1. Introduction

1.1 Background

Some of the Core Values of the American Library Association are Diversity, Access, and the Public Good. In other words, the library profession's ideal goal is to provide equal access to public library services for any community regardless of that community's socioeconomic conditions. Libraries offer many free resources which may improve opportunity for underprivileged communities, such as access to computers and internet, job-seeking tools, and classes for developing new skills, so it is important that all communities should have fair and equal access to these services.

1.2 Research Question

In this report, I will examine the geographic distribution of libraries in New York City to determine how well the goal of equal access is actually being met. Specifically, I will explore whether libraries are distributed evenly among districts of higher and lower socioeconomic privilege, and if not, what variables lead to the presence of more libraries in a district.

1.3 Interest

This issue is of potential interest not only to librarians but also to local governments and civic planners for deciding how to allocate funding where it is most needed, since the research question can help determine if any communities have a gap in the availability of library access, which would imply that those communities lack the development resources that public libraries offer.

2. Methodology

2.1 Data Collection

The United States Census committee has divided New York City into a set of official sub-boroughs for data collection and analysis, called Public Use Microdata Areas (PUMAs). For ease of analyzing census data, I define "communities in New York City" in my research question to be the set of 55 PUMAs, as shown in Figure 1.

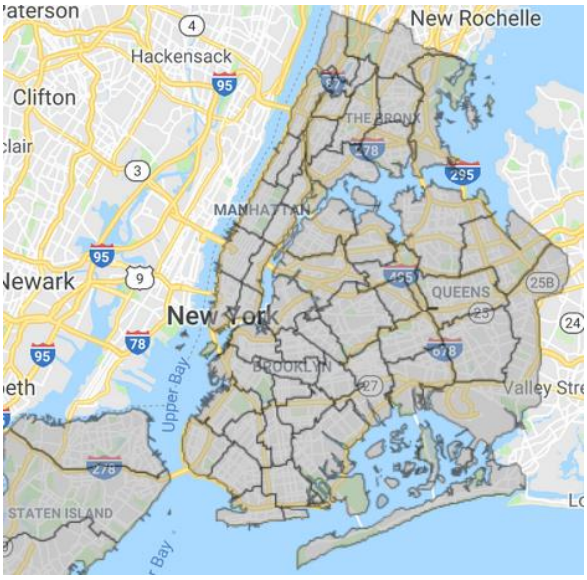

Figure 1: The PUMA boundaries superimposed on a Google Map, screenshot from NYC Open Data

The indicators I will use to measure socioeconomic status and other demographic characteristics are several of the fields captured by the United States Census for each PUMA and available for download in CSV format from Coredata.NYC under the "Demographics" menu option. I chose population size as a potential variable that might affect the number of libraries in an area, along with several additional variables related to populations that have potentially been underserved by public libraries (immigrants, minority ethnicities, people with disabilities, older adults, and people with low income.) The list of variables I selected from the Coredata.NYC data repository is as follows:

- Population size
- Median household income
- Percent of the population with a disability
- Percent of the population who is white
- Percent of the population below the poverty line
- Percent of the population born outside the United States
- Percent of the population over the age of 65

Each indicator can be selected individually from the Coredata.NYC page and downloaded as a spreadsheet in which each row represents a neighborhood, and the columns are the years from 2000 to 2017. I downloaded the 7 individual spreadsheets for each variable above and then merged them into a single new table using Excel: I pasted the list of neighborhood names as the first column and created subsequent columns that would hold the 2017 column from each of the variables' data sets (for each set, I used the neighborhood name as the search key in the VLOOKUP function to retrieve the corresponding value from that variable's 2017 column.)

Lastly, as a primary key to be used later with geodata, I added a column for the numerical ID of each PUMA as defined by the official Census listing. (This column had to be entered manually by means of comparison between this PUMA ID map and Coredata.NYC's equivalent map showing their versions of the PUMA names.)

