S = ‘ID,Title,Author,Year,Publisher,ISBN,Format,Language,URL’

OpenLibrary Schema	Google Books Schema
1. ID - Unique identifier with 'ol_' prefix 2. Title - Book title 3. Author - Primary author name 4. Year - First publication year 5. Publisher - First publisher listed 6. ISBN - First ISBN (either ISBN-10 or ISBN-13) 7. Format - Always set to 'Paperback' 8. Language - Language code (e.g., 'eng') 9. URL - Link to OpenLibrary book page	1. ID - Unique identifier with 'gb_' prefix 2. Title - Book title 3. Author - Author names (joined with commas if multiple) 4. Year - Publication year (first 4 digits of publishedDate) 5. Publisher - Publisher name 6. ISBN - Preferred ISBN (ISBN-13 if available, else ISBN-10) 7. Format - Always set to 'Paperback' 8. Language - Language code 9. URL - Google Books info link

I made a matching algorithm that looks specifically for:

- ISBN matching
- Title similarity
- Author name normalization
- Publication year proximity

I don’t believe there are any missing values in table_a. I believe, since I am looking for books on popular and reliable websites, there is no missing information. I’ve looked and could not find any null or missing values from table_a.

Classification:

ID Type: Textual

- Average length: 14.3 characters
- Min length: 11 characters
- Max length: 17 characters

Title Type: Textual

- Average length: 21.1 characters
- Min length: 2 characters
- Max length: 58 characters

Author Type: Textual

- Average length: 16.4 characters
- Min length: 6 characters
- Max length: 48 characters

Year Type: Numeric (4 digits)

- Average, Max, and Min are: 4 characters

Publisher Type: Textual

- Average length: 24.7 characters
- Min length: 3
- Max length: 52

ISBN Type: Textual

- Average length: 12.8 characters
- Min length: 10 characters (ISBN-10 format)
- Max length: 13 characters (ISBN-13 format)

Format Type: Categorical

- Average length: 9 characters (it seems that it only has “paperback”)
- Min length: 9
- Max length: 9

Language Type: Categorical

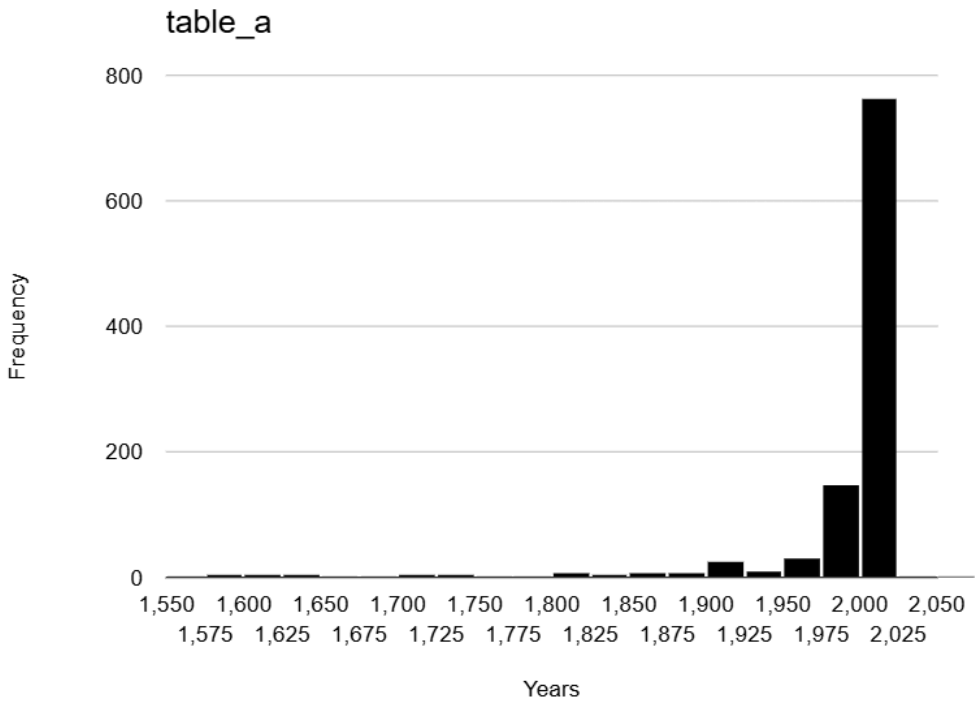
- Average length: 3 characters
- Min length: 3 (all language codes)
- Max length: 3 (all language codes)

URL Type: Textual

- Average length: 47.2 characters
- Min length: 39 characters
- Max length: 55 characters

Outliers and Anomalies:

For the year distributions, I noticed most were between 1900s and 2000s. Majority of the outliers were from the 1750s and earlier.



