GEOGRAPHIC DISTRIBUTION OF LIBRARIES IN NEW YORK CITY

Rachel Dols

Table of Contents

Abstract

This study investigates the locations of libraries in New York City and seeks to determine whether libraries are fairly distributed across the various communities, or whether the number of libraries in a community is related to socioeconomic factors like the wealth or ethnic makeup of that community. The study finds that most socioeconomic factors have no effect on the number of libraries in a community; however, the number of libraries is moderately correlated with the median income of a community, and moderately negatively correlated with the immigrant percentage of the population in that community. That is, communities with higher median income tend to have slightly more libraries than communities with lower median income, and communities with larger immigrant populations tend to have slightly fewer libraries than communities with smaller immigrant populations.

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

2.1.1 Socioeconomic Data

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.

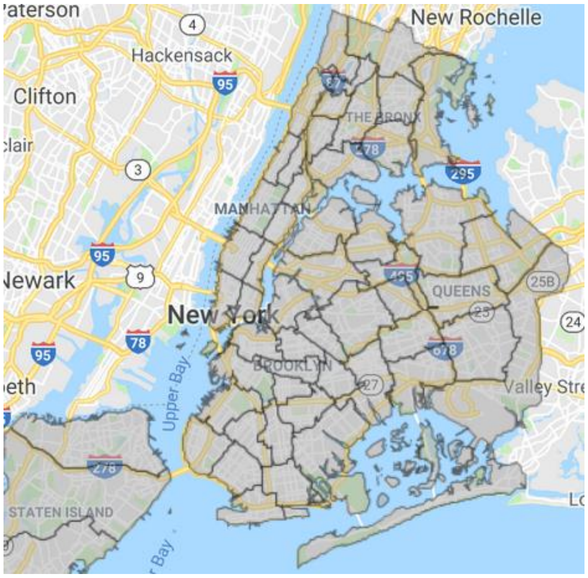


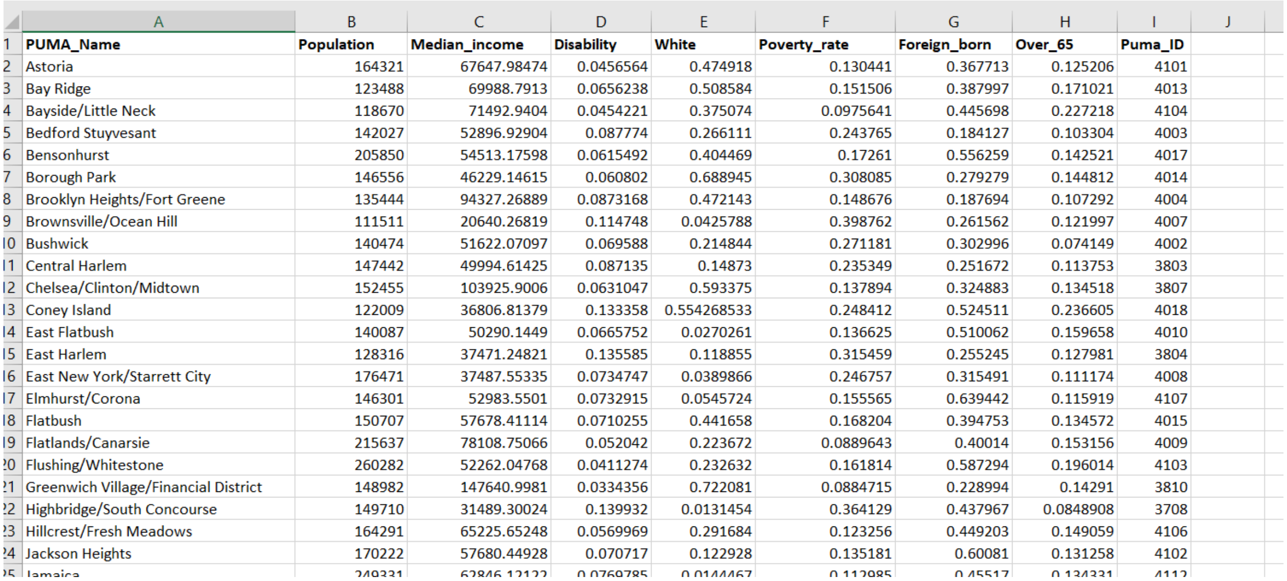
Figure 1: The PUMA boundaries superimposed on a Google Map, screenshot from NYC Open Data

The indicators I am using 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. I uploaded that spreadsheet to my IBM Watson project so that I could work with it as a Pandas dataframe.

