What Makes a Movie Successful : Using Analytics to Study Box Office Hits

Chancellor's Honors Program

SARAH JOSEPH

SPRING 2019

Table of Contents

Introduction……3

Today's Use of Analytics……4

Analysis Overview……6

Data Preparation……7

Descriptive Analytics……8

Rating……8

Release Month……10

Universe……11

Distributor……13

Director……15

Genre……16

Predictive Analytics……19

Will the movie be a success?……19

How much money will the movie make at the box office?……23

Conclusions……29

Works Cited……30

Introduction

Everyone loves a good movie! From animated classics to franchise action movies, movies appeal to many different groups of people and offer wonderful escapes from reality. The entertainment industry, American produced films and television shows, is a large part of our nation's economy, supporting around two million jobs in the country in 2016 (Busch). The industry feeds a large amount of money into other parts of the economy, for example it paid out around $50 million to local businesses in 2016 (Busch). The film industry alone generates a massive amount of revenue, with an estimated generated revenue of almost $43 billion in 2017 (Robb). Since the first movies were produced, people have been trying to figure out what exactly makes a movie a "success", what common factors. Now, with the growing field of business analytics, it is becoming easier to narrow down some common factors in successful movies.

As the demand for new, original content grows, in part due to the rise of streaming platforms such as Netflix, Hulu, Amazon Prime, Disney+, and others producing original movies and television shows, so too does the need for comprehensive analytics about what makes a movie successful (Busch). With the ever growing demand for new movies and the wide range of movie genres, analytics about the performance of movies can give useful information to studios so that they can make the most strategic decisions regarding production and financing.

Today's Use of Analytics

The use of analytics in the film industry to predict success or attempt to produce "box office hits" is different than using past intuition about movies; it is different than following a set formula or plan that has appeared to have worked in the past (Schlesinger). Rather, analytics in movies, as in any other industry is about finding useful patterns in the data that are not visible to the naked eye and can be exploited for gain (Schlesinger).

As with many other industries right now, analytics as common practice is still relatively new to the film industry. There are, however, some techniques and data collection being analyzed and considered in marketing and other decisions. IBM is one of the leaders in predictive analytics and it has been partnering with movie studios to collect data to determine what it is about a movie that audiences like ("Big Data and Hollywood: A Love Story"). Along with information like the studio, the actors, the budget, and more, IBM is also collecting data on audience sentiment, meaning the company is using audience and critic scores from multiple sources as well as key words in reviews to gather an audience's overall feelings towards a movie ("Big Data and Hollywood: A Love Story"). With this, IBM is learning better ways to market to certain demographics, smaller groups of people ("Big Data and Hollywood: A Love Story"). This technique is allowing companies to make minor adjustments to their marketing campaigns in order to specifically target certain groups in hopes of generating more revenue from that demographic ("Big Data and Hollywood: A Love Story").

Apart from audience reactions, other companies are focusing their studies and analyses on the actual characteristics of the film; Netflix has classified movies into certain genres based

on a group of 70,000 different characteristics ("Big Data and Hollywood: A Love Story"). It is the hope that in the future the analysis of characteristics of movies combined with the analysis of audience sentiment can be combined into a predictive model for studios to employ before the movie is even in production, hopefully saving studios valuable time and money ("Big Data and Hollywood: A Love Story").

One studio that is pursuing the marketing angle and the production angle is Legendary Entertainment (Krigsman, Marolda). The company is using the same data as described above, the characteristics of movies and the audience sentiment, as well as data about the audience itself to scale its marketing practices into micro-campaigns that are very focused on one group (Krigsman, Marolda). The analytics department is receiving this audience data from social media platforms like Facebook, Twitter, and Instagram and using the micro-campaigns and the previously collected data to predict ticket sales to movies for specific groups (Krigsman, Marolda). One way the company is measuring audience reactions to movies is using biometrics, in which audience members wear a device comparable to a Fitbit to track certain bodily reactions like heart rate, blood pressure, and more (Krigsman, Marolda). Incredibly, Legendary Entertainment's intense use of data and analytics only began around 2012 and 2013 and they have already developed technologies for analysis and data storage that other companies are envious of (Krigsman, Marolda).

Analysis Overview

In this analysis, I will perform an analysis on a large data set of approximately 3,000 movies. The data includes many different types of information about each movie, ranging from the release date, the director, the studio, to other information like the budget, the box office earnings, the audience and critic scores from different sites. I will analyze this data set to determine what contributes to a movie being a "success". For this analysis, I will define success in financial terms, basing the success of a movie on the amount of money it earns in comparison to the movie's budget.

I will begin my analysis by performing some descriptive analytics on the data set. Through this analysis, I will show some interesting trends in the data pertaining to what successful movies have in common. This analysis will mainly be done through the examination of charts, which will be produced using the software Tableau.

I will then perform some predictive analytics on the data set. Through this analysis, I will provide a model for predicting whether a movie will be a success or not, then examine the factors that are most important to making that prediction. I will also provide a model for predicting how much money a movie will earn at the box office as well as the factors that are important to those predictions. This model building and predicting will be done through the use of the software R-studio, and then some graphics will be produced in Tableau to further understanding.

Data Preparation

The data for this analysis was provided to me by Adam Milliken, a graduate student in the Masters of Science in Business Analytics program at the University of Tennessee. The original data set contained information about 3,159 movies released between January 8, 1999 and October 5, 2018. There were thirty-two different pieces of information recorded for each movie, this included release date, budget, studio, director, genre, audience and critic scores, top-billed actor, and box office earnings.

To prepare the data for analysis, I removed some of the more recent movies because a number of the variables could still change after the data was collected due to different international premiere dates and when movies finally leave all theaters around the world and stop being reviewed. So, I removed any movies that were released after June 28, 2018, making the total number of movies in the data set 3,118. I then removed some variables that I did not feel were important for analysis, making the variable count twenty-five. I removed the unwanted movies and variables in Excel before transferring the data to R-studio to be cleaned further.

As I will be looking at success or failure in a categorical sense and a numerical sense, the cleaning of the data will be different for the two criteria. With the categorical success or failure descriptive and predictive analytics, other categorical variables will be used. With the numerical success or failure descriptive and predictive analytics, other numerical and categorical variables will be used.

Descriptive Analytics

For the descriptive analytics portion, I used categorical variables to look at whether a movie was successful or not; the basis for identifying the movie as a success was whether the movie earned twice as much money or more in profit as the movie's budget. I also looked at some average values of numerical variables.

Rating

Figure 1 shows a chart depicting the percent of the total number of successful movies grouped by rating. This means that of those movies in the data set that were identified as successful, earning at least double the movie's budget in profit, approximately forty percent were rated R, followed by approximately thirty-seven percent being rated PG-13. *Figure 2* shows the count of the number of movies classified as each rating within the data set. From this, it can be seen that PG-13 movies are the most frequent in the data set, followed by R rated movies. This aligns with the results found in *Figure 1* except the top two positions are switched between PG-13 and R ratings.

