Milestone Report Capstone 2

I. Problem statement

According to Global English Editing, India, Thailand, and China are the countries with best reading habits in 2018, spending on reading between 8:00 and 10:42 hours per week. The U.S., with 5:42 hours, is in 22nd position. 37% of American adults with a high school degree or less and 7% of college graduates have reported not reading a book in any format in the past year.

In the age of Netflix and instantaneous entertainment, seems to be people is still reading, but it is hard to keep the people interested in reading classic, historical and universal literature. How we can improve this reading behavior? Past year, Netflix developed around 50 literary projects (turning novels into series). Cinematographic adaptations bring books to people and could inspire them to read the books later.

This project tries to find relations between movies and books: genres, type of fictitious worlds, the extension of the stories, historical contexts, physical places where the plots are developed, for instance. This kind of relationships could help us to recommend books depending on the people preferences or give us a guide about how to predict the success of the metamorphosis from novels to cinematographic adaptations and discovering what genres would achieve more acceptance by the audience.

Additionally, we'll use historical contexts and physical spaces where books were built for mapping the documents in innovative and educational ways: if you don't show curiosity of books, at least you can connect them with historical events and understands the links with the present. The goal is to achieve a representation of different cultures throughout the literature and looking for innovative visualizations to show the results.

The inspiration for this project is educational. We are trying to understand the impressions that books generate in population to find innovative ways to teach them and bring books to them. Also, we'll use all the available information on our database to predict trends related to how to decide what books could be interesting purposes to cinematographic adaptations. The educational sector, companies associated with book sell and interested in turning books into audio-visual projects, could be interested in these results.

II. Description of the dataset

Data Wrangling in Movies

In this section, we import two files of **the movies Dataset: *movies_metadata*** and ***keywords***.

After importing the metadata, useless columns are dropped off the dataFrame as *budget*, *home page*, *poster path*, *revenues* and information about *videos*. Instead, we put attention in *movie id*, *original languages*, *title, overview, release date*, *genres*, *runtime*, *spoken languages*, *vote average*, *vote count* and *popularity*.

Some regular expression techniques are applied to get the *movie id* and *genres* because all of them are included as part of dictionaries of information. We note that one particular movie has two or more *genres*, then we pivot them, adding 18 columns to the dataFrame with the respective genre names: *Action, Adventure, Animation, Comedy, Crime, Documentary, Drama, Family, Fantasy, Foreign, History, Horror, Thriller, Music, Mystery, Romance, War, Western.*

Secondly, the *keywords* associated with the *movie id* are imported. Using regular expressions we extract a list of words for every movie and finally, both dataFrames are merged and saved in Data interim folder. An illustrative pipeline of the process is displayed in Fig. 1.

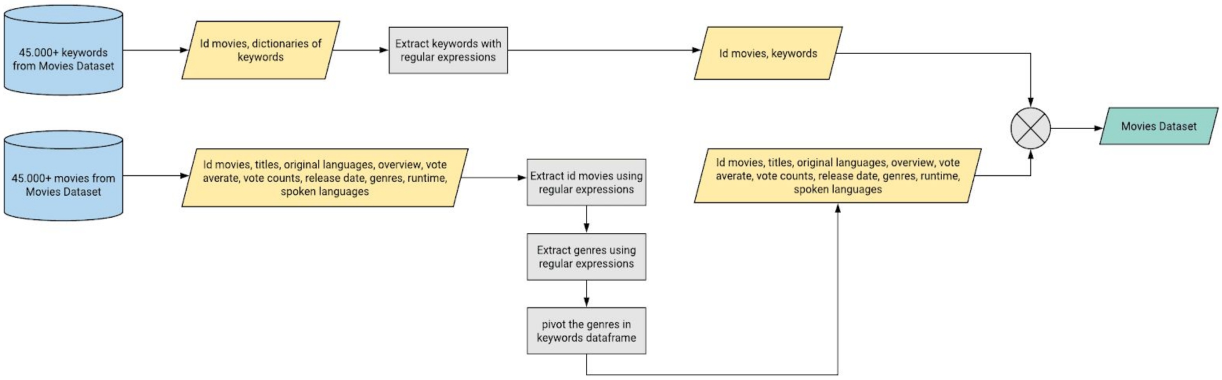

45.000+ keywords from Movies Dataset
Id movies, dictionaries of keywords
Extract keywords with regular expressions
Id movies, keywords
Movies Dataset
45.000+ movies from Movies Dataset
Id movies, titles, original languages, overview, vote averate, vote counts, release date, genres, runtime, spoken languages
Extract id movies using regular expressions
Extract genres using regular expressions
pivot the genres in keywords dataframe
Id movies, titles, original languages, overview, vote averate, vote counts, release date, genres, runtime, spoken languages

Fig. 1: Pipeline Movies Wrangling

Data Wrangling in Books

Book data consists of a dataset with general information about books, useful to discover languages, number of pages, publication years and an average rating of a wide variety of books. Additionally, we applied a pre-processing on the reader reviews dataset to get some NLP metrics thought sentiment analysis.

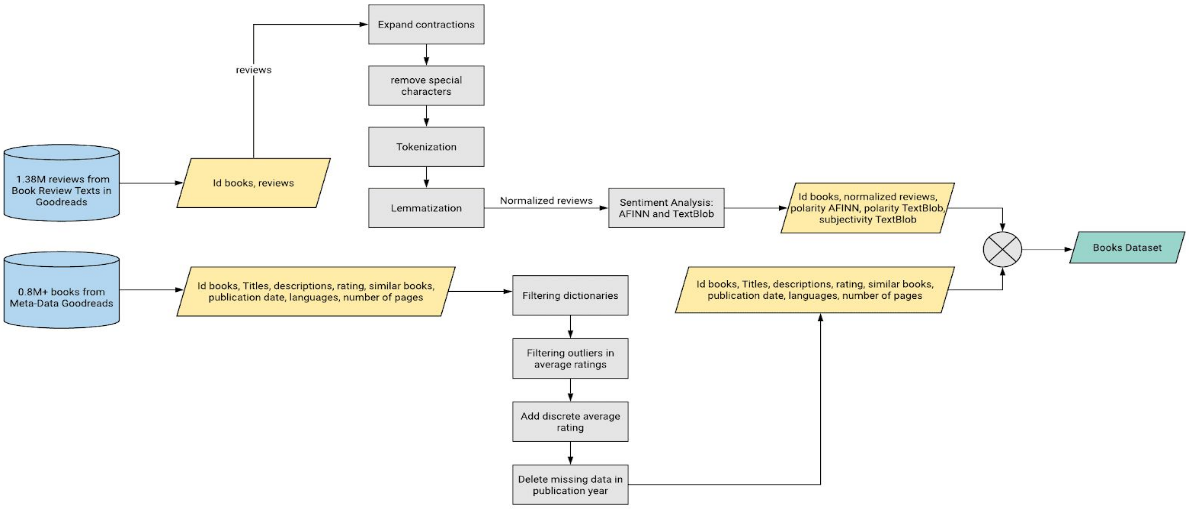
About the general information database, we focus on *titles, authors, average rating*, *book id*, *description of books*, *languages code*, *number of pages, publication year*, *rating counts* and *similar books*. The wrangling process, step by step, is the following:

1. Are we including dictionaries in this analysis? We must consider that the number of pages of dictionaries is larger than common books and for that reason dictionaries are filtered using regular expressions. We found almost 500 documents belongs to this category.
2. If we are curious to work with the *average rating* of books, we need to consider how many *rating count* has everyone and delete **outliers**. For instance, the *average rating*

of books with just one *rating count* can't be compared with others with one hundred counts. To deal with these differences, we calculate the **z score** of every *rating count* and filter books with z higher than 3 (standard threshold).

3. We add an absolute rating rounding ratings of books to deal with ratings in linear and discrete version.
4. Delete missing data in the publication year

Then, a NLP preprocessing is applied to read reviews before calculating the sentiment patterns. The techniques used are expanding contractions, subtraction of special characters, tokenization, and lemmatization of the words. We define a *pre-processing* function as a pipeline of the methods mentioned and a *sentiment parameters Pattern* function that calculate the polarity and subjectivity pattern of every review. In this case, we measure polarity using TextBlob and AFINN lexicon. The new metrics, patterns according to AFINN, TextBlob and normalized text are included in reviews of movies dataset. The pipeline of the process is displayed in Fig. 2.



Expand contractions
reviews
remove special characters
Tokenization
1.38M reviews from Book Review Texts in Goodreads
Id books, reviews
Lemmatization
Normalized reviews
Sentiment Analysis: AFINN and TextBlob
Id books, normalized reviews, polarity AFINN, polarity TextBlob, subjectivity TextBlob
Books Dataset
0.8M+ books from Meta-Data Goodreads
Id books, Titles, descriptions, rating, similar books, publication date, languages, number of pages
Filtering dictionaries
Id books, Titles, descriptions, rating, similar books, publication date, languages, number of pages
Filtering outliers in average ratings
Add discrete average rating
Delete missing data in publication year

Fig. 2: Pipeline Books Wrangling

III. Initial findings from the exploratory analysis

Exploratory Data Analysis for Movies

The following analysis was applied to movie dataset previously mentioned, that contains more than 40 thousand films. We are curious about release dates, languages, average score, genres and time series of movies.

1. **What are the most popular languages in movies?**

As we can observe in Fig. 3, English is the most recurrent language in movies, followed by French, Italian, Japanese, Germain, Spanish and Russian. A little more down are Hindi, Korean and Chinese.

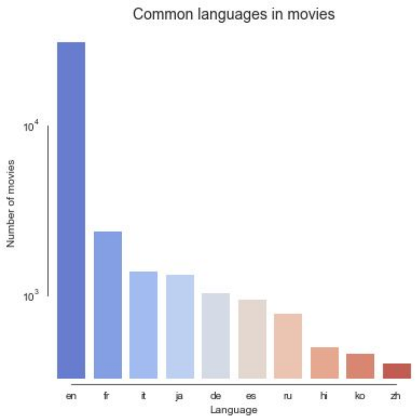


Fig. 3: Original languages in movies

2. **How loved or hated are long movies?**

For Hitchcock, *"the length of a film should be directly related to the endurance of the human bladder"*, but certainly, some directors would disagree with him. In this section, we analyze the relation between vote rating (from 0 to 10 scores) and runtime of films. We only consider the movies with more than 4 vote counts and a runtime superior than 15 minutes, that represent three superior quartiles of data.

The mean of the runtime in movies is around 100 minutes and there are some movies that exceed the three hours.

We appreciate that the range of runtime is spreading as the vote rating increases. In other words, the movies little liked have a runtime range shorter than the more liked movies. On the other hand, we get high scores in short and large movies. Then, there is not a clear trend to love or hate movies with a determine length, but films with excessive runtimes seem to have a good reception (vote rating lower than 4 only belongs to films with a maximum runtime of 200 minutes, meanwhile almost all the films with duration superior at 300 minutes have scores upper than 5 scores). If we measure the **z score** of the runtime to every movie, we get two classes: outliers or longer movies and regular movies, as is shown in Fig. 4a. Regular movies have a mean of 93 minutes and a maximum value of 204 minutes, meanwhile, when we treat all data as one, the mean is 100 minutes and the maximum value is 1256 minutes. Details are displayed in boxplots in Fig. 4b, where the percentiles 25 to regular movies are 86 and 107 minutes and for longest movies, 230 and 350 minutes.

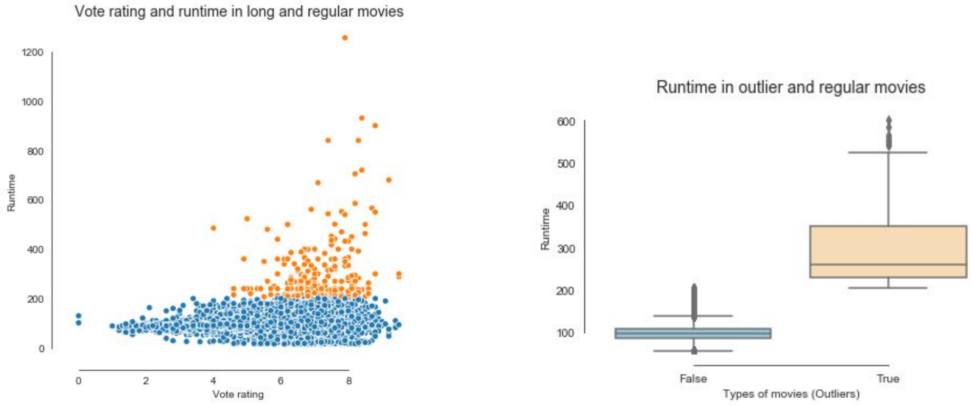
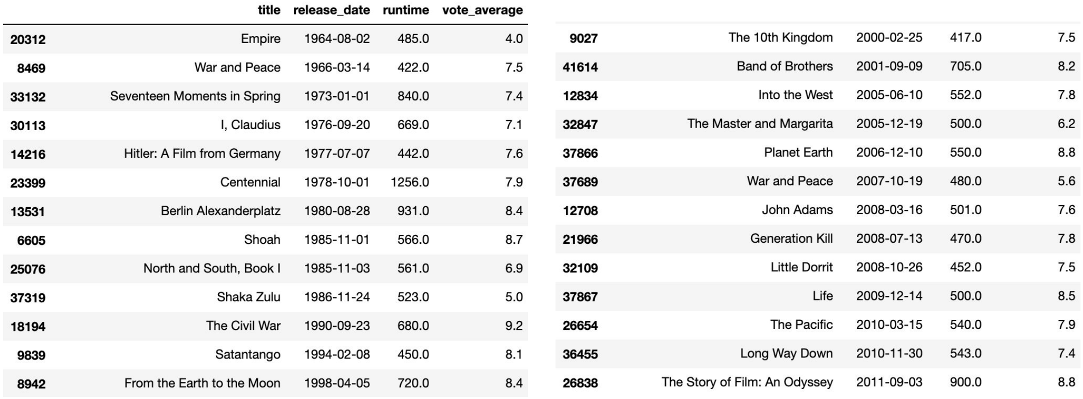


