BIG DATA PROJECT

Presented by:
Margarita Dymshits
Natasha Shnaidman

Introduction

The purpose of this project is to demonstrate an analytical skills on large amount of data: book crossing dataset.

The analysis is performed with Pyspark and Tableau.

Data Overview

Here some initial information about dataset:

- **Ratings – CSV file, 1,048,575 rows, 3 columns, only 1 column contains unfiltered data, nullable.**
- **Books – CSV file, 271,379 rows, 1 column with unfiltered data**
- **Users – CSV file, 278,858 rows, 1 column with unfiltered data**

Process Stages

There are several steps in the project:

1	2	3	4	5	6
Import data	**Data Processing**	**First Analysis**	**Star Scheme Model**	**Tableau**	**Challenges and Conclusions**
Import data from csv files to Spark	Cleaning and organizing datasets	Understanding data and making initial assumptions	Building a Star Scheme Model	Analysis and visualizations	Challenges and Conclusions

Import data

After creating an environment, we imported data from 3 csv files and got 3 datasets named:

1. **Ratings – dataset that provides ratings by book number**
2. **Books – information about books: titles, authors, publishers and year of publishment**
3. **Users – information about users that rated the books (age, location)**

Data Processing

Ratings transformations:

- **Split by delimiter “;” first column that contained the data to 3 columns**
- **Remove “” characters from all the columns**
- **Changed columns formats**
- **Nulls check up – no nulls**

Data Processing

Books transformations:

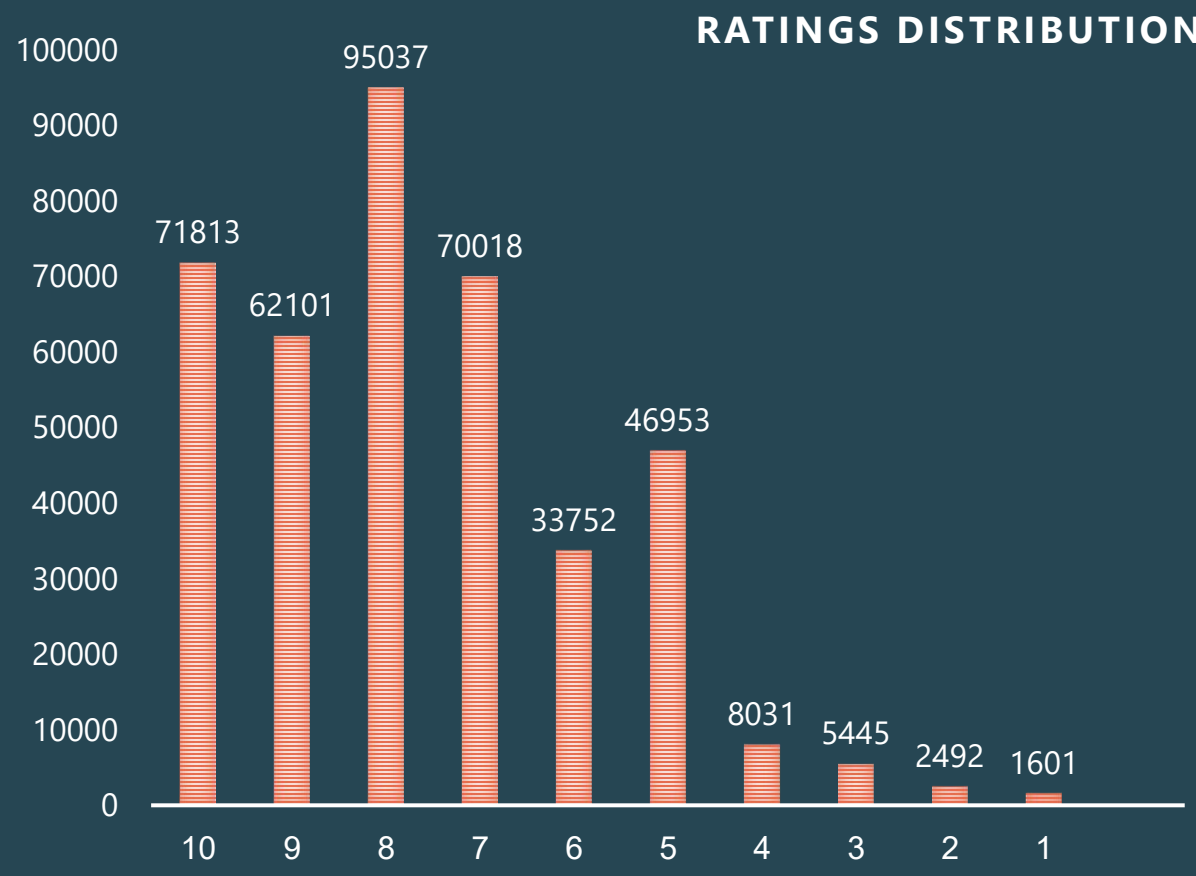
- Split by delimiter “;” first column that contained the data to 4 columns
- Remove “” characters from all the columns
- Nulls check up – cleaning nulls in author, publisher, year columns
- Reorganizing data that moved to different columns, for example titles that divided to two different columns, publisher names in year columns etc.
- Taking care of empty values in different columns
- Changing formats
- Unknown year and zeros in year column replaced by 9999

