**Capstone Project – The Battle between two cities**

By

Shaoni Chakravarty

December 25,2019

Contents

Introduction

* Background
* Problem

Data

* New York City dataset
* Toronto dataset
* Foursquare API

Methodology

* Explore the Toronto dataset
* API calls to Foursquare
* Explore the NYC dataset
* API calls to Foursquare

Data Analysis

* Data Cleaning
* Data Visualization
* Analytical views

Machine Learning

* K-Means
* The Elbow Method for NYC dataset

Results

* For Toronto Dataset
  + 1. Cluster 0 set
    2. Cluster 1 set
    3. Cluster 2 set
    4. Cluster 3 set
    5. Cluster 4 set
* For New York Dataset
  + 1. Cluster 0 set
    2. Cluster 1 set
    3. Cluster 2 set
    4. Cluster 3 set
    5. Cluster 4 set
    6. Cluster 5 set
    7. Cluster 6 set
    8. Cluster 7 set

Discussion

Conclusion

Reference

Introduction

Background

The city we are working on are : New York City & Toronto. Both the cities are very popular city of USA. Both the cities are very diversified. It is the home of more than 8.1 million people. Many immigrants also live here from various parts of the world. More than 800 languages can be found in both the cities.

As both the cities are based on the diverse culture, so people from different traditional cuisine can be found there. From Asian Cuisine to Middle Eastern , American authentic food can be found there. Now to open a new food chain or restaurant will be a challenge from business point of view as well as people should accept the taste of the food being served.

Problem

While working on this, we have faced certain issues. The idea of this project to categorically segment the neighborhoods of NYC & Toronto into major clusters & examine the cuisine. This project will help to understand the diversity of neighborhood by leveraging venue data from Foursquare API calls & K-means Clustering , an unsupervised ML algorithm. It will be helpful for a new vendor who is willing to open a new restaurant.

Data

Toronto Dataset Details:

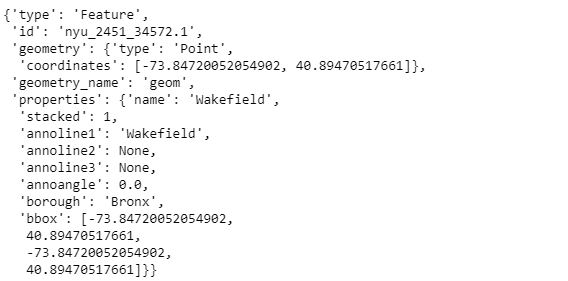
Toronto link : https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M

This link helps to get Toronto data including location geographical locations. This helps in getting idea about the neighborhoods of Toronto.

**New York Dataset Details:**

**New York link :** <https://cocl.us/new_york_dataset>

This link acts as a guide to New York Citys neighborhood that appear on the web resource. The json extract is as follows:



Foursquare API:

Link: <https://developer.foursquare.com/docs>

Foursquare API, a location data provider, will be used to make API calls to retrieve data about venues in different neighborhoods.

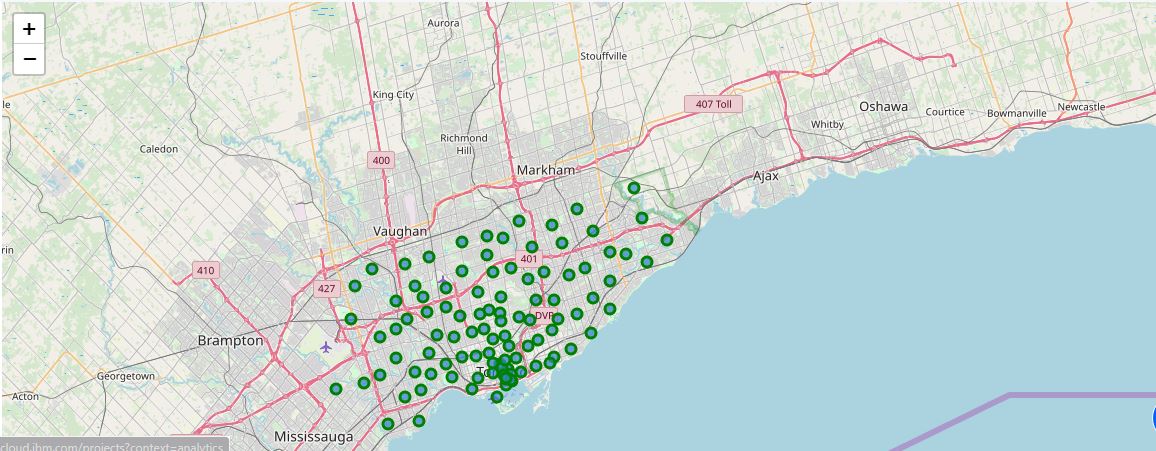
**Methodology**

Download & Explore Toronto Dataset :

In order to segment neighborhoods in Toronto, we found a dataset that contains 12 boroughs and 103 neighborhoods, that exist in each borough, with respective latitude & longitude coordinates. The dataset was downloaded from a wiki table using Beautiful Soup. We found a data with the neighborhoods. We have added the geospatial column to get the geographical co-ordiantes as well :

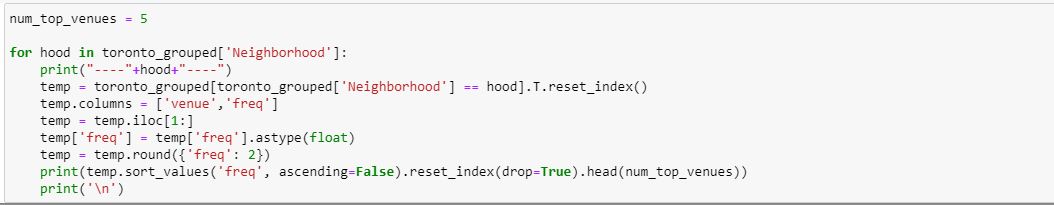


Upon analysis, we have 12 boroughs & 103 neighborhoods. Further , ‘geopy’ library was used to get the latitude & longitude values of Toronto dataset. The ‘folium’ library is used to depict the map generation.

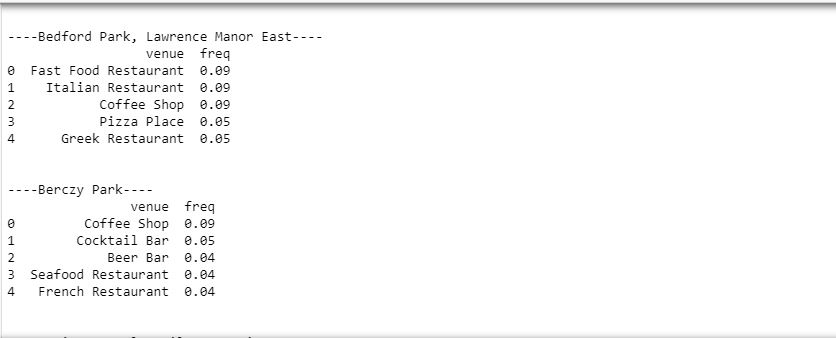


RESTful API calls to Foursquare for Toronto dataset:

The Foursqure api is used to explore neighborhoods & segment them. To access the API CLIENT\_ID, CLIENT\_SECRET , VERSION needs to be defind. There are many endpoints available on Foursquare for various GET requests. But to explore the cuisines, we need the data for the ‘Food’ category. As stated, The function takes restaurant names.

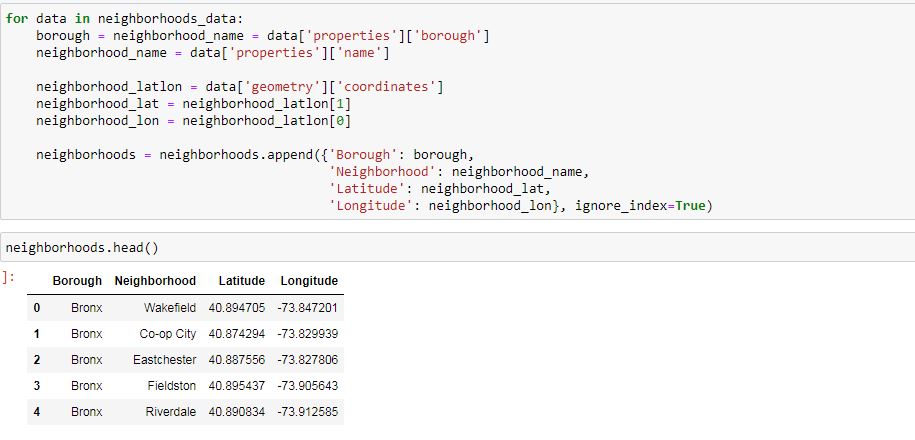


The restaurant output results in:

