University of Tartu

Faculty of Science and Technology

Institute of Computer Science

Janar Ojalaid

The power of stars: An empirical analysis of successful and flop movies

Master's Thesis (30 ECTS)
Software Engineering Curriculum

Supervisor(s):

Rajesh Sharma PhD

Tartu 2020

Abstract:

The power of stars: An empirical analysis of successful and flop movies

Production companies collectively produce thousands of new movies every year. Some of the movies perform incredibly well but some of them turn out to be flops. Having a movie flop can be fatal to a production company because of the costs of producing a movie. Therefore, it is essential to produce movies with a higher likelihood of success. In this thesis, we explore the features of successful and unsuccessful movies by performing empirical analysis on a large dataset of movies. The data about movies was collected from The Internet Movie Database (IMDb), The Movie Database (TMDb) and Box Office Mojo, and the data about movie trailers from YouTube Data API v3. The final dataset contains 470,743 movies, has 26 features, and is about 2.1 GB large. As expected, a single feature can not be used to determine whether a movie is successful or not, the outcome depends on many different factors. However, by performing association rule mining, we found that there are lots of combinations of features that affect the outcome of the movie with the crew, cast, production company, belonging to a collection, trailer, genre, maturity rating all playing a major role.

CERCS: P170 Computer science, numerical analysis, systems, control

Keywords: successful movies, unsuccessful movies, movie characteristics, empirical analysis, association rule mining

Tähtede võim: empiiriline analüüs edukatest ja läbikukkunud filmidest

Filme tootvad produktsioonifirmad loovad tuhandeid uusi filme aastas. Mõndadel filmidel läheb väga hästi ja neist saavad hitid, kuid mõned kukuvad läbi. Filmi läbikukkumine võib suure produktsioonikulu tõttu saada produktsioonifirmale saatuslikuks, mistõttu on väga oluline produtseerida filme, mille läbilöömise tõenäosus on kõrge. Käesolevas töös uurime nii edukatele kui ka läbikukkunud filmidele omaseid tunnuseid, viies läbi empiirilise analüüsi suuremahulise andmestiku peal. Andmed filmide ja nende treilerite kohta on kogutud järgnevatest allikatest: The Internet Movie Database (IMDb), The Movie Database (TMDb), Box Office Mojo ja YouTube Data API v3. Kogutud andmestik sisaldab 470,743 filmi, millel on 26 karakteristikut ning on 2.1 GB suurune. Nagu eeldatud, ei saa filmi tulemust hinnata ainult ühe karakteristiku põhjal - tulemus sõltub mitmetest erinevatest faktoritest. Assotsiatsioonireeglite kaevandamise tulemusel leidsime, et eksisteerib suur hulk erinevaid tunnuste kombinatsioone, mis mõjutavad filmi tulemust. Põhilisteks sellisteks tunnusteks olid filmi meeskond, näitlejad, produktsioonifirma, kollektsiooni kuuluvus, treiler, žanr ja sisu hinnang.

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Võtmesõnad: hittfilmid, läbikukkunud filmid, filmi karakteristikud, empiiriline analüüs, assotsiatsioonireeglite leidmine

Contents

1 Introduction

The movie business is an enormous industry, with global box office revenue having reached an all-time high of $42.5 billion worldwide in 2019 according to [39]. Every year thousands of movies are released worldwide. Some of them turn out to be box office hits and some of them flops. According to [21], the average cost of movies released in US cinemas between 1999 and 2018 was $18 million. Moreover, for major studio movies, the average cost is even higher – up to $100 million according to [36]. Having a movie flop can turn out to be detrimental for the production company due to the high cost of producing a movie. Knowing the characteristics of successful and unsuccessful movies can help to avoid that.

Various research has been done about analyses of movies using social media platforms such as IMDb, Twitter, blogs, etc., but most of these works are around predictive analytics, i.e., predicting the outcome of box office results of movies. The features of successful and unsuccessful movies are usually not explored. Moreover, the size of the datasets has usually been quite small. There were only a few papers that had analysed more than 1,000 movies – [27] analysed 2,506 movies, [15] analysed 4,260 movies, and [20] analysed 6,590 movies. The rest of the papers had much smaller amounts such as 312 movies in [34], 311 movies in [45], 200 movies in [12], 35 movies in [6], and 24 movies in [9], while [2] only analysed Bollywood movies. Similarly, the amount of analysed features has been quite low. Most of the papers analysed less than 15 features, e.g., [15] used 11 features, [34] used 8 features, and [45] used 11 features.

In this thesis, we applied a data science approach to extract features, which can be used to infer the reasons behind successful and flop movies. We concentrated on the following research questions:

- **RQ1:** How do the characteristics between successful and unsuccessful movies differ?

  RQ1.1: What are the characteristics of successful movies?

  RQ1.2: What are the characteristics of unsuccessful movies?

We applied the following approach:

1. **Dataset collection:** We first collected data from multiple data sources including IMDb, TMDb, YouTube and Box Office Mojo, using data dumps, application programming interfaces (APIs) and web scraping. The final dataset is held in a PostgreSQL database, contains 470,743 movies, has 26 features, and is about 2.1 GB large. This dataset included movies from all over the world.

2. **Categorisation of movies:** We divided the movies into three classes – hits, flops and neutral. Our definition of hits and flops is based on two criteria, namely, i) income and ii) rating. By income, we considered the movie a hit if it made at least triple its budget and a flop if it made less than its budget. By rating, we considered the movie a hit if it had an IMDb rating above 7.1 and a flop if the rating was less than 5.5.

3. **Approaches:** In particular, we used association rule mining and clustering to find out the characteristics of both successful and unsuccessful movies and the differences between them.

To the best of our knowledge, this is the first work which has analysed such a large dataset of movies. The major findings of our analysis on this dataset are the following:

1. **Features of successful movies:** We found out that successful movies usually belong to a collection, have good trailers, and unrestricted maturity rating. Good cast and crew are also very important.

2. **Features of flop movies:** We found out that flop movies usually don't belong to a collection, have restricted maturity rating, and a very short length. Additionally, we found that documentaries are more likely to flop.

The rest of the thesis is organized as follows:

1. In Chapter 2, we discuss related works.

2. In Chapter 3, we describe the collection and available features of the data.

3. In Chapter 4, we describe the data preparation steps and approaches to data analysis.

4. In Chapter 5, we give an overview of the results.

5. We conclude the thesis with a discussion of future directions in Chapter 6.

2 Related work

In this chapter, the research questions are discussed by researching relevant papers. Firstly, we describe how movie success can be defined (Section 2.1). Secondly, we explore how the outcome of the movie and the characteristics of successful and unsuccessful movies can be found using data science approaches (Section 2.2). Finally, we talk about the main findings in this context (Section 2.3).

2.1 Defining movie success

The way success is defined is a matter of the uttermost importance in these kinds of problems. Success can be defined in different ways and the results can differ greatly depending on the definition. In the researched papers, the success was measured in the following parameters:

1. Box-office revenue – Box-office revenue is the income a movie generates before any expenses are taken out. Most of the works such as [6, 22, 30, 45] have primarily focused on using revenue as the metric of success. However, the downside of using revenue as the measure is that it does not take the costs of producing a movie into account.

2. Profit and profitability – Profit and profitability are closely related. Profit is an absolute number, which is usually referred to as the net income, i.e., the income a movie generates after expenses are taken out. Profit was used as the measure in works such as [16, 27]. Profitability, on the other hand, is a relative metric used to determine profit in relation to the size of the business. The most common profitability ratio is the return on investment (ROI). ROI was used as the measure in [17].

3. The number of admissions – The number of admissions shows how many movie tickets were sold. This metric was used in works such as [10, 33]. For example, [33] classified a movie as a hit if it had at least 400,000 admissions, which was a threshold value that only the top 20% of German films released in the researched time period (1991-2006) reached.

4. Critics rating – Movie criticism is the analysis and evaluation of movies. Critics review the movies and publish the results on a variety of platforms such as magazines, blogs, podcasts, newspapers. There are sites that aggregate movie reviews and provide averaged scores. One such site is Metacritic. Critics rating was used as a measure in papers such as [13, 18].

5. IMDb rating – IMDb allows registered users to rate all movies in their database on a scale of 1 to 10. The individual votes are aggregated and summarized as a single IMDb rating, which is displayed on the movie's main page. IMDb rating was used as a metric in [8].

2.2 Predicting movie success

There have been many approaches taken in order to predict movie success. Most of the works have been around predicting the movies' box-office success using regression models. Different regression models have been used in various papers such as [9, 12, 27, 28, 29, 34, 45, 46, 49]. For example, a stable distribution regression model was used in [46] while Multivariate Linear Regression was used in [34] on movies' Wikipedia activity. Many different regression analysis methods such as Lasso Regression, Support Vector Regression, Ridge Regression were used in [27]. In [20], two different supervised learning algorithms were used on data from IMDb, Rotten Tomatoes, and Wikipedia. The algorithms used were Locally weighted Linear Regression and Support Vector Machines. Similarly, Support Vector Machines were used in [15] on data from IMDb. In addition to regression analysis methods, decision tree methods such as CART, M5P trees, REP tree were used in [27]. Moreover, the J48 decision tree algorithm was used in [6]. Probabilistic classifiers such as Naive Bayes classifier have also been used in [6, 15]. Chi-squared distribution analysis was used to predict the success and failure of the upcoming movies by finding correlations between various attributes such as actors-vs-genres, genres-vs-ratings, ratings-vs-actors in [2]. Clustering has also been used in various research papers. In [49] k-nearest neighbours algorithm was used to group movies by similarity. The distance of the two movies was computed by normalizing the reference and sentiment counts. However, in [6] k-means clustering was used instead.

Social network analysis techniques, automatic sentiment analysis, and text mining have also been widely used to predict movie success. The most common data sources for that have been Twitter, IMDb, and YouTube. [9] performed sentiment analysis on 2.89 million Twitter tweets to predict movies' box-office results, [41] analyzed the effects of tweets on movie sales using sentiment analysis on 4.2 million tweets, and [47] analyzed 1.77 million tweets to find whether movie reviews on Twitter can predict the outcome of the movie. Similarly, social network analysis techniques and sentiment analysis have been used on IMDb comments in [6, 26] to predict movies' box-office performance. Additionally, [6] also performed sentiment analysis on YouTube comments. [27] used text mining to extract features from IMDb and Box Office Mojo in an automated fashion to predict movie success.

In addition to the aforementioned methods, methods such as neural networks in [38, 44, 50], Graph Network and transductive algorithms in [37], Markov Chains in [19], and parsimonious models in [43] have been used.

2.3 Identifying important features

Really important features for a movie's success were found to be genre and cast according to [2]. Similarly, [45] found that genre, cast, and maturity rating (The Motion Picture Association film rating) had the most prominent role, and [20] also got similar results. In addition to the aforementioned features, [20] found that running time, studio, and budget were also important.

[6] found that the popularity of leading actress and the combination of past successful genre and a sequel movie are patterns for success. They also found that new movies in the not popular genres and with low popularity actors could be a pattern for a flop. [27] found that the average profit of actor-director collaboration, director gross, winter release dates, total actor profit, and annual profit percentage by genre were key factors for determining success. They also found key factors for determining failure, which were movies rated "R", drama and foreign genres, and plot topics related to wars and music. Similarly, [8] found that the drama genre was the most significant factor with release year and director also playing a role. [26] concluded that high intensity of discussion about a movie is an indicator of success. Similarly, [49] found that news references highly correlate to grosses. [34] who analysed Wikipedia activity concluded that the combination of Wikipedia article page views, number of editors, number of edits, collaborative rigour, and number of theatres that screen the movie were the key factors for determining success. [29] found that familiarity with actors, characters, and story, as well as positive reviews from reviewers were the most important factors. [30] found a strong correlation between large-scale social media content and box-office revenue and [13] concluded that it is profitable to release a movie if it has favourably good critical reviews and effective marketing strategies.

3 Data description

In this chapter, the collection and available features of the data are described. Firstly, we explain different methods of data collection, how such methods were used, and the raw data itself (Section 3.1). Secondly, we describe how raw data from different sources was transformed into usable data (Section 3.2). Lastly, we give an overview of the collected data and how it is stored (Section 3.3). The flowchart of the process can be seen in Figure 1.

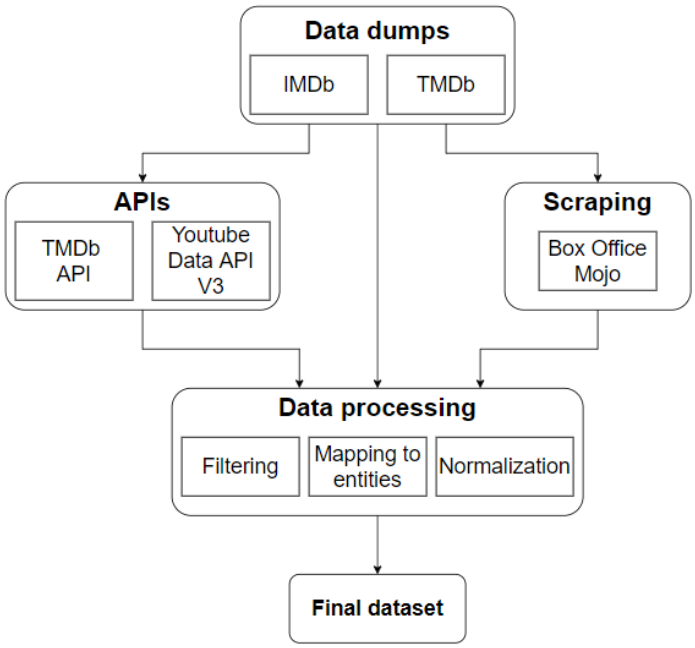

Figure 1: The data collection flowchart.

3.1 Collecting the data

We aimed at analysing a large dataset, which is missing in previous studies. However, there was no initial dataset available therefore data had to be gathered. To accomplish that different methods of data collection were used:

1. Data Dumps – Initial data was gathered from different data dumps. A data dump is the transfer of a large amount of raw data between systems. In a technical sense, it is a way to describe a dataset obtained from a system or an application. A very common way to dump data is through database dumps – that is exporting the data of the database usually in the form of SQL statements or CSV files. The biggest advantage of using data dumps is the simplicity of getting all the data – the whole dump can be downloaded fully and at once. The biggest disadvantages are that the data is often raw, unprocessed, might be out-of-date, and not contain all the needed features. Many websites publish their data for everyone to use and two such websites were used in the thesis, i.e. The Internet Movie Database (IMDb) and The Movie Database (TMDb).

2. Application Programming Interfaces (APIs) – A large amount of data was gathered from different APIs. An API is a server-side interface, which provides means of communication between the website and its users. Public APIs expose dedicated URLs, which accept requests and provide responses. The responses are stripped of presentational overhead and are well-structured – the data is often provided in a machine-readable format such as JSON, XML or CSV. The simplicity of both making such requests and processing the responses are the biggest advantages of using APIs. The biggest disadvantage, however, is that APIs are usually not designed for transferring truly large amounts of data. There are often rate limits in place by providers that stop users from overconsuming their resources. In addition to that, APIs might not provide all the available information but only a subset. In the thesis, two APIs were used, i.e. The Movie Database API and YouTube Data API v3.

3. Web Scraping – Some of the data, that was otherwise unavailable, was gathered by web scraping. Web scraping is used to extract data from websites. It is done by downloading the web page and extracting wanted information directly from it. It can be done both manually and automatically, but collecting large amounts of data manually is very impractical and error-prone. Automatic web scraping is implemented by web crawlers and bots. The biggest advantage of using web crawlers is that they have access to everything that an ordinary user sees on a page – that is, even if some of the information is omitted from the data dumps and API responses, it can still be gotten by scraping the web page. The biggest disadvantages are the relative complexity compared to using data dumps and APIs, and it might often be against the website's terms of service. Some of the web pages even use methods that prevent scraping by detecting and disallowing bots. During data collection, a crawler was built and used to scrape data from Box Office Mojo.

3.1.1 The Internet Movie Database data dumps

The Internet Movie Database is an online database of information related to movies, television and video games. They provide access to subsets of IMDb data for personal and non-commercial use. They publish 7 different datasets called title.akas, title.basics, title.crew, title.episode, title.principals, title.ratings, and name.basics. The datasets are gzipped and contain values in TSV format. Missing values are denoted with a '\N'. Each file contains a header that describes what the column contains. The files are updated daily and can be freely downloaded from `datasets.imdbws.com`. The available datasets contain the following information:

1. titles.akas – Contains localized information about movie and TV episode title, region, language, type (e.g., movie, short, video, etc.), and whether it is an original title or not. The file is almost 1 GB large and contains approximately 20.4 million rows. The localized information was not necessary for the thesis and thus this file was not used.

2. title.basics – Contains information about movie and TV episode primary title, original title, type (e.g. movie, short, video, etc.), start year, end year, length, genres, and whether the title is an adult title or not. The file is about 500 MB large and contains approximately 6.3 million rows. Almost all of the features in this file were needed for the research and thus this file was used. An example of the contents of the file can be seen in Table 1.

Table 1: An example of the contents of title.basics.

tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary,Short
tt0000009	movie	Miss Jerry	Miss Jerry	0	1894	\N	45	Romance
tt0793190	tvEpisode	Horse Trail	Horse Trail	0	1952	\N	\N	Comedy
tt5618938	movie	SWA-TECH	SWA-TECH	0	2018	\N	55	Sport

3. title.episode – Contains data about TV episodes including season and episode numbers and information about parent TV series if the series has one. The file is about 100 MB large and contains approximately 4.6 million rows. This file was not used because we were only interested in movies and not TV episodes.

4. title.crew – Contains references to directors and writers for all the titles in the IMDb. The file is about 200 MB large and contains approximately 6.5 million rows. Although the data in this file is useful for the thesis it was not used because title.principals also includes the same information.

5. title.principals – Contains the principal cast and crew for all the titles including unique identifiers of the persons involved, the order they were cast, the category of the job they were in, their job title, and characters played. The file is almost 1.6 GB large and contains about 36 million rows. Almost all of the features were necessary for the research and thus this file was used. An example of the contents of the file can be seen in Table 2.

Table 2: An example of the contents of title.principals.

tconst	ordering	nconst	category	job	characters
tt0000005	2	nm065304	actor	\N	["Assistant"]
tt0000007	3	nm000569	director	\N	\N
tt0000007	4	nm037465	director	\N	\N
tt0000001	1	nm158897	self	\N	["Herself"]
tt0000001	3	nm037465	cinematographer	director of photography	\N

6. title.ratings – Contains the average IMDb rating and the number of votes for the titles. The file is only about 16 MB large and contains approximately 1 million rows. The rating is useful for determining whether the movie was a success or not and thus was used in the thesis. An example of the contents of the file can be seen in Table 3.

Table 3: An example of the contents of title.ratings.

tconst	averageRating	numVotes
tt0000001	5.6	1543
tt0000002	6.1	186
tt0000003	6.5	1202
tt0000004	6.2	114

7. name.basics – Contains information about persons including their name, birth year, death year, primary profession, and titles they are known for. The file is about 500 MB large and contains approximately 9.6 million rows. The file was used in the thesis to connect the person's names to their identifiers in principals file. An example of the contents of the file can be seen in Table 4.

Table 4: An example of the contents of name.basics.

nconst	primaryName	birthYear	deathYear	primaryProfession	knownForTitles
nm0000001	Fred Astaire	1899	1987	soundtrack,actor,miscellaneous	tt0053137,tt0043044,tt0072308
nm0000002	Lauren Bacall	1924	2014	actress,soundtrack	tt0038355,tt0037382,tt0071877
nm0000003	Brigitte Bardot	1934	\N	actress,soundtrack,producer	tt0049189,tt0057345,tt0059956
nm0000004	John Belushi	1959	1982	actor,writer,soundtrack	tt0078723,tt0077975,tt0080455

3.1.2 The Movie Database data dumps

The Movie Database is a community built movie and TV database. They publish ID file exports, which are not meant to be full data exports. Instead, the files contain valid IDs that can be found on TMDb and some higher-level attributes that help to filter items. Each line in the file is formatted as a valid JSON object. The files are updated daily and are currently available for download without any authentication. They publish 7 files – movie_ids, tv_series_ids, person_ids, collection_ids, tv_network_ids, keyword_ids and production_company_ids. The files are available for download at `files.tmdb.org`.

