Final Capstone Project

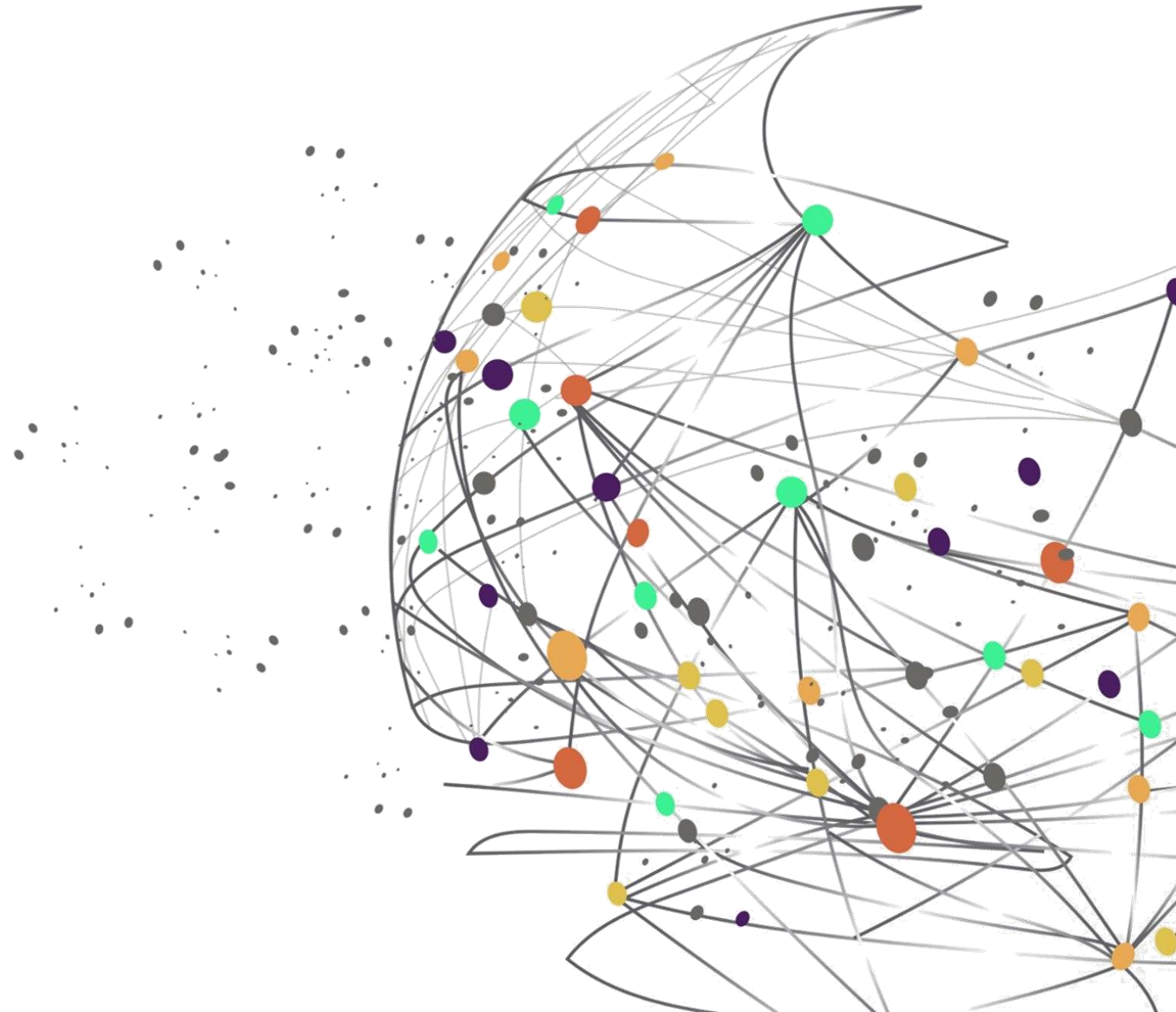# Clustering Italian restaurants in NYC

A data-driven approach for opening a new restaurant

Andrea Sarni, Aug 2021
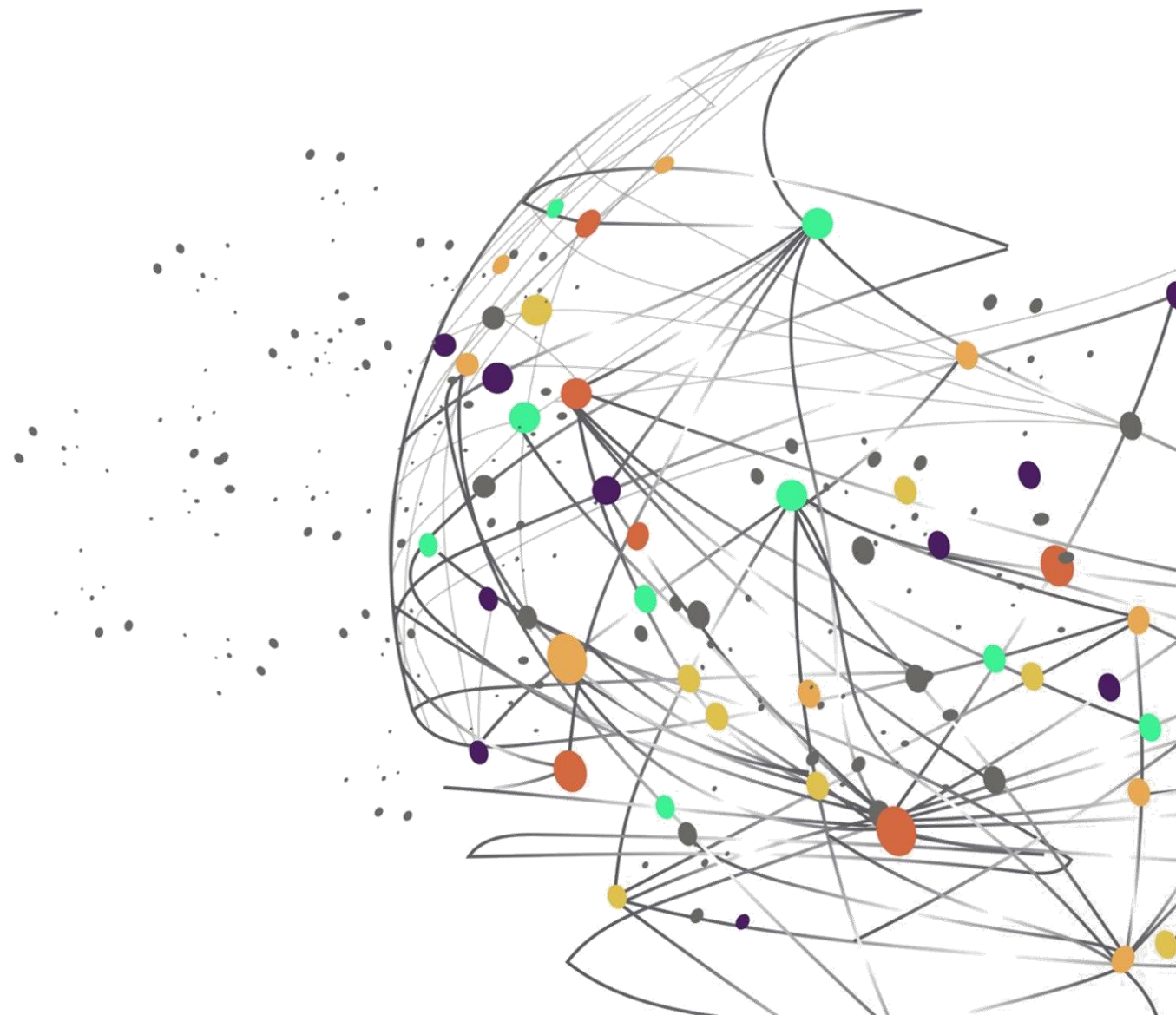
Agenda

1. **Context**
2. **Approach**
3. **Analytical model**
4. **Discussion**

Clustering Italian restaurants in NYC│ Context

New York City

New York is the **most populous city and** the **most densely populated** major city in the United States:

- Estimated 2020 **population**: 8,253,213
- **Area**: 784 km², New York City is also major city in the United States
- **5 Boroughs**: Brooklyn, Queens, Manhattan, the Bronx, and Staten Island

Business Requirement

Italian cuisine is one of the top three cuisines in the United States, according to the National Restaurant Association (around 1k restaurants out of 20k total restaurants in NYC (Indian, French or Thai cuisines count only 300 restaurants each!)

