Battle of Neighborhoods

Pablo Fernandez Cid

01/05/2020

1. Introduction

1.1. Background

Toronto is a Canadian city with a population of more than 2.6 million and divided in 140 neighborhoods. Currently, for this study case, a person is living in Etobicoke borough, more specifically in the neighborhood of Thistletown, with 10,360 residents and multiple services, such as three public school boards and three surrounded parks.

The majority of the residents live in houses, are married and have at least one children, according to the data from Neighborhood Profiles. The person is not married and doesn't have any child, and currently is living alone. He was born in Europe and he enjoys the venues inside his neighborhood.

1.2. Problem

Due to a job opportunity in Toronto East, our subject is thinking in moving to a new neighborhood, closer to this new possible job, which is located in Guildwood, East Toronto.

As this job opportunity is an improvement in his career, he expects to have also a improvement in the neighborhood, having in mind services, such as restaurants or entertainment, and also be surrounded by people in the same situation as him (single people, currently working and from Europe). Besides, he expects the same security at least as in his previous neighborhood. Besides, he expects that the new neighborhood will be closer to the new job.

1.3. Interested People

This analysis could be interesting also for people in the same situation as our subject or people interested in the city of Toronto.

2. Data

2.1. Data Used

For this study, the following data is going to be used:

- Foursquare API: Extraction of information regarding the venues per each Neighborhood

- Neighborhood Profiles: information in order to understand the population, their income and their origin for each neighborhood. Data is extracted from multiple sources and merged by Open Data Toronto. See link here for more information

- Police Report information: Information regarding police report from 2014 to 2019 (link here). This data will be used to understand the security of each neighborhood, having in mind the number per each type of crime (Assault, Auto Theft, Break and Enter, Robbery and Theft Over) per year.

2.2. Data Preparation

Per each database, it is necessary a different preparation:

- Neighborhood profile:
    1. Extract only the necessary rows:
        - Population Data
        - Age of the population
        - Type of dwelling: apartment, single-detached house among other
        - Family Characteristics, couples with/without children,
        - Income of individuals in 2015
        - Immigrants by continent
    2. Format the numbers and convert them to integer in order to operate with them

    Example of data for Neighborhood profiles

		City of Toronto	Thistletown-Beaumond Heights
Topic	**Characteristic**		
Immigrants by selected place of birth	**Asia**	674490	2655
	Americas	212010	1205
	Europe	298270	1140
	Africa	77445	460
	Oceania and other places of birth	3780	10
	Total	1265995	5470

- Police Report information:
    1. Some of the Neighborhoods names are different from Neighborhood profile, so these names have been corrected. Also, it is necessary to remove a space at the end of all the Neighborhoods names
    2. Group by Neighborhood
    3. Calculate per each Neighborhood the quantity of crimes for a specific year and per 1000 habitants in that Neighborhood.

    Example of crimes per 1000 habitants for each Neighborhood in 2019:

Neighbourhood	Population	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
Agincourt North	29113	2.713564	1.442654	1.923539	1.202212	0.068698
Agincourt South-Malvern West	23757	5.177421	2.609757	4.209286	1.220693	0.631393
Alderwood	12054	2.986560	1.161440	2.156960	0.414800	0.580720
Annex	30526	9.434580	0.982769	6.944899	1.015528	1.375876
Banbury-Don Mills	27695	2.671962	1.516519	2.924716	0.361076	0.505506

- Foursquare:
    1. Get data regarding the location per each Neighborhood using Geolocator. There are some errors during this procedure:
        - Error due to the mix of several names in the name of the Neighborhood, such as Bedford Park-Nortown: solved separating the names and calculating the average of the coordinates
        - Missing data: only 3 neighborhoods affected (“Humbermede”, “Mimico (includes Humber Bay Shores)” and “Woodbine Corridor”) . Solved manually checking the coordinates in google maps.

    2. From the list of venues, remove the venues not interesting for our subject due to their interest (maintain restaurants and fitness venues), situation(single) and gender (male).

