SIADS 591/592: Milestone 1 Project

Chalse Okorom-Achuonye, Cameron Grams

Motivation

The goal of our project was to look at the use of resources within the Seattle Public Library system and try to offer recommendations to better serve local communities. The fact that not all patrons of a library use the resources the same way is clear to anyone who visits a modern library. Some patrons are there for the books or other physical materials, others for community sponsored events, and others for internet access or simply access to electrical power or shelter. Others even prefer to access digital resources from the comfort of their homes.

Libraries serve a variety of functions, and continue to diversify as our needs develop and change. We hope to be able to communicate this and better clarify this through our research. At this stage, we aim to focus on the physical aspects of the library, the physical resources, circulation, and branch locations accessed by the patrons.

The Seattle Public Library system is made up of 27 branch locations.These branches are spread throughout the city to better afford the citizens access to the resources. These locations vary in terms of the demographics of the surrounding areas. Originally we hoped to be able to examine the financial health of the branch locations and draw conclusions about the level of support to the surrounding community, however this was not possible given the level of non-PII (personally identifiable information) obtained from the city of Seattle.

Further, the Seattle Public Library is largely treated as a single system where physical resources are allocated by designation to specific collections and subjects. Subjects refer to the topic inside physical materials while the collection is a more broad descriptor and also refers to the location of materials within a branch. By analyzing the volume of materials being accessed, we hope to instead gain a high level of understanding of the items patrons utilize. Additionally, by using the collection information, we can further see how to align and allocate resources towards materials patrons are gravitated towards.

These collections, the material circulation, and their access from the 27 branch locations is the focus of this report. The goal was to better understand the level and type of service that is needed for the greater Seattle community.

Data Sources

City of Seattle Open Data Portal

The City of Seattle has an Open Data Portal that contains thousands of published and regularly updated datasets by various city departments. Data generated by the public available for anyone to use in order for the public to increase transparency, and accountability. Through this program, you can find historical checkout data from the Seattle Public Library dating back to April 2005 as well as a dataset containing descriptions and categories for various codes used within the primary dataset. These data sets are available via download as .csv or .json or programmatically through the Socrata open data API.

1. Seattle Public Library - Checkouts By Title (Physical Items): This dataset includes a log of all physical item checkouts from Seattle Public Library. The dataset begins with checkouts that occurred in April 2005 and is regularly updated with new checkouts. Additionally, within this dataset, renewals are not included and the checkout times are rounded to the nearest minute. This dataset includes over 108 million entries and is over 27 GB in size.
   Primary Features:
   a. Checkout Date Time - Utilized month and year to identify trends and/or usage
   b. Item Type - A code from the catalog record that describes the type of item. Some of the more common codes are, for example: acbk (adult book), acdvd (adult DVD), jcbk (children's book), accd (adult CD)
   c. Collection - A collection code from the catalog record which describes the item. For example: nanf (adult non-fiction), nafic(adult fiction), ncpic(children's picture book), nycomic (Young adult comic books).
2. Integrated Library System (ILS) Data Dictionary: This dataset is useful for understanding the codes used in some of Seattle Public Library's other datasets. These codes (namely "ItemType" and "ItemCollection") are systematically used in the cataloging of items within the Integrated Library System (ILS).
   Primary Features:
   a. Code - A code used to describe SPL items by the following dimensions: branch location, item collection, item medium, etc.
   b. Description - The full description of what an item’s code means
   c. Format (Sub) Group - The high-level description of an item or thing (i.e. print, periodical, etc.)

Seattle Public Library
Our original intention was to examine the financial health of the branches but this proved to be impossible since the financial information on the branches is maintained in aggregate at the city level. However, when we reached out to the Staff at the Seattle Public Library and explained our intentions, they were very helpful. The Director of Public Communications, Laura Gentry, provided us with the door count information for each of the branch locations. The dataset was

extracted from 10 PDF files and composed into a csv file of about 40 KB in size.

