Clustering NYC Italian Restaurants

Andrea Sarni

August 04, 2021

Contents

[1. Introduction 2](#_Toc79342979)

[1.1 Background 2](#_Toc79342980)

[1.2 Problem 2](#_Toc79342981)

[1.3 Interest 2](#_Toc79342982)

[2. Data acquisition and cleaning 2](#_Toc79342983)

[2.1 Data sources 2](#_Toc79342984)

[2.2 Data definition and cleaning 3](#_Toc79342985)

[3. Python libraries and Descriptive Analysis 4](#_Toc79342986)

[3.1 Python libraries 4](#_Toc79342987)

[3.2 Geographic data 4](#_Toc79342988)

[3.3 Demografic data 5](#_Toc79342989)

[3.4 Restaurants data 6](#_Toc79342990)

[4. Neighborhoods Clustering 9](#_Toc79342991)

[5. Final Discussion 11](#_Toc79342992)

[6. Future Developments 13](#_Toc79342993)

Clustering NYC Italian Restaurants

Andrea Sarni

August 04, 2021

# Introduction

## Background

New York is the most populous city in the United States. With an estimated 2020 population of 8,253,213 distributed over about 302.6 square miles (784 km2), New York City is also the most densely populated major city in the United States.

New York City is composed of five boroughs: Brooklyn, Queens, Manhattan, the Bronx, and Staten Island.

During the late 19th and early 20th centuries, many Italians coming from Naples and Sicily moved to large American cities, such as New York City, Philadelphia, Chicago, St. Louis, Boston, Los Angeles, and San Francisco. This immigration flow brought Italian cuisine in US (thus largely derived from Neapolitan and Sicilian tradition), which has gradually become more influential in the American diet. In fact, it is one of the top three cuisines in the United States, according to the National Restaurant Association.

A young American investor, Mister X would like to take this business opportunity and decided to open an Italian Restaurant in NY.

## Problem

Data may give a precious contribute for the selection of the best neighborhood in NY to locate an Italian restaurant considering 2 categories of variables for each neighborhood:

* **Demographic data**: qualitative and quantitative information about the population living in the neighborhood;
* **Restaurant data**: qualitative and quantitative information about restaurants in the city.

## Interest

Based on Food Establishment Inspections data from nyc.gov, Italian Restaurants is one of the most common category of restaurants in NY, around 1k restaurants out of 20k total restaurants in NYC (Indian, French or Thai cuisines count only 300 restaurants each!), it is very important for Mister X to investigate which should be the best borough and neighborhood to locate his new activity, in order to have the best solution in terms of attractivity and profitability.

# Data acquisition and cleaning

## Data sources

As stated in paragraph 1.2, the goal is to collect and analyze two categories of variables: demographic and restaurant.

The first category of variables is available on <https://www.data.cityofnewyork.us> (Population and Area) and through the Kaggle dataset <https://www.kaggle.com/muonneutrino/new-york-city-census-data?select=nyc_census_tracts.csv> (for income per cap data).

Note: the Kaggle dataset reports the income per cap by Census Tract (which is a different classification of the city, more census tracts constitute a neighborhood). For this reason, an equivalency table has been found on <https://www.data.cityofnewyork.us> to convert and aggregate Census Tracts by neighborhood.

The second group of information can be scraped from Foursquare, the most trusted, independent location data platform for understanding how people move through the real world, using different API calls.

A third dataset is needed to determine latitude and longitude of neighborhoods (for scraping restaurants in a specific radium). This data are available in a .json file at the following link: <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json>.

## Data definition and cleaning

Since in both datasets there are a lot of information included, it is needed to select the most important variables for perceiving the goal.

The best approach would be to perform a regression analysis in order to understand which variable impacts the most on the profitability of a restaurant. Unfortunately, data for the target variable are not public and so this study focuses on 4 KPIs, two for each category, which, based on the common sense, should have a good explanatory power on the success of a restaurant:

* **Demographic data:** population density and average income per person.
* **Restaurant data:** number of Italian restaurant and average restaurant ratings.

