Battle of Neighbourhoods

ALEX HONCHARUK
MAY 11, 2020

1. Introduction.

1.1. Background

«Co-ffee» - is an old coffee-shops brand, that have open branches all over the world. They specialise on small coffee-points near places of sightseeing and rest. For some time they were building a project of opening another branch in Toronto. It was decided to use modern data science methods to determine appropriate place for new restaurant.

1.2. Problem.

«Co-ffee» is looking for place to open new branch in Toronto. Toronto is perspective place, as it has a lot of parks and few sights, and also has good city transport system. It would be perfect city to open small coffee-shop.

1.3. Interest.

«Co-ffee» chose few main criteria to analyse perspective neighbourhoods to place new branch:

1. *Number of restaurants nearby.* Company is interested to place their branch in free of other restaurants neighbourhood.
2. *Availability of city sights.* Company is sure that this would attract more customers.
3. *City transport nearby.* This would allow to reach restaurants easier.
4. *Access to apartments.* Tourist accommodation places will grant the flow of some new customers.

2. Data acquisition and cleaning.

2.1. Data acquisition.

For the purpose of research were used 3 main datasources:

1. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M - database of Toronto neighbourhoods by postal code.
2. https://cocl.us/Geospatial_data - database with all geospatial data by postal codes.
3. https://api.foursquare.com/ - foursquare API with data regarding all venues in Toronto.

2.2. Data cleaning.

This three datasets were cleaned and concatenated into one simple database into next steps:

1. Data from Wikipedia was organised into database. All missing information regarding neighbourhoods was removed or replaced with proper values. After all manipulations I had near 100 neighbourhoods with postal codes.
2. Geospatial data was appended to original database. This would help to identify venues relation to any neighbourhood.
3. Having the list of neighbourhoods with geospatial data, I was able to load venues for each of it. I connected to Foursquare API and sent request for each neighbourhood.
4. After data was loaded, I extracted needed information into database.

5. There were a lot of duplications in data. I removed all duplicated venues and all without postal codes. I wouldn't be able to match venue with neighbourhood without it.

2.3. Feature selection.

There is a list of criteria, provided by «Co-ffee», so I used it as a guide. Venues database had information regarding the category of the venue. But it was hard to use because of unorganised structure.

There were over 200 different categories in database. This wasn't usable, so I grouped all of them into higher level of grouping. E.g. «Chinese restaurant», «Italian restaurant» were grouped into «Restaurants»; «Park» and «Museum» were grouped into «Sights» etc. This gave me opportunity to choose 4 crucial criteria for further research.

After I had a clean view over venues category, it was time to get data of venues by postal code (as the main identifier of neighbourhood and key feature to link data between multiple databases). So, 4 categories, that were chosen as interested («Restaurants», «Sights», «Apartments», «Transport») were separated and binarised. After that I created database of number of venues by category by postal code and appended it to the main Toronto Neighbourhoods database. You can check the appearance of the final table on the figure 1.

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Shops	Restaurants	Sights	Apartments	Transport
0	M3A	North York	Parkwoods	43.753259	-79.329656	1.0	3.0	1.0	2.0	1.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	1.0	1.0	1.0	0.0	0.0
2	M5A	Downtown Toronto	Regent Park / Harbourfront	43.654260	-79.360636	4.0	5.0	1.0	0.0	0.0
3	M6A	North York	Lawrence Manor / Lawrence Heights	43.718518	-79.464763	10.0	1.0	0.0	0.0	0.0
4	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government	43.662301	-79.389494	0.0	0.0	0.0	0.0	0.0

Figure 1. Database of Toronto neighbourhoods with venue number to each.

As you can see on the figure, there are 4 main features, that I researched and one was a bonus - «Shops». Shops nearby also could influence on the popularity of the coffee-shop, so it was included into the model to check this parameter too.

3. Analysing Data.

3.1. Exploring neighbourhoods.

Let's check what we are dealing with. Using geospatial data, I can prepare map of Toronto neighbourhoods (figure 2).

