Clustering Neighborhood Data in Toronto City: Cities Facilities

FIrman Insan Muhammad
March 2020

1. Introduction

1.1 Background

In a Big modern city, we can find a lot of open data that correspond our needs like location data of food and facilities venue. It may have varied in places from quiet to crowded one. It is important find and segment the different food venues and also facilities in a neighborhood according to venue category, and then we can group neighborhoods together that incorporate similar kind of neighborhoods to find pattern and similarities.

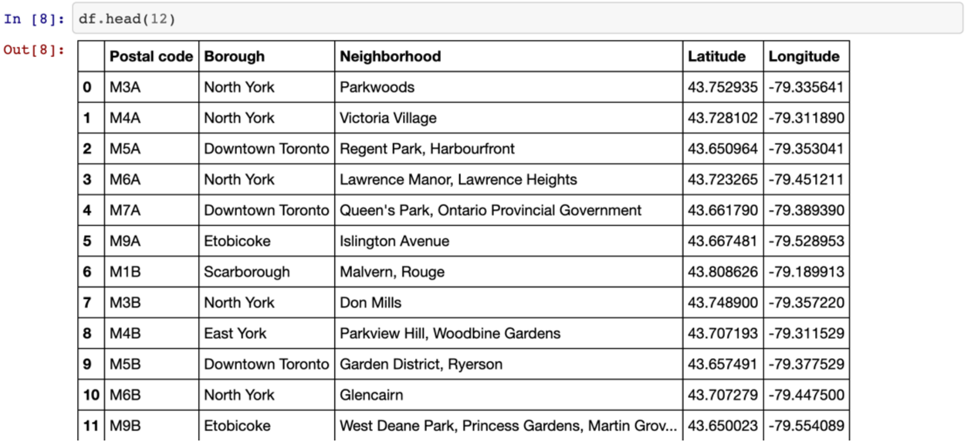
1.2 Problem

The challenging problem of any big cities is to provision this resource they have to get the best and optimum decision making to choose a place to start a business or may be life and choosing a neighbourhood.

2. Data Acquisition and Pre-Processing

2.1 Data Sources

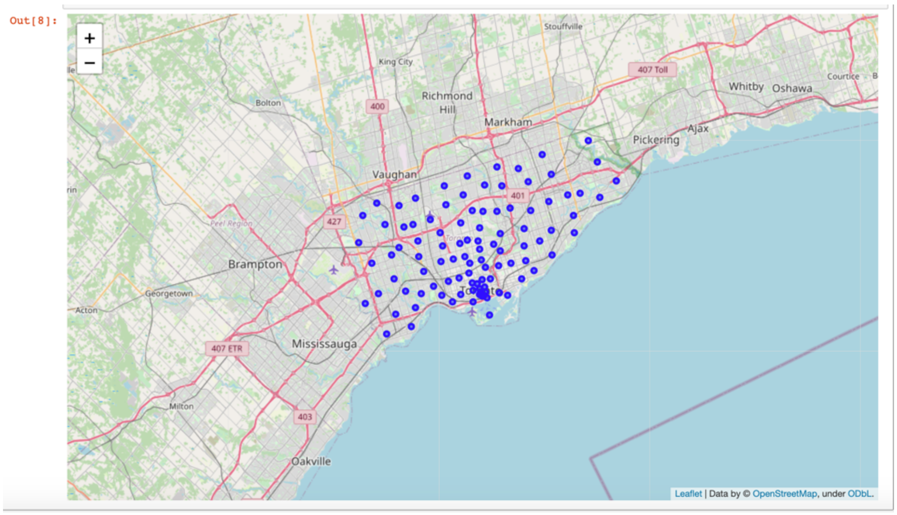
1. For neighborhood and borough naming in Toronto City we can get data from Wikipedia.



```
In [8]: df.head(12)
Out[8]:
```

	Postal code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.752935	-79.335641
1	M4A	North York	Victoria Village	43.728102	-79.311890
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.650964	-79.353041
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.723265	-79.451211
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.661790	-79.389390
5	M9A	Etobicoke	Islington Avenue	43.667481	-79.528953
6	M1B	Scarborough	Malvern, Rouge	43.808626	-79.189913
7	M3B	North York	Don Mills	43.748900	-79.357220
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.707193	-79.311529
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657491	-79.377529
10	M6B	North York	Glencairn	43.707279	-79.447500
11	M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...	43.650023	-79.554089

