IMDb Datasets Analysis - Proposal Report

Bruno Gonçalves Vaz
Faculdade de Ciências
Universidade do Porto
up201705247@fc.up.pt

Maria de Fátima Barros
Faculdade de Engenharia
Universidade do Porto
up201608444@fe.up.pt

Maria João Lavoura
Faculdade de Ciências
Universidade do Porto
up201908426@fc.up.pt

***Abstract*—It is common for people to choose their next movie or show through other viewers' experience statements, like the Internet Movie Database (IMDb) presents. In this proposal report, we will be inspecting the datasets publicly provided by the IMDb, cleaning and filtering them for further analysis. After this, exploratory data analysis will reveal some hypotheses worth exploring through the data, which will be described and supported. In the future, these hypotheses will be proven or refuted.**

***Index Terms*—Database, Exploratory Data Analysis, IMDb Rating**

I. Introduction

The Internet Movie Database (IMDb)[1] is a well-known source for obtaining information about movies, TV shows, celebrities, etc. People are accustomed to selecting their next film or program accordingly to their ratings and reviews. This database comprehends this and much more data. For instance, we have an insight of the crew members with details on some personal information. It also specifies different labels for each instance according to world region and even describes the duration and genres of each title.

In this proposal report, we will be analysing the IMDb datasets[2], which are publicly available and updated daily. First, a description of the dataset will be presented, detailing the features of each dataset for a better understanding of the material within it. Then, we will review the specifications of the processing techniques used for filtering and cleaning the data, making it useful for visual analysis. At last, we will introduce our proposed hypothesis based on the exploration of the data, including some of the problems we may have when trying to inspect the data in the future.

II. IMDb datasets

We were presented with 6 from the 7 available IMDb datasets, excluding the one from information about TV Series or episodes, which is not our focus. In Table I, a brief description of the information for each dataset is provided.

The original data contains up to 40M entries in some datasets, so it is mandatory that we conduct a process of filtering and cleaning of each dataset. In this section, an explanation of the filtering process for the entire dataset will be given and a following resolution of the cleaning techniques will then be described independently for each dataset, including the removal of less relevant variables and instances for our analysis.

TABLE I: Brief description of the IMDb datasets used in this project. Instances in italic in the description refer to variables' names.

Dataset	**Description**
Title Basics	This dataset contains the main information for each title. That is, the name of movie or show. It includes features such as the identity of the title (*tconst*), the most used title (*primaryTitle*), the title in the original language (*originalTitle*), the genres (*genres*) and if it's for adults (*isAdult*). It also includes the year it was released (*startYear*), and for TV series, the end year (*endYear*).
Title Akas	In this dataset, according to the *titleId* (of type *tconst*), we can find enumerated information (*ordering*) about the title (*title*) by region (*region*) and language (*language*), taking in consideration if it is the original title or not (*isOriginalTitle*). A set of numerated attributes (*types*) and not numerated (*attributes*) is also provided.
Title Ratings	For each title (*tconst*), we can find in this dataset the corresponding rating (*averageRating*), and number of votes (*numVotes*).
Title Crew	This dataset refers to the directors (*directors*) and writers (*writers*) for each title (*tconst*).
Title Principals	In each title (*tconst*), principal cast and crew are enumerated (*ordering*) and detailed with an Id (*nconst*), describing their job category in a given title (*category*), and specific job, if applicable (*job*). For actors, we may also find the name of the character played (characters).
Name Basics	IMDb also provides information about titles' celebrities. So in this dataset, for each person (*nconst*), their name (*primaryName*), birth year (*birthYear*), primary professions (*primaryProfession*), and titles the person is known for (*knowForTitles*) are described. The death year (*deathYear*) is also provided, if applicable.

A. Filtering

As stated previously, we are not going to analyse TV Series or episodes. Our main focus is on movies. In order to filter our datasets, we inspect the *Title Basics* tabular data, which contains 7 281 233 instances and includes the type of title, which can be one of the ten following options: short, movie, tvShort, tvMovie, tvSeries, tvEpisode, tvMiniSeries, tvSpecial, video and videoGame.

