# Clustering NYC Neighborhoods Based on Available Restaurants

Sumit Saha

July 3, 2019

## 1. Introduction

### 1.1 Background

The City of New York is the most populous city in the United States. New York City's food culture includes an array of international cuisines influenced by the city's immigrant history. The city is home to nearly one thousand of the finest and most diverse haute cuisine restaurants in the world. Good food is something thing that makes people happy and in any corner of the world people always try to find the best food available. The available Restaurants in a neighborhood also play a great role to attract people towards a neighborhood. Analyzing the available restaurant categories in different neighborhoods would give us an idea of the distribution of restaurant types across the city.

### 1.2 Problem

The problem here is to determine which place in the city has which type of restaurants more. The New York City neighborhoods need to be clustered  to find which type of restaurants dominate which part of the city.

### 1.3 Interest

Different stakeholders may be interested in a model which can identify different types of areas NY based on available restaurant types. People who are planning to move to a new place, they can determine which place to choose if good food plays a great role in their life. If some start-up is looking for opening a new restaurant they can find which place would be appropriate that means in that area same kind of restaurants (type of restaurant to be opened) are not available.

# 2. Data

## 2.1 Data Sources

To solve the above problem I are going to use data from mainly two sources –

1. **NYC neighborhood dataset with locations (latitude and longitude)** : This data set contains columns Borough, Neighborhood, longitude and latitude. The dataset was downloaded from IBM skills network labs. Few rows from the dataset is shown below.

| | 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 |

2. **Foursquare'Places API'** is used to pull data of the venues around different neighborhoods. Data retrieval was done based on venue category '4d4b7105d754a06374d81259' which represents venues related to food. Sending request to the api gave us data in json format. The json data then formatted into a pandas dataframe.

| | Neighborhood | Neighborhood_Latitude | Neighborhood_Longitude | Venue_Name | Venue_Category | Venue_Latitude | Venue_Longitude | Venue_city | Venue_State |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Wakefield | 40.894705 | -73.847201 | Cooler Runnings Jamaican Restaurant Inc | Caribbean Restaurant | 40.898276 | -73.850381 | Bronx | NY |
| 1 | Wakefield | 40.894705 | -73.847201 | Dunkin' | Donut Shop | 40.890459 | -73.849089 | Bronx | NY |
| 2 | Wakefield | 40.894705 | -73.847201 | SUBWAY | Sandwich Place | 40.890656 | -73.849192 | Bronx | NY |
| 3 | Wakefield | 40.894705 | -73.847201 | Pitman Deli | Food | 40.894149 | -73.845748 | Bronx | NY |
| 4 | Wakefield | 40.894705 | -73.847201 | Baychester Avenue Food Truck | Food Truck | 40.892293 | -73.843230 | Bronx | NY |

## 2.2 Exploratory Data Analysis

Now it's time to analyze the data to clean it and prepare for visualization and clustering.

1. From the data retrieved by the api call I will find what are the state names Ire returned. I could see that NJ, NY, New York, these names Ire found in the dataset. For our analysis I are going to keep data only related to NY and New York.
2. For some records city name parameter was not filled. 64 data entries Ire found where city name was not provided.
3. While checking for null values it is found that there Ire no such cell which have null values. So this helped us reducing one step in data cleaning.

4. The Dataset had 132 different types of venue categories but all the categories are not related to restaurants. As I fetched data for all the venues related to food, the venues are also there in the dataset which are not restaurants .

### 2.3 Data Cleaning

1. Records for which venue states are not related to ‘NY’ and ‘New York’ are dropped and all the New Yorks are replaced with NY. One record was removed.

2. 64 Entries with city name as ‘N/A’ Ire removed.

3. From the whole dataset I wanted to concentrate of venue categories related to restaurants. So I created a list of restaurant types('Chinese Restaurant', 'Italian Restaurant', 'Mexican Restaurant', 'American Restaurant', 'Fast Food Restaurant', 'Sushi Restaurant', 'Japanese Restaurant', 'Latin American Restaurant', 'Thai Restaurant', 'Spanish Restaurant', 'Caribbean Restaurant', 'Seafood Restaurant', 'Korean Restaurant', 'Indian Restaurant', 'French Restaurant'). After filtering the data I removed 4254 entries from the dataset.

