Applying Data Science Methodology to Discover Business Opportunities in Toronto, Ontario, Canada

Qusay Sellat

Nov. 2019

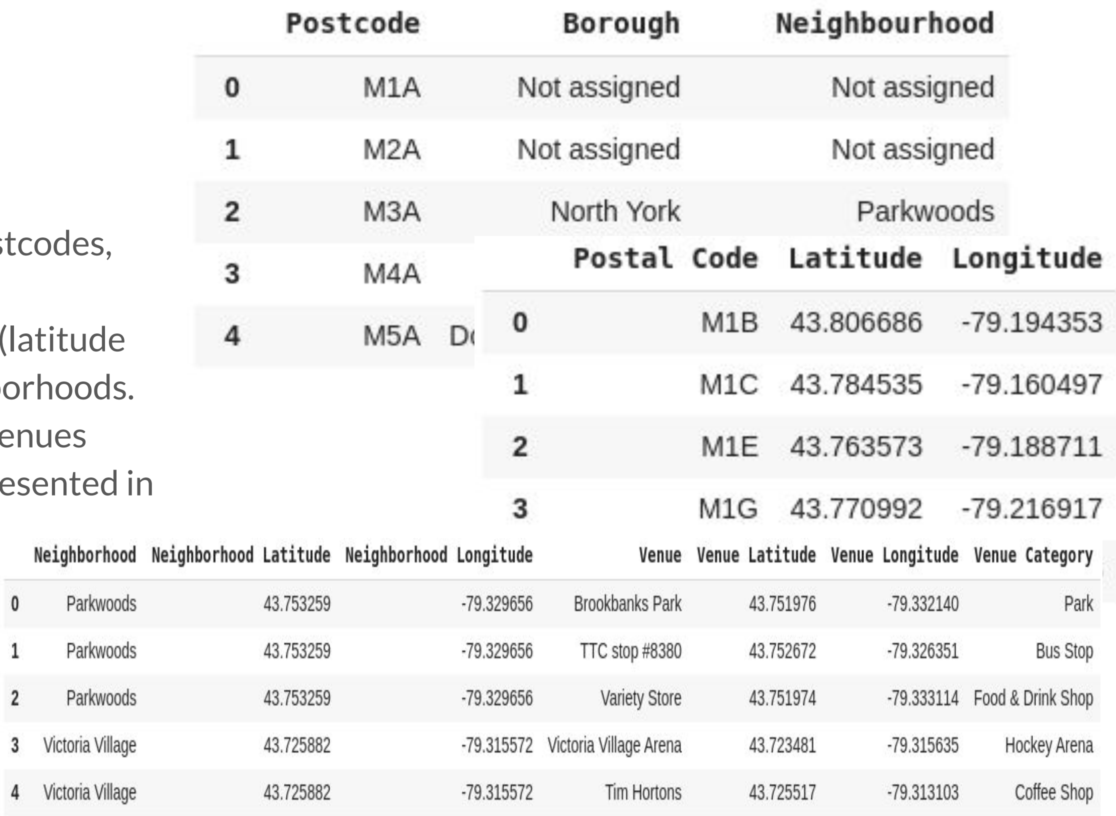
Introduction

- Toronto is the provincial capital of Ontario and the most populous city in Canada.
- Some people outside Toronto have plans to invest in the city but they don’t know how and where. Therefore, this project can be helpful for those who have dreams of having a successful business into the city of Toronto.
- This project analyses the data of Toronto’s various neighborhoods in order to discover the best businesses that can be done in each of the respective neighborhoods by doing a cluster analysis.

Data Collecting

Three initial datasets:

- **Neighborhoods**: Data about Toronto's postcodes, boroughs, and neighborhoods.
- **Coordinates**: Data about the coordinates (latitude and longitude) of various Toronto's neighborhoods.
- **Venues**: Data about the various types of venues located in each of the neighborhoods represented in the data collected in the above steps.