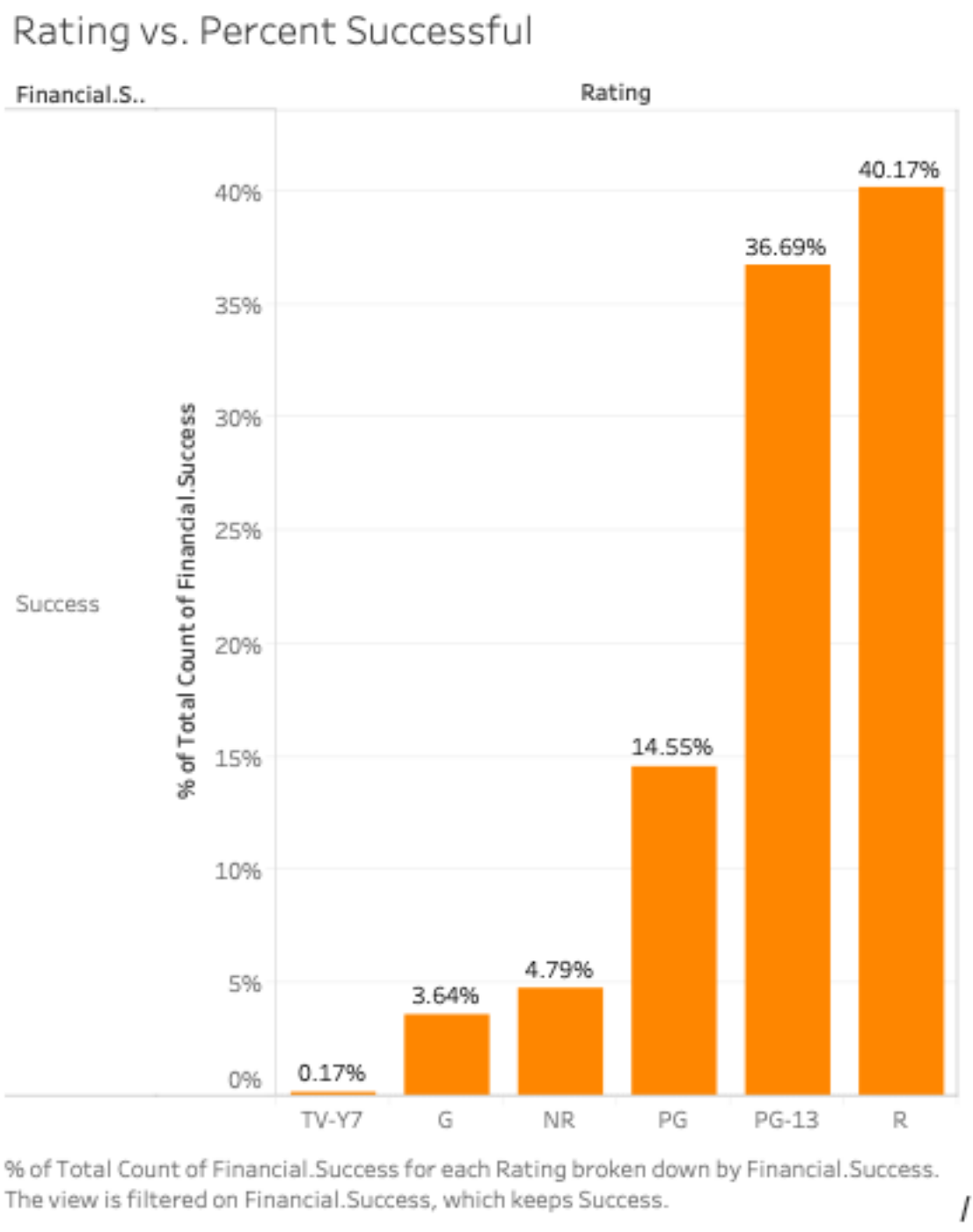

Rating vs. Percent Successful
Financial.S..
Rating
Success
% of Total Count of Financial.Success
40%
35%
30%
25%
20%
15%
10%
5%
0%
0.17%
3.64%
4.79%
14.55%
36.69%
40.17%
TV-Y7
G
NR
PG
PG-13
R
% of Total Count of Financial.Success for each Rating broken down by Financial.Success.
The view is filtered on Financial.Success, which keeps Success.

Figure 1

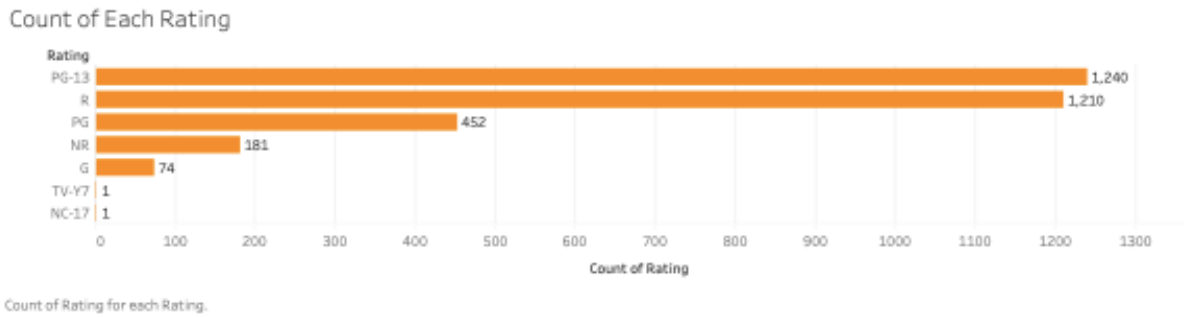


Figure 2

Release Month

Figure 3 shows a chart depicting the percent of the total number of successful movies that were released each month. It shows that the largest percentage of successful movies were released in the month of July, with a percentage of approximately ten and a half percent, with the month of November following close behind. The smallest percentage of successful movies were released in the months of January and April, approximately seven percent each month. This is especially interesting because, as *Figure 4* shows, the most movies have been released in the month of December, while the number of movies released in July is significantly lower.