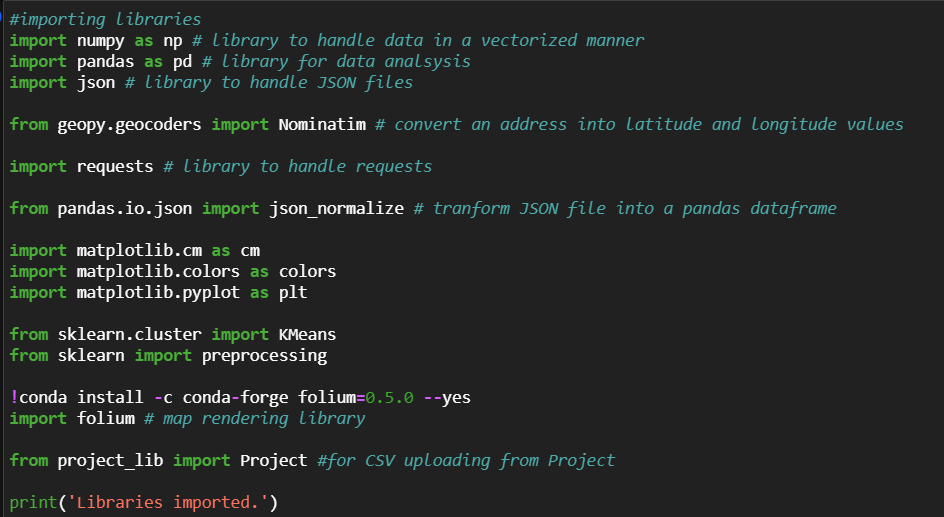
For Demographic data there are 2 issues which have been balanced by making 2 hypotheses:

1. Data are not updated to 2021 (2020 for Population Density, 2018 for Income Per Cap and 2010 for Census Tract – Neighborhood equivalency) 🡪 Hp: population and income are stable over time or the decrease/increase by neighborhood is evenly distributed (in this case the numbers could be different, but the proportion and the conclusion should be the same). The equivalence between census tracts and neighborhoods is still true.
2. Missing data for some neighborhoods 🡪 Hp: missing data have been replaced with the average value of the related borough.

For Restaurant data the only issue is about data completeness (the database could not include some restaurants or radius defined could not cover the correct area) and biases due to personal judgement (ratings are determined by humans so they could be based also on subjective reasons). In this case the only hypothesis is that the dataset, also if incomplete, is in average responding to the real world and that the noise linked to human bias could be excluded.

# Python libraries and Descriptive Analysis

## Python libraries



## Geographic data

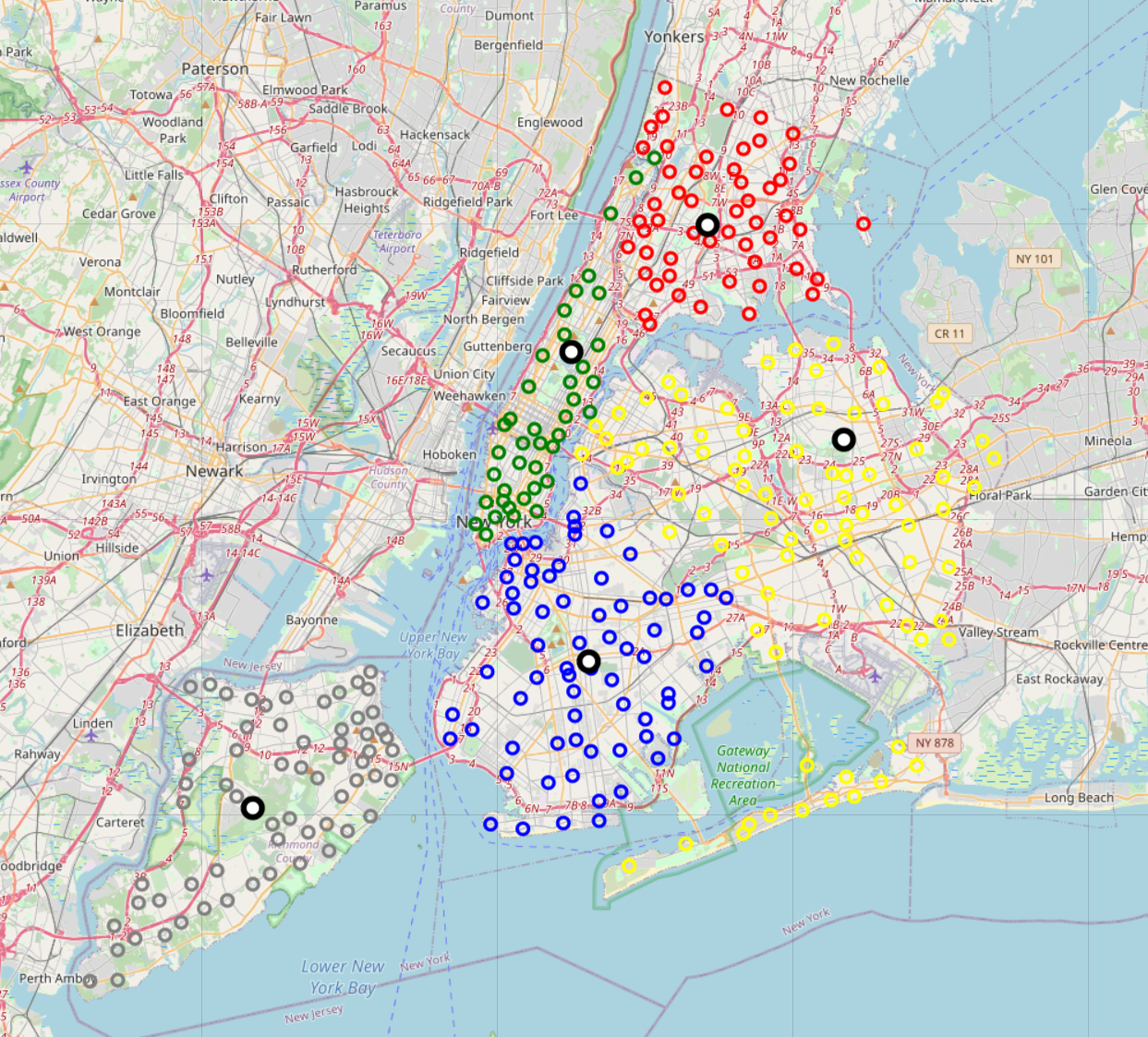
Data have been downloaded and imported into a Pandas DataFrame called *geo\_df,* with the following structure:

Borough – Neighborhood – Latitude – Longitude

The number of Neighborhoods by Borough is descripted in the table below:

|  |  |
| --- | --- |
| **Borough** | **N. Neighborhoods** |
| Bronx | 52 |
| Brooklyn | 70 |
| Manhattan | 40 |
| Queens | 81 |
| Staten Island | 63 |

Using Folium (a Python package for map rendering), it has been produced a visualization of the distribution of neighborhoods around New York City:



The black and white circles identify the center of each borough, which can be distinguished by the colour of the circles representing the neighborhoods:

* Bronx: red,
* Brooklyn: blue,
* Manhattan: green,
* Queens: yellow,
* Staten Island: grey.

## Demografic data

Demographic data have been retrieved on https://www1.nyc.gov/ and <https://www.census.gov/>.

4 different dataset have been merged into a unique .csv file:

* Income per Cap: CensusTract - <https://www.kaggle.com/muonneutrino/new-york-city-census-data?select=nyc_census_tracts.csv>
* Population Density (Population/Area):
  + Population: New York City population by neighborhood tabulation area – <https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp>
  + Area: New York NTA - <https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas-NTA-/cpf4-rkhq>
* Census Tract to Neighborhood Equivalency table: https://data.cityofnewyork.us/City-Government/2010-Census-Tract-to-Neighborhood-Tabulation-Area-/8ius-dhrr/data

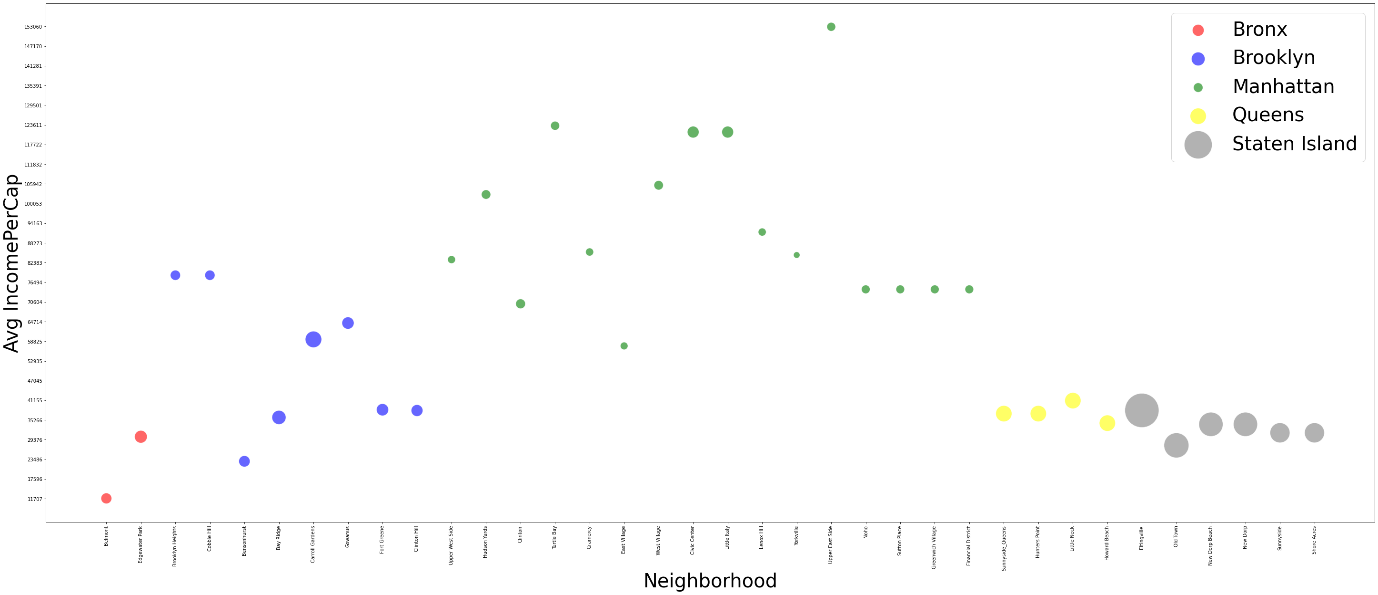
This file has been uploaded into a pandas DataFrame (*demo\_df*) using the Project library (to get the file from the project folder in IBM Cloud Park).