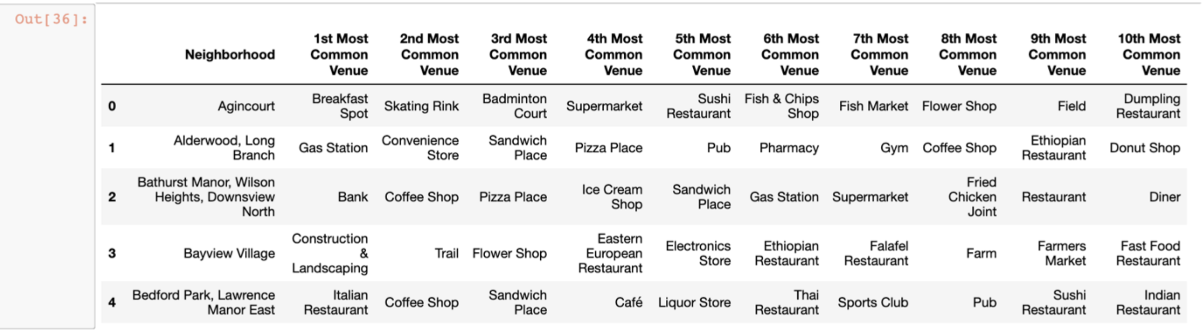
2. For location and geo tagging we collect data from foursquare.



3. For food venues and other facilities, we collect data from Toronto open data.

2.2 Web Scrapping

To get our first data sources which we can get data from Wikipedia using web scrapping, Xpath Python Library, then wrangling it into pandas Data Frame. And also for coordinate data, which we get from foursquare Application Programming interface (API) in JSON format and the combine it with csv data from Toronto open data. Once the data are already tidied, we can see the pattern that we want to research on.



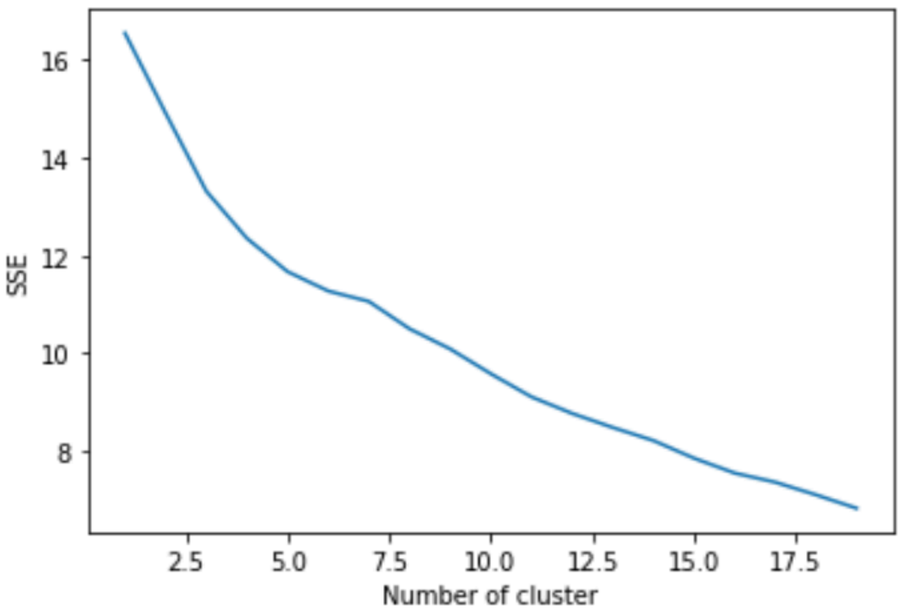
Out[36]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Breakfast Spot	Skating Rink	Badminton Court	Supermarket	Sushi Restaurant	Fish & Chips Shop	Fish Market	Flower Shop	Field	Dumpling Restaurant
1	Alderwood, Long Branch	Gas Station	Convenience Store	Sandwich Place	Pizza Place	Pub	Pharmacy	Gym	Coffee Shop	Ethiopian Restaurant	Donut Shop
2	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Pizza Place	Ice Cream Shop	Sandwich Place	Gas Station	Supermarket	Fried Chicken Joint	Restaurant	Diner
3	Bayview Village	Construction & Landscaping	Trail	Flower Shop	Eastern European Restaurant	Electronics Store	Ethiopian Restaurant	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
4	Bedford Park, Lawrence Manor East	Italian Restaurant	Coffee Shop	Sandwich Place	Café	Liquor Store	Thai Restaurant	Sports Club	Pub	Sushi Restaurant	Indian Restaurant

3. Methodology of Unsupervised Machine Learning

3.1 Optimization of Cluster Number

First to calculate the optimal number of clusters, I am experimenting with several cluster and iteration to look optimal cluster with sensible SSE.

