Trends in Seattle Library Usage

Taylore Prinsen
taylore.prinsen@colorado.edu
Colorado University at Boulder
Boulder, Colorado

Sky Johnson
sky.johnson@colorado.edu
Colorado University at Boulder
Boulder, Colorado

Peter Minwegen
peter.minwegen@colorado.edu
Colorado University at Boulder
Boulder, Colorado

Figure 1: The Seattle Public Library

ACM Reference format:
Taylore Prinsen, Sky Johnson, and Peter Minwegen. 2019. Trends in Seattle Library Usage. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 6 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT/MOTIVATION

The Seattle Public Library is a landmark of the city for it's architectural style, and unique community gatherings. However we wanted to look further into it's functions as an actual library.

It is currently unclear whether or not trends in rentals of library books (or other materials) in the Seattle area are impacted by the weather trends in the area. We intend to use a data set with weather statistics in Seattle as well as a data set with statistics related to the library check outs in the area. We would like to explore whether or not people check out books more or less when the weather is rainy, sunny, etc. We would also like to see if there is a correlation between winter and summer months and the topics of books that people check out. This could be very interesting to help us better understand how people tend to cope with varying weather conditions.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Many people suffer from Seasonal Affective Disorder (SAD) and are left facing varying side effects from weather impacting their mood and demeanor. We are intrigued by the ways that weather can impact people's moods and behaviors and are intrigued to begin to understand how this may impact people's daily habits, going to the library, etc. We are also intrigued to learn if these weather trends tend to influence the genre of books or music that people check out.
https://www.mayoclinic.org/diseases-conditions/seasonal-affective-disorder/symptoms-causes/syc-20364651

2 LITERATURE SURVEY

There has been very little research done with these data sets in the past. The only work we came across that used the Seattle Public Library data set only referenced data specific to that database - what time of day is most popular for rentals at the library, which type of rental is most popular at certain times, etc. At the end of this quick study the author came to the conclusion that 4pm, Fridays in January and July are the busiest days for activity within the library's checkout system dataset. However this is limited and doesn't ask much of why that is the case. We intend to extend this research by combining it with weather data from the area to better understand how weather might impact rental trends.

Some of the trends that previously have been studied in the Seattle Public Library Data set include:

- Number of checkouts across the years (separated by books and DVD/CD)
  A steady number of book checkouts (around 4 million) observed, while CD/DVD checkouts decreased over time, peaking around 2009 - probably from the emergence of Netflix, Spotify, etc.

- Checkout temporal trends:
  - Time of Day
    4 p.m. is the most popular hour for checkouts - possibly because people get off work and then head to the library?
  - Day of the Week
    Saturday has remained the most popular day for all years in the data set (Friday and Sunday are the least popular days).
  - Month of the Year
    "The beginning of the year and the summer seem to be the most popular months for checkouts, with September and December the least popular." (CITE - from toward data science link)

https://towardsdatascience.com/how-and-when-people-use-the-public-library-1b102f58fd8a

3 PROPOSED WORK

Due to the exploratory nature of this project, we will have to revise and plan to explore more options of analysis as we continue to work on this. However, we do know some basic parts that we will have to work on at some point

3.1 Data Pre-Processing

The nature of our data sets will involve a decent amount of effort within integrating. The checkout data doesn't inherently follow the same structure as the weather data, so we will have to figure out the best way to align those. Currently we plan to work with the checkout data initially to find basic statistics related to the libraries usage. (These will be nothing but high level observations and extrapolation). Then, we will compare those observations and directly add in weather data in a way that matches, that way we can begin to create more accurate observations that includes the weather data.

3.2 Time Series Analysis

The next big challenge that we will have to approach is how we handle time series within the data set. Currently we plan to use some basic mathematical statistics APIs to quickly find and visualize time data on how check out patterns change.

4 DATA SETS

We will be using two (possibly more) data sets for our main research. We will use one data set that tells us about the checkout trends at the Seattle Library, and concatenate this with a data set that tells us the weather trends in the same months. By looking at these together, we hope to answer complex questions about trends in weather and library rentals.

