Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Tony Talarico
February 2020

Ortega Branch Library, San Francisco, California

Which neighborhood of San Francisco has the greatest need for a multipurpose visual and performing arts venue?

Introduction

Public libraries are the beating heart of our neighborhoods. Children congregate at the library after school to do their homework. Government officials host town hall meetings at libraries to promote discussion of current events. Senior citizens meet outside libraries to exercise, to dance, and to visit with their neighbors. In addition to its main branch located next to City Hall, the city of San Francisco, California has 27 branch libraries that serve all corners of its 49 square mile expanse. While all of the branch libraries are located near schools and public parks, not all of the libraries are located near other cultural attractions, such as theaters, concert halls, or museums.

Some neighborhoods would benefit significantly if there was a multi-purpose event space in close proximity to the branch library, in the sense that:

- Branch libraries which currently have limited meeting space would be able to hold larger events in the new venue,
- Local artists and performing groups would have a new venue to exhibit their arts in close proximity to the library,
- Local residents would have a new option for cultural enrichment, and
- The library and the artists would each gain greater exposure by having access a common target audience.

A community service organization is interested in opening a multipurpose visual and performance arts venue near one of the branch libraries in San Francisco, and they have asked me as a data scientist to give them insight which neighborhoods would reap the greatest benefit from this event space. Specifically, they have asked me to answer the following questions:

1. Which of our branch libraries in San Francisco are used the most heavily?
2. Are our branch libraries used most by school-aged juveniles, adults, or senior citizens?
3. Of the heavily used branch libraries, which are located in areas which do not have museums, theaters or other cultural offerings nearby?
4. Where should the new multipurpose event space be located?

To answer these questions, I will follow the data science methodology, analyze a library usage dataset, complement that dataset with location data from Foursquare, and prepare a presentation of my findings to the community.

Data

This study will include the following datasets and sources of data:

- Library Usage dataset from the City Government of San Francisco (https://data.sfgov.org/Culture-and-Recreation/Library-Usage/qzz6-2jup) This dataset consists of 1062 rows of usage data by library patrons, items borrowed or renewed from the library catalog of inventoried resources. User information is anonymous, and specific resources borrowed or renewed are not identified. The dataset indicates the type of patron, the branch library of the patron, the age range of the patron, the total inception to date checkouts and renewals for that patron, and the supervisory district where the patron resides.
- Branch library addresses from the San Francisco Public Library website (https://sfpl.org/locations/#!/filters?sort_by=weight&sort_order=ASC&page=0) The branch libraries identified in the library usage dataset and the address data will enable us to capture latitude and longitude data for each branch library.
- Foursquare location data (https://foursquare.com) Location data from Foursquare will be used to identify venues and activities in the areas surrounding each branch library, in order to gain insight about which branch libraries already have cultural offerings in the vicinity of the branch library and which libraries do not have as many cultural offerings located near the branch library.

Methodology

I followed the following methodology during the course of this project:

Is the data adequate to make the project be feasible?

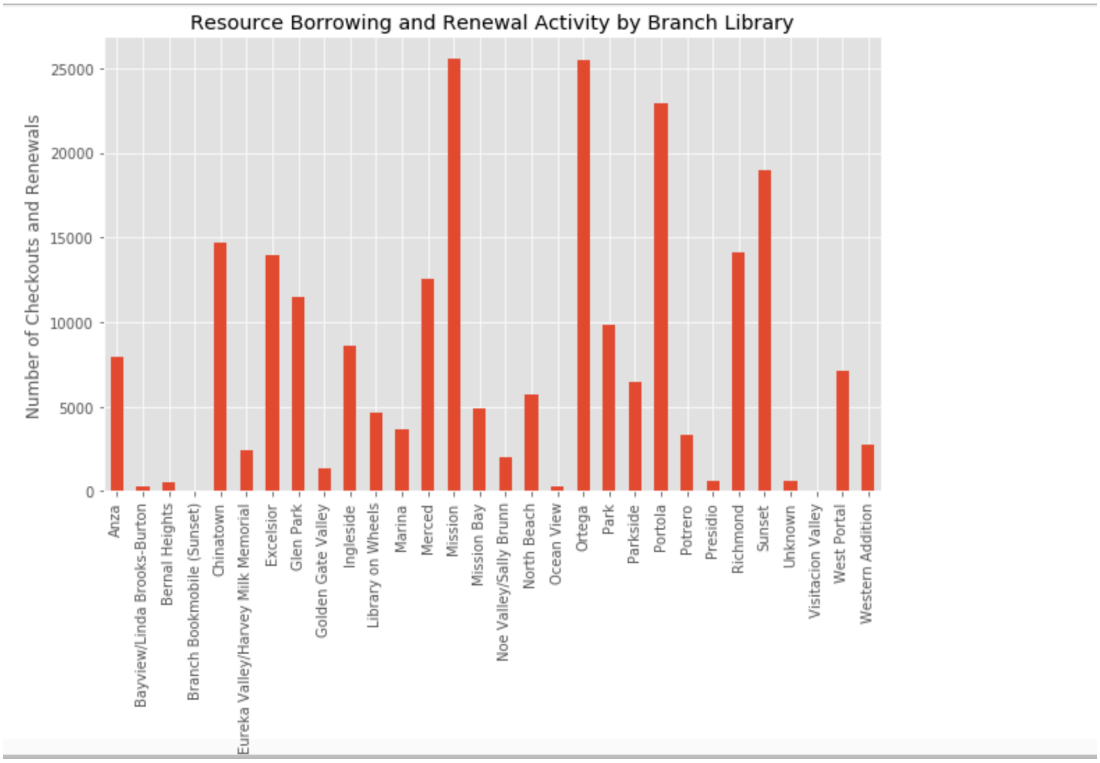
Before undertaking any analysis of the library data, I performed a preliminary examination of the source data to make sure that I could upload it into my development environment. I confirmed that I could get geo coordinates for the library addresses, and I confirmed that I could use the coordinates in Foursquare to explore the neighborhoods surrounding each library. All of this preliminary work was completed in week 1 of the project, before I submitted the initial GitHub notebooks that included the Introduction and Data sections. Completing this preliminary work enabled me to proceed into Week 2 of the project, where I followed the methodology outlined herein to perform the data analysis.

