Data Visualization Project Proposal – Group 10

Deborah Dore
up202202823@fc.up.pt

Rebekka Schade
up202203042@fc.up.pt

Mariana Barbosa
up201508455@edu.fc.up.pt

1 Introduction

Our research goal was to explore trends and changes in the movie industry during Second World War, with a focus on Germany and Italy. These two countries were the main European partners in the Axis powers and are viewed as paradigms of fascist cinema due to the strategic importance and prestige reached by their national film productions in the 30s and 40s (Pena-Rodríguez, 2012). After cleaning, filtering and merging the datasets, we proceed to an exploratory data analysis, which allowed us to identify which variables - besides the star year of the movies being within the WWII period – could give us insights on our research goal, thus orienting our hypothesis formulation. As for our hypothesis, we began by wondering if there was any interesting change in the distribution of genres during WWII period (4.1) This analysis led us to select less genres for the second hypothesis and introduce the analysis by country (4.2) which in turn led us for the specific exploration of the genre war, within genres and within countries (4.3)

2 Cleaning, filtering & merging

For each dataset, we performed cleaning and filtering. In every dataset for the columns `tconst` and `nconst` we checked that every entry consists of "tt" or "nm" followed by 7 or 8 digits and then transformed it to integers corresponding to the digits. Missing values denoted as "\\N" or empty character were transformed to *NA*. Unless stated otherwise, we didn't find missing values.

- The **Title Basic** dataset contains informations about tiles, in particular their *original title*, *run time minutes*, *genres*, *start year* and *end year* if we're talking about a tv series. It has 9.267897 million rows and 9 columns. This dataset contained a very high amount of NaN values that we decided to drop.
- The **Title Akas** on the other hand dataset contains an entry for every different version of the original title, including the different languages in which the title was distributed. It has 33.426498 million rows and 8 columns. An important problem that must be highlighted regarding this dataset is the high percentage of NaN values in the *language* column which we solved using a classifier that was able to understand the language based on the original title. We dropped the *ordering* column and the *isOriginalTitle* column for redundancy since if a title is original is already specified in the column *types*.
- The **Title Episode** dataset with 6.991827 million rows and 4 columns contains the TV episodes for every series present in the *title basics* dataset. As usual, a high number of errors and NaNs: the column *season* for example, should have contained the number of the season of the episode but in more than 30% of the cases, it contained the year it was hired. We solved that problem.
- The **Title Ratings** dataset contains the average rating for each title and the number of votes. The only problem we identifyied with this dataset was the high value of NaNs.
- The **Name Basic** dataset contains additional information about persons or institutions such as bands and choirs. It has 11.972689 million rows after cleaning and originally 6 columns. One row which had a missing `primaryName` was deleted. Of the columns `birthYear` and `deathYear` more than 95% are missing values, so we deleted the columns. The column `primaryProfession` has 21.5% and `knownForTitles` has 17.7% missing values. We decided to leave them in the dataset for now. One entity has up to 3 professions and is known for up to 6 titles.
- The **Title Principals** dataset links the titles with its principal persons and the role they played in the production of the title. It has 52.370246 million rows and 6 columns before, 3 columns after cleaning. We considered the column `ordering` redundant and deleted the column. For the column `category`, we renamed the entries "actress" and "self" to "actor" and merged the categories "archive_footage" and "archive_sound" to "archive". Afterwards, 12 unique categories remained. The columns `job` and `characters` specify the column `category` which is redundant for our quantitative analysis so we deleted the columns. Also, we created a boolean column indicating whether someone played themself in a title using information from both `category` and `characters`.
- The **Title Crew** dataset links each title with its director(s) and writer(s). Before cleaning, it had 9.267897 million rows and 3 columns. Many titles had neither a director nor a writer and after deleting these rows, 5.982843 remained. We added two columns indicating the number of directors and writers per title. Some title had an implausibly big amount of directors (up to 491) and/or writers (up to 1319) and we suspected wrong data here. If we had used this column, we would have deleted these rows but we decided to keep all information for now.

