**Data Science Capstone Project Report:**

**Clustering and Segmenting and Analysis the New York and Toronto neighbourhood**

Prepared By: Nilabja Haldar

Version: 1.0

Date: 26.11.2019

Contents

[1. Business Background 3](#_Toc25688884)

[2. Problem Definition 5](#_Toc25688885)

[3. Problem Definition and Target audience 5](#_Toc25688886)

[4. Data Requirement 6](#_Toc25688887)

[5. Data Pre-processing and Data Cleaning 11](#_Toc25688888)

[6. Methodology 12](#_Toc25688889)

[7. Results 18](#_Toc25688890)

[8. Discussion 21](#_Toc25688891)

[9. Conclusion 22](#_Toc25688892)

This file contains information about the project being executed, in this case, the Clustering and Analysis of Neighbourhood data of New York and Toronto. It is organized according to the capstone project process flow.

NOTE: This is a sample for a tutorial, so scope, plan etc., does not necessarily correspond to an actual data science project addressing a specific business question. In an actual project, the problem definition, scope, plan, personnel sections are likely to be much more detailed, based on discussions with the client (or business owner), the structure of the data science team etc.

# Business Background

The study include 2 of most popular tourist and industrial centre in the world.

Toronto

Toronto is the provincial capital of Ontario and the most populous city in Canada,

Toronto is a prominent centre for music, theatre, motion picture production, and television production, and is home to the headquarters of Canada's major national broadcast networks and media outlets.

Its varied cultural institutions, which include numerous museums and galleries, festivals and public events, entertainment districts, national historic sites, and sports activities, attract over 43 million tourists each year. Toronto is known for its many skyscrapers and high-rise buildings, in particular the tallest free-standing structure in the Western Hemisphere, the CN Tower.

Toronto has a diverse array of public spaces, from city squares to public parks overlooking ravines. Nathan Phillips Square is the city's main square in downtown, and forms the entrance to City Hall.

Large parks in the outer areas managed by the city include High Park, Humber Bay Park, Centennial Park, Downsview Park, Guild Park and Gardens, and Morningside Park. Toronto also operates several public golf courses. Most ravine lands and river bank floodplains in Toronto are public parklands.

The Royal Ontario Museum is a museum of world culture and natural history. The Toronto Zoo is home to over 5,000 animals representing over 460 distinct species. The Art Gallery of Ontario contains a large collection of Canadian, European, African and contemporary artwork, and also plays host to exhibits from museums and galleries all over the world. The Gardiner Museum of ceramic art is the only museum in Canada entirely devoted to ceramics, and the Museum's collection contains more than 2,900 ceramic works from Asia, the Americas, and Europe. The city also hosts the Ontario Science Centre, the Bata Shoe Museum, and Textile Museum of Canada.

City of New York

The City of New York, usually referred to as either New York City (NYC) or simply New York (NY), is the most populous city in the United States. New York is also the most densely populated major city in the United States. Located at the southern tip of the 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 and one of the world's most populous megacities.

New York City is often referred to collectively as the five boroughs, and in turn, there are hundreds of distinct neighborhoods throughout the boroughs, many with a definable history and character to call their own. If the boroughs were each independent cities, four of the boroughs (Brooklyn, Queens, Manhattan, and the Bronx) would be among the ten most populous cities in the United States

New York City is a global hub of business and commerce, as a center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theater, fashion, and the arts in the United States; while Silicon Alley, metonymous for New York's broad-spectrum high technology sphere, continues to expand.

Tourism is a vital industry for New York City, which has witnessed a growing combined volume of international and domestic tourists. Approximately 12 million visitors to New York City were from outside the United States, with the highest numbers from the United Kingdom, Canada, Brazil, and China.

Major tourist destinations include Times Square; Broadway theater productions; the Empire State Building; the Statue of Liberty; Ellis Island; the United Nations Headquarters; museums such as the Metropolitan Museum of Art; greenspaces such as Central Park and Washington Square Park; etc.

