,



Group 3: movie analysis report

Naheem Olaniyan, Nishchint Upadhyaya, Pierce Nordeng, Rakhi Karki, Zachary Vincent



April 28, 2019

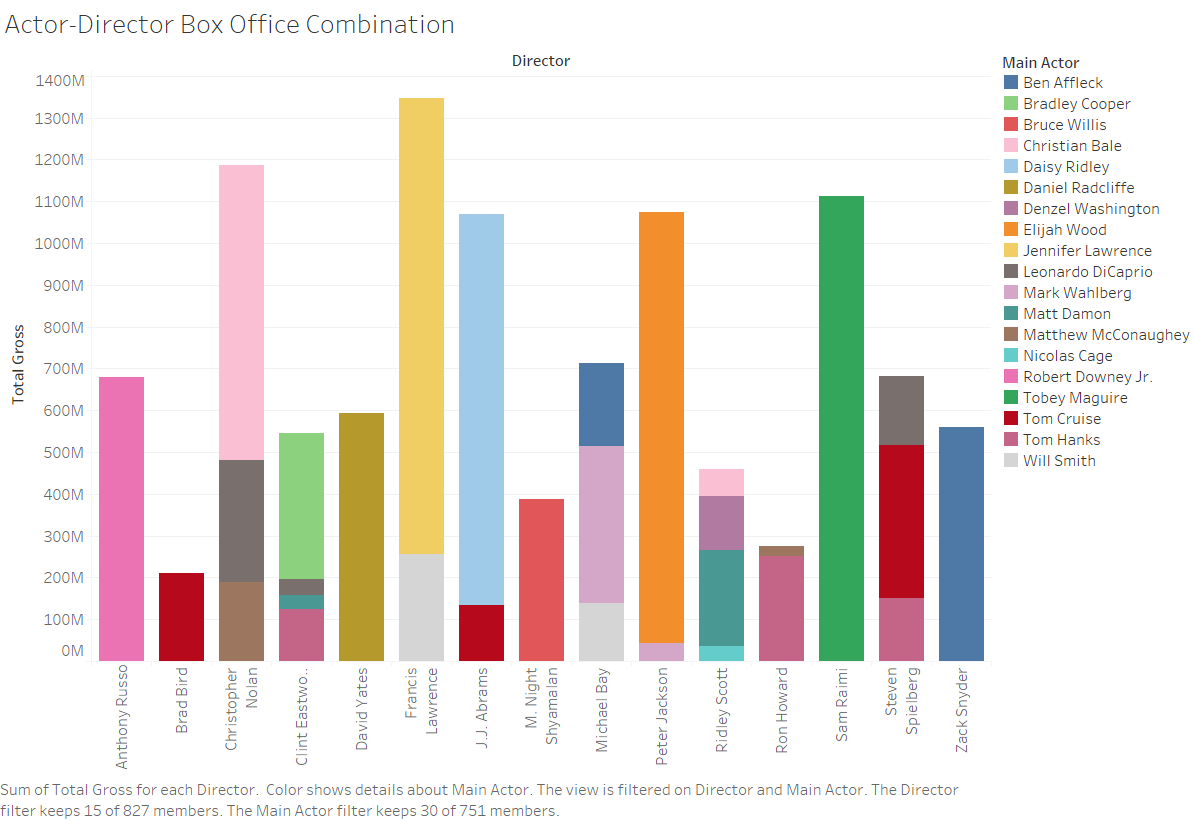
Business Problem

* We have a lot of cases today where a certain movie released is not as successful as anticipated or does not live up to the audiences’ expectations. In this study, we will try to see the correlation between variables such as movie release date, movie studio, actor, director and movie genres with the total box office to help determine how these factors can ensure better success of a movie.
* By looking at a combination of release date, actors, directors, studios to find a correlation with the total box office income of a given movie which can help determine the success of the movie.
* The data can be used by movie studios to help determine which actor-director combination along with the month of release, can bring success to a movie they plan on releasing.
* The data can also be used my smaller theaters with fewer screens, to help determine which movies can be shown for better revenue.

Research Methodology

* Methods to solve the business problem:
  + Data collection: We collected data regarding the top 100 highest grossing movies for the last 20 years (1999 – 2018). We pulled box office gross revenue, opening week revenue, release date, number of theaters it opened in, total number of theaters it opened to from boxofficemojo.com. We also got directors and main actors from imdb.com. Lastly, we queried ‘tmdb’ api to get genre and their rating system.
  + We used Tableau to create visualizations to go more in-depth in to our business problem and to find potential solutions.
* Data Cleaning:
  + We used Microsoft Excel and Microsoft Access to pair values like ratings and genre which were not available on our initial dataset acquired from BoxOfficeMojo and imdb. We also used Excel to search and remove any duplicates that were created in the pairing process.
* Model:
  + Using Jupyter we were able to make a model. Doing this took some time but in the end we found the most appropriate variables to predict total gross.

