IMDb Datasets Analysis

Final Report - Group 5

Bruno Gonçalves Vaz
up201705247@fc.up.pt

Maria de Fátima Barros
up201608444@fe.up.pt

Maria João Lavoura
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 report, we will be inspecting the IMDb public datasets, processing and analyzing them. The main exploration will be regions titles are translated to and investigate how the success of a title relates to its cast, crew, and awards nominations and wins.**

***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 titles, TV shows, celebrities, etc. People are accustomed to selecting their next film or program according to its 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 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 explore and prove, or refute, the proposed hypothesis based on the EDA.

All the implementation was done using `R` programming language, with libraries `dplyr`, `tidyverse`, `stringr`, `ggplot2`, `igraph`, `plotly` and `maps` and is available at https://github.com/fsbarros98/IMDb_analysis.

II. IMDb datasets

We were presented with 6 from the 7 available IMDb datasets, excluding the one with 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 these processes will be given.

A. Filtering

The dataset *Title Basics* was used to filter the data through the type of title, including the categories *short*, *movie* and *tvMovie*, and excluding the remaining (*tvShort*, *tvSeries*, *tvEpisode*, *tvMiniSeries*, *tvSpecial*, *video* and *videoGame*). *Filtered Title Basics* has now 1443759 rows.

After this, the filtering process for *Title Akas*, *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 people's Id's in the now filtered *Title Principals*.

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 *5752893 rows*	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 *19922791 rows*	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 *1086028 rows*	For each title (*tconst*), we can find in this dataset the corresponding rating (*averageRating*), and number of votes (*numVotes*).
Title Crew *7281233 rows*	This dataset refers to the directors (*directors*) and writers (*writers*) for each title (*tconst*).
Title Principals *40449024 rows*	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 *6283772 rows*	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.

B. Cleaning

1) ***Title Basics****:* We discovered that only 10319 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*,

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

genre2 and *genre3* - each one containing only one genre and a 0 whenever there's no data. The cleaned dataset was left with 1277962 rows.

2) ***Title Akas****: Title Akas* after filtering has 2211001 rows. The *ordering* feature does not add any additional information, so it was erased. For the *attributes* and *types* feature, most of the entries do not have this information (95% and 50%), so they were discarded. The localized title (*title*) feature does not include additional information and it was very difficult to analyse, since it comes in different types of alphabets. Hence, it was removed. For *region* feature, 13.6% of the entries don't have a value, so they were also removed. *Language* feature also has a lot of missing data (85.1%), so this column was removed. The *isOriginalTitle* feature has, now, 0.0016% of missing values, which were also removed. At last, after cleaning, *Title Akas* was left with 1911339 entries (decrease of approximately 13.6%).

3) ***Title Ratings****:* After filtering, the *Title Ratings* dataset was left with 423183 rows, these corresponding to the *movie*, *tvMovie* 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.

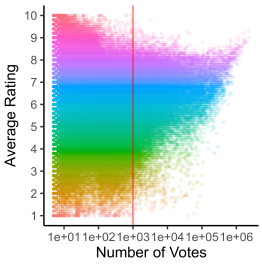


Fig. 1: Scatter Plot of Number of Votes vs. Average Ratings. Threshold of 1000 number of votes in red.

Besides this, and analysing the distribution of average ratings versus the number of votes (Figure 1, we decided to only consider entries with *numVotes* superior to 1000.

4) ***Title Crew****:* The filtering process left the *Title Crew* dataset with 1443759 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. We were left with 1330669 rows. Additionally, we created a new dataset from *Title Crew* transforming each column of writers and directors into rows, so each instance contained a *tconst*, an *nconst*, and a boolean for determining if the *nconst* referred to either a director or a writer.

5) ***Title Principals****:* In *Title Principals*, we have 4621598 instances (post-filtering). 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 1634635 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 away, 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, where empty entries are imputed with 0's). We removed empty instances (6.3%) of the *knownForTitles* feature, decreasing the rows to 1532134. 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. Exploratory Data Analysis and Hypotheses

In this section, the hypotheses we believe are worth exploring will be presented, along with the explanation of our line of thought to formulate them.

