# Group Project Report

##### Authors: Kheirie Kaderi, Mohamed Al Jalanji, Clemence ROLDAN

# Data import

The data was import from the csv file.

Several mistakes were detected and corrected :

* Data shifted to the left for one row (row 1847)
* Wrong dates
* Column type

# Data exploration

### General exploration

* The database doesn’t have empty cells
* There is a high correlation between ratings\_count and text\_reviews\_count (0.87)
* The average\_rating and the num\_pages are slightly correlated at a correlation of 0.15

## Average Rating

The average rating is really imbalanced, as 92% of the ratings are between 3.5 and 4.5.

This can be explained by several facts :

* Goodreads is a website used by people who enjoy reading
* Reading is a time consuming hobby, so most people will only read books they enjoy and stop reading books they don’t like

Average rating is our target, so we will have to address this imbalance for the machine learning part.

#### Ratings of 0

On the database, we notice that there is 26 books with an average\_rating of 0. By looking further, we notice that those books received no ratings.

## Number of pages

* Most of books have less than 400 pages
* The relationship between the number of pages and ratings cannot be clearly inferred, but we notice that books with less than 450 pages got highest ratings, and very large books do not have low ratings.

#### Books with few pages

We found that there is 76 books with 0 values, and 195 books with less than 10 pages. Some of those lines refer to audiobooks or CDs. By looking at the titles we can see that the vast majority of them should have more pages.

## Number of ratings and reviews

## Boxplots for the number of ratings and number of reviews show that, for both, the mean is significantly higher compared to the median. The vast majority of books don't have many ratings or written reviews, but there are some high extremes that influence the mean.

## Upon examining the higher numbers of reviews and ratings, we found that they correspond to well-known books. Therefore, even though these values are extreme, they are not errors.

## This pattern is consistent with the book market: there is an enormous number of new books each year, but very few of them become famous. Most books will only be read by hundreds of readers or fewer.

## The number of ratings and number of reviews are correlated with each other. There appears to be a link between the number of reviews and ratings, and the average rating, as books with a large number of ratings or reviews tend to have a higher average rating.

## Languages

* There are 27 different language codes, with 5 language codes for English.
* 94.76% of books are in English.

## Publisher

* There are 2292 different publishers.
* Among them, 1296 are cited only once.
* 230 are cited more than 10 times.
* 9 are cited more than 100 times.

## Author

* There is 6643 authors
* Some books have several authors names : it can be several authors, but also the name of the translator or the illustrator.

By looking at the dataset we see that the main author is almost always the first name to be written, so we make the choice of removing all the author after the first “/”.

If we analyze only the first author written :

* There is 4219 different authors
* Among them, 2756 have only one book (65.32%)
* 1251 have 2 to 10 books
* 21 have more than 10 books on the database

## Titles

* There are 511 books that have duplicates
* Most of them have the same average rating between duplicates, except for 59 books

By looking for more information on the Goodreads database, we understand that there is several possibilities :

* Some books have the same content but they differ in type (e.g. hardcover, CD, audio, paperback etc.) : on good reads when they give the overall rating (i.e. average\_rating) it is an aggregation of all these different versions of the book. That’s why they can have the same average\_rating, but different isbn, authors, publishers, review\_count, rating\_count etc.
* Some books have the same title, but are different books. We can see it by checking the first published date on internet. That’s when the average\_rating is different.

# Data engineering

## Lack of information

The data exploration showed several issues within the database :

* Books that are not books (audiobook, CD collection,…)
* Error in the number of pages
* Same titles for different books
* Same title and same average rating, but different isbn, authors, publishers, review\_count, rating\_count etc

To deal with those issues, we decided to scrap for more information online.