

# Capstone Project - The Battle of Neighborhoods (Week 2)

## 1. Business problem and Audience:

Food Locations are getting popular among the food lovers and travelers. Prior information about these locations helps a lot in setting up the budget for having multi-cultural restaurants experiences at one place.

It's unnecessary hassle for traveler to keep moving continuously if restaurants are not plenty and not distributed within shorter distance. As per my past traveling experiences, more time is spent on reaching to places rather than staying and enjoying the food.

**Problem Definition:** *Compare the boroughs (Manhattan and Queens) of the New York city and determine the “best cluster that offers a greater number of multi-cultural restaurant experience”.*

**Audience:** Multi-cultural Food Lovers

We will be clustering neighborhood of each borough by choosing top three most common restaurants categories. Then analyzing all clusters, we would like to ask following question.

- Which borough cluster has a greater variety of multi-cultural restaurant food.
- Discussion on cluster distribution of each borough (evenly or oddly spaced clusters)

## 2. Datasets:

For this project, two datasets are considered. *New York City dataset* will have borough, neighborhood, latitude and longitude information. In *Foursquare venue dataset*, we will have 'Venue', 'Venue Latitude', 'Venue Longitude', 'Venue Category' along with neighborhood, latitude and longitude information. These two datasets will be combined later using “neighborhood” values for the clustering purpose.

**New York City dataset:** New York city dataset that contain boroughs and their neighborhood along with latitude and longitude features.

[Dataset source](https://cocl.us/new_york_dataset): [https://cocl.us/new\\_york\\_dataset](https://cocl.us/new_york_dataset)

**Foursquare venue dataset:** Venue', 'Venue Latitude', 'Venue Longitude', 'Venue Category' along with neighborhood, latitude and longitude information.

[Dataset source](https://api.foursquare.com): <https://api.foursquare.com>

Foursquare venue dataset has Pizza Place, Yoga studio, Coffee Shop, Donut Shop, Restaurants and many more Venue Categories. We need to extract only restaurants venue related data values. Data fields other than “Restaurants Venue” category will be simply ignored.

## 2.1 New York City dataset:

New York City dataset has five different kinds of boroughs. But we are interested in only two boroughs (Manhattan and Queens). Two different data frames are created for Manhattan and Queens regions. Each data frame has related Borough, neighborhood, latitude and longitude information as shown below.

Manhattan Dataset				Queens Dataset				
Borough	Neighborhood	Latitude	Longitude	Borough	Neighborhood	Latitude	Longitude	
0	Manhattan	Marble Hill	40.876551	-73.910660				
1	Manhattan	Chinatown	40.715618	-73.994279	0	Astoria	40.768509	-73.915654
2	Manhattan	Washington Heights	40.851903	-73.936900	1	Woodside	40.746349	-73.901842
3	Manhattan	Inwood	40.867684	-73.921210	2	Jackson Heights	40.751981	-73.882821
4	Manhattan	Hamilton Heights	40.823604	-73.949688	3	Elmhurst	40.744049	-73.881656

Figure 1: Manhattan and Queens dataset

Let's Visualize these neighborhoods on Manhattan and Queens Map to get better demographic understanding.

