Applied Data Science Capstone - "Battle of the Neighborhoods"

Where to open a restautrant in New York City?

ABSTRACT

This report presents a data science project to provide assistance to a decision maker to decide where to open a restaurant in New York City.

Stefan Lemm

IBM Data Science Professional Certificate

1. Introduction/Business Problem

The problem to be discussed in this report is: In what neighborhood of New York City should a restaurant be opened?

Several criteria have to be taken into account in case your're iterested in opening a restaurant. As you will be starting a business, you need to

- prepare a business plan
- consider legal requirements, e.g. the type of business (LLC, partnership or cooperation)
- define a logo, cards, stationary
- get tax ID numbers, licenses and permits
- think of insurance
- prepare accounts
- get a business line of credit
- ready the workspace
- leace an office space and equipment.

A well-known saying regarding property is "location, location, loction". This definitely goes for where to open a business, too. All of these above activities are dependant on the location of your business. Only when you have decided on the state, town and neighborhood of your business, you will be able to start working on the other requirements / steps towards the actual opening.

Important factors of identifying the right location (in. thuis case neighborhood) for your restaurant include

a. Who are the potential customers and how many are may be available at a given location? Depending on the type of business you want to attract fifferent types of customers. For an high class restaurant with very high prices with reservation policy, you want to attract welthier customers, than if you want to open a Chinese low cost restaurant with walk-in customers.
b. How important is proximity?
   If you're a retail store that relies on the local community, this is vital. For other business models, it might not be. If you need people to come into your store, make sure that store is easy to find. Remember: even the best retail areas have dead spots.
c. For your employees, schools, recreational activities, cultural opportunities might be very important.
d. How many competing restaurants of what types are in a specific neighborhood? Sometimes having competitors nearby is a good thing. Other times, it's not. You've done the market research, so you know which is best for your business.
e. For a business idea that isn't completely new, it might make sense to think about the current offerings and focus on how to create something better, cheaper or faster.

Who would be interested in this project?

This project could provide support for a decision maker where to prioritise in investing time and ressources to find a location for a new restaurant. Even after determining a potential

neighborhood, an actual object has gto be found, but restricting the area where to look for it could make a difference.
A more generic approach could find potential areas for any kinf of business.

2. Data

2.1 Analytical Approach
The analytical approach for the problem, "finding the best neighborhood for a new restaurant", includes following steps:

a. Choose the country, state and town for the restaurant
Based on own interest, the project will start with New York City

b. Determine possible neighborhoods
New York City is split up into five boroughs, which are the Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Each borough has the same boundaries as a county of the state. The county governments were dissolved when the city consolidated in 1898, along with all city, town, and village governments within each county. The term *borough* was adopted to describe a unique form of governmental administration for each of the five fundamental constituent parts of the newly consolidated city.
To be more detailed, a level lower should be selected for the research.
The **community boards** of the New York City government are the appointed advisory groups of the community districts of the five boroughs. There are currently 59 community districts: twelve in Manhattan, twelve in the Bronx, eighteen in Broklyn, fourteen in Queens, and three in Staten Island. They are also called Community Districts.

The given dataset from following web site provided central locations in neighborhoods of New York City:
https://geo.nyu.edu/catalog/nyu_2451_34572

From the New York City open data, the Community Districts geojson file was dowladed for visualization. https://data.cityofnewyork.us/City-Government/Community-Districts/yfnk-k7r4

c. Collect data for each Community District the Number of potential customers
NYC neighborhoods were defined in terms of Public Use Microdata Areas (PUMAs).
PUMAs approximate NYC Community Districts (CDs)
Following web site has population data on PUMA level:
https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoodpop.htm

Another source of population data on a more detailed level can be found on Wikipedia:
https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City

d. The number of other restaurants in a neighborhood by type
Foursquare will be used to collect the data

e. Based on the collected data, a model will be build to then use clustering to compare the neighborhoods and identify the ranking of neighborhoods.

2.2 Data cleaning

The data from several sources were combined into one table.
A community district consists of more than one neighborhood. The consisting neighborhoods were identified with the right communit district.
To be able to be more detailed, the venue data was downloaded from Foursquare for neighborhoods with a radius of 500m. This might lead to a restaurant being counted for several community districts, but as the defined radius of interest for a potential customer was assessed to be within1km, this seemd to be logical.