In the thesis, we were only interested in the files containing movies and people. The movie IDs file is about 46 MB large and contains approximately 470,000 rows. An example of the contents of the file can be seen in Figure 2.


```
{"adult":false,"id":9,"original_title":"Sonntag im August","popularity":2.152,"video":false}
{"adult":false,"id":11,"original_title":"Star Wars","popularity":63.762,"video":false}
{"adult":false,"id":12,"original_title":"Finding Nemo","popularity":37.577,"video":false}
{"adult":false,"id":13,"original_title":"Forrest Gump","popularity":36.838,"video":false}
{"adult":false,"id":14,"original_title":"American Beauty","popularity":33.406,"video":false}
{"adult":false,"id":15,"original_title":"Citizen Kane","popularity":22.48,"video":false}
{"adult":false,"id":16,"original_title":"Dancer in the Dark","popularity":17.996,"video":false}
{"adult":false,"id":17,"original_title":"The Dark","popularity":7.159,"video":false}
{"adult":false,"id":18,"original_title":"The Fifth Element","popularity":45.747,"video":false}
{"adult":false,"id":19,"original_title":"Metropolis","popularity":19.116,"video":false}
```


Figure 2: An example of the contents of movie_ids.

The person's IDs file is about 100 MB large and contains approximately 1.46 million rows. An example of the contents of the file can be seen in Figure 3.


```
{"adult":false,"id":3129,"name":"Tim Roth","popularity":5.99}
{"adult":false,"id":62,"name":"Bruce Willis","popularity":16.454}
{"adult":false,"id":37336,"name":"Kimberly Blair","popularity":0.6}
{"adult":false,"id":2042,"name":"Stephen Hopkins","popularity":3.585}
{"adult":false,"id":52035,"name":"Lewis Colick","popularity":0.652}
{"adult":false,"id":71417,"name":"Gene Levy","popularity":0.6}
{"adult":false,"id":37,"name":"Alan Silvestri","popularity":2.561}
{"adult":false,"id":2044,"name":"Peter Levy","popularity":0.6}
{"adult":false,"id":2880,"name":"Emilio Estevez","popularity":1.985}
{"adult":false,"id":9777,"name":"Cuba Gooding Jr.","popularity":3.303}
```


Figure 3: An example of the contents of person_ids.

3.1.3 The Movie Database API

The Movie Database API is an API that provides access to TMDb data. They provide endpoints exposing data about movies, TV episodes, actors, and images. To query the API an API key is needed, which can be requested freely by signed-in users. The registration itself is also free. The API used to be rate-limited – at the time of data collection in November 2019 the limit was up to 4 requests per second meaning one could make up to 345,600 queries per day, but as of December 16, 2019, the rate limit has been removed. The API only supports JSON format, but it is possible to use the callback parameter to encapsulate the JSON response in a JavaScript function (JSONP). The API provides the possibility to make subrequests within the same namespace in a single HTTP request using the append_to_response query parameter. Using subrequests cuts down the number of queries considerably – for example, to query 500,000 movies and their keywords, reviews, credits, trailers, and release dates it would have taken 3,000,000 requests without using subqueries but only 500,000 requests using them. Each subrequest result gets appended to the response as a new JSON object.

Two of the provided endpoints were used during data collection – get movie and person details. The movie details provided include title, language, overview, budget, revenue, popularity, length, release year, genres, production companies, production countries, ratings, collection ID, maturity rating, tagline, status, posters, external IDs. In addition to movie details, the information about keywords, reviews, credits, videos, and release dates was also requested with movie details query using subqueries. For that, the following URL template was used `api.themoviedb.org/3/movie/{id}?api_key={key}&append_to_response=keywords,reviews,credits,videos,release_dates`, where *id* is the ID of the movie and *key* is the API key. An example of the response can be seen in Figure 4.

The person's details provided include name, birthday, death day, gender, biography, popularity, place of birth, what the person is known for, aliases, external IDs. In addition to the aforementioned details, the data about movie credits was also requested using subqueries by

```
{
  "adult": false,
  "backdrop_path": "/kpuTCMw3v2AuKjqGS7383uWbc8V.jpg",
  "belongs_to_collection": null,
  "budget": 0,
  "genres": [{"id": 18, "name": "Drama"},{"id": 80, "name": "Crime"}],
  "homepage": "",
  "id": 2,
  "imdb_id": "tt0094675",
  "original_language": "fi",
  "original_title": "Ariel",
  "overview": "Taisto is a coal miner who is framed for a crime he did not commit",
  "popularity": 9.342,
  "poster_path": "/ojDg0PGvs6R9xYFodRct2kdI6wC.jpg",
  "production_companies": [
    {"id":2303, "name": "Villealfa Filmproductions","origin_country": "FI"}
  ],
  "production_countries": [{"iso_3166_1": "FI","name": "Finland"}],
  "release_date": "1988-10-21",
  "revenue": 0,
  "runtime": 73,
  "spoken_languages": [{"iso_639_1": "fi", "name": "suomi"}],
  "status": "Released",
  "tagline": "",
  "title": "Ariel",
  "video": false,
  "vote_average": 6.8,
  "vote_count": 100,
  "keywords": {"keywords": [
    {"id": 240, "name": "underdog"},
    {"id": 378, "name": "prison"}
  ]},
  "reviews": {"page": 1, "results": [], "total_pages": 0, "total_results": 0},
  "videos": {"results": []},
  "release_dates": {"results": [
  {"iso_3166_1": "GB", "release_dates": [
    {"certification": "", "iso_639_1": "", "note": "", "type": 3,
       "release_date": "1989-06-30T00:00:00.000Z" }
  ]}]},
  "credits": {
    "cast": [{
        "cast_id": 3, "credit_id": "52fe420dc3a36847f8000029",
        "character": "Taisto Olavi Kasurinen", "name": "Turo Pajala",
        "gender": 2, "id": 54768, "order": 0
        }],
    "crew": [{
        "id": 1683985, ,"credit_id": "5e1cdc3312b10e0016942be4",
        "name": "Kjell Westman", "gender": 0,
        "department": "Sound", "job": "Sound Mixer"
        }]
}
```


Figure 4: An example of the response from TMDb API for movie details query.

appending movie_credits to the response. To do that, the following URL template was used `api.themoviedb.org/3/person/{id}?api_key={key}&append_to_response=movie_credits`, where *id* is the ID of the person and *key* is the API key. An example of the response can be seen in Figure 5.

```
{
  "birthday": null,
  "known_for_department": "Art",
  "deathday": null,
  "id": 11700,
  "name": "Keith Neely",
  "movie_credits": {"cast": [
    {
      "id": 673124, "credit_id": "5e4821061e92250015c0578d"
      "character": "",
      "original_language": "en",
      "original_title": "Shooting 'Panic Room'", "title": "Shooting 'Panic Room'",
      "overview": "Watch the filming process for Panic Room (2002) in this documentary.",
      "genre_ids": [99],
      "release_date": "2004-03-30",
      "popularity": 0.6, "vote_average": 0, "vote_count": 0,
      "adult": false, "video": true,
      "backdrop_path": null,
      "poster_path": "/fRZcZUfslVnKyHBgacRwKyDN6wz.jpg",
      }
    ],
    "crew": [
      {
        "id": 9923, "credit_id": "52fe454ac3a36847f80c5e33",
        "department": "Art", "job": "Art Direction",
        "original_language": "en",
        "original_title": "Domino", "title": "Domino",
        "overview": "The story of the life of Domino Harvey",
        "genre_ids": [28, 80],
        "release_date": "2005-10-14",
        "popularity": 13.404, "vote_average": 6, "vote_count": 803,
        "video": false, "adult": false,
        "backdrop_path": "/6Epyr9UJV4J1ptql7IyV6ni0N0h.jpg",
        "poster_path": "/lbLEiLoRycZSKRhjuZ4Zy8KQsct.jpg"
      }
    ]
  },
  "also_known_as": [],
  "gender": 0,
  "biography": "",
  "popularity": 0.6,
  "place_of_birth": null,
  "profile_path": null,
  "adult": false,
  "imdb_id": "nm0624213",
  "homepage": null
}
```


Figure 5: An example of the response from TMDb API for person details query.

3.1.4 YouTube Data API v3

The YouTube Data API v3 is an API that provides access to YouTube data. In total, the API provides access to 19 different resources, including channels, playlists, videos, comments. The API also supports methods to query, insert, update, or delete most of these resources. To get access to the API an API key or OAuth 2.0 token is required. The API key can be gotten from Google's Developer console at `console.developers.google.com`. The API is rate limited – the API key has a quota of 10,000 units that can be used per day. Depending on the resource and actions performed a different number of units is used up per query. In the thesis, we were interested in getting movie trailer statistics including the count of views, likes, dislikes, and comments. For that a GET query was performed on video resource querying only the statistics part

using the following URL template `content.googleapis.com/youtube/v3/videos?id={id}&key={key}&part=statistics`, where *id* is the ID of the video and *key* is the API key. Such query used up 3 units of the quota and thus 3,333 trailers could be queried per day. Over the period of a month, over 66,000 trailers were queried and results inserted into the database. The API's response is in JSON format and an example of the response querying statistics about The Avengers trailer can be seen in Figure 6.


```
{
 "kind": "youtube#videoListResponse",
 "etag": "\"Dn5xIderbhAnUk5TAW0qkFFir0M/KtSF7ALMhVIc41nBzj-x2fBbIo4\"",
 "pageInfo": {
  "totalResults": 1,
  "resultsPerPage": 1
 },
 "items": [
  {
   "kind": "youtube#video",
   "etag": "\"Dn5xIderbhAnUk5TAW0qkFFir0M/GNvcyWMWe0ZSGUL-vUfXVFpEIZc\"",
   "id": "eOrNdBpGMv8",
   "statistics": {
    "viewCount": "33340653",
    "likeCount": "214424",
    "dislikeCount": "4500",
    "favoriteCount": "0",
    "commentCount": "39430"
   }
  }
 ]
}
```


Figure 6: An example of the response from YouTube Data API V3.

3.1.5 Box Office Mojo scraper

Box Office Mojo is a website that tracks box office revenue. They track gross both domestically and internationally. In addition to that, they also provide cumulative worldwide gross. Box office data is provided separately for daily, weekly, monthly, quarterly, yearly, seasonally, and for weekends and holidays. They have data dating back to 1977. Box Office Mojo is owned by IMDb. In the thesis, we were interested in a movie's worldwide box office. The data is available at `www.boxofficemojo.com/year/world/{year}`, where *year* is substituted by the desired year. To get the necessary data, a Python script was created that queried the given URL with year values ranging from 1977 to 2020 and used BeautifulSoup to parse the resulting HTML. That allowed us to get movie titles and worldwide gross. The result was then saved as a TSV file. The resulting file is about 1 MB large and contains approximately 17,000 rows. An example of the resulting file can be seen in Table 5.

Table 5: An example of the contents of the data from Box Office Mojo.

id	title	worldwide	year
1	Saturday Night Fever	237,113,184	1977
2	Smokey and the Bandit	126,737,428	1977
3	A Bridge Too Far	50,750,000	1977

3.2 Data processing

Data processing is the process of manipulating raw data to produce useful information. It involves data entry, organization, modification, and storage. It is done to produce meaningful and informative results. Processed data is often in the form of tables, diagrams, and reports. A Java Spring Boot application was built to process raw data from different sources and insert processed output into a PostgreSQL database. The API of the built application can be seen in Figure 7.

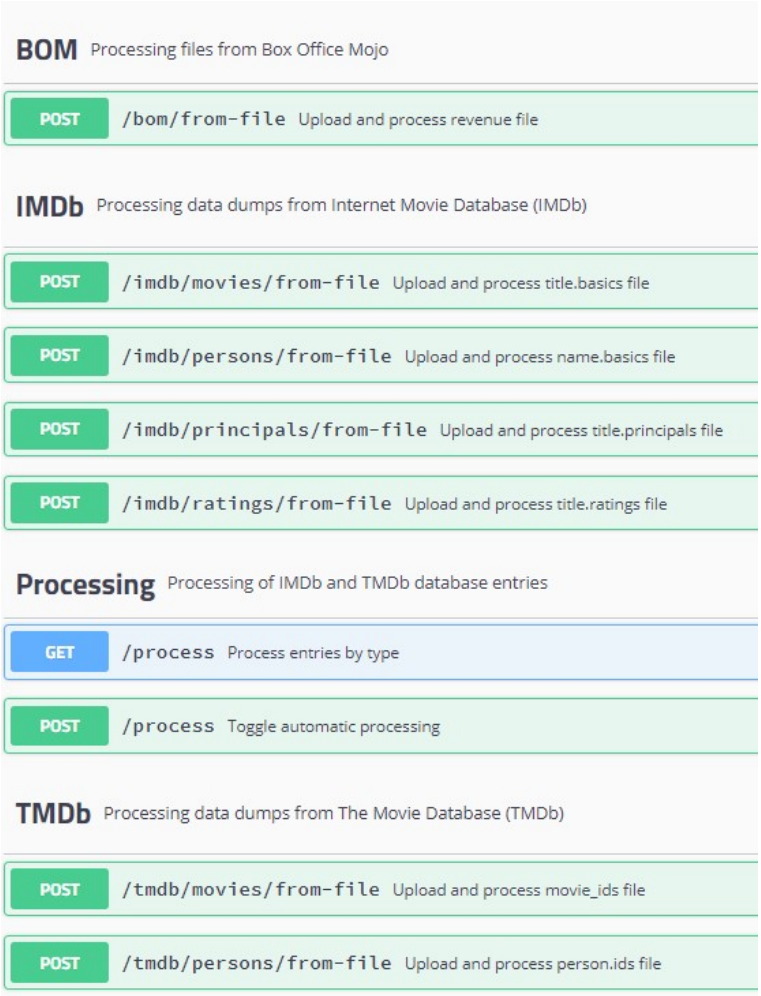

Figure 7: The API of the built application.

The processing of IMDb, TMDb, and Box Office Mojo raw data included mapping all the rows to entities, filtering out unwanted entities, and then inserting the remaining entities into the database tables. No filtering was necessary for IMDb person's and rating's file, for titles file the entities, which were not marked as "movie" type or were marked as "adult" were not kept and for principals file the rows, which contained information about titles that were not movies were filtered out. For both TMDb movies and persons file the entities that were marked as "adult" were filtered out. The Box Office Mojo data did not need any filtering and thus was simply mapped to entities and inserted into the corresponding table.

The data that was inserted into the database was then further processed to add more necessary features. Firstly, the TMDb API was queried with all of the identifiers from the movie_ids and person_ids file. The results were then mapped into entities and combined with the data from IMDb data dumps by comparing the IMDb external ID that the TMDb API returned to the ID of the IMDb entities. Secondly, YouTube Data API V3 was queried with the IDs of the trailers gotten from TMDb API and results saved to the corresponding table to get information about how popular and hyped the movie trailers were. Finally, the collected data from Box Office Mojo was added to movies by comparing normalized movie titles and release years. The normalization involved 4 steps:

1. Converting the titles to lowercase.
2. Only keeping letters from the English alphabet, numbers, and whitespaces.
3. Removing stopwords "the", "a" and "an".
4. Removing all whitespace.

By doing that, 13,011 of 17,286 scraped movies were mapped to existing entities. An example of the titles considered the same can be seen in Table 6.

Table 6: An example of the titles that were matched.

TMDb title	**scraped BOM title**	**year**
A Soldier's Story	A Soldiers Story	1984
The Underground	Underground	1997
Alléluia	Allluia	2014
Bat*21	Bat21	1988
(500) Days of Summer	500 Days of Summer	2009
The Conjuring	The Conjuring	2013

3.3 Final dataset

The final dataset contains 470,743 movies and has 26 features. The available features can be seen in Table 7. A single movie can only have a single feature value for features 1-15, but

multiple values for features 16-26.

Table 7: The available features.

1. title	8. maturity rating	15. BOM gross	22. number of trailer views
2. language	9. length	16. keywords	23. number of trailer likes
3. overview	10. release year	17. genres	24. number of trailer dislikes
4. budget	11. IMDb rating	18. cast	25. number of trailer comments
5. revenue	12. IMDb number of votes	19. crew	26. reviews from TMDb
6. popularity	13. TMDb rating	20. production companies	
7. collection ID	14. TMDb number of votes	21. production countries	

All the collected data is held in a relational PostgreSQL database which diagram can be seen in Figure 8. In total, the size of the database is about 2.1 GB.

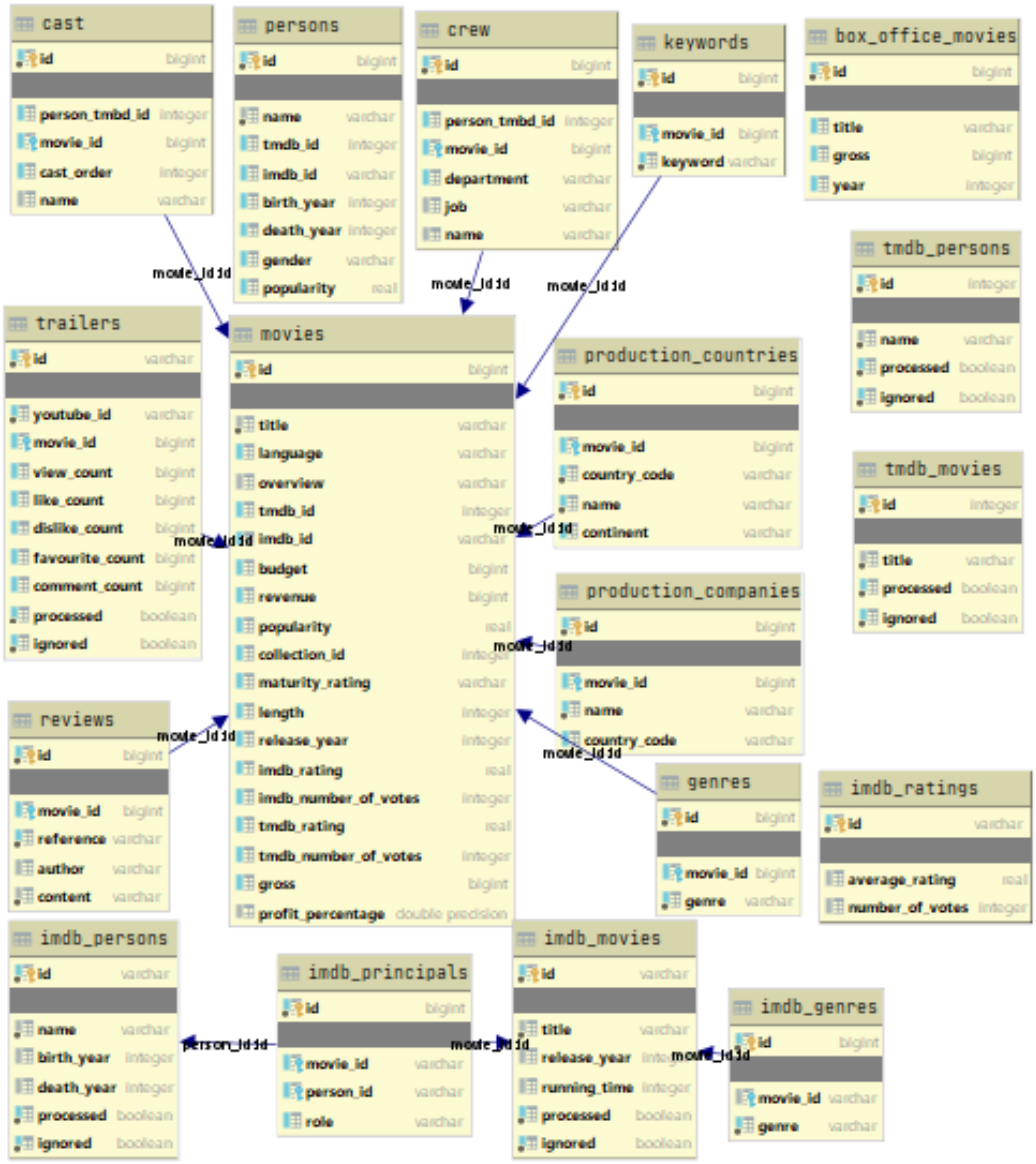

Figure 8: Database diagram.