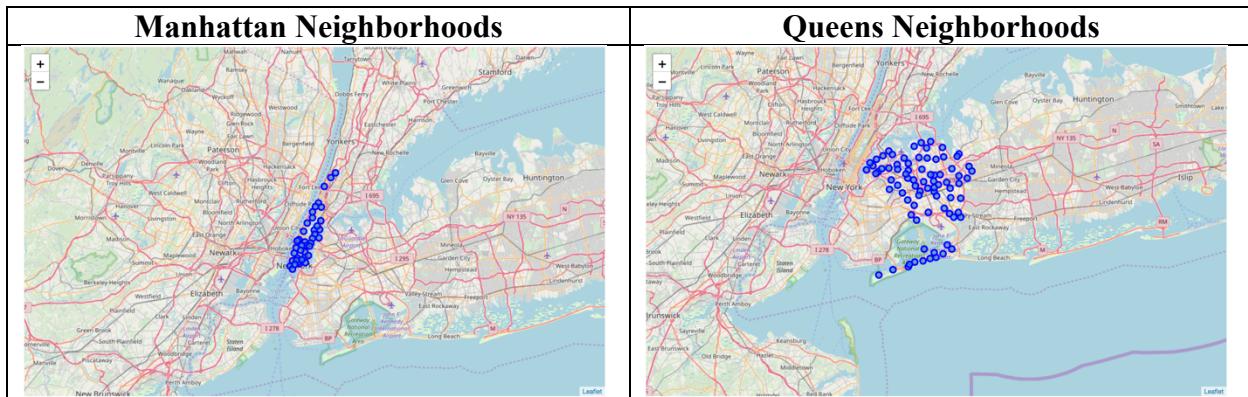


Figure 2: Visualization of Manhattan and Queens neighborhood on Map

## 2.2 Foursquare venue dataset:

Foursquare API provides the location-based information such as venues, users, photos. Relative venue information of Manhattan and Queens's neighborhoods can be easily extracted using Foursquare API platform. We separated “restaurants” **venue category** related data from Foursquare API venue dataset. Data fields other than “Restaurants Venue” category will be simply ignored or discarded.

Our interest is in only top 3 most common venue restaurants of each neighborhood. Filtered sample restaurant datasets of Manhattan and Queens are as shown below: Only first five rows of dataset are displayed, total 39 entries are available in Manhattan and 70 entries in Queens.

Manhattan neighborhoods top 3 most common restaurant venues				Queens neighborhoods top 3 most common restaurant venues			
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0 Battery Park City	Italian Restaurant	Sushi Restaurant	American Restaurant	0 Arverne	Thai Restaurant	Vietnamese Restaurant	Dumpling Restaurant
1 Carnegie Hill	French Restaurant	Japanese Restaurant	Indian Restaurant	1 Astoria	Middle Eastern Restaurant	Greek Restaurant	Seafood Restaurant
2 Central Harlem	African Restaurant	American Restaurant	Seafood Restaurant	2 Astoria Heights	Italian Restaurant	Indonesian Restaurant	Hotpot Restaurant
3 Chelsea	Italian Restaurant	American Restaurant	Seafood Restaurant	3 Auburndale	Italian Restaurant	American Restaurant	Korean Restaurant
4 Chinatown	Chinese Restaurant	Vietnamese Restaurant	American Restaurant	4 Bay Terrace	American Restaurant	Gluten-free Restaurant	Vietnamese Restaurant

Figure 3: Top 3 most common restaurants venue of Manhattan and Queens

Let's analyze top 3 common venue restaurants of each borough (Manhattan and Queens) and check their restaurant category popularity (highest count) using bar plot.