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 examples of modern New York street food.

The city is home to "nearly one thousand of the finest and most diverse haute cuisine restaurants in the world", according to Michelin. The New York City Department of Health and Mental Hygiene assigns letter grades to the city's restaurants based upon their inspection results. As of 2019, there were 27,043 restaurants in the city, The Queens Night Market in Flushing Meadows–Corona Park attracts over 10,000 people nightly to sample food from over 85 countries.

# Problem Definition

This project is to solve the below business cases:

1. First Business case is to target tourists, who are visiting the neighbourhood first time can get an understanding of the top venues surrounding in New York and Toronto. Also they will get a simple comparison between the top venues in different neighbourhood.
2. 2nd business case is, this analysis will help the entrepreneur to give an idea about the which type of business is popular in which area and which type of cuisine popular in which area and will give and idea of where to invest.

# Problem Definition and Target audience

The purpose of this sample is to show how to instantiate and execute a project using the data science structure and tools.

The intended audience of this project is usually the tourists who are first time visiting the neighbourhood of Toronto or New York City and don’t have an idea of the top venues around a place.

So they are sometime lost and didn’t get chance to visit the top places. Using this analysis we can help the tourists to self-guide to where to visit according to their choice.

Target audience of the 2nd use case is upcoming entrepreneurs who want to establish some business in these neighbourhoods.

The issue is most of the times the less experience budding entrepreneurs have less or no idea about the top running business in certain localities, which cuisine is most popular in which area, which shop is most trending in which area, and which area has what popular venues. So lack of this data often misguides them to enter in wrong business and can be a trouble to them.

# Data Requirement

.

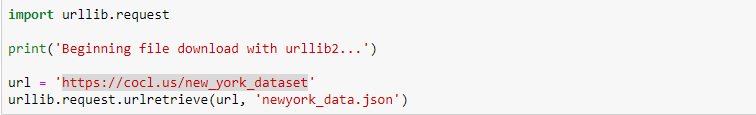
The dataset for this project is collected from various sources.

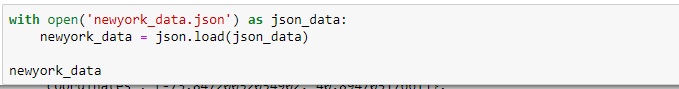
**New York City Data:**

New York City data is collected from the below link:

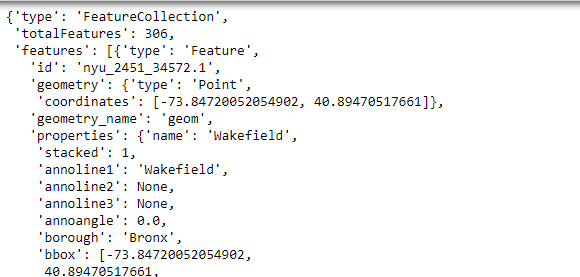
<https://cocl.us/new_york_dataset>

This is a json file. So this needs to read the data in json format and loaded locally in newyork\_data.json file. The read the data from the local json file.

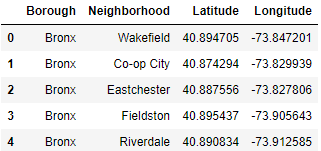




Sample json data:



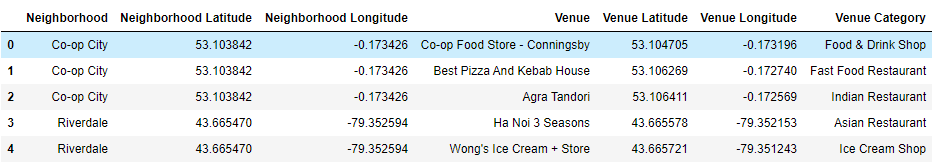
This json data then transformed and pre-processed into data frame with columns Borough, neighbourhood, latitude and longitude.