Data Processing

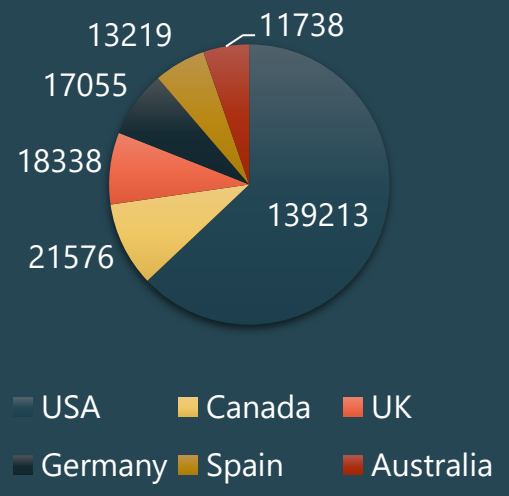
Users transformations:

- Create a new schema with 3 columns instead of one
- Split location column to city, region and country columns. There are missing values in all of them
- Replace nulls with “-1” value
- Replace nulls and “NaN” values with “unknown”
- Use country_clean function to clean and fix country column

First analysis



- Users tend to rate books they liked than books they didn't like. There are far more high rates than low ones.
- Rating "0" means user read the book but did not rate it. There are 651,327 zero ratings.
- Most of users are from USA, Canada, UK, Germany and Spain.



Star scheme model

Organizing datasets and creating fact table from ratings dataset and 2 dimensions based on books and users.

Saving and downloading data to a csv files for further analysis in Tableu Desktop

userSK	userBK	age	country
1	1.0	34	United States
2	2.0	18	United States
3	3.0	34	Russia
4	4.0	17	Portugal
5	5.0	34	United Kingdom
6	6.0	61	United States
7	7.0	34	United States
8	8.0	34	Canada
9	9.0	34	United States
10	10.0	26	Spain
11	11.0	14	Australia
12	12.0	34	United States
13	13.0	26	Spain
14	14.0	34	United States
15	15.0	34	Canada
16	16.0	34	United States
17	17.0	34	United States
18	18.0	25	Brazil
19	19.0	14	unknown
20	20.0	19	United States

only showing top 20 rows

isbnSK	isbnBK	title	author	year	publisher
1	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University...
2	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Ca...
3	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial
4	0374157065	Flu: The Story of...	Gina Bari Kolata	1999	Farrar Straus Giroux
5	0393045218	The Mummies of Ur...	E. J. W. Barber	1999	W. W. Norton
6	0399135782	The Kitchen God's...	Amy Tan	1991	Putnam Pub Group
7	0425176428	What If?: The Wor...	Robert Cowley	2000	Berkley Publishin...
8	0671870432	PLEADING GUILTY	Scott Turow	1993	Audioworks
9	0679425608	Under the Black F...	David Cordingly	1996	Random House
10	074322678X	Where You'll Find...	Ann Beattie	2002	Scribner
11	0771074670	Nights Below Stat...	David Adams Richards	1988	Emblem Editions
12	080652121X	Hitler's Secret B...	Adam Lebor	2000	Citadel Press
13	0887841740	The Middle Stories	Sheila Heti	2004	House of Anansi P...
14	1552041778	Jane Doe	R. J. Kaiser	1999	Mira Books
15	1558746218	A Second Chicken ...	Jack Canfield	1998	Health Communicat...
16	1567407781	The Witchfinder (...	Loren D. Estleman	1998	Brilliance Audio ...
17	1575663937	More Cunning Than...	Robert Hendrickson	1999	Kensington Publis...
18	1881320189	Goodbye to the Bu...	Julia Oliver	1994	River City Pub
19	0440234743	The Testament	John Grisham	1999	Dell
20	0452264464	Beloved (Plume Co...	Toni Morrison	1994	Plume

