Vancouver Public Library

- **World Languages**

Web scrapping Project
Rochita Sundar
Feb’22

- Introduction
  - Public Library System for the city of Vancouver, British Columbia
  - Third largest public library system in all of Canada
  - Over 20 branches, 1 central library location & online
  - Over 9.5 million materials: books, e-books, CDs, DVDs, magazines etc.
  - Serving >400,000 active members

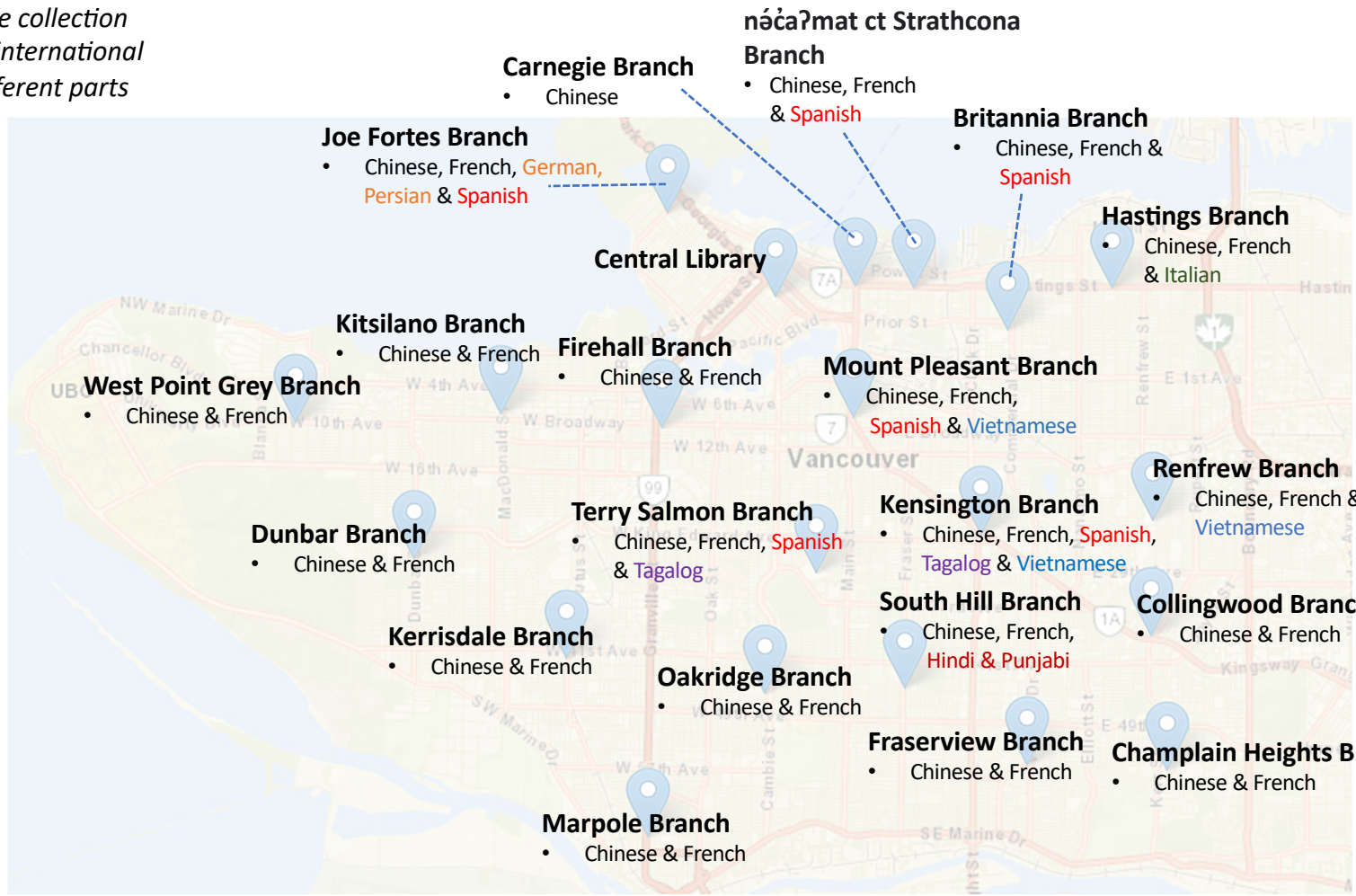
- Objective
  - Scrape the website of VPL using automation test software
    - Study international language collection carried by each VPL location
    - Gather information on language collections, titles, authors, categories, availability statuses & user ratings to draw insights