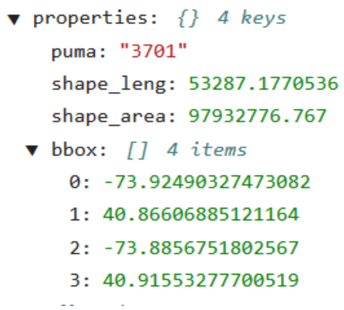


	A	B	C	D	E	F	G	H	I	J
1	**PUMA_Name**	**Population**	**Median_income**	**Disability**	**White**	**Poverty_rate**	**Foreign_born**	**Over_65**	**Puma_ID**	
2	Astoria	164321	67647.98474	0.0456564	0.474918	0.130441	0.367713	0.125206	4101	
3	Bay Ridge	123488	69988.7913	0.0656238	0.508584	0.151506	0.387997	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

2.1.2 Geographic Data

In order to carry out the Foursquare search that I would be doing in the next step, I needed to find the latitude/longitude coordinates 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, and the Properties portion of each venue contains the main information I needed for my data collection:



```
▼ 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 approximated the center coordinates of each PUMA 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 proceeded to use this set of centroids for my Foursquare searching.

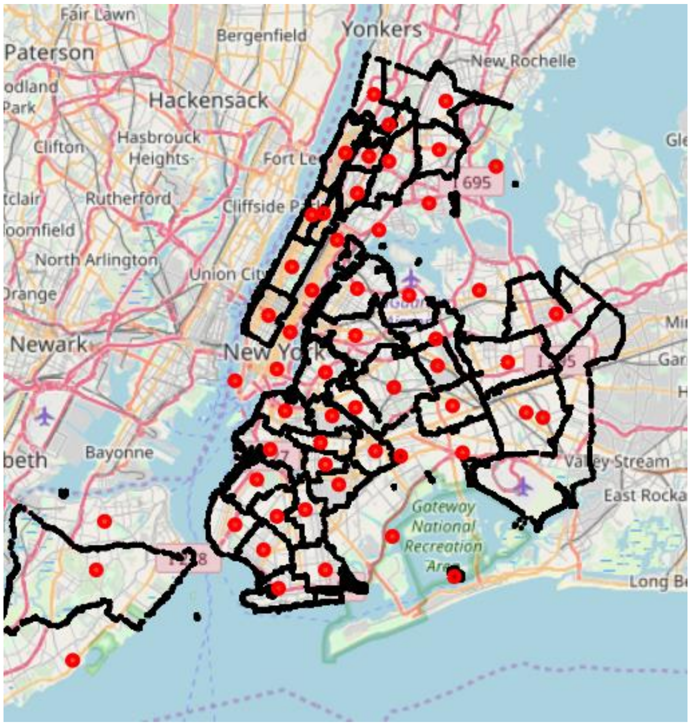
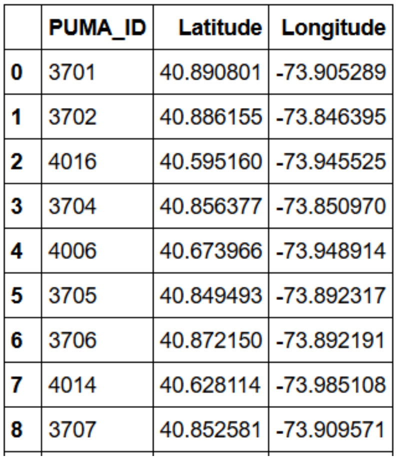


Figure 4: A preliminary Folium plot of the calculated center points (red) from the bbox values

I scraped the contents of the JSON file into a new dataframe to hold the PUMA ID, the averaged bbox latitude, and the averaged bbox longitude.