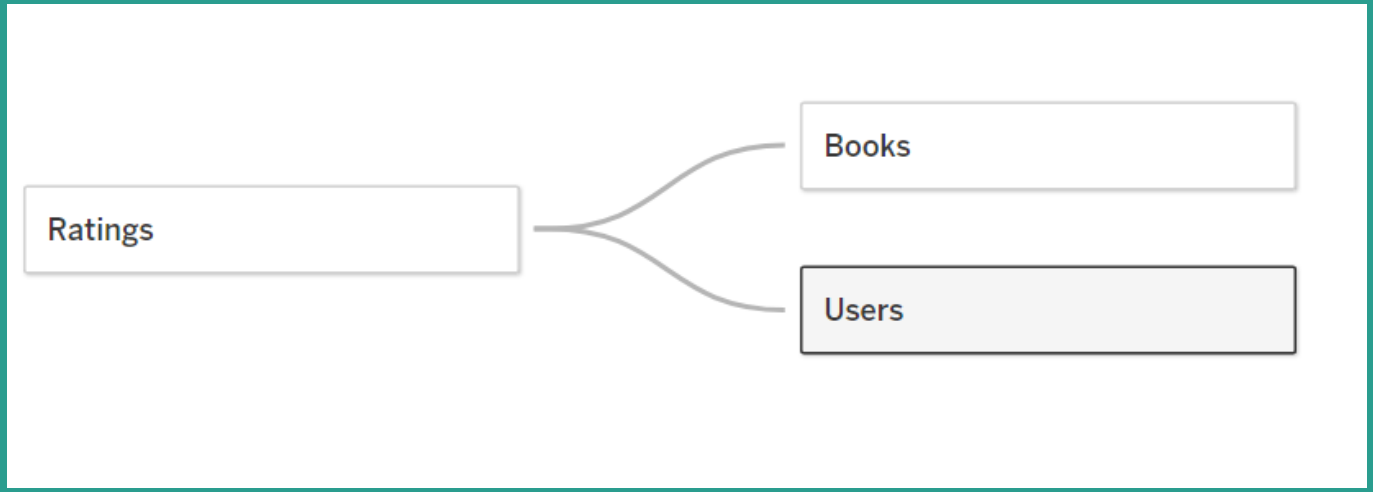
only showing top 20 rows

ratingSK	userSK	isbnSK	rating
1	276725	2967	0.0
2	276726	225830	5.0
3	276727	11055	0.0
4	276729	246855	3.0
5	276729	246856	6.0
6	276733	123646	0.0
9	276744	9296	7.0
11	276746	2031	0.0
12	276746	228	0.0
13	276746	1005	0.0
14	276746	597	0.0
15	276746	87285	0.0
16	276746	30986	0.0
17	276747	4780	9.0
18	276747	25798	0.0
19	276747	7155	0.0
20	276747	1837	9.0
21	276747	6277	8.0
22	276747	246857	7.0
23	276747	71880	0.0