Fig. 4: a) Scatter plot runtime (left) and b) vote rating in movies and boxplot (right)

But what kind of movies exceed the 400 minutes? The list below is showing us those films, with the vote average included.



	title	release_date	runtime	vote_average
20312	Empire	1964-08-02	485.0	4.0
8469	War and Peace	1966-03-14	422.0	7.5
33132	Seventeen Moments in Spring	1973-01-01	840.0	7.4
30113	I, Claudius	1976-09-20	669.0	7.1
14216	Hitler: A Film from Germany	1977-07-07	442.0	7.6
23399	Centennial	1978-10-01	1256.0	7.9
13531	Berlin Alexanderplatz	1980-08-28	931.0	8.4
6605	Shoah	1985-11-01	566.0	8.7
25076	North and South, Book I	1985-11-03	561.0	6.9
37319	Shaka Zulu	1986-11-24	523.0	5.0
18194	The Civil War	1990-09-23	680.0	9.2
9839	Satantango	1994-02-08	450.0	8.1
8942	From the Earth to the Moon	1998-04-05	720.0	8.4
9027	The 10th Kingdom	2000-02-25	417.0	7.5
41614	Band of Brothers	2001-09-09	705.0	8.2
12834	Into the West	2005-06-10	552.0	7.8
32847	The Master and Margarita	2005-12-19	500.0	6.2
37866	Planet Earth	2006-12-10	550.0	8.8
37689	War and Peace	2007-10-19	480.0	5.6
12708	John Adams	2008-03-16	501.0	7.6
21966	Generation Kill	2008-07-13	470.0	7.8
32109	Little Dorrit	2008-10-26	452.0	7.5
37867	Life	2009-12-14	500.0	8.5
26654	The Pacific	2010-03-15	540.0	7.9
36455	Long Way Down	2010-11-30	543.0	7.4
26838	The Story of Film: An Odyssey	2011-09-03	900.0	8.8

Fig. 5: More extended films order by release date

This list has only five movies, with clear historical contents and all of them before 1995: Empire, topping the list, is a black-and-white silent film of more than eight hours of slow-motion footage of an unchanging view of the Empire State Building. War and Peace, based on Leo Tolstoy's 1869 novel is about the Napoleonic era. Hitler: A Film from Germany is a 1977 Franco-British-German experimental film; Shoah, a french movie about the Holocaust and Satantango, a Hungarian film based on the novel of the same name, about authoritarianism during the Hungarian People's Republic.

The rest of the names on the list are miniseries based -almost all of them- on novels or history. Seventeen Moments of Spring is a 1973 Soviet twelve-part television series; Berlin Alexanderplatz is a 14-part West German television miniseries, Heimat is a series of films about life in Germany from the 1840s to 2000. Shaka Zulu is a 1986 South African television series based on the story of the king of the Zulu, Shaka. The Civil War is a 1990 American television documentary miniseries about the American Civil War.

Perhaps they are not films of one part, we include them in the analysis due to their historical value and because of on the data there are more series with diverse length included as movies.

3. Are movies getting longer?

In Figure 6 we display the relationship between the length and release date of the movies in the most common languages, with runtimes upper than 15 minutes. As we know, English is the most popular language and almost all the runtime outliers are in English, but we can find a pair of Germain (Berlin Alexanderplatz, 1980; Hitler: A Film from Germany, 1977), French (Napoleon, 1927; The French Revolution, 1989), and Japanese (The Guyver: Bio-Booster Armor, 1989) outliers films.

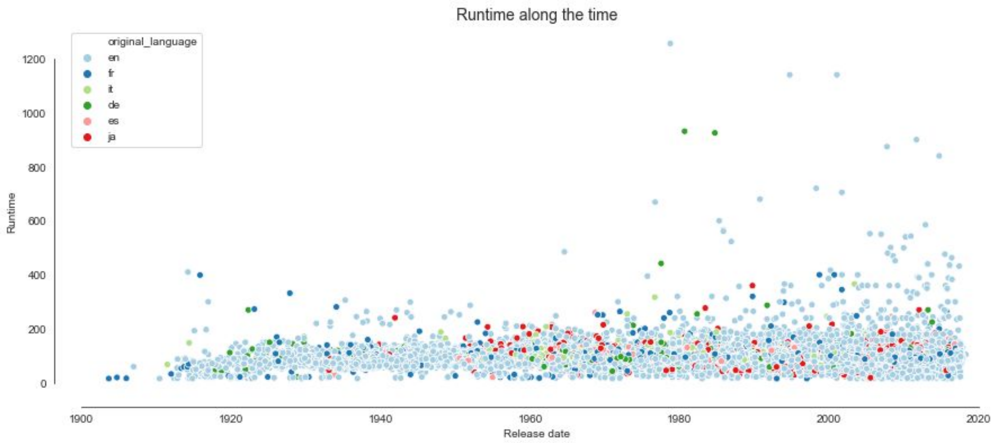


Fig. 6: Scatterplot of the length of movies by language and release date

What if we group the movies per year and determine the mean of the length as a time series? Let's have a look at the line chart below, where we can see a clear upward trend during the 20th century, but later, from the beginning of the present century until now, length of films keeps around the 100 minutes. We are using rolling means in windows of 5 years here.

