DECISION 520Q
Data Science for Business

Final Project

Hierarchical Clustering for Film Success and Application in Prediction

Section A, Team 31

Alice Lin, Karry Lu, Maria Olmos, Sherie Zhang, Oliver Zheng
Professor Natesh Pillai
Nov.16th, 2020

Abstract

Many companies collect information about customer interactions with their products. Yet in the traditional filmmaking industry, the application of data analytics is sparse and film companies are limited to a few generic standards, such as online reviews, to evaluate film success. Also, existing metrics are retrospective in nature, and film producers may wish to plan ahead when deciding how to allocate creative and financial resources, how their films look, and how their contents should be delivered. We attempt to create a comprehensive and systematic way to segment and predict film outcomes through machine learning, with the goal of helping film producers decide their target audience and tailor film content and marketing efforts.

Business Understanding

Competition in the film industry has increased dramatically over the past few years as entry barriers are lowered thanks to technology, making content creation accessible to everyday people, and increasing demand for original content. In this age of rapid change, producers stumble upon the question, “what makes a film successful?” Common metrics to determine a film's success are ratings and reviews, such as those on IMDb or Rotten Tomato. However, film ratings rely on viewers' personal opinions and do not consider all possible factors that make a movie “successful” or “rewarding” to the film producer. For example, Disney’s poorly reviewed Aladdin brought in a whopping $86 million in the opening weekend, but the highly-rated coming-of-age comedy, Booksmart, brought in a mere $6.5 million. We want to create a more holistic measure of “success” by aggregating factors such as box office, amount of public interest generated, recognition from professional critics, and long-lastingness (being regarded as a “classic”). Our model is tailored towards the filmmaking industry, and it is predictive and easy to apply. Film producers can use our model to assess a movie’s outlook prior to production to determine whether the movie will be worth their effort and investment. They can also use our model to identify and focus resources on aspects of a movie integral to their particular vision of “success.” Last but not least, by predicting and understanding where a film falls in terms of three key dimensions we identified - monetary return, quality, and popularity - film producers can more accurately target the right audience.

Data Understanding and Preparation

We retrieved our main dataset from Kaggle (Leone, 2020). The raw data was scraped from the publicly available movie rating website IMDb and last updated in January 2020. The dataset has 85,855 unique observations of films, 71 columns containing variables such as genre, year produced, and average rating, as well as two supporting tables that offer additional information related to production and cast members. We also joined film revenue and budget information from a second dataset (Shakir, 2018) and the Academy Awards record from a third dataset (Fontes, 2020).

For data cleaning, we used a two-stage approach, horizontal and vertical. We started by adjusting data type column-by-column and extracting useful information by separating or combining data. For columns with multiple values, we kept the first value as the primary attribute and created a dummy column to identify whether that attribute contains multiple values (e.g. pri_country and multi_country, pri_genre, and multi_genre). Then, we substituted or dropped abnormal values for each column after cross-checking the information on Google or other international sources. Several columns were dropped for reasons such as irrelevancy (i.e., Description column), having over 90% data entry errors (i.e., Gross income column), and lack of reference (i.e., Metascore column). For foreign currencies, we used 10-year average exchange rates to convert them into US dollars. To address the Null values in the budget column, we joined additional datasets with relevant information through the unique “imdb_id”. When mapping movies during exploratory data analysis, we encountered issues of countries' names no longer existing, given the dataset spans across a century. For these movies, we either assigned them to their successor countries or removed them if the associated number of movie records was small enough.

Several steps in data cleaning were completed in MySQL for maneuverability and speed. Additional minor cleaning, transformation, and calculation were performed along with exploratory data analysis and model construction.

Methodology

Typically, a predetermined categorical variable is needed as the target variable for classification predictions. Instead of arbitrarily choosing a variable that may be either too narrow in scope or already exploited by film producers, we wanted to create a new categorical variable that comprehensively assesses a film’s success. The specific attributes that fed into our new standard were selected based on descriptive value for envisioning and defining various types of success for films.

