Analysis of Academic Sciences Library Collection: Coding Temple Capstone Project

Kelli Michaels Nov 2024

Introduction: In this report, the goal is to analyze my own data set and draw useful conclusions using the tools I have learned throughout my Data Analytics course. The data set I have chosen is a selection of books housed in an academic sciences library at Indiana University.

Data and mission: The data set is composed of a large table with each row representing an item in the collection. I will investigate the statistics surrounding the books in the collection. By analyzing these statistics we can hopefully make some insightful observations about the library collection and use-history. For each item (book), we have access to a long list of attributes (columns). Some important selections include “Author”, “Home Location”, “Publication Year”, and “Place of Publication. Here is the full list of column titles (The information we are able to access about each book)

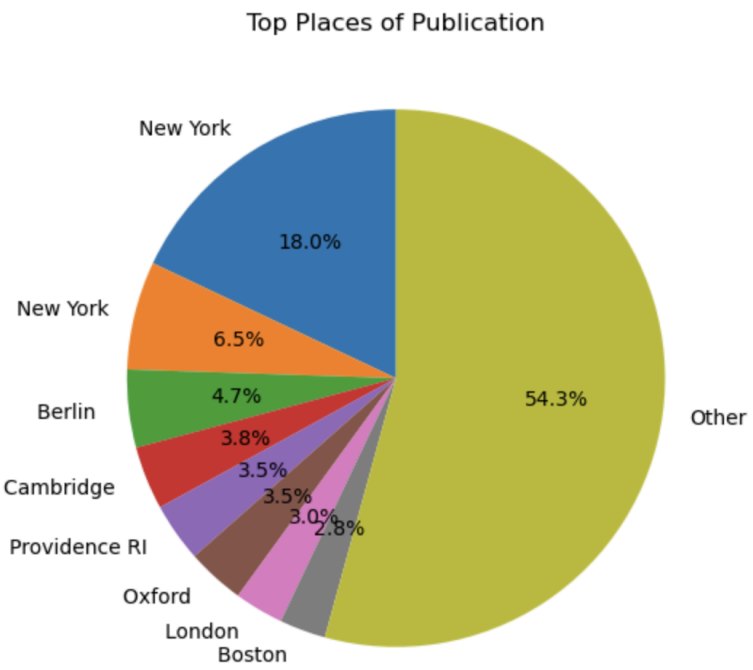
- 'Title Control Number', 'Catalog Key', 'Call Sequence', 'Copy Number', 'Format', 'Pub Year', 'BLvl', 'Type', 'Bib Form', 'Type/Form', 'Language', 'MARC key', 'Author', 'Title', 'Library', 'Call Number', 'Shelving Key', 'Class code (LC, SUDOC, NLM)', 'Call Number Range Key', 'Item Created Date', 'Item ID', 'Item Type', 'Home Location', 'Current Location', 'Item Category 1', 'Item Category 2', 'Cataloging department code', 'Cataloging Staff Code', 'Cataloging date code', 'Place of Publication (260a)', 'Publisher (260b)', 'Date of Publication (260c)', 'ISSN', 'ISBN', 'GMD (245h)', 'OCLC', 'Title Created Date', 'Title Cataloged Date', 'Created By', 'Pagination (300a)', 'Illustrations (300b)', 'Size (300c)', 'Total Charges', 'In-House Charges', 'Date Last Charged', 'Last Activity Date'
- We won’t use all of these but it is useful to see what kind of information we can access.
- NOTE: The content of these cells is not perfectly standardized throughout the spreadsheet. Some data is missing, filled with an invalid value, or has whitespace/nonsense surrounding the actual data. In my code I go through several cleaning steps to address these situations. I’m not going to address those details in this report but feel free to check it out in my Jupyter notebook. Just know that this cleaning was also a necessary step in this analysis.

Questions and hypotheses: Here are some discussion questions and associated hypotheses. I formed my hypotheses by taking a quick scroll through the data and making my best guess, which was often difficult and (as it turns out) not very accurate.

- How many books are in the collection?
    - I can scroll to the bottom and see the spreadsheet has 28377 books. Note this does not indicate whether they are all unique or not.
- Where are our books being published? Show a pie chart of the most common places of publication.
    - Scrolling through, I am seeing a lot of New York, I'm going to guess that is the most common. Hard to speculate about others though.
- What authors appear most often in the collection? List the top 5
    - This I really can't assess by eye. I cannot even find any two entries with the same author, the spreadsheet is too big.
- What percentage of books in the collection are illustrated?
    - From my scroll through the column I'd estimate about 85% of books have some illustrations
- What percentage of books are currently in their home location?
    - Also hard to visually assess. Maybe 90-95% seem to have matching current and home locations.
- Let's compare the publication years vs the year the book was added to the catalog. Make a histogram of the time differences (in years) between the year of publication and the year the book was added to the collections.
    - Really can't tell just by looking at the data. I am kind of expecting most books to be added to the catalog fairly close to the date of publication, let's find out if that is the case!
- Investigate the distribution of language for our collection as a function of time. Compare the language breakdown of the collection in the year 2024 and in the year 2004
    - Scrolling through I am seeing mostly English with also some French and German. Definitely need some code to compare the two time periods

Analysis: Below I list the outcomes of our analysis questions, compare to my hypotheses, and note anything else I discovered along the way. Some questions had surprising answers and some of my questions needed refining. I'm copying the questions from above for easy reference.

