# ­­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:

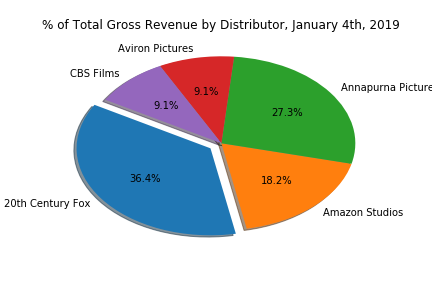
1. Jupyter Notebook (Python 3 and Python Libraries like Beautiful Soup, Pandas, and Matplotlib)
2. MongoDB

## Methods

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 analysis of movie reviews from different web sources

The base data of Top 100 Grossing movies were extracted from the-numbers.com web tables using the Pandas read\_html built-in function. The data from Metacritic and Rotten Tomatoes websites were extracted through web scraping using the BeautifulSoup library. Finally, IMDB data was extracted using the OMDB API. All extractions were performed using the requests library. Since the aim was to do the comparison of the top-grossing movies from three different websites, the extracted data required rigorous cleaning to bring it to the form where merging was possible. First, data from the Numbers website was extracted to get the top 100 grossing movies and cleaned to convert box office column datatype from str to float.

Movie titles from Numbers website are used to scrap data from Metacritic website. Similarly, data from IMDB website is extracted from OMDB API and data from Rotten Tomatoes website is extracted with web scraping.

Following Data cleaning were steps taken before merging the data:

* Stripping blank spaces.
* Converting box office data to decimal formatting.
* Re-formatting selected movie names to maintain uniformity throughout the datasets and ease of analysis.
* Merging data from different web sources

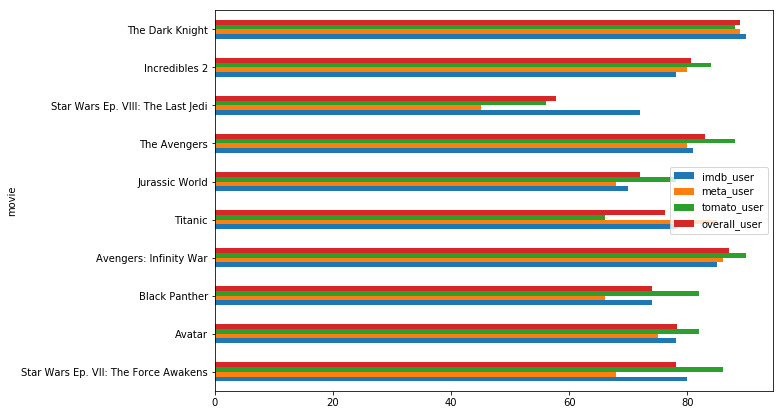
Data from four different website is then merged into a new data frame to have a collection of user ranking and critic ranking columns. The merged data is plotted to see the comparison between different web sources.

All the files are loaded to the MongoDB database for further analysis.

## Findings and Discussion

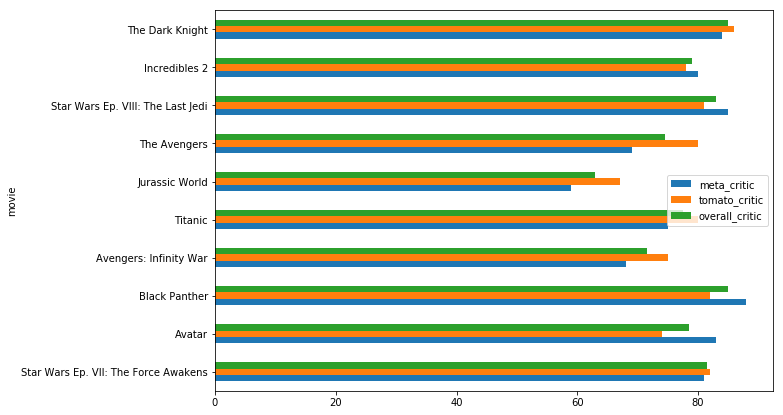
To Visualize user ranking data from different websites for top 10 grossing movies, horizontal bar graph was plotted The Figure 2 below shows the comparison of user ranking data from different web sources.

**Figure 2 Comparing user ranking from different web sources**



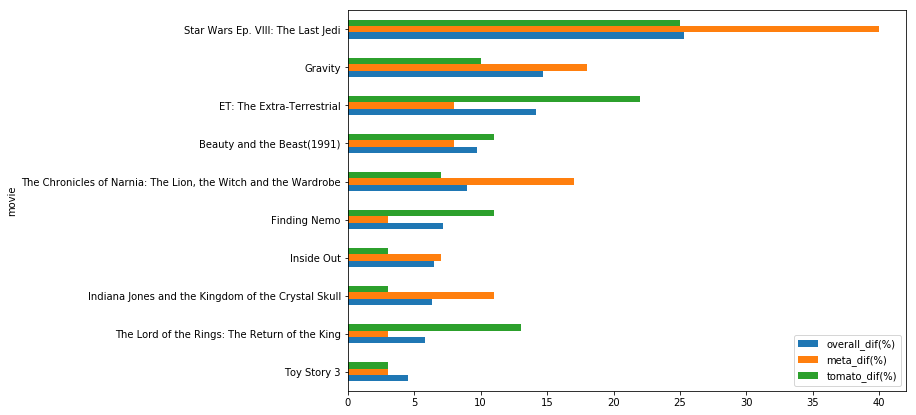
Similar graph (Figure 3) was created to visualize critic ranking data from different web sources for top grossing movies.