After cleaning the data set it had all the data related to our desired restaurant types.

| | Neighborhood | Neighborhood_Latitude | Neighborhood_Longitude | Venue_Name | Venue_Category | Venue_Latitude | Venue_Longitude | Venue_city | Venue_State |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Wakefield | 40.894705 | -73.847201 | Cooler Runnings Jamaican Restaurant Inc | Caribbean Restaurant | 40.898276 | -73.850381 | Bronx | NY |
| 7 | Co-op City | 40.874294 | -73.829939 | Arby's | Fast Food Restaurant | 40.870518 | -73.828657 | Bronx | NY |
| 10 | Co-op City | 40.874294 | -73.829939 | Guang Hui Chinese Restaurant | Chinese Restaurant | 40.876603 | -73.829710 | Bronx | NY |
| 12 | Co-op City | 40.874294 | -73.829939 | Kennedy's | Fast Food Restaurant | 40.876807 | -73.829627 | Bronx | NY |
| 16 | Eastchester | 40.887556 | -73.827806 | Fish & Ting | Caribbean Restaurant | 40.885539 | -73.829151 | Bronx | NY |

# 3. Encoding and Visualization

### 3.1 One-Hot-Encoding venue categories

To use Foursquare’s category values to find similar neighborhoods based on restaurant types, a one-hot-encoding representation of each entry was created using Pandas’ ‘get_dummies’ function. The result was a dataframe of New York City restaurant related venues where entry venue category is represented by a value of 1 in the column of matching venue category.

| | Neighborhood | American Restaurant | Caribbean Restaurant | Chinese Restaurant | Fast Food Restaurant | French Restaurant | Indian Restaurant | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Latin American Restaurant | Mexican Restaurant | Sea Restau |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Wakefield | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **7** | Co-op City | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **10** | Co-op City | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **12** | Co-op City | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **16** | Eastchester | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **19** | Eastchester | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **22** | Eastchester | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **23** | Eastchester | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **25** | Eastchester | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **26** | Eastchester | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **30** | Kingsbridge | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |

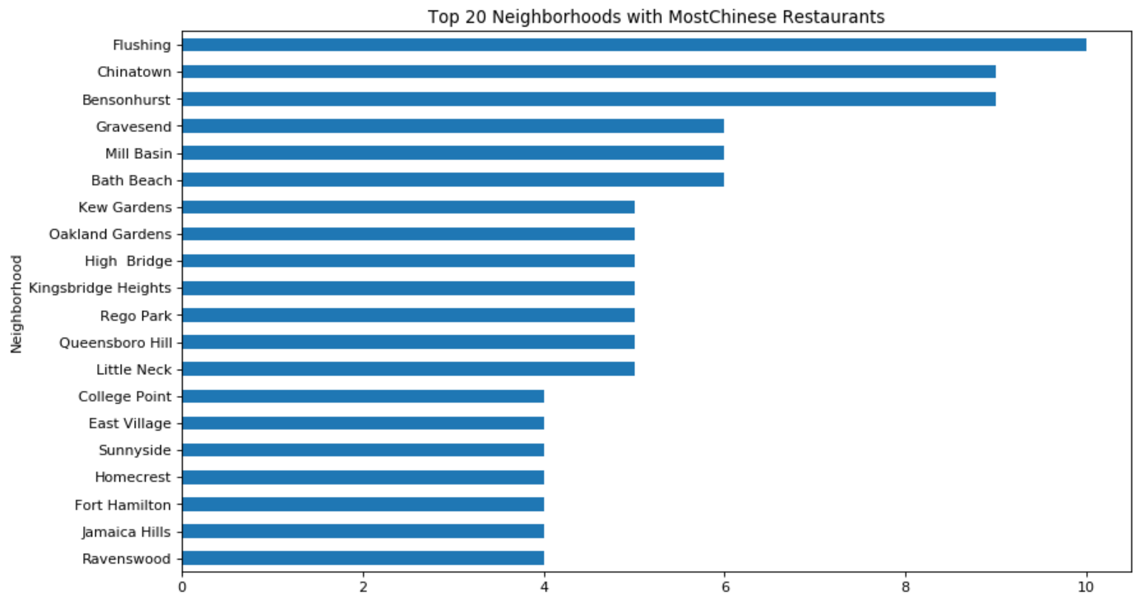
Counts of the venues Ire determined for each venue category and neighborhood in New York City using the one hot encoded DataFrame.