only showing top 20 rows

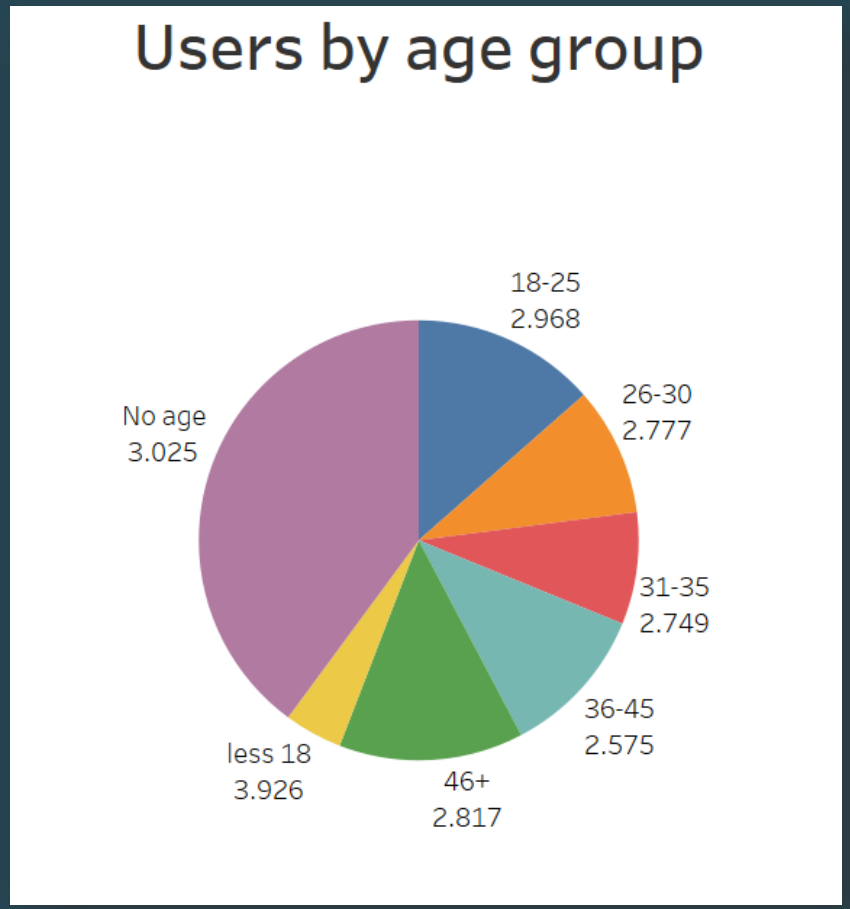
Tableau

Star scheme model built from fact table and has 2 dimensions: Books and Users

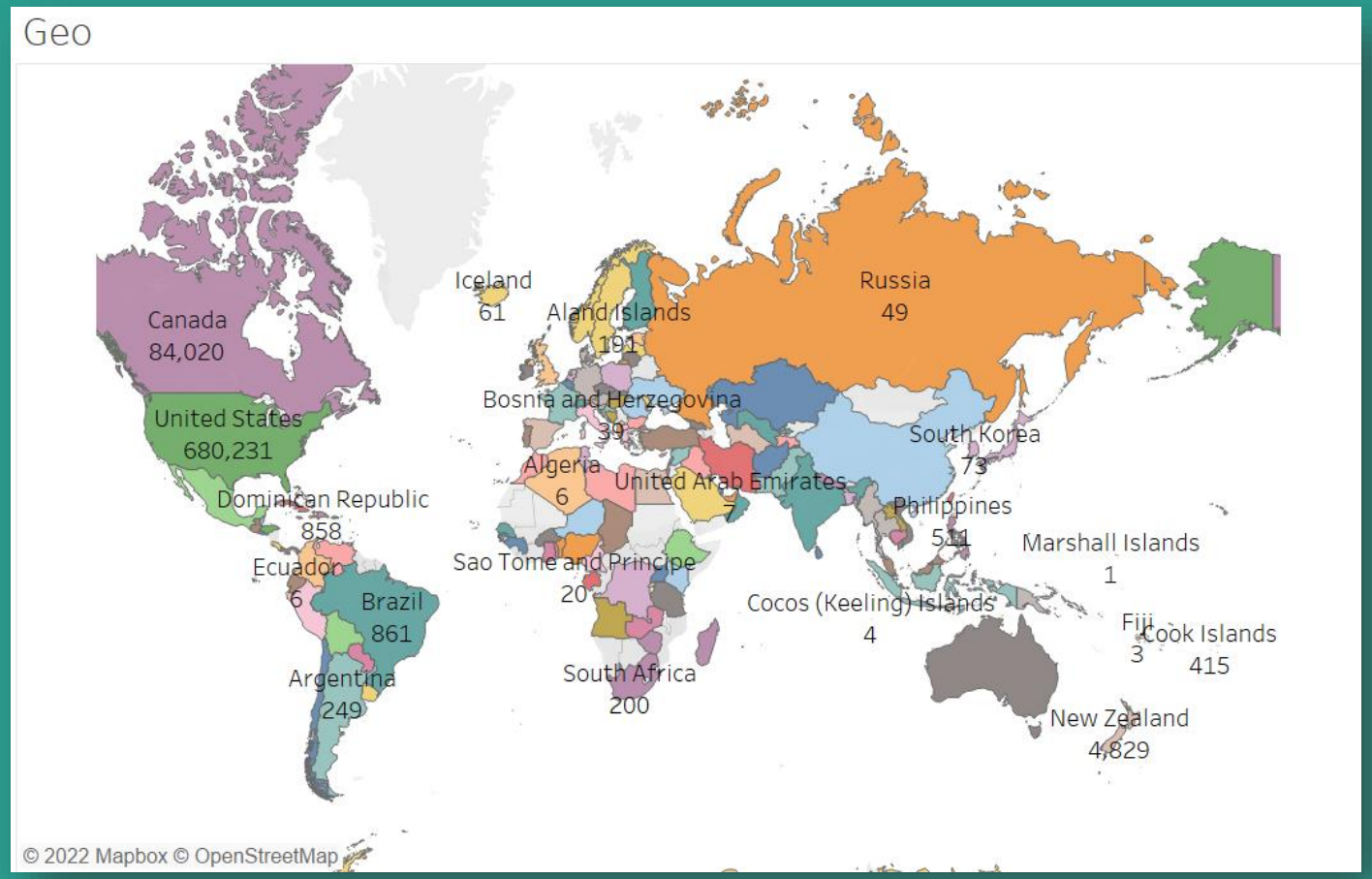


Tableau

Age group distribution and age group's average rating.



