Zoning of Health Districts: Segmentation of New York City Neighborhood per Restaurant and Sport Institutions.

Lydia BESSAI

January 22, 2021

INTRODUCTION

Obesity is a main issue in the United States. The prevalence of obesity was 42.4% in 2017/2018. Moreover, obesity-related conditions are composed of heart disease, stroke, type 2 diabetes, and certain types of cancer.
Obesity concerned physical condition but also medical cost. Indeed, the estimated annual medical cost of obesity in the United States was $147 billion in 2008 and people with obesity cost $1,429 more than people of normal weight [(1)]. In New York City, the percentage of adults who have overweight or are obese increased from 42% in 1997 to 62.7% in 2018 [(2)].

Many analyses study the relationship between obesity and race, income, gender, level of education, or age [(3)]. However, the impact of the environment on obesity is still incomplete [(7)]: obesity is mainly related to food and lack of sport. Thus, the number of restaurants and the number of sports institutions in a city could directly influence the obesity trend. High availability of restaurants compared to low availability of sports facilities has a psychological impact on food consumption. Indeed, predominant exposure to restaurants in the environment could increase body weight.

The goal of this study is to present a clustering analysis of the neighborhoods in New York City (NY) based on the number of restaurants and sports institutions in each district. The clustering analysis is the first approach to identify a "healthy index" of each neighborhood. To do so, we will first calculate a restaurant index and a sport index related to the number of restaurants or sports institutions in each neighborhood. Then, these indexes will be used to perform a clustering analysis to define different types of districts.

Lastly, we will correlate the restaurant index and the sport index with the income and the density population of each borough in NY.

This analysis could help governmental institutions to regulate the distribution of restaurants and sports facilities in order to act on body weight of NY residents. Further analyses as the correlation of the number of restaurants with the obesity rate could bring additional elements to understand the impact of the environment on obesity trend.

DATA

Description of data

To achieve this analysis, we have to combine three kinds of data.

The first part of our analysis aims to get the venues – restaurants and sports facilities – for each neighborhood in NY.
To do so, we needed to collect the names of boroughs, neighborhoods, latitudes, and longitudes in NY to segment all the districts. These values are presented in a table. Then, we found all the restaurants or sports institutions in each neighborhood. For that, we called the Foursquare API and we specified the needed category (category 1: Athletics & Sports; category 2: Food). Each GET request gathered all the venues thanks to the latitudes and the longitudes from the first table. Thanks to these new values, we calculated a restaurant and a sport index related to the number of respective categories.

The second part of the analysis consists to correlate the indexes with the income and the density population. This analysis is less precise than the first one because of a lower level of granularity. Indeed, the data available regarding the income or the density of population target boroughs instead of neighborhoods. Moreover, the data come from Wikipedia website and were presented through a table combining information such as Jurisdiction, Population, Gross Domestic Product, Land area, and Density per NY borough. To select income and density population, we scraped the table and we cleaned it.

Data Cleaning

Data cleaning relates exclusively to Wikipedia data. Please find the table below.

New York City's five boroughs V·T·E								
Jurisdiction		**Population**	**Gross Domestic Product**		**Land area**		**Density**	
Borough	*County*	*Estimate (2019)*	*billions (2012 US$)*	*per capita (US$)*	*square miles*	*square km*	*persons / mi²*	*persons / km²*
The Bronx	Bronx	1,418,207	42.695	30,100	42.10	109.04	33,867	13,006
Brooklyn	Kings	2,559,903	91.559	35,800	70.82	183.42	36,147	13,957
Manhattan	New York	1,628,706	600.244	368,500	22.83	59.13	71,341	27,544
Queens	Queens	2,253,858	93.310	41,400	108.53	281.09	20,767	8,018
Staten Island	Richmond	476,143	14.514	30,500	58.37	151.18	8,157	3,150
City of New York		**8,336,817**	**842.343**	**101,000**	**302.64**	**783.83**	**27,547**	**10,636**
State of New York		19,453,561	1,731.910	89,000	47,126.40	122,056.82	412	159

Table 1: Wikipedia table (4)

As shown, many details are not irrelevant to our analysis. We choose to work with "Density" and "Gross Domestic Product" (GPD). As the neighborhood could be narrow, we target "persons / mi2". Furthermore, GPD is the economic indicator that makes it possible to quantify the total value of the annual "production of wealth" carried out by economic agents (households, companies, public administrations) residing within a territory. We used GPD in billions to approximate income.
Please find the cleaned table below.