4 Methodology

In this chapter, the data preparation steps and approaches to data analysis are described. Firstly, we give an overview of descriptive analysis (Section 4.1). Secondly, we describe how we performed data cleaning and feature engineering (Section 4.2 and Section 4.3). Thirdly, we explore different ways to define hit and flop movies (Section 4.4). Finally, we give an overview of how we performed association rule mining and extracted the results (Section 4.5). The overview of the methodology can be seen in the flowchart in Figure 9.

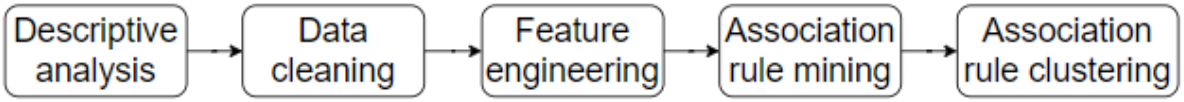

Figure 9: The Methodology flowchart.

4.1 Descriptive analysis

Descriptive data analysis is performed to describe the features of the data. It is done to understand the data better and to gather hidden insights from the data. It allows us to spot mistakes and missing data, anomalies and outliers, and to test hypotheses and check assumptions.

At first, we turned the attention to missing values. From Figure 10 it can be seen how many of the movies have values filled for the features and how much data is missing. It can be seen

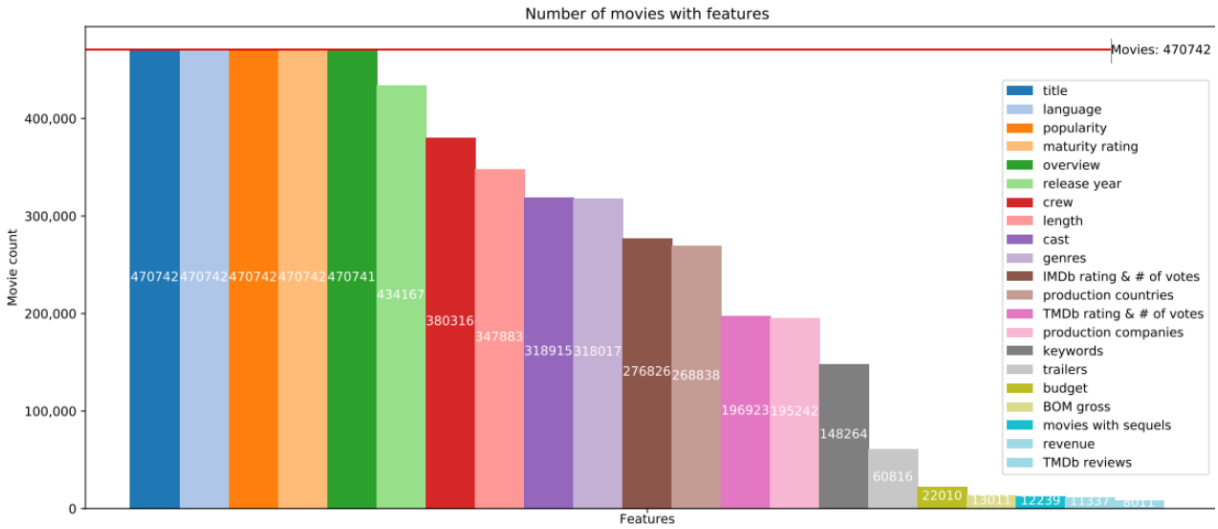

Figure 10: Number of movies with features.

that all of the movies have the title, language, maturity rating and popularity filled and only one of the movies is missing an overview. More than two-thirds of the movies also have release year, length, genres, crew, and cast information. About half of the movies have information about production countries, production companies, IMDb and TMDb ratings with their number

of votes. About third of the movies have data about keywords and one eight of the movies have data about trailers. There is very little data about budget, revenue, and reviews – less than 5% of the movies have that information filled. It is expected that a small number of movies belong to a franchise and thus have the collection ID filled, but it is unclear whether missing value means it does not belong to a franchise or it is left unfilled.

After that, the descriptive statistics of the dataset were checked. For that, the features were divided into two – numerical and categorical variables. For numerical variables, the minimum value, maximum value, quartiles, mean and standard deviation were checked. Those statistics for budget and revenue can be seen in Table 8. It can be seen that there are outliers and most likely incorrect values there that should be fixed before performing further analysis.

Table 8: The budget and revenue statistics.

	Budget	**Revenue**
count	22,010	17,903
mean	13,032,381	90,138,091
std	202,336,345	6,545,507,991
min	1	−12
25%	10,000	190,624
50%	761,550	3,167,681
75%	8,000,000	22,930,349
max	25,414,000,000	874,000,000,000

For categorical variables, the distributions and unique values were checked to spot any mistakes. Checking distributions also showed whether data was unbalanced or not. The distribution of movies by release year can be seen in Figure 11. The distribution is left-skewed which is ex-

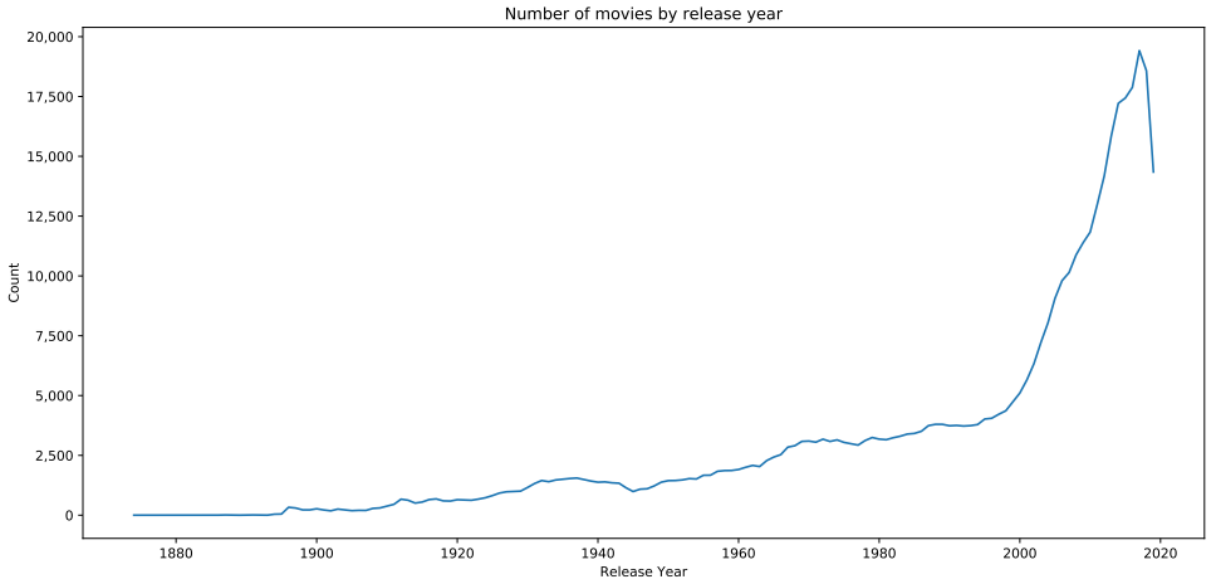

Figure 11: The number of movies in our dataset per year.

pected because nowadays there are way more movies than there were in the past. The drop at the end is due to the data collection taking place at the beginning of November 2019 and is thus missing movies from the last months of the year 2019.

From the genre distribution in Figure 12, it can be seen that nearly a quarter of the movies in the dataset are tagged as "drama" and about 15% of the movies are tagged as "comedy" and "documentary". Less than 1% of the movies are tagged as biography, musical, sci-fi, sports, and news.

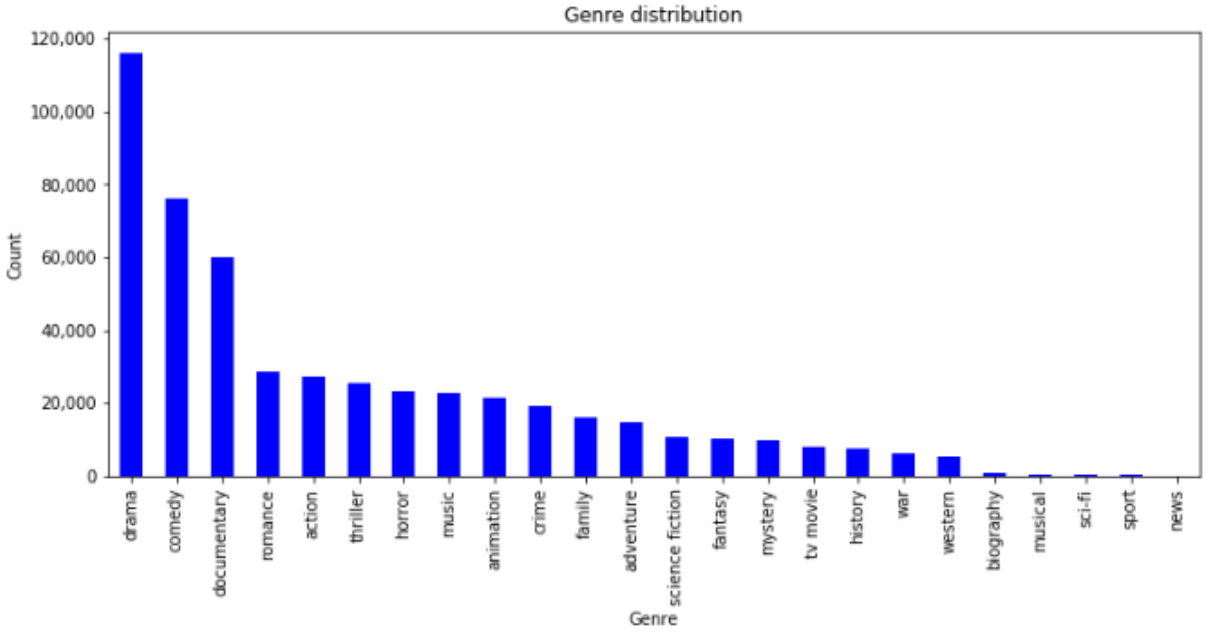

Figure 12: The genre distribution.

It should also be taken into consideration that one movie could be tagged with multiple genres. As can be seen in Table 9, most of the movies are tagged with only one or two genres, but there are a few outliers that have more than 7 tagged genres. Having a movie tagged with

Table 9: The number of genres on movies.

Genres on movie	**Number of movies**
1	186,088
2	84,123
3	36,448
4	9,022
5	1,898
6	383
7	49
8	3
9	3

that many different genres could mean that there has been an error in the tagging or a movie does indeed belong in all of them.

The distribution of production countries in Table 10 shows that about one-sixth of the movies in the dataset have been produced in the United States of America. Approximately 5% of movies have been produced in Germany, France, United Kingdom. In total, there were 240 different countries where movies had been produced. Out of those 240, there were 44 countries where more than 1000 movies had been produced and 30 countries where less than 5 movies had been produced.

Table 10: The production country distribution.

#	Country	Number of movies
1	United States of America	81,538
2	Germany	25,734
3	France	20,836
4	United Kingdom	20,235
5	Japan	14,228
...	...	...
236	St. Helena	1
237	Naunu	1
238	Tokelau	1
239	Svalbard & Jan Mayen Islands	1
240	Heard and McDonald Islands	1

Similarly to genres, one movie could have been produced in multiple countries. The number of production countries the movies have been produced in can be seen in Table 11. It is expected that most of the movies are produced in only a few countries, but there are a few outliers of movies that have been produced in more than 10 different countries.

Table 11: The number of production countries on movies.

Number of Production countries	1	2	3	4	5	6	7	8	9	10	11	12	14	15	19	23	25	26	30
Movie count	242,079	20,524	4,423	1,202	402	119	40	16	8	5	8	3	1	1	1	1	2	2	1

The movies in the dataset were produced by 81,500 different production companies. Most of the production companies have produced a very small amount of movies as can be seen in Table 12. There are a few major film studios like Warner Bros. Pictures, Columbia Pictures, Metro-Goldwyn-Mayer, Universal Pictures, Paramount that have produced over 2,000 movies each. The average amount of movies produced per company is 4 while the median is only 1.

Table 12: Movies produced per production company.

	Movies produced per company
count	81,500
mean	4
std	27
min	1
50%	1
99%	35
99.9%	225
max	2,867

In Table 13 it can be seen that there are movies in 155 different languages in the dataset, but a bit over half of them are in English. Most of the languages only appear a few times. It does not make sense to drop those languages and instead they can be grouped.

Table 13: The language distribution.

#	**Language**	**Count**
1	en	272,591
2	de	29,254
3	fr	28,023
4	es	20,728
5	ja	14,424
....	...	...
154	sd	1
155	ik	1

The Motion Picture Association (MPA) film rating system is used to rate the suitability of a movie for certain audiences based on the movie's contents. It is mostly used in the United States and its territories, but non-members of MPA can also submit their films for rating. The available film ratings are taken from [35] and are as follows:

1. G – General Audiences. All ages admitted, nothing that would offend parents for viewing by children.

2. PG – Parental Guidance Suggested. Some material may not be suitable for children, parents urged to give "parental guidance". May contain some material parents might not like for their young children.

3. PG-13 – Parents Strongly Cautioned. Some material may be inappropriate for children under 13, parents are urged to be cautious. Some material may be inappropriate for pre-teenagers.

4. R – Restricted. Under 17 requires accompanying parent or adult guarding, contains some adult material. Parents are urged to learn more about the film before taking their young children with them.

5. NC-17 – Adults Only. No one 17 and under admitted.

Looking at the distribution of maturity ratings in Table 14, it can be seen that most of the movies are rated as 'NR' which stands for 'Not rated' and only about 5% of the movies have a real classification. The MPA rating system is a voluntary scheme that is not enforced by law and thus movies can be exhibited without a rating, which explains why so many are not rated.

Table 14: The maturity rating distribution.

Maturity rating	**Number of movies**
NR	445,394
R	10,769
PG-13	5,072
PG	5,035
G	4,144
NC-17	328

The statistics for movie length in Table 15 show, that on average a movie is 72 minutes long with the median being 82 minutes. However, the minimum and maximum values seem to indicate there may be an error in the data, but upon further inspection, it turned out that those values are correct – in 2012 there was an experimental art film, which recorded the production cycle of a pedometer in real-time and the journey lasted for 37 days and nights, nonstop. Movies with a really small length are just short movies – there are even international one-minute movie festivals held for them.

Table 15: Movie length statistics.

	Movie length
count	347,883
mean	72
std	130
min	1
25%	29
50%	82
75%	96
max	51,420

By comparing the IMDb and TMDb ratings, it can be seen that both the mean and the median are higher for the IMDb ratings as seen in Figure 16. However, one thing to keep in

mind is that in the case of TMDb ratings the minimum a movie can be rated as is 0 while for IMDb the minimum is 1. Looking at the number of votes, it can be seen that IMDb has a lot more data – that is expected since IMDb is a lot more popular as opposed to TMDb. In the case of TMDb over half of the movies have two or fewer votes so its rating might not be that useful.

Table 16: The ratings statistics.

	IMDb rating	**IMDb number of votes**	**TMDb rating**	**TMDb number of votes**
count	276,826	276,826	196,923	196,923
mean	6.24	2,885.25	6.06	56.34
std	1.29	28,595.90	1.77	470.56
min	1.00	5	0.00	1
25%	5.50	19	5.00	1
50%	6.40	63	6.00	2
75%	7.10	283	7.00	8
max	10.00	2,150,648	10.00	23,657

There are over 2 million records available about the movie crew. The crew is divided into 12 departments and the number and percentage of movies that have those features filled can be seen in Table 17. About four-fifths of the movies have information about the director, about half have information about the writer and a quarter have information about the producer and sound mixer. Less than 10% of the movies have information about the crew in costume & make-up, lighting, and visual effects departments. Actors should not be in this table and should be deleted when cleaning the data.

Table 17: The crew departments.

Department	**Number of movies**	**% of movies (rounded)**
Directing	374,136	79%
Writing	211,056	45%
Production	118,372	25%
Sound	103,207	22%
Editing	92,986	20%
Camera	86,211	18%
Misc	58,027	12%
Art	53,836	11%
Costume & Make-Up	36,060	8%
Visual Effects	12,683	3%
Lighting	9,598	2%
Actors	21	0%

From the statistics about trailers in Table 18, it can be seen that while most of the trailers have a very low view, like, dislike and comment counts, the more popular ones bring up the average considerably. For likes, dislikes, and comments the average is even about 50 times higher than the median. It can be also seen that on average the trailers have 13.5 likes to 1 dislike.

Table 18: The trailer statistics.

	View count	**Like count**	**Dislike count**	**Comment count**
count	65,255	60,178	60,178	60,366
mean	696,435	4,524	335	428
std	4,526,587	47,691	5,899	5,221
min	1	0	0	0
25%	7,034	19	1	2
50%	36,596	79	6	9
75%	191,614	428	34	54
max	622,858,210	4,344,297	1,138,550	659,290

4.2 Data cleaning

Data cleaning is the process of detecting and correcting inaccurate or corrupt data from the dataset. After doing descriptive analysis four different types of errors were detected:

1. Typographical errors.
2. Invalid values.
3. Missing values.
4. Excessive or redundant values.

Most of the typographical errors were fixed manually by using SQL queries. These included fixing typos and casing in genres, maturity ratings, and keywords. For outliers and invalid values, the values were double-checked and corrected manually where possible, and if a correction was not possible then the whole row was deleted. This included removing movies that had release year set in the future or either impossibly high or low budget and revenue values. In the case of missing values deleting the whole row was undesirable and thus those rows were kept, but filtered out when deemed necessary. In the case of excessive and redundant values, the values were simply deleted.

4.3 Feature engineering

Feature engineering is the process of inventing or discovering new features from raw data using domain knowledge. It is used to prepare a proper input dataset that is compatible with machine learning algorithms and to improve the performance of machine learning models.

In addition to the initial features described in Chapter 3.3, the following additional features were created:

1. Language_engineered – Similar to the original language feature, but with grouping all non-English languages into one "Foreign" language.

2. In_collection – Feature showing whether a movie belonged to a collection or not. The value was set to true if the value of original feature collection_id was not null and set to false if it was null.

3. Length_engineered – Feature that was created by dividing the original length feature into 5 groups. The groups were created by looking at the percentiles of the original lengths and are as follows:

    - Very short – Original length less than 5 minutes (0.05 percentile).
    - Short – Original length between 5 (0.05 percentile) and 29 minutes (0.25 percentile).
    - Medium – Original length between 29 (0.25 percentile) and 96 minutes (0.75 percentile).
    - Long – Original length between 96 (0.75 percentile) and 135 minutes (0.95 percentile).
    - Very long – Original length more than 135 minutes (0.95 percentile).

4. Continent – The continent the movie was produced in. Created by mapping all production countries to their continents using the data from [25].

5. Person_and_department – Feature created by combining the person ID and the department they were in.

6. Trailer_views_group – Feature that divides the number of trailer views into groups. The following 5 groups were created by taking [31] as the basis:

    - Very popular – Amount of views higher than 1 million.
    - Popular – Amount of views between 100,000 and 1 million.
    - Average – Amount of views between 10,000 and 100,000.
    - Unpopular – Amount of views between 100 and 10,000.
    - Very unpopular – Amount of views less than 100.