	PUMA_ID	Latitude	Longitude
0	3701	40.890801	-73.905289
1	3702	40.886155	-73.846395
2	4016	40.595160	-73.945525
3	3704	40.856377	-73.850970
4	4006	40.673966	-73.948914
5	3705	40.849493	-73.892317
6	3706	40.872150	-73.892191
7	4014	40.628114	-73.985108
8	3707	40.852581	-73.909571

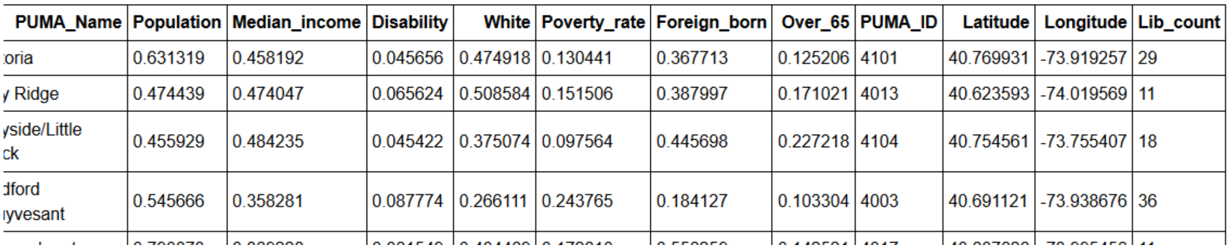
Figure 5: The first few rows of the scraped and calculated JSON data

I then joined the socioeconomic dataframe with this geographic dataframe using PUMA ID as the primary key and using a left join. That gave me a merged dataframe with the socioeconomic data for each PUMA as well as the latitude/longitude coordinates of its center.

2.1.3 Finding the Libraries

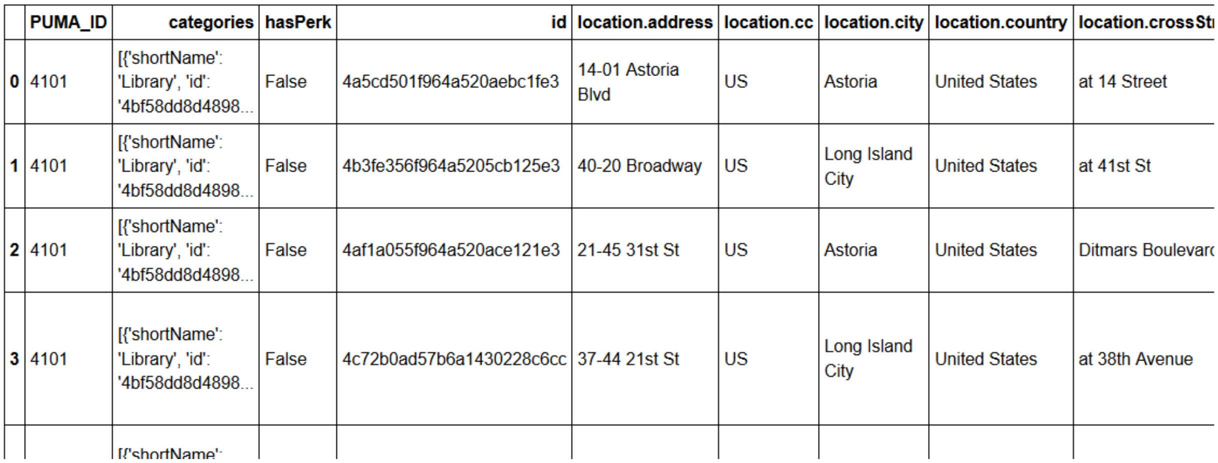
In order to determine how many libraries are in each PUMA, I used the Foursquare API to explore venues around the center of each PUMA. Since "Library" is an existing category type in Foursquare's venue results, I used "Library" as my query string for a search of a radius of 2500 meters centered around the central latitude/longitude 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.

