Applied Data Science Capstone

Capstone Project
The Battle of Neighborhoods

José Andrés Escobar Luzuriaga

May 12, 2020

Table of contents

Introduction

New York is one of the most popular cities in the United States, divided into 5 districts, each in various neighborhoods, having in total more than 300 neighborhoods and hundreds of points of interest, including its restaurants. As the city has a rich cuisine, not only its own, but from other countries such as Chinese, Mexican, Italian and many more, there are plenty of restaurants and it is one of the most common and easiest to start a small business.

But just as it's easy to start a restaurant, it's quite difficult to get people's attention and satisfaction when it comes to food, especially when there's competition with other restaurants or food chains with more time in the sector and greater popularity. The chance of success is affected by these factors, so choosing an opening location can determine the success or failure of a restaurant.

The goal of this project is to find some promising areas to open a restaurant where competition is low or zero for any gastronomic topics. The advantages of each area will be clearly expressed so that stakeholders can choose the best possible final location to open a restaurant in this city.

Data

Based on the problem, the following necessary data sets can be identified:

- Coordinates of the different neighborhoods of New York
- Restaurants near or within the city's neighborhoods.

For the neighborhoods, it is required a dataset that provides data about each neighborhood of New York, mainly its coordinates.

There is free dataset on the web [1] and it is the same used in previous labs of the course. This dataset was created as a guide to New York City’s neighborhoods in 2014, and provides Borough Name, Neighborhood Name, Neighborhood’s Latitude and Longitude. Latitude and longitude represent an estimated coordinate of the centroid of each neighborhood.

The dataset a is a GeoJSON file with 306 objects in total and has the following structure:

```
{'type': 'Feature',
   'id': 'nyu_2451_34572.2',
   'geometry': {'type': 'Point',
    'coordinates': [-73.82993910812398, 40.87429419303012]},
   'geometry_name': 'geom',
   'properties': {'name': 'Co-op City',
    'stacked': 2,
    'annoline1': 'Co-op',
    'annoline2': 'City',
    'annoline3': None,
    'annoangle': 0.0,
    'borough': 'Bronx',
    'bbox': [
       -73.82993910812398,
       40.87429419303012,
       -73.82993910812398,
       40.87429419303012]}},
```

The key ”coordinates” has neighborhood’s longitude and latitude. Inside “properties” key is “name” (Neighborhood name) and ‘borough’ (Neighborhood’s Borough).

For restaurant data, Foursquare API provides different endpoints to get data about venues around certain location. In this case, the explore endpoint returns a list of recommended venues near the current location and has some useful parameters like limit (number of result), category id (limit the results to the specified category) and radius (to search within).

The focus of this project is any type of restaurants such as Mexican, Italian or fast food, so the appropriate category ID must be used for the Foursqueare API to return all necessary data. From the documentation [2], the category id for Food is 4d4b7105d754a06374d81259 and includes all subcategories (Mexican, Chinese, Buffet, Dessert). Neighborhoods with food-related locations that are not categorized as restaurants, such as cafeterias, bakeries, and others, will not be considered.

The API responses with a JSON that has different keys, the most important being “items” because it contains all the venues related to some location:

```
"items": [{
    "reasons": {
       "count": 0,
       "items": [{
          "summary": "This spot is popular",
          "type": "general",
          "reasonName": "globalInteractionReason"}]},
   "venue": {"id": "49b6e8d2f964a52016531fe3",
      "name": "Russ & Daughters",
      "location": {"address": "179 E Houston St",
         "crossStreet": "btwn Allen & Orchard St",
         "lat": 40.72286707707289,
         "lng": -73.98829148466851, …], …},
         "categories": [{
            "id": "4bf58dd8d48988d1f5941735",
            "name": "Gourmet Shop", …}]}},
 …]
```

The "venue" key contains all the data related to a place, has a key "name" that refers to the restaurant name, the keys "lat" and "lng" are the coordinates, and the key "category" has an identifier and a name for the corresponding place category.

Methodology

For this project, it is proposed to identify neighborhoods in New York City where a restaurant can excel, either by:

- the absence of restaurants of some gastronomy or
- the low density of restaurants.

To this end, the process was divided into 3 main steps, which will enable the objective to be achieved.

The first step is to collect the required data identified previously and prepare it, so that it is useful for the following steps.

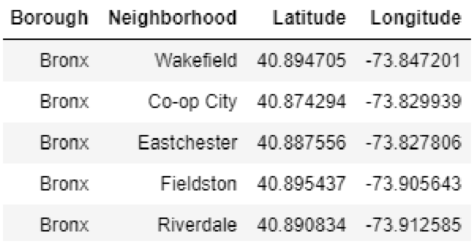
The second step of the project is an exploratory analysis of the data collected from step one, creating some graphs and maps to show the data behavior.

