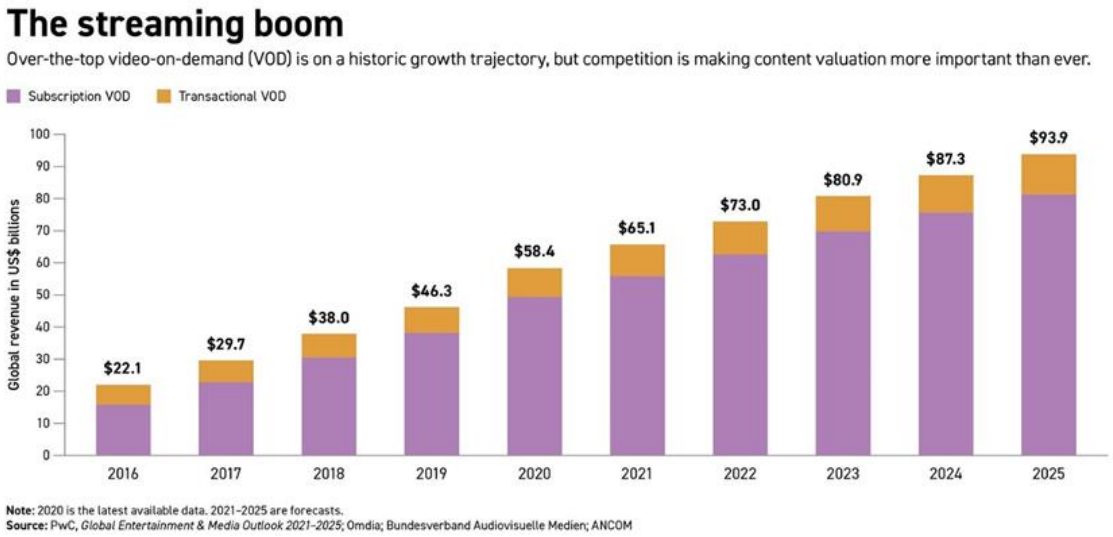
Raven Davis
Paige McLaughlin
Tabassum Shahid
Jialin Yuan

Analysis of Key Factors that Drive Movie Revenue

The Changing Movie Business

The US movie industry is a multi-billion-dollar industry. According to *Motion Picture Production & Distribution - Quarterly Update*, the US motion picture and distribution industry has an annual revenue of approximately $66 billion (pg. 2). However, the rise of streaming and video-on-demand have caused big shifts in how people consume media.

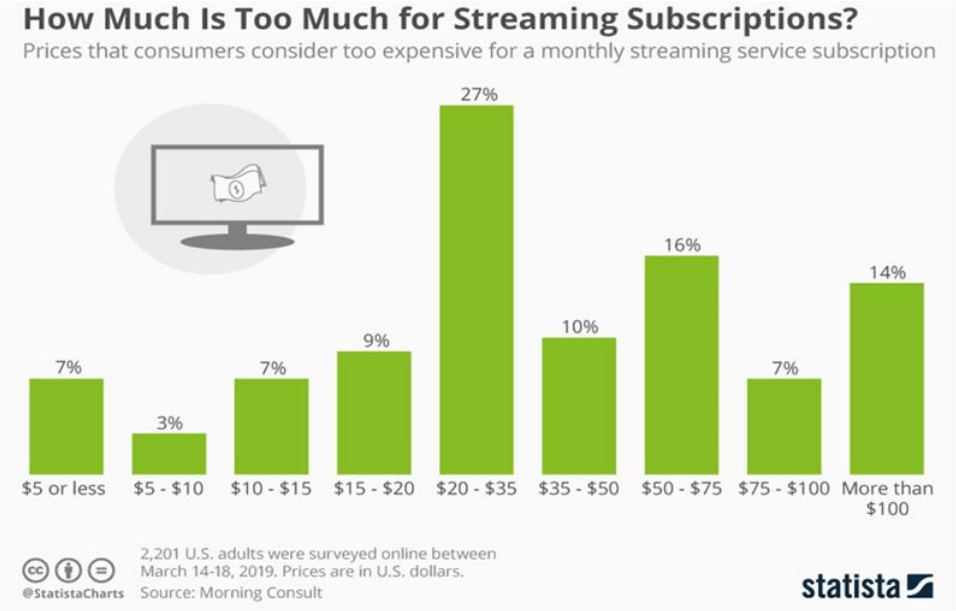


As *First Research* reports, “Even with higher movie ticket prices, total US box office revenue has reported slow growth over the last several years. Global box office decreased by about 72% in 2020, according to Motion Picture Association. As home entertainment systems become more advanced and more consumers subscribe to streaming services, audiences may be less likely to venture to the theater to view movies.” (pg. 8)

This decline was exacerbated by Covid-19. With the temporary closure of movie theaters and either delay or altogether stopping of certain production projects, the motion picture and distribution industry suffered heavily. According to the Encyclopedia of Global Industries, global box office revenue dropped by 72%, while, simultaneously, the home entertainment market grew 23%, with subscriptions to online video services growing 26%.