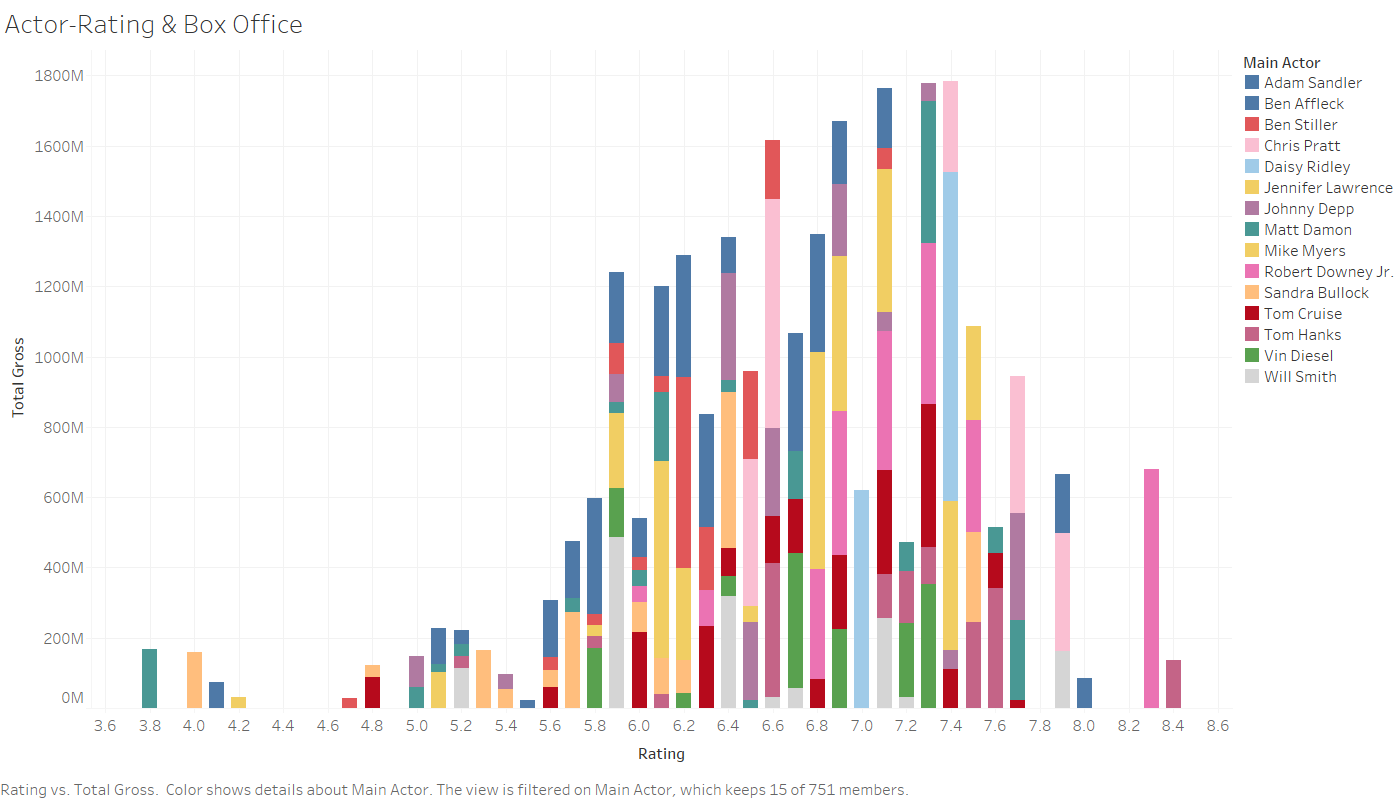
Results and Findings



This visualization tells us how a combination of actor and directors impacts the total box office income of a movie. It calculates the total money generated by each of the actor-director combinations. This visualization gives us the top 15 actor-director combinations. The x-axis has the names of the directors and the color code tells us the actor paired up with that director. The y-axis has the box office income.

We see that the combinations of Jennifer Lawrence & Francis Lawrence, Sam Raimi & Toby Maguire, Peter Jackson & Elijah Woods and Anthony Russo & Robert Downey Jr are some of the combinations where the movie generated the most revenue.