A. Titles per Region

The dataset *Title Akas* induced us to study the distribution of titles around the globe. In Figure 2, we can find a frequency distribution of translated titles to different regions. It's noticeable that the US have a much bigger value of translated titles than any other region.

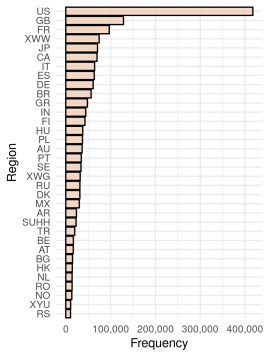


Fig. 2: Number of titles per region.

Moreover, for each title, only 25% are translated to other languages. We plotted the distribution of the number of regions per title of these 25% in a violin plot, which can be inspected in Figure 3. 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.

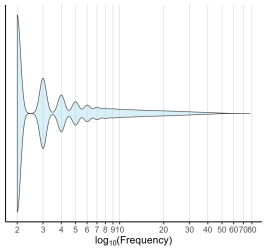
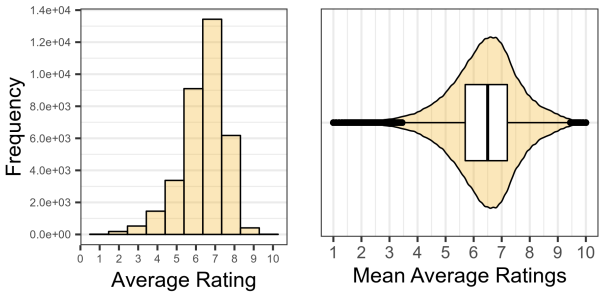


Fig. 3: Violin plot distribution of unique titles with 2 or more translations.

Thereafter, we created the following hypothesis: "Is the title's success (rating and number of votes) higher when it is translated to other languages?".

B. *Titles' Success*

1) Cast and Crew Influence: While analysing the *Title Ratings* dataset we saw that most titles have an average score. In Figure 4a 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 4b. Hence, we were lead to formulate



(a) Histogram of Average Ratings.

(b) Mean Average Ratings of *knownForTitles* per Person

Fig. 4: Average Ratings per Title/Person.

the following question: "What makes a movie successful?".

2) Title's Awards: We would like to go beyond the IMDb datasets to inspect the relationship between titles and their success. One way of doing that is to analyze another important feature of the film industry: movie awards. To do so, we researched publicly available datasets on the topic and three different datasets of interest were found : the *Emmy Awards* dataset[3] contains data from 1949 up to 2020. The *Golden Globe Awards*, 1944-2020 dataset[4],and finally, the *Oscar Awards*, 1927-2020 [5]. These datasets contain features such as the year of the film, the year and number of the ceremony, the category, the name of the nominee, the name of the title, and a boolean win.

We were able to intersect these datasets with *Title Basics* dataset through the name of titles and their premier year. The analysis of this dataset made us question if the titles are more recognized for the motion picture itself, or the people that make the movie. More than $\frac{3}{4}$ of nominees in the Golden Globes dataset are people, and not titles, since most categories are made for the crew and cast. This can be visualized in the pie chart present in Figure 5.

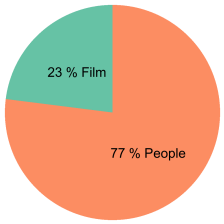


Fig. 5: Pie chart for the type of nominees' category (film or people) in the Golden Globes dataset.

Since most movie awards categories rely on the people making the movie come to life rather than the movie itself, we want 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. Hypotheses Assessment

After formulating our hypotheses, we will be exploring, proving, or refute them in this section.

A. *Titles per Region*

First and foremost, we think it's important to have a general overview regarding the distribution of the titles concerning the countries they were translated to. For such purposes, we are going to plot a world map with circles in the regions present in the dataset *Title Akas*, and whose area is given by the number of titles translated to that country. Moreover, the circles shall have different colours representing the most frequent translated genres.

