**An Insightful Look at IMDB Data between 2006 and 2016**

Team Members: Mahtab Karimipour, Luke Gonzales, Molly Eskelson

**Motivation & Summary**

\* Define the core message or hypothesis of your project.

\* Describe the questions you asked, and \_why\_ you asked them

\* Describe whether you were able to answer these questions to your satisfaction, and briefly summarize your findings

- What do people like in movies?

- Everyone has their favorite movies whether from their favorite trope, underlining message, or action scene. It’s the little bits of something that make a movie great. While this dataset doesn’t categorize and document every possible variable that a movie can have, this dataset does give some insights on what people like in a movie.

We found this dataset on Kaggle.com. The data set is comprised of 1,000 of the most popular movies on IMDB spanning from 2006 to 2016. The data fields included are Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascore.

* Columns
  + Rank- Movie rank order
  + Title- The title of the film
  + Genre- A comma-separated list of genres used to classify the film
  + Description- Brief one-sentence movie summary
  + Director- The name of the film's director
  + Actors- A comma-separated list of the main stars of the film
  + Year- The year that the film released as an integer.
  + Runtime (Minutes)- The duration of the film in minutes.
  + Rating- User rating for the movie 0-10
  + Votes- Number of votes
  + Revenue (Millions)- Movie revenue in millions
  + Metascore- An aggregated average of critic scores. Values are between 0 and 100. Higher scores represent positive reviews.

Some limitations of this data include that it only give insights to the voters from the IMDB website. It is also unknown when the revenue data was collected.

The dataset is a packaged dataset; not everything is there. There had to be a cut off somewhere. Unknown where that cut off was. USA films? First 100 films/percentage.

We can see just by going onto IMDB website that there are more films shown 10 2016 then we have in our dataset.

Going further with this study we would use the websites API/JSON to build our dataset.

**Questions & Data**

Some Questions we can answer with this dataset are:

* How much revenue did each movie make per a year?
* How much revenue did each genre make?
* What are the top 15 movies according to revenue and year?
* What is the average rating for movies? (by each genre?) then what would be a good or bad movie rating score (rating vs. Metascore)?
* Does the runtime of a movie affect the movie's rating?
* Is there a relationship between a movies revenue and the rating users give it?
* Analyzing the difference between meta score(critiques) and users(rating)
* Who are the top directors and actors?- Which actors and directors show up in the most films during these 10 years?

**Data Cleanup & Exploration**

\* Describe the exploration and cleanup process

\* Discuss insights you had while exploring the data that you didn't anticipate

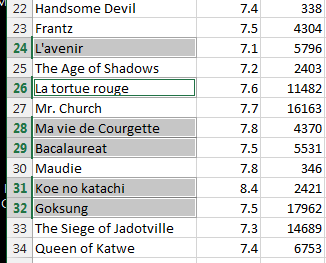
\* Discuss any problems that arose after exploring the data, and how you resolved them

\* Present and discuss interesting figures developed during exploration, ideally with the help of Jupyter Notebook

* Looking closely at the data some questions arose. What is the difference between music genre and the musical genre? La La Land vs. Les Mistrals vs. Mama Mia.

Interesting figures:

Created a separate excel sheet with high voter ratings by – in other words, underrated films. Not unsurprisingly a few of the movies to make this list have foreign titles.



**Data analysis**

\* Discuss the steps you took to analyze the data and answer each question you asked in your proposal

\* Present and discuss interesting figures developed during analysis, ideally with the help of Jupyter Notebook

