IBM Data Science Capstone Project: New Yorkers live in Toronto with Data Science

A. Introduction

The quality of life is one of pursuances in our life. Someone live in New York with high pressures may want to move to a place which is slow pace. At this moment, Toronto would be a good choice. There are three reasons to support this decision. First, the weather is similar with New York. Second, the distance between New York and Toronto is about 1.5 hours flight time. Third, the **cost of living index** in Toronto is 72.26 which is much lower than New York(100). However, How do we choose a community or a borough in Toronto if I originally live in New York? This report can provide a point of view to let you find similar Toronto places with the place you live in New York by Data Science. Moreover, the **Foursquare API** is used to explore neighborhoods in both cities. The **explore** function is to get the most common venue categories in each neighborhood, and then to group the neighborhoods into clusters. The k-means clustering algorithm is to complete this task. The 10 clusters would be generated to let users to find similar communities between New York City and Toronto. Finally, the Folium library is to visualize the neighborhoods in New York City and Toronto with emerging clusters.

B. Tools and Data

B.1 Tools

- Jupyter Notebook is used to execute the python program and describe the logics with figures and charts in the browser.

- Numpy is a library to handle data in a vectorized manner.

- Pandas is a library for data analysis.

- Geopy is to convert an address into latitude and longitude values.

- Matplotlib and Seaborn library is used to make the chart to show the correlation among the variables.

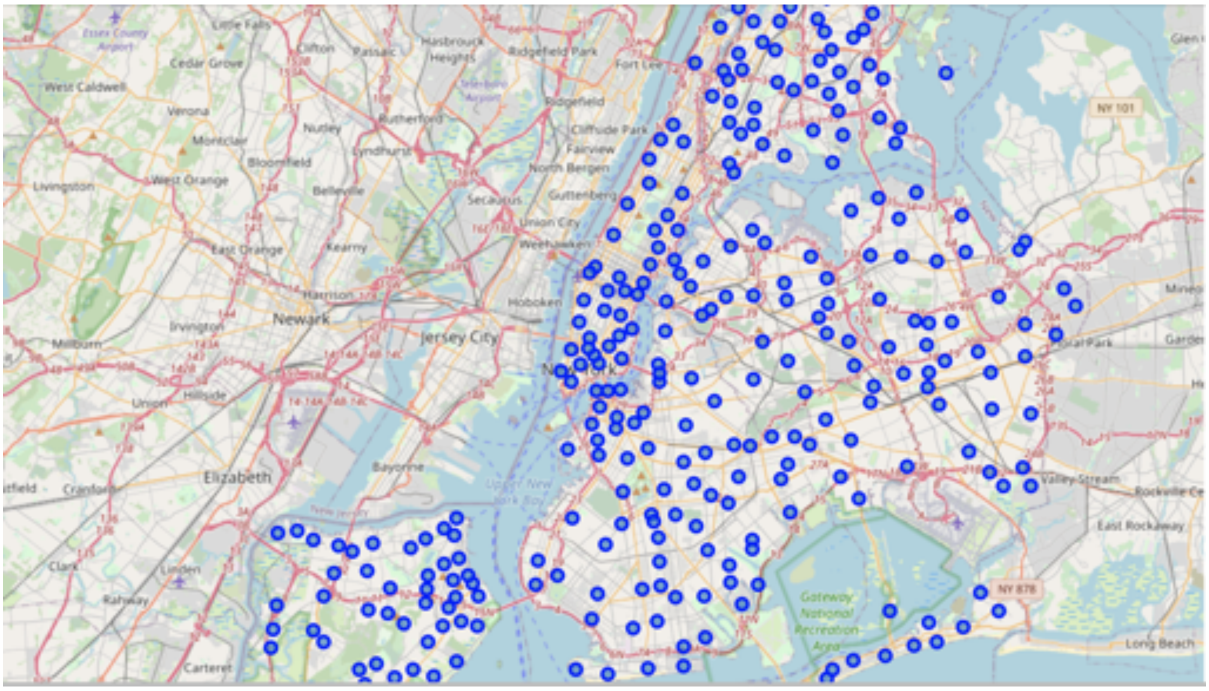
- Folium library is used to visualize geographic details of Seattle and its clustering marks can demonstrate the number of vehicle accidents.import numpy as np # library to handle data in a vectorized manner.