1<sup>st</sup> Most Common Restaurant Venue Category of Manhattan and Queens

2<sup>nd</sup> Most Common Restaurant Venue Category of Manhattan and Queens

3<sup>rd</sup> Most Common Restaurant Venue Category of Manhattan and Queens

### 1<sup>st</sup> Most Common Restaurant Venue Category of Manhattan and Queens:

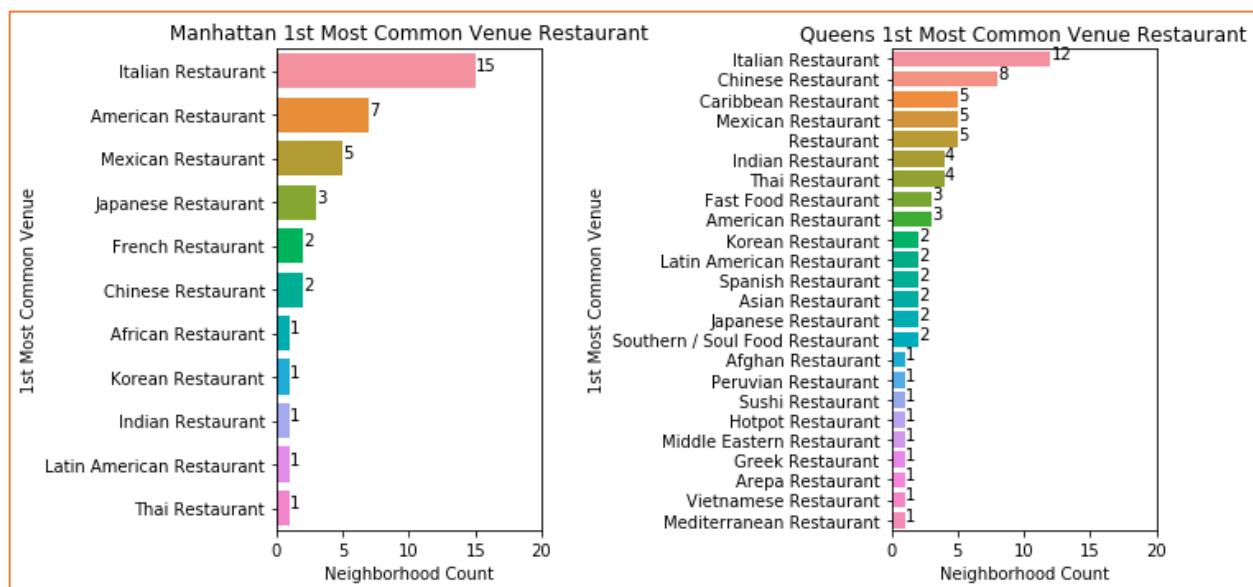


Figure 4: x-axis represents the neighborhood count and y-axis represents the category of restaurant

**Observations:** Queen has a greater number of “1<sup>st</sup> most common venue restaurant” categories compared to Manhattan. **Italian Restaurant** is a “1<sup>st</sup> most common venue restaurant” categories for Manhattan and Queens boroughs.

## 2<sup>nd</sup> Most Common Restaurant Venue Category of Manhattan and Queens:

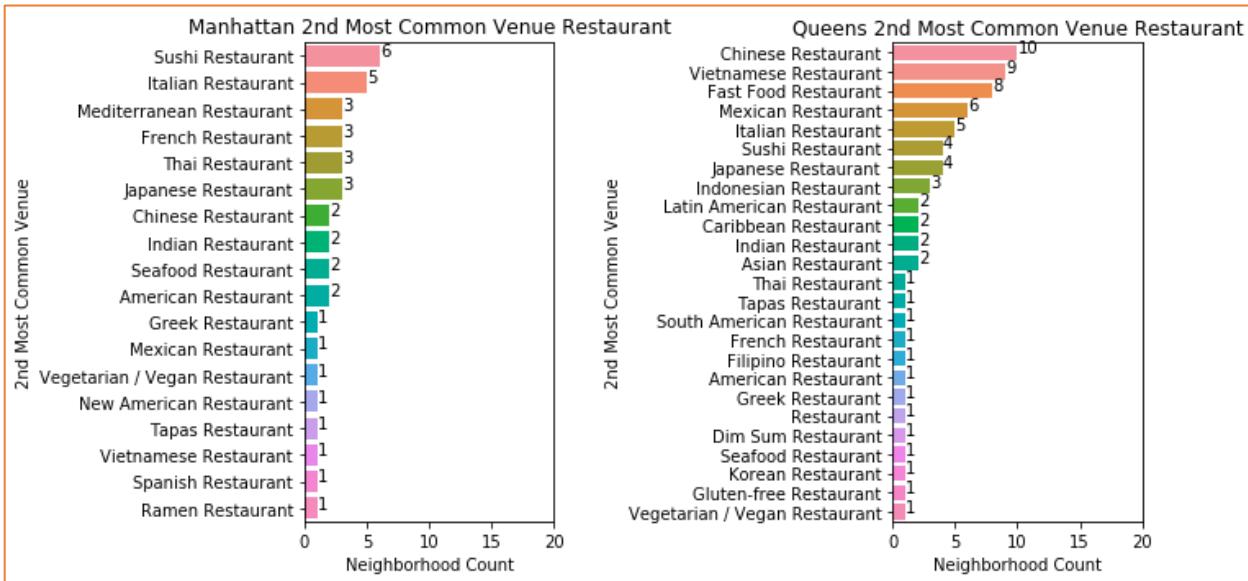


Figure 5: x-axis represents neighborhood count and y-axis represents restaurant category

