Finding similarity and grouping Neighborhoods in Toronto, based on venues

1. Introduction

In every age and time, people have been moving around to the world for one reason or the other. When they move to a new city, they need to select a place where they can buy new house or where they want to live. While selecting a place they have several factors in mind depending who is moving and from where. Such as an individual moving out due to his job, might look for place close to his job with some coffee shops, restaurants, banks, pharmacy and more. On the other hand, if a family is moving with their children one of the many important things for them will be a school nearby.

So, to find a place of their choice they have to physically go in different neighborhoods in the city or search thoroughly on the internet, which could be cumbersome. And it is also difficult to have a comparison of all neighborhoods in a big city. The aim of this project is to cluster neighborhoods based on the factors mentioned above and make it easy for the people to decide which neighborhood to choose to buy a house.

2. Data acquisition and Cleansing

2.1. Data Source

The solution is specifically provided for the Toronto City that is for the people who are moving to Toronto. For this data for neighborhood along with their postal codes has been collected from Wikipedia page:'https://en.wikipedia.org/w/index.phptitle=List_of_postal_codes_of_Canada:_M&oldid=1011037969' as shown in the Figure 1. While Longitude and Latitude of each neighborhood has been gathered from 'https://cocl.us/Geospatial_data' for each neighborhood and the data about the venues is collected by using the Foursquare API.

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Figure 1

2.2. Data Cleansing

From the data, the postal code that has not been assigned any neighborhoods was deleted. After that, this data has been merged with geospatial data and Longitude and Latitude are added to the table for every corresponding neighborhood. There were 103 neighborhoods in the Toronto City among them 5 are shown in the Figure below.

	Postal Code	Borough	Neighbourhood	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	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Figure 2

After gathering Toronto city neighborhoods with their Longitude and Latitude, Foursquare API has been used to get nearby venues for each neighborhood with venue name and its category. A total of 4883 venues are gathered from 1000 miles of each neighborhood. Among these venues there were 330 unique venue categories

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
2	Parkwoods	43.753259	-79.329656	Tim Hortons	43.760668	-79.326368	Café
3	Parkwoods	43.753259	-79.329656	A&W	43.760643	-79.326865	Fast Food Restaurant
4	Parkwoods	43.753259	-79.329656	Bruno's valu-mart	43.746143	-79.324630	Grocery Store
...	...	...	...	...	...	...	...
4878	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Mr.Sub	43.636174	-79.520655	Restaurant
4879	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Queensway Fish & Chips	43.621720	-79.524588	Fish & Chips Shop
4880	Mimico NW, The Queensway West, South of Bloor	43.628841	-79.520999	Global Pet Foods	43.621304	-79.526146	Pet Store

Figure 3

2.3. Feature Selection

To make venues as our features one hot encoding is done on venue categories and then grouped by neighborhoods. As there were 330 venues and those all cannot be selected as our features we have selected venues with top 10 frequencies as features for calculating similarity among neighborhoods. The Figure 4 shows those top 10 venues with their respective mean against each neighborhood.

Neighborhood	Coffee Shop	Café	Park	Restaurant	Pizza Place	Italian Restaurant	Bakery	Grocery Store	Sandwich Place	Japanese Restaurant	Longitude	Latitude
Agincourt	0.044444	0.022222	0.022222	0.022222	0.022222	0.000000	0.044444	0.022222	0.044444	0.000	-79.262029	43.794200
Alderwood, Long Branch	0.040000	0.000000	0.080000	0.000000	0.080000	0.000000	0.000000	0.040000	0.040000	0.000	-79.543484	43.602414
Bathurst Manor, Wilson Heights, Downsview North	0.064516	0.000000	0.064516	0.032258	0.032258	0.000000	0.000000	0.000000	0.032258	0.000	-79.442259	43.754328
Bayview Village	0.000000	0.062500	0.062500	0.062500	0.000000	0.000000	0.000000	0.125000	0.000000	0.125	-79.385975	43.786947
Bedford Park, Lawrence Manor East	0.081081	0.027027	0.027027	0.027027	0.027027	0.081081	0.027027	0.027027	0.054054	0.000	-79.419750	43.733283