The last step is to create a model that can group these locations based on the amount and category of restaurants in each neighborhood. Because the data is unlabeled, an unsupervised machine learning technique is required. With the resulting clusters, it is possible to get a group of potential neighborhoods for the opening of a restaurant.

Data collection and preparation

The data sources have been identified and are aligned with the project requirements. As mentioned, the neighborhood and restaurant datasets are JSON files and have additional data that is not required for the project, so it is necessary to clear them.

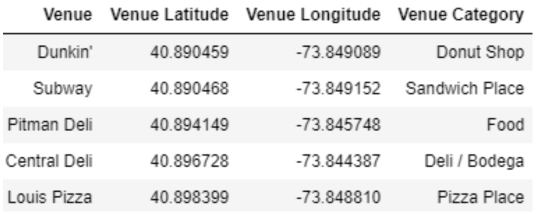
For the neighborhoods, four features are needed: Borough, Name, Latitude and Longitude. After converting these data into a panda's data frame, the result is:



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

Figure 1: Neighborhoods data frame

Restaurants need a similar treatment, five features are needed: Venue Name, Latitude, Longitude and Category.

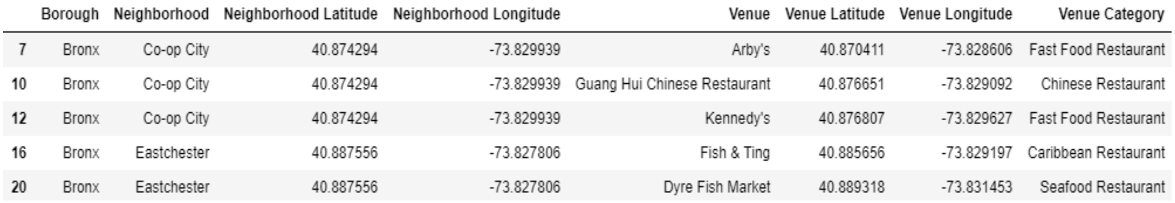


Venue	Venue Latitude	Venue Longitude	Venue Category
Dunkin'	40.890459	-73.849089	Donut Shop
Subway	40.890468	-73.849152	Sandwich Place
Pitman Deli	40.894149	-73.845748	Food
Central Deli	40.896728	-73.844387	Deli / Bodega
Louis Pizza	40.898399	-73.848810	Pizza Place

Figure 2: Restaurants data frame

This data frame has a size of 8032 and has different categories related to food, only those containing the word "Restaurant" are needed. The data frame has a total of 3949 restaurants. Almost half of the data has a different category.

Then neighborhoods and restaurants data frames are joined to create a new data frame.



	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
7	Bronx	Co-op City	40.874294	-73.829939	Arby's	40.870411	-73.828606	Fast Food Restaurant
10	Bronx	Co-op City	40.874294	-73.829939	Guang Hui Chinese Restaurant	40.876651	-73.829092	Chinese Restaurant
12	Bronx	Co-op City	40.874294	-73.829939	Kennedy's	40.876807	-73.829627	Fast Food Restaurant
16	Bronx	Eastchester	40.887556	-73.827806	Fish & Ting	40.885656	-73.829197	Caribbean Restaurant
20	Bronx	Eastchester	40.887556	-73.827806	Dyre Fish Market	40.889318	-73.831453	Seafood Restaurant

Figure 3: Result data frame

Here is a map showing all the coordinates collected. The green circles represent the neighborhoods and the blue dots represents the restaurants.