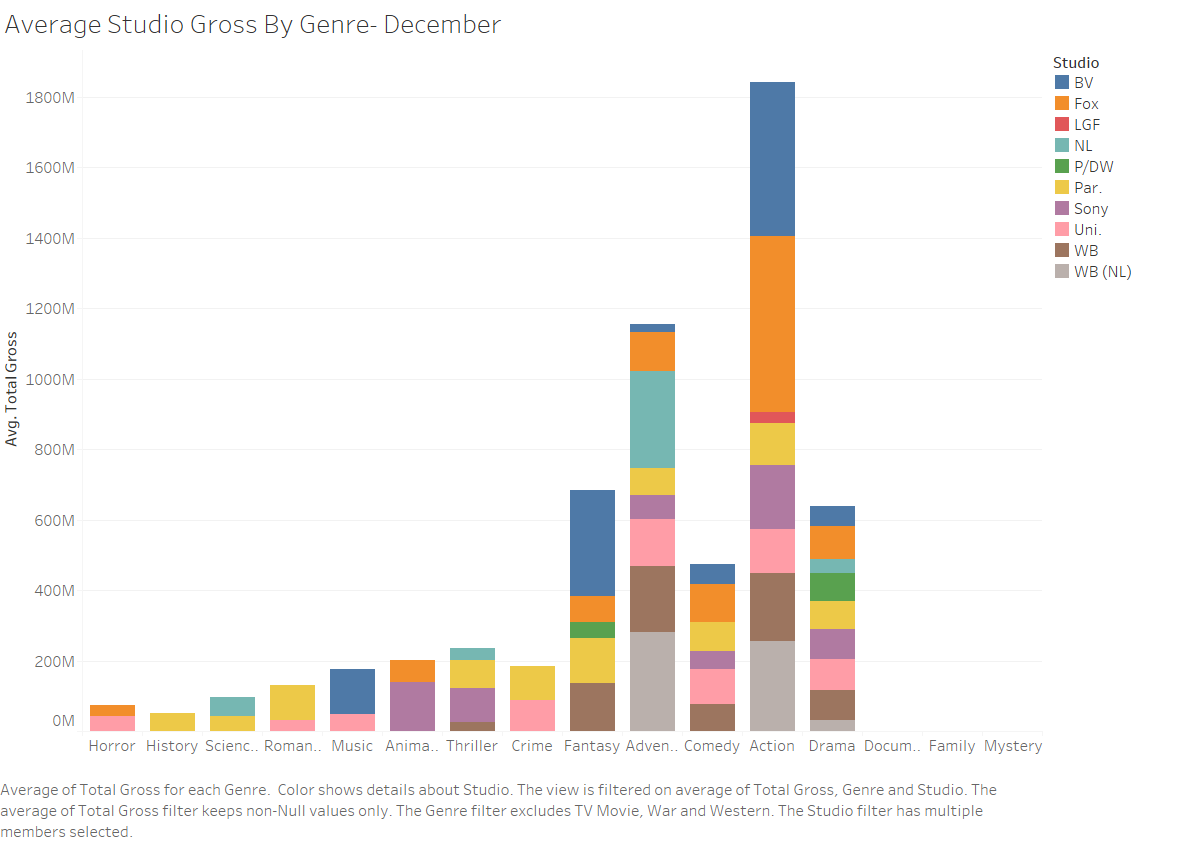
This shows that these actor-director combinations are a good choice if a studio is planning on releasing a movie. Due to the reputation and history of these combinations, the movie will be able to generate much more revenue.



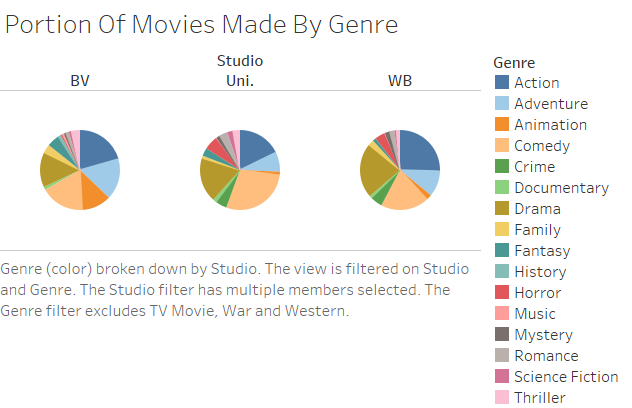
This visualization talks about the box office income for movies with low and high ratings along with the actors for those movies. The x-axis has all the ratings and the color code signifies the actors. The y-axis has the box office income. The visualization has the ratings for the top 15 highest box office earning actors.