To conclude this section, we merged the cleaned datasets `title.basics`, `title.akas`, `title.ratings` and `title.episode` by the column `tconst`. After filtering for original titles and removing duplicate cases, there were still 164 rows containing duplicate values for `tconst`. We fixed the issue and build an algorithm to decide which duplicates to drop.

3 Exploratory Data Analysis

For further analysis, we filtered our data to only original titles so we have only one row per unique production. Also, we selected only movies, shorts, TV movies and TV series because we especially TV episodes inflate the number of titles and every TV episode is already represented in a TV series. We dropped titles with a start year before 1897 because there were only a few titles (1133) in the data and we also dropped titles with a start year in the future after 2022. When we analyze the popularity of the titles, we have to remove missing values which constitute about 65% of the remaining rows.

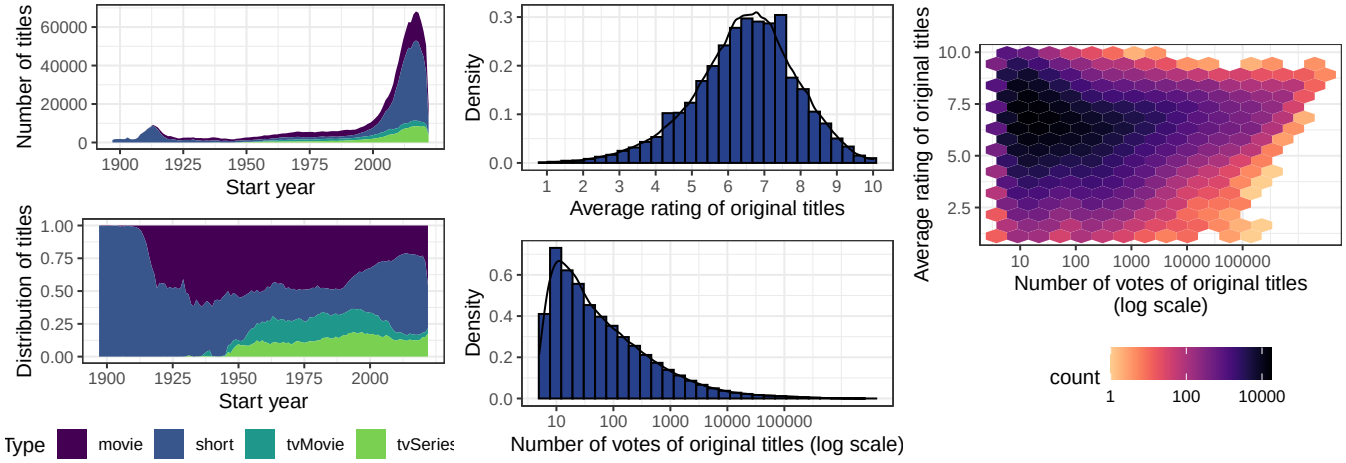

Figure 1: Exploratory Data Analysis

The number of movies, shorts, TV movies and TV series in the dataset increases drastically since the 2000s. Until the 1920s, almost every title is a short, then more movies come and since the 1950s the release of TV formats increases. The distribution of the average rating is relatively symmetrical with the mean at 6.5 while the distribution of the number of votes is extremely left skewed. Among the titles with more votes, there is a light exponential relation (positive linear on the log scale) between the number of votes on the log scale and the average rating.

4 Can we detect changes in the movie industry during 1933-1945 in Italy and Germany?

In the following plot we focus on the period from 1933-1945. In Italy, the fascist regime started in 1922 and in Germany the Nazi regime started in 1933. World War II started in 1939 and ended in 1945. We chose the 1933-1945 time period because it has included both the Nazi regime in Germany, the fascist regime in Italy and WWII.