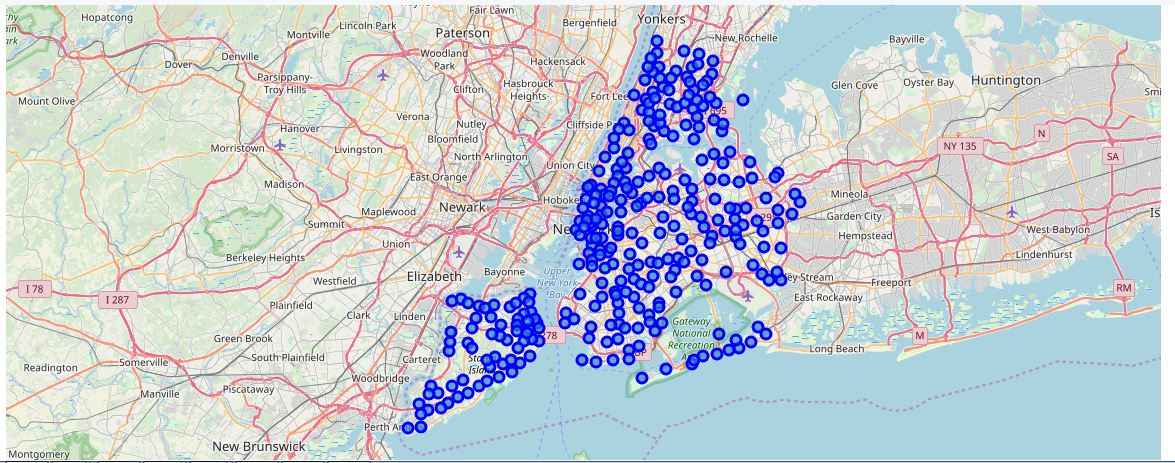


Download & Explore New York City Dataset :

In order to segment neighborhoods of the New York City, the dataset contains 5 boroughs & the neighborhoods, that exist in each borough with respective latitude & longitude co-ordinates. Once the json file is downloaded , we study the structure of the data then. Dataframe is created with Borough, Neighborhood, Latitude & Longitude.

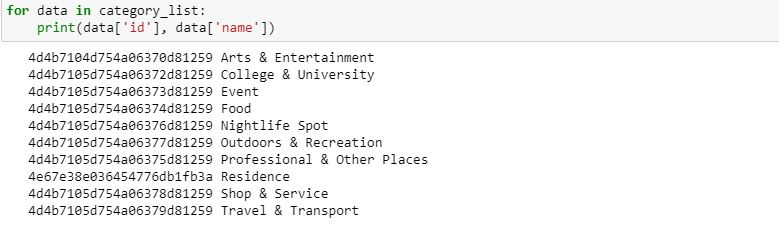


The curated dataframe is then used to visualize the map of the New York City with neighborhoods superimposed on the top. The map has been generated using python ‘folium’ library.



RESTful API calls to Foursquare for New York city dataset:

The Foursqure api is used to explore neighborhoods & segment them. To access the API CLIENT\_ID, CLIENT\_SECRET , VERSION needs to be defind. There are many endpoints available on Foursquare for various GET requests. But to explore the cuisines, we need the data for the ‘Food’ category. Upon analysis, it is found that there are 10 major categories of venues. Under which all other sub-categories are included .



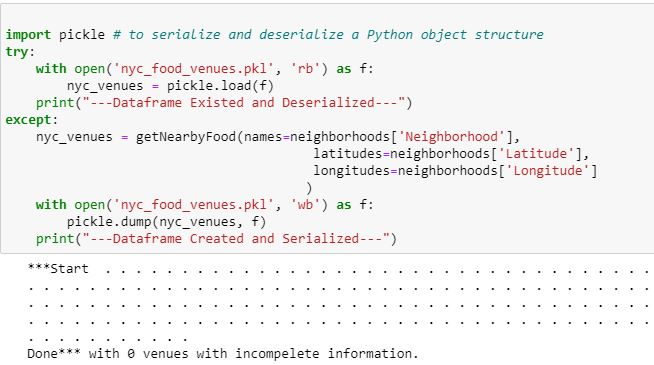
To further understand the result of GET request, the first neighborhood of ‘New York City dataset’ is explored. The first neighborhood returned is ‘Wakefield’. Then a GET request url is created to search for venue with Category\_id = ‘4d4b7105d754a06374d81259’ and radius = ‘500’ meters.

As the aim is to segment the neighborhoods of New York City with respect to ‘Food’ in the vicinity, it is further required to fetch the data from all 306 neighborhood venues. To overcome the redundancy of the process followed above, a function ‘getNearbyFood’ is created. This function loops in all the neighborhoods & creates API url with a limit = 100 & radius = 500.



Pickle :

Pickle is very important & easy-to-use library. It is used to serialize the data retrieved from GET requests, to make a ‘pkl’ file. This file can later be deserialized to retrieve an exact python object structure.



The dataframe is as follows:



As of now, two separate dataframes have created:

1.‘neighborhoods’ that contain neighborhood, Latitude,Longitude, Borough

2. ‘nyc\_venues’ is a merger between ‘neighborhood’ & ‘food’ category . Also eah venues has its own latitude,longitude & category.

Exploratory Data Analysis:

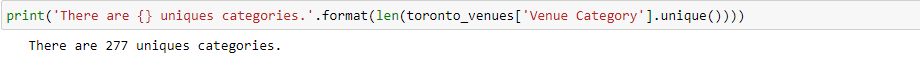
* Toronto Data Analysis:

To have a data analysis on Toronto dataset, we need to depict the dataframe

‘neighborhoods\_venues\_sorted’ . It shows the common venues in the dataframe.

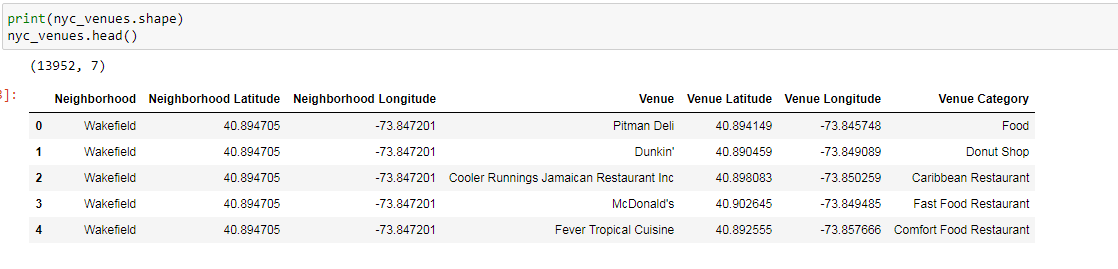


There are unique neighborhood in Toronto.

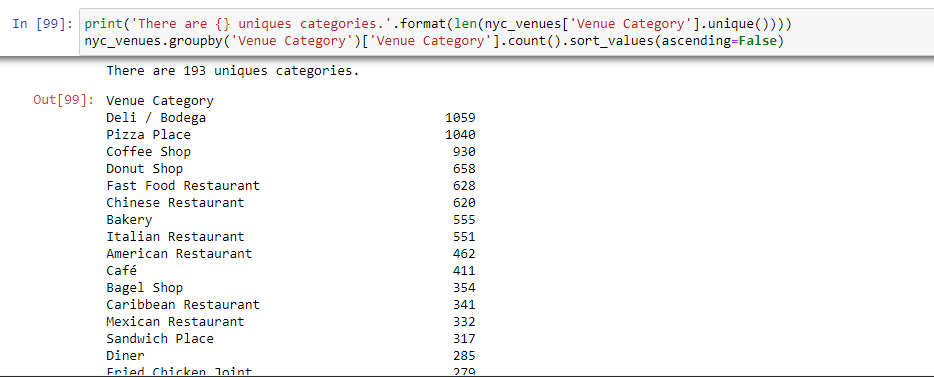


* New York City Data Analysis :

The merged dataframe ‘nyc\_venues’ has all required information. It has 13952 rows.

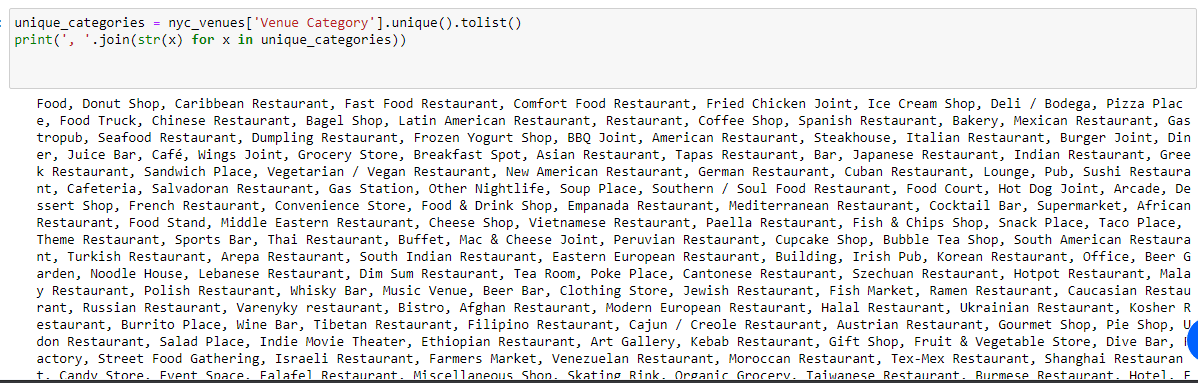


There are 193 unique categories with most occurring venues as follows:



Data Cleaning

As there is cultural diversity in the neighborhood, its important to clean the data with generalized categories. Here, by generalized categories, it means that these venues are common across different culture & different food habbits. So, first unique categories have been listed down.