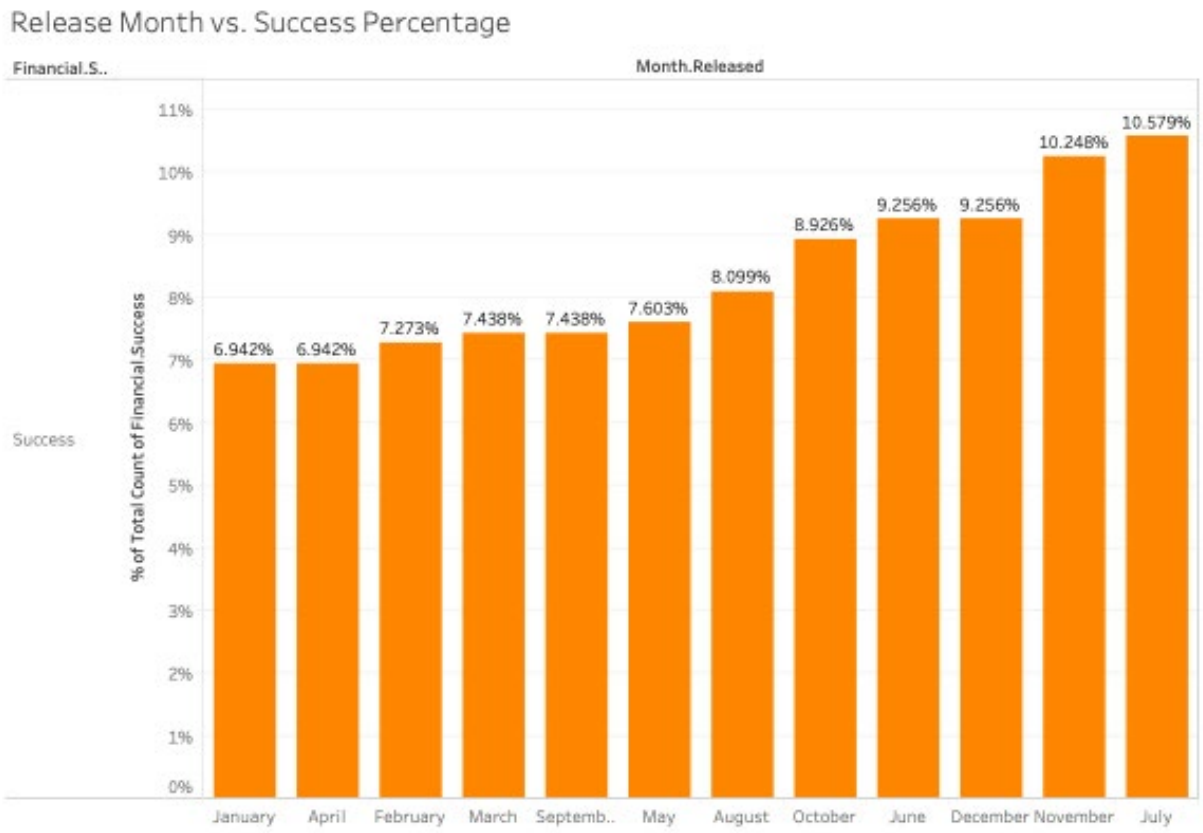


Figure 3

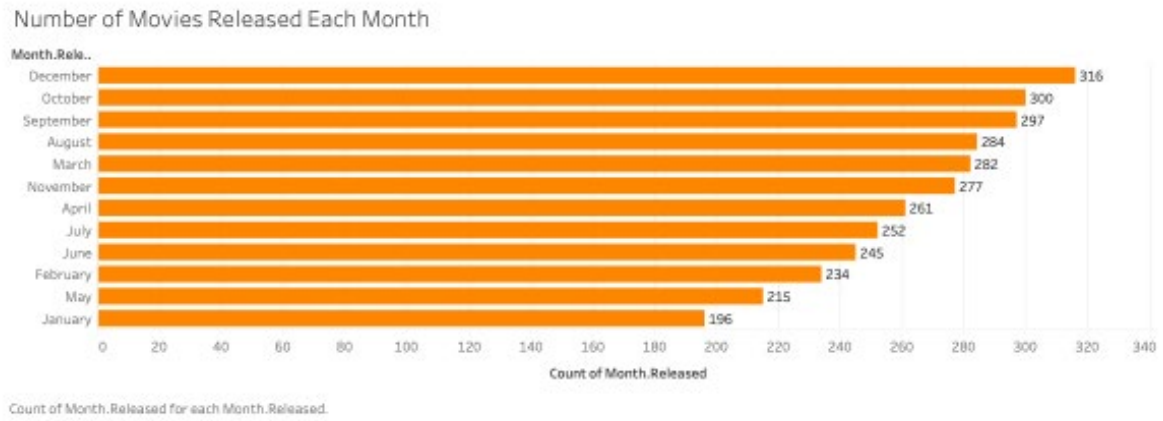


Figure 4

Universe

Figure 5 depicts some popular movie universes from the data set; the universes presented here are those in the data set that include five or more movies. The chart shows the percentage of the movies in that universe that were financial successes. Three franchises have a one hundred percent success rate, the Twilight saga, the Saw movies, and the Paranormal Activity franchise, meaning all of the movies in those franchises made more than double their budgets in profit. Universes that have been wildly popular in recent years have not been as financially successful in terms of their earnings in comparison to their budgets, such as the Marvel Cinematic Universe having only thirty-five percent of its movies as financial successes, the X-Men franchise only having approximately twenty-seven percent of its movies as successes, and only twenty percent of the James Bond films being successful. *Figure 6* shows an estimated average budget for a movie in each universe; it can be seen that the Saw franchise and the Paranormal Activity franchise, two of the three universes that have had all of their movies be successes, are the two franchises with the lowest estimated average budget, $14,357,143 and $6,904,500 respectively, meaning they needed the lowest profit to be a financial success. It can also be seen that the DCEU, one of the universes with the lowest percentage of financial successes, twenty percent, is the universe with the second highest estimated budget, $344,700,000, indicating each of its movies needed one of the highest estimated profits to be a financial success.

