# ­­Group 7 Project 2 - Movies Database

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

## Abstract

## The objective of this project was 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 would be 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. What better way to find a popular film that one missed years back?

This project was divided into three main components: Extract, Transform and Load. First, data was 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 ratings 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. However, we expect that the scores would be in the same ballpark (not more than two out of ten points apart). We also hypothesize that the user and critic scores are vastly different since the popular box office titles may not be critically-acclaimed.

## 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/>

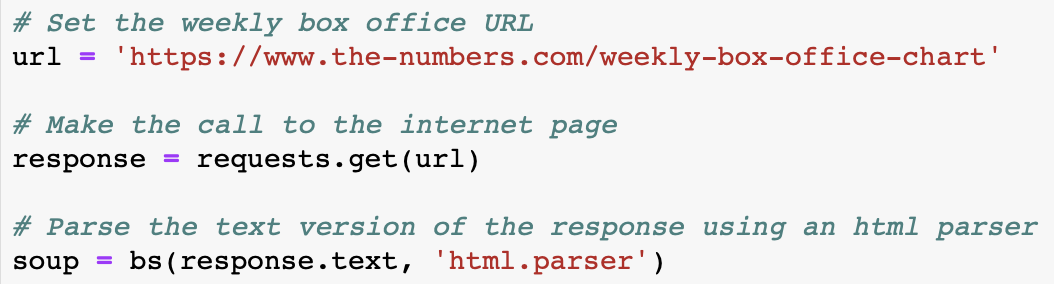
Technologies used are:

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

## Methods

1. Gross Revenue by Distributor

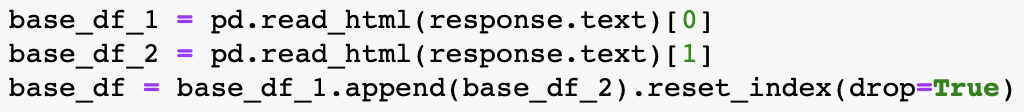
The weekly top grossing movies data from Numbers website was extracted with web scraping using the BeautifulSoup library in Python.



For data cleaning, column Headers were added manually to replace the column numbers. Subsequently, the ‘$’ signs were removed from the Gross and Total Gross columns to enable plotting analysis. The data was then grouped by the ‘Distributer’ column and gross earning and total movies per distributer data were extracted. The data was then merged into one table and plotted into the pie plot as shown in Figure 1 in the Findings section below.

1. Comparative Analysis of Movie Reviews from Different Web Sources

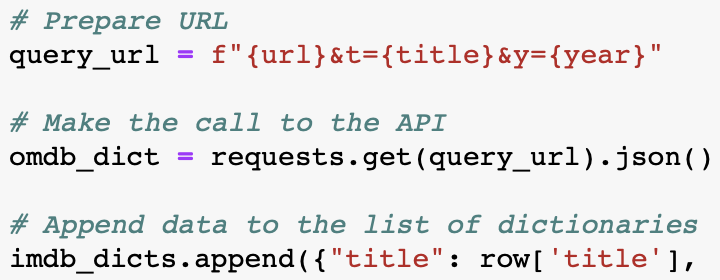
The base data of Top 100 Grossing movies were extracted from the-numbers.com web tables using the Pandas read\_html built-in function.



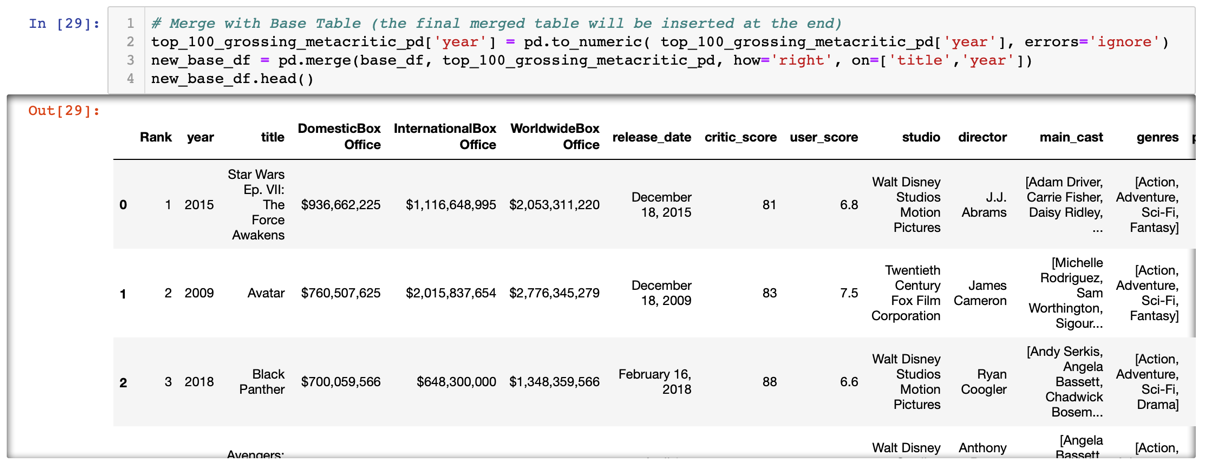
The data from Metacritic and Rotten Tomatoes websites were extracted through web scraping using the BeautifulSoup library.