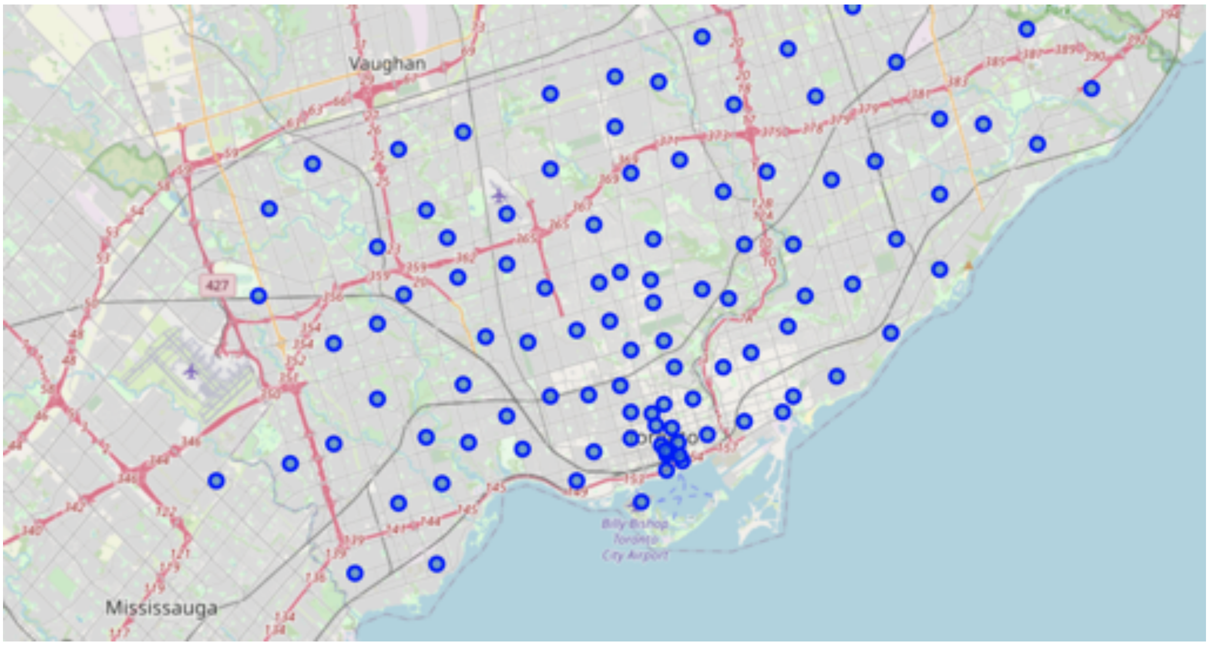
- Foursquare API is to explore neighborhoods.

- Sklearn library is to import the k-means clustering method.

B.2 Data

The Location Data is from previous Applied Data Science Capstone assignment. Neighborhood in New York 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 the latitude and longitude coordinates of each neighborhood. Here is **New York neighborhood dataset** from NYU. Neighborhood in Toronto has a total of 5 boroughs and 103 neighborhoods. There is no dataset for Toronto neighborhood, so the web scraping is used to get information from **wiki**. At the meantime, the coordinate information of Toronto is from Couresera content. Combining the web scraping result and coordinate data is to build the Toronto neighborhood dataset.