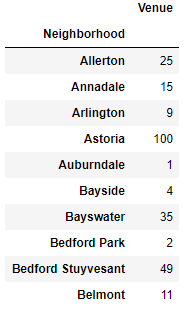
The dataframe has 5 boroughs and 306 neighborhoods.

Foursquare location data used along with this data to find out the venues, popularity, location etc.

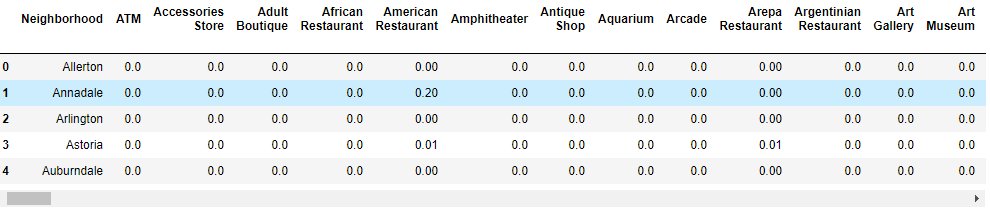
Foursquare location data is used to venue location, details and category.



How many venue in which location.



Mean venue count per neighbourhood:



Top 10 most common venues per neighbourhood:

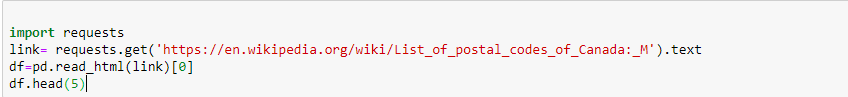


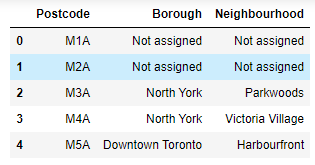
**Toronto Data:**

Toronto data is collected from the below link:

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

This Wikipedia contains the postal codes of the neighbourhood of Toronto. The table read from the Wikipedia as below and loaded as data frame.

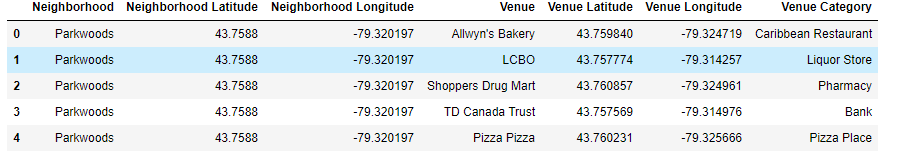




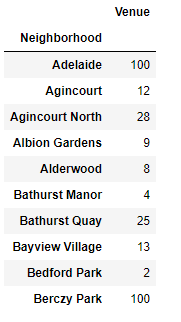
The dataframe has 287 records.

Next foursquare location data used along with this data to find out the venues, popularity, location etc. And used for further analysis, clustering and segmentation.

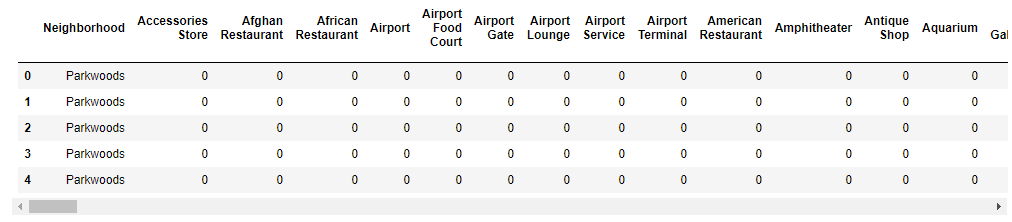
Foursquare location data is used to venue location, details and category.



How many venue in which location.