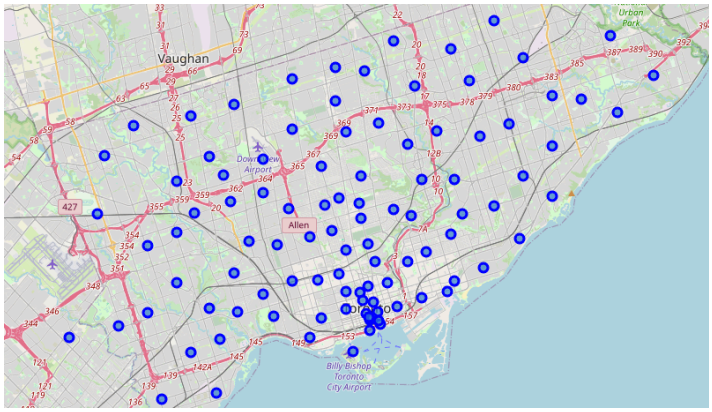

Figure 2. Toronto neighbourhoods map.

As we can see from the map, most of Toronto neighbourhoods are situated in the center. They are spread equally over other regions of the city. Visually, there are no possibilities to group those neighbourhoods on any clusters using only geospatial data.

3.2. Clustering Neighbourhoods.

As we prepared 5 parameters, we can use them to cluster neighbourhoods on more sophisticated groups.

I used k-cluster algorithm with number of clusters equal to 5 (it showed best clustering after several pre-starting experiments). After proceeding, I received next 5 clusters (figure 3):

Cluster	Shops	Restaurants	Sights	Apartments	Transport
0	62.0	49.0	26.0	14.0	10.0
1	24.0	16.0	4.0	22.0	6.0
2	100.0	77.0	8.0	4.0	3.0
3	19.0	60.0	2.0	7.0	0.0
4	104.0	48.0	3.0	1.0	3.0

	Latitude								Longitude		...	Apartments		Transport		
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	s
Cluster																
0	55.0	43.709159	0.051154	43.602414	43.670283	43.706876	43.749332	43.836125	55.0	-79.403113	...	0.0	2.0	55.0	0.181818	0
1	5.0	43.690448	0.051680	43.640816	43.662696	43.665860	43.712751	43.770120	5.0	-79.392730	...	4.0	7.0	5.0	1.200000	1
2	25.0	43.713771	0.057025	43.605647	43.669542	43.709060	43.754328	43.815252	25.0	-79.392546	...	0.0	1.0	25.0	0.120000	0
3	6.0	43.651651	0.008350	43.644771	43.647364	43.649249	43.651263	43.667967	6.0	-79.383716	...	2.5	3.0	6.0	0.000000	0
4	12.0	43.697038	0.050996	43.648429	43.657946	43.674281	43.722209	43.799525	12.0	-79.387997	...	0.0	1.0	12.0	0.250000	0

Figure 3. Toronto neighbourhoods clusters.

Using criteria of «Co-ffee» company, we can choose appropriate cluster for further investigation. As we can see, clusters 1-4 had A LOT of restaurants - more then 1 per neighbourhood. Looks like those are not suitable. Let's check cluster 0. Number of restaurants is 0.9 per neighbourhood - good chance of finding neighbourhood without restaurant. 26 sights is huge! This is very satisfying parameter. 10 transport and 14 apartments only adding confidence to our choice.

3.3. Exploring chosen cluster.

So, now we have cluster and criteria. We can combine those to find perfect places for a restaurant. Let's check if there are any neighbourhoods without restaurants (figure 4):

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Shops	Restaurants	Sights	Apartments	Transport	Cluster
14	M4C	East York	Woodbine Heights	43.695344	-79.318389	2.0	0.0	3.0	0.0	1.0	0
74	M5R	Central Toronto	The Annex / North Midtown / Yorkville	43.672710	-79.405678	0.0	0.0	1.0	1.0	1.0	0
101	M8Y	Etobicoke	Old Mill South / King's Mill Park / Sunnylea /...	43.636258	-79.498509	0.0	0.0	1.0	1.0	1.0	0

Figure 4. Satisfying neighbourhoods.