- Motivation
  - General curiosity in reading, languages & interest in exploring resources in my locality

Reference: https://en.wikipedia.org/wiki/Vancouver_Public_Library

Library location vs. language collection analysis gives an insight to international demographic residing in different parts of Vancouver

- All locations carry material in Chinese
- All but one, carry material in French
- 7 locations carry Spanish, 4 locations carry Vietnamese & 3 locations carry Tagalog materials
- Central library carries all international language materials



Reference: https://www.vpl.ca/hours-locations

World Languages

We collect reading material for adults, teens, and children in sixteen languages other than English.

See world language collections

العربية

Arabic Language Collection العربية

This collection has material in Arabic for all ages.

Explore this collection

language

中文

Chinese Language Collection 中文

This collection has material in Chinese for all ages.

Explore this collection

française

French Language Collection (Children) française

This collection has material in French for children.

Explore this collection

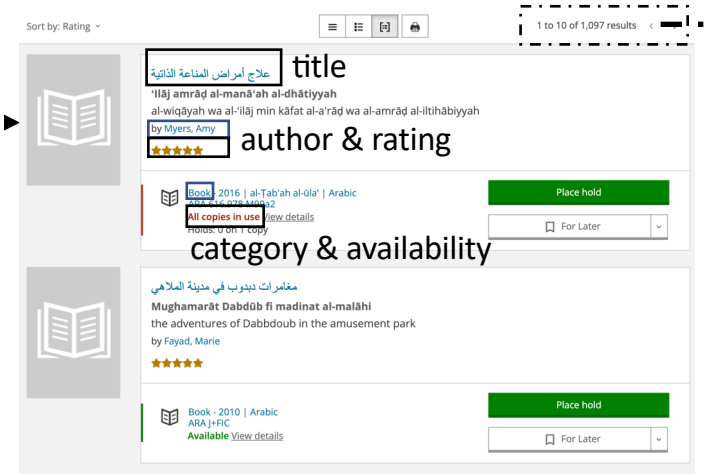
16 international languages

Tiếng Việt

Vietnamese Language Collection Tiếng Việt

This collection has material in Vietnamese for all ages.