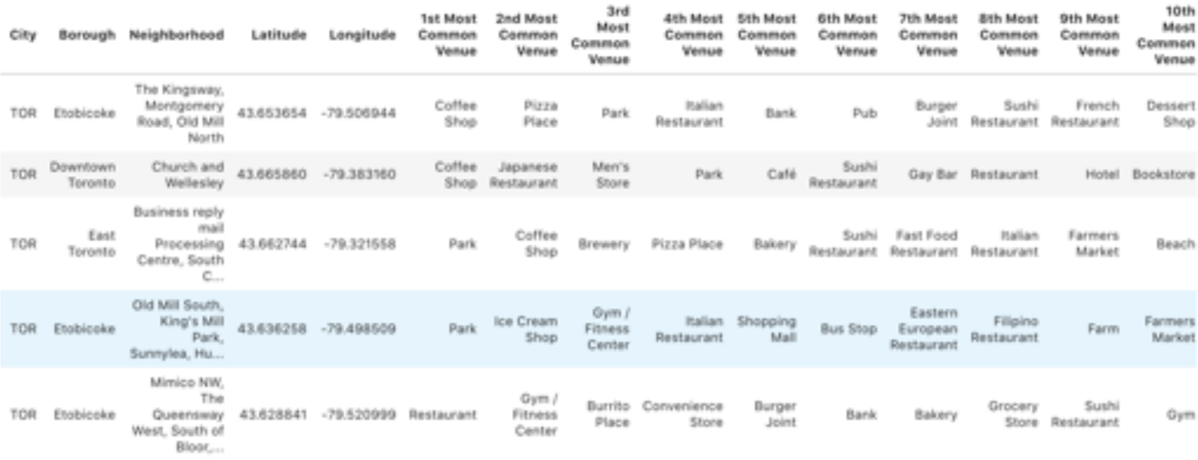


C. Results

C.1 Neighborhood Exploring in New York and Toronto

Exploring neighborhood is to identify the borough conditions. The Foursquare API is used to explore the categorized neighborhood. Using this API to get the top 100 venues are in specific coordinate within a radius of 5000 meters. Extracting the categories of top 100 venues is to build the features of this borough. In order to building the system for comparing New York and Toronto, Only common categories are selected.