Netflix, the first streaming company to gain real traction in space, is one of the biggest streaming platforms. It launched its streaming service in 2007, and, since it began releasing original content in 2012, it has gained almost 222 million subscribers in 190 countries (Grimes, 2022). However, despite the huge impact streaming platforms like Netflix, Hulu, Amazon, and now Disney, have made in the movie and television industry, true profitability and longevity in space is still a journey ahead.

Subscriptions are the predominant source of these companies' revenue, and both gaining and keeping consumers continues to be the largest business problem for these platforms. According to a survey by *Morning Consult*, 27% of people find paying between $20.00 -$35.00 too high for a monthly streaming service subscription. Finding a balance between price and holding on to subscribers, while still making a profit, seems to be one of the central conflicts of streaming services currently.



“Streaming services are struggling with subscription volatility as subscribers in the US sign up and cancel subscriptions over a short period of time. Deloitte’s 2022 Digital Media Trends survey reports streaming services churn rate in the US is 37%, while countries like the US, Germany, and Japan only see 30%. As such, streaming services have come up with options to combat the subscriber “churn and turn.” Streaming services can evaluate their content, ensuring there is fresh and relevant media for subscribers to keep them on despite the cost.”

In order to maintain relevancy as well as growth, streaming platforms and movie and television companies are now shifting focus toward producing original content on these platforms, in an effort to hold on to subscribers, as well as investors.

The crux of the issue appears to be content. Billions of dollars are now being invested into the production of original content, with media companies spending unprecedented amounts of money. As investors become wary and the streaming space becomes more competitive, content will be the huge differentiator if these platforms hope to generate profit and satisfy stakeholders.

We initially hypothesized that genre, director, and production company would be most impactful to overall profit and return on investment. Our experiment is focused on understanding any trends or patterns that appear to impact a movie’s ROI. We intend to create a model that will help to determine the characteristics that have the greatest impact on ROI. Our hope is that we can then use those characteristics in a predictive model that will ultimately help streaming companies in desperate need of quality content to generate the maximum ROI for their company and investors.

We intend to investigate the following questions to better aid in our understanding of the different factors that might impact a film's ROI and its success in general:

1. Is genre a good indicator over time, or only during certain time periods?
2. What variables influence revenue?
3. Is there a relationship between director/cast and revenue? Does the type of genre change this relationship?
4. What is the top genre of movie that the top ten directors produced in the United States?
5. Which years/seasons were the most successful for movies of a particular genre?
6. What are the top 10 movies by production companies?

