**TCBDA14- Big Data project**

TALYA DAVID 066092370

EINAT TEPPER 032291999

SIVAN LITVAK 065976029

In this project we had 3 databases to clean and implement:

1. Book rating
2. Books
3. Users

**Part A - Cleaning the data:**

1. **Book rating clean:**
   1. Read CSV file, splitting by delimiter (;)
   2. Split columns using function we implemented, function return new df

Def split\_column(df, col\_name, col\_delimiter=';', new\_cols\_names=None)

* 1. Remove initial, ending of (“) and (,) by creating functions and using them-
     + Remove prefix
     + Remove suffix
  2. Change numeric columns into INT type.
  3. Clean “User id” and "book Rating" columns of letters and other signals
  4. Clean “ISBN” column of letters and other signals except numbers and the letter X .
  5. Cleaning invalid “isbn” by using filter function (where), we filtered the rows from df based on a condition:

length(isbn) =10 or length(isbn) = 13

* 1. Fill empty values with null.
  2. Fill null value with (0) for "book rating column"

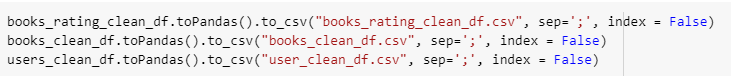
1. **Books clean**
   1. Read CSV file, splitting by delimiter (;)
   2. "Year Of Publication”-
      * Change data column into INT type.
      * Fill empty data’s column with (1900) when data is missing or not valid.
   3. Clean “ISBN” column of letters and other signals except numbers and the letter X .
   4. Cleaning invalid “isbn” by using filter function (where), we filtered the rows from df based on a condition:

length(isbn) =10 or length(isbn) = 13

* 1. Clean ASCII characters on string columns by using function (“Book-Title”, “Book-Author”, “Publisher”).
  2. Clean “Book- author” and “publisher” columns of signals as $@? Etc.
  3. Clean “Book- title” column of signals as @$ at the beginning of the name.

1. **Users clean**
   1. Read CSV file, splitting by delimiter (;)
   2. Change numeric column “Age” into INT type **.**
   3. Split "Location" column into 3 columns:
      * "District"
      * "City"
      * "Country"
   4. When age is above 120 or less than 0, we changed the data into (-1) .
   5. Clean column’s data City and District of signals and numbers.
   6. Fill “unknown” when column’s “City”, “District”, “Country” data is null.
   7. Step of cleaning the data of countries:

* Load CSV file filled with Countries data (official) and create dataframe.
* Join between the users dataframe and official countries dataframe and remove all countries that are not in the official list.
  1. Fill empty values with null.
  2. Fill null value in the “unknown” or (-1)