	Borough	County	Population (2019)	GDP billions (2012, US$)	GDP per capita (US$)	square miles	squarekm	persons/mi2	persons/km2
0	The Bronx	Bronx	1418207	42.695	30100	42.10	109.04	33867	13006
1	Brooklyn	Kings	2559903	91.559	35800	70.82	183.42	36147	13957
2	Manhattan	New York	1628706	600.244	368500	22.83	59.13	71341	27544
3	Queens	Queens	2253858	93.310	41400	108.53	281.09	20767	8018
4	Staten Island	Richmond	476143	14.514	30500	58.37	151.18	8157	3150

Table 2: Table from our analysis after scraping and cleaning Wikipedia values.

Source of data

Data type (NY)	Source	Additional information (weblink or technical request)
Borough, Neighborhood, Latitude, Longitude	Coursera Class	https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json
Restaurants and Sport Institutions	Foursquare	Call Request from API Foursquare
Income and Density Population	Wikipedia	https://en.wikipedia.org/wiki/Boroughs_of_New_York_City

Table 3: Data type, source and Additional information.

METHODOLOGY

We performed all the analysis thanks to Jupyter Notebook using Python language.

Determination of neighborhoods in NY

To execute analysis on NY neighborhoods, we need to define the latitude and the longitude of each district. After downloading the dataset, we get the table that gathers boroughs, neighborhoods, latitudes, and longitudes in NY *(Table 4).* Thanks to this table, we can create a first map showing all the districts in NY thanks to the Folium library *(Image 1).*

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Table 4: Head of the dataset gathering NY neighborhoods.

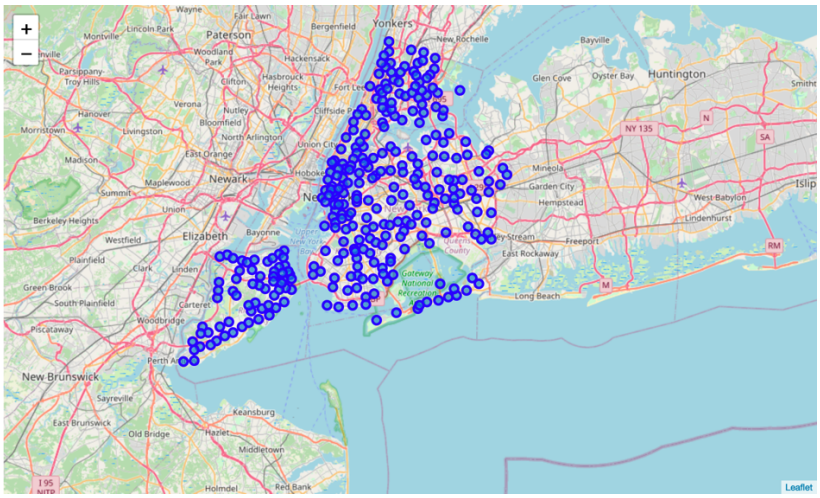

Image 1: Map of neighborhoods in NY.

Importation of Foursquare data

Foursquare is a social media that allows users to indicate their location using a geolocation system and, in doing so, recommend places to go out. Thanks to the developer's platform, we can request API call and get the venues for a specific location. We focused on venues regarding restaurants and sports institutions. To get all the venues for these categories, we created an URL request specifying the categoryID of "Athletics & Sports" and "Food" (5, 6). This filter allowed us to directly target the relevant venues. We fixed the limit to 100 venues.

We created two tables combining neighborhoods and venues (table 1: restaurant venues; table 2: sports institutions venues).

In order to calculate the restaurant index and the sport index, we executed the function "get_dummies" for one-hot encoding. The results are two tables with binary numbers (0: no venue; 1: recorded venue). Finally, we grouped by neighborhood to have the sum of venue_category for each district.

Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	Bagel Shop	Bakery	Belgian Restaurant	Bistro
Allerton	0	0	1	0	0	0	0	0	0	0	3	0	0
Annadale	0	0	3	0	0	0	0	0	0	0	2	0	0
Arden Heights	0	0	0	0	0	0	0	0	0	0	0	0	0
Arlington	0	0	2	0	0	0	0	0	0	0	0	0	0
Arrochar	0	0	0	0	0	0	0	0	0	2	0	0	0
Arverne	0	0	0	0	0	0	0	0	1	0	0	0	0
Astoria	0	0	1	0	0	0	0	0	1	3	4	0	0

Table 5: Example of the table for restaurants after one-hot encoding and groupby function

Calculation of Restaurant and Sport index

To calculate the restaurant and sport index, we need to count the total number of restaurants or sports institutions in each neighborhood. Then, we divided the results per the maximum value of restaurants or sports institutions (limit= 100) to normalize the results. The final results are represented by a range from 0 to 1. To create a contrast between restaurants index and sports facilities, we have reversed the order of the range for restaurants. Thus, the index range for sports facilities is from 0 to 1 (0: no presence of sports facilities in the neighborhood; 1: the neighborhood with the most sports facilities) and the index range for restaurants is for 1 to 0 (1: no presence of restaurants in the neighborhood; 0: the neighborhood with the most restaurants).

The final table combines boroughs, neighborhoods, sport index and restaurant index.

	Borough	Neighborhood	Sport index	Restaurant index
0	Bronx	Wakefield	0.01	0.96
1	Bronx	Co-op City	0.04	0.90
2	Bronx	Eastchester	0.00	0.88
3	Bronx	Fieldston	0.01	0.00
4	Bronx	Riverdale	0.02	0.99

Table 6: Final table combining Borough, Neighborhood, Sport index and Restaurant index

Clustering of neighborhoods based on sport and restaurant index

We choose the “kmeans” clustering method to segment NY districts.

Kmeans is a well-known non-supervised machine learning method. Kmeans clustering is a method of partitioning data and a combinatorial optimization problem. Given points and an integer k, the problem is to divide the points into k groups, often called clusters, to minimize the function. We consider the distance of a point from the average of the points of its cluster; the function to be minimized is the sum of the squares of these distances.

After trying "3" and "5" as numbers of clusters, we choose "4" as the final value. Indeed, the more we increase the number of clusters, the less the repartition was uniform. However, the interpretation of clusters with "4" types was more interesting than with "3".
The results were shown in the NY map where each neighborhood associated with a cluster label is represented.

Correlation of Sport and Restaurant index with income and density population

To finish the analysis, we operated a correlation. The objective of the correlation is to highlight potential links between the sport index or the restaurant index with the income or the density population.

For that purpose, we use the Wikipedia data, and we created a final table with borough, sport index, restaurant index, and GDP billions (2012, US$), persons/mi2. The sport index and the restaurant index were calculated with the average value of all the neighborhoods for each borough.

Thanks to all these data, we created a correlation matrix to have a clear vision of positive, neutral, or negative correlation.

RESULTS

Clustering of neighborhoods based on sport and restaurant index

	Cluster Labels	Borough	Neighborhood	Sport index	Restaurant index
0	1	Bronx	Wakefield	0.01	0.96
1	1	Bronx	Co-op City	0.04	0.90
2	1	Bronx	Eastchester	0.00	0.88
3	0	Bronx	Fieldston	0.01	0.00
4	1	Bronx	Riverdale	0.02	0.99

