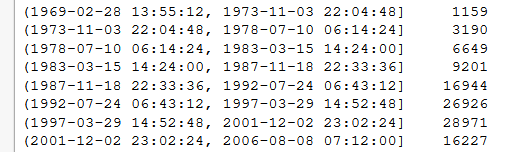
# Generic

nb lines: 114528 lines

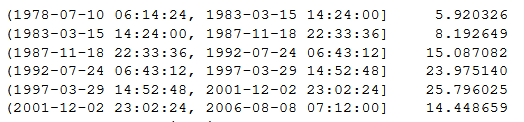
# Year of publication

## Lots of books published between 1983 and 2006



[…]

## Years intervals representing at least 5% of books in our dataset



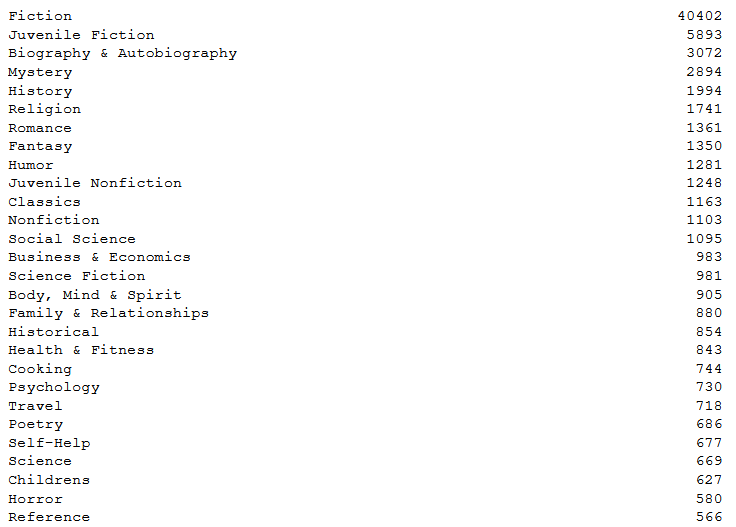
# Categories

## Before cleaning

5010 categories

* Categories with only one element saved in " cat\_one\_book.csv"

### Lots of Fiction : 35.28 % of all books



[…]

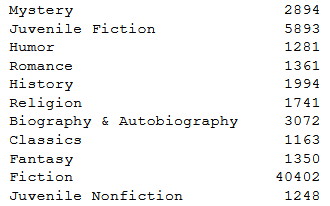
### Repartition of the number of elements insed categories : lots of categories with less than 80 books inside



[…]

### To keep 95% of the most important categories, we could keep the 3989 first categories (on 5010)

### Categories representing at least 1% of all the books



## Cleaning

The cleaning has been performed in several steps :

* 1st categories with ***"&"*** inside : a second column "Category\_other" has been created to take the 2nd term after the "&"
* Then the category words between ***"(…)"*** has been put inside "Category\_other" column, and the first term only remains on the "Category" column
* Then the category words with ***","*** has been splitted in two : the first term before the "," on "Category" column, and all the remaining termns in "Category\_other" column

## After cleaning

# Author genres

## Cleaning

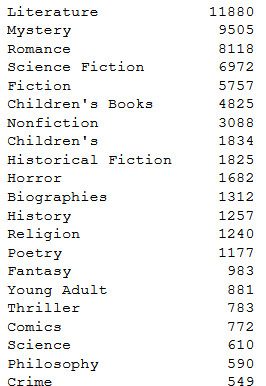
Cleaning (on Colab ) as multiple author genres in one cells, genres separated by &

* 2 columns : "author\_genres" and "author\_genres\_other"

## author\_genres

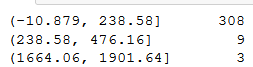
* author genres with only one element inside saved in "author\_gender\_one\_book.csv"

### Lots of Literature as author gender



[…]

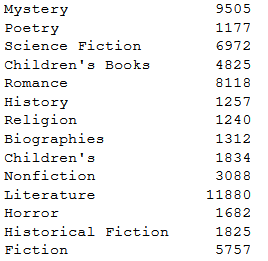
### Repartition of the number of elements insed author genres : lots of author genres with less than 240 books inside



[…]

### To keep 95% of the most important author genres, we could keep the 303 first author genres (on 338)

### author genres representing at least 1% of all the books

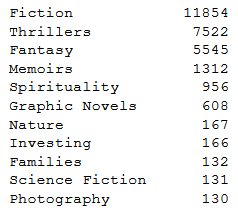


* Mystery, Romance, History, Religion, biography and fiction in common with categories.