3. Branch Door Counts: This dataset contains information about individual branches in relation to foot traffic, and open hours
   Primary Features:
   a. Time Period - Utilized month and year to identify trends and/or usage
   b. Branch - The code for an SPL branch
   c. Total Visits - Total number of visits to a branch within the given time period

Methodology
From the Seattle Open Data website, we were able to access checkout information for physical items from 2005 up until to present day. This was our first dataset and it provided a wealth of information but accessing the full dataset was quite resource intensive and couldn't be done over Deepnote as originally planned. To manage this resource we downloaded the full collection and managed the analysis locally using PySpark. Using PySpark we were able to pose very detailed queries, discovering for instance that "The Catcher in the Rye" has been checked out of Seattle Public Library branch locations 5,861 times since 2005!

This dataset proved to be a gold mine, since it showed the volume of people accessing the branch locations with numerous features per checkout. The door count data was provided in the form of 10 PDF files that contained a consistent arrangement of tables for the months of each year. We were able to convert these pdf to csv files using the Python library tabula. At that point managing the analysis from Pandas was straightforward, since each years' record was only 11 KB.

The third dataset was essential to understanding the other two; the dictionary of terms and acronyms used by the library provided a key to understanding the other entries. The dictionary was a csv file of 586 entries and had to be integrated at multiple points of the analysis to understand the many acronyms used by the library. This was the only way to match acronyms to the names between the collections of physical material and the branch locations. Importing and managing the csv file was done in Pandas.

***Note**
Since, the door count data is only available from 2010 to 2019, thus, the physical checkout analysis is limited in timeframe as well with an addition to the year 2020. This year was added as it seemed relevant to the current state of the world that both door counts and physical checkouts are severely limited in 2020 due to the COVID-19 pandemic.

Data Manipulation Methods

Physical Checkouts

Because of the size of this dataset we relied on PySpark for most of the manipulations. In PySpark we were able to create SQL queries as we searched for specific features. The key features that we were interested in working with were: the year, month and and overall counts for the checkouts of library items. The dates were stored as strings and had to be converted into datetime data types. Once in datetime format, the months and years were able to be used for querying the dataset by further categorical features from the third dataset such as age groups, physical item types available for checkout, checkout types, etc.

Due to the large size of the dataset, the first act of data manipulation was to cut the dataset down into the correct time period for our analysis and reduce it to the rows we aim to view and analyze. The data was saved in this state in spark and further cut down and filtered accordingly using spl special queries for each type of data used in the plots.

Once the data was filtered and aggregated using spark sql statements, we then converted the much smaller spark dataframe into a pandas dataframe. Once it was a pandas dataframe, we used the ILS data dictionary to convert the library item type and collection type codes into human readable text. Additionally, age groups and format types or subgroups were added to the data frame as well. The additionally categorical variables were then used either as facet columns or to increase plot expressiveness and readability with color. One issue with using collections or item types for the legend is that plots became less readable when the legend had more than thirty colors/variables. Finally, the visualizations for these manipulations were created using Plot.ly express.

Door Counts

The first step in managing the door counts was reading in each of the ten door count PDF files. The 'tabula' library returns a list of dataframes in which each PDF page is a separate dataframe. For the purposes of our analysis, this meant that individual months had to be extracted from sometimes multiple pages.

Additionally, these features were not uniformly translated; title years or totals did not replicate across rows and populated the dataframes with NaN values. The NaN values were replaced by the appropriate dates by aligning the other entries in the rows and adding the correct date. String-formatted months and years were integrated into datetime objects using the Python datetime library.

We produced the functions in the notebook, 'door_counts_all_years.ipynb' to manage the migration to a single year's csv file, which was then posted to GitHub. We also included an audit function that ensured that each branch had 12 months of records per year. Only one branch (the Rainier Beach Branch Library) in 2015 did not have the full year's worth of door counts in the data.

The 'Library_Use' notebook provides the code we used to produce the trends in overall library branch visits, trends in visits vs checkouts, and the correlation between library visits and material use. The process involved creating a Spark Session reading in the main checkouts csv file and formatting it with an appropriate datetime feature. Finally, the visit information was integrated from the door count csv file and lastly the dictionary in order to present the information most effectively.