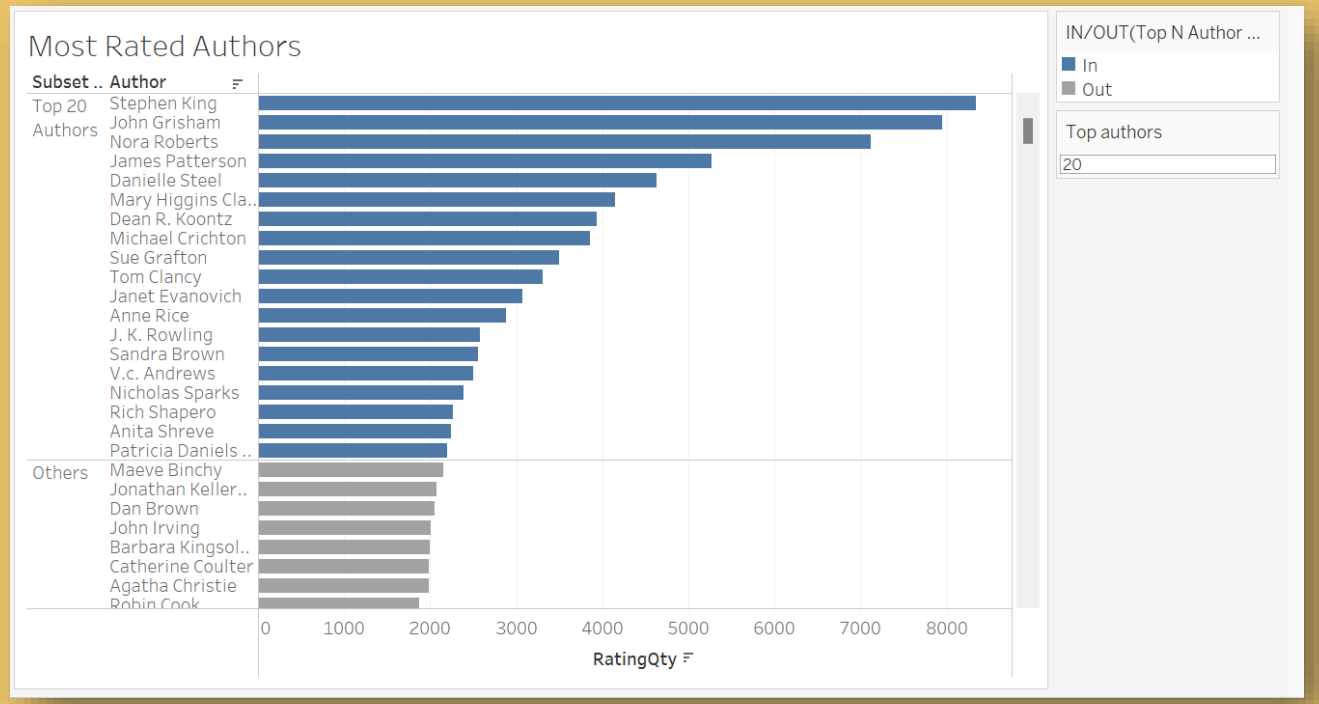
Tableau



Most of the readers live in North America and Europe

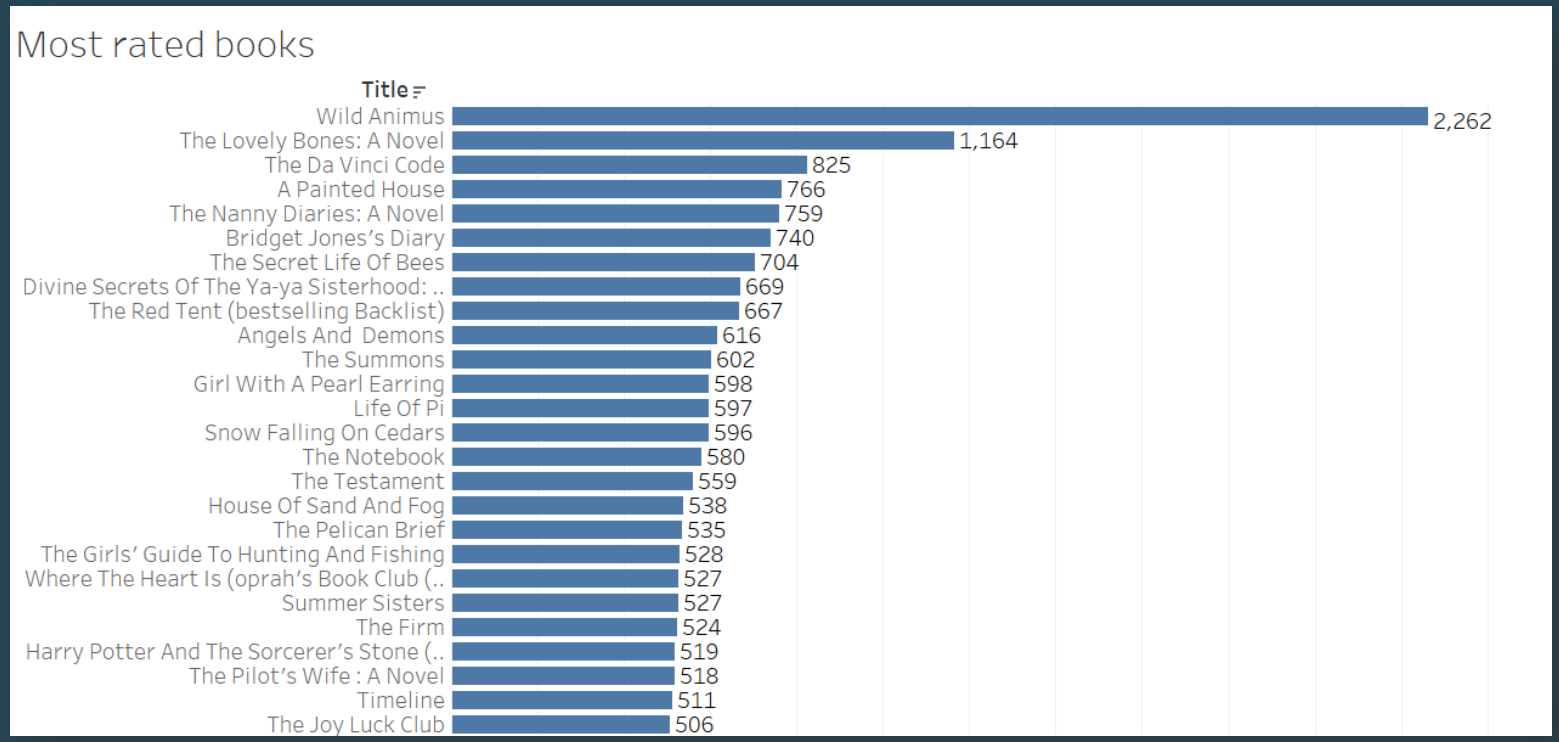
Tableau

