The Wealth of Literature

Marium Sultan (/users/175510-marium-sultan) (**New York University (/inst/15081)**)

+ Add Collaborator Manage

PUI2017 Extra Credit Project <Marium Sultan, MariumS, mas1300>

Abstract:

Access to libraries is important for everyone . This project aimed to find if libraries are equitably distributed by median household income, with Public Use Microdata Areas as the geographical unit.

Public libraries provide free access to books, technologies and classes . I also included special and academic libraries from the NYC Open Data facilities database, which may not be accessible to all, but are to a subset of the population.

I used linear regression to check for correlation between income and amount of libraries by PUMA, and Moran's I to check for clustering in space.

Results of spatial autocorrelation suggest that libraries do cluster in space, and looking at the map that clustering seems to be in middle and lower Manhattan (Fig. 2). The amount of libraries positively correlates with the median income of the PUMA it belongs to, but with too low an R^2 to draw any meaningful conclusions. Future work can include quality of library rather than just presence of one.

Introduction:

What sparked this idea is my memory of reading an article saying that there are no bookstores in the Bronx. The Bronx is known overall as a lower income than the other boroughs so I wondered if that was the key factor in this lack (1 (http://gothamist.com/2016/11/03/no_bookstores_the_bronx.php)). Steven Melendez examined over 60 years of bookstore data and noted the decreases (4 (http://gothamist.com/2015/01/30/rip_nyc_bookstores.php)). I predicted that neighborhoods with higher median incomes will be more likely to contain bookstores and libraries. I also wondered if there is a difference in which neighborhoods contain libraries versus which contain bookstores, and if one of these two institutions is more equitably distributed. In this version of the project I only looked at libraries. I decided not to include bookstores because the price of books is a deterring factor for their accessibility and places with higher income would allow more business to these institutions by fact. The data would also have to be webscraped, a skill I have not yet learnt.

Public libraries provide much more than books. They also have language classes, technology training programs and job search resources, among other services (5 (http:// https://nycfuture.org/data/libraries-teach-tech-building-skills-for-a-digital-world)). "Altogether, the Library offers 103,000 free programs annually, serving everyone from toddlers to teens to seniors" (2 (http://https://www.nypl.org/help/about-nypl)).

I did not find any previous work analyzing if there is equitable distribution of libraries in New York City.

Data:

In order to pursue my analysis I collected data on median household income and population from American Community Survey, extracted a list of libraries from the facilities database, and found the shapefile for Public Use Microdata Areas, the latter two both from NYC Open Data.

I initially intended to use census tract as my unit of analysis but examining the data I collected by census tracts I realized that numbers were repeated across regions, even within NYC itself. For example, even filtering the income data by county to only include the 5 boroughs census tract 13800 occurred 5 times. At that point I decided to start over with Public Use Microdata Areas instead as the labels are not repeated.

The ACS data I accessed through an API call after identifying the correct variables from the long list of possibilities. My previous work on assessing the distribution of broadband across the city provided the template for this call.

The income and population data come from American Community Survey Estimates (5 yr). Although I did not access the margin of error information I know that such estimates often have wide margins of error, calling their validity into account. The library data included pubic, special and academic libraries and I trust the facilities database to be fairly comprehensive. I am not sure how fully accessible special and academic libraries as compared to public ones, but they still add to the provision of knowledge, at least to a subset of the population.

After doing an inner merge on the income data (which was for the whole NY state) and the NYC puma shapefile, I did a spatial join with the libraries dataframe, which had point geometry for each institution.

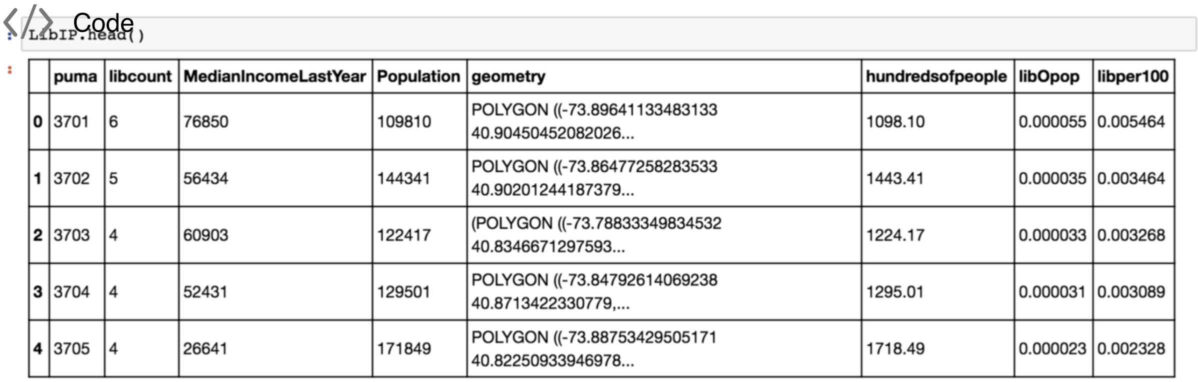
arkdown

```
#PwL = Puma with Libraries
#facl = Libary subset of Facilities Database, with Point Geometries
#PIncomeMap = Income by Puma, with Polygon Geometries

PwL = gpd.sjoin(facl, PIncomeMap, how="inner", op='intersects')
```

