Daten
Data today, Future awaits

Submitted to:

Yossi Nygate

RMIT University Vietnam

702 Nguyen Van Linh Street,

Tan Hung Ward, District 7,

Ho Chi Minh City, Vietnam

MOVIE ANALYTICS

A Comprehensive Exploratory Data Analysis on Collections of Modern Movies

Tran Dam Quan - s3678708

Tran Phi - s3636649

Le Nguyen Thien Phu - s3639855

Nguyen Huynh Anh Phuong - s3695662

Submitted by:

Group 10

Daten Analysis Team

Date: 27/09/2020

Executive Summary

Investing in the movie industry can be a wild, and sometimes unwise adventure. This also includes scouting for the right talent, managing production costs and finding the right distributors. The hardest factors to gauge are the personal whim of the movie audiences and the voice of the critic. More importantly, if a movie did a good performance at the box office, there would be an exciting possibility to make a franchise, however, if it flops, it can sabotage the studios as well as the career of the actors. Furthermore, a good movie without the niche definition of the customer segment can also be considered as inevitable failure. Therefore, the target moviegoers as well as the release date for the box office should be defined specifically in order to maximize the gross profit through means of data-driven methods. A good example is Netflix has been successful in increasing more viewers and enhancing better user (customer) experience, especially, in the series House of Cards [1] with their comprehensive usage of data.

This paper provides a comprehensive data analysis and evaluation of IMDb movie collections from 1980 to 2017, with the aim of formulating precise business plans and strategies to generate high profit movies at their development stage for the client by Daten. Particularly, the process of analyzing movie records from 3 different IMDb datasets is provided to demonstrate valuable business insights by following the CRISP-DM methodology framework. Methods of data pre-processing include dimensionality reduction, feature creation and transformation, and aggregation to clean the dataset after merging as preparation for the data modelling phase. The findings are provided by using data visualization through graphs, charts and scatter plots to display various outcomes of the analysis, and using machine learning models to classify the return of investment (ROI) as profitable or loss, and predicting the profit of the movie as well. Decision Tree, Random Forest, and Linear Regression are the 3 algorithms used in this study to handle the mentioned tasks. The team has successfully trained a model with acceptable accuracies of 73-80%, with only 10 out of 49 selected features of the dataset.

Furthermore, four key features have been revealed to affect the model the most: the number of US votes, number of female viewer's votes, the release date of the movie and finally the budget of a movie. Therefore, a conclusion has been made that the United States movie market, especially female viewers as the main target audience, should be preeminently focused on. In terms of movie attributes, key aspects such as specific words in the movie's title/overview, release date and month, popularity of movie participants and collaboration between these people, etc. have been found and observed to have a positive effect on the generated profit of that movie. As such, this dataset can also be used to build a movie recommendation system in the future.

Table of Contents

Table of Figures

Table of Tables

1 Overview

This paper is organized into the following way:

- **Section 2** describes and presents the project in a business perspective where business objectives/ business success criteria, available resources, assumptions and constraints and project planning is clearly defined.
- **Section 3** provides details of how 5Vs in data analytics are encountered and addressed.
- **Section 4** demonstrates the implementation of data cleaning methods on the merged dataset.
- **Section 5** illustrates the visualization of cleaned dataset through means of infographics.
- **Section 6** applies different modelling algorithms and chooses the most applicable ones by making comparisons.
- **Section 7** explains how the data is analyzed and evaluated based on the defined business objectives.
- **Section 8** specifies recommendations on what can be improved and makes conclusions on the findings.

2 Business Understanding

2.1 Determine Business Objectives

Background

Daten is a new data science and engineering consultancy firm that provides professional experience and great expertise in data analytics and builds high-quality data solutions in search of profitable growth for various enterprises across several domains. Daten carries a wide range of technical competencies, leading tools and end-to-end IT consultant to clients of any scale.

Our company specializes in implementation of advanced Big Data, Machine Learning and Visualization solutions to help companies analyze tremendous volumes of data in order to propose the most optimal solution for business. Additionally, data management services are offered such as data pre-processing, data mining, data amalgamation to further support in obtaining deep insights and providing detailed business consideration for the clients.

At Daten, we dedicate the utmost excellency of service delivered with professionalism, integrity, accountability and the most comprehensive data analysis solution for your business.

Business Objectives

Pyxar is a new feature film production company based in Emeryville, California. The company mainly focused on the production of feature films, animations and artist films for audiences of any age. With quality storytelling and striking visuals, the company has recently gained a massive increase in popularity from the general public and begun collaborating on producing and editing with other associated filmmakers, advancing their performance towards the international market.

With huge attractions from promising markets, it is imperative to inquire into categorizing high-gross movies to invest in. To formulate a precise business plan for auspicious movies at their development stage, it is critical to understand the data sources and the structure of the data itself, hence developing a set of hypotheses of what range of business questions that can be answered predictively for desirable outcomes.

In essence, from a business perspective, the primary objective is to optimize their marketing strategies and promote a movie that could potentially achieve the best profit in return and by extension increase the popularity of the movie. Specifically, we would want to gain insights on how to develop a movie that could seemingly suit people's preferences the most to maximize gained profit.

Business Success Criteria

The success for the business values critically depends on the variety of our findings to the ultimate question. In particular, it is inadequate to give an answer that limits the choices Pyxar can make for their movie (e.g. saying that you should make a particular genre that have these actors/actresses or directors in order to obtain a high stream revenue is not an acceptable solution). A smarter decision is to apply a divide-and-conquer approach to the problem where we divide the complication into smaller instances and propose our solvings correspondingly to improve the validity of the results. By explicitly giving multiple distinct combinations of attributes, there would be more choices for Pyxar to make based on their strategy, hence holding a high degree of confidence and accountability during their development phase.

2.2 Assess Situation

Inventory of Resources

Personnel. Daten's Data Analysis Team consists of 4 personnel proficient in the process of data mining and analyzing to provide critical business insights for the clients based on

Data. Main sources of data are taken from Kaggle - a famous online community for data analysts where it provides a large proportion of public datasets for personal uses:

1. **IMDb Movies Extensive Dataset** [2] is the world's most popular and authoritative source for movie, TV and celebrity content. The dataset provides beneficial information such as genre, directors/actors, user ratings and reviews, release dates and many more aspects. Last updated: 12/2019

2. **The Movies Dataset** [3] contains metadata for 45,000 movies released on or before July 2017 collected from TMDB and MovieLens website. The dataset provides crucial data points including cast, crew, budget, revenue, release dates, countries, TMDB vote counts and vote averages and ratings on a scale of 1-5. Last updated: 2017

3. **The IMDb Dataset** [4] provides information about the principal cast/crew for titles and following information for names of a particular person of a movie. Last updated: 11/2019

Software. Jupyter Notebook and relevant Python's data analysis, modelling and visualization libraries such as Numpy, Matplotlib or SKLearn are dominantly adopted to analyze the mentioned datasets.

Requirements, assumptions and constraints

Preliminary requirements include the schedule of completion where the project is estimated to be accomplished within 3 months from July - October 2020. Resources taken online have to be guaranteed available for personalized use, which Kaggle does not have any technical legislations on their datasets. Consequently, the provided datasets must satisfy the aforementioned business objectives where Pyxar is provided with multiple non-trivial insights to support them in developing their movie.

Initially, the profit of a movie is assumed to be the total revenue that the movie can generate from box office, however, that is not the case. In reality, a movie can have multiple revenue streams such as box office, merchandise and DVD & Blu-rays which all add up to the total revenue. Furthermore, the budget of a movie is also an independent variable that affects how profit is calculated.

During our exploration and examination of how a movie's profit is evaluated, 2 constraints regards to revenue and budget are found and described as following:

1. **Revenue**

   According to [5], the IMDB dataset only collects box office data from BoxOfficeMojo where it receives data from a variety of sources including film studios, distributors and production companies around the world. In other words, the revenue in this project primarily refers to theatrical receipts, which is the amount of ticket sales worldwide (merchandise or DVD streams are not included).

   Although the scope of revenue mentioned in this project is limited to box office data only, it does not stand as a major problem to the final findings since the revenue is still being maximized.

2. **Budget**

   A movie budget can be driven by a variety of costs including production cost, marketing cost, residuals, financing costs, overheads, etc. [6] , this budget data can be different from those listed in other sources as some studios are usually reticent when discussing the budget, especially when that movie performed poorly in cinema. Moreover, the reported budget can greatly vary depending on whether the studios include or exclude additional cost (printing and advertising cost) [5], which poses a problem. Therefore, for the sake of simplicity, the budget mentioned in this project only covers the production cost.

With the mentioned information about the budget and revenue above, the profit of a movie can be calculated as the subtraction between the international box office revenue (already including domestic revenue) and the budget.

$$\boldsymbol{Profit} = \boldsymbol{Revenue}\ (\boldsymbol{Box\ office}) - \boldsymbol{Budget}\ (\boldsymbol{Production})$$

Risks and Contingencies

Aside from the subsidiary constraints that may downgrade the accuracy of our findings, there is not a great deal of immediate risks in this venture. However, it is notable to mention again that time is a censorious factor so the project must be completed within scheduled time.

2.3 Determine Data Mining Goals

Data Mining Goals

While a business goal states objectives in business terminology, a data mining goal states project objective in technical terms. After going through the movie datasets, one can start to notice some correlations or trends between combinations of characteristics of the movie.

In conformity with stated business objectives, the focus of our investigation would revolve around:

"What kind of characteristics and traits of a movie can generate the highest possible profit? Is there a set of combinations of different attributes of a movie that could return higher profit than others?"

By using this pertinent technical question as our base, multiple prime sub-questions can be derived, and various trends can be explored to dig into those questions even further. In particular,

1. Does a specific release date have an impact on the generated revenue of a movie? Our basic understanding is that the popularity of the movie can be affected by seasons, but we would want to know which specific month and day that has the best possibility of generating the highest profit.

2. While genre is a critical independent variable to the success of a movie, does the popularity of specific contributors have a positive effect on the profit of the movie? It is understandable that famous actors/actresses have a significant impact on the outcome of the movie, but what about directors or writers? Do they create similar impacts and if so, to what extent? We would want to confirm if the common belief that these people can positively affect the generated profit is true or not.

3. For further exploratory analysis, do collaborations between mentioned actors/actresses, directors and writers correlate with the movie's revenue? If that is the case, which collaboration is the most beneficial of all? Understanding this trend can effectively help the client in considering which people to choose for their movie.

Data Mining Success Criteria

Success must also be defined in technical terms to keep the data mining efforts on track. In specific, certain benchmarks have to be reached in order for the data mining results to become satisfactory. This includes resolving above questions through graph visualizations of the dataset and displaying a set of features that when are used for modelling phase, reaches an acceptable precision of approximately 75-80%, with no more than 1-2% lower difference.

2.4 Produce Project Plans

TASKS	SUB-TASKS	TIMELINE												
		Wk 1 29/6 - 6/7	Wk 2 6/7 - 13/7	Wk 3 13/7 - 20/7	Wk 4 20/7 - 27/7	Wk 5 27/7 - 3/8	Wk 6 3/8 - 10/8	Wk 7 10/8 - 17/8	Wk 8 17/8 - 24/8	Wk 9 24/8 - 31/8	Wk 10 31/8 - 7/9	Wk 11 7/9 14/9	Wk 12 14/9 - 21/12	Wk 13 21/12 - 28/12
BUSINESS UNDERSTANDING	Define business objectives													
	Define business success criteria													
	Find data resources and evaluate their constraints													
	Define data mining goals													
	Define data mining success criteria													
DATA UNDERSTANDING	Determine total number of records													
	Determine types and number of dirty data													
	Provide summary statistics for each attributes													
DATA PREPARATION	Clean dataset													
DATA VISUALIZATION	Visualise data and provide insights (graphs, charts, etc.)													
DATA MODELLING	Choose suitable algorithms													
	Build models													
DATA EVALUATION	Evaluate insights with initial business objectives													
	Document the project													

Figure 1. Gantt Chart

The project plan comprises of 6 stages at intervals of approximately 3 months:

- **Business Understanding:** Business objectives and success criteria are defined with clients during the beginning week where both parties finalized on investigating which attributes of a movie can help generate the most profit and based on these agreements, we begin to find suitable data resources and evaluate their constraints in week 2. Finally, to be more detailed and technical, we construct data mining goals and success criteria (based on the stated business objectives) to deliver relevant questions that help us in the progress of producing valuable insights in week 3.

- **Data Understanding:** After defining goals and constructing proper development plans, we begin to explore our data in week 4. In detail, we determine the total number of records of the merged dataset and verify the data quality whether there are any missing values, illegal values, illogical values or outliers exist. At the beginning of week 5, we begin to provide summary statistics for each attribute (min, max, etc.).
- **Data Preparation:** From mid of week 5 to end of week 6, we begin the cleaning dataset phase based on the data quality verification in the previous week. This process mainly includes using dimensionality reduction methods to delete any unnecessary attributes or dirty values.
- **Data Visualization:** Simultaneously, during week 6, we start plotting cleaned data on graphs and charts to provide further insights on which attributes or a combination of attributes that are likely to generate higher profit.
- **Data Modelling:** In week 7 , we start testing out different modelling algorithms to see which produce the best accuracy or give some useful insights during the development time and for the next 2 weeks, up until week 9, we begin to build the models.
- **Data Evaluation:** From week 10-12, we begin evaluating our findings with comparisons to the initial business objectives to see whether the results are satisfactory. If not, then data tuning progress is conducted (if any) throughout these 3 weeks. Lastly, in the middle of week 11, we start to document our findings (including analysis and visualization) and inform the clients at the end of week 13.

3 Data Understanding

3.1 Data Description

There are numerous records and attributes to process when dealing with movie metadata. In particular, before the data preparation phase, the four V's of Big Data (excluding **value** since it is already defined in business objectives) is applied to provide detailed information about the datasets as follow:

Volume

	rows	columns
imdb_df	81273	12
meta_data	45466	3
merge_df	34521	14
dataframe	9377	18
year_df	7098	20
new_df	7098	19
imdb_cast	23811321	4
cast	47888	8
data	6896	51
imdb_rating	81273	33
final_df	6896	79

Initially, 4 tables are utilized from the IMDb Movie datasets where **imdb_df**, **meta_data**, **imdb_cast**, **imdb_rating** are merged into the **final_df** for later use. The number of records (rows) and attributes (columns) after merging is defined in *Figure 2.*

The final dataset available for the next phase contains a total of 6896 records and 79 different attributes.

Figure 2. Volume of movie datasets

Velocity

Unlike extremely continuous data such as SMS messages, Facebook status updates, or credit card swipes, the transmission speed of movie's metadata is considered to be very low as movie or TV shows require development time of at least weeks or months. Therefore, it takes a considerable amount of time, compared to data that can be produced every second or minute, to generate metadata for a movie, hence the project's datasets have a low velocity.

Variety

These data resources contain only structured data type, whether they are dates and times, the budget number of a movie or the total votes/ratings from viewers. In particular, some major attributes are selected to demonstrate the variety of the final dataset.