4.1 Was there a change in the distribution of genres during 1922-1945?

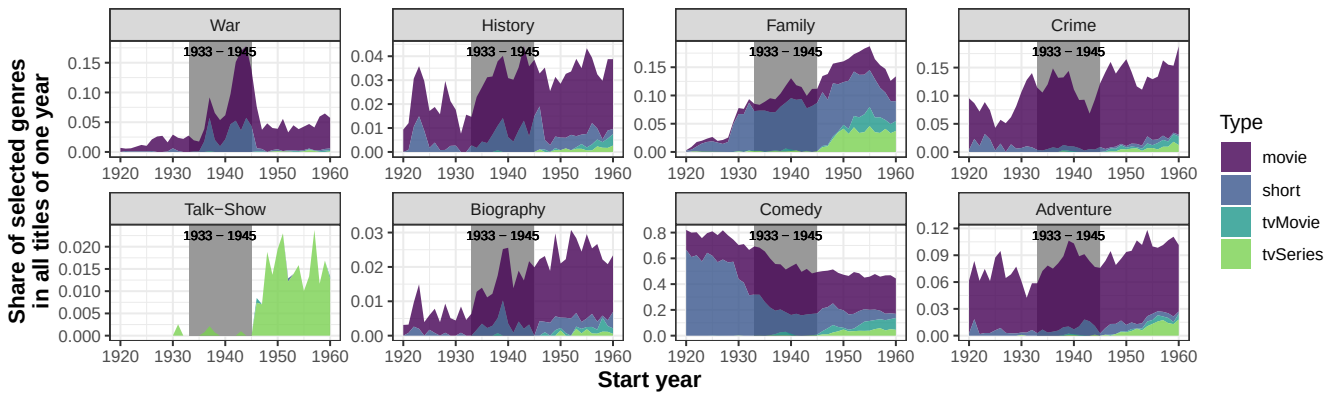

Figure 2: Development of the distribution of genres among original movies, shorts, TV movies and TV series

Most striking is the peak in the genre *War* during World War II. In particular, there were many shorts produced during this time. In hypothesis 4.3 we look further into this genre. The share of the genre *History* dropped around 1930 and then had a small peak in WWII. The share of the genre *family* increased promptly right before the period of interest. Interesting as well is the obvious short-term decrease in the share of the genre *Crime* during WWII. The genre *Talk-Show* became popular just after the ending of WWII.

4.2 Can we detect a different development in the period 1920-1960 between titles in Italian, German and other languages?

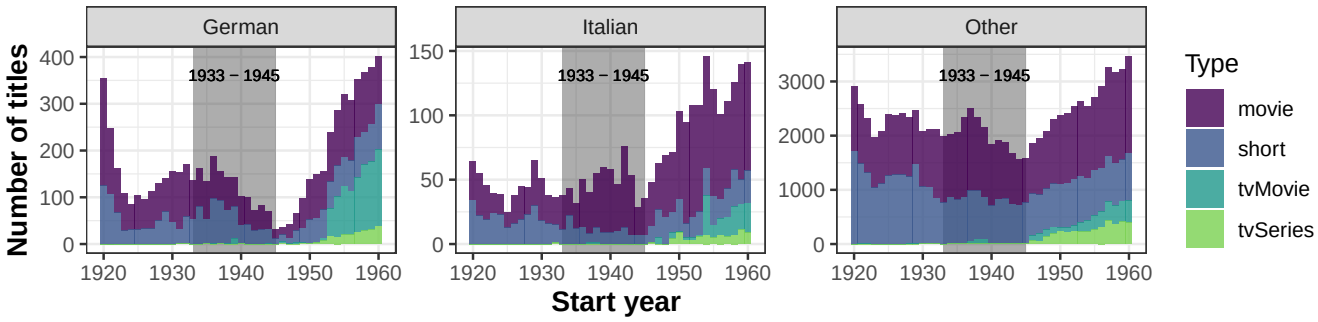

Figure 3: Distribution of each title with focus on Italian, German and other languages