Some of the neighborhoods from the original New York table were not in the Community district table and could not be identified. I chose to not use them.

Via Foursquare, all venues were retireved, then out of the 440 unique categories, only these with “restaurant” (“R” and “r” as a start possible) in the name were chosen to keep. The total number of restaurants per Community District was added.

To be able to visualize via chloreopath map, the geojson codes were derived from the community board names.

2.3 Feature selection

After cleaning, there were 58 community districts. Within 267 neighbourhoods, 2293 venues were retireved via Foursquare.
The final table consists of the Community board, the type and number of restaurants and the total number of restaurants.

Final decision of recommended neighborhoods was made taking into account the number of residents per square kilometer and the most restaurants in a CD, but the least Italian restaurants in a CD.

4. Exploratory Data Analysis

a. As a database I used an open database on NYC neighborhoods.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

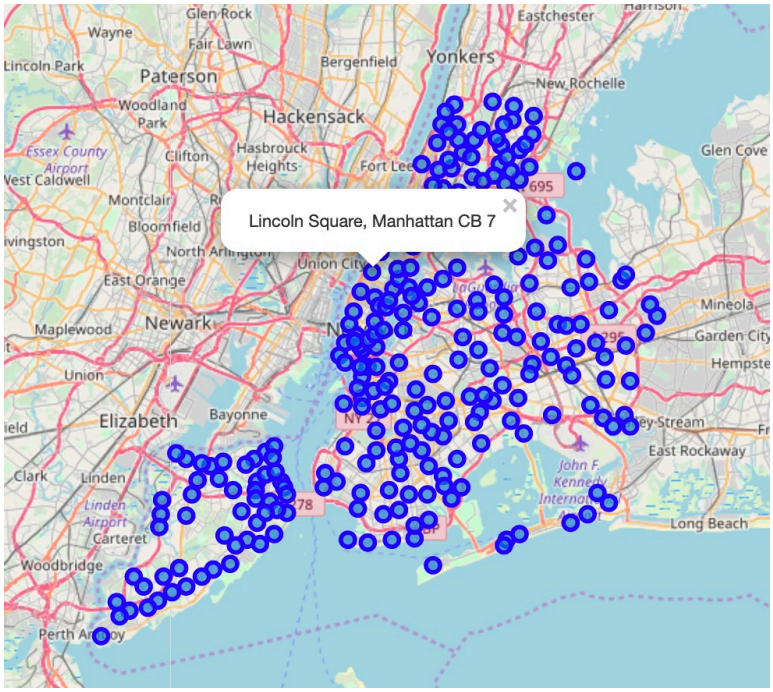
I added data on New York Community districts, found at Wikipedia. The result was following dataframe, which includes population data and the geojson code for each Community District.

	CB	CB_Area	CB_Population	CB_Pop_skm	Neighborhoods	boro_cd
0	Bronx CB 1	7.17	91,497	12,761	Melrose, Mott Haven, Port Morris	201
1	Bronx CB 2	5.54	52,246	9,792	Hunts Point, Longwood	202
2	Bronx CB 3	4.07	79,762	19,598	Claremont, Concourse Village, Crotona Park, Mo...	203
3	Bronx CB 4	5.28	146,441	27,735	Concourse, High Bridge	204
4	Bronx CB 5	3.55	128,200	36,145	Fordham, Morris Heights, Mount Hope, Universit...	205

As the data analysis was to be performed on a community district basis, for each neighborhood, the belonging CB (Community Board) had to be found and added to the dataframe. After extensive cleaning, the result was following dataframe:

	Borough	CB	Neighborhood	Latitude	Longitude	boro_cd
1	Bronx	Bronx CB 1	Melrose	40.819754	-73.909422	201
2	Bronx	Bronx CB 1	Mott Haven	40.806239	-73.916100	201
3	Bronx	Bronx CB 1	Port Morris	40.801664	-73.913221	201
4	Bronx	Bronx CB 10	City Island	40.847247	-73.786488	210
5	Bronx	Bronx CB 10	Co-op City	40.874294	-73.829939	210

I used folium, to visualize all the neighborhoods with CB's:



I then utilized Foursquare to download venue data for each neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Melrose	40.819754	-73.909422	Blink Fitness	40.819543	-73.910554	Gym / Fitness Center
1	Melrose	40.819754	-73.909422	Blink Fitness St Ann's	40.819470	-73.910522	Gym
2	Melrose	40.819754	-73.909422	Senshi Okami Martial Arts Center	40.819295	-73.914158	Martial Arts Dojo
3	Melrose	40.819754	-73.909422	Perry Coffee Shop.	40.823433	-73.910940	Diner
4	Melrose	40.819754	-73.909422	Cinco de Mayo	40.822600	-73.911586	Mexican Restaurant

8731 venues were found in total

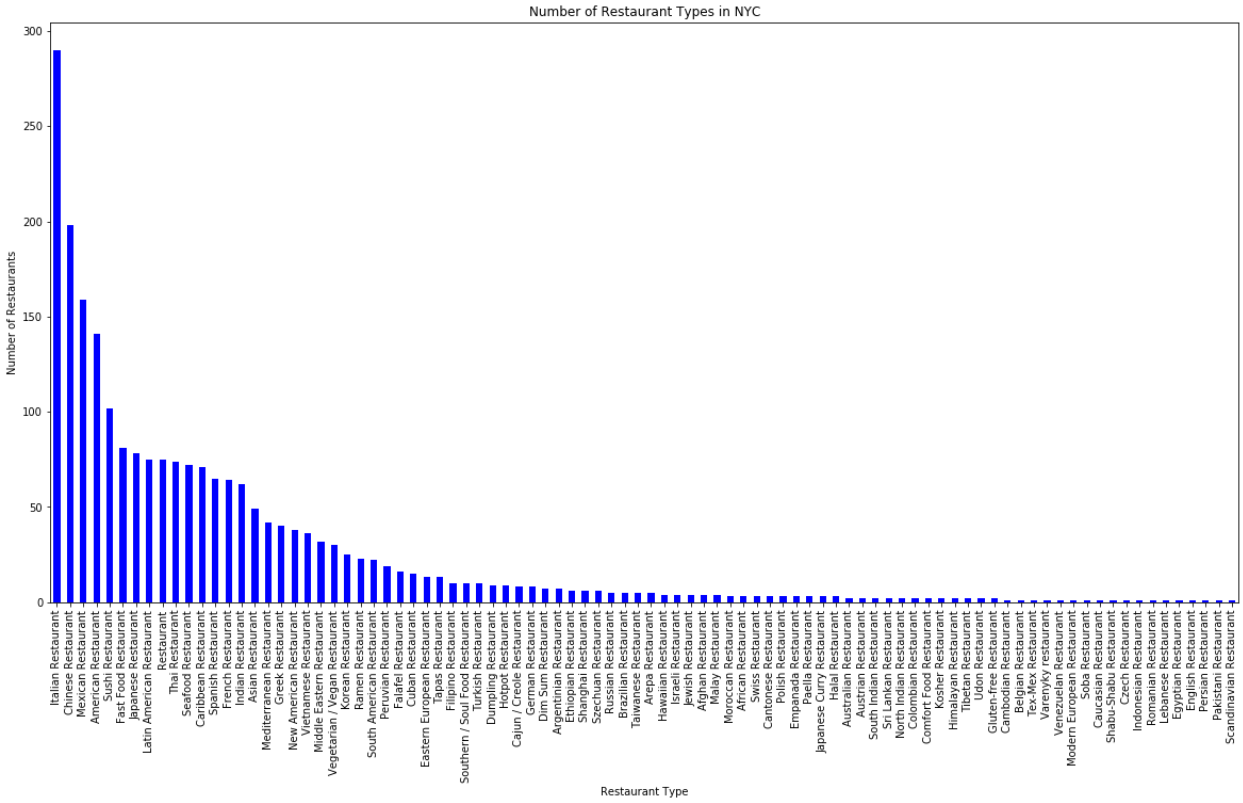
As I was only interested in Restaurants, I reduced the dataframe:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
4	Melrose	40.819754	-73.909422	Cinco de Mayo	40.822600	-73.911586	Mexican Restaurant
23	Mott Haven	40.806239	-73.916100	Pio Pio	40.806047	-73.914185	Peruvian Restaurant
35	Mott Haven	40.806239	-73.916100	Picanteria El Botecito	40.803614	-73.918906	Spanish Restaurant
38	Mott Haven	40.806239	-73.916100	Carmen And Cindy's	40.804310	-73.911988	Spanish Restaurant
40	Mott Haven	40.806239	-73.916100	Rincon Ecuatoriano	40.803689	-73.911951	Latin American Restaurant