Explore this collection



title

author & rating

category & availability

multiple pages

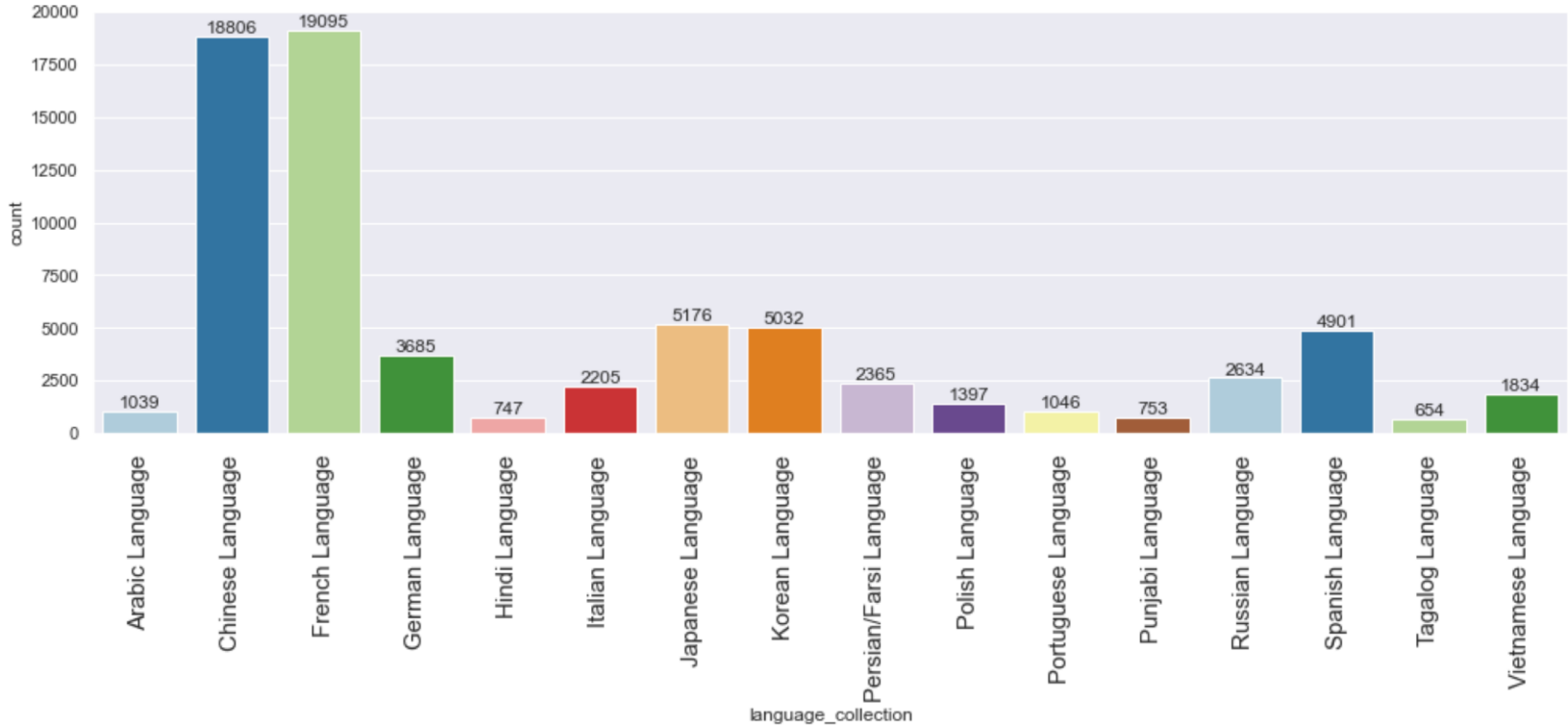
multiple records Per page

VPL Website & data collection:

Crawler scrapped more than 70K+ records (information on language, title, author, rating, category & availability)