    3. Calculate the frequency per each kind of venue per each neighborhood

    Example of the data

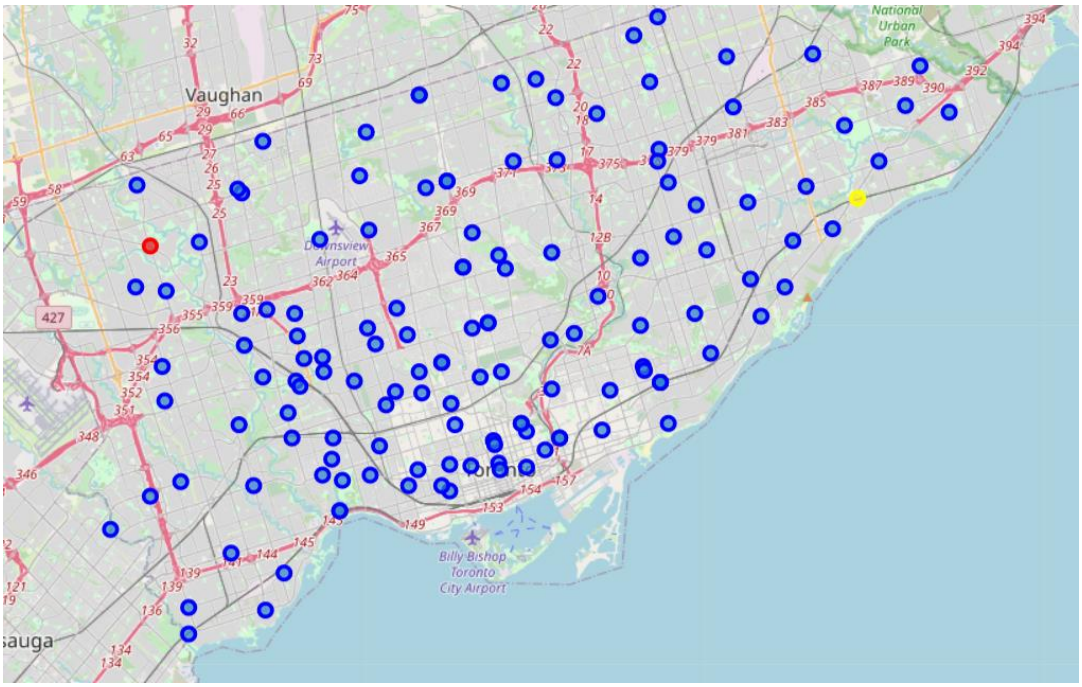
Neighborhood	Latitude	Longitude	Cluster Labels	ATM	Afghan Restaurant	American Restaurant	Amphitheater	Antique Shop	Arcade	Argentinian Restaurant	...	Transportation Service
Agincourt North	43.808	-79.2664	5.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0
Agincourt South-Malvern West	43.7892	-79.263	5.0	0.0	0.0	0.066667	0.0	0.0	0.0	0.0	...	0.0
Alderwood	43.6017	-79.5452	5.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0
Annex	43.6703	-79.4071	5.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0
Banbury-Don Mills	43.7348	-79.3572	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
Wychwood	43.6821	-79.424	9.0	0.0	0.0	0.020000	0.0	0.0	0.0	0.0	...	0.0
Yonge-Eglinton	43.7067	-79.3983	9.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0
Yonge-St.Clair	43.6881	-79.3944	9.0	0.0	0.0	0.017241	0.0	0.0	0.0	0.0	...	0.0
York University Heights	43.7588	-79.5194	5.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0

3. Methodology

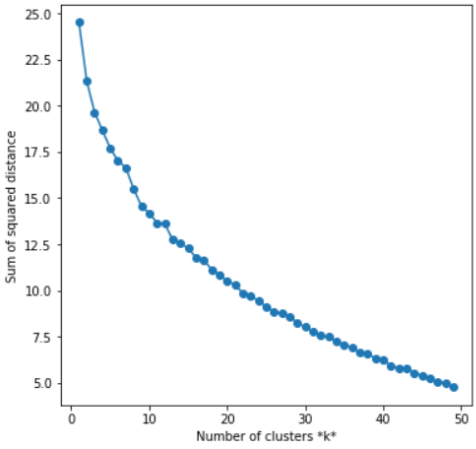
For the election of the best neighborhood, three criteria will be used based on category venues, criminality per neighborhood and profile of the neighborhood.