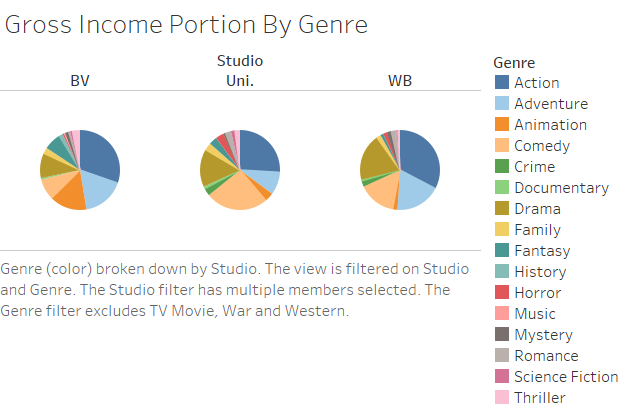
We see that for some movies with ratings like 6.5, actors like Chris Pratt still managed to earn the movie up to $400 million. The same case for rating of 6.6 where a movie(s) starring Chris Pratt made up to $650 million despite the poor ratings. Another good example is for a movie with a rating of as low as 5.9 where the movie(s) starring Will Smith still managed to make up to almost $500 million.

Therefore, we can say that even if a movie has low ratings, it can still manage to be a box office hit due to the actors it has cast. People will come to watch a movie just to see the performance of actors they love.

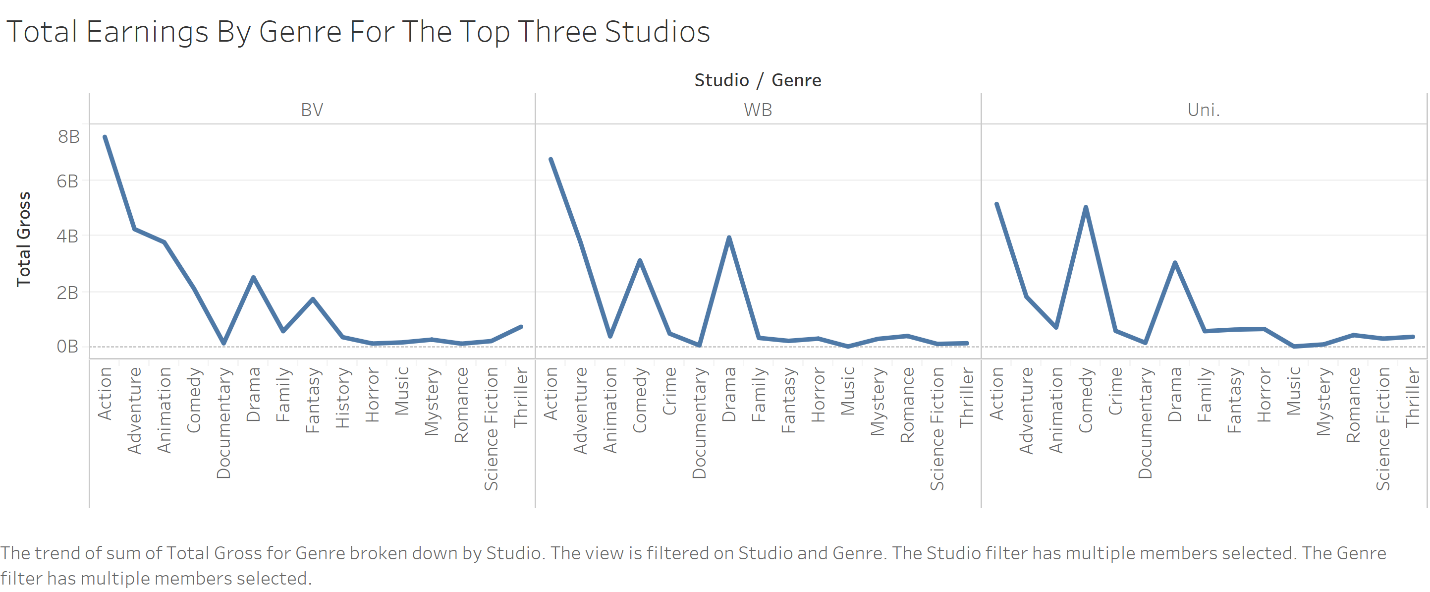


The graph above is one of twelve available for the average revenue per genre over the twenty-year dataset. We chose to include this one as it shows the relationship between month and average gross revenue. This is similar to the May-July graphs show a similar trend with a few genres like action and adventure grossing more in those months. This graph is narrowed down to the top 10 box office grossing studios as in our dataset we have many studios that appear around once or less times per.