Attribute	Data Type	Description	Statistics
Budget	Nominal Discrete	The budget (production cost) of a movie	**Min:** 20 **Max:** 300M **Mean:** 28.3M **Stdev:** 36.4M **Frequency:** 288 **Mode:** 10M
Profit	Nominal Discrete	The profit from a movie	**Min:** -111M **Max:** 2.6B **Mean:** 44.8M **Stdev:** 127.9M

			Frequency: 4 **Mode:** -1.34M
Revenue	Nominal Discrete	The revenue from a movie	**Min:** 17 **Max:** 2.79B **Mean:**73.1M **Stdev:** 152.63M **Frequency:** 4 **Mode:** 38.79M
Us_voters_votes	Nominal Discrete	Number of votes for a movie in US region	**Min:** 2 **Max:** 341457 **Mean:** 14009.05 **Stdev:** 23735.55 **Frequency:** 5 **Mode:** 138
non_us_voters_votes	Nominal Discrete	Number of votes for a movie outside US region	**Min:** 23 **Max:** 862970 **Mean:** 37049 **Stdev:** 66603 **Frequency:** 6 **Mode:** 614
Duration (minutes)	Nominal Discrete	The duration of a movie	**Min:** 63 **Max:** 271 **Mean:** 106.36 **Stdev:** 17.97 **Frequency:** 223 **Mode:** 90
Popularity	Nominal Discrete	The popularity point of a movie	**Min:** 0.000001 **Max:** 547.49 **Mean:** 8.56 **Stdev:** 12.54 **Frequency:** 4 **Mode:** 6.88
Year	Nominal Discrete	The year of a specific movie is released	**Min:** 1980 **Max:** 2017 **Mean:** 2003 **Stdev:** 9.49 **Frequency:** 309 **Mode:** 2008
Day_of_week	Nominal	The weekday in which a	**Total:** 6900

	Discrete	movie was released.	**Frequency:** 5432 **Mode:** Friday

Table 1. Volume of movie datasets

Veracity

The datasets used in this study are provided by IMDb. IMDb is a movie-related database owned by Amazon, which was established in 1990, with a large range of movie information from 1874 to the latest 2018. Due to the lack of documentation, the movies' data before 1980 could be lacking information or inaccuracy, however, it doesn't really have a strong impact towards the final findings. Moreover, due to marketing strategy, most of the movies that publish their data would be leaning towards profitable data. By combining with other datasets, the main dataset which was used to study maintains a balanced number of profitable and unsuccessful movies. The dataset is widely accepted and used by many other data analysis individuals and for many researches. With this high trustworthy dataset, the study was confidently developed.

3.3 Verify Data Quality

Two types of erroneous data are taken into consideration during the process of identifying and classifying all dirty data in the merged dataset where numbers of values (individual cells) are computed to show how disorganized the data is.

Missing Type

Data that is declared as 'NaN' is considered to be missing-type data. This can be caused by computational malfunctions or there are circumstances that it cannot be recorded. An example of missing data type is that only 0.09% (6/81273) of total movie records don't specify the overview (summary) of the movie.

The program detects a total of 12368 NaN values for the final dataset.

Illegal Type

Data is considered to be illegal when its data type is not the same as the type it is supposed to be (e.g. year should be numeric type but not string) or contains illegal characters. For example, the expected data type of a revenue or budget of a movie should be a float type where the individual cell displays a number. However, initially, both revenue and budget columns have different currency signs before the number which is a string type. Therefore, they need to be converted to float type in order to make calculations.

The program detects a total of 16919 illegal type values for the final dataset.

Type	Number of values
Missing Type	12368
Illegal Type	16919

Table 2. Illegal table

4 Data Preparation

4.1 Methods

Before merging the data files, some techniques are utilized to obtain a reliable dataset.

- **Dimensionality reduction:** By observing the final dataset, some of the attributes are considered to be discarded and selected. For example: “original_title”, “producer”, and “language” are not related in our case, so they were removed. Moreover, in the cast dataset, any category rows that are not actor, actress, director or writer are discarded such as camera man, and crew.
- **Transformation:** By converting the timestamp in date_published attributes to separate attributes like day_of_week, month and year. Moreover, the genre attribute can be separated to classify as 0 or 1 in per movie to count the total number of each genre.
- **Feature creation:** Create attributes such as profit and ROI rate
- **Aggregation:** To find the profit or revenue or popularity of participant attributes per movie, this method is used to group all the participants’ id and use the respective calculations for each attribute. For example: to find profit of an actor, we use cast id to group and calculate the mean of this participant’s profit.

4.2 Cleaning Process

Step 1: Remove movies that are before 1980 and after 2017

According to [6], the second golden age of Hollywood is after the 1970s. However, the IMDB dataset stated that the longer the information was published, the less reliable it might become [2]. Therefore, the information for movies over 15 years ago are quite sketchy. Besides, the second dataset was scrapped before July 2017. Hence, dropping movies that are from 2017 should also be considered. Here is the graph that counts the number of movies released from 1920 to 2019.

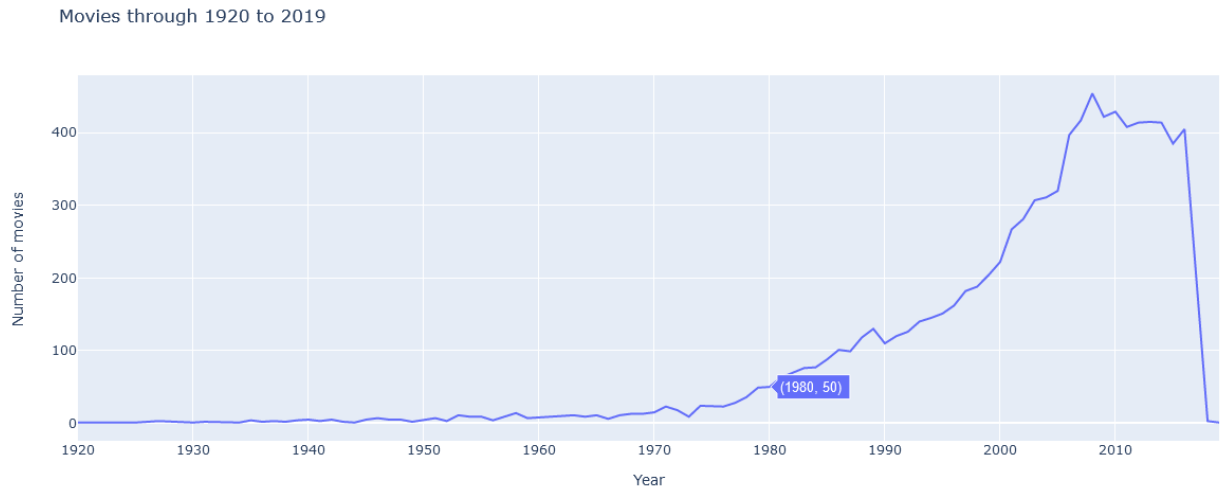


Figure 3. Number of movies from 1920-2017

Thus, using pandas DataFrame functions and logical conditioners, 433 records before 1980 and after 2017 were removed.

Step 2: Merge with IMDB rating

33 columns containing votes and rating related data are merged to the main dataset. After merging, the final dataset contains 6896 rows with 79 columns.

Step 3: Handling NaN data

Attributes (Column)	Total NaN Values	%
crew_pop	144	2.1
reviews_from_users	27	0.4
reviews_from_critics	21	0.3
overview	6	0.1
females_18age_votes	5	0.1
females_18age_avg_vote	5	0.1

females_45age_votes	3	0.0
females_45age_avg_vote	3	0.0
males_18age_avg_vote	1	0.0
males_18age_votes	1	0.0

Table 3. NaN values per column

For vote-related columns, the missing data is filled in with the mean vote of each row. Reviews from users and reviews from critics are used to create a new column which shows whether the user or the critic prefer that movie. For these two columns, the missing data is considered a lower number than the other, as an example, if the reviews for users is missing, then that movie is preferred by critics and vice versa.

Overview is then converted into a column that holds the counting of total words used in the overview. If the value is missing, the counting number is set to 0.

Crew popularity is calculated on the popularity of actors, actresses, and directors of the movies. These values cannot be replaced by any value that will not affect the result. So, the value is assumed to be zero and filled in with that value.

Step 4: Evaluate dataset after cleaning

Before the IMDB rating merge, the total rows and columns were 6896 and 51 respectively. However, after merging with the rating dataset, and dropping the unnecessary columns above, the new data frame is 6896 rows and 78 columns. Moreover, a new column is created to categorize between number of users' votes and critic's votes. In total, 5 duplicated and unnecessary columns were dropped, 1 column was created from user and critics vote, 2 columns used to create the mentioned feature were dropped, and 33 columns were added from the rating dataset. Not any row was dropped during the process

Step 5: Categorize movies into profitable or loss as a preparation for model training

To prepare for training a classification model, the target, which is ROI, is categorized into profitable and loss. If the ROI is positive, it is considered to be profitable and assign the value of 1. If the ROI is negative, it is considered to be loss and assign the value of 0.

5 Data Visualization and Analysis

This section provides a better understanding of the cleansed dataset by representing them in a graphical format, hence illustrating interesting findings on key characteristics and traits of a movie that can help generate better profit. Moreover, in this part, the attributes are evaluated to find the relationship between the profit in order to select the correct features for the machine learning data analysis.

5.1 Title and Overview Word Count

Figure 4. Title and Overview Word Cloud

The figure above illustrates the most occurrences of each word in titles and overview of the dataset using word cloud library. By observing the figures, it can be seen that words that have the most occurrences in the movie's title are "life", "find", "world", "family", "friend", etc. and words that have the most occurrences in the overview of a movie are "love", "night", "man", "girl", "last", "day", "life" and "movie". This information will not only provide some preliminary movie ideas for writers and producers but can also be used to build efficient marketing campaigns with the support of Search Optimization Engine (SEO) strategy.

5.2 Movie Release Date

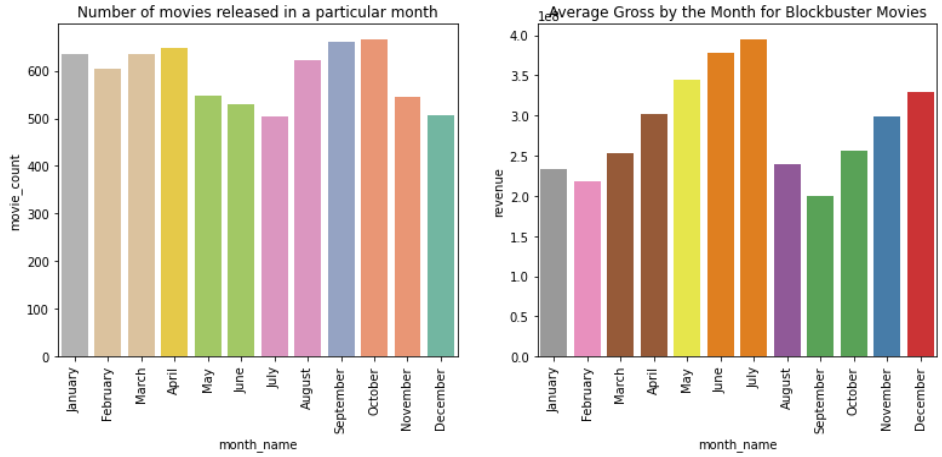


Figure 5. Number of movies released and average gross by month

Figure 5. shows the number of movie releases and average gross for blockbuster movies for each month. At first glance, May - July are months with the least amount of movie releases, however, when comparing with the average revenue generated per month, these 3 months are considered to be the best option for blockbuster movie release. A valid reason is that a large majority, especially kids and teenagers, are on vacation or have more free time than other seasons. Therefore, it is understandable that going to the cinema is more common during these months since watching movies is also a popular activity to do. On the contrary, during winter break, people tend to spend more time with their family and stay indoors due to the cold weather outside which may be one of the prime reasons the generated revenue in the last few months are lower than those before. With this being said, it is highly recommended for the client to **release a movie during the summertime, specifically, from May to July to achieve the highest generated profit possible.**

In addition to the release month, the released day of the week is also mentioned to demonstrate which weekdays or weekends should the clients release the movie to attract the highest number of viewers.

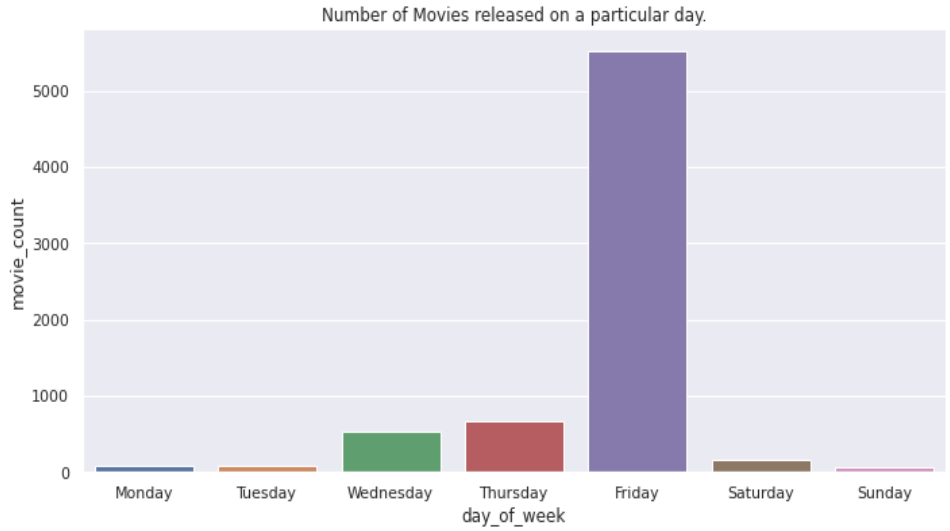


Figure 6. Number of movies released by days of week

It is evidently shown that more movies are released on Friday than remaining days of the week and this can be due to many reasons. First of all, people usually start hanging out with their friends and family on Friday night when they have finished their long working week, whereas people are more likely to stay at home on weekends. Another example is that according to CGV policy in Vietnam, ticket sales on weekends as well as on special occasions are more expensive than on weekdays [7]. Secondly, a business paycheck is likely to be released on Friday which people have more money to spend on. Lastly, Friday is also a good test run to decide whether to increase more shows or advertisements based on the response on Friday before the weekend begins. Therefore, it is highly recommended for our client to **release the movie on Friday.**

5.3 Popularity

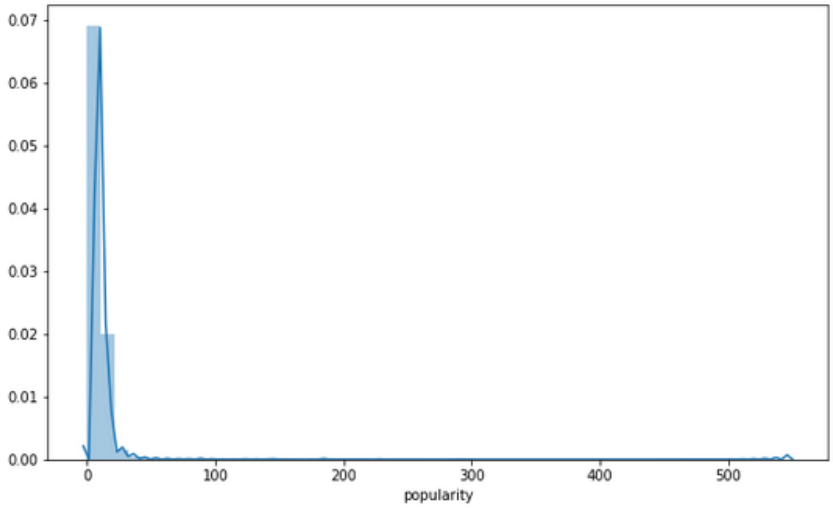


Figure 7. Popularity of movies

This section examines the correlation between the popularity of a movie with the profit it generates. From the graph above, we can see that the popularity data is absolutely skewed and not normally distributed. The following figures show the highest popularity movies as well as the correlation between the profit and movie popularity. From the second graph, unexpectedly, the correlation between the profit and the movie popularity is not strong. This is because the TMDB was founded in 2008, therefore, it can only show the connection between the recent years. Take the most popular movies based on the users' search in the table below as an example, it only shows movies from 2008 to 2017 as it was scrapped in 2017. Therefore, it cannot show the clear connection between the profit and the popularity.

	title	popularity	year
0	Minions	547.488298	2015
1	Wonder Woman	294.337037	2017
2	Beauty and the Beast	287.253654	2017
3	Baby Driver	228.032744	2017
4	Big Hero 6	213.849907	2015
5	Deadpool	187.860492	2016
6	Guardians of the Galaxy Vol. 2	185.330992	2017
7	Avatar	185.070892	2009
8	John Wick	183.870374	2015
9	Gone Girl	154.801009	2014

