The Film Industry between 1985 and 2019

Project Overview

Motivations:

An explorative project, focusing on the global film industry between 1985 and 2019, at the request of a small German film production company who wants to expand onto the global market.

The client wants to understand the trends in the global film industry in order to incorporate data driven insights in their decision-making process when greenlighting projects.

Objective:

Create a storyboard providing information about the top 200 movies released each year between 1985 and 2019.

It will be used by the stakeholders when defining their global editorial line and by the production team when evaluating individual projects and their potential.

Scope:

The analysis will focus on the top 200 movies released each year between 1985 and 2019, according to IMDb.

Data Source

This data is external, as it was collected by Daniel Grijalva, a Kaggle user who scraped the data from the IMDb website. The owner of the data is therefore Daniel Grijalva, who created the data set and made it available for use online, under the CC0 1.0 Universal license. The data sets can be accessed here.

I initially wanted to use the data sets made available for non-commercial use by IMDb on their website, but they did not include any geographical variable I could use. The only geographical variable was the one showing the country where specific localised titles were used, which would prove difficult to use in my analysis as all localised titles were bundled up together under one general film ID (primary key) in the other data sets.

The data was scraped from IMDb, which is a reputable source, and this data set has been used by many Kaggle users (it was downloaded 44,4k times, used in 125 projects published on Kaggle).

As the data was scraped by Daniel Grijalva, there is some risk linked to the collection methods and data quality, but the data set has gotten a grade of 10/10 on Kaggle in terms of usability, with a Credibility (Source/Provenance, Public Notebook, Update Frequency) score of 100%. It also had 397 upvotes on the platform on the day I downloaded it (20.02.2024). Because of this, I will consider this data set to be trustworthy.

One limitation of the data is that it focuses on the movies released between 1980 and 2020 who are listed on IMDb. It means that the data does not account for the movies that were released the past 3 years. Yet, it is not a major issue. This 41 years' timeframe allows me to clearly define the scope of my analysis and focus on these 4 decades.

I chose this data set because its content focuses on the film industry, which is a field I'm passionate about and have previously worked in. Additionally, I found it convenient that the data set would focus on the movies released between 1980 and 2020, as it provides a clean timeframe for my analysis.

Data Collection:

This data was scraped from the IMDb website on by Daniel Grijalva. It is unstructured data in HTML format that has been collected automatically and formatted into a structured data set. There is no time lag as the data set focuses on 4 decades, which were over once the data was collected. This data is now **3 years old**.

The owner of the data disclosed the collection method here:

“The data was automatically scraped using a Python script, using IMDb's advanced search tool. Here's an example query using just the year and type of content (feature film):

https://www.imdb.com/search/title?title_type=feature&release_date=1986-01-01,1986-12-31&count=100

This returns 100 films from the year 1986, ordered by popularity. The script simply selects all those films, one by one, from top to bottom.

That's the only criteria really, popularity.”

After conducting a data check, it appears that the top 200 movies released each year between 1985 and 2019 were included in the data set. For the years 1980 to 1984, less movies were included (from top 92 in 1980 to top 168 in 1984). The year 2020 only included the top 25 movies.

I will make the choice to exclude the years 1980, 1981,1982,1983, 1984 and 2020, **to keep only the years 1985-2019, for which the top 200 movies released each year are included in the data base.**

Another limitation of the data set is the fact that it only provides one value per record in each column. It therefore does not list co-directors, co-writers, co-stars, multiple genres, etc.

Data Content:

The original data set consisted of 15 columns (variables) and 7668 rows. Details about each variable was provided by the owner of the data on Kaggle.

After data cleaning and consistency checks, the prepared data set consists of 7000 columns and 13 columns. All the details of the data cleaning and wrangling can be found in the “1.FI_Exploring_and_Cleaning_Data” Jupyter notebook and “Data_profile” Excel file.

Additional Data:

GEOJSON showing countries of the world. Can be accessed here.

All data is licensed under the Open Data Commons Public Domain Dedication and License.

Note that the original data has been made available thank to Natural Earth, Lexman and the Open Knowledge Foundation. All source code is licenced under the MIT licence.