This data-preperation totally depends on ‘Data Analyst’ discretion & can be modified as required.

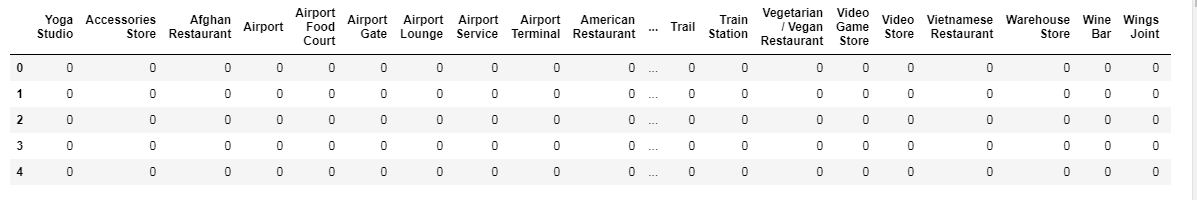


Engineering discussions:

* Toronto :

Now, each neighborhood is analyzed individually to understand the common cuisine served within 500 meters of the vicinity.





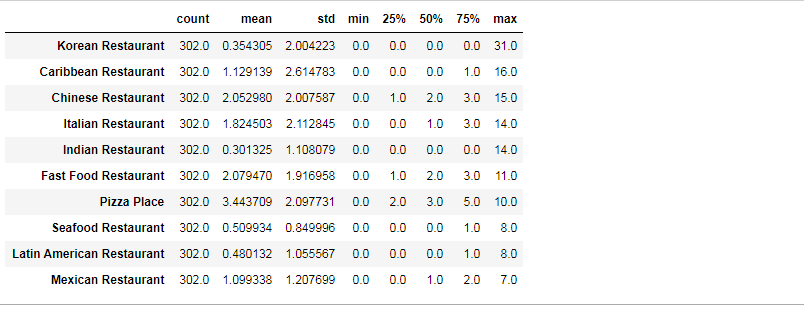
From this depiction it’s not clear which cuisine is popular for Toronto.

* New York City :

Now, each neighborhood is analyzed individually to understand the common cuisine served within 500 meters of the vicinity.

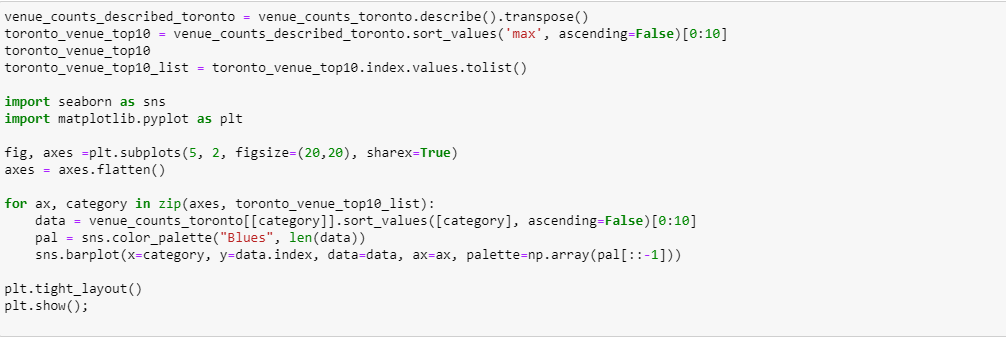


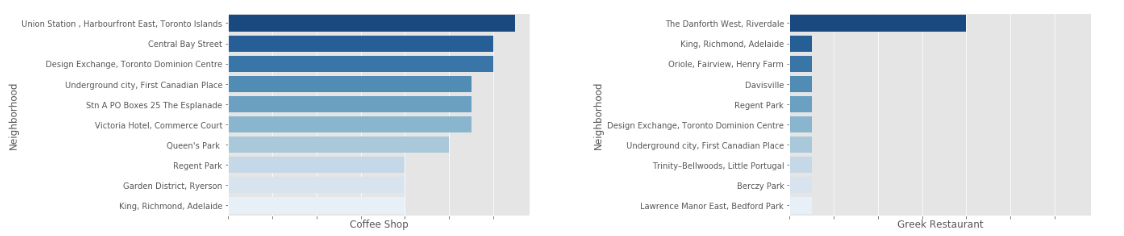
The top 10 venue-categories can also be depicted by counting its occurrences. According to findings, Korean , Carribean & Chinese Restaurants are the most found cuisine types.

