# Group Project

## Data import

While we started to convert the csv file, we noticed that there were error in 4 lines :

* Skipping line 3350: expected 12 fields, saw 13
* Skipping line 4704: expected 12 fields, saw 13
* Skipping line 5879: expected 12 fields, saw 13
* Skipping line 8981: expected 12 fields, saw 13

By looking them on the csv file, we manage to see that the author name had “,” on them, so the conversion of the csv file would separate the information “author name” in 2 cells.

For example, line 3350 we have : “Sam Bass Warner, Jr./Sam B. Warner”

In order to download all the information, we change those 4 names by removing the coma in the csv file.

## Data exploration

### Extreme or weird values

#### Book with 0 pages

There is 76 books with 0 values. By looking quickly at the lines, we see that there is mistakes (“The Da Vinci Code” is a well-known book, with 571 pages), but there is also line with “NOT A BOOK” as authors. We’ll investigate those line in the “Not a book” part.

We look further and see that there is 195 books with less than 10 pages, with good ratings, and a quick look inform us that there is also well know books there that are supposed to have more pages.

* There is mistakes on the num\_pages column

#### Not a book : audio books

There is 5 lines with written as author “NOT A BOOK”. After a quick google search, we see that it’s CDs, or radio collection.

There is also books with “audio” in the title.

Audiobooks have 0 or <5 pages so that explain some of the errors but not all of them.