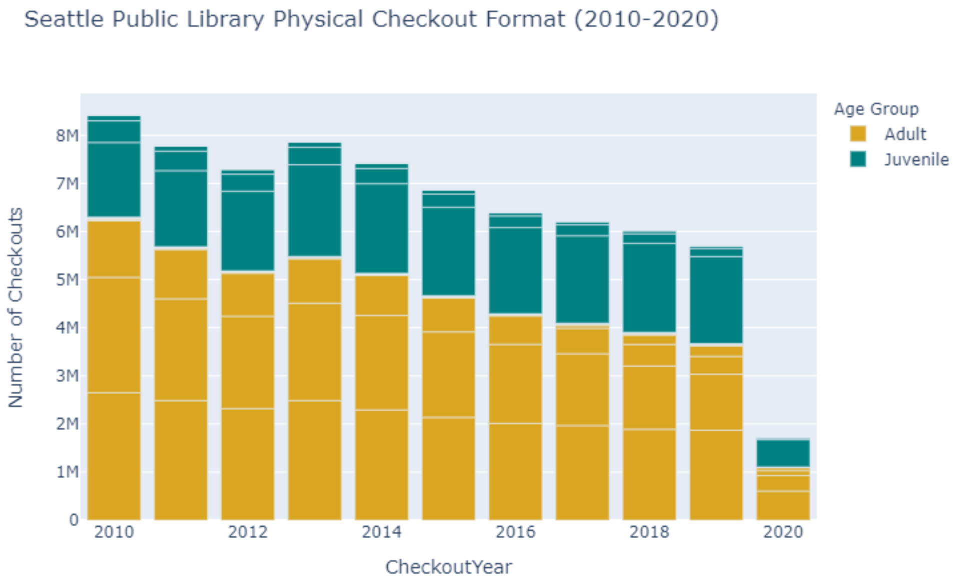
Dictionary

The last dataset used was the dictionary file that matched acronyms to the terms used by the library. The values were all strings and easily managed by Pandas once imported into the different notebooks. This was no more difficult than importing a csv file and then converting or filtering the features that were needed.

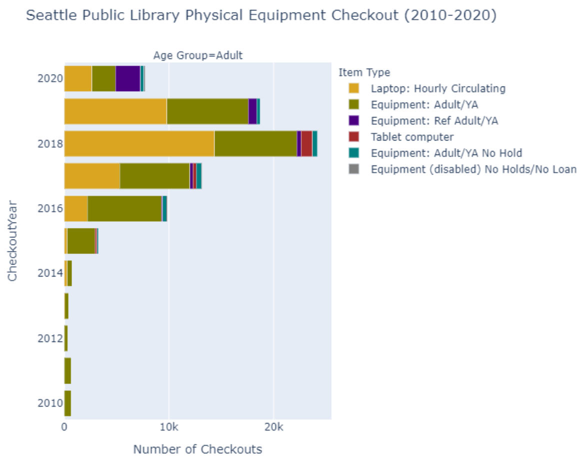
Analysis & Visualization

Physical Checkout of Materials

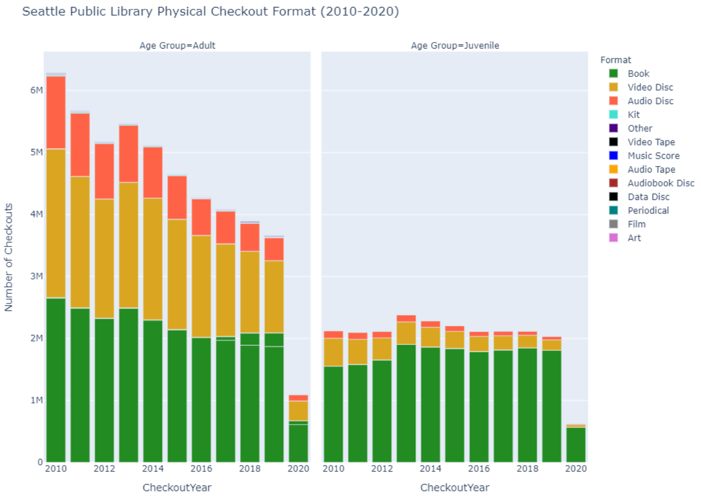
These visualizations are best viewed in the accompanying notebook, “spl_checkouts.ipynb”.