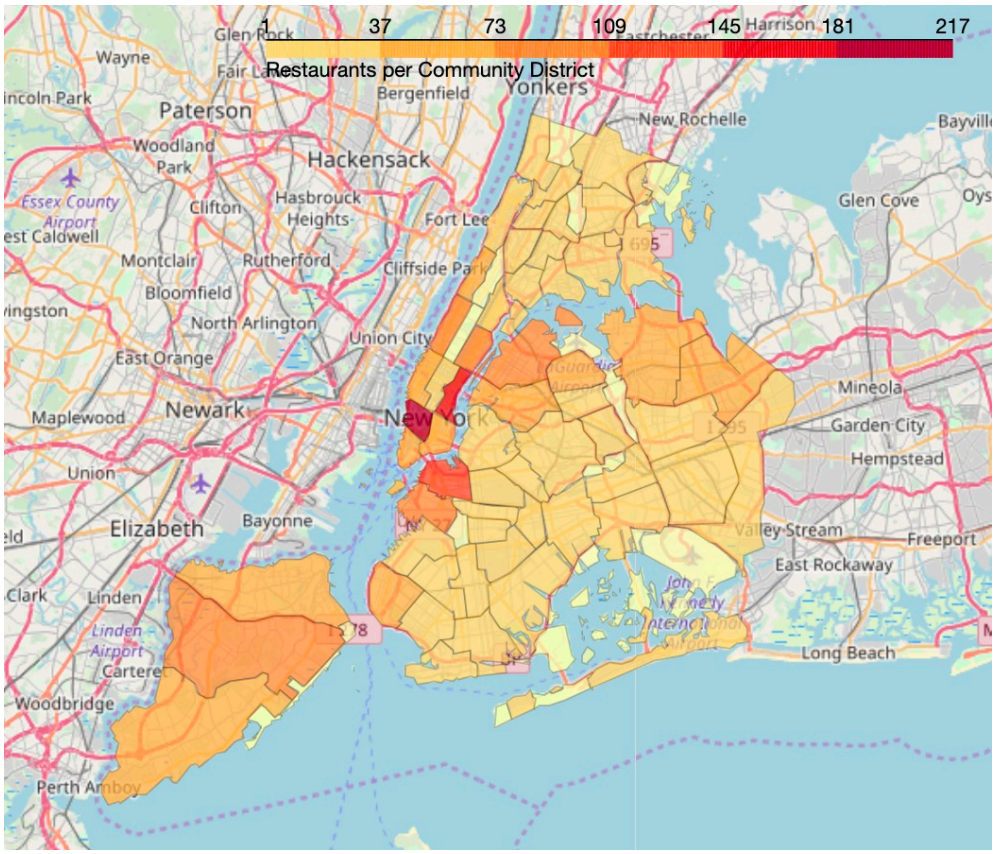
2,225 Restaurants were found.