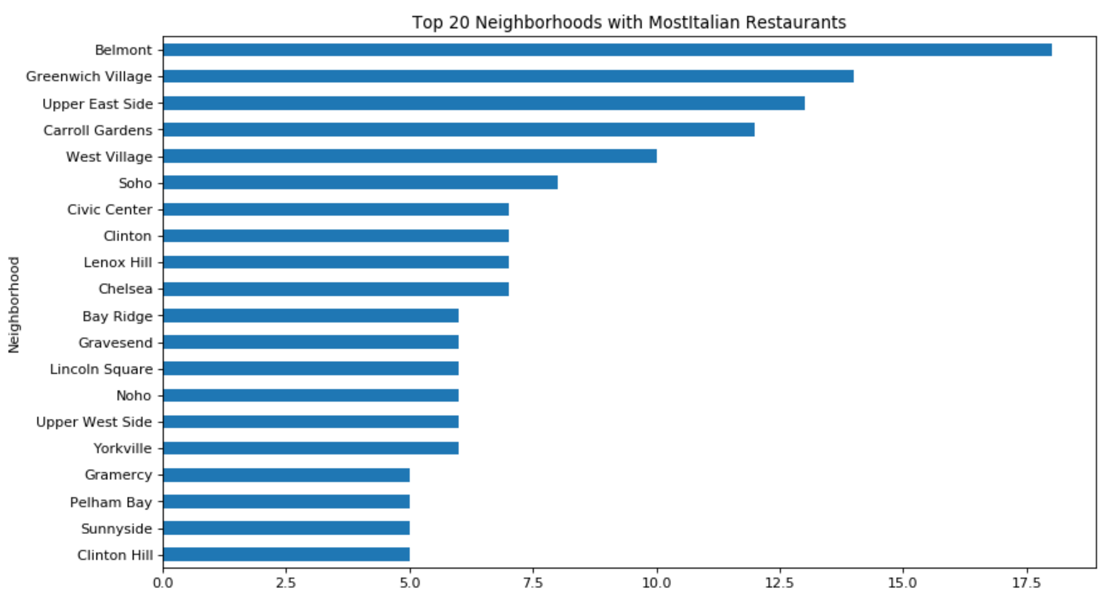
| Neighborhood | American Restaurant | Caribbean Restaurant | Chinese Restaurant | Fast Food Restaurant | French Restaurant | Indian Restaurant | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Latin American Restaurant | Mexican Restaurant | Seafood Restaurant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Allerton** | 1 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **Annadale** | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Arlington** | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Arrochar** | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| **Arverne** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Astoria** | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 0 | 1 | 2 | 1 | 4 |
| **Astoria Heights** | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| **Auburndale** | 1 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 3 | 0 | 0 | 0 |