Reference: https://www.vpl.ca/borrowing/world-languages

Language Collection : Count of materials across VPL locations

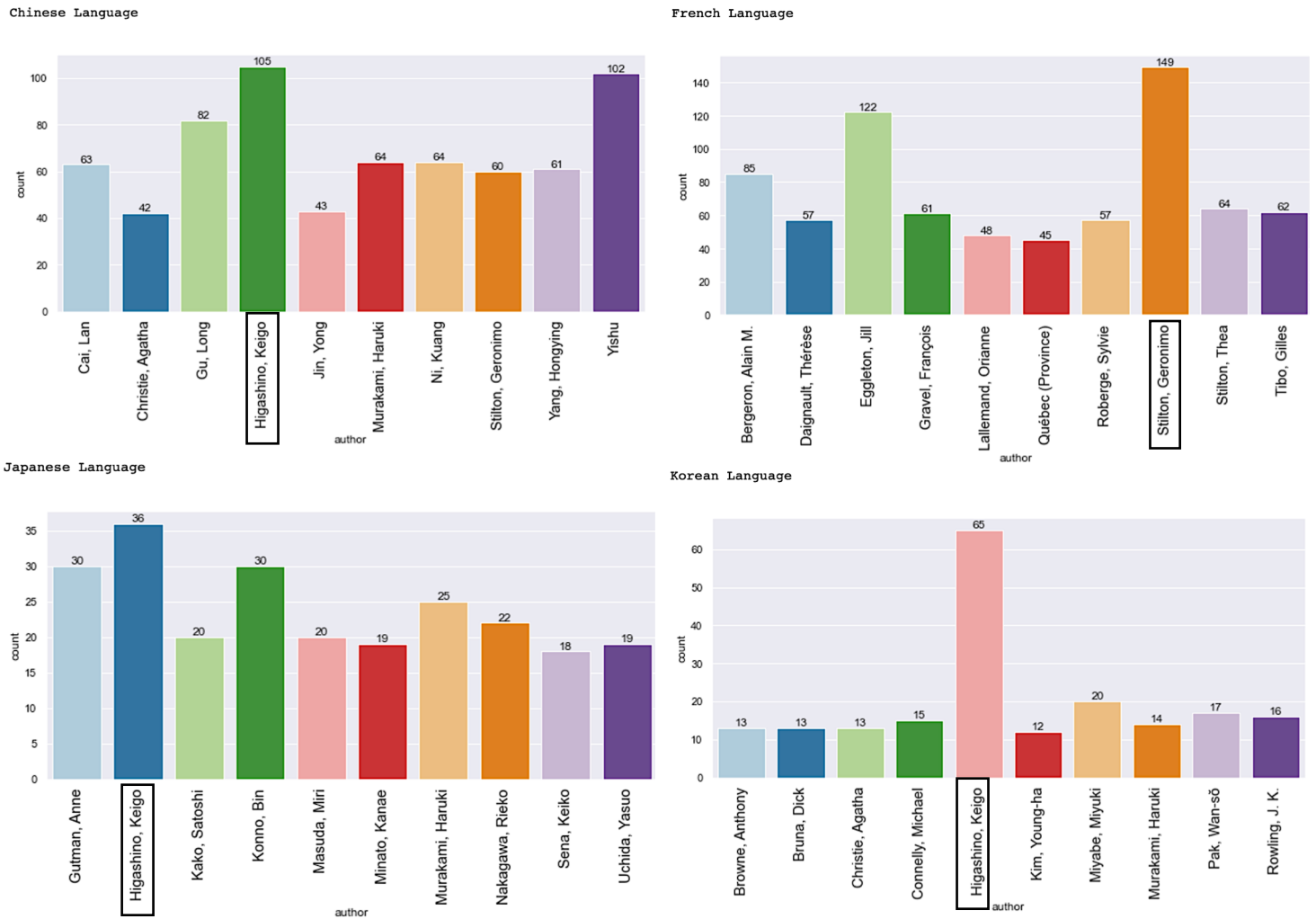


- Maximum material is available in French & Chinese languages followed by Japanese & Korean languages
- Tagalog, Hindi & Punjabi languages have some of the lowest number of material available

Let's explore top authors…..

Top 10 authors by number of published materials at VPL for international languages

- Interestingly, 'Higashino , Keigo' makes an appearance (# 1) under all Chinese, Japanese & Korean languages for max. materials at VPL
- Stilton, Geronimo' & 'Eggleton , Jill' are top French artists with 140+ & 120+ materials available at VPL locations respectively



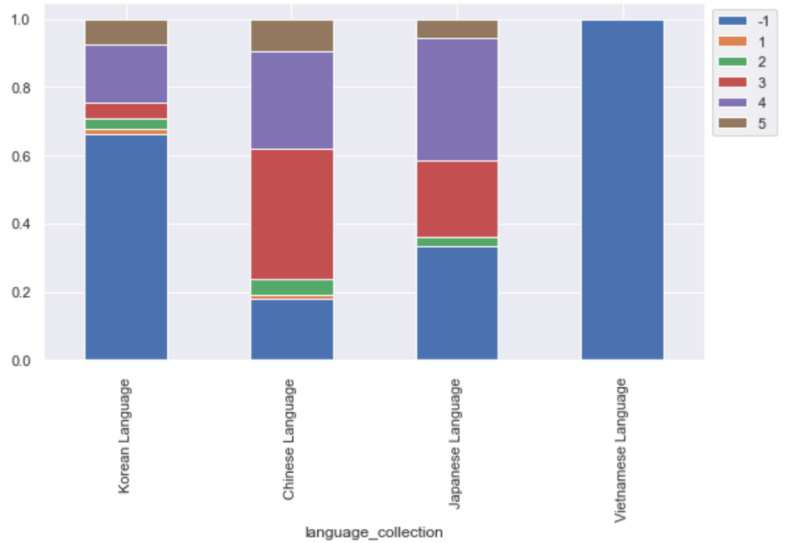
Let's explore Kiego's work's popularity …..