3.1. Clustering based on venues

Having the information prepared, the first step is the clustering in order to find similar patterns between neighborhoods and services in these neighborhoods. The neighborhoods for Toronto are displayed below. The red circle is the current neighborhood for our subject and the yellow circle is the neighborhood of the new job.

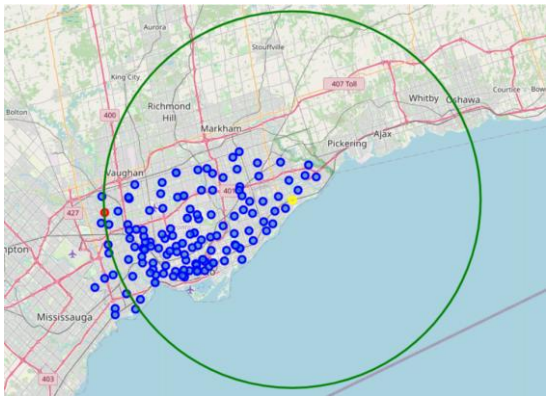


The clustering will be done based on the frequency of each venue. The method to be used for the cluster is K-means because the database is medium size and this method is efficient for this kind of cases. The value of K will be calculated based on the elbow method. If we use all the categories venues, the inertia, or within-cluster sum-of-squares criterion, there is no a really defined elbow, as It is shown below, so the data will be simplified in order to achieve better results.

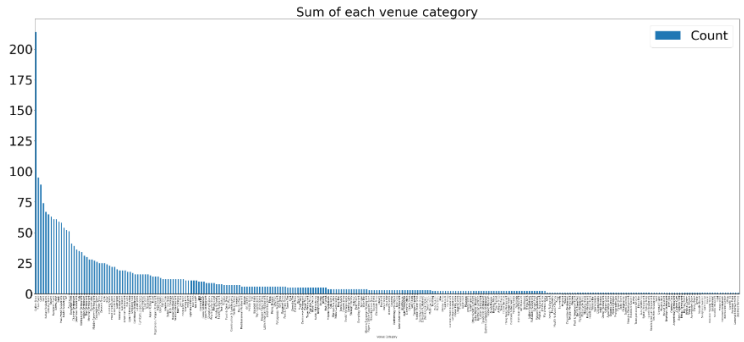


The simplification will be done as follows:

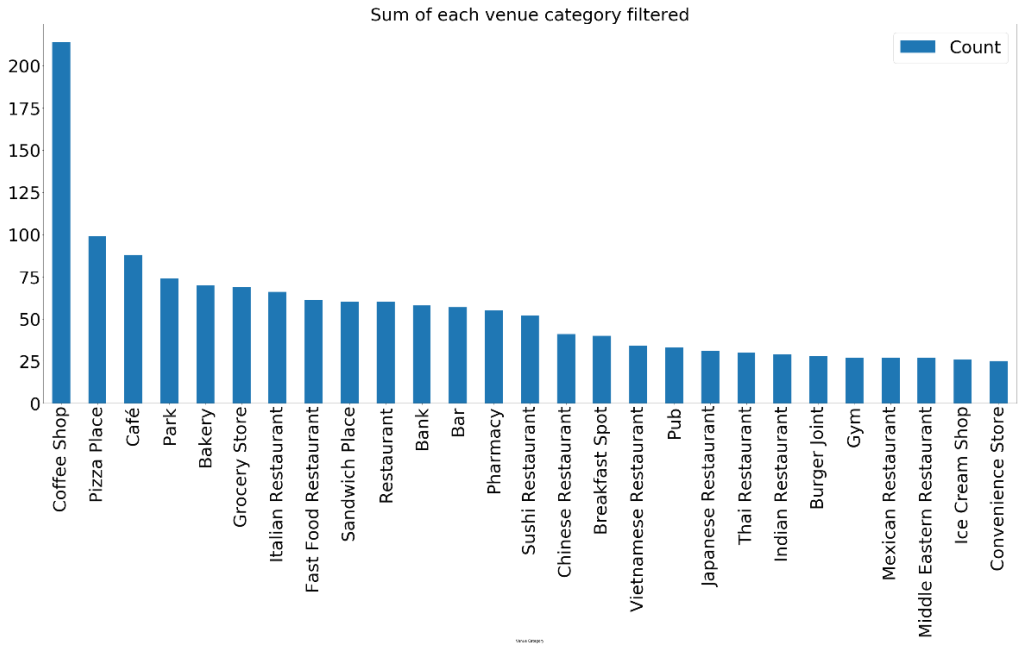
- Remove all the neighborhoods which are farther to the new job neighborhood than the origin neighborhood



