Intelligence Analytics Challenge 4.0

Project Report

Forecasting Box Office Collections

Prepared for:

Intelligence Analytics Society

Prepared by:

Shubham Murari | UTD

Maruthi Nandan Pulletikurthi | UTD

Teenaz Ralhan | UTD

Jinglin Zhao | UTD

March 26th, 2019

Team Spartans
March 26th 2019

Table of Contents

1. Description

We have had a positive experience working with this project. As part of our analysis, we used Python to read in the data and manipulate it such that we could bring it to the stage of analyzing it to answer the questions in a meaningful manner and make meaningful inferences from the data regarding various types of movies and their box office performances.

One of the first tasks that we performed was to put both the sheets in the given excel workbook together in a data frame and to clean the data. As part of the data cleaning process, we dropped those columns that did not have any title and removed those columns which more than 90% of values as N/A's.

In order to make the data more viable in terms of reading and analyzing, we split columns like director, actors, and genre to several columns based on the number of comma-separated values in each of the records.

We also changed the format of data fields like runtime, release date, and dvd release date so that each facet of each column would be more readable. Splitting the release date into date, month, and year has helped us to determine the effect on movies of their release dates.

Determining the month of release for a movie helps to glean the impact of the time of year, for example movies that release during the summer or during the holiday season tend to perform well in the box office.

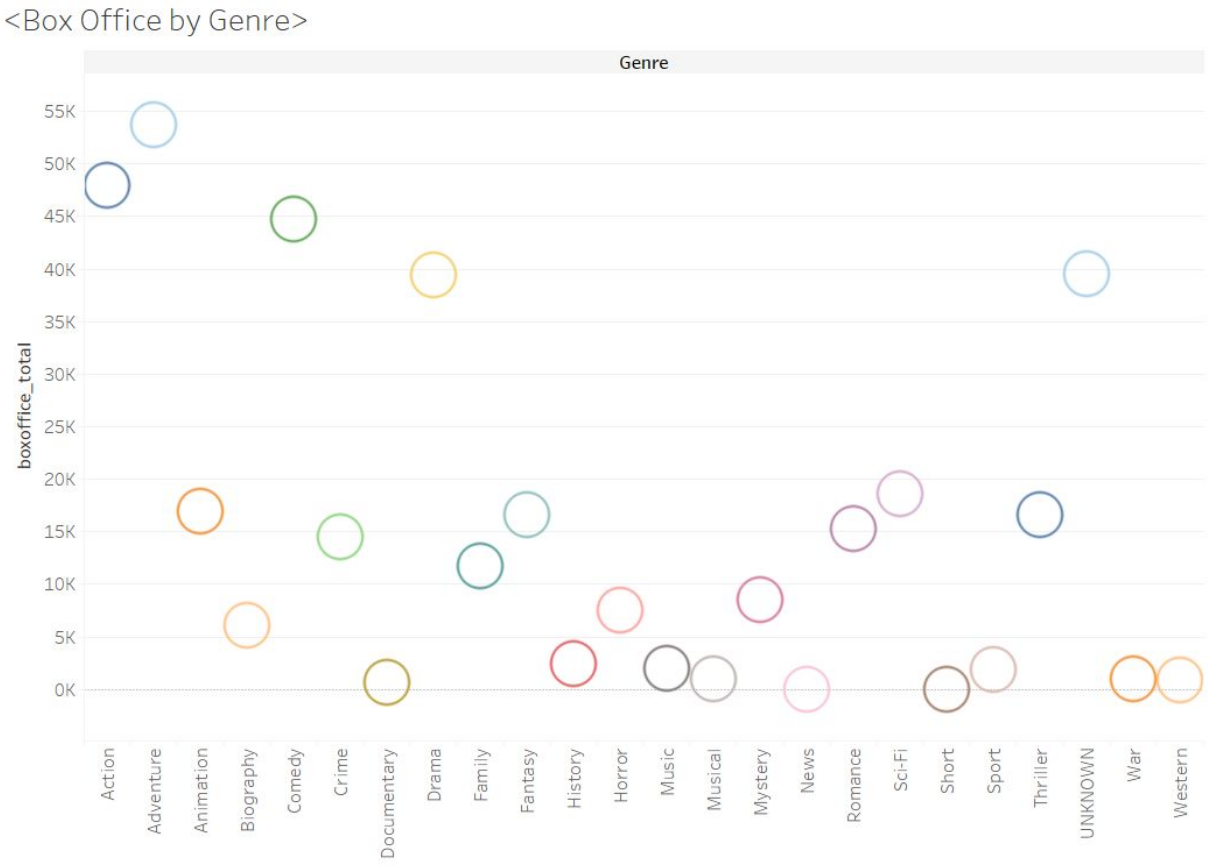
We performed an analysis dealing with inflation as part of our work for this project. In this, we adjusted the budget and box office earning for each movie based on the differences of inflation as the years have progressed. This helps us develop a holistic analysis of the data in terms of earnings based on different parameters.

The adjusted income also helps to ensure that we get accurate results. The models we have used as part of our analysis include lasso, stochastic gradient descent, random forest, and gradient boosting. Each of these models has performed fairly well on the data. We believe that the best model we have is Gradient Boosting with an accuracy of around 66%. We fed 85 variables into this model and we created dummy variables for month, genres, and movie ratings in order to reach this performance.