To start with, it's important to remove some regions from the *Title Akas* dataset, for they do not represent countries (e.g. XWW, which means world-wide, or XEU, which represents the European continent). They represent about 8% of the rows. We also used a dataset (*Countries*[6]) with two coordinates representing each country (latitude and longitude), for placing the circles in an appropriate position for visualization purposes. Besides, we also merged the *Title Akas* with the *Merged Basics Ratings Crew* by *tconst* and transformed the resulting dataset so each region had one and only one row, it's respective coordinates and the number of titles for each of the genres. The resulting plot is represented in Figure 6.

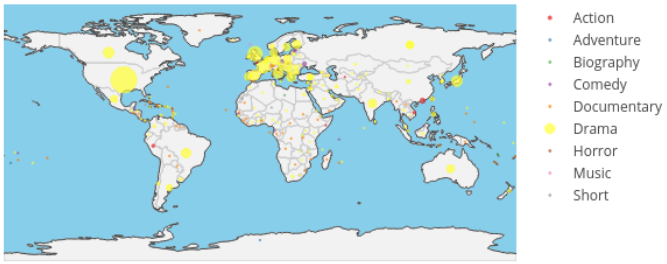


Fig. 6: World map and most frequent genres.

By analyzing the graphic, one comes immediately across the big yellow circle placed in the USA. This means that the USA are the country to which most titles are translated to and the predominant genre is drama. But drama is not only the most frequent genre in the USA. As one can notice from the number of yellow dots, this genre is the most translated worldwide. One can also observe that the more developed

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

[6] https://www.kaggle.com/paultimothymooney/latitude-and-longitude-for-every-country-and-state

countries tend to have more titles translated to, in comparison to the undeveloped countries.

Having this type of graphic led us to the idea of representing the links between regions. That is, given a movie, we assess to which regions it was translated to and we connected them by an edge, 2 by 2.

From the *Title Akas*, we removed countries only translated to one region, since they wouldn't be useful for our purposes and we transformed the resulting dataset, so each movie had one and only one row, the regions to which it was translated to and the number of such regions. Having this, we were able to create a new dataset which represented the connections between regions and the frequency of these connections. That is, each row represented a unique connection: as attributes we had two regions and the number of links between them (the number of titles that were translated to both regions). In Figure 7, we can see the final result.

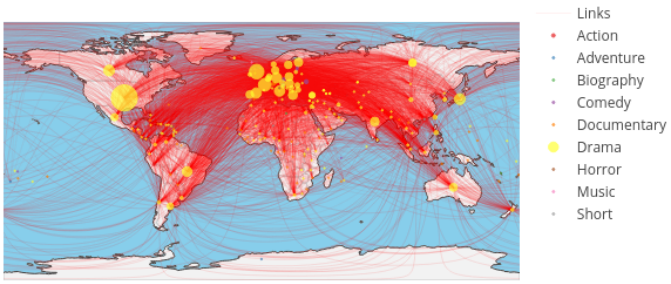


Fig. 7: World map with connections between countries and most frequent genres.

As we can see, the European continent has a lot of connections with the remaining countries. This means that the titles that are translated to more than one region are typically translated to some European country. An interactive version of this graphic and a spherical one is also available[7].

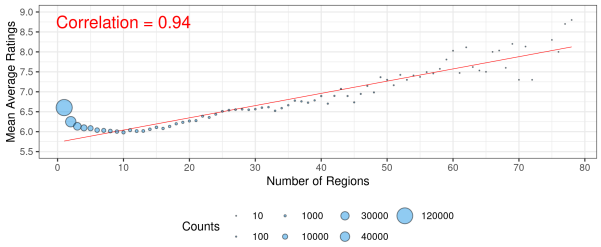


Fig. 8: Correlation between number of regions and mean average ratings.

Having this general overview of the distribution of translated titles and the respective connections between countries, it is important to analyze how the number of regions and the mean average ratings correlate. By bringing the *Title Ratings* dataset, we can, for each region, make the mean of the *averageRatings* of all the titles that were translated to that region, to plot the graphic in the Figure 8. In this graphic, the area of the circles represents the number of titles that were translated to a certain number of region.

We first take notice of the big blue circle, representing the number of titles translated to only one region, which are the majority. As the number of regions increases fewer titles are translated to that many regions, but the mean of their ratings increases. If we focus just on the titles that were translated to 40 or more regions, we expect (by the analysis of the world maps) that most of them will be translated to European countries. Indeed, that its true, as we can see in Figure 9.