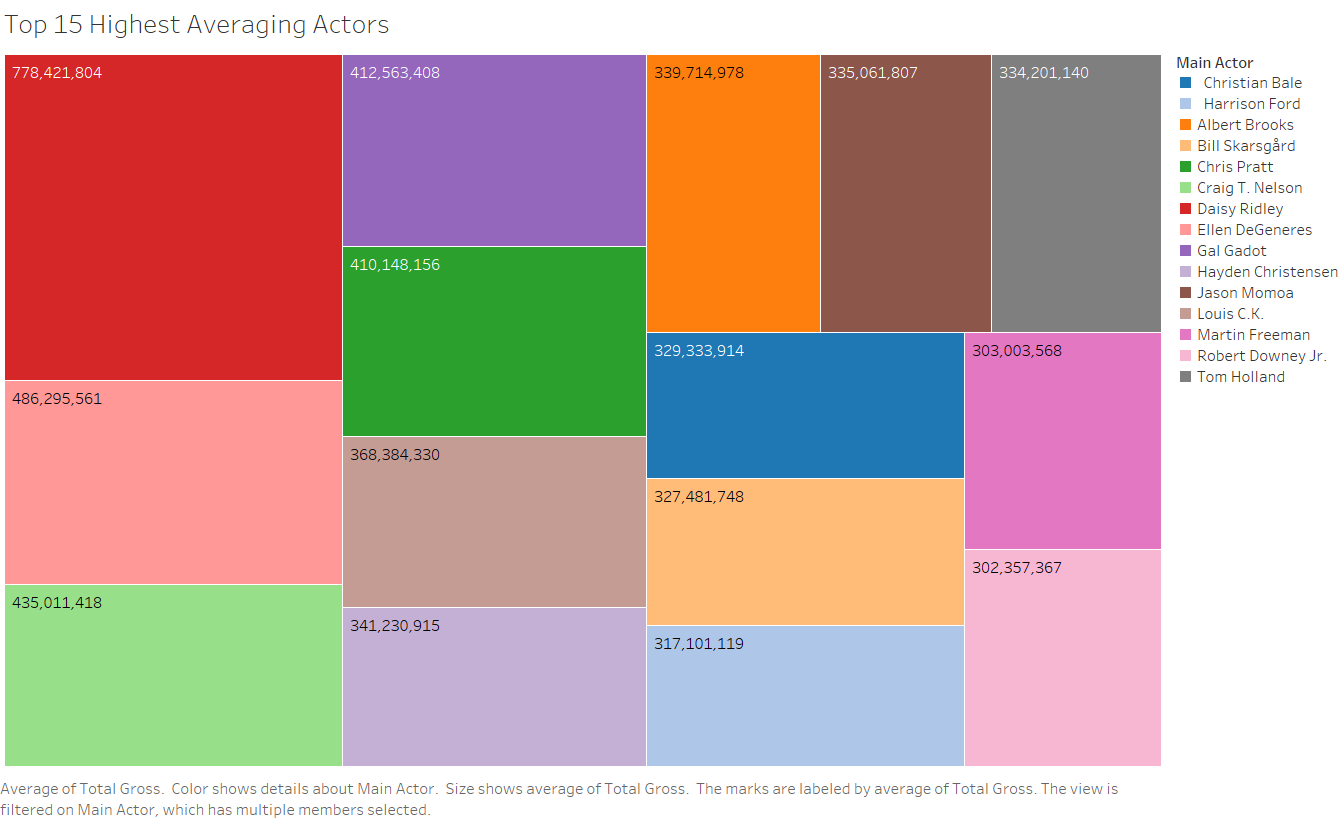


The two charts pictured above depict the total gross revenue of each movie genre for the top three grossing studios. The studios depicted are BV (Buena Vista), Universal studio, and Warner Brothers. We chose to exclude Westerns, TV Movies and War movies because neither of those had a large portion or impact on the revenue. The topmost graph, the portion graph, is the count of appearance from each genre for each studio. One thing that we can notice is that although animation is the fifth most made genre for Buena Vista (Disney) it is the third largest grossing genre for Disney. For BV action movies also have a similar effect in that they generate a larger portion than they occupy for count of the genre. Both Buena Vista and Warner Brothers generate their highest gross revenue from action movies while Universal Studio’s largest gross revenue was produced by the comedy genre. The graph pictured below gives the total gross revenue generated by each genre a numerical value.