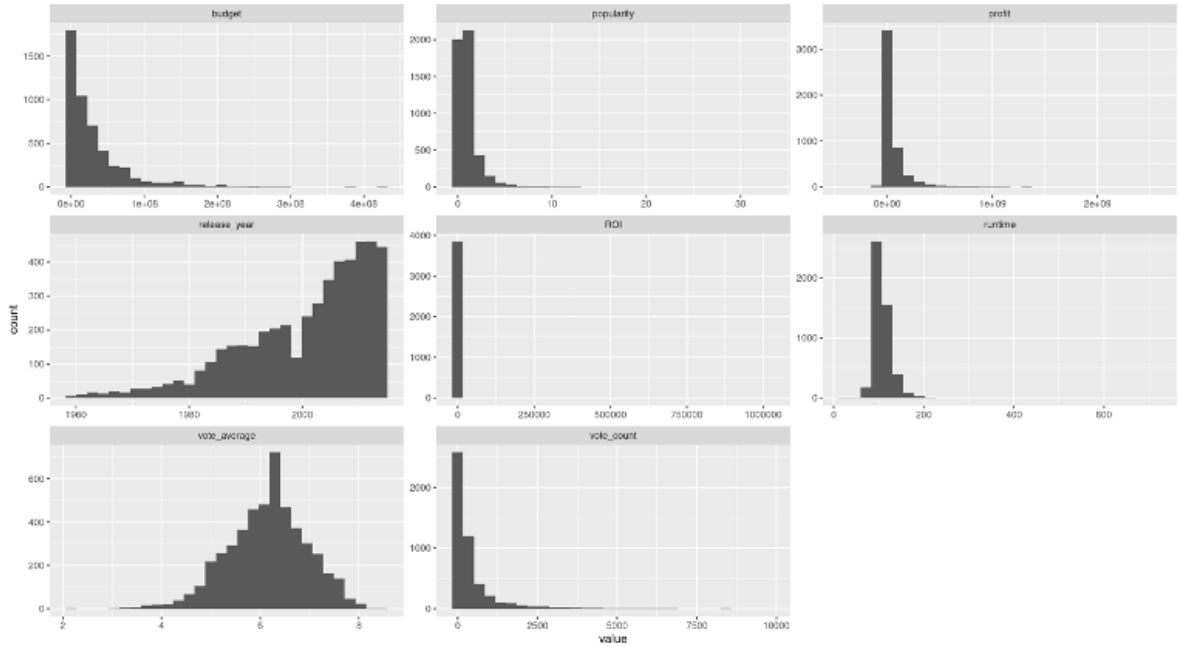
Exploring the Data

There are many free and open-source data sources that can be found in websites like Data.gov, Kaggle, GitHub, Google Cloud. We looked into some of the recommendations provided by the TAs of MGT 6203 course. After confirming a topic and discussing its business need, we found many datasets on Netflix, Amazon, and many other streaming services. Our goal was to use datasets that have revenue and budget columns that can be used to result in the greatest ROI between key variables of popular movies both on and off streaming sites.

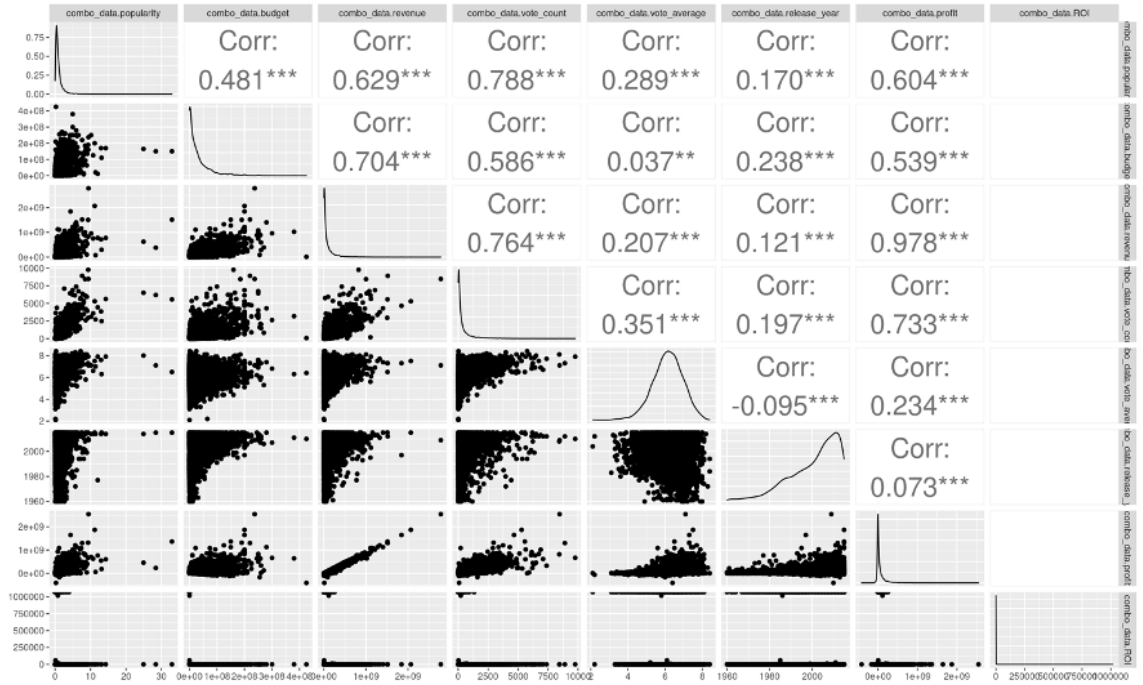
We used a dataset of 10,000 observations from TMDb (The Movie Database). This dataset includes detailed information of movies including title, director, release date, revenue, and budget. We also used a dataset from Kaggle, "TV Shows on Netflix, Prime Video, Hulu and Disney+," which has detailed information about movies and TV shows including title, release year, age demographic, IMDb ratings, Rotten Tomato ratings, and the platforms the movies are available on. We also found a dataset created by an analyst online called "Movie Stream," that includes detailed information about movies and TV shows, including streaming services, rating, directors, countries available, and language.