4.1 Seattle Checkouts by Title

This data set contains information for the Seattle Public Library for physical and digital data. There are 1,431,563 unique objects in the data set, each with data for 11 attributes:

- UsageClass (Nominal)
  Denotes if item is âĂIJphysicalâĂİ or âĂIJdigitalâĂİ
- CheckoutType (Nominal)
  Denotes the vendor tool used to check out the item.
- MaterialType (Nominal)
  Describes the type of item checked out (examples: book, song movie, music, magazine)
- CheckoutYear (Numeric/Interval)
  The 4-digit year of checkout for this record.
- CheckoutMonth (Nominal)
  The month of checkout for this record.
- Checkouts (Numeric/Interval)
  A count of the number of times the title was checked out within the âĂIJCheckout MonthâĂİ.
- Title (Nominal)
  The full title and subtitle of an individual item
- Creator (Nominal)
  The author or entity responsible for authoring the item.
- Subjects (Nominal)
  The subject of the item as it appears in the catalog.
- Publisher (Nominal)
  The publisher of the title.
- PublicationYear (Numeric/Interval)
  The year from the catalog record in which the item was published, printed, or copyrighted.

This data set begins in 2005. Link: https://www.kaggle.com/city-of-seattle/seattle-checkouts-by-title

4.2 Seattle Weather Data

4.2.1 Did It Rain in Seattle? (1948-2017). "Besides coffee, grunge and technology companies, one of the things that Seattle is most famous for is how often it rains. This dataset contains complete records of daily rainfall patterns from January 1st, 1948 to December 12, 2017."
This data was collected at Seattle Tacoma International Airport, WA.
https://www.kaggle.com/rtatman/did-it-rain-in-seattle-19482017
https://www.wunderground.com/hourly/us/wa/seattle/KSEA
https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND

4.2.2 Seattle Weather csv. This data set contains 1463 unique objects. Each is a day included in the range from January 1 2012 to December 31 2015. This data was found on Github:
The attributes for the objects include:

- Date (Numeric/Continuous)
  The unique day, month, year for each entry
- Precipitation (Numeric/Ratio)
  Amount of precipitation.
- $\text{Temp}_Max(Numeric/Ratio)$
  $The maximum temperature that day. Temp_Min(Numeric/Ratio)$
  $The minimum temperature that day.$
- Wind (Numeric/Ratio)
  The wind speed in mph.
- Weather (Nominal)
  Drizzle, Rain, Sun, or Snow

This data was found on Github: https://github.com/domoritz/maps/blob/master/dat
weather.csv

4.3 Seattle Traffic Data

If we have enough time left on the project, we may also be able to include traffic data within our analysis. To see how weather can influence the effects of people's travels within the city, and if that has a direct influence on when people utilize the library.

The attributes for the objects include:

- ID (Nominal)
  Unique ID for entry
- stname (Nominal)
  Name of the street
- countlocation (Nominal)
  Specific street location
- year (Ordinal)
  year that data was taken.

- aawdt (Numeric/Ratio)
  Annual Average Weekday Traffic

https://data.seattle.gov/resource/38vd-gytv.json

5 EVALUATION METHODS

How we evaluate and approach the data will obviously change as the project continues.

We will start however, with the aim to look into discovering interesting correlations between data based on time series analysis techniques. One of the biggest challenges we will face is trying to differentiate between when there is a significant event that has a correlation to another pattern, vs random noise.

I think it will be appropriate to measure how successful our project is based on how many interesting correlations we can find, and how we can support them with statistical significance. However we won't be able to figure out this amount, or if it's possible at all, without more work on the data set.

6 TOOLS

For this project we will use the standard libraries with Python and Jupyter Notebooks. Specifically, we plan to import and handle our data using Panda's DataFrames so we can quickly sort through and evaluate large portions of the data at a time.

We can then use Numpy's tools and methods to create any confusion matrices and arrays, while also letting us handle the large data sets without too much trouble.

Finally for Time Series, we can make use of the statsmodels API library to quickly get started in looking for statistically interesting events within the time series data, and finally we will use MatPlotLib to visualize.

7 MILESTONES

We plan to have the project completely by the Due date obviously, August 9th, 2019.

However we can set much more internal dates by then. Setting these milestones will allow us to also evaluate how our project is going so far, as well as help manage time.

Currently we plan to have a basic exploration and visualization of the data set done by the next submission, July 20th. This will include data integration, and analysis of trends within the library (basically leaving off where the last study left off).

The next section of the study will involve looking at how weather data and library data can influence one another, July 27th.

Finally, we will use the remainder of the time to address any additional questions or possible paths of exploration that might be interesting to pursue. This will include adding the Traffic Data to our analysis if it is possible.

This all aligns with our final due date less than a month out on August 9th.