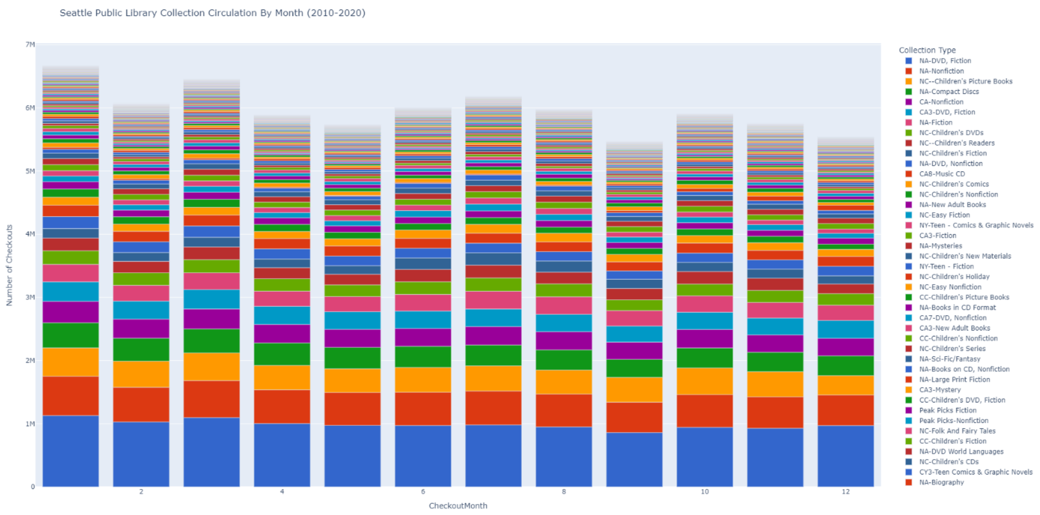
This chart is an overall glance at total circulation within our referenced time period, as well as 2020. It shows the total circulation across branches by checkout record and breaks down usage by age group. In general, there is a negative trend in checkouts. Additionally, there is a steep dropoff in physical item usage in 2020, likely a cause of the COVID-19 pandemic.



This plot shows the physical equipment checkouts, which are only available for use by adults. By 2014, other physical equipment usage has diversified, with an overall trend of increased usage for laptop checkouts. There is an increase in usage of reference materials, or equipment only usable within the library as well as normal equipment check outs (speakers, projectors, media carts, etc.). Overall, this plot shows an increasing trend in physical equipment checkouts sans 2020. This growth in usage over the past five years may suggest that patrons have a vested interest in borrowing laptops for personal/professional purposes or perhaps using (or learning how to use) AV equipment. These insights could inform further library resource acquisition and programming.



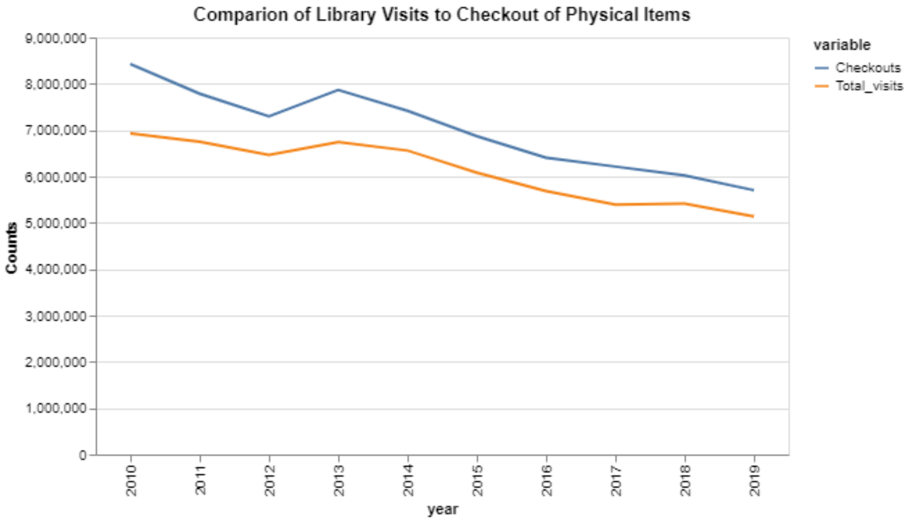
This plot shows the ratio of reference material formats for both adults and children. All patrons prefer to check out books, DVDs, and CDs. The decreasing trend in checkouts is also visible here for adults, but isn't there for juvenile patrons. Juvenile patrons may have more influence by parents, librarians, and educators to use the library while adults don't necessarily have that driver. Perhaps this may suggest additional programming that may be needed to increase these numbers. On the other hand, if we analyze digital checkouts (outside the scope of this analysis), there may be a trend that could infer directing library resources towards further fostering digital circulation.