**Observations:** Queen has a greater number of “2<sup>nd</sup> most common venue restaurant” categories compared to Manhattan. **Sushi Restaurant** and **Chinese Restaurant** are “2<sup>nd</sup> most common venue restaurant” categories for Manhattan and Queens respectively.

## 3<sup>rd</sup> Most Common Restaurant Venue Category of Manhattan and Queens:

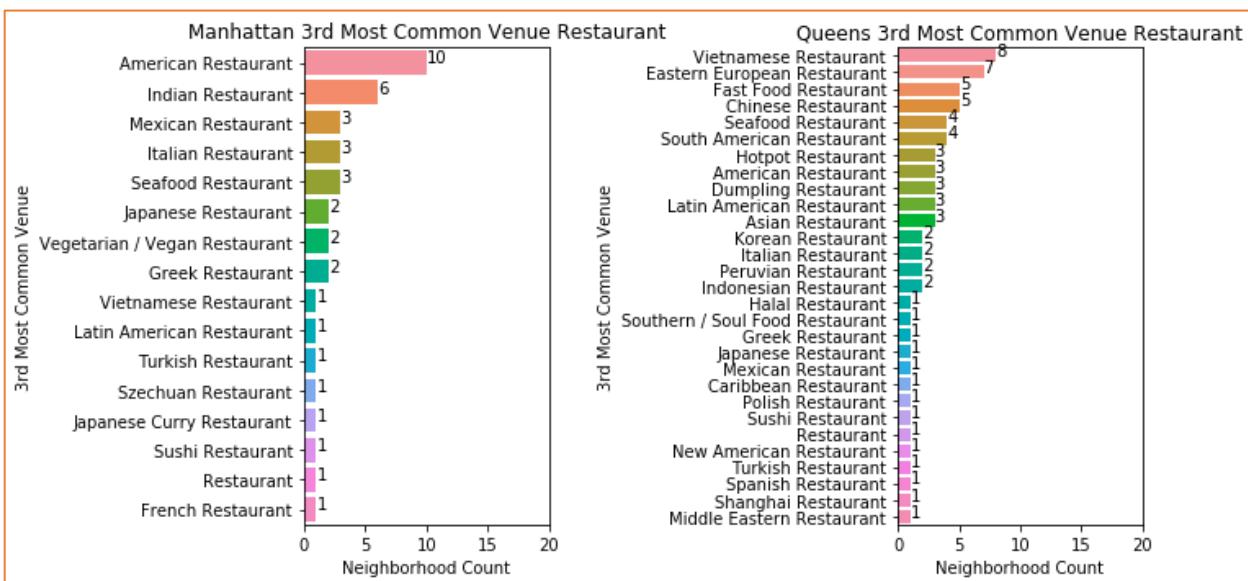


Figure 6: x-axis represents neighborhood count and y-axis represents restaurant category

**Observations:** Queen has a greater number of “3<sup>rd</sup> most common venue restaurant” categories compared to Manhattan. **American Restaurant & Vietnamese Restaurant** are “3<sup>rd</sup> most common venue restaurant” categories for Manhattan and Queens respectively.

Overall, **Queen** has a wide variety of food restaurants categories in top 3 most restaurant venues compared to Manhattan.

### 3. K-means Clustering Algorithm:

We would like to formulate five clusters using K-mean algorithm for top 3 most common venue restaurant datasets. Assigned cluster value **K=5** of K-mean algorithm will produce five different clusters. Clustering results of Manhattan and Queens “top 3 most common venue restaurant” datasets are displayed on each map separately, and clustering will be analyzed in next session.