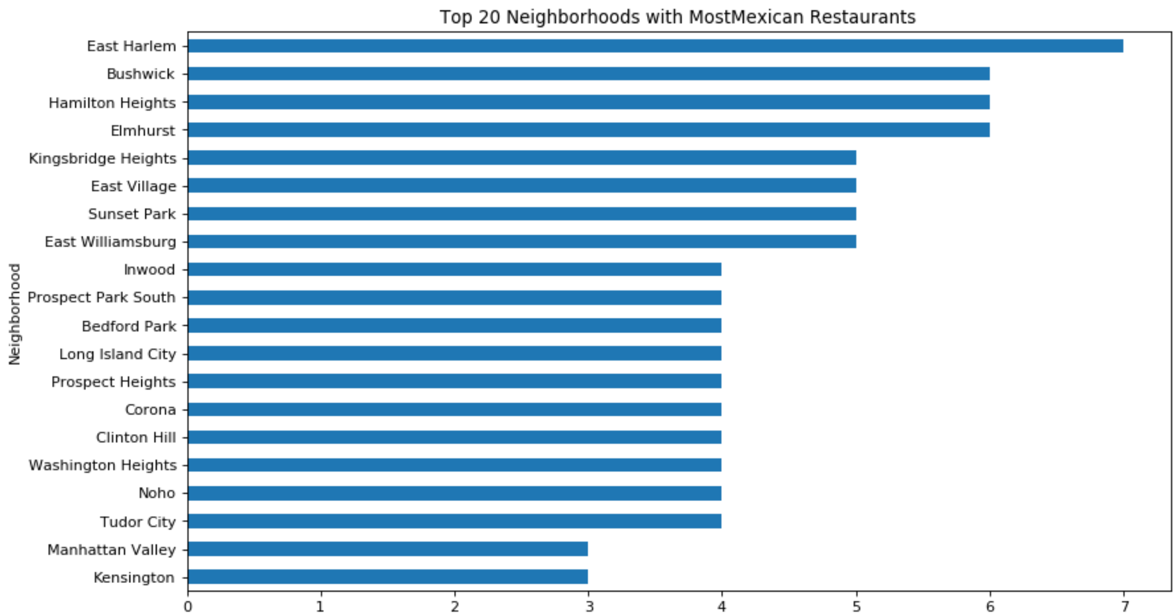
## 3.2 Data Visualization

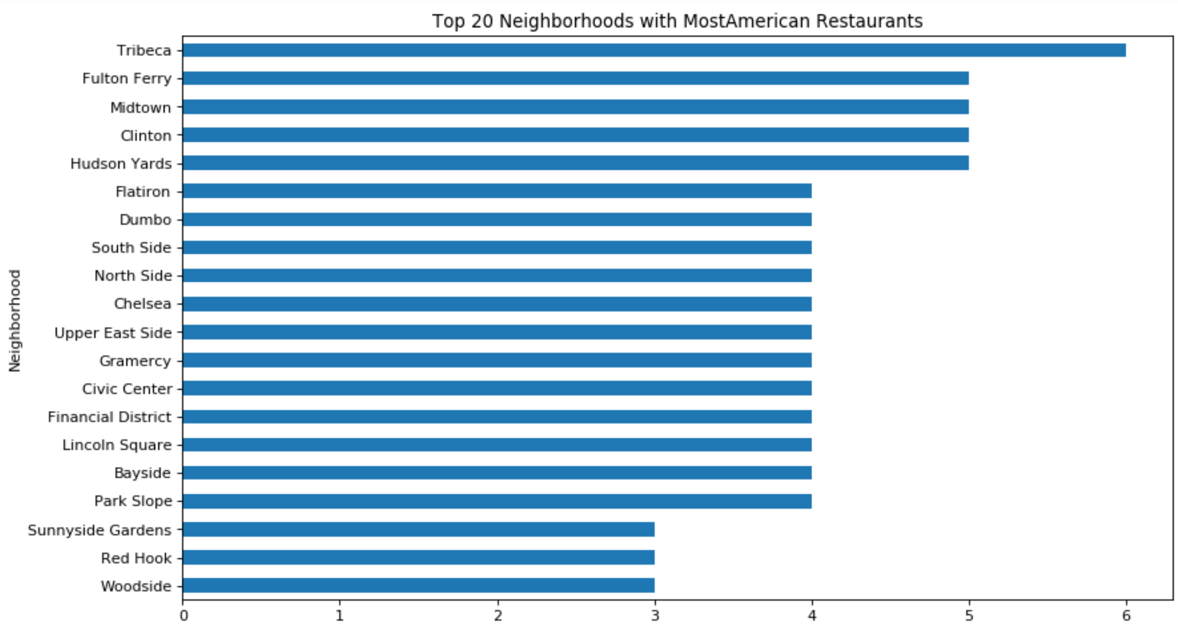
To visualize the data I plotted bar plots of top 20 neighborhoods with most of the category available

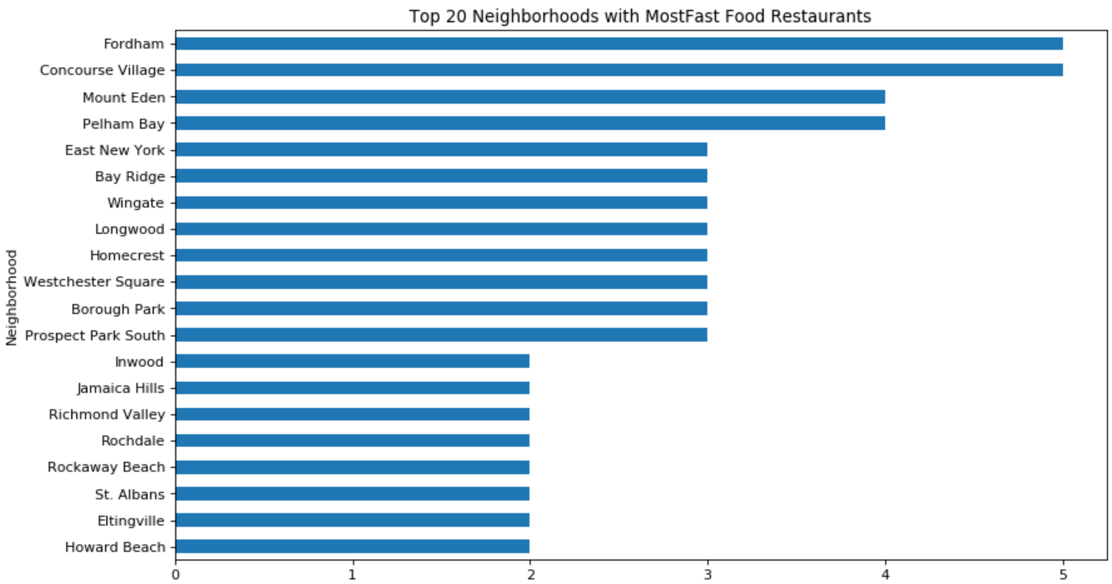
Here I have shown to visualize data for this list of categories(('Chinese Restaurant', 'Italian Restaurant', 'Mexican Restaurant', 'American Restaurant', 'Fast Food Restaurant'))