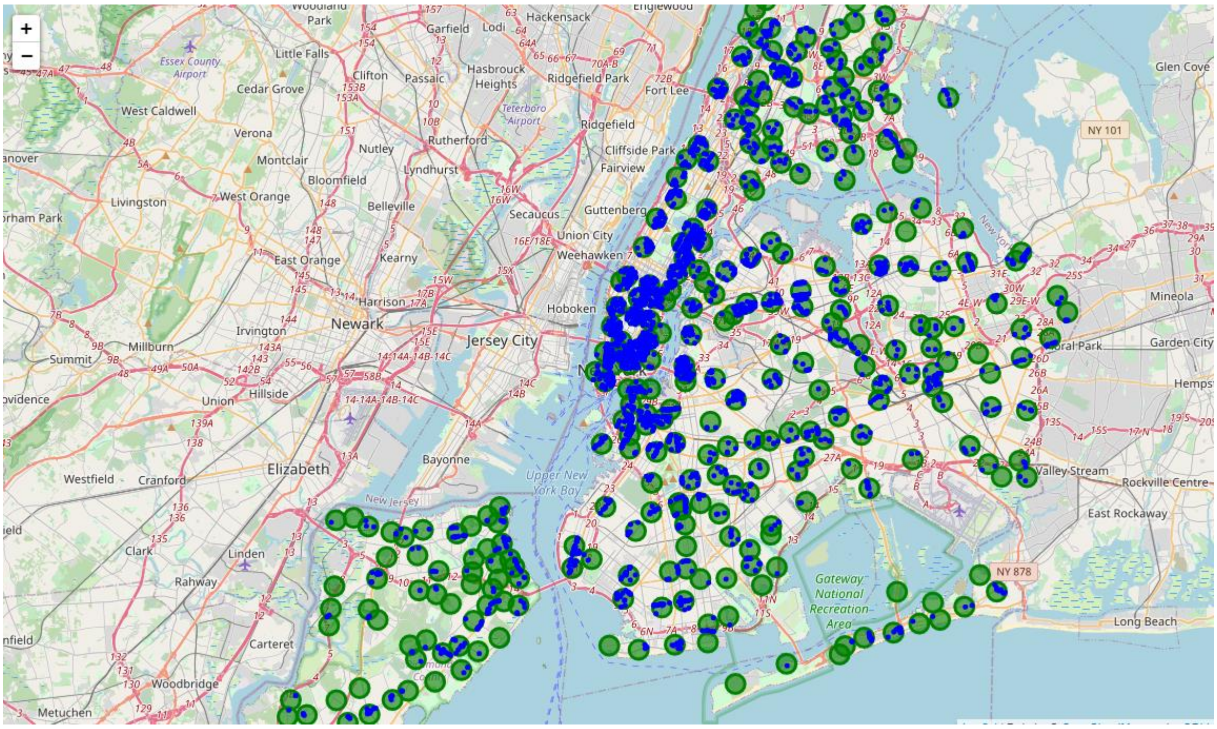


Figure 4: New York Neighborhoods and Restaurants Map

Exploratory Data Analysis

The first thing to note on the map is that some neighborhoods do not have restaurants within 500 meters radius. Those neighborhoods are:

Table 1: New York neighborhoods without restaurants nearby

Borough	**Neighborhood**	**Borough**	**Neighborhood**
Bronx	Wakefield	Staten Island	Todt Hill
Bronx	Fieldston	Staten Island	South Beach
Bronx	Riverdale	Staten Island	Port Ivory
Brooklyn	Mill Island	Staten Island	Oakwood
Brooklyn	Manhattan Beach	Staten Island	Park Hill
Brooklyn	Sea Gate	Staten Island	Graniteville
Brooklyn	Midwood	Staten Island	Butler Manor
Queens	Whitestone	Staten Island	Arden Heights
Queens	Broad Channel	Staten Island	Greenridge
Queens	Breezy Point	Staten Island	Heartland Village
Queens	Neponsit	Staten Island	Bloomfield
Queens	Holliswood	Staten Island	Emerson Hill
Queens	Somerville	Staten Island	Randall Manor
Queens	Brookville	Queens	Bayswater

The dataframe has no numeric values, so a count of different features will be performed to obtain some graphs where trends can be observed. Based on the map (Figure 4), Manhattan has a lot of neighborhoods and restaurants. And in fact, that is the case (Figure 5) with approximately 1500 restaurants, being the most popular borough in New York and the third most populous and the smallest of all. On the other hand, Staten Island has the fewest restaurants (less than 250) and less populated than the 5 boroughs but the third largest.

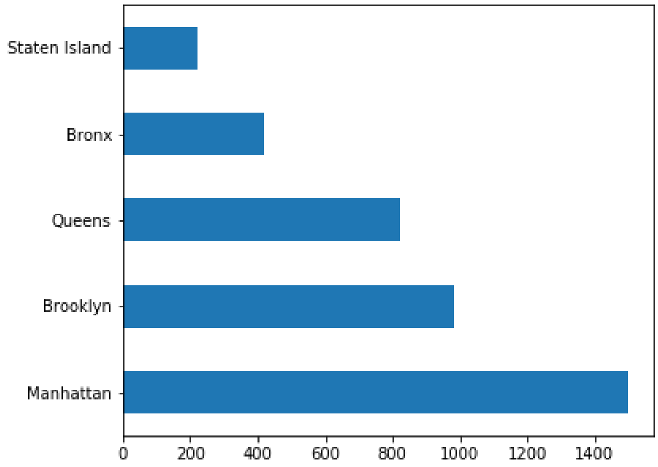


Figure 5: Number of restaurants per Borough

Across all districts, Chinese and Italian restaurants are the two most common categories with a quantity close to 500 and 450 respectively. It is followed by Mexican, American and Sushi restaurants (Figure 6).