2. Answers to Questions

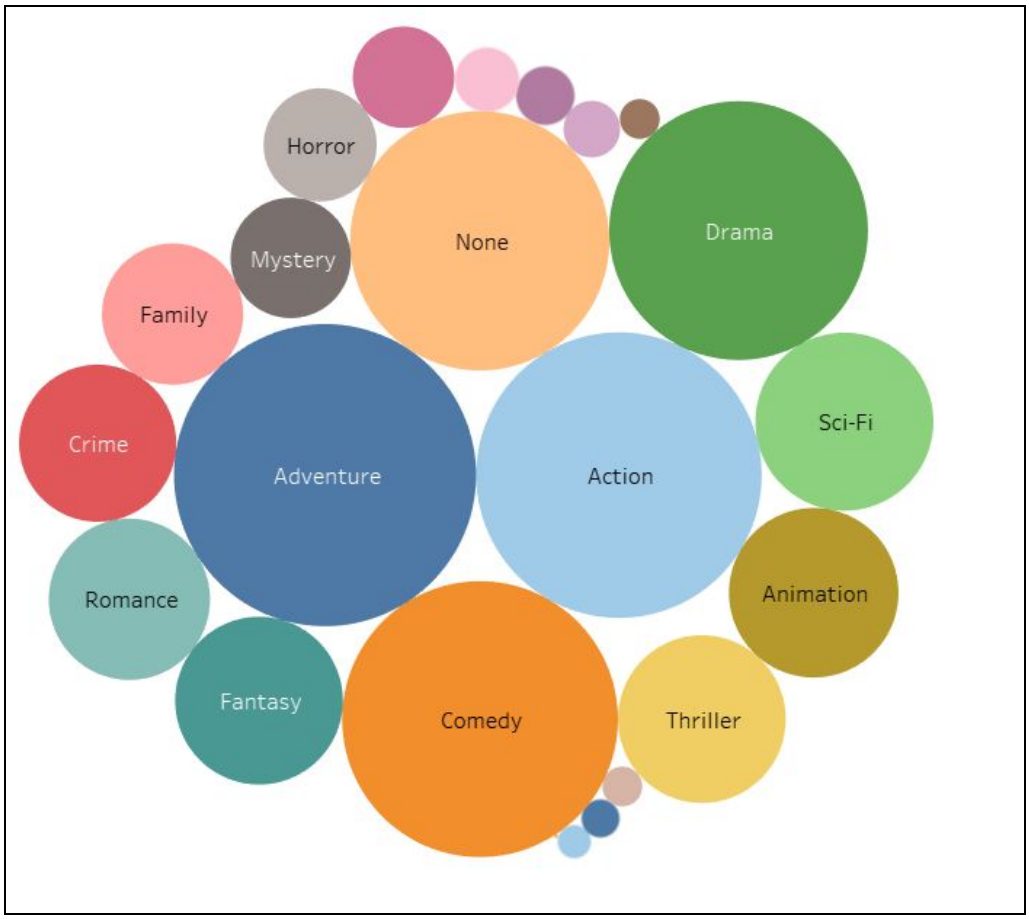
1. How do movies fare in terms of genre? Comedy, science, fiction?

   The results being presented below are based on absolute values and not on adjusted pricing indices.



Box Office by Genre

This chart can give us a better demarcation of the boxoffice_total earnings by each of the genre we see that the top earners are Action,Adventure and Comedy.



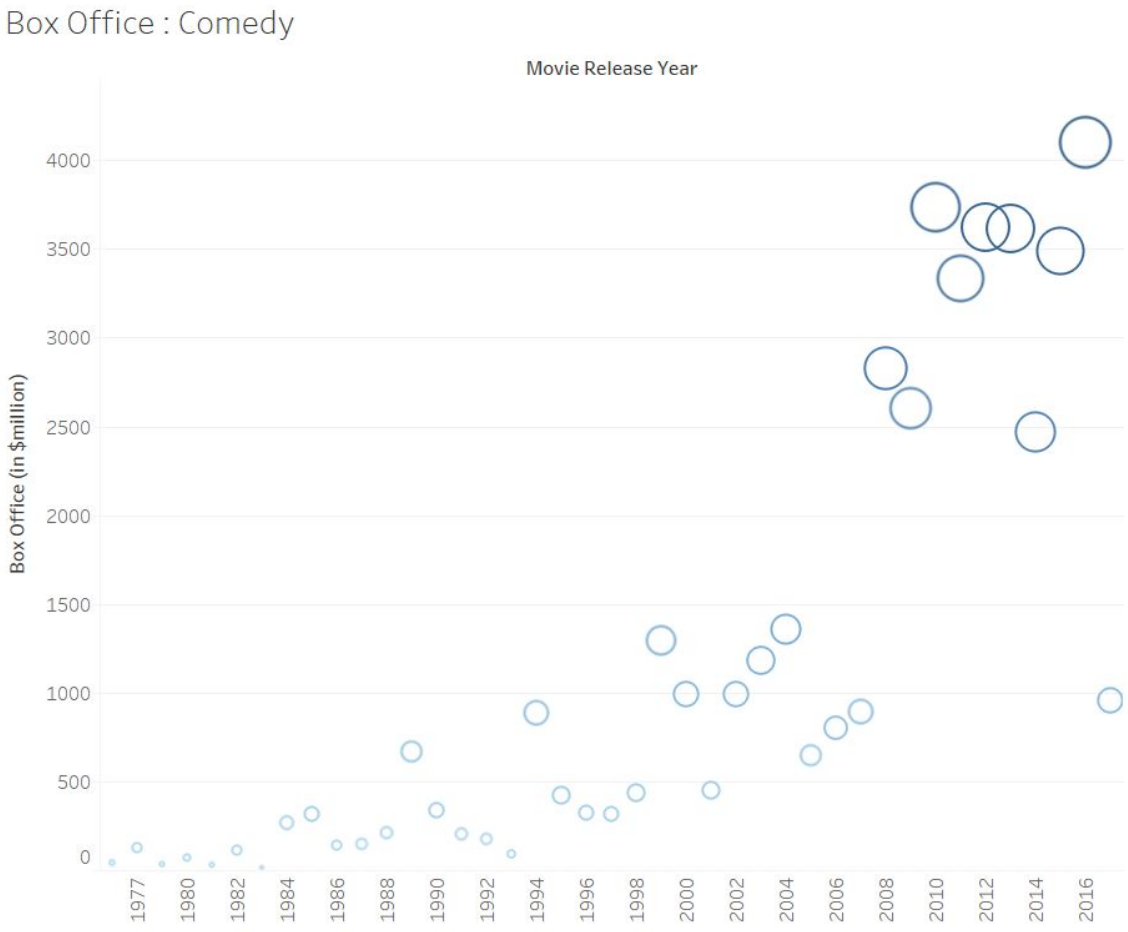
Bubble Graph for Genres(by Box Office)

Graph shows the effect of Genre on the Movie Earnings. The size of the Bubble shows us few of the top earning genres.