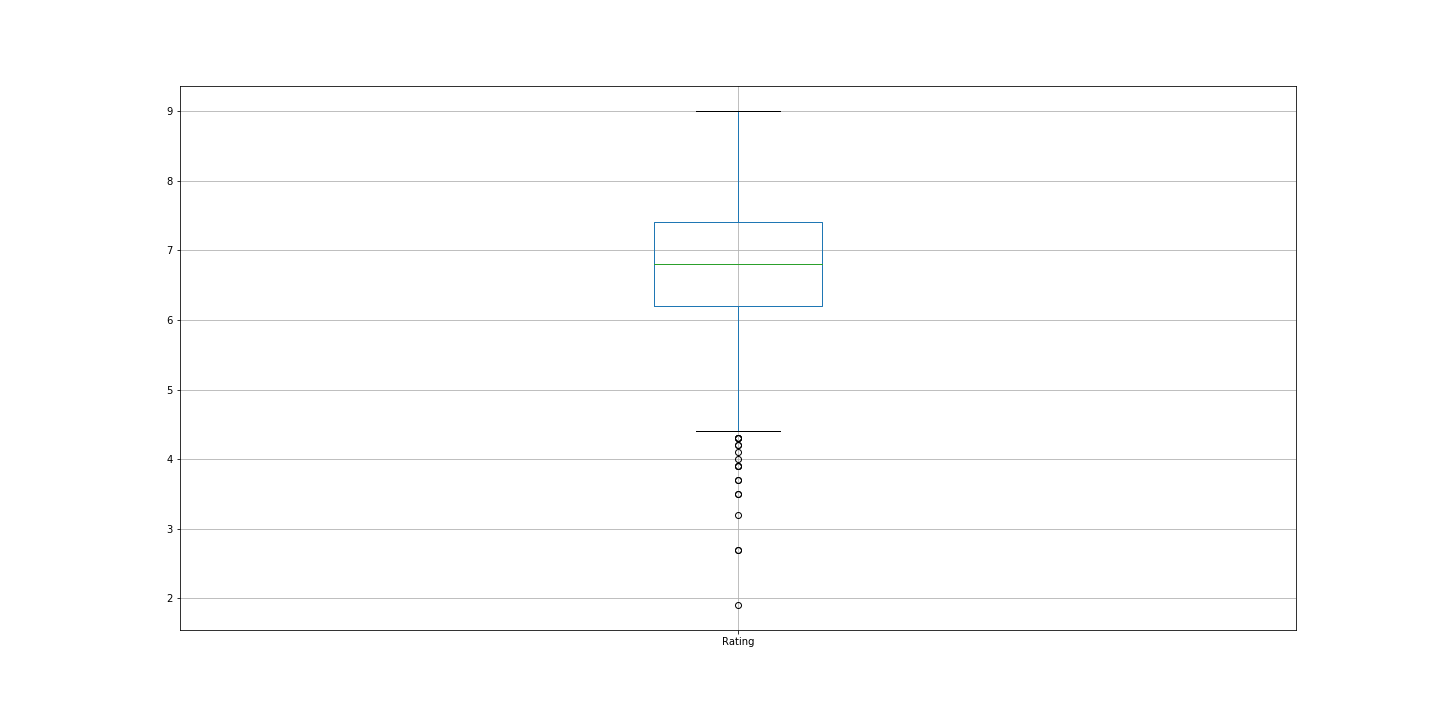
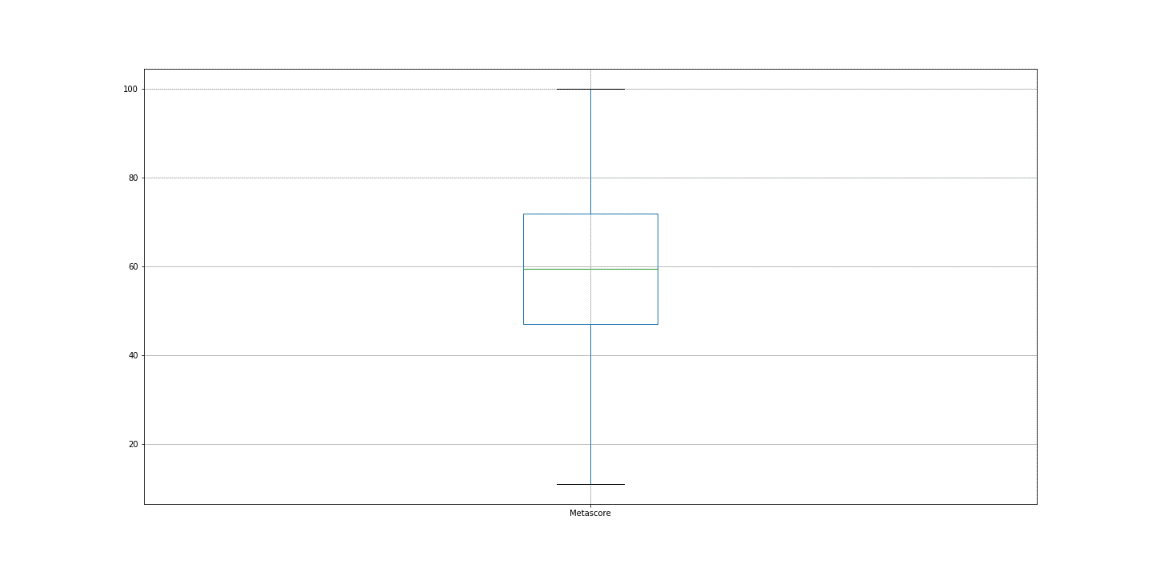
**Statistics**