I analysed, how many types and number rof restaurants we have.estaurants

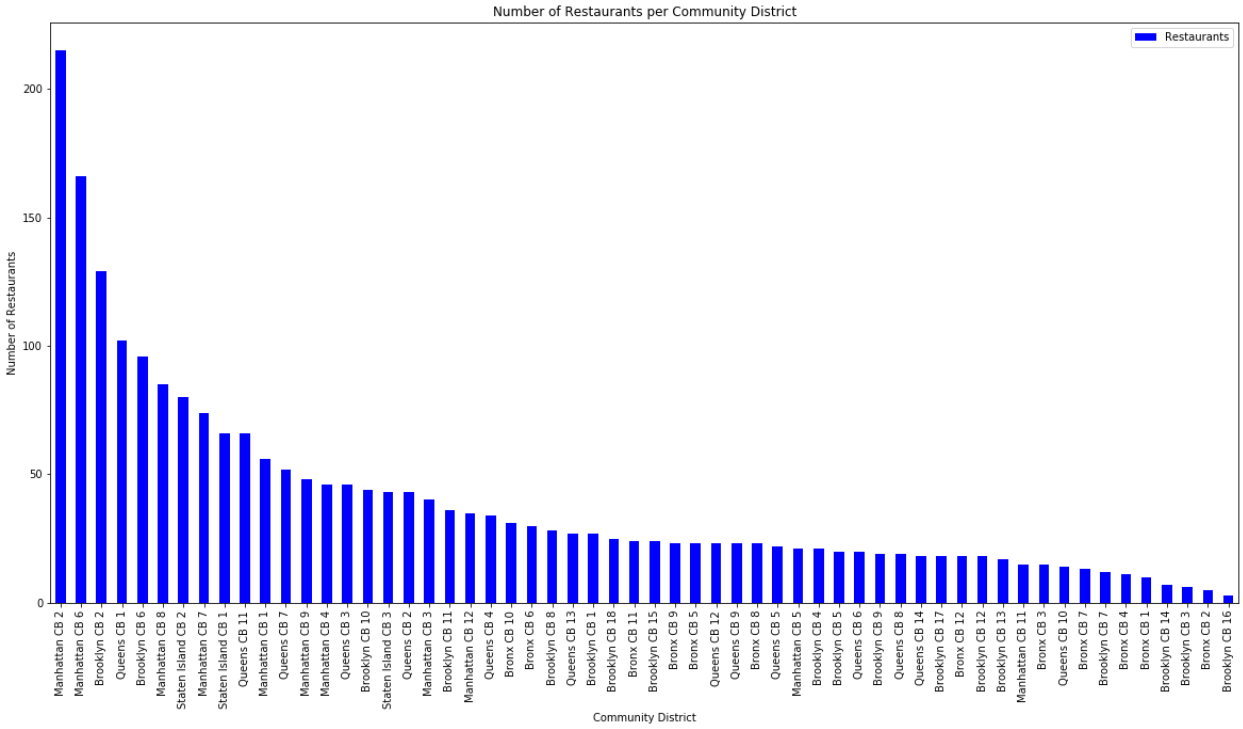
Italian Restaurant	282
Chinese Restaurant	216
Mexican Restaurant	161
American Restaurant	144
Sushi Restaurant	105
Fast Food Restaurant	83
Latin American Restaurant	79
Thai Restaurant	79
Japanese Restaurant	75
Restaurant	75
Seafood Restaurant	67
Caribbean Restaurant	67
Spanish Restaurant	67
Indian Restaurant	65
French Restaurant	58
Asian Restaurant	57
Korean Restaurant	47
Vietnamese Restaurant	38
New American Restaurant	37
Mediterranean Restaurant	37
Greek Restaurant	36
Middle Eastern Restaurant	32



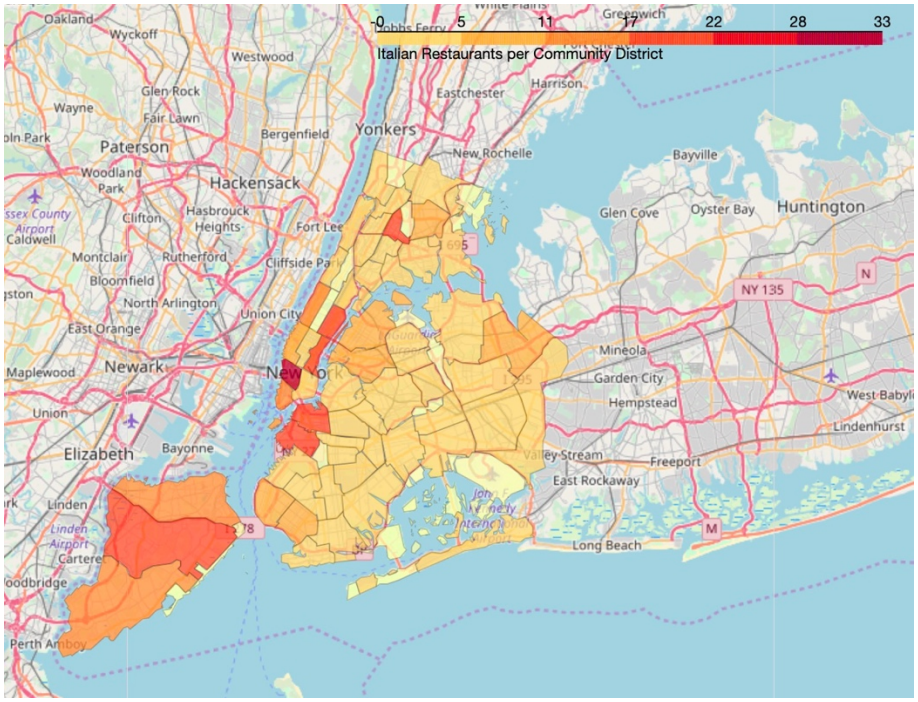
Then, how the restaurant were distributed in New York City

