Project 1 - Data Exploration Concept & MySQL Query: Write-up

Yanchen Lu

@January 24, 2023

Concept Description

This project is an exploration of books and media regarding LGBTQ topics in the Seattle Public Library, how they are represented and categorized by the library, as well as how the public's interest in them have changed throughout the years.

As a bisexual person growing up, I always felt that queer literature and media were limited or insufficiently represented compared to other subjects. Access to these resources is an important way for both members of the LGBTQ community and people outside of it to learn about, understand, and accept these marginalized identities and come to appreciate queer sub-cultures. Fortunately, as more countries decriminalized and/or legalize same-sex marriage, and as (western/US) culture shifts to become more accepting of the LGBTQ community, I have started to see more and more authors and creators advocate for queer representation in literature and media. Therefore, I set out to investigate what types of queer books/media can the public access from the Seattle Public Library, and how has the public's interest in them has grown and shifted.

MySQL Queries

First, I need to figure out how to extract a more or less complete collection of books/media on queer topics from the library's database. One way the topic or the subject of a book is reflected is through its title. My first query selects all books/media containing LGBTQ keywords in their titles, from the `spl_2016.outraw` database of checkout records.

The keywords include 'lesbian', 'gay', 'bisexual', 'transgender', 'queer'. These keywords are part of the LGBTQ acronym, and so they should cover a good amount of relevant books/media. (The acronym can be extended to LGBTQIA+ to include more identities. I added more keywords in later queries correspondingly.) The keyword 'homosexual' is also included because it's often used as an umbrella term to describe non-straight relationships.

```
SELECT
    DISTINCT(bibNumber), title, deweyClass
FROM
    spl_2016.outraw
WHERE
    title LIKE '%homosexual%'
  OR title LIKE '%lesbian%'
    OR title LIKE '%gay%'
    OR title LIKE '%bisexual%'
    OR title LIKE '%transgender%'
    OR title LIKE '%queer%'
ORDER BY bibNumber
LIMIT 1000;
```

However, this query has two main issues:

- A word in the book title matches the keywords, typically a person's name, but the book isn't actually about queer topics. For example, *Enola* ***Gay***, *Hili**gay**non lessons*.
- Books/media that do cover queer topics don't always contain these keywords in their titles. For example, *Meditations in an emergency* (categorized as Gay Poetry), *Sita* (categorized as Lesbians United States Biography)

Therefore, I decided to explore how the `spl_2016.subject` database can help me better capture and extract queer books/media.

I found that the same set of keywords is able to cover a wide range of categories that should help me extract a more complete collection. Joining the `spl_2016.outraw` database with the `spl_2016.subject` database on each row's bibNumber column matches the book/media to the subjects it was assigned.

I explored 2 ways to join the tables, in anticipation of missing/mismatched records from 2 separate databases

- `spl_2016.outraw` LEFT JOIN `spl_2016.subject`
- `spl_2016.subject` LEFT JOIN `spl_2016.outraw`

to see how the coverage may change.

```
# spl_2016.outraw LEFT JOIN spl_2016.subject
# query result has 8000+ rows
SELECT
    *
FROM(
  SELECT DISTINCT(bibNumber) AS distinct_bib, title, deweyClass, itemtype
    FROM
        spl_2016.outraw
    WHERE
        title LIKE '%homosexual%'
    OR title LIKE '%lesbian%'
    OR title LIKE '%gay%'
    OR title LIKE '%bisexual%'
    OR title LIKE '%transgender%'
    OR title LIKE '%queer%'
        OR title LIKE '%intersex%'
    OR title LIKE '%asexual%'
    ORDER BY bibNumber
) queer_titles
LEFT JOIN spl_2016.subject
ON distinct_bib = spl_2016.subject.bibNumber;
```

```
# spl_2016.subject LEFT JOIN spl_2016.outraw
# query result has around 14K rows
# though it does involved duplicates
SELECT *
FROM(
  SELECT DISTINCT(bibNumber) AS distinct_bib, subject
  FROM
    spl_2016.subject
  WHERE
    spl_2016.subject.subject LIKE '%homosexual%'
    OR spl_2016.subject.subject LIKE '%lesbian%'
    OR spl_2016.subject.subject LIKE '%gay%'
    OR spl_2016.subject.subject LIKE '%bisexual%'
    OR spl_2016.subject.subject LIKE '%transgender%'
    OR spl_2016.subject.subject LIKE '%queer%'
        OR spl_2016.subject.subject LIKE '%intersex%'
    OR spl_2016.subject.subject LIKE '%asexual%'
  ORDER BY bibNumber
) bib_subj
LEFT JOIN (
  SELECT DISTINCT(bibNumber) AS out_bib, title, itemtype, deweyClass
  FROM
    spl_2016.outraw
)outraw_titles
ON distinct_bib = out_bib;
```