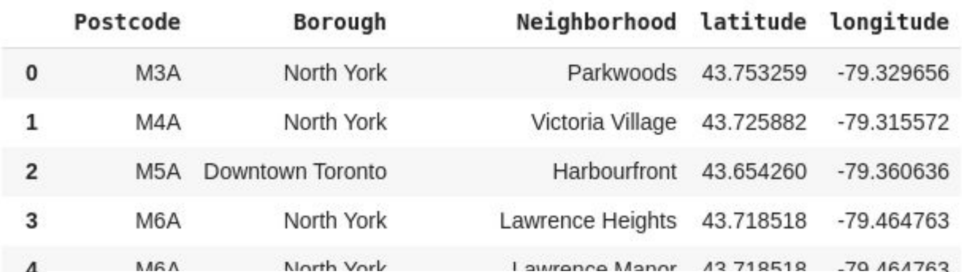
	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A		
4	M5A	D	

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917

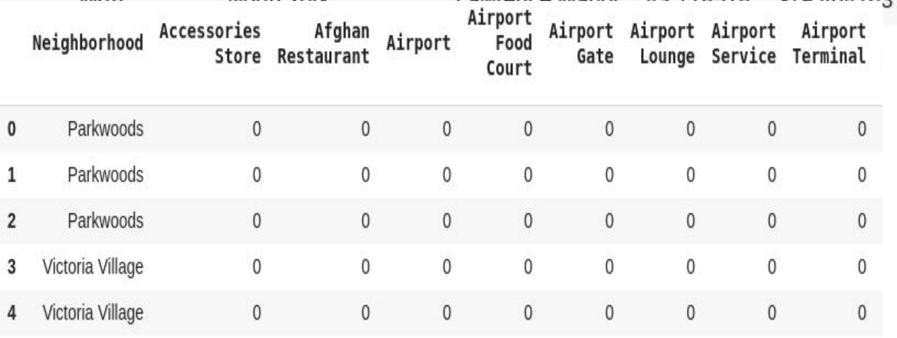
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	TTC stop #8380	43.752672	-79.326351	Bus Stop
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

Data Pre-processing

- For convenience, any row containing Not assigned Borough was dropped. Also, any Not assigned Neighborhood is replaced by the corresponding Borough.
- we prepare Neighborhoods data set to contain latitude and longitude information by using Coordinates data set.
- In order to do the clustering process on the neighborhoods of Toronto, we must first make a dataset that can be fit into a clustering algorithm like KMeans. One approach is to use one-hot encoding to represent each of the returned avenues with the corresponding neighborhoods.



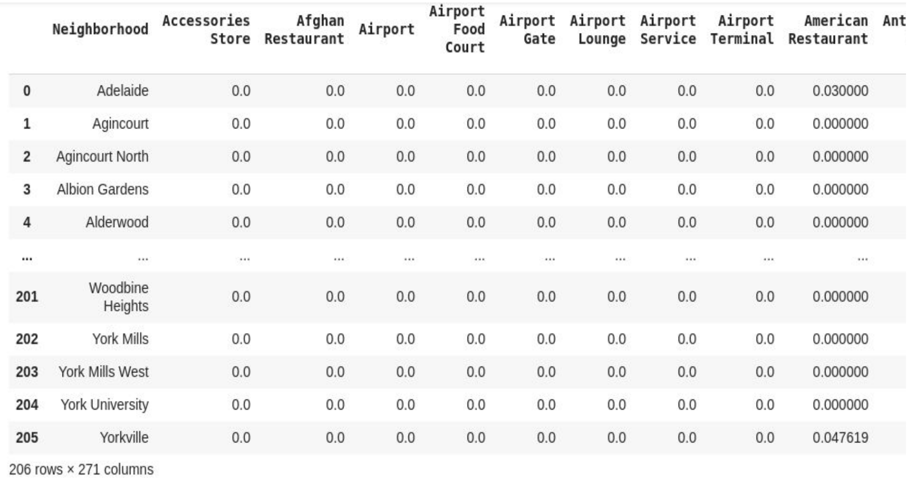
	Postcode	Borough	Neighborhood	latitude	longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights	43.718518	-79.464763
4	M6A	North York	Lawrence Manor	43.718518	-79.464763



	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal
0	Parkwoods	0	0	0	0	0	0	0	0
1	Parkwoods	0	0	0	0	0	0	0	0
2	Parkwoods	0	0	0	0	0	0	0	0
3	Victoria Village	0	0	0	0	0	0	0	0
4	Victoria Village	0	0	0	0	0	0	0	0

Data Pre-processing

- We notice that the number of avenues is 4379 of 270 type. For each neighborhood, in order to fit the K-Means algorithm, it’s important to know what avenues located in each neighborhood. For this reason we group the one-hot encoded dataset of the avenues by the neighborhood by applying the mean function to represent the importance of each avenue in describing the neighborhood. So the final dataset looks like the following, we will call it toronto_grouped dataset (because of sparsity nature of data, most entries are zero), and it will be the input of our clustering algorithm.