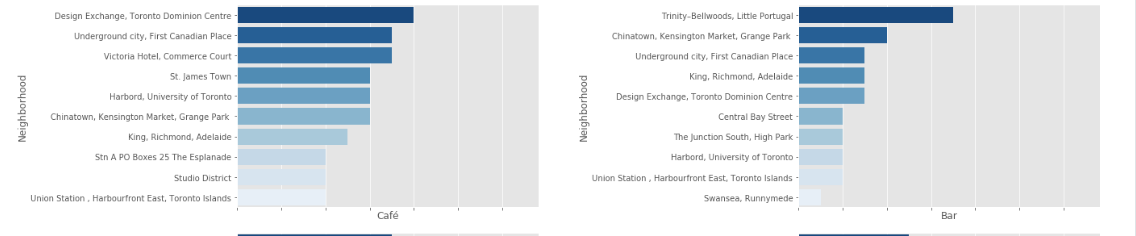


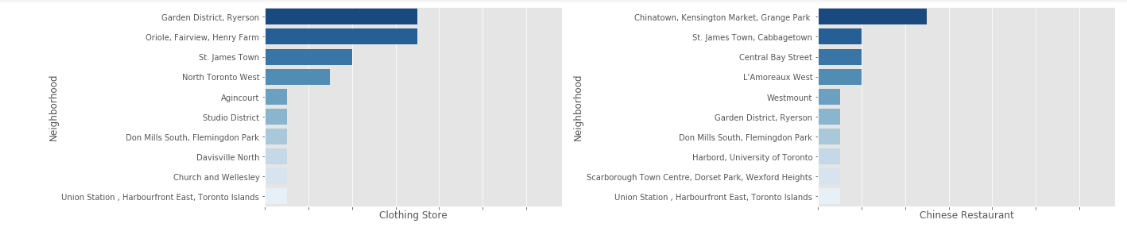
Data Visualization

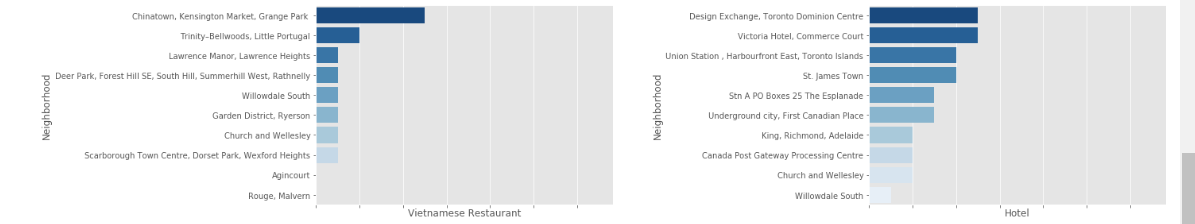
* Toronto city







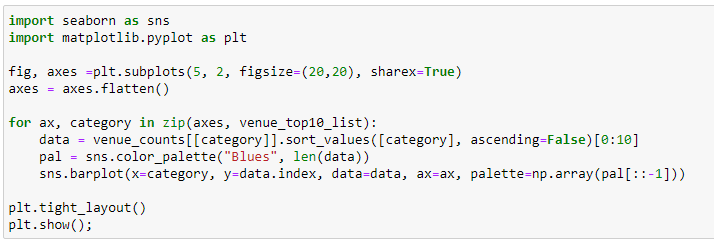




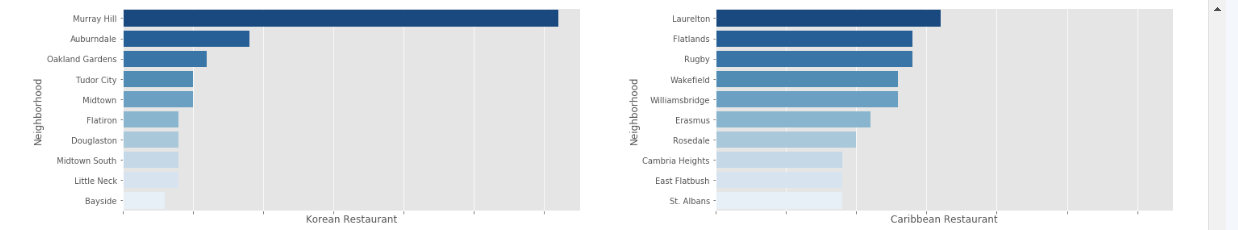


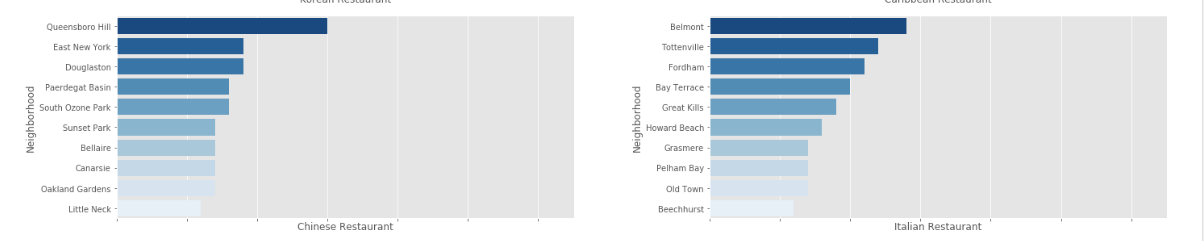
* New York City

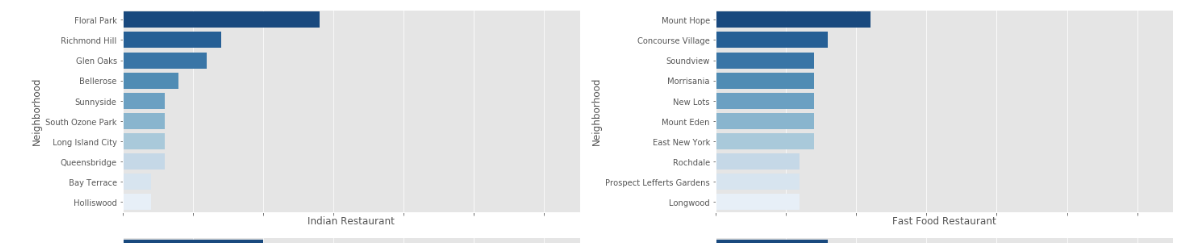
We have plotted the top 10 categories using python ‘seaborn’ library.

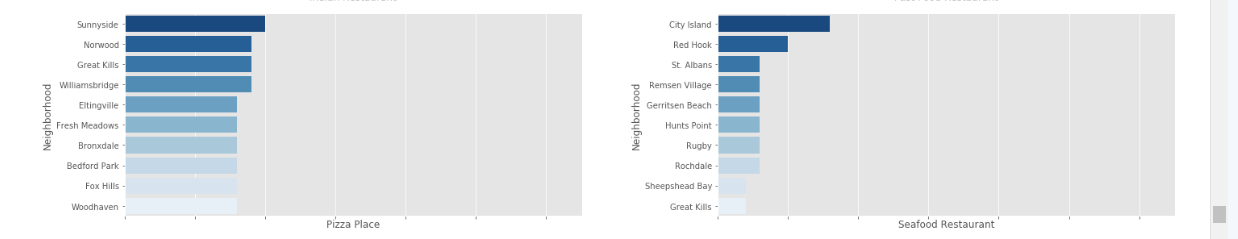


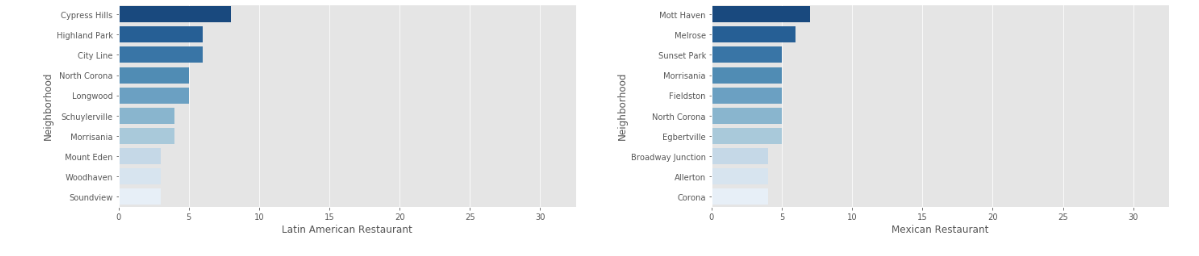
The bar graph is shown as below:











Machine Learning :

Toronto City:

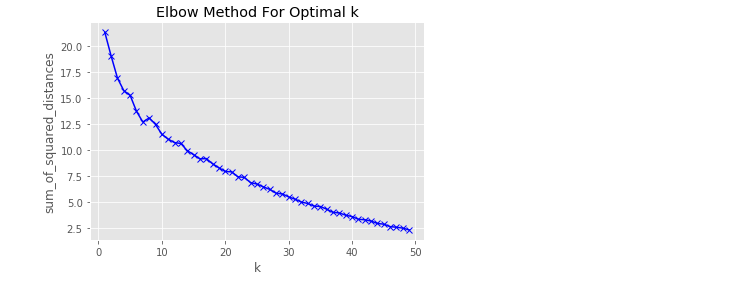


k-means is an unsupervised machine learning algorithm, which creates clusters of data points aggregated together because of certain similarities. This algorithm will be used to count neighborhoods for each cluster label for variable cluster size.

To implement this algorithm, it is very important to determine the optimal number of clusters. One of the popular method is ‘The Elbow Method’

The Elbow Method :





Sometimes , Elbow method does not give required result. As there is gradual decrease in sum of squared distances, optimal of clusters can not be determined.

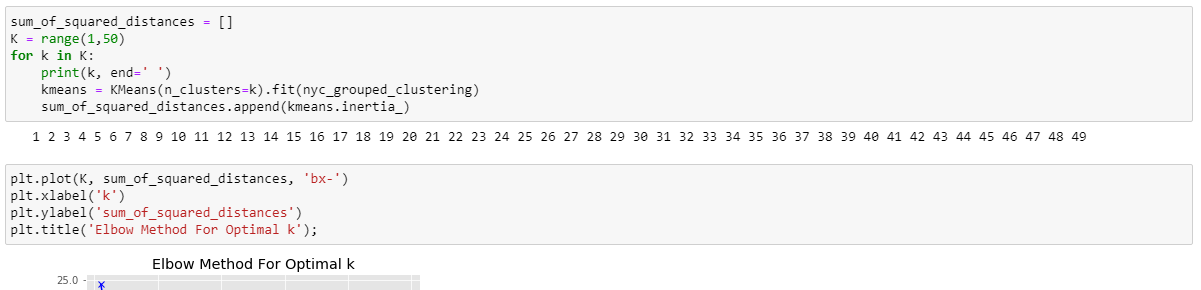
New York City :

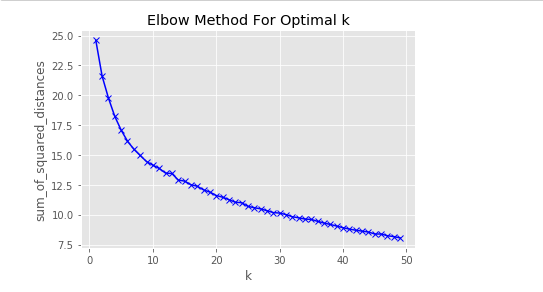
k-means is an unsupervised machine learning algorithm, which creates clusters of data points aggregated together because of certain similarities. This algorithm will be used to count neighborhoods for each cluster label for variable cluster size.

To implement this algorithm, it is very important to determine the optimal number of clusters. One of the popular method is ‘The Elbow Method’

The Elbow Method

The Elbow method calculates the sum of squared distances of samples to their closest cluster centre for different values of ‘k’. The optimal number of clusters is the value after which there is no significant number of decrease in the sum of squared distances.