The file contains a lot of columns and so only the relevant ones has been kept: Neighborhood , Borough, PopByNeigh (Density) and IncomePerCap.

Missing data have been replaced with the average value of the related borough.

Finally, it has been performed an inner join between rest\_df and demo\_df, to keep only neighborhoods involved in which are common between the two datasets.

The following bubble plots show the average IncomePerCap (y) and the Population Density (bubble sizes) by neighborhoods (x - note that the colours identify a unique borough).



As shown in the plot above, Staten Island has the highest density but low income per cap, while Manhattan neighborhoods have the highest income per cap with a low density. In a first look, Brooklyn and Bronx seem to be the worst performing regarding this group of variable because they show low density and low income per cap.

## Restaurants data

This DataFrame (rest\_df) has been created in 2 steps.

In the first step the API “venues/explore” has been used to scrape from Foursquare all the venues for each neighborhood. The API has been defined as follows:

'https://api.foursquare.com/v2/venues/explore?&client\_id={}&client\_secret={}&v={}&ll={},{}&radius={}&limit={}'

Where the {} have been filled respectively with the following parameters:

* CLIENT\_ID,
* CLIENT\_SECRET,
* VERSION,
* lat,
* lng,
* radius,
* LIMIT

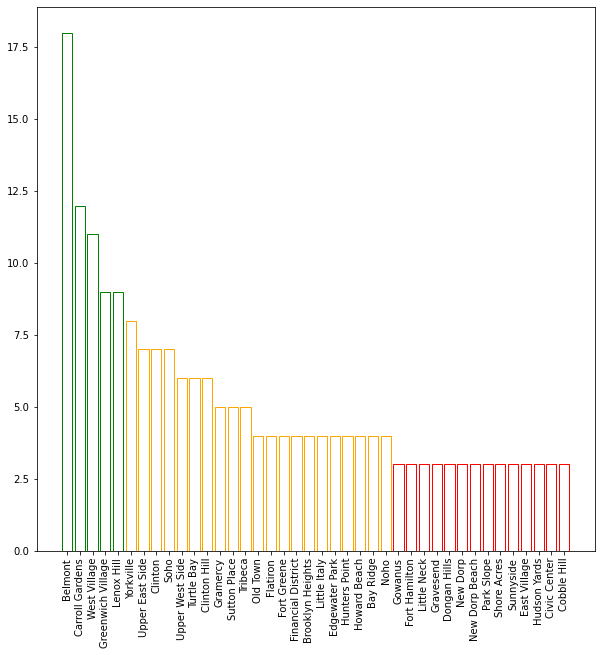
The resulting df has the following columns:

| **Neighborhood** | **Neighborhood Latitude** | **Neighborhood Longitude** | **Venue** | **Id** | **Venue Latitude** | **Venue Longitude** | **Venue Category** |
| --- | --- | --- | --- | --- | --- | --- | --- |

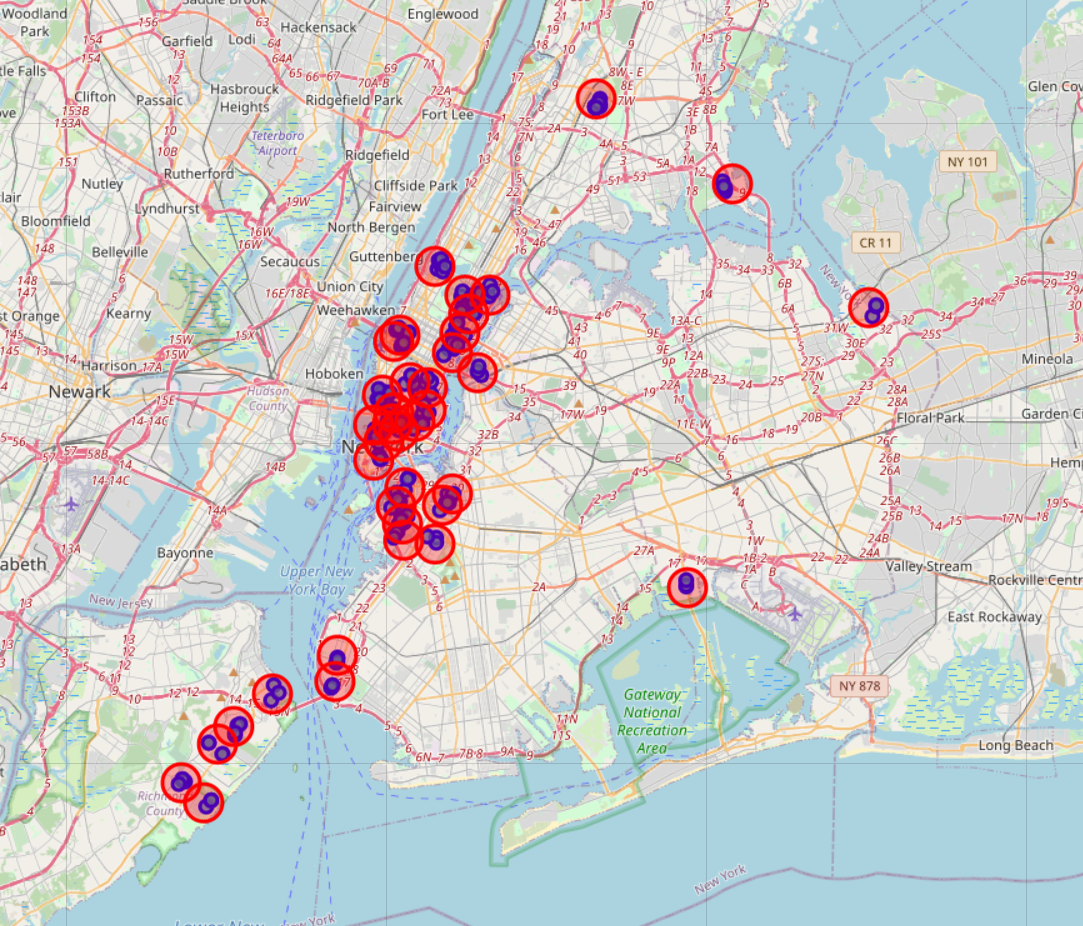
Data have been filtered by Venue Category = “Italian Restaurant”.