Mean venue count per neighbourhood:



Top 10 most common venues per neighbourhood:

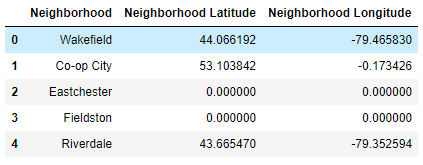


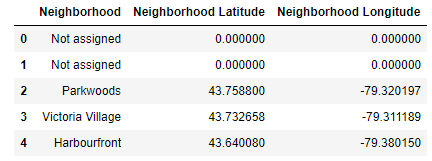
# Data Pre-processing and Data Cleaning

**Feature Engineering:**

**Data cleanup: Removing columns and rows:** In data science project the data we get is not always perfect, error free and in right format. So it is very often that we need to pre-process the data after we receive the data and also we need to clean the data.

Here I noticed that there are many rows which does not have any latitude and longitude. So if we continue with the missing data then it can cause error in further analysis. So we removed these rows.

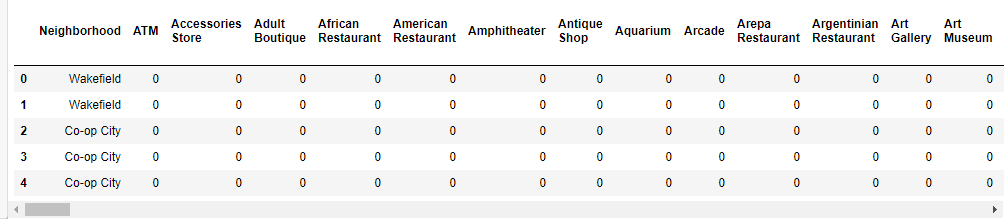




The New York data set before missing data was 306 rows and 3 columns and Toronto was 287 rows and 3 columns. Data set size after data cleaning is for New York 135 Rows and for Toronto 199 rows.

**One-hot encoding categorical features**

Following categorical features were one-hot encoded using dummy column.

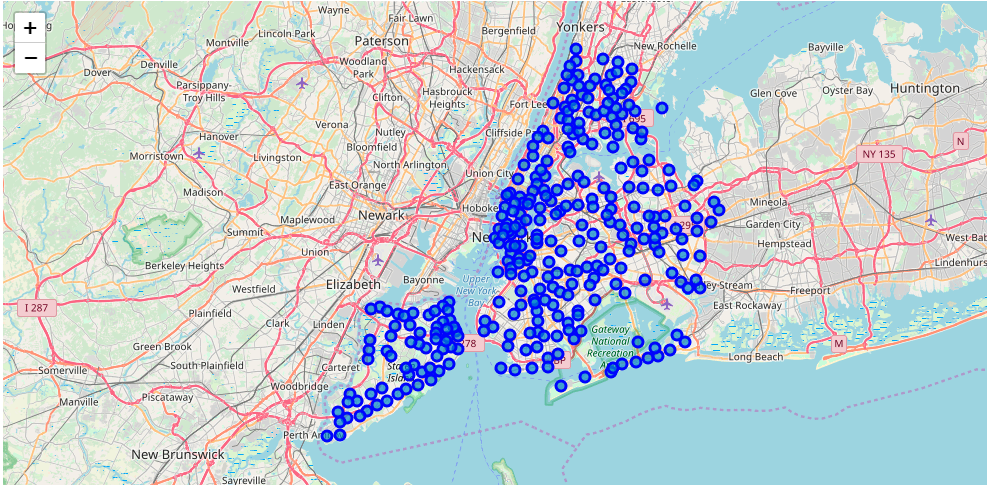


# Methodology

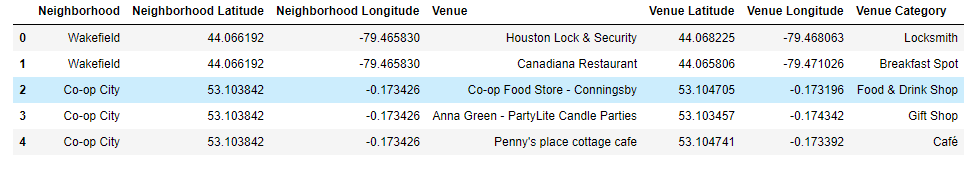
As a database, I used GitHub repository in my study. I have created a public repository in my git account, name as [Capstone\_Neighborhood\_Analysis](https://github.com/nilabja-OpenSC/Capstone_Neighborhood_Analysis).

