Recommendation System for Grocery Vendors

Sohan Tirpude

May 23rd, 2020

1. Introduction:

New York City (NYC), often called the simply New York (NY), is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States. Located at the southern tip of the U.S. state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass. With almost 20 million people in its metropolitan statistical area and approximately 23 million in its combined statistical area, it is one of the world's most populous megacities. New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

Situated on one of the world's largest natural harbors, New York City is composed of five boroughs, each of which is a county of the State of New York. The city and its metropolitan area constitute the premier gateway for legal immigration to the United States. As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world. New York is home to more than 3.2 million residents born outside the United States, the largest foreign-born population of any city in the world as of 2016.

2. Problem Background:

Above introduction gives a vague idea that people from various parts of the world is living in the city. Hence the diversity in the food industry is also required. As it is highly developed city so cost of doing business is also one of the highest. Thus, any new business venture needs to be analyzed carefully. The insights derived from analysis will give good understanding of the business environment which help in strategically targeting the market. This will help in reduction of risk. And the Return on Investment will be reasonable.

3. Problem Description:

In 2019, there were 27,043 restaurants in the city, up from 24,865 in 2017. New York City's food culture includes an array of international cuisines influenced by the city's immigrant history.

Central and Eastern European immigrants, especially Jewish immigrants from those regions, brought bagels, cheesecake, hot dogs, knishes, and delicatessens (or delis) to the city. Italian immigrants brought New York-style pizza and Italian cuisine into the city, while Jewish immigrants and Irish immigrants brought pastrami and corned beef, respectively. Chinese and other Asian restaurants, sandwich joints, trattorias, diners, and coffeehouses are ubiquitous throughout the city. Some 4,000 mobile food vendors licensed by the city, many immigrant-owned, have made Middle Eastern foods such as falafel and kebabs.

So this gives an idea that to supply groceries to all these restaurants, it needs to be studied carefully so one can maximizes its services to variety of the restaurants in less efforts.

So here in this case, we are going to study following problem:

- If you are a groceries supplier, then which part of the city will be better in terms of providing services as well in earning higher profits?

4. Target Audience:

Grocery supplier companies which can provide groceries to restaurants more efficiently.

5. Data Acquisition:

For this case study, we will be using the below datasets:

- New York City has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.
  Link to the dataset is: https://geo.nyu.edu/catalog/nyu_2451_34572
- Along with the above dataset, we will utilizing New York City's geographical coordinates with the Foursquare API which will help us provide details regarding all the restaurants in the area.
  URL of the Foursquare developer portal: URL of the Foursquare developer portal: https://developer.foursquare.com/

6. Data Utilization:

Using the first dataset, we'll get all the neighborhood names of New York City. And the result would look like this:


```
{
  'type': 'FeatureCollection',
  'totalFeatures': 306,
  'features': [
    {
      '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]

      }

    }
  ]
}
```


Using the geographical co-ordinates of a neighborhood locations, we will utilize the API service provided by the Foursquare to get all the restaurants nearby the neighborhood. The API response will look like this:


```
{
  'reasons': {
    'count': 0,
    'items': [
      {
        'summary': 'This spot is popular',
        'type': 'general',
        'reasonName': 'globalInteractionReason'
      }
    ]
  },
  'venue': {
    'id': '5012c967e889cf0567c9e2d4',
    'name': 'Grill 26 at TCR',
    'location': {
      'address': '2600 Netherland Ave',
      'crossStreet': 'off Kappock Street',
      'lat': 40.878802209517126,
      'lng': -73.9156723022461,
      'labeledLatLngs': [
        {
          'label': 'display',
          'lat': 40.878802209517126,
          'lng': -73.9156723022461
        }
      ],
      'distance': 490,
      'postalCode': '10463',
      'cc': 'US',
      'city': 'Bronx',
      'state': 'NY',
      'country': 'United States',
      'formattedAddress': [
        '2600 Netherland Ave (off Kappock Street)',
        'Bronx, NY 10463',
        'United States'
      ]
    },
    'categories': [
      {
        'id': '4bf58dd8d48988d14c941735',
        'name': 'American Restaurant',
        'pluralName': 'American Restaurants',
        'shortName': 'American',
        'icon': {
          'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/default_',
          'suffix': '.png'
        },
        'primary': True
      }
    ],
    'photos': {
      'count': 0,
      'groups': []
    }
  },
  'referralId': 'e-0-5012c967e889cf0567c9e2d4-26'
}
```