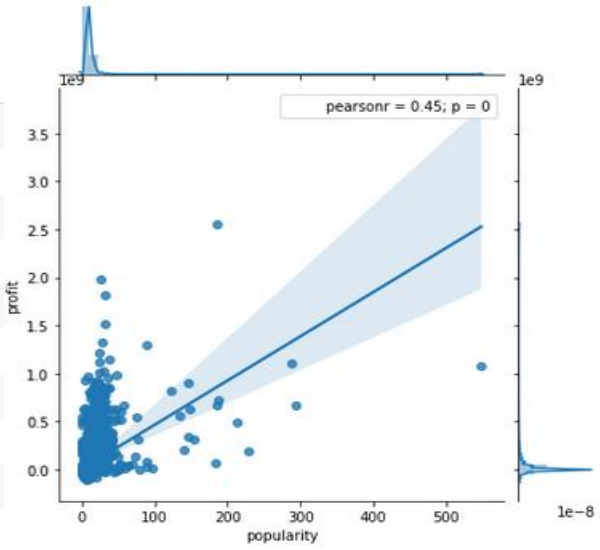


Figure 8. Profit versus popularity

5.4 Duration

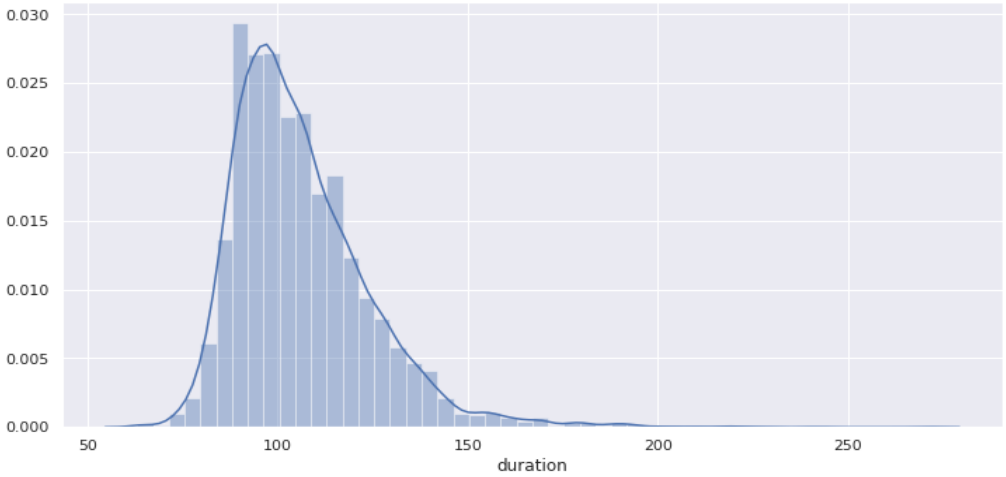
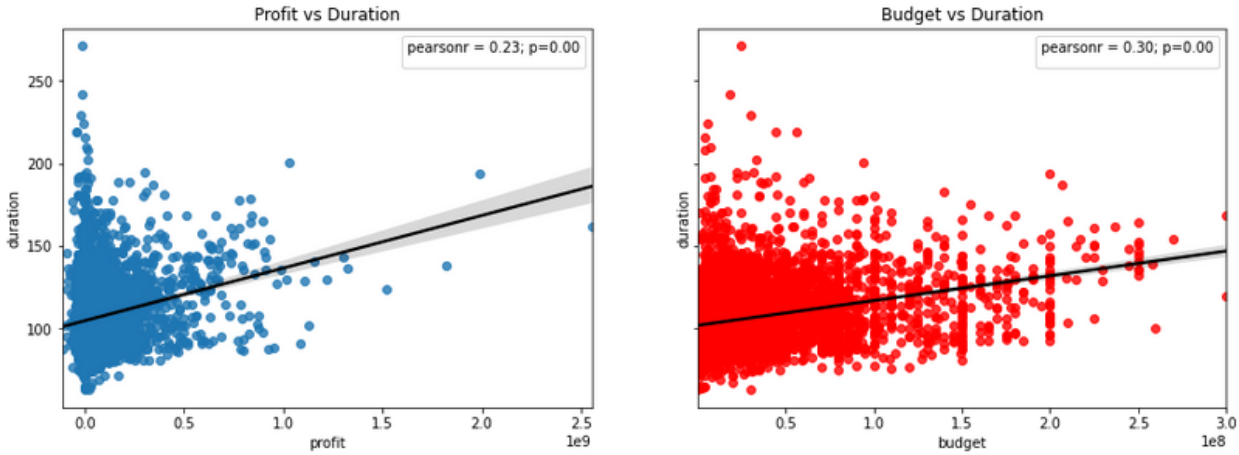


Figure 9. Duration of movies

The duration of the movie has the maximum of 271 minutes, where the duration is more distributed in the range of 60 to 190 minutes, thus it is common for movies to be in this acceptable range.

Figure 10. Profit/Budget versus duration



The above figures are displayed to show the correlation between the budget vs duration and profit vs duration. It is expected that the longer the duration of the movie, the bigger the budget, however, the correlation between these two attributes is only 0.3 (Pearson), which is not positively strong. Therefore, the budget attribute also affects the correlation between the profit vs duration. By linking with reality, the duration tends to be relying on the genre of the movie more than the budget. For example, although a movie art house like The pianist and Sci-Fi movie like Sign were produced in the same year (2002), the sooner duration is 150 minutes and cost 35 million USD to make, while the latter duration is only 107 minutes, however, it cost 72 million USD to make, hence confirming the point that **a longer movie does not mean needing more budget to produce.**

Furthermore, it is observable that the correlation between duration of a movie and the generated profit is even lower, with only 0.23 using Pearson correlation which implies that **duration does not significantly affect the profit of the movie either**.

In conclusion, duration is not a great independent variable that can be used to predict the generated profit of a movie.

5.5 Budget

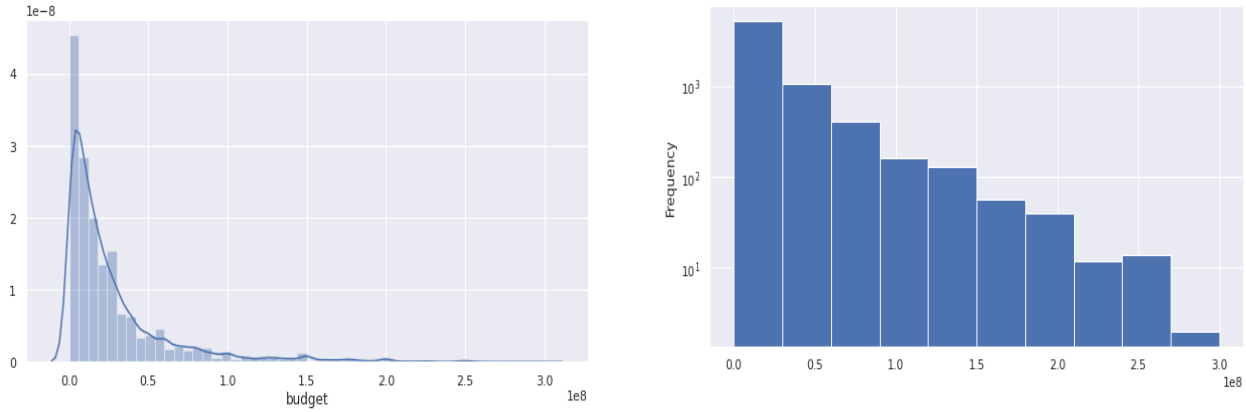


Figure 11. Budget and budget frequencies

As expected, the movie budget is not normally distributed as well, since the budget can be dependent on many factors. The following tables and graphs are displayed below to illustrate the movie with highest budget and the correlation between budget vs revenue respectively.

	title	year	budget	revenue	profit	ROI
0	Pirates of the Caribbean: At World's End	2007	300000000	960996492	660996492	220.332164
1	Justice League	2017	300000000	657924295	357924295	119.308098
2	Superman Returns	2006	270000000	391081192	121081192	44.844886
3	Tangled	2011	260000000	591794936	331794936	127.613437
4	Spider-Man 3	2007	258000000	890871626	632871626	245.299080
5	John Carter	2012	250000000	284139100	34139100	13.655640
6	The Fate of the Furious	2017	250000000	1236005118	986005118	394.402047
7	Avengers: Age of Ultron	2015	250000000	1402805868	1152805868	461.122347
8	The Dark Knight Rises	2012	250000000	1081041287	831041287	332.416515
9	Harry Potter and the Half-Blood Prince	2009	250000000	934326396	684326396	273.730558

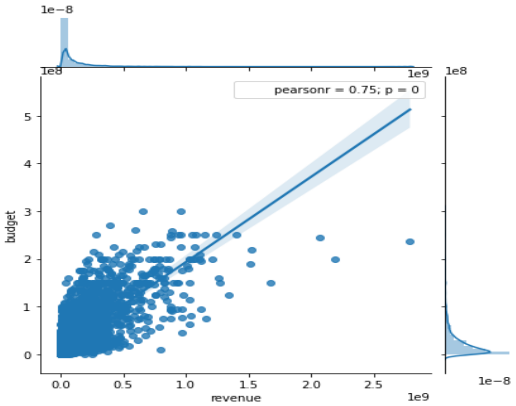


Figure 12. Budget versus revenue

The graph shows an extremely strong correlation (0.75) between revenue vs budget using Pearson correlation. Therefore, the budget is chosen as an attribute to classify ROI rate and predict revenue later on.

5.6 Actor, actress, directors, writer's vs revenue and profit

The following graphs display the average profit of movies that each actor, actress, director and writer participate in correspondingly from left to right. Based on graphed information, the client can know who they should prioritize over others to invite based on their roles.

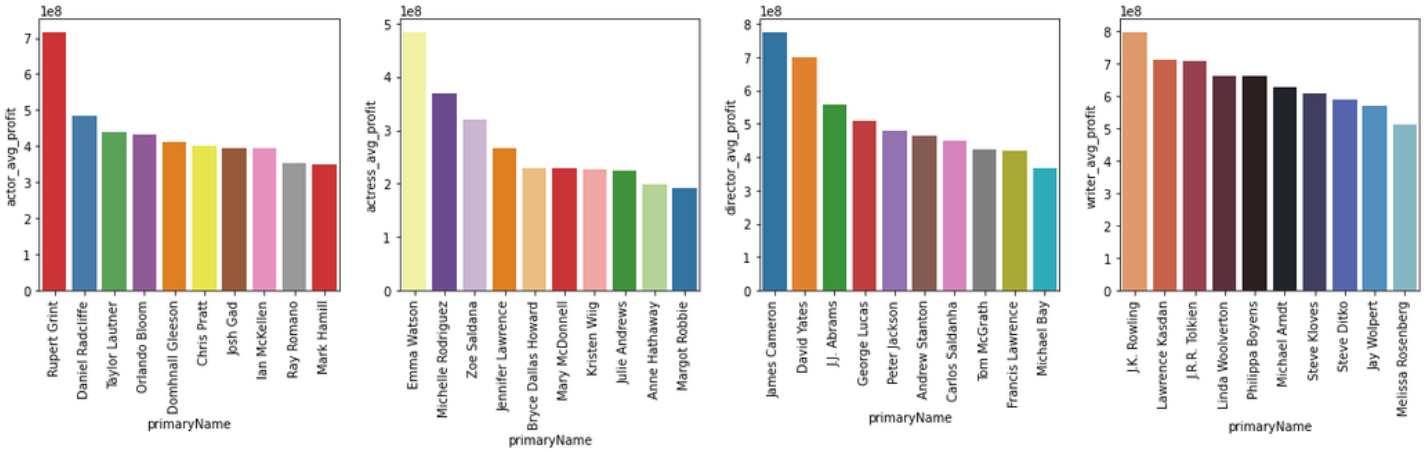


Figure 13. Cast members with highest gross

To illustrate, the actor Rupert Grint in the left-most graph participates in many movies which the average profit that those movies generated is approximately 700 million USD. This means that, for many reasons, whatever the movie Rupert casts in, they are anticipated to generate higher profit. Therefore, Rupert is likely to be the top candidate for being the leading male character in the movie, or Daniel Radcliffe can be auditioned since he is only behind Rupert. This applies to other roles as well, where the client can choose whoever they desire based on their participation in highest grossing movies.

Additionally, as a side note, we also take into consideration in investigating whether inviting multiple famous actors/actresses/directors/writers can have a better impact on the movie's profit, rather than one person only. The following graphs depict 3 types of correlation - popularity of individual participant versus profit of the movie, popularity of total participants versus profit of the movie, and popularity of total participants versus budget of the movie (participants are actor, actress, director, and writer).

