Coursera - Applied Data Science Capstone Project

# The Battle of Neighborhoods: Cuisines of New York City

Mayuresh Bakshi

Jul 2, 2020

# Table of Contents

# 1. Introduction & Business Problem

## 1.1 Introduction

New York City is the most populous city in the United States with an estimated population of around 8.4 Million. It is also the most densely populated city in the United States. With over 3.2 Million residents born outside the US, New York City is one of the most ethnically diverse cities. As we all know, with ethnic diversity comes a diversity of cuisines. New York City is home to more than 27,000 restaurants with Queens alone serving food from around 85 Countries.

## 1.2 Problem

Owing to cast diversity of food venues in New York City, we feel a need to explore the similarities between various neighborhoods in terms of Food Venues/Restaurant Types to determine which neighborhoods serve similar types of cuisines. As a part of the Coursera Applied Data Science Capstone Project by IBM, we are going to explore the various restaurants in all the neighborhoods in New York City and try to cluster them based on their types and location.

## 1.3 Target Audience

Curious foodies who want to explore various neighborhoods in New York City based on restaurants present in those neighborhoods.

# 2. Data

## 2.1 Data Requirement

Let's get a brief overview of the structure of New York City. New York City has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude of each neighborhood. Alongside this data, we are going to use the Foursquare API to get venue details for each neighborhood in the above dataset classified under Food Category.

## 2.2 Data Source

The dataset mentioned below contains each Neighborhood stored as a Key: Feature in JSON format.

Location Data:

- NYC Boroughs/Neighborhood Geospatial Dataset (Raw):

  *https://geo.nyu.edu/catalog/nyu_2451_34572*

- NYC Boroughs/Neighborhood Geospatial Dataset (Cleaned):

  *https://cocl.us/new_york_dataset*

Venue Data:

- Venue details: *https://developer.foursquare.com/*
- Food Category Id: 4d4b7105d754a06374d81259
- Geospatial Coordinates: Geopy Library: *https://geopy.readthedocs.io/en/stable/*

## 2.3 Data Cleaning

Data from NYC Boroughs/Neighborhood Geospatial Dataset was downloaded to a JSON file and then stored in a pandas DataFrame. Following is the raw JSON data retrieved from the file.

Note that *type: Feature* contains the information we need like Neighborhood name, lat-long, borough.


```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```


This data was then extracted and converted to a DataFrame as shown below:

| 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 |

The DataFrame contains information for all 306 neighborhoods. This data will be used later on to fetch nearby venues using Foursquare API.

# 3. Methodology

## 3.1 Food Venue Data

Foursquare is a technology company that built a massive dataset of location data. We used the Foursquare API to fetch venues in the proximity of each neighborhood center. For this, we created a Foursquare API request URL –

```
url = 'https://api.foursquare.com/v2/venues/explore?&client_id=
{}&client_secret={}&v={}&ll={},{}&radius={}&categoryId={}&limit=
{}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            categoryId,
            LIMIT)
```

This URL fetches venue data for the specified lat, lng and category id as shown below.