Business Question:
Giving the great extension of the city, is there a data-driven approach to determine which among all neighborhoods (division of each borough) are the best solution to locate a new activity, in order to have the best solution in terms of attractivity and profitability?

Clustering Italian restaurants in NYC | Approach

Problem setting

Understanding the neighborhoods of NYC in terms of population and category of commercial activities, to define the best neighorhood to place an italian restaurant. The answer should find a neighborhood balancing specific *principles**

*Principles
- **Max number of Italian Restaurants**: this ensures that in the neighborhood the italian cuisine is really appreciated
- **Max average salary**: this ensures that a lot of people would be able to spend their money in the restaurant
- **Max size of population**: this ensures that a lot of people could enter in the restaurant and talk about it around the city
- **Min average rate of Competitors**: this ensures that the level of competition will be not so hard

Data Assessment & Availability

For this problem we would need 2 groups of information:
- Geographic data
- Demographic data
- Restaurants data

Bottom-up: from data to real solution

Real Problem solution
- Results Discussion
- Story Telling

Data Analysis Design
- Descriptive Analysis
- Clustering problem → K-Means algorithm

Top-down: from real problem to data

Clustering Italian restaurants in NYC | Analytical Model

Demographic data (y=IncomePerCap; size=Density)
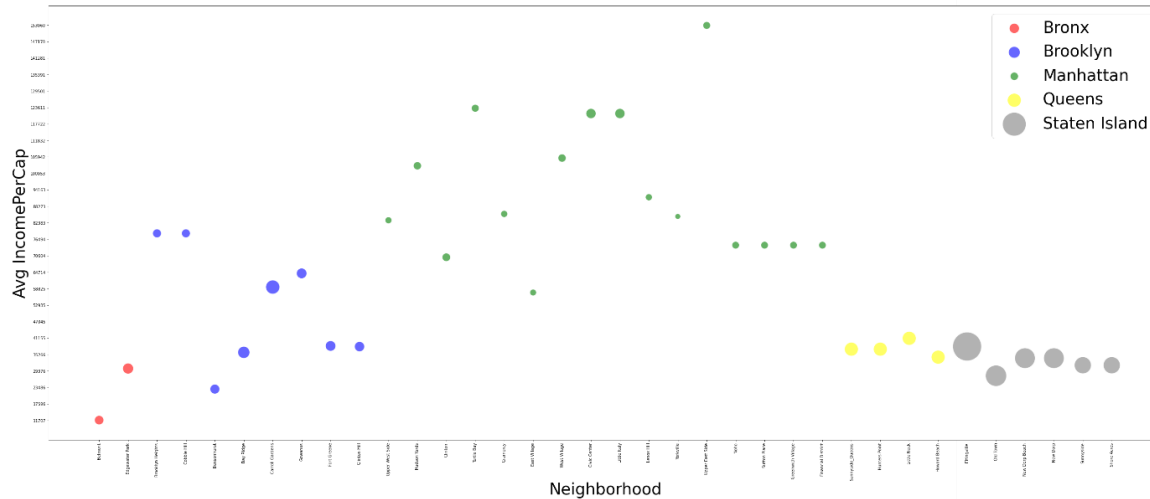
Comments
- **Bronx**: *Low Income – Low Density*
- **Brooklyn**: *Mid Income – Mid Density*
- **Manhattan:** *High Income – Low Density*
- **Queens**: *Low Income – Mid Density*
- **Staten Island**: *Low Income - High Density*

Neighborhoods in Manhattan and Brooklyn could be good candidates based only on this variables.

Queens and Bronx (especially Belmont) seem to be the worst candidate.

Staten Island shows good value in terms of density but low average income per cap.

