The Battle of Neighborhoods

Yann PERRIN

23/11/2020

Introduction

An entrepreneur who owns a coffee shop in Flatbush neighborhood (in New York City) wants to open another coffee shop. And he wants to know the best location to do so.

Because his first coffee shop goes quite well, he wants us to open his new shop in a neighborhood quite similar to Flatbush. That way, he expects that similar neighborhood will give similar success for his new coffee shop.

However, he would also like to go in a neighborhood where the life standard is greater or equal to Flatbush, in order to be sure that the quality standard of his coffee shop will be still adequate.

Last, he considers that if there are too many coffee shops in a same area, the business will go bad. So he wants the new neighborhood to have a smaller density of coffee shop than Flatbush.

So the study we will make is basically to list neighborhoods that are similar to Flatbush, and to plot on a map those neighborhoods where he could open a new coffee shop.

In a second time, we will work on that list to keep only the neighborhoods where the median income is higher than in Flatbush, in order to avoid neighborhood with lower life standard. Last we will identify in this short list the neighborhoods where we could consider there is a a smaller density of coffee shop than in Flatbush.

Code used for the project :

https://github.com/YannPerrin/Applied-Data-Science-Capstone-YP/blob/main/The%20Battle%20of%20Neighborhoods%20YP.ipynb

Data used and global steps for the study

Links to data used

We will use data on New York city neigborhoods to do the study. We will need to have data on the population and income in each neighborhoods. We will also need to have geographic localization of neighborhoods in order to get venues nearby and also for plotting neighborhoods on a map.

We will use the newyork_data.json file of the Module 3 to get the list of neighborhoods and their geographic coordinates. We can find this file at the following link :
https://github.com/YannPerrin/Applied-Data-Science-Capstone-YP/blob/main/newyork_data.json

We will also use data from the following site to have the population and median income by neighborhood：https://geodacenter.github.io/data-and-lab/NYC-Nhood-ACS-2008-12/
The file itself is saved at the following link : https://github.com/YannPerrin/Applied-Data-Science-Capstone-YP/blob/main/NYC_Nhood%20ACS2008_12.dbf

Because neighborhoods are cited by NTA code, we will also use a table that gives use the correspondance between NTA code and neighborhood name ：
https://www1.nyc.gov/assets/planning/download/office/data-maps/nyc-population/census2010/nyc2010census_tabulation_equiv.xlsx
The file used (after removing some headlines) is also saved at the following link :
https://github.com/YannPerrin/Applied-Data-Science-Capstone-YP/blob/main/nyc2010census_tabulation_equiv.xlsx

Global steps of the study

For the study, we will proceed as follow:

Import libraries

We will first import libraries required for analysing and plotting the data

Import input datas

We will then read the neighborhoods data and import it in a dataframe
We will also read the population and income data and import it in a dataframe
We will then merge those dataframes, and check we don't loose too many neighborhood in the process
As a result of this step, we will have a dataframe with for each neighborhood : its latitude and longitude, its population, the median income

Get the venues

We will then get the venues in each neighborhood through the Foursquare API, and count the number of venues of each type in each neighborhood
The dataframe we will now have will have the number of venues of each type, for each neighborhood

Cluster the neighborhoods

We will then scale the number of venues with the population of the neighborhood, and then work with the number of venues per million inhabitants
We will last cluster the neighborhoods using the number of venues for each type of venue and be able to answer to the first question and plot the required map

Go further with the list of similar neighborhoods

For the second question, we will remove the neighborhoods where the median income is lower than in Flatbush in order to respect the life standard criteria
We then will remove the neighborhoods where there are more coffee shops per million people than in Flatbush, which will respect the second criteria of density of coffee shops
At the end, we will have a short list of neighborhoods respecting all the criterias

Methodology

Importing and Cleaning Data

The first step in the study is to import and clean the data used.