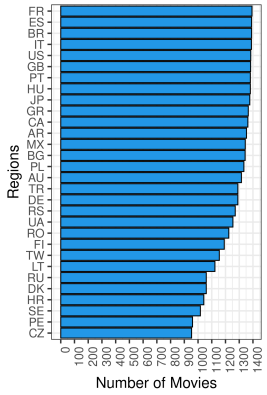


Fig. 9: Frequency of regions for titles translated to more than 40 regions.

Returning to Figure 8, we can see the correlation is very close to 1, meaning that titles with higher ratings are translated to more regions. This gives a very clear answer to our question. Indeed, the success of a movie is higher when translated to other languages.

B. Title's Success

1) Cast and Crew Influence: To explore the hypothesis "what makes a movie successful?" we analyzed the datasets and focused on answering more specific questions related to the hypothesis like "Does the cast of a movie influences its success?" and "Is there a specific profession that influences the success of a movie?".

To answer these questions, we started by defining "success". We looked into the distribution of the number of votes (Figure 1) and the histogram of the average ratings (Figure 4a). Since the rate goes from 1 to 10, a successful movie is one that has an average rating equal to 7 or more and an unsuccessful movie is one that has an average rating equal to 3 or less. Therefore, we filtered the *Merge Basics Ratings Crew* dataset with instances that had *numVotes* superior to 1000 and *averageRating* superior or equal to 7 or inferior or equal to 3.

Usually, the most commonly known cast professions for a movie are actors, actresses, directors and writers. To know how a cast influences a movie's success, we filtered the *Name Basics* dataset keeping at least one of the 3 columns of primary profession equal to one of previously enumerated most commonly known professions.

We explored this dataset deeper and realized that there are some huge values regarding the number of people that are known for a specific movie. People being known for a single movie goes from 1 to 3087. Even though the whole crew of a movie can be as big as 3087, is difficult to accept that the whole crew gets to be known by it. Moreover, for visualization purposes, we filtered it by the top 10 frequencies. In other words, we kept a maximum number of 10 people being known for a movie (Figure 10).

[7]https://github.com/fsbarros98/IMDb_analysis/tree/main/graphic_results/regions_images

This histogram represents, in the x-axis, the number of people known for a movie and, on the y-axis, the number of titles for each frequency of people. So, there are roughly 4.3×10^5 titles with 1 person that is known for that movie, 2×10^5 titles with 2 people known for it, and so on. In the same line of thought, for visualization purposes, we also looked only at the cast that is still alive.

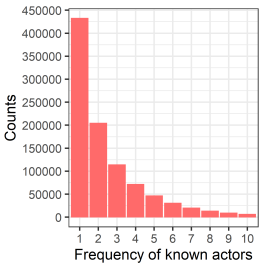


Fig. 10: This histogram is the top 10 frequencies of the original one, that goes up to 3087 people known for a movie.

In Figure 11 is represented a graph where each node is a movie and each link is a casting person in common. The red nodes represent unsuccessful titles and the green nodes represent successful titles. The links are colored based on the primary profession of each person, with actor/actress being red, writers being blue, directors being black, producers being orange and other professions being grey. To analyze titles that have more connections between them, a filter was applied where a person has to be known for 3 or 4 titles. In this graph are only represented 11 unsuccessful titles against 753 successful ones. There are a lot of smaller groups of 3, 4, 5, 6, 7, or 8 titles bonded together, some median chains bonded by 9, 10, 12, 13, or 15 titles, and 2 bigger chains of 21 and 27 titles connected. The unsuccessful titles are present in small size chains (groups of totals of 3, 4, 5, 6, or 8 titles connected). The successful titles are present in all sizes of chains. Therefore, we can conclude that a successful movie tends to have more common people with other successful titles, than with unsuccessful ones.

