IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

CAPSTONE PROJECT REPORT

Jacqueline (Xuan) Guo

TABLE OF CONTENTS

Contents

IBM DS CAPSTONE PROJECT

Introduction

IDEA

When people want to start a new business in a city like a restaurant, it always a good idea to learn the details about distribution of the city which they want to start their business. The detail distribution which means according to different district or neighborhood such as what kind of restaurant located more or less, what is one particular neighborhood generally used for, what kind of people living here, and what kind of service they need most. By knowing these detail information, it would be easier for merchant to decide the location for their restaurant business.

FEATURES IMPORTANT FOR RESTAURANT LOCATION PICKUP

- Restaurant distribution: there's lots of different restaurant categories such as Chinese restaurant, fast food, Italian restaurant, and Mexican food. Avoiding duplicate food categories when there was already existing plentiful in a same neighborhood would reduce a fierce competition.
- Safety: choose a safety neighborhood is also very important for restaurant business. Since people willing to pick a safe place to hang out.
- Parking: Since some city like Los Angeles has limited parking lot which lead to very expensive parking fees. It will lead to an increased cost of both customers and business owners.
- Surrounding environment: surrounding environment can be considered directly related to customer amount. For example, a shopping mall around may lead to more customers.

These features have a very close connection to restaurant profits. There are still a lot more features need to be considered when start a new restaurant business. However, due to the limitation of data set, we only analyze several features that could be showed by our clustering method.

Data

The dataset for this capstone project consists of neighborhood information of Toronto obtained from Wikipedia website: *List of postal codes of Canada: M*. By using the BeautifulSoup package of Python, transforming the website data to pandas data frame which include postcode, borough, and neighborhood features. Then according to the transformed data frame obtained latitude and longitude by applying geopy package. Then using the latitude and longitude we can collect venues near each neighborhood for cluster analysis by using Foursquare API.

Figure 1 shows the first five rows of website data from Wikipedia website:

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront, Regent Park
3	M6A	North York	Lawrence Heights, Lawrence Manor
4	M7A	Queen's Park	Queen's Park

Figure 1

This data frame only contains index numbers and three features which are postcode, borough, and neighborhood. I used folium package to map the data point on map, which shows in figure 2:

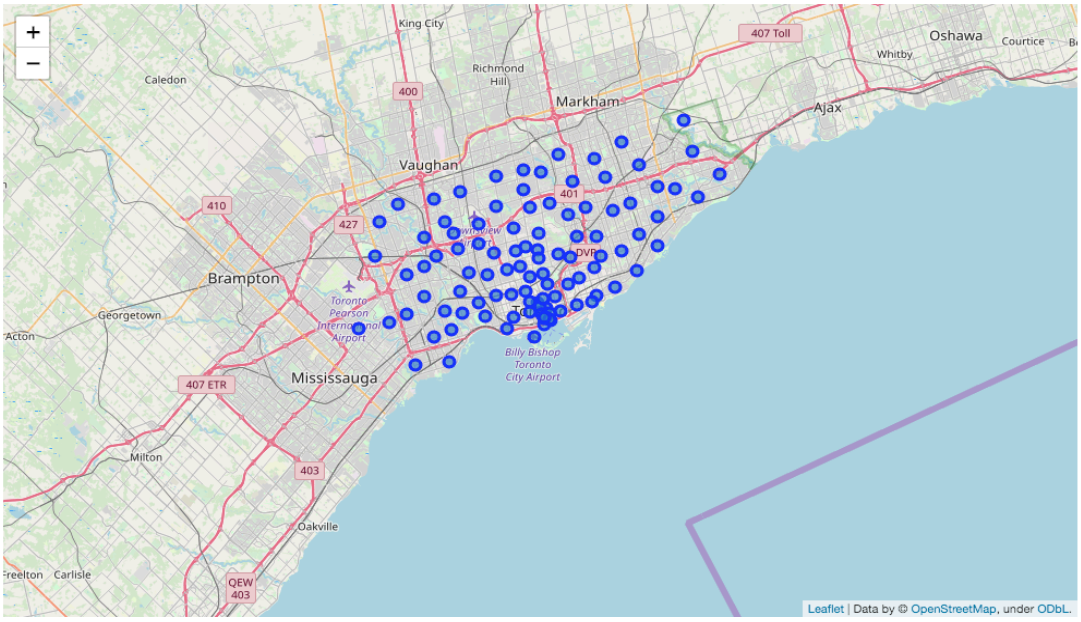

Figure 2

We can see there's boroughs not in Toronto, and the table does not contain latitude and longitude. The next step is to obtain latitude and longitude, then filtering the unnecessary values.