We split our primary dataset into 95% (A) and 5% (B). First, we used Set A for unsupervised learning, applying Hierarchical Clustering analysis to find the most appropriate clusters. This created a way of categorizing films for success. Using the newly defined categorical variable from clustering, we applied supervised learning and further split Set A into 70% training and 30% testing. We then built a Random Forest Classification model that predicts film success using the training set and used the testing set to examine the robustness of our predictive model. Finally, we used Set B as a hypothetical set of new movies to validate our classification model and generate insights for model deployment.

Exploratory Data Analysis

Our cleaned dataset contains over 11,000 film records, with incrementally more films available in more recent years. As shown by the frequency histograms in Figure 1, film durations are normally distributed, with median and mean durations being 107 and 103 minutes, respectively. This implies that our data is a representative sample of the typical 90-minute film population of interest, although skewed by some outliers like the Lord of the Rings trilogy. Over half the films in our dataset are US-produced, and English is the primary language in 71% of the movie. Notably, the film budget distribution is heavily skewed to the right due to some ultra-high-budget adventure and superhero films like the Avengers. The median budget is USD 8.8M, which equates to a medium-low-budget film in our industry context. Highly correlated with the film budget is film revenue as shown by the correlation matrix. Revenue is even more skewed than budget, with a few large outliers and underreported revenue data from smaller films. As a result, we did not rely on revenue data and chose budget as a proxy for monetary return on film production since revenue is 0.755 correlated with the budget. The number of written reviews from critics and regular viewers is understandably concentrated at zero and right-skewed, with a high number of lesser-known films getting no written reviews. Though unevenly distributed, the number of written reviews provides an imperfect but intuitive measure of film popularity. Lastly, the weighted average rating score ranges from 1 to 10 and provides a fair measure of movie quality based on the judgment of the crowd. Rating scores are normally distributed, with a mean of 6.2 and a median of 6.4 – in line with what we would expect.

To understand the relationship between different variables of interest, we proceeded to two-dimensional exploratory data analysis. We created two heat maps (Figure 2.1 and 2.2) to compare the number of critic reviews and average ratings across different countries. Based on the first map, we were surprised to observe that films from New Zealand are the most popular (highest average number of critic reviews) followed by North American (the U.S. in particular), Asian and European films. Digging deeper into films by country, we realized that films from New Zealand mainly consisted of Lord of the Ring and The Hobbit films which brought up the popularity overall; these films were in fact produced in New Zealand but financed by American Investors. Aside from New Zealand, other trends were both expected and acceptable, as our data comes from IMDb, which is more widely used in the U.S. and our model mainly caters to American film producers. Interestingly when it comes to the average rating, foreign films, especially those from the U.K. and Algeria, dominate American films. The highly-rated films from Africa had very few but high ratings; this is likely biased, as poorly made and lesser-known foreign films may not get documented on IMDb. We used a boxplot (Figure 3) to examine the relationship between average score and different genres of movies, and there appears to be a pattern: documentary, Film-Noir, and musical all have fairly high average scores with small standard deviations in comparison to other categories of films. Genres including horror, sci-fi, and thriller appear to have lower average scores and larger variations. Thus, based on the plot it is evident that there is some degree of correlation between movie genre and the average rating.

To understand how the average rating of a movie would impact its popularity, we segmented the population-based on three different demographics: gender, residency, and age group. Based on the plots Figure 4) we can conclude that in general, regardless of demographic information, people tend to seek out and review films with higher average ratings, contributing to these films' higher popularity. When examining what kinds of people are filling out film reviews, we found that there are far more males leaving ratings on IMDb than females. There are also many non-US users active on the platform. The plot also shows that the age group between 30 to 45 years old gives more ratings compared to other age groups. And not surprisingly, people over 45 years old give the least amount of ratings among all age groups.