Is the data in the proper format to enable analysis?

After uploading raw data, what data wrangling did I do in order to analyze the data? Initially I tried to use a csv file provided by the City of San Francisco government website, but I was unable to read this file after it was uploaded. It appears that there were many http labels and tags that were invisible to me upon viewing the file. Instead of using a csv file, I accessed a file of json data directly from the city government website. After accessing the json file, I read the json data and loaded it into a pandas dataframe. The made columns that I required for the project, checkout and renewal data, had to be converted from string type to integer data type so that analysis could be performed. Since I was most interested in data about the branch libraries, I excluded data from the Main Library from my analysis. (The Main Library is located downtown near the city's major cultural offerings, and it has many multipurpose event spaces, so it did not make sense to consider the Main Library as a candidate for a new adjoining multipurpose event space.). I was most interested in focusing on the usage at the Branch libraries, so after excluding data from the Main Library, I was able to analyze branch library usage.

What did the data show?

From my dataset, I was able to identify the branch libraries which have the heaviest usage, in terms of the total number of checkouts and renewals made by patrons who identified that branch as their home-library-definition.



I reviewed this data and made sure that there were no outliers or anomalies, and focused the rest of my analysis on the top ten branches by total usage.

The age range of library patrons was also of interest to me during this study, and so I grouped the data by patron-type in order to see a relative breakdown of customers that use the various branch libraries. I highlighted the branches which had high use by juveniles and by senior citizens, and these findings factored into the conclusions.

As part of the data analysis, I prepared many visualizations of the data, and all of those are provided in this notebook. The bar charts and folium map were the most helpful visualization tools for this project.

How did the location data complement the library usage data?

Once I had completed the analysis of the library usage data, I used Foursquare location data to provide insight about the neighborhoods surrounding each branch library. I was looking for libraries which did not have many other cultural offerings in their vicinity.

I defined "cultural offerings" as any of the following categories:

- Art gallery
- Concert Hall
- Opera House
- Dance Studio
- Performing Arts Venue
- Art Museum
- Theater
- Music School
- Event Space

The Foursquare analysis required several iterations, to find the proper mix of parameters to use in the Foursquare queries. A vicinity (referred to as a 'range' in Foursquare) of 200 meters was too small to produce meaningful results, while a vicinity of 1000 meters was too large. After several rounds of experimentation, I settled on a range of 750 meters. Similarly, how many surrounding venues was I interested in considering? A limit of ten items was too small, 500 items seemed to resource-intensive, and after a few rounds of trial I settled on a limit of 100.)

During the study, I encountered challenges using Nominatim and Geolocator to access coordinate data. I noticed that my functions were returning incorrect coordinates in some cases. This may be due to lack of precision in my address data, although the coordinates were correctly identified at the Nominatim website. More inconsistencies were observed when I plotted the first folium map, I noticed that the marker for the Ortega library was in the incorrect location.

In these examples, I chose to override the latitude and longitude of Ortega library and Mission library with coordinates that I knew to be accurate. The function 'get coordinates' above worked better for me than geolocator.geocode(address) After my overrides, the markers appeared in the correct spot. All of the markers for other branch libraries were in the correct spots when plotted in Folium.