Table 7: Result of label clustering

Cluster label	Number of neighborhoods
0	16
1	214
2	55
3	29

Table 8: Distribution of labels

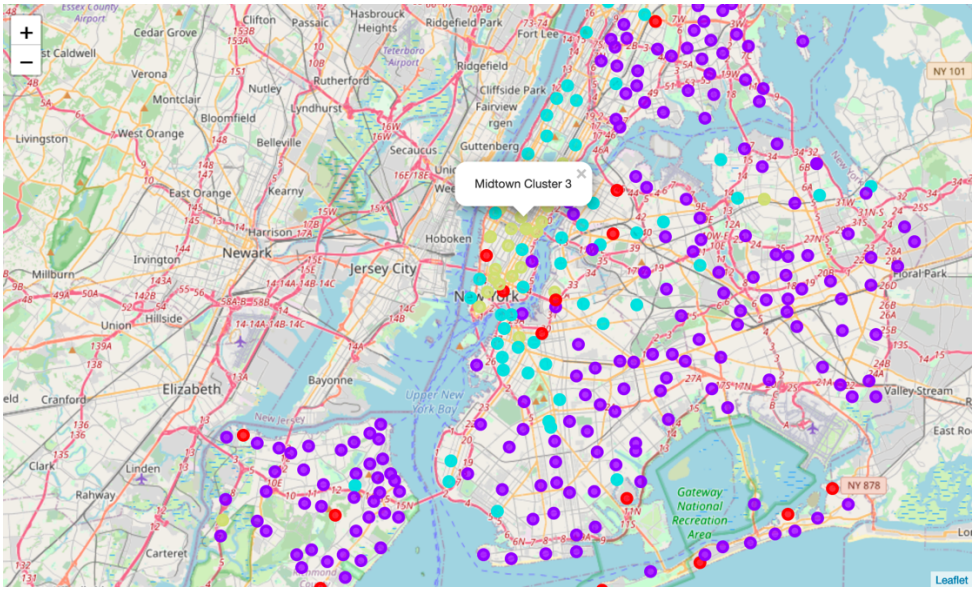

Image 2: Map representing the clusters in NY based on the restaurant index and the sport index.

Examination of clustering

Cluster 0 – the restaurant neighborhoods

Represented in red in the map *(Image 2),* the cluster 0 has the particularity to have a very low restaurant index (mean: 0.089375; median: 0) and a very low sport index (mean: 0.088125; median: 0.015000). Moreover, the repartition is quite homogenous but with more neighborhoods involved in the Queens borough.

	Cluster Labels	Sport index	Restaurant index
count	16.0	16.000000	16.000000
mean	0.0	0.088125	0.089375
std	0.0	0.125072	0.117102
min	0.0	0.000000	0.000000
25%	0.0	0.000000	0.000000
50%	0.0	0.015000	0.000000
75%	0.0	0.137500	0.145000
max	0.0	0.340000	0.310000

Table 9: Statistic results for the cluster 0

Borough of NY	Number of neighborhoods (cluster 0)
Queens	6
Brooklyn	3
Staten Island	3
Manhattan	2
Bronx	2

Table 10: Distribution of the cluster 0 in NY

Cluster 1 – the low activity neighborhoods

Represented in purple in the map *(Image 2)*, the cluster 1 has a high restaurant index (mean: 0.879626; median: 0.885000) and a very low sport index (mean: 0.029486; median: 0.030000). The repartition is homogenous but with few neighborhoods involved in Manhattan.

	Cluster Labels	Sport index	Restaurant index
count	214.0	214.000000	214.000000
mean	1.0	0.029486	0.879626
std	0.0	0.025494	0.078147
min	1.0	0.000000	0.700000
25%	1.0	0.010000	0.820000
50%	1.0	0.030000	0.885000
75%	1.0	0.040000	0.950000
max	1.0	0.170000	0.990000

Table 11: Statistic results for the cluster 1

Borough of NY	Number of neighborhoods (cluster 0)
Queens	63
Brooklyn	45
Staten Island	59
Manhattan	3
Bronx	44

Table 12: Distribution of the cluster 1 in NY

Cluster 2 – the medium activity neighborhoods

Represented in blue in the map (Image 2), the cluster 2 has a medium restaurant index (mean: 0.554000; median: 0.580000) and a low sport index (mean: 0.142727; median: 0.110000). The neighborhoods from cluster 2 are particularly displayed in Brooklyn, Manhattan, and Queens.

	Cluster Labels	Sport index	Restaurant index
count	55.0	55.000000	55.000000
mean	2.0	0.142727	0.554000
std	0.0	0.109738	0.104058
min	2.0	0.000000	0.350000
25%	2.0	0.075000	0.475000
50%	2.0	0.110000	0.580000
75%	2.0	0.185000	0.630000
max	2.0	0.420000	0.730000

Table 13: Statistic results for the cluster 2

Borough of NY	Number of neighborhoods (cluster 0)
Queens	13
Brooklyn	20
Staten Island	2
Manhattan	14
Bronx	6

Table 14: Distribution the cluster 2 in NY

Cluster 3 – the active neighborhoods

Represented in yellow in the map *(Image 2)*, the cluster 3 has a medium restaurant index (mean: 0.182759; median: 0.120000) and a low sport index (mean: 0.707586; median: 0.670000). The neighborhoods from the cluster 2 are mainly represented in Manhattan.