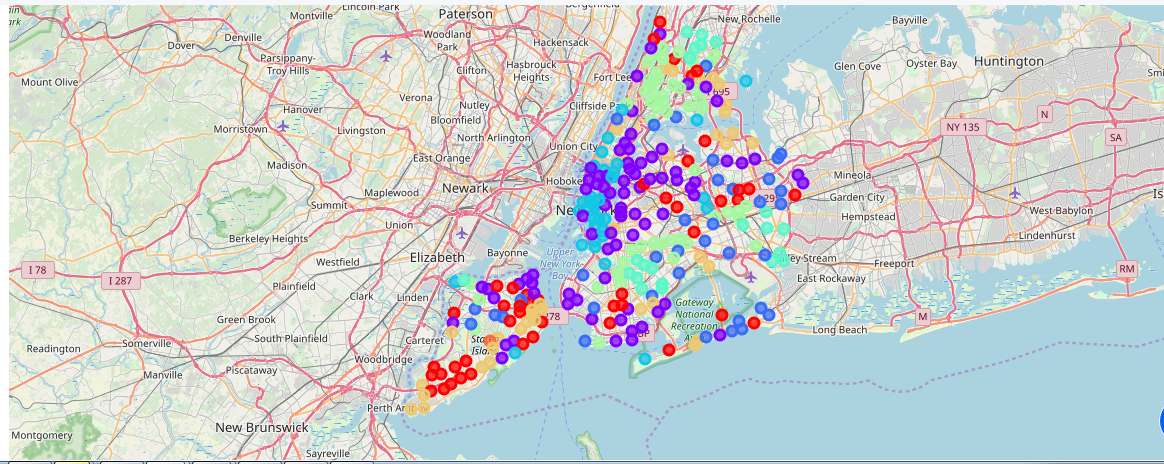
Sometimes , Elbow method does not give required result. As there is gradual decrease in sum of squared distances, optimal of clusters can not be determined.

K-Means:

The following code block runs k-means algorithm with number of clusters = 8 and prints the counts of neighborhoods assigned to different clusters :



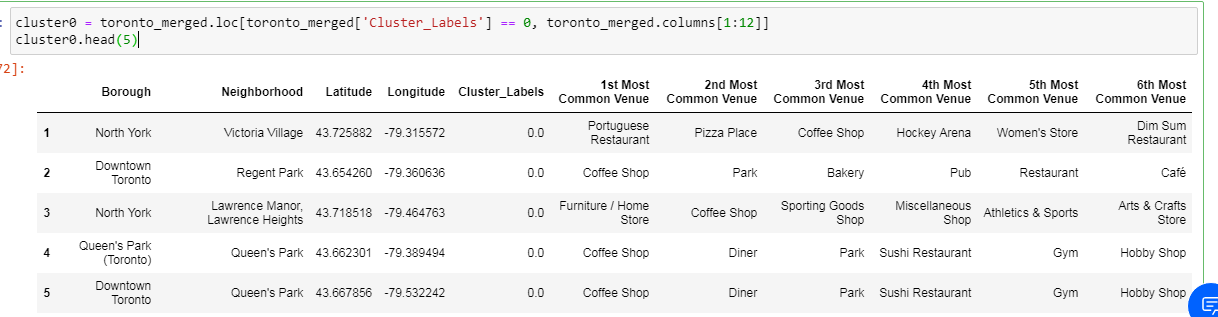
Again, New York City’s neighborhoods are visualized by using code block , which utilizes python ‘folium’ library

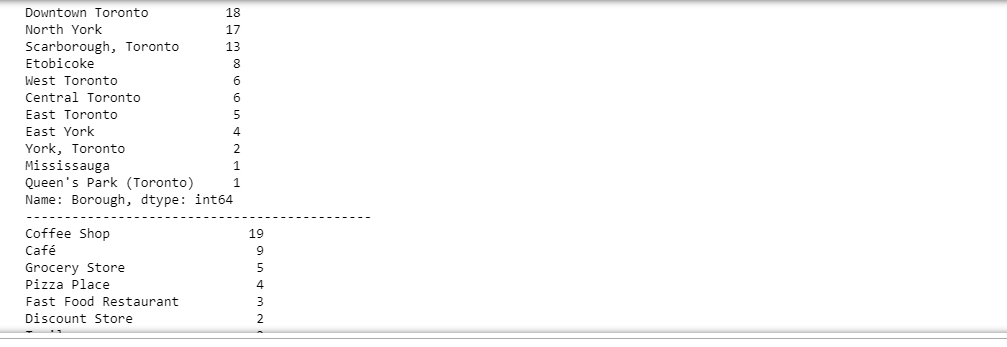


Results :

Toronto Result:

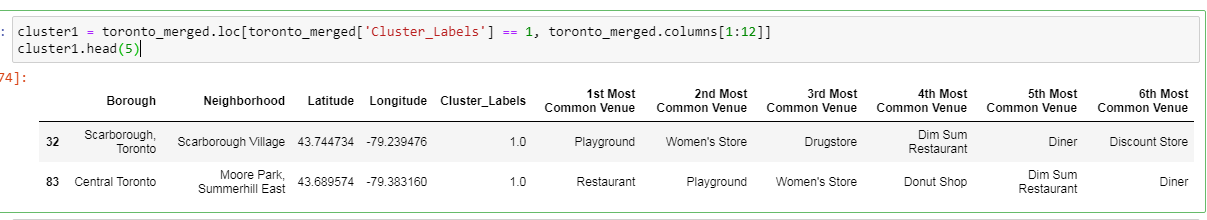
Cluster 0

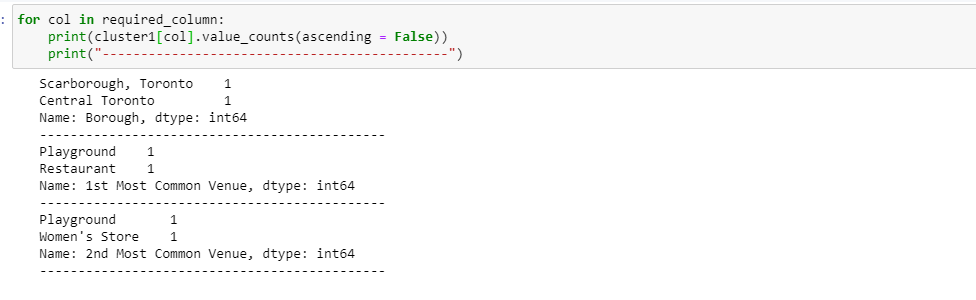




Downtown Toronto has the maximum Coffee shop & Café with 19 & 9 occurrences.

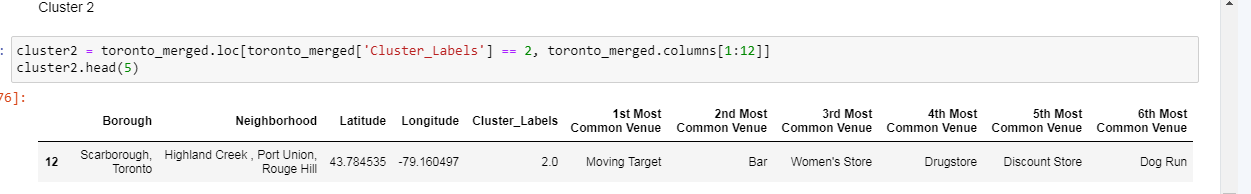
Cluster 1

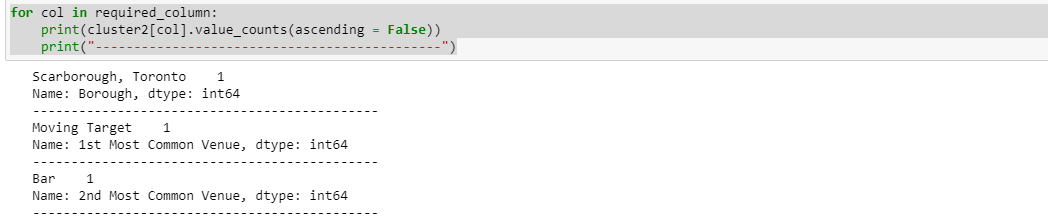




Cluster 1 does not contain any kind of eateries.

Cluster 2





Cluster 2 has 1 bar only.