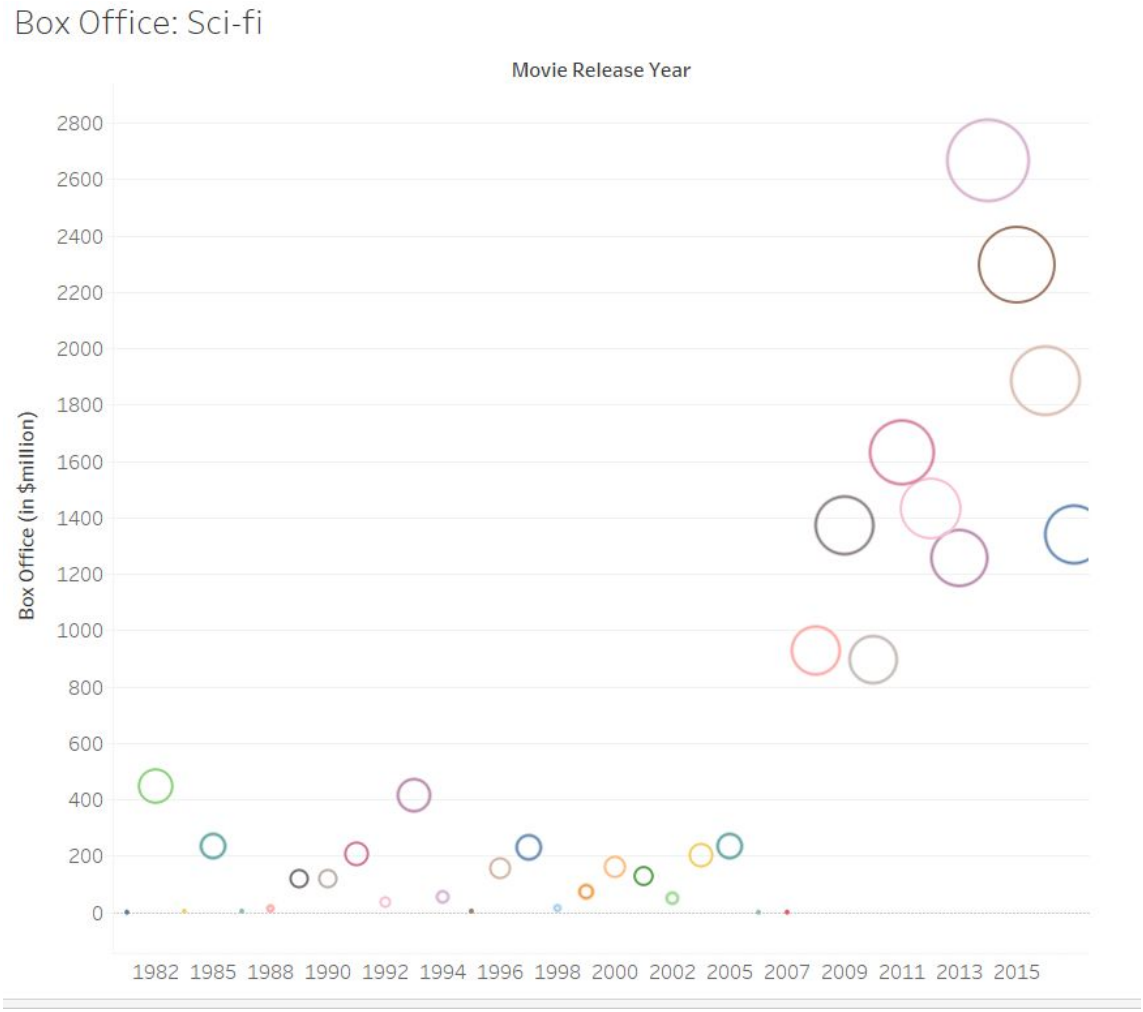
The top four genres are Adventure,Action,Comedy,Drama.
So there is high probability that a movie made on Comedy will perform well at the Box-office when compared to Science Fiction.So, as a person who will be investing in movies we recommend him/her to invest in Comedy movie rather than a science fiction movie.

The science fiction movie performs better when compared with Animation,Thriller,Romance ,Crime,Family,Fantasy,Horror and Mystery

Movies with genres like adventure, action, drama, and comedy are the best performers as seen in the graphic above.



From the graph we can clearly say that in case of the comedy films that were released in 2014-15 are the highest grossers.And the next league of highest grossers from comedy came up at 2010 to 2014.



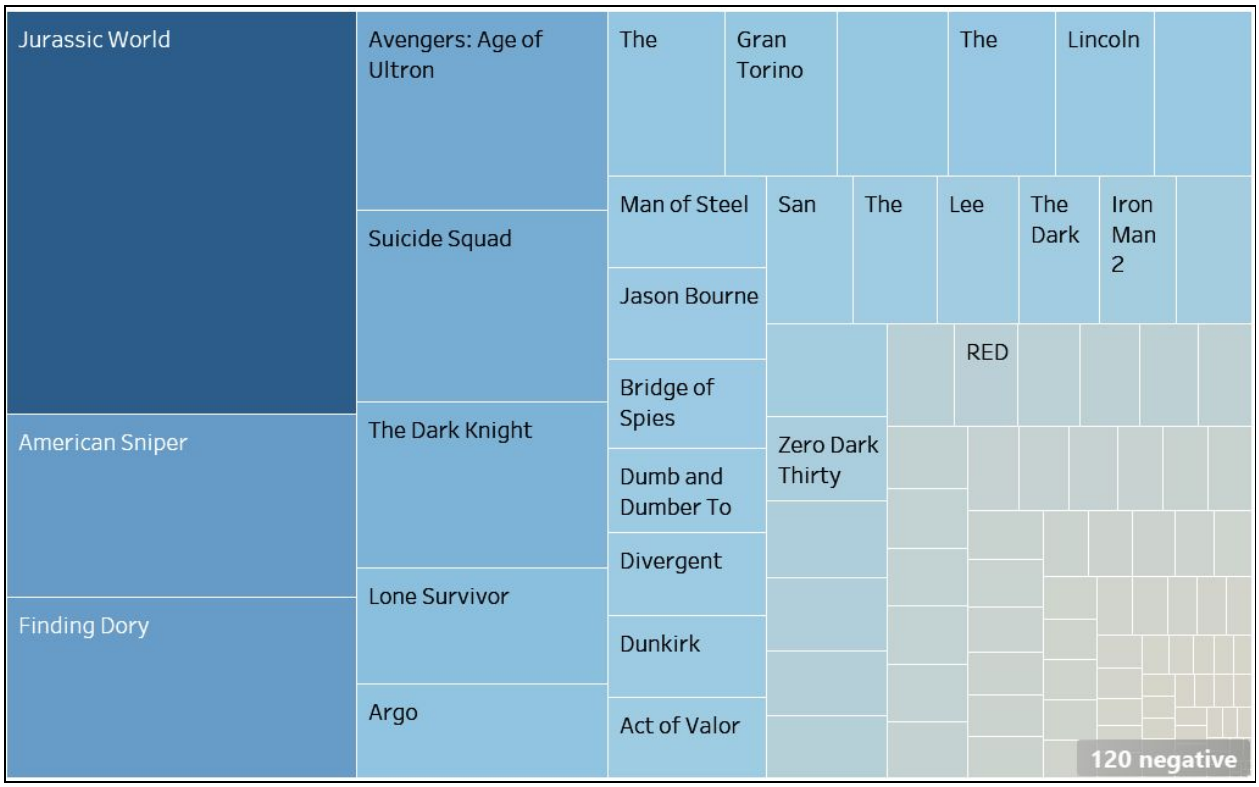
In the 2010s, Sci-fi movies are observed to obtain huge boosts in the market. Especially in 2014 and 2015, when the annual total box offices grow up to $2,666 million and $2,299 million, leading by the famous sequel *The Hunger Games: Mockingjay - Part 1(2014),* and *Guardians of the Galaxy (2014),* winning $337 million and $333 million respectively.