	Cluster Labels	Sport index	Restaurant index
count	29.0	29.000000	29.000000
mean	3.0	0.707586	0.182759
std	0.0	0.212681	0.128366
min	3.0	0.390000	0.000000
25%	3.0	0.560000	0.120000
50%	3.0	0.670000	0.120000
75%	3.0	0.870000	0.270000
max	3.0	1.000000	0.520000

Table 15: Statistic results for the cluster 3

Borough of NY	Number of neighborhoods (cluster 0)
Queens	2
Brooklyn	2
Staten Island	2
Manhattan	23
Bronx	0

Table 16: Distribution the cluster 3 in NY

Correlation of Sport and Restaurant index with incomes and density population

	Borough	Cluster Labels	Sport index	Restaurant index	GDP billions (2012, US$)	persons/mi2
0	Bronx	1.076923	0.030577	0.807500	42.695	33867
1	Brooklyn	1.300000	0.080143	0.712857	91.559	36147
2	Manhattan	2.380952	0.491667	0.361905	600.244	71341
3	Queens	1.130952	0.063452	0.753333	93.310	20767
4	Staten Island	1.045455	0.044091	0.834394	14.514	8157

Table 18: Table after grouping by Borough and adding Wikipedia data

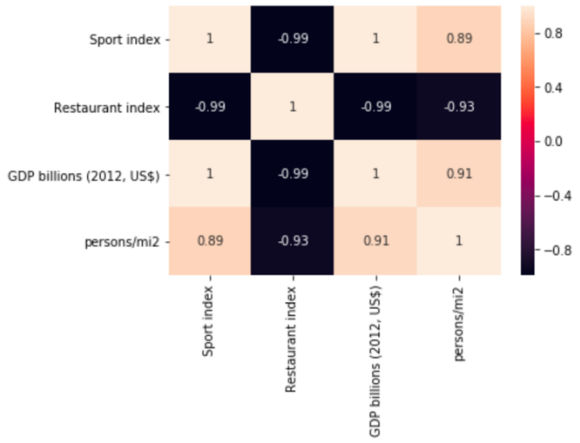

Image 3: Correlation between Sport index, Restaurant index, GPD in billions and persons/mi2.

The correlation results show that the sport index is positively correlated to GDP (correlation coefficient: 1) and density population (correlation coefficient: 0.89). Regarding the restaurant index, the correlation with GDP (correlation coefficient: - 0.99) and density population (correlation coefficient: -0.93) is negative.

DISCUSSION

Examination of clustering

Cluster 0 – the restaurant neighborhoods

The restaurant index and the sport index demonstrate that the neighborhoods in the cluster 0 are the neighborhoods with the most restaurant spots in NY but with a very low quantity of sport facilities.

Cluster 1 – the low activity neighborhoods

The restaurant index and the sport index demonstrate that the neighborhoods in the cluster 1 are the neighborhoods with the fewer restaurant spots in NY and a very low quantity of sport facilities. Moreover, we can see that all the NY districts are represented except Manhattan. Manhattan is a well-known high density and activity district. Taking into consideration all these arguments, we can assume that the cluster 0 is composed of a low activity area and the neighborhoods involved are probably more residentials.
This hypothesis has to be verified by engaging a new analysis on other venues in the Cluster 1.

Cluster 2 – the medium activity neighborhoods

The restaurant index and the sport index demonstrate that the neighborhoods in the cluster 2 have low activity. However, there is a predominance for the restaurant spot. The neighborhoods in the cluster 2 seem to have more restaurant spots than sport facilities.

Cluster 3 – the active neighborhoods

The restaurant index and the sport index demonstrate that the neighborhoods in the cluster 3 have the highest activity in NY. The neighborhoods are mainly based in Manhattan. This confirms that Manhattan is

a high-density area. Moreover, the cluster 3 has the highest sport index. This shows that sports facilities are mainly present in Manhattan, giving less possibility to practice sport for other district residents.

Correlation of Sport and Restaurant index with incomes and density population

The results of the correlation are not surprising. Indeed, taking into consideration that the Manhattan borough is the most active district with the highest level of sport and restaurant spots, we can predict that the relation between density population and incomes is directly linked to facilities because of supply and demand.
The correlation coefficients have shown us this assumption. The number of sport spots and the number of restaurants correlate with the number of residents in NY boroughs and the boroughs' incomes. This last result opens a new questioning on the relation between income and body weight.