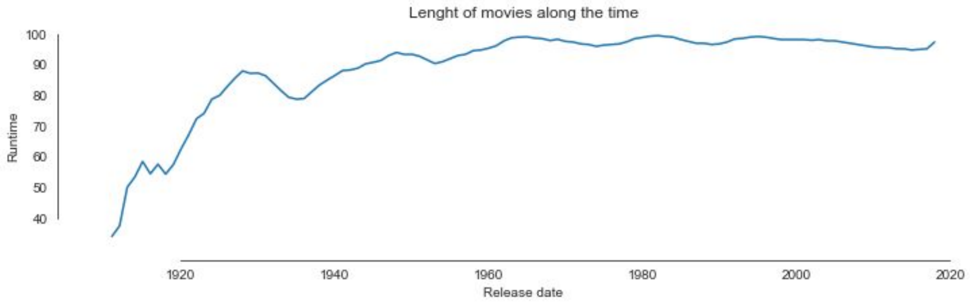


Fig. 7: Runtime of movies during the XX and XXI centuries

4. Historical release of movies

If you look at the following graph (see Fig. 8), you will notice an expected exponential growth of movies released every year from the beginnings of movies until now. Most interesting is to study the curves of movies released in different countries during the same period. Fig. 9a displays the most popular languages in movies in logarithmic scale, to include English movies in the same plot and Fig. 9b is the same analysis in linear scale excluding the English movies.

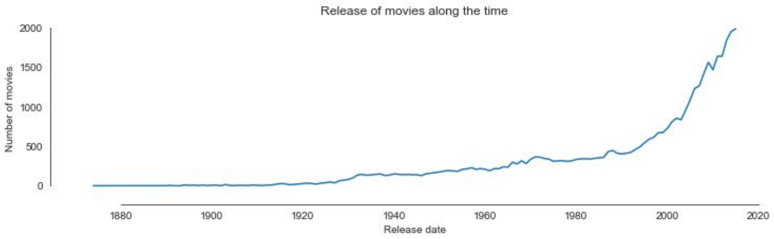


Fig. 8: Release of movies during the XX and XXI centuries

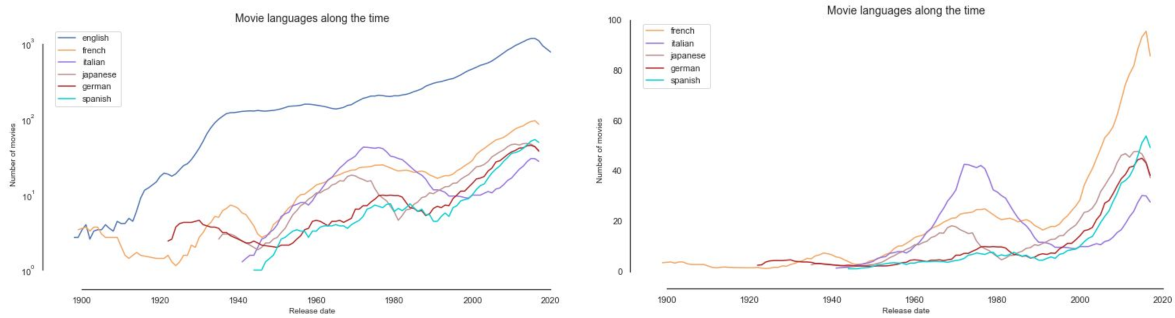


Fig. 9: Release of movies by language in a) Logarithmic scale and b) Linear scale

English and French movies have early development, but the first experiment a dramatic increase until 1940. World Wars and repercussions of the Great Depression could have impacted the growth of the french seventh art during the first half of the past century. Italian movies have an explosion in the second half of the centuries, probably due to the same reasons. In the case of Germain movies, they played a fundamental role in the politic campaigns after the WWI and we can observe a slight evolution of movies between 1920-1930 and a clear decay during the second war and after that. Finally, all the languages except English show a decrease in the movie evolution during the early 1990s recession.

In Fig. 10 we split English movies according to the production country. The curves of Britain U.S have a bigger and younger growth than Canadian and Australian movies.

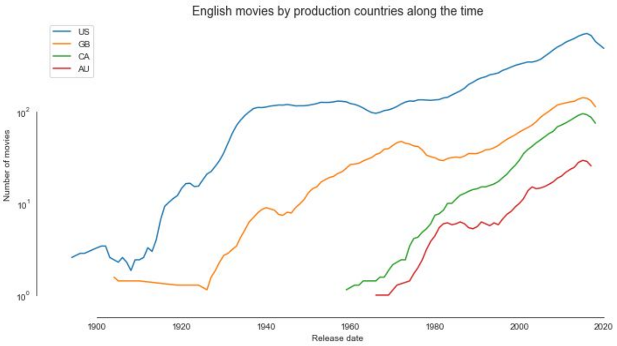


Fig. 10: Release of English movies by production country

5. How many languages can we find in the same movie?

Some movies have a list of spoken languages because they develop their stories in different countries or cultures. Can we measure the frequency and how many languages appear in every movie?

One or two languages are the most typical case, according to Fig. 11, that displays the histogram of languages in every movie using a logarithmic scale. We found less than ten movies including more than 9 languages and one film with 19 languages, called *Vision of Europe*, from 2004 and correspond to an anthology film that contains 25 short films about 25 directors from Europe.

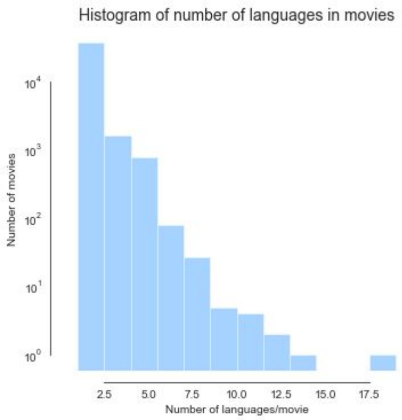


Fig. 11: Spoken languages in movies

6. Distribution of movies by language

In this section, statistical tests using a significance level $\alpha = 0.05$ are applied to determine if distributions of movies split by language have the same population mean. The one-way ANOVA test is used because we try two or more groups of different size. We build distributions using the most popular languages mentioned above. The histogram and respective KDE of English movies are displayed in Fig12a, where we observe that the mean of release does not exceed 250 movies every year. Deviation reaches 308 movies and the curve is leptokurtic (higher peak and profusion of outliers) and highly positively skewed. The same analysis to French and Italian films (Fig12b) reveals curves leptokurtic and highly skewed, but in this case, means are smaller (around 20 movies in French and 16 in Italian films) and the maximum value of release in French movies duplicate the biggest release in Italian productions.