```
'venue': {'id': '4b4429abf964a52037f225e3',
 'name': "Arturo's",
 'location': {'address': '5198 Broadway',
  'crossStreet': 'at 225th St.',
  'lat': 40.87441177110231,
  'lng': -73.91027100981574,
  'labeledLatLngs': [{'label': 'display',
    'lat': 40.87441177110231,
    'lng': -73.91027100981574},
   {'label': 'entrance', 'lat': 40.874401, 'lng': -73.910339}],
  'distance': 240,
  'postalCode': '10463',
  'cc': 'US',
  'city': 'New York',
  'state': 'NY',
  'country': 'United States',
  'formattedAddress': ['5198 Broadway (at 225th St.)',
   'New York, NY 10463',
```

We limited the results to 100 venues within 500m of the neighborhood center. A function was then defined to call this API for each neighborhood and store the data in a DataFrame.

| Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| Wakefield | 40.894705 | -73.847201 | Dunkin' | 40.890459 | -73.849089 | Donut Shop |
| Wakefield | 40.894705 | -73.847201 | Subway | 40.890468 | -73.849152 | Sandwich Place |
| Wakefield | 40.894705 | -73.847201 | Pitman Deli | 40.896744 | -73.844398 | Food |
| Wakefield | 40.894705 | -73.847201 | Central Deli | 40.896728 | -73.844387 | Deli / Bodega |
| Wakefield | 40.894705 | -73.847201 | Louis Pizza | 40.898399 | -73.848810 | Pizza Place |

This DataFrame does not contains data for all the 306 neighborhoods as some of them do not have any food venue within 500m of their centers.

## 3.2 Clustering

In order to cluster the neighborhoods together, we first need to understand the methodology we are going to use. K-means is a type of partitioning clustering. It is an unsupervised algorithm. That is, it divides the data into k non-overlapping subsets or clusters without any cluster internal structure or labels. This was implemented using Scikit Learn library in Python. It was found in previous data gathering activity that there are 135 different types of venue present in total from all neighborhoods. These types act as the features based on which the clustering was performed. In order to achieve this, each feature was encoded using one hot encoding. This assigned a 0 or 1 value to each venue in each neighborhood. The neighborhoods were then grouped and a frequency of occurrence of each type was calculated as a means of standardizing the data.

| Neighborhood | Afghan Restaurant | African Restaurant | American Restaurant | Arepa Restaurant | Argentinian Restaurant | Asian Restaurant | Australian Restaurant | Austrian Restaurant | BBQ Joint | ... |
|---|---|---|---|---|---|---|---|---|---|---|
| Allerton | 0.000000 | 0.00 | 0.041667 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.00 | 0.000000 | ... |
| Annadale | 0.000000 | 0.00 | 0.200000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.00 | 0.000000 | ... |
| Arden Heights | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.00 | 0.000000 | ... |
| Arlington | 0.000000 | 0.00 | 0.500000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.00 | 0.000000 | ... |
| Arrochar | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.00 | 0.000000 | ... |
| Arverne | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.00 | 0.000000 | ... |
| Astoria | 0.000000 | 0.00 | 0.022989 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.00 | 0.011494 | ... |
| Astoria Heights | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.00 | 0.000000 | ... |
| Auburndale | 0.000000 | 0.00 | 0.181818 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.00 | 0.000000 | ... |

Next step was to cluster the data using k-means and display it on a map for better understanding the data.

The clustering was done using following code:

```
# set number of clusters
kclusters = 6

neighborhoods_grouped_clustering = neighborhoods_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(neighborhoods_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

# 4. Results

## 4.1 Result Data

The cluster label was added to the DataFrame containing neighborhood venue frequency. This DataFrame was joined with the first DataFrame containing Location and Borough Data for each neighborhood to create a DataFrame as shown below:

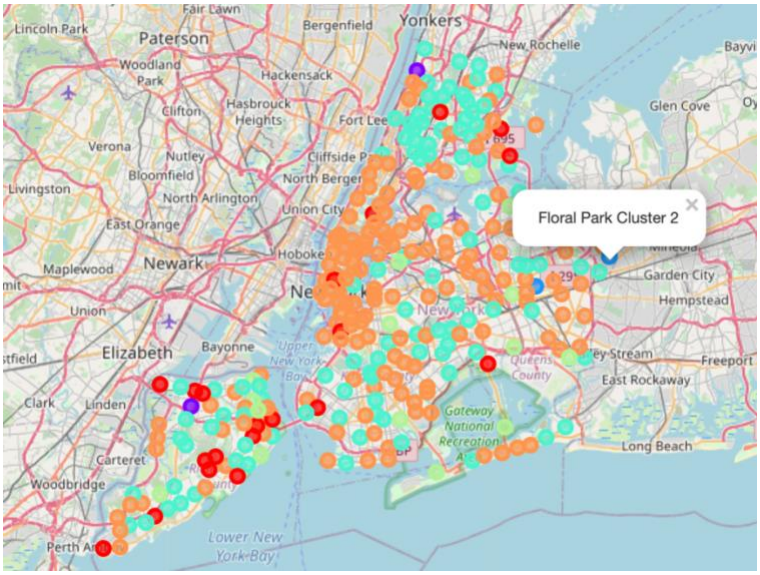
| Borough | Neighborhood | Latitude | Longitude | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|
| Bronx | Wakefield | 40.894705 | -73.847201 | 3 | Deli / Bodega | Sandwich Place | Donut Shop | Pizza Place | Food |
| Bronx | Co-op City | 40.874294 | -73.829939 | 3 | Restaurant | Fast Food Restaurant | Pizza Place | Deli / Bodega | Fried Chicken Joint |
| Bronx | Eastchester | 40.887556 | -73.827806 | 5 | Caribbean Restaurant | Diner | Deli / Bodega | Seafood Restaurant | Fast Food Restaurant |
| Bronx | Riverdale | 40.890834 | -73.912585 | 1 | Food Truck | English Restaurant | Food | Fish & Chips Shop | Filipino Restaurant |
| Bronx | Kingsbridge | 40.881687 | -73.902818 | 3 | Pizza Place | Deli / Bodega | Sandwich Place | Donut Shop | Bakery |