Top 20 most rated authors

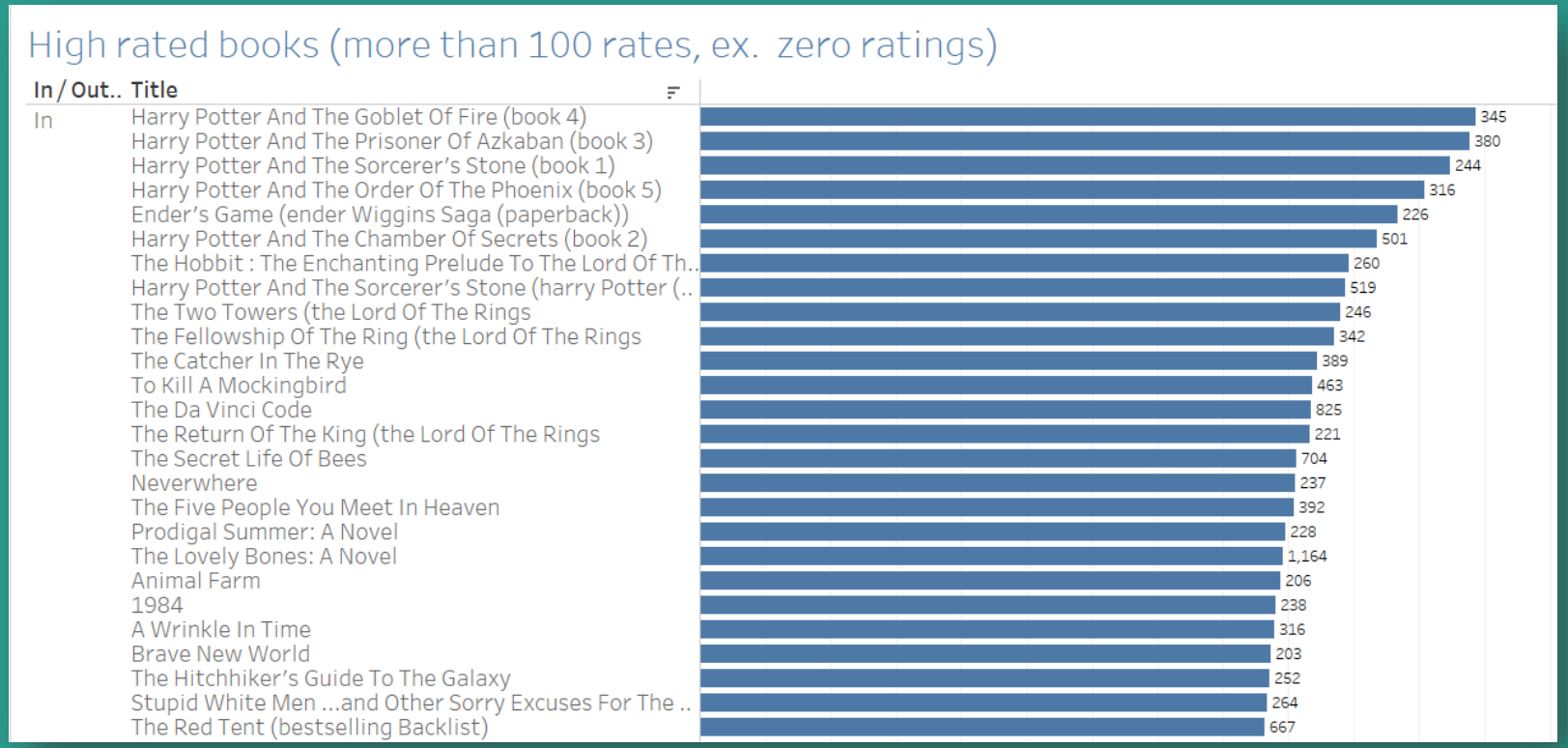


Tableau

Most rated books.
“Wild Animus” is obvious lieder in quantity of ratings