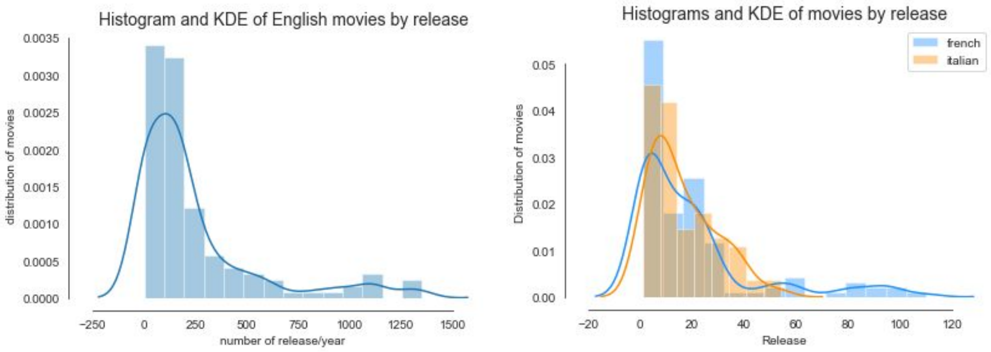


Fig. 12: Histograms and KDE of movies by release/year in a) English and b) French and German languages

The first test applied **rejects the null hypothesis about English, French and German distributions have the same population mean**, with a $p \ll 0.05$ and as we discuss in historical time series section, the release of English movies have had a different and more speed up evolution. But, certainly, there are relationships between the other distributions analyzed. In the second test, the null hypothesis is **French and German distributions have the same population mean** and in this scenario, we **fail to reject the null hypothesis**, with $p = 0.22$. Finally, the third test is related to the other popular languages (Japanese, German and Spanish films). The same ANOVA test **fails to reject the null hypothesis** with $p = 0.059$. Fig. 13 displays the KDE of these languages.

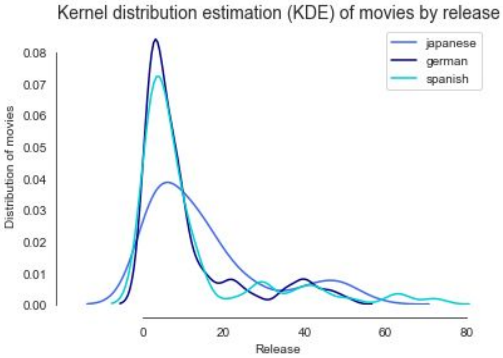


Fig. 13: KDE of Japanese, German and Spanish movies by release/year

7. Recurrent genres in movies

Every movie on dataset belongs at least to two genres of the following categories. The graph in Fig.14 shows how recurrent is every genre and as we could expect, *Drama* and *Comedy* are the most popular, followed by *Action*, *Horror-Thriller*, and *Romance*. The following table lists some of the movies of every genre order by popularity.

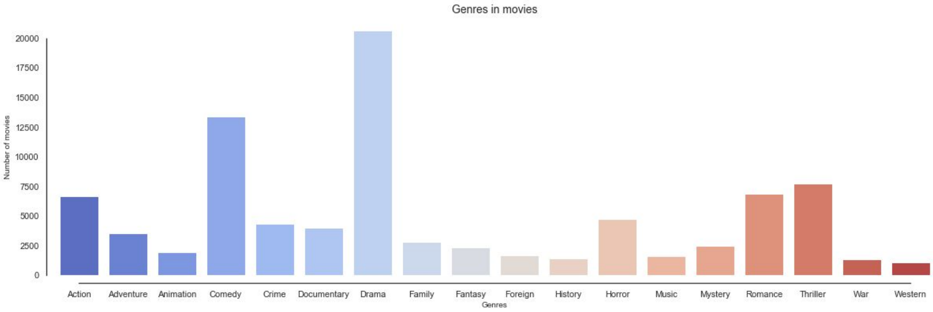


Fig. 14: Genres in movies

Genre	Movies
Comedy	Minions, Big Hero, Deadpool, Guardians of the Galaxy, Forest Gump, Ted, Pirates of the Caribbean
Drama	Gona Girl, War of the Planet of the Apes, Blade Runner, Whiplash, Logan, The Shawshank Redemption, Schindler's List, Life is Beautiful
Romance	Beauty and the Beast, The Twilight Saga, Fifty Shades of Grey, Titanic, Cinderella, La La Land
Thriller	John Wick, Gona Girl, The Hunger Games, Pulp Fiction, The Circle, Alien, Transformers, Get Out, Jurassic World
Horror	Alien, The Dark Tower, Get Out, World War Z, Don't Breathe, Rings, Saw, Annabelle, Black Mirror, Amityville
Action	Wonder Woman, Avatar, Captain America, The Avengers, Thor, Doctor Strange, Suicide Squad, Star Wars

Table 1: Examples of movies by genre

8. How likely are movies by genre?

In this section, we explore the average score of films by genre, displaying histograms and KDE to compare visually and then statistically the distributions of movies.

Action and Horror are the first genres chosen and as we note in Fig. 15a., distributions are negatively skewed. Action distribution is leptokurtic ($k > 3$) and Horror distribution is platykurtic ($k < 3$). We apply the Levene test, which null hypothesis is that **all input samples are from populations with equal variances**. Levene is used because we are dealing with significant deviations from normality (therefore, the assumptions of the ANOVA test are violated).

The first null hypothesis proposes that **Action and Horror movies are from populations with the equal variances** and we **fail to reject the null hypothesis** with a p-value $p = 0.26$

The second null hypothesis adds one of the less frequent genres, War. From Fig.15b., we chose previously two genres with similar frequency. How much likely are War movies? Apparently, they concentrate more positive scores because the distribution shows a more high skew than Action and Horror movies. And the same previous test, but adding one more genre **fail to reject the null hypothesis** with a p-value $p = 0.33$

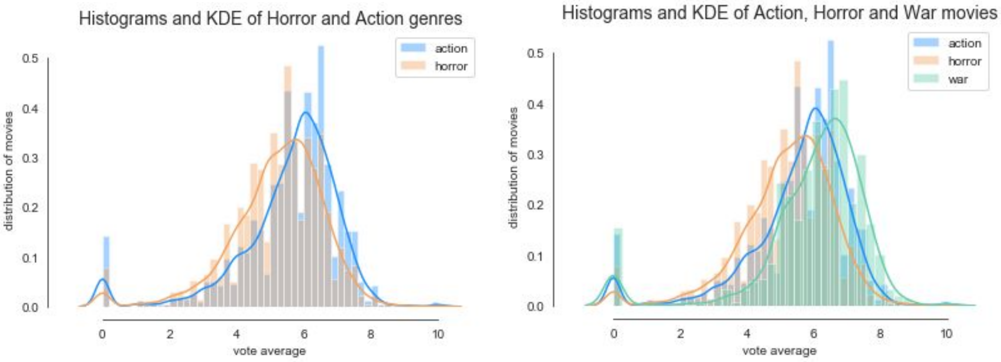


Fig. 15: Distribution of movies in genres Action, Horror (a) and War (b) by vote average

Finally, we apply a third test using Drama, Comedy, Romance. **The null hypothesis asserts that Drama, Comedy and Romance movies are from populations with the equal variances** and we **fail to reject the null hypothesis** with a p-value $p = 0.09$. Figure 15 shows that KDE of these genres are similar. If we include some of the previous genres to this list (action, horror or war) the result is to reject the null hypothesis. Therefore, we conclude that Drama, Comedy and Romance belong to one cluster of liked movies and Action, Horror and War to another.

