**An Insightful Look at IMDB Data**

**between 2006 and 2016**

Team Members: Mahtab Karimipour, Luke Gonzales, Molly Eskelson

**Motivation & Summary**

-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.

How do people define a good movie? How can we tell a good movie from bad movies?

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.

* + 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 can only give insights to the population that uses and rates movies on the IMDB website. It is unknown at what moment in time the revenue data was taken, though it appears that the data was collected at the end of 2016 as films that were released on dates such as December 16th or later have little or no data for fields such as revenue or votes.

The dataset is a packaged dataset, and after a cursory fact checker, not everything is there for the claimed period. We understand that there has to be a cutoff, but we are unable to determine where. It is unclear if the cut off was films released in the USA, films with ranking in (x) percentile, etc.

That is why if we were to continue with this project we would collect our data from the IMDB website through JSON.

**Questions & Data**

Some Questions we can answer with this dataset are:

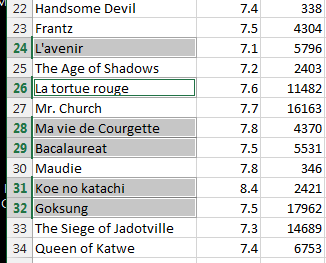
* How much revenue do movies make per a genre? Year?
* What are the best/top 15 movies according to revenue? Year?
* What is the average rating for movies? (by each genre?)
* What would be a good or bad movie rating score?
* Does the runtime of a movie affect the movie’s rating?
* Is there a relationship between a movie’s revenue and the rating users give it?
* Analyzing the difference between meta score(critics) 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**

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

Interesting findings:

We also created a separate excel sheet with high voter ratings by votes less than 20,000– in other words, underrated films. Not unsurprisingly a few of the movies to make this list have foreign titles.