The resulting spreadsheet is a single table where each neighborhood is a row, with the columns holding the 2017 values of each of the socioeconomic indicators for that neighborhood:

	A	B	C	D	E	F	G	H	I	J
1	PUMA_Name	Population	Median_income	Disability	White	Poverty_rate	Foreign_born	Over_65	Puma_ID	
2	Astoria	164321	67647.98474	0.0456564	0.474918	0.130441	0.367713	0.125206	4101	
3	Bay Ridge	123488	69988.7913	0.0656238	0.508584	0.151506	0.387997	0.171021	4013	
4	Bayside/Little Neck	118670	71492.9404	0.0454221	0.375074	0.0975641	0.445698	0.227218	4104	
5	Bedford Stuyvesant	142027	52896.92904	0.087774	0.266111	0.243765	0.184127	0.103304	4003	
6	Bensonhurst	205850	54513.17598	0.0615492	0.404469	0.17261	0.556259	0.142521	4017	
7	Borough Park	146556	46229.14615	0.060802	0.688945	0.308085	0.279279	0.144812	4014	
8	Brooklyn Heights/Fort Greene	135444	94327.26889	0.0873168	0.472143	0.148676	0.187694	0.107292	4004	
9	Brownsville/Ocean Hill	111511	20640.26819	0.114748	0.0425788	0.398762	0.261562	0.121997	4007	
10	Bushwick	140474	51622.07097	0.069588	0.214844	0.271181	0.302996	0.074149	4002	
11	Central Harlem	147442	49994.61425	0.087135	0.14873	0.235349	0.251672	0.113753	3803	
12	Chelsea/Clinton/Midtown	152455	103925.9006	0.0631047	0.593375	0.137894	0.324883	0.134518	3807	
13	Coney Island	122009	36806.81379	0.133358	0.554268533	0.248412	0.524511	0.236605	4018	
14	East Flatbush	140087	50290.1449	0.0665752	0.0270261	0.136625	0.510062	0.159658	4010	
15	East Harlem	128316	37471.24821	0.135585	0.118855	0.315459	0.255245	0.127981	3804	
16	East New York/Starrett City	176471	37487.55335	0.0734747	0.0389866	0.246757	0.315491	0.111174	4008	
17	Elmhurst/Corona	146301	52983.5501	0.0732915	0.0545724	0.155565	0.639442	0.115919	4107	
18	Flatbush	150707	57678.41114	0.0710255	0.441658	0.168204	0.394753	0.134572	4015	
19	Flatlands/Canarsie	215637	78108.75066	0.052042	0.223672	0.0889643	0.40014	0.153156	4009	
20	Flushing/Whitestone	260282	52262.04768	0.0411274	0.232632	0.161814	0.587294	0.196014	4103	
21	Greenwich Village/Financial District	148982	147640.9981	0.0334356	0.722081	0.0884715	0.228994	0.14291	3810	
22	Highbridge/South Concourse	149710	31489.30024	0.139932	0.0131454	0.364129	0.437967	0.0848908	3708	
23	Hillcrest/Fresh Meadows	164291	65225.65248	0.0569969	0.291684	0.123256	0.449203	0.149059	4106	
24	Jackson Heights	170222	57680.44928	0.070717	0.122928	0.135181	0.60081	0.131258	4102	
25	Jamaica	249331	62846.12122	0.0769785	0.0144467	0.112985	0.45517	0.134331	4112	

Figure 2: a screenshot of my working spreadsheet

In order to determine how many libraries are in each PUMA in the above table, I will use the Foursquare API for searching. Since "Library" is an existing category type in Foursquare's venue results, I will use "Library" as my query string for a search of a radius of 2500 meters centered around a latitude and longitude that represents the approximate center of each PUMA. The value of 2500 meters was selected by examining this map of the PUMAs by eye and determining that their typical size and shape is approximately a square of side length 5 km, so a search radius of half that amount should approximately cover the PUMA area.