Actor

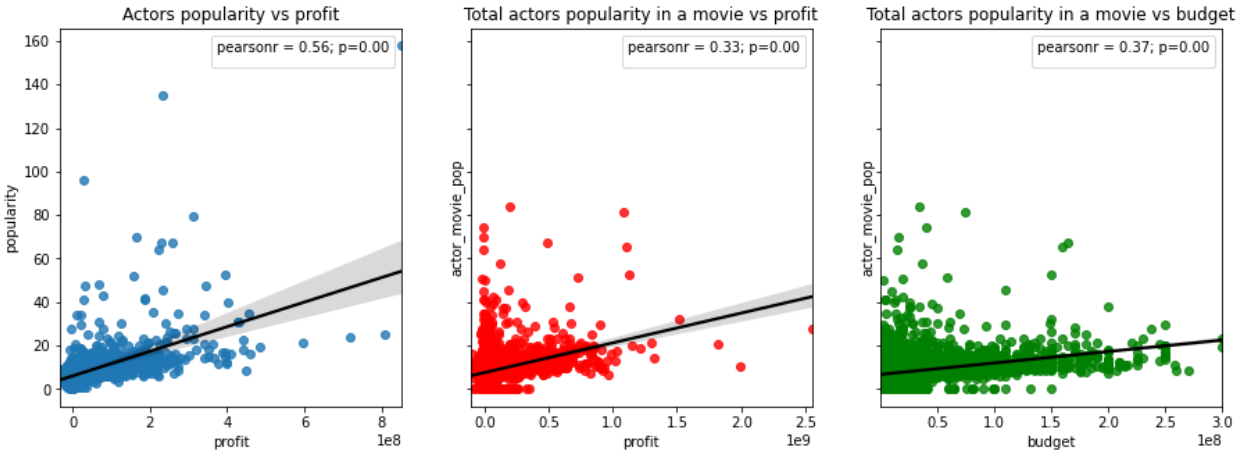


Figure 14. Actor popularity versus profit/budget

Actress

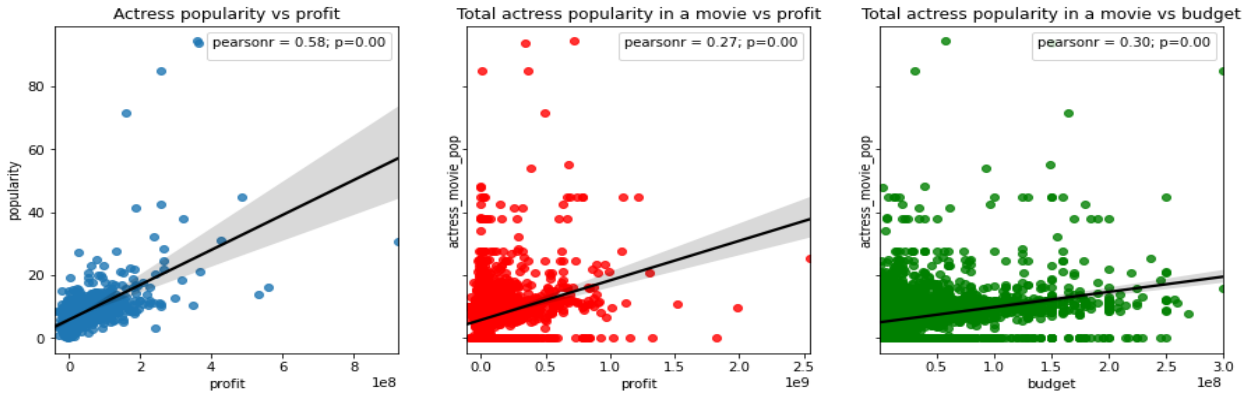


Figure 15. Actress popularity versus profit/budget

Director

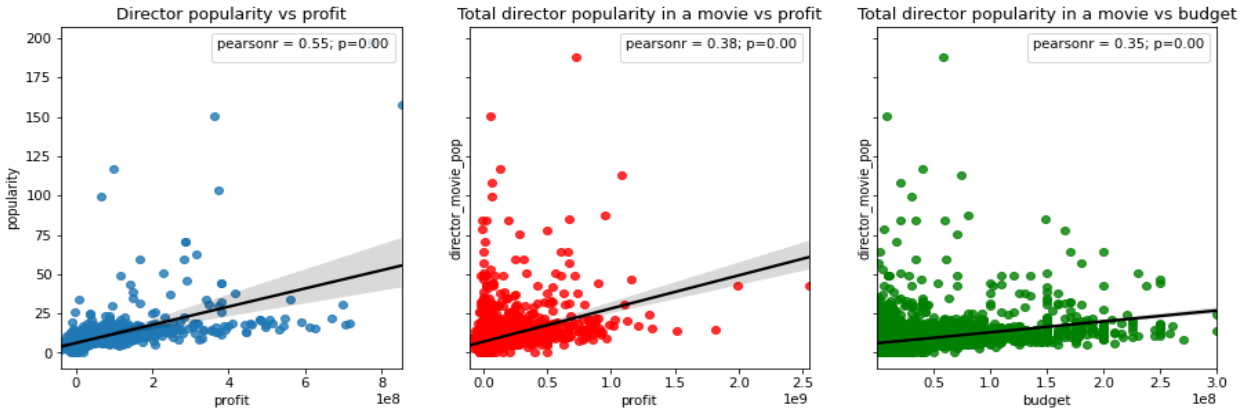


Figure 16. Director popularity versus profit/budget

Writer

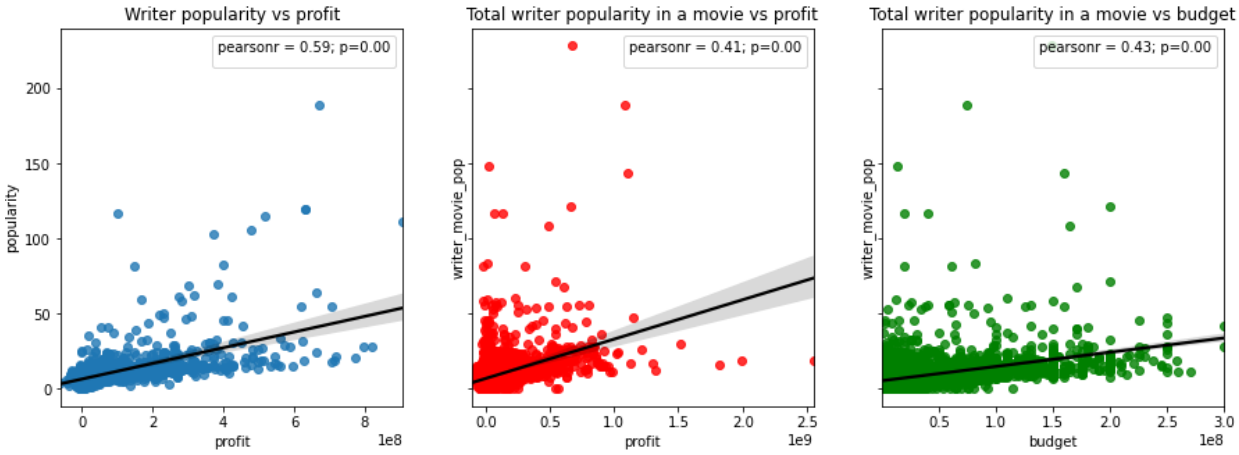


Figure 17. Writer popularity versus profit/budget

Looking at the first graph where it portrays the popularity of individual participant versus the profit of the movie, the Pearson correlation for each participant is approximately from 0.55 to 0.59, which can be considered to be positively strongly correlated, meaning that the more popular the participant is, the more profit that movie can generate. However, looking at the second graph, it can be seen that the Pearson correlation between all participants versus the profit of the movie is relatively less than the previous one. This implies that, for example, inviting Rubert Grint and Daniel Redcliffe and other actors does not necessarily increase the percentage of generating higher profit. In fact, it is evidently concluded that inviting 2 or more famous actors for the same movie can actually lessen the generated profit than inviting only 1 for each role. This condition applies to the budget of the movie as well - more famous actors (and other roles) does not mean that more budget is needed for that movie. Notably, the accuracy of these findings may not be high since the popularity attribute is relatively skewed as TMDB was founded in 2008 so their data volume is small compared to our dataset (from 1980-2017).

5.7 Casts (Actor and Actress)

These following graphs show the correlation between the number of actors or actresses in a movie. As expected, it does not show a strong relationship towards profit as well as popularity because clearly the audience does not care about how many actors or actresses in a movie to watch. However, it is illuminated that actors have a higher positive correlation (roughly 0.1) than actresses, which indirectly implies that male casts have higher influences to the profit and popularity of the movie than actresses do, but is considered to be unknown as the correlation is too weak to justify.

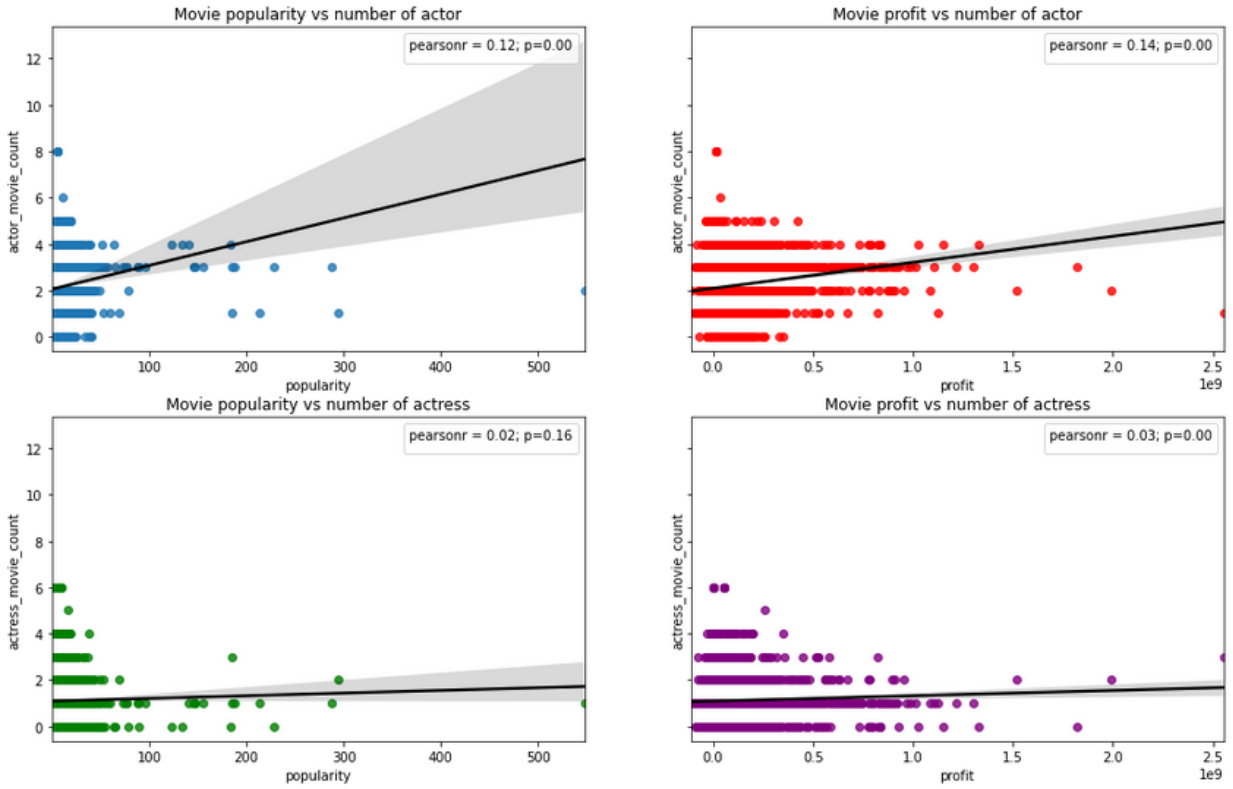


Figure 18. Number of actors/actresses versus Movie popularity/profit.

Furthermore, the below graph shows the maximum number of actors and actresses during 1980 to 2017. By observing the graphs, the number of actors in the same year is mostly higher than that of actresses, which somewhat explains why actors have a better correlation than actresses mentioned above.

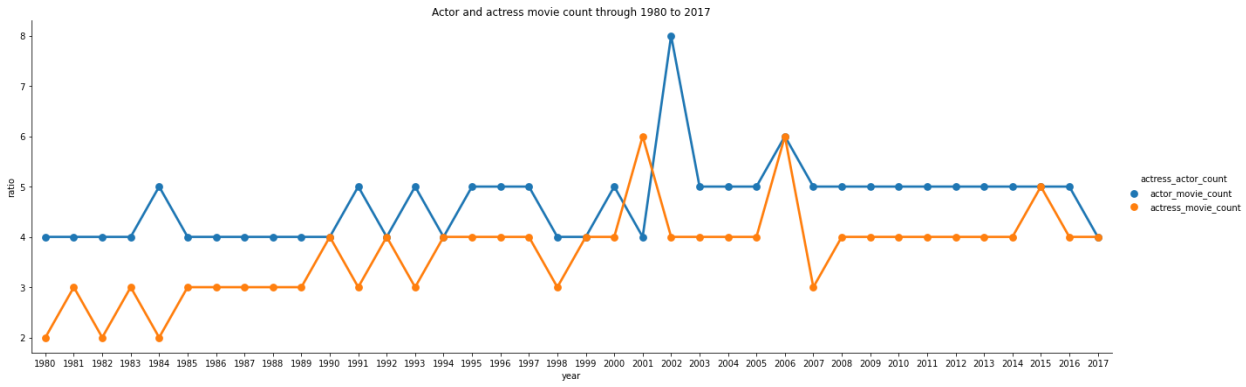


Figure 19. Number of actors/actresses participates in movie from 1980-2017

5.8 Collaboration

The following tables exemplifies the collaborations between participants with over 5 movie collaborations with each other. This observation can potentially help producers and casting directors decide which combination of people they decide to work with in order to generate good profit.

Actor vs Actress/Director/Writer

	actor	actress	count	profit
0	George Clooney	Julia Roberts	7	9.954275e+07
1	Daniel Radcliffe	Emma Watson	6	8.136241e+08
2	Johnny Depp	Helena Bonham Carter	6	2.591897e+08
3	Rupert Grint	Emma Watson	6	8.136241e+08
4	Eric Stuart	Rachael Lillis	5	6.865795e+07
5	Eric Stuart	Veronica Taylor	5	6.865795e+07
6	Robert Pattinson	Kristen Stewart	5	5.920635e+08
7	Sam Rockwell	Drew Barrymore	5	1.499766e+06

	actor	director	count	profit
0	Adam Sandler	Dennis Dugan	8	1.256039e+08
1	Johnny Depp	Tim Burton	8	1.982080e+08
2	Mel Gibson	Richard Donner	6	1.512086e+08
3	Ethan Hawke	Richard Linklater	6	8.158182e+06
4	George Clooney	Steven Soderbergh	6	1.385729e+08
5	Ian McKellen	Peter Jackson	6	8.193179e+08
6	Antonio Banderas	Robert Rodriguez	6	7.369857e+07
7	Tom Hanks	Ron Howard	5	2.958502e+08
8	Russell Crowe	Ridley Scott	5	1.397367e+08
9	Robert De Niro	Martin Scorsese	5	4.424447e+07

	actor	writer	count	profit
0	Adam Sandler	Tim Herlihy	10	1.065890e+08
1	Daniel Radcliffe	J.K. Rowling	7	8.192618e+08
2	Rupert Grint	J.K. Rowling	7	8.192618e+08
3	Ian McKellen	Philippa Boyens	6	8.193179e+08
4	Vin Diesel	Gary Scott Thompson	6	6.475866e+08
5	Rupert Grint	Steve Kloves	6	8.243753e+08
6	Robert Englund	Wes Craven	6	3.810193e+07
7	Paul Walker	Gary Scott Thompson	6	5.099775e+08
8	Leonard Nimoy	Gene Roddenberry	6	8.981277e+07
9	Johnny Depp	Terry Rossio	6	5.704197e+08

Figure 20. Collaborations between actors and actresses/directors/writers

For example, if the client has chosen George Clooney as the main actor then they would want to choose Julia Roberts as the main actress to participate together since they have 7 collaborations with each other and the movies they participated in have the highest profit. Similarly, if Dennis Dugan is chosen to be the film director then they would want to hire Adam Sandler as their main actor in order to generate higher profit. However, this collaboration would depend on the genre of the movie as well. As for the combination of George Clooney and Julia Roberts, 4 of their collaboration movies genres are mainly Comedy, Crime and Drama [8]. This also applied for Adam Sandler and Dennis Dugan collaboration, where all of their collaboration's movie genres are Comedy, Family and Drama [9]. Overall, the genre of the movie has to be decided first before choosing the suitable people to collaborate with.

The following tables between remaining participants apply the same mechanism as explained above:

Actress vs Director/Writer

	actress	director	count	profit
0	Helena Bonham Carter	Tim Burton	6	2.813155e+08
1	Mia Farrow	Woody Allen	6	9.148305e+05
2	Veronica Taylor	Kunihiko Yuyama	6	7.453984e+07
3	Veronica Taylor	Michael Haigney	6	7.453984e+07
4	Catherine Keener	Nicole Holofcener	5	7.012507e+06
5	Jessica Alba	Robert Rodriguez	5	3.606144e+07
6	Milla Jovovich	Paul W.S. Anderson	5	1.629779e+08
7	Rachael Lillis	Kunihiko Yuyama	5	6.865795e+07
8	Rachael Lillis	Michael Haigney	5	6.865795e+07
9	Rica Matsumoto	Kunihiko Yuyama	5	6.271887e+07

	actress	writer	count	profit
0	Emma Watson	J.K. Rowling	6	8.136241e+08
1	Veronica Taylor	Norman J. Grossfeld	6	7.453984e+07
2	Veronica Taylor	Takeshi Shudo	6	7.453984e+07
3	Emma Watson	Steve Kloves	5	8.186329e+08
4	Jamie Lee Curtis	Debra Hill	5	3.057284e+07
5	Kristen Stewart	Melissa Rosenberg	5	5.920635e+08
6	Kristen Stewart	Stephenie Meyer	5	5.920635e+08
7	Rachael Lillis	Norman J. Grossfeld	5	6.865795e+07
8	Rachael Lillis	Takeshi Shudo	5	6.865795e+07
9	Rica Matsumoto	Norman J. Grossfeld	5	6.271887e+07

Figure 21. Collaborations between actresses and directors/writers

Director vs Writer

	director	writer	count	profit
0	Peter Jackson	Fran Walsh	11	4.802452e+08
1	Ang Lee	James Schamus	8	3.690469e+07
2	Peter Jackson	Philippa Boyens	8	6.610058e+08
3	Kunihiko Yuyama	Norman J. Grossfeld	6	7.453984e+07
4	Peter Jackson	J.R.R. Tolkien	6	8.193179e+08
5	Michael Haigney	Takeshi Shudo	6	7.453984e+07
6	Michael Haigney	Norman J. Grossfeld	6	7.453984e+07
7	Kunihiko Yuyama	Takeshi Shudo	6	7.453984e+07
8	James Ivory	Ruth Prawer Jhabvala	6	-2.581750e+06
9	Danny Boyle	John Hodge	6	2.117331e+07

Figure 22. Collaborations between directors and writers

5.9 Genre

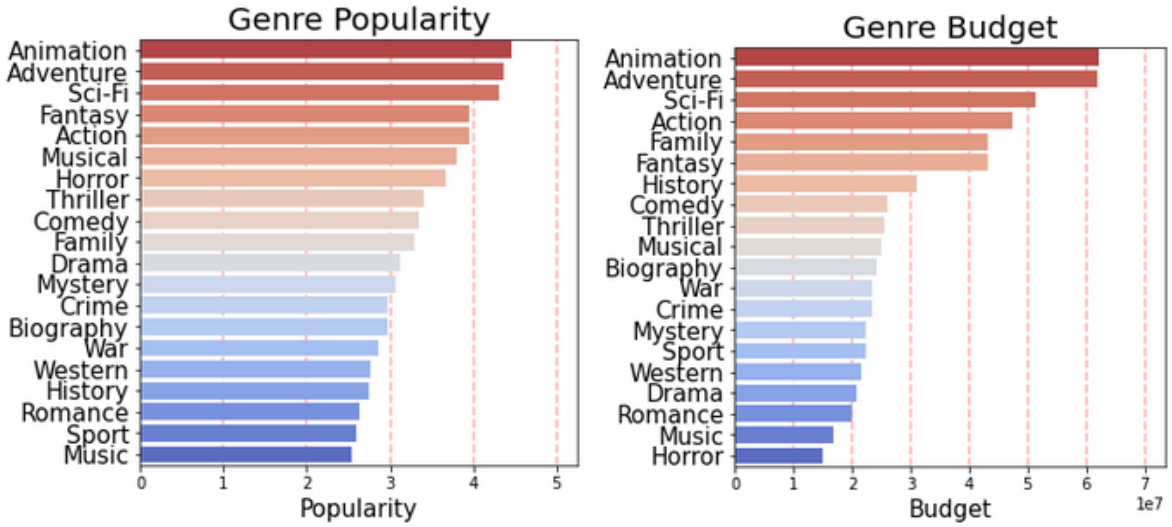


Figure 23. Genre popularity and budget

The above figures elucidate that the popularity and budget of the genre is directly proportional. Animation, adventure, sci-fi, and action have higher popularity than the other genres but also come with higher cost of budget.