Restaurants data (y=Rating; size=N. of Restaurants)
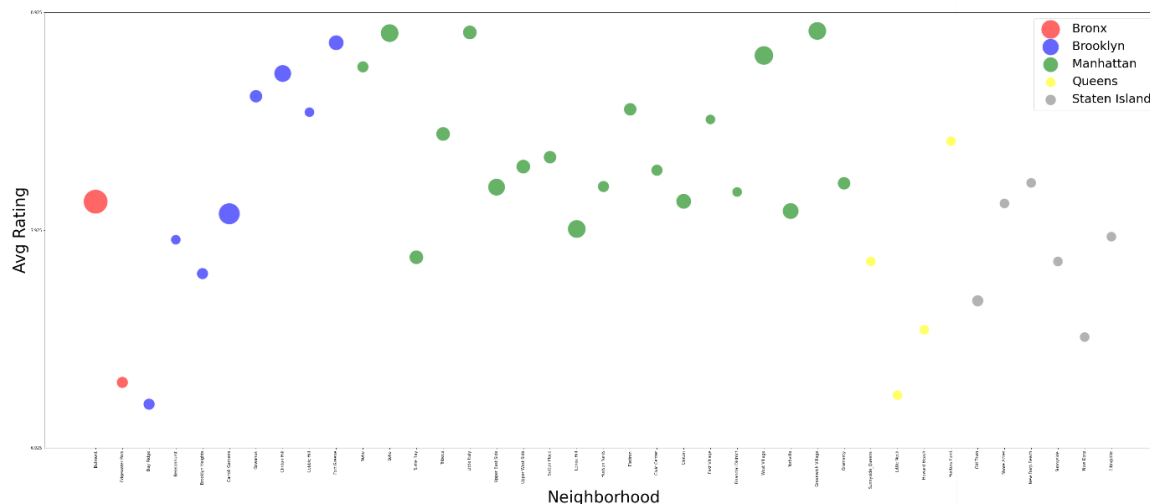
Comments
- **Bronx**: *Low Rating – Mid N. of Restaurants*
- **Brooklyn**: *Mid Rating – Mid N. of Restaurants*
- **Manhattan**: *High Rating – Mid N. of Restaurants*
- **Queens**: *Low Rating – Low N. Restaurants*
- **Staten Island**: *Low Rating – Low N. Restaurants*

Neighborhoods in Manhattan and Brooklyn show high rated restaurants but high likelihood for Italian cuisine.

Queens and Staten Island show low rated restaurants but low likelihood for Italian cuisine.