The key independent variables we focused on were genre, directors, release dates, movie rating, production company, and budget. We are creating a new variable, ROI, that was the difference in revenue and budget, which we used as our dependent variable.

After cleaning the data, we performed an initial exploration of the data to gain a high-level understanding of its structure using the head and str functions. We then used the summary function to compute descriptive statistics for the variables in the dataset. This function provides information about the fields, minimum and maximum values, median, standard deviation, percentage of missing data, the number of unique values, and the mean. All of these statistics can provide value and necessary context during our analysis. To comprehend the value distribution, we used a histogram. All of the variables appeared to be skewed. The only variable that appears to be close to normally distributed is "vote average."

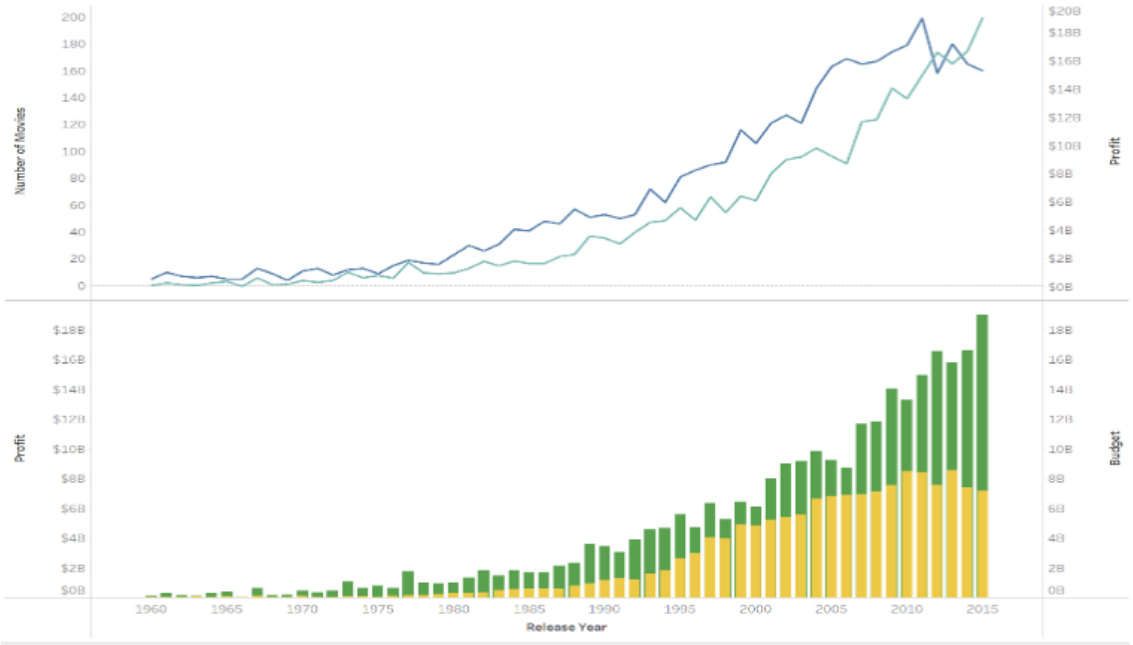


We also used pairs plots to visualize the data and understand the correlation between variables.