2. How well do you think that remakes, tent-poles and sequels perform?

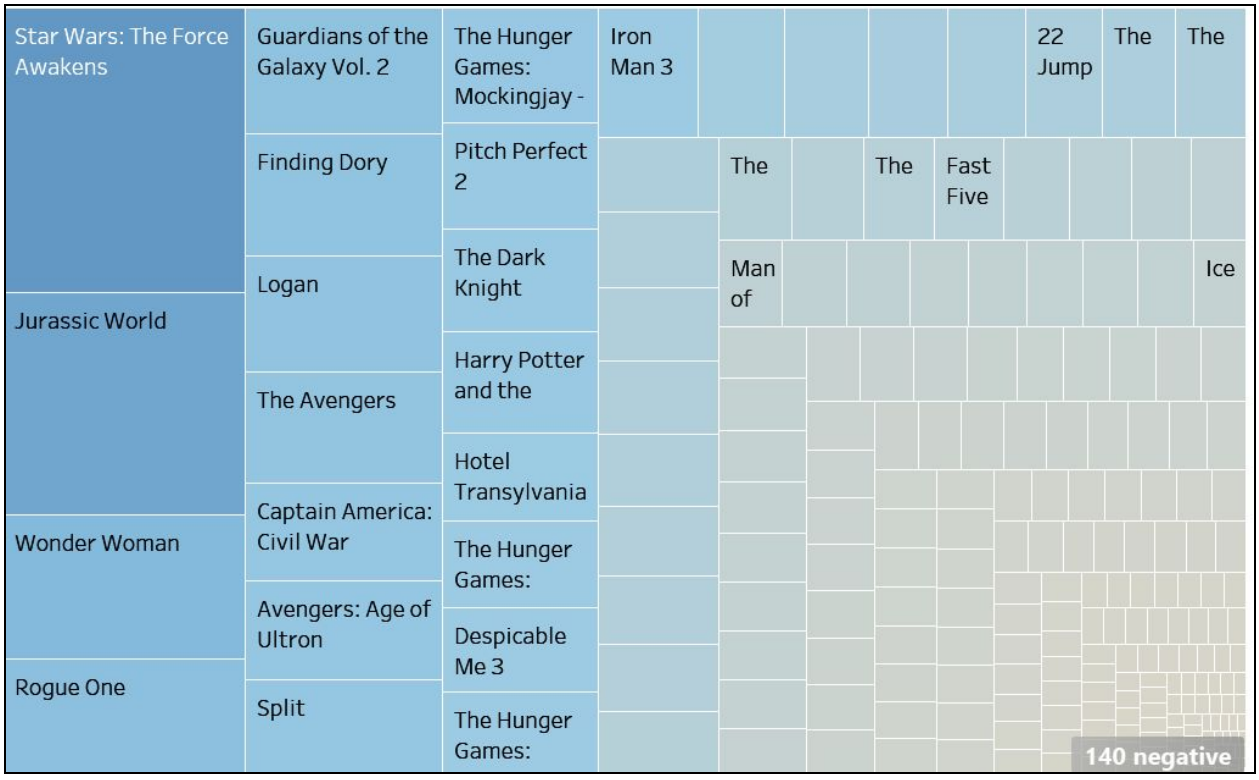
```
Action           88.425329
Adventure       103.632997
Animation       150.881667
Biography        44.491021
Comedy           45.532683
Crime            33.271690
Documentary       7.438495
Drama            26.020338
Family          829.562121
Fantasy          48.717221
History           0.721626
Horror           40.966535
Mystery          79.217792
Romance          23.422755
Sci-Fi           30.012626
Sport             0.164376
Thriller          4.838858
War               0.200442
Western           0.245131
Name: boxoffice adjusted, dtype: float64
```

The image above shows the average box office collection for each genre of movie (in millions of dollars). These values have been adjusted according to inflation over the years as have the values in the three graphics below. Tentpole movies tend to perform well in the box office. Examples of high performance tentpole movies are Jurassic World ($7032M), American Sniper ($3205M), Finding Dory ($3149M), and Avengers: Age of Ultron ($2508M). All four of these movies surpass the mean gross box office collections in each of their respective genres. Movies in the sequel category, too, greatly surpass the average box office collections in their respective genres. Some movies like Star Wars: The Force Awakens collect as much as $8992 million just by virtue of them being part of the gripping Star Wars Movie series. Remakes also tend to surpass expectations in

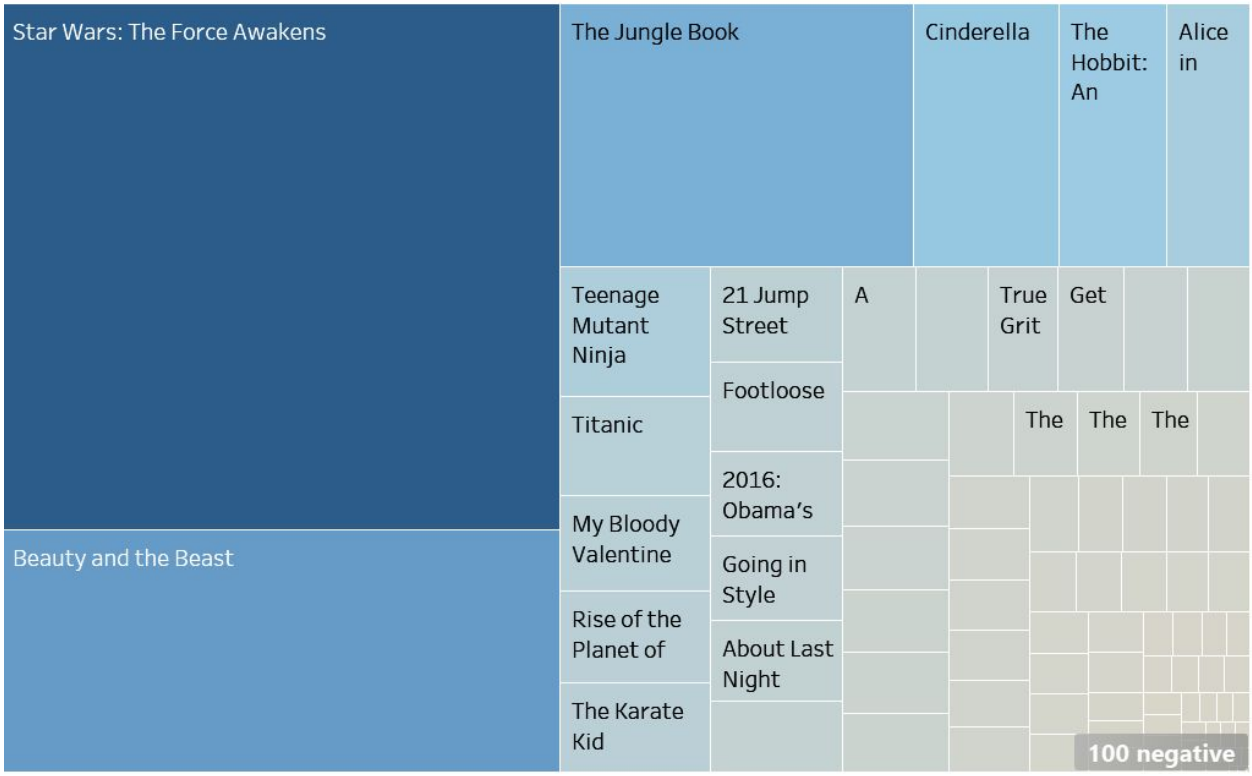
terms of box office collections. The images below show treemaps for each of three categories of movies (made in Tableau).