Great, we have not only one, but three of them! That's always good to have a choice. First even have access to 3 sights, that's more then OK. Each has access to transport. Those are just perfect. Let's explore venues of each neighbourhood (figure 5).

	name	categories_group	postalCode
911	The Annex	Apartments	M5R
1272	2 kinsdale blvd	Apartments	M8Y
185	The best backyard ever!	Shops	M4C
186	The Beer Store	Shops	M4C
684	Tim Hortons	Shops	M3M
687	Country Style	Shops	M3M
688	St.Pio Bakery	Shops	M3M
689	WIND Mobile	Shops	M3M
55	Queen's Park	Sights	M5R
188	Stan Wadlow Park	Sights	M4C
196	Les Anthony Parkette	Sights	M4C
198	Everett Park	Sights	M4C
1268	Park Lawn Park	Sights	M8Y
194	Mortimer Avenue At Woodbine Avenue Bus Stop	Transport	M4C
917	Dupont Subway Station	Transport	M5R
1271	Humber Bay Yatch Cub	Transport	M8Y

Figure 5. Neighbourhood venues.

Well, mosts of sights are parks, but that's can be promising. People, after having good walk on the fresh air, would wish to have a meal. And there are no restaurants nearby any of those. Each neighbourhood has transport station. So This should deal. Lets check results on map (recommended neighbourhoods will be highlighted).

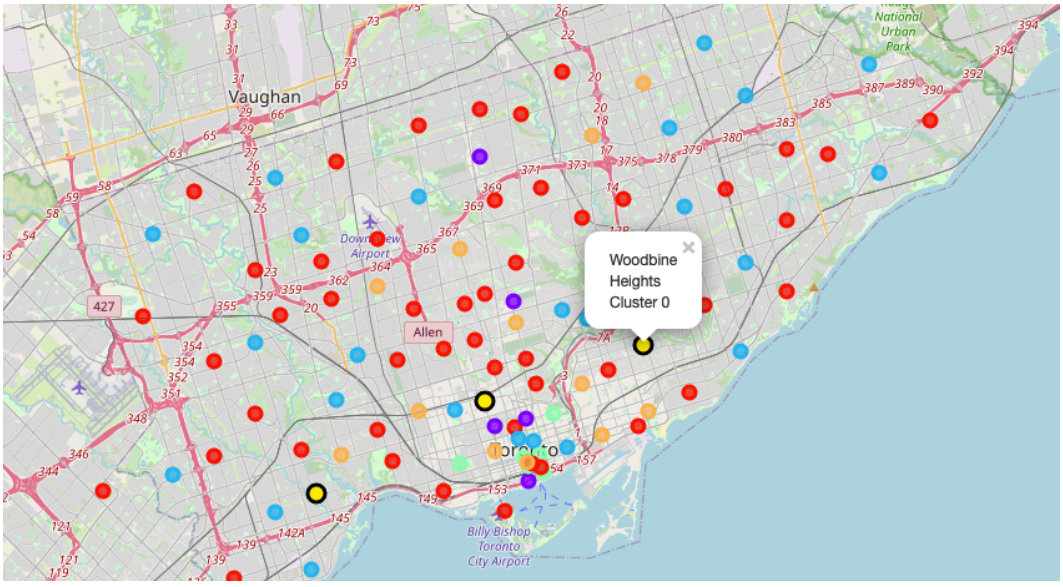

Figure 6. Clustered neighbourhoods (with chosen ones highlighted with gold)

4. Results and discussion

So, I chose few neighbourhoods (some of them are nearby to each other) - Woodbine Hights, The Annex/North Midtown/Yorkville and Old Mill South/King's Mill Park/Sunnylea.

All this neighbourhoods are situated in the center, which means more tourists and visitors. Also, those were neighbourhoods without any restaurant, and with access to transport and a lot of sights (mostly parks).

Looks like those are really promising places to open coffee-shop. Most interesting is the The Annex/North Midtown/Yorkville one, that can have business places nearby, and also would be accessable from Annex Apartments and Queens park.

5. Conclusion

Purpose of this project was to identify good places to open a coffee-shop in Toronto, following some criteria and using Python data analysis methods.

During the research, was built a database, that included information regarding all venues in Toronto, grouped by Neighborhoods. K-cluster analysis helped to choose promissing group of places, which were analyzed farther.

After additional research, 3 candidates were chosen. All those places satisfied criteria.
Using the research material, was built a map with the results.