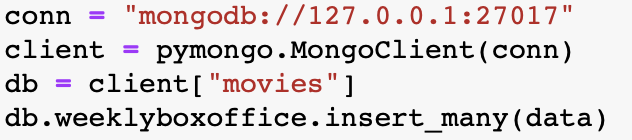
Finally, IMDB data was extracted using the OMDB API. All extractions were performed using the requests library. Since the aim was to do the comparison of the top-grossing movies from three different websites, the extracted data required rigorous cleaning to bring it to the form where merging was possible.

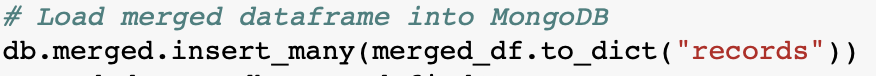


Data from the four different websites were then merged into a new dataframe by title and year to have a collection of user rating and critic rating columns. Merging on both title and year was critically as four or five movies had multiple versions that were released in different years. The merged data was then used for plotting analyses to see the comparison of ratings between the different rating systems.



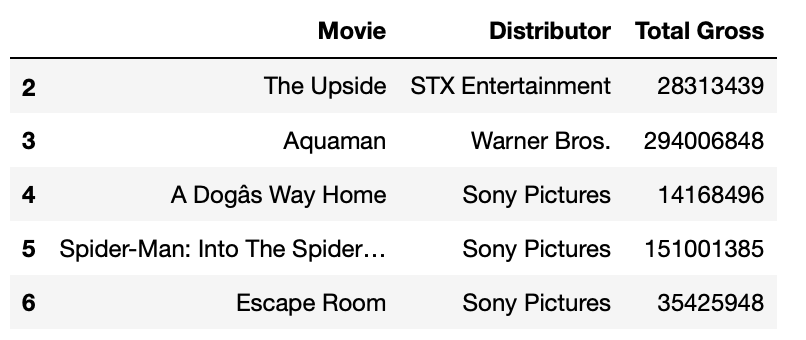
The prepared tables from each rating system and the final composite table with all four sources were loaded into a “movies” MongoDB database under different collection names.





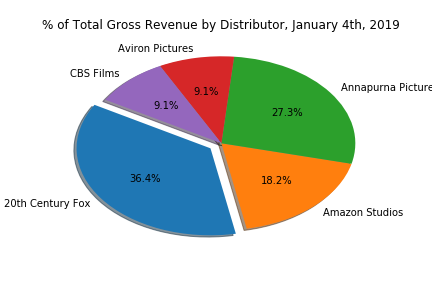
## Findings and Discussion

1. The top weekly box office titles of the week of January 4th are:



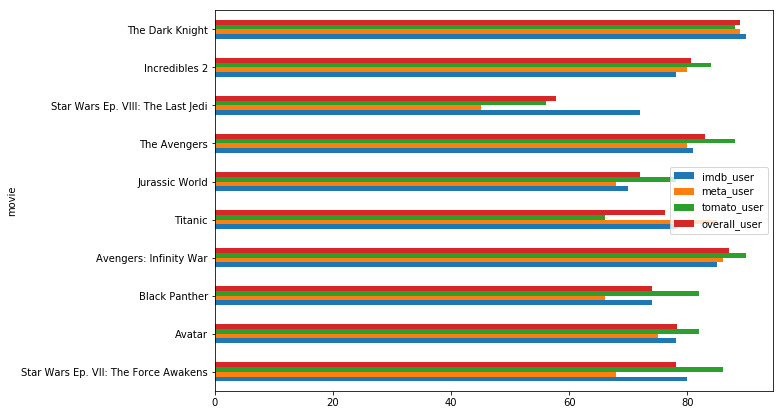
Out of interest, Gross Revenue by Distributor was plotted and 20th Century Fox was the leader surprisingly followed by Amazon Studios seen in Figure 1 below.