7. Trailer_comments_to_views_ratio_group – Feature that divides the trailer's ratio of comments-to-views into groups. The following 5 groups were created by taking [40] as the basis:
    - Superb – The ratio of comments-to-views higher than 0.005.
    - Great – The ratio of comments-to-views between 0.001 and 0.005.
    - Average – The ratio of comments-to-views between 0.000,1 and 0.001.
    - Bad – The ratio of comments-to-views between 0.000,01 to 0.000,1.
    - Very bad – The ratio of comments-to-views less than 0.000,01.
8. Trailer_likes_to_views_ratio_group – Feature that divides the trailer's ratio of likes-to-views into groups. The following 5 groups were created by taking [40] as the basis:
    - Superb – The ratio of likes-to-views higher than 0.004.
    - Great – The ratio of likes-to-views between 0.001 and 0.004.
    - Average – The ratio of likes-to-views between 0.000,1 and 0.001.
    - Bad – The ratio of likes-to-views between 0.000,01 and 0.000,1.
    - Very bad – The ratio of likes-to-views less than 0.000,01.
9. Trailer_likes_to_dislikes_ratio_group – Feature that divides the trailer's ratio of likes-to-dislikes into groups. The following 5 groups were created by taking [32] as the basis:
    - Superb – The ratio of likes-to-dislikes higher than 30 or if a movie had more than 50 likes and no dislikes.
    - Great – The ratio of likes-to-dislikes between 20 and 30 or if a movie had less than 50 likes and no dislikes.
    - Average – The ratio of likes-to-dislikes between 10 and 20.
    - Bad – The ratio of likes-to-dislikes between 1 and 10.
    - Very bad – The ratio of likes-to-dislikes less than 1.
10. Revenue_engineered – Feature that coalesced the original revenue and gross features.
11. Income – Feature that was created by subtracting the budget from engineered revenue.
12. Result – Feature that classified the movie into successes and flops according to the definitions in Chapter 4.4.

4.4 Defining popularity – hit and flop movies

There is no one definition of whether a movie was a hit or a flop. To classify movies two different angles were considered – income and rating.

4.4.1 By income

According to [5] a movie needs to make at least double its budget to break even. That is because there are hidden costs that are usually not added to the budget like advertising. Based on that, a movie was considered successful if the revenue was at least triple the budget. The movie was considered a flop if the budget exceeded the revenue. However, in the dataset there were almost 10,000 movies that did not have information about the budget, but did have information about revenue. In those cases the movie was considered a success if it made more than 3rd quartile of such movies, which equated to $6.16 million and considered as a flop if it made less than 1st quartile, which equated to $77,000. For the remaining cases the movies were classified as neutral (AVG). The distribution of the classification can be seen in Table 19.

Table 19: The distribution of movies classified by income.

Classification	Count
HIT	4,181
FLOP	4,019
AVG	7,955

4.4.2 By rating

Looking at the IMDb ratings the movies were classified by quartiles – if the movie had a rating less than the 1st quartile, which was a rating of 5.5 then it was considered a flop and if the movie had a rating higher than 3rd quartile, which was a rating of 7.1 then it was considered a success. For the remaining cases – a rating between 1st and 3rd quartiles – the movies were classified as neutral (AVG). The distribution of the classification can be seen in Table 20.

Table 20: The distribution of movies classified by rating.

Classification	Count
HIT	65,347
FLOP	72,710
AVG	138,743

4.5 Association rule mining

Association rule mining is a machine learning method for discovering patterns, correlations, and associations in the data. It is used to find features, which occur together. Association rule is defined as an implication of the form antecedent -> consequent. The effectiveness of an association rule can be measured by different measures. The most well-known measures are:

1. Support – Indication of how frequently the itemset appears in the dataset.

2. Confidence – Indication of how often the rule is true.

3. Lift – The ratio of the observed support to that expected if the antecedent and the consequent were independent.

To find feature combinations of successful and flop movies it was important to find association rules with high confidence and lift. Rules with such characteristics were important even if they had low support because a certain combination of features could only occur a few times, but made the movie always a success. In addition to that, the dimensionality of the features is very large due to there being hundreds of thousands of actors, production companies, etc. Due to that, an algorithm that performed well with low support constraint and high-dimensional datasets had to be used.

There are many algorithms for finding association rules. A very widely used algorithm is the Apriori algorithm that was proposed in 1994 by R. Agrawal and R. Srikant in [1], which uses the breadth-first search strategy and candidate generation function, which uses the downward close property of support. Over time there have been many improvements suggested to the algorithm such as more efficient filtering in [14] or reducing the time consumed in transactions scanning for the generation of candidate itemsets by reducing the number of transactions to be scanned in [4]. There are other algorithms such as the Eclat, which is a depth-first search algorithm proposed in [48] and the FP-Growth algorithm, which builds an FP-tree structure by inserting transactions into a trie as proposed in [24]. Research has also been conducted on mining frequent itemsets for Big Data using Hadoop- and Spark-based scalable algorithms in [7]. In addition to the aforementioned, many algorithms have been proposed that work well on high-dimensional datasets such as the algorithms proposed in [3] and algorithms that work well without minimum support such as the algorithms proposed in [11, 42].

To perform association rule mining an implementation of the Apriori algorithm that used some improvements (e.g., a prefix tree and item sorting) was used. The process of association rule mining was as follows:

1. Querying the data from the database. To keep the dimensionality a little bit lower the engineered values that were described in Chapter 4.3 were used instead of their original values, which helped by dividing similar values into groups.

2. One-hot encoding the data. This was needed to make the dataset suitable for the algorithm.

3. Creating a pivot table using the movie ID as index and max as the aggregator function. This was used to combine rows of the same movie.

4. Running the algorithm on the created dataset to get the frequent itemsets.

5. Generating association rules from the frequent itemsets.

Finally, the generated association rules were grouped by lift using k-means clustering as described in [23].

5 Results

In this chapter, an overview of the results of the data analysis is given. Firstly, we give an overview of the characteristics of successful movies (Section 5.1). The results are explored for both definitions of success described in Chapter 4.4. Secondly, we give an overview of the characteristics of unsuccessful movies (Section 5.2). Similarly, the results are explored for both definitions of flops described in Chapter 4.4. Finally, we take a look at the similarities and differences between the successful and unsuccessful movies by comparing the results (Section 5.3).

5.1 Characteristics of successful movies

5.1.1 By income

To be able to tell whether a movie was successful when considering its income only rows that had information about the revenue available were kept. In total there were 16,034 such movies in the dataset and they had 222,090 unique feature values. On an average, a movie had 45 feature values, with the median being 33, minimum 4 and maximum 993. Movies with many feature values usually had lots of cast and crew. Association rule mining was performed with a minimum support of 10 divided by the length of the dataset, which equated to about 0.0006. Low support was chosen to find combinations of feature values that might have occurred rarely but always made a movie successful. Only rules that had confidence above 0.8 and the consequent equal to "result_HIT", which denoted that the movie had been successful were kept. This was done to only find rules that were true most of the time. There were 247,325 such rules found. The summary of quality measures can be seen in Table 21. It can be seen that the median confidence of the rules was 0.93 and at least 25% of the rules had confidence equal to 1.

Table 21: The summary of quality measures of successful movies by income.

	Support	**Confidence**	**Lift**	**Count**
min	0.0006	0.80	2.61	10
25%	0.0006	0.91	2.97	10
50%	0.0007	0.93	3.03	11
avg	0.0007	0.94	3.06	11.8
75%	0.0007	1.00	3.27	12
max	0.0051	1.00	3.27	82

The association rules with the highest support can be seen in Table 22. It is no surprise that movies with very popular and liked trailers that belong to collections (i.e., have sequels made) are likely to be successful (confidence of 0.83) because one would not make a sequel if the movie had not been a hit in the first place. Examples of movies that had these features include

Table 22: The rules with the highest support.

Antecedent	Consequent	Support	Confidence	Lift	Count
inCollection_TRUE, trailerLikesToDislikesRatio_SUPERB, trailerViews_VERY-POPULAR	result_HIT	0.0051	0.83	2.7	81
continents_North America, inCollection_TRUE, length_VERY-LONG	result_HIT	0.0051	0.80	2.6	82

Star Wars, Top Gun, Deadpool, Kingsman: The Secret Service.

There were 77 association rules with the antecedent length being equal to 1 meaning that the appearance of these feature values meant the movies were successful more than 80% of the time. The distribution of these features and some example values can be seen in Table 23. It can be seen that the cast, crew, and production companies all play a major role in the outcome of the movie. In addition to that, it can be seen that movies released in 1972 have done well.

Table 23: The distribution of features with 1 antecedent.

Feature	Count	Examples
Crew (production)	15	Craig Perry, Walt Disney, James Wan, Bradley Fuller, Andrew Form, Jason B. Stamey, Victoria Alonso, Robert Daley, Walter Hamada, Oren Peli
Crew (sound)	10	Kim Tae-seong, Michael Hedges, Harry Manfredini, Daniel Laurie, Andrew Tay, Ron Judgins, Eun Hui-su Gang Hye-yeong, Choi Tae-young, Robert Jackson
Crew (costume & make-up)	9	Chae Kyung-hwa, Dorothy Jeakins, Stuart Freeborn, Keith G. Lewis, Kwon Yoo-jin, Kim Seo-young, Eleanor Sabaduquia, Kwak Tae-young, Choi Hye-lim
Crew (visual effects)	7	Tim Burke, Milt Kahl, Frank Thomas, Eric Larson
Companies	6	Beijing Enlight Pictures, Marvel Studios, Heyday Films, Iranian Independents, IMAX, Fuji Television Network
Crew (art)	6	Dan Davis, Andy Park, Cho Hwa-sung, Charles Rosen
Crew (editing)	6	Shin Min-kyung, Richard A. Harris, Bud Molin
Cast	5	Rica Matsumoto, Peter Sellers, Oh Dal-su, Ikue Otani
Crew (directing)	3	Kunihiko Yuyama, Carl Reiner, Alfred Hitchcock
Crew (writing)	3	J.K. Rowling, Satoshi Tajiri, George Lucas
Crew (camera)	3	Bobby Greene, David M. Walsh, Jan de Bont
Crew (misc)	2	Yoon Dae-won, Andy Thomas
Crew (lighting)	1	Ross Dunkerley
Release year	1	1972

The rules were then grouped into 3 groups by lift using k-means clustering. The resulting groups can be seen on the balloon plot in Figure 13. The colour of the balloon represents the aggregated median lift of the group and the size of the balloon shows the aggregated support. In addition to that, the number of rules in the group, the number of unique features, and 5 of the

most important (frequent) features can be seen.

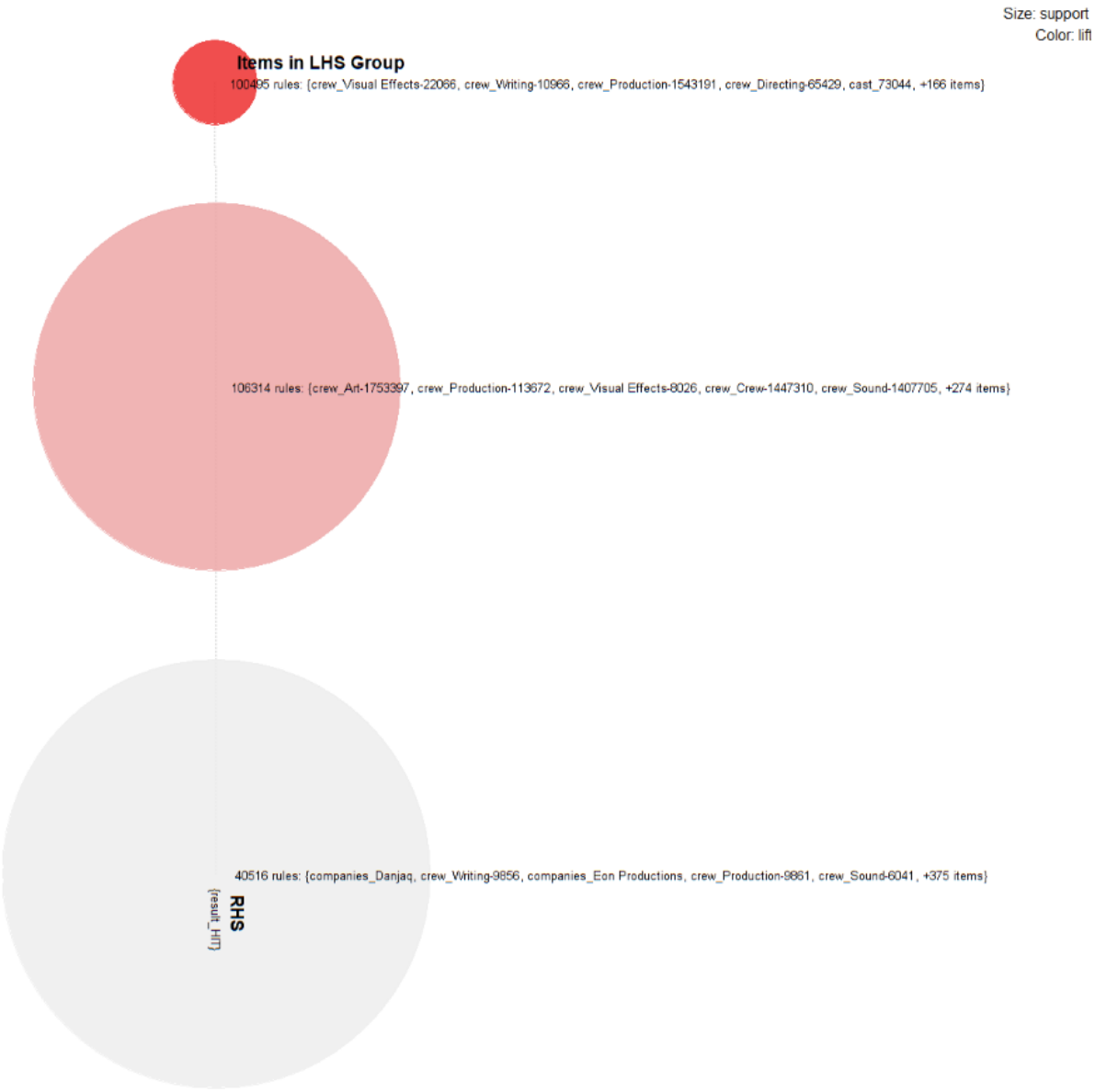

Figure 13: The association rules of successful movies by income grouped by lift.

In the first group, there were 100,495 rules. Some of the most frequently found features of successful movies in that group were:

1. **crew_Visual Effects-22066** – The ID corresponds to **Eric Larson** who was an animator for the Walt Disney Studios starting in 1933. He has been involved as an animator in 12 movies in our dataset (e.g., Cinderella, Mary Poppins, Alice in Wonderland, Sleeping Beauty) of which have all been successful. Aside from his participation, the movies have the following in common:

   (a) **Family animation genres:** In total 427 movies in our dataset have these genres of which 49% have been successful and 15% have flopped.

(b) **Produced in North America:** In total 10,249 movies in our dataset have been produced in North America of which 33% have been successful and 23% have flopped.

(c) **Produced by Walt Disney company:** In total 288 movies in our dataset have been produced by Walt Disney company of which 47% have been successful and 9% have flopped.

(d) **In the English language:** In total 11,905 movies in our dataset are in English of which 31% have been successful and 26% have flopped.

(e) **Intended for general audiences i.e., maturity rating G:** In total 367 movies in our dataset have that maturity rating of which 54% have been successful and 15% have flopped.

There are 45 movies in our dataset with these characteristics and 30 (67%) of them have been successful while only 1 has flopped.

2. **crew_Writing-10966** – The ID corresponds to **J. K. Rowling** who is best known for writing the Harry Potter fantasy series. She has been involved as a writer in 10 movies in our dataset (e.g., Harry Potter and the Philosopher's Stone and Fantastic Beasts and Where to Find Them) of which have all been successful. Aside from her participation, the movies have the following in common:

   (a) **Fantasy adventure genres:** In total 399 movies in our dataset have these genres of which 41% have been successful and 20% have flopped.

   (b) **Produced in Europe:** In total 6,123 movies in our dataset have been produced in Europe of which 17% have been successful and 17% have flopped.

   (c) **Produced by Heyday Films and Warner Bros. Pictures:** In total 13 movies in our dataset have been produced by them of which have all been successful.

   (d) **In the English language:** In total 11,905 movies in our dataset are in English of which 31% have been successful and 26% have flopped.

   (e) **Belongs to a collection:** In total 2,108 movies in our dataset belong to a collection of which 58% have been successful and 9% have flopped.

   (f) **Parental guidance suggested or cautioned, i.e., Maturity rating PG or PG-13:** In total 3,995 movies in our dataset have these maturity ratings of which 42% have been successful and 19% have flopped.

   Excluding the production companies, there are 36 such movies in our dataset of which 47% have been successful and 20% have flopped.

3. **crew_Production-1543191** – The ID corresponds to **Jason B. Stamey** who is best known for his work on WandaVision and Avengers. He has been involved as a casting associate

for 17 movies in our dataset (e.g., Guardians of the Galaxy, Doctor Strange, Avengers: Endgame, Alpha) of which 15 (88%) have been successful and only 1 has flopped. Aside from his participation, the successful movies have the following in common:

(a) **Action-adventure genre:** In total 878 movies in our dataset have these genres of which 43% have been successful and 18% have flopped.

(b) **Produced in North America:** In total 10,249 movies in our dataset have been produced in North America of which 33% have been successful and 23% have flopped.

(c) **Produced by Marvel Studios:** In total 23 movies in our dataset have been produced by Marvel Studios of which 87% have been successful and 0% have flopped.

(d) **In the English language:** In total 11,905 movies in our dataset are in English of which 31% have been successful and 26% have flopped.

(e) **Very popular trailers having more than 1 million views:** In total 2,711 movies in our dataset have very popular trailers of which 43% have been successful and 16% have flopped.

(f) **Belongs to a collection:** In total 2,108 movies in our dataset belong to a collection of which 58% have been successful and 9% have flopped.

There are 22 movies in our dataset with these characteristics of which 20 (91%) have been successful and 0 that have flopped. Excluding the Marvel Studios production company, there are 137 such movies in our dataset of which 99 (72%) have been successful, and only 2 (9%) have flopped.

In the second group, there were 106,314 rules. Some of the most frequently found features of successful movies in that group were:

1. **crew_Art-1753397** and **crew_Production-113672**- The IDs correspond to **Andy Park** who is a Korean-American comic book artist, illustrator, concept artist in the Visual Development team in Marvel Studios and **Alan Fine** who is a former Marvel Entertainment executive. Andy Park has been involved as an illustrator and Alan Fine as an executive producer in 13 movies in our dataset (e.g., The Avengers, Doctor Strange) of which 11 (85%) have been successful and only 1 that has flopped. Aside from their participation, the successful movies have the following in common:

   (a) **Produced by Marvel Studios:** In total 23 movies in our dataset have been produced by Marvel Studios of which 87% have been successful and 0% have flopped.

   (b) **Producer Kevin Feige:** In total 41 movies in our dataset have been produced by Kevin Feige of which 66% have been successful and 5% have flopped.

   There are 23 movies in our dataset with these characteristics of which 20 (87%) have been successful and 0 that have flopped.

2. **crew_Visual Effects-8026** – The ID corresponds to **Brett Coderre** who is an award-winning animator. He has been involved as an animator in 14 movies in our dataset (e.g., Cars, Up, Finding Nemo, Toy Story 2) of which 11 (79%) have been successful and none that have flopped. Aside from his participation, the successful movies have the following in common:

   (a) **Michael Silvers in the sound department:** In total 31 movies in our dataset have him in the sound department of which 58% have been successful and 6% have flopped.

   (b) **Produced by Pixar:** In total 22 movies in our dataset have been produced by Pixar of which 73% have been successful and 0% have flopped.

   (c) **Produced in North America:** In total 10,249 movies in our dataset have been produced in North America of which 33% have been successful and 23% have flopped.

   (d) **Family animation genres:** In total 427 movies in our dataset have these genres of which 49% have been successful and 15% have flopped.

   There are 13 movies in our dataset with these characteristics of which 11 (85%) have been successful and 0 that have flopped.

3. **crew_Crew-1447310** – The ID corresponds to **Jeannine Berger** who is a production manager. She has been involved as a post-production supervisor in 19 movies in our dataset (e.g., Minions, The Secret Life of Pets, Despicable Me) of which 13 (68%) have been successful and 3 (16%) that have flopped. Aside from her participation, the successful movies have the following in common:

   (a) **Christopher Meledandri in the production department:** In total 16 movies in our dataset have him in the production department of which 88% have been successful and 0% have flopped.

