Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods (Week 2)

Important Information

It is especially important to submit this assignment before the deadline, May 11, 1:59 AM -05, because it must be graded by others. If you submit late, there may not be enough classmates around to review your work. This makes it difficult - and in some cases, impossible - to produce a grade. Submit on time to avoid these risks.

Instructions

In this week, you will continue working on your capstone project. Please remember by the end of this week, you will need to submit the following:

1. A full report consisting of all of the following components (15 marks):
    - Introduction where you discuss the business problem and who would be interested in this project.
    - Data where you describe the data that will be used to solve the problem and the source of the data.
    - Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
    - Results section where you discuss the results.
    - Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
    - Conclusion section where you conclude the report.
2. A link to your Notebook on your Github repository pushed showing your code. (15 marks)

3. Your choice of a presentation or blogpost. (10 marks)

Is there any room for a new Coffee Shop in the island of Manhattan (New York)?

1. Introduction/Business Problem

Without a shadow of a doubt, we all have heard -at least once in our lifetime- that "New York is the City that never sleeps" and we all can bet, for sure, that Coffee has played a major role in keeping the city awake. Although there seems to be a countless number of Coffee shops spread throughout the City of New York, the renewed interest of the younger generations and the Hipsters[1] on the taste and the correct preparation of the stimulant drink, opens the question of whether it would be a good idea or not to open a new Coffee Shop in any of the neighborhoods[2] of the island of Manhattan, especially after taking into consideration that some newspapers report that in some Coffee Shops like Eleven Madison Park, people are willing to pay up to US$24 for a simple cup of the drink.

Nowadays, business decisions are taken, not only based on the mere instinct or heart-beats of the people interested in obtaining a good profit , but on the increasingly amount of data that is available from organizations like Foursquare, where customers publish their opinions an ratings about the different venues they have the opportunity to visit, and from public offices or institutes that collect the demographic information from the people of New York.

In the following Peer-graded assignment, the data collected of the different Coffee Shops that are in the Borough of Manhattan, New York, is going to be complemented with some demographics and some income data from the city of New York to answer a simple business question: "Is there any room for a new Coffee Shop in the island of Manhattan?" And if so, in which neighborhood should this new Coffee Shop be located?

2. Data section

The provided datasets are spread across two tables, that are relevant to the chosen question/topic. The two datasets are the following:

Demographics

[1] The hipster subculture is comprised of 20-to-30-year old, highly educated and socio-politically informed millennials. They are middle to upper class, educated, nonconformists who typically live in urban areas.

[2] According to Wikipedia, the neighborhoods of the island of Manhattan are the following: Chinatown, Greenwich Village, Little Italy, Lower East Side, NoHo, SoHo, West Village, Alphabet City, Chinatown, East Village, Lower East Side, Two Bridges, Chelsea, Clinton, Hell's Kitchen, Hudson Yards, Midtown, Gramercy Park, Kips Bay, Rose Hill, Murray Hill, Peter Cooper Village, Stuyvesant Town, Sutton Place, Tudor City, Turtle Bay, Waterside Plaza, Lincoln Square, Manhattan Valley, Upper West Side, Lenox Hill, Roosevelt Island, Upper East Side, Yorkville, Hamilton Heights, Manhattanville, Morningside Heights, Harlem, Polo Grounds, East Harlem, Randall's Island, Spanish Harlem, Wards Island, Inwood and Washington Heights.

Demographic data (population, age, income, etc.) organized alphabetically by Neighborhood Tabulation Area (NTA), with 188 rows & 33 columns (See detail).

demographics

Demographic data (population, age, income, etc.) organized alphabetically by NTA.
188 rows & 33 columns. Size: ~0.1MB.