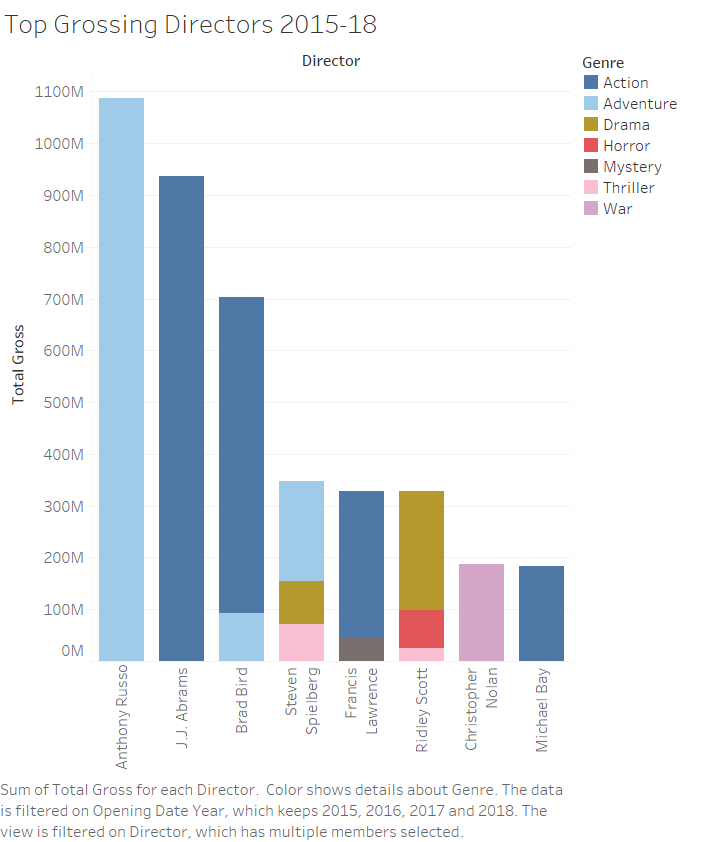
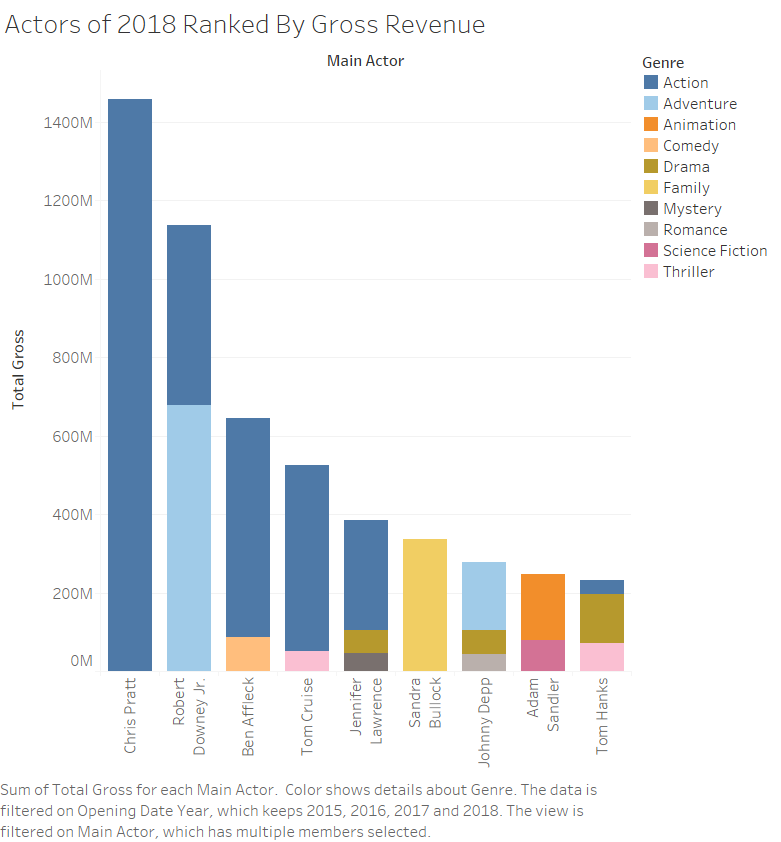


The visualization shows the filtered data studio and genre affecting the total gross. We can see that the genre action movie by studio BV is contributing to generate the highest total gross. Similarly, for the other studio WB the genre action is on top in terms of making high total gross followed up by the genre Drama. Lastly the studio Uni has genre action on top with not much difference with genre comedy. After analyzing all three diagram we can find one common thing is the genre action is directly related to the highest total Gross.

