IBM DS Professional Certificate

Capstone Project

Xuan (Jacqueline) Guo

Business Idea – Restaurant Location Pickup

- This capstone project is focus on restaurant location pickup by using location data from Wikipedia website and venues data from Foursquare API.
- A good location has the direct relationship with the restaurant business profit.
- This project using machine leaning method such as k-means clustering to find out the best location for restaurant business in Toronto.

Data Acquisition and Cleaning

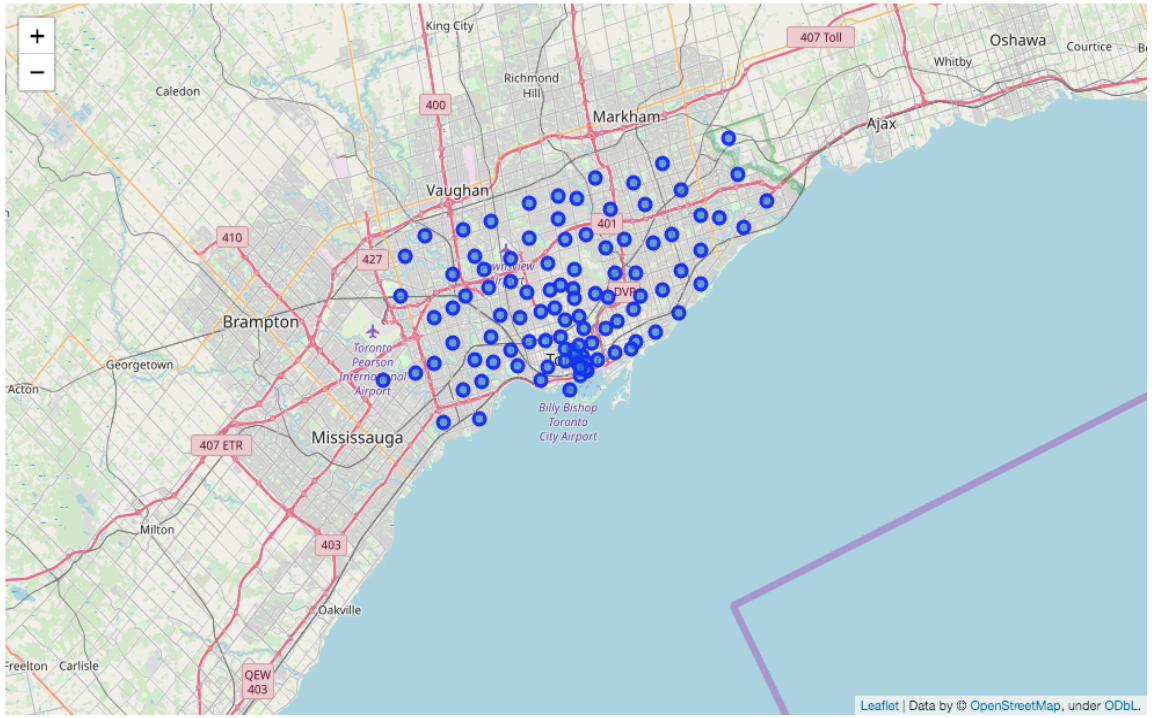
- Using BeautifulSoup to scrape Toronto neighborhood data from Wikipedia website
- Using Python Pandas package to resemble data to pandas data frame and do data cleaning
- Using Geopy package to obtain the latitude and longitude information
- Using one hot encoding to encode categorical data

Cleaned Table

This table contains location feature of Toronto such as postcode, borough, neighborhood, latitude and longitude.

	Postcode	Borough	Neighbourhood	Latitude	Longitude
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
9	M5B	Downtown Toronto	Ryerson, Garden District	43.657162	-79.378937
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
19	M4E	East Toronto	The Beaches	43.676357	-79.293031
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306

Location Point Map



Using Python Folium package to map the location point.

Foursquare API Venues

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Berczy Park	57	57	57	57	57	57
Business Reply Mail Processing Centre 969 Eastern	19	19	19	19	19	19
Central Bay Street	84	84	84	84	84	84
Chinatown, Grange Park, Kensington Market	100	100	100	100	100	100
Christie	15	15	15	15	15	15

Using Foursquare API obtain neighborhood venues. This is counted venues information.

