# Group 7 Project 2 - Movies Database

Members: Lisa Santonato, Minna Wu, Dian Wang, Suchitra Sharma

## Abstract

## The objective of this project is to extract data from multiple sources of online movie ratings and compile them into a central database. Using this database, one can conveniently compare user and critic ratings of top films across multiple sources (i.e. IMDB.com, RottenTomatoes.com, and Metacritic.com). Our process involves extracting gross earnings from the-numbers.com, then triangulating ratings and other movie details from the previously mentioned three sources.

## Introduction

There are many websites online that offer movie reviews. A typical process many film seekers perform is to check the typical IMDB rating, the audience score on RottenTomatoes.com, and/or the Metacritic.com score. Some people check two or three of these one at a time through a Google search! Cumbersome, indeed. This project strives to make rating comparisons of top films throughout time easily accessible. The project is divided into three main components: Extract, Transform and Load. For this, data is extracted from four different websites using the BeautifulSoup and Pandas library, transformed into the required format using the Pandas library and then loaded into MongoDB. The analysis is done also using the Pandas library. Finally, plotting is done using the Matplotlib library.

The questions that this project proposes to answer are:

1. What are the top grossing films per week?
2. What are the top 100 grossing films of all time?
3. What are the rankings of these 100 films on Metacritic.com, IMDB.com, and RottenTomatoes.com?
4. What does the comparison between critic scores and user scores look like?

We expect that user and critic scores would be different on each website, which is why there is value to checking each source in the first place.

## Resources

The websites that are used in this project are:

1. <https://www.the-numbers.com/>
2. <https://www.metacritic.com/>
3. <https://www.imdb.com/>
4. <https://www.rottentomatoes.com/>

Technologies used are:

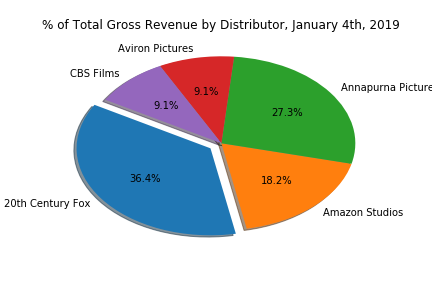
1. Jupyter Notebook (Python 3 and Python Libraries like Beautiful Soup, Pandas, and Matplotlib)
2. MongoDB

## Methods, Findings, and Discussion

1. Gross Revenue by Distributor

The data from Numbers website is extracted with web scraping using BeautifulSoup. The box office chart table is directly extracted from the website and used for analysis. For data cleaning, column Headers are added manually replacing the column numbers. Subsequently, ‘$’ sign is removed from the Gross and Total Gross columns to enable analysis. The data is then grouped by ‘Distributer’ column and gross earning and total movie per distributer data is extracted. The data is then merged into one table and then plotted into the pie plot as shown in Figure 1 below:

Figure 1: Total Gross Revenue by Distributor



1. Comparative analytics from Metacritic, IMDB and Rotten Tomatoes data

The data from Metacritic and Rotten Tomatoes website is extracted using web scraping, while imdb data is extracted from OMDB API. Since the aim here was to do the comparison between the ranking of top grossing films from three different websites, the extracted data required rigorous cleaning to bring it to the form where merging is possible.

## Conclusions

…