My master data which has the main components Borough, Neighbourhood information of the city is taken from Wikipedia and online repository.

I used python folium library to visualize geographic details of New York and Toronto and its boroughs and I created a map of New York and Toronto with boroughs superimposed on top. I used latitude and longitude values to get the visual as below:

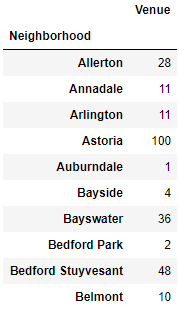


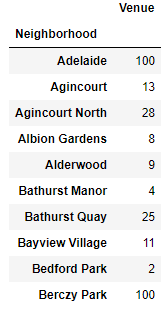
I utilized the Foursquare API to explore the boroughs and segment them. Here is a head of the list Venues name, category, latitude and longitude information from Forsquare API.





In summary of this data below no of venues were returned by Foursquare. Here is a table of no of venues by boroughs.





We can see that Adelaide, Berczy Park reached the 100 limit of venues. On the other hand; Albion Gardens, Alderwood, Bathurst Manor, Bedford Park boroughs are below 20 venues in our given coordinates with Latitude and Longitude.

The result doesn’t mean that inquiry run all the possible results in boroughs. Actually, it depends on given Latitude and Longitude informations and here is we just run single Latitude and Longitude pair for each borough. We can increase the possibilities with Neighborhood informations with more Latitude and Longitude informations.

In summary of this graph 256 unique categories were returned by Foursquare, then I created a table which shows list of top 10 common venue category for each borough in below table.

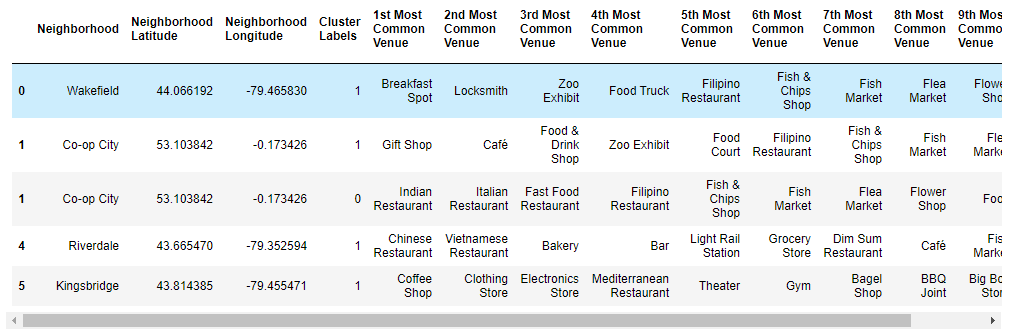




We have some common venue categories in boroughs. In this reason I used unsupervised learning K-means algorithm to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning.

First, I will run K-Means to cluster the boroughs into 5.

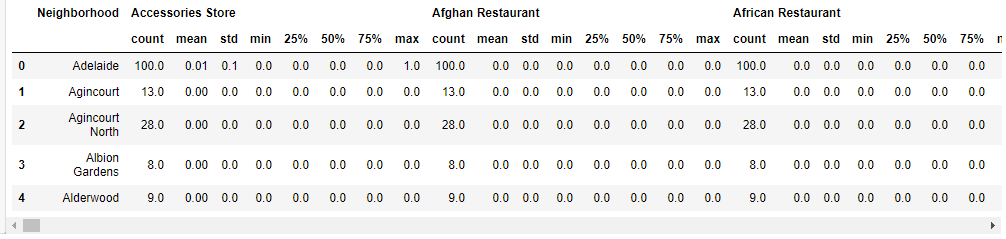
Here is my merged table with cluster labels for each borough.