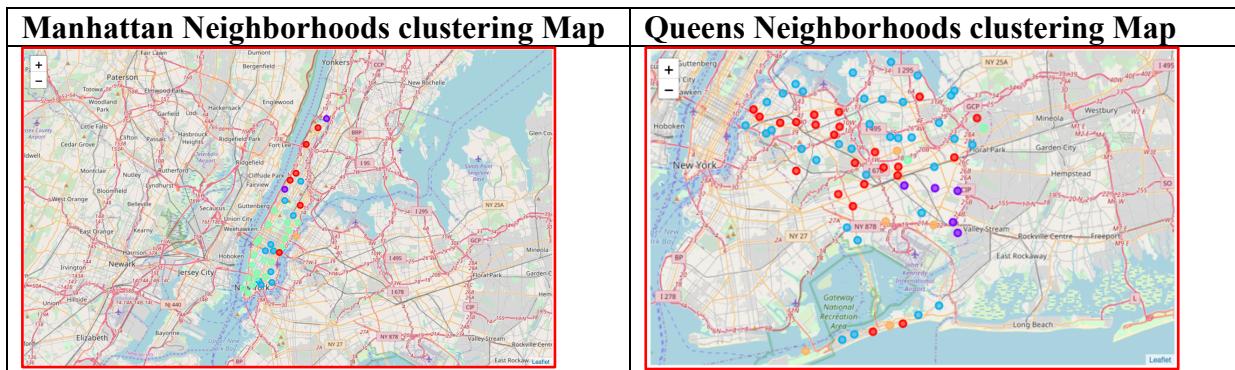


Figure 7:K-mean clustering results of Manhattan and Queens

#### Observations:

1. On Manhattan map, “light green” colored cluster contains a greater number of neighborhoods compared to other clusters.
2. On Queens map, “light blue” colored cluster contains a greater number of neighborhoods compared to other clusters.

### 4. Analyze Clusters of Manhattan and Queens:

Following **Figure 8** shows the distribution of clustered labels for the Manhattan and Queens boroughs.

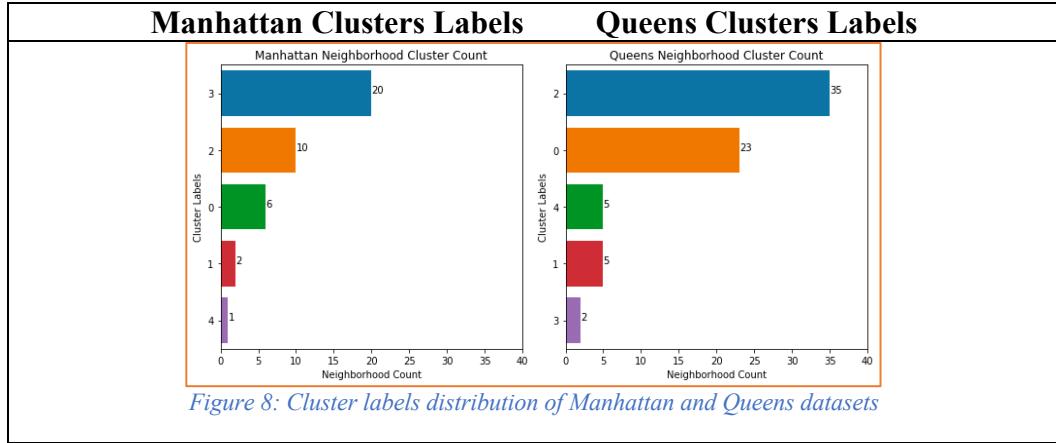
#### Observations:

1. **Cluster number 3 of Manhattan** (as shown **Figure 8**) contains maximum number of neighborhoods compared to others. Cluster number 3 has total number of 20 neighborhoods.
2. **Cluster number 2 of Queens** (as shown **Figure 8**) contains maximum number of neighborhoods compared to others. Cluster number 2 has total number of 35 neighborhoods.