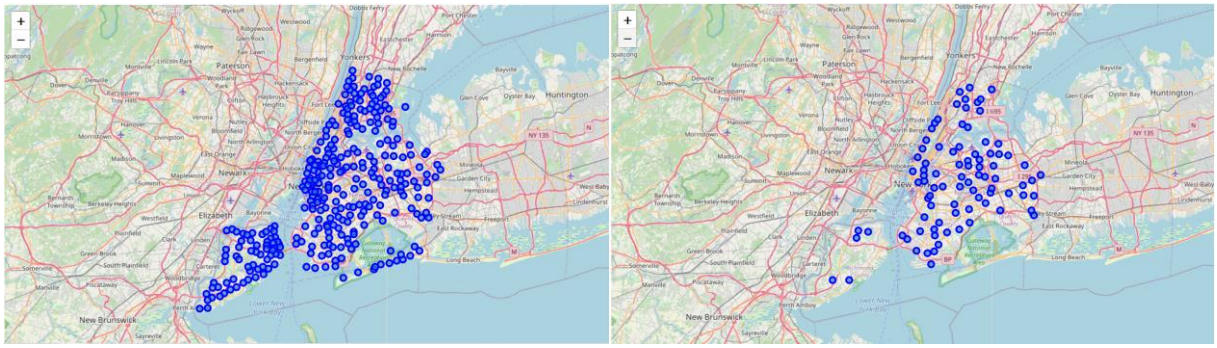
We first have a list of 306 neighborhoods with their latitude and longitude, thanks to the json file.

On another hand, we have a list of 195 neighborhoods with population and income data. In this list, the neighborhoods are called by their NTA code. So we need to make the correspondence between this NTA code and the neighborhoods of the first dataframe. This is done through a census excel.

When we have done this process, we can merge all the data together, in order to get a dataframe showing the neighborhoods with their latitude, longitude, population and median income. The dataframe looks like this:

	NTA	Total population	Median income	Neighborhood	Borough	Latitude	Longitude
0	BK90	33155	519058	East Williamsburg	Brooklyn	40.708492	-73.938858
1	QN23	24199	354073	College Point	Queens	40.784903	-73.843045
2	SI54	43427	718593	Great Kills	Staten Island	40.549480	-74.149324
3	QN05	28201	506307	Rosedale	Queens	40.659816	-73.735261
4	BX27	27562	149520	Hunts Point	Bronx	40.809730	-73.883315

However, through this process, we have lost a lot of neighborhood. From 306 neighborhoods in the json file (which is the correct number of neighborhood in NYC), we have come down to 88. This is due to lack of data in the data used. In this study we want to have a good coverage of the city, so we plot the neighborhoods on maps and visually compare the maps obtained. We can see below the map with 306 neighborhoods on the left, and the 88 neighborhood we will base our study on on the right.



The loss of data is obvious. However we still have neighborhoods representing all of the city, even though staten island is poorly represented.

This point is a weakness of the study, but having no better data right now, we will go on with that.

Exploring and Analysing Data

First answer, based on clustering

The first part of the study is based on clustering the neighborhoods in order to identify the ones that are similar to Flatbush.

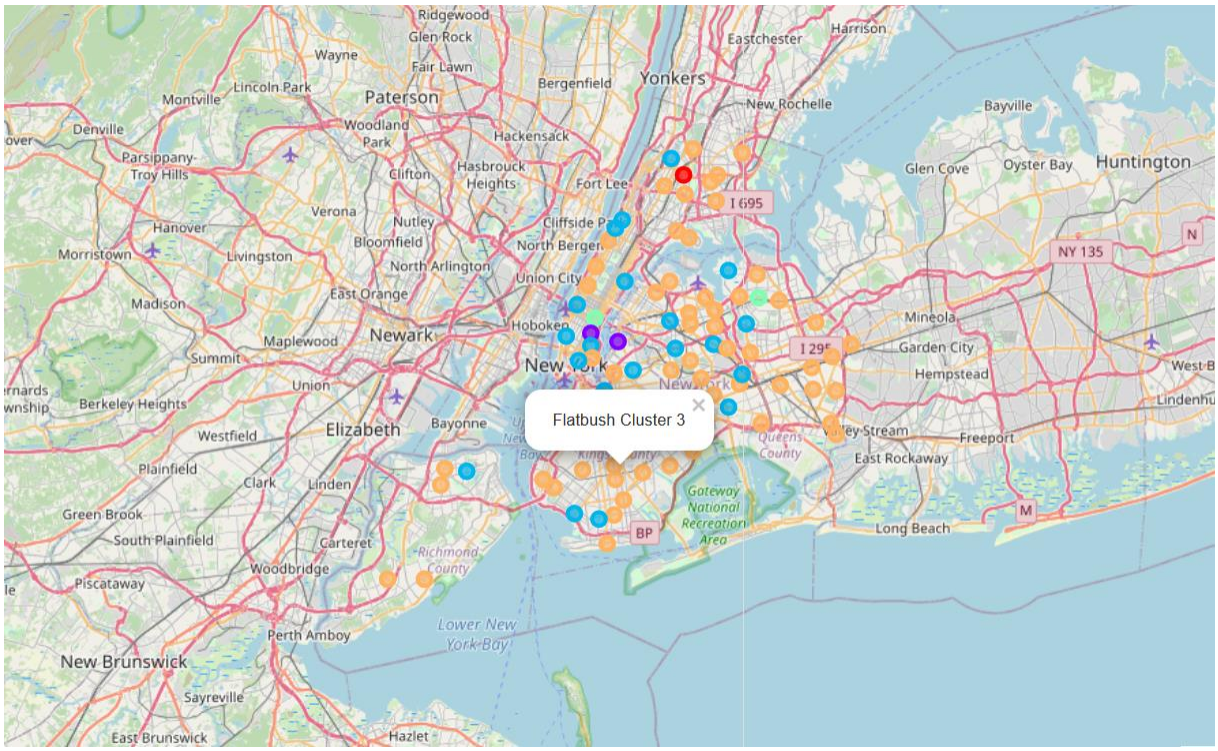
To do so, we will base our clustering on the venues that are located in the neighborhood.

