IBM Applied Data Science Capstone Project -- Final Report

Daisy Zhu

April, 2020

Clustering and segmenting neighborhood in Toronto

Table of content:

Daisy Z.

I. Introduction

This report is a part of IBM applied data science capstone project, a 9 course series for data science learning created by IBM on Coursera platform. The overall background and data source are given from the course whereas the rest of analysis and approach are left for course students to explore and develop. Through the project, students need to access data from website using scraping and to visualize data on Foursquare API. Furthermore, students can further explore data by using various Python packages and machine learning techniques to draw conclusions. Before conducting analysis and modeling, data will be collected, wrangled and cleaned. Students will find best data format and features for machine learning and modeling phase. And then, students need to find out the best fit model through trying different algorithms and tuning. Along with the whole process, students would visualize data to help understand and improve the modeling process.

The original idea for this project is that someone would like to open a new restaurant in the city Toronto. He/She is wondering where the location should be. The objective of the project is to explore neighborhood in Toronto and give a sound suggestion.

The audience for this report are:

- ✓ Potential investors who plan to run business
- ✓ Potential real estate buyers
- ✓ Potential real estate renters
- ✓ Course peers and instructors

II. Data description

The dataset for this project mainly comes from two parts:

- ✓ The Foursquare API: geographical data with related information will be access via Python scraping to get most venues for each neighborhood in the city of Toronto. By doing so, we can visualize geographical data on the map and clearly see how venues are distributed in neighbors.

- ✓ Other relational data: a csv file containing geospatial coordinate data. We can join coordinate data with dataset above.

You can search and look at data at
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

III. Methodology

The project will conduct several following methodology:

✓ Exploratory data analysis to see frequency of venues by neighbor

```
----Adelaide, King, Richmond----
         venue  freq
0  Coffee Shop  0.06
1   Restaurant  0.06
2         Café  0.05


----Agincourt----
                      venue  freq
0             Clothing Store  0.25
1  Latin American Restaurant  0.25
2                     Lounge  0.25


----Agincourt North, L'Amoreaux East, Milliken, Steeles East----
         venue  freq
0         Park   0.5
1   Playground   0.5
2  Yoga Studio   0.0
```

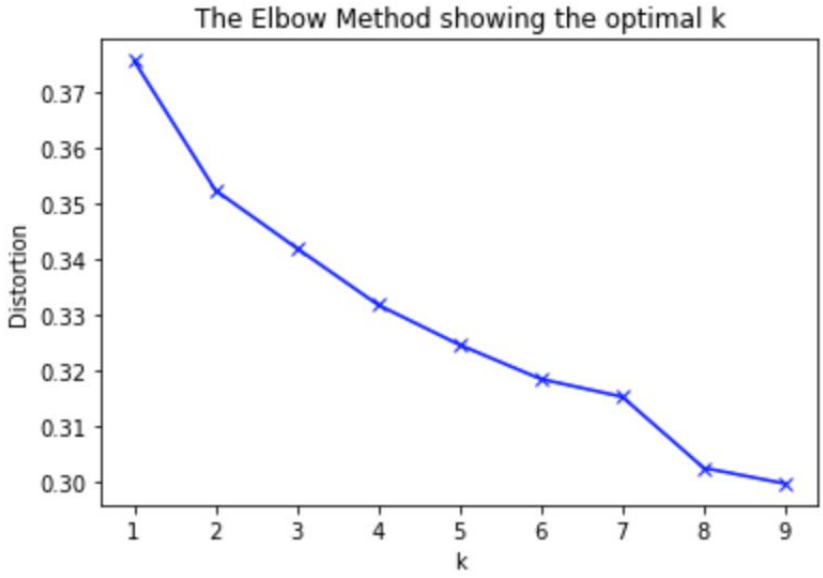
Figure 1 frequency of venues by neighborhood

✓ Feature Engineering - one-hot coding

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	...	Trail	Train Station
0	Adelaide, King, Richmond	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
4	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

Figure 2 Result of one-hot coding

✓ Elbow method to determine the best K for the K-means clustering



When K increases, the centroids are closer to the clusters centroids.

The improvements will decline, at some point rapidly, creating the elbow shape.

That point is the optimal value for K. In the image above, K=8

Figure 3 Result of Elbow method

✓ K-means clustering

To explore features of different neighbors, we are going to deploy clustering techniques, particularly K-means algorithm. For segmentation and clustering purpose, K-means is a simple and quick way to do so. And we have created 15 features, including 10 most common venues, postcode, geospatial coordinates, borough etc.

IV. Modeling

As EDA shown above, before conducting K-means clustering, we have to decide the best k for the number of clusters. As the result, when k equals to 8, it created the elbow shape according to the rule.