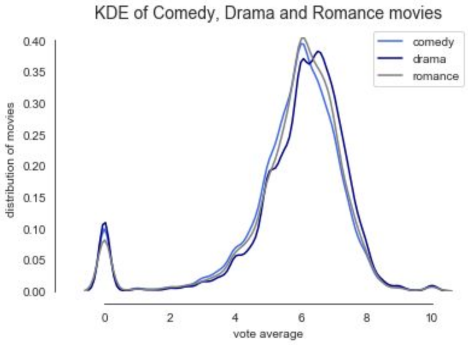


Fig. 16: Distribution of movies in genres Comedy, Drama and Romance by vote average

Exploratory Data Analysis for Books

"Books and movies are like apples and oranges. They both are fruit but taste completely different." (Stephen King).

In this section, we'll try to understand and extract the most relevant and useful information about books, preparing us for the next step, where we'll do some experiments to find the similarities between these kinds of apples and oranges. Book dataset contains more than 550 thousand of items. To display some figures and answer proposal questions, samples of the whole data are used, chosen randomly and applying bootstrapping and statistical tests to assure the legitimacy of results.

1. Relation between publication year and average rating: along the time, could we detect some trends?

As we could expect, we find more book publications in recent years with respect to the past century. How is the evaluation of readers for books from past century? How is right now? The average rating is measured between 1 and 5 scores. Older books publicated before 1960 have a good reception from the readers and from 1980 we observe a more wide range of scores, but the distribution of rating shows a clear more positive reception, with a concentration of books mainly between 3 and 5 scores. It means, that in general, books are qualified as Neutral, Good and Excellent moving scores to a categorical scale.

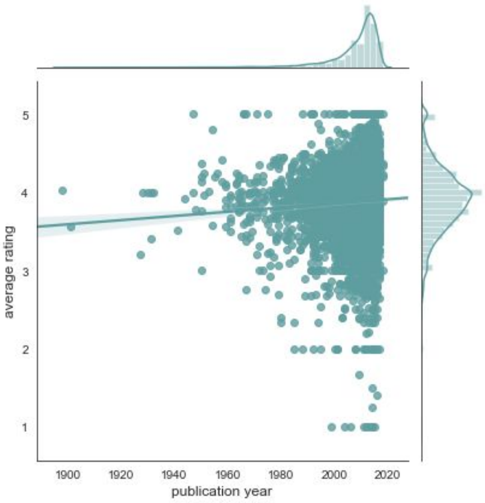


Fig. 17: Scatter plot average rating and publication year

To know if this sample is representative of the population, we compute the Kolmogorov-Smirnov statistic on 2 samples: the observed sample distribution and the alleged original distribution for the **null hypothesis that the rating average sample distribution is drawn from the population of 550k books and then is representative**. With a p-value $p = 0.32$ **we fail to reject the null hypothesis.** Furthermore, a pairs bootstrapping is applied to measure the Pearson coefficient between the rating average and publication year. The relationship found in this sample of the original population is a weak correlation of $p_{sample} = 0.0518$, but is it really illustrative? After 1000 trails of 5000 data points chosen randomly with replacement (the size of these samples is comparative with the size of the original sample analyzed), we calculate the Pearson coefficient and determine how many times we get a coefficient equals or higher than p_{sample} . We fail to reject the null hypothesis with a p-value $p = 0.358$.

2. Authors are writing more long novels?

Firstly, we inspect the histogram and KDE for the number of pages in every book (see Fig. 18). The mean is 260 pages and the percentiles 25% and 75% are 152 and 346 pages respectively. The maximum value is 1078 pages and correspond to the books The Nietzsche Anthology, published in 2013, The Count of Monte Cristo by the french Alexandre Dumas, The Source, a historical novel about Jewish people before the monotheistic era, published in 1965, to mention just a few of them.

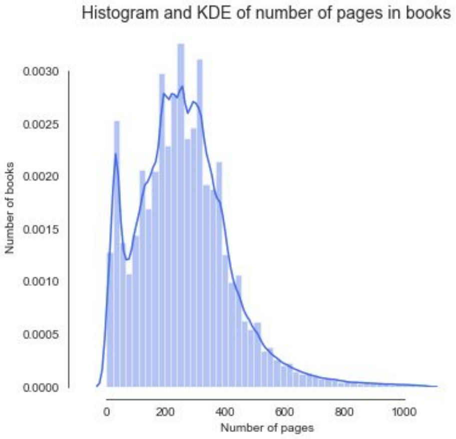


Fig. 18: Histogram of the length of books

Now, using the sample distribution previously exposed, we could graph a scatter plot of the length of books and publication year. And this time, we need to check, for the new variable studied (number of pages) if the sample randomly chosen is representative of the population. Therefore, the **null hypothesis**, in this case, is that t**he sample distribution of the number of pages belongs to the same distribution that the original population** and if we couldn't reject the hypothesis, we assert that the **sample is illustrative**.

Let’s turn to the graph in Fig. 19. We detect that the range of length of books spreads out from the second half of the past century until now. Inspecting the longest books in this period, we found titles as Quixote by Cervantes published in 1605, 2004 and 2013; Ulisse by

James Joyce published in 1922 and 2013. Then, if we found longer books in this century some of them could be new editions of older books, but it doesn't mean that authors are not writing long books right now. It only limits us in the sense that we can't assert that current books are longer or not. But certainly, according to the graph, long stories, anthologies, re-edition of classic books still being sold and read today.

Using a significance level of 0.05, and a Kolmogorov-Smirnov statistic on 2 samples, a p-value $p = 0.23$ was got and we fail to reject the null hypothesis. Therefore, the **sample is illustrative**.

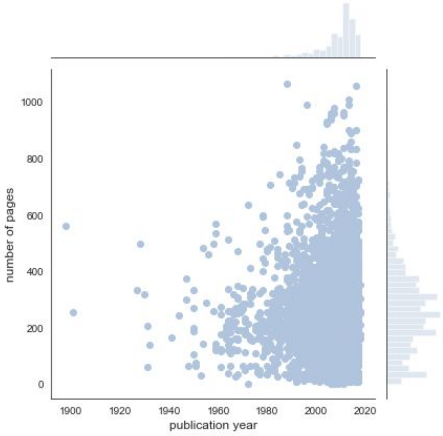


Fig. 19: Scatter plot number of pages and publication year

3. Readers prefer shortest or longest novels?

If you are thinking write a book, this is an interesting question that you must be able to answer. Could improve your reception the decision about short or long stories? People still love both, but certainly, we can find interesting results..