Then it has been deleted all rows referring to neighborhoods with a number of restaurants below the total average (because if there are so few Italian restaurants, probably this cuisine is not appreciated in that area).

Looking at the distribution of the venues, 3 clusters of neighborhoods are identified: High level (more than 8 restaurants), Mid Level (between 4 and 8 restaurants), Low Level (less than 4 restaurants).



A map of all the venues has been plotted using Folium (red circles are nighborhoods and blue circles are restaurants):



The second step is the usage of the API “venues/venue\_id” to scrape the rating for each restaurant.

'https://api.foursquare.com/v2/venues/{}?&client\_id={}&client\_secret={}&v={}',

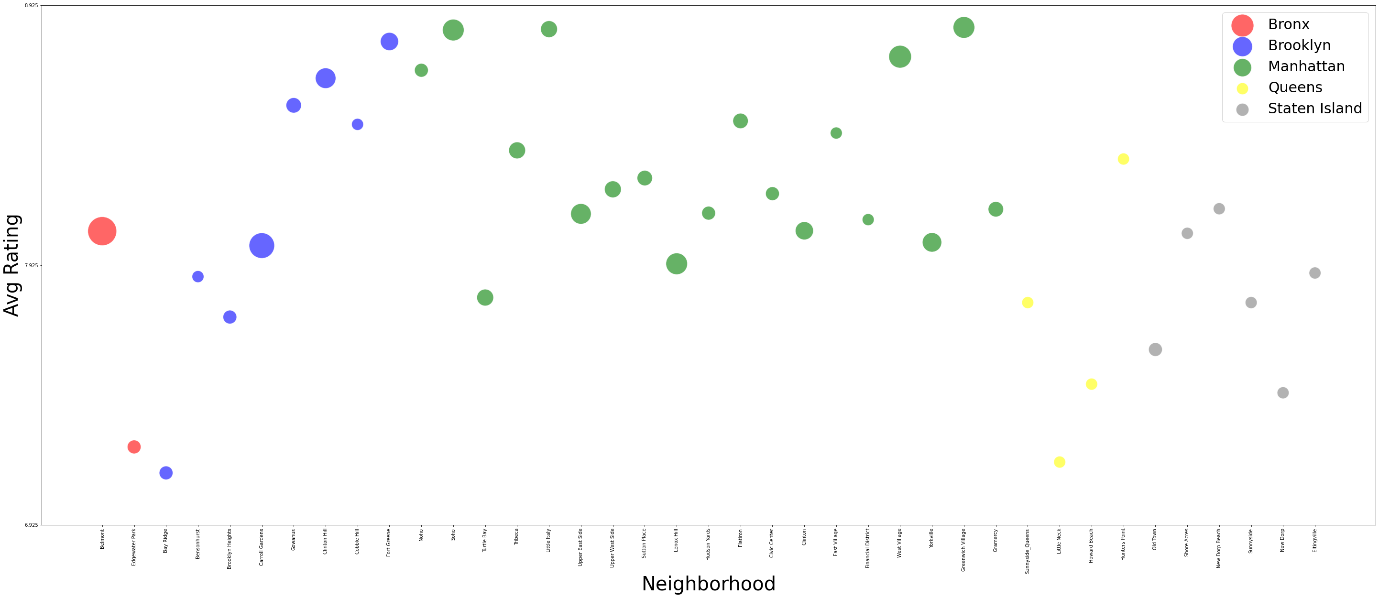
Where the {} have been filled respectively with the following parameters:

* VENUE\_ID
* CLIENT\_ID,
* CLIENT\_SECRET,
* VERSION

The average rating by neighborhood has been performed and the dataframe has been merged with the first one, to obtain the final rest\_df with the following columns:

| **Neighborhood** | **Borough** | **Rating** | **Latitude** | **Longitude** | **Restaurants** |
| --- | --- | --- | --- | --- | --- |

The following bubble plots show the average rating (y) and the n. of restaurants (bubble sizes) by neighborhoods (x - note that the colours identify a unique borough).



As shown in the plot above, Manhattan have the highest number of restaurants, but also high ratings. Brooklyn and Bronx has some neighborhoods with a high number of restaurants and high ratings and another group with few restaurants and low ratings. Queens and Staten Island low number of restaurants and low ratings. Belmont (Bronx) and Carroll Garden (Brooklyn) are examples of good performer because they show high number of restaurants with low average rating.

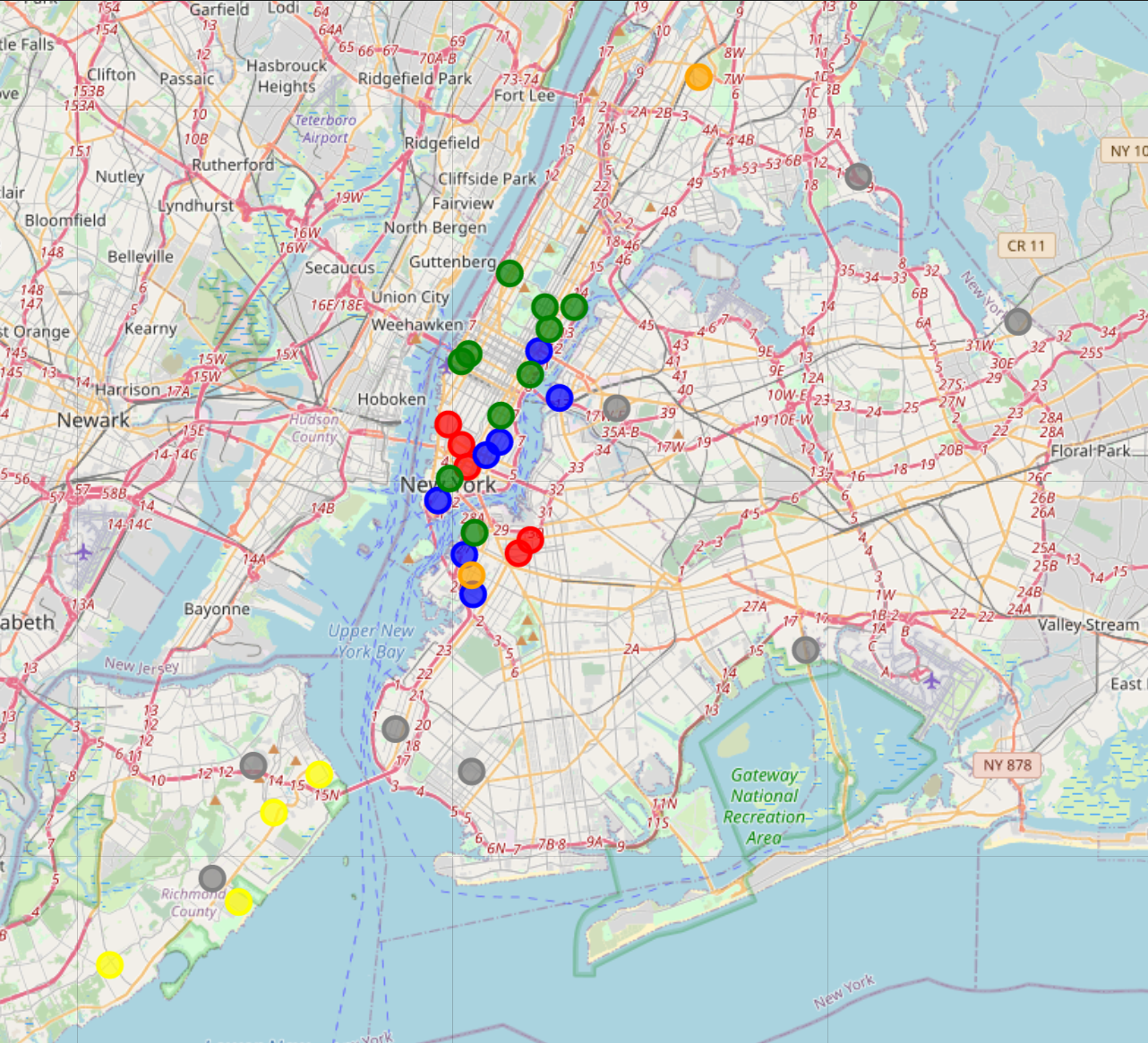
# Neighborhoods Clustering

The first step is the creation of the dataset for the K-Means algorithm.