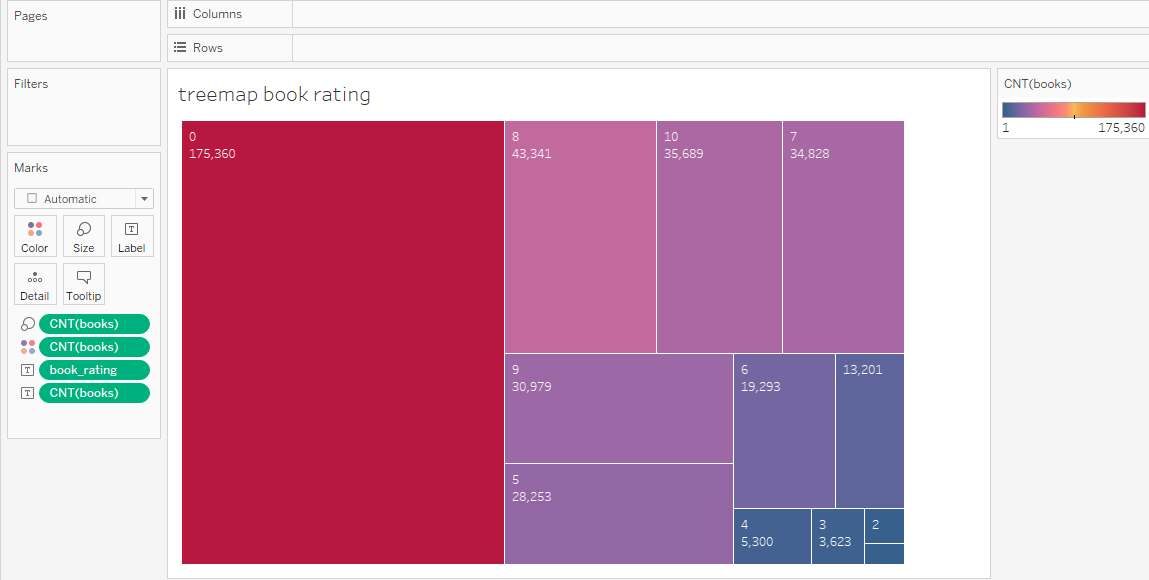
**Part B- put the data into CSV file**

And now we have 3 CSV files in order to work it Tableau:

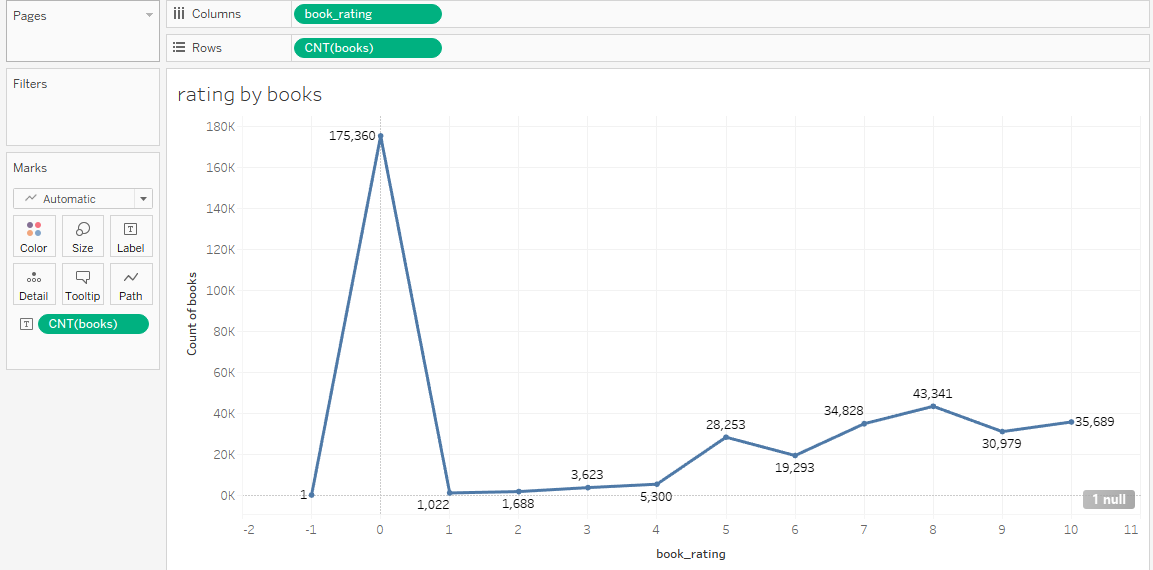
* books\_rating\_clean\_df.csv
* books\_clean\_df.csv
* user\_clean\_df.csv

