The Battle of the Neighborhoods

Finding the beast area in Manhattan to open an Italian restaurant

Date: Dec 5, 2019

Author: Aurélien Robert

Table of contents

Introduction: Business Problem

The objective is to help an entrepreneur to find the best place to open a new gastronomic **Italian restaurant** in **Manhattan**. For this study we consider the "restaurant market" is very mature in Manhattan. It means that we will consider that places where the number of restaurants is very low are locations with low potential (e.g. residential places) and not locations with high potential. For Manhattan, the notion of city center does not really exist, so we cannot measure the distance between potential locations and a "city center". As a result, the approach taken to solve the problem will be the following one:

1. **find zones with a certain density of restaurants**
2. segment those zones in order to **find different profiles** (e.g. Asian/Chinese profile, Latin/Italian profile, TexMex...)
3. within each profile/segment, **assess the potential of each zone** (e.g. a zone being tagged as Latin/Italian with fewer Italian restaurant than average)

Data

The first need is to have a **map of Manhattan** (with neighborhoods, coordinates of Manhattan frontiers...).
Source: https://raw.githubusercontent.com/Aurelrobert/Coursera_Capstone/master/NY.geojson
Example :

	Borough	Neighborhood	Latitude	Longitude	Polygon
8	Manhattan	Battery Park City	40.71125	-74.01608	POLYGON ((-74.01375400000018 40.71369, -74.014...
49	Manhattan	Central Park	40.78247	-73.96557	POLYGON ((-73.94965699268657 40.79706872127114...
51	Manhattan	Chelsea	40.74661	-73.99981	POLYGON ((-73.99346400122367 40.75218999709955...
53	Manhattan	Chinatown	40.71578	-73.99571	POLYGON ((-73.99999096973443 40.71797674709695...
56	Manhattan	Civic Center	40.71387	-74.00285	POLYGON ((-73.9979936899674 40.71272235466727,...

Based on this map, we will first define zones: a zone being defined as a perimeter around a given place.
We will then create a grid, within Manhattan frontiers, with a certain distance (e.g. d = 500m) between nodes and make sure zones cover the full Manhattan area (at this stage we do not care if zones overlap): the radius of each zone will be equal to d/sqrt(2).
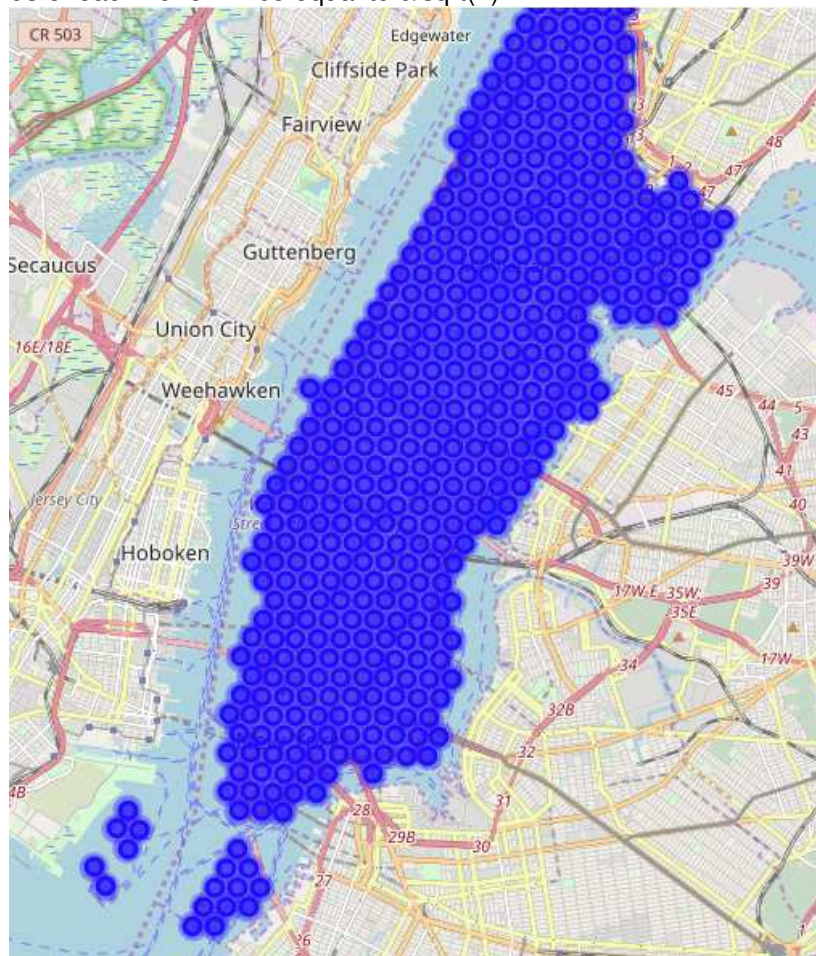
For each zone we will retrieve the **list of restaurants** (whatever the type) using **Foursquare** data.
Source: Foursquare API
Example:

	Node	Node Latitude	Node Longitude	Venue ID	Venue	Venue Latitude	Venue Longitude	Venue Category	Node Number Venues
0	1	40.68270	-74.02170	4d6ee122b45b59417c6a6778	Kimchi Taco Truck	40.68216	-74.02331	Food Truck	1
1	5	40.68914	-74.04527	4c3f973dcc410f47db6aa861	Statue of Liberty Crown Cafe	40.68986	-74.04635	American Restaurant	3
2	5	40.68914	-74.04527	558c3c89498eb8fa9fad3f2f	Bill's "World Famous" Lemonade	40.69017	-74.04691	Food Truck	3
3	5	40.68914	-74.04527	4caf5c8c75ebb60c74a3bbad	Liberty Island Food Court	40.68985	-74.04640	Food Court	3
4	6	40.68894	-74.02161	582270dab078be59c611c926	Jolly Fat Pig	40.68917	-74.02233	Food Truck	2

We will then compute the density of each type of restaurant and segment/cluster the different zones.
Here is what we obtain when clustering the zones:
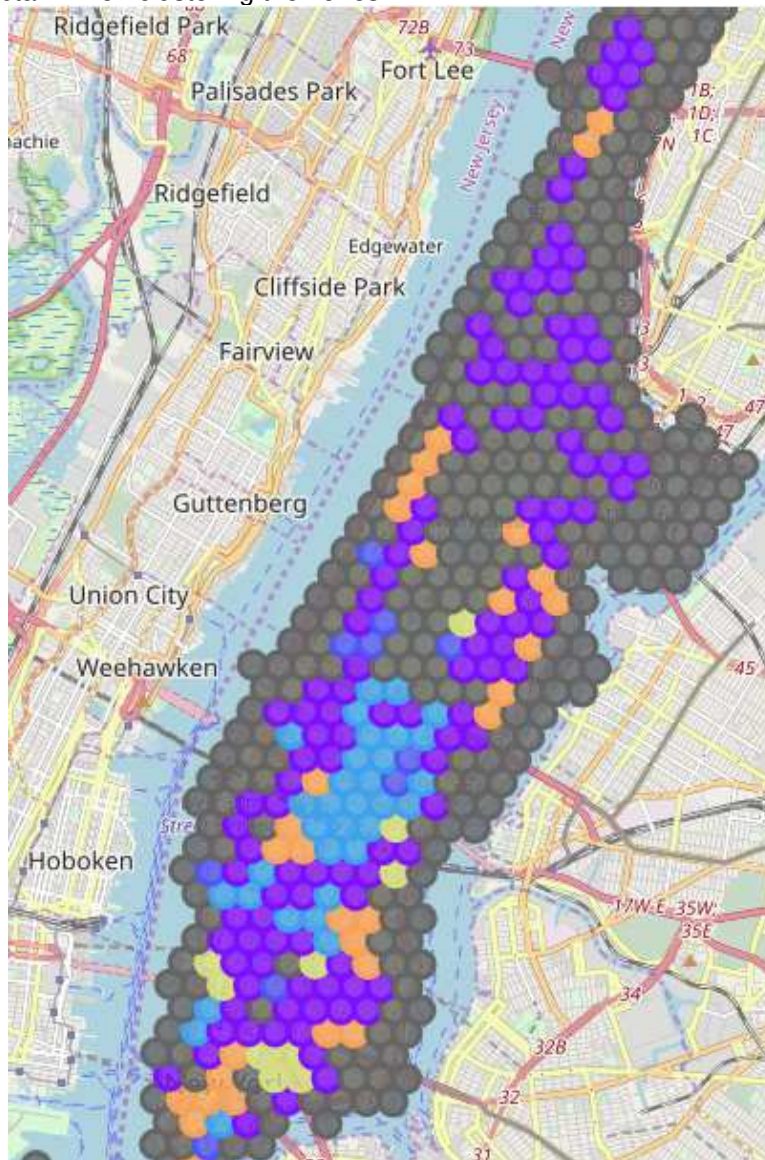