I then looped through the PUMA dataframe and performed the Foursquare search for each PUMA's pair of latitude/longitude coordinates, and I computed the length of the results list as the number of libraries in that PUMA. I also created a new dataframe to hold the Foursquare venue results, that is, all the libraries found in the venue search.



PUMA_Name	Population	Median_income	Disability	White	Poverty_rate	Foreign_born	Over_65	PUMA_ID	Latitude	Longitude	Lib_count
oria	0.631319	0.458192	0.045656	0.474918	0.130441	0.367713	0.125206	4101	40.769931	-73.919257	29
y Ridge	0.474439	0.474047	0.065624	0.508584	0.151506	0.387997	0.171021	4013	40.623593	-74.019569	11
yside/Little ck	0.455929	0.484235	0.045422	0.375074	0.097564	0.445698	0.227218	4104	40.754561	-73.755407	18
dford yvesant	0.545666	0.358281	0.087774	0.266111	0.243765	0.184127	0.103304	4003	40.691121	-73.938676	36

Figure 6: The count of libraries in each PUMA was appended as the final column, Lib_count.



	PUMA_ID	categories	hasPerk	id	location.address	location.cc	location.city	location.country	location.crossSt
0	4101	[{'shortName': 'Library', 'id': '4bf58dd8d4898...	False	4a5cd501f964a520aebc1fe3	14-01 Astoria Blvd	US	Astoria	United States	at 14 Street
1	4101	[{'shortName': 'Library', 'id': '4bf58dd8d4898...	False	4b3fe356f964a5205cb125e3	40-20 Broadway	US	Long Island City	United States	at 41st St
2	4101	[{'shortName': 'Library', 'id': '4bf58dd8d4898...	False	4af1a055f964a520ace121e3	21-45 31st St	US	Astoria	United States	Ditmars Boulevar
3	4101	[{'shortName': 'Library', 'id': '4bf58dd8d4898...	False	4c72b0ad57b6a1430228c6cc	37-44 21st St	US	Long Island City	United States	at 38th Avenue
		[{'shortName':							

Figure 7: The table of libraries, scraped from the Foursquare results

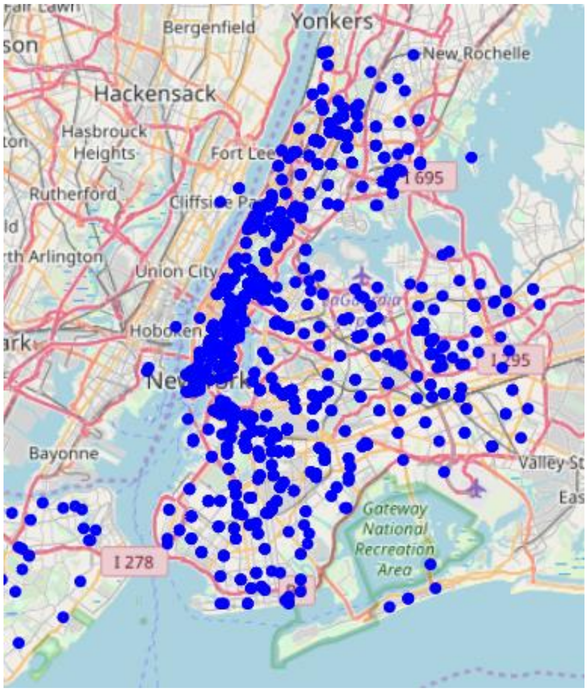
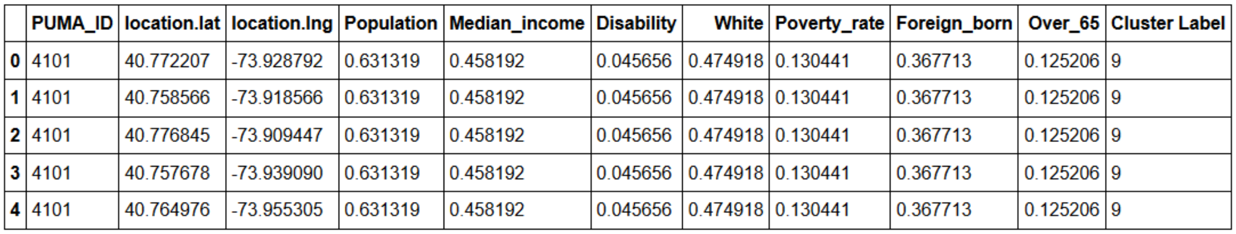


Figure 8: A map of all the libraries retrieved by the Foursquare search

3. Results

To visualize the distribution of libraries by PUMA, I used k-means clustering on the set of libraries, using 55 clusters because there are 55 PUMAs. To ensure that libraries from the same PUMA would be grouped into the same cluster, I created a new dataframe that merged the library list with the socioeconomic dataframe, but in the new dataframe I kept only the socioeconomic data columns and the column holding the library's location coordinates. That way, the data rows for each library in the same PUMA would be identical in all columns except for the libraries' unique latitudes and longitudes, thus giving a good chance that the k-means clustering algorithm would consider libraries in the same PUMA to be more similar than libraries from different PUMAs and would group libraries from the same PUMA together.