Tentpoles

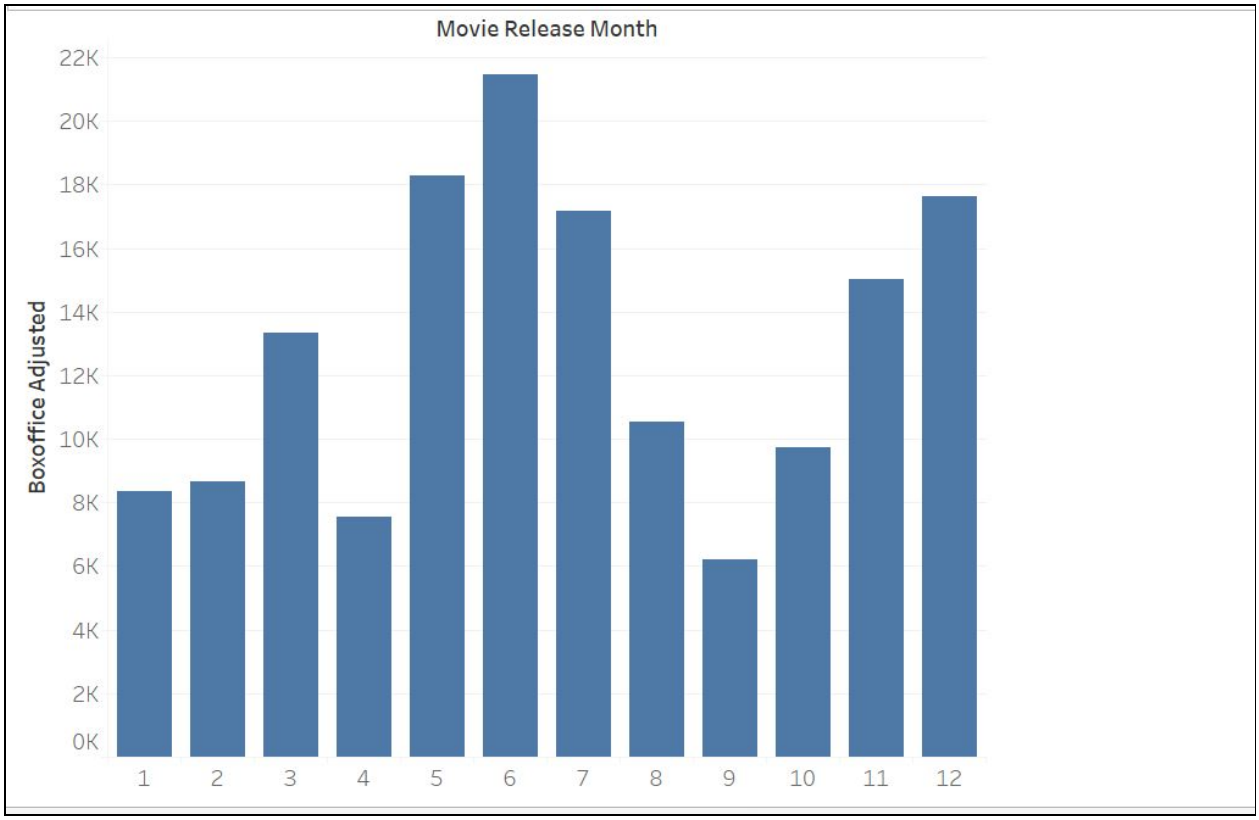


Sequels



Remakes

3. **Does release date influence box office performance? Do movies released at the same time perform differently?**



As seen in the image above, the month of June tends to see the highest earnings in terms of box office collection at $21,466 million. Reasons for this could be that the month of June is when students and are off from school and more families tend to watch movies. Other high earning months (above $16,000 M) are May and December.
The data above has been adjusted for inflation over the years.

4. If you were building a model to predict profitability of a movie at the box office. Which key variables would you consider? Bring out the relationships of these variables and their effect on profitability of a movie. Once you have studied the various relationships and have a robust model, what kind of model would you advise to a movie distributor on how he can maximize profit?

   After our analysis we found out that the key variables determining the movie success depends on the following factors and model performance is listed below:
   # Summary:

   Our Coefficient of Correlation (R-Squared) after applying xgBoost is coming around 68%.

   This means independent variables are explaining the target variable i.e. Box Office Gross optimally.

   RMSE value for our model is 1.7 which is again pretty good.

   So in this case XGBoost performs well for us.

   Top Features that are explaining the performance and Gross Collection of movie are the following based upon the above analysis if important features:

   1. Movie_Votes:

   More the votes movie is getting means people are more likely to watch the movie that is its going to perform well in the box office.

   2. Metacritic and Movie_ratings:

   These are also playing significant role.

3. Other Attributes like Directors and Actors:

These are also significant which we have presented in the Tableau Dashboards.

4. Genre wise 'Thriller', 'Action', 'Drama' and 'Comedy':

These are some of the top performing genres.