Clustering Italian restaurants in NYC| Analytical Model

Legend
- 🔴 Cluster 0
- 🔵 Cluster 1
- 🟢 Cluster 2
- 🟡 Cluster 3
- ⚪ Cluster 4
- 🟠 Cluster 5
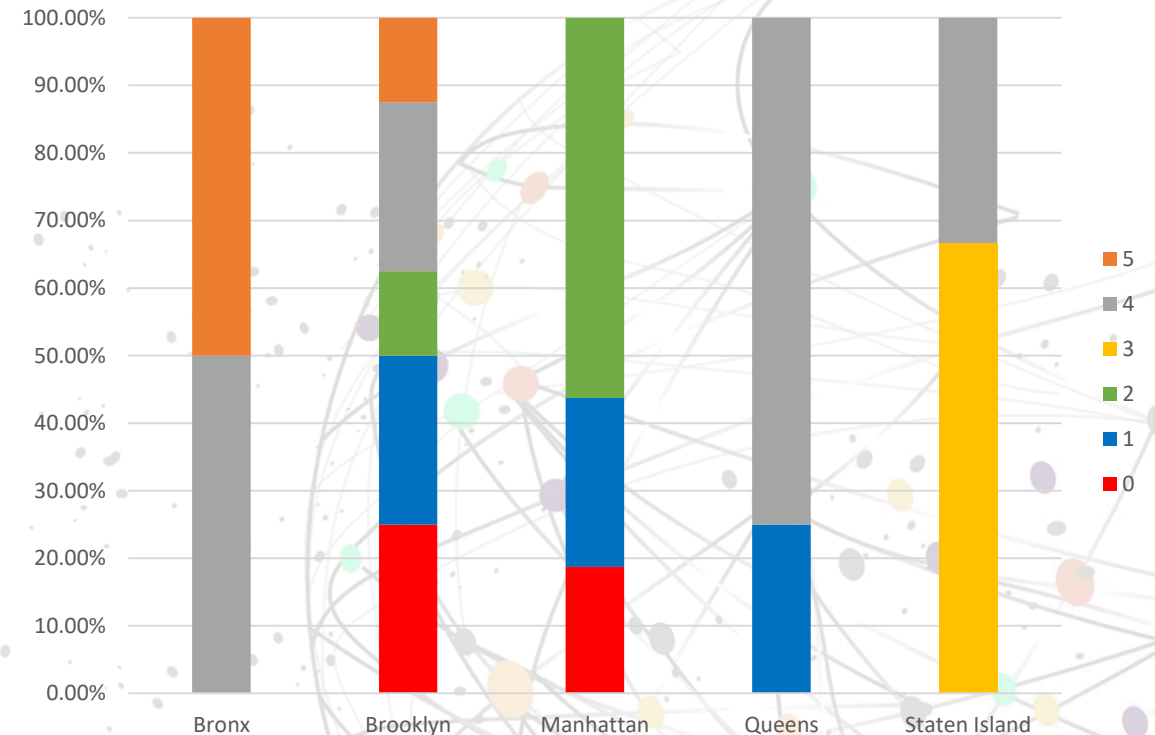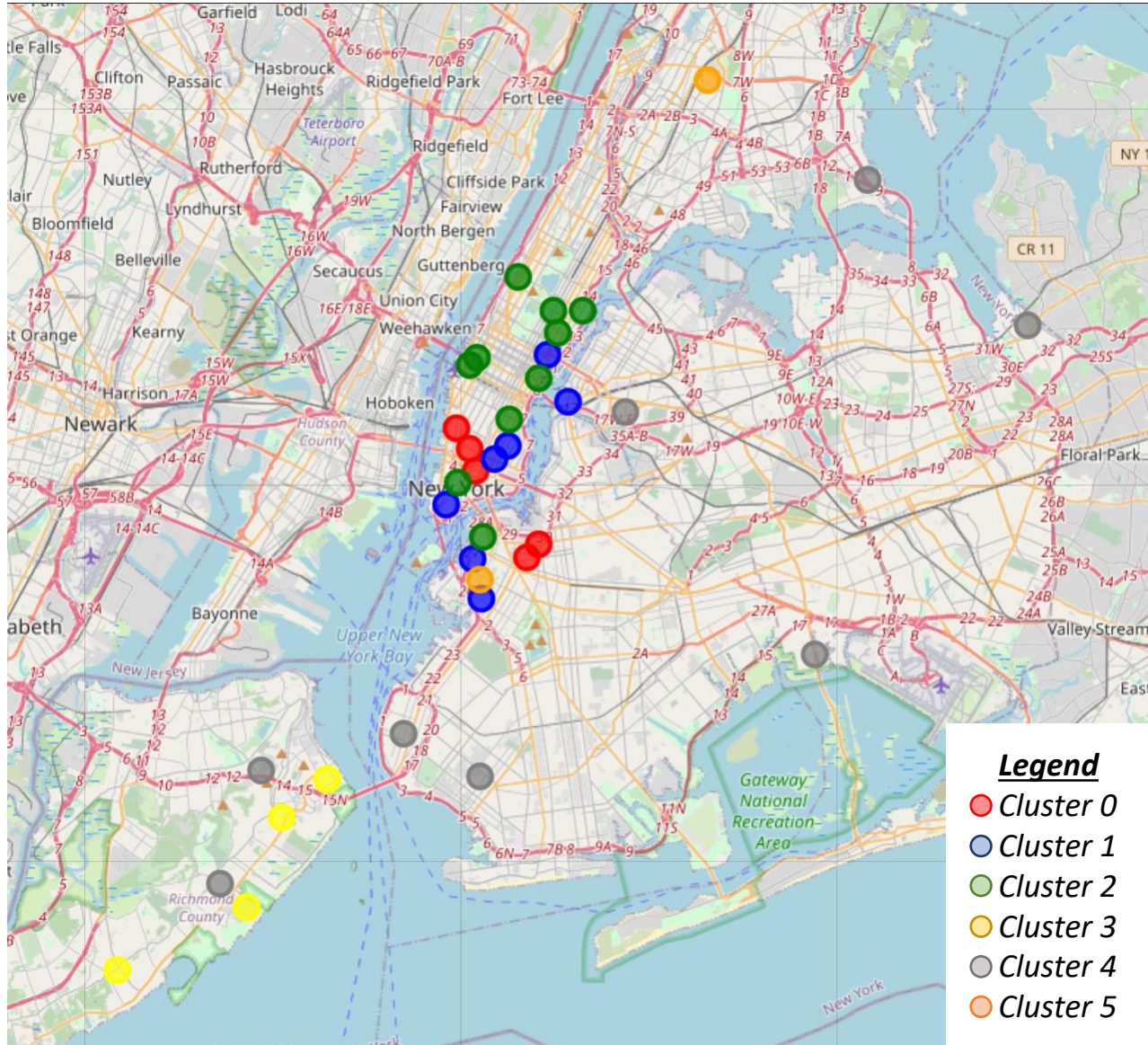
Each borough is composed by neighborhoods of different clusters.
- **Bronx and Staten Island**: equally split in 2 clusters (respectively 4-5 and 3-4)
- **Brooklyn**: Mix of 4 clusters (0-1-2-4-5)
- **Manhattan and Queens**: mostly 1 cluster + minor clusters (respectively 2+0-1 and 4+1)
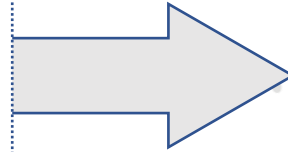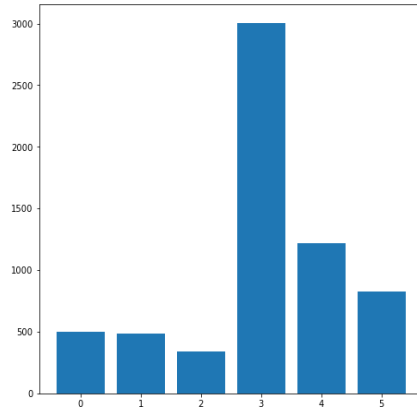
Clustering Italian restaurants in NYC| Discussion