- How many unique books are in the collection?
    - Result: 25,790 books
    - Comment: There were 28,377 entries in the spreadsheet, so we can conclude that most of the entries were unique but that there were around 3000 repeats.
- Where are our books being published? Show a pie chart of the most common places of publication.



- Result:
- Comment: I filtered out all the data with no place of publication information, so all those that said ‘other’ do have a location outside of those listed in the chart.
- Comment: I have a few comments about this one. First of all, I was right that New York was the most common place of publication. But I'm not sure that a pie chart was the best approach to this question. There are so many unique places. Additionally, since the city is listed, I think it would be useful to sort these into countries of publication instead. That would be hard to implement given that the data set lists only the city, not a whole address.
- Comment: If we were really interested in learning more about the origins of the collection I would see if another data set is available with full addresses.

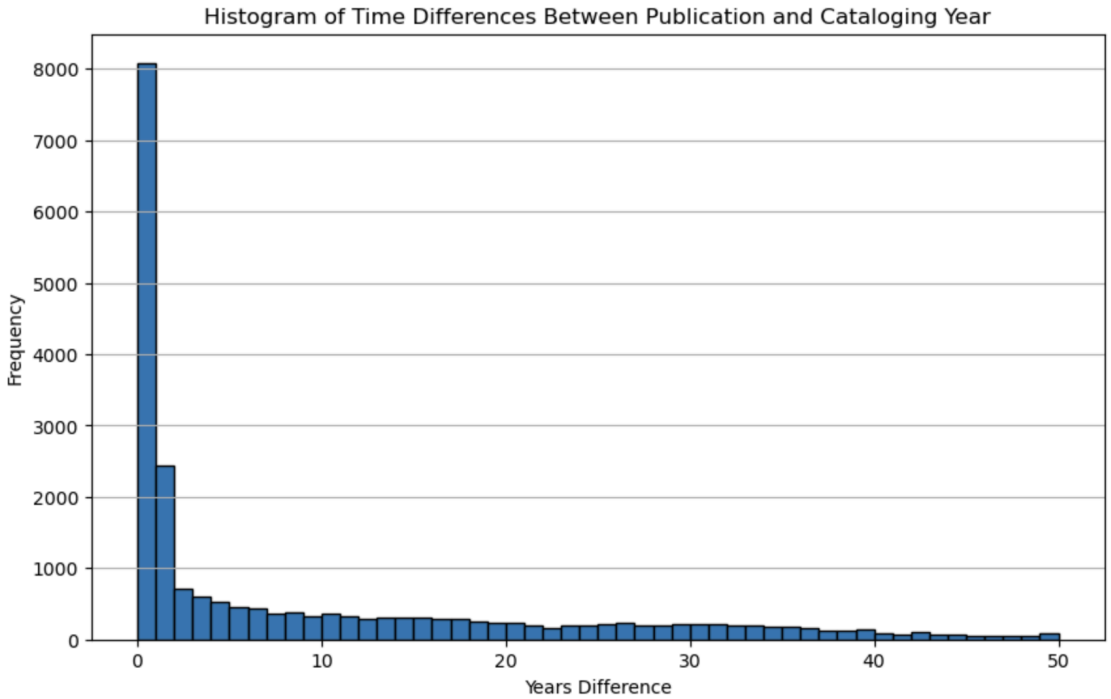
- What authors appear most often in the collection? List the top 5
  - Result:

Author	Book Count
Bourbaki Nicolas	63
Lang Serge	46
Knuth Donald Ervin	37
Feynman Richard P	32
Krantz Steven G	28

  -

- Comment: We successfully found the top 5 authors. I think it is interesting that no author has more than 63 books in the collection of over 25,000. Another interesting query would be how many unique authors there are.

- What percentage of books in the collection are illustrated?
  - Result: 79.45%
  - Comment: I really over-estimated how many books are illustrated in my hypothesis. It seems that a disproportionate number of books towards the top of the spreadsheet were illustrated. This shows the importance of doing the math rather than estimating by eye.
- What percentage of books are currently in their home location?
  - Result: 90.87%
  - Comment: This was very hard to estimate, but we were surprisingly close. We confirmed that a large majority of the collection is currently housed in its home location.
- Let's compare the publication years vs the year the book was added to the catalog. Make a histogram of the time differences (in years) between the year of publication and the year the book was added to the collections.
  - Result:



  - Comment: We see the histogram is heavily skewed to the left, indicating that most books were added to the collection within a few years of their publication. I did the calculation and specifically 50.59% of books in the collection were added within two years of their publication date.
  - Worth noting that some books were added to the catalog up to 50 years after their publication year.

- Investigate the distribution of language for our collection as a function of time. Compare the language breakdown of the collection in the year 2024 and in the year 2004
    - Result:

Language (2024)	Percentage (2024)
English	98.49%
French	1.04%
German	0.38%

    - Result:

Language (2004)	Percentage (2004)
English	97.42%
French	1.74%
German	0.70%

    - In both time ranges the collection was mostly (over 97%) English. However, in the collection from 2004, there were slightly more texts from French and German relative to the 2024 collection.

Conclusion and recommended future studies: Our analysis provided valuable insight into the composition of the collection at the Sciences Library. We learned the collection has over 28,000 books; we learned that a majority of them are in English as well as illustrated. We saw that most books are added to the collection within a few years of their publication date, but there are a notable number of exceptions where the gap was as long as fifty years. We also had a few ideas for future studies, like gathering more information on place of publication. We could also investigate the number of unique authors, since our analysis here didn't directly answer that. There is still much left to explore!