CONCLUSION

Clustering of neighborhoods based on sport and restaurant index

The first goal of this analysis was to cluster neighborhoods in NY based on sports and restaurant facilities. This clustering was performed thanks to Foursquare data and Wikipedia data. We also used data from Coursera to get the longitudes and the latitudes of the NY neighborhood. During this first analysis, we calculated a sport index and a restaurant index. These indexes were based on the total number of sports institutions or restaurants in NY (with a limit of 100 venues for Foursquare). After joining neighborhoods to sport and restaurant index, we achieved a Kmeans clustering analysis to segment NY neighborhoods.

The results of clustering analysis revealed 4 types of neighborhoods in NY:

- Cluster 0 – the restaurant neighborhoods
- Cluster 1 – the low activity neighborhoods
- Cluster 2 – the medium activity neighborhoods
- Cluster 3 – the active neighborhoods

Each cluster has a different sport and restaurant mean index. Moreover, all the neighborhoods are homogeneously represented except for the cluster 3 where Manhattan is the main borough involved.
This analysis shows disparities in sport and restaurant access in neighborhoods. This could have a direct impact on the body weight of residents depending on their living area. Indeed, sports facilities are mainly present in neighborhoods in Manhattan. This means that residents from other boroughs have less accessibility to sports institution.

Correlation of Sport and Restaurant index with incomes and density population

The second analysis aimed to correlated sport and restaurant index with incomes and density population.
The results demonstrate that the sport index is positively correlated to GDP and density population. Regarding the restaurant index, the correlation with GDP (correlation coefficient: - 0.99) and density population (correlation coefficient: -0.93) is negative. This means that the number of sport spots and the number of restaurants correlate with the number of residents in NY boroughs and the respective income.

General Conclusion

Obesity and poverty have already been linked [2,3,7]. Besides, several studies have shown disparities in neighborhood access to food and fitness in US cities. However, few studies present the relationship between obesity and food and fitness access.

The main objective of our analysis was to give a first approach to the following research questions: could disparities between fitness and restaurant facilities access in NY neighborhoods have an impact on obesity? Is it possible to act on body weight by regulating food and sports facilities per neighborhood?

First, our analysis demonstrated that disparities between NY neighborhoods do exist. The different kinds of clusters and the correlation with incomes and density population have also shown that these disparities come from several factors.
Also, the findings of Black et al. noticed a significant association between neighborhood disparities and obesity due to different availability of food stores, fitness facilities, percent of commercial land use, and area income.
In conclusion, further analyses need to be realized to investigate the direct link between restaurant/sports facilities with obesity. To do so, a precise dataset gathering the obesity rate of NY residents in each neighborhood is necessary. Explore the type of restaurants (fast-food restaurants, healthy restaurants, …) and the access to other obesity-related facilities (food store or snacking place) could be relevant to conclude about the impact of disparities on obesity in NY neighborhoods.

Limitations and future directions

Our analysis has limitations due to the lack of obesity data and the specific characteristic of each neighborhood.
First, it was impossible to conclude a link between obesity and restaurant/sport index because the obesity rate per district was not available.
Second, the calculation of the restaurant index included all the restaurant facilities. This means that salad bars or other healthy restaurants were taking into account. It is also very difficult to define how healthy is a restaurant basing on the name of the category (i.e., how Mexican restaurant is healthy?) To more specify the analysis, we have to apply a filter on each type of restaurant. For example, the glycemic index could be an interesting value to categorize the restaurant.
Third, due to a free Foursquare account, we could not access all the venues in NY. The limit of 100 for the API call request could also modify the value of sport and restaurant index per neighborhood.
Lastly, to perform a complete analysis on obesity, we have to include all the factors that are related to body weight. For instance, race, gender, and education are other factors that need to be linked to the neighborhoods. Otherwise, the interpretation of the analysis will be biased.

BIBLIOGRAPHY

1. https://www.cdc.gov/obesity/data/adult.html
2. https://www.health.ny.gov/prevention/obesity/
3. https://www.health.ny.gov/prevention/obesity/statistics_and_impact/docs/2000-2010_adult_obesity.pdf
4. https://en.wikipedia.org/wiki/Boroughs_of_New_York_City
5. https://developer.foursquare.com/docs/build-with-foursquare/categories/
6. https://www.xspdf.com/resolution/11049375.html
7. Black, Jennifer L., James Macinko, L. Beth Dixon, and Jr. Fryer George E. 'Neighborhoods and Obesity in New York City'. *Health & Place* 16, no. 3 (1 May 2010): 489–99. https://doi.org/10.1016/j.healthplace.2009.12.007.