Data Profile

Data Description

Variables	Time-variant / Time-invariant	Structured / Unstructured	Qualitative / Quantitative	Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous
movie_name	time-invariant	Structured	Qualitative	Nominal
MPAA_rating	time-invariant	Structured	Qualitative	Ordinal
genre	time-invariant	Structured	Qualitative	Nominal
release_year	time-invariant	Structured	Qualitative	Ordinal
grade	time-variant	Structured	Quantitative	Continuous
number_of_votes	time-variant	Structured	Quantitative	Discrete
director	time-invariant	Structured	Qualitative	Nominal
writer	time-invariant	Structured	Qualitative	Nominal
main_star	time-invariant	Structured	Qualitative	Nominal
production_country	time-invariant	Structured	Qualitative	Nominal
gross_revenue	time-invariant	Structured	Quantitative	Continuous
production_company	time-invariant	Structured	Qualitative	Discrete
runtime	time-invariant	Structured	Quantitative	Continuous

Data Cleaning and Consistency Checks

Details about the data cleaning and consistency checks can be found in the "1.FI_Exploring_and_Cleaning_Data" Jupyter notebook and "Data_profile" Excel file.

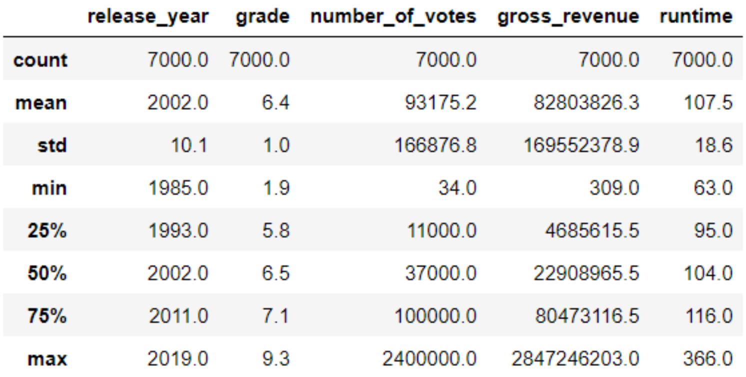
Number of columns that were deleted: 2

- budget
- released

Number of rows that were deleted: 668

- row 466 (White Star)
- row 471 (Last Plane Out)
- row 7664 (The Robinsons)
- row 7665 (More to life)
- The records with a release year between 1980 and 1984
- The records with 2020 as release year

Descriptive statistics:



	release_year	grade	number_of_votes	gross_revenue	runtime
count	7000.0	7000.0	7000.0	7000.0	7000.0
mean	2002.0	6.4	93175.2	82803826.3	107.5
std	10.1	1.0	166876.8	169552378.9	18.6
min	1985.0	1.9	34.0	309.0	63.0
25%	1993.0	5.8	11000.0	4685615.5	95.0
50%	2002.0	6.5	37000.0	22908965.5	104.0
75%	2011.0	7.1	100000.0	80473116.5	116.0
max	2019.0	9.3	2400000.0	2847246203.0	366.0

The values in the number_of_votes and gross_revenue columns differ greatly, but the minimum and maximum values were checked online and found to be correct.

No inconsistency is found here.

Data Limitations

As previously explain in the "Data collection" section of this document, I have made the choice to focus on the years 1985-2019 (35 years) to make up for irregularities in the data collection method for the other 6 years.

Another limitation of the data set is the fact that it only provides one value per record in each column. It therefore does not list co-directors, co-writers, co-stars, multiple genres, etc. From an ethical standpoint, this could create collection and exclusion biases.

I will still use this data set, as it meets the requirements for my project, but this limitation should be stated in my deliverables.

Some countries do not exist anymore (Yugoslavia, Federal Republic of Yugoslavia and West Germany). The entries listing West Germany have been replaced by Germany. The entries listing Yugoslavia or the Federal Republic of Yugoslavia have been replaced by Serbia, which can lead to some distortion and could lead to Serbia being overrepresented (Bosnia and Herzegovina, Croatia, Macedonia, Montenegro, Serbia (including the regions of Kosovo and Vojvodina) and Slovenia were all part of Yugoslavia).

Questions to explore

Clarifying questions:

- Why are movies popular?
- What kind of movies are included in the top 200 popular movies released each year?
- Why do some movies get better grades than others on IMDb?
- What kind of movies IMDb users engage with the most? *The engagement measure here would be the number of votes.*

Funnelling questions:

- Are there some genres that are more popular than others?
- Is there a length of movies that is more popular than others?
- Are specific movie ratings more popular than others?
- Does the runtime have an impact on the grades given to movies?
- Does the runtime have an impact on the number of votes a movie collected?

Adjoining questions:

- Are the grades given by viewers on other platforms (Rotten Tomatoes, Letterboxd, etc.) consistent with the ones from IMDb?
- Are there other ways of rating the success of a movie that might be more of importance for a small company like Pearplex (festival nominations, cinema releases, streaming rights being bought, etc.)?

Elevating questions:

- Why is learning about popular movies on an international scale important for the production strategy of Pearplex?

Resources

- Link to the data set here.
- Jupyter notebook used for data exploration and cleaning.
- “Data_profile” excel file.