The query that searched for matching keywords from subjects produced a more complete collection with fewer irrelevant entries upon manual inspection.

Next, I want to investigate how queer books/media are classified and identify interesting patterns.

I selected distinct titles that have a Dewey Classification, to see what these works are classified as. Unfortunately, there is a non-negligible amount of books/media not categorized in the Dewey Classes, which may cause the distribution to be skewed.

```
SELECT DISTINCT(title), deweyClass
FROM (
  SELECT *
  FROM(
    SELECT DISTINCT(bibNumber) AS distinct_bib, subject
    FROM
      spl_2016.subject
    WHERE
      spl_2016.subject.subject LIKE '%homosexual%'
      OR spl_2016.subject.subject LIKE '%lesbian%'
      OR spl_2016.subject.subject LIKE '%gay%'
      OR spl_2016.subject.subject LIKE '%bisexual%'
      OR spl_2016.subject.subject LIKE '%transgender%'
      OR spl_2016.subject.subject LIKE '%queer%'
      OR spl_2016.subject.subject LIKE '%intersex%'
      OR spl_2016.subject.subject LIKE '%asexual%'
    ORDER BY bibNumber
  ) bib_subj
  INNER JOIN (
    SELECT title, itemtype, deweyClass, bibNumber as out_bib
    FROM
      spl_2016.outraw
    WHERE deweyClass != ''
  )outraw_titles
  ON distinct_bib = out_bib
)inner_join_table
ORDER BY deweyClass;
```

Lastly, I aggregated checkout data from 2006 to 2023 of books/media under each queer subject keyword, grouped and ordered by year and month, in order to explore how people’s interest has evolved through the two decades.

For example, the query for the checkout data for books/media on the bisexual subject as follows:

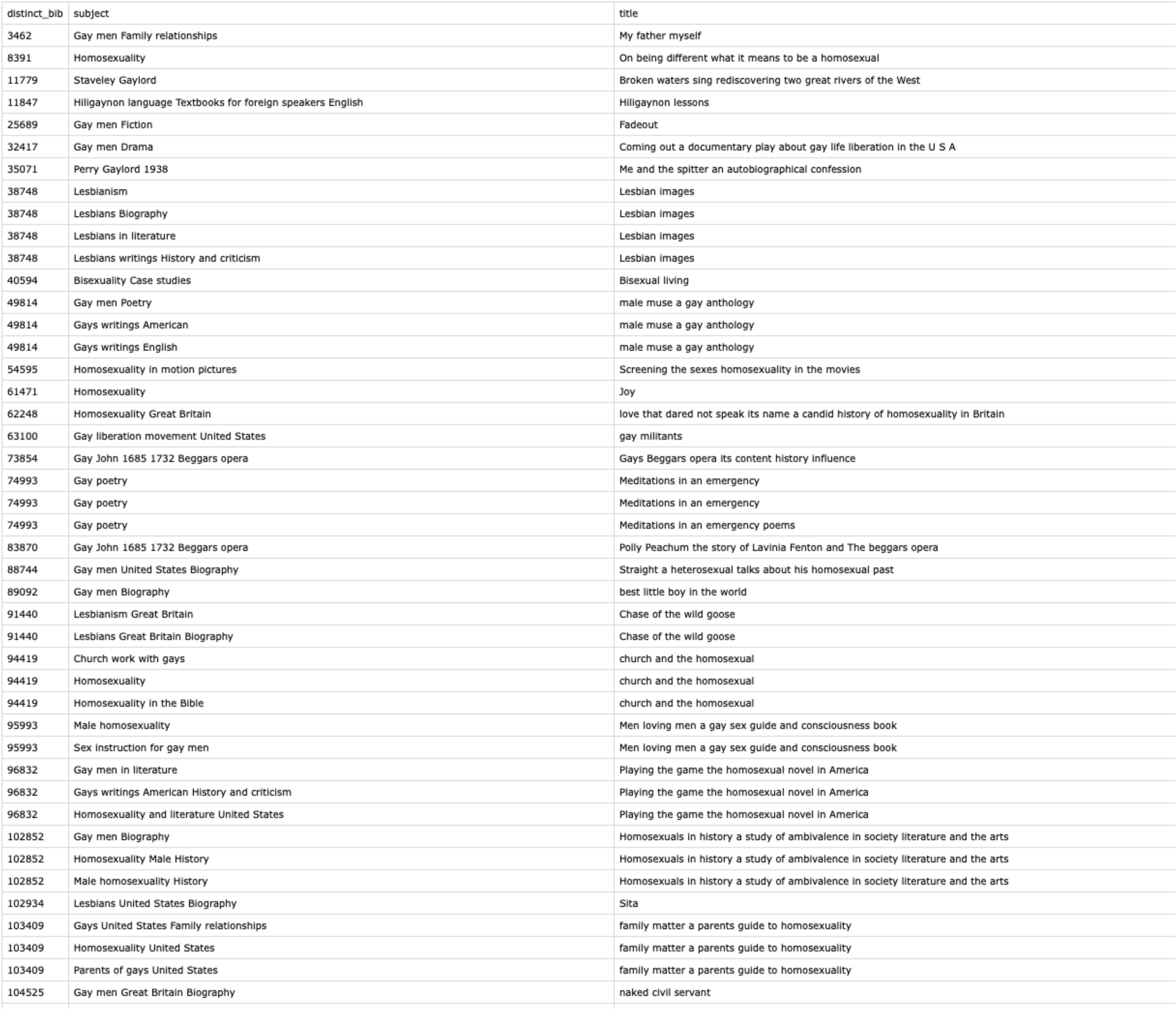
```
SELECT
    YEAR(cout) AS YR, MONTH(cout) AS MO, COUNT(cout) AS NUM_CHECKOUT
FROM(
  SELECT *
```

```
    FROM(
      SELECT DISTINCT(bibNumber) AS distinct_bib, subject
      FROM
        spl_2016.subject
      WHERE
        spl_2016.subject.subject LIKE '%bisexual%'
      ORDER BY bibNumber
    ) bib_subj
    INNER JOIN (
      SELECT title, itemtype, deweyClass, bibNumber as out_bib, cout
      FROM
        spl_2016.outraw
    )outraw_titles
    ON distinct_bib = out_bib
) inner_join_table
GROUP BY YR, MO;
```

Data and Results

The queries resulted in quite interesting results. While I wasn’t able to avoid all irrelevancy and errors in the produced book/media collection, when I manually check the CSV file, it didn’t look like they were a majority or would drastically affect data aggregation.

It’s interesting to see that each book is often assigned multiple subjects. From the dataset, it seems that these subjects are often related and overlap with each other.