Then we run the K-means algorithm by importing sklearn.cluster package, and we get the result table with clusters:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Clusters	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5t Co
0	M3A	North York	Parkwoods	43.753259	-79.329656	6	Park	Bus Stop	Food & Drink Shop	Distribution Center	Desse
1	M4A	North York	Victoria Village	43.725882	-79.315572	0	Intersection	Coffee Shop	Pizza Place	Hockey Arena	Port Res
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636	4	Coffee Shop	Park	Pub	Bakery	
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763	4	Clothing Store	Furniture / Home Store	Women's Store	Coffee Shop	Miscell
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494	4	Coffee Shop	Yoga Studio	Distribution Center	Mexican Restaurant	

Figure 4 Result of K-means clustering

And the distribution of the clusters:

```
Clusters
0      8
1      1
2      1
3      1
4     74
5      1
6     11
7      1
Name: Neighbourhood, dtype: int64
```

Figure 5 Distribution of clusters

We can see from tables above that clusters being 0, 4 and 6 have number of venues much more than others. Then we begin to explore three clusters deeply to see which cluster of neighbor are we going to choose. We checked three clusters and display the result respectively:

	Borough	Clusters	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	North York	0	Intersection	Coffee Shop	Pizza Place	Hockey Arena	Portuguese Restaurant
8	East York	0	Pizza Place	Fast Food Restaurant	Gastropub	Café	Athletics & Sports
10	North York	0	Park	Pub	Pizza Place	Japanese Restaurant	Distribution Center
50	North York	0	Empanada Restaurant	Pizza Place	Dog Run	Department Store	Dessert Shop
63	York	0	Pizza Place	Bus Line	Caribbean Restaurant	Brewery	Women's Store
70	Etobicoke	0	Pizza Place	Middle Eastern Restaurant	Chinese Restaurant	Coffee Shop	Discount Store
77	Etobicoke	0	Park	Bus Line	Pizza Place	Sandwich Place	Discount Store
93	Etobicoke	0	Pizza Place	Coffee Shop	Pharmacy	Sandwich Place	Skating Rink

Figure 6 Cluster 0

	Borough	Clusters	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Downtown Toronto	4	Coffee Shop	Park	Pub	Bakery	Theater
3	North York	4	Clothing Store	Furniture / Home Store	Women's Store	Coffee Shop	Miscellaneous Shop
4	Downtown Toronto	4	Coffee Shop	Yoga Studio	Distribution Center	Mexican Restaurant	Bank
7	North York	4	Café	Baseball Field	Gym / Fitness Center	Caribbean Restaurant	Japanese Restaurant
9	Downtown Toronto	4	Clothing Store	Coffee Shop	Café	Cosmetics Shop	Japanese Restaurant
13	North York	4	Beer Store	Gym	Restaurant	Coffee Shop	Italian Restaurant
14	East York	4	Skating Rink	Dance Studio	Spa	Diner	Curling Ice

Figure 7 Cluster 4

	Borough	Clusters	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	North York	6	Park	Bus Stop	Food & Drink Shop	Distribution Center	Dessert Shop
21	York	6	Park	Market	Women's Store	Gluten-free Restaurant	Gift Shop
35	East York	6	Park	Convenience Store	Coffee Shop	Dessert Shop	Dim Sum Restaurant
40	North York	6	Park	Airport	Doner Restaurant	Dessert Shop	Dim Sum Restaurant
49	North York	6	Park	Bakery	Construction & Landscaping	Doner Restaurant	Dim Sum Restaurant
64	York	6	Park	Convenience Store	Empanada Restaurant	Electronics Store	Eastern European Restaurant
66	North York	6	Park	Bank	Convenience Store	Bar	Women's Store
85	Scarborough	6	Park	Playground	Doner Restaurant	Dessert Shop	Dim Sum Restaurant
91	Downtown Toronto	6	Park	Playground	Trail	Eastern European Restaurant	Dumpling Restaurant

Figure 8 Cluster 6

By interpreting results of each cluster, we summarized features and named each cluster

Cluster 0 -- fast food, coffee shop, restaurant etc

Cluster 4 -- convenient store, coffee shop, area that suitable for women

Cluster 6 -- park area, transportation area, and area that suitable for sports people

V. Conclusion

As the result, we decide to put the location in the area of cluster 0, where lots of restaurant, coffee shop around here, which means that flow of people are huge around here.

However, other cluster having only 1 venue does not mean in reality there is only one venue located around that area. We have a very bias result here.

Furthermore, when choosing location of a new venue, especially restaurant, we not only consider the number of people around the area, but also their power of consumption, which means their income level. If we open three Michelin star restaurant surrounding by fast food restaurant, we can simply conclude that people who eat fast food daily wouldn't have buying power of dining at high-end restaurant. So for further exploration, we can add income level of local people, marital status, and even more stats about people so that we can draw more precise and sound conclusion.

References：

- ✓ Wiki website
- ✓ Coursera course website
- ✓ Foursquare API