   (b) **Family animation comedy genres:** In total 214 movies in our dataset have these genres of which 52% have been successful and 12% have flopped.

   (c) **Produced in North America:** In total 10,249 movies in our dataset have been produced in North America of which 33% have been successful and 23% have flopped.

   (d) **In the English language:** In total 11,905 movies in our dataset are in English of which 31% have been successful and 26% have flopped.

   There are 12 movies in our dataset with these characteristics of which 11 (92%) have been successful and 0 that have flopped.

In the third group, there were 40,516 rules. Some of the most frequently found features of successful movies in that group were:

1. **crew_Writing-9856**, **companies_Danjaq** and **companies_Eon Productions** – The ID corresponds to **Ian Fleming** who was an English author and journalist. He is best known for his James Bond series of spy novels. **Danjaq** is the holding company responsible for the copyright and trademarks related to James Bond on the screen while **Eon Productions** is a British film production company that is a subsidiary of Danjaq. They have been involved in 22 movies in our dataset (e.g., Live and Let Die, Diamonds Are Forever, Skyfall, Spectre) of which 17 (77%) have been successful and none that have flopped. Aside from them, the successful movies have the following in common:

   (a) **Popular trailers having between 100,000 and 1,000,000 views:** In total 4,464 movies in our dataset have popular trailers of which 34% have been successful and 24% have flopped.

   (b) **Great likes-to-views ratio on trailers, i.e., ratio between 20 and 30:** In total 6,998 movies in our dataset have the ratio between that interval of which 35% have been successful and 23% have flopped.

   (c) **Action-adventure genre:** In total 878 movies in our dataset have these genres of which 43% have been successful and 18% have flopped.

   (d) **In the English language:** In total 11,905 movies in our dataset are in English of which 31% have been successful and 26% have flopped.

   (e) **Belongs to a collection:** In total 2,108 movies in our dataset belong to a collection of which 58% have been successful and 9% have flopped.

   There are 68 movies in our dataset with these characteristics of which 48 (70%) have been successful and 1 that has flopped. Including **Ian Fleming**, **Danjaq** production company and **Eon Productions** production company in these characteristics brings the success rate up to 81% and the flop rate down to 0.

2. **crew_Sound-6041** – The ID corresponds to **Brian Tyler** who is an American composer, musician, and producer. He has been involved as a music composer in 49 movies in our dataset (e.g., The Expendables, Fast & Furious, Avengers: Age of Ultron) of which 24 (49%) have been successful and 10 (20%) have flopped. Aside from him, the successful movies have the following in common:

   (a) **Belongs to a collection:** In total 2,108 movies in our dataset belong to a collection of which 58% have been successful and 9% have flopped.

   (b) **Produced in North America:** In total 10,249 movies in our dataset have been produced in North America of which 33% have been successful and 23% have flopped.

   (c) **In the English language:** In total 11,905 movies in our dataset are in English of which 31% have been successful and 26% have flopped.

(d) **Very popular trailers having more than 1 million views:** In total 2,711 movies in our dataset have very popular trailers of which 43% have been successful and 16% have flopped.

There are 530 movies in our dataset with these characteristics of which 375 (71%) have been successful and 20 (4%) have flopped. Including **Brian Tyler** in the characteristics brings the success rate up to 80% and the flop rate down to 0.

Additionally, some feature values were manually picked to find association rules with a bit lower support, but which could still provide interesting insights. To do that, a subset of the mined rules were used by filtering the antecedent using the selected feature values. The most important features of such rules can be seen below:

1. **Movies in a foreign language** – Animated adventure movies produced in Asia and belonging to a collection. Such movies had a median confidence of 0.91 and a lift of 3.

2. **Movies not belonging to a collection** – Long movies produced in North America, in the English language with popular trailers that have a great likes-to-views ratio. Such movies had a median confidence of 0.85 and a lift of 2.8.

3. **Movies with restricted access** – Horror movies produced in North America, in the English language, and belonging to collections with trailers having a great likes-to-views ratio. Such movies had a median confidence of 0.87 and a lift of 2.9.

4. **Movies shorter than 5 minutes** – No rules found.

5. **Movies produced in Oceania continent** – Very long action-adventure movies in the English language and belonging to a collection. Such movies had a median confidence of 0.84 and a lift of 2.7.

6. **The cast who were in the rules that had confidence equal to 1** – There were 10 such actors in total:

   (a) **Peter Sellers** – All of the 12 movies he has been involved in (e.g., Lolita, The Pink Panther, Casino Royale) have been successful.

   (b) **Ikue Otani** – All of the 13 movies she has been involved in (e.g., Pokémon: The Rise of Darkrai, One Piece: Stampede) have been successful.

   (c) **Rica Matsumoto** – All of the 13 animated movies that she has been involved in and which have been written by **Satoshi Tajiri** (e.g., Pokémon: The Movie 2000, Pokémon Heroes: Latios and Latias, Pokémon: The Rise of Darkrai) have been successful. Looking at all of the movies she has been involved in an impressive 93% have been successful and none that have flopped.

(d) **Colin Firth** – All of the 10 long romantic movies that he has been involved in and which have been produced in Europe (e.g., Mamma Mia!, Love Actually, Bridget Jones's Diary) have been successful. However, looking at all of the movies he has been involved in 59% have been successful and 15% have flopped.

(e) **Dustin Hoffman** – All of the 10 movies rated PG that he has been involved in (e.g., Kung Fu Panda, Agatha, Hook) have been successful. However, looking at all of the movies he has been involved in 59% have been successful and 22% have flopped.

(f) **Daniel Radcliffe** – All of the 10 movies that belong to a collection and he has been involved in (e.g., Harry Potter and The Goblet of Fire, Now You See Me 2, The Woman in Black) have been successful. However, looking at all of the movies he has been involved in 59% have been successful and 29% have flopped.

(g) **Héctor Elizondo** – All of the 10 comedy movies that he has been involved in and which have been directed by **Garry Marshall** (e.g., Valentine's Day, Pretty Woman, The Princess Diaries) have been successful. However, looking at all of the movies he has been involved in 54% have been successful and 29% have flopped.

(h) **Robert Downey Jr.** – All of the 11 adventure movies that belong to a collection and he has been involved in (e.g., Iron Man, Sherlock Holmes, The Avengers) have been successful. However, looking at all of the movies he has been involved in 49% have been successful and 25% have flopped.

(i) **Al Pacino** – All of the 10 movies with great trailer likes-to-dislikes ratio that he has been involved in (e.g., Dog Day Afternoon, Heat, The Merchant of Venice) have been successful. However, looking at all of the movies he has been involved in 44% have been successful and 26% have flopped.

(j) **Christopher Lloyd** – All of the 11 long comedy movies that he has been involved in (e.g., Back to the Future, The Dream Team, Who Framed Roger Rabbit) have been successful. However, looking at all of the movies he has been involved in 44% have been successful and 31% have flopped.

7. **The crew who were in the rules that had confidence equal to 1 –** There were 88 such persons in total, the persons who appeared in the rules most frequently and were not at the top of the groups seen previously were:

   (a) **Stan Lee**, **Sarah Finn**, and **Victoria Alonso** – All of the movies **Sarah Linn** produced and **Stan Lee** wrote were successful. Similarly, all of the movies **Victoria Alonso** produced and **Stan Lee** wrote were successful. However, out of all the movies **Sarah Finn** has been part of only 43% have been successful and 10% have flopped, for **Stan Lee**, 60% have been successful and 5% have flopped, and for **Victoria Alonso**, 77% have been successful and none have flopped.

(b) **Louis D'Esposito** – All of the movies that he produced and **Jack Kirby** wrote were successful. Out of all the movies he has been part of 69% have been successful and only 13% have flopped.

(c) **Jason Blum** – All of the movies he produced that belong to a collection were successful. Out of all the movies he has been involved in 71% have been successful and 9% have flopped.

5.1.2 By rating

To be able to tell whether a movie was successful when considering its IMDb rating only rows that had information about IMDb rating available and at least 50 votes were kept. In total there were 151,770 such movies in the dataset and they had 894,088 unique feature values. On an average, a movie had 24 feature values, with the median being 21, minimum 5 and maximum 994. Movies with many feature values usually had lots of cast and crew. Association rule mining was performed with a minimum support of 10 divided by the length of the dataset, which equated to about 0.00007. Low support was chosen to find combinations of feature values that might have occurred rarely but always made a movie successful. Only rules that had confidence above 0.8 and the consequent equal to "result_HIT", which denoted that the movie had been successful were kept. This was done to only find rules that were true most of the time. There were 2,850,408 such rules found. The summary of quality measures can be seen in Table 24. It can be seen that the median confidence of the rules was 1, which means over half of the found rules were true all the time.

Table 24: The summary of quality measures of successful movies by IMDb rating.

	Support	**Confidence**	**Lift**	**Count**
min	0.00007	0.80	3.98	10
25%	0.00007	0.92	4.60	11
50%	0.00008	1.00	4.98	12
avg	0.00010	0.96	4.79	14.9
75%	0.00010	1.00	4.98	15
max	0.00175	1.00	4.98	266

The association rule with the highest support can be seen in Table 25. In total there are 326

Table 25: The rule with the highest support.

Antecedent	**Consequent**	**Support**	**Confidence**	**Lift**	**Count**
genres_music, genres_documentary, length_LONG	result_HIT	0.00175	0.82	4.1	266

movies with these features with 266 (82%) having been successful and only 5 (2%) that have flopped. Successes include movies such as Buena Vista Social Club (1999), Imagine: John

Lennon (1988), This Is It (2009), U2: Rattle and Hum (1988), Beats Rhymes & Life: The Travels of A Tribe Called Quest (2011). Flops include movies such as Justin Bieber: Never Say Never (2011) rated at 1.6 and Tony Bennett Celebrates 90 (2016) rated at 2.5.

There were 270 association rules with the antecedent length being equal to 1 meaning that the appearance of such feature values meant the movies were successful more than 80% of the time. The distribution of these features and some example values can be seen in Table 26. It can be seen that the cast, crew and production companies all play a major role in the outcome of the movie.

Table 26: The distribution of features with 1 antecedent.

Feature	Count	Examples
Cast	98	Roger Waters, Tito Ortiz, Jerry Garcia, Bruce Springsteen, Ozzy Osbourne, Jimmy Carr, Bill Hicks, David Gilmour, Kirk Hammett
Crew (directing)	35	Tony Dow, Ken Burns, David Maysles, Tim Kirkby, Paul Wheeler, David Lean, Bob Smeaton, Kirk Browning, Brian Large
Crew (writing)	34	Eddie Izzard, George Carlin, Michael Palin, J. K. Rowling, Alan Bennett, Billy Brown, Martha Williamson, Louis Theroux
Crew (production)	26	Helen Parker, Terry Shand, Bill Melendez, Brian Klein, Raj Kapoor, Jerry Hamza, Steve Levitan, Orson Welles, Adam Somner
Companies	23	Ellipsanime Productions, Universal Pictures UK, National Theatre Live, Picture Palace, BBC Wales, Universal Music, Disneynature, Universal Music
Crew (sound)	20	Jim Morgan, Wolfgang Amadeus Mozart, Jean Nény, Charlie Chaplin, Ray Parker, Kim Foscato, Charles Paley, Manna Dey, Tom Szczesniak
Crew (visual effects)	14	Ray Patterson, Pete Burness, Kenneth Muse
Crew (art)	8	Robert Gentle, Philip DeGuard, Tatsuo Hamada
Crew (editing)	5	Akira Kurosawa, Ewa Smal, Jordana Berg
Crew (misc)	3	Andy Thomas, Wayne Fitzgerald, Charles M. Schulz
Crew (costume & make-up)	2	Amanda Knight, Alberto De Rossi
Crew (camera)	2	Yûharu Atsuta, Takao Saitô

The rules were then grouped into 3 groups by lift using k-means clustering. The resulting groups can be seen on the balloon plot in Figure 14. The colour of the balloon represents the aggregated median lift of the group and the size of the balloon shows the aggregated support.

In addition to that, the number of rules in the group, the number of unique features and 5 of the most important (frequent) features can be seen.

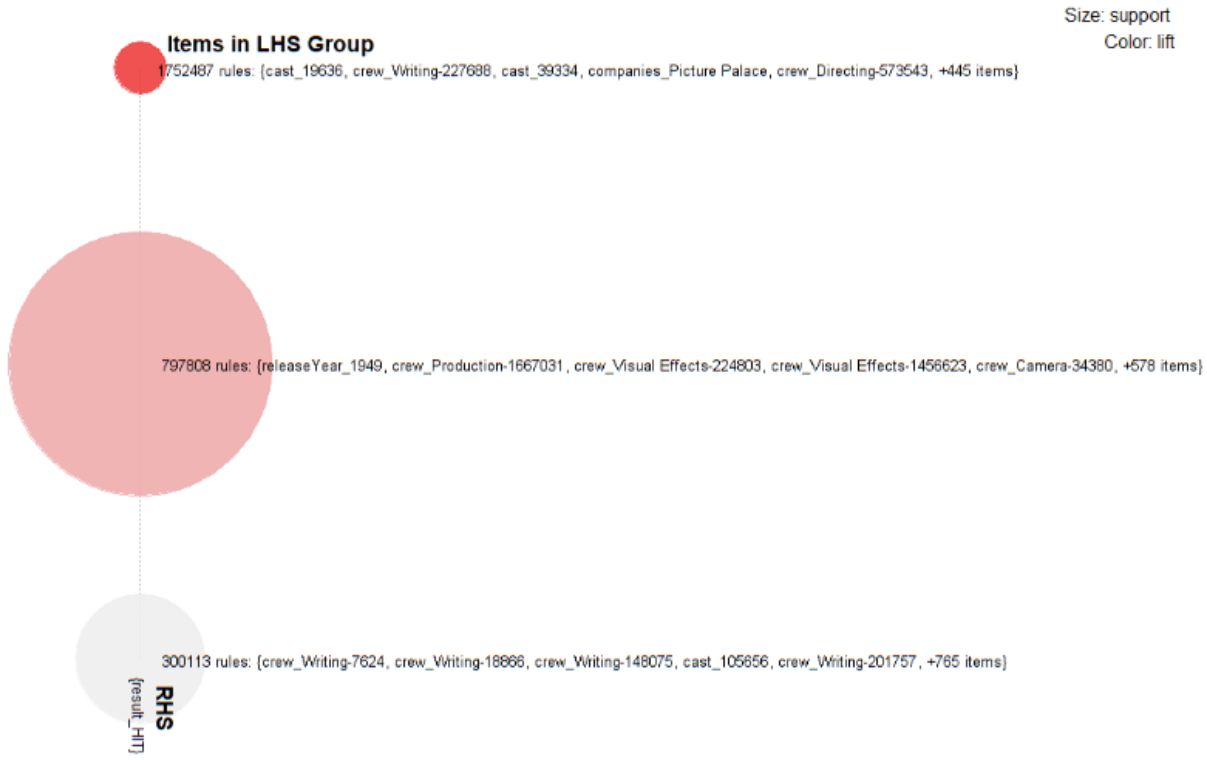

Figure 14: The association rules of successful movies by IMDb rating grouped by lift.

In the first group, there were 1,752,487 rules. Some of the most frequently found features of successful movies in that group were:

1. **cast_19636** and **cast_39334** – The IDs correspond to **Jean-Pierre Moulin** and **Yves Barsacq** who are French film actors. They have been involved in movies such as Prisoners of the Sun, Land of Black Gold, Destination Moon. Of the 26 movies **Jean-Pierre Moulin** has been involved in 18 (69%) have been successful and 2 (8%) have flopped and of the 21 movies **Yves Barsacq** has been involved in 18 (86%) have been successful and none have flopped. Aside from them, the successful movies have the following in common:

   (a) **Animation adventure genres:** In total 1,185 movies in our dataset have these genres of which 29% have been successful and 18% have flopped.

   (b) **Belongs to a collection:** In total 9,287 movies in our dataset belong to a collection of which 18% have been successful and 31% have flopped.

   (c) **Produced in Europe:** In total 46,740 movies in our dataset have been produced in Europe of which 21% have been successful and 22% have flopped.

   (d) **In a foreign language:** In total 58,519 movies in our dataset are in a foreign language of which 21% have been successful and 22% have flopped.

There are 84 movies in our dataset with these characteristics and 39 (46%) of them have been successful while 12 (14%) have flopped. Including the aforementioned actors in these characteristics brings the success rate up to 100%.

2. **crew_Writing-227688** – The ID corresponds to **Bernard Cornwell** who is an English author of historical novels. He has been involved as a writer in 16 movies in our dataset (e.g., Sharpe's Gold, Sharpe's Honour, Sharpe's Waterloo) and they have all been successful. Aside from him, the movies have the following in common:

   (a) **Historical war films:** In total 752 movies in our dataset have these genres of which 37% have been successful and 9% have flopped.

   (b) **Produced in Europe:** In total 46,740 movies in our dataset have been produced in Europe of which 21% have been successful and 22% have flopped.

   (c) **In the English language:** In total 93,251 movies in our dataset are in the English language of which 20% have been successful and 31% have flopped.

   There are 144 movies in our dataset with these characteristics and 68 (47%) of them have been successful while only 9 (6%) have flopped.

In the second group, there were 797,808 rules. Some of the most frequently found features of successful movies in that group were:

1. **releaseYear_1949** – The feature denotes movies released in 1949. There are 555 movies released in 1949 in our dataset (e.g., Late Spring, White Heat, Twelve O'Clock High) of which 24% have been successful and 6% that have flopped. Aside from the release year, the successful movies have the following in common:

   (a) **Short animated movies:** In total 3,750 movies in our dataset have these characteristics of which 36% have been successful and 6% have flopped.

   (b) **Produced by Metro-Goldwyn-Mayer:** In total 2,132 movies in our dataset have been produced by Metro-Goldwyn-Mayer of which 16% have been successful and 11% have flopped.

   There are 182 movies in our dataset with these characteristics and 105 (58%) of them have been successful while only 6 (3%) have flopped. Including the 1949 release year brings the success rate up to 92% and the flop rate down to 0%.

2. **crew_Production-1667031** – The ID corresponds to **Steve Levitan** who is an American director, screenwriter, and producer. He has been involved as a producer in 13 movies in our dataset (e.g., Striker's Mountain, Goosebumps) of which 11 (85%) have been successful and 2 (15%) that have flopped. Aside from him, the successful movies have the following in common:

(a) **Family horror genres:** In total 100 movies in our dataset have these genres of which 28% have been successful and 34% have flopped.

(b) **Movies with the length between 29 and 96 minutes:** In total there are 72,484 movies in our dataset with the length between that interval of which 16% have been successful and 34% have flopped.

(c) **In the English language:** In total 93,251 movies in our dataset are in the English language of which 20% have been successful and 31% have flopped.

(d) **Not belonging to a collection:** In total 142,483 movies in our dataset do not belong to a collection of which 20% have been successful and 27% have flopped.

There are 57 movies in our dataset with these characteristics and 20 (35%) of them have been successful while 16 (28%) have flopped. Including **Steve Levitan** in the characteristics brings the success rate up to 92% and the flop rate down to 8%.

3. **crew_Visual Effects-1456623** – The ID corresponds to **Shinji Ôtsuka** who is a Japanese animator. He has been part of 12 movies in our dataset (e.g., The Boy and the Beast, Howl's Moving Castle, Princess Mononoke) of which 11 (92%) have been successful and none that have flopped. Aside from him, the successful movies have the following in common:

   (a) **Animated movies:** In total 8,059 in our dataset have that genre of which 31% have been successful and 12% have flopped.

   (b) **Produced in Asia:** In total 18,586 movies in our dataset have been produced in Asia of which 23% have been successful and 21% have flopped.

   (c) **In a foreign language:** In total 58,519 movies in our dataset are in a foreign language of which 21% have been successful and 22% have flopped.

   (d) **Trailers having superb likes-to-views and likes-to-dislikes ratios:** In total 3,542 movies in our dataset have trailers with such great ratios of which 45% have been successful and 10% have flopped.