Each neighborhood along with the top 5 most common venues can be printed as following:

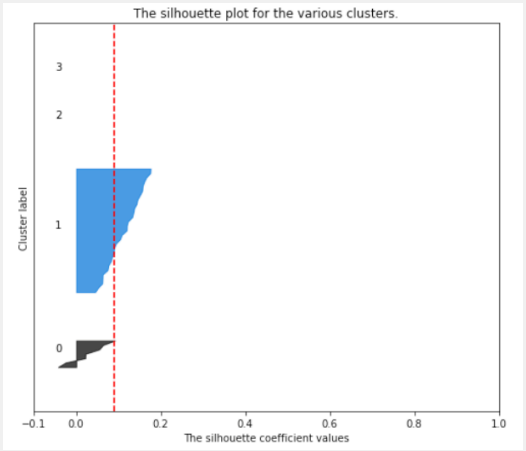
```
----Berczy Park----
              venue  freq
0      Coffee Shop  0.09
1     Cocktail Bar  0.05
2  Farmers Market  0.04
3       Steakhouse  0.04
4             Café  0.04
```

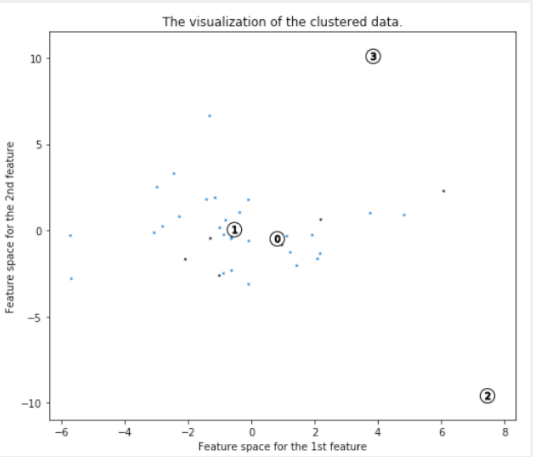
```
----Central Bay Street----
                 venue  freq
0          Coffee Shop  0.14
1  Italian Restaurant  0.05
2       Ice Cream Shop  0.05
3       Sandwich Place  0.04
4         Burger Joint  0.04
```

```
----Christie----
              venue  freq
0             Café  0.20
1  Grocery Store  0.20
2             Park  0.13
3            Diner  0.07
4       Baby Store  0.07
```

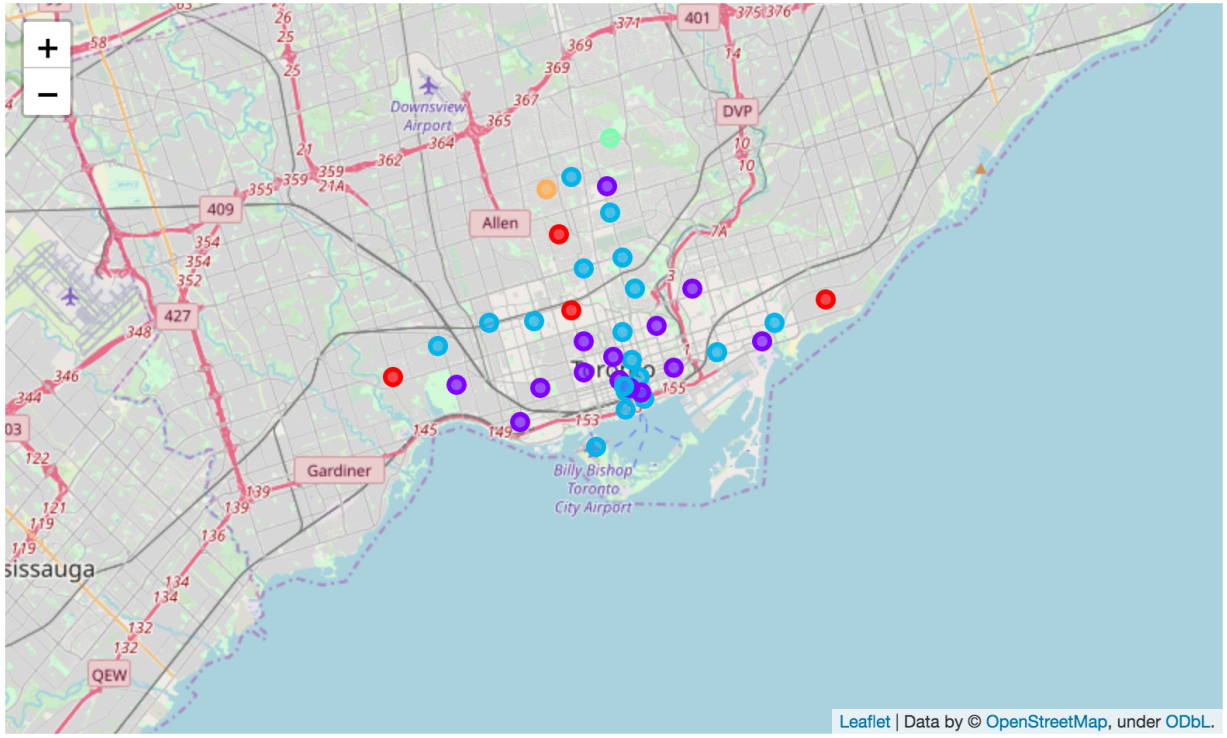
Dimension Reduction for Better K-means

After got the most common from Foursquare API, the next step is to apply k-means clustering model. Since k-means is a search strategy to minimize the squared Euclidean distance, and Euclidean distance is not a good metric in high dimensions. Therefore, applying PCA and Tsne to reduce data dimension.