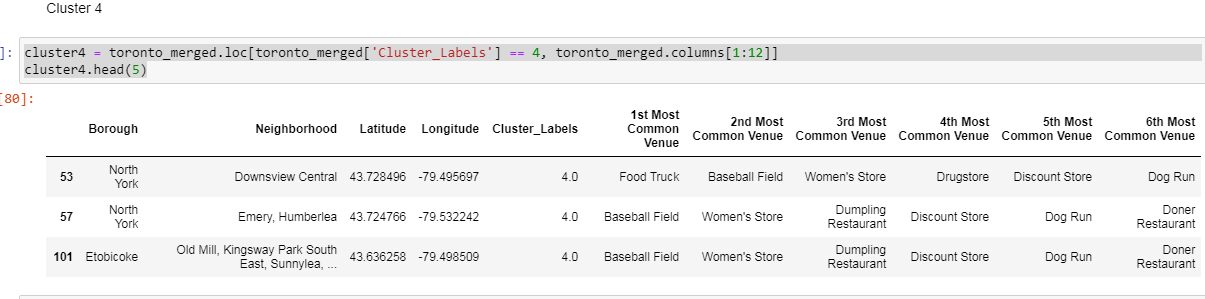
Cluster 3

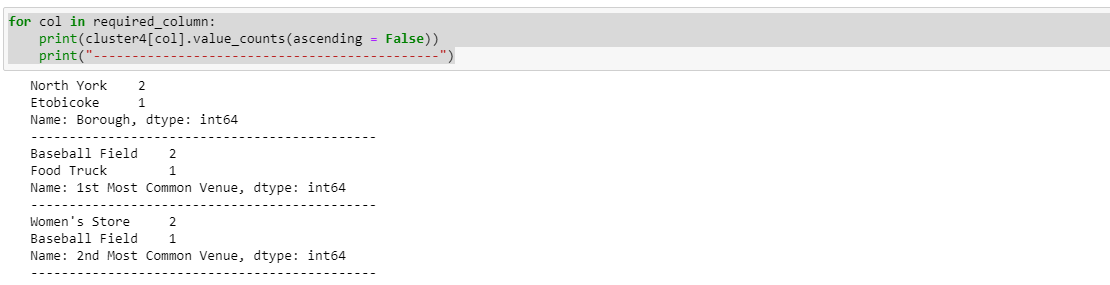




For Cluster 3, Coffee shop is the 2nd common venue with only 1 occurance.

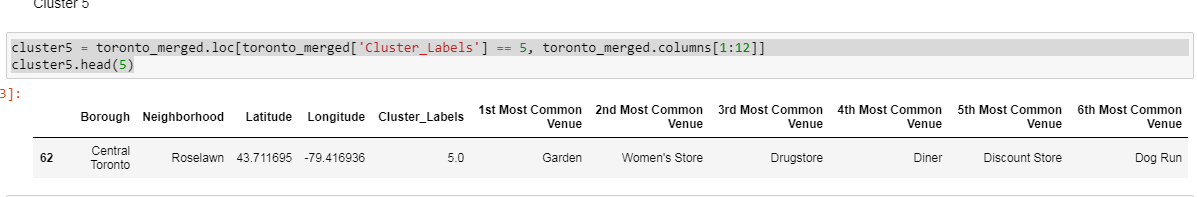
Cluster 4

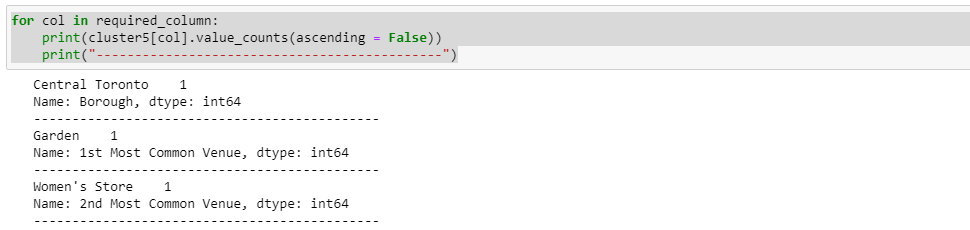




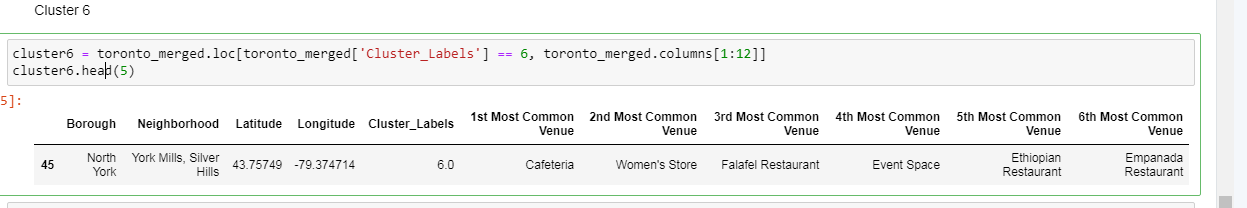
Only 1 food truck is available in Cluster 4 .

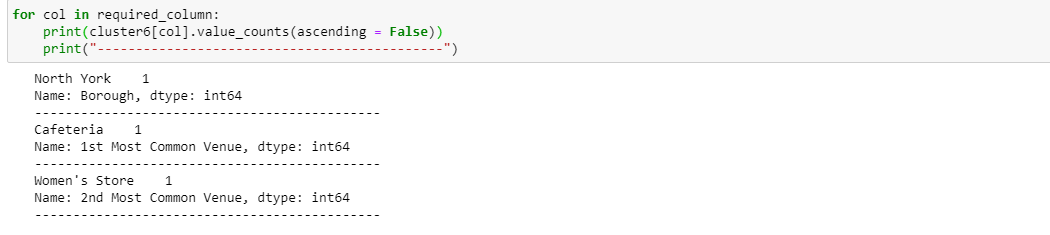
Cluster 5





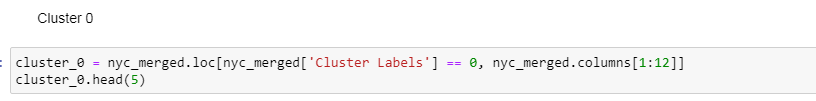
Cluster 6





New York City Result :

Cluster 0



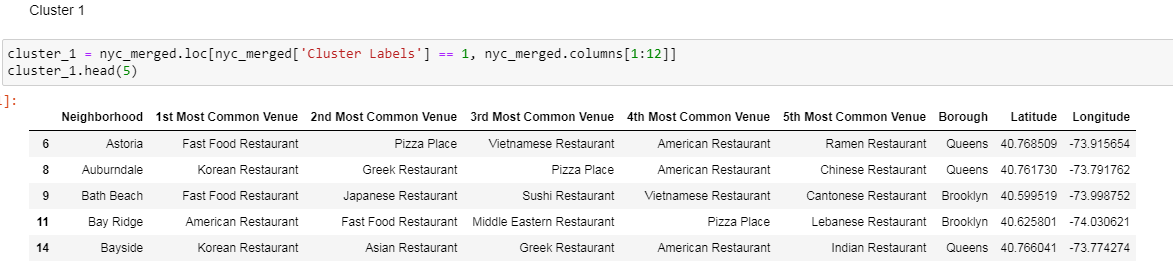
Following are result of cluster 0 analysis:



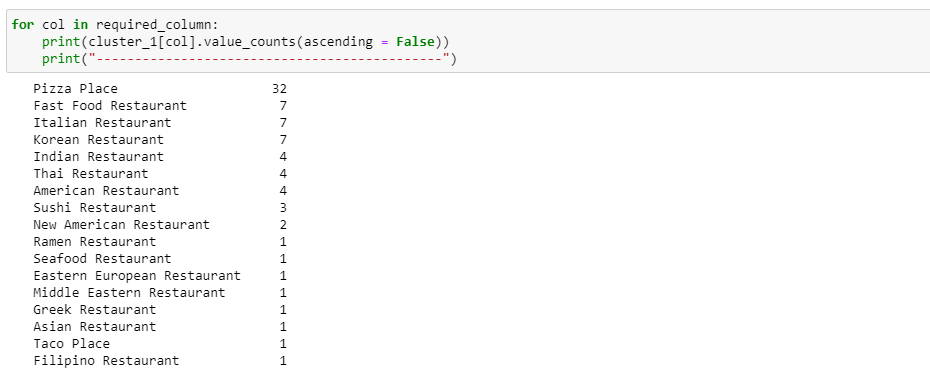
‘Pizza Place’ holds the massive accountability for this cluster with 45 occurrences in ‘1st common venue’ across different neighborhood followed by ‘American restaurant’ & ‘Italian Restaurant’ with 11 & 10 occurrences respectively . It is very inquisitive to know that majority of these neighborhoods are in ‘Staten Island’ borough of New York City.

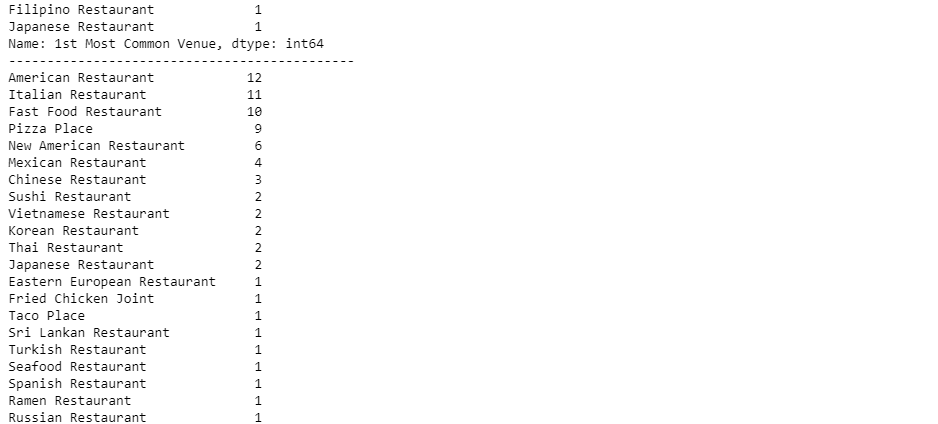
So, Cluster 0 is a ‘Pizza Place’ dominant cluster.

