Book – Crossing Dataset

Big Data Project

Submitting: Alina Greenberg, Natalie Elkin & Larisa Platitcin

The Book Crossing Dataset

- The data was collected from the Book-Crossing Community in a 4-week crawl (August-September 2004).
- It contains 278,858 users (anonymized but with age), providing 1,048,575 ratings (explicit / implicit) over 271,379 books.

Data Upload To PySpark

Uploaded 3 csv files with the following columns:

- **BX-Users.csv** :User-ID, Location, Age

- **BX-Books.csv**: ISBN, Book-Title, Book-Author, Year-Of-Publication, Publisher, Image-URL-S, Image-URL-M, Image-URL-L

- **BX-Books-Ratings.csv**: User-ID, ISBN, Book-Rating

Project Timeline Template

Obtaining the relevant data from multiple sources. In this case, CSV files.

Cleaning and arranging the data using Pyspark, connecting to the data through Tableau Prep and building a star schema data model.

Creating measures and calculated fields through Tableau, while defining which data we wish to present to users.

Creating visualizations and deriving conclusions.

Data Cleaning - BX-Users.csv

User-ID

- Checking for null , empty or like null values (n/a, None, etc.) and deleting these rows.

Age

- Replacing nulls/ empty/ like null values to (-1)
- Replacing Age > 113 to (-1) (the age of the oldest person alive in 2004).

Data Cleaning - BX-Users.csv

Location

- ✧ Column extraction: City, State, Country
- ✧ Replacing null , empty or like null values (n/a, None, etc.) to ‘Unknown’.
- ✧ Replacing all special characters in the three columns with spaces, deleting unnecessary spaces, etc.

```
regexp_replace("State",r"[^a-zA-ZÀ-ÿ $]", " ")
```

- ✧ Importing an up-to-date list of countries and comparing the original list of countries to the up-to-date list using the following functions:

```
SequenceMatcher(None,word,candidate).ratio()
```

and replacing to correct country name (if similarity > 0.75 otherwise unknown).

Data Cleaning - BX-Books.csv

Cleaning steps:

- Drop **URL** columns – will note be used in visualizations or analysis.
- Clean special characters from all columns
- Clean invalid **ISBN** values
  - Filtering by length [10 digits]
  - Creating a function that checks validity
- Replace incorrect **publication year** values with ‘9999’
  - Year = 0
  - Year > 2004
- Replace null values in **book author**, **publisher** and **book title** columns with ‘Unknown’

Data Cleaning – Books Ratings

- User-ID
  - Checking for null , empty or like null values (n/a, None, etc.) and deleting these rows.
  - It was noticed that the number of UserID in USERS data frame is different from the distinct number of UserID in BOOK RATING data frame – using left join in Tableau.

ISBN

- Checking for null , empty or like null values (n/a, None, etc.) and replace with ‘Unknown’.

ISBN

- Creating a function that checks if we have an ISBN with a length different than 10.
- ISBN which is 13 in length, we imported a special library called isbntools and we created a function that turns 13 into 10 - turn the ISBN into valid with a length of 10.
- Creating a function that checks if there are incorrect characters in the ISBN (it was defined that it can contain 10 numbers or 9 with an X at the end, and if the final number leaves no remainder when divided by 11 – Otherwise replace with 'Unknown'.

ISBN

- It was noticed that the number of ISBN in BOOKS data frame is different from the distinct number of ISBN in BOOK RATING data frame – using left join in Tableau.

Book Rating

- Using the function to see the special values that exist in the list.
- The defined range of values should be between 0 and 10 and otherwise will be replaced with -1.