Figure 3 shows the table with latitude and longitude added and only included the borough which contains Toronto:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
9	M5B	Downtown Toronto	Garden District, Ryerson	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

Figure 3

Then with obtained table, we can get the venues from Foursquare, and doing k-means clustering to obtain the most popular venue at each borough.

Methodology

I. Exploratory Data Analysis

Exploratory data analysis was used to analyze data sets to summarize their main characteristics, and visual data sets.

i. First, transforming online data table to Pandas data frame. Using Pandas command to eliminate missing values.

ii. Use Geopy package to obtain the latitude and longitude of Toronto, then mapping table point on Toronto area. Thus, we can visualize data point on map.

II. Segmenting and Slicing, and dimensional reduction

Cluster analysis is the task of grouping a set of objects in an unsupervised way that objects in the same group, and more similar objects will be clustered in a same group. Thus, it's important to know how many different features (here is different venues) in each neighborhood. We use Foursquare API to obtain venues, and there are 72 venues were returned. Then exploring neighborhoods in Toronto. As a result, we convert a Toronto venues data frame in size (1699,7), and the counted information shows by figure 4:

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Adelaide, Richmond, King	100	100	100	100	100	100
Berczy Park	57	57	57	57	57	57
Brockton, Parkdale Village, Exhibition Place	21	21	21	21	21	21
Business Reply Mail Processing Centre 969 Eastern	19	19	19	19	19	19
CN Tower, Bathurst Quay, King and Spadina, Island airport, Railway Lands, Harbourfront West, South Niagara	14	14	14	14	14	14

Figure 4

In order to fit the venues data and neighborhood into the clustering model, we have to group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. Figure 5 shows the result:

	Neighborhood	Yoga Studio	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium
0	Adelaide, Richmond, King	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.030000	0.0	0.0
1	Berczy Park	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	Business Reply Mail Processing Centre 969 Eastern	0.052632	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	Cabbagetown, St. James Town	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Central Bay Street	0.011905	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.011905	0.0	0.0

Figure 5

III. Use both Tsne and PCA to reduce data dimension

For this step, since k-means is a search strategy to minimize the squared Euclidean distance, and Euclidean distance is not a good metric in high dimensions [III]. Thus, I decided to reduce the dimension of each feature and apply a better k-means clustering.

Therefore, I chose PCA and Tsne to reduce data dimension. PCA reduces dimensionality, but it does not change the number of observations and order of the data. Also, for PCA, I drop all components with their eigenvalue less than 1 according to Kaiser criterion and used the remaining components as the reduced dimension. Figure 6 shows part of results for reduction:

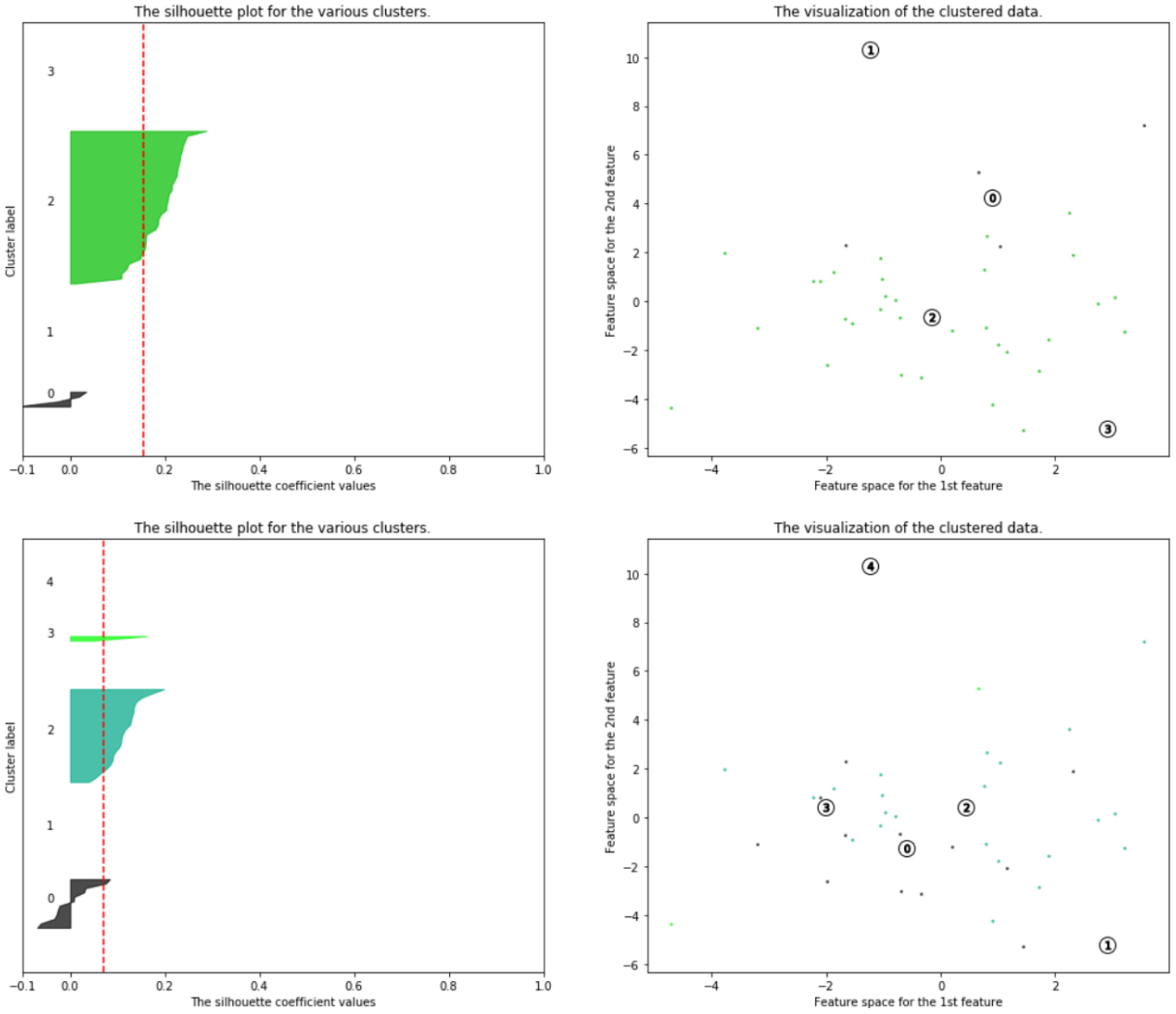

Figure 6

IV. Neighborhood Exploration and Clustering

After dimension reduction process, we set number of clusters equal to 5, and run k-means to cluster the neighborhood into 5 clusters. Figure 7 shows the part of result:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th M Comn Ve
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	1	Coffee Shop	Park	Pub	Bakery	C
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	2	Coffee Shop	Clothing Store	Cosmetics Shop	Middle Eastern Restaurant	C
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	2	Hotel	Italian Restaurant	Restaurant	Café	Co Sh
19	M4E	East Toronto	The Beaches	43.676357	-79.293031	0	Health Food Store	Trail	Pub	Coffee Shop	Farm Mar
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	2	Coffee Shop	Cocktail Bar	Cheese Shop	Bakery	Steakho

Figure 7

V. Cluster of Neighborhoods in Toronto

Then visualize the result in figure 8:

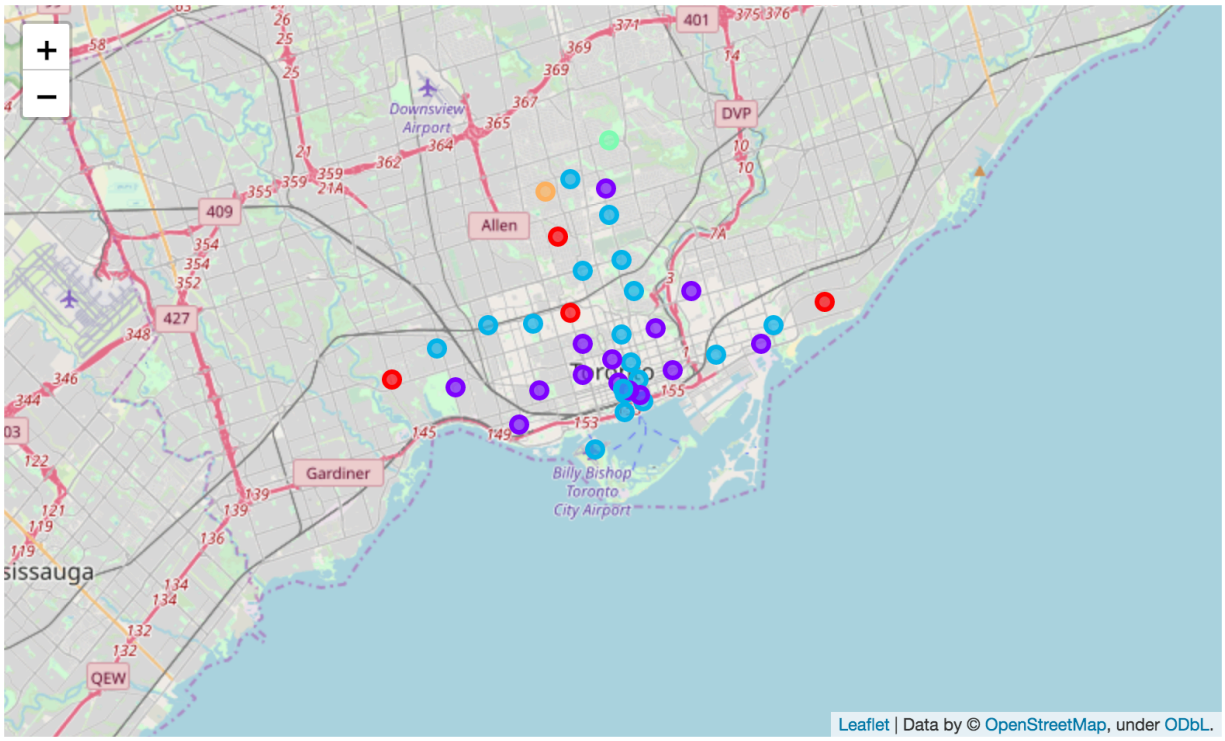

Figure 8

Result

After applying clustering method, I printed each clustering result to 5 tables shown as following:

CLUSTERING 1

Figure 9 shows part of the first clustering table which contains only top three most common venues:

	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

Figure 9

CLUSTERING 2

Figure 10 shows part of the second clustering table which contains only top three most common venues:

	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

Figure 10

CLUSTERING 3

Figure 11 shows part of the third clustering table which contains only top three most common venues:

	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

Figure 11

CLUSTERING 4

Figure 12 shows part of the fourth clustering table which contains only top three most common venues:

	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

Figure 12

CLUSTERING 5

Figure 13 shows part of the fifth clustering table which contains only top three most common venues:

	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

Figure 13

According to these five clustering results, the most popular venue in downtown Toronto is coffee shop. Half percent of east Toronto's most common venues is food supply. 2/9 of central Toronto's most common venues is food supply. ½ of west Toronto's most common venue is food supply.

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. More specifically, for cluster #1, places are clustered together because they are geologically close and have similar distribution of venues, and sushi restaurant are most common venues in place #2 and #4. Thus, place #1 and #3 are great places to open sushi restaurant because they are similar based on geo location and venue distribution but there aren't so many sushi restaurants there yet and therefore a newly opened sushi restaurant may be popular there.

Conclusion

This capstone project focus on using cluster modeling to fit the location data and find out the most popular venues around each location. Then using the clustered information make business decisions such as which location is the best location to start a specific restaurant business.

By using Wikipedia website data obtain the location information in Toronto and using Foursquare API to obtains venues information. Transforming the data to pandas data frame, then applying the cleaned data to fit the k-means clustering model. I obtained top 10 most common venues for each cluster as shown above.

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.

Reference

I. 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.

II. 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.

III. 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.