After confirming the correct coordinates, the Foursquare queries returned 561 items, when the range was set to 500 meters, and when I set the range to 750 meters, the queries returned 750 rows for further consideration. I loaded the foursquare data into a pandas dataframe, filtered to include only the "cultural offerings" defined above, and this enabled me to present the findings that were the primary objective of this study.

With the filtered dataframe that focused on cultural offerings, I could look at summary views of cultural offerings for each branch library and combine that with the library usage analysis to make my conclusion and recommendations to the project stakeholders.

Results

The analysis of library usage data indicated that:

1. Portola Branch Library was the third most heavily used library branch in the San Francisco Public Library system, 2. Richmond Branch Library was the sixth most heavily used branch,
3. Excelsior was the seventh most heavily used library branch, but among this group of four Excelsior had the highest usage by juvenile patrons. and
4. Glen Park was the ninth most heavily used library branch, but among this group of four Glen Park had the highest usage by senior citizen patrons.

Foursquare location analysis indicated that:

1. Excelsior Branch had no cultural offerings within 750 meters of the library, and
2. Portola, Richmond and Glen Park branch libraries had only one cultural offerings within 750 meters of the library.

Findings from Library Usage Dataset Analysis

The analysis of the library usage dataset enabled me to group and rank the branch libraries based on their checkout and renewal data from the dataset, to identify the ten branch libraries with the highest borrowing and renewal activity. They will be good candidates for us to look at Foursquare data, to see if there are other cultural venues surrounding these libraries.

Analysis of Age-range usage

When I analyzed the library usage dataset for insight about the ages of library patrons in the branch libraries, I found that the 'patron-type-definition' field was more insightful than the 'age-range' field.

Findings from Foursquare Analysis

Some branch libraries have many cultural offerings nearby, but the branch libraries which have the fewest cultural offerings nearby are

- Excelsior, which shows no nearby venues
- Glen Park, Portola and Richmond, which each have one nearby venue
- With 15 offerings returned, Mission initially appeared to be an outlier. I reviewed the coordinates of Mission Branch library and I reran the entire notebook using correct coordinates for Mission Branch, and this time 7 offerings were returned. Although this branch is close to City Hall, I am still surprised about the number of venues found near this branch.

Challenges with Nominatim and Geolocator

In some cases, I noticed that my functions were returning incorrect coordinates. This may be due to the lack of precision in the library addresses. In these examples, I chose to override the latitude and longitude of Ortega library with coordinates that I knew to be accurate. The function 'get coordinates' above worked better for me than geolocator.geocode(address)

More inconsistencies were observed when I plotted the first folium map, I noticed that the marker for the Ortega library was in the incorrect location. After my override it is now in the correct spot. All of the other markers appear to be in the correct spots for their branch libraries.

Discussion

Each branch library in San Francisco would benefit from having more cultural venues in its vicinity. But at this time the project stakeholders only have the budget to create one new multipurpose event space. So while my analysis identifies four branch libraries that would reap the greatest benefit, the stakeholders ask me to select one branch, based on the data that I have reviewed in this study.

If I were to continue this study, I would try to expand the library usage dataset, incorporate more loops to make my code more efficient, and build a choropleth plot that includes geodata and usage data for all branches in the city. I would also consider incorporating the crime statistics data that we reviewed earlier in the course, to see if the crime data would affect the analysis or the conclusions presented in this report.

Conclusion

The analysis of library usage data indicated that:

1. Portola Branch Library was the third most heavily used library branch in the San Francisco Public Library system,
2. Richmond Branch Library was the sixth most heavily used branch,
3. Excelsior was the seventh most heavily used library branch, but among this group of four Excelsior had the highest usage by juvenile patrons. and
4. Glen Park was the ninth most heavily used library branch, but among this group of four Glen Park had the highest usage by senior citizen patrons.

Foursquare location analysis indicated that:

1. Excelsior Branch had no cultural offerings within 750 meters of the library, and
2. Portola, Richmond and Glen Park branch libraries had only one cultural offerings within 750 meters of the library.

The new multipurpose event space should be located near Excelsior Branch

While any of these four branches is worthy and deserving of having a new multipurpose event space locating nearby, I recommend that Excelsior Branch be the library where the new event space should be located. Among all of the branch library options across San Francisco, adding an event space near Excelsior Branch will:

- Enable the Excelsior Library to hold larger events in the new venue,
- Enable local artists and performing groups to exhibit their arts in close proximity to Excelsior Library,
- Provide local residents with a new option for cultural enrichment, and
- Give Excelsior Library and the local artists a new space to attract a common target audience.