Popularity (rating) of Higashino, Keigo's materials by languages

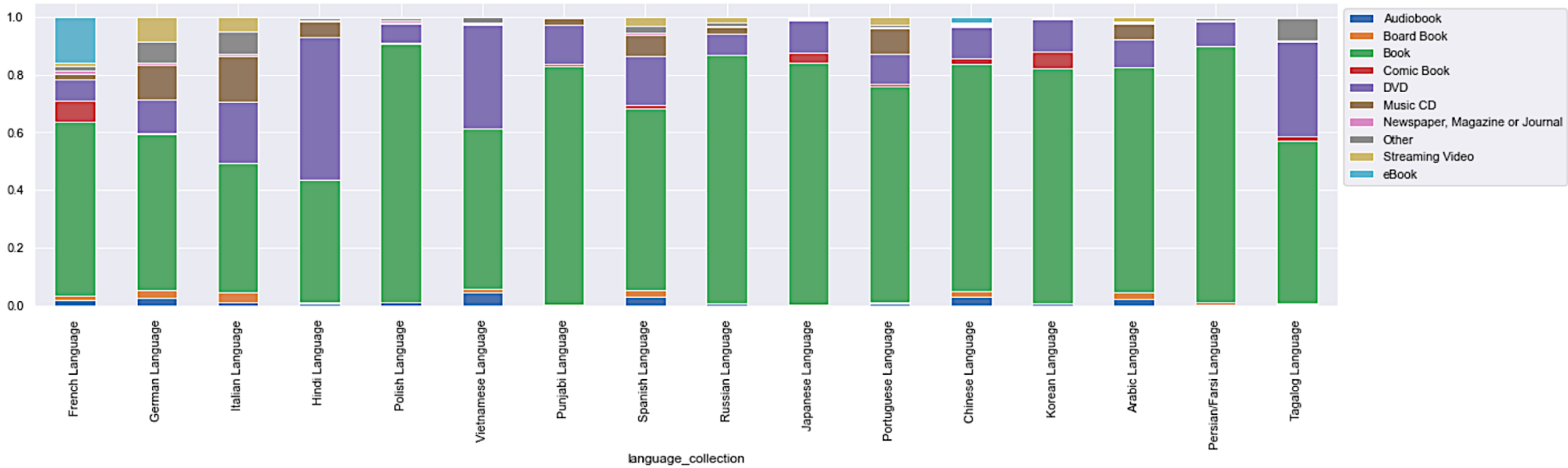
- "-1" rating corresponds to "no rating information available". No rating information is available for any of his 4 materials in Vietnamese language. ~60% of his work in Korean language, ~20% of his work in Chinese language & ~30% of his work in Japanese language is not rated

- ~40% of his work in Chinese language has a rating of 3 stars, ~25% has a rating of 4 stars and ~5% a rating of 5 stars

rating	-1	1	2	3	4	5	All
language_collection							
All	78	2	8	51	54	17	210
Chinese Language	19	1	5	40	30	10	105
Korean Language	43	1	2	3	11	5	65
Japanese Language	12	0	1	8	13	2	36
Vietnamese Language	4	0	0	0	0	0	4

