955464 Lei Sin Ian

MB955048 Chao Lok Kio

Topic: Movie Analyzation

**1. Introduction**

In this project, we are interested in analyzing the behaviour of movie selection of people. We have chosen some critical elements that are important to a movie selection for our analysis. In our work, we have applied the techniques we have learnt in Python and some Python libraries to analyze and visualize the relationship of each element.

**2. Data Source**

Data Source from Kaggle: movies\_metadata.csv

(<https://www.kaggle.com/rounakbanik/the-movies-dataset>)

**3. Python Libraries**

We have make used of different libraries in our work:

* Numpy: it is a fundamental package for scientific computing with Python
* Pandas: it is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language
* Matplotlib.pyplot: it is a state-based interface to matplotlib and provides a MATLAB-like way of plotting
* Ast (Literal\_eval): it is used for safely evaluating strings containing Python expressions from untrusted sources without the need to parse the values oneself
* Seaborn: a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics
* Plotly: a Python library that is known for developing and providing online analytics, statistics and graphing tools

**4. Procedures**

I. Import Dataset

This dataset contain a few csv files and we have just chosen two files for our analyzation after reading the description of each dataset. Also, in one of our datasets, the column ID is different from the others, so when are importing the dataset, we have renamed the column ID and calculated the mean rating for each movie. Then, as we are using 2 csv files, we have merged these two files in our next step.

II. Data Cleaning

First, we have chosen the corresponding elements for our data analysis, then we need to check whether these columns contain any null values. Also, we have checked the datatypes for each column and we need to change the data type for "popularity", "budget", "vote\_count", "vote\_average" to "float64" and "release\_date" to "datetime64[ns]".

|  |  |
| --- | --- |
| Before | A picture containing text  Description automatically generatedAfter |

For the data in "genres" and "production\_countries", as they are presented in dictionaries in list, we have applied the following codes (Fig.1) to get back the values of dictionaries.

A screenshot of a cell phone

Description automatically generated Fig. 1

III. Analysis and Result

**The Correlations Between Elements**

From Fig. 2, we have found out that budget, revenue, profit and vote\_count are correlated. It states that the higher budget movies can result in a higher revenue, but it doesn't represent that they have a high rating as the vote\_count is not correlated with the ratings.

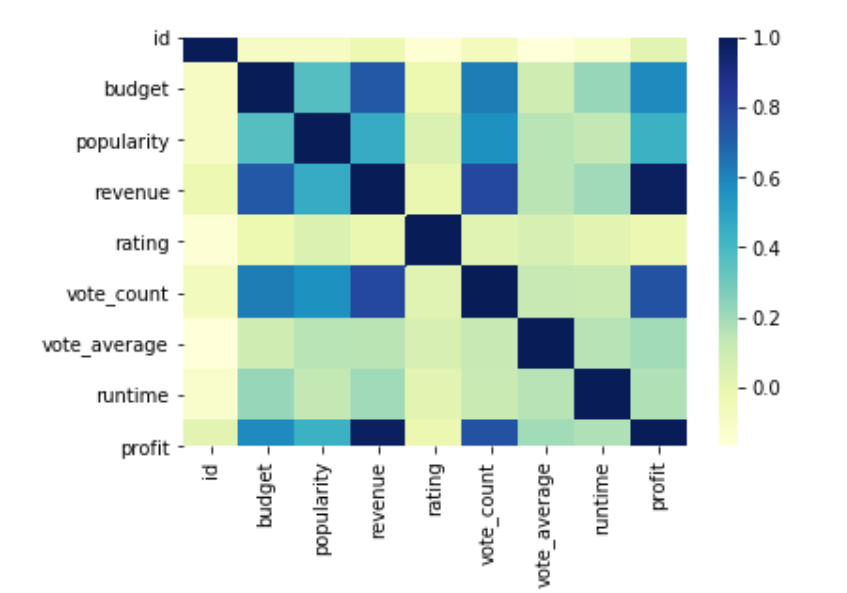


Fig. 2 The Correlations Between Elements

**Top 10 Movie Types**

From Fig. 3, we have found out the most common movie types that are being produced and Drama is the most favorite among all types of movie.