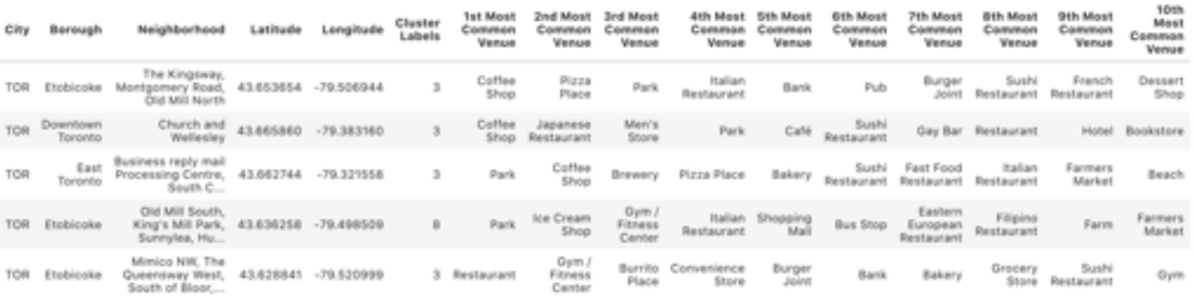
C.2 Analyze Each Neighborhood



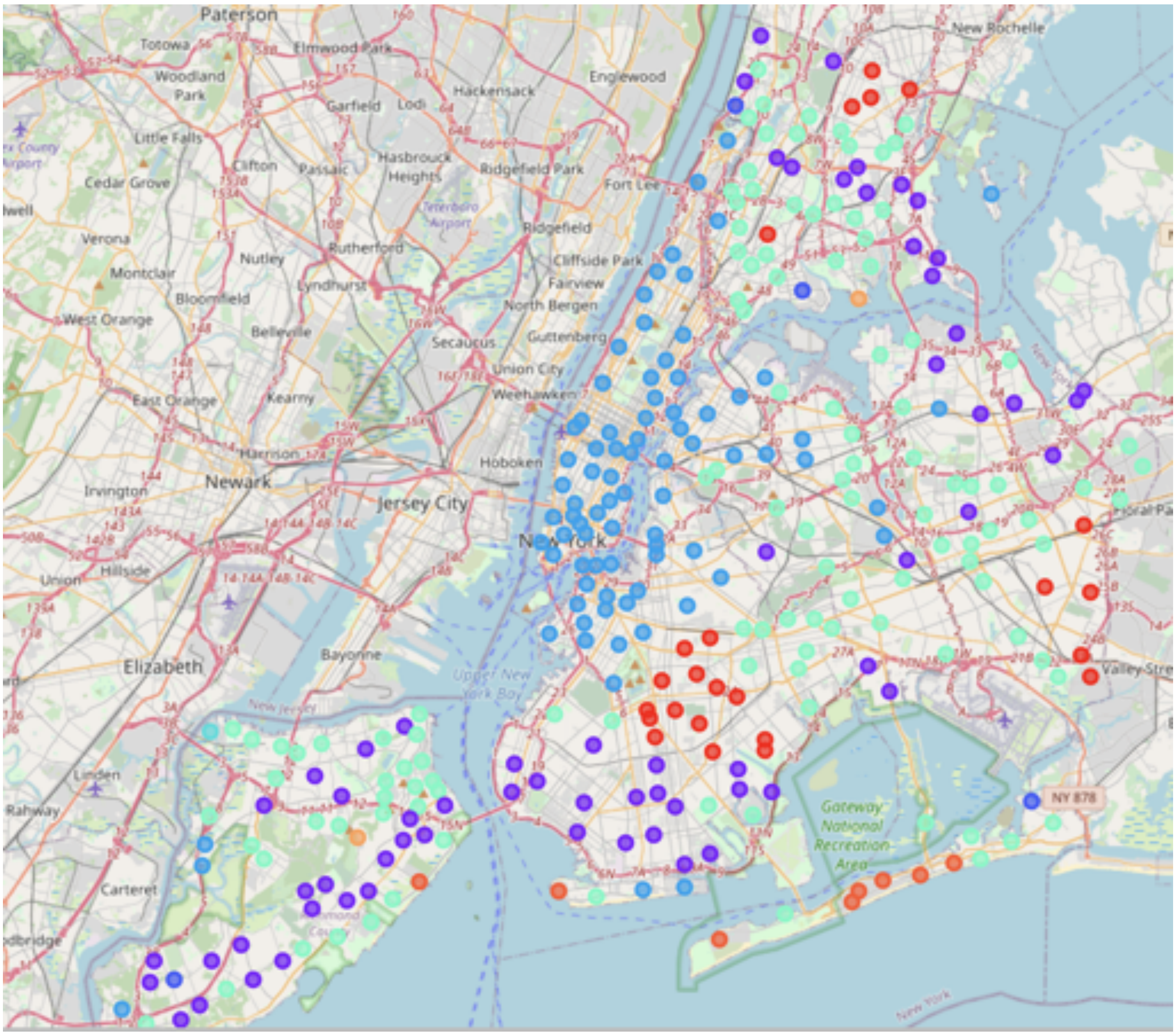
City	Borough	Neighborhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
TOR	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944	Coffee Shop	Pizza Place	Park	Italian Restaurant	Bank	Pub	Burger Joint	Sushi Restaurant	French Restaurant	Dessert Shop
TOR	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	Coffee Shop	Japanese Restaurant	Men's Store	Park	Café	Sushi Restaurant	Gay Bar	Restaurant	Hotel	Bookstore
TOR	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558	Park	Coffee Shop	Brewery	Pizza Place	Bakery	Sushi Restaurant	Fast Food Restaurant	Italian Restaurant	Farmers Market	Beach
TOR	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509	Park	Ice Cream Shop	Gym / Fitness Center	Italian Restaurant	Shopping Mall	Bus Stop	Eastern European Restaurant	Filipino Restaurant	Farm	Farmers Market
TOR	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Restaurant	Gym / Fitness Center	Burrito Place	Convenience Store	Burger Joint	Bank	Bakery	Grocery Store	Sushi Restaurant	Gym