7. Data Cleaning:

First we have downloaded the New York City's 5 boroughs and 306 Neighborhood dataset which contains the latitude and the longitude of each neighborhood. This data is in JSON format.

After initial exploration, we found out that the feature section of this dataset is what we want. From this feature section, we will be needing only name of the Borough, name of the neighborhood, latitude and longitude of the neighborhood. So, using only these columns we will create the dataframe.

8. Data Pre-processing:

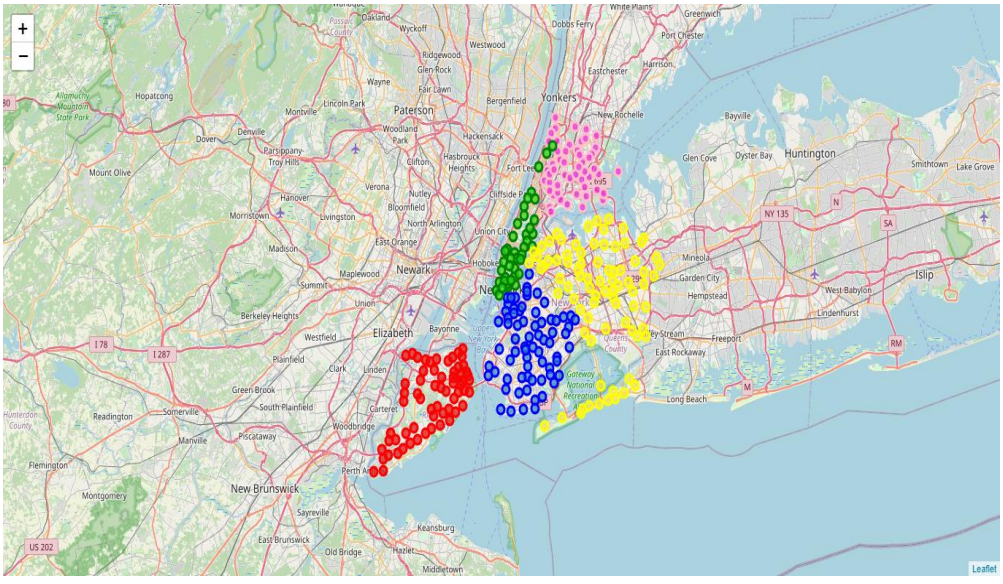
One we got latitudes and longitudes of all the neighborhoods, we'll utilize the Foursquare API to fetch the 100 restaurants in each neighborhood area within the 500 meters.

Once we get fetch data using the Foursquare API, the resultant dataset will look like this:

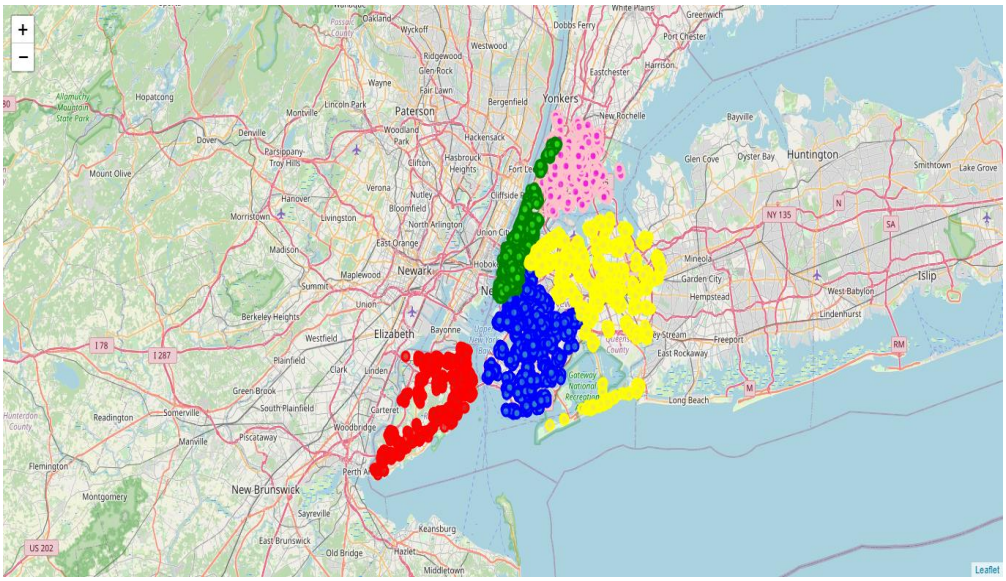
	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bronx	Wakefield	40.894705	-73.847201	Pitman Deli	40.894149	-73.845748	Food
1	Bronx	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
2	Bronx	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant
3	Bronx	Wakefield	40.894705	-73.847201	Chef Central	40.891625	-73.844531	Diner
4	Bronx	Wakefield	40.894705	-73.847201	New China Gardens	40.897796	-73.853388	Asian Restaurant

9. Data Exploration:

Now, we have acquired the data we want, we'll try to plot the map using the latitude and longitude of all the neighborhoods using the Folium library. Here we'll try to plot the map of all the neighborhoods with respect to the Borough it belongs. And the result looks like this:



And this is map of all the venues of all the neighborhoods in the New York City with respect to the Borough it belongs. And the result looks like this:



10. Feature Selection:

To accomplish our goal, we'll need total no. of venues/restaurant against each neighborhood. For this, we'll take count of the all the Venues respective to the neighborhood. And this will be our Final dataset which we will use to train our cluster model on it.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Category
0	Bronx	Allerton	40.865788	-73.859319	43
1	Bronx	Baychester	40.866858	-73.835798	21
2	Bronx	Bedford Park	40.870185	-73.885512	44
3	Bronx	Belmont	40.857277	-73.888452	49
4	Bronx	Bronxdale	40.852723	-73.861726	42

11. Clustering Algorithm

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them.

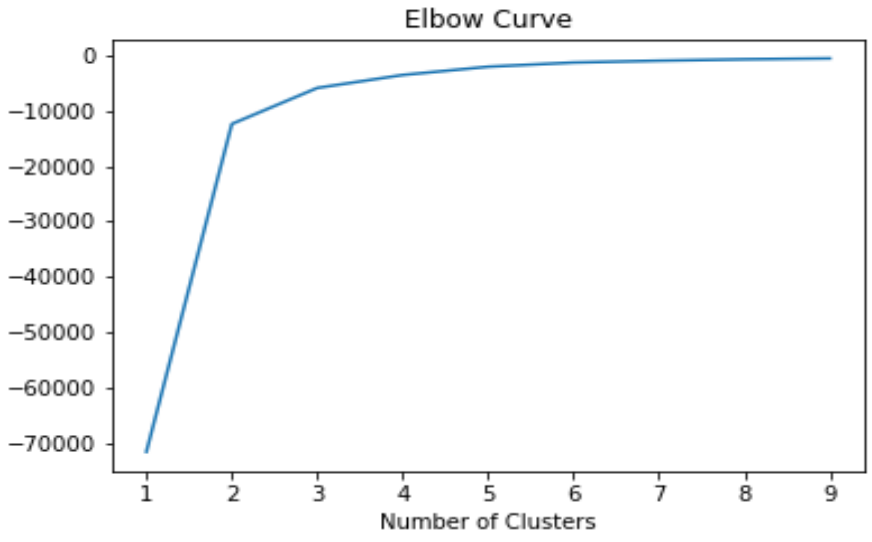
Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including parameters such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results.

Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

To accomplish our goal, we are going to utilize *k*-means clustering algorithm to cluster the neighborhoods. K-means clustering is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs.

k-means is somewhat naive — it clusters the data into **k** clusters, even if **k** is not the right number of clusters to use. When we come to clustering, it's hard to know how many clusters are optimal. Therefore, when using *k*-means clustering, we need a way to determine whether we are using the right number of clusters.

One method to validate the number of clusters is the elbow method. The idea of the elbow method is to run *k*-means clustering on the dataset for a range of values of **k** (say, k from 1 to 10), and for each value of **k** calculate the Sum of Squared Errors (SSE). When **k** increases, the centroids are closer to the clusters centroids. The improvements will decline rapidly at some point, creating the elbow shape. That is the optimal value for **k**.



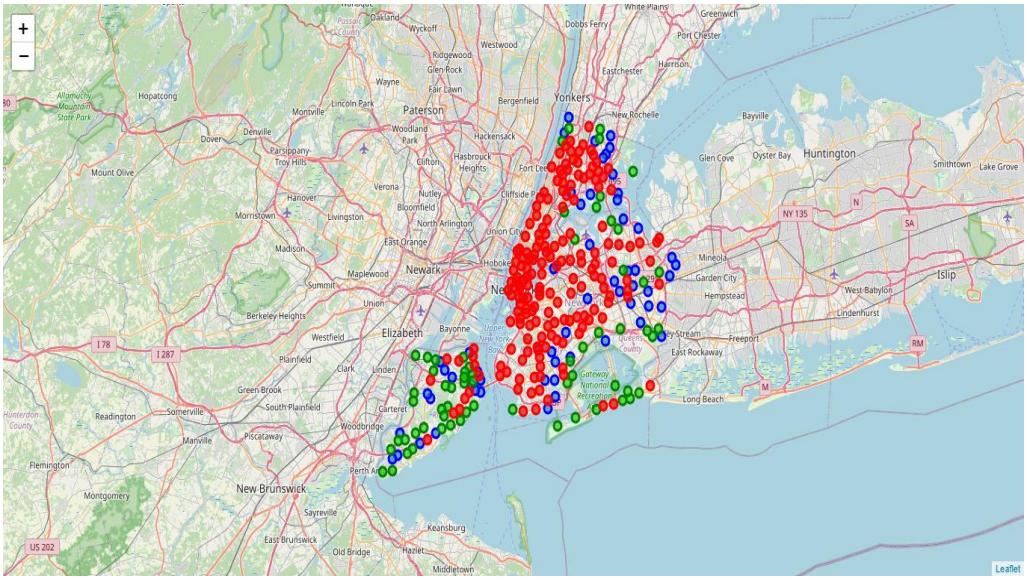
When we graph the plot, we see that the graph levels off slowly after 3 clusters. This implies that addition of more clusters will not help us that much. Hence, we'll be clustering our data into 3 partitions.

After training the *k*-means model to cluster our dataset into 3 partitions, we get the result which looks like this:

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Category	cluster_labels
0	Bronx	Allerton	40.865788	-73.859319	43	2
1	Bronx	Baychester	40.866858	-73.835798	21	0
2	Bronx	Bedford Park	40.870185	-73.885512	44	2
3	Bronx	Belmont	40.857277	-73.888452	49	2
4	Bronx	Bronxdale	40.852723	-73.861726	42	2

12. Analyzing our Clusters:

Once we get the cluster labels to each of our neighborhood dataset, we'll start analyzing our cluster by plotting them in the map.