   There are 296 movies in our dataset with these characteristics and 168 (57%) of them have been successful while only 10 (34%) have flopped.

In the third group, there were 300,113 rules. Some of the most frequently found features of successful movies in that group were:

1. **crew_Writing-7624** and **crew_Writing-18866** – The IDs correspond to **Stan Lee** and **Jack Kirby** who were American comic book writers, and editors. The duo created many of the Marvels' major characters such as the X-Men, Thor, Iron Man. They have been involved as writers in 56 movies in our dataset (e.g., Spider-Man, Fantastic Four, The

Avengers) of which 21 (38%) have been successful and 13 (23%) have flopped. Aside from them, the successful movies have the following in common:

(a) **Action-adventure genre:** In total 3,270 movies in our dataset have these genres of which 13% have been successful and 38% have flopped.

(b) **Trailers having more than 1 million views and superb likes-to-views ratio:** In total 1,720 movies in our dataset have trailers with these characteristics of which 34% have been successful and 19% have flopped.

(c) **Belongs to a collection:** In total 9,287 movies in our dataset belong to a collection of which 18% have been successful and 31% have flopped.

There are 76 movies in our dataset with these characteristics and 37 (49%) of them have been successful while only 3 (4%) have flopped. Including **Stan Lee** and **Jack Kirby** in the characteristics brings the success rate up to 67% and the flop rate down to 0%.

2. **crew_Writing-148075** – The ID corresponds to **Nick George** who was an author of comic book stories and animation scripts. He has been involved as a writer in 33 movies in our dataset (e.g., Lucky Number, Out on a Limb, Winter Storage) of which 18 (55%) have been successful and none that have flopped. Aside from him, the successful movies have the following in common:

   (a) **Short animated movies:** In total 3,750 movies in our dataset have these characteristics of which 36% have been successful and 6% have flopped.

   (b) **Produced in North America:** In total 53,739 movies in our dataset have been produced in North America of which 17% have been successful and 32% have flopped.

   There are 1,962 movies in our dataset with these characteristics and 705 (36%) of them have been successful while 121 (6%) have flopped. Including **Nick George** in the characteristics brings the success rate up to 52% and the flop rate down to 0%.

Additionally, some feature values were manually picked to find association rules with a bit lower support, but which could still provide interesting insights. To do that, a subset of the mined rules were used by filtering the antecedent using the selected feature values. The most important features of such rules can be seen below:

1. **Movies with restricted access** – Very long dramas with very popular trailers. Such movies had a median confidence of 0.81 and a lift of 4.

2. **Movies shorter than 5 minutes** – No rules found.

3. **Movies produced in Oceania continent** – No rules found.

4. **The cast who were in the rules that had confidence equal to 1** – There were 115 such actors in total, the three who appeared in the rules most frequently and were not at the top of the groups seen previously were:

   (a) **Thierry Wermuth** – All of the 18 medium-length movies that he has been involved in (e.g., Cigars of the Pharaoh, The Seven Crystal Balls) have been successful. Looking at all of the movies he has been involved in 90% have been successful and none have flopped.

   (b) **Christian Pelissier** – All of the 13 medium-length movies that he has been involved in (e.g., The Secret of the Unicorn, Prisoners of the Sun) have been successful. Looking at all of the movies he has been involved in 93% have been successful and none have flopped.

   (c) **Henri Labussière** – All of the 12 movies he appeared in together with **Christian Pelissier** (e.g., Destination Moon, The Red Sea Sharks) have been successful. Looking at all of the movies he has been involved in 65% have been successful and only 9% have flopped.

5. **The crew who were in the rules that had confidence equal to 1** – There were 212 such persons in total, the persons who appeared in the rules most frequently and were not at the top of the groups seen previously were:

   (a) **Joseph Barbera (directing)** and **William Hanna (directing, acting)** – They were the creators of Tom & Jerry. All of the 11 short movies they have co-directed and in which **William Hanna** also acted in (e.g., Solid Serenade, The Lonesome Mouse) have been successful. Individually, the movies **Joseph Barbarra** has directed have a success rate of 75% and a flop rate of 6%, and the movies **William Hanna** has directed have a success rate of 78% and a flop rate of 5%.

   (b) **Robert Gentle (art)** and **Fred Quimby (production)** – All of the 19 movies where **Robert Gentle** was the background designer and which were produced by **Fred Quimby** (e.g., Pet Peeve, That's My Pup!) have been successful. Looking at all of the movies **Robert Gentle** has been in the art department 87% have been successful and none have flopped, and looking at all of the movies **Fred Quimby** has produced 81% have been successful and none have flopped.

   (c) **Stéphane Bernasconi (directing)** and **Jim Morgan (sound)** – All of the 13 movies which **Stéphane Bernasconi** directed, **Jim Morgan** produced music, and **Jean-Pierre Moulin** acted in were successful. Both **Stéphane Bernasconi** and **Jim Morgan** have been part of the same movies in our dataset with the success rate of 90% and no flops.

5.1.3 Comparison of successful movies by income and rating

To compare the similarities and differences between successful movies by income and rating 25% of the most frequent features from the found association rules were taken. 16 features appeared in both association rule lists. The distribution of these features and their values can be seen in Table 27. It can be seen that a good crew and trailer are very important.

Table 27: The distribution of common features in successful movies association rules.

Feature	Count	Values
Genres	12	Adventure, Action, Science fiction, Thriller, Animation, Family, Comedy, Horror, Fantasy, Drama, Romance, Crime
Crew (production)	8	Walt Disney, Victoria Alonso, Stan Lee, Steven Spielberg, Sarah Finn, Kevin Feige, Jason B. Stamey, Louis D'Esposito
Companies	6	Marvel Studios, Warner Bros. Pictures, Columbia Pictures, Paramount, Pixar, Walt Disney Productions
Crew (visual effects)	5	Eric Larson, Frank Thomas, Milt Kahl, John Lounsbery, Ollie Johnston
Continents	3	North America, Europe, Asia
Length	3	Medium, Long, Very long
Maturity rating	3	PG-13, G, R
Trailer views	3	Very popular, Popular, Average
Crew (writing)	2	Stan Lee, Jack Kirby
In collection	2	True, False
Language	2	English, Foreign
Trailer comments-to-views ratio	2	Great, Average
Trailer likes-to-dislikes ratio	2	Superb, Great
Trailer likes-to-views ratio	2	Superb, Great
Crew (art)	1	Andy Park
Crew (sound)	1	John Williams

There were 13 features with values that appeared in successful movies association rules by income but did not appear in successful movies association rules by rating. The distribution of these features and their values can be seen in Table 28. It can be seen that there are companies, crew, and cast whose movies have done well income-wise, but not that good rating-wise. Additionally, it can be seen that some of the movies that haven't had good trailers have still done well income-wise.

There were 15 features with values that appeared in successful movies association rules by rating but did not appear in successful movies association rules by income. The distribution of

Table 28: The distribution of features in successful movies association rules by income but not by rating.

Feature	Count	Values
Companies	17	Eon Productions, United Artists, Blumhouse Productions, Universal Pictures, 20th Century Fox, Lucasfilm, Platinum Dunes, Sony Pictures Animation, StudioCanal, DreamWorks Animation, New Line Cinema, Working Title Films, Heyday Films, Walt Disney Pictures, Amblin Entertainment, Danjaq, Sony Pictures
Crew (production)	12	Albert R. Broccoli, Terri Taylor, Avi Arad, Jason Blum, Andrew Form, Michael Bay, George Lucas, Tim Bevan, Eric Fellner, Bradley Fuller, Kathleen Kennedy, Alan Fine
Crew (sound)	9	Christopher Boyes, Dan O'Connell, John T. Cucci, Michael Silvers, Jana Vance, Dennie Thorpe, Dave Jordan, Brian Tyler, Ben Burtt
Crew (writing)	4	J.K. Rowling, George Lucas, Ian Fleming, Satoshi Tajiri
Cast	3	Robert Downey Jr., Rica Matsumoto, Ikue Otani
Crew (visual effects)	2	Daniel Sudick, Brett Coderre
Trailer likes-to-dislikes ratio	2	Average, Bad
Crew (art)	1	Stuart Craig
Crew (directing)	1	Kunihiko Yuyama
Genres	1	Mystery
Maturity rating	1	PG
Release year	1	2017
Trailer comments-to-views ratio	1	Bad

these features and examples of their values can be seen in Table 29. Similarly, it can be seen that there are cast, crew and companies whose movies have done well rating-wise, but not that good income-wise.

Table 29: The distribution of features in successful movies association rules by rating but not by income.

Feature	Count	Examples
Cast	40	Sean Bean, Yves Barsacq, Thierry Wermuth
Crew (writing)	28	Joseph Barbera, Michael Maltese, Krzysztof Piesiewicz
Crew (directing)	26	Chuck Jones, Stéphane Bernasconi, Friz Freleng
Companies	20	Ellipsanime Productions, Picture Palace, Metro-Goldwyn-Mayer, BBC, World Wrestling Entertainment (WWE)
Crew (visual effects)	18	Ray Patterson, Ed Barge, Ben Washam
Crew (production)	11	Hal Roach, Fred Quimby, John Lasseter
Crew (sound)	11	Ray Parker, Jim Morgan, Scott Bradley
Crew (editing)	7	Treg Brown, Ewa Smal, Michael Kahn
Release year	7	1943, 1948, 1949, 1954, 1989, 1992, 2012
Crew (art)	6	Robert Gentle, Hans Dreier, Philip DeGuard
Genres	5	TV movie, History, War, Documentary, Music
Crew (camera)	3	Janusz Kamiński, Art Lloyd, Yûharu Atsuta
Crew (costume & make-up)	2	Edith Head, Wally Westmore
Length	1	Short
Trailer views	1	Very unpopular

5.2 Characteristics of unsuccessful movies

5.2.1 By income

Similarly to characteristics of successful movies by income in Section 5.1.1, only the 16,034 rows that had information about the revenue were kept. Association rule mining was then performed with a minimum support of 10 divided by the length of the dataset, which equated to about 0.0006. Low support was chosen to find combinations of feature values that might have occurred rarely but always made a movie flop. Only rules that had confidence above 0.8 and the consequent equal to "result_FLOP", which denoted that the movie had been unsuccessful were kept. This was done to only find rules that were true most of the time. There were 207 such rules found. The summary of quality measures can be seen in Table 30. It can be seen that most of the rules had confidence between 0.8 and 0.85, which means there were often some cases where the rules were not true.

Table 30: The summary of quality measures of unsuccessful movies by income.

	Support	**Confidence**	**Lift**	**Count**
min	0.00062	0.80	3.33	10
25%	0.00062	0.83	3.47	10
50%	0.00062	0.83	3.47	10
avg	0.00071	0.85	3.53	11.3
75%	0.00075	0.85	3.52	12
max	0.00131	1.00	4.16	21

The association rule with the highest support can be seen in Table 31. In total there are 26 movies with these features and 21 of them have flopped. Flops include movies such as Moscow

Table 31: The rule with the highest support.

Antecedent	**Consequent**	**Support**	**Confidence**	**Lift**	**Count**
genres_action, genres_horror, inCollection_FALSE, language_English, trailerViews_AVERAGE	result_FLOP	0.0013	0.81	3.4	21

Zero, The Tenant, Hood of Horror, Beowulf, God of Vampires. Most of the flopped movies have very bad ratings, the average income of about -$9 million and the median income of about -$5.6 million. The 5 that did not flop (e.g., The Island, Blood: The Last Vampire) do not have budget information available, but at least they made on average about $7 million.

There was only 1 association rule with the antecedent length being equal to 1 meaning that the appearance of that feature value meant the movies were unsuccessful more than 80% of the time. The aforementioned feature value was producer **Andrew Stevens** – 83% of movies he produced flopped.

The rules were then grouped into 3 groups by lift using k-means clustering. The resulting groups can be seen on the balloon plot in Figure 15. The colour of the balloon represents the

aggregated median lift of the group and the size of the balloon shows the aggregated support. In addition to that, the number of rules in the group, the number of unique features and 5 of the most important (frequent) features can be seen.

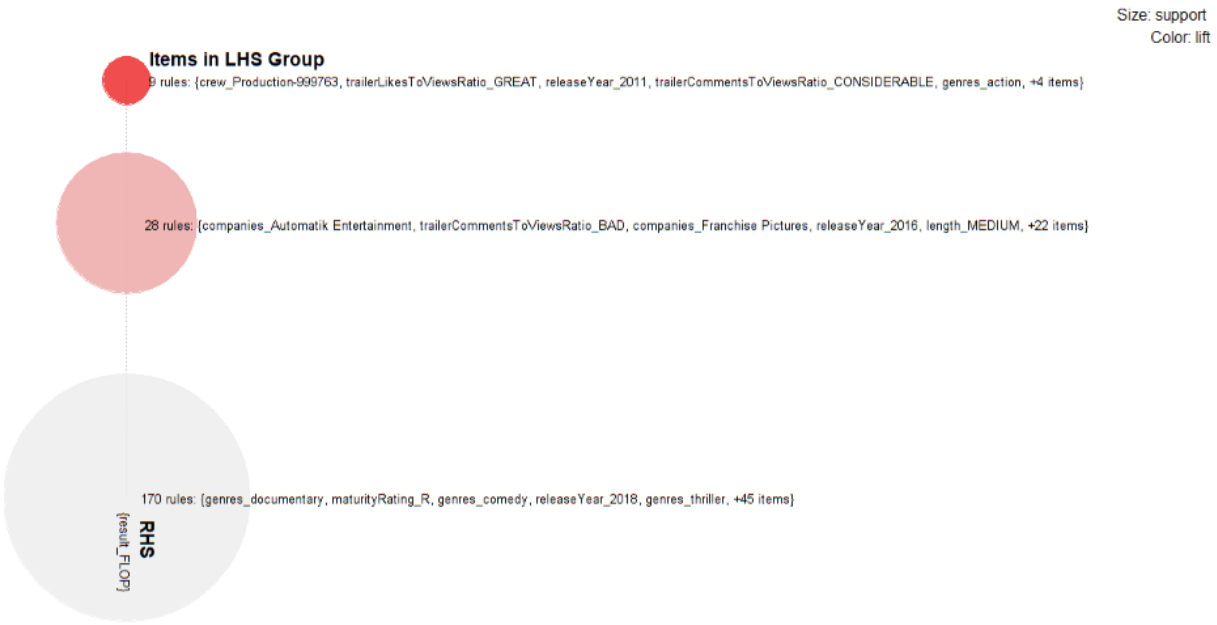

Figure 15: The association rules of unsuccessful movies by income grouped by lift.

In the first group, there were 9 rules. Some of the most frequently found features of unsuccessful movies in that group were:

1. **crew_Production-999763** – The ID corresponds to **Brian Kavanaugh-Jones**. He has been involved as a producer in 26 movies in our dataset (e.g., Skyline, At First Light, Boundaries) of which 12 (46%) have flopped. Aside from his participation, the movies that flopped have the following in common:

   (a) **Not belonging to a collection:** In total 13,926 movies in our dataset do not belong to a collection of which 26% have been successful and 26% have flopped.

   (b) **Trailers having a comments-to-views ratio between 0.001 and 0.0001:** In total 8,120 movies in our dataset have the ratio between that interval of which 34% have been successful and 24% have flopped.

   There are 6,858 movies in our dataset with these characteristics and 1,989 (29%) of them have been successful while 1,783 (26%) have flopped. His participation, however, has brought the success rate down to 9% and the flop rate up to 91%.

2. **trailerLikesToViewsRatio_GREAT** and **trailerCommentsToViewsRatio_AVERAGE** – **Great likes-to-views ratio** means the ratio is between 0.001 and 0.004 or in other words 1 like for every 250-1,000 views and **average comments-to-views ratio** means the ratio is between 0.001 and 0.0001 or in other words 1 comment for every 1,000-10,000 views.

There are 5,657 such movies in our dataset (e.g., Vertigo, Jurassic Park, Malcolm X) of which 35% have been successful and 23% have flopped. Aside from these, the flopped movies have the following in common:

(a) **Not belonging to a collection:** In total 13,926 movies in our dataset do not belong to a collection of which 26% have been successful and 26% have flopped.

(b) **Produced by Automatik Entertainment production company:** In total there are 23 movies in our dataset that have been produced by Automatik Entertainment of which 22% have been successful and 48% have flopped. One of the owners of the company is the aforementioned **Brian Kavanaugh-Jones**.

There are 20 movies in our dataset with these characteristics and only 2 (10%) of them have been successful while 11 (55%) have flopped. Including the trailer likes-to-views and comments-to-views ratio characteristics, the success rate drops to 9% and the flop rate increases to 91%.

3. **genres_action** – Movies in the **action genre**. There are 2,889 such movies in our dataset (e.g., Shaft, Heist, RoboCop) of which 35% have been successful and 22% have flopped. Aside from the genre, the flopped movies have the following in common:

   (a) **In a foreign language:** In total 4,129 movies in our dataset are in a foreign language of which 29% have been successful and 19% have flopped.

   (b) **Produced by Canal+ production company:** In total 346 movies in our dataset have been produced by Canal+ of which 25% have been successful and 23% have flopped.

   There are 179 movies in our dataset with these characteristics of which 43 (24%) have been successful and 36 (20%) have flopped. Including the action genre characteristic brings the success rate down to 7% and the flop rate up to 80%.

In the second group, there were 28 rules. Some of the most frequently found features of unsuccessful movies in that group were:

1. **trailerCommentsToViewsRatio_BAD – Bad comments-to-views ratio** means the ratio is between 0.0001 and 0.00001 or in other words 1 comment for every 10,000-100,000 views. There are 1,554 such movies in our dataset (e.g., Four Rooms, Miami Vice, Spartan) of which 34% have been successful and 22% have flopped. Aside from this, the flopped movies have the following in common:

   (a) **Thriller genre:** In total 3,121 movies in our dataset have this genre of which 30% have been successful and 26% have flopped.

   (b) **Released in 2018:** In total 684 movies in our dataset have this release year of which 31% have been successful and 22% have flopped.

(c) **Trailers having a likes-to-dislikes ratio between 1 to 10:** In total 2,702 movies in our dataset have the ratio between that interval of which 27% have been successful and 29% have flopped.

(d) **Trailers having a likes-to-views ratio between 0.0001 and 0.001:** In total 1,696 movies in our dataset have the ratio between that interval of which 29% have been successful and 26% have flopped.

There are 18 movies in our dataset with these characteristics and only 2 (11%) of them have been successful while 12 (67%) have flopped. Including the bad comments-to-views ratio brings the success rate down to 0% and the flop rate up to 91%.

2. **companies_Franchise Pictures – Franchise Pictures** was an independent motion picture production and distribution company that went bankrupt. There are 23 movies in our dataset produced by them (e.g., Get Carter, Half Past Dead, The Pledge) of which 4% have been successful and 70% have flopped. Aside from the company, the flopped movies have the following in common:

   (a) **Produced by Elie Samaha.:** In total 22 movies in our dataset have been produced by him of which 14% have been successful and 64% have flopped.

   (b) **Not belonging to a collection:** In total 13,926 movies in our dataset do not belong to a collection of which 26% have been successful and 26% have flopped.

   There are 17 movies in our dataset with these characteristics and only 3 (18%) of them have been successful while 12 (71%) have flopped. Including the company brings the success rate down to 9% and the flop rate up to 91%.

3. **length_MEDIUM** – The **medium length** denotes movies with the length between 29 and 96 minutes. There are 5,418 movies in our dataset with the length between that interval of which 26% have been successful and 29% have flopped. Aside from the length, the flopped movies have the following in common:

   (a) **Action-horror genre:** In total 154 movies in our dataset have this genre of which 22% have been successful and 32% have flopped.

   (b) **Not belonging to a collection:** In total 13,926 movies in our dataset do not belong to a collection of which 26% have been successful and 26% have flopped.

   (c) **Trailers having between 10,000 and 100,000 views:** In total there are 3,577 movies in our dataset having views between that interval of which 29% have been successful and 27% have flopped.