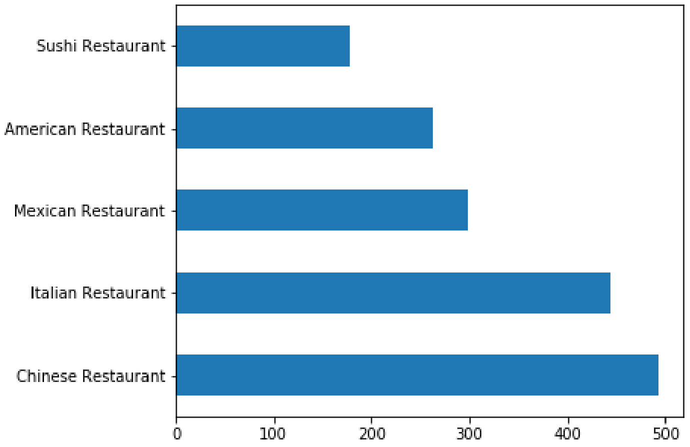
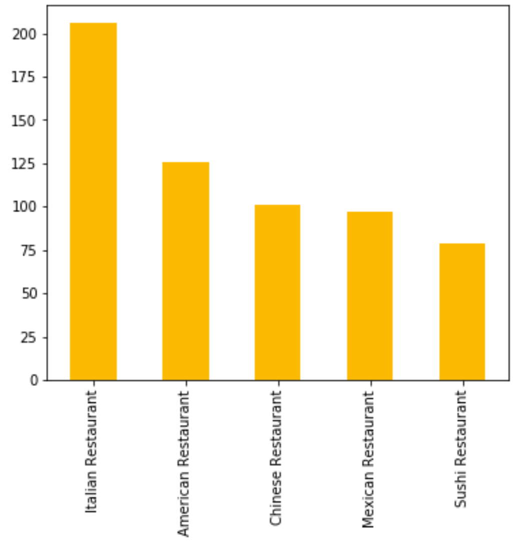
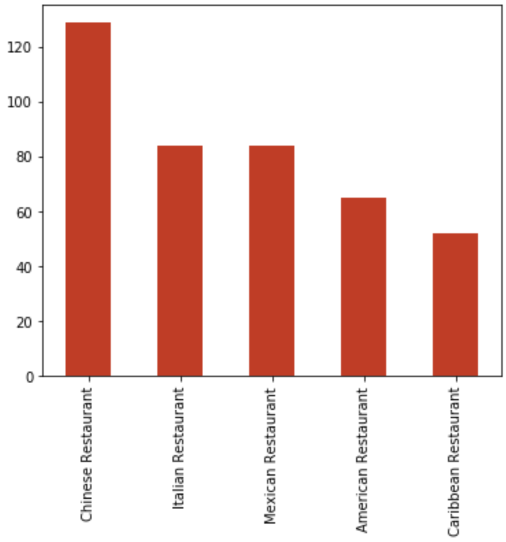


Figure 6: Top 5 of most common restaurants' categories of New York

Let's see how these restaurants are distributed in all 5 boroughs. In Manhattan, the most common restaurant is Italian with more than 200, followed by nearly 125 American restaurants and about 100 Chinese restaurants. Brooklyn has nearly 130 Chinese restaurants and more than 80 Italian and Mexican restaurants. In Queens, Chinese, Korean and Mexican are the 3 most common restaurants with approximately 135, 65 and 55 each. The Bronx has more than 95 Chinese restaurants and more than 40 Italian and Mexican restaurants, For Staten Island, there are more than 55 Italian restaurants, approximately 30 Chinese restaurants and nearly 25 American restaurants.

It is clear with this that Chinese restaurants are found in large numbers in each of the city's borough, entering the top 3 of the most common restaurants in each borough. The same with Italian restaurants, being present in the top 5 of the most common restaurants in each of the borough.