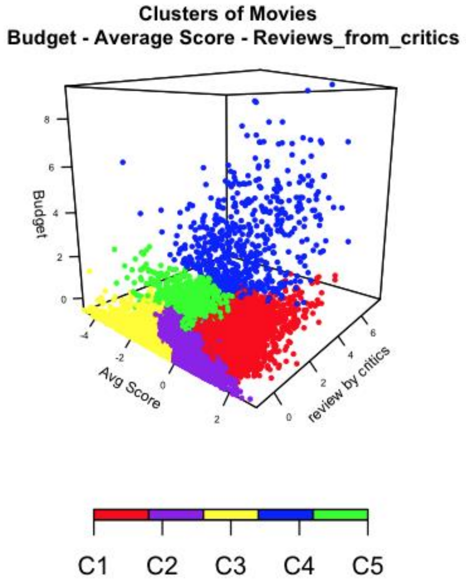
To expand our dataset and capitalize on our availability of actor, director, and writer information, we constructed a cast and crew score utilizing an outside dataset that documents Oscar Award nomination and winning records. We intended to examine whether the fame and credibility of cast and crew influence a movie's popularity and quality. Our score takes into account how many Oscar nominations a film's cast and crew have received in the past, and assigns higher weights (four times, since winning an Oscar means defeating four other nominees) for any actor or crew member that wins an Oscar. The blue trendline in the resulting scatter plot (Figure 5.1) shows that greater cast and crew fame, as measured by our score, does bring out higher movie ratings, and the congregation of rating scores in the high range among movies with high cast and crew score suggests that famed actors and directors may guarantee an above-average level of film quality. However, the red trendline in Figure 5.1 suggests that critics don't pay more attention to movies with higher cast and crew score. In fact, they avoid reviewing movies that may have been overhyped due to an "all-star" lineup, but rather prefer movies with low cast and crew fame, or those with up-and-coming actors, directors, or writers.

To analyze the information on the director and production company in further detail, we structured a three-dimensional visualization with popularity metrics measured both by the average number of users review and critic review, and the average rating score of the films they direct or produce. Then, we ranked the directors and production companies by the rating scores and picked the top 50 of each to visualize. According to the graph, several highly rated "user-friendly" production companies (those getting more than 1400 user reviews), such as Orion-Nova Productions and Cecchi Gori Pictures, are not well received by critics. Similarly, two critics-favored production companies attract mediocre popularity among ordinary viewers. As interesting as the companies, several median rated directors are actually spoiled by critics, getting more than 300 critic reviews. And two terribly rated directors, Lenny Abrahamson and Damian Szifron may have triggered some serious criticism and public outcry.

Models

Part 1 - Clustering Model Analysis and Interpretation

To construct our standard of success for films, we implemented an Agglomerative Hierarchical Clustering model. We were explicitly using Ward's minimum variance method with Euclidean distance between data points for a three-dimensional clustering model. We tried 4 and 6 dimensions for clustering our dataset, which would achieve better performance. However, for visualization and business interpretability purposes, we decided to keep dimensions at 3 as a safe and manageable starting point. Future improvements will be discussed in later sections.



We tried out different combinations of features and eventually settled on budget, average rating score, and a number of critic reviews. From the business point of view, we believed these three attributes come together to best support and represent the reputational and financial trajectory of films. From a modeling perspective, this combination provided optimal clustering results from the information we possessed. By examining the dendrogram (Figure 6.1) of the clustering model, we decided to define five different clusters according to the output's business interpretability and clustering statistics.

We performed relative cluster validation by trying out different values of k, meaning the number of clusters, and decided on k=5. Then, we further validated our model by internal cluster validation. The Silhouette width of 0.3 being closer to 1 instead of -1 suggests that the dataset is well clustered in the sense that films in each of the four groups are alike (Figure 6.3 and 6.4). However, the Dunn Index indicates that the ratio between inter-cluster and intra-cluster separation is not optimal (Index value is considerably small). For example, cluster 4 shows a large intra-cluster separation, whereas the inter-cluster separation between clusters 2 and 3 is relatively small. Given the clusters were created by us, meaning there are no previously determined same cluster labels from this dataset, we were not able to perform external cluster validation on this model. Interactive 3-D visualization for our clustering model can be downloaded at https://github.com/KL98/Project/blob/main/movie_5_score.% 23review_critic.budget_11.12.html