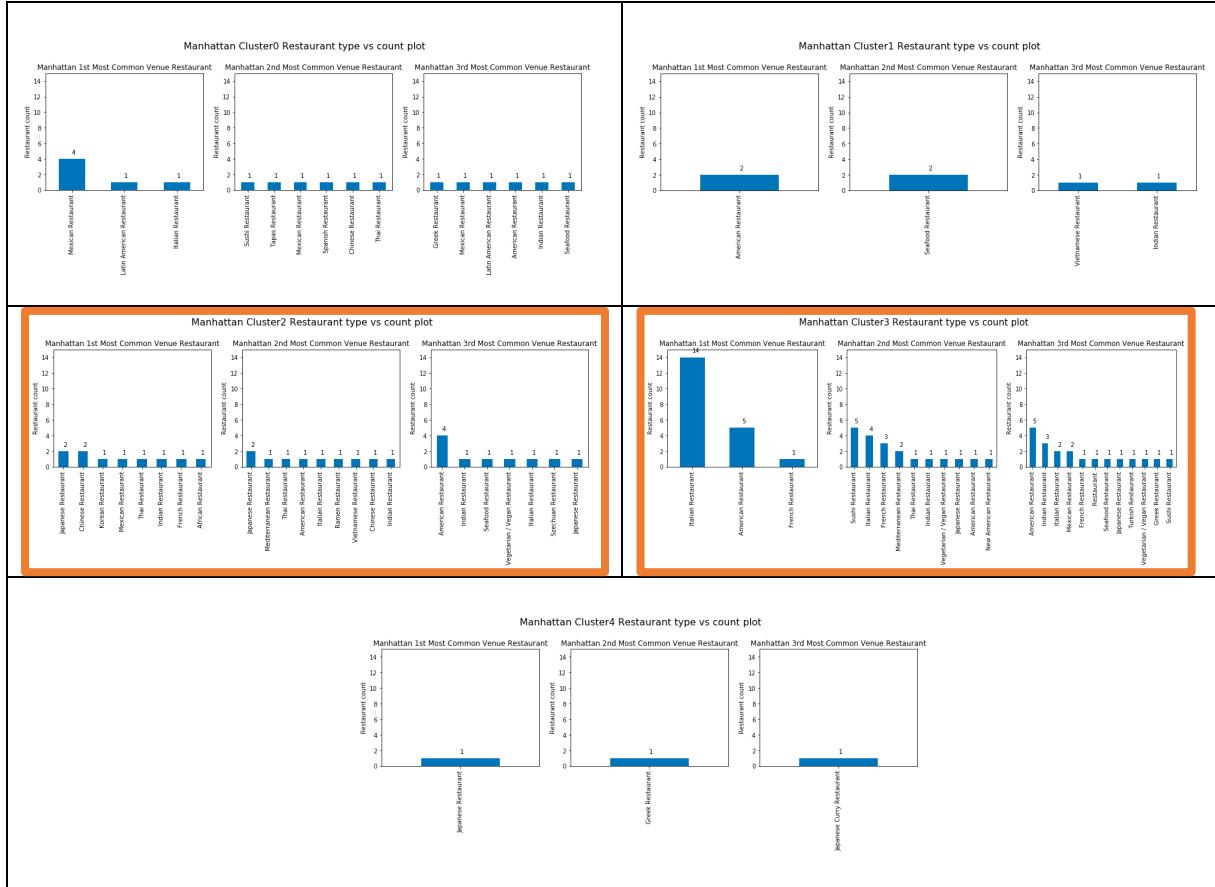
#### Guessing solution of our problem:

- **Cluster number 2** of Queens has a maximum number of neighborhoods among all clusters of both boroughs (Manhattan and Queens).

- A guess can be made that the cluster 2 may offer a greater number of multi-cultural food restaurants due to its maximum number of neighborhoods
- Again, It just a guess. Let's see the correct results.