We can also estimate the number of 1st Most Common Venue in each cluster.

**Inferential Statistics**

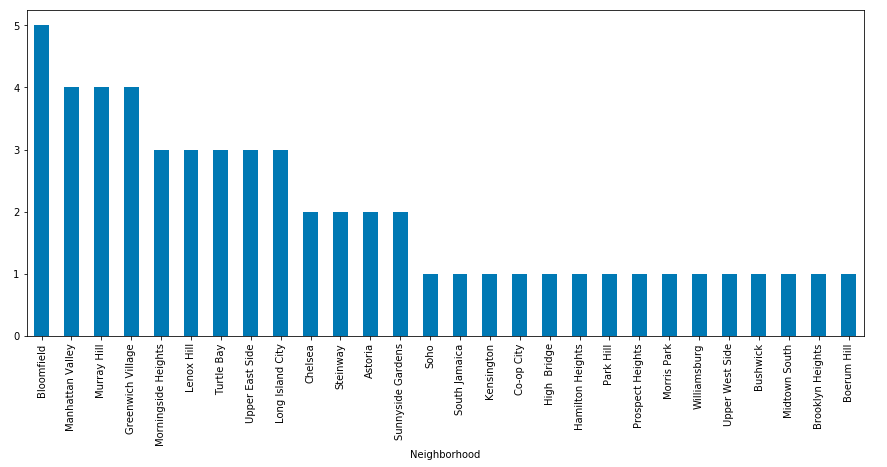
I used inferential statistics to find out the statistics about Indian Restaurant in both the location. First below is the statistical analysis of venue category and Neighbourhood wise.

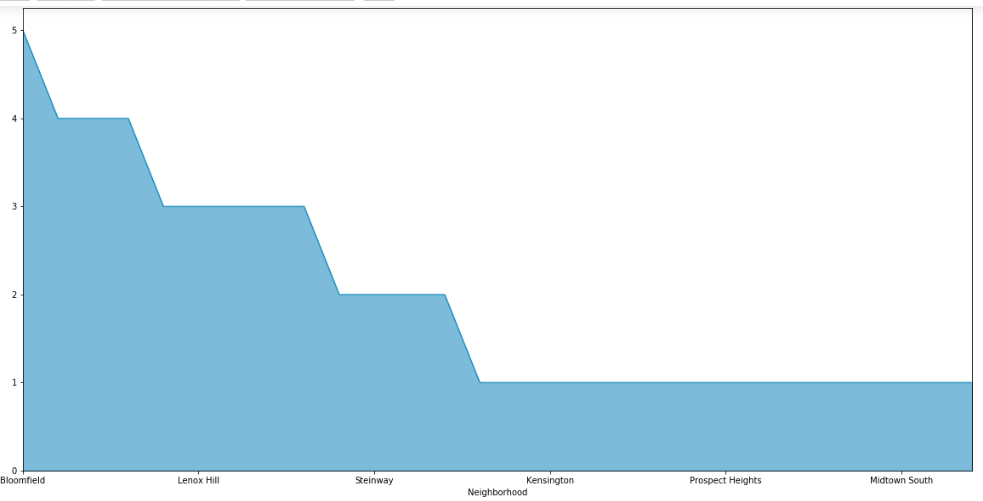


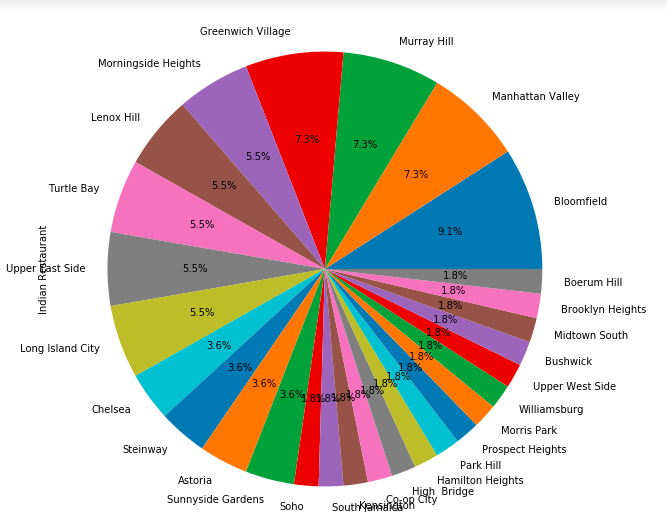
Below is the count of Indian restaurants in different locations.