We filtered *Title Basics* so it only includes short, movie and tvMovie categories, resulting in a *Filtered Title Basics* with 1 443 759 rows. After this, the filtering process for *Title Akas*,

[1] https://www.imdb.com/pressroom/

[2] https://www.imdb.com/interfaces/

Title Principals, *Title Crew* and *Title Ratings* was based on the title Id's (*tconst*) present in the *Filtered Title Basics*. Then, for filtering *Name Basics*, we used the instances of persons' Id's in the now filtered *Title Principals*.

B. Cleaning

In this subsection, we will be detailing the cleaning of each dataset, as well as how to deal with some complications, such as missing values.

*1) **Title Basics**:* As stated before, *Title Basics* has now 1 443 759 rows. The feature *originalTitle* does not add relevant or useful information, since its analysis results in high computational costs. Hence, they were discarded. If we do need the name for a specific title, we can use the Id (*tconst*) and locate the titles in dataset *Title Akas*. We discovered that only 10 319 titles are adult movies. Entries without a start year were also removed (6.1%). The column *endYear* only referred to series, so it was deleted. Entries with the duration (*runtimeMinutes*) of more than 4 hours were also removed, since they are considered to be outliers and only represent around 0.067% of the instances. Finally, the instances without any value for the attribute *genres* (around 5.7%) were removed and, since the column *genres* has up to 3 values, it was divided into 3 different columns - *genre1*, *genre2* and *genre3* - each one containing only one genre and a 0 whenever there's no data. The cleaned dataset was left with 1 277 962 rows.

*2) **Title Akas**: Title Akas* after filtering has 2 211 001 rows with a variety of features depicted in Table I. The *ordering* feature is useless for our purposes, not adding any additional information, and for that, it was erased. For the *attributes* feature, most of the entries do not have this information (95%), so it was discarded. Considering the *types* feature, half of them have missing values, so it was also removed. The localized title (*title*) feature does not include additional information to the *region* and *language* features and has too many languages. It is also very difficult to analyse, since it comes in different types of alphabets, where different symbols may cause problems when reading the files. Hence, it was removed. For *region* feature, 13.6% of the entries don't have a value, so they were removed. *Language* feature also has a lot of missing data (85.1%), so this column was removed. The boolean that determines if the title is the original or not (*isOriginalTitle*) has, now, 0.0016% of missing values, which were also removed. At last, after cleaning, *Title Akas* was left with 1 911 339 entries (decrease of approximately 13.6%).

*3) **Title Ratings**:* After filtering, the Title Ratings dataset was left with 423 183 rows, these corresponding to the *movies*, *tvShort* and *short*. A careful analysis was conducted and we can state there are no *averageRatings* outside the interval [0,10], nor are negative values for the attribute *NumVotes*. Since there are also no missing values, no pre-processing was made for this dataset.

*4) **Title Crew**:* The filtering process left the Title Crew dataset with 1 443 759 instances. We observed there were around 7.8% observations with neither a director nor a writer. Since this is not useful, they were removed. Then, two new columns were created, one containing the number of directors per instance and the other the number of writers, also per instance. This was done since the only relevant information we could obtain from this dataset was, indeed, the number of writers and directors. Note that, the directors and the writers in the *Title Crew* either are in the *Title Basics*, or they are useless by themselves, since we can not apprehend any important information from them alone. The columns containing the directors and writers were, therefore, removed. We were left with 1 330 669 rows.

*5) **Title Principals**:* In *Title Principals*, we have 4 621 598 instances. The *ordering* column did not give any relevant information, so we discard it. Also, 85.5% and 62.7% of the population did not have a *job* or *characters*, respectively, so these features were also disposed. After this, we were left with the variables *tconst* for identifying the movie, *nconst* for the person in question, and their *category*. Since no rows were removed, we remained with the starting number of instances.