We wanted to further investigate if there was a difference between titles in Italian, German and other countries during 1933-1945. We can see from Figure 3 that the production of movies for all the languages started to increase in the second half of the 1920s and decreased considerably until the end of World War II. Italy's decline is irregular: in 1940, when the production of movies continues to rise, it joins the war as a German ally and in 1943, when the production starts to decrease, it switches sides and fights alongside the Americans.
For other languages, the production in 1960 has increased by about a third compared to the level before World War II, the production of Italian and German titles has about tripled. In particular we can see that during the Nazi regime (1933-1945), the number of German movies per year was higher compared to Italian movies during the fascist regime (1922-1945).

4.3 Is the genre *War* represented differently in German and Italian titles compared to other languages during 1933-1945 and recently?

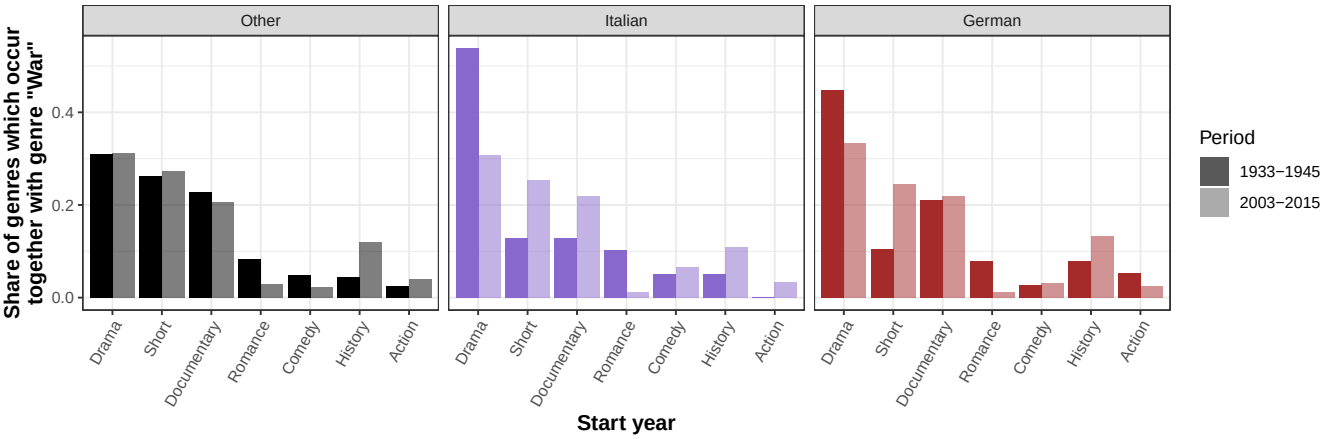

Figure 4: Representation of the genre *War* in titles in German, Italian and other languages

In other languages, the representation of the genre *War* was not very different between the two periods of interest. We can see that for all languages, the genre *History* had a larger proportion in 2003–2015 than it had in 1933–1945. For Italian and German titles, the plot shows a huge difference. During 1933-1945, *War* was much more represented in Drama and Romance titles while during 2003-2015, its share in Shorts, Comedy and History increased. We can assume that the representation of *War* in the languages of countries with fascist regimes was more emotional.