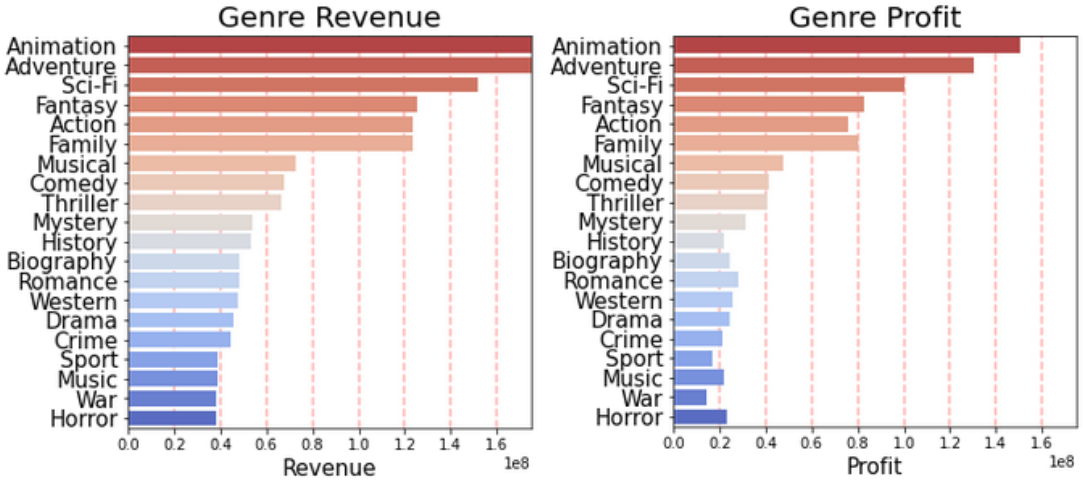


Figure 24. Genre revenue and profit

It also denotes that genres including animation, adventure and sci-fi are also the top three genres that can generate highest profit, with minor differences in the ranking of action genres (from 4th to 5th in both figures).

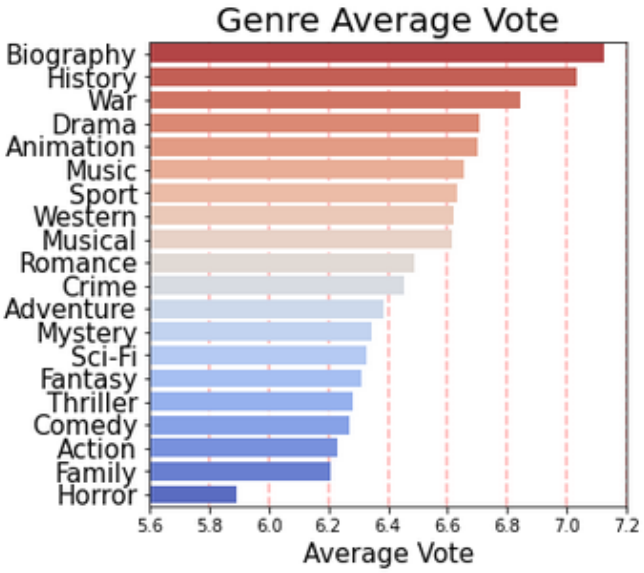
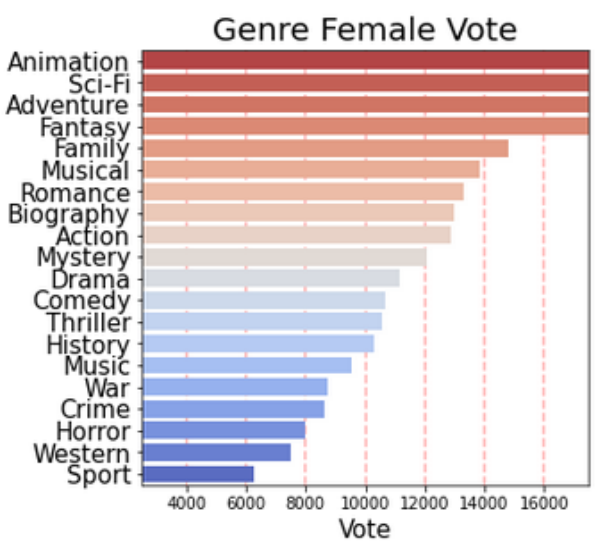
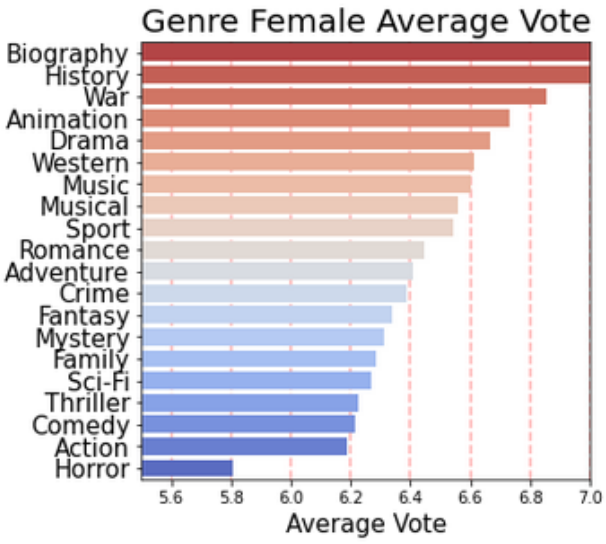


Figure 25. Genre average vote

As mentioned above, the rating of biography, history, and war genre has a higher rating than the 4 mentioned genres above even though it has low popularity, revenue, and profit. It is due to the fact that the number of movies with these genres is very low compared to the total amount of movies - about 60 movies. This has caused the bias in the visualization of our dataset.





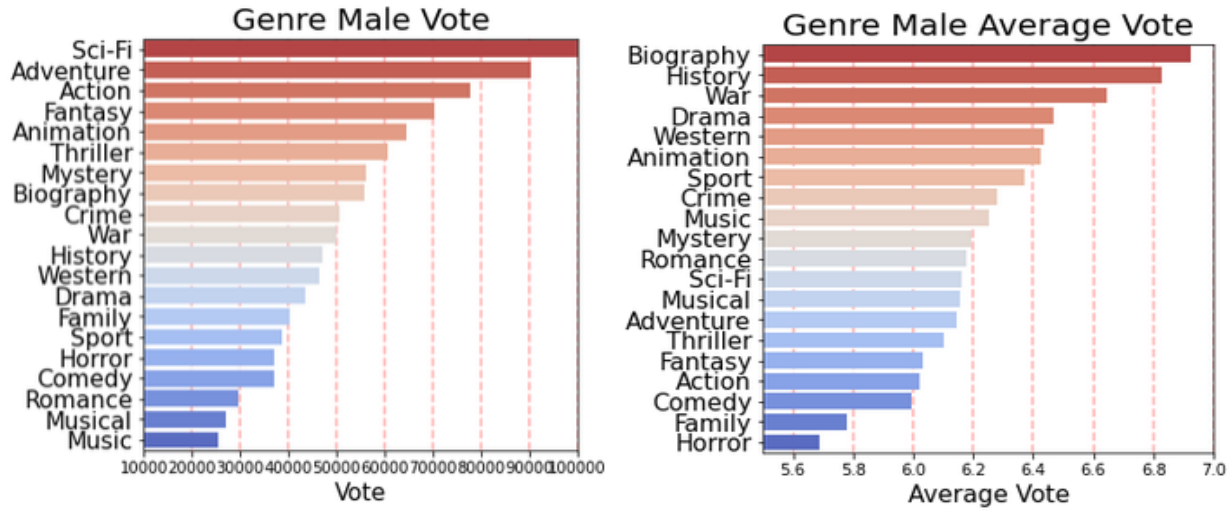
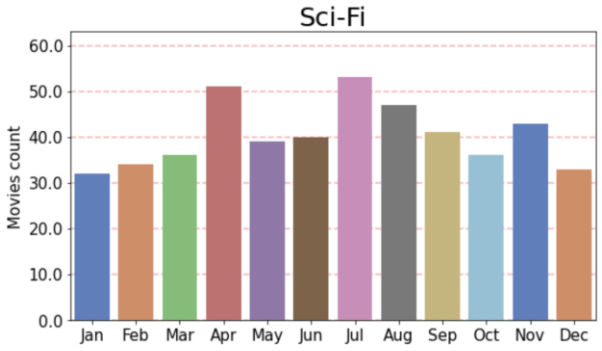


Figure 26. Genre versus female/male votes/rating.

These four graphs show the popularity of the genre based on gender's votes. Excluding biography, history and war genres (as they are biased), it is clearly shown that female viewers favor genres such as animation, science fiction, adventure and fantasy, whereas male viewers prefer science fiction, adventure and action genres. Consequently, our client is recommended to write a screen about science fiction, animation, adventure, action, or fantasy genres to capture the most attention from both genders.

To be more specific, we also make investigations in finding suitable release date for 4 topmost preferable genres (Science Fiction, Animation, Adventure, Action) as follow:

Science Fiction

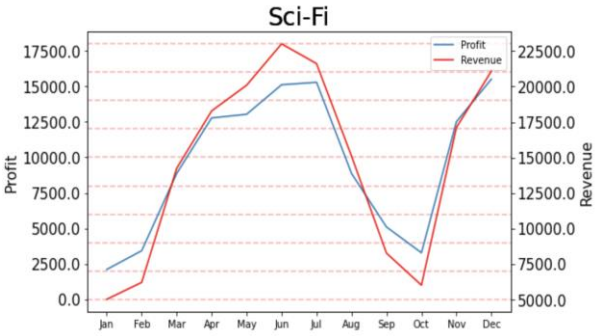


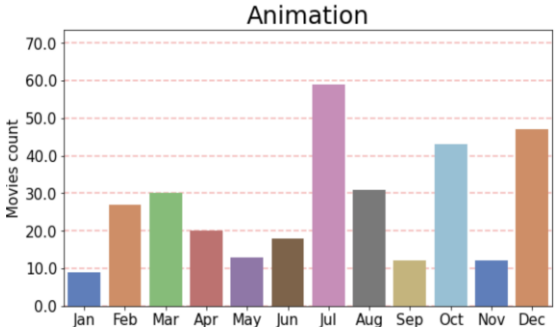
As shown in the first graph, Sci-Fi movies are released more in the middle course of the year, especially in April and July. In the second graph, the profit and revenue of these movies peak in May, June and July, where profit increases from \$130M to \$153M and revenue increases from \$200M to \$230M. It is also noticeable that science fiction movies released in December also have high profit and revenue, at \$150M and

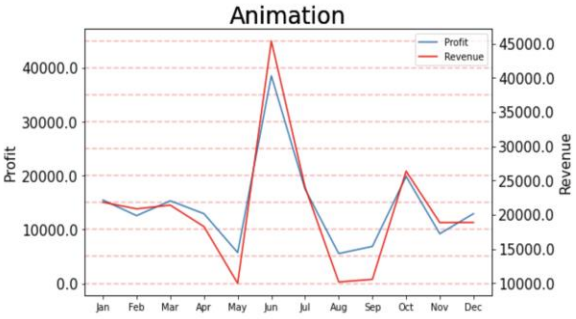
$200M respectively. Therefore, advisably, a science fiction movie should be released in April or May so that the profit and revenue of that movie can generally peak at 1-2 months later.

Figure 27. Number of Sci-Fi movies and their profit by months

Animation



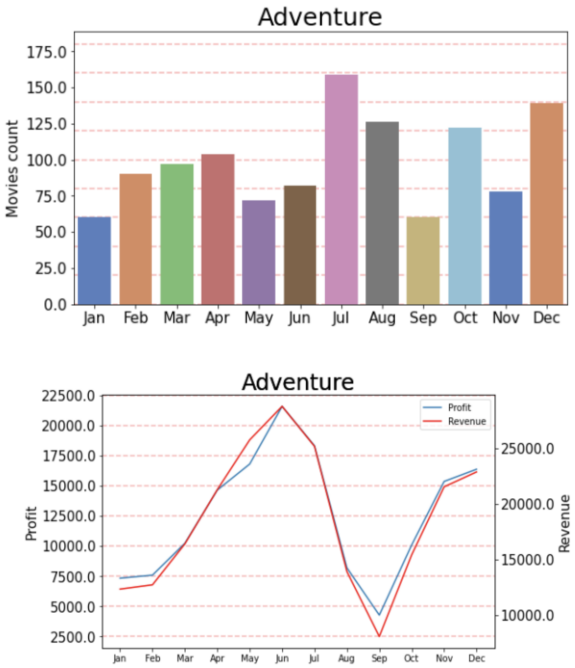




Animation movies are more likely to be released in July, nearly 60 movies on average. December came in second place with nearly 50 movies on average. The reason is that during summer and winter break, children and possibly teenagers, who are the main target audience of this genre, are on holiday. So, the number of viewers will increase if the movies are released in these months. Even though there are not a lot of movies released in May or June, the profit and revenue in June peak at $386M and $454M respectively. It is noticeable that this genre brings in the most profit and revenue at $150M and $213M, on average, respectively. Therefore, it is best to release the animation movie in June.

Figure 28. Number of animation movies and their profit by months

Adventure



Adventure genre movies have the second highest profit and revenue value next to animation. It is also the second in the number of movies produced. As can be seen in the first graph, July is the best time to release this genre of movies based on the frequencies of releases throughout the year, at nearly 150 movies on average. The profit and revenue peaked in June, at $216M and $288M respectively. It is surprising to see that in September, both the profit and revenue is bottom at $43M and $80M respectively.

Figure 29. Number of adventure movies and their profit by months

Action

Action genre has the largest number of movies produced compared to the other genres. The movies are released in uniform order throughout the year. The number of movies released peaked in August with 180 movies on average. The profit and revenue of this genre is highest in July with $145M and $206M respectively. It bottomed in September at only $29M and $61M respectively. This statistic has revealed that the Action and Adventure genres often come together, hence the common in peak and bottom data. Therefore, it is best to release an action movie near the end of June or the beginning of July to gain the maximum profit and revenue.