Tableau Prep Option No.1

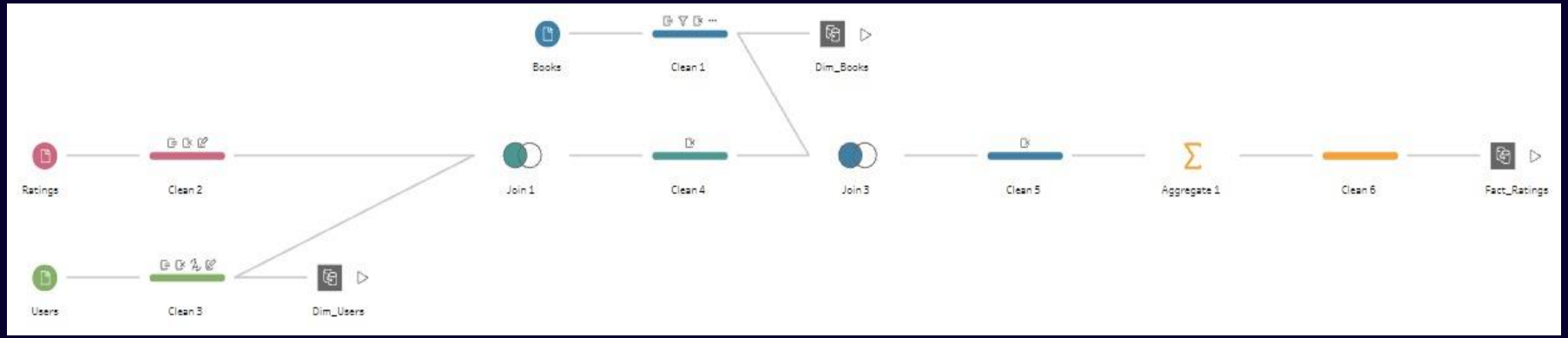