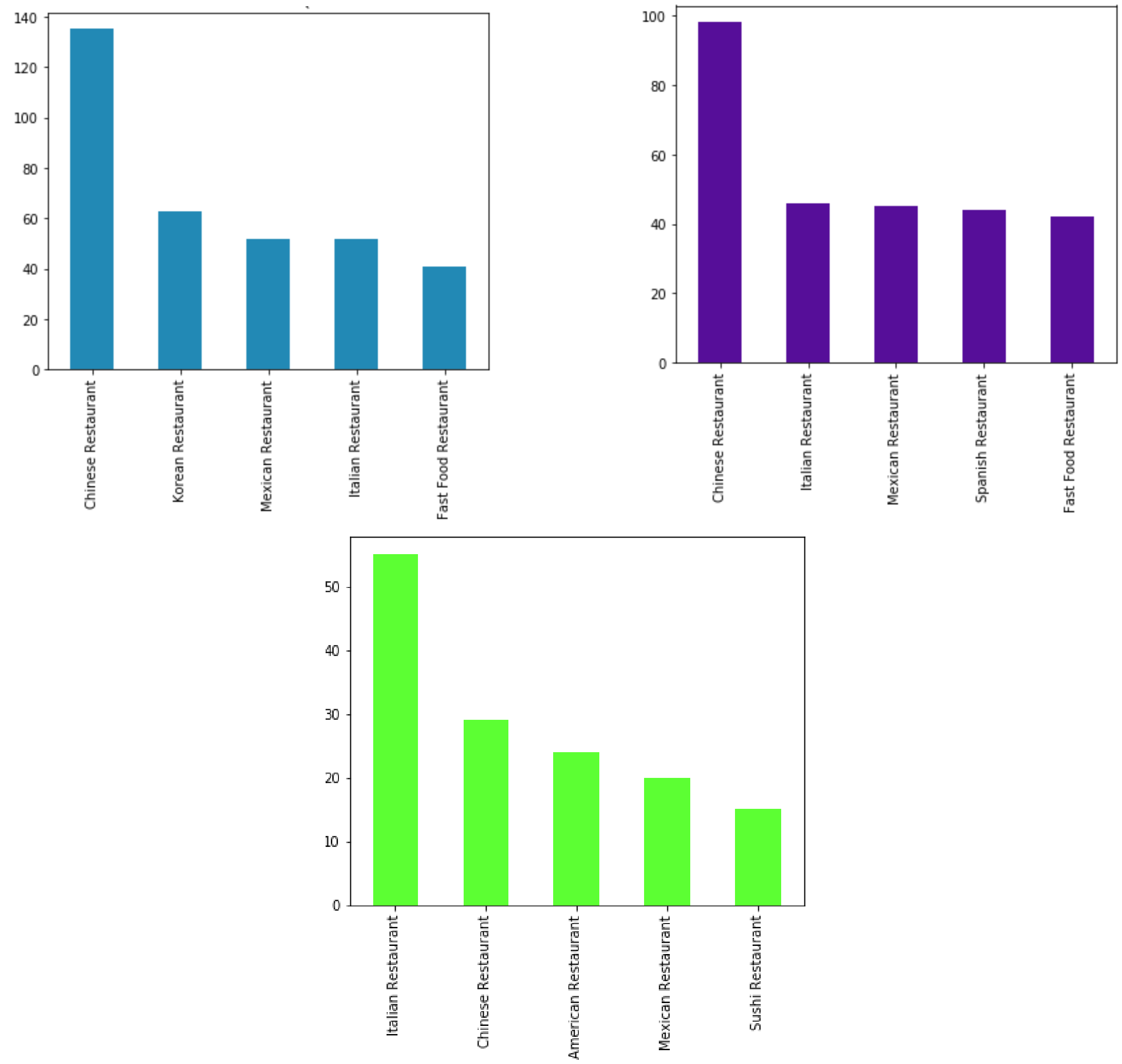


Figure 7: Top 5 of most common restaurants' categories of Manhattan (top left), Brooklyn (top right), Queens (middle left), The Bronx (middle right) and Staten Island (bottom)

Model development

As mentioned before, an unsupervised machine learning technique is used for this project because data is unlabeled and the purpose is to group neighborhoods that have similar characteristics, in this case, similar restaurants. A good algorithm for portion-based clustering is K-means, it divides the data into non-overlapping clusters minimizing the intra-cluster distance and maximizing the inter-cluster distances. This means that each cluster has neighborhoods that are dissimilar from neighborhoods in other clusters.

Before using the K-means, some data normalization in needed. The dataframe only contains categorical variables, the first step is to convert them to numeric variables using indicators. An indicator is a numerical value (0 or 1) used to label categories, in this case, restaurant categories. Grouping by neighborhoods and using Simple Feature Normalization, those values are converted into a similar range, from 0 to 1.

For the model, a value of 10 is used for the cluster number that the algorithm will produce. After training the model with the data, it returns an array of numbers that represent the cluster label for each neighborhood. Using those labels, a new map of New York was created showing all clusters with different colors.

