By Frank Schliephacke, 2019-07-24

REPORT: Final Project – socio-economic evaluation of neighborhood

Introduction

Can the socio economic degree of a neighborhood be determined based on the number of coffee and pizza places ?

The target audience for the answer to the above question is anyone wanting to know about the wealth and age distribution of certain boroughs/neighborhood. This simple correlation may be of interest to a person wanting to move to a new city.

Background information:

Given the rising cost of housing and apartments in todays major cities, at least in Stockholm, Sweden, a trend for smaller apartments and houses is seen. Also a trend of moving the living room to coffee shops can be noticed, in Stockholm Sweden, at least. As apartments/houses get smaller it is easier for people to meet in coffee shops instead.

The project will investigate the cost of housing/apartments and compare it to the number of coffee shops and pizza places in the surrounding area of major cities.

The goal is to be able to predict not only the cost of housing and apartments based on the availability of coffee shops and pizza places, but also correlate to the age of the people that live in the area defined by the number of coffee shops and pizza places. As we also want to be able to predict something about the age of people living in the neighborhood the project is about the social economic aspect of a neighborhood.

Data description

A json-file [2] holding the AreaNames/Neighborhoods including geolocation data was used.

Due to time constraints and availability only data for New York City has been used. The used data was extracted from data provided by the City under the open data concept [1].

The used average age per neighborhood has been calculated based on population data; the mid value of the age range bins have been used. For the 85 and above bin the age was set to 91 years.

The average income was calculated based on the economic information for the neighborhoods. For the highest income bin, being unlimited upwards an assumption was made to use a value of twice the low value for this bin.

The housing cost was calculated based on the housing economy information for the neighborhoods. For the highest bin, being unlimited upwards an assumption was made to use a value of twice the low value for this bin.

The foursquare database (foursquare.com) has been used to determine the availability of coffee and pizza places.

Methodology

For the data analysis Python 3.6 was used. The code was assembled using a jupyter notebook [3] on IBM Watson Studio. In the end some work was done on a local PC as well.

The income, housing cost and age data have been extracted from the csv files. The neighborhood names used in the files where connected to a neighborhood ID, thus I chose to use the AreaID (= neighborhood ID) to track and unify the data, not only the AreaName (=Neighborhood Name).

Some simple plot figure where created using this data. The relevance of the income, housing cost data was evaluated by creating a ratio income (per year) divided by housing cost. Using the 2-sigma (=mean+2*standard deviation) approach to identify the significant peaks in this ratio. The analysis was limited to use Manhattan neighborhoods only. The identified Manhattan AreaID (Neigborhoods) where manually compared to the results of the amount of coffee and pizza places for an AreaID, see also table 1.

Finally a clustering using the top 10 venues as extracted from foursquare using a 200m radius per AreaID (Neighborhood) was done. The number of clusters was manually evaluated to not change significantly above 7.

For the purpose of simpler understanding interactive folium maps have been created showing relevant data for each AreaID (Neighborhood). The interactivity does not work in Github or local PC jupyter.

Results

Below figure show the relation between parameters of interest; the parameters dependency on the AreaID/Neigborhood.

X	AreaID	Neighborhood	Coffee	Pizza	SocioRatio (2Sigma limit)
0			0	0	0
1			0	0	0
2	MN35	Washington Heights	2	2	0
3	MN01	Inwood	2	0	0
4			0	0	0

5	MN06	Manhattanville	0	1	0
6			0	0	0
7	MN33	East Harlem	2	0	0
8	MN40	Upper East Side	1	0	76
9	MN32	Yorkville	1	0	65
10			0	0	0
11			0	0	0
12			0	0	0
13			0	0	0
14	BK69	Clinton	1	0	0
15	MN17	Midtown	1	2	65
16	MN20	Murray Hill	0	2	65
17	MN13	Chelsea	0	1	62
18	--	Greenwich Village	0	1	0
19	MN22	East Village	3	2	0
20			0	0	0
21	--	Tribeca	1	1	0
22	MN24	Little Italy	1	5	0
23			0	0	0
24	MN23	West Village	0	3	72
25			0	0	0
26	MN21	Gramercy	0	1	70
27	MN25	Battery Park City	0	1	68
28	--	Financial District	2	1	0
29	MN40	Carnegie Hill	1	1	76
30	--	Noho	2	3	0
31	MN24	Civic Center	0	2	0
32	MN17	Midtown South	0	4	65
33	--	Sutton Place	1	1	0
34	MN19	Turtle Bay	0	2	77
35			0	0	0
36			0	0	0
37	--	Flatiron	1	3	0
38	MN13	Hudson Yards	0	3	62

Table 1. Bar chart data Fig 1. – Only Manhattan Coffe & Pizza places.