This chart echoes the findings from the previous chart but offers further breakdowns of the more general types such as “Book” or “Periodical”. For example, from this chart, we can see the most popular collections for checkout each month are ‘Adult Fiction DVDs’, followed by ‘Nonfiction Books’ and Children's books. This high usage of DVDs may point to needs by those with less (or less updated) technology access that may need to be assessed for effectiveness.

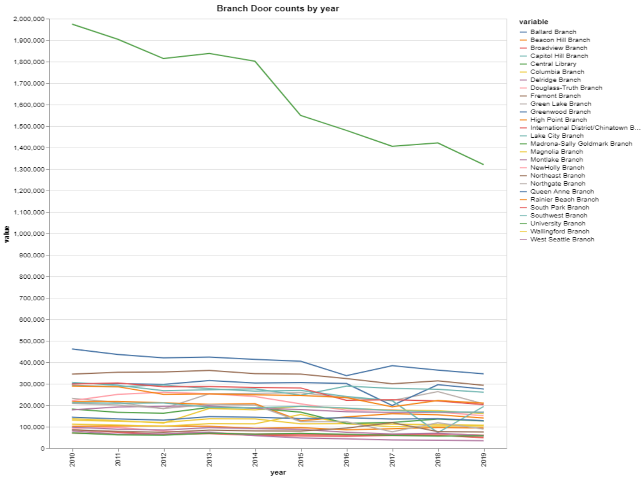
Branch Visits

These visualizations are best viewed in the accompanying notebook, “Library_Use.ipynb”. There the tooltip information in the graphics is functional and branch names are more legible.

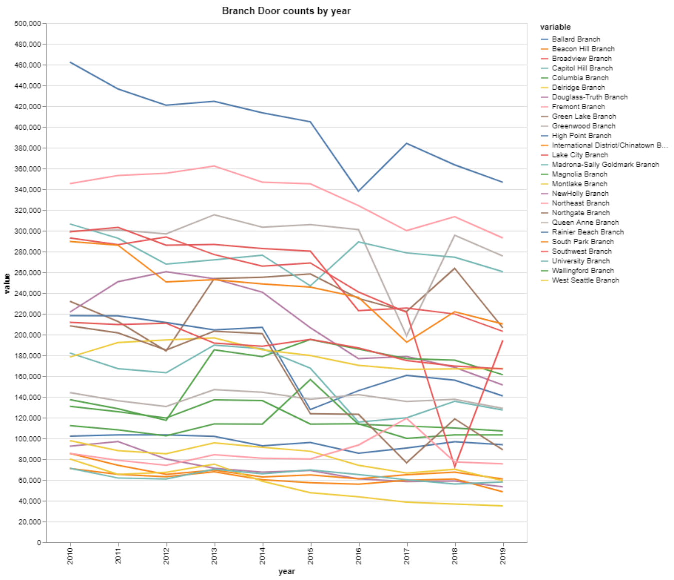


Over the 10 year time period there was a decrease in the volume of visitors to the branch locations that was closely correlated to the decrease in physical item checkouts. The difference between the volume of checkouts and the number of visitors decreased over time; patrons made progressively less use of physical items less each year over the 10 year period (the narrowing of the gap between the two lines). The composition of the other materials that the patrons were using was not part of the analysis (see Conclusion).