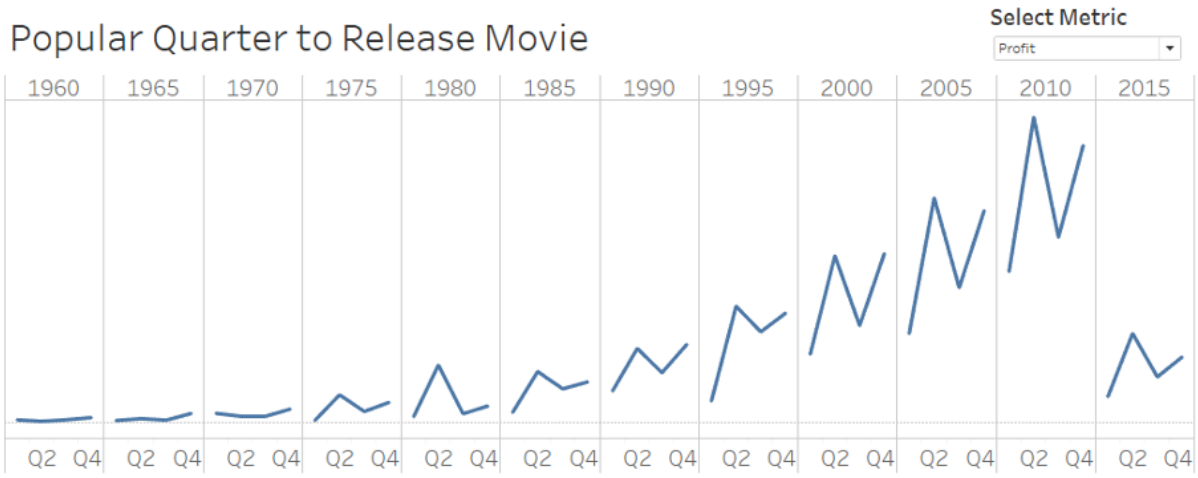


Profitability appears to be highly correlated with revenue, which makes sense. We noted that vote count was more highly correlated with popularity and revenue than the other variables, which also seemed logical. Movies with a high level of popularity appear ed generate more profit and revenue. High-budget films generated more revenue, but not necessarily more profit. These relationships, however, are merely correlations and do not imply causation.

We developed a tableau dashboard to gain further insights into our data with the hope of gleaning valuable information about any patterns or relationships between various attributes that might exist.