Field	Type	Description
nta_name	STRING	Name of NTA
borough	STRING	Borough that NTA is located in
nta_code	INTEGER	Identifying code for NTA
population	INTEGER	Total number of people in NTA
age brackets (14 total)	INTEGER	Number of people in given age bracket
median_age	FLOAT	Median age of people in NTA
people_per_acre	INTEGER	Number of people per acre
households	INTEGER	Total number of households in NTA
income brackets (10 total)	INTEGER	Number of households in given income bracket
median_income	INTEGER	Median household income
mean_income	INTEGER	Mean household income

New York Dataset

The New York Dataset has a total of 5 boroughs and 306 neighborhoods and luckily, exists for free on the web. Here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

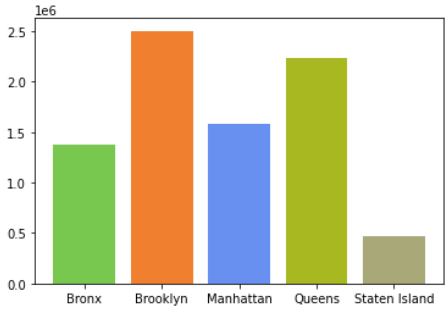
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

3. Methodology

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

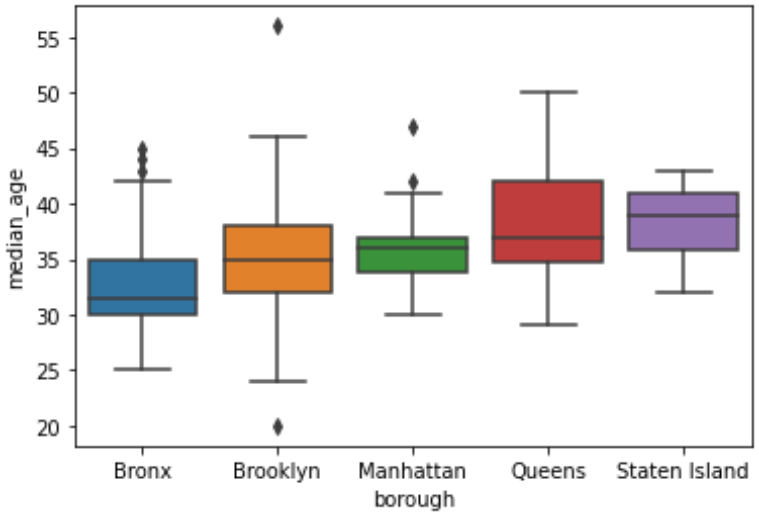
For establishing if a particular neighborhood of New York was a good investment opportunity or not, an Exploratory Data Analysis was realized on some demographics of the city of New York, specifically to determine in which areas were located the people who were younger and had higher economic revenues than the rest.

In the first place, we found that the two most populated boroughs from New York are Brooklyn and Queens.

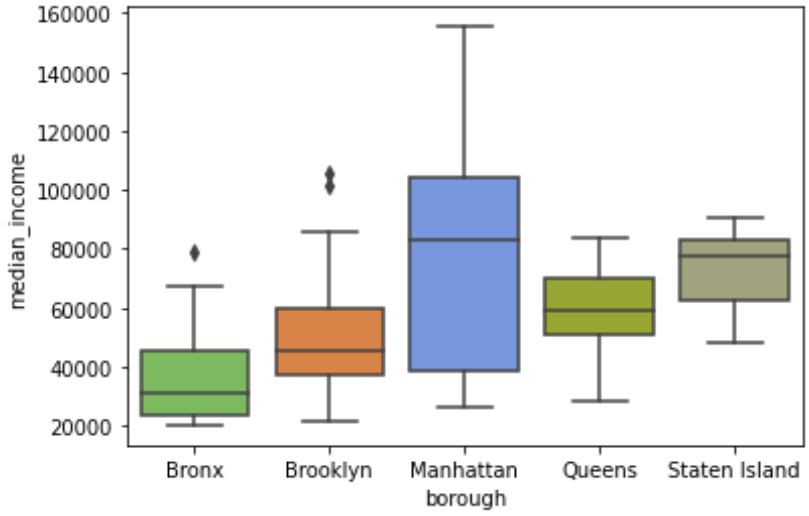


	borough	population
0	Bronx	1372111
1	Brooklyn	2503518
2	Manhattan	1584024
3	Queens	2230286
4	Staten Island	468730

After some manipulations of the data, it was possible to establish that the borough with the bigger number of youths, was Manhattan, followed by Brooklyn and the Bronx.