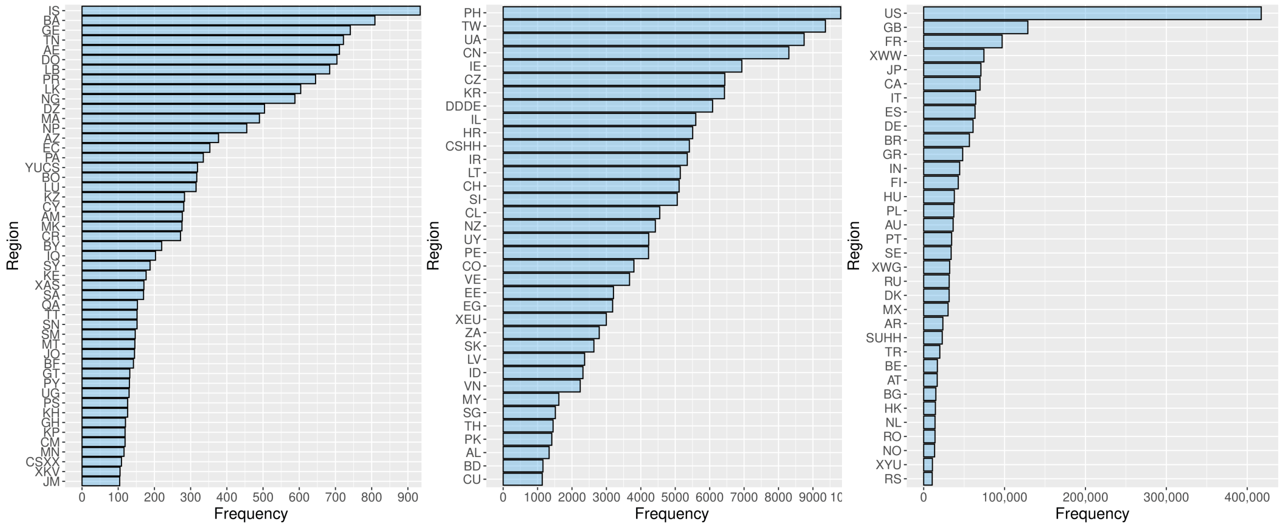
*6) **Name Basics**:* After filtering, *Name Basics* was left with 1 634 635 instances. We discovered that 85% of these did not have a birth year entrance, so we decided to discard this variable. As for the death year, we transform the data into a new column *isDead*, a boolean, that has value 1 for people that have already passed, and 0, otherwise. Considering that each person may have up to 3 professions, the *PrimaryProfession* column was replaced by 3 new columns, each with only one value (with only one *Profession*, where empty entries are imputed with 0's. This dataset also includes information about the titles the people are most known for (*knownForTitles*). We removed instances without any data with respect to this attribute (6.3%), decreasing the rows to 1 532 134. The *knowForTitles* attribute has a frequency varying between 1 and 6 different values per row. However, since only 23 rows have 5 different values of *knownForTitles* and only 2 rows have 6 different values, we discarded these and replaced the column by 4 new ones with the first four *knownForTitles* values per instance.

C. Merging

After filtering and cleaning, the *Title Basics*, *Title Crew* and *Title Ratings* datasets were left joined by the *tconst* attribute. First, we left joined *Title Basics* and *Title Crew*. Then, we left joined the resulting dataset and *Title Ratings*. Therefore, the resulting merged dataset has some instances in which the values of the attributes of the *Title Crew* and *Title Ratings* datasets are missing. In such cases, a global constant (-1) was imputed.

III. Hypotheses and Exploratory Data Analysis

After filtering and cleaning our datasets, we started to explore them and visualizing some of the features. Given that, we formulated some hypotheses we believe are worth exploring. In this section, these hypothesis will be presented, along with the explanation of our line of thought to formulate them.



(a) Range]100, 1 000].