Below are the different charts created to show the distribution of Indian restaurants location wise.



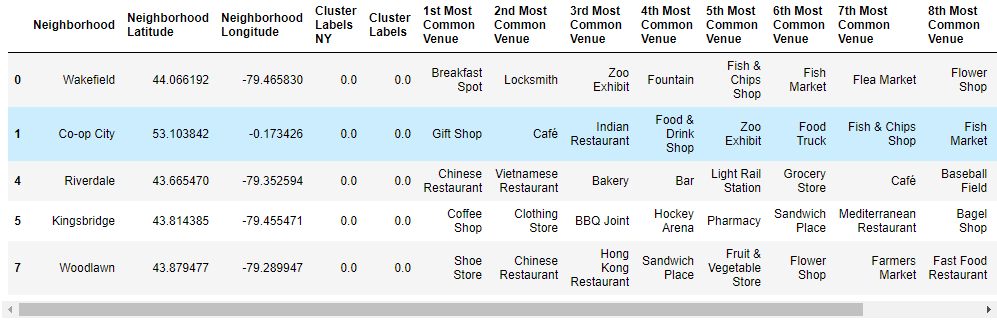




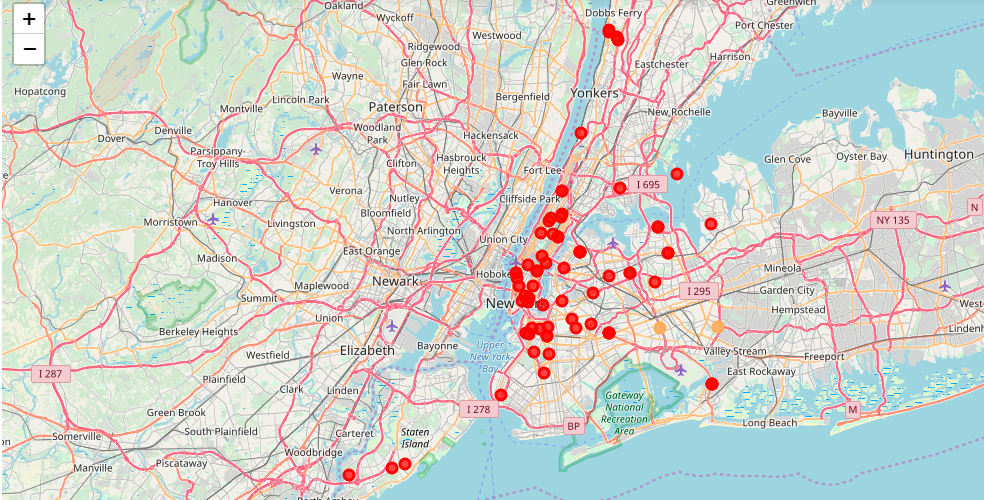
# Results

We follow the stages for the capstone lifecycle, and organize documentation and code according to the stages of the lifecycle. Documentation about the work and findings in each of the lifecycle stages is included.