7.1 Completed (As of July 25, 2019)

We are slightly behind our initial Milestone plans however we feel like we have achieved a substantial amount of work/gained a lot of knowledge so far into the project on dealing with real world data sets and integration. For our update, we will initially talk about this basic exploration in 8.2.

However, for seciont 9 and beyond we had to pivot to slightly to use different data sets that were more encompassing for our later explorations of the code, we realized that the original data sets did not provide enough information, or had to little unique data points to make many analytically conclusions outside of what could be observed. Specifically for Checkouts and Weather.

For these parts of the study, which are discussed after our initial explortation, we used the following data sets.

We have included a more in-depth data set for Library checkouts that includes hourly checkouts for everyday since 2005 April 13th to now. Located : https://data.seattle.gov/dataset/Checkouts-by-Title-Physical-Items-/3h5r-qv5w

We also have decided to transition to the NOAA Land based Climate Stations weather data set for Tacoma International Airport in Seattle. Which we can find here : https://www.ncdc.noaa.gov/cdo-web/datatools/lcd

The milestones we have been able to reach as of the current date of submission include the following.

- Basic exploration and visualization of the data set including data integration and an analysis of trends within the library. Essentially understanding and visualizing where the previous study left off.
- As well as a quick a quick analysis with the old data sets with whether or not it rained each day in Seattle and if that influenced checkout numbers.
- In depth analysis of rain and weather patterns within Seattle to be combined later.
- In depth analysis of checkout patterns and usage on the Seattle Public Library System as well.

7.2 To Be Completed

- Study weather data and library data in combination to further understand how weather can potentially affect library checkout data.
  For example:
  - do digital checkouts increase with rainy weather
  - do checkouts increase or decrease overall with weather fluctuations?
- We should also spend more time brainstorming potential questions we could answer using our additional expanded data sets now.
- Address additional questions and/or possible paths of exploration that pertain to the initial area of study. If time permits, we would like to add the Traffic Data to our analysis.

8 INITIAL DATA SET RESULTS AND EXPLORATION

8.1 Basic Exploration and Visualization of Data

Our first goal and area of work was to understand and visualize the Seattle Library Checkouts Data Set individually of the weather data. By doing so, we could begin to understand the typical trends and patterns present in the library checkouts of Seattle. And if we would need to rethink any of our approaches and datasets. We started by diving right in, and trying to find any pattern between rain and checkout numbers.

When analyzing checkout numbers, we began by following prior work here: https://towardsdatascience.com/how-and-when-people-use-the-public-library-1b102f58fd8a . We soon identified number of checkouts, checkout object type (Book, DVD, CD, VHS, Other), checkout time by day, checkout time by month, and checkout day of the week as points of interest to focus on. From each day, week, and month analysis; there will be 3 graphs. graph overall, graph by object type, and graph over all years. These 3 types of graph for each day, week, and month should provide excellent analysis over the checkout data. Another possible avenue to explore could be titles of each object type to determine a category (such as Fiction, Non-fiction, Mystery, Kids, Etc...). We could cross examine the category by time of year to see if peoples interest in Books, DVD's, CD's, or VHS change depending on the time of year (or even week).

8.2 Quick Analysis of Rain + Checkouts

We wanted to get a quick and brief look in to whether or not weather (more specifically, rain) drastically affected the number of checkouts in the Seattle library. Since the Seattle library checkouts data has millions of data points, to cut back on run time, we looked just at the years of 2016 and 2017, as they seem to have just around the same amount of checkouts each year (right around 6 million). We also used the "Did it Rain in Seattle?" data set that has data on whether or not there was rain in Seattle each day between the years 1948 to 2017. We pruned both the Seattle Library Checkouts as well as the "Did it Rain in Seattle?" data to include just the years 2016 and 2017. At first, we were looking to include 2011 to 2017, however, once calculating the run time for certain functions, we decided to look at a smaller data set. We understand that this is a very small amount of data and will provide no conclusive results, however, we still wanted to get a glimpse into whether or not the two interact at first glance.

The Seattle Library checkouts has two data sets provided - one that looks at both physical and digital checkouts (the one we used to get a quick glimpse) and one that looks at only physical checkouts (which we will use for further exploration). One caveat of the digital/physical data set is that the data for each checkout only includes a month and year for checkout, rather than a day, month, and year. Because of this, we decided to create a function to determine whether each month (January - December) in both 2017 and 2016 was **on average** rainy or not. To do so, we created a nested for loop that would look through the amount of rainy days that we counted in each month (in each year) and returned a boolean True/False value indicating whether or not each month was **on average** rainy or not.

