Best books database

Extraction-Transformation-Loading (ETL)

Ana Kelyna Siliceo
Margarita Garza
Cristina Carmona
Marcelo García

Abstract

For the use of the public in general and for those who love books, a database is constructed and loaded into a server. The main difference between this and other existing databases is that more fields are considered as part of the junction of three already cleaned sources of information: "List of best-selling books" retrieved from Wikipedia through web-scraping, Kagle website provides a CSV file (Goodreads-books) with clean information from Good Reads Api, and Britannica's website is the source of the 12 Novels Considered the "Greatest Book Ever Written". A database with 3 tables, 174 individual books, and 9 main characteristics is available in pgAdmin4 SQL format.

Methodology

According to some Guru99 the process for ETL follows three main steps:

1) Extraction

This step can be done by full extraction, partial extraction-without update notification, and partial extraction-with update notification. To validate extraction it is recommended to:

1.1 Reconcile records with the source data
1.2 Make sure that no spam/unwanted data loaded
1.3 Data type check
1.4 Remove all types of duplicate/fragmented data
1.5 Check whether all the keys are in place or not

2) Transformation

After extracting data, most information needs to be cleansed, mapped and transformed in order to be useful for reporting, it is advisable to consider:

2.1 Different spelling of the same concept
2.2 Different account numbers are generated by various applications for the same customer.
2.3 If some data requires blanks
2.4 Invalid entries or capture mistakes

Useful validation for data transformation includes:

2.5 Standardization by filtering, using rules and lookup tables
2.6 Set characters and units of measurements
2.7 Data threshold and required fields
2.8 Cleaning
2.9 Split, merge or transpose data
2.10 Using complex data validation like error management

3) Loading

When this final step is made in a high data volume that needs to be loaded in a short period of time, the loading process should be optimized. It is also advisable to set a recovery mechanism that ensures no data is lost. This methodology resumes three types of loading: Initial Load refers to the population of Data Warehouse tables in its totality, Incremental Load can apply ongoing changes periodically and Full Refresh, updates by erasing and replace when needed.

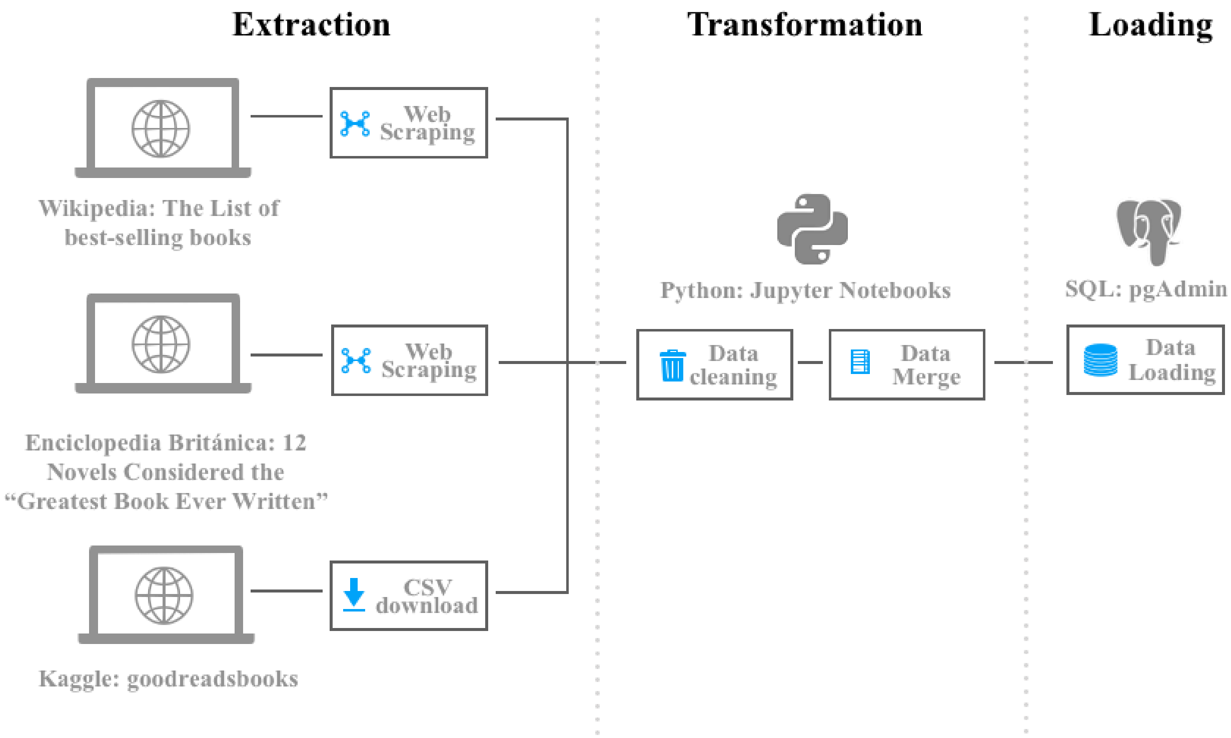
Load verification
3.1 Ensure key field data is not missing or null
3.2 Model views on main tables and BI reports
3.3 Check dimension and the history table

