|  |  |
| --- | --- |
| A close up of a map  Description automatically generated  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

1. 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.

1. 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.

1. For your employees, schools, recreational activities, cultural opportunities might be very important.
2. 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.
3. 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.

1. **Data**
   1. Analytical Approach

The analytical approach for the problem, “finding the best neighborhood for a new restaurant”, includes following steps:

1. Choose the country, state and town for the restaurant

Based on own interest, the project will start with New York City

1. 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>

1. 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>

1. The number of other restaurants in a neighborhood by type

Foursquare will be used to collect the data

1. Based on the collected data, a model will be build to then use clustering to compare the neighborhoods and identify the ranking of neighborhoods.
   1. 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.

* 1. 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.

1. Exploratory Data Analysis
2. As a database I used an open database on NYC neighborhoods.

A screenshot of a cell phone

Description automatically generated

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.

A screenshot of a cell phone

Description automatically generated

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:

A screenshot of a cell phone

Description automatically generated

I used foliukm, to visualize all the neighborhoods with CB’s:

A close up of a map

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

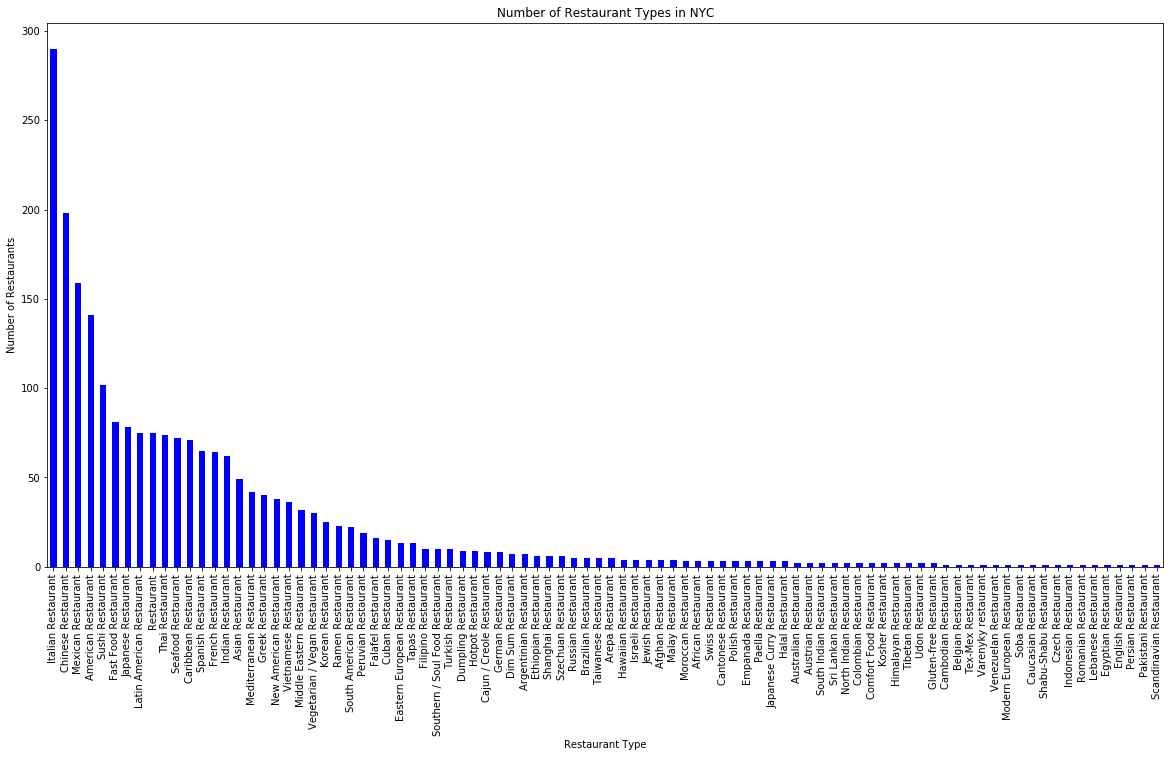
As I was only interested in Restaurants, I reduced the dataframe:

A screenshot of a cell phone

Description automatically generated

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



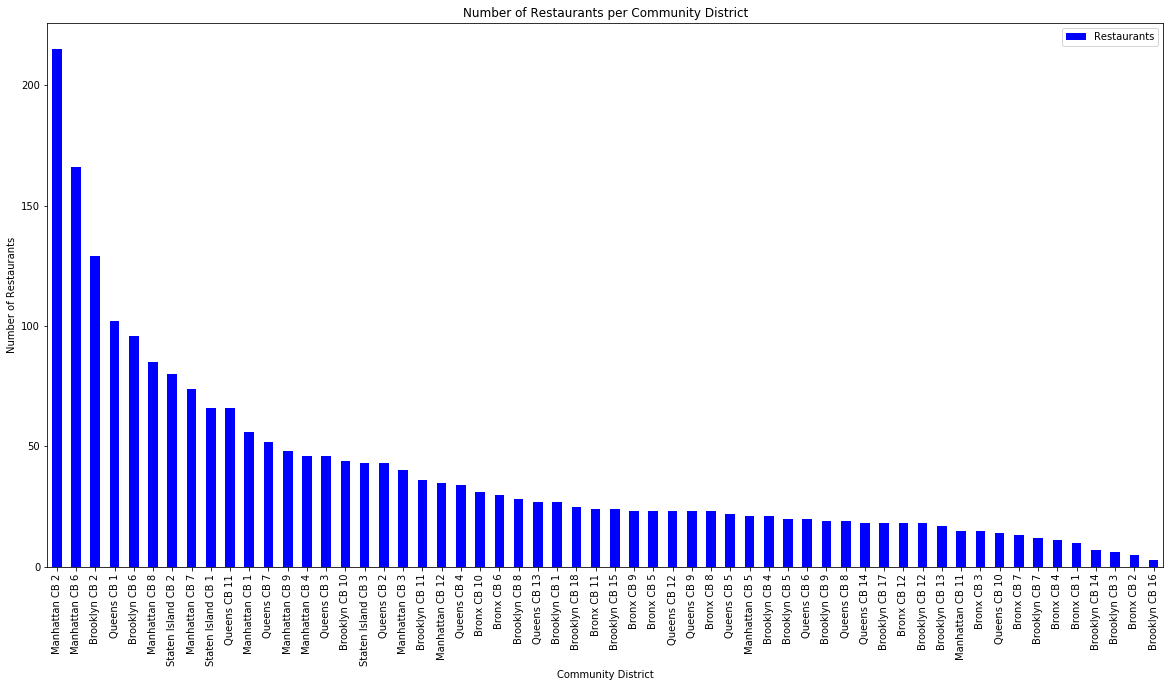


Then, how the restaurant were distributed in New York City

A picture containing text, map

Description automatically generated

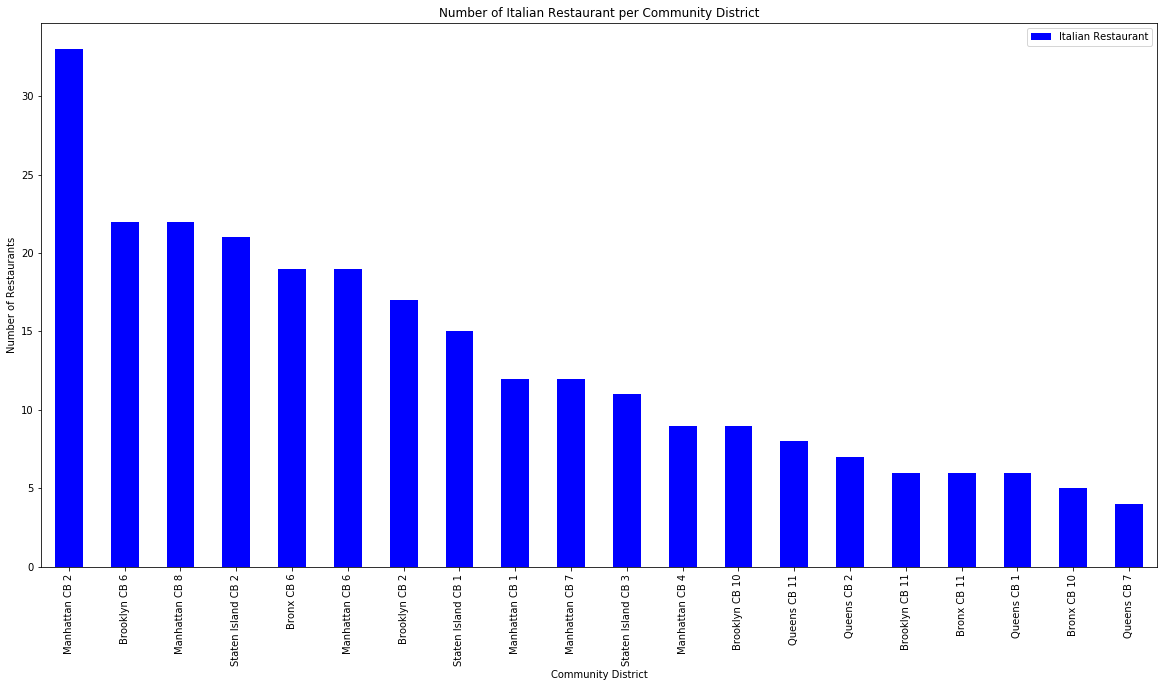
The most restaurants are in following districts:



How about Italian Restaurant Distribution in New York City

A picture containing text, map

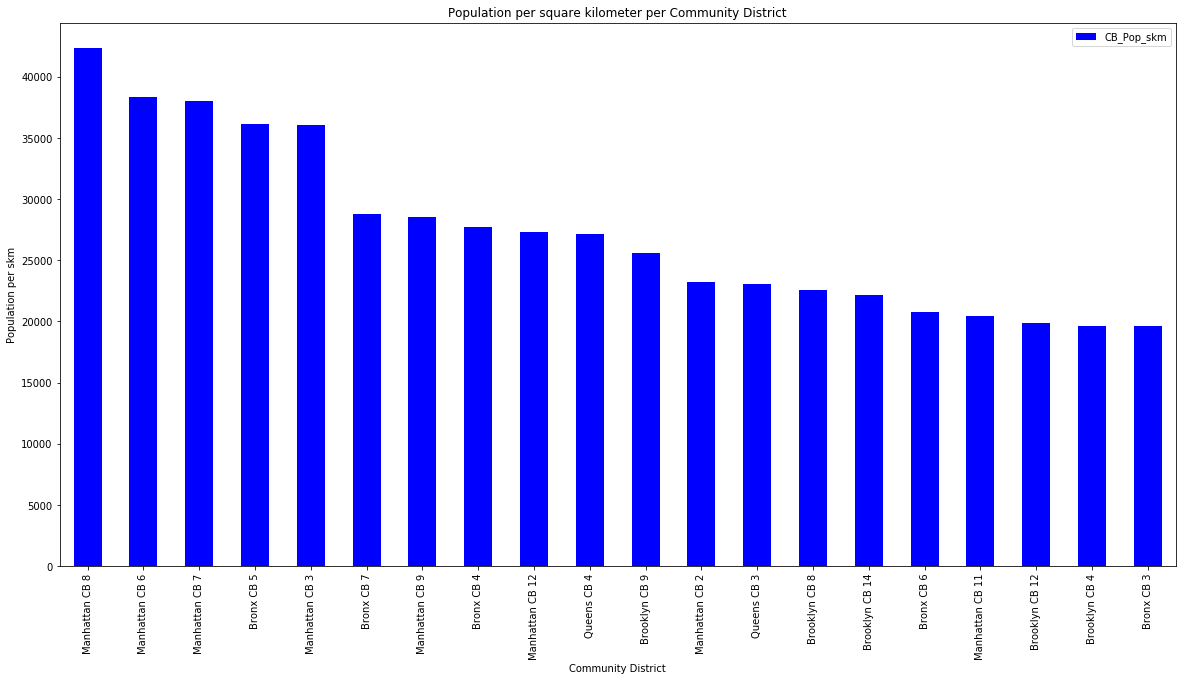
Description automatically generated



Community Districts with the most residents per square kilometer

A picture containing text, map

Description automatically generated



1. Modeling

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

A screenshot of a cell phone

Description automatically generated

### A picture containing text, map Description automatically generated

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.

1. **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.

1. **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.