Clustering



Appling k-means clustering and showing it on map as the figure shown:

Result – Clustering 1

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
19	East Toronto	0	Health Food Store	Trail	Pub
68	Central Toronto	0	Jewelry Store	Sushi Restaurant	Park
74	Central Toronto	0	Coffee Shop	Sandwich Place	Café
81	West Toronto	0	Coffee Shop	Café	Sushi Restaurant

Result – Clustering 2

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
2	Downtown Toronto	1	Coffee Shop	Park	Pub
24	Downtown Toronto	1	Coffee Shop	Ice Cream Shop	Italian Restaurant
30	Downtown Toronto	1	Coffee Shop	Café	Bar
37	West Toronto	1	Bar	Coffee Shop	Asian Restaurant
41	East Toronto	1	Greek Restaurant	Coffee Shop	Italian Restaurant
43	West Toronto	1	Breakfast Spot	Café	Coffee Shop
48	Downtown Toronto	1	Coffee Shop	Café	Hotel
67	Central Toronto	1	Breakfast Spot	Hotel	Sandwich Place
75	West Toronto	1	Breakfast Spot	Gift Shop	Coffee Shop
80	Downtown Toronto	1	Café	Bakery	Bar
84	Downtown Toronto	1	Café	Vegetarian / Vegan Restaurant	Chinese Restaurant
92	Downtown Toronto	1	Coffee Shop	Restaurant	Café
96	Downtown Toronto	1	Park	Coffee Shop	Café
100	East Toronto	1	Light Rail Station	Yoga Studio	Auto Workshop

Result – Clustering 3

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
9	Downtown Toronto	2	Coffee Shop	Clothing Store	Cosmetics Shop
15	Downtown Toronto	2	Hotel	Italian Restaurant	Restaurant
20	Downtown Toronto	2	Coffee Shop	Cocktail Bar	Cheese Shop
25	Downtown Toronto	2	Café	Grocery Store	Park
31	West Toronto	2	Bakery	Pharmacy	Supermarket
36	Downtown Toronto	2	Coffee Shop	Hotel	Aquarium
42	Downtown Toronto	2	Coffee Shop	Café	Hotel
47	East Toronto	2	Pizza Place	Fast Food Restaurant	Italian Restaurant
54	East Toronto	2	Café	Coffee Shop	Bakery
69	West Toronto	2	Café	Mexican Restaurant	Bookstore
73	Central Toronto	2	Sporting Goods Shop	Coffee Shop	Clothing Store
79	Central Toronto	2	Sandwich Place	Dessert Shop	Pizza Place
83	Central Toronto	2	Playground	Gym	Restaurant
86	Central Toronto	2	Coffee Shop	Pub	Pizza Place
87	Downtown Toronto	2	Airport Service	Airport Terminal	Airport Lounge
91	Downtown Toronto	2	Park	Playground	Trail
97	Downtown Toronto	2	Coffee Shop	Café	Hotel
99	Downtown Toronto	2	Coffee Shop	Japanese Restaurant	Sushi Restaurant

Result – Clustering 4

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
61	Central Toronto	3	Park	Bus Line	Dim Sum Restaurant

Result – Clustering 5

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
62	Central Toronto	4	Garden	Women's Store	Flea Market

Result Discussion

The clusters cluster locations with similar factors such as venues and geo locations. This could be a good guide for venue owner to choose opening places. For example, if you want to open some venue at places people may like to go to but not places that already full of same kind of venue, you could use this clustering.

Conclusion

According to the result, I found for starting a new restaurant: First we have to choose a location which did not contain lots of restaurant. Then the restaurant category should be noticed by avoiding duplicate restaurant. Moreover, we can start new restaurant business by looking up similar nearby environment such as same clustering information.
Thus, k-means cluster modeling can be used very widely when facing such problem.


THANK YOU

Reference

1. List of Postal Codes of Canada: M.” Wikipedia, Wikimedia Foundation, 15 July 2019, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
2. H.HH.H 1111 silver badge22 bronze badges, et al. “PCA before K-Mean Clustering.” Data Science Stack Exchange, 1 Aug. 1967, https://datascience.stackexchange.com/questions/17216/pca-before-k-mean-clustering.
3. Mathieu, et al. “How Do I Know My k-Means Clustering Algorithm Is Suffering from the Curse of Dimensionality?” *Cross Validated*, 1 Feb. 1967, https://stats.stackexchange.com/questions/232500/how-do-i-know-my-k-means-clustering-algorithm-is-suffering-from-the-curse-of-dim.