- Remove all the venues with a very small frequency. In graph below it can be seen that the half of the venue categories have a small frequency compared with the other half. So these venue categories will be removed in order to simplify the clustering



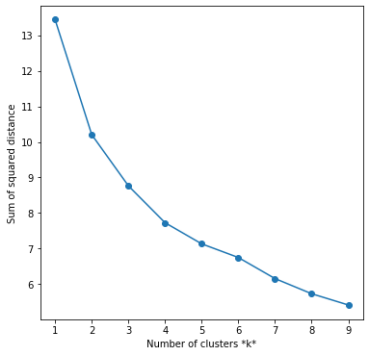
Removing category venues with less than 25 locals for all the neighborhoods this is the result.



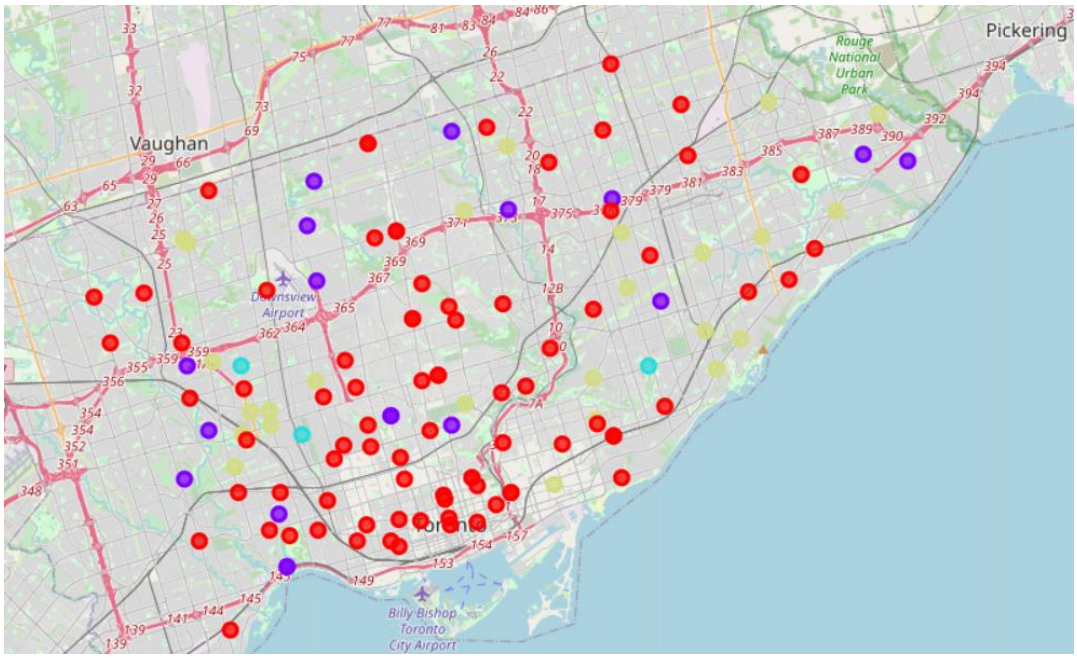
- Group similar categories to reduce to 15 categories

New Category	Category Venues				
Coffee	Coffee Shop	Café	Breakfast Spot		
Food Store	Grocery Store	Convenience Store			
Fast Food	Sandwich Place	Fast Food Restaurant	Burger Joint	Pizza Place	
Asian Restaurant	Sushi Restaurant	Vietnamese Restaurant	Thai Restaurant	Chinese Restaurant	Japanese Restaurant
Mexican Restaurant	Mexican Restaurant				
Ice Cream Shop	Ice Cream Shop				
Italian Restaurant	Italian Restaurant				
Restaurant	Restaurant				
Bar-Pub	Pub	Bar			
Oriental Restaurant	Indian Restaurant	Middle Eastern Restaurant			
Park	Park				
Gym	Gym				
Pharmacy	Pharmacy				
Bank	Bank				
Bakery	Bakery				