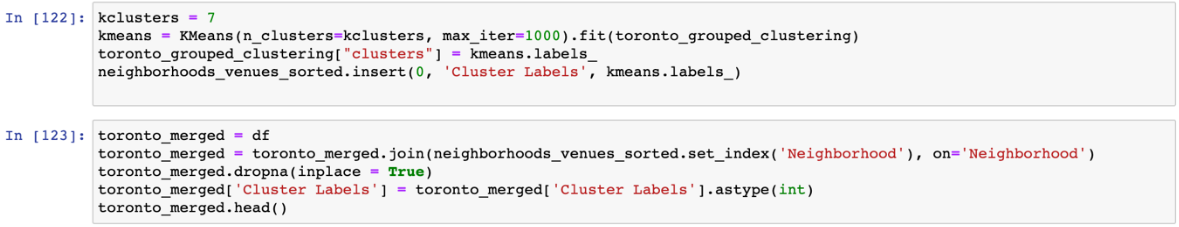


With the figure above we can look at optimal number of clusters is up to 5- 7 cluster.

3.2 K- means

According to Wikipedia, k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

And after we found optimal clusters, we can invoke the code like below to start unsupervised machine learning algorithm.



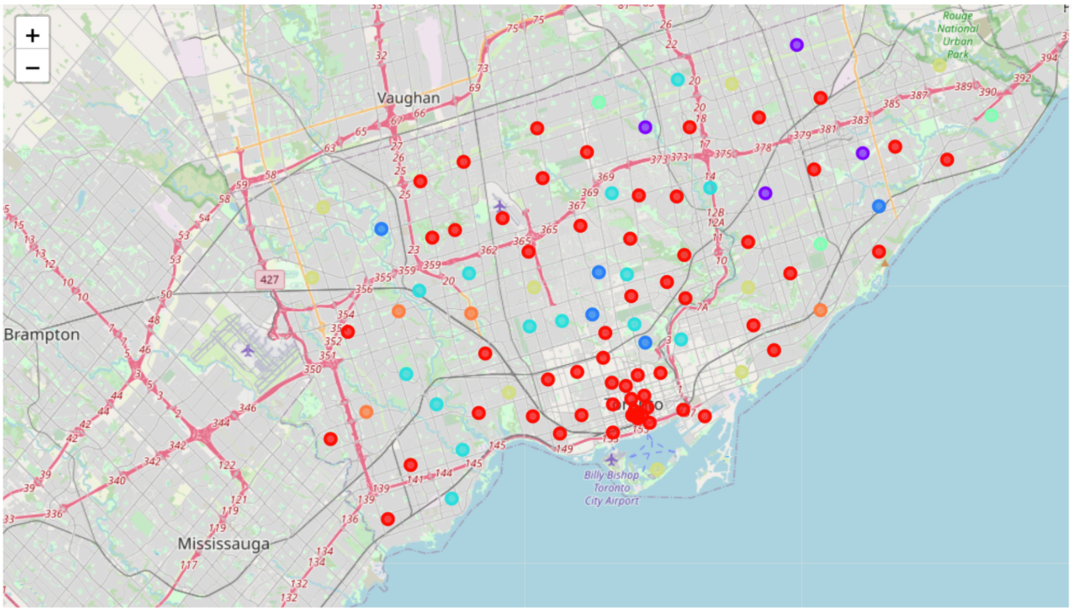
```
In [122]: kclusters = 7
          kmeans = KMeans(n_clusters=kclusters, max_iter=1000).fit(toronto_grouped_clustering)
          toronto_grouped_clustering["clusters"] = kmeans.labels_
          neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

In [123]: toronto_merged = df
          toronto_merged = toronto_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')
          toronto_merged.dropna(inplace = True)
          toronto_merged['Cluster Labels'] = toronto_merged['Cluster Labels'].astype(int)
          toronto_merged.head()
```

4. Result

4.1 Visualization on the Geo Map

Out[124]:



The different color that available on the Toronto map above is the cluster of venues from foursquare API based on the most common venue on its region / borough.

We can see there are 7 clusters that we can use to determine our decision whether to invest on property or to start a business. For example, if we know food and beverages venue is too common in the area, we can try to start a business to complement the market there. And so on.

5. Further works

In an Unsupervised clustering the room of improvement can be overlooked on venues there and enrich our data as much as possible so we can get the optimal results of our investigations.