This placed the libraries within their PUMA polygons, which allowed for counting libraries by PUMA.

</> Code



```
: LibIP.head()
```

	puma	libcount	MedianIncomeLastYear	Population	geometry	hundredsofpeople	libOpop	libper100
0	3701	6	76850	109810	POLYGON ((-73.89641133483133 40.90450452082026...	1098.10	0.000055	0.005464
1	3702	5	56434	144341	POLYGON ((-73.86477258283533 40.90201244187379...	1443.41	0.000035	0.003464
2	3703	4	60903	122417	(POLYGON ((-73.78833349834532 40.8346671297593...	1224.17	0.000033	0.003268
3	3704	4	52431	129501	POLYGON ((-73.84792614069238 40.8713422330779,...	1295.01	0.000031	0.003089
4	3705	4	26641	171849	POLYGON ((-73.88753429505171 40.82250933946978...	1718.49	0.000023	0.002328

Fig. 1

Dataset with all relevant variables for analysis. Last 3 are calculated by me (left to right): hundreds of people, library count divided by population, and libraries per 100 people

Methodology:

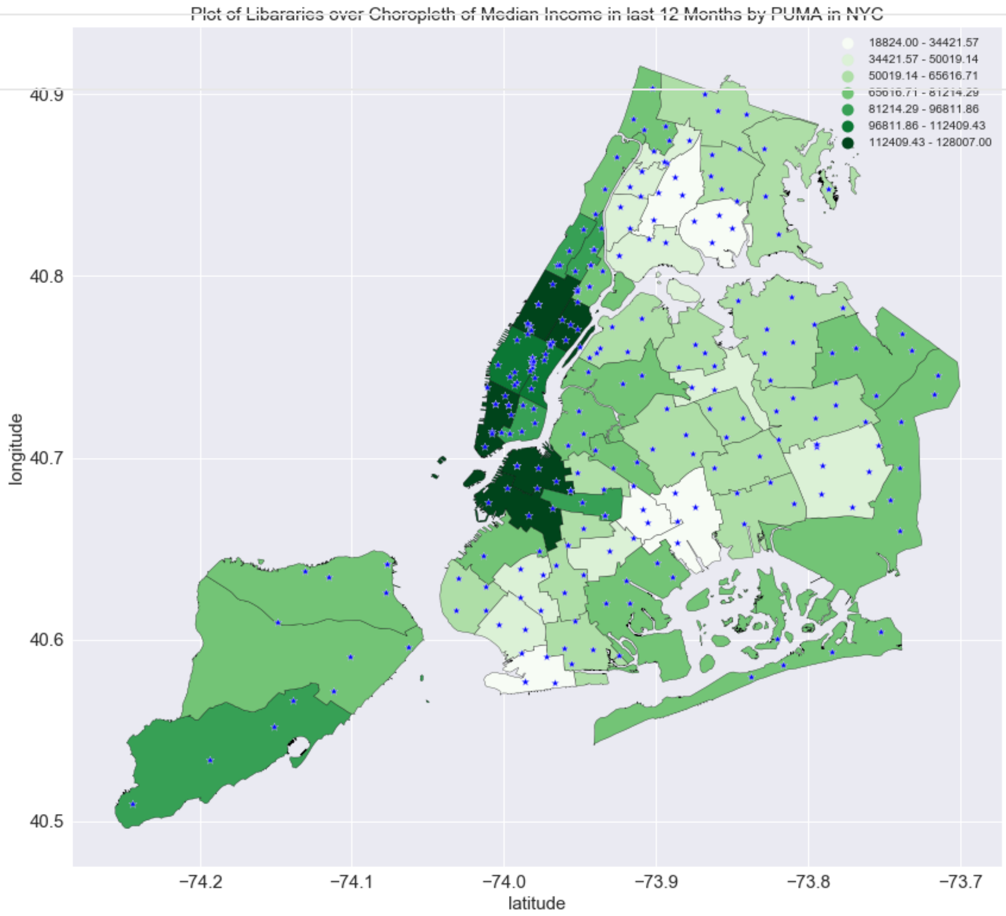


Fig. 2

Libraries plotted as stars over choropleth of median income by PUMA

Conclusions:

I did not find a strong correlation between the amount of libraries in a certain PUMA and the income of that PUMA. I tried measuring libraries per 100 people plotted against income (Fig. 5), the log of libraries over population against income (Fig .4) , pure library count against income (Fig. 3), and library count by population (Fig. 6). In all cases the R^2 was below .30.

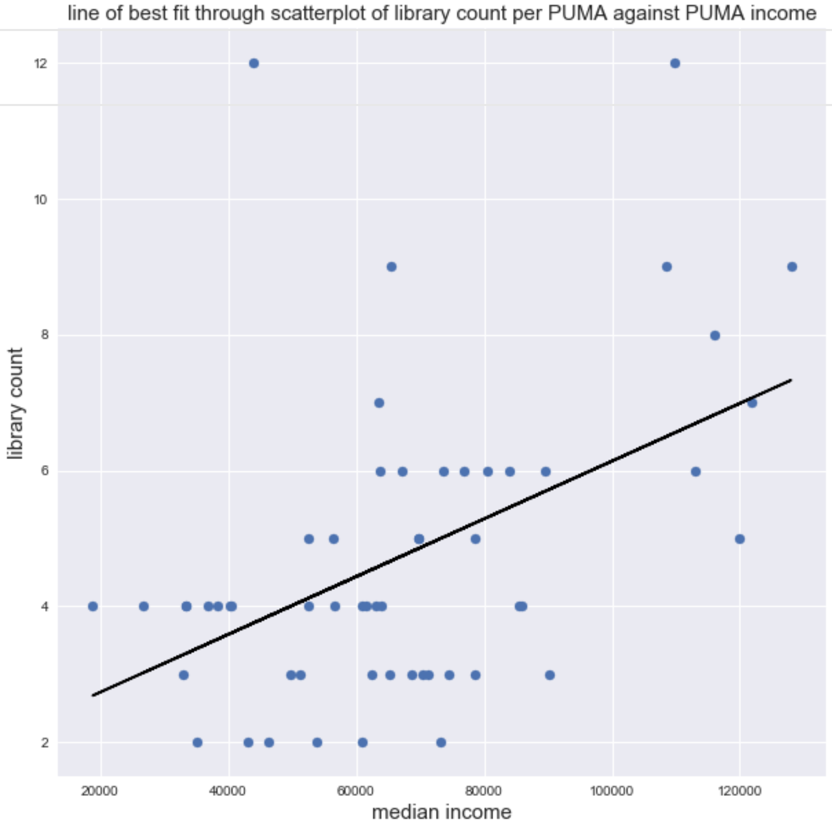