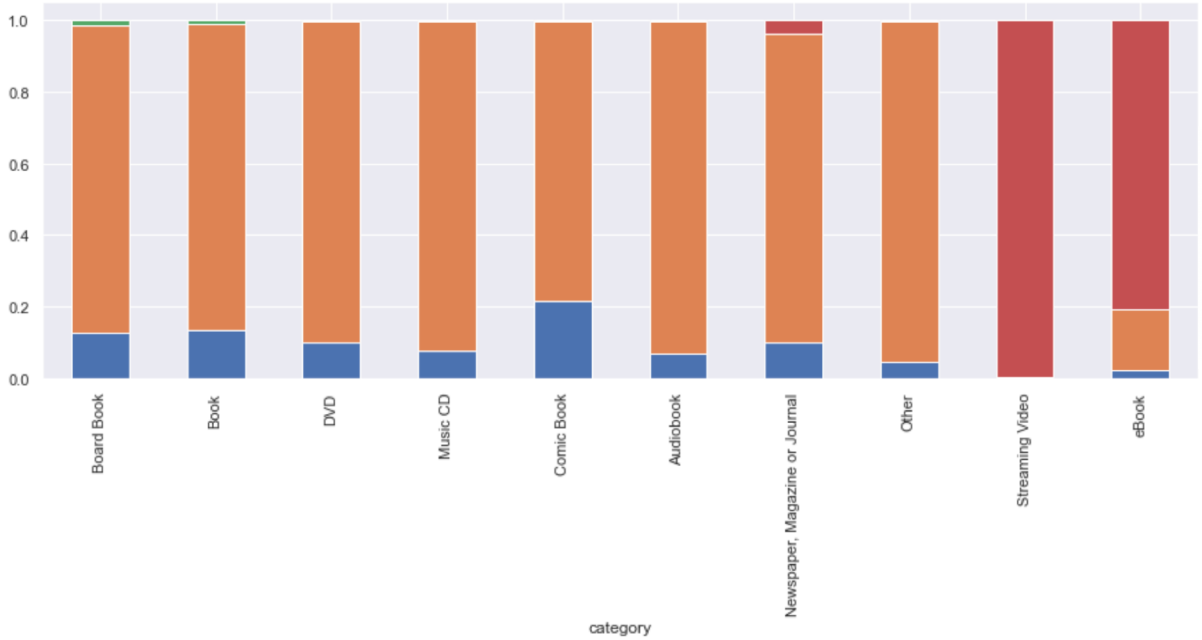


Language Collection : For each language, what is the percentage of material in different categories (Audiobook, DVD, Books etc.…)?



- ~60% of material in Hindi language is DVDs. Italian, Vietnamese, & Tagalog languages have ~20%, ~40% and ~30% of their material as DVDs
- Italian and German languages have ~10% of the material as Music CDs and ~5% of the material as streaming video
- French has ~20% of the material available as eBooks, more so than any other languages
- Korean (~5% of material), French & Japanese languages have more comic books than other languages

Analyzing the availability status & category of all materials



status	All copies in use	Available	On order	not available	All
category					
All	8801	58218	507	3843	71369
Book	6894	42950	434	32	50310
DVD	839	7463	42	1	8345
Board Book	135	922	18	0	1075
Comic Book	515	1857	7	4	2383
Music CD	153	1772	6	1	1932
Audiobook	91	1249	0	2	1342
Newspaper, Magazine or Journal	41	350	0	15	406
Other	52	1074	0	5	1131
Streaming Video	1	1	0	1018	1020
eBook	80	580	0	2765	3425