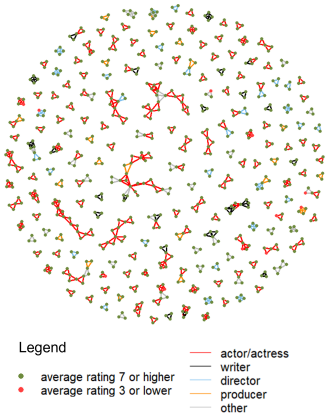


Fig. 11: Graph showing the relationships between cast and crew of the titles they are in it.

Regarding the links in the graph, there is a clear majority of actors and actresses, followed by other professions and directors, writers and lastly producers. The 3 and 4 node chains have all their links of the same color, in other words, the nodes are connected by the same profession. The 5 to 8 node chains are connected by 2 different link colors, so 2 different professions. The medium and big chains are the ones that gather up to 3 different professions within their connections. The bigger chains (21 and 27 titles connected) and medium chains (from 9 to 15 titles connected) are dominated by actors/actresses. Therefore, we can conclude that the most relevant profession in successful titles is the actor/actress.

We also explored the relationships amongst the crew (directors and writers) within a title and their relationships in successful titles, by merging the contents from *Title Crew* and *Title Ratings* and making a tree graph with the contents present in Figure 12. Each vertex represents either a director (red square) or a writer (green circle), and the nodes are titles with an average rating above 7, differing in color according to this feature. This graph excludes all isolates and portraits only the maximum crew members that are connected through at least one title. A division is clear amongst this data: on the left, the directors relate themselves through some titles, and then these also relate to some writers, and this part comprehends the titles that include more than one director or writer per title, mostly. On the other side (right), a significant part of the directors only relate themselves to one writer, which makes the realization that most of the titles portrayed on this side of the graph only have one director and one writer. Given that most of the connections are blue (average ratings between 7 and 8), we conclude that there is no specific kin between the number of members of a crew and the title's average rating above 8. Despite this, one specific director (second on the top line) is heavily related to many other directors, which can indicate that making a movie with this specific director may lead to success since all titles represented can be considered successful.

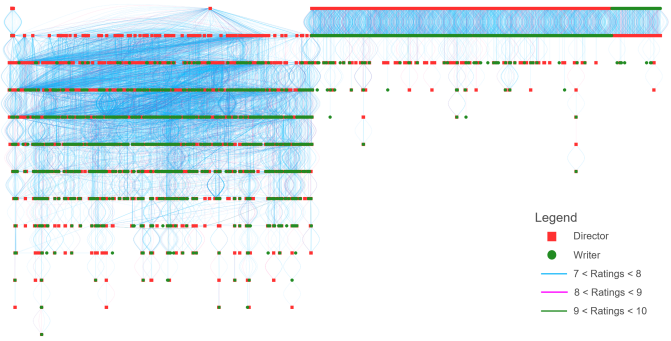


Fig. 12: Tree Graph showing the relationships amongst the crew of successful titles.

2) Title's Awards: The *Emmy*, the *Golden Globes*, and the *Oscar* awards datasets have 1649, 3808, and 8997 correspondences with *Title Basics* dataset, respectively. In Golden Globes dataset, 727 correspondences are nominations to titles itself, and 3081 nominations are for people in a certain title.

After this merge, we also full joined the dataset with Title Ratings. Hence, we were able to display the distribution of the average ratings for the titles with zero nominations, one or more nominations, and one or more winnings. This is present in Figure 13 and shows that titles with no nominations have a distribution of ratings with much wide range than those which had nominations, where the majority of titles with nominations and wins have their IMDb average ratings above 5. Also, the distribution of titles with nominations and with winnings is very similar, although titles with wins only account for 27% of the titles with nominations.

It is interesting to notice how outliers from the distribution of titles with nominations do not appear as much in the titles with wins.

Considering only the titles that won an award, we also inspected how their genres may vary. To do so, we plotted the number of titles with each genre (Figure 14, on the left), and concluded that drama is the most popular genre for titles that won any type of award studied here, followed by romance, comedy, and biography. We separated these features into titles with 1, 2, or 3 genres, resulting in the middle graphic from Figure 14. This division only affects the distribution inside each bin. From this, we can conclude that drama appears mostly in titles with three genres, but it also has a considerable part in titles with two or even one genres. It is also the most popular genre across titles with only one genre. After this, we evaluated if the position of the genre (first, second, and third) is an important feature of this analysis. This division is present in the right graphic of Figure 14. From this, we can take that most drama titles have this genre as their first genre, but a considerable amount has genre in the second position. Also, most of the biography titles have this genre as their first genre, which indicates that the position of the genre is important for defining the type of the movie. Romance does not appear frequently as the first genre of a movie.