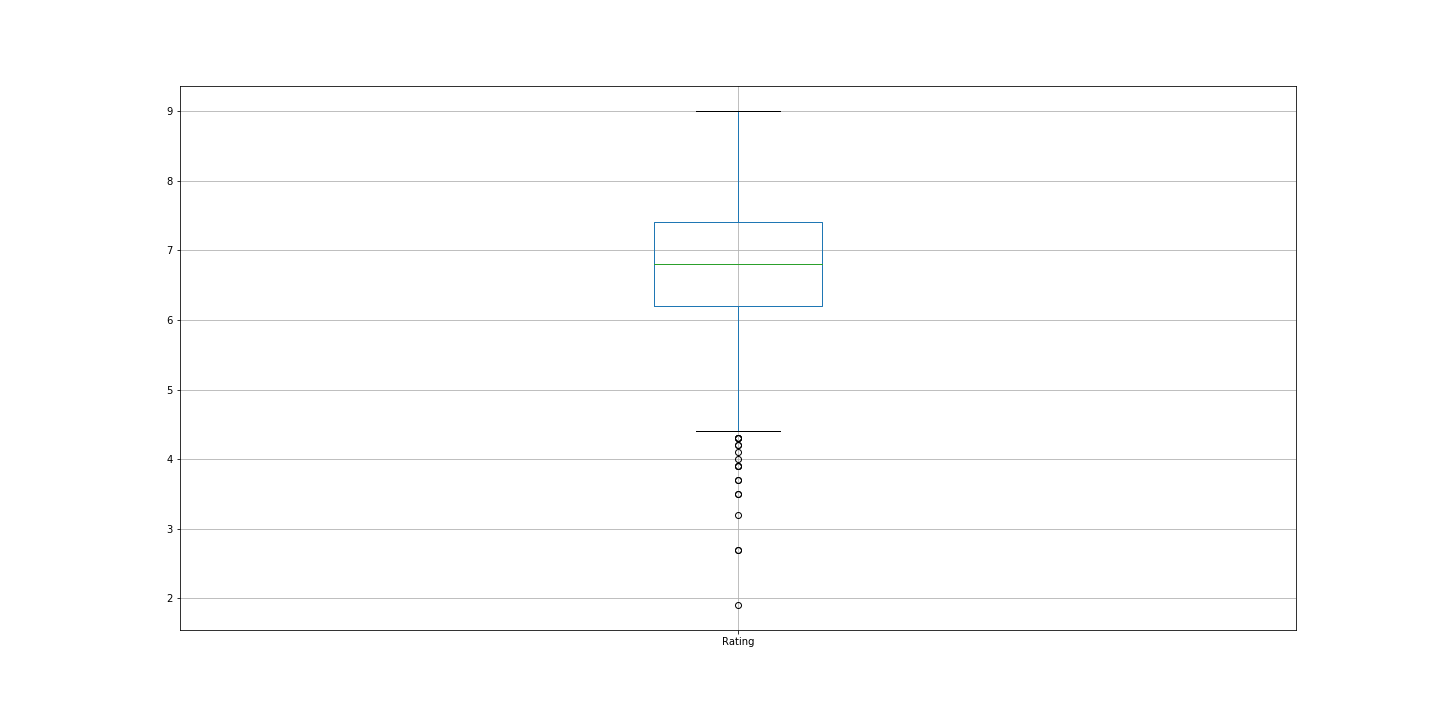
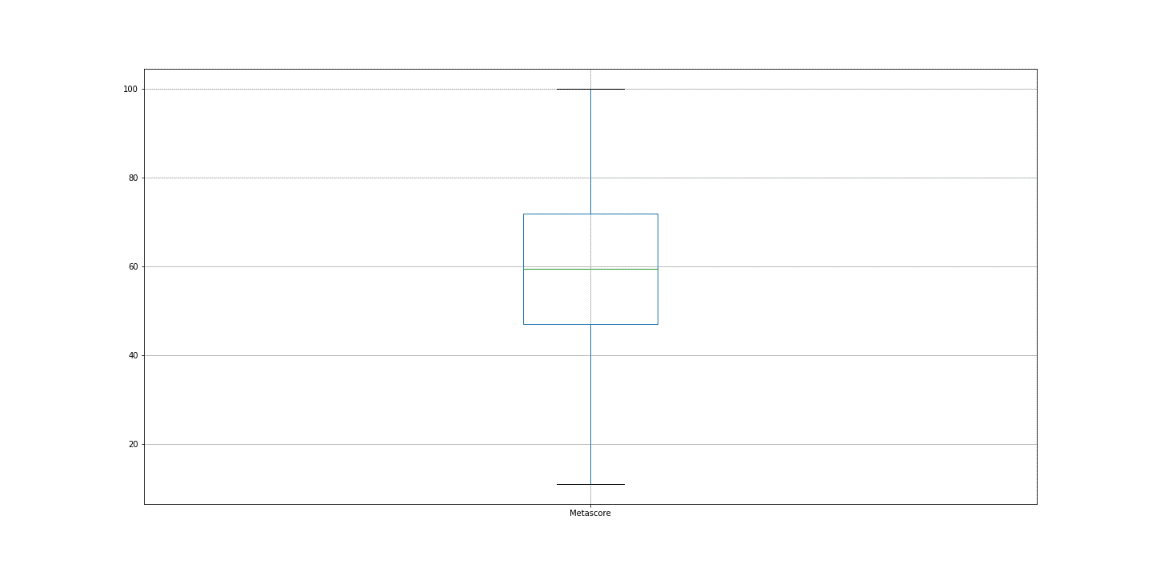
**Statistics**

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.

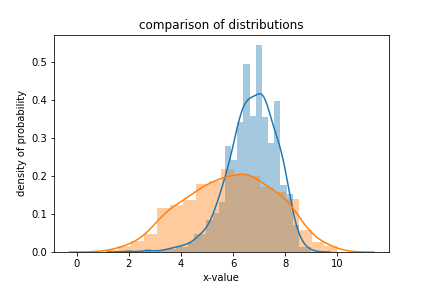
Upper quartile and Lower quartile are also specified on the plots. 

(in this study we divided all the Metascore 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**

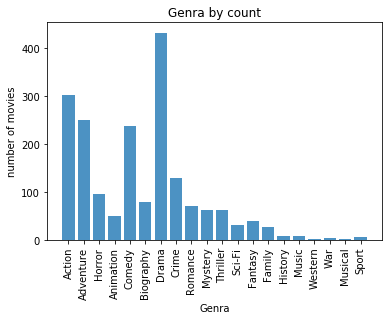
The list of directors that made the most movies in 10 years period.

|  |  |
| --- | --- |
| 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 in 10 years period.

|  |  |
| --- | --- |
| 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.