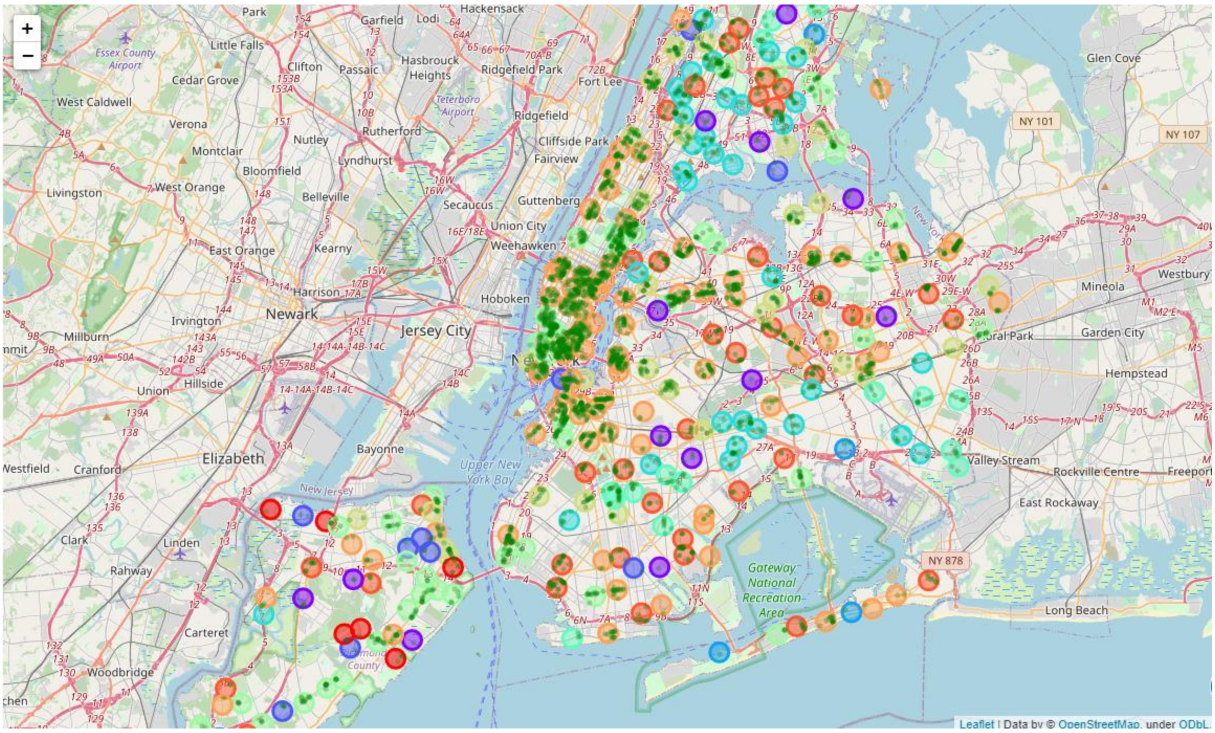


Figure 8: New York Map with Neighborhood Clusters

Most of the neighborhoods are in clusters from 4 to 8 (more than 15). Cluster 3, 0, 2 and 1 have the lowest number of neighborhoods (less than 15).

Table 2: Number of Neighborhoods per Cluster Label

Cluster Labels	N° Neighborhoods
3	5
0	6
2	11
1	13
5	15
7	31
4	33
9	38
6	47
8	79

Since the project interest is neighborhoods with few restaurants, it is important to know how many restaurants exist in each cluster (Table 3). With fewer neighborhoods, fewer restaurants obviously existed, but these cluster may be related by the number of restaurants and share the same gastronomy.

Cluster 3 (Table 4) has 5 neighborhoods where the most common restaurants are fast food restaurants within a 500-meter radius. 4 out of 5 neighborhoods have fast food restaurants as their only restaurants.

Cluster 0 (Table 5) has 6 neighborhoods where the most common restaurants are Italian. All of them located in Staten Island. 5 out of 6 neighborhoods have only Italian Restaurant within the 500-meter radius. Italian food is the second most common gastronomy in NY and the most common in Staten Island.

Cluster 2 (Table 6) has a wider variety of restaurants but the most common are American restaurants. That group doesn't look as good as the previous ones.

Like Cluster 2, Cluster 1 (Table 7) has a greater variety of restaurants. This groups correspond to the Chinese restaurants that are the most common and saturated gastronomy in New York.

Table 3: Number of Restaurants per Cluster Label

Cluster Labels	N° Restaurants
3	8
0	13
2	41
1	32
5	99
7	341
4	255
9	357
6	775
8	2139

To segment the groups a bit, a value of 2 was used as the maximum of restaurants per neighborhood to consider it for the analysis.

