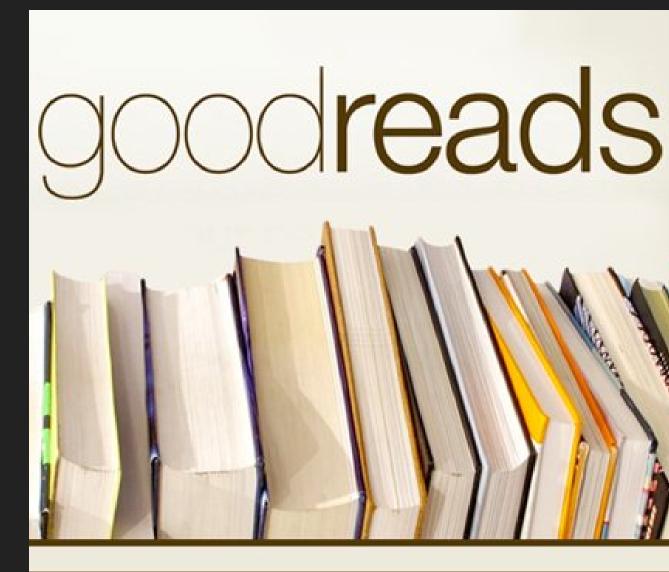
Books Analysis

Data profiling and cleaning of a data set about books



Stages

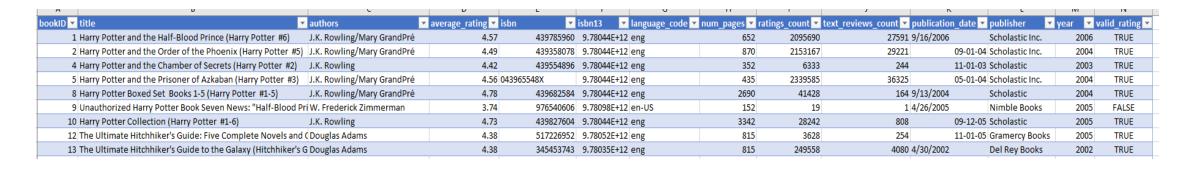
- Choose dataset;
- 2. Unclean the dataset to simulate real life scenarios of messy data;
- 3. Identify 3 user stories the dataset should answer;
- 4. Proceed with data profiling and build a quality dimensions table;
- 5. Clean the data;
- 6. Answer the user stories

1. Dataset



Additional columns:

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



2. Uncleaning

Columns affected:

```
title, authors, publisher, average_rating, language_code, year, valid_rating
```

Quality dimensions affected:

```
conformity, validity, consistency
```



Example:

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

Unclean dataset

bookI	title 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	0439785960	9780439785969	eng	652	2095690	27591	9/16/2006	Scholastic Inc.	2006	1
2	HARRY POTTER AND THE ORDER OF THE PHOENIX (HARRY POTTER #5)	J.K. Rowling/Mary GrandPré	-4.49	0439358078	9780439358071	eng	870	2153167	29221	9/1/2004	Scholastic Inc.P	2004	1
4	Harry Potter and the Chamber of Secrets (Harry Potter #2)	J.K. Rowling	4.42	0439554896	9780439554893	eng	352	6333	244	11/1/2003	Scholastic	2003	Yes
5	Harry Potter and the Prisoner of Azkaban (Harry Potter #3)	J.K. Rowling/Mary GrandPré	4.56	043965548X	9780439655484	eng	435	2339585	36325	5/1/2004	Scholastic Inc.	2004	1
3	Harry Potter Boxed Set Books 1-5 (Harry Potter #1-5)	J.K. Rowling/Mary GrandPré	-4.78	0439682584	9780439682589	ENG	2690	41428	164	9/13/2004	Scholastic	2004	1
9	Unauthorized Harry Potter Book Seven News: "Half-Blood Prince" Analysis and Speculation	W. Frederick Zimmerman	3.74	0976540606	9780976540601	en-US	152	19	1	4/26/2005	Nimble Books	2005	No
10	Harry Potter Collection (Harry Potter #1-6)	J.K. Rowling	-4.73	0439827604	9780439827607	eng	3342	28242	808	9/12/2005	Scholastic	2005	1
12	The Ultimate Hitchhiker's Guide: Five Complete Novels and One Story (Hitchhiker's Guide to the Galaxy #1-5)	Douglas Adams	4.38	0517226952	9780517226957	eng	815	3628	254	11/1/2005	Gramercy BooksP	2005	Yes
13	The Ultimate Hitchhiker's Guide to the Galaxy (Hitchhiker's Guide to the Galaxy #1-5)	Douglas Adams	-4.38	0345453743	9780345453747	eng	815	249558	4080	4/30/2002	Del Rey Books	2002	1
14	THE HITCHHIKER'S GUIDE TO THE GALAXY (HITCHHIKER'S GUIDE TO THE GALAXY #1)	Douglas Adams	-4.22	1400052920	9781400052929	eng	215	4930	460	8/3/2004	Crown	2004	1
16	The Hitchhiker's Guide to the Galaxy (Hitchhiker's Guide to the Galaxy #1)	Douglas Adams/Stephen Fry	4.22	0739322206	9780739322208	eng	6	1266	253	3/23/2005	Random House Audio	2005	1
18	The Ultimate Hitchhiker's Guide (Hitchhiker's Guide to the Galaxy #1-5)	Douglas Adams	-4.38	0517149257	9780517149256	ENG	815	2877	195	1/17/1996	Wings Books	2000	1
21	A SHORT HISTORY OF NEARLY EVERYTHING	Bill Bryson	4.21	076790818X	9780767908184	eng	544	248558	9396	9/14/2004	Broadway BooksP	2004	1
22	Bill Bryson's African Diary	Bill BrysonU	3.44	0767915062	9780767915069	eng	55	7270	499	12/3/2002	Broadway BooksP	2003	1
23	Bryson's Dictionary of Troublesome Words: A Writer's Guide to Getting It Right	Bill Bryson	3.87	0767910435	9780767910439	ENG	256	2088	131	9/14/2004	Broadway Books	2004	1
24	IN A SUNBURNED COUNTRY	Bill Bryson	4.07	0767903862	9780767903868	ENG	335	72451	4245	5/15/2001	Broadway Books	2001	1
25	I'm a Stranger Here Myself: Notes on Returning to America After Twenty Years Away	Bill Bryson	3.9	076790382X	9780767903820	eng	304	49240	2211	6/28/2000	Broadway Books	2008	1
26	THE LOST CONTINENT: TRAVELS IN SMALL TOWN AMERICA	Bill BrysonU	3.83	0060920084	9780060920081	eng	299	45712	2257	8/28/1990	William Morrow Paperbacks	1990	Yes
27	NEITHER HERE NOR THERE: TRAVELS IN EUROPE	Bill Bryson	-3.86	0380713802	9780380713806	ENG	254	48701	2238	3/28/1993	William Morrow Paperbacks	1993	1
28	Notes from a Small Island	Bill Bryson	3.91	0380727501	9780380727506	eng	324	80609	3301	5/28/1997	William Morrow PaperbacksP	1997	Yes

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





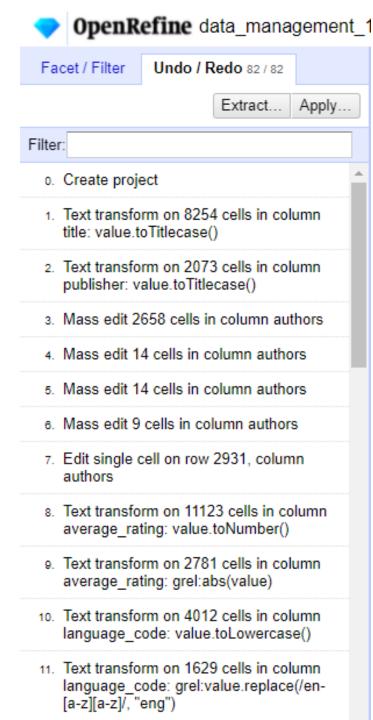
Quality dimensions table

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



Data Management 1 • Prof. Dr. Ajinkya Prabhune • Prof. Ashish Chouhan • Std: Gonçalo Coutinho • Id: 11016322