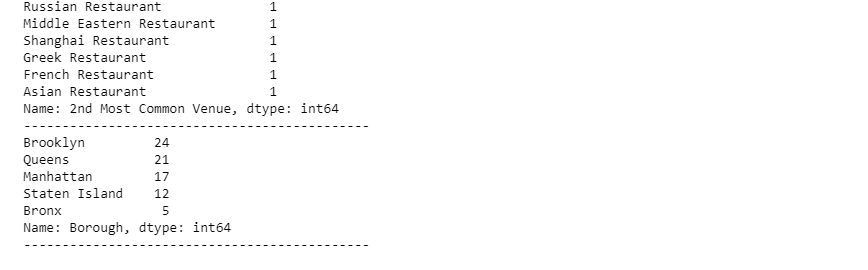
Cluster 1 :



Following are result of cluster 1 analysis:







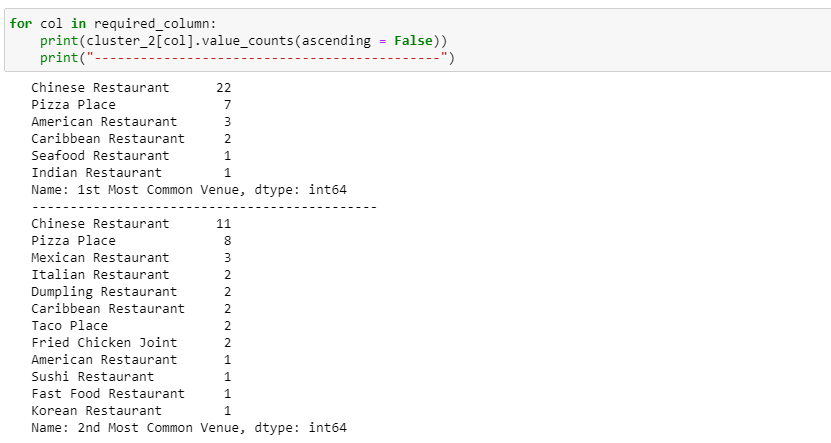
‘Pizza Place’ holds the massive accountability for this cluster with 32 occurrences in ‘1st common venue’ across different neighborhood followed by ‘American restaurant’ & ‘Italian Restaurant’ & ‘Fast Food Restaurant’ with 12, 11 & 10 occurrences respectively . It is very inquisitive to know that majority of these neighborhoods are in ‘Brooklyn’ borough of New York City.

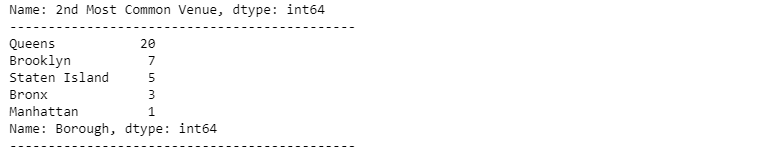
So, Cluster 1 is a ‘Pizza Place’ dominant cluster.

Cluster 2 :



Following are result for cluster 2 analysis:

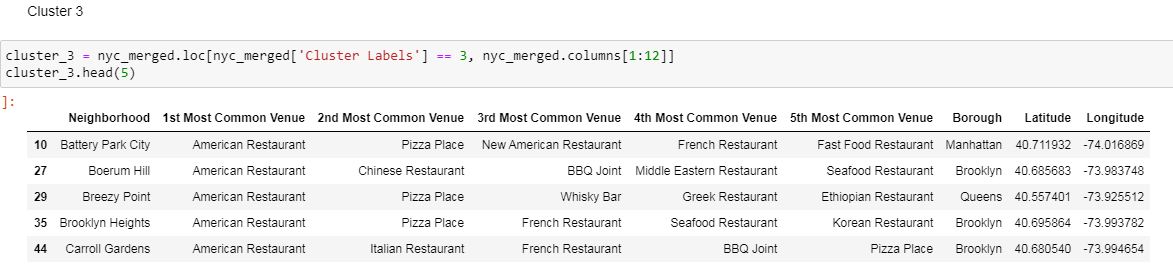




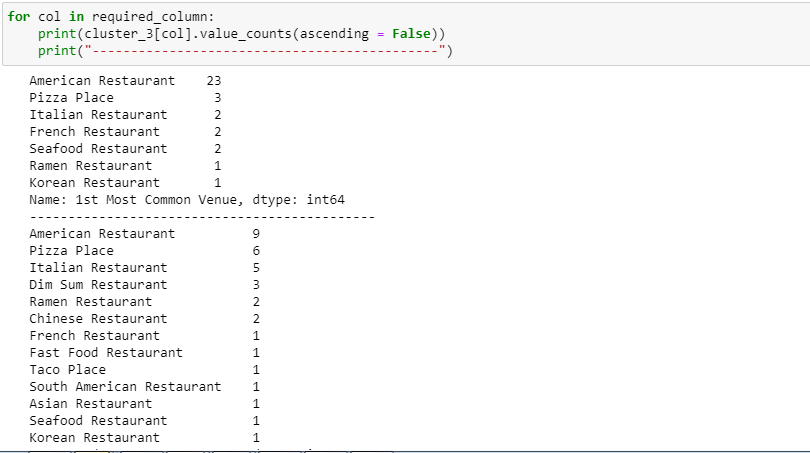
‘Chinese Restaurant’ holds the massive accountability for this cluster with 22 occurrences in ‘1st common venue’ across different neighborhood followed by ‘Chinese restaurant’ & ‘Pizza Place’ with 11 & 8 occurrences respectively . It is very inquisitive to know that majority of these neighborhoods are in ‘Queens’ borough of New York City.

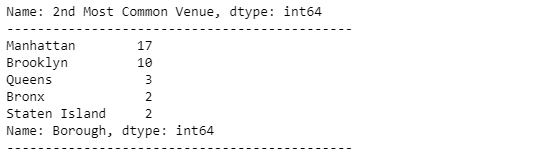
So, Cluster 2 is a ‘Chinese Restaurant’ dominant cluster.

Cluster 3 :



Following are result of cluster 3 analysis:





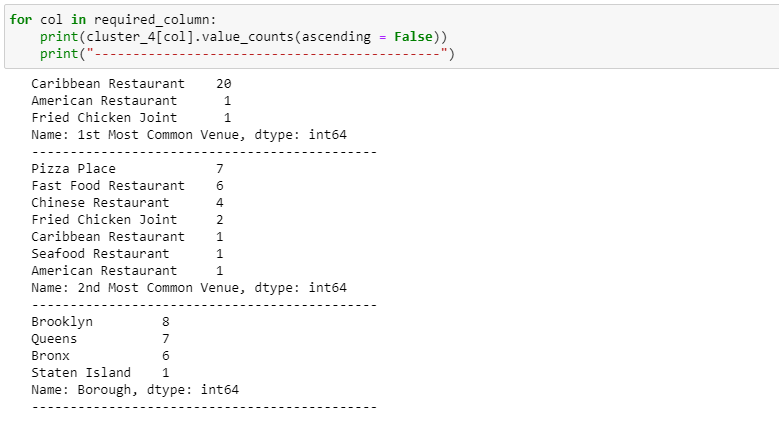
‘American Restaurant’ holds maximum accountability for this cluster with 23 occurances as ‘American Restaurant & ‘Pizza Place’ with 9 & 6 occurances in ‘1st common venue’ across different neighborhood followed by the majority neighborhood in ‘Manhattan’ & ‘Brooklyn’ of NYC.

So, Cluster 3 is ‘American Restaurant’ dominant.