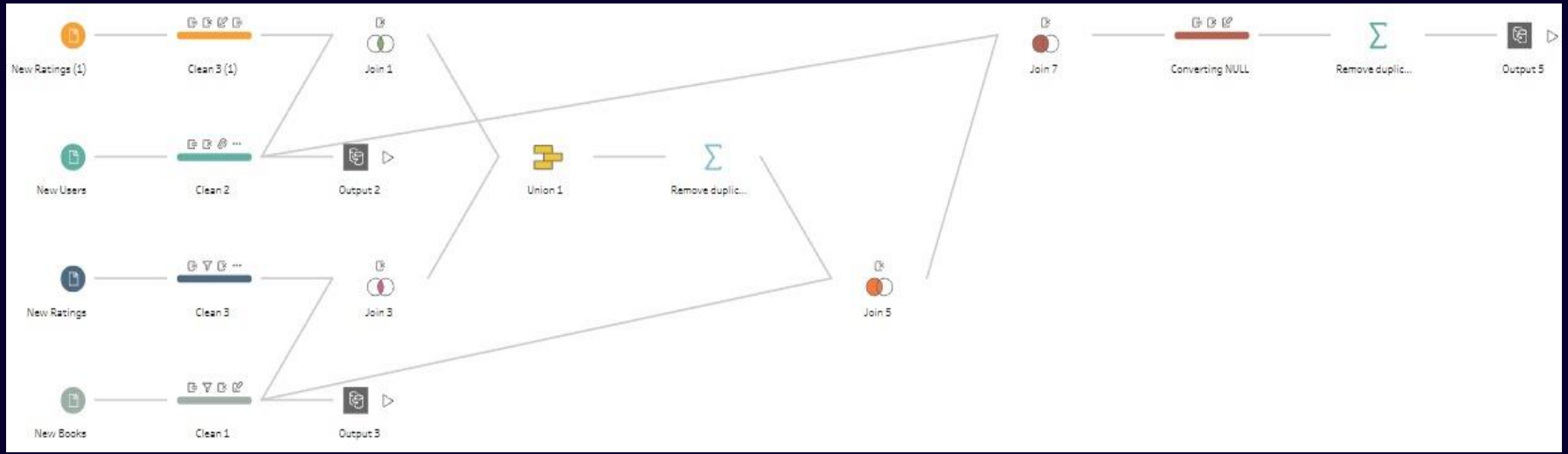
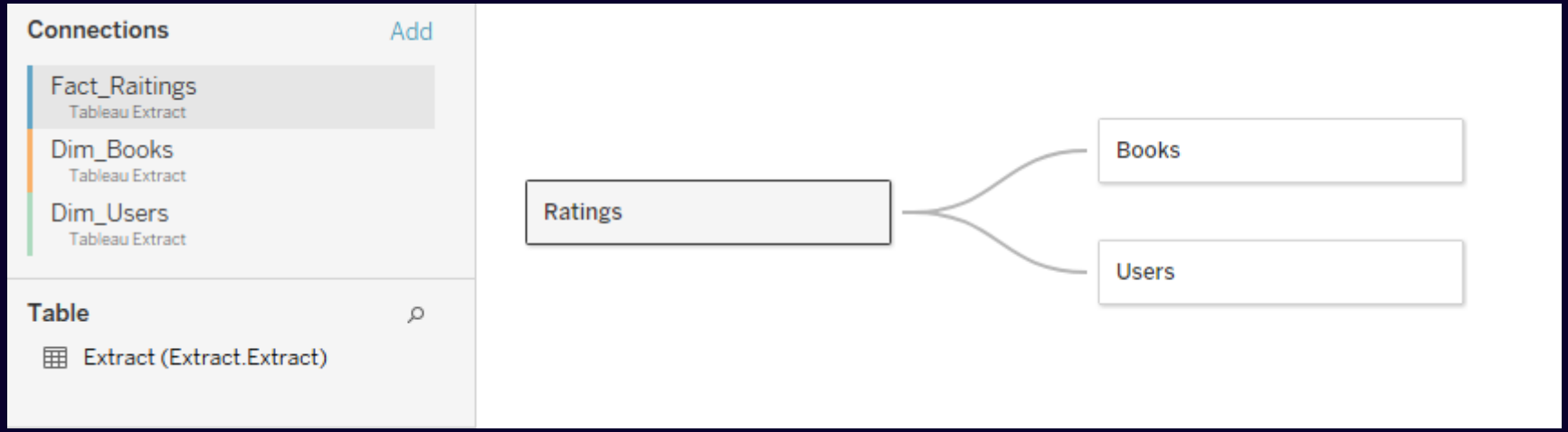