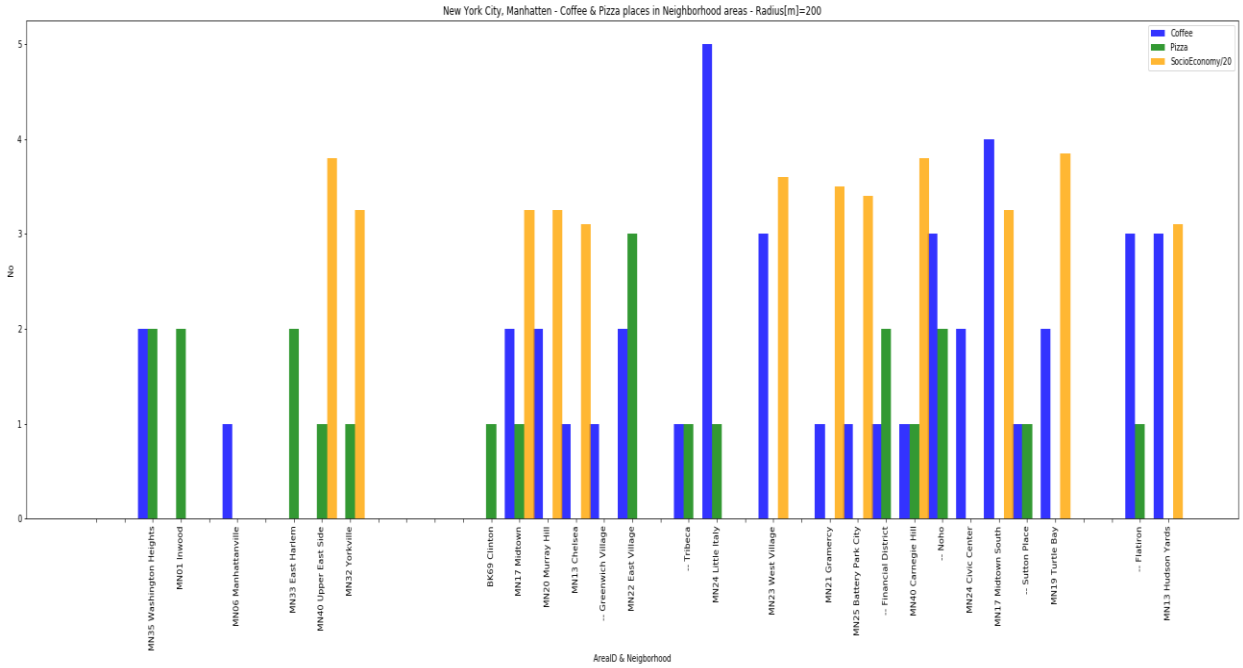

Fig.1. Manhattan Coffe & Pizza places for neighbarhoods (Radius=200m) and Socio-economic ratio.