In the same way, it was possible to establish that the people most affluent people lives equally in Manhattan.

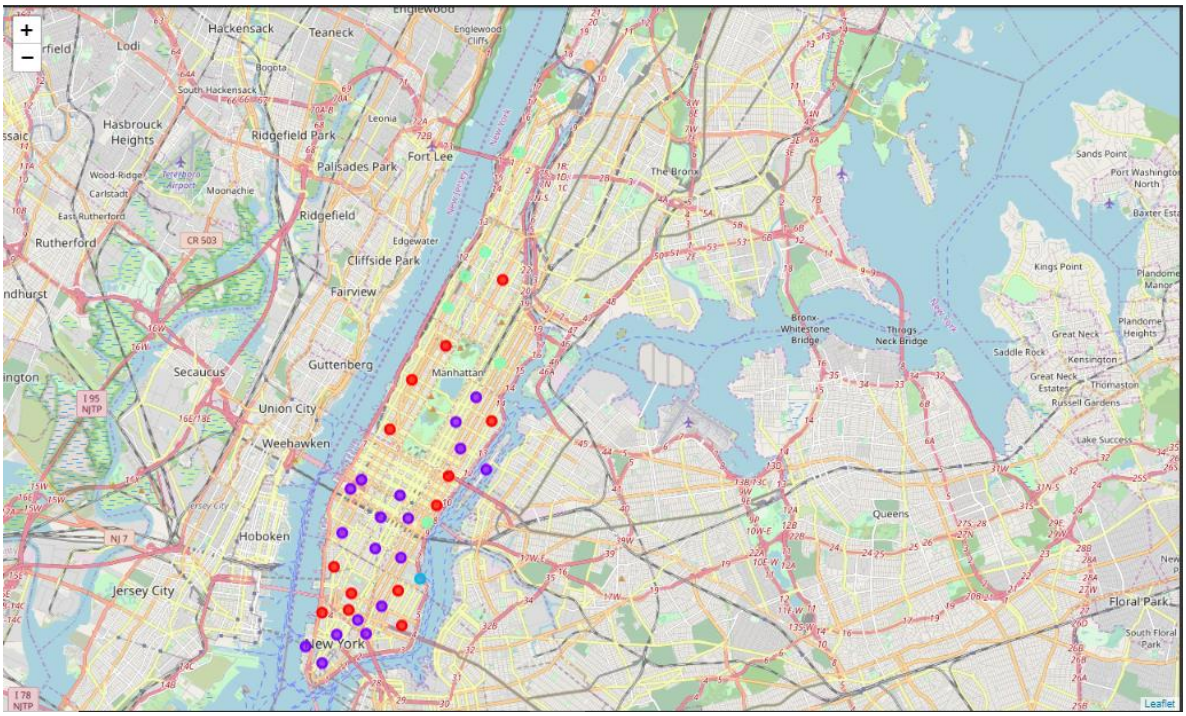


Although the target population lives in Manhattan, a hierarchical clustering algorithm (K-means) was applied to other two different Boroughs of the city (Brooklyn and Queens), to determine if besides the abundant Coffee shops on the island, there were some other market niches in which a new investment was possible. Nevertheless, after checking the different boroughs, it was confirmed that the offer is already aimed to targets with different necessities or desires, like enjoying international foods or going out for a few drinks. The major outcomes of the implementation of the Machine Learning tool and the principal traits of the different traits of each of the three boroughs, are going to be presented on the next section.

4. Results

Manhattan

After doing a segmentation of Manhattan, Brooklyn and Queens to determine in which of the three areas it was more feasible to open a new Coffee Shop, it was found that although there were some scattered shops throughout all the city, they were specially located in the borough of Manhattan.