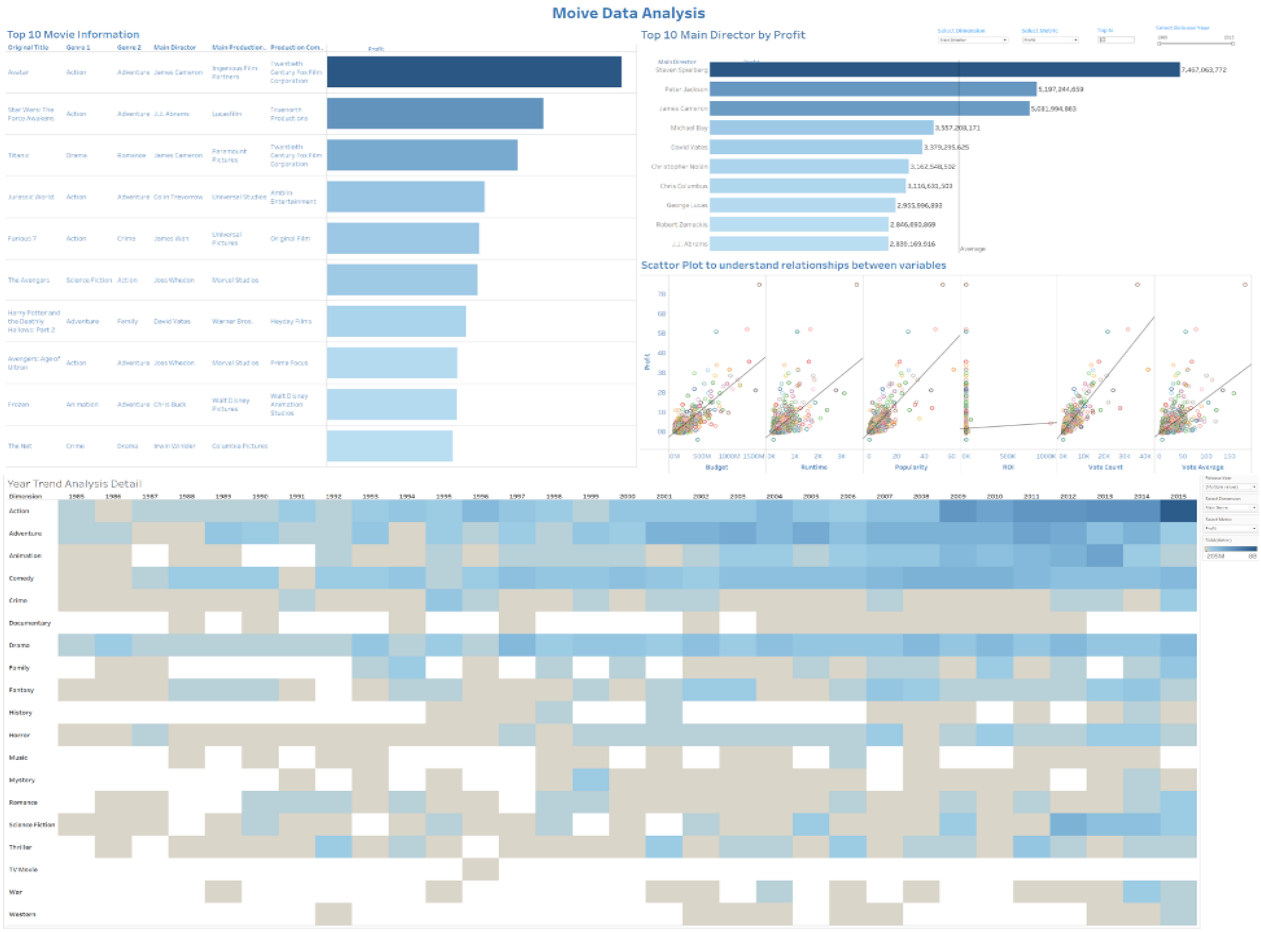


The above graph shows that the number of movies produced per year has increased over time. The frequency of films has increased from five films in 1960 to 160 in 2015, with a peak of 199 films in 2011. Profits have also risen as more films have been produced. However, the cost of producing movies has also increased, necessitating a larger budget. The number of films being released finally began to decline around 2012.



We used the graph above to investigate when movies are more likely to be released. According to the graph, Q2 and Q4 are the times of the year that movies are most likely to make a profit.

Using various metrics, we analyzed the top “n” dimensions, including genre, production company, and director. We also used a scatter plot to understand how the relationships between the various variables change as the dimensions change. Finally, we developed a heatmap to show the change based on the metrics and dimensions.



Action, adventure, comedy, drama, and animation are the top five most lucrative genres. The majority of the top ten films, including *Avatar*, *Star Wars*, *Jurassic World*, and *The Avengers*, are action-adventure films. It's important to keep in mind that most movies were classified as more than one genre, so it is critical for production companies to understand which combinations of genres are best for investment.

Comedy, drama, and adventure have a high popularity. Profits by genre are shown on the heatmap above from 1960 to 2015. Higher profits are indicated by darker blue fields, while lower profits are indicated by a lighter gray. Profits are generally rising over time, particularly for the genres of action, adventure, animation, family, and drama. The top three most profitable film production companies over the course of the time period were Universal Pictures, Paramount, and Walt Disney. It appears that Paramount held the top spot early on, but the company has steadily lost market share over time.

Many successful films were made by the afore-mentioned production companies, such as the high-grossing *Furious 7*. Several production companies, including Paramount Pictures and Twentieth Century Fox Film Corporation, worked together on several of the movies. The most successful director is Steven Spielberg, based on profit. James Cameron and Peter Jackson follow behind Spielberg. The director of the highly profitable *Jurassic Park* is Steven Spielberg. The director of the successful *Lord of the Rings* franchise is Peter Jackson. James Cameron is best known for directing *Titanic* and *Avatar*.

To see if there is a linear relationship between different variables, we displayed the correlations in scatterplots shown above. When profits are grouped by genre, they are highly correlated with budget, popularity, and vote count. Run time and vote average show a

moderately linear positive relationship when production companies change. The relationship between those variables appears to have been reawakened by the change in director.

These relationships are merely correlations, not indications of underlying causes. Since the strength of relationships had not been evaluated statistically at this point in our experiment, we still needed to investigate further to really understand the relationships between each of the variables.

The validity of our recommendations was also impacted by several constraints, biases, and data quality issues. For instance, many films had a $0.00 budget and $0.00 revenue. Additionally, there were some movies that had a budget of only a few hundred dollars or a revenue of around the same amount, which stand out as anomalies. After removing this data, the analysis may only be a best guess based on the data that is available. We also need to learn more about the specifics of the budget and income projection, as well consider the movies that have not yet been released by the survey date but have the potential for massive sales.
We have a link to our Tableau dashboard here:
https://public.tableau.com/shared/3MPGC7XG3?:display_count=n&:origin=viz_share_link

Data Cleaning and Data Transformation

To begin processing our data for our model, we had several factors that needed to be transformed in order to be helpful. We wanted to explore the relationship between a film's director and its ROI, but the director variable included multiple directors per film, in the same column of the dataset. We decided to extract the first director listed, assuming that the directors were listed in order of importance since they were not listed alphabetically. We put this director's name into a new column called "main_director" to better model the relationship this variable might have with our outcome variable. We did the same process with genre, extracting the first genre for a movie that was listed in the column, and putting those genres into a new column called "main_genre" to better analyze its relationship with ROI. We also attempted to do this with the "production companies" variable, and transferred those primary production companies into a new "main_productioncompany" column. However, this proved hard to work with as, many times, the same production company would be listed in the dataset under variations of the same name.