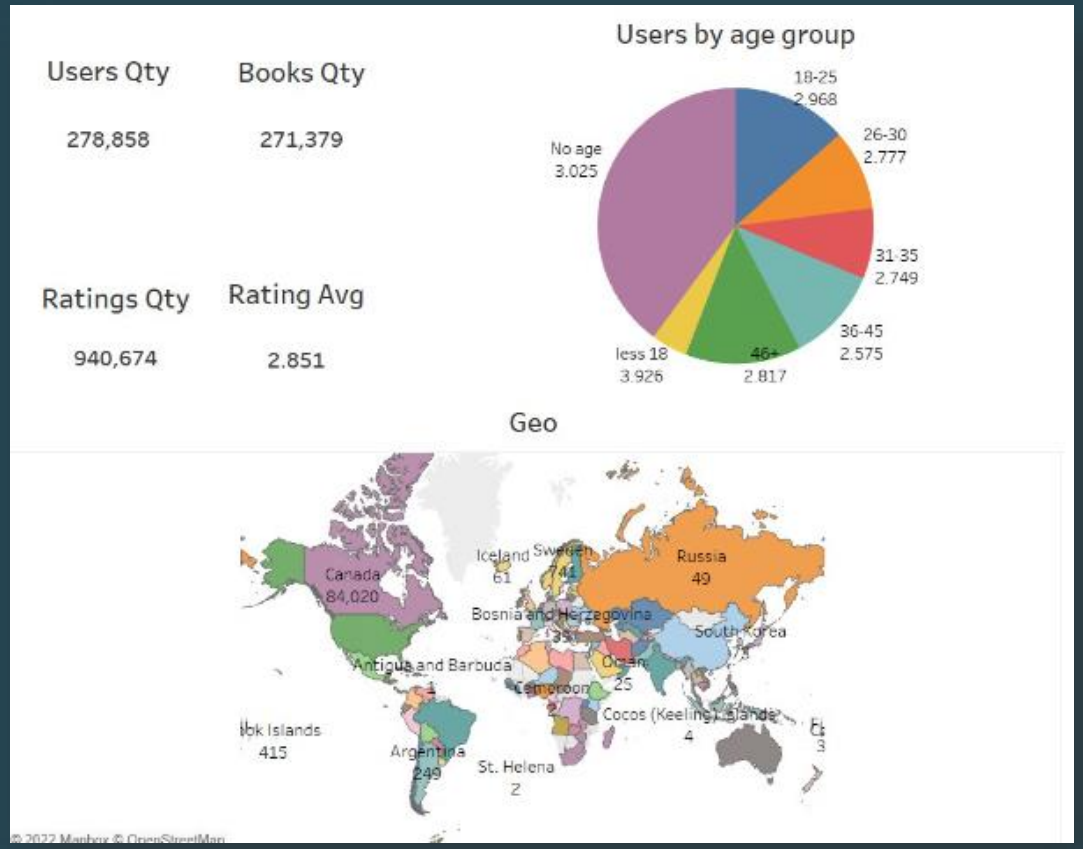


Tableau



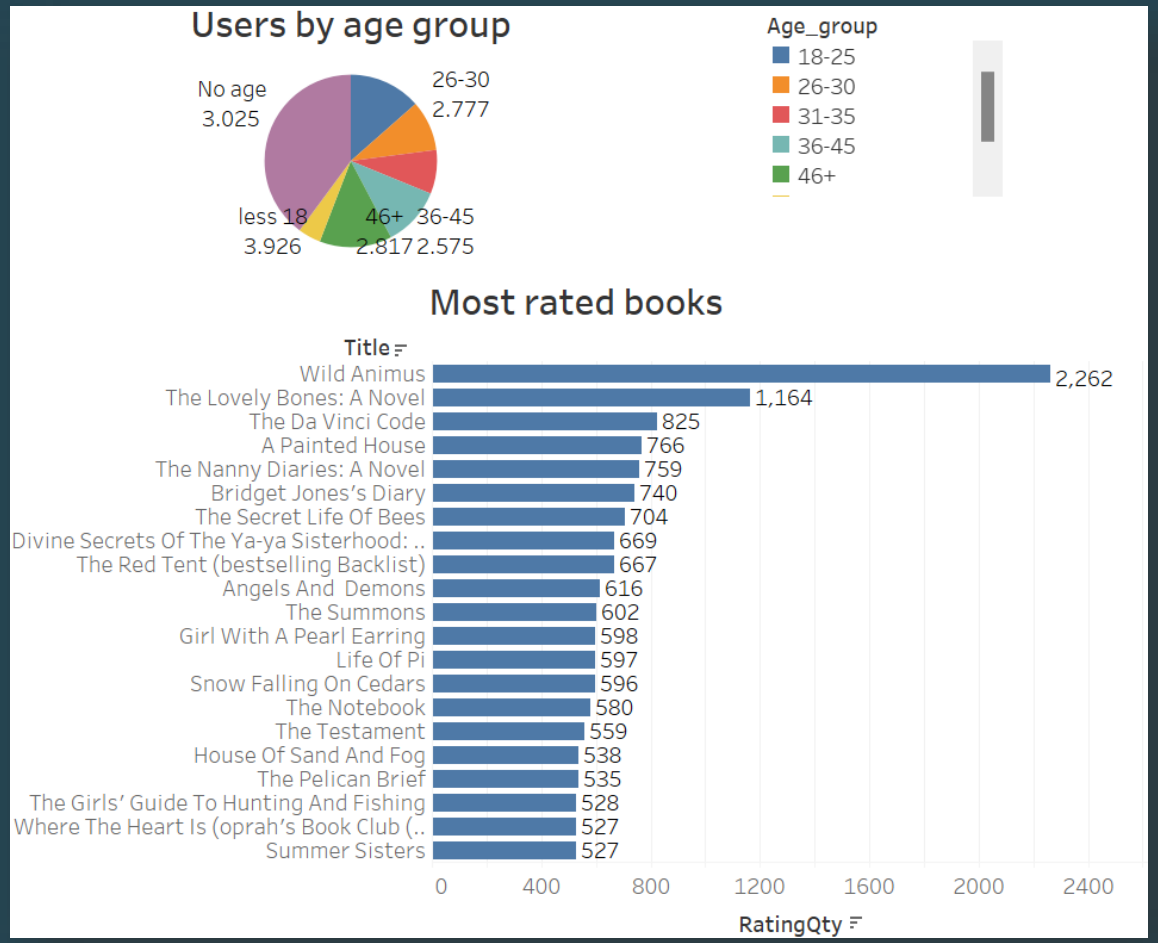
High rated books with more than 100 ratings (zero ratings excluded)