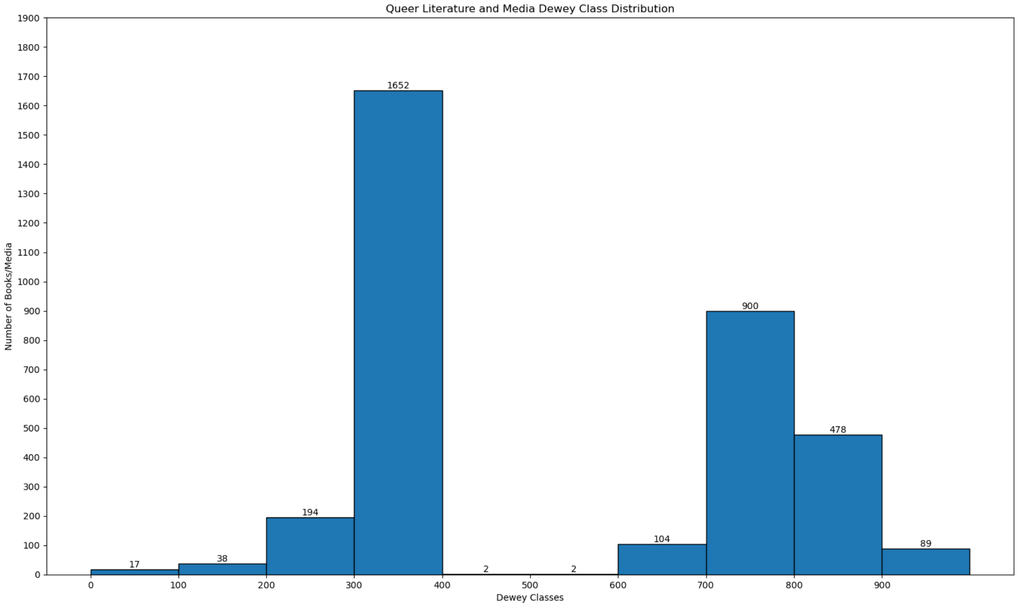


distinct_bib	subject	title
3462	Gay men Family relationships	My father myself
8391	Homosexuality	On being different what it means to be a homosexual
11779	Staveley Gaylord	Broken waters sing rediscovering two great rivers of the West
11847	Hiligaynon language Textbooks for foreign speakers English	Hiligaynon lessons
25689	Gay men Fiction	Fadeout
32417	Gay men Drama	Coming out a documentary play about gay life liberation in the U S A
35071	Perry Gaylord 1938	Me and the spitter an autobiographical confession
38748	Lesbianism	Lesbian images
38748	Lesbians Biography	Lesbian images
38748	Lesbians in literature	Lesbian images
38748	Lesbians writings History and criticism	Lesbian images
40594	Bisexuality Case studies	Bisexual living
49814	Gay men Poetry	male muse a gay anthology
49814	Gays writings American	male muse a gay anthology
49814	Gays writings English	male muse a gay anthology
54595	Homosexuality in motion pictures	Screening the sexes homosexuality in the movies
61471	Homosexuality	Joy
62248	Homosexuality Great Britain	love that dared not speak its name a candid history of homosexuality in Britain
63100	Gay liberation movement United States	gay militants
73854	Gay John 1685 1732 Beggars opera	Gays Beggars opera its content history influence
74993	Gay poetry	Meditations in an emergency
74993	Gay poetry	Meditations in an emergency
74993	Gay poetry	Meditations in an emergency poems
83870	Gay John 1685 1732 Beggars opera	Polly Peachum the story of Lavinia Fenton and The beggars opera
88744	Gay men United States Biography	Straight a heterosexual talks about his homosexual past
89092	Gay men Biography	best little boy in the world
91440	Lesbianism Great Britain	Chase of the wild goose
91440	Lesbians Great Britain Biography	Chase of the wild goose
94419	Church work with gays	church and the homosexual
94419	Homosexuality	church and the homosexual
94419	Homosexuality in the Bible	church and the homosexual
95993	Male homosexuality	Men loving men a gay sex guide and consciousness book
95993	Sex instruction for gay men	Men loving men a gay sex guide and consciousness book
96832	Gay men in literature	Playing the game the homosexual novel in America
96832	Gays writings American History and criticism	Playing the game the homosexual novel in America
96832	Homosexuality and literature United States	Playing the game the homosexual novel in America
102852	Gay men Biography	Homosexuals in history a study of ambivalence in society literature and the arts
102852	Homosexuality Male History	Homosexuals in history a study of ambivalence in society literature and the arts
102852	Male homosexuality History	Homosexuals in history a study of ambivalence in society literature and the arts
102934	Lesbians United States Biography	Sita
103409	Gays United States Family relationships	family matter a parents guide to homosexuality
103409	Homosexuality United States	family matter a parents guide to homosexuality
103409	Parents of gays United States	family matter a parents guide to homosexuality
104525	Gay men Great Britain Biography	naked civil servant