## author\_genres\_other

* author genres with only one element inside saved in "author\_other\_gender\_one\_book.csv"

### Lots of Fiction as author other gender



[…]

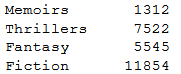
### Repartition of the number of elements insed author other genres : lots of author other genres with less than 120 books inside



[…]

### To keep 95% of the most important author other genres, we could keep the 34 first author other genres (on 39)

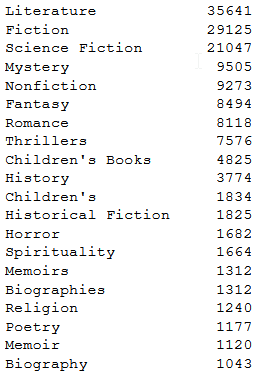
### author other genres representing at least 1% of all the books



## Both columns

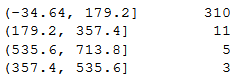
* author unified genres with only one element inside saved in "author\_unified\_gender\_one\_book.csv"

### Lots of Literature as unified author genres



[…]

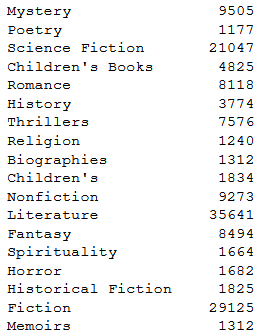
### Repartition of the number of elements insed unified author genres : lots of author genres with less than 180 books inside



[…]

### To keep 95% of the most important unified author genres, we could keep the 324 first author other genres (on 352)

### unified author genres representing at least 1% of all the books



* Thriller, Fantasy, Spirituality and Memoirs is appearing from the author\_genres column
* Mystery, Romance, History, Religion, Biography, fantasy, fiction in common with categories.

# Languages

## Cleaning

Some languages were coded in several words :

* "Catalan; Valencian"
* "English, Middle (1100-1500)"
* "Multiple languages"
* "zh-CN"
* "zh-TW"
* Replaced by their ISO 639 coding :

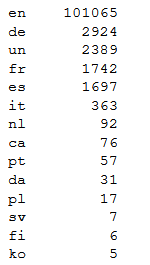


[…]

Then some languages are coded with words instead of ISO 639 coding

* Replaced by their ISO 639 coding

## Mainly english books

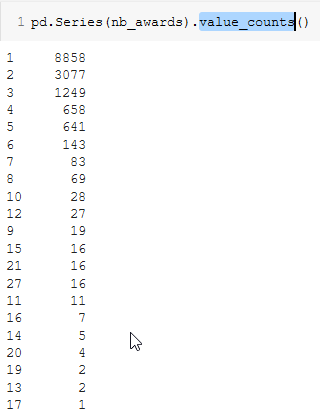


[…]

# Awards

## Books with lost of awards (before cleaning)

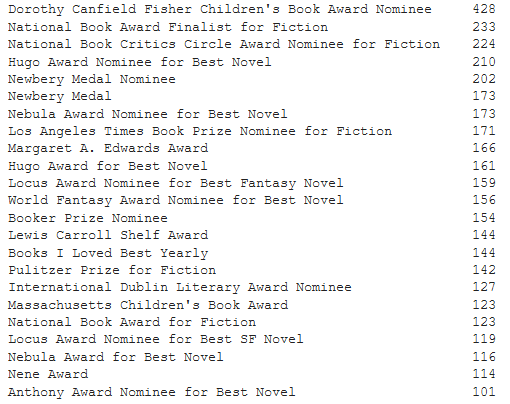
The maximum number of awards for a same book is 27 !



* Harry Potter and the Sorcerer's Stone (at indice 2809) !!!!

## Kind of awards (before cleaning)

* list saved in " awards\_names.csv"



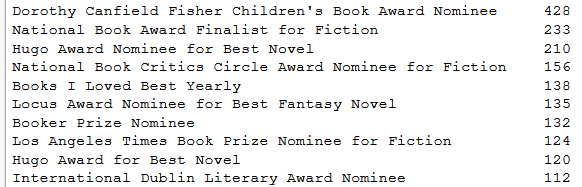
[…]

## Cleaning

I kept only the most famous awards in each cells.

* It remains 613 kind of awards.

## Kind of awards (after clening)



[…]