Tableau Prep Option No.2 – Using Union



This was the chosen option.

Star Schema – Tableau Desktop



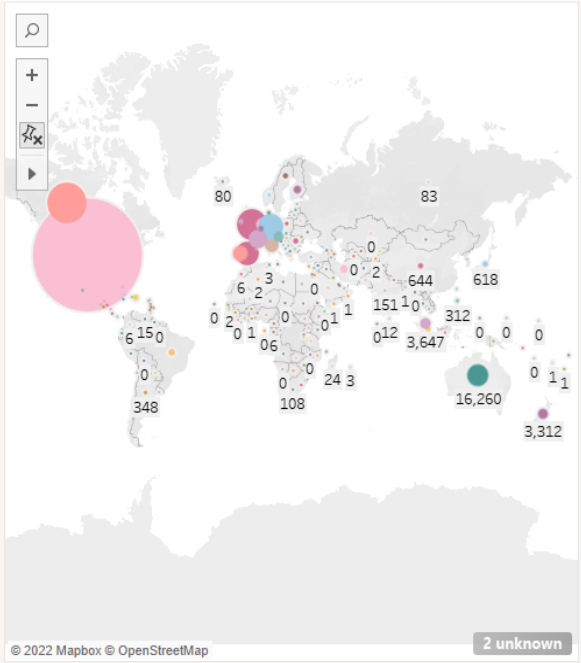
Data Visualiztion

Top 10 Most Favorite Books

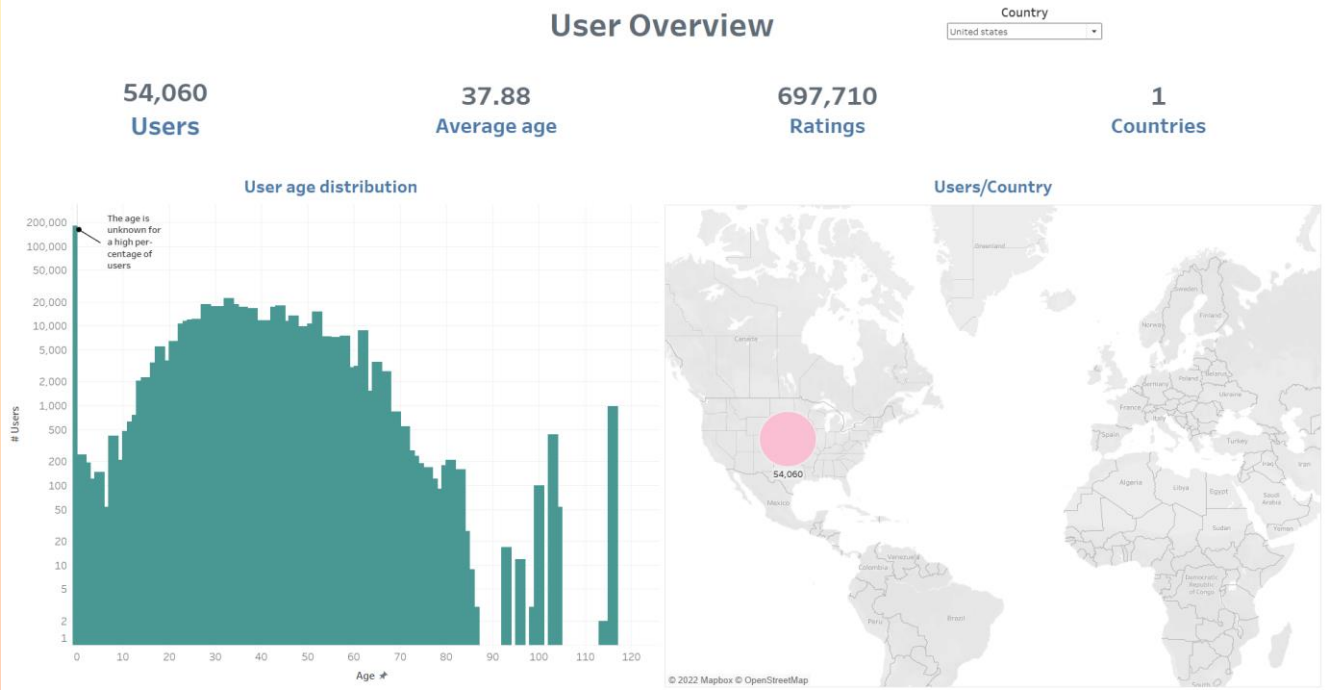
Publication decade	Country	Age (group)
(All)	(All)	(Multiple values)

INDEX	Book Title	Book Author	Avg. Book Ratin
1	Tales and Sketches: Including Twice-Told Tales, Mosses from an Old Manse, and the Snow-Image (Library of America College Editions)	Nathaniel Hawthorne	9.91
2	The Cowboy Rides Away	Betsy Thornton	9.84
3	Skipping Christmas	JOHN GRISHAM	9.79
4	Born of the Sun	Joan Wolf	9.76
5	We've Got Blog: How Weblogs Are Changing Our Culture	Editors of Perseus Publishi..	9.73
6	Red Odyssey: A Journey Through the Soviet Republics	Marat Akchurin	9.73
7	Bait	Karen Robards	9.67
8	The Indian Cemetery (Sugar Creek Gang, No 11)	Paul Hutchens	9.66
9	The Future of Success	Robert B. Reich	9.66
10	Monsieur MalaussÃÂne	Daniel Pennac	9.65

Map



User Overview



- More than 50% of users are from the US, providing more than 50% of ratings.
- Their age distribution is similar to the one of all users.