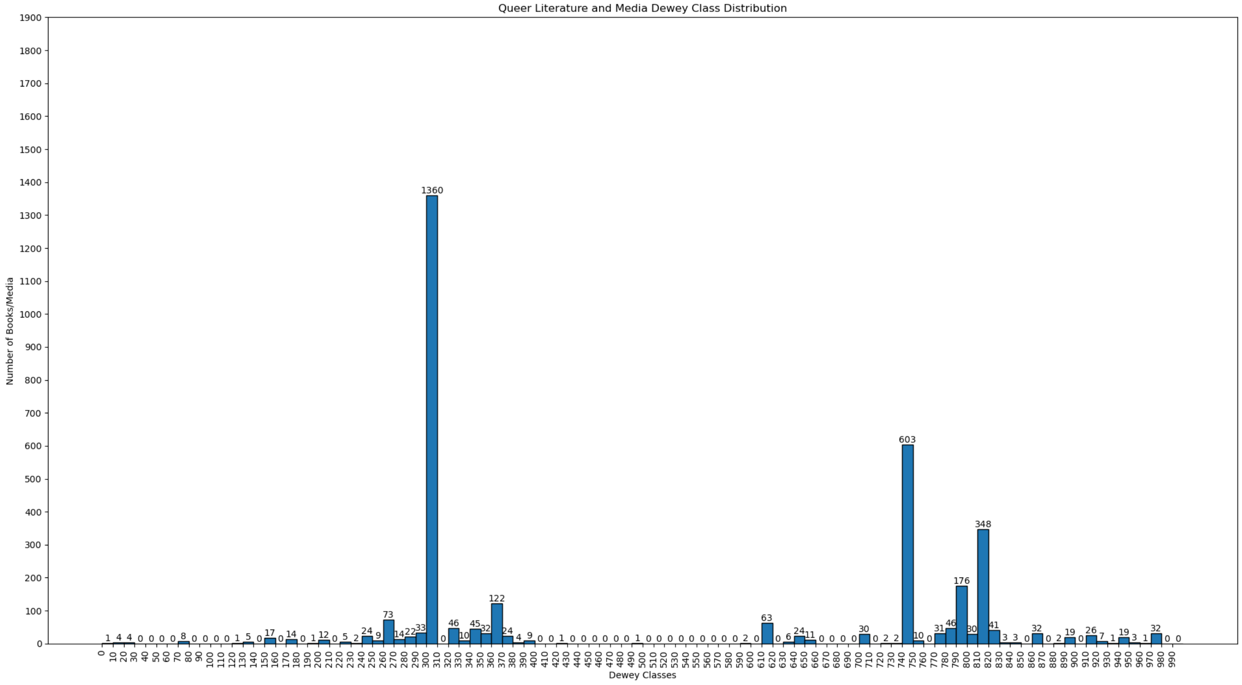
Collection of queer books/media and their assigned subjects by SPL

This is part of the reason why I decided to work with queer books/media that are already categorized into Dewey Classes. As I've mentioned in the MySQL Queries section, the drawback is that those entries without a Dewey Class are unfortunately discarded from the dataset, which can skew distribution.

The generated distribution bar chart revealed significant representation in the 300 Dewey classes, specifically Social Sciences, Sociology and Anthropology classes. The next two most significant categories include the 700-800 classes on Arts and Recreation, and the 800-900 classes on Literature.



Queer books/media is most represented in the 300-400 categories (Social Sciences).
The next most popular are the 700-800 (Arts and recreation) categories.