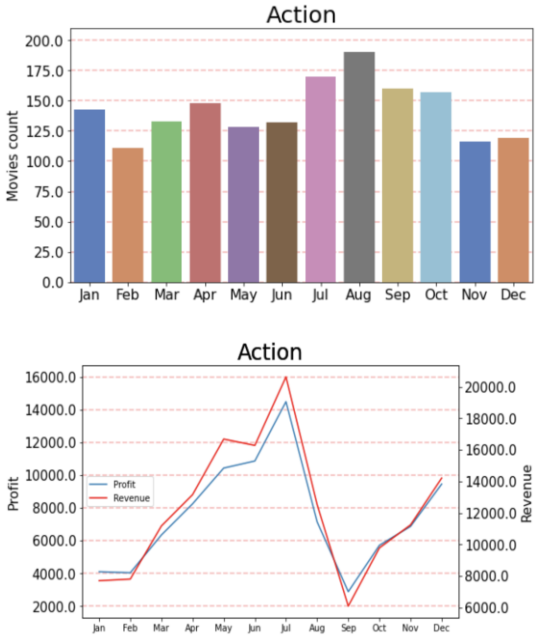
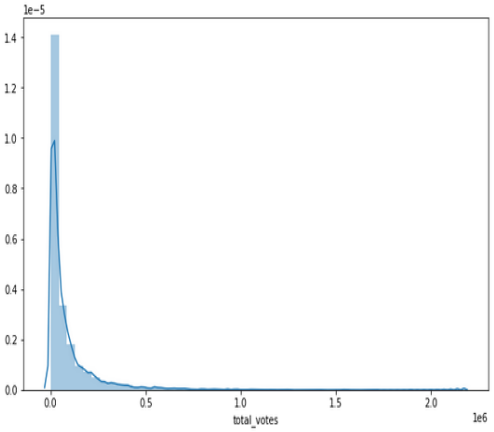


Figure 30. Number of action movies and their profit by months

5.10 Votes

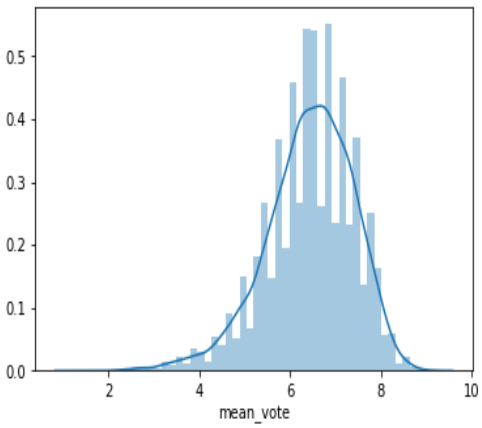
Votes are the number of votes that viewers rate for a specific movie. This section does not investigate any findings between the votes of viewers and the profit of the movie, but rather emphasizes on illustrating the correlation between specific age group and gender versus the profit of the movie for the data modelling phase.



	mean_vote
count	7100.000000
mean	6.432493
std	0.969595
min	1.300000
25%	5.900000
50%	6.500000
75%	7.100000
max	9.100000

The graph shows the distribution of total votes in the dataset. As seen from the graph, it is not normally distributed, however, the normal distribution in the **mean_vote** graph is shown quite clearly.

Figure 31. Number of votes distribution



	median_vote
count	7100.000000
mean	6.474437
std	1.138377
min	1.000000
25%	6.000000
50%	7.000000
75%	7.000000
max	10.000000

The reason for choosing the **mean_vote** attribute over **median_vote** is that it has a smaller number of standard deviations than the latter one.

Figure 32. Mean vote distribution

The two tables below show the highest total number of total_votes and mean_votes. As presented on the tables, a movie with a big number of total votes does not clarify that it would have a high score.

	title	total_votes	year
0	The Shawshank Redemption	2159628	1995
1	The Dark Knight	2134569	2008
2	Inception	1892929	2010
3	Fight Club	1725365	1999
4	Pulp Fiction	1695085	1994
5	Forrest Gump	1662528	1994
6	The Matrix	1554261	1999
7	The Lord of the Rings: The Fellowship of the Ring	1548863	2001
8	The Lord of the Rings: The Return of the King	1533574	2003
9	The Dark Knight Rises	1421494	2012

	title	mean_vote	year
0	The Shawshank Redemption	9.1	1995
1	The Dark Knight	8.9	2008
2	The Lord of the Rings: The Return of the King	8.8	2003
3	The Lord of the Rings: The Fellowship of the Ring	8.7	2001
4	Forrest Gump	8.7	1994
5	Inception	8.7	2010
6	Schindler's List	8.7	1994
7	Pulp Fiction	8.7	1994
8	Fight Club	8.7	1999
9	Interstellar	8.6	2014

Figure 33. Movies with the highest number of votes/ratings.

Additionally, the following figures are presented to gain more insights about the correlation between profit vs all the details in votes.

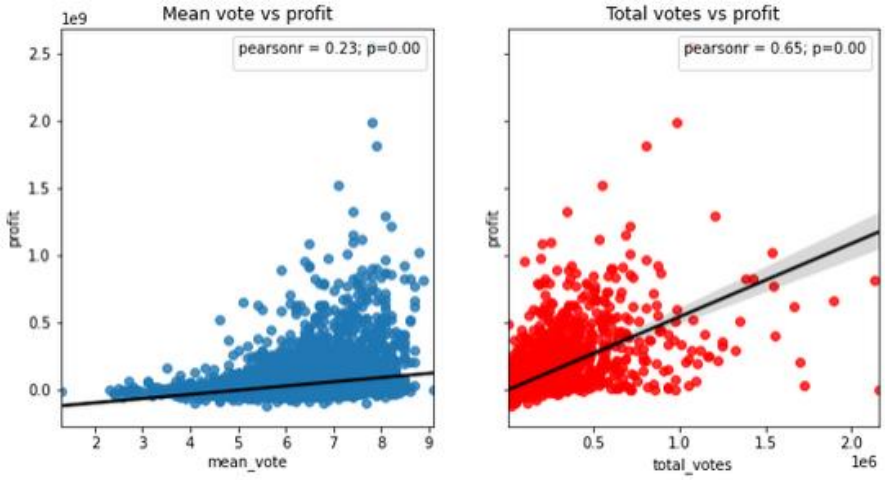


Figure 34. Mean/Total votes versus profit.

As presented above, the figure shows strong correlation between the total votes versus profit, while the mean vote versus the profit only shows 0.23/1. The following tables will find out if the total votes in gender or age can follow the same trend as the total votes or mean vote attributes or not.

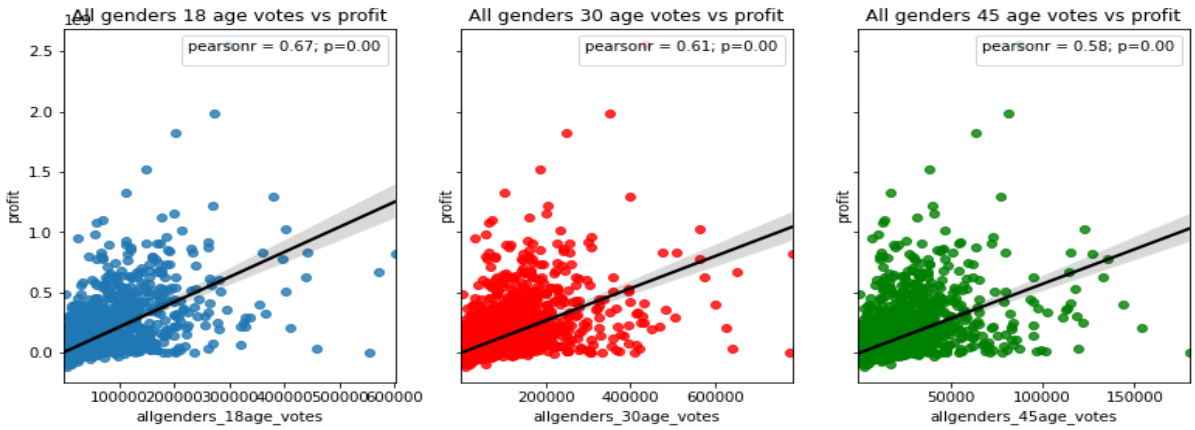


Figure 35. Total votes by age versus profit.

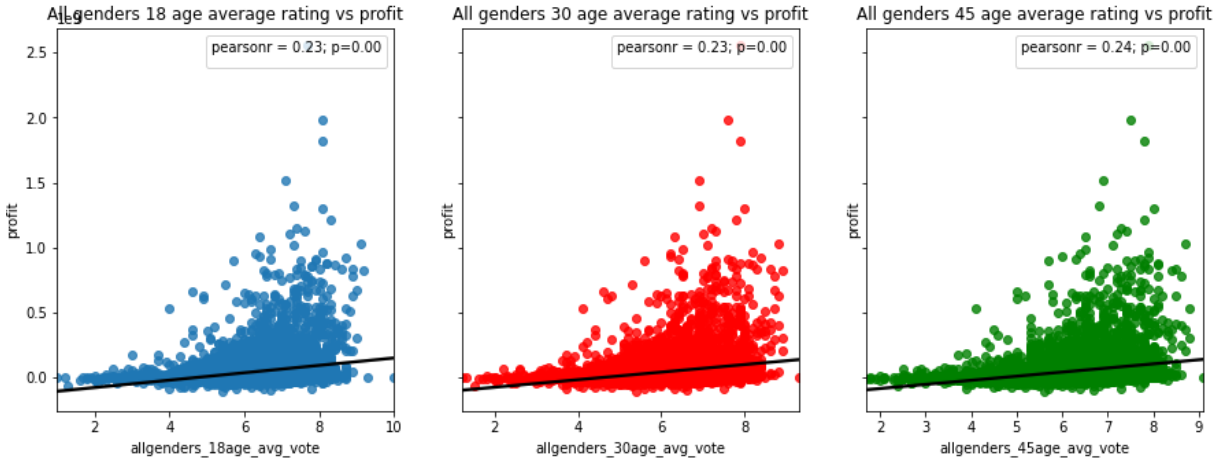


Figure 36. Average votes by age versus profit.

From all the graphs above, we can see that the total number of votes of each age group demonstrate strong correlation, while the mean votes of each age group are not. The following graphs are also showing the same correlation trends for mean vote and total votes regardless of the age range and the gender of users.

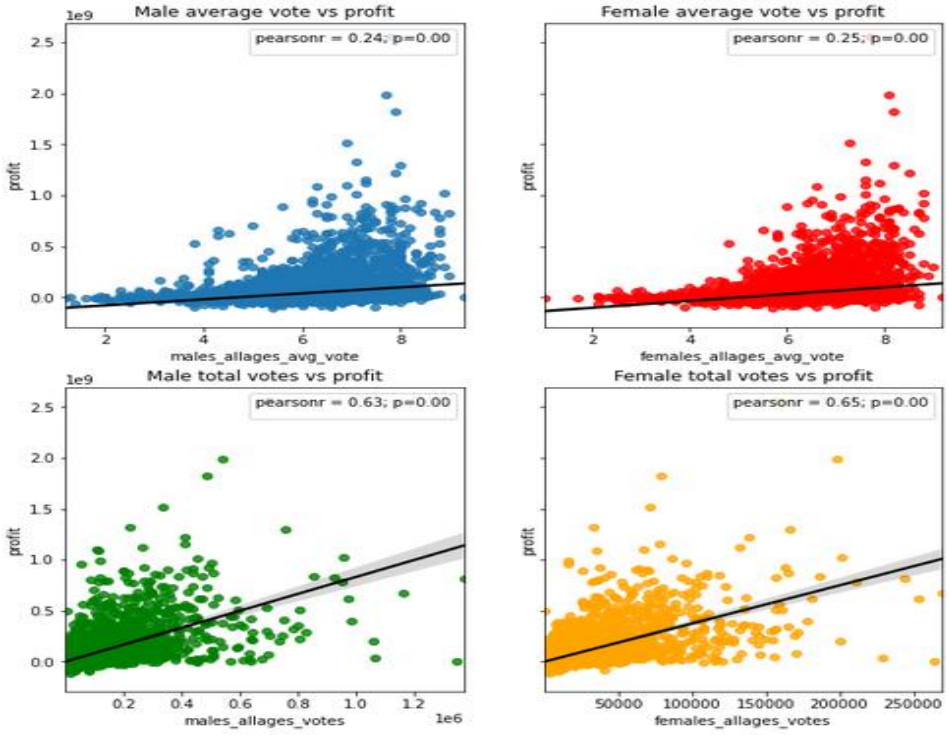


Figure 37. Male/Female total/average votes versus profit

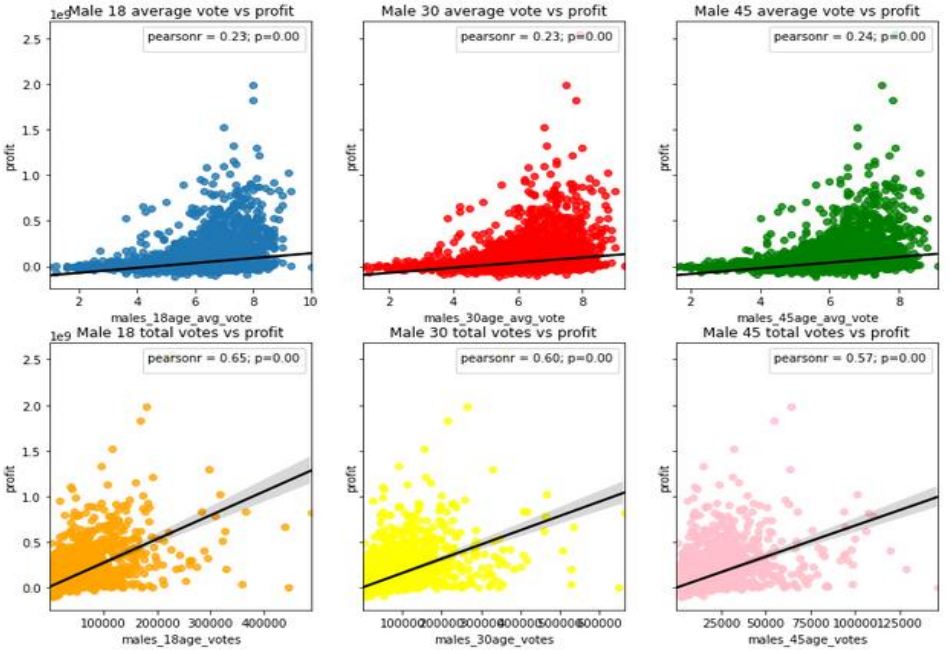


Figure 38. Male vote by age versus profit

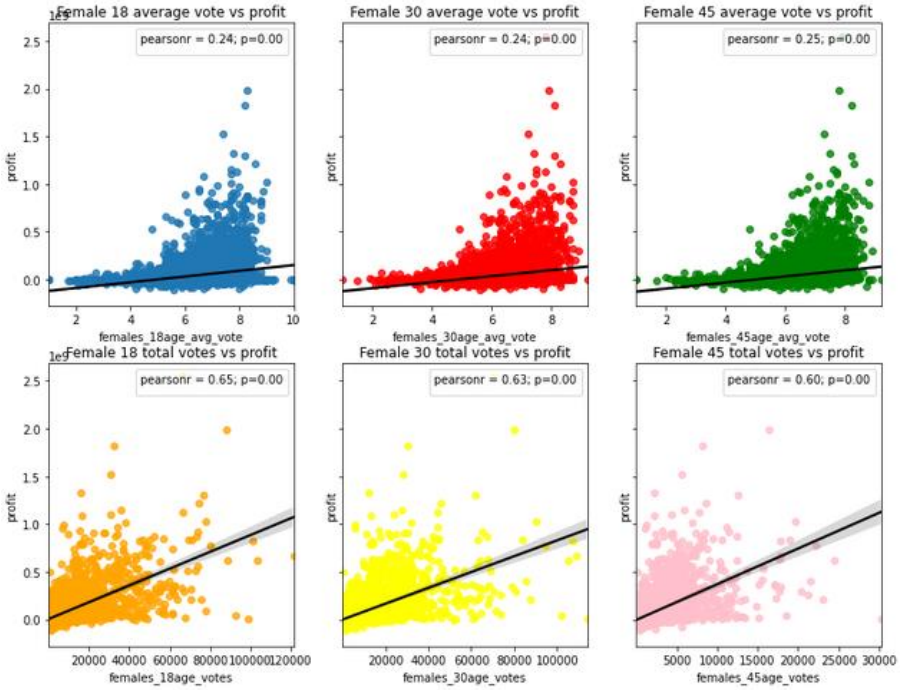


Figure 39. Female vote by age versus profit

This trend also continues with a number of us voters and non us voters:

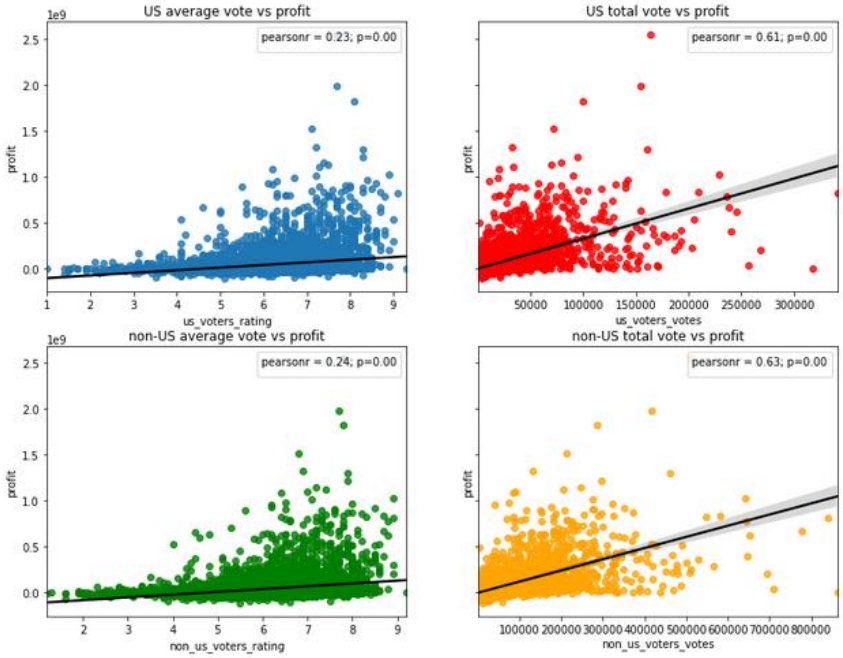


Figure 40. US/Non-US total/average votes versus profit

From all the following strong correlations, a heat map is displayed to show attributes for selecting in the machine learning model.