This DataFrame contains a Cluster Label for each Neighborhood and Top 5 most common venue types in that neighborhood.

## 4.1 Result Visualization

It is difficult to understand all the clusters from the above data view. One cannot visualize how the clusters exit in reality as they are just a number in this data. To visualize it properly, we plotted all these neighborhoods on the map using Folium library in Python.

Clustering of Neighborhoods based on Food Venue Types:




# 5. Discussion

The map gives us a better understanding of how the clusters exist in New York City. However, we still do not understand which cluster contains what types of restaurant. To check that, we displayed data of each cluster separately as shown below:

## 5.1 Cluster 1

| Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|
| Country Club | Sandwich Place | Italian Restaurant | Chinese Restaurant | Wings Joint | English Restaurant |
| Belmont | Italian Restaurant | Deli / Bodega | Pizza Place | Bakery | Spanish Restaurant |
| Edgewater Park | Italian Restaurant | Deli / Bodega | Pizza Place | Donut Shop | Japanese Restaurant |
| Carroll Gardens | Italian Restaurant | Deli / Bodega | Pizza Place | Bakery | Thai Restaurant |
| Dyker Heights | Italian Restaurant | Bagel Shop | Food | Food Truck | Hunan Restaurant |
| Upper East Side | Italian Restaurant | American Restaurant | Pizza Place | Diner | French Restaurant |
| Soho | Italian Restaurant | Mediterranean Restaurant | Café | French Restaurant | Sandwich Place |
| Howard Beach | Italian Restaurant | Bagel Shop | Deli / Bodega | Sandwich Place | Chinese Restaurant |
| Mariner's Harbor | Deli / Bodega | Italian Restaurant | Pizza Place | Donut Shop | Dosa Place |
| Tottenville | Italian Restaurant | Deli / Bodega | Sandwich Place | Mexican Restaurant | Wings Joint |
| Old Town | Italian Restaurant | Greek Restaurant | Deli / Bodega | Pizza Place | Middle Eastern Restaurant |
| New Dorp Beach | Italian Restaurant | Deli / Bodega | Food | Diner | Restaurant |

We can clearly conclude that Cluster 1 consists of Italian Restaurants in majority.

## 5.2 Cluster 2

| Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|
| Riverdale | Food Truck | English Restaurant | Food | Fish & Chips Shop | Filipino Restaurant |
| Graniteville | Food Truck | English Restaurant | Food | Fish & Chips Shop | Filipino Restaurant |

Cluster 2 is mostly Food Trucks and English Restaurants.

## 5.3 Cluster 3

| Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|
| Floral Park | Indian Restaurant | Dosa Place | Chinese Restaurant | Pizza Place | Wings Joint |
| Jamaica Estates | Indian Restaurant | Wings Joint | Empanada Restaurant | Fish & Chips Shop | Filipino Restaurant |

Cluster 3 hosts Indian Restaurants in majority.

## 5.4 Cluster 4

| Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|
| Wakefield | Deli / Bodega | Sandwich Place | Donut Shop | Pizza Place | Food |
| Co-op City | Restaurant | Fast Food Restaurant | Pizza Place | Deli / Bodega | Fried Chicken Joint |
| Kingsbridge | Pizza Place | Deli / Bodega | Sandwich Place | Donut Shop | Bakery |
| Woodlawn | Deli / Bodega | Pizza Place | Bakery | Food Truck | Donut Shop |
| Norwood | Pizza Place | Deli / Bodega | Chinese Restaurant | American Restaurant | Mexican Restaurant |
| Pelham Parkway | Chinese Restaurant | Deli / Bodega | Italian Restaurant | Pizza Place | Donut Shop |
| Bedford Park | Pizza Place | Deli / Bodega | Chinese Restaurant | Fried Chicken Joint | Diner |

Cluster 4 is a mix of Deli/Bodegas and Pizza Places.

