

Books Analysis

Data profiling and cleaning of a data set about books



Stages

1. Choose dataset;
2. Unclean the dataset to simulate real life scenarios of messy data;
3. Identify 3 user stories the dataset should answer;
4. Proceed with data profiling and build a quality dimensions table;
5. Clean the data;
6. Answer the user stories

1. Dataset



Additional columns:

year = year from the publication_date column
valid_rating = Boolean value based on text_reviews_count and ratings_count

bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher	year	valid_rating
1	Harry Potter and the Half-Blood Prince (Harry Potter #6)	J.K. Rowling/Mary GrandPré	4.57	439785960	9.78044E+12	eng	652	2095690	27591	9/16/2006	Scholastic Inc.	2006	TRUE
2	Harry Potter and the Order of the Phoenix (Harry Potter #5)	J.K. Rowling/Mary GrandPré	4.49	439358078	9.78044E+12	eng	870	2153167	29221	09-01-04	Scholastic Inc.	2004	TRUE
4	Harry Potter and the Chamber of Secrets (Harry Potter #2)	J.K. Rowling	4.42	439554896	9.78044E+12	eng	352	6333	244	11-01-03	Scholastic	2003	TRUE
5	Harry Potter and the Prisoner of Azkaban (Harry Potter #3)	J.K. Rowling/Mary GrandPré	4.56	043965548X	9.78044E+12	eng	435	2339585	36325	05-01-04	Scholastic Inc.	2004	TRUE
8	Harry Potter Boxed Set Books 1-5 (Harry Potter #1-5)	J.K. Rowling/Mary GrandPré	4.78	439682584	9.78044E+12	eng	2690	41428	164	9/13/2004	Scholastic	2004	TRUE
9	Unauthorized Harry Potter Book Seven News: "Half-Blood Pri W. Frederick Zimmerman		3.74	976540606	9.78098E+12	en-US	152	19	1	4/26/2005	Nimble Books	2005	FALSE
10	Harry Potter Collection (Harry Potter #1-6)	J.K. Rowling	4.73	439827604	9.78044E+12	eng	3342	28242	808	09-12-05	Scholastic	2005	TRUE
12	The Ultimate Hitchhiker's Guide: Five Complete Novels and C Douglas Adams		4.38	517226952	9.78052E+12	eng	815	3628	254	11-01-05	Gramercy Books	2005	TRUE
13	The Ultimate Hitchhiker's Guide to the Galaxy (Hitchhiker's G Douglas Adams		4.38	345453743	9.78035E+12	eng	815	249558	4080	4/30/2002	Del Rey Books	2002	TRUE

2. Uncleaning

Used python pandas package to transform the dataset.

Created functions to unclean at least 25% of the data.

Example:

```
def unclean_title(df):  
    sample = select_rows(df)  
    return sample.title.str.upper()
```

3. User Stories

- As a librarian I want to know which are the top rated books by year for the last 10 years so that I can make recommendations to our users.
- As a student I want to know which are the top 10 books and respective authors that have the most text reviews so that I can study what leads to the most user engagement in books.
- As a new fan of Harry Potter I would like to know how many Harry Potter related books exist in english.

4. Data profiling

Structural analysis:

1. Connection overview analysis

Column analysis:

1. Simple statistics;
2. Functional dependency analysis;
3. Nominal analysis;
4. Pattern frequency analysis;
5. Discrete data analysis;
6. Summary statistics analysis



Quality dimensions table

Columns	Primary key	Uniqueness	Completeness	Consistency	Accuracy	Conformity	Validity	Currency & Timeliness	Reliability & Credibility
book ID	Primary key	Yes	Yes	Yes	-	-	Yes	-	Yes
title	-	-	Yes	No	Yes	Yes	Yes	-	Yes
authors	-	-	Yes	Yes	Yes	No	Yes	-	Yes
average_rating	-	-	Yes	Yes	No	Yes	No	No	Yes
isbn	Possible primary key	Yes	Yes	Yes	-	-	Yes	-	Yes
isbn13	Possible primary key	Yes	Yes	Yes	-	-	Yes	-	Yes
language_code	-	-	Yes	No	Yes	No	Yes	-	Yes
num_pages	-	-	Yes	Yes	Yes	Yes	Yes	-	Yes
ratings_count	-	-	Yes	Yes	Yes	Yes	Yes	No	Yes
text_reviews_count	-	-	Yes	Yes	Yes	Yes	Yes	No	Yes
publication_date	-	-	Yes	No	Yes	No	Yes	-	Yes
publisher	-	-	Yes	Yes	Yes	No	Yes	-	Yes
year	-	-	Yes	Yes	Yes	Yes	No	-	Yes
valid_rating	-	-	Yes	No	Yes	No	No	No	Yes