Figure 41. Attributes correlation versus profit heatmap

From the heat map and the findings above, the attributes that show highest correlation with profit are

- males_all_age_votes.
- us_voters_vote
- female_all_age_votes
- budget
- all_gender_30age_votes
- male_18age_votes
- female_18age_votes
- all_gender_18age_votes
- non_us_voters_votes
- total_votes.
- female_30age_votes

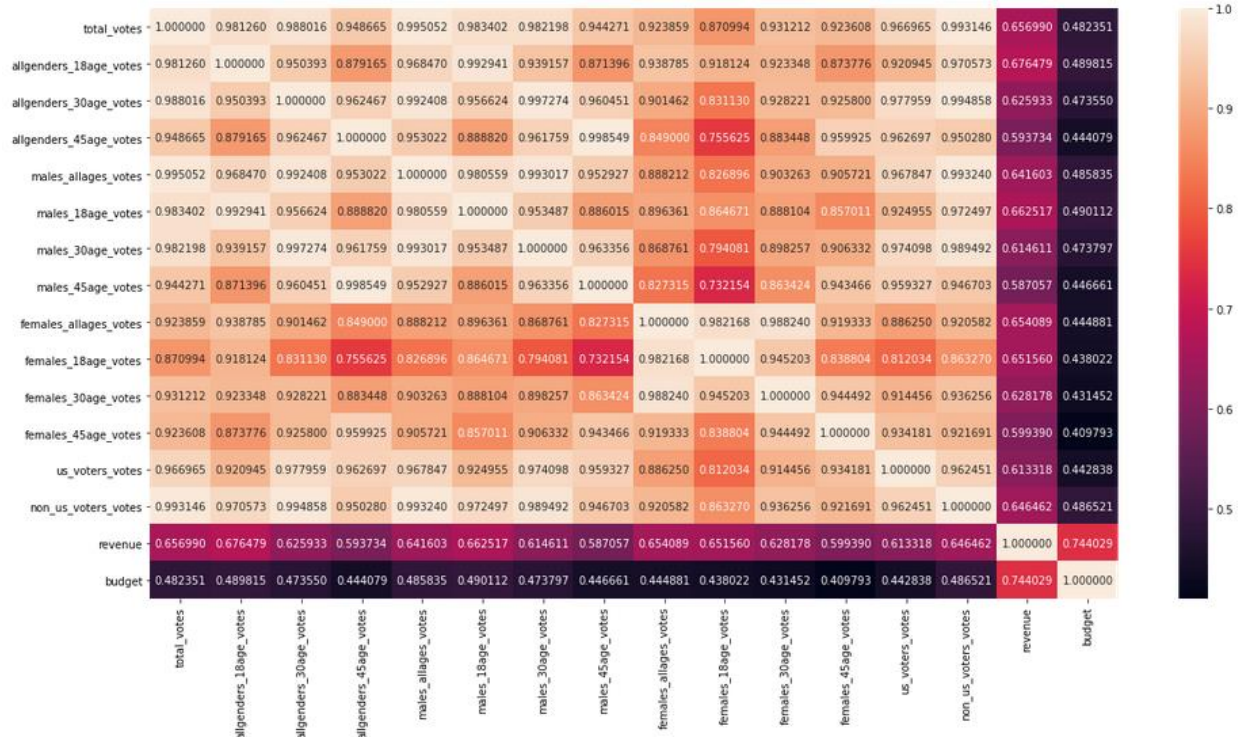


Figure 42. Attributes correlation versus revenue heatmap

From the heatmap above, these attributes for the revenue selection to give in to the model are:

- budget
- non_us_voters_votes
- us_voter_votes
- female_30age_votes
- female_18age_votes
- female_all_age_votes
- male_30age_votes
- male_18age_votes
- male_all_age_votes
- all_gender_30age_votes
- all_gender_18age_votes
- total_votes

With these attributes, they can make contributions to the feature selections in the following data modelling section.

6 Data Modelling

6.1 Return of Investment (ROI)

The return of investment (ROI) [10] value represents the profitability of the movie production. The movie is profitable if the value is positive, and vice versa. The ROI formula can be calculated as the division between the movie profit and the budget cost, where the result is expressed in percentage (%). Moreover, the ROI results can be interpreted as either the positive or negative of the results as well as the number of each digit occurring in the results.

$$ROI\ (\%)\ =\ (\ Profit\ /\ Budget\ )\ *\ 100\%$$

Machine Learning classification models are applied to classify profitability of the product. The target is mapped into 2 values: 0 and 1, where 0 stands for loss (negative value) and 1 stands for profitable (positive value). Specifically, 8 different classification algorithms are applied and arranged in the order of descending accuracy score in the table below.

Modelling Algorithms	Accuracy score (%)
Random Forest	80.14
Decision Tree	79.13
Logistic Regression	73.91
KNN	73.40
Perceptron	72.39
Naive Bayes	66.96
Support Vector Machine	56.96
Stochastic Gradient Descent	56.96

Table 4. Classification model algorithms' accuracy comparison

From the results above, we decided to apply the Random Forest and Decision Tree classification since they have the highest accuracy score and near the acceptable range (80%) the most.

After continuously training model with various combinations of features, the following set of features are proven to generate the highest accuracy score:

- **budget**: The budget of the production.

- **us_voters_votes**: Number of votes from the U.S.
- **non_us_voters_votes**: Number of votes from outside of the U.S.
- **Total_vote**: Number of the total votes of a movie.
- **Gender-Age-related vote attributes:**
    - **allgenders_18age_votes**: Number of votes from users who are 18-30 years old.
    - **allgenders_30age_vote**s: Number of votes from users who are 30-45 years old.
    - **females_allages_votes**: Number of votes from female users.
    - **females_18age_votes**: Number of votes from female users who are 18-30 years old.
    - **females_30age_votes**: Number of votes from female users who are 30-45 years old.
    - **males_18age_votes**: Number of votes from male users who are 18-30 years old.

The importance of each feature is shown in Figure 43.

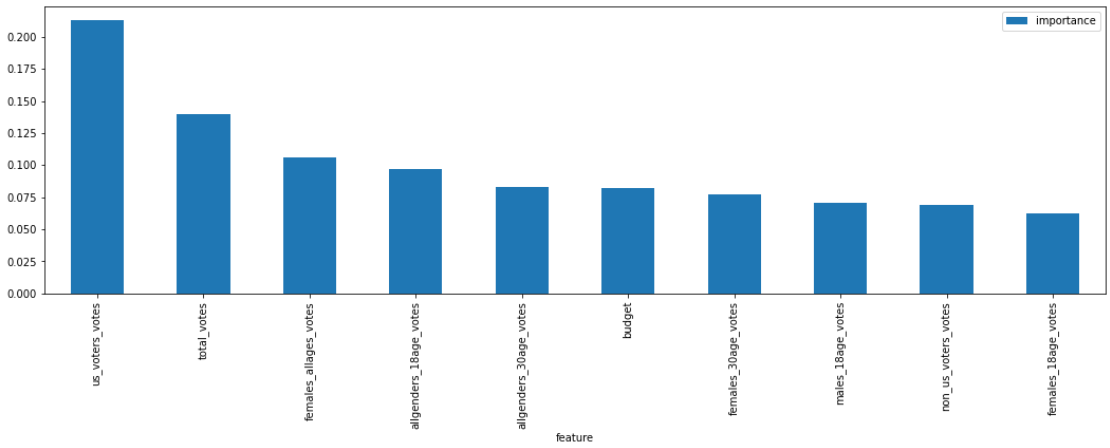


Figure 43. Attributes' importance for classification

Decision Tree

Decision Tree is a non-parametric supervised learning method used for classification and regression.

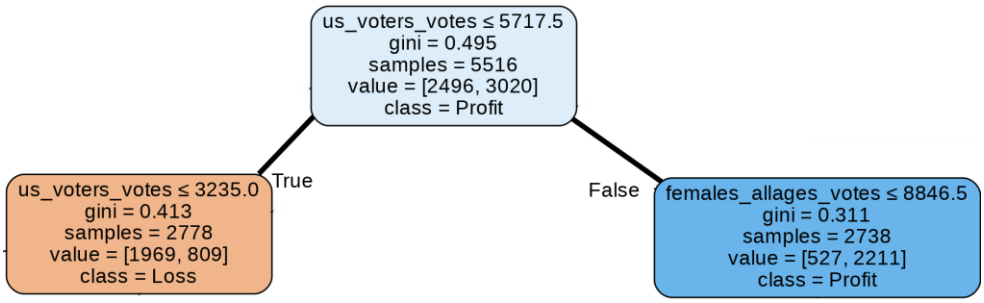


Figure 44. Parent Node

The root node is decided by the **number of US voters**, which is divided into 2 main branches from the pivot data point at 5717.5 votes.

For each of the following branches, 1 sample rule is provided:

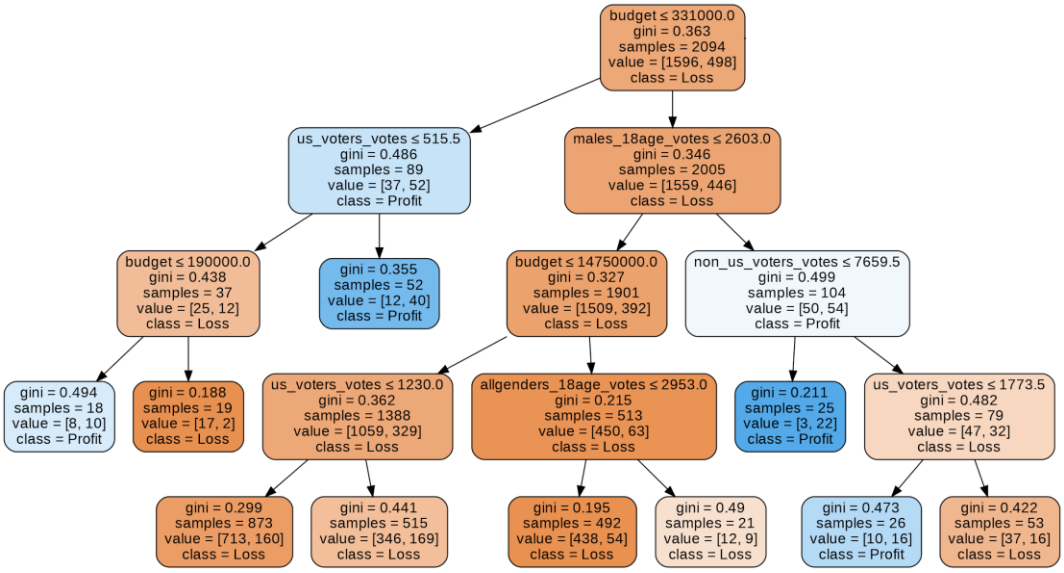


Figure 45. First major branch

- If the **number of US voters** is less than or equal to 3235 votes (from Figure 42), and the **movie budget** is greater than $331000, and the **male voter (18-30 years old) votes** less than or equal to 2603, and the **movie budget** is greater than $14750000, and the **all voter (18-30 years old) votes** less than or equal to 2953 then the movie has the profitable rate of 10.9% (54/492).

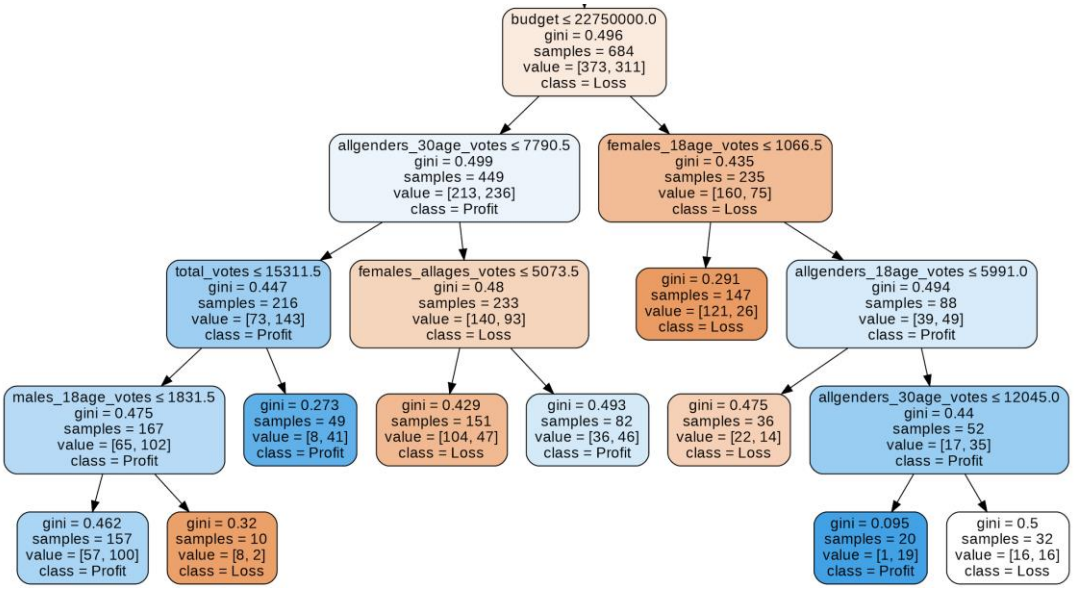


Figure 46. Second major branch

- If the **US voter votes** greater than 3235 votes (from Figure 42), the **movie budget** is greater than $22750000, the **female voter (18-30 years old)** is greater than 1066, the **all voter (18-30 years old) votes** less than or equal to 5991, the **all voter (30-45 years old) votes** greater than 12045 then the movie has the profitable rate of 50% (16/32).

