**Choosing the correct review source.**

1. **Abstract**

The objective of this project is to compare gross movie earnings data with ratings and popularity rankings from various online resources. Our process involves extracting the Movie earnings and rating data from our sources, transform it into a format that is comparative and then load it to the MySQL workbench or MangoDB. The project aims to answer several possible questions as outlined below and will include a summative and comparative analysis.

1. **Introduction**

There are many websites online that offer movie reviews. However, how each website is unique in its review content is not known. This project is a small effort to compare the review ranking from different websites and conclude which review is linearly correlated with the gross earnings. The project is divided into three main components: Extract, Transform and load. For this, data is extracted from four different websites, transformed into the required format and then load it to the MongoDB. The analysis is done on Pandas IDE and plotting is done using 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 ranked films?
3. Do the film rankings differ between websites?
4. Are the film ratings for each website normally distributed? (Or skewed/bimodal)
5. What does the comparison between critic scores and user scores look like?
6. **Resources**

The websites that are used in this project are:

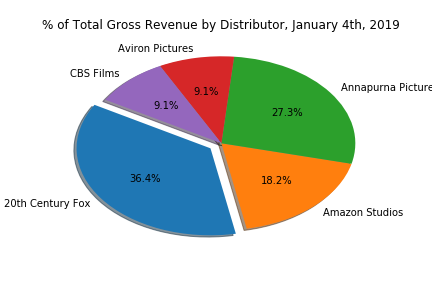
1. <https://www.the-numbers.com/>
2. <https://www.imdb.com/>
3. <https://www.rottentomatoes.com/>
4. <https://www.metacritic.com/>

Technologies Used are:

1. Pandas,Jupyter
2. Matplotlib
3. MongoDB
4. Beautiful Soup
5. **Methods 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

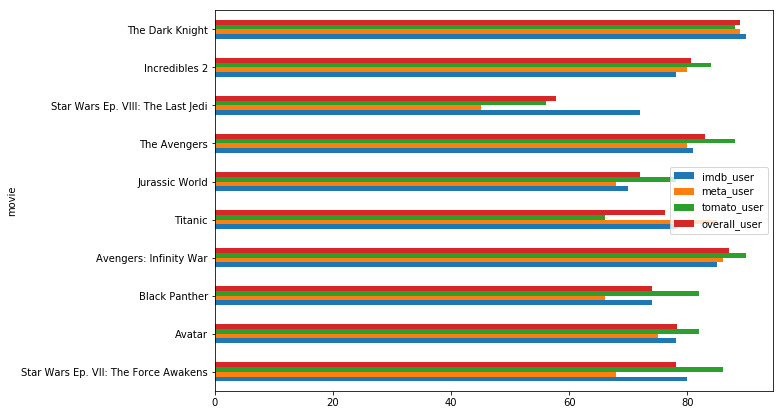


* 1. Comparative analytics from Metacritic, IMDB and Rotten Tomatoes data.

The data from Metacritic, Numbers and Rotten Tomatoes websites are extracted using web scraping, while imdb data is extracted using OMDB API. Since the aim was to do the comparison between the top-ranking movies from three different websites, the extracted data required rigorous cleaning to bring it to the form where merging is possible. First, data from Numbers website is extracted to get 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 scrapped form OMDB API and data from Rotton Tomatoes website is scraped. Data from four different website is then merged into a new data frame to have a collection of user ranking and critic ranking columns.

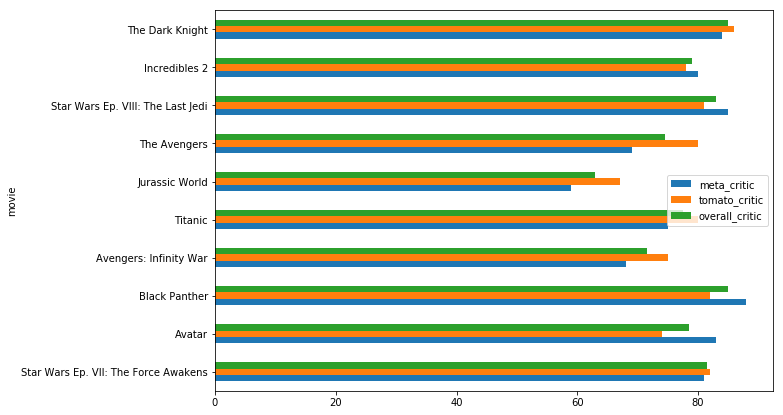
The merged data is plotted to see the comparison between different web sources. The Fig.2 below shows the comparison of user ranking data from different web sources.

Fig.2 Comparing user ranking from different web sources



The Fig 3. Below shows the comparison of critic ranking from different web sources.

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 is critic ranking 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.

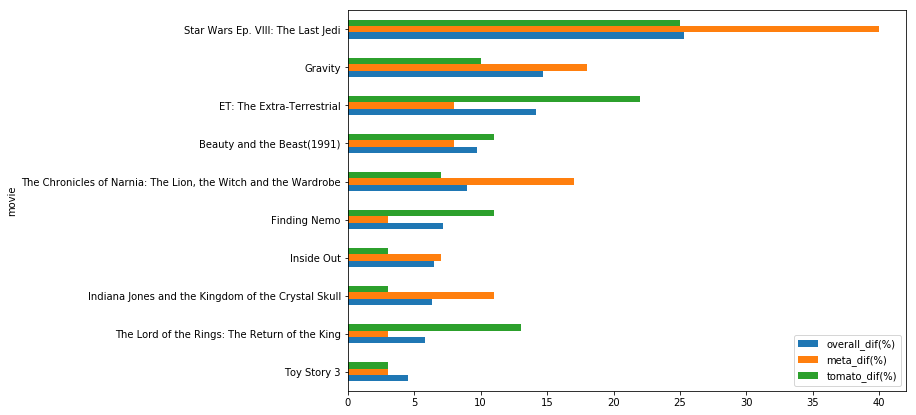
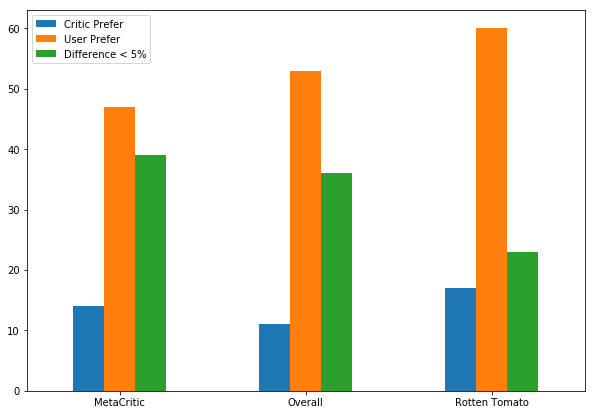


Figure 4 below shows the comparison of difference between user and critic score. The plot concludes that the difference between them is less than 5%.



**Conclusion:**