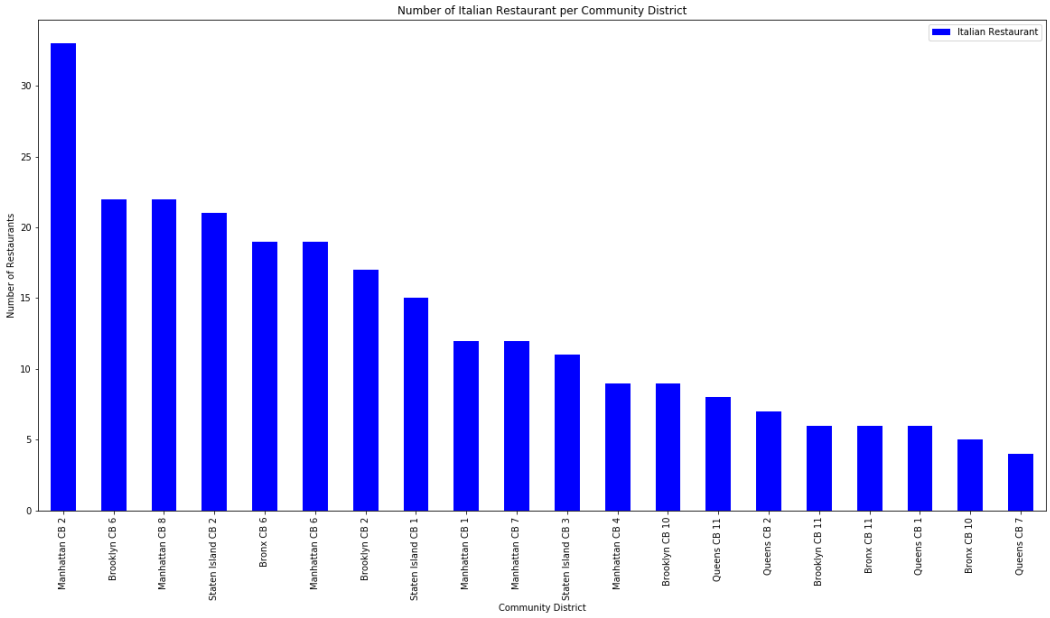


The most restaurants are in following districts:

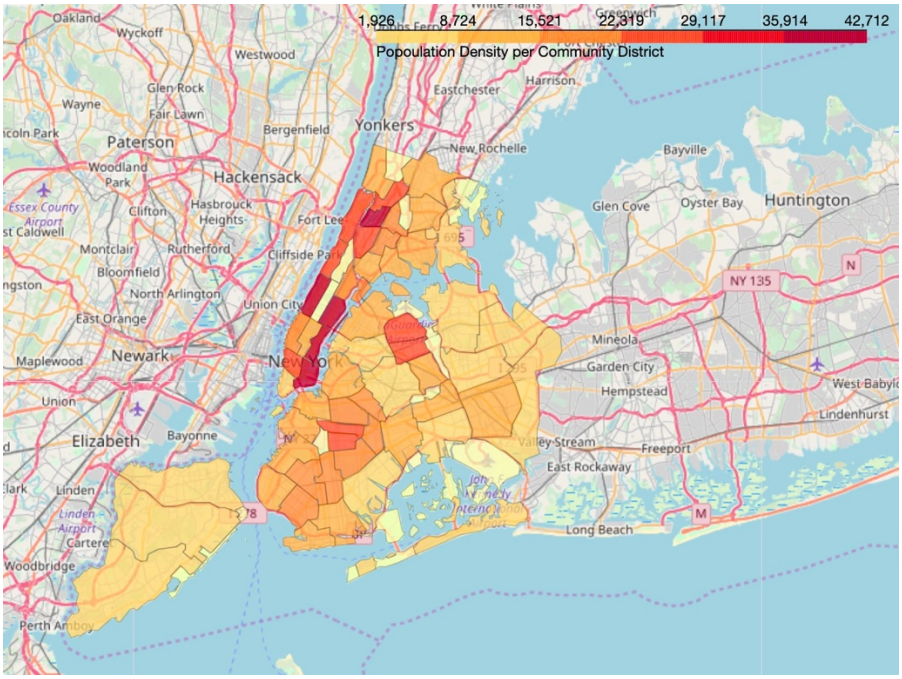


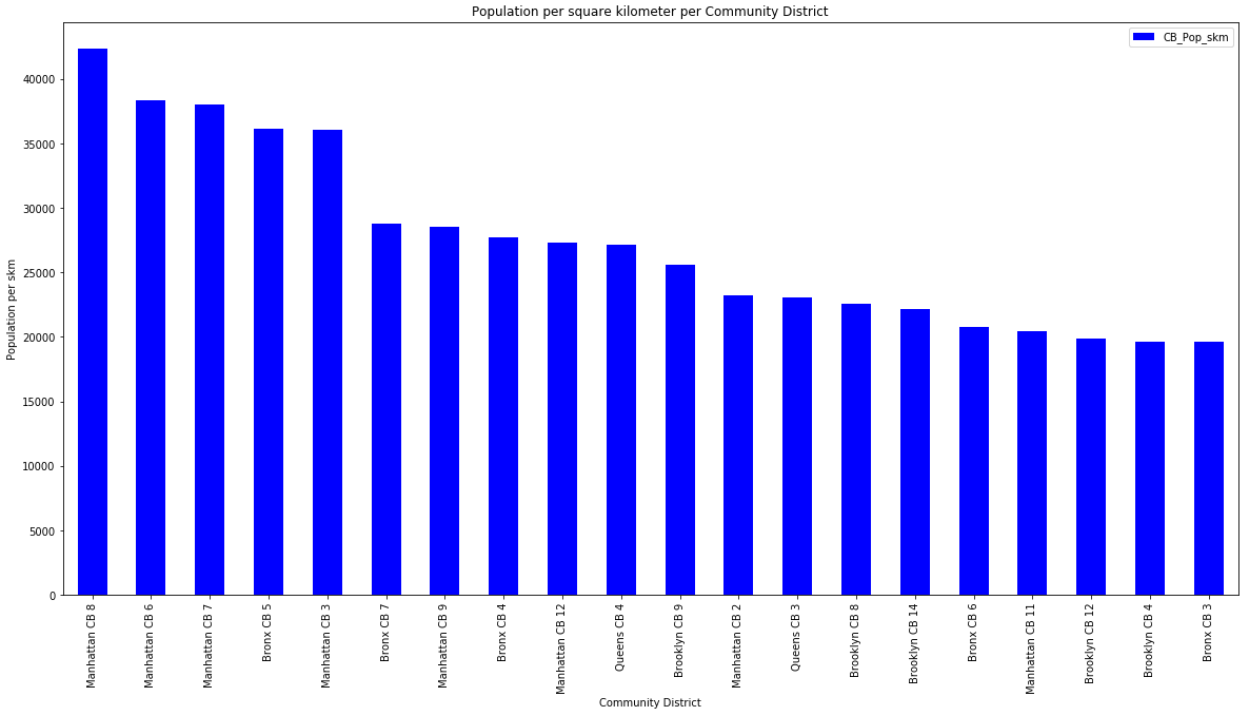
How about Italian Restaurant Distribution in New York City