Tableau



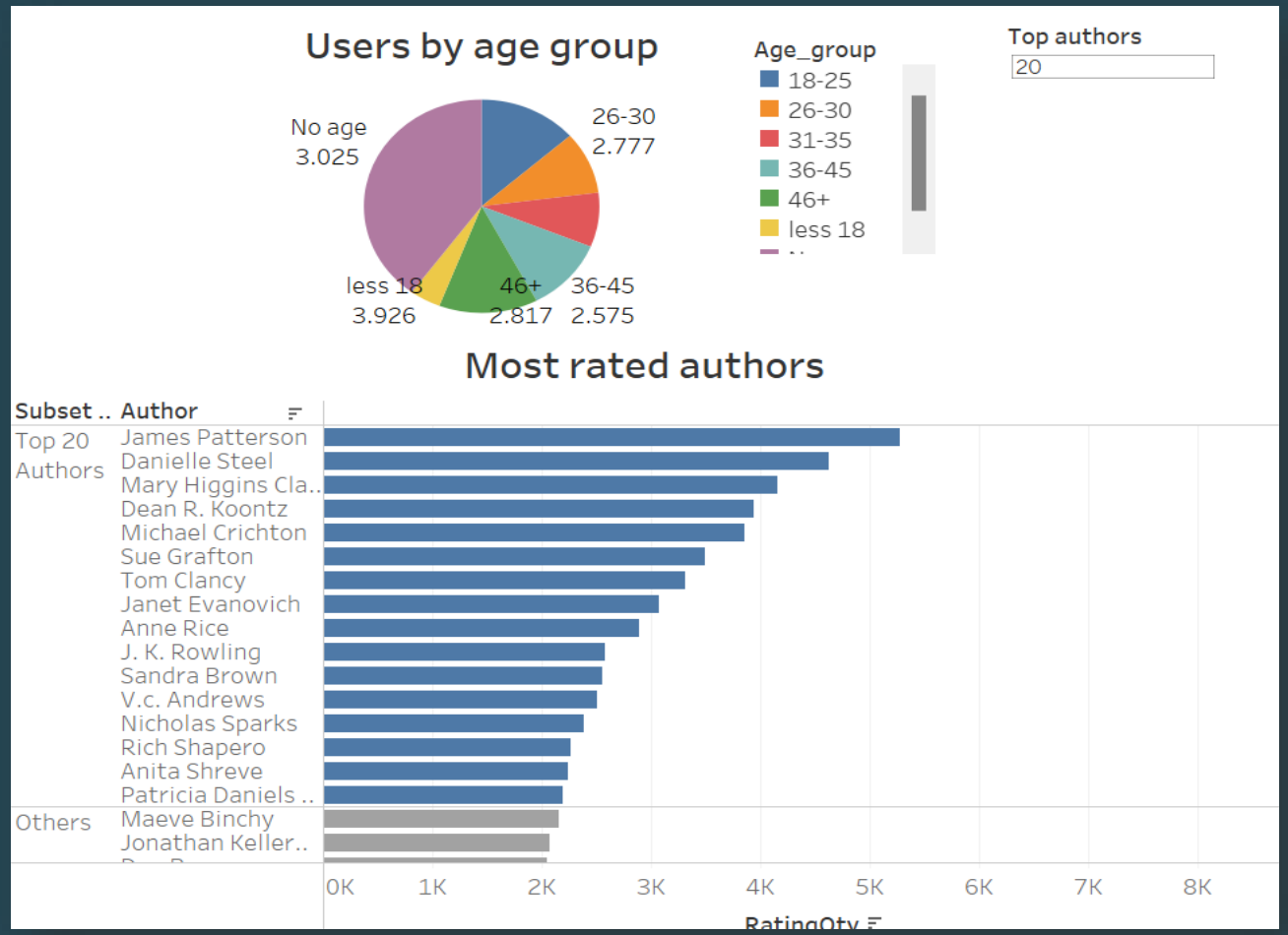
First dashboard shows main data indicators and user's distribution by age groups and country.

Tableau



Second dashboard shows what books are most rated in every user age group.

Tableau



Third dashboard shows what authors are most rated in every user age group.

Challenges

In this project we've met lots of challenges:

- Understanding what problems need to be solved in data clearing process
- Cleaning unfiltered data is quite hard when dealing with big data:
    - ISBN – a string column with letters and numbers. In this case letters may have a meaning, so we decided not to clear letters and not to convert whole column to int.
    - Books table was messy, after splitting columns we saw that data is not organized and therefore we tried to find solutions.
    - Country column – cleaned by function clean_country. We've tried to find similar ways to clean region and city, but unfortunately without success

- Dealing with loosing some data
- Finding a way to download data to local pc
- Dim Date – there is no date dimension in our model because in our fact table key metric is rating, and ratings have no dates. For this reason, we didn't include date dimension in a model

Conclusion

During work on this project, we've concluded that analyzing big data differs from regular data analysis.
Working with big amount of data we had to realize, that we probably won't be able to fix and organize every row in every dataset we had.
Therefore, our goal was to fix main issues in order to see major trends and insights from our data.

Thank you