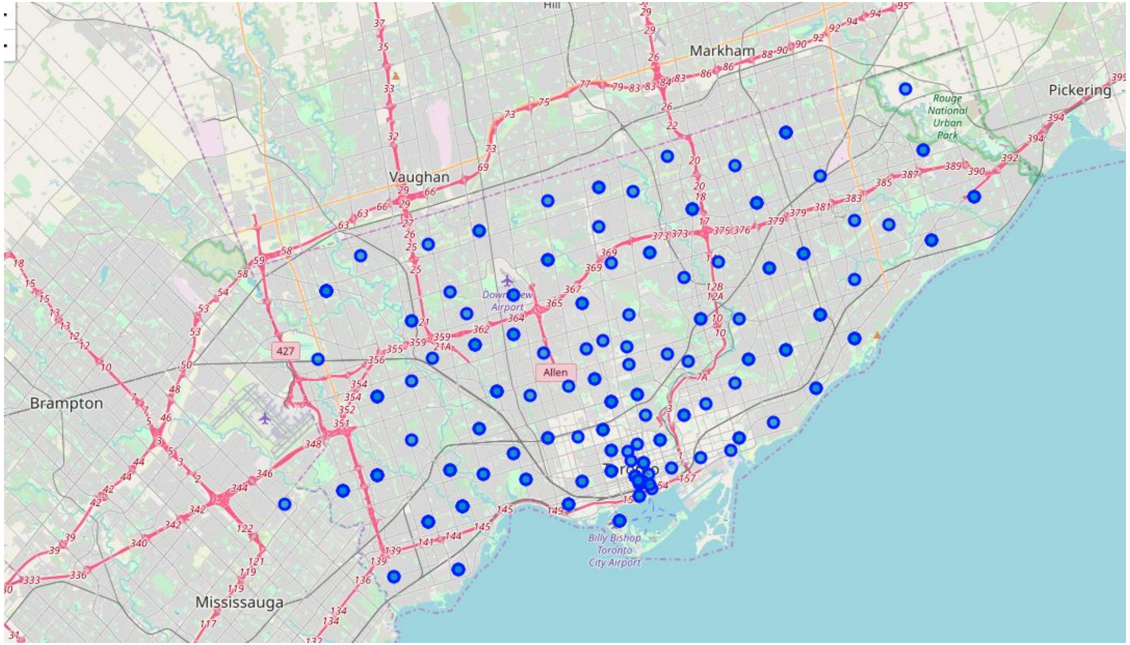


	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Ant
0	Adelaide	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.030000	
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
2	Agincourt North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
3	Albion Gardens	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
4	Alderwood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
...	...	...	...	...	...	...	...	...	...	...	
201	Woodbine Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
202	York Mills	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
203	York Mills West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
204	York University	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
205	Yorkville	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.047619	

206 rows × 271 columns

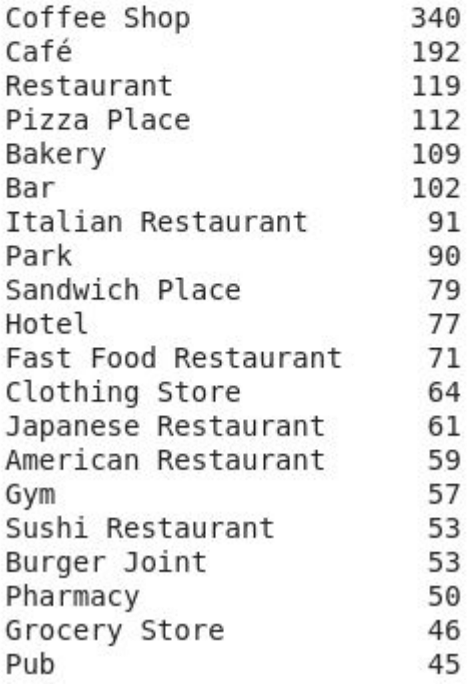
Data Analysis

The 210 neighborhoods of Toronto belong to 11 boroughs and there are 103 distinct postcodes for them.