Each of the three axes of our 3-D clustering model represents an important aspect of movie success. The "budget" dimension corresponds to a film's monetary success and return on investment. Due to the budget variable's high correlation and slight upward quadratic relationship with revenue, we can assume that a low-budget film will bring in lower profit too, and a high-budget film will bring in higher profit. Exceptional circumstances are unlikely, such as when a low-budget film achieves high box office; even when that happens, we would expect the popularity or quality dimensions to provide additional insights and complete the story for us. When the budget dimension is interpreted along with the two other dimensions, it provides a quick way to gauge a film's return on investment. The "average score" dimension is more self-explanatory and corresponds to a film's quality and audience reception. Finally, the "reviews from critics" dimension corresponds to a film's popularity. Film critics are supposedly alert and aware of popular films; so, the more written reviews a film receives from critics, the more popular it is likely to be, irrespective of the reviews' actual sentiment (which should be more appropriately captured by the quality dimension). Putting the three dimensions together, we were able to verbalize what each of the clusters represents, detailed below:

Cluster Nickname:	**1. Quality Works**	**2. Indie Rare Gems**	**3. Complete Flops**	**4. Mediocre Big Releases**	**5. Safe Bets**
Monetary	Medium budget and profit	Low budget and profit	Low budget and profit	High budget and profit	Medium budget and profit
Quality	High score	High score	Low score	Medium score	Medium score
Popularity	Medium/high	Low	Low	Medium/high	Medium
Example	The Green Book (2018)	Once (2007)	Sharknado (2013)	Captain America: The First Avenger (2010)	Good Boys (2019)

We gained two critical insights from the above cluster interpretations. First, budget positively correlates with popularity across the board. So the more film producers spend, the more likely their productions will invite public attention. Especially for the "Mediocre Big Releases" cluster, the stretched out shaping of blue dots in the 3-D cluster diagram demonstrates this cluster's strong potential for creating blockbusters if producers were willing to spend more, although this has unclear implications on film quality. Secondly, there's a thin line between "Quality Works" and "Safe Bets," but the former's outcome is significantly better. With a medium budget, film producers need to really control and maximize the quality of their works in order to have a chance at getting their names out; once they establish quality, recognition and future opportunities may follow.

Part 2 - Classification Model Analysis and Validation

With the newly created cluster label 1 to 5, we then built our classification model to predict the target cluster, or in our business context, the standard of success of films. After comparing the prediction accuracy and performance of different algorithms including SVM, K-NN, Decision Tree, and Random Forest, we decided to build a Random Forest Classification Model due to its robust predictive accuracy and capability for covering a large number of features.

We set up 18 columns of film information from the original dataset and randomly split the dataset into a training set and test set by 70% to 30% proportions. Then, we trained the model using 100 trees, which proved to be optimal by trying different values, and applied predictions on the test set. Consequently, we achieved a prediction accuracy of 93.3% and error rate of 6.7% as shown by the confusion matrix below. To deal with bias and further validate our model, we applied k-folds cross validation with k=10; we were able to improve the prediction accuracy to an average of 97.18%. Our model is performing well with an AUC of 0.918 (See ROC curve in Figure 7).

CM	1	2	3	4	5	Pr
1	543	46	0	3	12	89.9%
2	17	1243	43	0	14	94.4%
3	0	6	930	0	17	97.6%
4	10	0	0	171	5	91.9%
5	26	13	14	6	337	85.1%
Re	91.1%	95.0%	94.2%	95.0%	87.5%	93.3%
k-Folds (k=10) Accuracy	97.1%	97.3%	97.7%	96.0%	95.9%	
	97.5%	96.7%	98.9%	97.6%	97.2%	

Prediction Evaluation and Model Deployment

Our Classification model was designed as a tool for film producers to predict the outcome of films, and to adjust investment and positioning of their films accordingly to optimize returns and film success. As discussed before, the Classification model predicts the cluster labels generated from our Clustering model. Looking at the prediction results, we are confident that our proposed 5 clusters of categorizing movies are comprehensive and intuitive for film producers or investors to set up goals and expectations; and our prediction model is robust in helping them make strategic decisions and detailed projections. Moreover, our model is able to handle a large number of proposed or potential films that producers are designing, and aid their decisions on what each of their movie attributes should be like, in order to achieve specific goals.