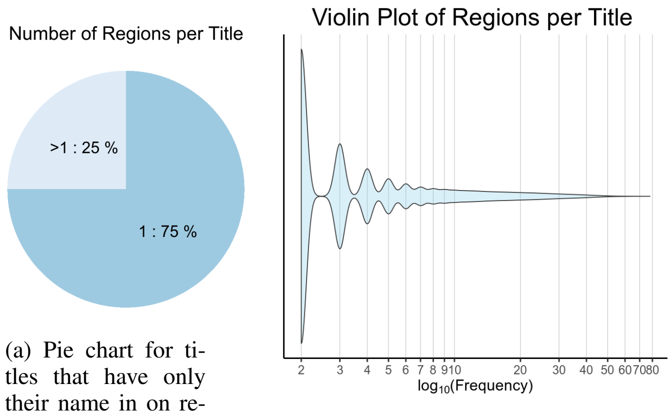
(b) Range]1000, 10 000].

(c) Greater than 10 000.

Fig. 1: Frequency of translated titles to that region. Regions with number of titles equal or less than 100 were discarded.

A. *Titles per Region*

The dataset *Title Akas* induced us to study the distribution of titles around the globe. In Figure 1, we can find a frequency distribution of translated titles to different regions. We can already detect some of the most important regions to this case, such as the United States (US), Great Britain (GB), or France (FR). The US having a much bigger value of translated titles than any other. Regions with 100 or fewer distributed titles we're not displayed, since there were too many of them and it would jeopardize the visualization.



(a) Pie chart for titles that have only their name in on region vs. more than 1 region.

(b) Violin plot distribution of unique titles with 2 or more translations.

Fig. 2: Count of regions per title exploration.

Moreover, for each title, only 25% are translated to other languages (see Figure 2a). We plotted the distribution of the number of regions per title of these 25% in a violin plot, which can be inspected in Figure 2b. From the violin plot we understand that most of the titles are only translated to two or three regions, but some of them are translated to dozens of other languages, relating to other regions in the world.

Thereafter, we created the following hypothesis: "Is the title's success (rating and number of votes) higher when it is translated to other languages?". This analysis could be achieved when combining this data with the information on *Title Ratings* dataset.

B. *Titles' Success*

1) Audience Success: While analysing the *Title Ratings* dataset we saw that most movies can be considered average (not a very high nor low score). In Figure 3 we show a histogram which supports our statement. There we can detect that most titles have ratings between 6 and 7. In addition, the average ratings of titles per person, which results from the combination of *averageRating* and *knownForTitles* features, in *Title Ratings* and *Name Basics*, respectively, is also considered average, as we can inspect when analysing Figure 4.

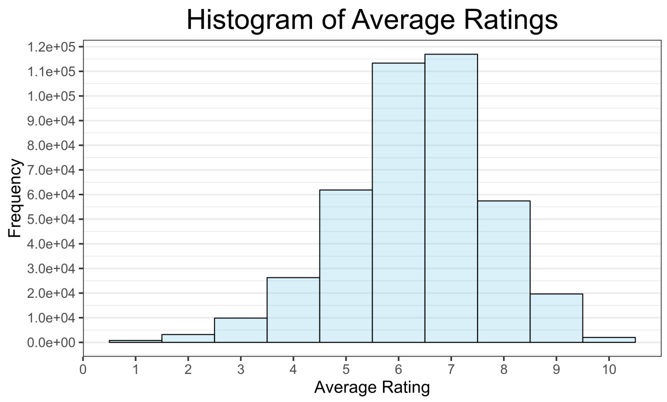


Fig. 3: Histogram of the number of Average Ratings.

Hence, we were lead to formulate the following question: ”What makes a movie successful?”. This has a lot to be said, and our future work will try to answer the question very thoroughly by analysing the importance of the cast, and the characters they impersonate in the movie, of the director(s), writer(s), the genre of the movie and the year of release, etc., by combining the data present in the IMDb datasets.