With the simplification, the k result is as follows, so now we can take 5 as value of the k for K-means

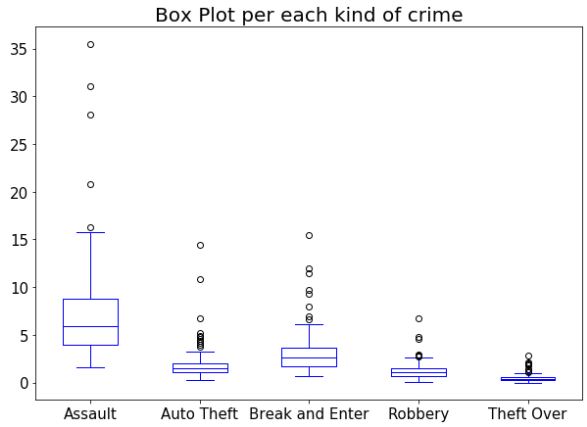


The clusters after all this are shown below

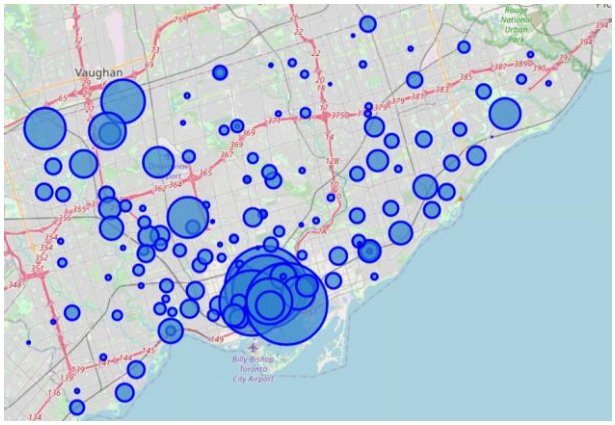


3.2. Criminality per Neighborhood

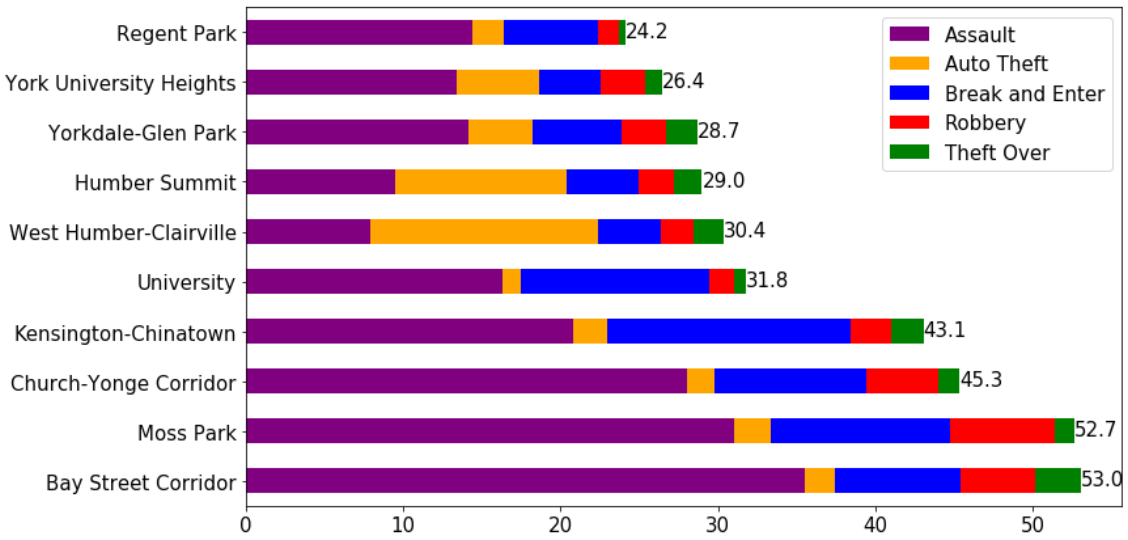
The statistical values per each kind of crime is represented in the graph below, we can see that the most common crime is the Assault, with some Neighborhood with a high isolated value.