**Part C- Tableau analysis conclusions:**

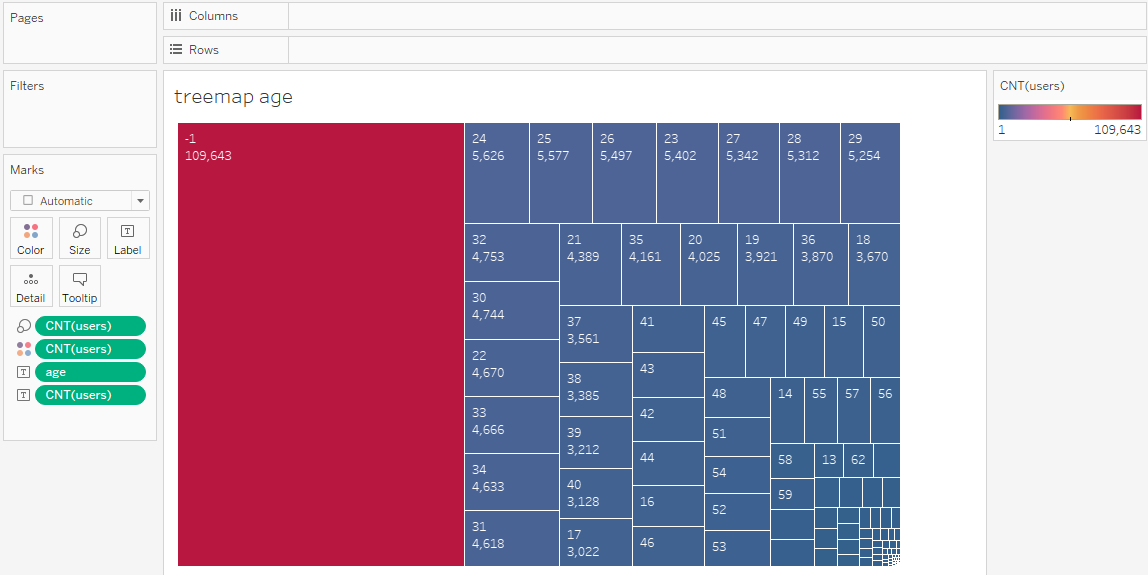
1. It seems that approximately, the same number of books received either the rating “0” or a rating higher than 7 (the sum of rating of 7 and up). That may indicate that if a book is being rated – it would probably get a relatively high rating (either “not rated” or high rating)



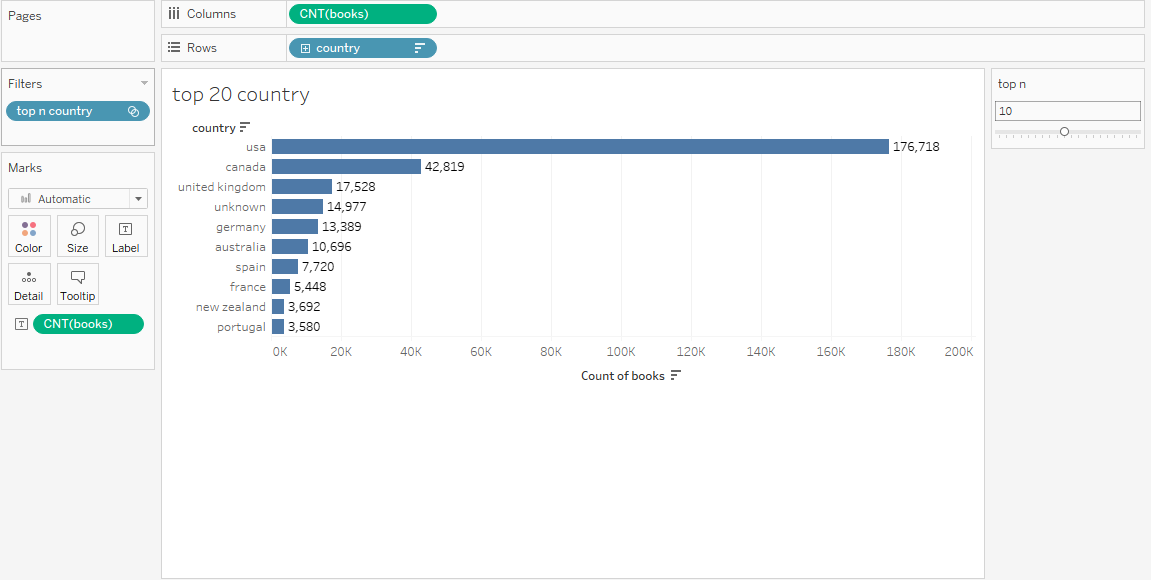
1. Most of the users rates the book they read, and mainly rate it high



1. Most of the users prefer not to indicate their age. Among those who did indicate their ages – most of the “Raters” are between the ages of 23-35



1. The number of books in the USA is significantly higher from all other countries (the second country to the USA is Canada and the number of books there is more than 4 times less)



1. The number of books published, grew along the years with a bloom of publications in the beginning of the 21st century.