Results

Best books database SQL pgAdmin 4		
Number of books	Number of elements in the database	174

Book information	Columns containing information about a book: Title, Author, Original Language, First Published, Approximate sales, Genre, Rating count, Average Rating, Text Reviews Count	9

Best books database is the result of the following process:


Extraction
Transformation
Loading
Web Scraping
Wikipedia: The List of best-selling books
Web Scraping
Enciclopedia Británica: 12 Novels Considered the “Greatest Book Ever Written”
CSV download
Kaggle: goodreadsbooks
Python: Jupyter Notebooks
Data cleaning
Data Merge
SQL: pgAdmin
Data Loading

Extract

Original data sources include Kaggle as the source of a CSV file, extraction is made by downloading. Wikipedia is the source for the List of best-selling books and Britannica’s website is the source of the 12 Novels Considered the “Greatest Book Ever Written” for both sources extraction is made by web scraping.

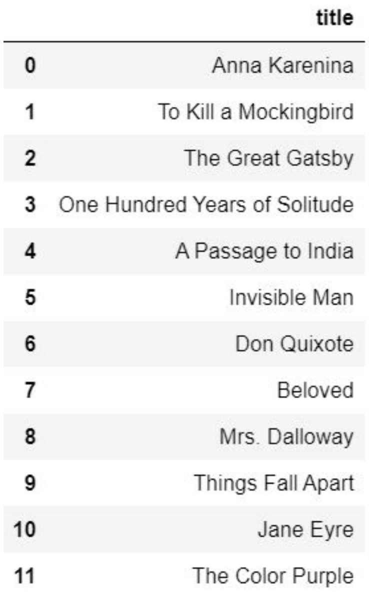
1.1 Reconcile records with the source data

Kaggle
A prepared CSV file was found on kaggle.com that contains the rating information about the books, primarily. This CSV was loaded into a Jupyter notebook for information to be explored and cleaned

usnghbsdfhkbksdbk pandas (pd.read_csv())ing pandas(pd.read_csv). The CSV file contained a list of columns with details of a vast list of book titles including columns such as: ISBN, author, number of pages, and ratings.

Britannica

The simplest data set of the three: a simple web scrape of the site to obtain a list of books that according to the articles are "12 Novels Considered the "Greatest Book Ever Written". However by using Web-Scraping, only the title of the book was obtained. Therefore- more information was needed to obtain additional information on these 12 novels.

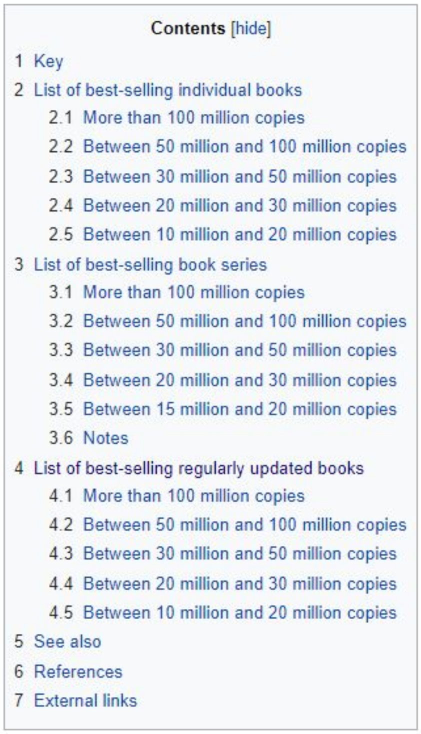


	title
0	Anna Karenina
1	To Kill a Mockingbird
2	The Great Gatsby
3	One Hundred Years of Solitude
4	A Passage to India
5	Invisible Man
6	Don Quixote
7	Beloved
8	Mrs. Dalloway
9	Things Fall Apart
10	Jane Eyre
11	The Color Purple