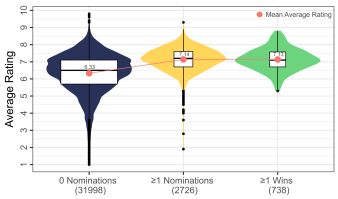


Fig. 13: Distribution of Average Ratings across titles with no nominations (right), with one or more nominations (center) and with one or more wins (left).

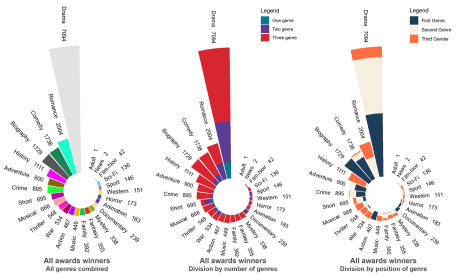


Fig. 14: Circular bar plots of genres in titles that won any award.

Going further, we separated the right graphic of Figure 14 according to the awards, resulting in the graphics from Figure 15. Drama is still clearly the number one. The second place is dominated by romance unless for the Emmy winners, where short comes after the drama. Hence, not considering drama, a movie is more likely to get a Golden Globe for the people if it is a romance and for the movie, if it is a comedy.

Since we could separate the nominees that were people from the Golden Globes dataset, we wanted to inspect how they relate to each other through the titles they were nominated for. The resulting graph showing these relations is present in Figure 16, where each vertex is a nominee and each node is

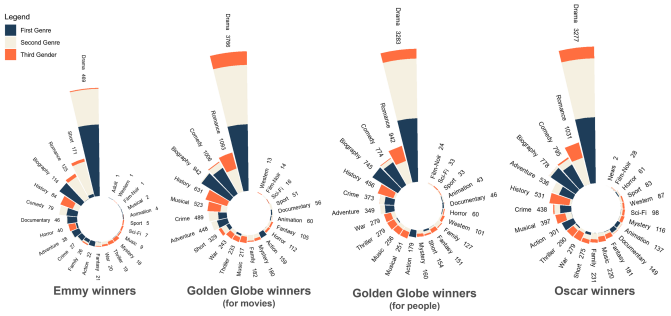


Fig. 15: Circular bar plots of genres in titles that won different awards.

a title. Nominees that did not win the award are more likely to be connected to other nominees that also did not win the award, regardless of the movie ratings, as we can see some aggregated red vertexes. Also, some nominees participate in titles with high ratings but never won an award, which is interesting. Besides this, as previously mentioned, there is no relevant connection between the rating and the winning, all nominees have medium-high ratings.

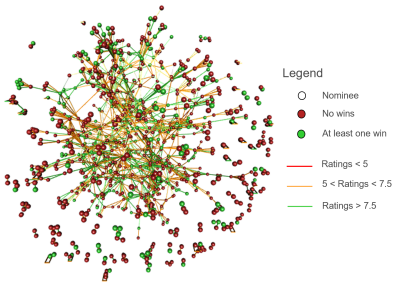


Fig. 16: Graph showing the relationships between nominees by the nominated titles.

V. Conclusions

Concerning the regions of titles, the US is the country to which most titles are translated. Furthermore, titles that are translated to more than one region are typically translated to a European country. As the number of regions increases, the average ratings get higher and the frequency of translated titles diminishes.

Concerning cast and crew, from the analysis performed we concluded that actors are the most relevant people for characterizing the title's success, that is, the average ratings. Despite this, there may be some specific members of the crew, more exactly directors, that can be related to a title's success.

Titles that win any type of the awards explored (Emmy, Golden Globes, or Oscars) have significantly higher ratings. Drama is clearly the predominant genre for titles that win awards, and genre features may determine the type of nominee of an award.

Taking all of this into consideration, our analysis responds to how can a title be successful according to the studied terms.