5. Data Cleaning

Recipe methods used:

- Text transformations
- Regular expression transformations
- Column based transformations
- Clustering
 - nearest neighbor -> levenshtein & ppm
 - key collision -> fingerprint & ngram-fingerprint

The screenshot shows the OpenRefine web application interface. At the top, the title bar reads "OpenRefine data_management_1". Below this, there are two tabs: "Facet / Filter" and "Undo / Redo 82 / 82". To the right of these tabs are two buttons: "Extract..." and "Apply...". Below the tabs is a "Filter:" input field. The main area of the interface displays a list of 11 recipes, each with a number and a description of the transformation applied to a specific column and number of cells. The recipes are as follows:



0. Create project
1. Text transform on 8254 cells in column title: value.toTitlecase()
2. Text transform on 2073 cells in column publisher: value.toTitlecase()
3. Mass edit 2658 cells in column authors
4. Mass edit 14 cells in column authors
5. Mass edit 14 cells in column authors
6. Mass edit 9 cells in column authors
7. Edit single cell on row 2931, column authors
8. Text transform on 11123 cells in column average_rating: value.toNumber()
9. Text transform on 2781 cells in column average_rating: grel:abs(value)
10. Text transform on 4012 cells in column language_code: value.toLowerCase()
11. Text transform on 1629 cells in column language_code: grel:value.replace(/en-[a-z][a-z]/, "eng")

6. Answering the user stories

1st user story:

As a librarian I want to know which are the top rated books by year for the last 10 years so that I can make recommendations to our users.

Answer:

Result Grid  Filter Rows: <input type="text"/> Export:  Wrap Cell Content:			
	title	average_rating	year
▶	A Quick Bite (argeneau #1)	3.91	2020
	The Wish Giver: Three Tales Of Coven Tree	3.83	2019
	Ariel: The Restored Edition	4.27	2018
	Rick Steves' Europe Through The Back Door	4.24	2017
	When The Heart Waits: Spiritual Direction For Li...	4.13	2016
	Hamlet's Mill: An Essay Investigating The Origin...	4.29	2015
	Libraries	4.28	2014
	Soul Mates: Honouring The Mysteries Of Love A...	3.99	2013
	J.R.R. Tolkien 4-book Boxed Set: The Hobbit An...	4.59	2012
	Sand And Foam	4.08	2011

2nd user story:

As a student I want to know which are the top 10 books and respective authors that have the most text reviews so that I can study what leads to the most user engagement in books.

Answer:


Result Grid   Filter Rows: <input type="text"/> Export:  Wrap Cell Content:  Fetch rows: 			
	title	authors	text_reviews_count
▶	Twilight (twilight #1)	Stephenie Meyer	94265
	The Book Thief	Markus Zusak/Cao Xuân Việt Khương	86881
	The Giver (the Giver #1)	Lois Lowry	56604
	The Alchemist	Paulo Coelho/Alan R. Clarke/Özdemir İnce	55843
	Water For Elephants	Sara Gruen	52759
	The Lightning Thief (percy Jackson And The Olympians #1)	Rick Riordan	47951
	Eat Pray Love	Elizabeth Gilbert	47620
	The Glass Castle	Jeannette Walls	46176
	The Catcher In The Rye	J.D. Salinger	43499
	Harry Potter And The Prisoner Of Azkaban (harry Potter #3)	J.K. Rowling/Mary GrandPré	36325

3rd user story:

As a new fan of Harry Potter I would like to know how many Harry Potter related books exist in english.

Answer:

```
1 • USE booksdb;
2 • SELECT COUNT(title) FROM clean_books
3   WHERE title LIKE '%Harry Potter%' AND language_code = 'eng'
```



The screenshot shows a SQL query execution interface. At the top, the query is displayed in a code editor with line numbers 1, 2, and 3. Below the query, there is a toolbar with icons for 'Result Grid', 'Filter Rows', 'Export', and 'Wrap Cell Content'. The 'Result Grid' icon is active. Below the toolbar, a table displays the results of the query. The table has two columns: 'COUNT(title)' and a value '18'.

COUNT(title)
18

Thank you! Questions?

Github can be found [here](#) and original dataset [here](#)