I’m unsure how to make a histogram based on textual frequency data…

ENG	314	31.4 %
SPA	159	15.9 %
FRE	81	8.1 %
GER	80	8 %
CHI	64	6.4 %
POL	51	5.1 %
DUT	50	5 %
ITA	27	2.7 %
TUR	25	2.5 %
RUS	24	2.4 %
SWE	20	2 %
POR	11	1.1 %
BUL	9	0.9 %
HEB	9	0.9 %
JPN	9	0.9 %
KOR	9	0.9 %
CAT	8	0.8 %
DAN	7	0.7 %
UND	4	0.4 %
CZE	3	0.3 %
HRV	3	0.3 %
HUN	3	0.3 %
PAN	3	0.3 %
PER	3	0.3 %
URD	3	0.3 %
ALB	2	0.2 %
GRE	2	0.2 %
LIT	2	0.2 %
RUM	2	0.2 %
SRP	2	0.2 %
BAQ	1	0.1 %
EST	1	0.1 %
FAO	1	0.1 %
FIN	1	0.1 %
GLE	1	0.1 %
IND	1	0.1 %
MAL	1	0.1 %
MUL	1	0.1 %
TLH	1	0.1 %
UKR	1	0.1 %
VIE	1	0.1 %

From looking at this frequency chart of languages used for each book, the outliers seem to be in categories of 1-digit frequencies. The most being used, contains double to triple digits.

Formats

ID Format:

- Expected format: "ol_" followed by alphanumeric characters
- All values follow this format
- Standardization is not needed

Title Format:

- Free text
- No need for any specific format requirements

Author Format:

- Inconsistent formats found:
    - "Lastname, Firstname"
    - "Firstname Lastname"
- Multiple authors separated by commas
- Some include titles (Dr., Mr., etc.)
- Some include foreign characters
- Will need standardization if current algorithm isn't precise enough (however, I think it is)

Year Format:

- Expected format: 4-digit year
- All years are 4 digits

Publisher Format:

- Free-form text
- Inconsistencies:
    - Some include location
    - Some include "Ltd", "Inc", etc.
    - Some include additional, unneeded information that the other table might not include
    - Some are missing
- Would benefit from standardization if current algorithm isn't precise enough with the others

ISBN Format:

- Should be ISBN-10 or ISBN-13
- Potential problems:
    - Mix of ISBN-10 and ISBN-13 between each table for the same book?

Format:

- All values are "Paperback"
- Consistent format

Language:

- Uses 3-letter language codes
- Consistent format
- All lowercase

URL Format:

- All start with "https://openlibrary.org/books/"
- Consistent format

There are no synonyms for ANY of the headers. Unexpectedly, for every book taken from the Open Library- with the specific filters- they are returned back as “Paperback”.

Sometimes, attribute values are "sprinkled" all over the item. For example, a book may have an attribute "publisher", but its value is missing. Instead, the book title contains the publisher (e.g., "Principles of Data Integration by Springer"). Do you have this problem with this attribute?

1) No instances found where publisher information is "hidden" in the title field
2) Publisher information is either:
    a) Present in the Publisher field
    b) Completely missing
    c) Or properly formatted in the Publisher field

Do you see any other data quality problems with this attribute?

- No, everything else seems fine.

List any software tools you have used to understand and clean the above data. For example, if you have used a particular Python package, list the name of the package.

‘Import csv’ for getting the data for the histograms
‘from difflib import SequenceMatcher’ and ‘import re’
I also used a histogram and frequency analysis websites to get data.