Applying **z score** criteria to the complete population of data, we could split the data points in two groups: regular and long books. Let's turn to the boxplot below, where every absolute rating is displayed per group. As we could expect, long books have higher means than regular books (around 800 vs 200 pages) and further, the mean number of pages (in regular and long books) increase according as the absolute rating. Observing the ratings 3, 4 and 5 (that means, neutral, good and excellent books in a categorical scale), it's clear that they include the books with higher number of pages.

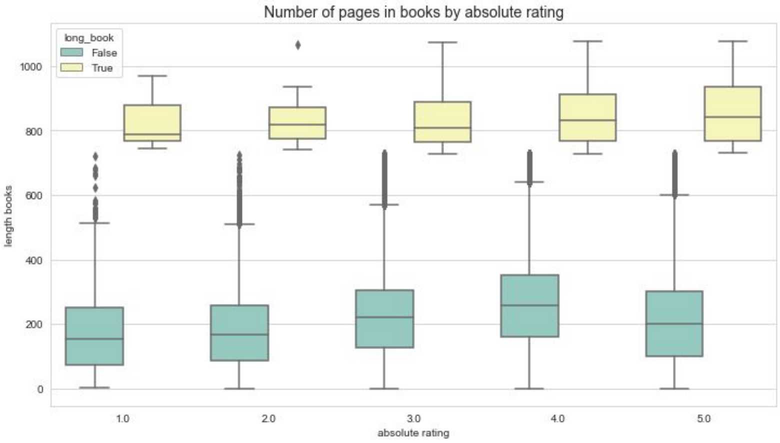


Fig. 20: Boxplot of books by absolute rating and type of book

4. **What are the most popular languages in books?**

We found 146 different languages in data and this is a small value thinking about there are more than 7 thousand of languages in the world. Despite this variety, some of them are more frequent. To analyze them, we work with the complete population. As we can see in Fig. 21, English is the most popular language, followed by italian, german, spanish, french and finally, portugal and aragon. We use a logarithmic scale because English literature cover more than the 60%. and is in a higher magnitude scale than the rest.

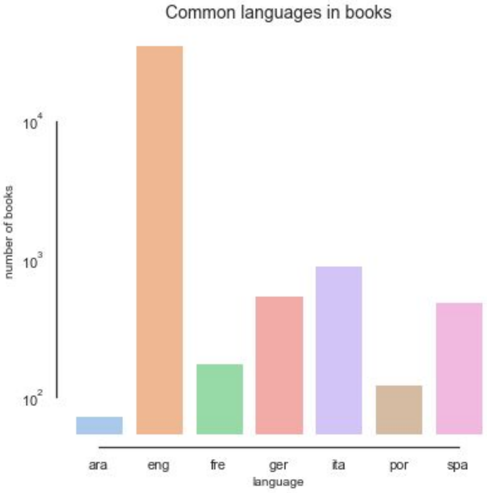


Fig. 21: Bar chart languages in books

To inspect books by language through the time and average scores, we get a sample of 20% of data. But again, how do we know that this sample is representative? We did a chi-square bootstrapping test, chosen randomly 1000 subsamples of data (with replacement) for getting absolute frequence of the ten most popular languages and applying a chi-square test using these results as data expected and the frequence on the sample analyzed as data observed. Finally, we calculate how many times we fail to reject the null hypothesis about the **sample data of languages is representative of the population** with a confidence level of 0.05 and we get that 73% of tests fail to reject the null hypothesis.

As we mentioned early, the rating scores in books are concentrated in positive categories and, splitting them by languages, we can conclude the same. Fig. 22 displayed English books (including Canadien, British) and Romance languages books (including French, Spanish, Portuguese, Italian, Romanian, Catalan, Aragon languages). And it's interesting note that distribution of data points is similar between categories and in both case the population of Excellent and Good books still being higher than Fair and Poor items. Besides, this dataset in particular has more older English books than Romance books. We highlight Romance languages because they have a particular interest in a lot of undergraduate and postgraduate programs now and certainly it’s because their rich cultural heritage and because it’s a key to understand the multicultural societies in the world.

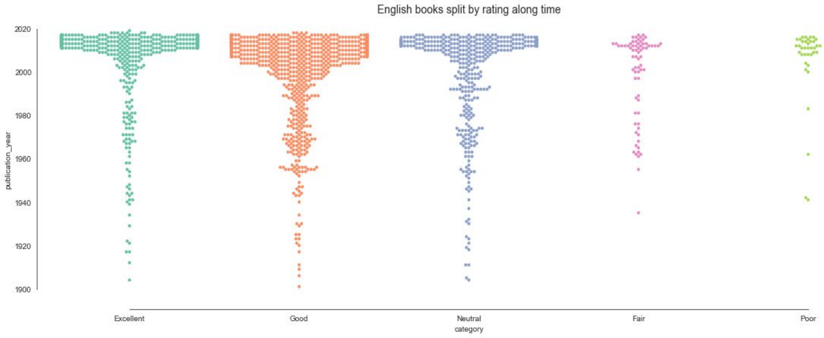


Fig. 22a: English books through time split by rating category

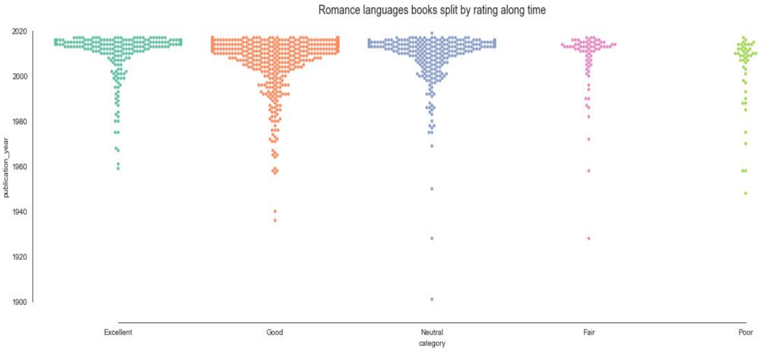


Fig. 22b: Romance books through time split by rating category

5. Evolution of publications over time

The following graph (see Fig. 23) shows the evolution of books belong to the English language, German language (except English) and Romance languages. The lasts groups has a similar growth through the time. Germanic language family includes German, Swedish, Danish, Dutch, Norwegian, Afrikaans and Icelandic languages.