This dataset has to contain only the explanatory variables which have been also normalized (with MinMax Scaler).

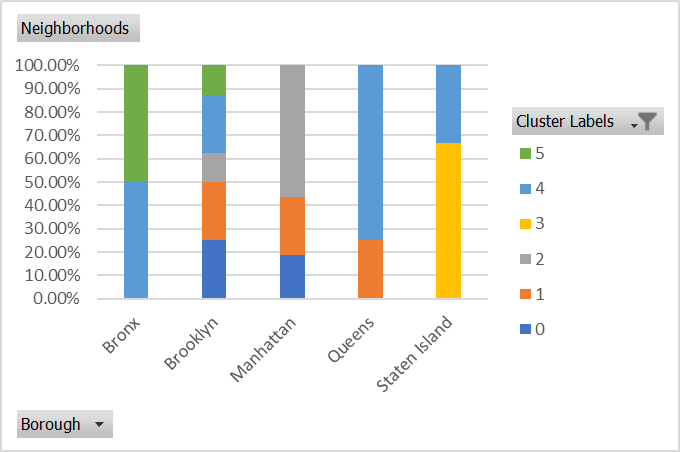
The number of clusters (k=6) has been determined by testing different values of k and getting the lowest value which gives at least 2 clusters for each borough (so that also intra-borough differences could be described).

Below the representation of the final result of the clustering:



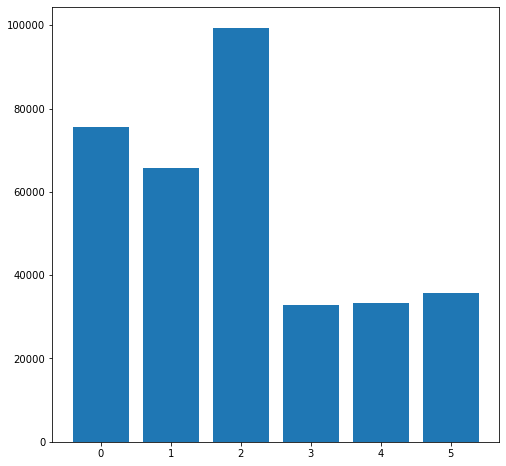
*LEGEND*:

* Cluster 0: red
* Cluster 1: blue
* Cluster 2: green
* Cluster 3: yellow
* Cluster 4: grey
* Cluster 5: orange

The following chart describes the distribution of clusters into each borough.

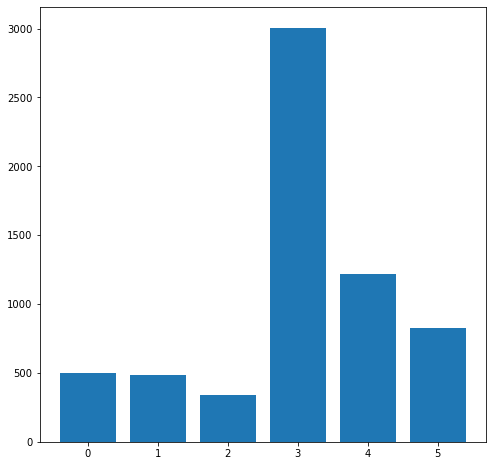
# Final Discussion

*Income Per Cap*



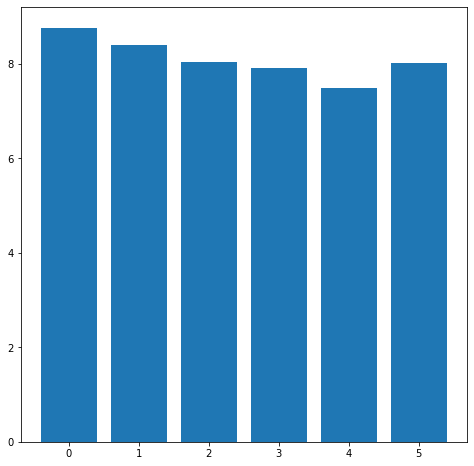
Clusters 0,1,2 offer a good opportunity in terms of income per cap, while clusters 3,4,5 seem to be less attractive

*PopbyNeigh*



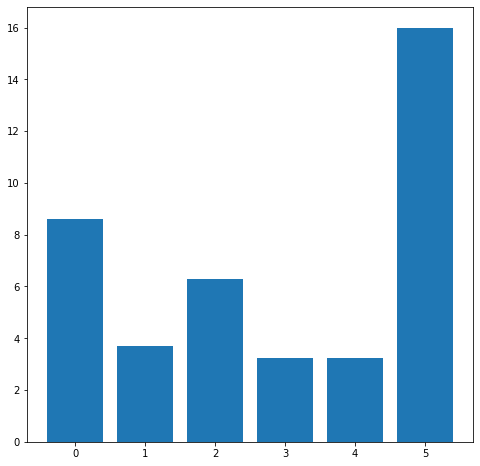
Neighborhoods in Clusters 0,1 and 2 have a low population density while clusters 3,4,5 are the densest.