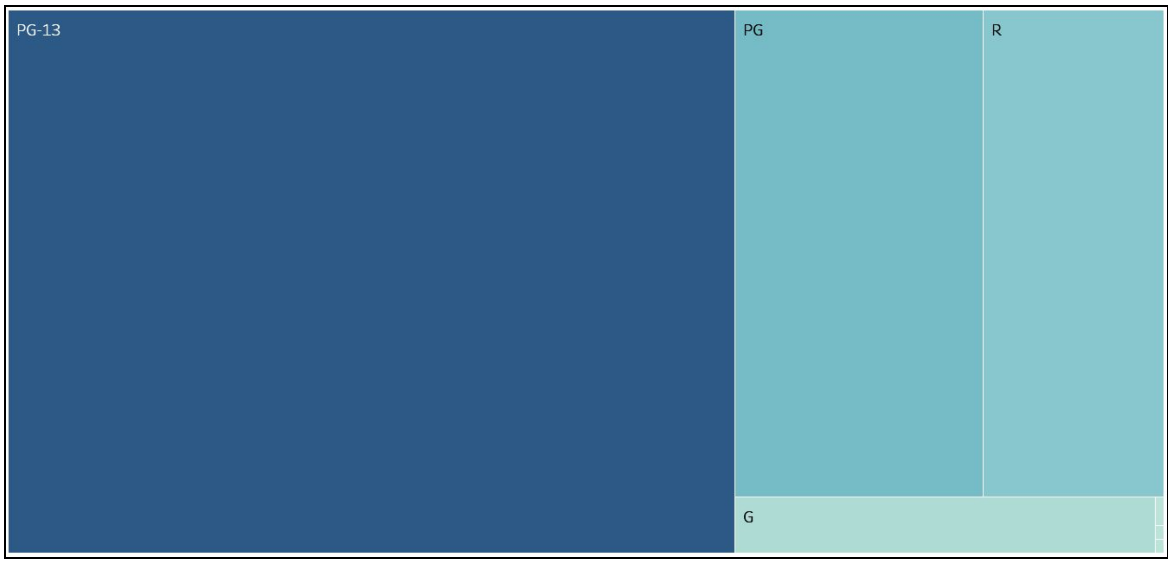
5. Rating wise R -rated:

These movies are more likely to do good in Box Office (as it has higher significance level and it is okay to watch for everyone)

5. Does rating of a movie have any effect on movie earnings?

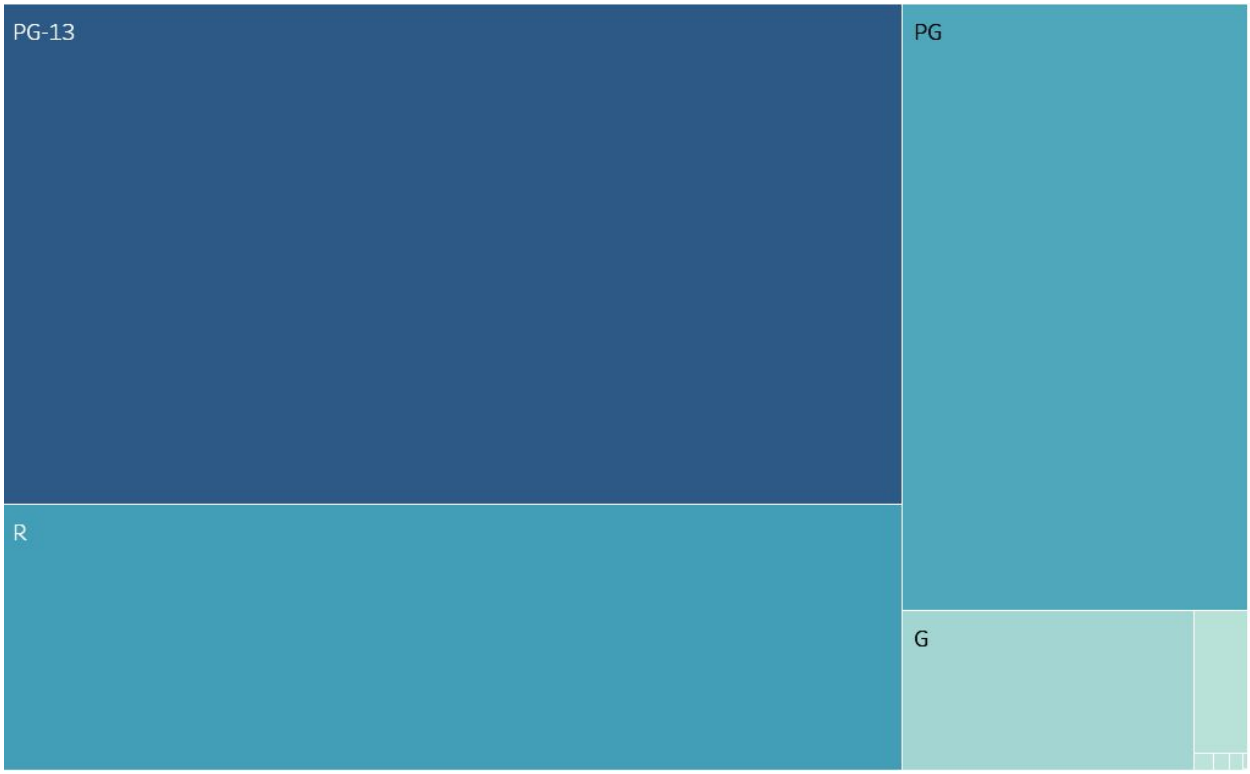
The results being presented below are based on absolute values and not on adjusted pricing indices.

Here, we see that PG-13 movies earn the most in terms of gross earning at $26,390 million. The second best ratings in terms of performance at the box office are PG and R.