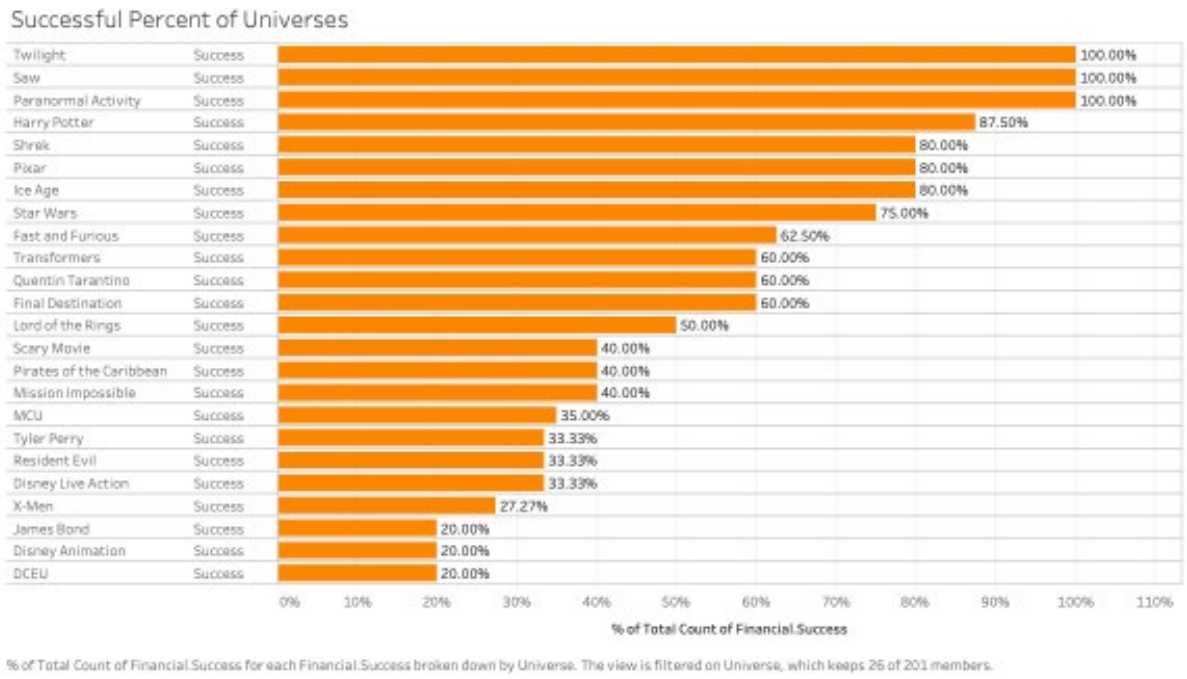


Figure 5

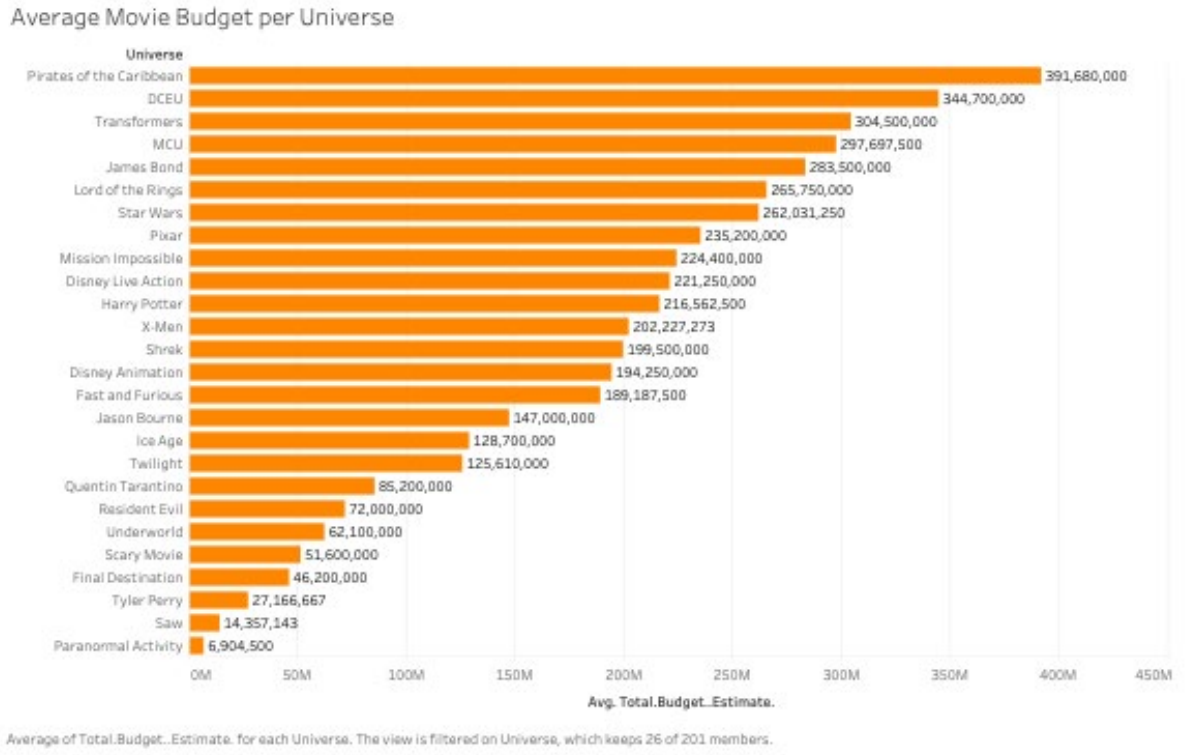


Figure 6

Distributor

Figure 7 shows a list of distributors that appear at least ten times in the data set that have had some financial success with the movies they have released. The percentages shown are the percent of the movies released by that distributor that were identified as financial successes. As can be seen, no studio had a majority of its releases be financially successful, with the highest percentage of successful movies being approximately thirty-six percent. It can be noted that the studios with the four highest percentages released fewer than one hundred movies, as seen in *Figure 8*. *Figure 9* shows the average estimated budget for a movie released by each distributor, and it can be noted that the studio with the lowest estimated budget, $9,483,871, is A24, the studio with the second highest percentage of successful movies, meaning that the studio on average had to earn less money than others in order to be a success.