4.4 Are movies and shorts in Italian, German and other languages made during 1933-1945 rated differently?

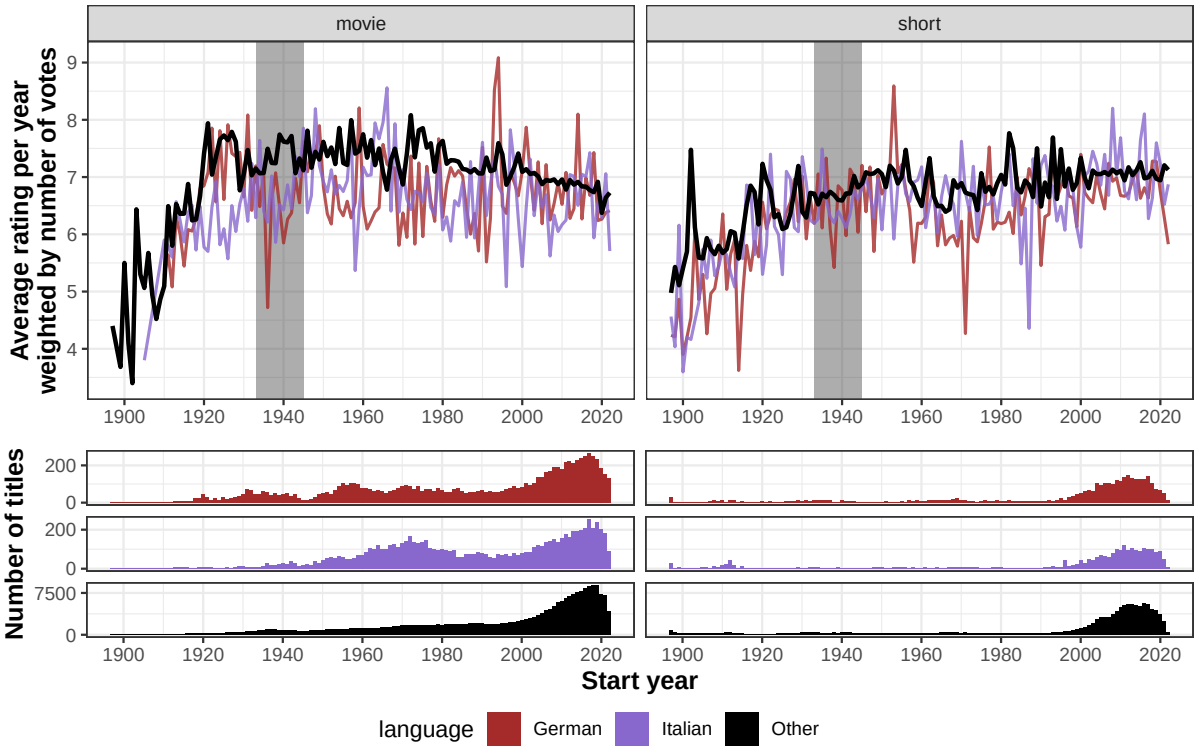

Figure 5: Popularity of movies and shorts

We were interested if the patterns we detected in the number of titles were also visible in the popularity. As shown in Figure 5, the ratings of German and Italian movies during their fascist regimes were considerably lower than movies in other languages. The ratings of movies in other languages slightly decreased over time. In recent decades, the average ratings of German and Italian movies have reached about the same level as other languages. As for shorts, there is no obvious difference in the development of the average ratings between Italian, German and other languages. The information about the ratings in recent years are more reliable because there were many more titles rated as we can see in the histograms below. This also explains why the variance in the yearly average ratings decreases in recent years, especially among shorts.

5 Conclusion & Future Work

Given our results, it seems that there are visible differences for the period 1933-1945 for titles in Italian, German and other languages. In section 4, Figure 4, as already discussed, we can see that titles of the genre *War* had a preference of being Dramas in Italy during 1933-1945 compared to 2003-2015. This phenomenon is weaker in Germany and not present in other languages. In Figure 5 we analyzed the popularity of the titles and the plots showed that there is not a big difference in the popularity of German and Italian movies compared to other languages. In conclusion, our work had the goal to find patterns in original titles with focus on Italian and German titles, specifically during fascist regimes in these countries.
For future work, it would be interesting to analyze the frequency of each genre per country during WWII compared to recent years, to find out if there is a peak of drama movies during the war period (independently of intersecting with the genre war), since it could be related with the financial effects of war (Drama is the cheapest genre to produce as movies don't necessarily require special/visual effect).