For the Foursquare search, I need the latitude and longitude values that represent the approximate center of each PUMA. The official PUMA boundaries are defined in a GeoJSON file here, and each entry in the file consists of a long list of latitude/longitude points defining the lines that make up its boundary. The portion of the file that I will use is the Properties section of each PUMA, which contains several useful fields that I can scrape:

▼ properties: {} 4 keys
  puma: "3701"
  shape_leng: 53287.1770536
  shape_area: 97932776.767
  ▼ bbox: [] 4 items
    0: -73.92490327473082
    1: 40.86606885121164
    2: -73.8856751802567
    3: 40.91553277700519

Figure 3: The 'Properties' entry for the first PUMA in the GeoJSON file

This Properties dictionary allows me to pull out the PUMA ID to use as a primary key with my census data spreadsheet, and the 'bbox' list enables me to calculate the coordinates of the approximate PUMA centers. The bbox, or bounding box, gives the minimum and maximum latitude and longitude in the geographical shape of the PUMA, so I will approximate the center coordinates by taking the means of these pairs of extrema. That is, the center latitude is the mean of the two bbox latitudes, and the center longitude is the mean of the two bbox longitude.

I tested the accuracy of this method by plotting a preliminary Folium map of the calculated center points (shown in red) alongside a subset of the actual PUMA boundary lines (shown in black; please note that the boundaries along bodies of water are implicitly assumed and therefore not plotted). The plot indicates that most of the calculated center points do fall approximately in the middle of their boundary lines, so I will use this set of centroids for my Foursquare searching.

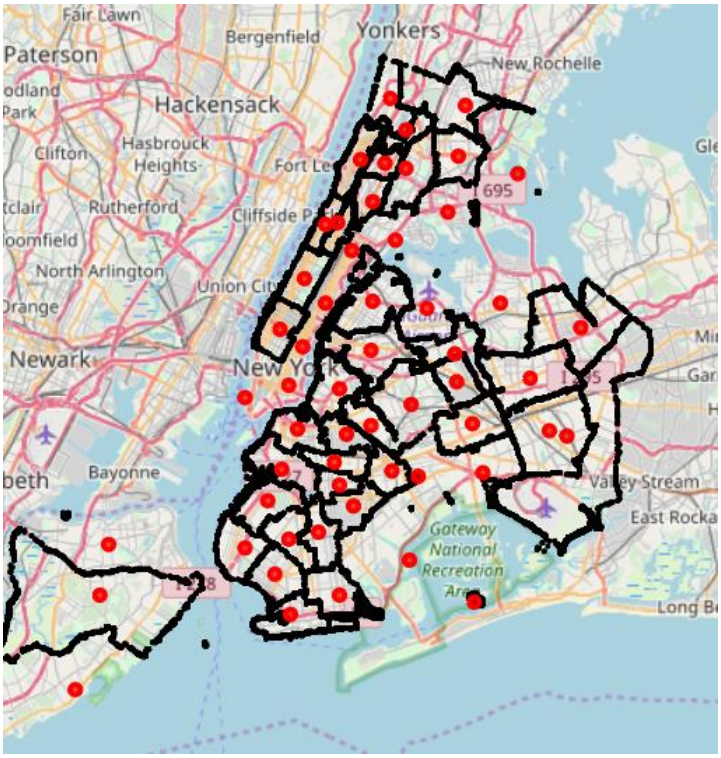

Figure 4: A preliminary Folium plot of the calculated center points (red) from the bbox values.

2.2 Next Steps

In Week 5, I will proceed to collect and analyze the library data using the following steps:

1. Loop through the list of PUMAs and conduct a Foursquare search centered at the calculated PUMA center coordinates. This search will give me:
    - A count of how many libraries are in each PUMA, and
    - A dataframe of all libraries found by Foursquare in New York city
2. Use k-means clustering to cluster the set of libraries on their similarity based the socioeconomic data associated with the PUMA each library is located in
3. Use linear regression analysis to determine if there is any correlation, and if so, how significant, between each socioeconomic variable (e.g., median income, percent white), and the number of libraries present in a PUMA.