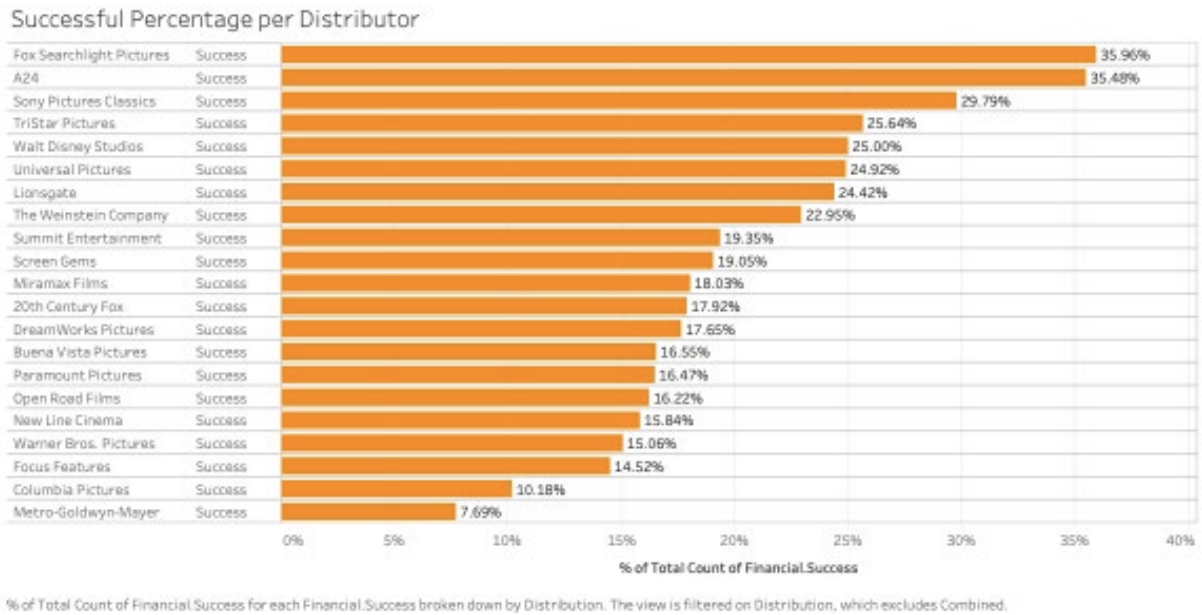


Figure 7

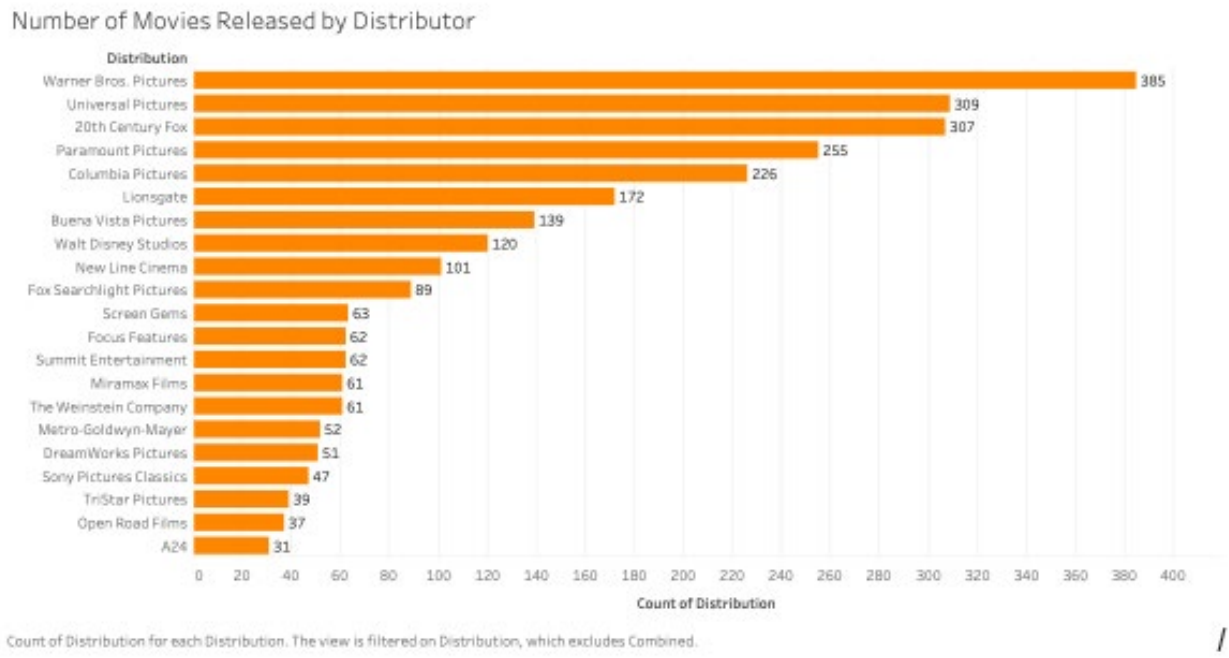


Figure 8

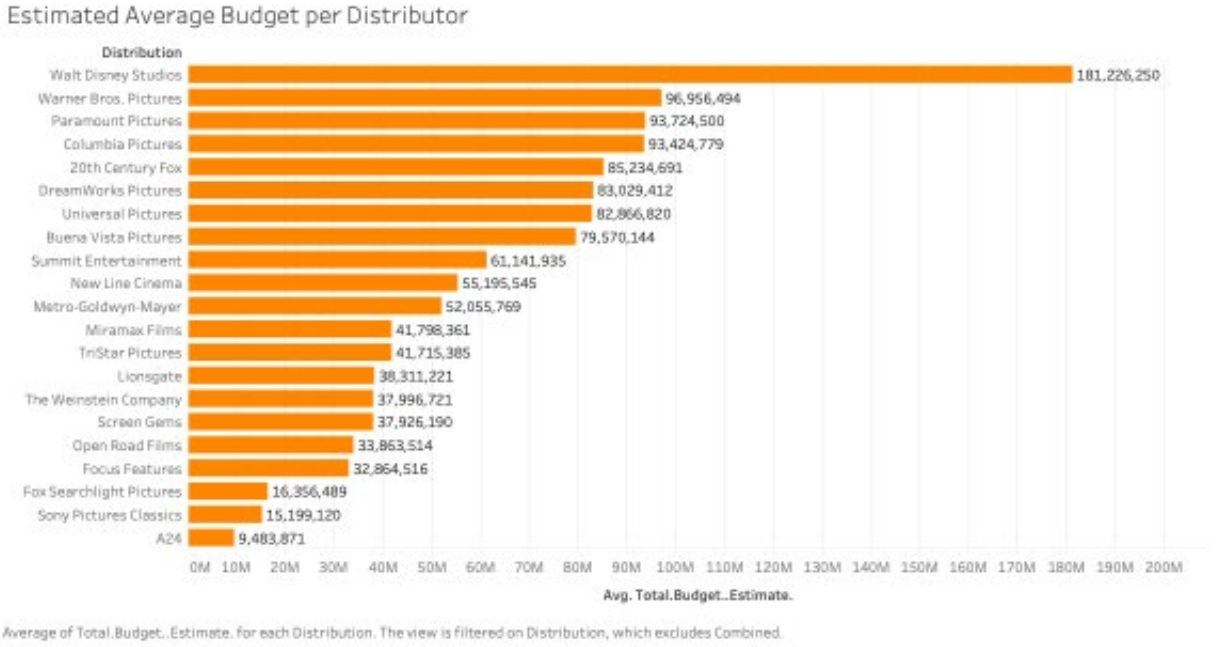


Figure 9

Director

Figure 10 shows a list of directors that appear at least ten times in the data set that have had some percent of their movies be financial successes. The director with the largest success rate is M. Night Shyamalan with a success rate of forty percent; Ridley Scott has the lowest success rate with approximately six percent of his movies being successful. Interestingly, there is not a correlating distribution when examining the average estimated budget for a movie by each director, shown in *Figure 11*. While Ridley Scott's movies cost more to make on average, $142,650,000, than M. Night Shyamalan, $101,850,000, both directors are in the middle of the range of average estimated budgets for the listed directors.



Figure 10

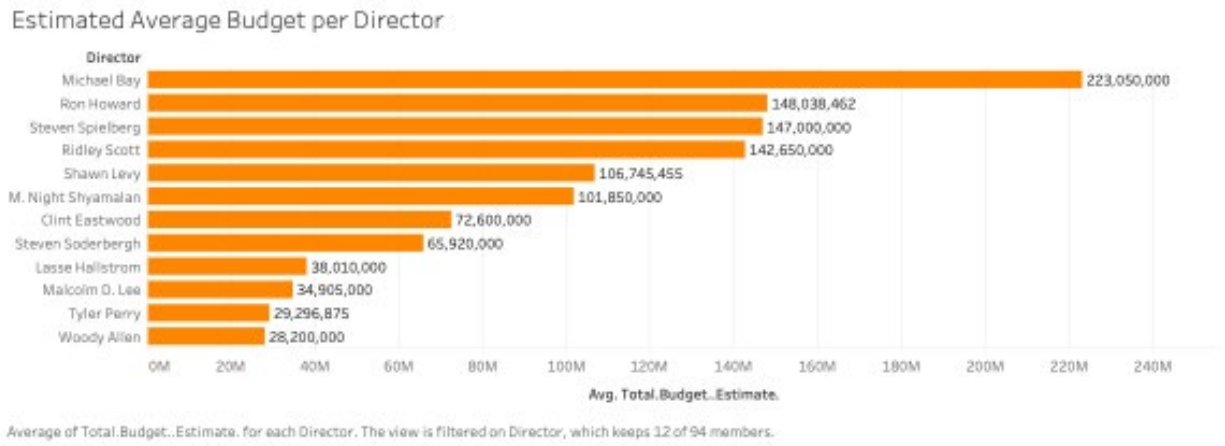


Figure 11

Genre

Figure 12 shows the percentage of the successful movies in the data set that are classified into each genre. The most frequent genre in the successful movies in the data set is Comedy while Space, Slasher, and Disaster movies appear the least. *Figure 13* shows the average estimated budget for a movie in each genre with Space movies being the most expensive at $292,500,000 and Exploitation movies costing the least at an average of $4,500,000. Space movies were one of the three least occurring genres in the successful movies; this correlates with the estimated average budget because the movies classified as Space movies had to make a much greater amount of money to be considered a financial success while Comedy movies did not have to make as much money to be financial successes because the average estimated budget for Comedy movies is the fourth lowest as shown in *Figure 13*.