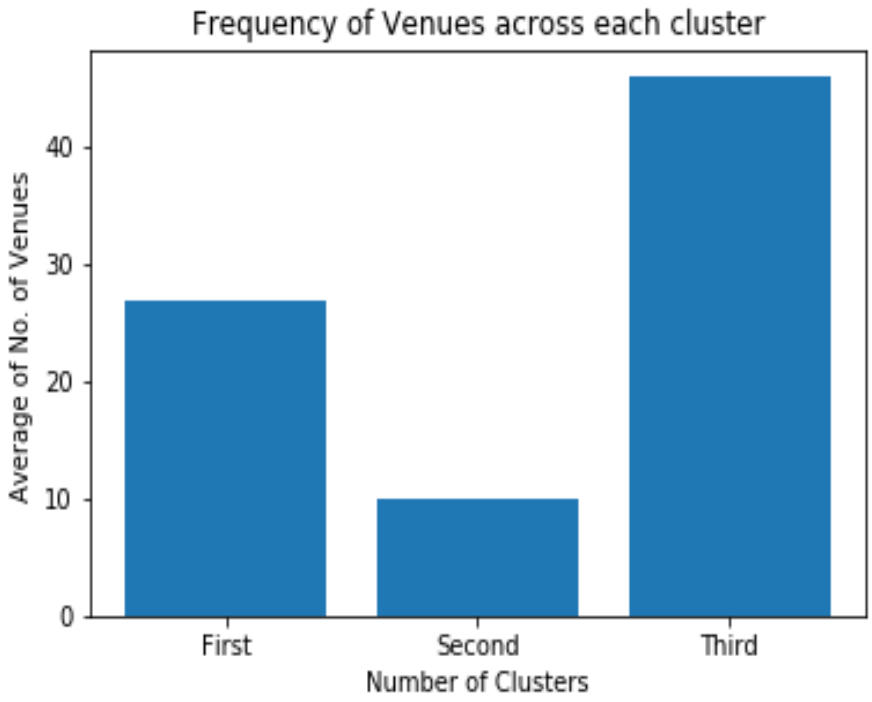


From the map we can say one thing is that the neighborhood area around Manhattan is pretty much covered in a singly cluster. But we will analyze our dataset more deeply to prove our conclusion.

Also, we will see how the data spread in each cluster is by getting average no. of venues/restaurant respective to the clusters. And this is the resultant data looks like:

	cluster_labels	Venue Category
0	0	26.877193
1	1	9.984375
2	2	46.022222

We will utilize the Matplotlib library to visualize the same analysis through graphical manner:



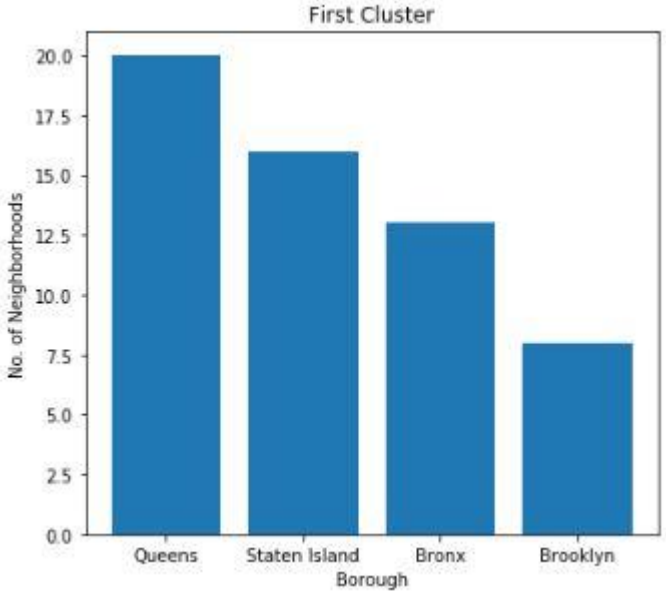
From above mean distribution graph, we can clearly see the distinguishing features about the clusters. Each cluster is itself different from the others.

At this point we can pretty much conclude that the third cluster which is the cluster no. 2 is the cluster of neighborhoods which has on average more than 40 numbers of restaurants and these neighborhoods any grocery vendors should target to build/expand the business.