Table 4: Number of Restaurants per Neighborhood in Cluster Label 3

Borough	Neighborhood	N° Restaurants
Queens	South Ozone Park	2
Staten Island	Pleasant Plains	1
Queens	Roxbury	1
Queens	Hammels	1

Table 5: Number of Restaurants per Neighborhood in Cluster Label 0

Borough	Neighborhood	N° Restaurants
Staten Island	Mariner's Harbor	2
Staten Island	Howland Hook	1
Staten Island	Egbertville	1
Staten Island	Lighthouse Hill	1

Table 6: Number of Restaurants per Neighborhood in Cluster Label 2

Borough	Neighborhood	N° Restaurants
Bronx	Clason Point	2
Staten Island	Grymes Hill	1
Staten Island	Silver Lake	1
Staten Island	Arlington	2
Staten Island	Richmond Town	2
Brooklyn	Madison	1
Staten Island	Fox Hills	2

Table 7: Number of Restaurants per Neighborhood in Cluster Label 1

Borough	Neighborhood	N° Restaurants
Bronx	Edenwald	1
Brooklyn	Marine Park	1
Queens	Glendale	1
Staten Island	Midland Beach	1
Queens	Blissville	1
Staten Island	Willowbrook	2

A map with these locations, including those without restaurants, are show in the following figure

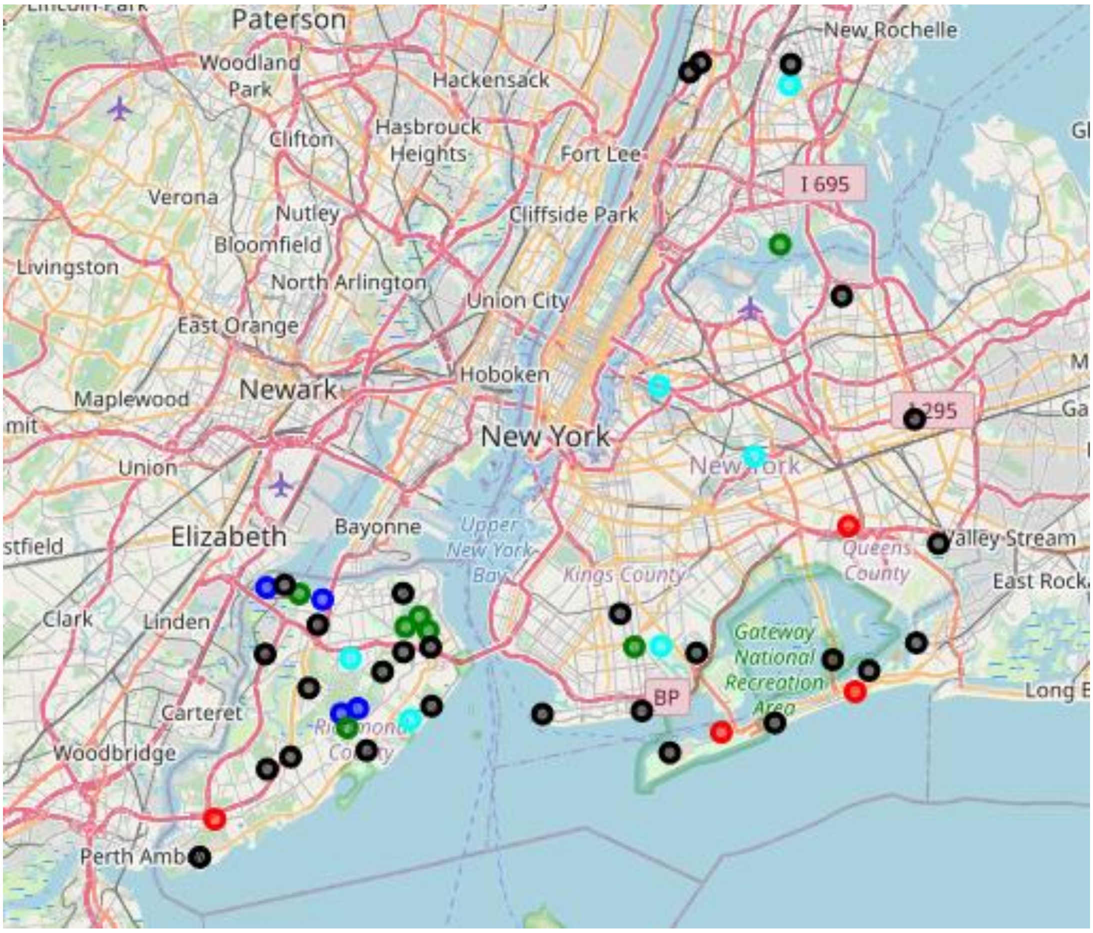


Figure 9: Map of New York with promising neighborhoods.