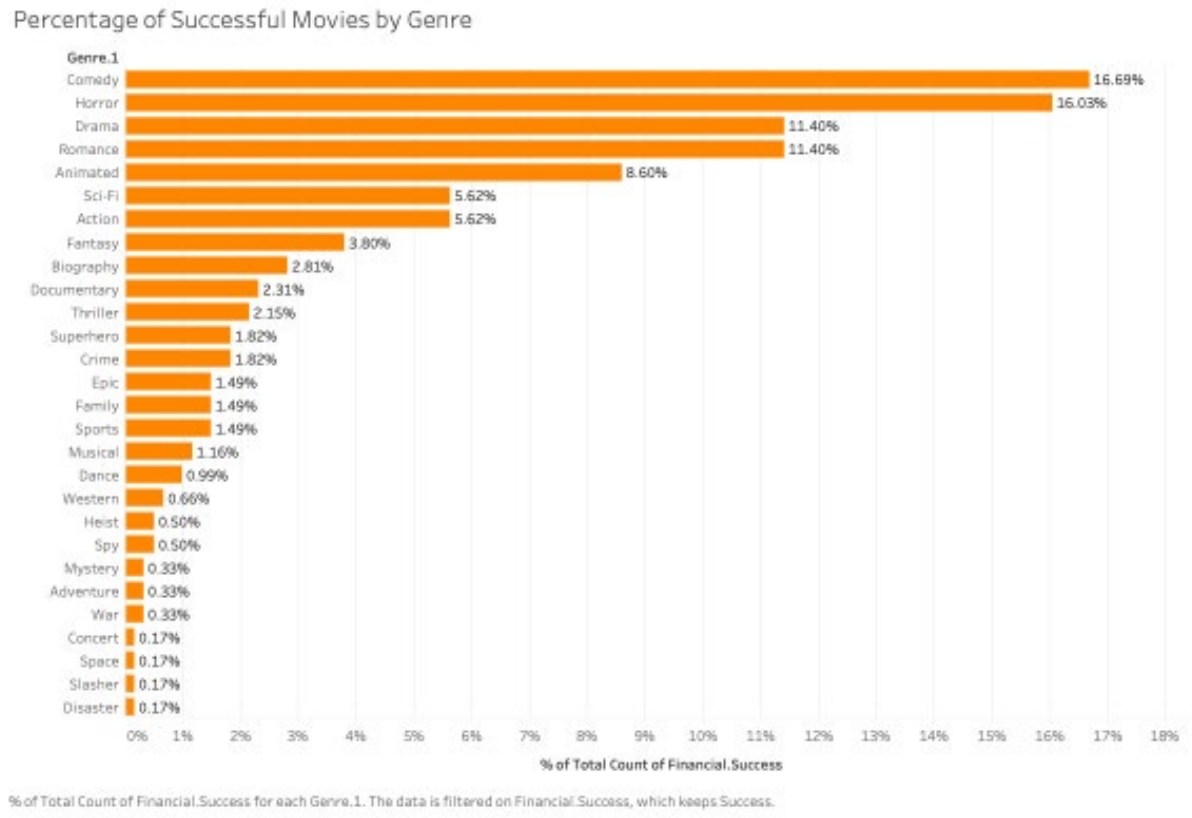

Percentage of Successful Movies by Genre
Genre.1
Comedy 16.69%
Horror 16.03%
Drama 11.40%
Romance 11.40%
Animated 8.60%
Sci-Fi 5.62%
Action 5.62%
Fantasy 3.80%
Biography 2.81%
Documentary 2.31%
Thriller 2.15%
Superhero 1.82%
Crime 1.82%
Epic 1.49%
Family 1.49%
Sports 1.49%
Musical 1.16%
Dance 0.99%
Western 0.66%
Heist 0.50%
Spy 0.50%
Mystery 0.33%
Adventure 0.33%
War 0.33%
Concert 0.17%
Space 0.17%
Slasher 0.17%
Disaster 0.17%
0% 1% 2% 3% 4% 5% 6% 7% 8% 9% 10% 11% 12% 13% 14% 15% 16% 17% 18%
% of Total Count of Financial.Success
% of Total Count of Financial.Success for each Genre.1. The data is filtered on Financial Success, which keeps Success.

Figure 12

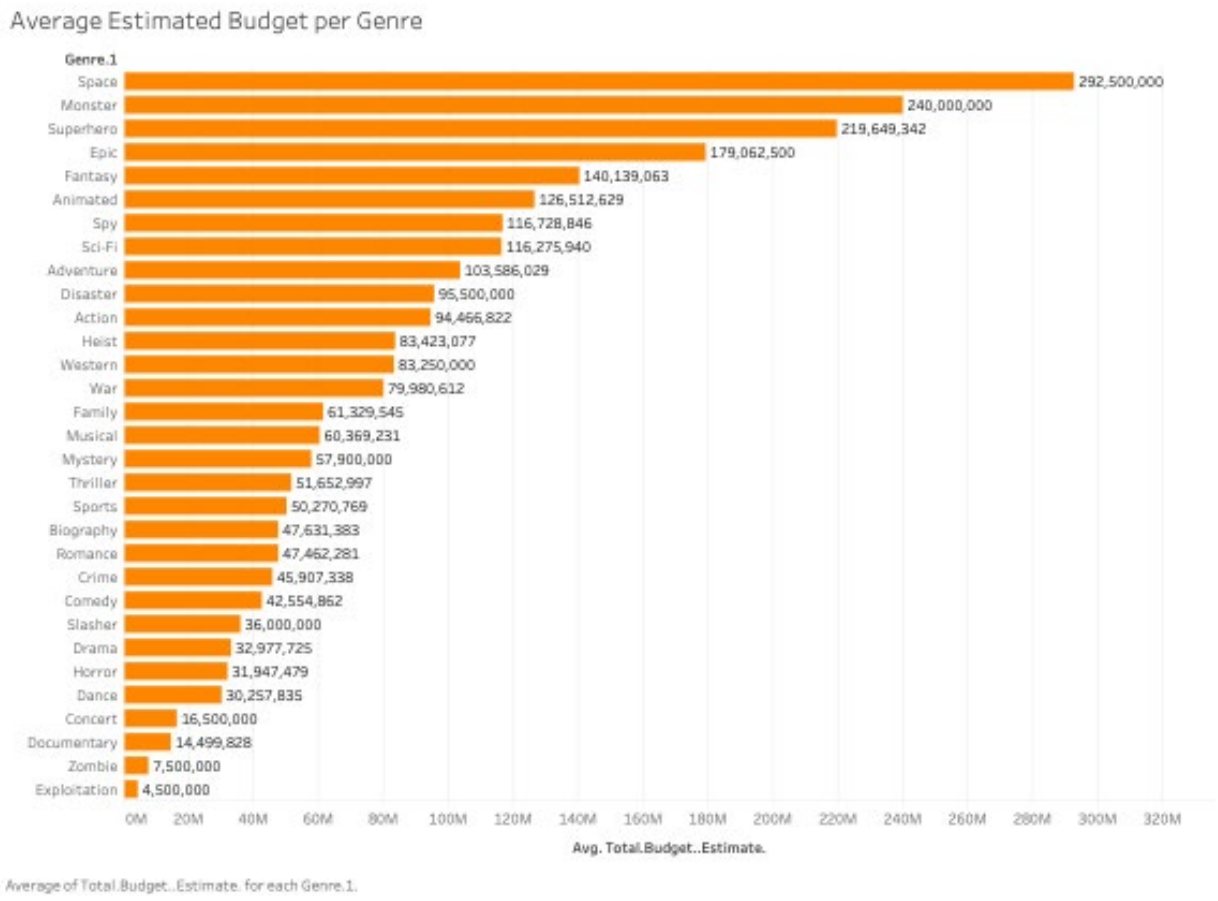

Average Estimated Budget per Genre
Genre.1
Space 292,500,000
Monster 240,000,000
Superhero 219,649,342
Epic 179,062,500
Fantasy 140,139,063
Animated 126,512,629
Spy 116,728,846
Sci-Fi 116,275,940
Adventure 103,586,029
Disaster 95,500,000
Action 94,466,822
Heist 83,423,077
Western 83,250,000
War 79,980,612
Family 61,329,545
Musical 60,369,231
Mystery 57,900,000
Thriller 51,652,997
Sports 50,270,769
Biography 47,631,383
Romance 47,462,281
Crime 45,907,338
Comedy 42,554,862
Slasher 36,000,000
Drama 32,977,725
Horror 31,947,479
Dance 30,257,835
Concert 16,500,000
Documentary 14,499,828
Zombie 7,500,000
Exploitation 4,500,000
0M 20M 40M 60M 80M 100M 120M 140M 160M 180M 200M 220M 240M 260M 280M 300M 320M
Avg. Total.Budget..Estimate.
Average of Total.Budget..Estimate. for each Genre.1.

Figure 13

Predictive Analytics

For the prescriptive analytics, I decided to build two models, one predicting whether a movie is a financial success or not and one predicting the estimated profit of a movie.

Will the movie be a success?

To predict whether a movie will be a financial success, I began with cleaning the data to include only categorical variables. This meant eliminating any critic or audience ratings from the data, the running time, the box office earnings, and the budget. I calculated the return on investment (ROI) for a movie by diving the estimated profit from the movie by the total estimated budget and then defining if the movie is a financial success by whether or not the ROI was greater than or equal to two, if yes then the movie was a success, otherwise it was a failure. Once this was defined, I eliminated the total estimated budget, estimated profit, and return on investment from the data set, leaving just the identification of whether the movie was successful or not. I then cleaned some of the categorical variables such as universe, genre, and rating to make the data easier to process. I then split the data set into a training sample of data and a holdout sample. The training data set was for building and testing the different models and the holdout set was for testing the chosen model.