Among the different clusters, we will be interested by zones with a "lack" of Italian restaurants: this lack will be computed as the distance between the density of Italian restaurant of the zone and the average of the same density for the cluster, only for zones being under the average.
Here is what we obtain: zones in **Green** have a higher potential than zones in **Red**
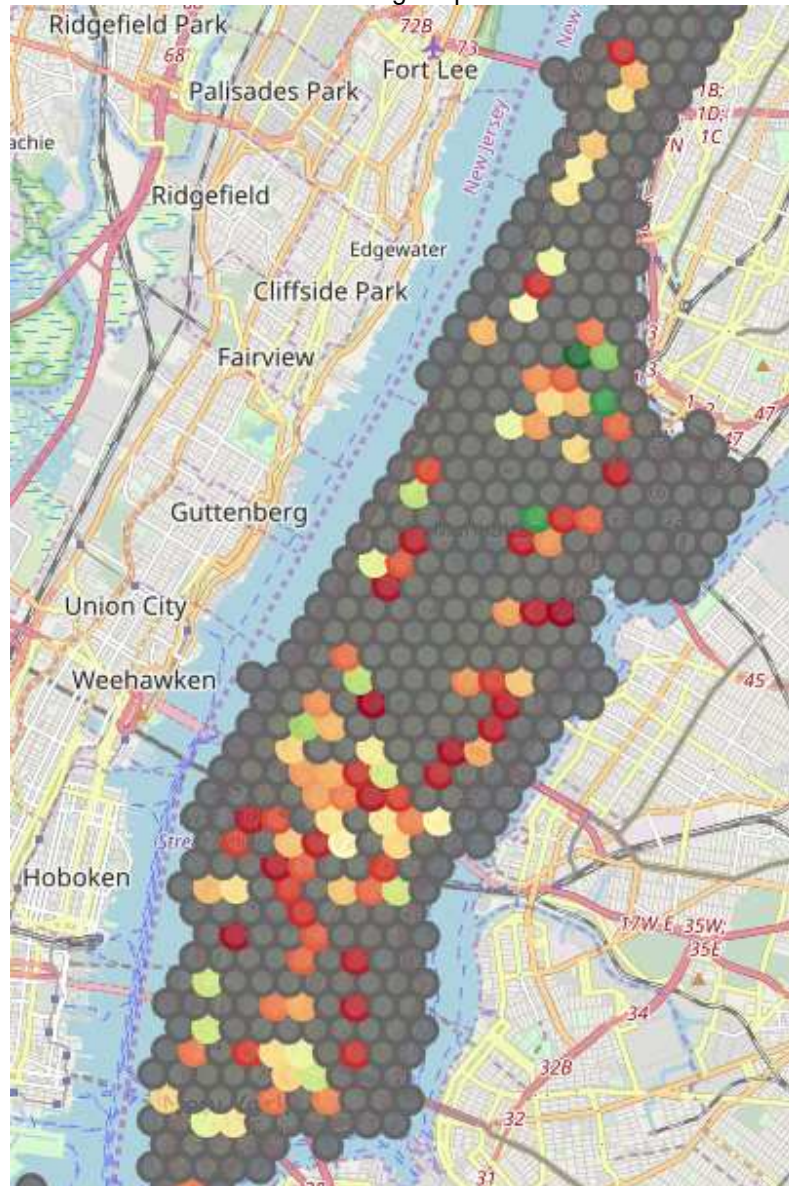
Once zones with the best potential have been found, we will **get the ratings of the Italian restaurants** of the top 20 zones in order to assess the quality of the future competitors. In order to let the entrepreneur decide, the results of the study will be **displayed as a map**, with zones colored depending on their potential, and restaurants with **good ratings (>8) displayed on the map** for the top 20 zones