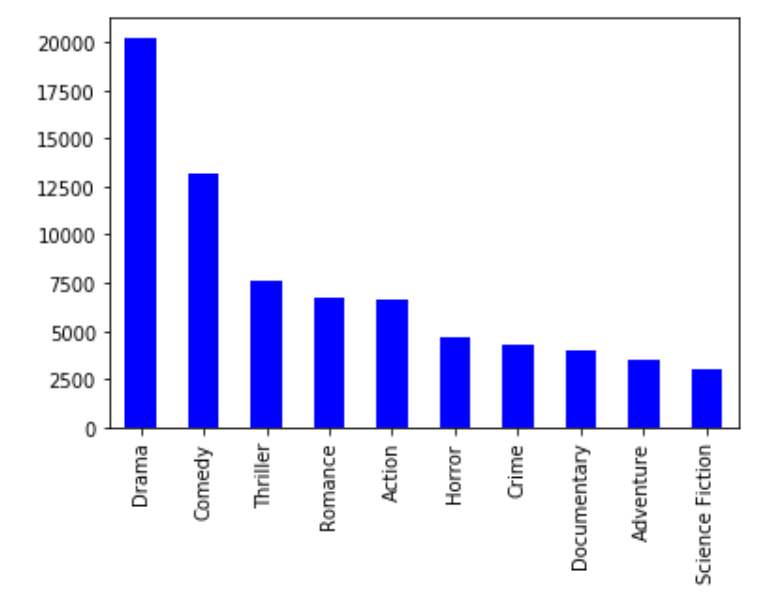


Fig. 3 Top 10 Movie Types

**Movies Around the World**

From Fig. 4, we have found out the number of movies produced in each country and America has the greatest number of movies produced.