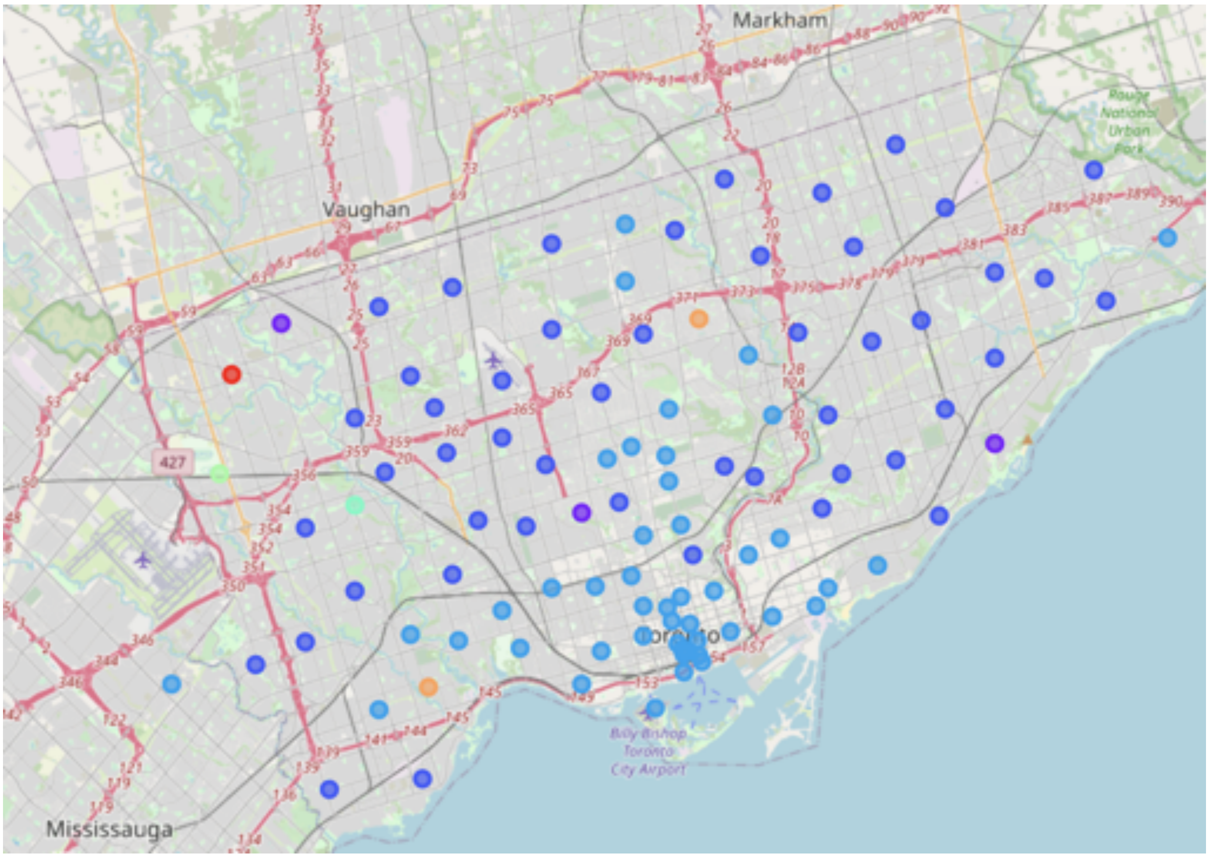
C.3 Cluster Neighborhoods

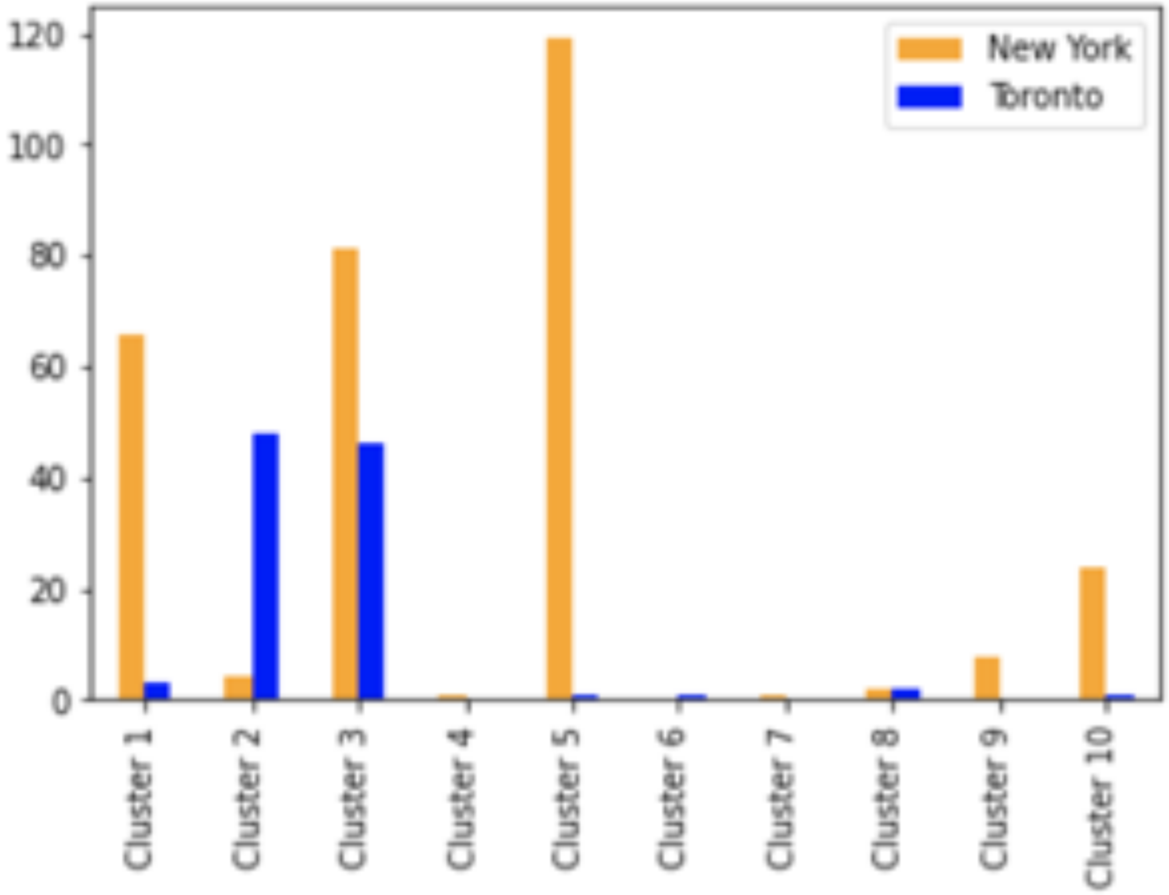
Each Neighborhood with cluster labels

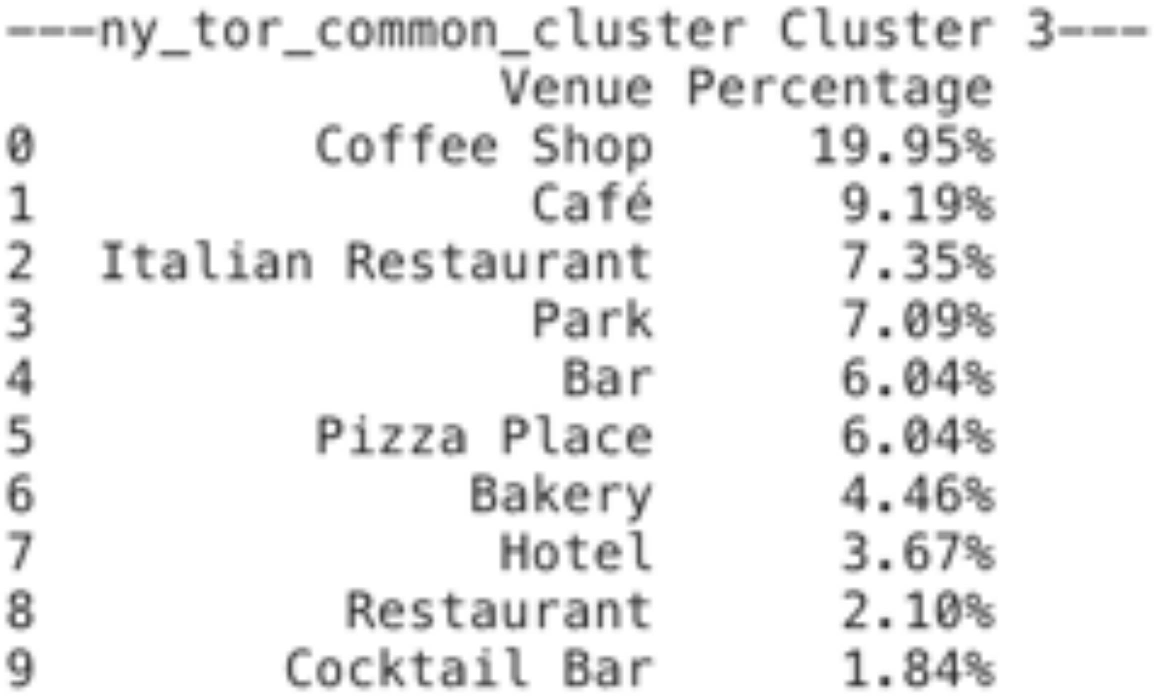


City	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
TOR	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944	3	Coffee Shop	Pizza Place	Park	Italian Restaurant	Bank	Pub	Burger Joint	Sushi Restaurant	French Restaurant	Dessert Shop
TOR	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	3	Coffee Shop	Japanese Restaurant	Men's Store	Park	Café	Sushi Restaurant	Gay Bar	Restaurant	Hotel	Bookstore
TOR	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558	3	Park	Coffee Shop	Brewery	Pizza Place	Bakery	Sushi Restaurant	Fast Food Restaurant	Italian Restaurant	Farmers Market	Beach
TOR	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509	8	Park	Ice Cream Shop	Gym / Fitness Center	Italian Restaurant	Shopping Mall	Bus Stop	Eastern European Restaurant	Filipino Restaurant	Farm	Farmers Market
TOR	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	3	Restaurant	Gym / Fitness Center	Burrito Place	Convenience Store	Burger Joint	Bank	Bakery	Grocery Store	Sushi Restaurant	Gym