	PUMA_ID	location.lat	location.lng	Population	Median_income	Disability	White	Poverty_rate	Foreign_born	Over_65	Cluster Label
0	4101	40.772207	-73.928792	0.631319	0.458192	0.045656	0.474918	0.130441	0.367713	0.125206	9
1	4101	40.758566	-73.918566	0.631319	0.458192	0.045656	0.474918	0.130441	0.367713	0.125206	9
2	4101	40.776845	-73.909447	0.631319	0.458192	0.045656	0.474918	0.130441	0.367713	0.125206	9
3	4101	40.757678	-73.939090	0.631319	0.458192	0.045656	0.474918	0.130441	0.367713	0.125206	9
4	4101	40.764976	-73.955305	0.631319	0.458192	0.045656	0.474918	0.130441	0.367713	0.125206	9

Figure 9: The dataframe used for clustering, with the cluster label appended to the end. Note that for a single PUMA, all columns hold identical data except for the individual libraries' latitude and longitude.

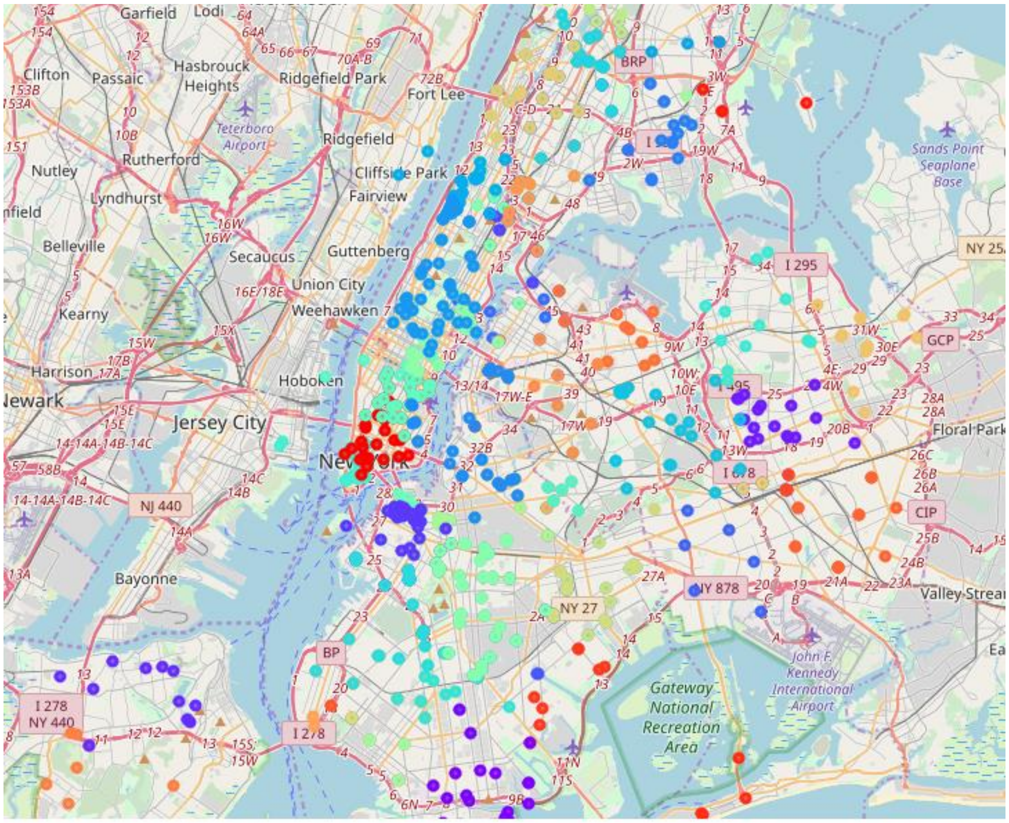
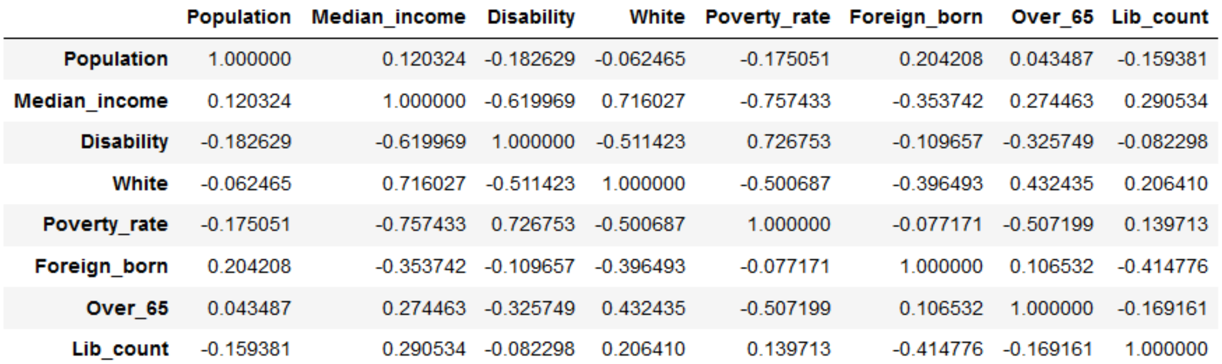


Figure 10: A map of the libraries colored by cluster (and therefore colored by PUMA)

The visualization in Figure 10 indicates that some clusters (i.e., some PUMAs) have more libraries or a denser distribution of libraries, and others have very few libraries that are sparsely distributed. This leads again to the original research question: is there any correlation between the socioeconomic privilege of the PUMAs and the number of libraries they have?

To answer that question, I examined the Pearson correlation coefficients between each of the socioeconomic variables and the library count, which are shown in the following table:



	Population	Median_income	Disability	White	Poverty_rate	Foreign_born	Over_65	Lib_count
Population	1.000000	0.120324	-0.182629	-0.062465	-0.175051	0.204208	0.043487	-0.159381
Median_income	0.120324	1.000000	-0.619969	0.716027	-0.757433	-0.353742	0.274463	0.290534
Disability	-0.182629	-0.619969	1.000000	-0.511423	0.726753	-0.109657	-0.325749	-0.082298
White	-0.062465	0.716027	-0.511423	1.000000	-0.500687	-0.396493	0.432435	0.206410
Poverty_rate	-0.175051	-0.757433	0.726753	-0.500687	1.000000	-0.077171	-0.507199	0.139713
Foreign_born	0.204208	-0.353742	-0.109657	-0.396493	-0.077171	1.000000	0.106532	-0.414776
Over_65	0.043487	0.274463	-0.325749	0.432435	-0.507199	0.106532	1.000000	-0.169161
Lib_count	-0.159381	0.290534	-0.082298	0.206410	0.139713	-0.414776	-0.169161	1.000000

Table 1: The Pearson correlations between variables. Note the bottom row.