   There are 30 movies in our dataset with these characteristics and only 2 (7%) of them have been successful while 22 (73%) have flopped. Including the medium length brings the success rate down to 0% and the flop rate up to 77%.

In the third group, there were 170 rules. Some of the most frequently found features of unsuccessful movies in that group were:

1. **genres_documentary** – Movies in the **documentary genre**. There are 1,313 such movies in our dataset (e.g., Back to Normandy, The Tillman Story, The Living Sea) of which 12% have been successful and 37% have flopped. Aside from the genre, the flopped movies have the following in common:

   (a) **Trailers having between 10,000 and 100,000 views:** In total there are 3,577 movies in our dataset having views between that interval of which 29% have been successful and 27% have flopped.

   (b) **Trailers having a comments-to-views ratio between 0.001 and 0.0001:** In total 8,120 movies in our dataset have the ratio between that interval of which 34% have been successful and 24% have flopped.

   (c) **Released in 2018:** In total 684 movies in our dataset have this release year of which 31% have been successful and 22% have flopped.

   There are 59 movies in our dataset with these characteristics and only 10 (17%) of them have been successful while 31 (53%) have flopped. Including the documentary genre brings the success rate down to 0% and the flop rate up to 83%.

2. **maturityRating_R** – The **maturity rating R** means that under 17 requires accompanying parent or adult guardian. There are 3,940 such movies in our dataset (e.g., Memento, Blade Runner, Psycho) of which 32% have been successful and 27% have flopped. Aside from the maturity rating, the flopped movies have the following in common:

   (a) **Horror thrillers:** In total 674 movies in our dataset have these genres of which 37% have been successful and 24% have flopped.

   (b) **Not belonging to a collection:** In total 13,926 movies in our dataset do not belong to a collection of which 26% have been successful and 26% have flopped.

   (c) **Produced in Europe:** In total 6,123 movies in our dataset have been produced in Europe of which 17% have been successful and 17% have flopped.

   (d) **Long movies with the length between 96 and 135 minutes:** In total 8,511 movies in our dataset have the length between that interval of which 33% have been successful and 22% have flopped.

   There are 73 movies in our dataset with these characteristics and 9 (12%) of them have been successful while 32 (44%) have flopped. Including the R maturity rating brings the success rate down to 12% and the flop rate up to 58%.

3. **genres_comedy** – Movies in the **comedy genre**. There are 5,390 such movies in our dataset (e.g., Beverly Hills Cop, Something Wild, 50 First Dates) of which 36% have been successful and 20% have flopped. Aside from the genre, the flopped movies have the following in common:

   (a) **Produced by Kerry Barden:** There are 123 movies in our dataset that he has produced of which 31% have been successful and 33% have flopped.

   (b) **Trailers having between 10,000 and 100,000 views:** In total there are 3,577 movies in our dataset having views between that interval of which 29% have been successful and 27% have flopped.

   There are 25 movies in our dataset with these characteristics and only 3 (12%) of them have been successful while 16 (64%) have flopped. Including the comedy genre brings the success rate down to 8% and the flop rate up to 85%.

Additionally, some feature values were manually picked to find association rules with a bit lower support, but which could still provide interesting insights. To do that, a subset of the mined rules were used by filtering the antecedent using the selected feature values. The most important features of such rules can be seen below:

1. **Movies belonging to a collection** – No rules found.

2. **Movies, which trailers had more than 1 million views** – No rules found.

3. **Movies, which trailers likes-to-dislikes ratio was higher than 30** – No rules found.

4. **Movies, which trailers likes-to-views ratio was higher than 0.004** (i.e., more than 1 like for every 250 views) – Documentaries released in 2016-2018, which trailers had between 100,000 and 1,000,000 views had a flop rate of 87%. This shows that even if people like the documentaries, they do not generate that much revenue.

5. **The cast who were found in the rules** – There was only 1 such actor found – **Christian Slater**. 81% of the thrillers he has been in have flopped. However, not counting the thrillers, 48% of the movies he has been in have been successful.

6. **The crew who were found in the rules** – There were 5 such persons found. It's interesting to note that all of them are producers:

   (a) **Andrew Stevens** – Out of the 12 films he has produced 10 (83%) have flopped and none have been successful.

   (b) **Elie Samaha** – 91% of the movies that have been produced by him and **Franchise Pictures** production company and which do not belong to a collection have flopped. Out of all the movies he has been part of 14% have been successful and 64% have flopped.

(c) **Brian Kavanaugh-Jones** – All of the 10 movies produced by him in North America, which do not belong to a collection and which trailers have the comments-to-views ratio between 0.0001 and 0.001 have flopped. Out of all the movies he has been part of 31% have been successful and 46% have flopped.

(d) **Avi Lerner** – 83% of the dramas, which have the length between 96 and 135 minutes and were produced by him have flopped. Out of all the movies he has been part of 12% have been successful and 52% have flopped.

(e) **Kerry Barden** – 85% of the comedys, which trailers have between 100,000 and 1,000,000 views and were produced by him have flopped. Out of all the movies he has been part of 31% have been successful and 33% have flopped.

5.2.2 By rating

Similarly to characteristics of successful movies by rating in Section 5.1.2, only the 151,770 rows that had information about the IMDb rating and at least 50 votes were kept. Association rule mining was then performed with a minimum support of 10 divided by the length of the dataset, which equated to about 0.00007. Low support was chosen to find combinations of feature values that might have occurred rarely but always made a movie flop. Only rules that had confidence above 0.8 and the consequent equal to "result_FLOP", which denoted that the movie had been unsuccessful were kept. This was done to only find rules that were true most of the time. There were 194,040 such rules found. The summary of quality measures can be seen in Table 32. It can be seen that the median confidence of the rules was 0.91 and at least 25% of the rules had confidence equal to 1.

Table 32: The summary of quality measures of unsuccessful movies by rating.

	Support	**Confidence**	**Lift**	**Count**
min	0.00007	0.80	2.89	10
25%	0.00007	0.85	3.06	11
50%	0.00009	0.91	3.28	13
avg	0.00012	0.91	3.30	19
75%	0.00012	1.00	3.61	18
max	0.02132	1.00	3.61	3,235

The association rules with the highest support can be seen in Table 33. Out of 4,028 movies with these characteristics, only 2% have been successful and 80% have flopped. Flops include

Table 33: The rule with the highest support.

Antecedent	**Consequent**	**Support**	**Confidence**	**Lift**	**Count**
genres_horror, language_English, length_MEDIUM, trailerViews_VERY-UNPOPULAR	result_FLOP	0.021	0.8	2.9	3,235

movies such as The Worm Eaters (1977), Bitten (2008), The 8th Plague (2006) and successes include movies such as Murrain (1975), Goosebumps: The Ghost Next Door (1998), Back to the Black Lagoon: A Creature Chronicle (2000).

There were 983 association rules with the antecedent length being equal to 1 meaning that the appearance of these feature values meant the movies were unsuccessful more than 80% of the time. The distribution of these features and some example values can be seen in Table 34. It can be seen that the cast, crew, and production companies all play a major role in the outcome of the movie. In addition to that, it can be seen that people have lowly rated movies made at the end of 19th century and it is not surprising that a movie, which trailer was not liked has also been rated lowly.

Table 34: The distribution of features with 1 antecedent.

Feature	Count	Examples
Cast	325	Enzo Salvi, Darian Caine, Daniel Bernhardt, Lydie Denier, Olivier Gruner, Lee Bane, Jennifer Blanc, Randy Wayne, Beverly Lynne
Crew (directing)	168	David A. Prior, Brett Kelly, Al Adamson
Crew (writing)	121	Robert Vince, Andrew Jones, Erich Tomek
Crew (production)	109	Paul Bales, Paul Ruddy, Gerald Webb
Companies	101	PM Entertainment Group, Feifer Worldwide, Elite Film, Action International Pictures, Full Moon Pictures, Rapid Heart Pictures
Crew (camera)	44	Mark Atkins, Ken Blakey, Klaus Werner
Crew (sound)	40	Lisa Ries, Chris Cano, John Gonzalez
Crew (editing)	30	Vanick Moradian, Danny Draven, Mark Polonia
Crew (costume & make-up)	12	Laura Gemser, Ralitsa Roth, Oksana Shevchenko
Crew (art)	11	Baron Shaver, Asen Bozilov, Pavlos Xenakis
Crew (misc)	10	Richard Pepin, Bob Bragg, Joe Castro
Crew (visual effects)	5	Scott Wheeler, Glenn Campbell, Sandell Stangl, Joseph J. Lawson, Aleksandar Yochkolovski
Release year	4	1891, 1896, 1897, 1899
Crew (lighting)	2	Todor Kostov, Sotiris Adamopoulos
Trailer likes-to-dislikes ratio	1	ratio less than 1 (more dislikes than likes)

The rules were then grouped into 3 groups by lift using k-means clustering. The resulting groups can be seen on the balloon plot in Figure 16. The colour of the balloon represents the aggregated median lift of the group and the size of the balloon shows the aggregated support. In addition to that, the number of rules in the group, the number of unique features and 5 of the most important (frequent) features can be seen.

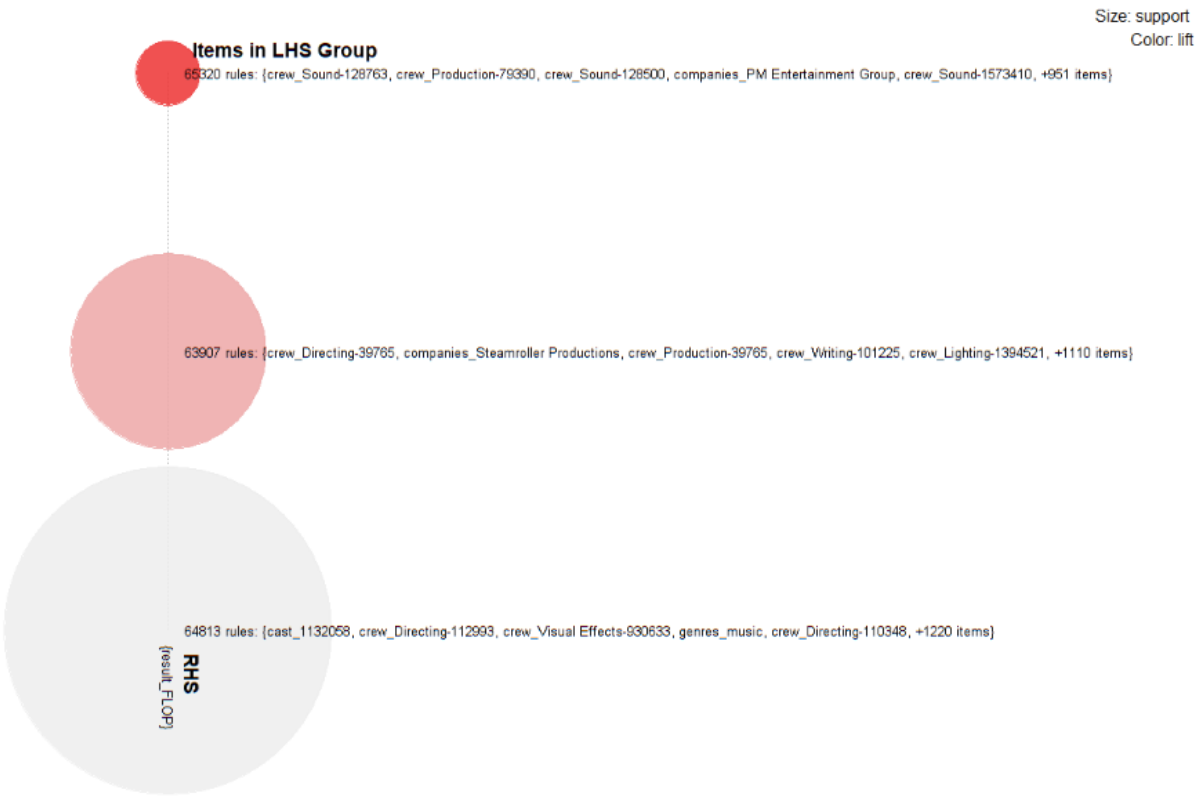

Figure 16: The association rules of unsuccessful movies by rating grouped by lift.

In the first group, there were 65,320 rules. Some of the most frequently found features of unsuccessful movies in that group were:

1. **crew_Sound-128763** – The ID corresponds to **Chris Ridenhour** who is a film and tv composer. He has been involved as the music composer in 64 movies in our dataset (e.g., Dragonquest, Born Bad, Super Cyclone) and they have all flopped. Aside from him, the movies have the following in common:

   (a) **Action genre:** In total 16,458 movies in our dataset have this genre of which 12% have been successful and 41% have flopped.

   (b) **Produced by The Asylum:** In total 176 movies in our dataset have been produced by The Asylum of which 99% have flopped.

   There are 91 movies in our dataset with these characteristics and they have all flopped.

2. **crew_Production-79390** – The ID corresponds to **Paul Bales** who is an American director, screenwriter, producer. He has been involved as a producer in 43 movies in our dataset (e.g., Ice Sharks, Monster Islands, Isle of the Dead) of which have all flopped. Aside from him, the movies have the following in common:

   (a) **Science fiction action genres:** In total 2,158 movies in our dataset have these genres of which 12% have been successful and 52% have flopped.

(b) **Not belonging to a collection:** In total 142,483 movies in our dataset do not belong to a collection of which 20% have been successful and 27% have flopped.

(c) **The length between 29 and 96 minutes:** In total 72,484 movies in our dataset have the length between that interval of which 16% have been successful and 34% have flopped.

There are 966 movies in our dataset with these characteristics and only 43 (4%) of them have been successful while 693 (72%) have flopped.

3. **companies_PM Entertainment Group** – The **PM Entertainment Group** was an independent American film production company that was closed in 2002. The company has produced 60 movies in our dataset (e.g., Alien Intruder, The Silencers, Final Impact) of which have all flopped. Aside from the company, the movies have the following in common:

   (a) **Action thrillers:** In total 4,558 movies in our dataset have these genres of which 9% have been successful and 47% have flopped.

   (b) **Not belonging to a collection:** In total 142,483 movies in our dataset do not belong to a collection of which 20% have been successful and 27% have flopped.

   (c) **Produced in North America:** In total 53,739 movies in our dataset have been produced in North America of which 17% have been successful and 32% have flopped.

   (d) **In the English language:** In total 93,251 movies in our dataset are in the English language of which 20% have been successful and 31% have flopped.

   There are 1,920 movies in our dataset with these characteristics and only 101 (5%) of them have been successful while 1,121 (58%) have flopped.

In the second group, there were 63,907 rules. Some of the most frequently found features of unsuccessful movies in that group were:

1. **crew_Directing-39765** and **crew_Producing-39765** – The ID corresponds to **Larry Buchanan** who was a film director, producer and writer. He has been involved in 26 movies in our dataset (e.g., Under Age, Sam, Strawberries Need Rain) of which none have been successful and 24 (92%) have flopped. Aside from him, the unsuccessful movies have the following in common:

   (a) **Not belonging to a collection:** In total 142,483 movies in our dataset do not belong to a collection of which 20% have been successful and 27% have flopped.

   (b) **In the English language:** In total 93,251 movies in our dataset are in the English language of which 20% have been successful and 31% have flopped.

(c) **Produced in North America:** In total 53,739 movies in our dataset have been produced in North America of which 17% have been successful and 32% have flopped.

There are 46,623 movies in our dataset with these characteristics and 7,863 (17%) of them have been successful while 15,066 (32%) have flopped.

2. **companies_Steamroller Productions** – **Steamroller Productions** is an American production company started by **Steven Seagal**. The company has produced 16 movies in our dataset (e.g., Street Wars, Code of Honor, Marked for Death) of which none have been successful and 88% have flopped. Aside from the company, the unsuccessful movies have the following in common:

   (a) **Starring Steven Seagal:** In total 63 movies in our dataset star him of which 2% have been successful and 79% have flopped.

   (b) **Maturity rating R:** In total 9,953 movies in our dataset have that maturity rating of which 12% have been successful and 46% have flopped.

   There are 33 movies in our dataset with these characteristics and none of them have been successful while 25 (76%) have flopped.

3. **crew_Writing-101225** – The ID corresponds to **Bert I. Gordon** who is an American filmmaker and visual effects artist. He has been involved as a writer in 17 movies in our dataset (e.g., The Magic Sword, King Dinosaur, Necromancy) of which 16 (94%) have flopped. For most of these, he has also been the producer and director. Aside from him, the unsuccessful movies have the following in common:

   (a) **Not belonging to a collection:** In total 142,483 movies in our dataset do not belong to a collection of which 20% have been successful and 27% have flopped.

   (b) **Produced in North America:** In total 53,739 movies in our dataset have been produced in North America of which 17% have been successful and 32% have flopped.

   (c) **In the English language:** In total 93,251 movies in our dataset are in the English language of which 20% have been successful and 31% have flopped.

   (d) **The length between 29 and 96 minutes:** In total 72,484 movies in our dataset have the length between that interval of which 16% have been successful and 34% have flopped.

   There are 26,500 movies in our dataset with these characteristics and 3,299 (12%) of them have been successful while 10,738 (40%) have flopped.

In the third group, there were 64,813 rules. Some of the most frequently found features of unsuccessful movies in that group were:

1. **cast_1132058** – The ID corresponds to **Johnny Murray** who was an American voice actor. He has been involved in 27 movies in our dataset (e.g., Hold Anything, Congo Jazz, Dumb Patrol) of which none have been successful and 21 (78%) have flopped. Aside from him, the unsuccessful movies have the following in common:

   (a) **Short animated comedies:** In total 1,136 movies in our dataset have these characteristics of which 40% have been successful and only 7% have flopped.

   (b) **Not belonging to a collection:** In total 142,483 movies in our dataset do not belong to a collection of which 20% have been successful and 27% have flopped.

   (c) **Friz Freleng in the visual effects department:** In total 39 movies in our dataset have him in the visual effects department of which none have been successful and 59% have flopped.

   There are 33 movies in our dataset with these characteristics and none of them have been successful while 23 (70%) have flopped.

2. **genres_music** – Movies in the **music genre**. There are 5,489 movies with that genre in our dataset (e.g., Billy Elliot, Mariah Carey: #1's) of which 2,305 (42%) have been successful and only 718 (13%) have flopped. The feature that makes films in this genre flop is the composer – **Frank Marsales**. There are 61 movies in which he was involved in in our dataset of which none have been successful and 74% have flopped.

3. **crew_Directing-110348** – The ID corresponds to **William Heise** who was a German-born American film cinematographer and director. He has been involved in 66 movies in our dataset (e.g., Fencing, A Hand Shake, Pillow Fight) of which none have been successful and 56 (85%) have flopped. Aside from him, the unsuccessful movies have the following in common:

   (a) **Documentary genre:** In total 13,815 movies in our dataset have this genre of which 53% have been successful and only 8% have flopped.

   (b) **Very short movies with the length less than 5 minutes:** In total 2,979 movies in our dataset are this short of which 12% have been successful and 43% have flopped.

   There are 472 movies in our dataset with these characteristics and only 17 (4%) of them have been successful while 322 (68%) have flopped.

Additionally, some feature values were manually picked to find association rules with a bit lower support, but which could still provide interesting insights. To do that, a subset of the mined rules were used by filtering the antecedent using the selected feature values. The most important features of such rules can be seen below:

1. **Movies belonging to a collection** – Horror comedies in the English language and with the length between 29 and 96 minutes.

2. **Movies, which trailers had more than 1 million views** – No rules found. Additionally, no rules were found for trailers with less than 100 views either.

3. **Movies, which trailers likes-to-dislikes ratio was higher than 30** – No rules found.

4. **Movies, which trailers likes-to-views ratio was higher than 0.004** – No rules found.

5. **The cast who were found in the rules that had confidence equal to 1** – There were 287 such persons in total, the persons who appeared in the rules most frequently and were not at the top of the groups seen previously were:

   (a) **Fernando Esteso** – All of the movies he appeared in have flopped.

   (b) **Lina Romay** – All of the movies in the English language she acted in have flopped. Out of all the movies she appeared in 96% flopped.