Once the data was cleaned, I ran many different kinds of models on the training set in order to determine the best practice. I determined the best practice by looking for the highest accuracy between the models when predicting a movie's success based on the training data set. The different kinds of models that I tried were a vanilla partition model, a random forest model,

a boosted tree model, a support vector machine model, an XG Boost model, and an R part model.

The model that had the highest accuracy for predicting whether the movie was a financial success or failure on the training data set was the random forest model which had an accuracy of 0.802. A concern with building models with training set and choosing from accuracy is that the model may be "overfit" to the training data, meaning it may predict great on the training set but terribly on the holdout sample. So, the model needs to be tested on the holdout data set and the accuracy examined. When I tested the model on the holdout data, I received an accuracy of 0.83, indicating the model was not "overfit". *Figure 14* shows the code used to build and test the model.

```
#####################################
#Random Forest
#####################################

forestGrid <- expand.grid(mtry=c(1,2,5,10))  #put in values


FOREST <- train(Financial.Success~.-Movie,data=TRAIN,method='rf',tuneGrid=forestGrid,
                trControl=fitControl, preProc = c("center", "scale"),importance=TRUE)

FOREST$results[rownames(FOREST$bestTune),]  #Just the row with the optimal choice of tuning parameter
postResample(predict(FOREST, newdata=HOLDOUT),HOLDOUT$Financial.Success)
```

Figure 14

Once the model had been built and tested, I wanted to find what variables were important in this model for making the predictions. *Figure 15* shows a list of the twenty most important variables in the model for predicting if the movie is a financial success or failure. It shows that the movie being a part of a movie universe, such as the MCU or DCEU or a Tyler Perry movie, is the most important factor in predicting if the movie is successful, followed by

the top billed star of the movie being Ryan Reynolds and the movie being classified as a horror film.

```
> varImp(FOREST)
rf variable importance

  only 20 most important variables shown (out of 239)

                                Importance
UniverseYES                         100.00
Top.Billed.StarRyan Reynolds         74.23
Genre.1Horror                        71.92
Genre.1Documentary                   51.26
DirectorJay Roach                    48.80
DirectorAng Lee                      46.50
Genre.1Dance                         45.40
Year2018                             44.85
DirectorDavid Yates                  41.07
DirectorDavid O. Russell             39.29
DirectorM. Night Shyamalan           35.57
DistributionUniversal Pictures       35.50
Top.Billed.StarKristen Stewart       34.30
Month.ReleasedJuly                   32.84
Genre.1Animated                      32.56
DirectorRichard Linklater            31.16
Genre.1War                           30.91
DirectorAndy Fickman                 30.79
DirectorGuy Ritchie                  30.33
Month.ReleasedOctober                29.81
```

Figure 15

Before moving on from predicting whether or not a movie would be a financial success, I wanted to build a decision tree. With this decision tree, I wanted to be able to identify the probability of a movie with certain characteristics being financially successful. In order to build

a more readable and easier to understand tree, I decided to remove variables with many levels and levels that had very few observations, so I removed director, top billed star, and distributor from my data. I built my tree using the code in *Figure 16* and it produced the tree shown in *Figure 17*. Because the tree is so large, it is hard to read in that figure, but I will list some of the paths the tree depicts. One path says that if a movie is a part of a movie universe and is classified as an action, adventure, animation, crime, disaster, epic, exploitation, family, sci-fi, spy, superhero, thriller, war, western, or zombie movie it is there is a 0.921 probability of the movie being a failure. Another path says that if a movie is not part of a movie universe, is not classified as an action, comedy, crime, dance, family, fantasy, heist, monster, sci-fi, slasher, space, spy, superhero, or western movie, and was not released in December, February, January, July, June, March, or November there is a 0.86 probability the movie will be a success. Another path states that if a movie is not part of a movie universe, is not classified as an action, comedy, crime, dance, family, fantasy, heist, monster, sci-fi, slasher, space, spy, superhero, or western movie, was released in December, February, January, July, June, March, or November, was not released in 2000, 2005, 2007, 2008, 2011, or 2014, is classified as an animation, horror, or musical movie, was not released in the fourth or fifth week of the month, was released in 1999, 2001, 2002, 2004, 2009, 2012, 2013, 2015, or 2017, and is not rated NR, G, or R, there is a 0.714 probability the movie will be a success. There are many other paths on the decision tree that can help a company determine if a movie will be a success or not.

```
MYTREE <- rpart(Financial.Success~.,data=TRAIN[,-1],cp=0.001)
visualize_model(MYTREE)
```

Figure 16

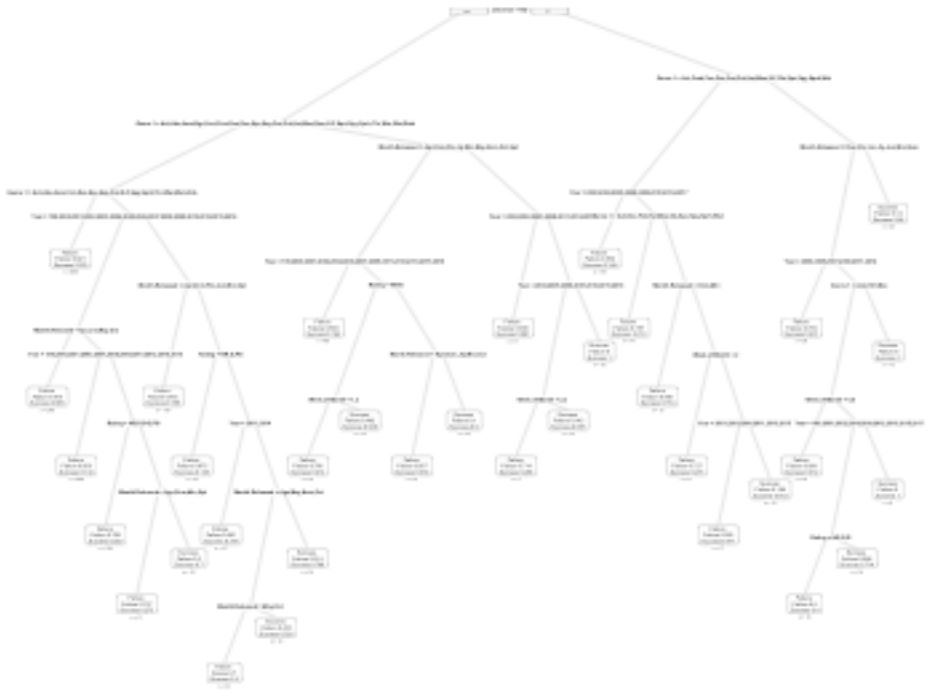


Figure 17

How much money will the movie make at the box office?

In order to predict how much money a movie will make at the box office, I cleaned the data so that it contained a variety of categorical and numerical factors. I removed audience and critic scores from the data set since there were so many movies missing values for those variables; I also removed the data set's original ROI, release date, estimated profit, total estimated budget, and month number. Also, I cleaned some other categorical variables such as universe, genre, and rating in order to make processing easier. In order to have an more normal distribution and better predict the box office earnings, I transformed the values of box office earnings and budget into logarithmic values. I then split the data into a training data set

and a holdout data set so that I could run models on the training set and test the model against the holdout set.