Suppose we have been approached by a few producers, each of whom is designing or planning a new film. They all have an expectation regarding which of the 5 clusters their film will be in, but like constructing a stock portfolio, they are not sure what combination of attributes will give them what they want. If one of the producers wants her film to end up in category 2, which is an “Indie Rare Gem”, she needs to decide which director to call, which actors to invite, how much money to invest, where in the world they should produce, etc. All these factors need to be carefully considered, which will involve hundreds of combinations (think about permutation/ combination of multiple choices). If the producer comes to us with her expectation and ideally some possible combinations of target directors, actors, or budget level in mind, we will first conduct a quick interview with her to record as many attributes as she can provide. With the use of our model, we will be able to instantaneously tell her whether the combinations will work, and we can then work with her further to test out possibilities and configure each of the attributes to increase likelihood of creating a film that matches her vision of success.

Cost-Benefit Analysis

As discussed above, our model can help producers create their own film portfolio based on their specific needs and help decide whether a script they have on hand will yield a positive return on investment (Mueller, 2020). To further illustrate the benefits of our model, we conducted a cost benefit analysis. For our analysis, we will focus on implementing our model in Hollywood (Follows, 2020). On average Hollywood makes 600 films a year with a profit-loss ratio of 51-49 and average budget of a film at around $100 million. While the profit and loss of movies vary greatly, for our case we will assume that a successful film will generate triple the budgeted cost: $300 million and a non-successful film will only return half of the budget, or a $50 million loss. Under our assumptions, 294 movies produced will incur a total loss of $14.7 Billion dollars. Based on our cluster model interpretation, we assume movies in categories 3 are unsuccessful and the remaining are successful. Validation data from testing our classification model indicates that we can identify unsuccessful films at a false positive rate of 5.78% and false negative rate of 0.93% Therefore:

Opportunity Cost (FN): 0.93%* (51% * 600) * 300 million = 855,164,034.02

Cost Saving(Total loss - FP): 14.7 billion - (5.78%* (49% * 600) * 50 million) = 13,851,063.829.79

Net Benefit: 13,851,063.829.79 - 855,164,034.02 = 12,995,899,795.77

From the calculation above, the implementation of our model can save Hollywood producers around $13 billion from producing non-successful films. This analysis only looks at the cost saving aspect; however, in reality, our model can also create value by guiding producers into making highly successful films.

Conclusion and Discussion

Producing films is an extremely complicated process that involves numerous decisions, which lead to risks that are hard to predict, given art and expression are highly subjective matters. However at the same time, film production is also a fierce race in this deeply commercialized industry, in which our model will come into play. By quantifying attributes and modeling the decision-making process, we can largely reduce time and cost spent on preliminary designing and positioning, and subsequently mitigate the risk of investment waterloo.

There are certainly drawbacks or possible improvements to our approach and models. Firstly, the film industry has developed well over a century, during which both technology and production procedures have changed frequently, as well as financial metrics like currency exchange rate and inflation. To comprehensively take care of these factors would require additional analysis or methods to make the model adaptable with these developments continuously. Moreover, as discussed in the Models section, we built our clustering model with three supporting attributes of films to optimize visualization and business interpretability. Yet, we find that higher dimensions (for example taking 6 attributes of the movie, including factors such as film duration) will achieve better Dunn index and thus make the clusters more precise and uniquely identifiable. We will further investigate how to interpret and visualize multidimensional clustering so that clients will be able to understand without modeling knowledge. Also, with more dimensions, we will be able to generate more clusters, meaning more detailed categorization of films. Last but not least, we believe that algorithms such as Regression Trees would facilitate our prediction model so that our clients, film producers, can acquire knowledge of marginal effect of each attribute on film outcome, thus providing them the additional benefit of being able to finetune each attribute to see what specific value will help them achieve their expectations.