So we first get the venues in a radius of 1000 m around the location of the neighborhood. We do it with the Foursquare API. We then process the results to get a dataframe where for each row representing one neighborhood, we have for each venue type the number of venues in this neighborhood.

Then we decide to scale the number of venues with the population of the neighborhood. It seems more logic to compare the number of venues per million inhabitants, rather than the number of venues itself.

This data processed is then used in a Kmeans clustering method in order to group all the neighborhoods in 5 clusters.

We can plot the neighborhood with a color by cluster as shown below. We can see on this map the neighborhoods similar to Flatbush in orange (Cluster 3).



We also get the list of similar neighborhoods :

Great Kills
Rosedale
Hunts Point
Whitestone
Starrett City
East New York
Canarsie
Westerleigh
Steinway
Homecrest
St. Albans
Bronxdale
Corona
East Tremont
Longwood
Borough Park
East Elmhurst
Auburndale
Parkchester
Cambria Heights
Hollis
Jackson Heights
Morningside Heights
Madison
South Jamaica
Co-op City
Brighton Beach
Port Richmond
Pelham Parkway
Mount Hope
Upper West Side
Richmond Hill
Lincoln Square
Arden Heights
Woodhaven
Laurelton
Queens Village
Bellerose
South Ozone Park
Williamsburg
Oakland Gardens
Midwood
Windsor Terrace
Flatlands
Norwood
Erasmus
Ridgewood
Lower East Side
Dyker Heights
Bay Ridge
Flatbush
Middle Village
Elmhurst
Glendale
Flushing
Kew Gardens Hills
Ocean Hill
Brownsville
Forest Hills
North Corona
Astoria

Second answer, adding criterias

In the second part of the study, we will get the previous list and remove neighborhoods that do not meet following criterias :

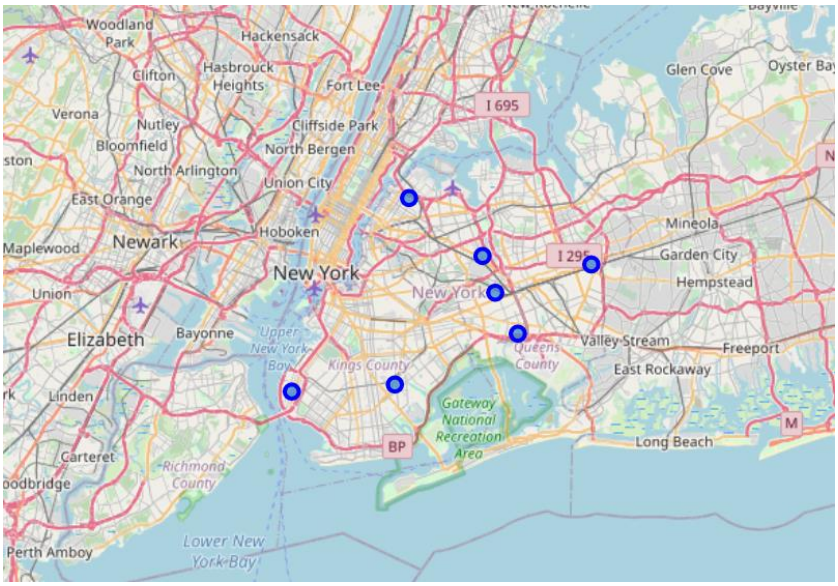
- A median income equal or higher than the one in Flatbush : this is supposed to keep only the neighborhood with higher life standard
- A density of coffee shops smaller than in Flatbush : the idea is to avoid opening a new coffee shop in an area where there are already (too) many of them