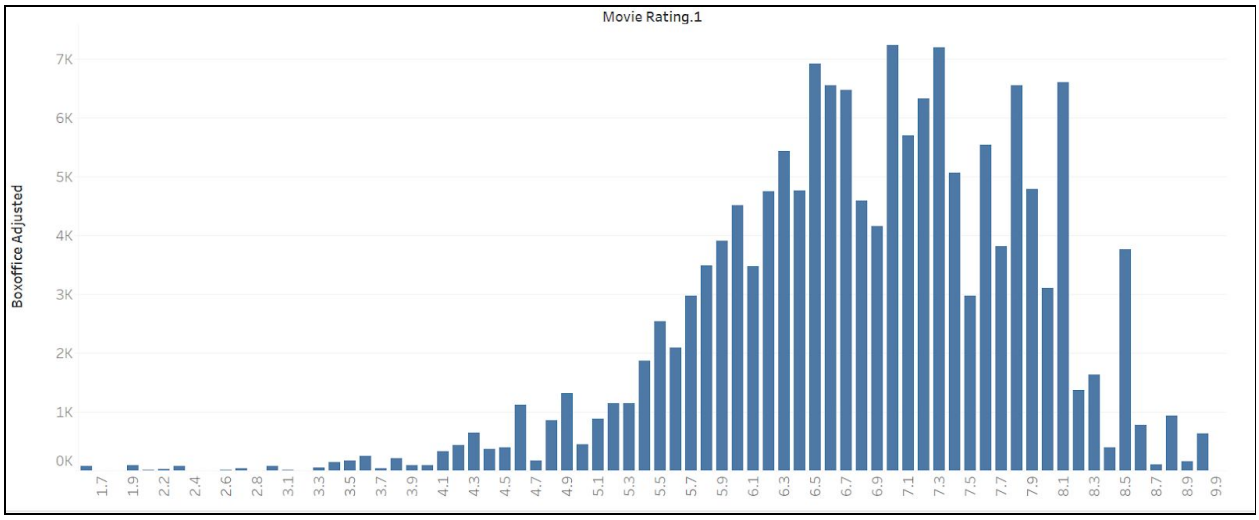
Rating in terms of PG-13, PG, R, and G does have an effect on movie earnings. We notice that movies with a rating of PG-13 earn the highest amount of money at $26,390 million. The second best ratings in terms of performance at the box office are PG ($33,716 million) and R ($38, 729) . The graphic above shows the amount of money earned by each "category" of rating. The smallest chunks in terms of money collected go to ratings like X, Not Rated, and Unrated.

The information below is based on adjusted pricing indices.



In the graphic above, we notice that the amount of money earned by PG 13 movies has increased to $72,645 million has increased from the previous un-adjusted box office collection. The amount earned by the next best, which are R ($38,729M) and PG ($33, 716M) have not changed from the previously stated values.

Below, we analyze the effects of critic ratings (the second type of rating) on the movie's earning potential. Here, we see that movies which have been given higher critic scores on a scale of 10 tend to earn more in terms of box office collections. The outliers on the extreme ends of the scale include movies which have been critically acclaimed but are not strong box office performs.