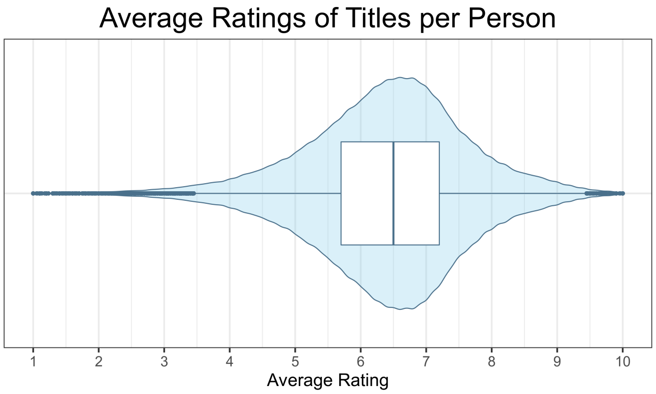


Fig. 4: Average Ratings in Titles per Person.

2) Title’s Awards: We would like to go beyond the IMDb datasets to inspect the relationship on titles and their success. One way of doing that is to analyse another important feature of the film industry: movie awards. To do so, we researched on publicly available datasets on the topic and an Emmy Awards dataset[3] was found. The Emmy Award is an American award that recognizes the excellence in the television industry. The dataset found contains data from 1949 up to 2019 and includes the following features:

- Id of the movie or show, a primary key (not related to IMDb *tconst*).
- Year of the award.
- Category of the award.
- Nominee (the name of the movie or show).
- Staff (the crew members).
- Company.
- Producer.
- Win, a boolean type variable indicating if the nominee won the award.

We were able to intersect this dataset with *Title Basics* dataset through the name of titles and their premier year.

Another popular award for film recognition is the Golden Globe. We would also like to investigate data from the previous Golden Globes for evaluating the titles. The Golden Globe Awards, 1944-2020 dataset[4] will be used for testing this. It contains similar data to the Emmy Awards dataset:

- Year of the film.
- Year of the ceremony.
- Number of the ceremony.
- Nomination (category of the award).
- Nominee (name of actor, member of crew, or title).
- Name of the title (when nominee is not title).
- Win (boolean indicating if the nominee won).

The analysis of this dataset made us question if the movies are more recognized for the motion picture itself, or the people that make the movie. In fact, more than $\frac{3}{4}$ of nominees in the Golden Globes dataset are people, and not movies, since most categories are made for the crew and cast. This can be visualized in the pie chart present in Figure 5.

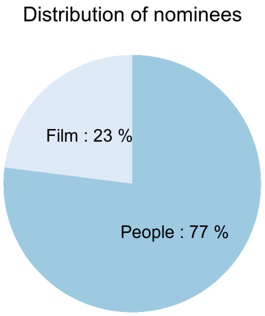


Fig. 5: Pie Chart for the type of nominees’ category (film or people) in the Golden Globes dataset.

Since most movie award categories rely on the people making the movie come to life rather than the movie itself, we wan’t to understand the following: ”Does the cast/crew of a movie make it successful?”, when it comes to ratings, votes, and even awards. For that, we will be combining the two datasets previously detailed and the IMDb datasets.

IV. Conclusion and Future Work

Dealing with large datasets is not always a straightforward task, where many processing techniques must be applied for obtaining a dataset that can be worked on properly and meaningfully. Our team transformed the IMDb datasets’ raw data into data that can be used for performing analysis and visualization in a relevant way, leaving space for further decisions upon a possible merging of the datasets, accordingly to our exploratory conduct.

From then on, we developed two hypotheses worth exploring, supported by some of the features present in these datasets. First, we would like to explore the regions the titles are translated to, and how can these be related to the IMDb success ratings. We already discovered that there are some preferences for title regions in the world, especially the US. Also, most movies (75%) only have one region in the dataset. The other hypothesis formulated is related to the success and attributes of the title itself, focusing on the cast and crew of them, which will be more explored in the future, by combining their relationships in proper visualizations. Additionally, we discovered that most movie awards are targeting people and not the movie itself, so we will find out if some people from the movie industry have a more related correlation with IMDb rating and movie awards, such as Emmys or Golden Globes, with the usage of additional datasets.

[3] https://www.kaggle.com/unanimad/emmy-awards
[4] https://www.kaggle.com/unanimad/golden-globe-awards