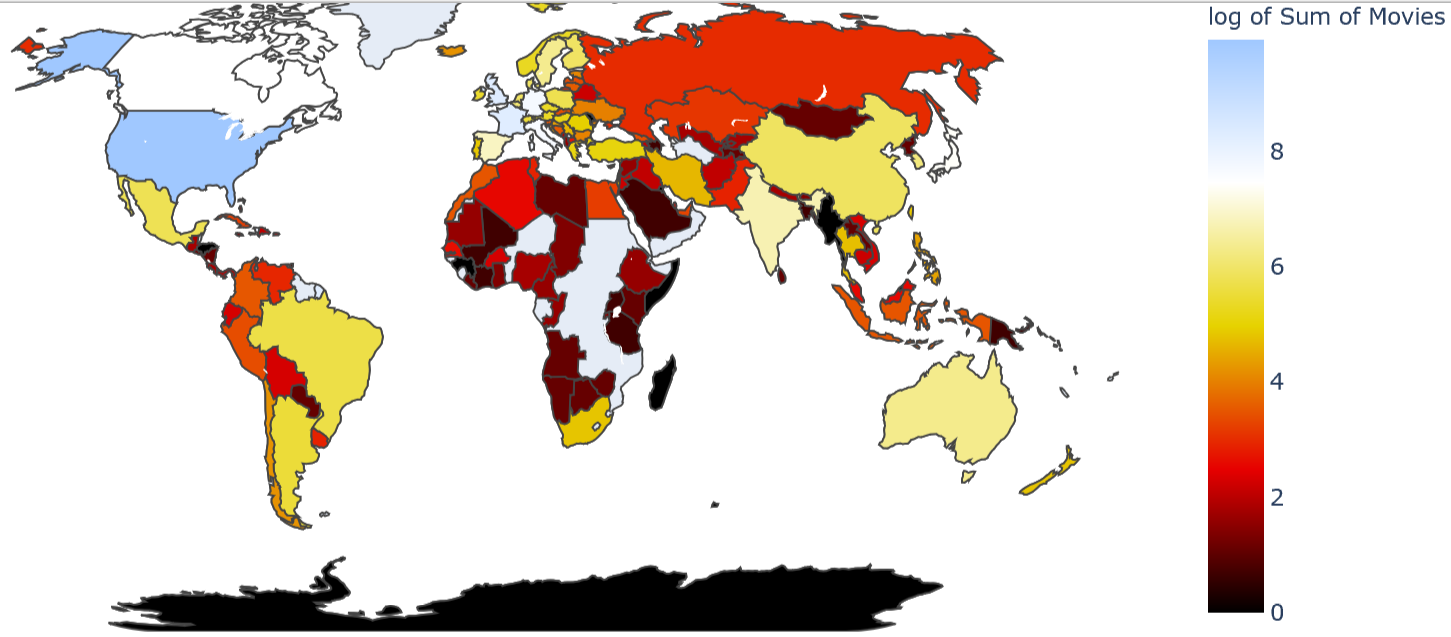


Fig. 4 Movies around the world

**Top 10 popular movies**

From Fig. 5, we have found out the most top 10 popular movies.

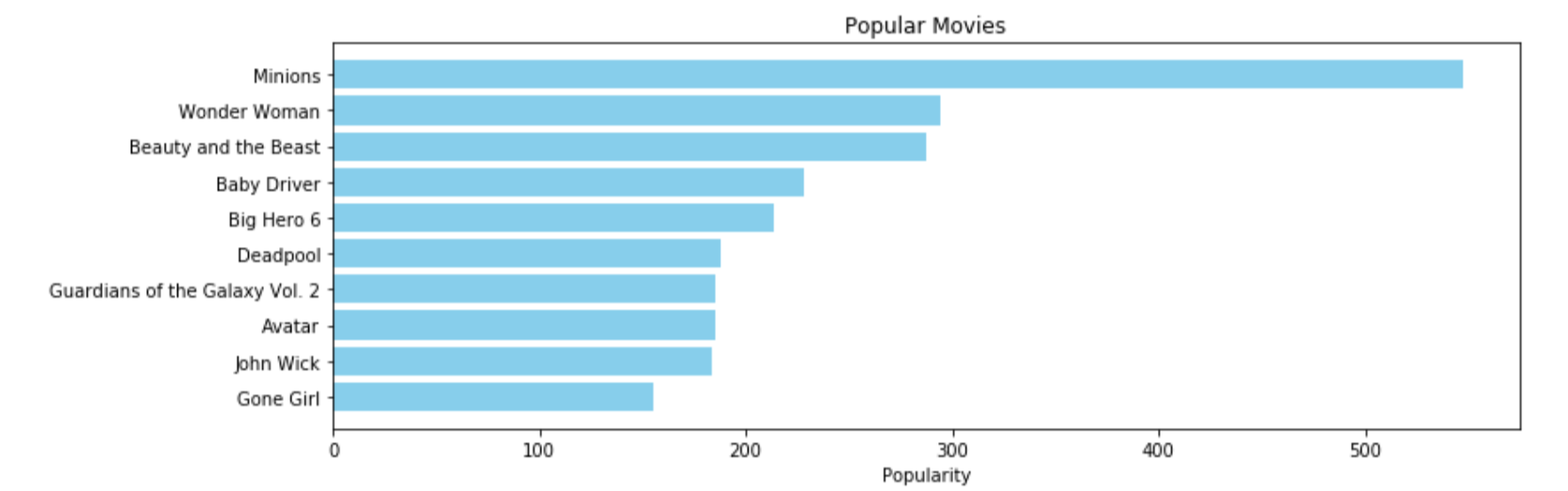


Fig. 5 Top 10 popular movies

**5. Conclusion**

To conclude, we have made use of the past data to analyze and visualize the behaviour of movie selection of people and we can keep update with the live data so as the movie producer can choose what types of movie they will produce in the future.

**6. Future studying**

As our computer science skills are limited, we suggest that some future work can be done with improved techniques:

* We can keep update with the live data as the dataset that we have now is limited to the current and past data;
* We can analyze the information on the actors and crew so as to produce the best movie for people;
* The behaviour of movie selection can be personalized and tailor-made from the past habits of each person.