Wikipedia

To retrieve the list of best-selling books we used pandas.read_html() to retrieve a list of tables from a Wikipedia webpage. This webpage was divided into tables depending on the number of sales and had lists of data for best-selling book series and regularly updated books which included elements like dictionaries, world atlas and alike. We decided to focus on Individual books only.

There were a total of 17 tables found by running this command. For our project we focused on tables 1 to 5 which correspond to best selling individual books with 10 to more the 100 million copies.
We then made a loop in order to append each of the 5 tables to a unique table.

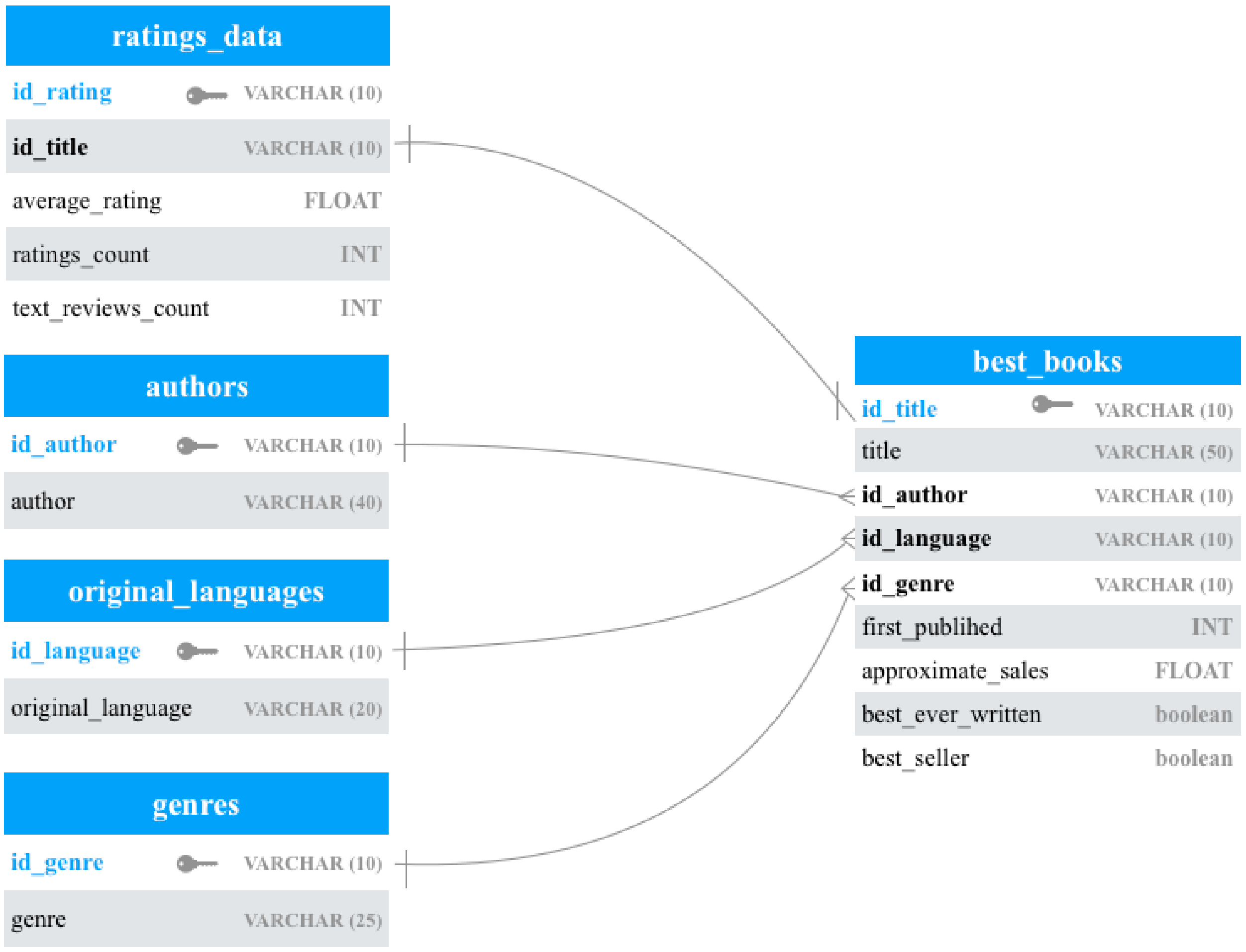


Contents [hide]