The bottom row is the row of interest for this study: the correlation coefficients between the library count and each of the other variables. Note that all the values in this row have fairly small magnitudes, less than 0.5, which means there are no variables here that have a strong correlation with library count.

For each variable, I also calculated the p-value to determine the confidence of the Pearson coefficient (see Table 2) and used a regression plot to visualize the fit of the data (see Figure 11).

Variable	Pearson Coeff	P-Value	Correlation size	Certainty
Median Income	0.2905	0.0314	Small-medium	Moderate
Disability	-0.0823	0.5503	Very small	Very Weak
White	0.2064	0.1305	Small	Weak
Poverty Rate	0.1397	0.3090	Small	Very Weak
Foreign-Born	-0.4148	0.00164	Medium	Strong
Over 65	-0.1692	0.2170	Small	Weak

Table 2: The Pearson correlations between variables, the p-value, and the qualitative descriptions of the correlations and confidence levels.

The plots in Figure 11 show why the Pearson coefficients in Table 2 are relatively low: the points are scattered widely in all cases, so many of them fall far from the fit line.

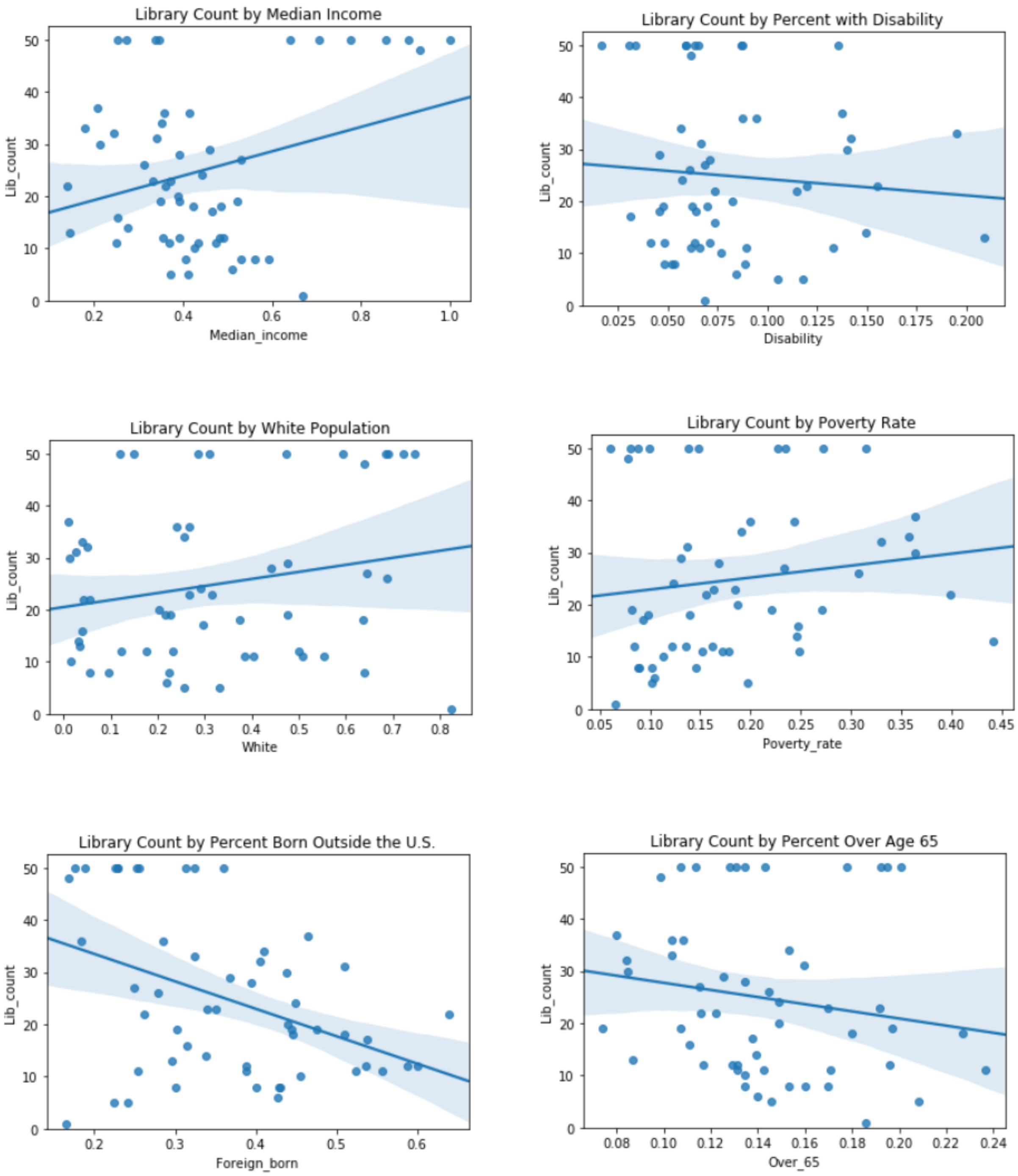


Figure 11: Regression plots of each of the socioeconomic variables versus library count

4. Discussion

The plots in Figure 11 indicate that the two variables with the steepest fit lines are Median Income and Percent Foreign-Born (the first and fifth plots in Figure 11, respectively.) The fact that the fit lines are steepest means that these two variables have the strongest correlations with the number of libraries located in a community. Indeed, these two variables have the highest Pearson coefficients and lowest p-values in Table 2 as well, which means that out of all the variables, we have the most certainty about their fit, and the fit shows more correlation with library count than the other variables.

In the case of Median Income, the correlation is positive; that is, as the median income of a community goes up, the more libraries that community tends to have. Conversely, the Foreign-Born variable has a negative correlation with library count; namely, if a community has a higher percentage of the population comprised of immigrants, that community tends to have fewer libraries.