Fig. 3

Note upward trend. Adjusted R^2= 0.211

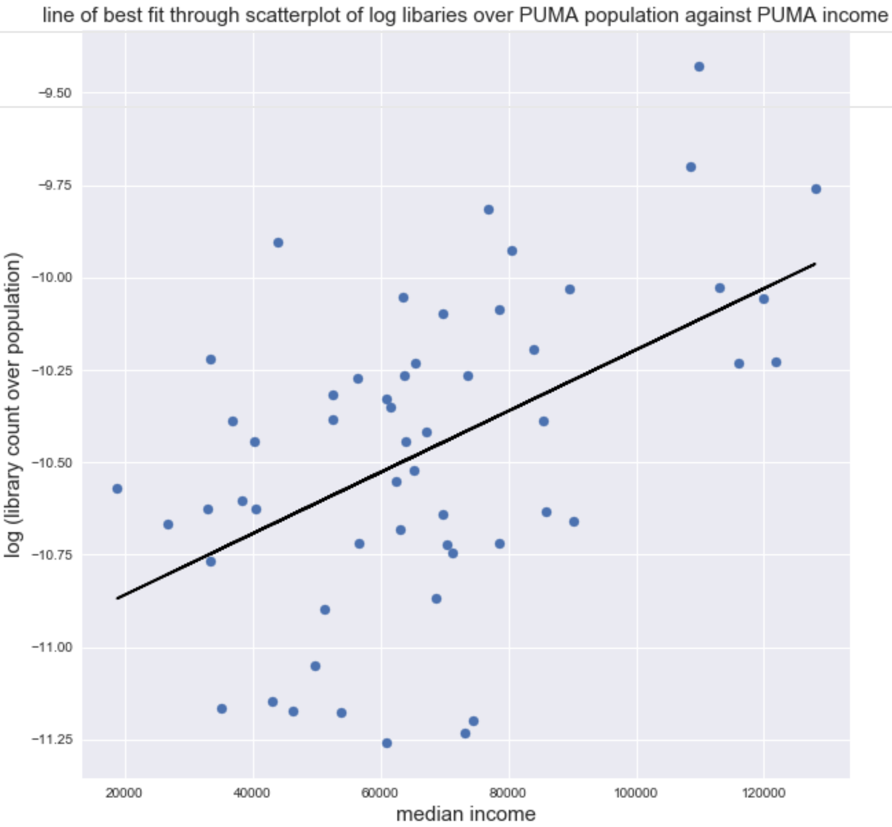


Fig. 4

Note upward trend. Adjusted $R^2 = 0.247$

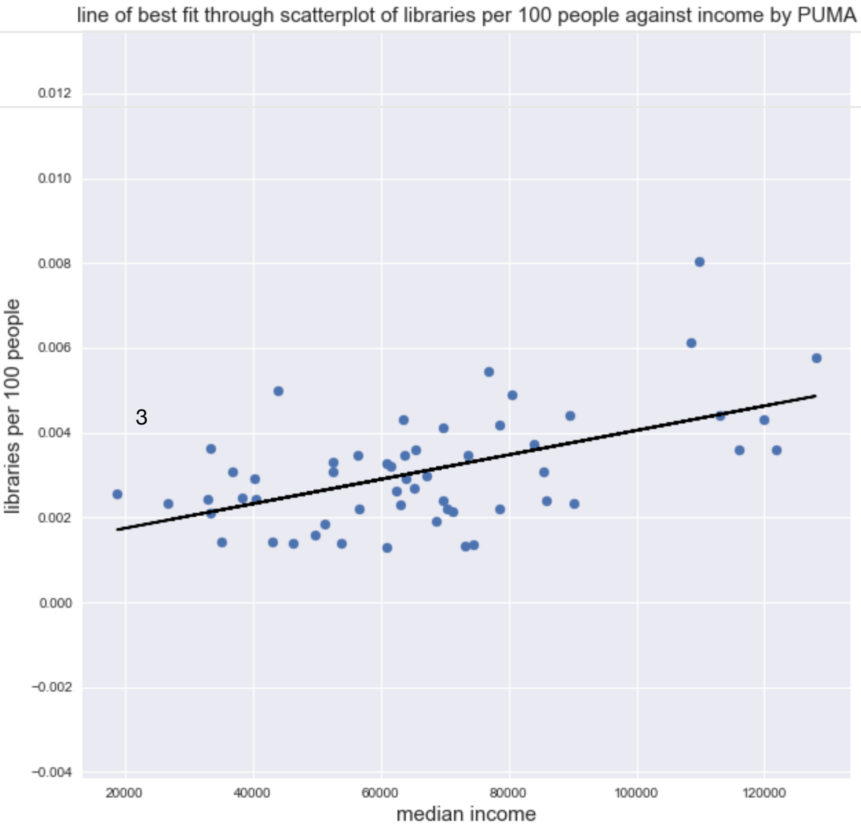


Fig. 5

Scatterplot of Income against libraries per 100 people, by PUMA. Note lack of trend. Adj. R^2 = 0.281

The adjusted R^2 indicates only 28.1% of the variability is accounted for by this model (not so good fit).