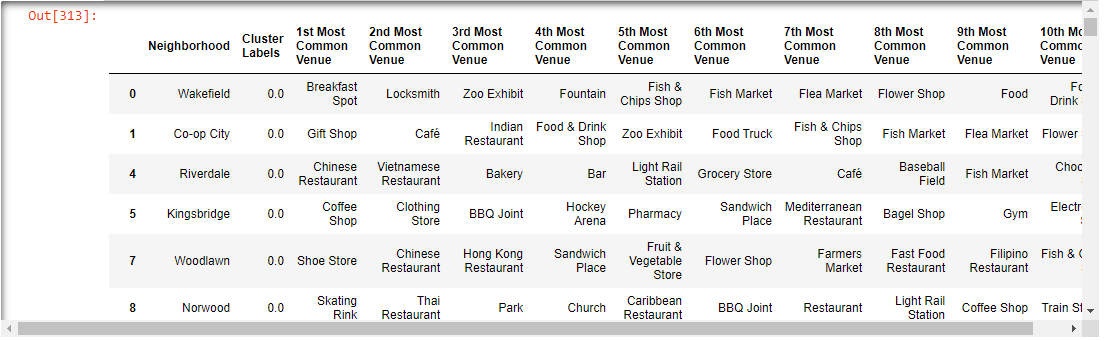
I merge new variables with related cluster informations in our main master table.

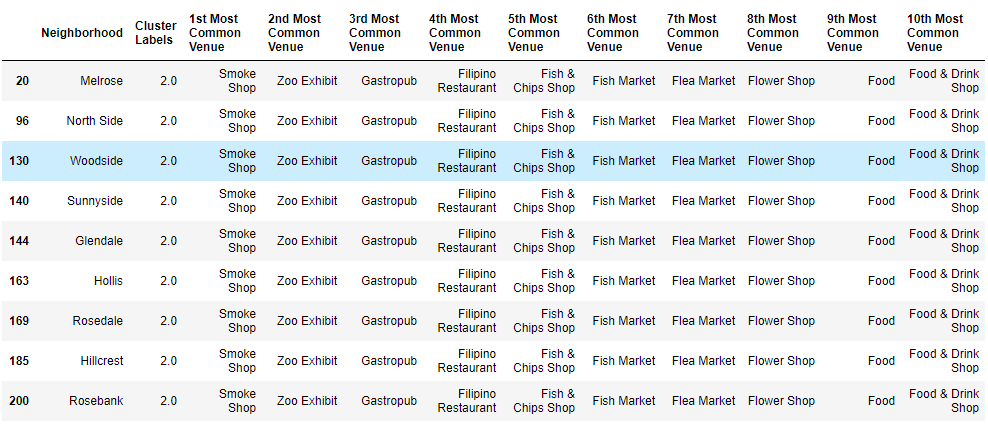


And created the below map.

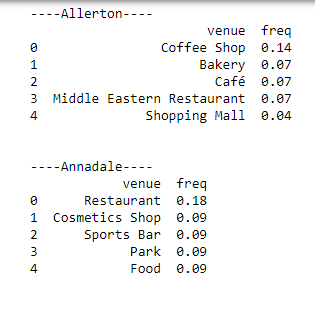


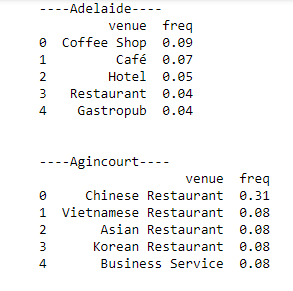
One of my aim was to show the most common venues and the popular venues. These are shown with the clusters and the category.



**Top 5 Venues:**





# Discussion

As I mentioned before, New York and Toronto are big cities with a high population density in a narrow area. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for this metropol.

I used the Kmeans algorithm as part of this clustering study. For more detailed and accurate guidance, the data set can be expanded and the details of the neighborhood or street can also be drilled.

I also performed data analysis through this information by adding the coordinates of districts and codes and documents stored GitHub. In future studies, these data can also be accessed dynamically from specific platforms or packages.

I ended the study by visualizing the data and clustering information on the New York and Toronto map. In future studies, web or telephone applications can be carried out to direct investors.

# Conclusion

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

As an entrepreneur a chef can try to start a restaurant in a city and can use this information to get the overall view and statistics of which cuisine is popular is which area. And according to that he can take decision.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.