## 5.5 Cluster 5

| Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|
| Clason Point | Deli / Bodega | South American Restaurant | Wings Joint | English Restaurant | Fish & Chips Shop |
| Bedford Stuyvesant | Deli / Bodega | Fried Chicken Joint | Café | Pizza Place | BBQ Joint |
| Marine Park | Deli / Bodega | Chinese Restaurant | Pizza Place | Dosa Place | Dumpling Restaurant |
| South Ozone Park | Deli / Bodega | Sandwich Place | Donut Shop | Food Truck | Fast Food Restaurant |
| Whitestone | Deli / Bodega | English Restaurant | Food | Fish & Chips Shop | Filipino Restaurant |
| Briarwood | Deli / Bodega | Indian Restaurant | Sushi Restaurant | Diner | Donut Shop |
| Broad Channel | Deli / Bodega | Sandwich Place | Pizza Place | Empanada Restaurant | Fish & Chips Shop |
| Brookville | Deli / Bodega | English Restaurant | Food | Fish & Chips Shop | Filipino Restaurant |
| New Brighton | Deli / Bodega | Food | English Restaurant | Fish & Chips Shop | Filipino Restaurant |
| Grymes Hill | Deli / Bodega | American Restaurant | English Restaurant | Food | Fish & Chips Shop |
| South Beach | Deli / Bodega | English Restaurant | Food | Fish & Chips Shop | Filipino Restaurant |

Cluster 5 is home to mostly Deli/Bodegas.

### 5.6 Cluster 6

| Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|
| Eastchester | Caribbean Restaurant | Diner | Deli / Bodega | Seafood Restaurant | Fast Food Restaurant |
| Marble Hill | American Restaurant | Sandwich Place | Deli / Bodega | Diner | Donut Shop |
| Williamsbridge | Caribbean Restaurant | Deli / Bodega | Restaurant | Soup Place | Filipino Restaurant |
| Baychester | Donut Shop | American Restaurant | Sandwich Place | Mexican Restaurant | Pizza Place |
| City Island | Deli / Bodega | Seafood Restaurant | Spanish Restaurant | French Restaurant | Diner |
| West Farms | Chinese Restaurant | Donut Shop | Sandwich Place | Diner | Fast Food Restaurant |
| Mott Haven | Donut Shop | Pizza Place | Food | Spanish Restaurant | Bakery |
| Port Morris | Latin American Restaurant | Food | Chinese Restaurant | Food Truck | Spanish Restaurant |
| Longwood | Deli / Bodega | Chinese Restaurant | Fast Food Restaurant | Mexican Restaurant | Café |

Cluster 6 displays the true diversity in terms of Food Venue Types.

This gives us much better insights in terms of clusters. We are able to understand how the features impacted the neighborhood partitioning in certain clusters.

## 6. Conclusion

As discussed in the previous section, we discovered the preferences for clusters based on Cuisine/Food Venue types as-

Cluster 1 — Satisfy your Italian cravings in the Red neighborhoods

Cluster 2 —Grab a bite on the go from Food Trucks in Purple neighborhoods

Cluster 3— Craving some spicy curry? Hop on to Blue neighborhoods for some Indian Food

Cluster 4— Grab New York Style Pizzas and fresh, hot food from Delis in the Turquois Zones

Cluster 5— Fresh meat products get the first priority in Light Green neighborhoods

Cluster 6 — *I don't know what I want to eat Cluster* (Orange neighborhoods)

In this Applied Data Science Capstone Project, we applied the various data manipulation techniques learned in *Data Analytics with Python* using the *Pandas* library. We also used the k-means clustering taught in *Machine Learning with Python* to cluster various neighborhoods based on various types of Food Venues.

## 7. Future Scope

These clusters can be further refined by trying out different values of k for clustering or by changing the features used. The backend code can be added using a framework to provide users with the facility to enter what they want to eat and then displaying those results similar to food near me feature on Maps.

## 8. References


- NYC Information: https://en.wikipedia.org/wiki/New_York_City
- Pandas Library: https://pandas.pydata.org/
- Foursquare API: https://developer.foursquare.com/docs/places-api/
- Geocode Library: https://geopy.readthedocs.io/en/stable/
- Scikit-Learn Library: https://scikit-learn.org/stable/
- Folium Library: https://python-visualization.github.io/folium/modules.html