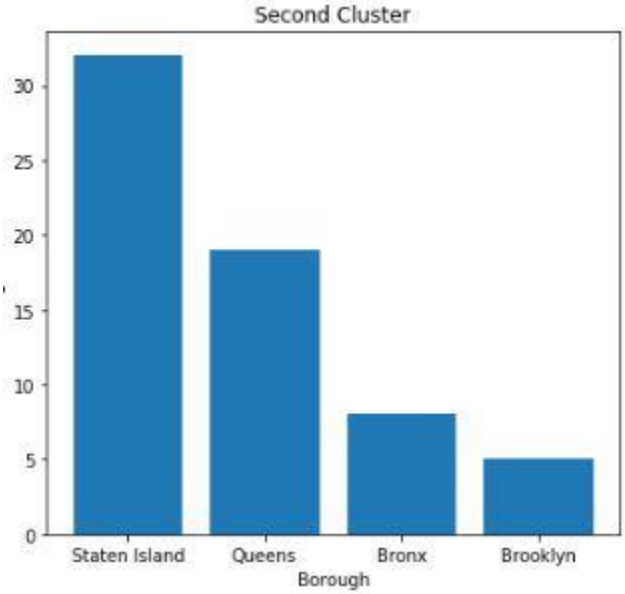
So this kind of analysis is highly recommended because it gives us the important feature about the clusters.

Now, we'll try to analyze the frequency of the venues/restaurants in each boroughs with respect to the clusters.

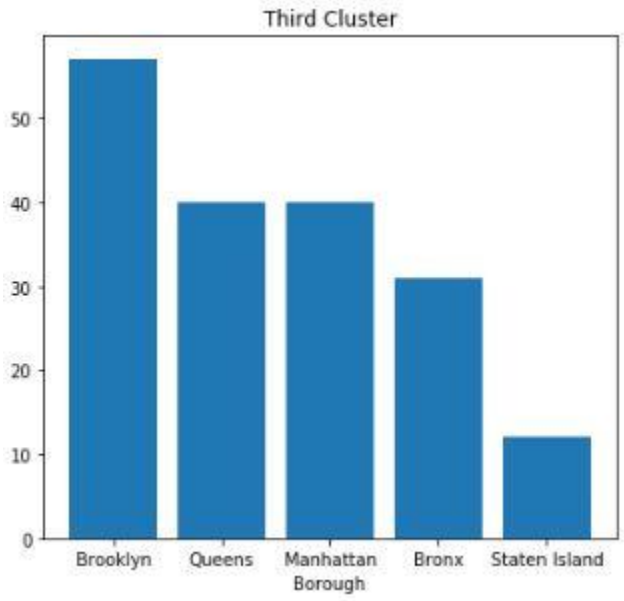
Frequency of the venues/restaurants in each Borough in First Cluster:



Frequency of the venues/restaurants in each Borough in Second Cluster:



Frequency of the venues/restaurants in each Borough in Third Cluster:



Now, if we analyze carefully, we will find out that the **Manhattan** is the only Borough which all neighborhoods has almost 40 venues/restaurants and it features only in the Third cluster which is a cluster no. 2. And this cluster no. 2 is the cluster of all neighborhoods which has on average more than 40 venues/restaurants.

13. Conclusions:

As mentioned in our Problem statement, our goal is to find the cluster of neighborhoods which has large no. of venues/restaurants which will be used to recommend any grocery vendors to start/expand their business.

Also the above study using the clustering algorithm pretty much accomplish our goal by suggesting the **Manhattan** borough of New York City has large no. of venues/restaurants. Along with this, in each borough there are many neighborhoods which has large no. of venues/restaurants.

14. Future Directions:

Using the above study, we can further go deep into what kind of restaurants each neighborhood has and this feature has any relationship with the diverse population of the New York City. Also, we can provide more analysis on If one wants to open a new restaurants of a particular type, which part of the city will be less competitive and highly profitable.

Also, we can provide analysis on what part of the city anyone can start a new restaurants and what kind of competition it will have from the other restaurants.