Once the data was cleaned, I ran many different models in order to determine the best for predicting profit. I chose the model based on the root mean squared error (RMSE) that the model returned; a model with an RMSE close to 1 is a model that is predicting well and making few errors. The different kinds of models that I tried were a vanilla partition model, a random forest model, a boosted tree model, a support vector machine model, an XG Boost model, and an R part model. The model that returned the highest RMSE was the R part model, however the RMSE for this model was only 0.53 and the RMSE when predicting on the holdout sample was 0.5; this means that the model is only predicting the box office earnings accurately for approximately half of the movies. *Figure 18* shows the code used to build the model and test it on the holdout data set.

```
library(caret)
fitControl <- trainControl(method = "cv", number = 5, allowParallel = TRUE, verboseIter = TRUE)

#Rpart model
rpartGrid <- expand.grid(cp=10^seq(-4,-3,length=100))
TREE <- train(Box.Office~.-Movie,data=TRAIN,method='rpart', tuneGrid=rpartGrid,
              trControl=fitControl, preProc = c("center", "scale"))

TREE$results[rownames(TREE$bestTune),]
plot(TREE)

postResample(predict(TREE,newdata=HOLDOUT),HOLDOUT$Box.Office)
```

Figure 18

While this model did not predict all that well, I was curious what variables it was using and what was the level of importance each variable. As I expected, the most important variable for predicting box office earnings was the movie's budget, and the budget's importance was much higher than the next variable which was the movie being in a movie universe, as seen in *Figure 19*. After finding the most important variable for predicting box office earnings, the budget, I wanted to test if there was a significant difference in the means of the box office earnings of different categorical variables' levels, such as universe, month released, and the week of the month that the movie was released. This code can be seen in *Figure 20*; this code produced a connecting letters report for each of the variables. A connecting letters report can be used to identify categories that have statistically significant differences by looking for the categories that do not have a letter in common.

```
> varImp(TREE)
rpart variable importance

  only 20 most important variables shown (out of 252)

                                    Overall
Budget                             100.000
UniverseYES                         40.191
Genre.1Horror                       18.767
DistributionCombined                18.345
RatingR                             16.242
RatingPG-13                         10.602
DistributionThe Weinstein Company    9.854
Genre.1Comedy                        9.329
Month.ReleasedAugust                 8.542
Week.of.Month3                       8.509
Month.ReleasedJanuary                7.980
Month.ReleasedDecember               7.735
RatingPG                             7.143
DistributionWalt Disney Studios      6.533
DistributionA24                      5.348
DistributionUniversal Pictures       4.931
Month.ReleasedOctober                4.725
Week.of.Month2                       4.714
Genre.1Drama                         4.469
Genre.1Biography                     4.447
>
```

Figure 19

```
COMPS <- aov(TRAIN$Box.Office~TRAIN$Universe)  #test of significance of difference in means
summary(COMPS)  #the p-value is less than 5% so some groups have statistically significance differences
TUKEY <- TukeyHSD(COMPS)  #set up multiple comparisons
TUKEY  #pairwise differences, if p adj < 5% then statistically significant difference
library(multcompView)
multcompLetters4(COMPS,TUKEY)

COMPS <- aov(TRAIN$Box.Office~TRAIN$Month.Released)  #test of significance of difference in means
summary(COMPS)  #the p-value is less than 5% so some groups have statistically significance differences
TUKEY <- TukeyHSD(COMPS)  #set up multiple comparisons
TUKEY  #pairwise differences, if p adj < 5% then statistically significant difference
library(multcompView)
multcompLetters4(COMPS,TUKEY)

COMPS <- aov(TRAIN$Box.Office~TRAIN$Week.of.Month)  #test of significance of difference in means
summary(COMPS)  #the p-value is less than 5% so some groups have statistically significance differences
TUKEY <- TukeyHSD(COMPS)  #set up multiple comparisons
TUKEY  #pairwise differences, if p adj < 5% then statistically significant difference
library(multcompView)
multcompLetters4(COMPS,TUKEY)
```

Figure 20

The connecting letters report for the box office means of movies that are a part of a universe and those that are not is shown in *Figure 21*. With the two groups having no letter in common, this means that the means for the box office for movies in a movie universe and not in a movie universe are statistically significant. The connecting letters report for the box office means of movies released in each month is shown in *Figure 22*. The report shows that there are statistically significant differences between certain months. For instance, November, July, and December, which all have the same letter in the report, have no statistically significant difference between their means, but all three months do have statistically significant differences in their means when compared to April, October, January, and September, which do not share a common letter with the first three months. The last connecting letters report I ran was for the week of the month that the movie was released, shown in *Figure 23*. The report shows that there is really only one pair of weeks that have a statistically significant difference in box office means, the third week of the month and the fifth week of the month.

```
> multcompLetters4(COMPS,TUKEY)
$`TRAIN$Universe`
YES  NO
"a" "b"
```

Figure 21

```
> multcompLetters4(COMPS,TUKEY)
$`TRAIN$Month.Released`
 November      July       May  December      June     March  February    August     April
      "a"       "a"      "ab"       "a"      "ab"     "abc"     "abc"      "bc"       "c"
  October   January September
      "c"       "c"       "c"
```

Figure 22

```
> multcompLetters4(COMPS,TUKEY)
$`TRAIN$Week.of.Month`
   3    4    2    1    5
 "a" "ab" "ab" "ab"  "b"
```

Figure 23

While the model I built would not be truly helpful to a company trying to predict the box office earnings of a movie accurately, it did provide some useful insight about variables that can be helpful in making the predictions. The model showed that knowing the budget of a movie is extremely important in predicting the box office earnings. The model also provided a list of important variables that could be further examined and used to create reports that showed some useful information about there being differences in the release month and week of a movie.

Conclusions

The use of data analytics in the movie industry will only continue to grow in the future. The industry has only scratched the surface of all of the possibilities for using analytics to make successful movies. Using descriptive analytics can help a company decide on what month to release a movie, what a movie should be rated, whether or not to hire a famous director, or what genre movie to produce. Likewise, a screenplay writer can use descriptive analytics to help choose what distributors he or she wants to pitch a movie to and much more.

Once a movie is already made and ready for release, studios can use predictive analytics to predict how much money the movie will earn at the box office. With information like the budget, the genre, the top billed star, the planned month of release, and the planned week of release, studios would be able to estimate box office earnings, so when the movie is finally released, the studio could study whether the movie earned more or less than expected. If a movie earned less money than expected, the studio can go back to descriptive analytics to try to find some commonalities between the movie and movies that earned a similar amount; the same could be done if the movie earned more than expected. Then, this data can be added to the predictive model to better tune the model so that it might predict more accurately the next time.

This analysis only covered data about the movies themselves and did not include any outside data such as countries released or ticket sale demographics. With movie data and demographic data, studios could use descriptive analytics to target marketing towards specific groups, as mentioned in the introduction.

Works Cited

"Big Data and Hollywood: A Love Story." *The Atlantic*, Atlantic Media Company, www.theatlantic.com/sponsored/ibm-transformation-of-business/big-data-and-hollywood-a-love-story/277/.

Busch, Anita. "MPAA: Hollywood Is Key Driver Of U.S. Economy With $49 Billion Payout, Salaries Much Higher Than Nat'l Average." *Deadline*, 16 Jan. 2018, deadline.com/2018/01/film-and-tv-industry-49-billion-payout-u-s-economy-average-salary-42-percent-higher-than-national-average-1202244209/.

Krigsman, Michael, and Matthew Marolda. "'Moneyball' for Movies: Data and Analytics at Legendary Entertainment." *CxOTalk*, 26 July 2018, www.cxotalk.com/episode/moneyball-movies-data-analytics-legendary-entertainment.

Robb, David. "U.S. Film Industry Topped $43 Billion In Revenue Last Year, Study Finds, But It's Not All Good News." *Deadline*, 13 July 2018, deadline.com/2018/07/film-industry-revenue-2017-ibisworld-report-gloomy-box-office-1202425692/.

Schlesinger, Scott. "Using Analytics to Predict Hollywood Blockbusters." *Harvard Business Review*, 7 Aug. 2014, hbr.org/2012/10/using-analytics-to-predict-hollywood-blockbusters.