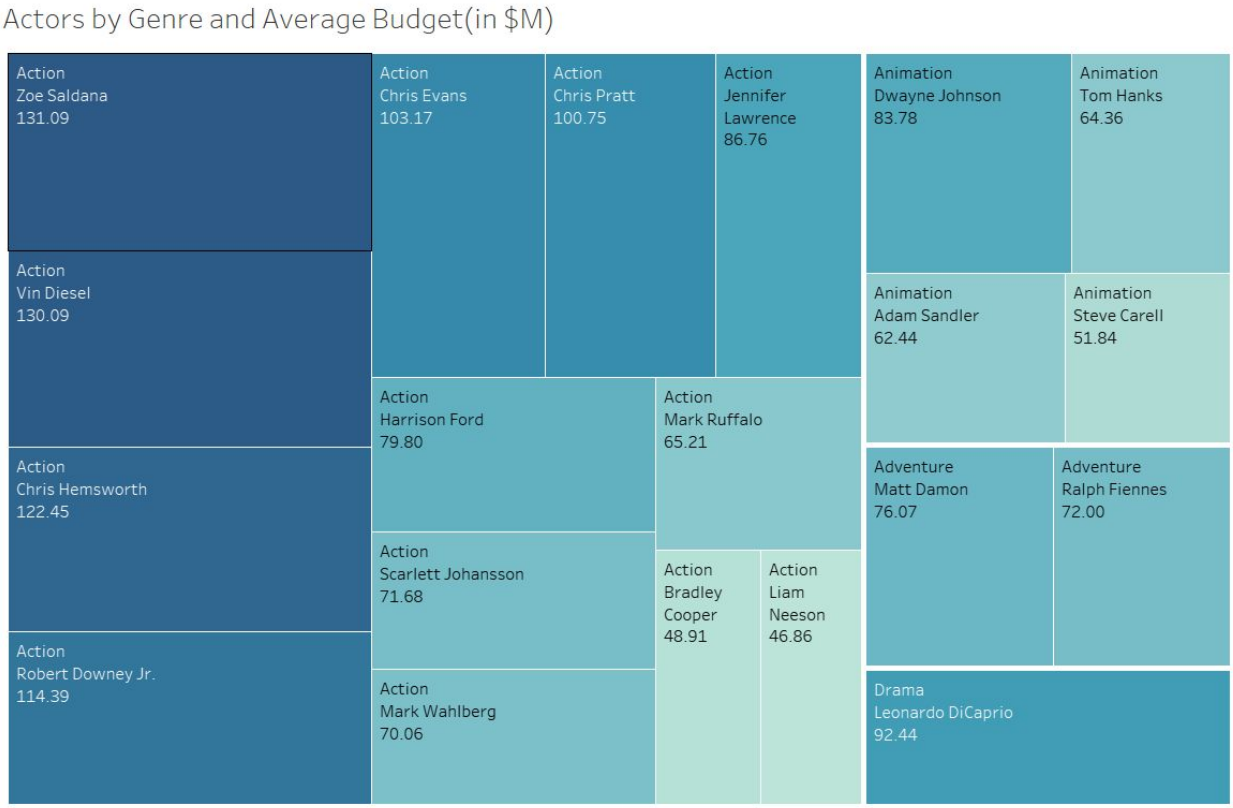


6. If you are running a movie production house, which Actor(s)/Director would you like to cast in your movie?

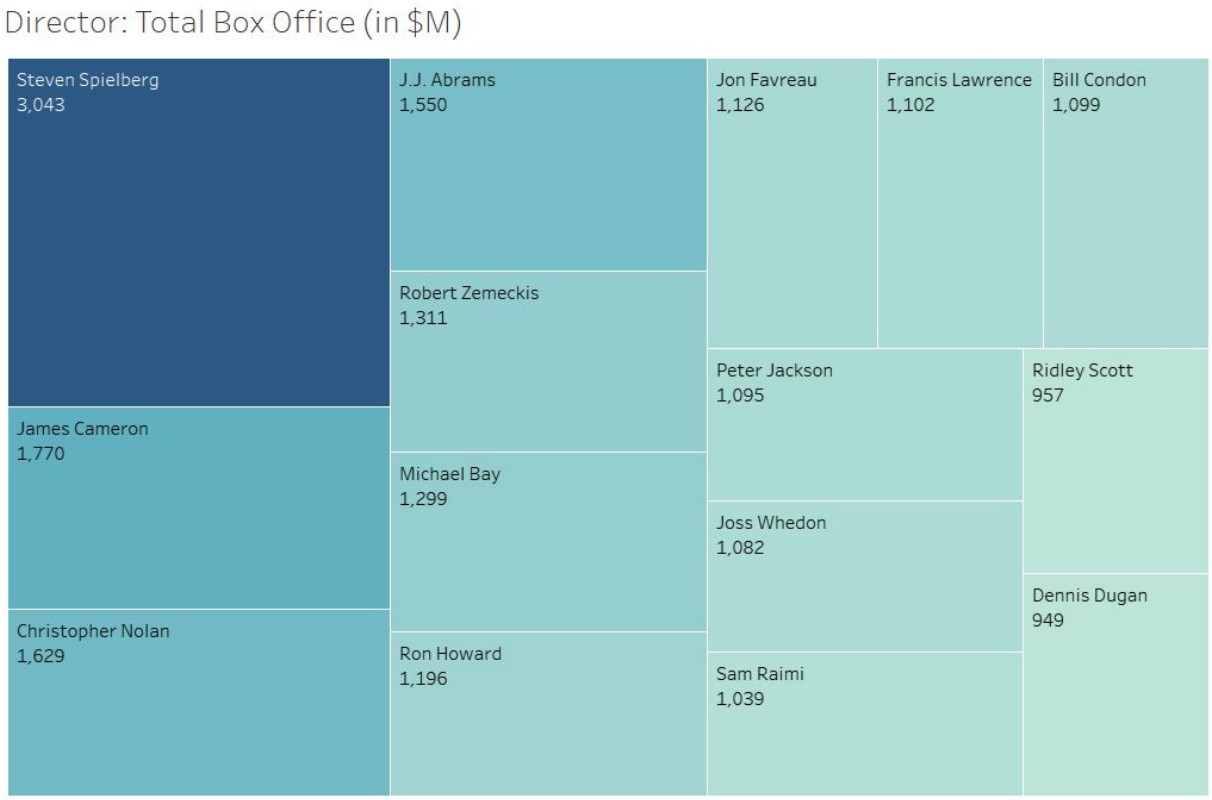
When thinking about film making, the popularity of actors first jumps into mind. Taking box office and movie genre into consideration, for action films, most of our most valuable actors comes from the Marvel superhero series, for example Robert Downey Jr.($3,540 M), Chris Evans($2,374 M) etc., followed by Vin Diesel ($2.191M) and Jennifer Lawrence($2.115M). If you are thinking about

animation films, Tom Hanks($3,014M) and Steve Carell($2,074M) are the best choices.

Simultaneously, budget of the filming should also be taken into consideration.

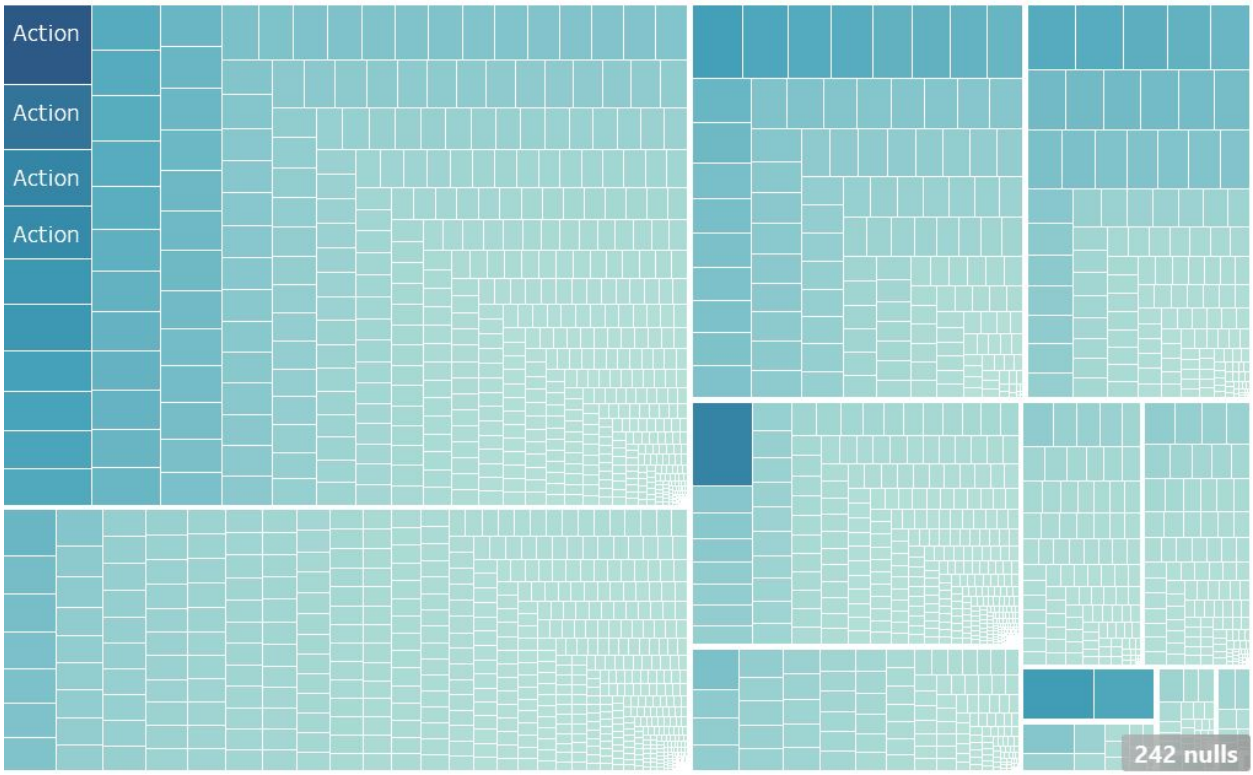


Considering about directors, Steven Speielberg produces the most box of $3,040M, followed by James Cameron and Christopher Nolan.



7. Does a plot of a movie affects its earnings in any way?

The plot of movie has an impact on earnings. Movies that have a plot relating to genres like action, adventure, and drama tend to perform better at the box office. In the graphic below, we see that action movies tend to perform best in terms of plot.


Action
Action
Action
Action
242 nulls

8. Why do some small budget films become blockbuster hits? Alternatively, why do some large budget films fail?

   Based on the data that we have analyzed, we are assuming that movies that have a budget lesser than $30 million and those that have a budget of more than $100 million to be large. Factors that contribute to this occurrence of small budget films earning more and large budget movies not earning as much could include the actors, director, plot, genre, critically acclaimed movies.
   Considering our model the best determiner of the box office gross small movies become blockbuster hits purely on the basis of the plot and also the time or duration the movie has been released.That might be in the holidays season of Christmas and the summer especially.