Data Analysis

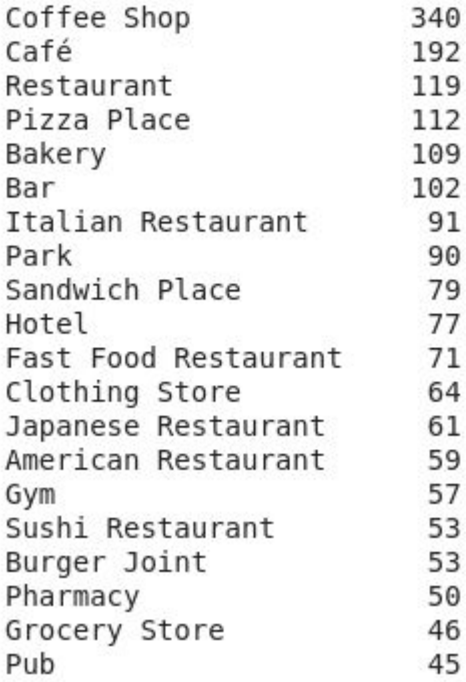
There are 4379 venues of 270 types. Different types of restaurants, playgrounds, bars, shops, transportation stations, .. etc. We also notice that venues belong to 206 neighborhoods - 4 neighborhoods less than we have in neighborhoods dataset. This is due to the fact that FourSquare API didn't retrieve any venues for some neighborhoods. However, the number is very small and will not affect the final result that much and we can continue with the dataset we obtained.



```
Coffee Shop                 340
Café                        192
Restaurant                  119
Pizza Place                 112
Bakery                      109
Bar                         102
Italian Restaurant           91
Park                         90
Sandwich Place               79
Hotel                        77
Fast Food Restaurant         71
Clothing Store               64
Japanese Restaurant          61
American Restaurant          59
Gym                          57
Sushi Restaurant             53
Burger Joint                 53
Pharmacy                     50
Grocery Store                46
Pub                          45
```

Data Analysis

The various kinds of restaurants dominates the venue list. This is a sign that one of our clusters (the cluster which have the neighborhoods with a large number of restaurants like avenues) will dominate other clusters in number of avenues.

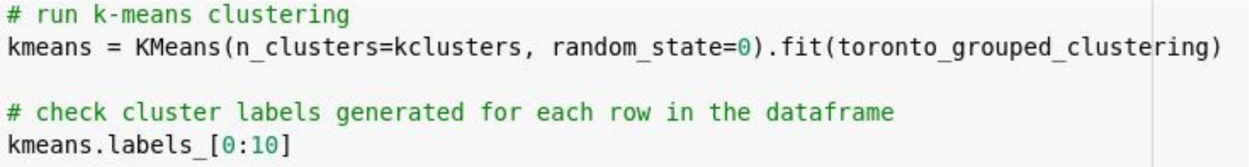


```
Coffee Shop            340
Café                   192
Restaurant             119
Pizza Place            112
Bakery                 109
Bar                    102
Italian Restaurant      91
Park                    90
Sandwich Place          79
Hotel                   77
Fast Food Restaurant    71
Clothing Store          64
Japanese Restaurant     61
American Restaurant     59
Gym                     57
Sushi Restaurant        53
Burger Joint            53
Pharmacy                50
Grocery Store           46
Pub                     45
```

K-Means Clustering

In order to give people good information about the characteristics of each neighborhood, it's good to cluster the neighborhoods into groups and find the characteristics of each group and try to generalize those characteristics to each neighborhood in the cluster.

We will consider the data frame we built (toronto_grouped) as the input of the K-Means algorithm.