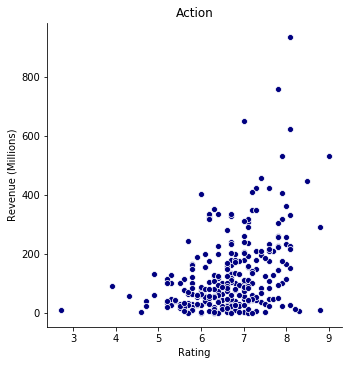
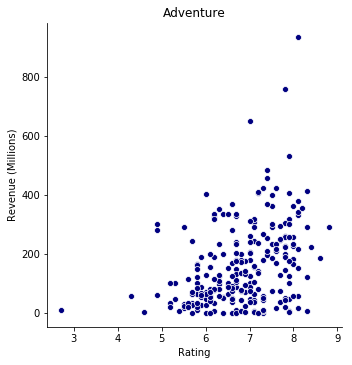
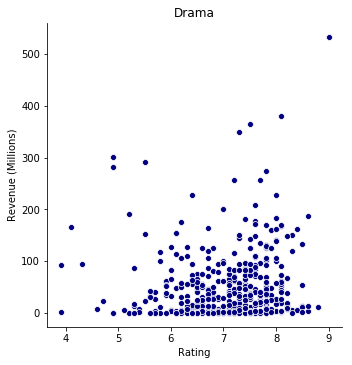
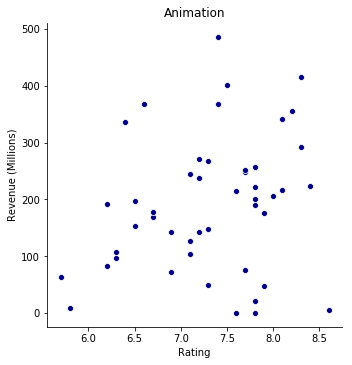


Is there a relationship between a movie’s revenue and the rating moviegoers/critics give it?

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

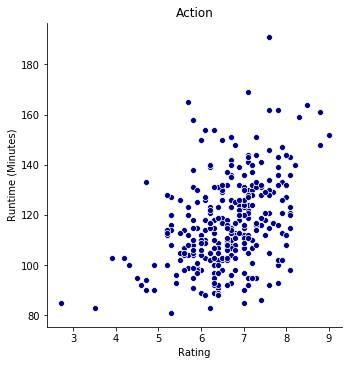
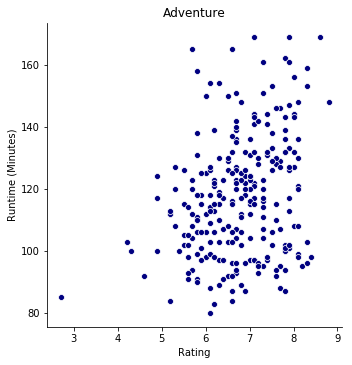
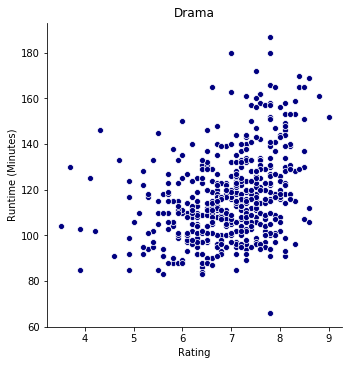
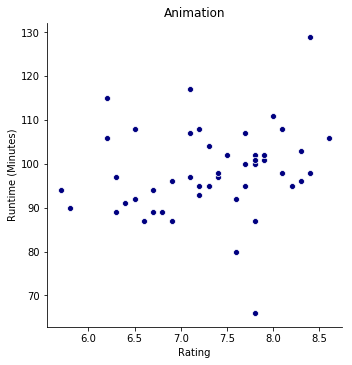
It is easy to see how anyone might assume that movie ratings are connected to movie revenues, with potential audiences more likely to buy tickets for a movie with a higher rating. However, that is not the whole story. Moreover, as many people in the movie industry believe, the correlation between movie 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 four of the most popular movie genres does not necessarily increase because of higher ratings or decrease because of lower ratings.

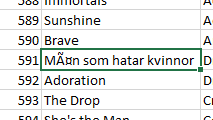
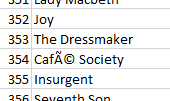
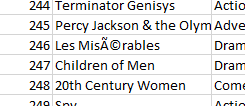
 

Does the runtime of a movie affect the movie's rating?

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 four of the most popular movie genres.

**Post Mortem**

* Originally, we were doing a project on yelp dataset but after experiencing difficulties with both the Excel files 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 Misérables and dates (aka forging keys) as titles weren’t an issue.
* 
* The Movies thrown out in the last month of the year are not included in the release date field.
* Our next research question was do these results/rankings stand true across all major user voting websites (rotten tomatoes, Fandango, etc.) We would need to cross-reference other website rating systems against our dataset from IMDB website.

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)