Top 20 Neighborhoods with MostChinese Restaurants
Neighborhood
Flushing
Chinatown
Bensonhurst
Gravesend
Mill Basin
Bath Beach
Kew Gardens
Oakland Gardens
High  Bridge
Kingsbridge Heights
Rego Park
Queensboro Hill
Little Neck
College Point
East Village
Sunnyside
Homecrest
Fort Hamilton
Jamaica Hills
Ravenswood
0
2
4
6
8
10


Top 20 Neighborhoods with MostItalian Restaurants
Neighborhood
Belmont
Greenwich Village
Upper East Side
Carroll Gardens
West Village
Soho
Civic Center
Clinton
Lenox Hill
Chelsea
Bay Ridge
Gravesend
Lincoln Square
Noho
Upper West Side
Yorkville
Gramercy
Pelham Bay
Sunnyside
Clinton Hill
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Top 20 Neighborhoods with MostMexican Restaurants
Neighborhood
East Harlem
Bushwick
Hamilton Heights
Elmhurst
Kingsbridge Heights
East Village
Sunset Park
East Williamsburg
Inwood
Prospect Park South
Bedford Park
Long Island City
Prospect Heights
Corona
Clinton Hill
Washington Heights
Noho
Tudor City
Manhattan Valley
Kensington
0
1
2
3
4
5
6
7


Top 20 Neighborhoods with MostAmerican Restaurants
Neighborhood
Tribeca
Fulton Ferry
Midtown
Clinton
Hudson Yards
Flatiron
Dumbo
South Side
North Side
Chelsea
Upper East Side
Gramercy
Civic Center
Financial District
Lincoln Square
Bayside
Park Slope
Sunnyside Gardens
Red Hook
Woodside
0
1
2
3
4
5
6




# 4. Results and Discussion

From the above analysis on the data I am very close to the solution of our problem. I mainly focused on two findings.

First, I had to find which type of restaurants dominates a particular neighborhood by its availability. To find that we grouped our encoded dataset and transformed into a table which displays the top five types of restaurants for each neighborhood. After processing the data I got the result something like this-

| | Neighborhood | 1st Top Venue Category | 2nd Top Venue Category | 3rd Top Venue Category | 4th Top Venue Category | 5th Top Venue Category |
|---|---|---|---|---|---|---|
| 0 | Allerton | Spanish Restaurant | Chinese Restaurant | Fast Food Restaurant | American Restaurant | Mexican Restaurant |
| 1 | Annadale | American Restaurant | Sushi Restaurant | Thai Restaurant | Spanish Restaurant | Seafood Restaurant |
| 2 | Arlington | Caribbean Restaurant | American Restaurant | Thai Restaurant | Sushi Restaurant | Spanish Restaurant |
| 3 | Arrochar | Italian Restaurant | Thai Restaurant | Sushi Restaurant | Spanish Restaurant | Seafood Restaurant |
| 4 | Arverne | Thai Restaurant | Sushi Restaurant | Spanish Restaurant | Seafood Restaurant | Mexican Restaurant |

Second, I had to find which types of restaurants are least available around a particular neighborhood. I have the same process to group the dataset but this time in a different manner to find the bottom five restaurant types for each neighborhood. After processing the data I got the result something like this-

click to scroll output; double click to hide

| | Neighborhood | 1st Bottom Venue Category | 2nd Bottom Venue Category | 3rd Bottom Venue Category | 4th Bottom Venue Category | 5th Bottom Venue Category |
|---|---|---|---|---|---|---|
| 0 | Allerton | Caribbean Restaurant | French Restaurant | Indian Restaurant | Italian Restaurant | Japanese Restaurant |
| 1 | Annadale | Caribbean Restaurant | Chinese Restaurant | Fast Food Restaurant | French Restaurant | Indian Restaurant |
| 2 | Arlington | Chinese Restaurant | Fast Food Restaurant | French Restaurant | Indian Restaurant | Italian Restaurant |
| 3 | Arrochar | American Restaurant | Caribbean Restaurant | Chinese Restaurant | Fast Food Restaurant | French Restaurant |
| 4 | Arverne | American Restaurant | Caribbean Restaurant | Chinese Restaurant | Fast Food Restaurant | French Restaurant |