Appendix

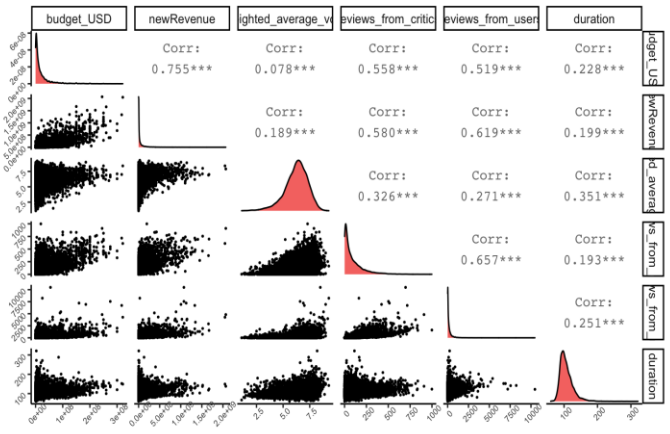


Figure - 1

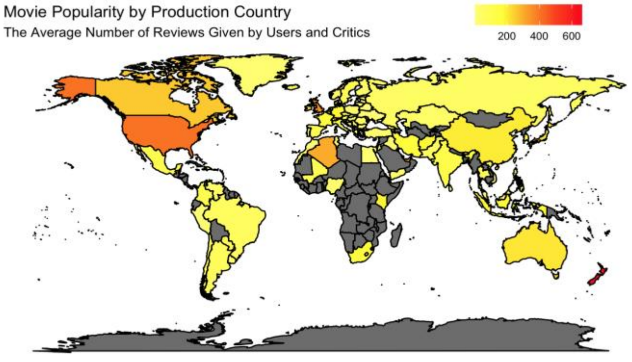

Movie Popularity by Production Country
The Average Number of Reviews Given by Users and Critics
200
400
600

Figure - 2.1

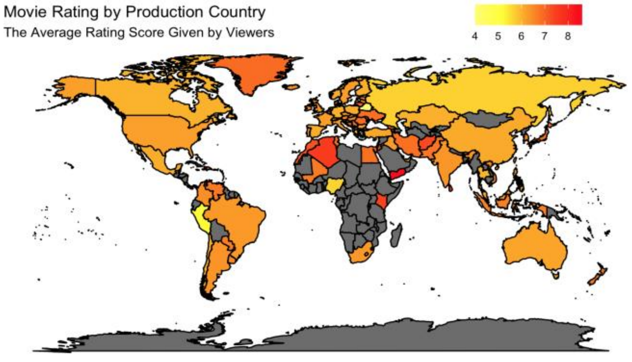

Movie Rating by Production Country
The Average Rating Score Given by Viewers
4
5
6
7
8