Although five clusters were build for the borough of Manhattan, three main segments were identified: Cluster 1, where the Coffee Shops are the second most popular venue after the Italian restaurants, Cluster 2, where the Coffee Shops are the most popular venue in neighborhoods like Midtown, Little Italy, the Financial District and the Civic center, and Cluster 5, where the Coffee Shops compete in popularity with ethnic restaurants, specialized in Mexican and Chinese recipes.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	Central Harlem	Seafood Restaurant	Bar	African Restaurant	American Restaurant	Chinese Restaurant
8	Upper East Side	Italian Restaurant	Coffee Shop	Bakery	Gym / Fitness Center	Exhibit
9	Yorkville	Italian Restaurant	Coffee Shop	Gym	Bar	Deli / Bodega
10	Lenox Hill	Italian Restaurant	Coffee Shop	Sushi Restaurant	Pizza Place	Cocktail Bar
12	Upper West Side	Italian Restaurant	Bakery	Bar	Coffee Shop	Thai Restaurant

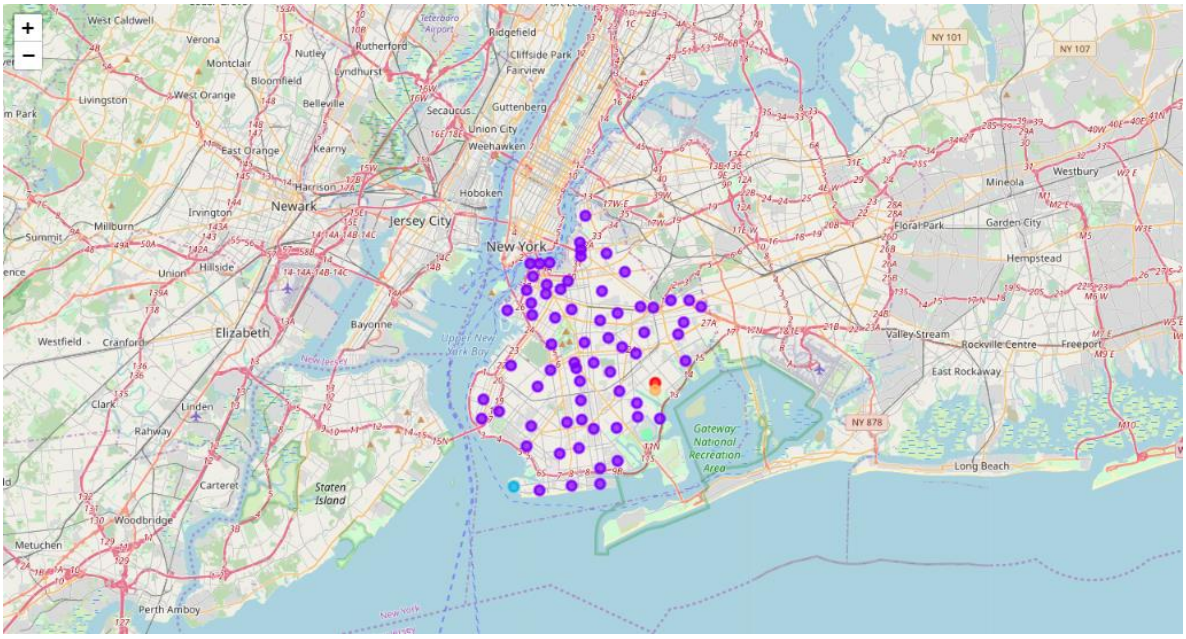
Cluster 2

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Chinatown	Chinese Restaurant	Cocktail Bar	Bakery	American Restaurant	Spa
14	Clinton	Italian Restaurant	Gym / Fitness Center	Theater	Coffee Shop	Spa
15	Midtown	Coffee Shop	Hotel	Theater	Sandwich Place	Sporting Goods Shop
16	Murray Hill	Sandwich Place	Hotel	Bar	Coffee Shop	Sushi Restaurant
22	Little Italy	Café	Bakery	Hotel	Bubble Tea Shop	Sandwich Place
28	Battery Park City	Park	Hotel	Gym	Boat or Ferry	Coffee Shop
29	Financial District	Coffee Shop	Bar	Pizza Place	Hotel	Cocktail Bar
32	Civic Center	Coffee Shop	French Restaurant	Hotel	Cocktail Bar	Yoga Studio