In this section, we want to compare the two-rating system that we have, which is Metascore and rating.

As we know, movies are ranked by the difference between critic scores and user scores. Metascore is a weighted average in that we assign more importance, or weight, to some critics and publications than others, based on their quality and overall stature on the other hand rating is a weighted average that each user assigned to the movies.

Based on this data set, the overall average for Metascore is 5.898, and the overall average for the rating is 6.723.

From these box plots, the median (middle quartile) marks the mid-point of the data and is shown by the line that divides the box into two parts. Inter-quartile range. The middle “box” represents the middle 50% of scores for the group.

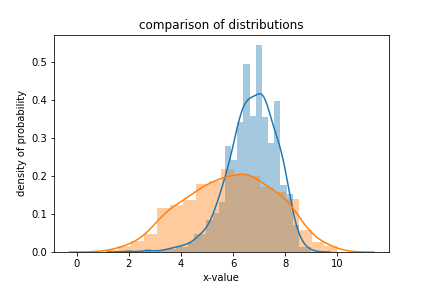
Upper quartile and Lower quartile are also specified on the plots. 

(in this study we divided all the Metascores to 10 to be able to compare the Metascore and rating)

The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. A t-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.

In these two graphs, we can see that the distribution of these two rating systems is slightly different, they both have a normal distribution but with different average. We run the t-test for these to the rating system for comparison.

The p-value was smaller than α ( p = 1.47919e-96) confirm that the two rating systems are different, meaning that we were able to reject the null hypothesis.



**Discussion**

* \* Discuss your findings. Did you find what you expected to find? If not, why not? What inferences or general conclusions can you draw from your analysis?
* Limited data-

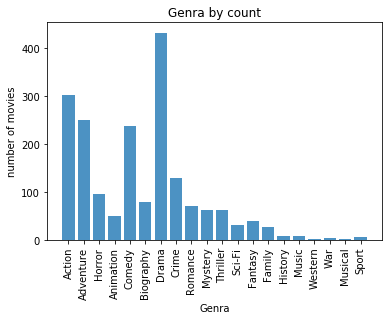
The list of directors that made the most movies during these 10 years.

|  |  |
| --- | --- |
| directors | movies |
| Ridley Scott | 8 |
| David Yates | 6 |
| M. Night Shyamalan | 6 |
| Michael Bay | 6 |
| Paul W.S. Anderson | 6 |
| J.J. Abrams | 5 |
| David Fincher | 5 |
| Antoine Fuqua | 5 |
| Christopher Nolan | 5 |
| Denis Villeneuve | 5 |

The list of actors that were involved the most during these 10 years.

|  |  |
| --- | --- |
| Actors | movies |
| Christian Bale | 11 |
| Mark Wahlberg | 11 |
| Will Smith | 9 |
| Denzel Washington | 9 |
| Brad Pitt | 9 |
| Leonardo DiCaprio | 9 |
| Jake Gyllenhaal | 9 |
| Adam Sandler | 9 |
| Johnny Depp | 8 |
| Tom Hanks | 8 |

The bar graph below shows us the number of movies created in each genre.