It is important to note that both of these correlations have Pearson coefficients below 0.5, so they are certainly not strong correlations, but they do exist nevertheless. What this means for librarians and for city planners is that New York City has done a reasonable job with its distribution of libraries in the sense that there are no very strong correlations between privilege and library availability to citizens; however, we cannot claim that the distribution of libraries is fully equal when it comes to wealth and immigration status: wealthier communities do tend to have more libraries, and communities with larger immigrant populations tend to have fewer libraries.

5. Conclusion

This research was an examination of library distribution in one U.S. city, but it would be valuable to replicate the same study for other cities as well to see how the results compare. It is possible that since New York City has a long history of ethnic and social diversity, it may have had time to establish resources like libraries in less-privileged areas, so that there were no strong correlations between the socioeconomic features of a community and the number of libraries in that community. Future research should investigate whether the same result holds true in other major cities that have distinct community areas with varying socioeconomic diversity.

6. References

American Library Association. "Core Values of Librarianship." Retrieved from http://www.ala.org/advocacy/intfreedom/corevalues

NYC OpenData. "Public Use Microdata Areas (PUMAs)." Retrieved from https://data.cityofnewyork.us/Housing-Development/Public-Use-Microdata-Areas-PUMA-/cwiz-gcty

NYC.gov. "New York City PUMAs and Community Districts." https://www1.nyc.gov/assets/planning/download/pdf/data-maps/nyc-population/census2010/puma_cd_map.pdf

NYU Furman Center. "CoreData.nyc." Retrieved from http://app.coredata.nyc/?mlb=true&ntii=hh_inc_med_adj&ntr=Sub-Borough%20Area&mz=14&vtl=https%3A%2F%2Fthefurmancenter.carto.com%2Fu%2Fnyufc%2Fapi%2Fv2%2Fviz%2F691a2b7c-94d7-46ac-ac4d-9a589cb2c6ed%2Fviz.json&mln=true&mlp=true&mlat=40.718&ptsb=&nty=2017&mb=roadmap&pf=%7B%22subsidies%22%3Atrue%7D&md=table&mlv=false&mlng=-73.996&btl=Sub-Borough%20Area&atp=neighborhoods

NYU Spatial Data Repository. "2016 New York City Public Use Micro Areas (PUMAs)." Retrieved from https://geo.nyu.edu/catalog/nyu-2451-34562