Methodology

Now we have all Manhattan venues, we will first segment the different zones in order to identify zones profiles.
Segment zones based on the different restaurant types density. In order to have a significant number of restaurants in the zone, we keep only the **2 first quartiles** (i.e. **zones with more than 13 restaurants**)
Then, for each cluster we will consider that zones having a **density under the cluster average** (for a given type of restaurant) **has a potential to increase**.
This will give the **list of 20 zones** (with the highest potential) we will focus on.
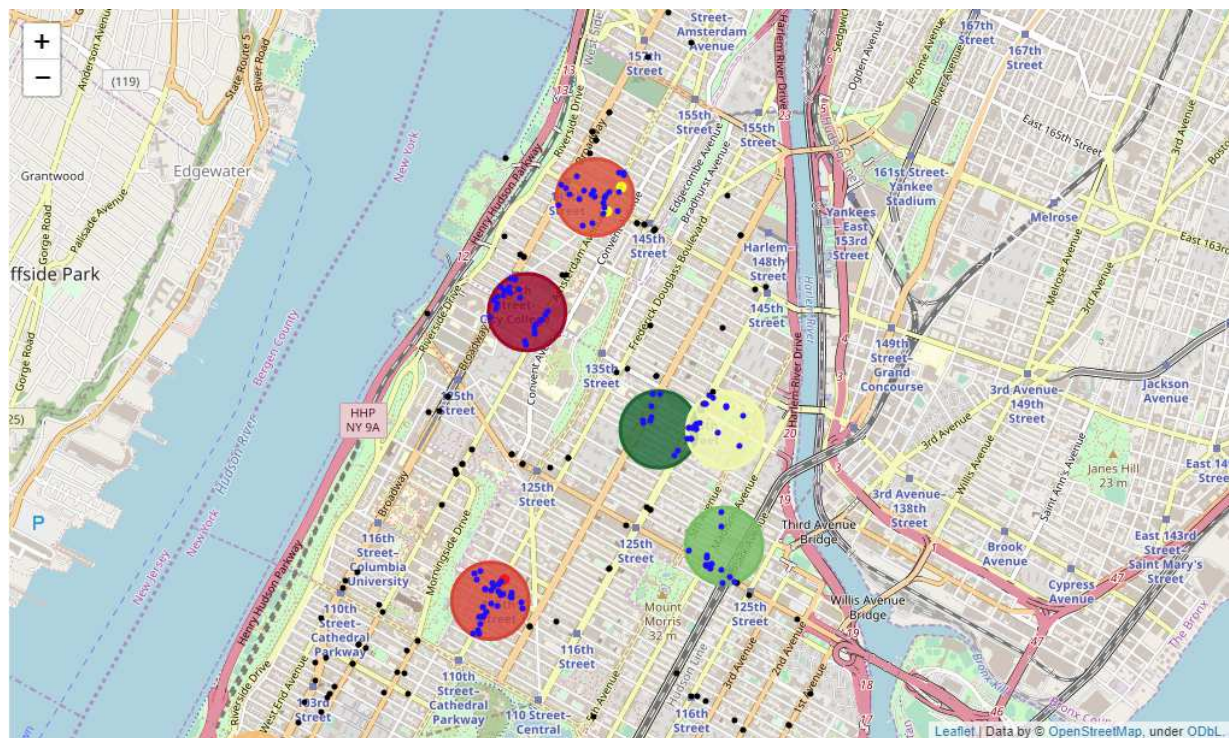For the selected zones, we will then retrieve the **ratings of the potential competitors** in order to find the best place to open a new restaurant (i.e. not too close from a competitor with a good rating - i.e. more than 8).
Results will be **displayed in a map**, with the 20 selected zones and the places where the competitors are located.