Cluster 5

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Chinatown	Chinese Restaurant	Cocktail Bar	Bakery	American Restaurant	Spa	Dessert Shop	Salon / Barbershop	Vietnamese Restaurant	Hotpot Restaurant	Optical Shop
2	Washington Heights	Café	Bakery	Grocery Store	Deli / Bodega	Chinese Restaurant	Mobile Phone Shop	Mexican Restaurant	Supplement Shop	Coffee Shop	Latin American Restaurant
3	Inwood	Mexican Restaurant	Café	Lounge	Pizza Place	Restaurant	Chinese Restaurant	Park	Bakery	American Restaurant	Frozen Yogurt Shop
4	Hamilton Heights	Pizza Place	Café	Coffee Shop	Deli / Bodega	Mexican Restaurant	Sushi Restaurant	Cocktail Bar	Sandwich Place	Bakery	Yoga Studio
5	Manhattanville	Coffee Shop	Seafood Restaurant	Italian Restaurant	Park	Mexican Restaurant	Chinese Restaurant	Spanish Restaurant	Check Cashing Service	Bank	Bar
7	East Harlem	Mexican Restaurant	Bakery	Deli / Bodega	Thai Restaurant	Latin American Restaurant	Convenience Store	Spanish Restaurant	Liquor Store	Gas Station	Taco Place

Brooklyn



Although we also identified five clusters in Brooklyn, we found, on the other hand, that its clusters are mainly dedicated to the nightlife, Pizzas and Bars (Cluster 1), but also to Caribbean foods, as it turned out to be the case of Cluster 4.

Cluster 1

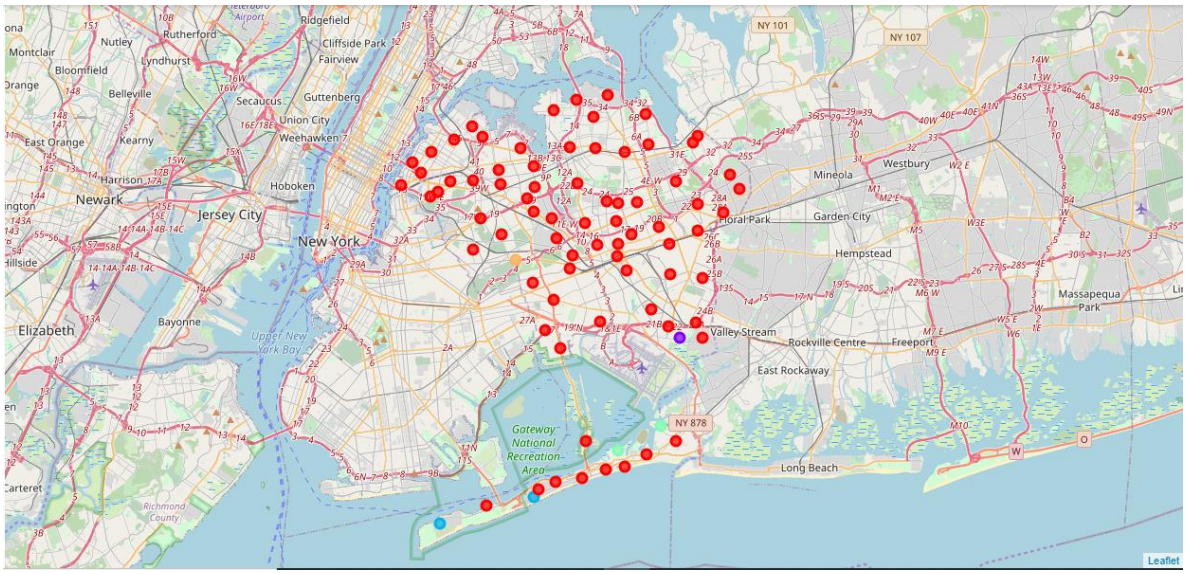
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bay Ridge	Italian Restaurant	Pizza Place	Spa	Bar	Greek Restaurant
3	Greenpoint	Bar	Pizza Place	Cocktail Bar	Coffee Shop	Yoga Studio
12	Windsor Terrace	Deli / Bodega	Café	Park	Grocery Store	Diner
13	Prospect Heights	Bar	Mexican Restaurant	Wine Shop	Gourmet Shop	Bakery
14	Brownsville	Restaurant	Park	Chinese Restaurant	Moving Target	Pool
15	Williamsburg	Bar	Coffee Shop	Grocery Store	Bagel Shop	Yoga Studio