Community Districts with the most residents per square kilometer

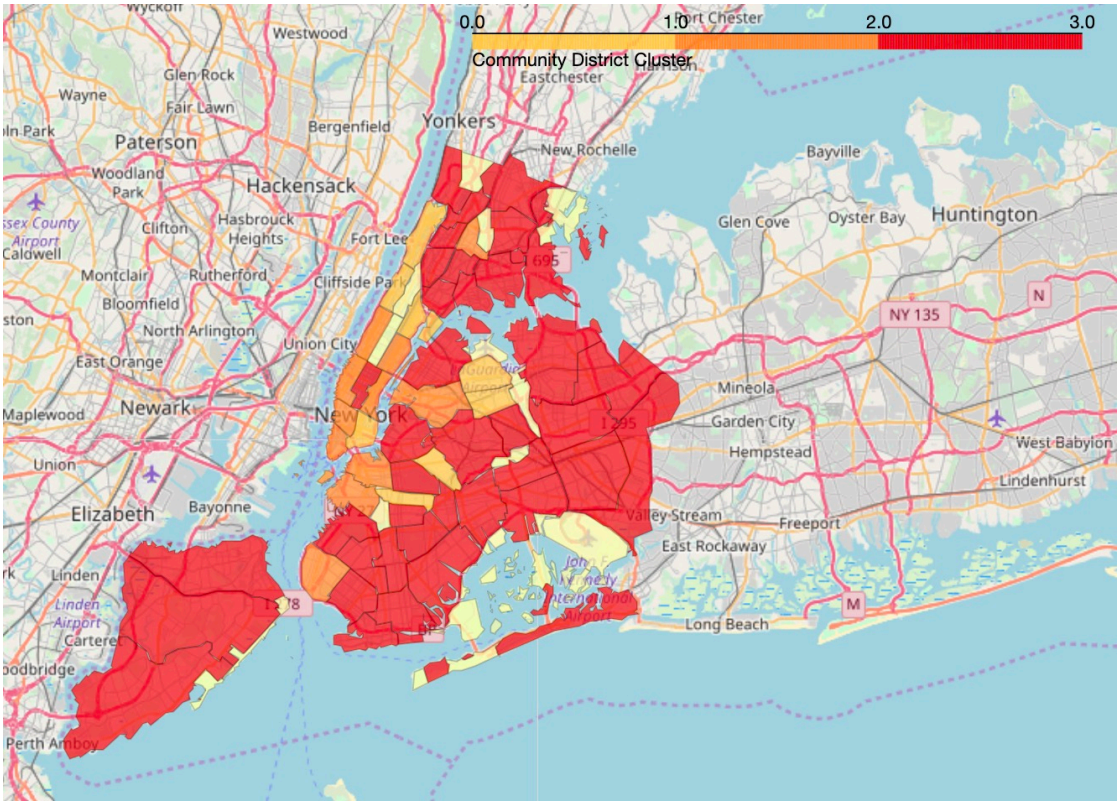




5. Modeling

To determine the most common restaurant types per community district, a kmeans clustering was used and 3 clusters were determined.

	CB	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bronx CB 1	Latin American Restaurant	Spanish Restaurant	Peruvian Restaurant	Restaurant	Mexican Restaurant
1	Bronx CB 10	Chinese Restaurant	Italian Restaurant	Fast Food Restaurant	American Restaurant	Asian Restaurant
2	Bronx CB 11	Italian Restaurant	Spanish Restaurant	Chinese Restaurant	Mexican Restaurant	Sushi Restaurant
3	Bronx CB 12	Caribbean Restaurant	Fast Food Restaurant	American Restaurant	Italian Restaurant	Seafood Restaurant
4	Bronx CB 2	Restaurant	Seafood Restaurant	Latin American Restaurant	Fast Food Restaurant	Spanish Restaurant



When examining the above graphs, I have labeled the clusters as follows:

Cluster 0 - Very dense population, Latin American and Chinese Restaurants

This cluster contains 8 Community Districts, which have a very high population per skm ratio and a predominantly Mexican, Latrin American and Chinese Restaurant present.

Cluster 1 - dense population, Italian and American Restaurants

This cluster contains 11 Community Districts, which have a mixed number of population per skm and many restaurants. These are predominantly Italian ans American Restaurants.

Cluster 2 - medium dense, differentiated restaurants

This cluster contains with 39 Community Districts the majorioty of NYC Community Districts. , which have a low to medium number of population per skm and mixed number of restaurants. The types of restaurants are generally different and no general structure can be determined.

6. Discussion

The analysis shows that although there is a great number of restaurants in New York City (~2300 in all Boroughs), there are pockets of lower restaurant density. The Highest concentration of restaurants was detected in Mahattan and in Brooklyn and Queens closer to the East River/ Manhattan.

7. Conclusion

Purpose of this project was to identify New York City Community Districts, that with lower number of restaurants (particularly Italian restaurants) in order to aid stakeholders in narrowing down the search for optimal location for a new restaurant. By calculating restaurant density distribution from Foursquare data we have identified districts with high, medium and lower number of restaurants. Clustering of those locations was then performed in order to create more information on community district restaurant information to be used as starting points for final exploration by stakeholders.

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