Analysis
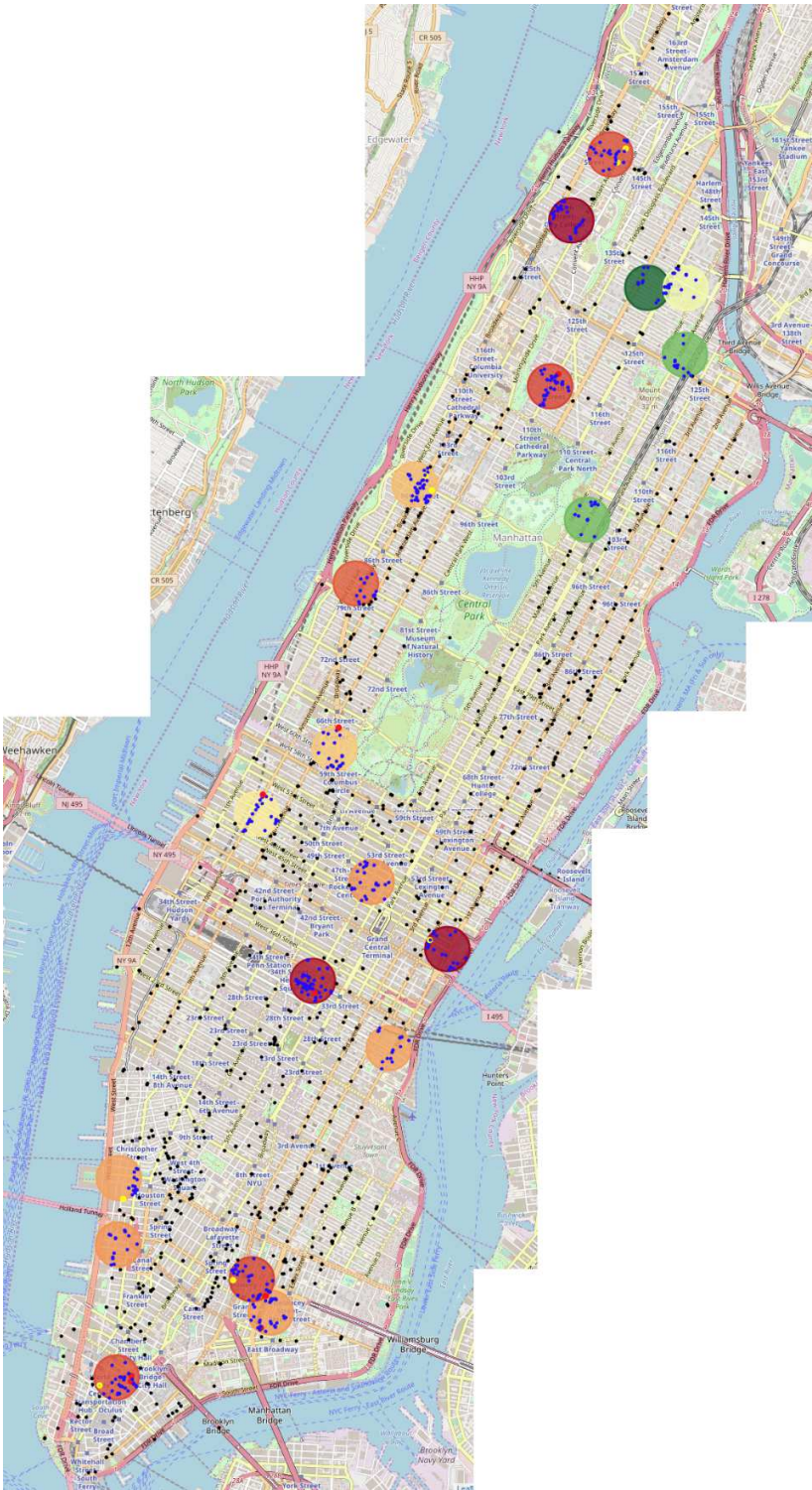
Zoom on Harlem:

The maps display:

1. **Top 20 zones** with highest potential are display, with a **color based on their potential**.
   Green is the highest potential
2. **In the top 20 zones**
   A. Competitors with a **good rating (>= 8)** are displayed with a **red marker**
   B. Competitors with a **low rating (< 8)** are displayed with a **yellow marker**
   C. **Non-italian** restaurants are displayed in **blue**
      It shows that zones are not in a no restaurant zone
3. all the Italian Restaurants and Pizza Places of Manhattan (in black).
   It shows the density of italian restaurant

Results and Discussion

Our analysis shows that we have zones with an interesting potential in different neighborhoods of Manhattan: the entrepreneur will have the choice between really different options.
The study also shows that there are zones with very few (or no) Italian restaurants but we other kind of restaurants. In those places, the entrepreneur will have few competitors. For zones with few Italian restaurants it is also interesting to check where the good competitors are. It is why we have displayed Italian restaurants with a rating greater than 8.

This study is just a starting point. It gives an overview of the potential interesting zones for opening an Italian Restaurant, but a deeper analysis is needed to understand why there is a lack of Italian Restaurants. For example, we see that the density of Italian restaurants is higher in the south than in the north (Harlem). It is clear that there are more potential clients Downtown than in Harlem (an not only for Italian restaurants) but before opening a restaurant in Harlem we have to make sure that people in Harlem are interested enough by Italian restaurants.

Conclusion

Purpose of this project was to identify Manhattan areas with a density of Italian restaurants under the average in order to find the optimal location for a new Italian restaurant. By calculating restaurant density distribution from Foursquare data we have first identified area profiles (by clustering the zones), and then computer for each zone if the density in Italian restaurant is under the average of its cluster or not. The distance between the density of the zone and the density of the cluster has given the growth potential of the zone. This analysis has given a list of 20 interesting zones for a deeper analysis. We have then analyzed the location of major competitors in order to make sure we do not recommend a zone surrounded by Italian restaurants with a very good rating.

Final decision on optimal restaurant location will be made by the entrepreneur based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location, levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.