Which is the best Neighborhood?
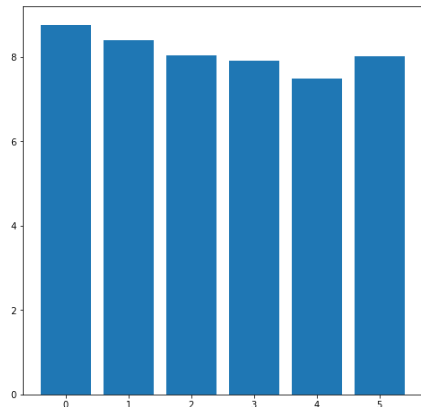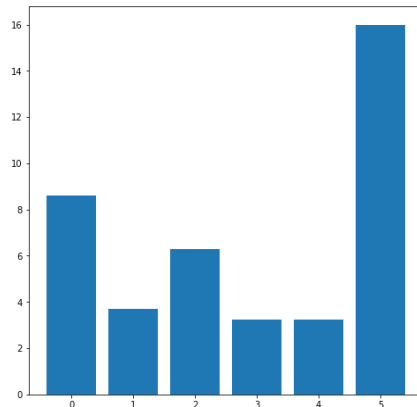
Demographic

Income per Cap
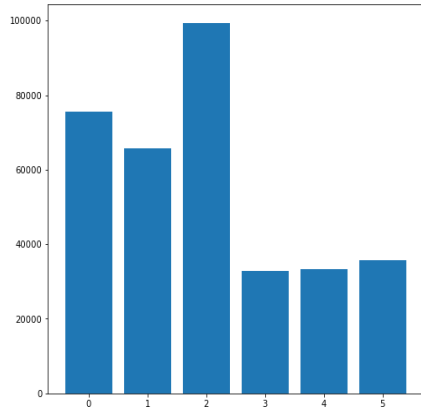
Density

	Demographic		Restaurants	
	Income	**Density**	**Rating**	**Number**
Cluster 0	Mid	Low	High	Mid
Cluster 1	Mid	Low	High	Low
Cluster 2	High	Low	Mid	Mid
Cluster 3	Low	High	Mid	Low
Cluster 4	Low	Mid	Low	Low
Cluster 5	Low	Mid	Mid	High

Restaurant

Rating

N. Of Restaurants

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).