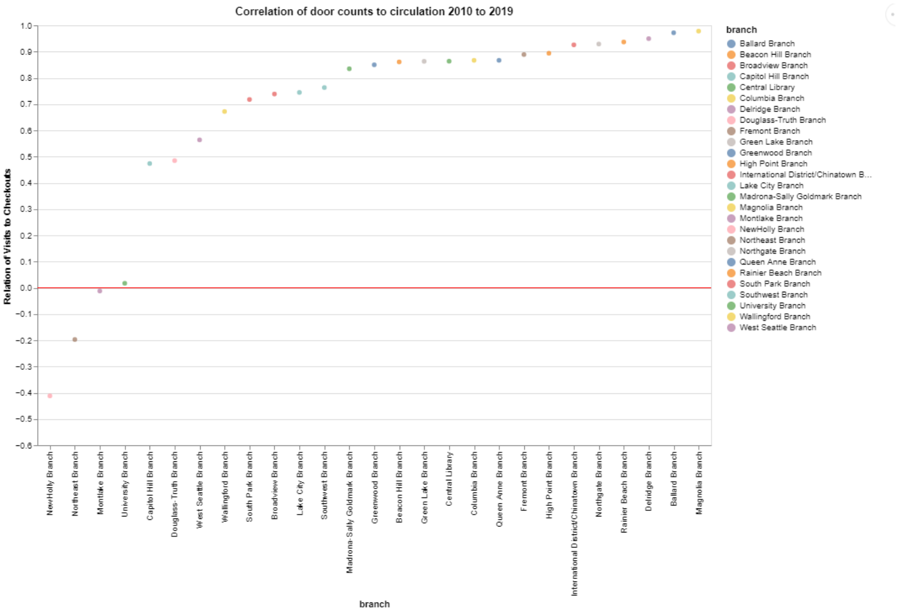
Another interesting characteristic of the analysis of branch attendance is the increase in visits and use of physical material that took place in 2013. There was a greater change in physical material use than the change in increased attendance. Both trends were positive in a way that was counter to the prevailing trend, but the 2013 increase in physical item use was greater than the increase in attendance.



The decrease in visits to branch locations was not seen to the same degree at each of the branches. The decrease was most apparent at the Central Library Branch location, but all of the branches had some degree of decrease in the number of visitors.



Filtering out the Central Library data from the other branches reveals a less dramatic decrease in branch attendance but still a negative trend. When seen at this scale the sharp dip for the Ballard Branch, Greenwood Branch, and Lake City Branch stand out in sharp contrast.



One of the main motivations for this project was to better understand the use of library resources. The correlation between visits to the branches and the number of physical items checked out from the library branches clarifies some differences among the branches. The 4 branches on the left have an inverse relationship between the number of visitors and the number of physical items checked out from the libraries. This indicates that the patrons who frequent these locations make greater use of other library resources.

Conclusions

There has been a clear decrease in the volume of physical resources used in the Seattle Public Library system. This trend and the composition of what materials are in highest demand varies between the individual libraries. There are three aspects of the data that we would like to highlight; first the use of physical materials from the libraries has remained steady across years for the Juvenile collections, second two branches (the NewHoly Branch, and the Northeast Branch) have an inverse relationship of visitors and the expected volume of physical material checked out, and third 2013 had a level of both visitors and materials checked out that was higher than anticipated by previous trends.

This analysis did not look at access to digital media or services at the branch locations. Future research into how often and in what volume use of internet access takes place in the branches is one area that would help provide a better understanding of modern library use. The same sort of analysis for services and activities at the branch locations would provide a more nuanced understanding.

It is clear that not all branches are used by patrons in the same manner, and some materials are more popular than others. More research on the resources used and activities at each of the branches is likely required to fully understand the dynamics at play within the Seattle Public Library system.

Statement of Work

Chalse, physical checkout analysis

- Prepare primary dataset, add the abbreviations, and condense checkouts by year within the overlapping time frame (2010-2020).
- Analyse checkout volume by year and material type, with additional filters

Cameron, branch visit analysis

- Prepare a secondary dataset, extract yearly door count information from PDFs
- Associate traffic records with a branch location and correlate circulation with branch location

Both

- Visualize findings, and describe analyses in final report