Figure - 2.2

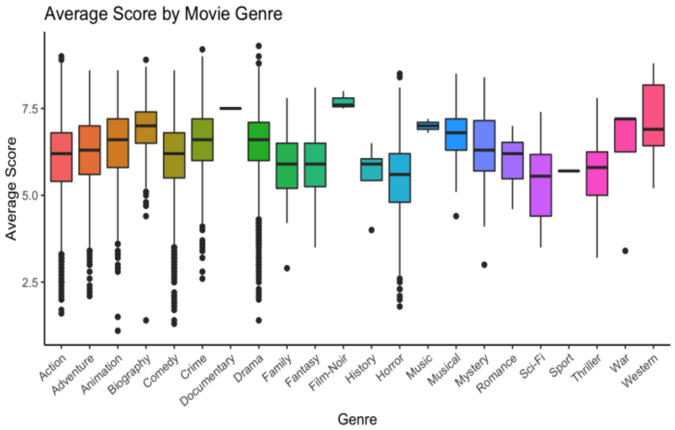


Figure - 3

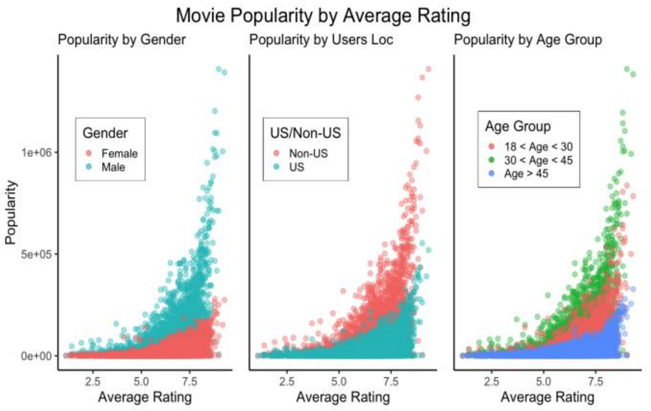


Figure - 4

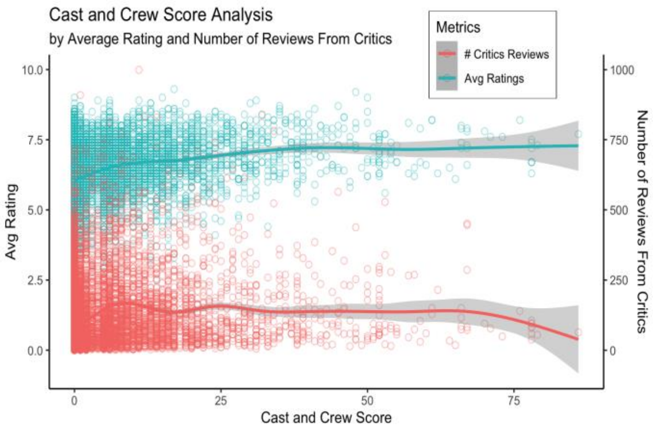


Figure - 5.1

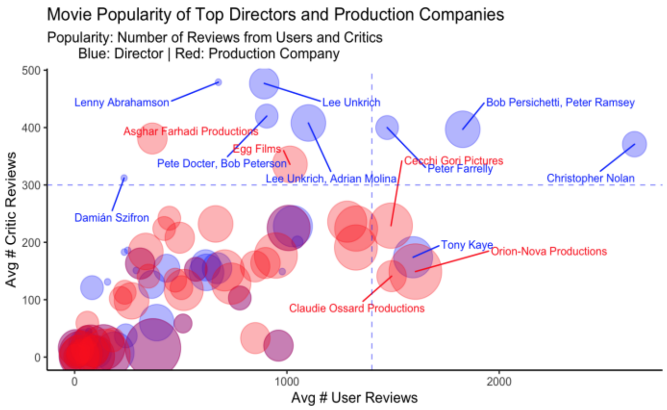


Figure - 5.2

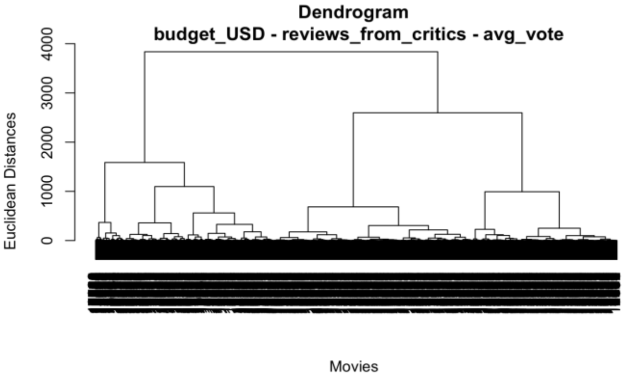
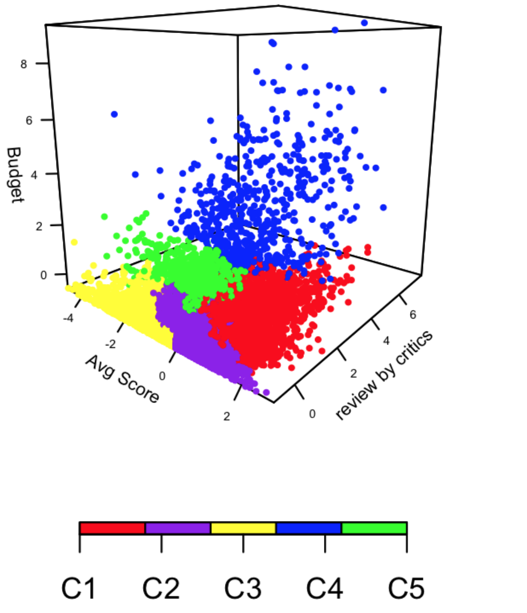