1 Key
2 List of best-selling individual books
  2.1 More than 100 million copies
  2.2 Between 50 million and 100 million copies
  2.3 Between 30 million and 50 million copies
  2.4 Between 20 million and 30 million copies
  2.5 Between 10 million and 20 million copies
3 List of best-selling book series
  3.1 More than 100 million copies
  3.2 Between 50 million and 100 million copies
  3.3 Between 30 million and 50 million copies
  3.4 Between 20 million and 30 million copies
  3.5 Between 15 million and 20 million copies
  3.6 Notes
4 List of best-selling regularly updated books
  4.1 More than 100 million copies
  4.2 Between 50 million and 100 million copies
  4.3 Between 30 million and 50 million copies
  4.4 Between 20 million and 30 million copies
  4.5 Between 10 million and 20 million copies
5 See also
6 References
7 External links

Data availability and interrelationships found on data sources are organized as shown in the next figure:

Information types and relationships



ratings_data	
id_rating	VARCHAR (10)
id_title	VARCHAR (10)
average_rating	FLOAT
ratings_count	INT
text_reviews_count	INT

authors	
id_author	VARCHAR (10)
author	VARCHAR (40)

original_languages	
id_language	VARCHAR (10)
original_language	VARCHAR (20)

genres	
id_genre	VARCHAR (10)
genre	VARCHAR (25)

best_books	
id_title	VARCHAR (10)
title	VARCHAR (50)
id_author	VARCHAR (10)
id_language	VARCHAR (10)
id_genre	VARCHAR (10)
first_publihed	INT
approximate_sales	FLOAT
best_ever_written	boolean
best_seller	boolean

Transform

For the data cleaning process, the three sources of information had cleaning performed on the different sources in order to prepare the data to be able to be merged properly.

2.1 Different spelling of the same concept

Data Cleaning for the CSV file from Kaggle

Data identification inside the CSV file, led to the fact that further data cleaning was needed. A simple exploration of the CSV file once loaded into Python, showed that if it was going to be merged with other book titles- the title of the book had to be compatible with the title in other databases. The main changes done to the CSV file was a simple clean up of the title and the author name.

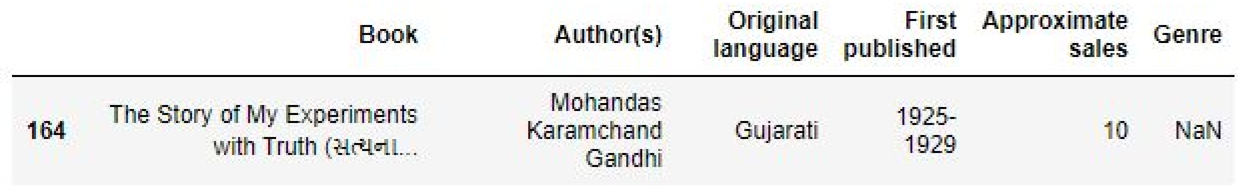
See example below-using the split function and identifying different delimiters, the imported data frame was able to be modified to a simpler more compatible Book Title and Author.

Original CSV Title/ Author	Clean Title / Author
Harry Potter and the Half-Blood Prince (Harry Potter #6)	Harry Potter and the Half-Blood Prince
J.K. Rowling/Mary GrandPrÃ©	J.K. Rowling

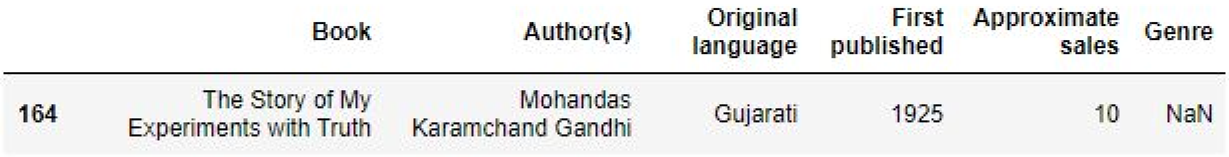
2.5 Standardization by filtering, using rules and lookup tables

Data Cleaning for Wikipedia: List of Best Selling Books

For the "First published" field, we had some data inputs where a range of years was given instead of the first published year. For this the commands **df.str.split('-').str[0]** along with **df.['Approximate sales'].unique().** were used.