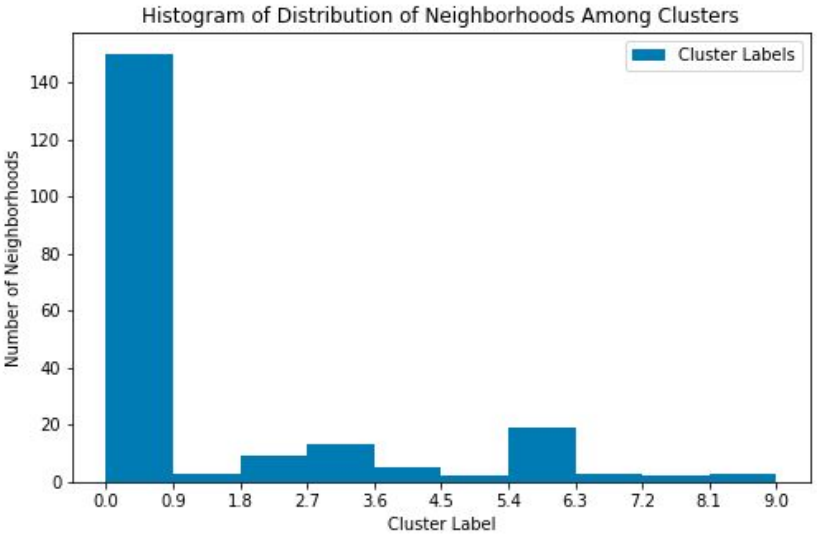


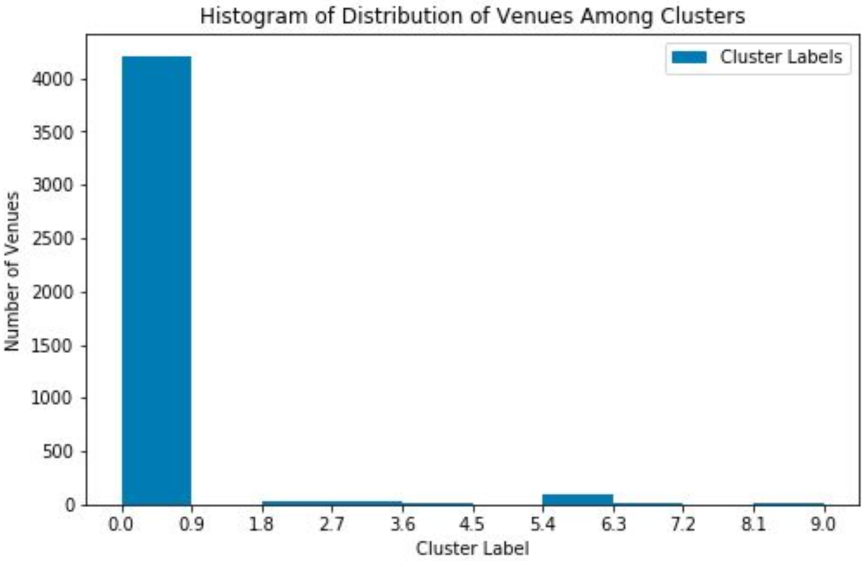
```
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

Result Analysis

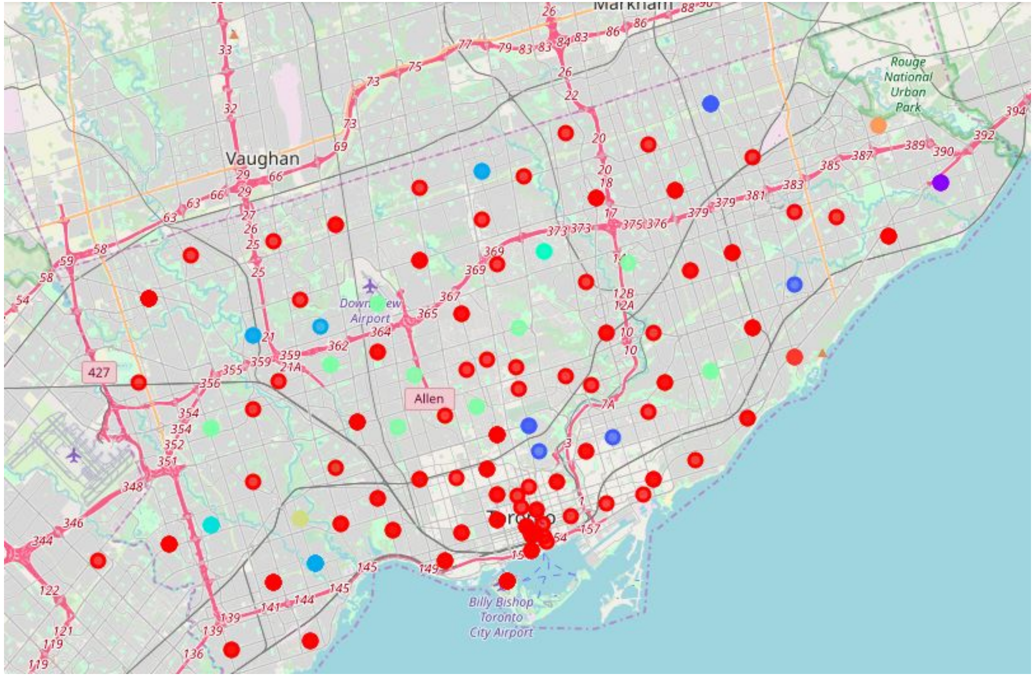
Cluster 0 has the biggest number of neighborhoods and venues. This is an indication that Toronto city is very homogeneous in nature of venues (most are some kind of restaurants). However, there are some neighborhoods that belong to other clusters.





Result Analysis

The geographical distribution of different clusters:



Conclusion

Basic clustering process clearly finds that most of Toronto’s neighborhoods are homogeneous and don’t have that much diversity in the nature of venues.

Perhaps, in the future, more data will result in more diversity, especially if it took into consideration the demographics of those neighborhoods.