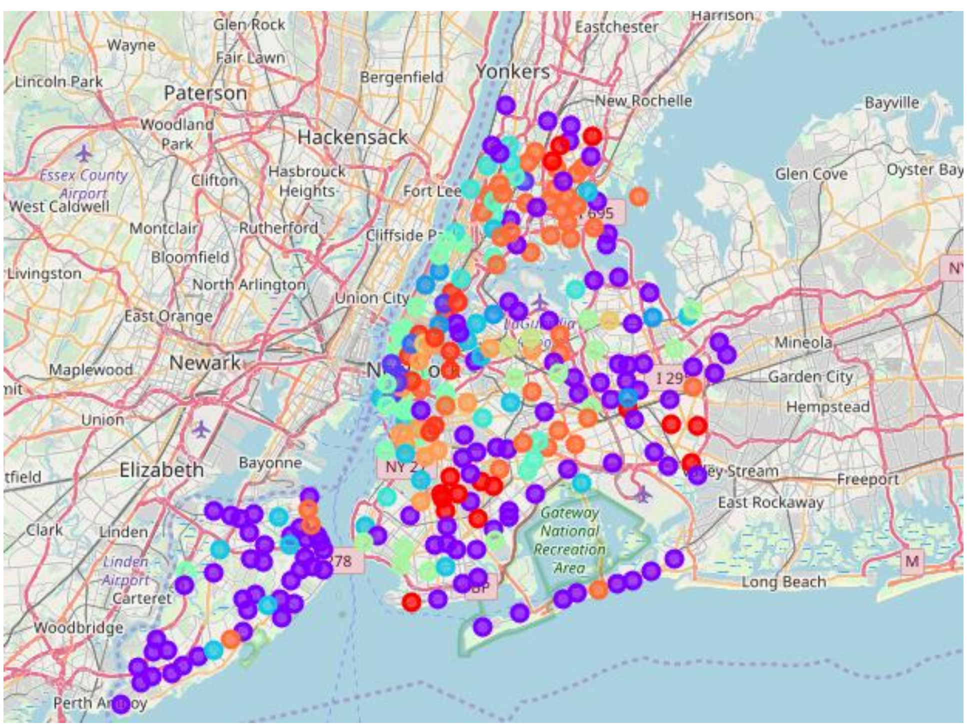
# 5. Clustering

**Now to visualize the diversity of food culture in NYC** I will cluster the neighborhoods based on available restaurants.

Using Scikit learn's K means clustering algorithm I clustered the neighborhoods based on the available restaurant types. Here I have created 15 different clusters. Below table shows which cluster is having how many neighborhoods.

| Cluster Label | Number of Neighborhood |
| --- | --- |
| 1 | 115 |
| 13 | 40 |
| 5 | 19 |
| 8 | 17 |
| 0 | 16 |
| 9 | 13 |
| 12 | 12 |
| 14 | 11 |
| 6 | 9 |
| 7 | 7 |
| 4 | 6 |
| 2 | 6 |
| 11 | 2 |
| 10 | 2 |
| 3 | 1 |

Now To visualize the clusters better I have plotted different clusters on NYC map with different colors.



## 6. Conclusion

From the above analysis I can see the similar neighborhoods based on available restaurant types and their distribution.

Most of the places in Staten Island are falling under cluster one. Most of the neighborhoods in Manhattan are in cluster 8.

In this analysis I have listed top restaurant categories for each neighborhood. Based on this list someone can identify which neighborhood will be appropriate to find a particular type of restaurant.

The purpose of this project was to identify venues in the city of New York where a particular type of restaurant is most available and which one is least available. Based on this data people interested in different cuisines can find their place of interest.

Stakeholder how are trying to find a place to open a restaurant can find a place where a particular type of restaurant is least available.

## 7. Future directions

This project was able to find the locations based on our searching conditions. Data was refined and formed tables to know the top and bottom restaurant categories. But this is not the optimal evaluation. Foursquare API could be used more extensively to filter and investigate the data more. In future different methods of analysis and different algorithm to cluster the data can be used to get more accurate results.