Cluster 4 :



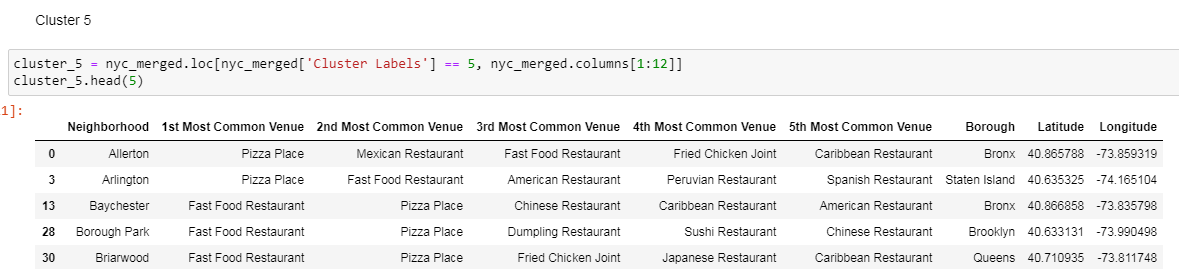
Following are result of Cluster 4 analysis:



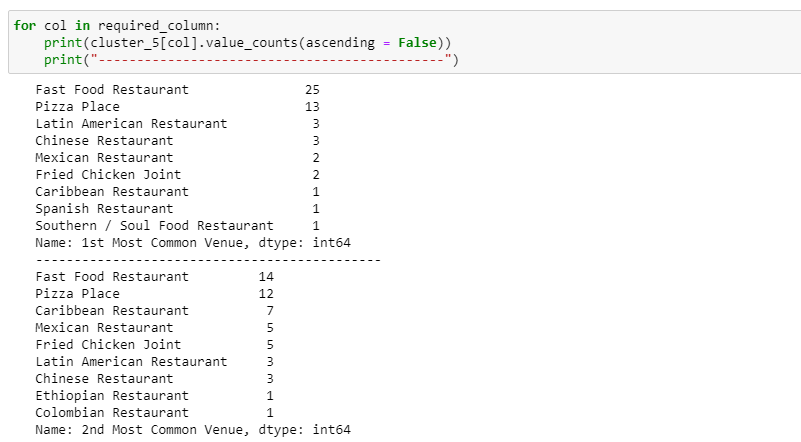
‘Korean Restaurant’ holds maximum accountability for this cluster with 8 occurances as ‘Korean Resturant & ‘American Resturant’ with 4 & 4 occurances in ‘1st common venue’ across different neighborhood followed by the majority neighborhood in ‘Manhatten’ of NYC.

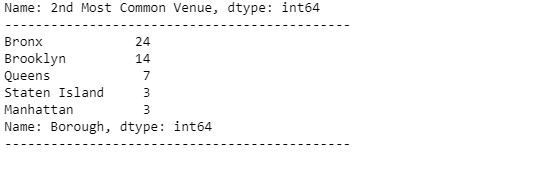
So, Cluster 4 is ‘Korean Restaurant’ dominant.

Cluster 5 :



following are result of cluster 5 analysis:





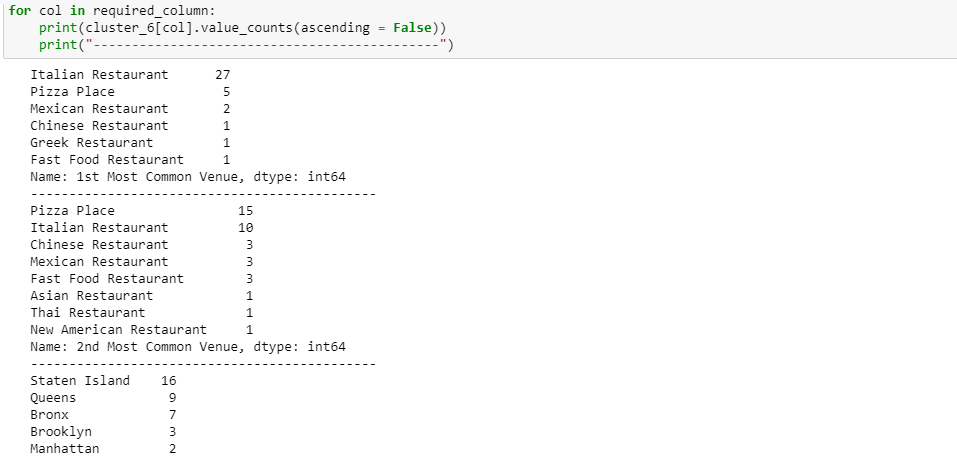
‘Fast Food Restaurant’ & ‘Pizza Place’ holds maximum accountability for this cluster with 25 & 13 occurrences as ‘Fast Food Restaurant & ‘Pizza Place’ with 14 & 12 occurances in ‘1st common venue’ across different neighborhood followed by the majority neighborhood in ‘Bronx’ & ‘Brooklyn’ of NYC.

So, Cluster 5 is ‘Fast Food Restaurant’ & ‘Pizza Place’ dominant.

Cluster 6 :



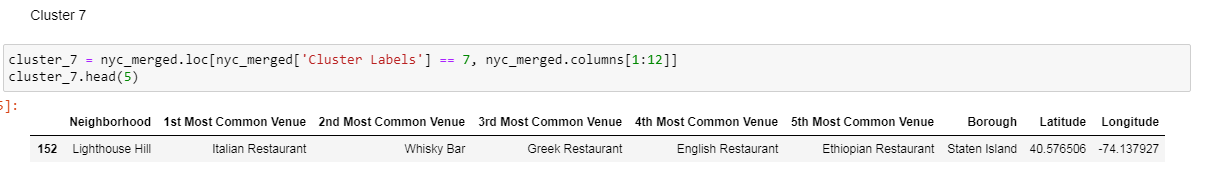
Following are result of cluster 6 cluster:



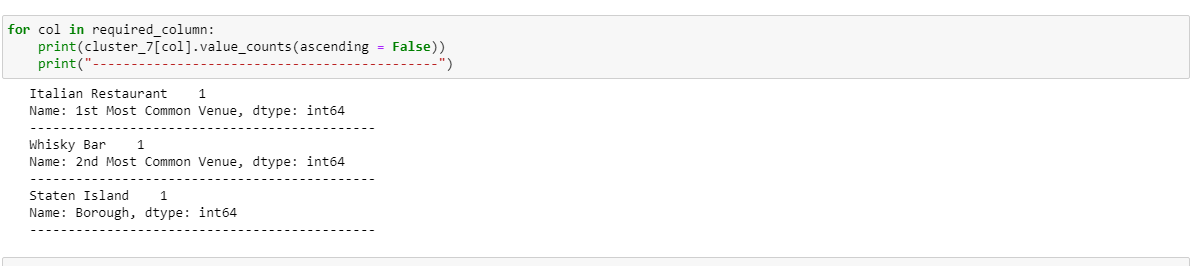
‘Italian Restaurant’ & ‘Pizza Place’ holds maximum accountability for this cluster with 27 & 5 occurrences as ‘Pizza Place’ & ‘Italian Restaurant’ with 15 & 10 occurances in ‘1st common venue’ across different neighborhood followed by the majority neighborhood in ‘Staten Island’ & ‘Queens’ of NYC.

So, Cluster 6 is ‘Italian Restaurant’ & ‘Pizza Place’ dominant.

Cluster 7:



Following are result of cluster 7 analysis:



Cluster 7 is only on ‘Staten Island’ borough of ‘New York City’ , which has Italian Restaurant & Whisky bar .

Discussion:

* Discussion on Toronto

On analyzing the food diversity, we came across a shocking fact, that the food diversity in Toronto is not much effective. While studying our clusters, we found that Toronto people are more on Coffee shops & café. Big joint restaurants are not available in Toronto.

* Discussion on New York City

To understand the clusters, analysis are done on three perspective –

Count of ‘Borough’

Count of ‘1st common value’

Count of ‘2nd common value’

Tabulating the results of k-means ML algorithm :

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | 1st common venue | 2nd common venue | Borough |
| 0 | Pizza Place | American Restaurant, Italian Reataurant | Staten Island, Queens |
| 1 | Pizza Place | American Restaurant, Italian Restaurant, Fast Food Restaurant, Pizza Place | Brooklyn, Queens, Manhatten, Staten Island |
| 2 | Chinese Restaurant, Pizza Place | Chinese Restaurant, Pizza Place | Queens |
| 3 | American Restaurant | American Restaurant ,  Pizza Place ,  Italian Restaurant | Manhattan, Brooklyn |
| 4 | Carribean Restaurant | Pizza Place, Fast Food Restaurant, Chinese Restaurant | Brooklyn, Queens, Bronx |
| 5 | Fast Food Restaurant, Pizza Place | Fast Food Restaurant, Pizza Place, Carribean Restaurant | Bronx, Brooklyn, Queens |
| 6 | Italian Restaurant, Pizza Place | Pizza Place, Italian Restaurant | Staten Island, Queens, Bronx |
| 7 | Italian Restaurant | Whisky bar | Staten Island |

From the conclusion, its clear that Pizza Place is the most common place in New York city.

Conclusion:

On application of cluster algorithm, a very inquisitive result can be curated which helps in understanding & visualization on the data. Pizza Place is always a ready-to-go restaurant in New York. Other than that, there are American, Italian & Chinese restaurant found in the diversity. Whereas, Toronto is not in big restaurants.

Now, the company , who wants to open a new restaurant, need to play strategically, to earn a good business. New York is more dependable to open a new food chain as diversity is much more available there. Whereas, it may happen that Toronto will go good in big fat restaurants , as there is no competitor in the market for the company.