Once we had data regarding whether or not each month was on average more or less rainy, we appended the boolean value on to the original data set in a new attribute labeled IsRainyMonth and created a new csv file ('checkoutWithRainy.csv') to be used to further look at the correlation between the data. We used various pandas functions to look at the data in various ways visually. We plan to work a little more on this as time permits, however, we have found that there seems to be no initial correlation between rain and library checkouts in this small data set. The graph below shows the months January through February (labeled 1-12) as well as the two years of 2016 and 2017. You can see that the only spike in checkouts with the presence of rain was during the months of January and February. The following graph shows the months January through December (labeled 1-12) as well as the years 2016 and 2017 (labeled 16 and 17). The only slight trend that we could pick out was the possible correlation with rain and checkouts in January and February in both years - though the results prove inconclusive at the current state.

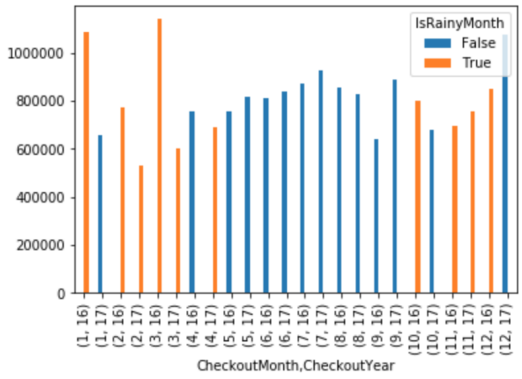

Figure 2: Rain and Checkout numbers

9 IN DEPTH LOOKS WITH BETTER DATA SETS

After our initial exploration we realized some things about our data sets, and had decided to switch our approach slightly, we realized some limits with the initial Data set provided by kaggle for library checkouts.

9.1 In depth look at Physical Checkout trends of Seattle Public Library

We found another data set that included hourly checkouts since 2005, so to gain a better understanding of the library data we decided we would split and partially focus on dealing with that data (See below). We also found that this data we could pull directly from the Seattle Government site, and as a result we would have much more up to date data. The new data is located here : https://data.seattle.gov/dataset/Checkouts-by-Title-Physical-Items-/3h5r-qv5w

Another data set utilized for looking at all the physical item checkouts from Seattle Public Library was found here: seattle-public-library-checkouts-kaggle" This data set begins with checkouts occurring in April of 2005 throughout September of 2017.

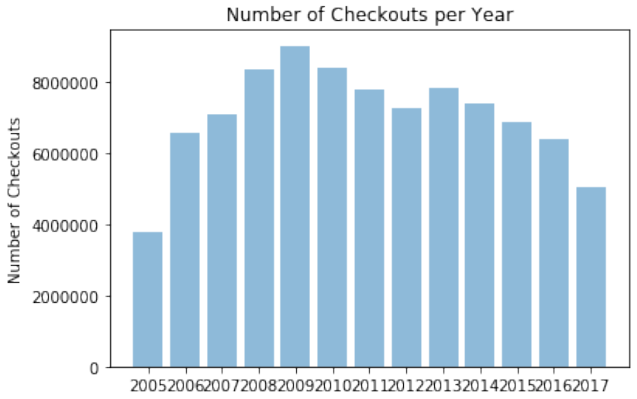

Figure 3: Number of Checkouts

In the graph above, it is important to note that this data set began in April of 2005 and ended in September of 2017. This explains the vastly reduced number of checkouts apparent in the years 2005 and 2017. We can see that the largest number of checkouts occurred in 2009. The number of checkouts is the total count of checkouts for any object type (Book, DVD, CD, VHS). We notice an increasing number of checkouts from 2005-2009. Afterwards, number of checkouts steadily declined yearly (except 2013; can possibly analyze special weather circumstances for the year 2013) Possibly a growing use of technology is responsible for the recent declines in the Seattle library checkouts. It should also be noted that the Seattle library recording this data has different hours based on different days of the week. Friday and Saturday have slightly reduced hours of operation while Sunday has greatly reduced hours of operation.