Results

As we can see, Manhattan has no markers, as Manhattan is one of the most popular boroughs in New York and is saturated with restaurants, so it would not be an ideal area to open a restaurant in front of so much competition.

Brooklyn is probably the second most popular borough and the second most populated. There are 4 neighborhoods with no restaurants nearby and 2 neighborhoods with less than 3 restaurants nearby. Marine Park seems to be a good area as it only has one Chinese Restaurant.

Queens is the third most populated borough and it is home of two airports, so a lot of tourists pass through there. Blissville is a good neighborhood since only have one Chinese restaurant and is within Long Island City, a residential and commercial area.

The Bronx has many middle-income neighborhoods, such as Wakefield, Fieldstone and Riverdale. These are good areas to open a restaurant since there are no restaurants nearby and are part commercial areas.

Staten Island is the least populated and mostly suburban of the five boroughs. Although the map shows a lot of locations, almost all of them are residential. Although, South Beach may be a good spot for some Sea food restaurants.

Table 8: Neighborhoods

N°	Borough	Neighborhood	N°	Borough	Neighborhood
1	Bronx	Wakefield	26	Staten Island	South Beach
2	Bronx	Fieldston	27	Staten Island	Grymes Hill
3	Bronx	Clason Point	28	Staten Island	Silver Lake
4	Bronx	Riverdale	29	Staten Island	Arlington
5	Bronx	Edenwald	30	Staten Island	Richmond Town
6	Brooklyn	Sea Gate	31	Staten Island	Fox Hills
7	Brooklyn	Marine Park	32	Staten Island	Midland Beach
8	Brooklyn	Midwood	33	Staten Island	Willowbrook
9	Brooklyn	Madison	34	Staten Island	Port Ivory
10	Brooklyn	Mill Island	35	Staten Island	Oakwood
11	Brooklyn	Manhattan Beach	36	Staten Island	Park Hill
12	Queens	Glendale	37	Staten Island	Graniteville
13	Queens	Blissville	38	Staten Island	Butler Manor
14	Queens	Whitestone	39	Staten Island	Arden Heights
15	Queens	Broad Channel	40	Staten Island	Greenridge
16	Queens	Breezy Point	41	Staten Island	Heartland Village
17	Queens	Neponsit	42	Staten Island	Bloomfield
18	Queens	Holliswood	43	Staten Island	Emerson Hill
19	Queens	Somerville	44	Staten Island	Randall Manor
20	Queens	Brookville	45	Staten Island	Pleasant Plains
21	Queens	South Ozone Park	46	Staten Island	Mariner's Harbor
22	Queens	Bayswater	47	Staten Island	Howland Hook
23	Queens	Roxbury	48	Staten Island	Egbertville
24	Queens	Hammels	49	Staten Island	Lighthouse Hill
25	Staten Island	Todt Hill			

Discussion

The project basically consisted of grouping neighborhoods in New York City to identify areas to open a restaurant. One of the drawbacks is that for such problems and data characteristics, unsupervised machine learning techniques are used, which are less controllable since the process is done by the algorithm itself and there are not many evaluation methods to know how good the result is.

For certain groups that the K-means algorithm obtained, there are several where the number of neighborhoods is exaggeratedly large and for others it is too small. Those small groups were grouped in a certain way correctly because they contained restaurants of the same category and similar amount like cluster 0 and 3.

Another drawback is that the data used are from neighborhoods and some of these are entirely residential areas (as in Staten Island). Geo-data could be used for tourist or points of interest where there is a greater influx of people.

Conclusion

In this project, it was possible to obtain a group of neighborhoods form New York City where the number of restaurants is low or none. Using neighborhood geodata and the Foursquare API, explored the nearby restaurants within a radius of approximately 500 meters from the center of each neighborhood. An unsupervised machine learning model was created using K-means algorithm to cluster all the neighborhoods based on the restaurants nearby.

Then a map was created showing the locations of the selected neighborhoods with colored indicators for easy identification of those places that do not have any restaurants.

Even though the model created is not the best and can be improved, many potential neighborhoods were obtained to choose from, and it is left to the interested parties to choose the neighborhood that best suits their needs.

References

[1] NYU, "2014 New York City Neighborhood Names," [Online]. Available: https://geo.nyu.edu/catalog/nyu_2451_34572. [Accessed 11 May 2020].

[2] Foursquare, "Venue Categories," [Online]. Available: https://developer.foursquare.com/docs/build-with-foursquare/categories/. [Accessed 11 May 2020].