   (c) **Antonio Ozores** – All of the movies he appeared in along with **Andrés Pajares** have flopped. Out of all the movies he appeared in 88% flopped.

6. **The crew who were found in the rules that had confidence equal to 1** – There were 469 such persons in total, the persons who appeared in the rules most frequently and were not at the top of the groups seen previously were:

   (a) **Lisa Ries** – All of the 37 movies she was the sound editor of flopped.

   (b) **David Michael Latt** and **David Rimawi** – They are the co-founders of **The Asylum**. Of all the movies they have produced 99% have flopped.

   (c) **Rudolf Ising** – He was an American animator who created **Warner Bros. Cartoons**. All of the 13 movies he directed and which sound was composed by **Bernard B. Brown** and were animated by **Rollin Hamilton** flopped. Out of all the movies he directed 3% were successful and 48% flopped.

5.2.3 Comparison of unsuccessful movies by income and rating

To compare the similarities and differences between the flop movies by income and rating 25% of the most frequent features from the found association rules were taken. 11 features appeared in both association rule lists. The distribution of these features and their values can be seen in Table 35. The genre, length, maturity rating, belonging to a collection, trailer are all important factors in the outcome of the movie and the values seen in the table can affect the outcome of the movie negatively.

There were no features with values that appeared in unsuccessful movies association rules by income but did not appear in unsuccessful movies association rules by rating.

There were 24 features with values that appeared in unsuccessful movies association rules by rating but did not appear in unsuccessful movies association rules by income. The distribution of these features and examples of their values can be seen in Table 36. It can be seen that

Table 35: The distribution of common features in unsuccessful movies association rules.

Feature	Count	Values
Genres	2	Drama, Documentary
Continents	1	North America
In collection	1	False
Language	1	English
Length	1	Medium
Maturity rating	1	R
Release year	1	2018
Trailer comments-to-views ratio	1	Average
Trailer likes-to-dislikes ratio	1	Bad
Trailer likes-to-views ratio	1	Superb
Trailer views	1	Average

there are many features such as cast, crew, release year, production company, genre, maturity rating, length, trailer, language, which certain value combinations can affect the IMDb rating of the movie negatively, but did not seem to have that much of an effect on income.

Table 36: The distribution of features in unsuccessful movies association rules by rating but not by income.

Feature	Count	Examples
Cast	86	Eric Roberts, Danny Trejo, Lina Romay
Crew (production)	70	Paul Bales, David Rimawi, David Michael Latt
Companies	65	Warner Bros. Cartoons, The Asylum, Dania Film
Crew (directing)	51	Rudolf Ising, Hugh Harman, David DeCoteau
Release year	40	1896, 1897, 1991, 1992, 2008, 2009, 2017, 2019
Crew (writing)	34	Tyler Perry, Jesús Franco, Mariano Ozores
Crew (sound)	24	Lisa Ries, Frank Marsales, Chris Ridenhour
Crew (camera)	19	Juan Soler, Gary Graver, Alexander Yellen
Genres	17	Horror, Action, Comedy, Thriller, Science fiction, War Animation, Adventure, Family, TV movie, Western, Fantasy, Mystery, Crime, Romance, Sci-fi, Music
Crew (editing)	8	Rob Pallatina, Paul G. Volk, Cameron Hallenbeck
Continents	5	Europe, South America, Asia, Oceania, Africa
Crew (visual effects)	5	Sandell Stangl, Friz Freleng, Joseph J. Lawson
Length	4	Very short, Short, Long, Very long
Trailer views	4	Very popular, Popular, Unpopular, Very unpopular
Crew (art)	3	Petros Kapouralis, Pavlos Xenakis, Kes Bonnet
Crew (misc)	3	Richard Pepin, Vasilis Vasileiadis, William Heise
Maturity rating	3	PG-13, PG, G
Trailer likes-to-dislikes ratio	3	Great, Average, Very bad
Crew (costume & make-up)	2	Oksana Shevchenko, Athina Tseregof
Crew (lighting)	2	Sotiris Adamopoulos, Mihalis Spanoudakis
Trailer comments-to-views ratio	2	Great, Bad
Trailer likes-to-views ratio	2	Great, Average
In collection	1	True
Language	1	Foreign

5.3 Comparison of results

5.3.1 By income

To compare the similarities and differences between successful and unsuccessful movies by income 25% of the most frequent features from the association rules of successful and unsuccessful movies by income were taken. There were 19 features with values that appeared in successful movies and not in unsuccessful movies. The distribution of these features and examples of their values can be seen in Table 37. It can be seen that the production company, cast, crew, genre, maturity rating, trailer, and belonging to a collection all play a role in the success of a movie.

Table 37: The distribution of features in successful movies but not in flop movies by income.

Feature	Count	Examples
Companies	23	Marvel Studios, Danjaq, Eon Productions, United Artists, Universal Pictures
Crew (production)	20	Walt Disney, Steven Spielberg, Stan Lee, Michael Bay
Genres	12	Adventure, Action, Science fiction, Thriller, Animation, Family, Comedy, Horror, Fantasy, Romance, Crime, Mystery
Crew (sound)	10	John Williams, Dennie Thorpe, Ben Burtt
Crew (visual effects)	7	Ollie Johnston, Milt Kahl, Frank Thomas, John Lounsbery, Eric Larson
Crew (writing)	6	George Lucas, Ian Fleming, Jack Kirby, J. K. Rowling, Satoshi Tajiri, Stan Lee
Cast	3	Robert Downey Jr., Rica Matsumoto, Ikue Otani
Maturity rating	3	PG-13, PG, G
Trailer likes-to-dislikes ratio	3	Superb, Great, Average
Continents	2	Europe, Asia
Crew (art)	2	Stuart Craig, Andy Park
Length	2	Long, Very long
Trailer comments-to-views ratio	2	Great, Bad
Trailer views	2	Very popular, Popular
Crew (directing)	1	Kunihiko Yuyama
In collection	1	True
Language	1	Foreign
Release year	1	2017
Trailer likes-to-views ratio	1	Great

There were only 2 features with values that appeared in unsuccessful movies but not in successful movies by income. These feature values were documentary genre and 2018 release year. Although movies in documentary genre have really great IMDb ratings, the income of such movies is small.

5.3.2 By rating

To compare the similarities and differences between successful and unsuccessful movies by rating 25% of the most frequent features from the association rules of successful and unsuccessful movies by rating were taken. There were 14 features with values that appeared in successful movies and not in unsuccessful movies. The distribution of these features and examples of their values can be seen in Table 38. It can be seen that the production company, cast, crew, genre all play a role in the success of a movie.

Table 38: The distribution of features in successful movies but not in flop movies by IMDb rating.

Feature	Count	Examples
Cast	40	Sean Bean, Mel Blanc, Kristin Booth, Emil Sitka, Larry Fine, Chris Irvine, Stan Laurel, Oliver Hardy, David Jason
Crew (writing)	30	Woody Allen, Joseph Barbera, Jack Kirby, Felix Adler, Sten Laurel, R. L. Stine, Billy Brown, Nick George, Stan Lee
Crew (directing)	26	Tony Dow, Tex Avery, Steven Spielberg, Chuck Jones, Kevin Dunn, Hawley Pratt, Jules White, Tom Clegg, Maurice Noble
Companies	25	Ellipsanime Productions, Nelvana, Metro-Goldwyn-Mayer, Picture Palace, Warner Bros. Pictures, Marvel Studios
Crew (visual effects)	23	Phil Monroe, Milt Kahl, Ken Harris, Irven Spence, Virgil Ross, Abe Levitow, Eric Larson, Frank Thomas, John Lounsbery
Crew (production)	19	Stan Lee, Walt Disney, Sarah Finn, Kevin Feige, Fred Quimby, Hal Roach
Crew (sound)	12	Ray Parker, Jim Morgan, John Williams
Crew (art)	7	Andy Park, Hans Dreier, Robert Gentle
Crew (editing)	7	Treg Brown, Ewa Smal, Dulal Dutta
Release year	4	1943, 1948, 1949, 1954
Crew (camera)	3	Art Lloyd, Janusz Kamiński, Yûharu Atsuta
Crew (costumer & make-up)	2	Edith Head, Wally Westmore
Genres	1	History
Trailer likes-to-dislikes ratio	1	Superb

There were 22 features with values that appeared in unsuccessful movies and not in successful movies by rating. The distribution of these features and examples of their values can be seen in Table 39. Similarly, it can be seen that production company, cast, crew, genre, and trailer all play a role in the outcome of the movie.

Table 39: The distribution of features in flop movies but not in successful movies by IMDb rating.

Feature	Count	Examples
Cast	86	Bruce Willis, Vinnie Jones, Nicolas Cage, Billy Zane, James Russo, Michael Madsen, Luke Goss, Steven Seagal, Eric Roberts
Crew (production)	70	Randall Emmett, Kirk Shaw, Joseph Lai, Avi Lerner, Erich Tomek, David Winters, John Davis, Paul Bales, Charles Band
Companies	64	The Asylum, Enjoy Movies, Cannon Group, Incendo Productions, PM Entertainment Group, Castel Film, Rapid Heart Pictures, CineTel Films
Crew (directing)	51	Rudolf Ising, Robert Vince, Paul Ziller, Al Adamson, Eddie Romero, Tyler Perry, Larry Buchanan, Brett Piper, Brett Kelly
Release year	38	1896, 1925, 1978, 1986, 1994, 2009, 2018
Crew (writing)	34	Mark Atkins, Neri Parenti, Joe D'Amato, John Dunn
Crew (sound)	24	Lisa Ries, Gert Wilden, Eric Wurst, David Wurst
Crew (camera)	19	Ken Blakey, Juan Soler, Gary Graver, Cheung Hoi
Crew (editing)	8	Danny Draven, Jesús Franco, Rob Pallatina
Crew (visual effects)	5	Friz Freleng, Sandell Stangl, Rollin Hamilton
Continents	3	South America, Oceania, Africa
Crew (art)	3	Petros Kapouralis, Pavlos Xenakis, Kes Bonnet
Crew (misc)	3	William Heise, Vasilis Vasileiadis, Richard Pepin
Genres	3	Mystery, Sci-fi, Western
Trailer likes-to-dislikes ratio	3	Very bad, Bad, Average
Crew (costumer & make-up)	2	Oksana Shevchenko, Athina Tseregof
Crew (lighting)	2	Sotiris Adamopoulos. Mihalis Spanoudakis
Length	1	Very short
Maturity rating	1	PG
Trailer comments-to-views ratio	1	Bad
Trailer likes-to-views ratio	1	Average
Trailer views	1	Unpopular

6 Conclusion

6.1 Summary

The features of successful and unsuccessful movies are very important for production companies due to the cost of producing a movie. Knowing those features can help to make smart decisions in order to make a movie succeed. In the thesis, the characteristics of both successful and unsuccessful movies and differences between them were explored. Firstly, the association rules and the most prominent features of successful movies were explored. Secondly, the association rules and the most prominent features of unsuccessful movies were explored. Finally, the similarities and differences between the most frequently occurring features of successful and unsuccessful movies were explored.

6.2 Answers to research questions

The research question RQ1.1 concentrated on finding the characteristics of successful movies. It was found that movies that belong to a collection and have trailers with lots of views and likes and a small number of dislikes were often very successful. Movies that had less restrictive maturity ratings such as G, PG, or PG-13 were also more likely to succeed. In addition to that, it was found that having a good production company, cast and crew are also very important.

The research question RQ1.2 concentrated on finding the characteristics of unsuccessful movies. It was found that movies that did not belong to a collection or were in certain genres such as documentary, produced by certain companies, or included certain crew and cast had a higher likelihood to flop.

The main research question RQ1 concentrated on finding the differences between the characteristics of successful and unsuccessful movies. The successful movies had features such as certain genres, belonging to a collection, maturity ratings for general audiences, good trailers, and certain cast and crew in common. The unsuccessful movies, on the other hand, had features such as not belonging to a collection, R maturity rating, very short length, certain genres, cast and crew in common. However, a single feature usually can not be used to determine whether a movie is successful or not – it depends on a combination of features and other external factors.

6.3 Limitations

Although the thesis studies a large number of movies there are some limitations to our approach. Firstly, there was a lot of missing data. This was especially bad if the revenue or the IMDb rating was missing because that meant dropping the whole row. Secondly, the correctness of the budget values reported by the production companies are questionable – they often do not include some costs such as advertising. Due to the way the outcome of the movie is defined, an incorrectly reported value could change the result. Finally, the division of features into groups

in Chapter 4.3 and definition of successful and flop movies could have been done differently, which could have changed the results.

6.4 Future work

Although many features were used in the thesis, there are still a lot of additional variables that could be used. For example, reviews from Metacritic or data from Twitter. Sentiment analysis could be performed on the reviews and tweets, the follower counts of actors and crew could be extracted for estimating the popularity of the movie, etc.

References

[1] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[2] J. Ahmad, P. Duraisamy, A. Yousef, and B. Buckles. Movie success prediction using data mining. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–4. IEEE, 2017.

[3] D. Ai, H. Pan, X. Li, Y. Gao, and D. He. Association rule mining algorithms on high-dimensional datasets. *Artificial Life and Robotics*, 23(3):420–427, 2018.

[4] M. Al-Maolegi and B. Arkok. An improved apriori algorithm for association rules. *arXiv preprint arXiv:1403.3948*, 2014.

[5] C. Anders. How much money does a movie need to make to be profitable. *io9*, 2011.

[6] K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. de Gregorio. Prediction of movies box office performance using social media. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 1209–1214. IEEE, 2013.

[7] D. Apiletti, E. Baralis, T. Cerquitelli, P. Garza, F. Pulvirenti, and L. Venturini. Frequent itemsets mining for big data: a comparative analysis. *Big Data Research*, 9:67–83, 2017.

[8] N. Armstrong and K. Yoon. Movie rating prediction. Technical report, Citeseer, 1995.

[9] S. Asur and B. A. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, volume 1, pages 492–499. IEEE, 2010.

[10] M. Baimbridge. Movie admissions and rental income: The case of james bond. *Applied Economics Letters*, 4(1):57–61, 1997.

[11] S. Bashir, Z. Jan, and A. R. Baig. Fast algorithms for mining interesting frequent itemsets without minimum support. *arXiv preprint arXiv:0904.3319*, 2009.

[12] S. Basuroy, S. Chatterjee, and S. A. Ravid. How critical are critical reviews? the box office effects of film critics, star power, and budgets. *Journal of marketing*, 67(4):103–117, 2003.

[13] A. Bhave, H. Kulkarni, V. Biramane, and P. Kosamkar. Role of different factors in predicting movie success. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–4. IEEE, 2015.

[14] C. Borgelt. Efficient implementations of apriori and eclat. In *FIMI'03: Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations*, 2003.

[15] D. Cocuzzo and S. Wu. Hit or flop: Box office prediction for feature films. *Stanford University*, 2013.

[16] A. De Vany and W. D. Walls. Uncertainty in the movie industry: Does star power reduce the terror of the box office? *Journal of cultural economics*, 23(4):285–318, 1999.

[17] A. Elberse. The power of stars: Do star actors drive the success of movies? *Journal of marketing*, 71(4):102–120, 2007.

[18] J. Eliashberg and S. M. Shugan. Film critics: Influencers or predictors? *Journal of marketing*, 61(2):68–78, 1997.

[19] J. Eliashberg, J.-J. Jonker, M. S. Sawhney, and B. Wierenga. Moviemod: An implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Science*, 19(3):226–243, 2000.

[20] J. Ericson and J. Grodman. A predictor for movie success. *Stanford University*, 2013.

[21] S. Follows. How much does the average movie cost to make? `www.stephenfollows.com/how-much-does-the-average-movie-cost-to-make`, July 2019. Accessed: 2020-04-30.

[22] S. Gopinath, P. K. Chintagunta, and S. Venkataraman. Blogs, advertising, and local-market movie box office performance. *Management Science*, 59(12):2635–2654, 2013.

[23] M. Hahsler and S. Chelluboina. Visualizing association rules: Introduction to the r-extension package arulesviz. *R project module*, pages 223–238, 2011.

[24] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12, 2000.

[25] John Snow Labs. Country and continent codes list. `www.datahub.io/JohnSnowLabs/country-and-continent-codes-list`, 2018. Accessed: 2020-04-30.

[26] J. Krauss, S. Nann, D. Simon, P. A. Gloor, and K. Fischbach. Predicting movie success and academy awards through sentiment and social network analysis. In *ECIS*, pages 2026–2037, 2008.

[27] M. T. Lash and K. Zhao. Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems*, 33(3):874–903, 2016.

[28] B. R. Litman. Predicting success of theatrical movies: An empirical study. *The Journal of Popular Culture*, 16(4):159–175, 1983.

[29] B. R. Litman and L. S. Kohl. Predicting financial success of motion pictures: The'80s experience. *Journal of Media Economics*, 2(2):35–50, 1989.

[30] T. Liu, X. Ding, Y. Chen, H. Chen, and M. Guo. Predicting movie box-office revenues by exploiting large-scale social media content. *Multimedia Tools and Applications*, 75(3): 1509–1528, 2016.

[31] C. Marshall. How many views does a youtube video get? average views by category. www.tubularinsights.com/average-youtube-views, February 2015. Accessed: 2020-04-30.

[32] A. McCulloch. Youtube videos: What's not to like? www.socialbakers.com/blog/2234-youtube-videos-what-s-not-to-like, September 2014. Accessed: 2020-04-30.

[33] B. Meiseberg and T. Ehrmann. Diversity in teams and the success of cultural products. *Journal of Cultural Economics*, 37(1):61–86, 2013.

[34] M. Mestyán, T. Yasseri, and J. Kertész. Early prediction of movie box office success based on wikipedia activity big data. *PloS one*, 8(8), 2013.

[35] Motion Picture Association. Film ratings. www.motionpictures.org/film-ratings, 2020. Accessed: 2020-04-30.

[36] A. Mueller. Why movies cost so much to make. www.investopedia.com/financial-edge/0611/why-movies-cost-so-much-to-make.aspx, April 2020. Accessed: 2020-04-30.

[37] R. Parimi and D. Caragea. Pre-release box-office success prediction for motion pictures. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 571–585. Springer, 2013.

[38] T. G. Rhee and F. Zulkernine. Predicting movie box office profitability: A neural network approach. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 665–670. IEEE, 2016.

[39] N. Ripley. Comscore reports highest ever worldwide box office. www.comscore.com/Insights/Press-Releases/2020/1/Comscore-Reports-Highest-Ever-Worldwide-Box-Office, January 2020. Accessed: 2020-04-30.

[40] M. R. Robertson. 3 metrics ratios to measure youtube channel success. www.tubularinsights.com/3-metrics-youtube-success, September 2014. Accessed: 2020-04-30.

[41] H. Rui, Y. Liu, and A. Whinston. Whose and what chatter matters? the effect of tweets on movie sales. *Decision support systems*, 55(4):863–870, 2013.

[42] A. Salam and M. S. H. Khayal. Mining top- k frequent patterns without minimum support threshold. *Knowledge and information systems*, 30(1):57–86, 2012.

[43] M. S. Sawhney and J. Eliashberg. A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science*, 15(2):113–131, 1996.

[44] R. Sharda and D. Delen. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2):243–254, 2006.

[45] J. S. Simonoff and I. R. Sparrow. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3):15–24, 2000.

[46] W. D. Walls. Modeling movie success when 'nobody knows anything': Conditional stable-distribution analysis of film returns. *Journal of Cultural Economics*, 29(3):177–190, 2005.

[47] F. M. F. Wong, S. Sen, and M. Chiang. Why watching movie tweets won't tell the whole story? In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pages 61–66, 2012.

[48] M. J. Zaki. Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, 12(3):372–390, 2000.

[49] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 301–304. IEEE, 2009.

[50] Y. Zhou, L. Zhang, and Z. Yi. Predicting movie box-office revenues using deep neural networks. *Neural Computing and Applications*, 31(6):1855–1865, 2019.

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Janar Ojalaid**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   The power of stars: An empirical analysis of successful and flop movies,

   supervised by Rajesh Sharma.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Janar Ojalaid
15/05/2020