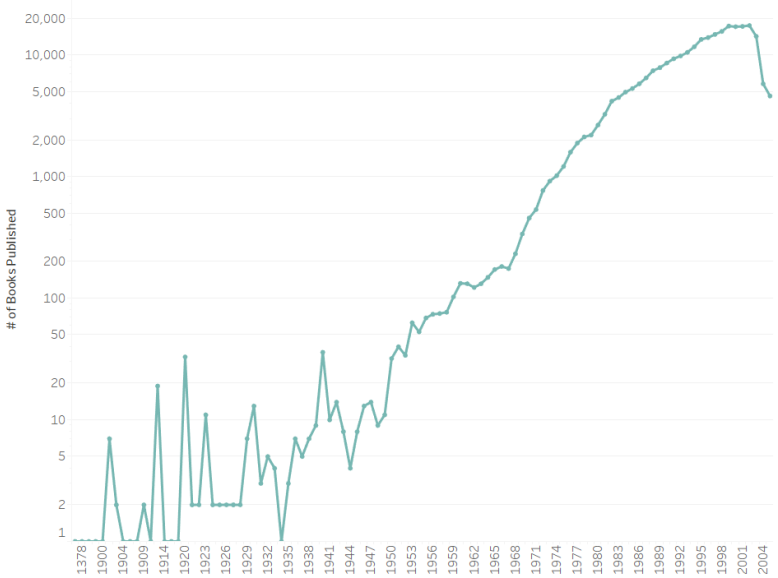
Books Overview

270,949
Books in DB

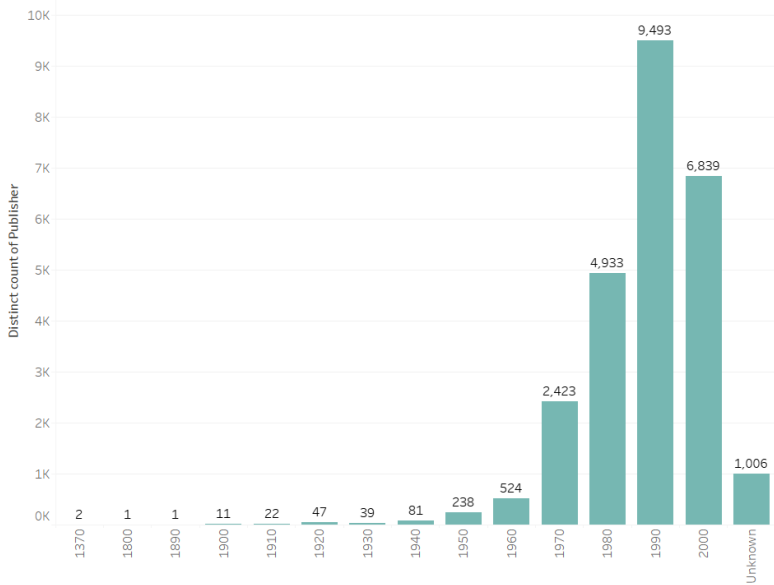
102,002
Authors

16,717
Publishers

Published Books/Year



Active Publishers/Decade

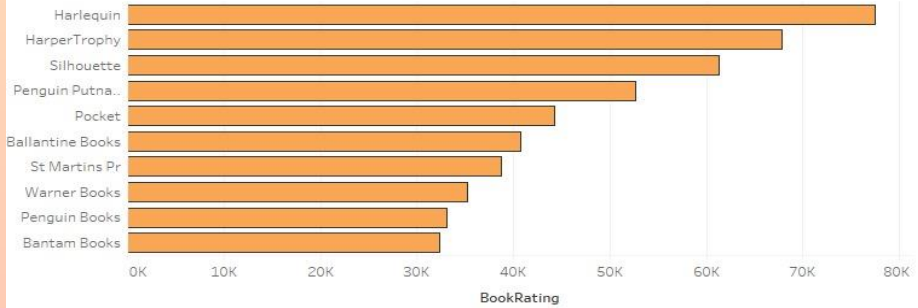


General Information Per Country

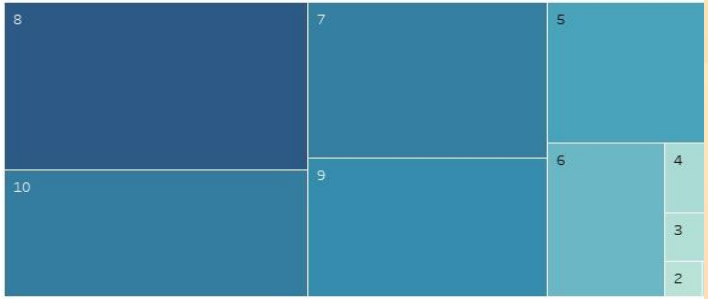
Country
(All)

From which countries the rating is - explicit\implicit

Top 10 Publisher



Which rating is the most given according to the User



What is the highest rated book?