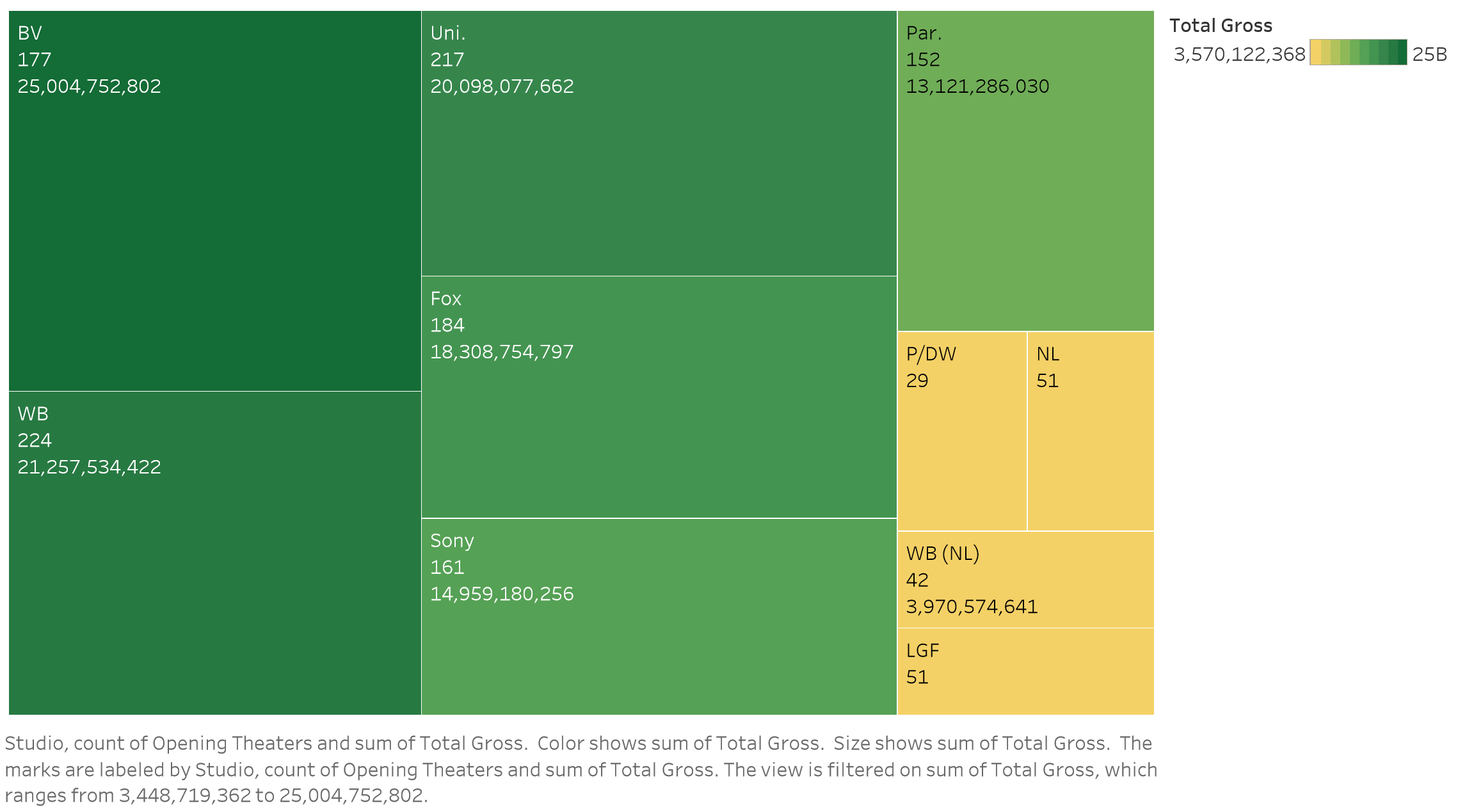
This diagram will help directors to decide on investing capital on the particular genre movie which turns out to be action movie according to the diagram. Another application of this insight is to smaller theaters in determining what genres from what studios to show that might generate the most revenue.