**Figure 1. Total Gross Revenue by Distributor**



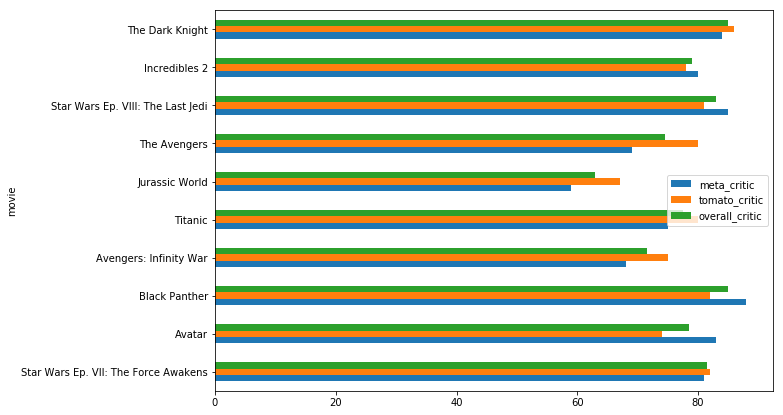
1. To visualize user rating data from different websites, we chose to plot the user and critic scores of the Top 10 Grossing movies of all time. A horizontal bar graph is shown in Figure 2 below. It details the differences between user ratings from different web sources. We can see that the data supports our hypothesis in that the scores are different between websites but not more than two points apart (or 20 in this case since the total was made to be out of 100).

**Figure 2. Comparing user ratings from different web sources**



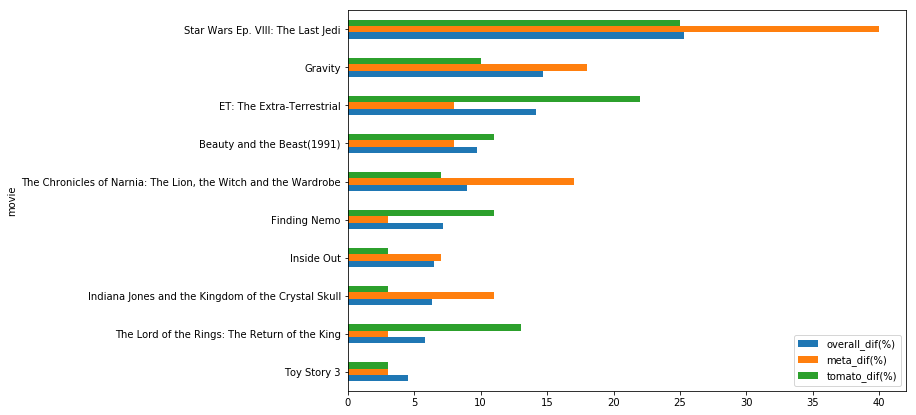
1. A similar graph (Figure 3) was created to visualize critic scores from the different web sources for the same Top 10 Grossing movies of all time.