Lastly, we processed the "cast" variable in the same way, extracting the first actors, presumed to be most important, from the column for each observation and putting it into a new column, "main_actor."

The dataset already had a "release_date" variable that gave the month, day, and year a film was released. We decided to transform this variable into a "season" variable, taking each release date and categorizing it into the appropriate season based on the month and day the movie was released. We felt this would be more helpful, as releasing a movie in a particular season is a variable that production companies can better control. Once we looked further into the data, we released that the ROI column included "Inf" values that we could not include in the model, so we filtered those values out.

Finally, we removed redundant columns and columns that we did not feel would have any bearing on ROI. The dataset included a "budget" and "budget_adj", "revenue" and "revenue_adj", "tagline", "keywords", "overview", "vote_count", "vote_average", and "profit" and "profit_adj." Since we were attempting to explore relationships with ROI, we removed the profit variables and revenue variables. Also, we removed columns describing the film, since they are unique to each film and would be difficult to examine in terms of impact on ROI. We also removed the variables related to votes, since votes would be collected after a movie was made and would not be controlled by a production company to influence ROI.

It was difficult to find a method to clean our data, and a strategy for which variables to use, and what values to extract. We finally determined that the first name listed in the director, production company, and cast columns were the most significant, and worked on putting them in their own column so that we could investigate their relationships with ROI more easily. Looking a step further showed that there was too much variability in director to have it be a good variable in the model.

We also had trouble finding a way to clearly represent the data in the production company column. With so many production companies listed multiple times under variations of their name, we struggled to find a way to accurately represent the production company variable in our data. We finally ended up removing the column, and focusing more on the other variables we had that were easier to model.

A Model that Predicts ROI

We used a multiple linear regression model, several logistic regression models, lasso regression and ridge regression models, a forward-stepwise regression model, and a random forest model for our experiment.

We chose linear regression to see if we could predict ROI in general. We fell short with none of the independent variables being significant. The business problem we were exploring was a classification problem—whether a set of factors led to a certain threshold of ROI or not—so we decided to use classification models for our prediction.

We tried creating a logistic regression model using two binary variables, one variable to represent an ROI greater than the median ROI in our data, and one variable to represent a positive ROI. Both logistic regression models yielded significant results, with above 50% accuracy. We also looked at a random forest model to see the importance of each variable that we found significant in our logistic regression models.

A challenge we all face when modeling is determining what is most important to focus on and which model is best. Should we focus on accuracy, specificity, or sensitivity? Both logistic regression models we ran had benefits to their model. We also weren't sure which features to focus most on. There were certain features, like "main_genre" and "runtime", that seemed to be significant in most of the models we created, but there were others, such as "main_productioncompany", that we intuitively felt should be significant, but were not in our model.

We divided our clean data set into two parts, 70% being used for our training data. This was to avoid training bias. We ran the logistic regression model a handful of times on a few different training sets to see if there was consistency in significant factors. After creating our models, we evaluated Goodness of Fit, and viewed which coefficients were significant at an alpha of 0.05. We also ensured the models fit the data by calculating the models' overdispersion, which was less than two for each model. We used the remaining 30% of our data to test how well the model performed. We used our model to see how well it would predict if ROI would be positive.

We used a confusion matrix to view accuracy, sensitivity, and specificity. We were especially concerned with accuracy and sensitivity, as the confusion matrix automatically counted any observations classified as "0" as a "success", while "0's" in our model were actually indicators that the movie did not yield ROI. We wanted to limit false positives, concluding that the cost of production companies creating a movie that was falsely predicted as a success would be much more valuable to production companies and investors.

Overall, we decided to choose the model with the highest accuracy and highest sensitivity. The first logistic regression model, we used "budget_adj," "runtime," "main_genre,"

“main_director,” and “season” as the predictors. “Runtime,” “animation_genre”, “comedy_genre,” “family_genre,” “science fiction_genre,” “horror_genre,” “Summer_season,” and “main_director” were all significant at the 0.05 level or greater. The AIC for the model was 3649.3. Another logistic regression model added “main_genre x season” as an interaction term, and a third logistic regression did a log transformation of the “budget_adj” variable. The forward-stepwise regression model identified “main_director”, “main_genre”, “runtime”, and “season” as significant, and had an AIC slightly lower than the first model at 3648.9. The lasso regression model identified “budget_adj”, “runtime”, "main_director", “main_genre”, “main_productioncompany”, and “season” as significant.