9.1.1 Data set graphs on time of day in creation... When looking at the time of day, it is crucial to understand that the library opens at 10am but closes at 8pm. It closes at 6pm on Friday and Saturday, while on Sunday it closes in range of (1pm-5pm). However, these library hours are only the CURRENT library hours. They may or may not have been different in the past (the graph could potentially point to this)

graph:Checkouts by hour overall

graph:Checkouts by hour by object type

graph:Checkouts by hour across all years

Data set graphs on time of week in creation... The library is open every single day of the week

graph:Checkouts by week overall

graph:Checkouts by week by object type

graph:Checkouts by week across all years

Data set graphs on time of month in creation...

It is important to understand the potential impact of the months containing a varying number of days (28-31). Each month also contains a different number of each day of the week, this could impact the total number of open hours for the library.

graph:Checkouts by month overall

graph:Checkouts by month by object type

graph:Checkouts by month across all years

9.2 In depth look at weather trends of Seattle

We also came to a similar conclusion involving weather data, we found it extremely hard to find accurate data from Wunderground since they no longer have an API, and web crawlers were taking too long, so we moved our focus to use NOAA's Local Climate Station data sets, specifically, the data set from Tacoma Airport, just south of the majority of the Seattle Public Libraries. This data set has 100% coverage of temperature and other various data points since 1944! (Although, we only used from 2005, the date the library data set goes back too). You can also find the data set here : https://www.ncdc.noaa.gov/cdo-web/datatools/lcd

Before we combine the results of the Library analysis, with weather data, we decided it would be helpful to understand some basic trends of the climate at Seattle. And if there has been any noticeable changes within the last 5 years.

First, we will look at how the temperature averages changed over the last few years to see if there is anything to be discovered there. Below is a graph of the Monthly average temperatures for the last few years. Made by re-sampling the "HourlyDryBulbTemperature" to monthly using Pandas.

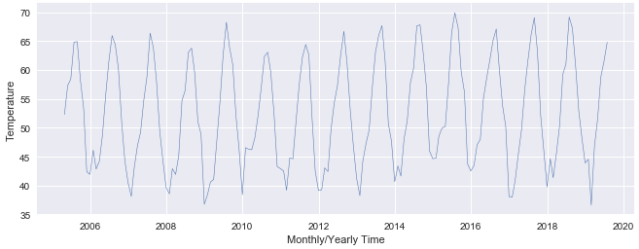

Figure 4: Average Monthly Temperature

As we can see from the graph (Figure 3) there appears to be a fairly expected trend, where December will report the coldest months, and July/June the hottest months. We can compare these two seasons to see if there is a noticeable change in checkouts in the next portion of this study.

Next, we can look at the rainfall/precipitation in more depth since 2005. To do this we will need to look at the hourly precipitation column. Originally values of "T" will indicate trace amounts, so we just replaced these values with 0.

Anyway, re-creating a similar graph to the temperature yields less intuitive results!

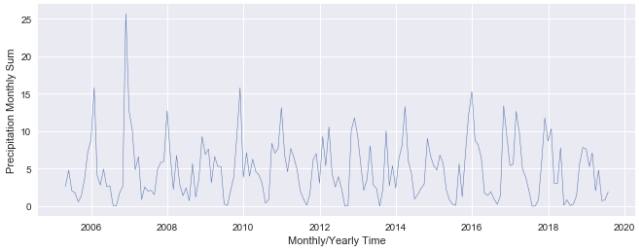

Figure 5: Monthly Sum of Precipitation

Observation of the graph shows that the months and amounts of rain vary a whole lot more than the temperature, which will lead to some more interesting analysis when we combine them with checkout numbers.

To get a better idea of the "rainy" Months, we broke it down into weeks, and plotted all the years against each other.

https://www.overleaf.com/project/5d373bb8f20ed216cac5c92b

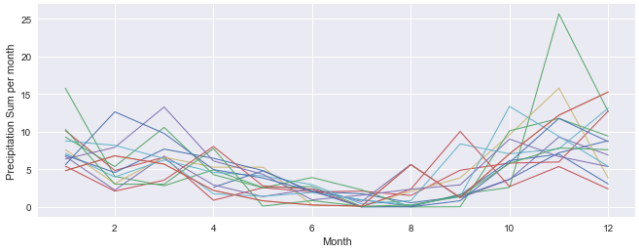

Figure 6: Month Sum of Precipitation

This graph is a little bit hard to interpret but we can see that the months with the least amount of precipitation are mostly during the summer, while November and January tend to receive substantially more. It is important to note that precipitation includes rain AND snow in this case.

We will have to include the different types of weather in future analysis for this project since it is likely to play an important role in determining attendance to Libraries.