**­­­Figure 3. Comparing critic ratings from different web sources**



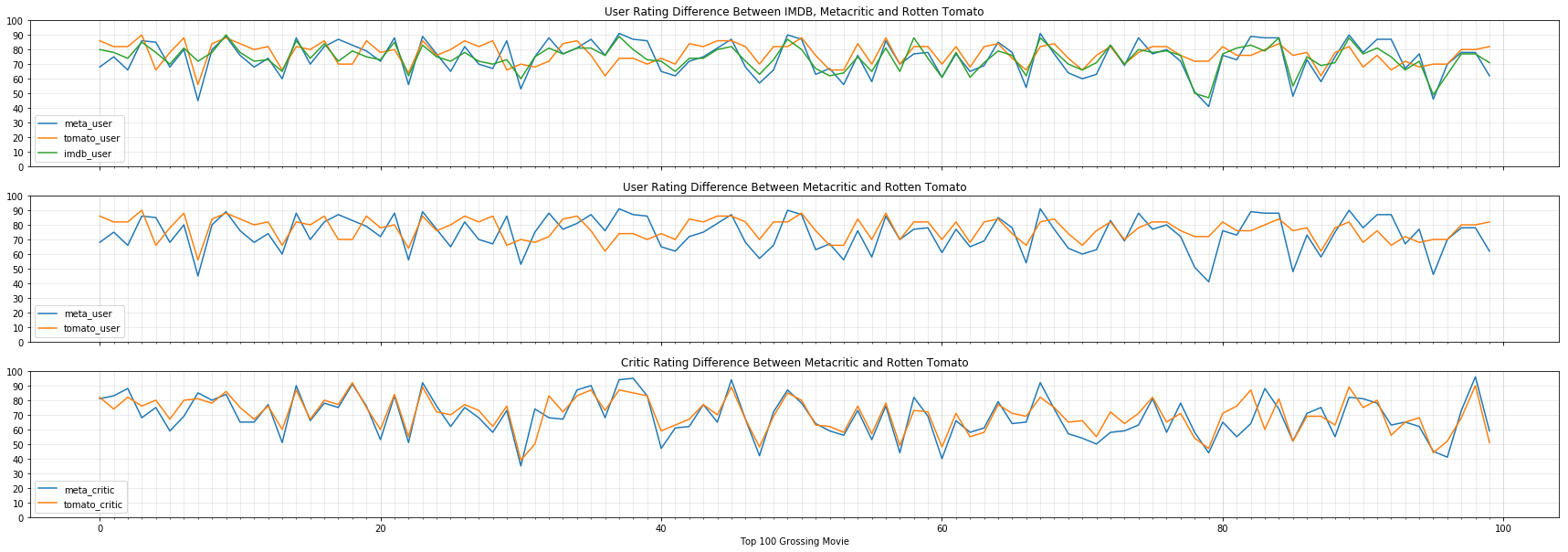
1. Movies with the largest rating difference between user and critic ratings were tabulated and plotted to see how different the critic ratings were from the user ratings. As critic data was available only from Rotten Tomatoes and Metacritic.com, we plotted the scores of these two as well as the average of the two in Figure 4 below. This shows the Top 10 movies with the largest critic and user rating discrepancies (overall on Metacritic and Rotten Tomatoes).

**Figure 4. Difference between critic and user ratings from different web sources**



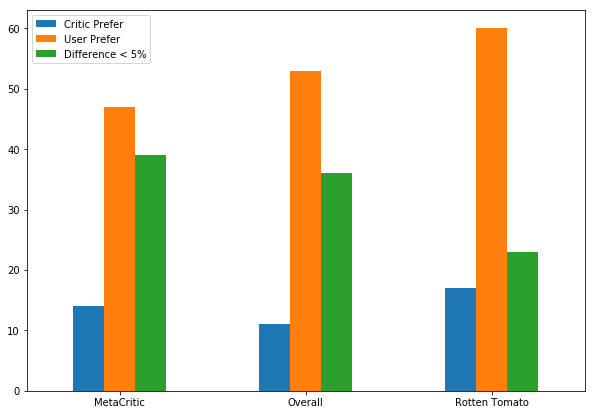
Rating data with differences between three movie rating sites were plotted. Figure 5 below shows the difference in ratings. (For a higher resolution image, please visit the subfolder, wangdian). As one can tell, the three lines are quite tightly wound together, not differing by more than 20 points (which supports our hypothesis).

**Figure 5. Rating difference between movie rating sites**



Lastly, if the critics on Rotten Tomatoes actually gave a fresh (rather than rotten) vote for a movie in our Top 100 Grossing Movies of All Time list, we added that to a count. In Figure 6 below, one can see that total is about 10% to 20%, which is quite low. Contrastingly, 50% to 60% of users gave a fresh vote to these movies. Finally, the green bars show that only 20% to 40% of Top 100 Grossing Movies had user and critic ratings that were less than 5% different. This supports our hypothesis that user and critic ratings are vastly dissimilar.

**Figure 6. Differences between user and critic preferences**



## Conclusion

Our hypotheses were supported by the data. This includes the ratings being different but not too different across the three major sites, IMDB, RottenTomatoes, and Metacritic. This also includes the fact that user and critic ratings are vastly dissimilar in each rating system.

Other interesting findings include the following:

In general, critic ratings were more consistent across the different websites. Figure 5 showed that critic scores were closely related compared to the user scores between different sites. This suggests that the user demographics on each site are quite different. This would be an interesting area for future research.

After analyzing the differences between user and critic preferences in Figure 6, it was seen that over 50% of the Top 100 Grossing Movies of All Time received low ratings by critics. In short, these movies may be good to watch for pure action-type entertainment but not for in-depth film appreciation.