The random forest model identified “runtime,” “main genre,” and season as significant, and performed the best of all our models in regards to accuracy and sensitivity. Our model has an accuracy of 75% and specificity of 99%. Very few profitable movies are predicted to have a non-positive ROI.

```
          Reference
Prediction   0   1
         0   4   8
         1 290 872

               Accuracy : 0.7462
                 95% CI : (0.7202, 0.7708)
    No Information Rate : 0.7496
    P-Value [Acc > NIR] : 0.621

                  Kappa : 0.0066

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.013605
            Specificity : 0.990909
         Pos Pred Value : 0.333333
         Neg Pred Value : 0.750430
             Prevalence : 0.250426
         Detection Rate : 0.003407
   Detection Prevalence : 0.010221
      Balanced Accuracy : 0.502257

       'Positive' Class : 0
```

We chose to run a logistic regression model with a variable indicating if the ROI is positive as the dependent variable. This model proved to have the highest accuracy and specificity.

Our findings differed a little bit from our initial hypothesis. We did believe genre would play a role in determining ROI. We also thought that the director and production company would play a role, but we did not include those in our model. We found that the genres horror and drama play a role in determining ROI. We found that runtime and season also played a role in determining ROI, but that was not something we initially hypothesized.

It’s tough to determine if a movie will be successful. Return on Investment is just one way to determine a movie's success. Our model showed us that releasing a movie in the summer

and creating a drama movie could lower your chances of having a positive ROI. On the other hand, creating a horror movie and increasing the runtime could increase chances of a positive ROI. The median runtime of movies with a positive ROI is 106 minutes.

Like mentioned above, it's hard to determine if a movie will be successful. Although it's not a perfect science, focusing on genre and runtime could help produce a profitable movie.

PROJECT TIMELINE/PLANNING

Phase 1 - Completed

i. Team formation

Phase 2: Due Sunday 9 October (3 weeks later) – Completed

Team due date: (10/04)

i. Project proposal or plan in text format

a. The team identified the problem together and decided to analyze key factors starting off by working on the proposal template.

Phase 3: Due Sunday 30 October (3 weeks later) – Completed

Team due date: (10/26)

i. Plan presentation video (3-5 min)

a. The team will work on creating a presentation in putting together the results of the initial EDA, business justification, modeling, literature review, and data visualizations by 10/26.

ii. Progress report in text format (3-5 pages)

a. Raven will gather literature and begin adding relevant quotes and sources to the progress report by 10/18.

b. Callie will explore the data and provide basic knowledge of the data by 10/17.

c. Paige will begin modeling the data and report back to group any meaningful insights by 10/17.

d. Tabassum will gather business justification and include an overview of the problem in general by 10/15.

e. The team will be putting together the results of the initial EDA, modeling, literature review, and data visualizations by 10/25 in the progress report.

Phase 4: Split Due Dates (2.5 to 3 weeks later) - In Progress

Team due date: (11/20) - On Track

i. Final presentation video: Due Wednesday 16 November - Completed

a. The team will work on the final presentation in putting together the results of the final EDA, modeling, and data visualizations by 11/10.

ii. Final Report text (10-12 pages) Due Sunday 20 November

a. The team will be putting together the results of the EDA, modeling, and data visualizations by 11/18 in the final report.

iii. Other Final Report components: slides from your video presentation, R code, data, charts, graphics…

Phase 5: Due 5 December

Team due date: (11/27) - On Track

i. Peer reviews of three Final Videos done by other groups

a. Everyone in the team will participate in peer revieing of Final videos done by other groups.

ii. Peer reviews of your teammate's performance in your group

Works Cited

“Motion Picture Production & Distribution - Quarterly Update 4/11/2022. (2022).” Mergent, 2022. First Research Industry Profiles. Web. 4 April 2022.

"Motion Picture and Video Production and Distribution." Encyclopedia of Global Industries. Farmington Hills, MI: Gale, 2021. Business Insights: Global. Web. 26 Oct. 2022.

Grimes, Christopher, and Anna Nicolaou. “Are You Still Watching? Netflix and the Future of Streaming.” Financial Times, 22 Apr. 2022, FT.com. Web. 27 Oct. 2022.