*Rating*



Rating is in average very similar between clusters. Cluster 4 seem to offer the best improvement opportunity, while cluster 0 identify neighborhood which already have the greatest Italian restaurants in NY.

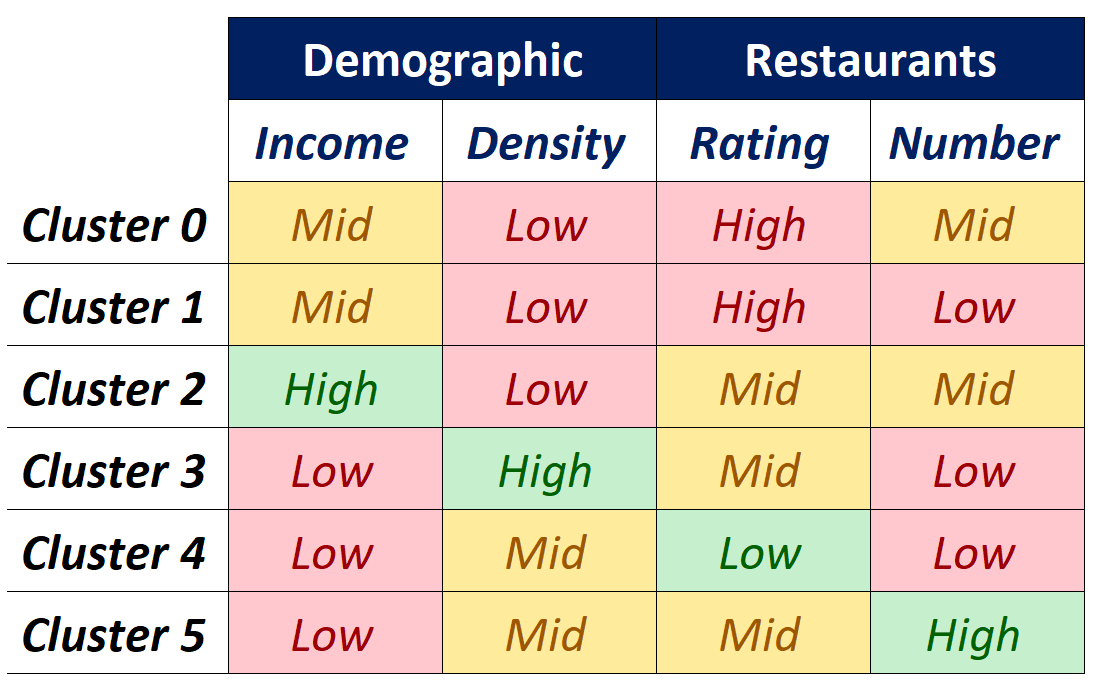
*Number of Restaurants*



Cluster 5 show the highest value and a big difference against all the others. Clusters 0 and 2 have an intermediate number of restaurants, while clusters 1,3,4 have the lowest.

Summary:





The most attractive clusters of neighborhood seem to be Cluster 2 and 5 with 1 High Performance, 2 Mid e 1 Low.

In detail:

* **Cluster 2** represent areas where to open an italian restaurant based on quality, raw material of first choice and high target of people. This because in those neighborhoods there is a low density of high-class population so the restaurant should be based on quality more than quantity. Moreover, the competition is middle level, this means that Italian Restaurants are quite appreciated but since the number of restaurants is mid, probably potential customers could prefer other cuisines. For this reason, I would suggest introducing some fusion cuisine elements in the menu in order to be more attractive in the market.
* **Cluster 5** represent areas where to open an Italian restaurant more based on a fast and friendly service and with a business model based on high quantities and low prices. This because in this area there is a Mid density of low-class population so the restaurant should attract a lot of people taking also into account their economic power. Moreover, the competition shows a high number of Italian restaurants with mid average rating. This means that Italian cuisine is very appreciated in these neighborhoods but there are already a lot of appreciated venues. For this reason, I would suggest applying marketing strategies for customer retention and brand awareness (for example membership cards, “bring a friend” promo, social media marketing campaigns and so on).

# Future Developments

This work has different improvement opportunities in terms of:

* Data integration: define new variables for clustering related to mobility data, number and type of commercial activities and buildings and so on.
* Modeling:
  + testing other clustering algorithm like DBSCAN.
  + Try to determine the weight of each variable on the success of a restaurant with a regression model, using rating as target variable. In this way only the most effective variables could be included and the clustering would be more efficient.