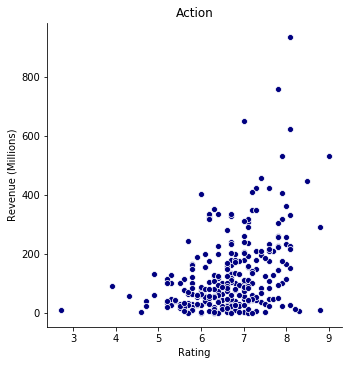
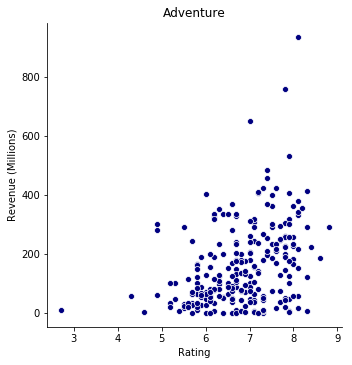
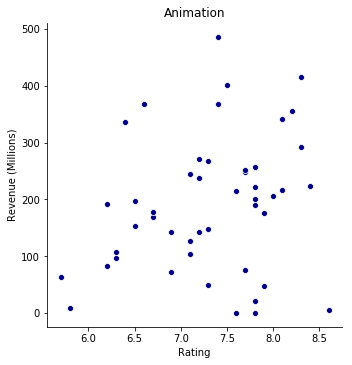


Is there a relationship between a movie’s revenue and the rating moviegoers/critics give it?

It can, but not necessarily as significant as you might think.

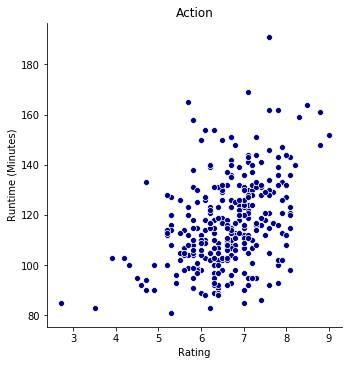
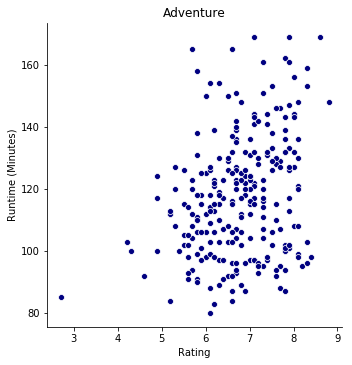
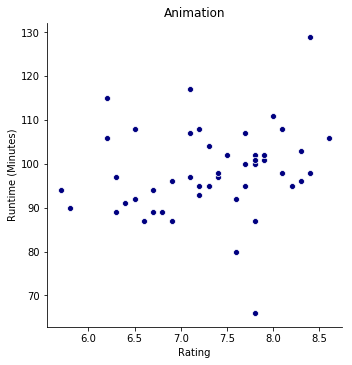
It is easy to see how anyone might assume that movie ratings are connected to movie revenues, with potential audiences more likely to buy tickets for a movie with a higher rating. However, that is not the whole story. Moreover, as many people in the movie industry believe, the correlation between movie ratings, critical opinion, marketing strategies, and actual box office returns is complicated. It’s not a simple cause-and-effect situation.

An excellent rating indicates strong critical consensus, and that can be good for films, smaller films in particular. However, when it comes to blockbusters, it is much less clear how much a film’s rating affects its box office total. A positive rating, for example, does not necessarily guarantee a film will be successful. The scatterplots below illustrates how the revenue of movies in three of the most popular movie genres does not necessarily increase because of higher ratings or decrease because of lower ratings.

Does the runtime of a movie affect the movie's rating?

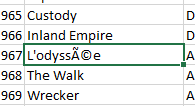
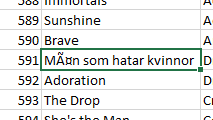
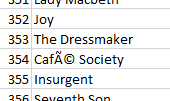
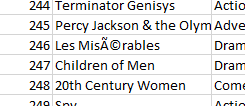
Similar to the previous question, the runtime of a movie does not positively or negatively affect a movie’s rating. Once again, the scatterplots below show the correlation between a movie’s runtime and rating in three of the most popular movie genres.

**Post Mortem**

Discuss any difficulties that arose, and how you dealt with them

\* Discuss any additional questions that came up, but which you didn't have time to answer: What would you research next if you had two more weeks?

* Initially, we were doing a project on yelp dataset but after experiencing difficulties with both the excel and JSON versions of the dataset we unanimously agreed to move on to another project.
* We could have normalized the CSV file, so situations like Les Mistrals and the use of dates as titles weren’t an issue.
* 

In conclusion, do our rankings match up with their ranks?

Do people put their money where their votes are?

* Our next research question was do these results/rankings stand true across all major user voting websites (Rottentomatos)