	Book	Author(s)	Original language	First published	Approximate sales	Genre
164	The Story of My Experiments with Truth (સત્યના...	Mohandas Karamchand Gandhi	Gujarati	1925-1929	10	NaN



	Book	Author(s)	Original language	First published	Approximate sales	Genre
164	The Story of My Experiments with Truth	Mohandas Karamchand Gandhi	Gujarati	1925	10	NaN

Then, several book titles were written in two languages, its original language and English. With **df.str.split('(').str[0]** we get rid of the original language book title since all were written between parenthesis.

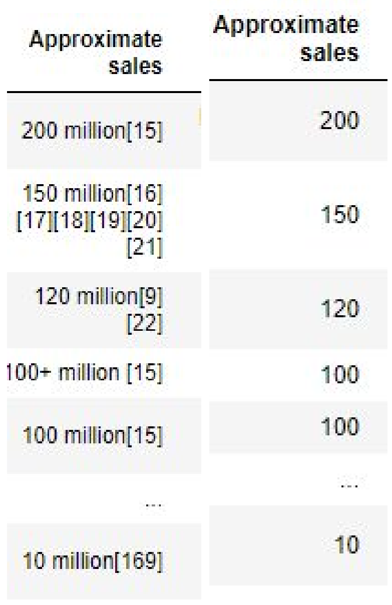
We then saved the Data frame into a CSV file with **df.to_csv** function in order to view all of the fields and inputs to identify any other input discrepancy. We then fixed these particular discrepancies using **df.at[df.index[df['col'] =='error'],['col'] = 'new value'.**

Finally we changed the column names in order to match with our SQL column names to avoid having problems on the loading part.

2.6 Set characters and units of measurements

Data Cleaning for Wikipedia: List of Best Selling Books

Once we had a single table, a few cleaning needed to be done. On the approximate sales, we had data inputs with several formats, some written in word form and others in number form with special characters like > or + to indicate more than x number of sales. We used **df.str.split(‘’).str[0]** along with **df.[‘Approximate sales’].unique()** to fix each of these cases.



Approximate sales	Approximate sales
200 million[15]	200
150 million[16][17][18][19][20][21]	150
120 million[9][22]	120
100+ million [15]	100
100 million[15]	100
...	...
10 million[169]	10

2.8 Cleaning

Data Cleaning for the CSV file from Kaggle

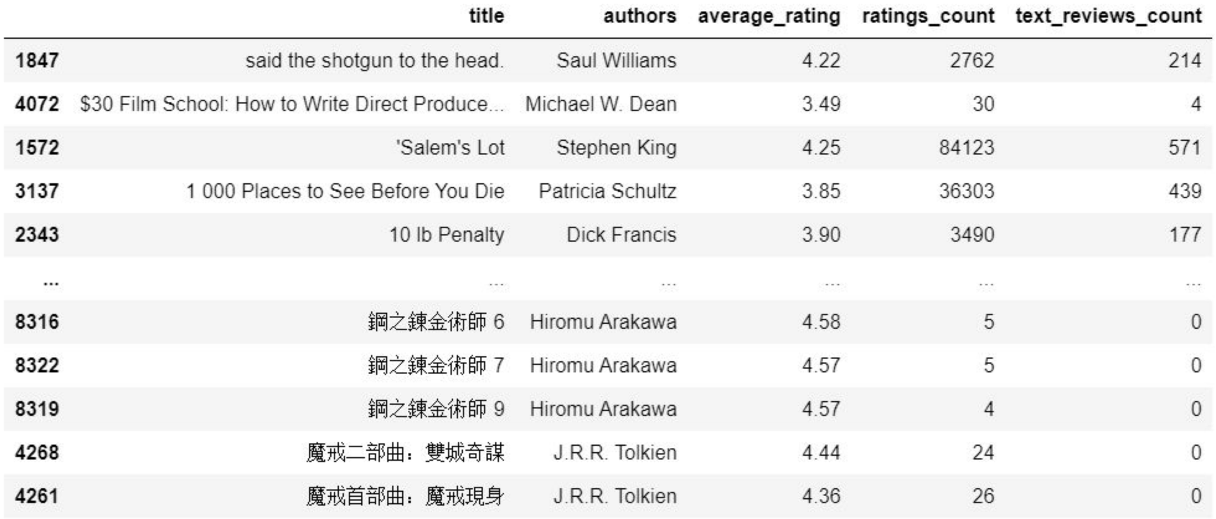
Furtherly, columns were eliminated from the data frame obtained from Kagle, and a table was created with columns with information that would add value in creating the database. The columns that were kept were the following:

'Title': The name under which the book was published.
'Author': Names of the authors of the book
'average_rating': The average rating of the book received in total.
'Ratings_count': Total number of ratings the book received.
'Text_reviews_count': Total number of written text reviews the book received.

Data Cleaning for the CSV file from Kaggle

Lastly, the original CSV file listed different publications of each book. As the greatest novels are popular books - each novel in the database has more than one publication, each with different ISBN. The information on each novel that was kept was the published version of the novel that had the largest ratings count.

The result of clean data with only one entry per book is the following:



	title	authors	average_rating	ratings_count	text_reviews_count
1847	said the shotgun to the head.	Saul Williams	4.22	2762	214
4072	$30 Film School: How to Write Direct Produce...	Michael W. Dean	3.49	30	4
1572	'Salem's Lot	Stephen King	4.25	84123	571
3137	1 000 Places to See Before You Die	Patricia Schultz	3.85	36303	439
2343	10 lb Penalty	Dick Francis	3.90	3490	177
...	...	...	...	...	...
8316	鋼之鍊金術師 6	Hiromu Arakawa	4.58	5	0
8322	鋼之鍊金術師 7	Hiromu Arakawa	4.57	5	0
8319	鋼之鍊金術師 9	Hiromu Arakawa	4.57	4	0
4268	魔戒二部曲：雙城奇謀	J.R.R. Tolkien	4.44	24	0
4261	魔戒首部曲：魔戒現身	J.R.R. Tolkien	4.36	26	0

Transform: Prepare dataframes to load into the Database

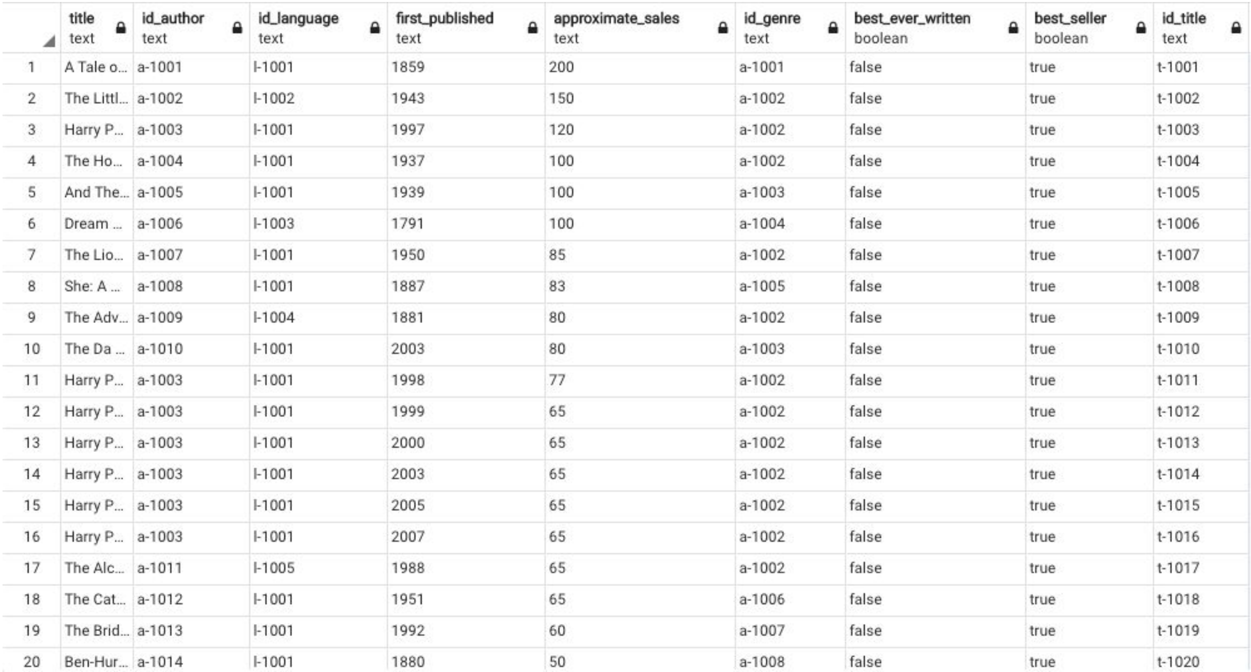
As different data sets were used to create a final dataset, the names of the columns were equalized. Other activities that were done to transform the data is the following:

- Add a column to identify those books that are best ever written according to Britannica Encyclopedia and Best Selling books according to Wikipedia
- Join with ratings Data Frame to fill the missing authors
- Eliminate blank spaces
- Eliminate NULLs from the data once joined
- Create catalogs and ids for primary keys before doing the final loading process

Load

PgAdmin was selected as the storage method for the “Best books database” since it is an open source option and it is also a relational database.

The main table was loaded and we checked for missing data and mistakes on the 174 rows successfully loaded



	title text	id_author text	id_language text	first_published text	approximate_sales text	id_genre text	best_ever_written boolean	best_seller boolean	id_title text
1	A Tale o...	a-1001	l-1001	1859	200	a-1001	false	true	t-1001
2	The Littl...	a-1002	l-1002	1943	150	a-1002	false	true	t-1002
3	Harry P...	a-1003	l-1001	1997	120	a-1002	false	true	t-1003
4	The Ho...	a-1004	l-1001	1937	100	a-1002	false	true	t-1004
5	And The...	a-1005	l-1001	1939	100	a-1003	false	true	t-1005
6	Dream ...	a-1006	l-1003	1791	100	a-1004	false	true	t-1006
7	The Lio...	a-1007	l-1001	1950	85	a-1002	false	true	t-1007
8	She: A ...	a-1008	l-1001	1887	83	a-1005	false	true	t-1008
9	The Adv...	a-1009	l-1004	1881	80	a-1002	false	true	t-1009
10	The Da ...	a-1010	l-1001	2003	80	a-1003	false	true	t-1010
11	Harry P...	a-1003	l-1001	1998	77	a-1002	false	true	t-1011
12	Harry P...	a-1003	l-1001	1999	65	a-1002	false	true	t-1012
13	Harry P...	a-1003	l-1001	2000	65	a-1002	false	true	t-1013
14	Harry P...	a-1003	l-1001	2003	65	a-1002	false	true	t-1014
15	Harry P...	a-1003	l-1001	2005	65	a-1002	false	true	t-1015
16	Harry P...	a-1003	l-1001	2007	65	a-1002	false	true	t-1016
17	The Alc...	a-1011	l-1005	1988	65	a-1002	false	true	t-1017
18	The Cat...	a-1012	l-1001	1951	65	a-1006	false	true	t-1018
19	The Brid...	a-1013	l-1001	1992	60	a-1007	false	true	t-1019
20	Ben-Hur...	a-1014	l-1001	1880	50	a-1008	false	true	t-1020

Catalogs included in this database can be used to filter and make further analysis like: gender with highest sales and language with more books written since catalogs include authors, genres, original languages and ratings data.

Findings

A database construction like “Best book database” required multiple transformations to create a final database that was considered complete and clean enough for the final loading.

As a relational database, Best books database allows gathering information from 5 tables of variables like rating, genre, and approximate sales. However, when the source comes from different users of origins, objects are different in the slightest degree and data cleaning is required to create an output that can create a reliable database.

It was found that the type of information merging had differences depending on the source. The database creating process was iterative, imported data was checked thoroughly several times and cleaned to make sure merges were inclusive of all data and all data was merged properly and that the quality of data generated would create an accurate output database. Further data cleaning would benefit the database even more, if titles were standardized furtherly, or more authors were included in the final data set. For the project purpose, Best book database provides a simple set of information that can be enriched with more cleaning or complementary data, if needed.