The first criterion is straightforward because we have the median income for each neighborhood.

The second one is checked with the number of coffee shops per million inhabitants in the neighborhood.

Having processed that, we get the following neighborhoods that meet both the similarity criterion and the two additional criterions, also plotted on a map :

Richmond Hill
Queens Village
South Ozone Park
Flatlands
Bay Ridge
Forest Hills
Astoria

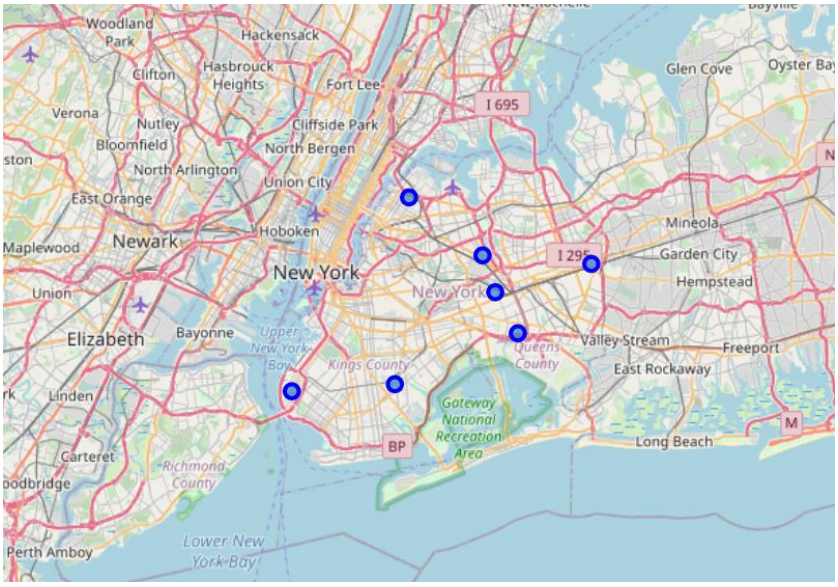


Results and Discussion

At the end of the study we have a short list of 7 neighborhoods that meet all criterias.

It is then up to the entrepreneur to choose the location he wants to consider. For instance, he could choose Astoria in the Queens. That would enable to open the new coffee shop in a different borough.

Richmond Hill
Queens Village
South Ozone Park
Flatlands
Bay Ridge
Forest Hills
Astoria



When printing the numbers of coffee shops in the list of proposed neighborhoods, it appears that all neighborhoods but Astoria has no coffee shop. That respects the criterion of not having a high density of coffee shops, but it raises the question of why there is none.

It might be interesting either to go to Astoria (where there is some), or at least to understand why there is no coffee shops in other neighborhoods and whether it could signal a risk of a business not going well.

Conclusion

In this study, we have clustered neighborhoods based on their number of venues per million inhabitants and by type of venues. It enabled to identify neighborhoods similar to Flatbush, which is interesting for a business extension problem.

We have also shorten the list by adding filter such as median income and number of coffee shop venues per million people in the neighborhood.

That has enabled to identify a short list of neighborhood meeting the criterias set for choosing a neighborhood for opening a new coffee shop.

At the end of the study, we need to keep in mind that not all 306 neighborhoods of New York City have been considered, because of a lack of data. This is a way of improvement for this study.

Also, we could improve the study by adding some data to understand if the neighborhoods that meet criterions but have no coffee shop so far are at risk for running such a business.