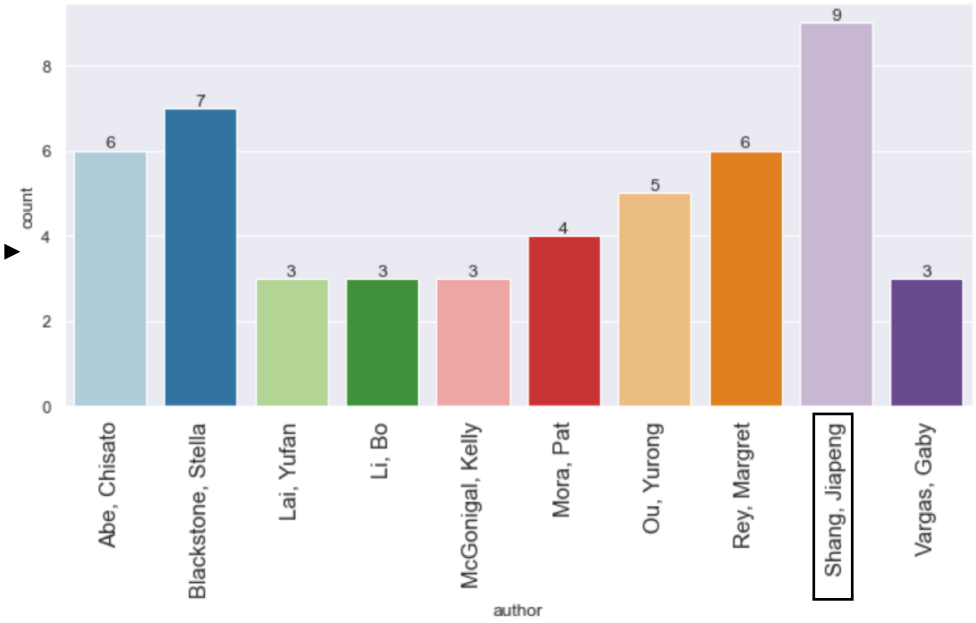
- No status information is available for a major percentage of Streaming video & eBooks and small percentage of newspaper, journals
- Out of 8801 materials with all copies in use, 6894 are books (78%), 839 are DVDs (9.5%), 515 are comic books (5%), 135 are board books (1.5%), 153 are music CDs, 91 audiobooks, & 80 eBooks...

Let's visualize top 10 authors who have an average rating of 4 or higher across all their published materials & the most number of published materials at VPL

List of author's with avg. rating 4 or higher across all their material at VPL

	author	rating
0	Li, Bubai	5.0
1	Martin, Catherine	5.0
2	Binghe	5.0
3	Binghan	5.0
4	Guillebeau, Chris	5.0
...	...	...
1705	Zhang, Mali	4.0
1706	Kikuchi, Arata	4.0
1707	Shore, Howard	4.0
1708	Zhang, Lihui	4.0
1709	Dieter, George	4.0

Top 10 author's in terms of count of material at VPL



& countless other visualizations possible…

Summary

- The languages carried by a VPL branch provides an insight into the international demographic of that particular region!
  - All locations carried Chinese language material, & all but 1 carried French language. This was followed by 7 locations carrying Spanish & 4 locations carrying Vietnamese language material

- Across all VPL locations & online, maximum material is available in French & Chinese languages (19K+), followed by Japanese & Korean languages (5K+), and then Spanish language. Tagalog, Hindi & Punjabi languages have some of the lowest number of material available

- Analysing top author's by most material at VPL, we discovered an author Higashino, Keigo leading, across several Asian languages. Further analysis was performed to see how his work is rated by users/readers across languages

- Language vs category analysis revealed the following:
  - ~60% of material in Hindi language is DVDs
  - ~10% of material in Italian & German languages is Music CD, & ~5% is DVDs
  - ~20% of material in French language is eBooks, more so than any other languages
  - Korean, French & Japanese languages have a higher percentage of material as comic books than other languages

- Similarly, availability status vs category information revealed that a major percentage of streaming video & eBooks, and a minor percentage of newspapers & journals do not have any availability information available. In terms of material with status "All copies in use", majority are books (78%), followed by DVDs(9.5%) and then comic books (5%)

- Author's were further categorized by popularity (i.e., user rating) & number of published material at VPL

Challenges & Next Steps

- Because of the nature of the website & type of information, all of the data collected was categorical & not numerical in nature. As a result not much scope for feature engineering

- The data collected has limited analysis & inference scope as it pertains to only those records available on VPL website. More data needs to be collected on popular international authors & merged with the data collected at VPL to see whether VPL carries popular material or not

- User information is not private, so unable to determine what the user preference is. Having this information can enable interesting predictions on what material is more popular & in demand for library to plan it's stocking & sourcing accordingly