Cluster 4

47	Prospect Park South	Caribbean Restaurant	Pizza Place	Fast Food Restaurant	Mobile Phone Shop	Grocery Store
54	Ditmas Park	Caribbean Restaurant	Pizza Place	Deli / Bodega	Women's Store	Chinese Restaurant
56	Rugby	Caribbean Restaurant	Bank	Grocery Store	Salon / Barbershop	Pizza Place
57	Remsen Village	Caribbean Restaurant	Fast Food Restaurant	Deli / Bodega	Fish Market	Salad Place
69	Erasmus	Caribbean Restaurant	Grocery Store	Yoga Studio	Food Truck	Supermarket

Queens

Queens, on its hand, turned out to be the are where most of the clusters are made from venues like Bodegas, parks and hotels and latino restaurants (Cluster 1), so mingled, that we could also consider the possibility of treating the whole borough as a single cluster.



Cluster 1

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Astoria	Bar	Middle Eastern Restaurant	Greek Restaurant	Hookah Bar	Seafood Restaurant
1	Woodside	Grocery Store	Latin American Restaurant	Bakery	Thai Restaurant	Filipino Restaurant
2	Jackson Heights	Latin American Restaurant	Peruvian Restaurant	South American Restaurant	Bakery	Mexican Restaurant
3	Elmhurst	Thai Restaurant	Mexican Restaurant	South American Restaurant	Vietnamese Restaurant	Chinese Restaurant
4	Howard Beach	Italian Restaurant	Pharmacy	Bagel Shop	Deli / Bodega	Sandwich Place
5	Corona	Mexican Restaurant	Convenience Store	Bakery	Deli / Bodega	Pizza Place
73	Sunnyside Gardens	Bar	Grocery Store	Pizza Place	Turkish Restaurant	Thai Restaurant
74	Blissville	Deli / Bodega	Hotel	Donut Shop	Rental Service	Movie Theater
75	Roxbury	Baseball Field	Deli / Bodega	Irish Pub	Beach	Trail
76	Middle Village	Playground	Diner	Sandwich Place	Farmers Market	South American Restaurant
77	Malba	Rest Area	Tennis Court	Latin American Restaurant	Vegetarian / Vegan Restaurant	Rock Club
78	Hammels	Beach	Deli / Bodega	Building	Fried Chicken Joint	Bus Station
80	Queensbridge	Hotel	Hotel Pool	Athletics & Sports	Spanish Restaurant	Beer Garden

5. Discussion

After finishing the Exploratory Data Analysis of the different boroughs of New York and finishing also, the clustering exercise in which the K-means algorithm played an outstanding role, it became evident that neither all boroughs nor or neighborhoods are good places to open a new Coffee Shop in New York, since the habit of going to a place and asking a cup of Coffee, contrary to our initial beliefs, seems to be popular only in the island of Manhattan and in the same neighborhoods where there are also restaurants dedicated to Italian food.

The Clustering exercise prove to be a very powerful tool to identify which neighborhoods should be considered as the most appropriate to open a new Coffee Shop (like Midtown, Little Italy, the Financial District and the Civic center) for a very simple reason: the people have a mind map of the city in which they know where they can go to take a cup of Coffee and where not. Besides, after doing the exploratory analysis of the Demographic Data of the City of New York, it became evident that the people who have the age and the money to pay for an expensive cup of Coffee, live mainly in the island of Manhattan.

6. Conclusion

After finishing the K-means clustering exercise for the venues of the city of New York to see where we could open a new Coffee Shop, we could conclude that:

In Manhattan, most of the venues are traditional restaurants (i.e. restaurants dedicated, for example, to Italian foods) and the second most popular venues are Coffee shops, although there are some neighborhoods (like Midtown, Little Italy, the Financial District and the Civic center) where everybody in New York, now they can go to get a cup of Coffee. Besides, after analyzing some demographics, it became evident that the people who have the age and the money to pay for an expensive cup of Coffee, live mainly in the island of Manhattan.