The Battle of Neighborhoods

# 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.

# Data used and global guidelines for the study

We will use data on New York city neigborhoods.  
We will use the newyork\_data.json file of the Module 3 to get the list of neighborhoods and their geographic coordinates.  
We will also data from <https://geodacenter.github.io/data-and-lab/NYC-Nhood-ACS-2008-12/> to have the population and median income by neighborhood  
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>

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  
A the end, we will have a short list of neighborhoods respecting all the criterias