Clean dataset

ind	ex	bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher	year	valid_rating
0		1	Harry Potter And The Half-blood Prince (harry Potter #6)	J.K. Rowling/Mary GrandPré		0439785960	9780439785969	eng	652	2095690	27591	16/09/2006	Scholastic	2006	
1		2	Harry Potter And The Order Of The Phoenix (harry Potter #5)	J.K. Rowling/Mary GrandPré	4.49	0439358078	9780439358071	eng	870	2153167	29221	01/09/2004	Scholastic	2004	1
2		4	Harry Potter And The Chamber Of Secrets (harry Potter #2)	J.K. Rowling	4.42	0439554896	9780439554893	eng	352	6333	244	01/11/2003	Scholastic	2003	1
3	1	5	Harry Potter And The Prisoner Of Azkaban (harry Potter #3)	J.K. Rowling/Mary GrandPré	4.56	043965548X	9780439655484	eng	435	2339585	36325	01/05/2004	Scholastic	2004	1
4		8	Harry Potter Boxed Set Books 1-5 (harry Potter #1-5)	J.K. Rowling/Mary GrandPré	4.78	0439682584	9780439682589	eng	2690	41428	164	13/09/2004	Scholastic	2004	1
5		9	Unauthorized Harry Potter Book Seven News: "half-blood Prince" Analysis And Speculation	W. Frederick Zimmerman	3.74	0976540606	9780976540601	eng	152	19	1	26/04/2005	Nimble Books	2005	0
6		10	Harry Potter Collection (harry Potter #1-6)	J.K. Rowling	4.73	0439827604	9780439827607	eng	3342	28242	808	12/09/2005	Scholastic	2005	1
7		12	The Ultimate Hitchhiker's Guide: Five Complete Novels And One Story (hitchhiker's Guide To The Galaxy #1-5)	Douglas Adams	4.38	0517226952	9780517226957	eng	815	3628	254	01/11/2005	Gramercy Books	2005	1
8		13	The Ultimate Hitchhiker's Guide To The Galaxy (hitchhiker's Guide To The Galaxy #1-5)	Douglas Adams	4.38	0345453743	9780345453747	eng	815	249558	4080	30/04/2002	Del Rey	2002	1
9		14	The Hitchhiker's Guide To The Galaxy (hitchhiker's Guide To The Galaxy #1)	Douglas Adams	4.22	1400052920	9781400052929	eng	215	4930	460	03/08/2004	Crown	2004	1
10		16	The Hitchhiker's Guide To The Galaxy (hitchhiker's Guide To The Galaxy #1)	Douglas Adams/Stephen Fry	4.22	0739322206	9780739322208	eng	6	1266	253	23/03/2005	Random House	2005	1
11		18	The Ultimate Hitchhiker's Guide (hitchhiker's Guide To The Galaxy #1-5)	Douglas Adams	4.38	0517149257	9780517149256	eng	815	2877	195	17/01/1996	Wings Books	1996	1
12		21	A Short History Of Nearly Everything	Bill Bryson	4.21	076790818X	9780767908184	eng	544	248558	9396	14/09/2004	Broadway Books	2004	1
13		22	Bill Bryson's African Diary	Bill Bryson	3.44	0767915062	9780767915069	eng	55	7270	499	03/12/2002	Broadway Books	2002	1
14		23	Bryson's Dictionary Of Troublesome Words: A Writer's Guide To Getting It Right	Bill Bryson	3.87	0767910435	9780767910439	eng	256	2088	131	14/09/2004	Broadway Books	2004	1
15		24	In A Sunburned Country	Bill Bryson	4.07	0767903862	9780767903868	eng	335	72451	4245	15/05/2001	Broadway Books	2001	1
16		25	I'm A Stranger Here Myself: Notes On Returning To America After Twenty Years Away	Bill Bryson	3.9	076790382X	9780767903820	eng	304	49240	2211	28/06/2000	Broadway Books	2000	1
17		26	The Lost Continent: Travels In Small Town America	Bill Bryson	3.83	0060920084	9780060920081	eng	299	45712	2257	28/08/1990	William Morrow	. 1990	1
18		27	Neither Here Nor There: Travels In Europe	Bill Bryson	3.86	0380713802	9780380713806	eng	254	48701	2238	28/03/1993	William Morrow	. 1993	1
19		28	Notes From A Small Island	Bill Bryson	3.91	0380727501	9780380727506	eng	324	80609	3301	28/05/1997	William Morrow	. 1997	1

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:



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.

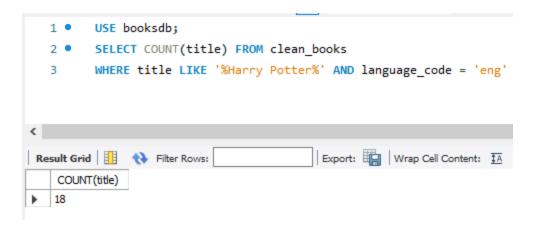
Answer:



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:



Thank you! Questions?

Github can be found here and original dataset here