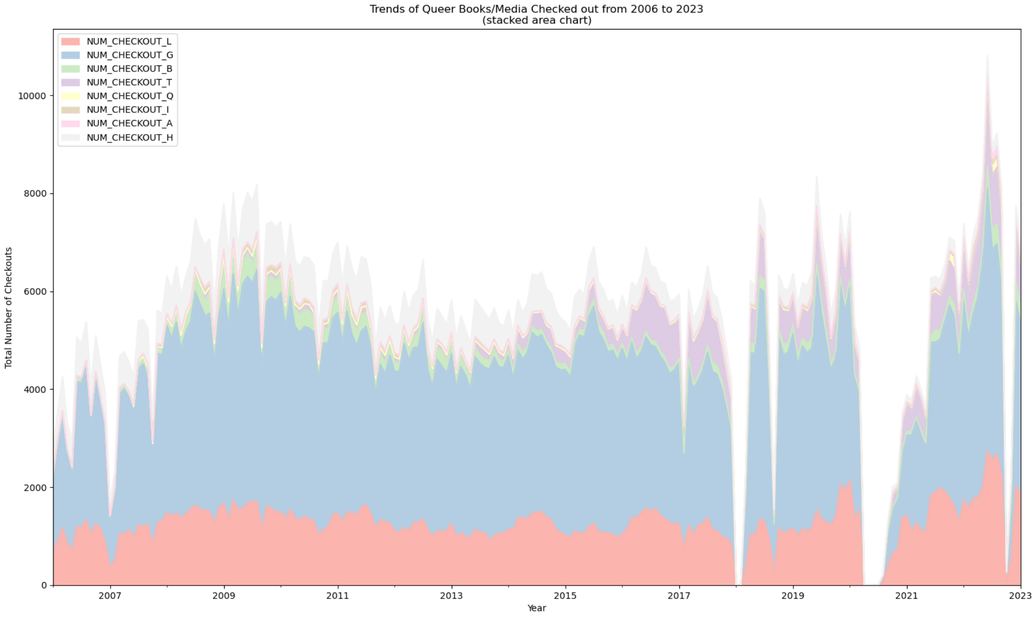


Let's take a finer look, the 300-310 Social sciences, sociology and anthropology categories are the most represented, with 1360 entries.
The next categories are 740-750 Graphic arts and decorative arts, and 810-820 American literature in English.

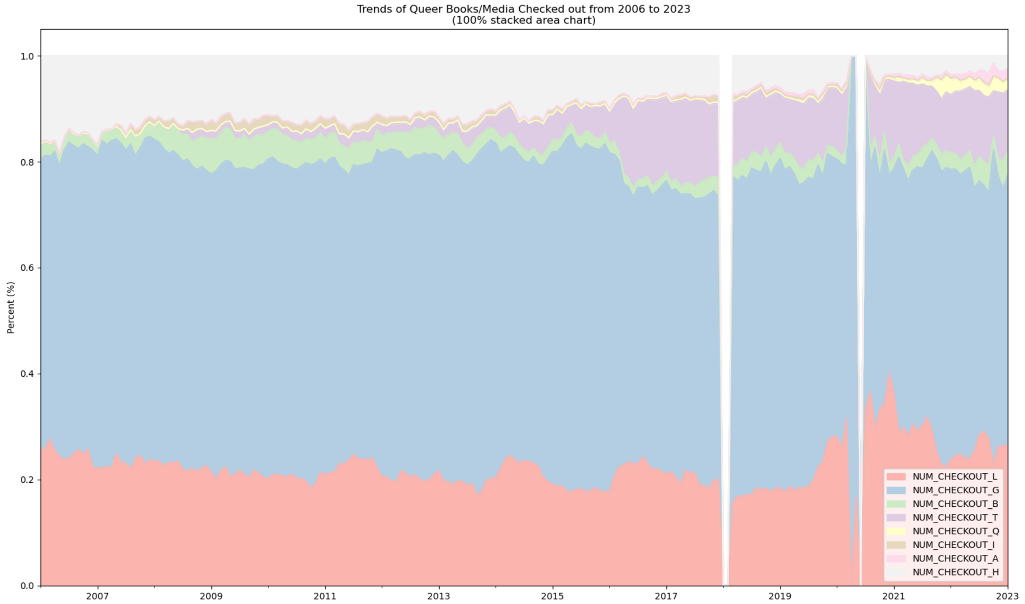
Lastly, the aggregated checkout records over the years showed an overall upward trend of increased checkout records of all queer books and media. I've separated into 8 subgroups, each letter in the LGBTQIA acronym as a subgroup, plus the broader 'homosexuality' subgroup. While most subgroups experience smaller increase in checkouts, there are two notable subgroups: the 'transgender' subgroup and the 'homosexuality' subgroup.

Since early 2013, checkout records of books/media on transgender subjects have experienced a more significant increase compare to other subgroups, suggesting the topic gaining more public interest. On the contrary, checkout records gradually decreased for the 'homosexual' subgroup over the two decades, as the usage of the word as a broad category of sexual and gender minorities falls out of favor.