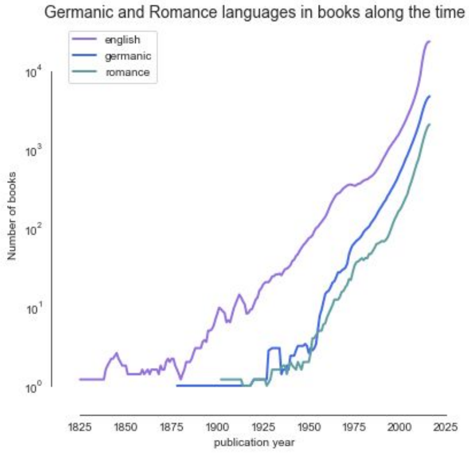


Fig. 23. Curves of growth in publications according to language families

6. Sentiment Analysis: AFINN lexicon versus TextBlob

Polarity and subjectivity patterns are some useful tools to analyze reviews. As we explained previously, we decide to try two techniques to measure the polarity of patterns: the textBlob pattern analyzer and AFINN. The first one has the advantage of compute both (polarity and subjectivity pattern) of text, to study the text data from two perspectives: we determine how much positive or negative is data and additionally, the linguistic intention of the message or how much informative/descriptive or argumentative is. The second criteria only measure the polarity, but, as we'll check later, use a different and more wide scale that allow us split levels of positive and negative information.

AFINN patterns are around -200 and 200 scores and TextBlob between -1 and 1. We re-scaled both to put them together in the following polarity distributions graph. It seems to be that TextBlob patterns are more homogeneously distributed in positive and negative scores (see Fig. 24).. Does it mean that the analyzer is better?

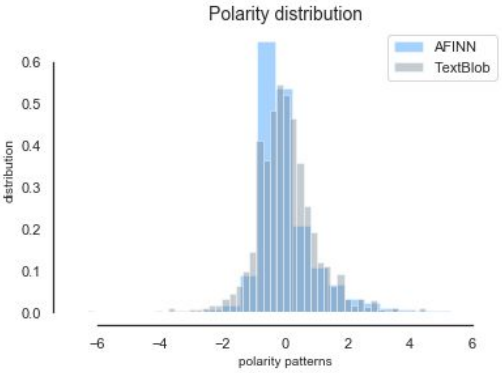


Fig. 24. Polarity distribution from AFINN and TextBlob analyzers

In the Fig. 25, we identify positive, negative and neutral information. Some conclusions about these pictures are that AFINN criteria concentrate more positive messages on the highest scores and more negative messages on the opposite. Otherwise, the TextBlob pattern analyzer gets coefficient along all the possible negative and positive scale independently of the rating score. Therefore, we'll use the polarity patterns from AFINN criteria and we'll study the subjectivity patterns measured by the analyzer of TextBlob.

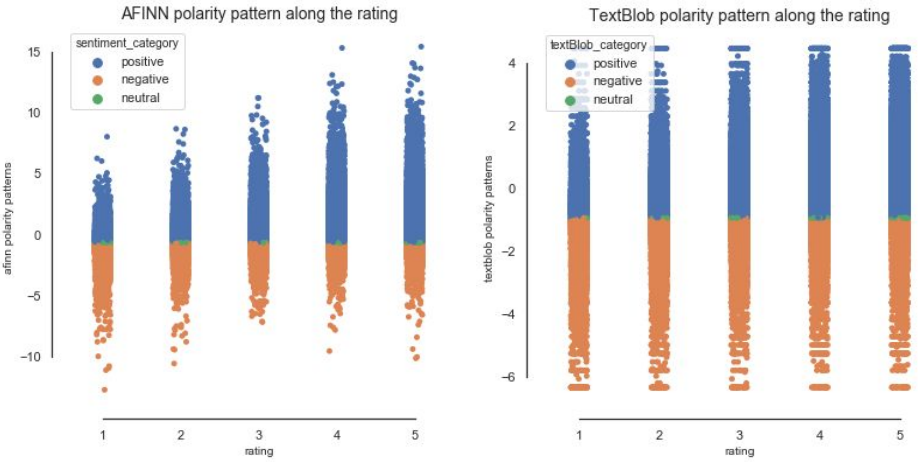


Fig. 25. Sentiment Analysis Comparative

7. Sentiment Analysis: subjectivity patterns

The polarity patterns for a sample corresponding to 1% of book reviews are displayed in Fig.26. The reviews qualified as neutral are more descriptive because they have fewer subjectivity scores (around 0) and the positive and negative reviews are in higher positions (between 0.4 and 1), that means, from mediumly to completely subjective. Furthermore, from rating 3 to 5, we see the increase of the subjectivity pattern.

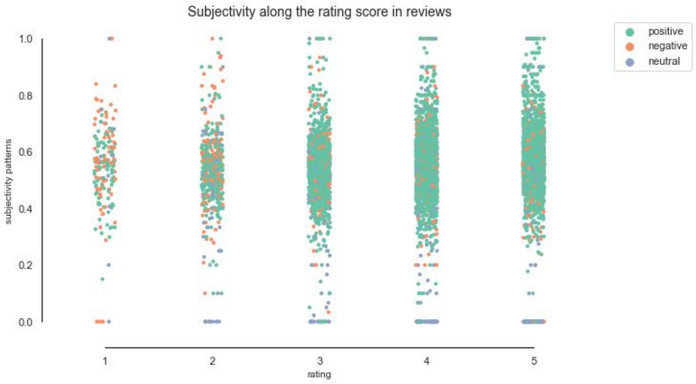


Fig. 26. Subjectivity patterns from a sample of 1% data

The polarity pattern to the same sample is displayed in Fig. 27. The polarity score is ascending through the rise of the rating, as it can be expected. We check that this a representative sample of population of reviews testing the distribution of polarity through of Kolmogorov-Smirnov statistic on 2 samples. We got a p-value $p >> 0.05$ we **fail to reject the null hypothesis** about **the sample is representative of the population.**

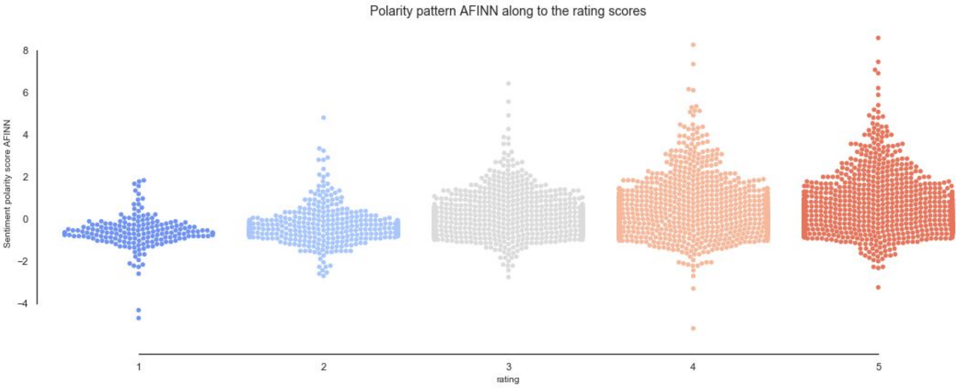


Fig. 27. Polarity patterns from a sample of 1% data