Conclusions

From which countries the rating is - Explicit\Implicit

Asia

It can be seen that the users from these countries gave an explicit rating

Africa

Most of those countries did not rate explictly

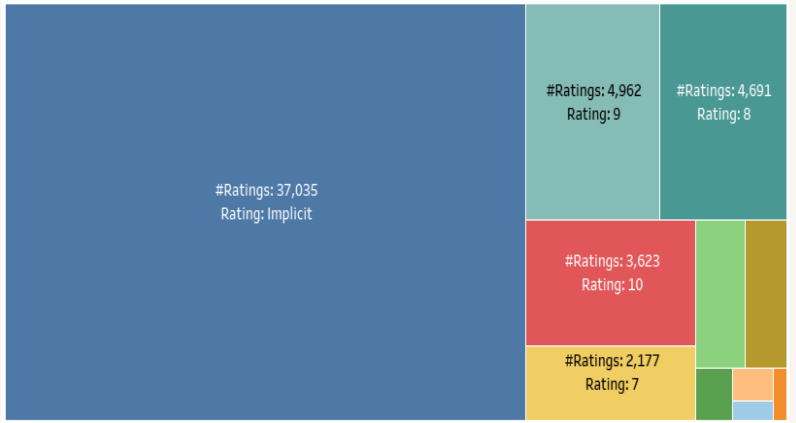
America

These countries gave an explicit rating

- Differences in colors indicate: blue - Explicit, light blue - Implicit

Top 10 Most Rated Books

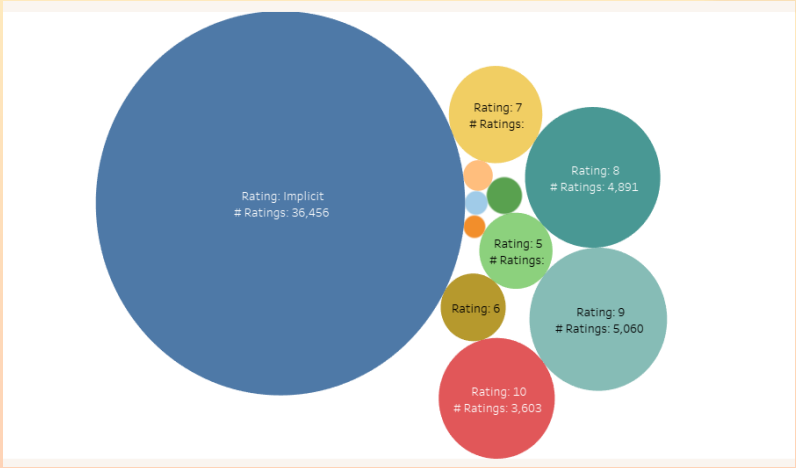
INDEX	Book Title	Book Author	# Rating..
1	Mama Don't Allow	Thacher Hurd	13324
2	The World Encyclopedia of Comics	Maurice Horn	7430
3	The love of Chinese cooking	Kenneth H. C Lo	6089
4	King of the Night: The Life of Johnny Carson	Laurence Leamer	5829
5	2001 a Space Odyssey	Arthur C Clarke	5789
6	A Man Cannot Cry	Gloria Keverne	4662
7	How to Make Basic Investment Decisions (Real Life, Real ..	Neal Ochsner	3364
8	The American Century Cookbook	JEAN ANDERSON	3071
9	Vancouver	Gail Sattler	3057
10	Crossing the Unknown Sea: Work As a Pilgrimage of Iden..	David Whyte	2973



- Relates to books that were mentioned in a review [with implicit and explicit ratings]
- A significant amount of the ratings was implicit
- Most of the explicit ratings of these books are high, >7.

Top 10 Most Rated Authors

INDEX	Book Author	# Ratings per author
1	Thacher Hurd	13,389
2	Maurice Horn	7,430
3	Kenneth H. C Lo	6,089
4	Laurence Leamer	5,834
5	Arthur C Clarke	5,789
6	Gloria Keverne	4,678
7	Neal Ochsner	3,364
8	JEAN ANDERSON	3,071
9	Gail Sattler	3,058
10	Paul Theroux	3,043



- There is a match between most of the top-rated authors and top-rated books

Challenges

- A significant challenge was dealing with inconsistent data documentation that hindered the data cleansing process.
- Data cleaning required dealing with inconsistent number of delimiters that made it challenging to extract into a data frame, and unique characters that had to be addressed.
- An additional challenge is the multiple NULL and unknown values in the entire dataset that had to be addressed during the cleaning and visualization process.
- Using python based UDF in pyspark caused extremely long run time. We tried to optimize the run time and found relevant solutions but couldn’t implement them.