#### 4.1. Analyze Manhattan's all five clusters along with their restaurant category:



**Observation:** Select cluster that has maximum number of restaurant categories (on x-axis of Figure 9) for top 3 most common venues: To make decision simple, we will choose these two highlighted clusters with orange colored box among Manhattan's five clusters. Manhattan cluster number 2 and cluster number 3 selected for further comparison because they have maximum number of restaurant categories on their x-axis.

## 4.2. Analyze Queens all five clusters along with their restaurant category:

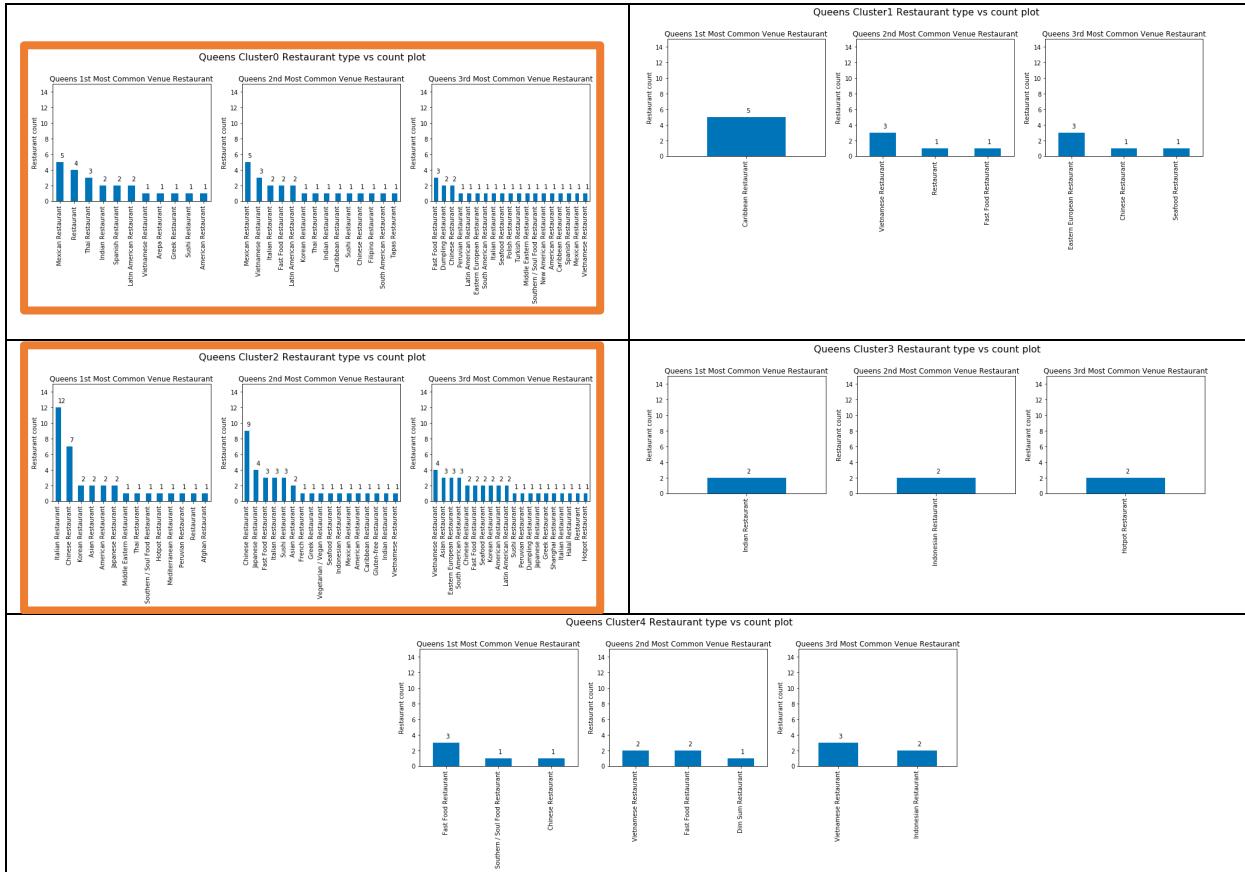


Figure 10: Queens's all five clusters with their restaurant venue category

**Observation:** Select the cluster that has maximum number of restaurant categories (on x-axis of Figure 10) for top 3 most common venues: To make decision simple, we will choose again these two highlighted clusters with orange colored box among Queens five clusters. Queens cluster number 2 and cluster number 0 selected because they have maximum number of restaurant categories on their x-axis. In total, we have selected **four clusters**, two from Manhattan and two from Queens for further comparison as shown in Figure 11.