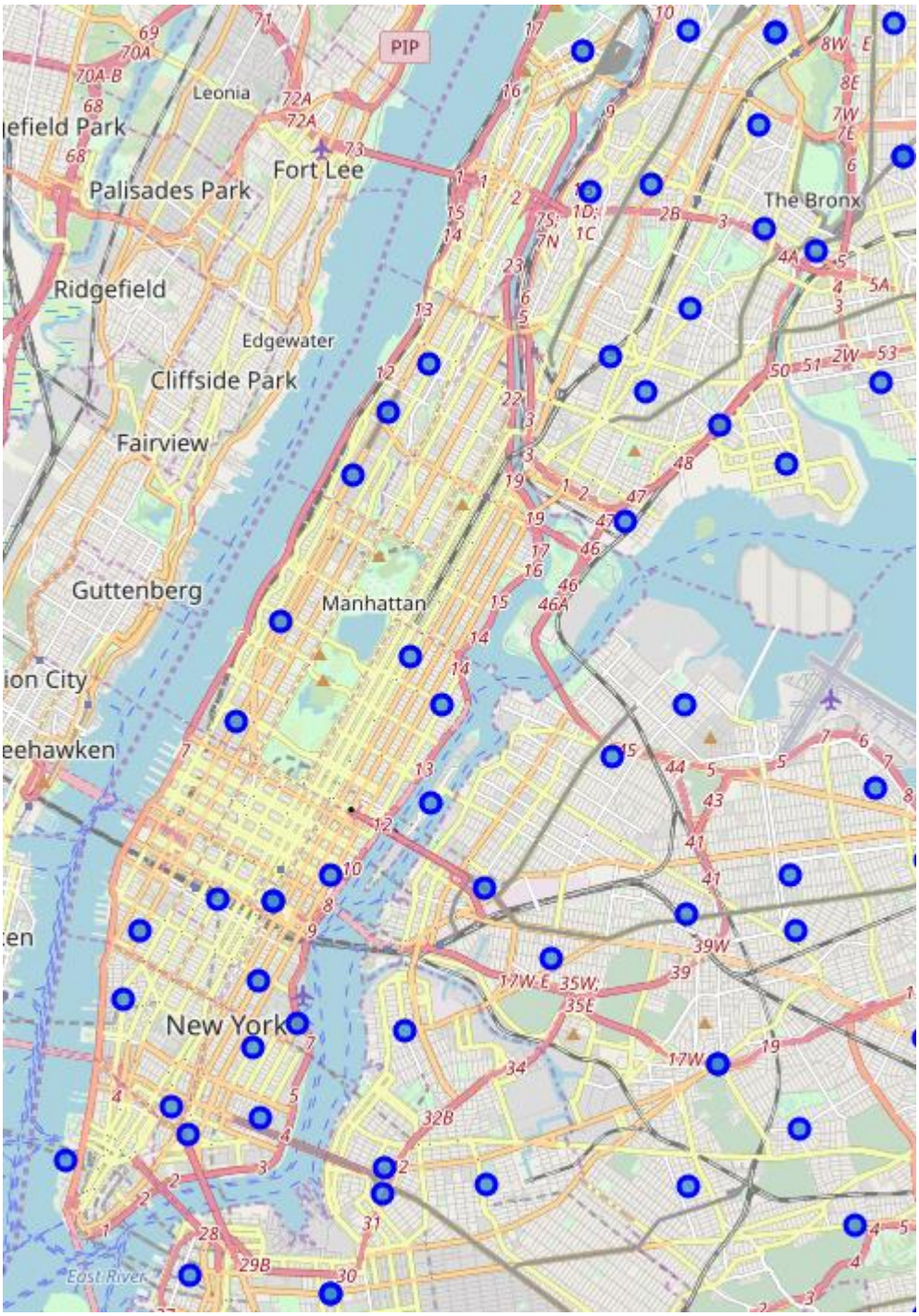

Fig.2. Manhattan neighborhoods

Discussion

One of the major obstacle to handle in this effort was to identify data with a high resolution geographic dependency, this was resolved using the AreaNames(Neighborhood Names) and associated AreaID as used by New York City.

The foursquare data extraction used a 200 m radius, as the initial suggested 564 m, would allow for an overlap of extracted foursquare location data.

As shown in table 1 and fig 1 there are some Area/Neighborhoods which also show a deviating number of coffee and pizza places, thus the initial question of the relation of coffee & pizza places to the income-housing_cost-neighborhood seems to have some correlation.

Conclusion

The amount of correlation as claimed in the initial question ("Can the economic degree of a neighborhood be determined based on the number of coffee and pizza places") seems however not clear.

In order to prove the correlation to actually exist at least two other cities should be investigated, preferably at least one major city in the US and another major city in a different country and cultural environment such as EU, China, India.

References

[1] NYC OpenData
https://data.cityofnewyork.us/browse?q=Demographic%20and%20Housing%20Profiles%20by%20Borough&sortBy=relevance

[2] New York neighborhood geolocation data json format
https://cocl.us/new_york_dataset

[3] Github – authors python jupyter notebook
https://github.com/KnarfSkidbacke/Coursera_Capstone