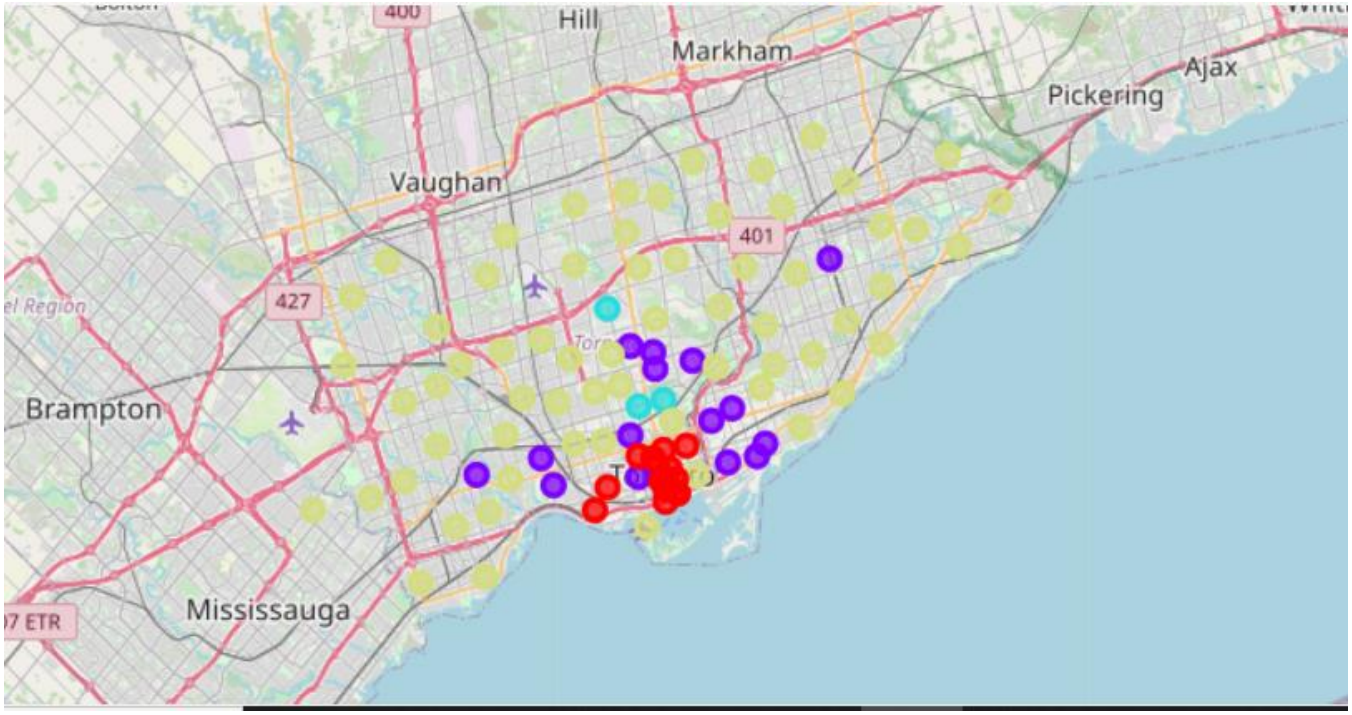
Figure 4

3. Methodology

To cluster or group neighborhoods in the city based on their venues, initially KMean clustering algorithm was selected with varying number of clusters. But those clusters were hard to interpret so DBSCAN algorithm has been selected. The DBSCAN algorithm has been selected instead because it was difficult to specify correct number of cluster in KMean even with using elbow method.

4. Results

With DBSCAN algorithm, 4 different clusters were formed. The cluster shown on the map below with red color has higher number of coffee shops, cafes, parks, restaurants and specifically Japanese Restaurants as compare to other clusters. So, these neighborhoods are most likely having a community of Japanese people around and Japanese might prefer this place.



Second cluster of neighbors shown with purple color has higher number of coffee shops, some parks, pizza places and restaurants but very few other venues.

The cluster with blue colors has only three neighborhoods and these neighborhoods have many coffee shops and Italian restaurants but at the same time it has also more cafes, parks, pizza places and all other venues compared to neighborhoods in other clusters. These three neighborhoods could be a good choice for Italian food lovers but simultaneously for any individual as this cluster is a good combination of all features.

The neighborhoods with mustered color are considered as noise by the DBSCAN. Looking at the data set of this cluster we can see that DBSCAN consider this cluster as noisy because most of the neighborhoods in this cluster has zero feature values.

5. Discussion

The DBSCAN has made four clusters with the EPS 1.5. Changing this number has a great impact on the clusters formed. With value smaller than 1.5 it considers all neighborhoods as noise and value greater than 1.5 include almost all neighborhoods in a single cluster. The top ten features are selected for clustering but increasing feature set makes it difficult to interpret the results.

Moreover, there are some important features that people usually consider them while moving to new neighborhood such as housing cost, crime rate in the area and more. These features are not been considered in this project because the features are mainly incorporated through Foursquare API. Including these features could change the clusters and could help users in better selection.

6. Conclusion

In this study, I have used DBSAN model to cluster neighborhoods of Toronto city. To group neighborhoods in the city, the similarity is calculated based on the venue's information gathered through Foursquare API. The model is chosen over other clustering model such as KMean because it does not require to specify number of clusters and formed clusters on its own. The model formed four different clusters and observing these cluster reveals insight into the groups. Such as a person who wants to have all the facilities nearby, can choose Bedford Park, Lawrence Manor East, one of the three neighborhoods in the group to live.