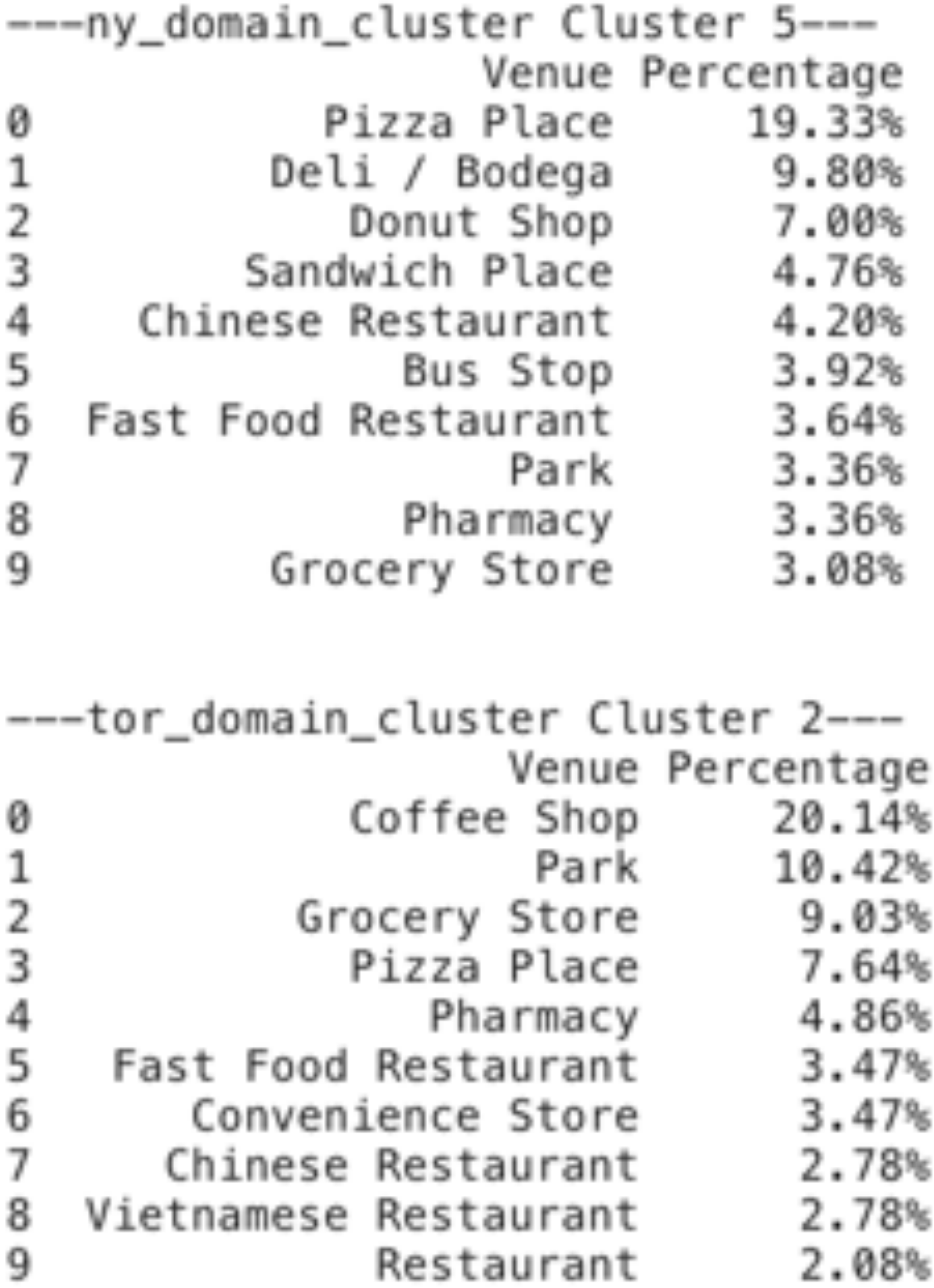



Markham
Vaughan
Rouge National Urban Park
427
Toronto
Billy Bishop Toronto City Airport
Mississauga





```
---ny_tor_common_cluster Cluster 3---
                  Venue Percentage
0           Coffee Shop     19.95%
1                  Café      9.19%
2   Italian Restaurant      7.35%
3                  Park      7.09%
4                   Bar      6.04%
5           Pizza Place      6.04%
6                Bakery      4.46%
7                 Hotel      3.67%
8            Restaurant      2.10%
9          Cocktail Bar      1.84%
```



```
---ny_domain_cluster Cluster 5---
                     Venue Percentage
0              Pizza Place     19.33%
1            Deli / Bodega      9.80%
2               Donut Shop      7.00%
3           Sandwich Place      4.76%
4       Chinese Restaurant      4.20%
5                 Bus Stop      3.92%
6     Fast Food Restaurant      3.64%
7                     Park      3.36%
8                 Pharmacy      3.36%
9            Grocery Store      3.08%
```

```
---tor_domain_cluster Cluster 2---
                      Venue Percentage
0               Coffee Shop     20.14%
1                      Park     10.42%
2             Grocery Store      9.03%
3               Pizza Place      7.64%
4                  Pharmacy      4.86%
5      Fast Food Restaurant      3.47%
6         Convenience Store      3.47%
7        Chinese Restaurant      2.78%
8     Vietnamese Restaurant      2.78%
9                Restaurant      2.08%
```

D. Discussion

The environment of borough is explored by analyzing the venue combination. The borough neighborhood data including the coordination and neighborhoods is from wikipedia or Couresera by web crawling or from open data. In our dataset, there are 5 boroughs and 306 neighborhoods in New York; there are 10 boroughs and 103 neighborhoods in