Figure - 6.1

Figure - 6.2
Clusters of Movies
Budget - Average Score - Reviews_from_critics



Cluster	1	2	3	4	5
Size	3562	1053	4800	780	747
Diameter	3.586	4.006	4.262	10.639	7.236
Avg distance	0.871	1.015	1.177	2.743	1.474
Separation	0.023	0.064	0.023	0.159	0.080
Sil-width	0.396	0.473	0.121	0.140	0.374
Avg Sil-width	0.301		Dunn	0.002	

Figure - 6.3

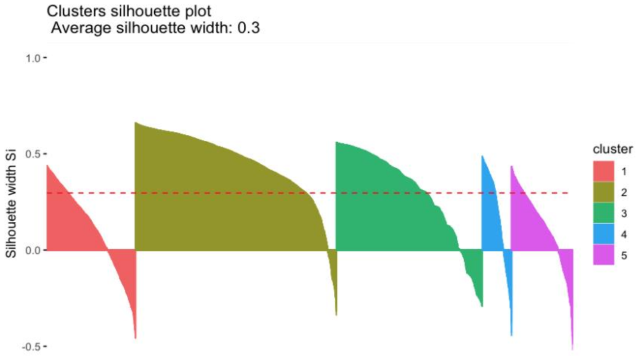
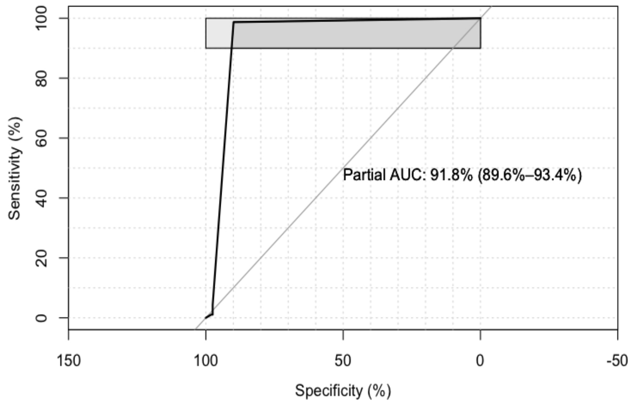


Figure - 6.4



95% CI: 89.85%-93.53% (2000 stratified bootstrap replicates)

Figure – 7

Team Member Roles

	Business Understanding	Data Preparation	EDA	Modeling /Interpretation	Prediction /Deployment
Alice	**X**		**X**	**X**	
Karry		**X**		**X**	**X**
Maria	**X**		**X**		**X**
Sherie	**X**		**X**		**X**
Oliver		**X**	**X**		

Reference

Follows, S. (2020, January 10). Do Hollywood movies make a profit? Retrieved November 14, 2020, from https://stephenfollows.com/hollywood-movies-make-a-profit/

Fontes, R. (2020, February 19). The Oscar Award, 1927 - 2020. Retrieved November 14, 2020, from https://www.kaggle.com/unanimad/the-oscar-award

Leone, S. (2020, September 14). IMDb movies extensive dataset. Retrieved November 14, 2020, from https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

Mueller, A. (2020, August 28). Why Movies Cost So Much To Make. Retrieved November 14, 2020, from https://www.investopedia.com/financial-edge/0611/why-movies-cost-so-much-to-make.aspx

Shakir, J. (2018, March 24). TMDb Movies Dataset. Retrieved November 14, 2020, from https://www.kaggle.com/juzershakir/tmdb-movies-dataset

Webmaster@fxtop.com, L. (n.d.). Historical rates. Retrieved November 14, 2020, from https://fxtop.com/en/historical-exchange-rates.php

Yearly Average Rates & Forex History Data. (n.d.). Retrieved November 14, 2020, from https://www.ofx.com/en-us/forex-news/historical-exchange-rates/yearly-average-rates/