**Fig.3 Comparing Critic ranking from different web sources.**



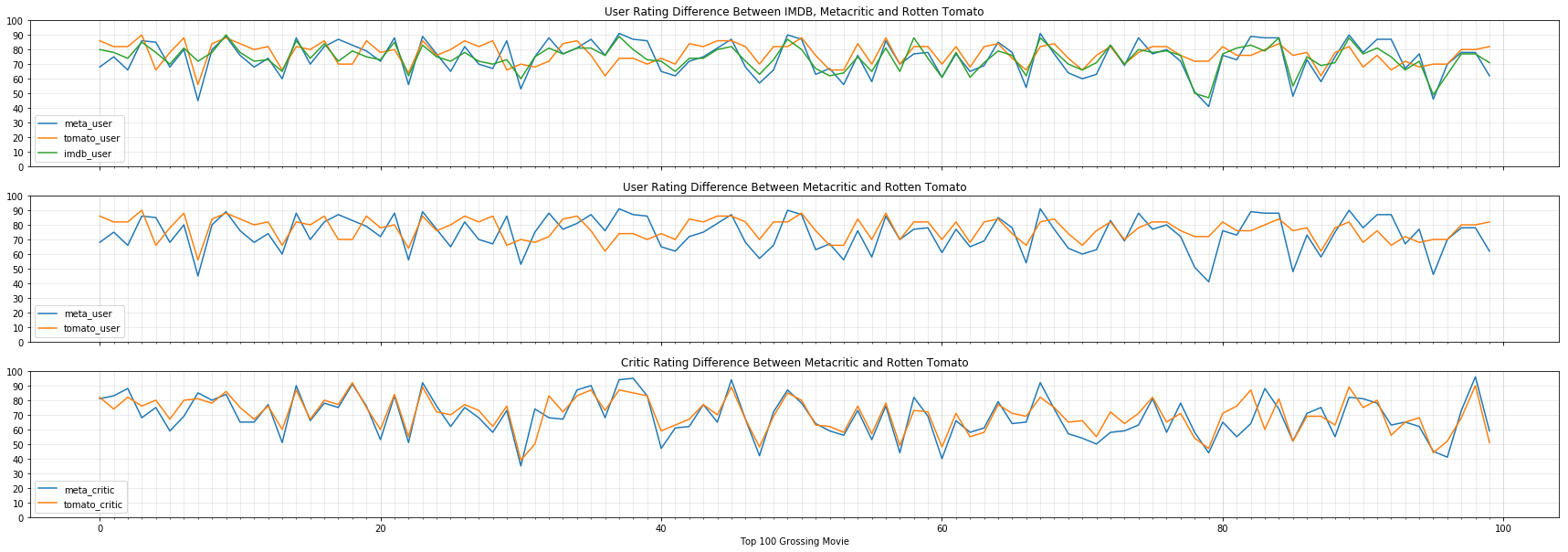
Data with the difference between user ranking and critic ranking is tabulated and plotted to see how different the critic ranking is from user ranking. As critic data was available only from rotton tomatoes and Metacritic, data from these websites are taken to plot the difference. Fig.4 below shows the plot for the same.

**Fig.4 Difference between critic ranking and user ranking from different web sources.**



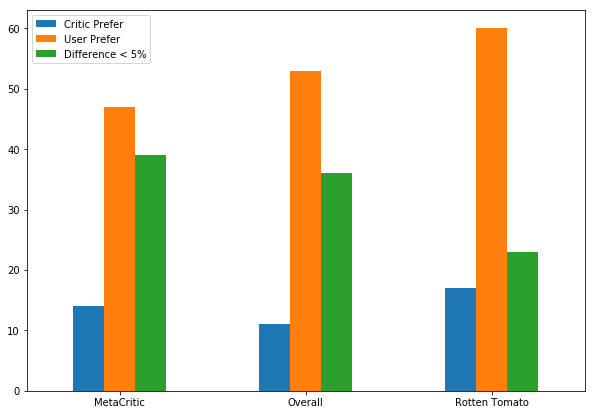
Rating Data with the difference between three movie rating sites. Fig.5 below shows the rating difference (for better view, please check wangdian’s subfolder).

**Fig.7 Rating Difference between Movie Rating Sites**



Lastly, the difference between critic and user preference is plotted on the bar graph with as shown in figure 4. Difference between critic preference and user preference is less than 5 %.

**Fig.6 Difference between user preference and critic preference**



## Conclusion

1. The top 100 grossing movies(Top100) gained more success in market than the re-watch value. After analyzing the rating difference between user and critic (based on MetaCritic, Rotten Tomato, and IMDB sites), it is clear that over 50% movies in Top100 are getting lower rate by critic (**Fig.6**). However, this does not affect these 100 movies getting huge commercial success in movie market. So, Top100 are probably more commercial successful products rather that re-watchable artwork.
2. We could not find a very clear relationship between rank and average rating in Top100, which proves the conclusion in point one again - Top100 are probably more commercial successful products rather than re-watchable artwork.
3. In general, critic rating and review are more reliable than amateur users. Even though movie rating sites have different algorithm to calculate the average rating, **Fig.5** shows that critic’s scores are closely related compared with user’s score between different sites. So, critic may intend to give a reasonable view based movie background knowledge, while users prefer to rate a movie randomly based on their individual preference.