Toronto. Moreover, The foursquare API is used to find the venues and their catgories to build the environment features. There are 471 uniques categories in New York, and 334 uniques categories in Toronto. After intersecting these two category set, there are 300 common categories between New York and Toronto. In addition, the one hot encode method is to build the frequency map of venue categories. The frequency map can be clustered by k-means and 10 clusters are generated. The top 10 venues are appended to the borough neighborhood information. At the meantime, top 3 venues are counted to further analyze. In the bar chart, the Cluster 5 is dominate on New York, the Cluster 2 is dominate on Toronto, and the Cluster 3 is both common on New York and Toronto. The top 1 coverage of venues in Cluster 3 is Coffee shop(almost 30%), these places in Cluster 3 may be in business quarter to let people to get coffee easily. Although the top 1 coverage of venues is Coffee shop(20%) in Cluster 2, the second one is Park, 10%. We can observe there are more green places in Toronto. However, The top 1 coverage of venues in Cluster 5 is Pizza Place(almost 20%). These results demonstrate the different types of these places including the culture and life style.

E. Conclusion

Immigration or long-stay in other places are important issues. The environment analysis in this report provides a point of view to select the destination. By analyzing the neighborhood combination of targeted boroughs, the similar location with the place of departure could be choosed. This way can reduce the pressure of changing. In this report, if someone want to transfer from New York to Toronto or from Toronto to New York, they can use this map to select the targeted place they can live. Not only changing the living country but finding the familiar place to live would benifit the person or the family.