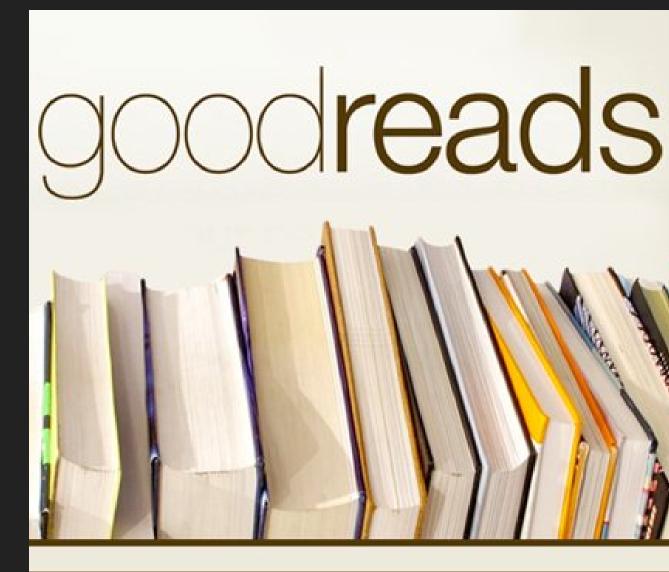
Books Analysis

Data profiling and cleaning of a data set about books



Stages

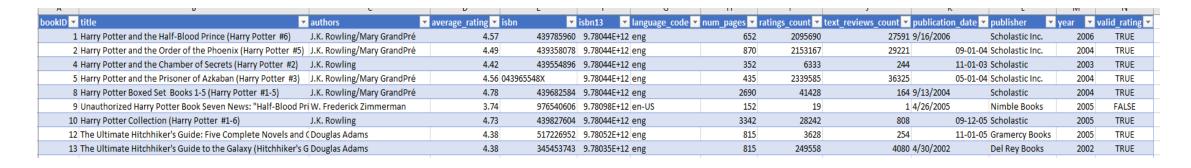
- Choose dataset;
- 2. Unclean the dataset to simulate real life scenarios of messy data;
- 3. Identify 3 user stories the dataset should answer;
- 4. Proceed with data profiling and build a quality dimensions table;
- 5. Clean the data;
- 6. Answer the user stories

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

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

OpenRefine data_management_1 Undo / Redo 82 / 82 Facet / Filter Extract... Apply... Filter: Create project 1. Text transform on 8254 cells in column title: value.toTitlecase() Text transform on 2073 cells in column. publisher: value.toTitlecase() 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 Text transform on 11123 cells in column average rating: value.toNumber() 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")

Data Management 1 • Professors: Prof. Dr. Ajinkya Prabhune, Prof. Ashish Chouhan • Student: Gonçalo Coutinho • StudentId: 11016322

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.

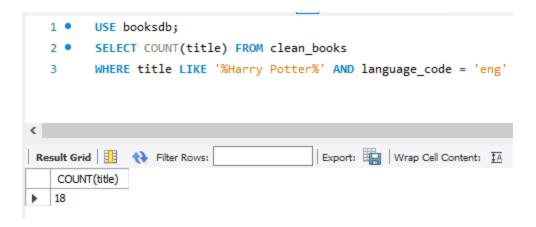
Answer:



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