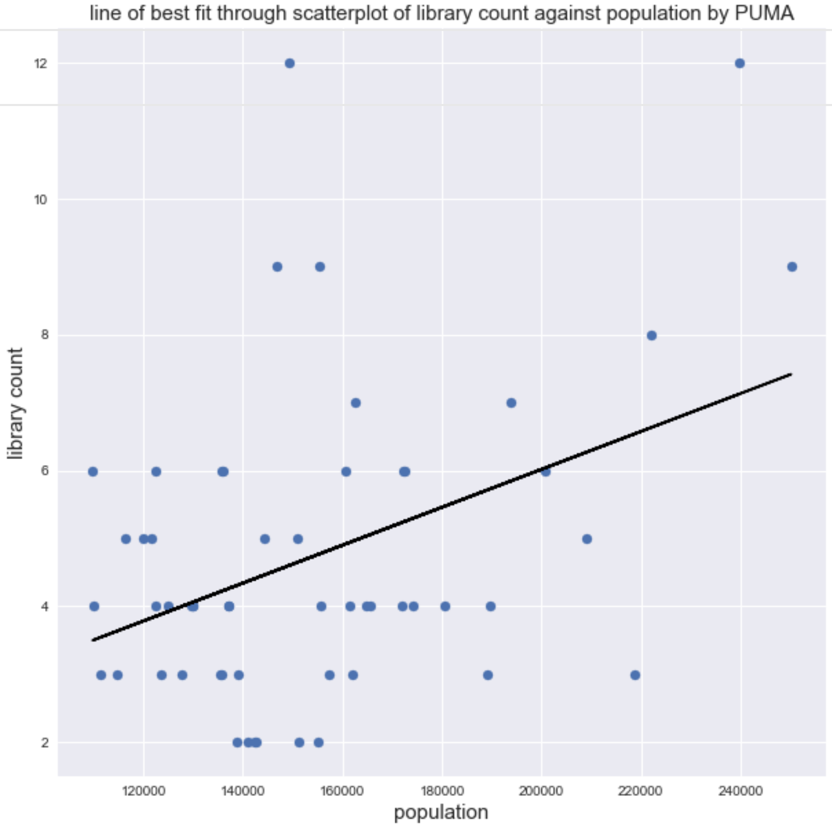


Fig. 6

Scatterplot of population against library count, by PUMA. Adj. R^2 = 0.143

The amount of libraries correlates even worse with population than with income.

The fact that there is a lower R^2 for libraries by population than the log of library count over population by income, seems to indicate that income does affect the distribution of libraries to some degree, although future multivariate regression and feature selection would have to be done here.

Lastly, I used Moran's I to test for spatial autocorrelation of library locations, setting an alpha of 0.05 and a null hypothesis that there is no spatial autocorrelation of libraries in NYC. For Moran's I, when using the Rook technique, the P value is 2.84e-5. For Moran's I, when using the Queen technique, the P value is 2.015e-5. Both of these are much less than the alpha of 0.05 so we can reject the null that there is no spatial autocorrelation. Looking at Fig. 1 we can see a clustering of libraries in middle and lower Manhattan, which may be what this algorithm is picking up on. Both methods showed a positive Z value which means similar values cluster together. (6 (http://http://www.statisticshowto.com/morans-i/))

Future work:

Other work can go back to the original question and include bookstores in the analysis, I suspect they are more correlated with income than libraries because they need an influx of sales to maintain themselves. Sales depend on having disposable income.
What also matters that I did not add to my analysis is how much each individual library offers. This could include hours, frequency of classes, types of classes, technologies and size of collection. Libraries vary greatly in these components and to group them all into a simple binary of existing or not ignores nuance. Although there isn't a strong R^2 for the presence of libraries and income there may be for the quality of those libraries.

Links:

https://github.com/MariumS/PUI2017_mas1300/tree/master/ExtraCreditProject (https://github.com/MariumS/PUI2017_mas1300/tree/master/ExtraCreditProject)

Bibliography

1. http://gothamist.com/2016/11/03/no_bookstores_the_bronx.php (http://gothamist.com/2016/11/03/no_bookstores_the_bronx.php)
2. https://www.nypl.org/help/about-nypl (https://www.nypl.org/help/about-nypl)
3. https://www.bustle.com/p/7-reasons-libraries-are-essential-now-more-than-ever-43901 (https://www.bustle.com/p/7-reasons-libraries-are-essential-now-more-than-ever-43901)
4. http://gothamist.com/2015/01/30/rip_nyc_bookstores.php (http://gothamist.com/2015/01/30/rip_nyc_bookstores.php)
5. https://nycfuture.org/data/libraries-teach-tech-building-skills-for-a-digital-world (https://nycfuture.org/data/libraries-teach-tech-building-skills-for-a-digital-world)
6. http://www.statisticshowto.com/morans-i/ (http://www.statisticshowto.com/morans-i/)