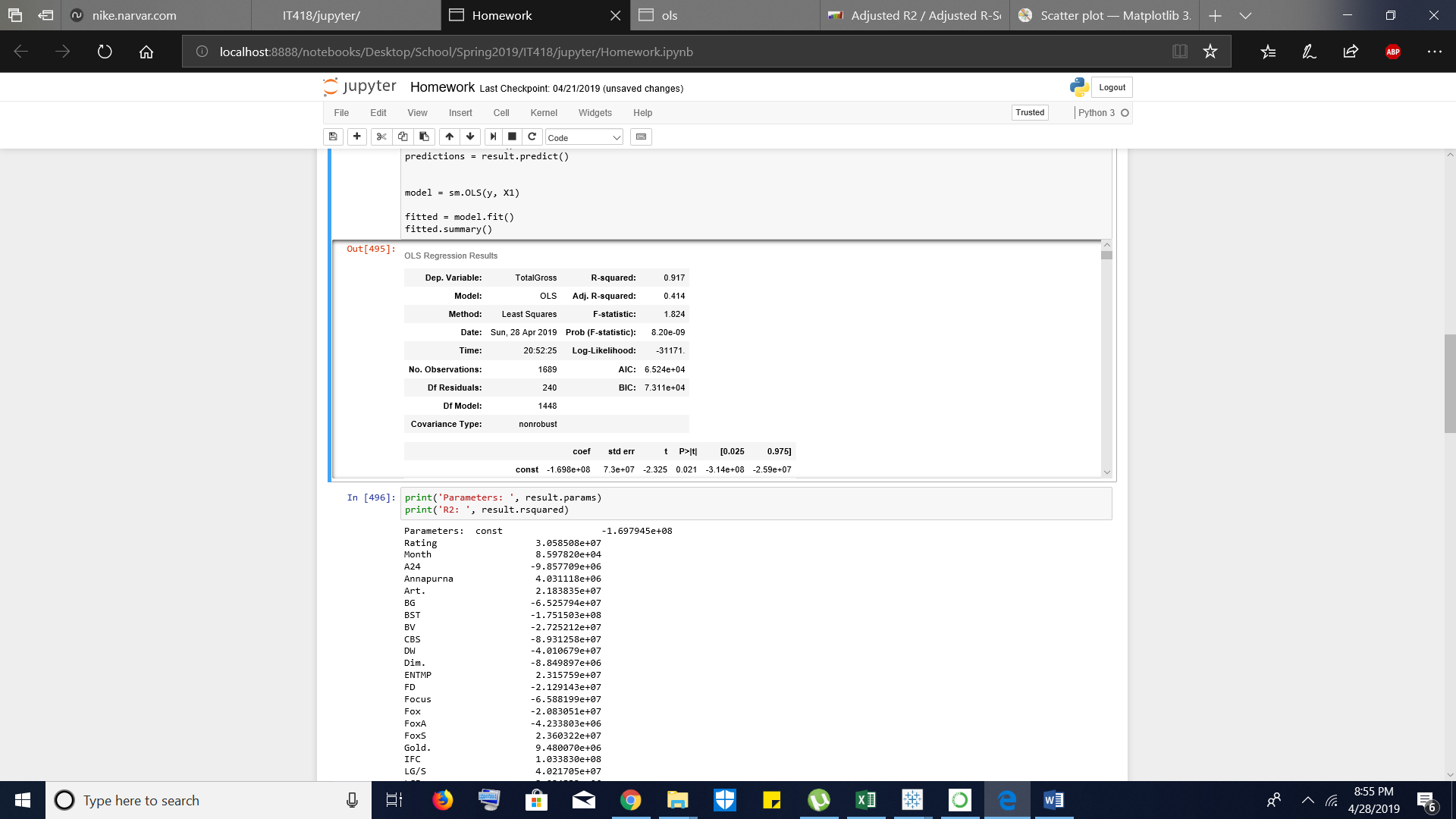
In the diagram we can see that the top actor among the top 15 Highest averaging is Daisy Ridley with the total gross of 778,421,804. The second and third position is followed up by Ellen DeGeneres and Craig T. Nelson respectively. Looking at this diagram any director can analyze and decide on which actor to choose for their upcoming new movie. Not only director but the studio and theatre can make the decision on making their business to make maximum revenue by selecting top actor for the movie and the theater by casting a movie played by the top actor.



These two graphs show the total gross breakdown by main actor and director for the last four years. We can see that Chris Pratt had the highest gross for actor and that Anthony Russo had the highest gross for director. The two graphs are color coded by genre so one assumption that we could make is that an action movie featuring Chris Pratt and J.J. Abrams might do well as they are the highest earning action actors/directors. One other combination that might do well is featuring Jennifer Lawrence in a Steven Spielberg or Ridley Scott directed drama. An adventure movie directed by Anthony Russo featuring Robert Downey Jr. might gross highly as well.



The visualization above shows the studio count of movie and their total gross in US dollar. It shows that there is little positive correlation between number of movies made by a studio and its total gross. BV is ranked 4th by number of movies produced but ranked 1st by total gross. Also, P/DW produced the least number of movies but ranked 7th by total gross. The major factor is the genre. We noticed that Buena Vista (Disney) held majority of the gross income for the animation genre which might have played a role in boosting them to the total leader gross domestic income.



Using Jupyter Notebook we were able to create a Logistical Regression Model. With this we were able to get a R-squared of .917 and an Adjusted R-squared of .414. Considering that the R-squared is way higher than the Adjusted R-squared it shows that some of the values tested were not useful. The values we determined that were best for the Adjusted R-squared were Ratings, Month released, Directors, Actors, Genres and Studios. Every variable other than Ratings and month released were converted to numerical data. We were also able to see that the predictions were pretty accurate using print('Predicted Values :',fitted.predict()).

Limitations

* One limitation we encountered was that a majority of the rating systems for movies refuse access for non-official projects. These include Google Movie Reviews, IMDB, Rotten Tomatoes and such. The site we used TMDB might not have been the best review source as it was user inputed but was our only choice.
* Not all movies had ratings or were able to be queried in the TMDB API so we lost about 300 movies records.
* Finding other independent variables was difficult so we had to Frankenstein different data sources together to form our monster.
* Movies had multiple genres ranging anywhere from 1 to 4 so we chose the first one.
* Same thing with ratings as with actors and directors, we ended up choosing the first one.
* TMDB’s rating can be questionable at best. Which might have had an effect on some of our visualizations.