### 4.3. Cluster Comparison: Find the best cluster offering a maximum number of multi-cultural food restaurants.

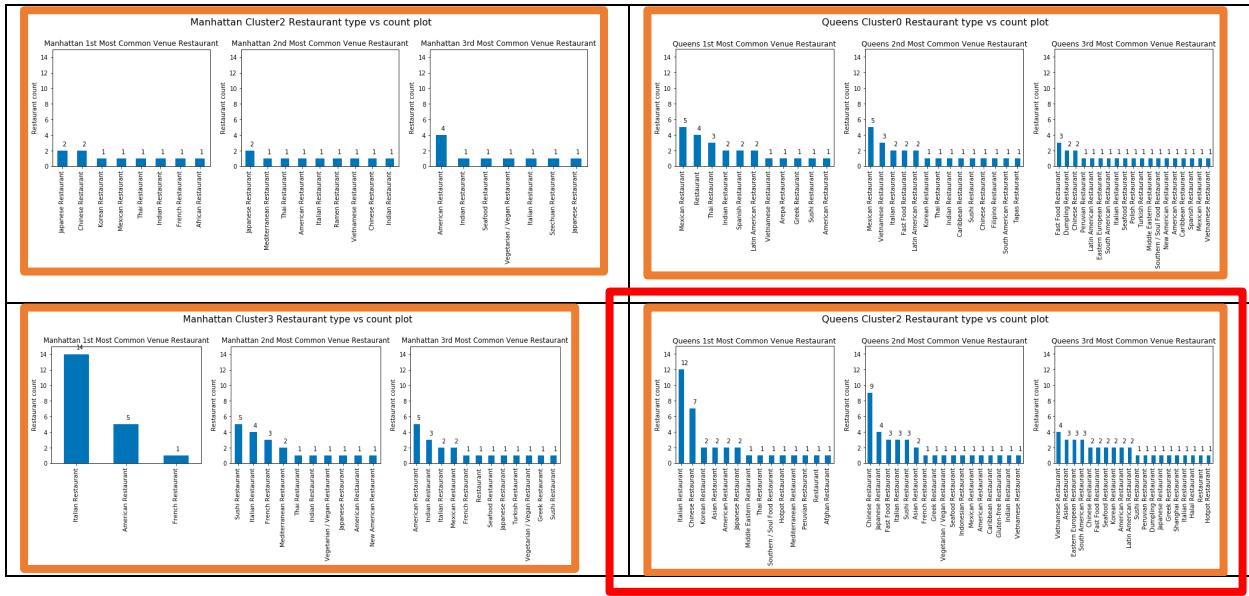


Figure 11: Cluster Comparison

From previously selected four clusters (as shown in [Figure 11](#)), it can be concluded that “**Queens cluster2**” (shown with red colored box in [Figure 11](#)) offers wide variety restaurant categories. This clustering information will help multicultural food lovers in making decision about the food locations and having all kinds of food at one place or within shorter distances.

## 5. Conclusion:

- “**Queens cluster2**” offers wide variety of restaurant categories
- **Odd distribution** of clusters observed for Queens and Manhattan boroughs.
- **Difference** between Queens and Manhattan **total number of neighborhoods** is 31.
- May produce **biased clustering results** with unbalanced neighborhoods **comparison**.