Apart from the 5 months of missing data in 2020 due to COVID-19, I've also found that in the early months of 2018, the SPL database had extremely low to no checkout records from any of the subgroups of books/media of queer subjects.



The stack area chart shows the overall trend of public interest in queer literature and media.



The 100% stacked area chart reveals relative trend differences between subgroups more clearly. Notably, interests in transgender and queer subjects increased, interests in homosexual and intersex subjects decreased, and a general trend of diversification of interests in LGBTQ+ literature and media.

Discussion and Future Work

The results from the MySQL queries I conducted for this project revealed interesting patterns in the data. Some were expected, while others were surprising. It appears that queer books/media with a Dewey Classification, are most represented in Social Sciences, followed by Arts and Recreation, and then Literature. This is consistent with my expectations prior to conducting the queries. LGBTQ identities and the community have become quite prominent in the larger social discourse, and surrounding them there are many important social issues worthy of attention, including civil rights, acceptance, and legal recognition, etc. It makes sense that there are a lot of works in Arts and Literature that describes and depicts the LGBTQ community. Moreover, Arts and Literature also provide important channels of self-expression for queer artists, authors, and the community.

For Dewey Classes such as Science and Technology, it’s not surprising that there is little to no representation. Lower representation in Religion and History topics reflects historical and religious marginalization of queer identities. I’m a little surprised that there is almost no entries in Language Classes, since queer Linguistics is definitely an area of interest of many scholars.

Perhaps it's because this area of research and publication is still quite new and a lot of it exist online instead of the Library.

For future work, I'm definitely interested in exploring the `spl_2016.subject` database in more detail, more than just using it as a way to filter and extract books/media on specific subjects. In particular, if I had more time, I would compile of a list of all subjects related to LGBTQ identities, explore overlaps in similar subject labels, and measure similarities between them and categorize groups of related queer subjects. Furthermore, since the subjects of each book/media entry are assigned manually by people who recorded the item into the SPL database, investigating if the subjects also experience changes in trends corresponding to the public's attitude towards LGBTQ topics could be interesting.

As for trends over the years, the query results also generally coincided with my own expectations and knowledge of how public attitudes toward queer identities and the LGBTQ community has evolved (in the United States). Historical events in the United State, such as the U.S. v. Windsor / Repeal of the Defense of Marriage Act – DOMA (Supreme Court Decision) in March of 2013, and the Obergefell v. Hodges (Supreme Court Decision) decriminalized and legalized same-sex marriage, sparking more positive attitudes and discourse on recognizing and accepting LGBTQ+ people. While the popularization of e-books and more available access to the internet and online resources may have impacted overall Library records, we can still see increasing checkout records of queer books/media since the 2010s.

Moreover, interests in different queer subjects has diversified. While the Gay subgroup continues to have most checkouts, the Lesbian subgroup slowly increased, especially after late 2019. The Transgender subgroup experienced the most increase in checkouts. A similar group, the Queer subgroup also experienced increased public interest, especially after 2021. Two subgroups, however, experienced decrease in interest, the Homosexual subgroup (overall and relative) and the Intersex subgroup (relative).

The word 'homosexual' was often used as a broader term to cover all non-heterosexual relationships. However, this term is not necessarily accurate nor does it represent queer identities inclusively. As attitudes toward queer identities become more positive, words and labels that better describe different queer identities emerged and are general more known and accepted. The trends reflected the usage of the broader and older term falling out of favor, as well as more recognition of the diversity within the LGBTQ community.

As awareness and discourse (and unfortunately controversy and transphobia) on one of the most marginalized sub-community — the transgender community increased, increase interests in books/media on transgender and gender-conforming people and topics are reflected in the trend. It’s possible that the transgender identity getting declassified as a mental illness and re-classified as 'gender dysphoria' in 2013 contributed to a more normalized perception of the word ‘transgender’, thus contributing to this increase as well.

Overall I think the data from Seattle Public Library reflected attitudes toward and interests in LGBTQ identities and the community rather accurately. The checkout data reveals increasing public interest in learning about the LGBTQ community, which may lead to more acceptance in our society. The data also validated that the increase in positive queer legislation, advocacy, and discourse in recent years was able to help diversity queer identities gain more recognition and acknowledgement.