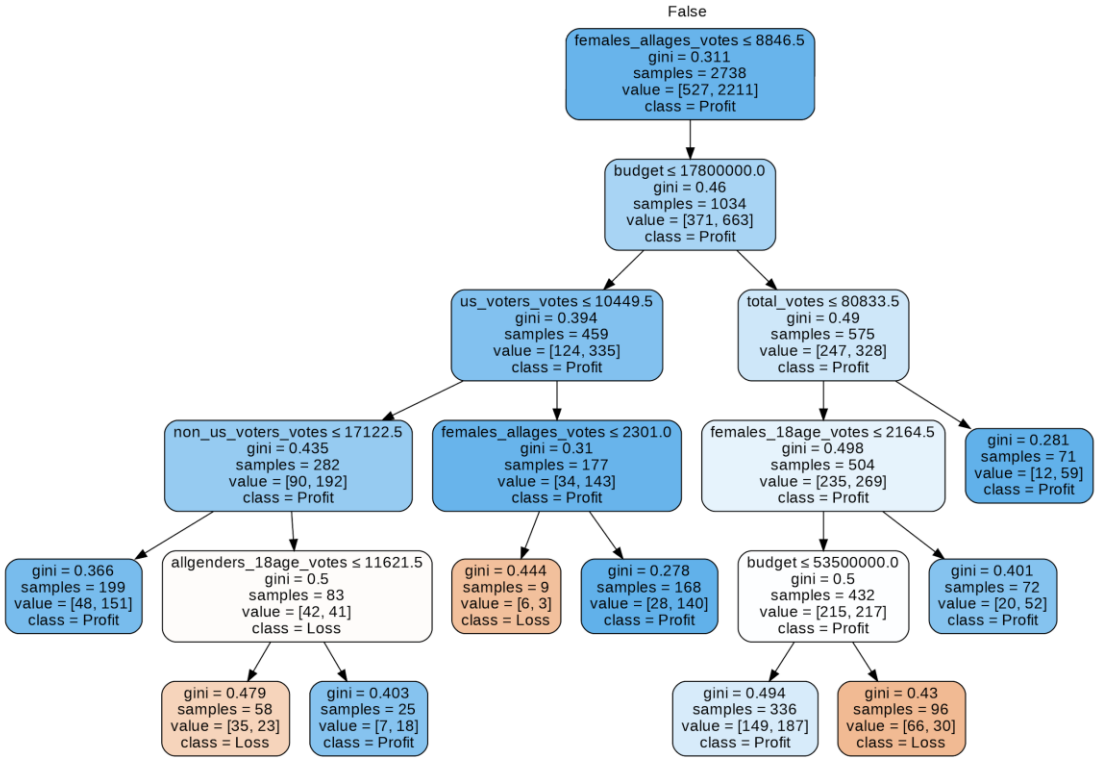


Figure 47. Third major branch

- If the **female total votes** less than or equal to 8846.5, the **movie budget** is greater than $\$178x10^5$, the **total votes** less than or equal to 80833.5, the **female voter (18-30 years old) votes** less than or equal to 2164.5, the **movie budget** is less than or equal to $\$6535x10^6$ then the movie has the profitable rate of 56% (187/336).

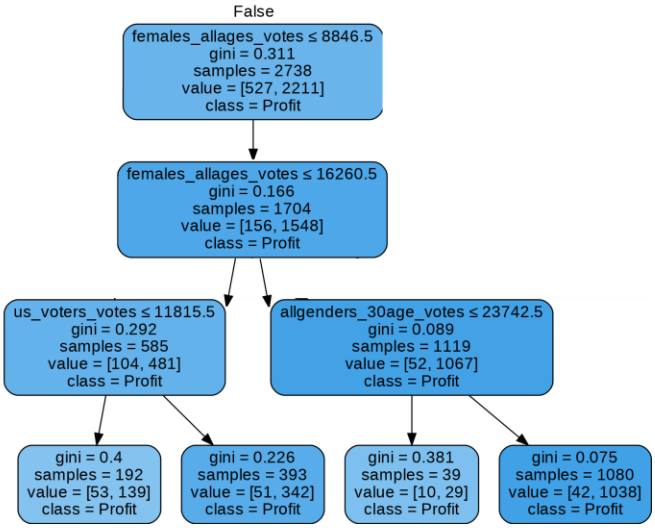


Figure 48. Fourth major branch

- If the **female total votes** greater than 8846.5, the **female total votes** greater than 16260.5, the **total votes (30-45 years old)** is greater than 23742.5 then the movie has the profitable rate of 96% (1038/1080).

The accuracy score of the model above is **79.13%**.

A confusion matrix is applied to evaluate the performance of this model:

		Predict	
		Loss	Profit
Actual	Loss	1892	604
	Profit	726	2294

Table 5. Decision Tree confusion matrix

Random Forest

Random Forest is an algorithm that combines multiple decision trees to build the most accurate model. Therefore, Random Forest usually has higher accuracy than Decision Tree. It uses multiple uncorrelated models to produce ensemble predictions. The upper advantage of this algorithm is that the models protect each other from their errors. The same set of features is fitted into the Random Forest Classifier model.

The accuracy score of the model is **80.14%**.

A confusion matrix is applied to evaluate the performance of this model:

		Predict	
		Loss	Profit
Actual	Loss	1902	594
	Profit	610	2410

Table 6. Random Forest confusion matrix

6.2 Revenue

Besides ROI, revenue is also an attribute which can determine the profitability of a movie production. Using regression analysis, we can predict the revenue number of a movie based on the following set of attributes. We used 2 different regression models to predict as close as the possible revenue number: Linear Regression and Random Forest Regression.

The following set of features are proven to generate the highest accuracy score:

- **Total_votes**: Number of total votes
- **Non_us_voters_votes**: Number of votes from outside of the U.S.
- **Budget**: The budget of the movie.
- **is_Friday**: If the movie is released on Friday, the value is 1 and 0 if it is not.
- **is_Holiday**: If the movie is released on a holiday, the value is 1 and 0 if it is not.

- **Gender-Age-related vote attributes:**
  - **Allgenders_18age_votes**: Number of votes from users who are 18-30 years old.
  - **Males_allages_votes**: Number of votes from male users.
  - **Males_18age_votes**: Number of votes from male users who are 18-30 years old.
  - **Males_30age_votes**: Number of votes from male users who are 30-45 years old.

- **Females_allages_votes**: Number of votes from female users.
- **Females_18age_votes**: Number of votes from female users who are 18-30 years old

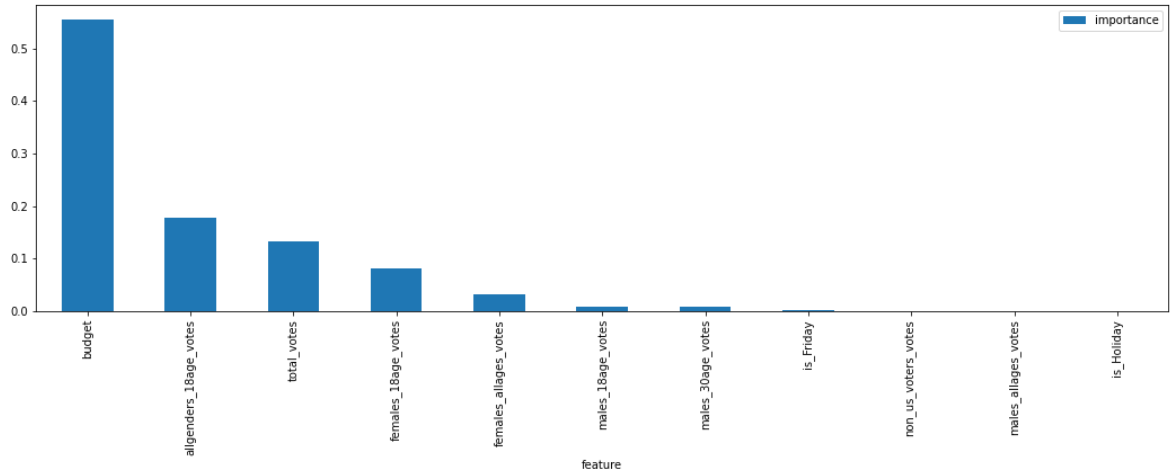


Figure 49. Attributes' importance for regression

The importance score of each attribute is displayed in the figure above. Even though the 4 last attributes have a small importance score in comparison with the other, if removed, the accuracy will be dropped by 2-4%.

7 Data Evaluation

7.1 Classification

Evaluation Metric	Decision Tree	Random Forest
Accuracy	79.13%	80.14%
True Positive	2294	2410
True Negative	1892	1902
False Positive	604	594
False Negative	726	610
Precision	0.79	0.80
Recall	0.76	0.79
F-Score	0.78	0.80

Table 7. Decision Tree and Random Forest evaluation

where:

True positive: The number of movies were correctly classified as profit.
True negative: The number of movies were correctly classified as loss.
False positive: The number of movies were wrongly classified as profit.
False negative: The number of movies were wrongly classified as loss.
Precision: The percentage of this model predicts a movie profitable correctly.
Recall: The percentage of this model predicted movie actually profitable.

From the comparison table above, the Random Forest model has a higher accuracy score than the Decision Tree model (1.81%). The difference between them is not significant.

From the Decision Tree model, we can see the model splitting the tree by Gini. The Attribute which has the greatest impact is the number of votes from the U.S. It is also our main target market so the client should consider which combination is in favor of the U.S. viewers. The female's vote is the second greatest impact on return of investment. The client should consider writing a screen targeting female viewers. According to [11], women viewers account for 26% of movie goers in Canada and US, however, they are just a small proportion in the online movie rating. Moreover, [11] also proved that the number of male audiences always outnumber the total number of female voters. Therefore, it would be understandable that the classification machine learning model would produce a better result on the female voters.

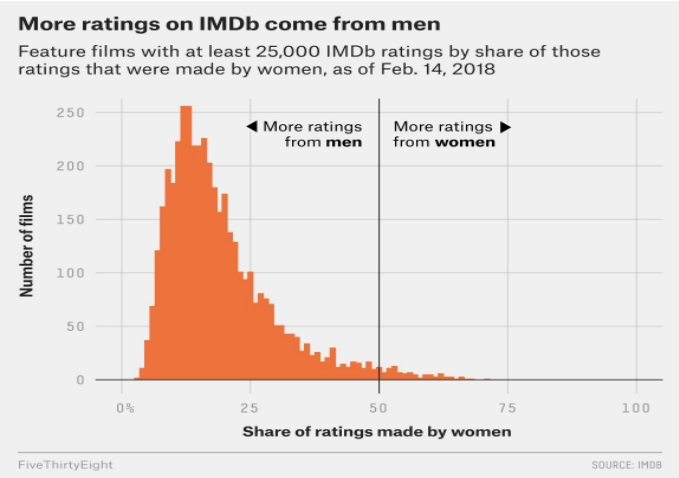


Figure 49. Share of ratings made by male and female voters [11]

7.2 Regression

Because the range of our revenue data is large, so instead of using mean squared error, the root mean square value divided by the range of revenue is used to evaluate the models. The value shows the mean difference of the test value versus the predicted linear line.

	Linear Regression	Random Forest Regression
RMS/range	0.04	0.04
R^2	0.73	0.76

Table 8. Linear Regression and Random Forest evaluation

The R^2 value determines the accuracy of our models. The Random Forest Regression model has more accuracy score than the Linear Regression model: 3%. The difference between them is insignificant and with the same value of RMS which comes to the conclusion that the revenue is predicted with the accuracy rate of 73-76%.
In general, by using Random Forest and Decision Tree as the best modelling algorithms for classification, and Random Forest and Linear Regression for regression, the data modelling phase satisfies the stated Data Mining Success Criteria as the precision of these two algorithms in both classification and regression are in the acceptable range.

8 Recommendations and Conclusion

Overall, the project concluded with great experience in the field of Data Analytics, with prime focuses on data management and analysis techniques. In data preparation, the process of cleaning dirty data using complexity reduction methods such as aggregation, feature creation, feature transformation and dimensionality reduction to remove anomalies and add additional attributes for the final dataset. This process is done with the support of Python dataset-related libraries called Pandas, Numpy, Scipy, Datetime and visualization libraries such as Seaborn, Plotly and Matplotlib to get a clear understanding of the data and the correlations between its attributes through means of bar graphs, scatter plots and trend lines. Later on, different modelling algorithms regarding classification and regression are implemented using the SKLearn package to compute the highest accuracy possible from a set of attributes.

Following the CRISP-DM process, the datasets of IMDb movies have been used to find valuable insights to meet stated business objectives. With the ultimate purpose of investigating multiple ways that can help Pyxar to promote a movie that could potentially achieve the best profit in return, this paper aims to investigate key aspects of a movie that can answer the question. In specific, the movie release date is proven to positively affect the end profit of that movie (of all genres), where May to July are observed to be the most profitable months and to be more specific, the movie should be released on Friday. Moreover, genres are unexpectedly to have a rather low to medium impact on the target value of our dataset. However, it is still useful to study these data to have a better tailored movie development experience. Genres such as science fiction, animation, adventure, and action are discovered to be the most successful genres in terms of generated profit.

While genre is one of the key independent variables to the profit, participants (actor, actress, director, writer) also have a strong positive correlation with the mentioned profit.

Tables with information about specific collaborations between these participants are also provided to let our client choose suitable combinations of people for their movie (based on the genre of the movie). Notably, more famous people, regardless of roles, contribute in one movie does not imply that the movie will make more profit, as well as the budget of that movie will increase. However, this might not be strongly accurate since the popularity attribute is evidently shown to be skewed in this case. In addition, other inferior findings are also included for Pyxar to become more flexible in the development stage. One good example is that based on the title and overview word cloud, preliminary movie ideas can be provided to the writers and producers and appropriate advertising strategies are conducted to capture people's attention more.

Through numerous trials of model training, a set of features is established to increase the accuracy of the models highest as possible. Initially, various modelling algorithms are computed to make comparisons to decide which ones can generate the highest accuracy, where Decision Tree and Random Forest meets the data mining success criteria, at 79% and 80% respectively for classification, and Random Forest and Linear Regression, at 76% and 73% respectively for regression.

Overall, Daten is fairly confident of both the accuracy and validity of the results of the project and thus continues the development stage. In the future, the dataset should be updated with newer data to eliminate bias results (popularity attribute or the balance between male vs female voters as examples) since the movie industry is currently developing at a fast pace. A good and viable direction is that the data should be collected directly from the IMDb and other movie websites using web crawling and scraping software programs (Selenium framework or Scrapy) as datasets on Kaggle might have been through some pre-processings beforehand. Lastly, from the foundation of these models, a movie recommendation could be developed to increase the profit of the company, where the system will collect user data to calculate a combination of movies that fit the clients the most.

9 References

[1] S. Burby, "Big data and creativity: What we can learn from 'House of Cards'", *The Next Web*, 2020. [Online]. Available: https://thenextweb.com/insider/2016/03/20/data-inspires-creativity/. [Accessed: 27- Sep- 2020].

[2]"IMDb movies extensive dataset", *Kaggle.com*, 2020. [Online]. Available: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset. [Accessed: 27- Sep- 2020].

[3]"The Movies Dataset", *Kaggle.com*, 2020. [Online]. Available: https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv. [Accessed: 27- Sep- 2020].

[4]"IMDb Dataset", *Kaggle.com*, 2020. [Online]. Available: https://www.kaggle.com/ashirwadsangwan/imdb-dataset. [Accessed: 27- Sep- 2020].

[5] "IMDb | Help", *Help.imdb.com*, 2020. [Online]. Available: https://help.imdb.com/article/imdb/discover-watch/box-office-faq/G4UCJ3GMFX6F23ZX?ref_=helpart_nav_15#. [Accessed: 25- Sep- 2020].

[6]"The 70s was the golden age of Hollywood. But why?", *the Guardian*, 2020. [Online]. Available: https://www.theguardian.com/film/filmblog/2007/jul/13/the70swasthegoldenageof. [Accessed: 26- Sep- 2020].

[7]2020. [Online]. Available: https://www.cgv.vn/en/terms-use/. [Accessed: 27- Sep- 2020].

[8]"george clooney and julia roberts movie - Google Search", *Google.com*, 2020. [Online]. Available: https://www.google.com/search?client=firefox-b-d&q=george+clooney+and+julia+roberts+movie. [Accessed: 27- Sep- 2020].

[9]"adam sandler and dennis dugan movie - Google Search", *Google.com*, 2020. [Online]. Available: https://www.google.com/search?client=firefox-b-d&sxsrf=ALeKk01i1kOBDCjz7IAB1GUPHuC5RWGdyQ%3A1601145298081&ei=0olvX4XSBNXprQGI4rmABQ&q=adam+sandler+and+dennis+dugan+movie&oq=adam+sandler+and+dennis+dugan+movie&gs_lcp=CgZwc3ktYWIQAzIECCMQJ1DqEFjqEGC8EmgAcAB4AIABkwGIAfUBkgEDMS4xmAEAoAEBqgEHZ3dzLXdpesABAQ&sclient=psy-ab&ved=0ahUKEwiFg-eFu4fsAhXVdCsKHQhxDlAQ4dUDCAw&uact=5. [Accessed: 27- Sep- 2020].

[10]"Return on Investment (ROI)", *Investopedia*, 2020. [Online]. Available: https://www.investopedia.com/terms/r/returnoninvestment.asp. [Accessed: 25- Sep- 2020].

[11] W. Hickey, "What If Online Movie Ratings Weren't Based Almost Entirely On What Men Think?", *FiveThirtyEight*, 2020. [Online]. Available: https://fivethirtyeight.com/features/what-if-online-movie-ratings-werent-based-almost-entirely-on-what-men-think/. [Accessed: 27- Sep- 2020].