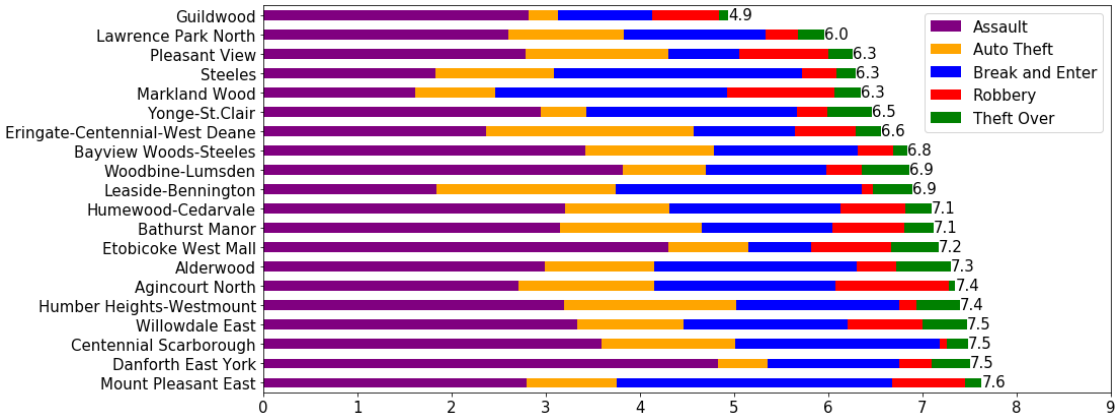
Having the data of the crime, the criminality per neighborhood can be represented with a bubble chart as follows. We can see that most of the crimes are in the center of the city and in the northwest.



Summarizing, this is the top 10 Neighborhood most insecure which should be avoided in case of decide a new neighborhood



On the other hand, the top 20 safest neighborhoods are this one, which can be a good option for the new Neighborhood



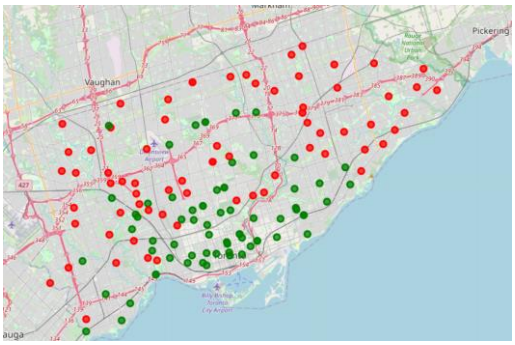
3.3. Profiles Neighborhood

Here we have three kind of data:

- People Age and Dwelling Characteristics of the neighborhood: having in mind the profile of our subject, the percentage of people living alone with less than 65 years per Neighborhood is interesting for this study. Below the results in descending order.

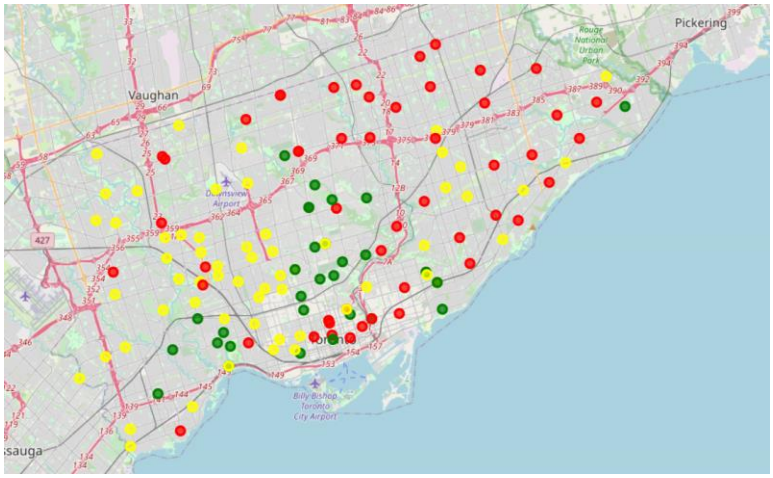
Neighbourhood	Liv. alone age less than 65 (percentage)
Church-Yonge Corridor	31.716656
Waterfront Communities-The Island	30.714730
Niagara	30.676716
Moss Park	29.308495
Bay Street Corridor	25.448696
Mount Pleasant West	25.355722
South Parkdale	22.037622
Cabbagetown-South St. James Town	20.353072
Annex	20.032104
North St. James Town	19.983884

With the calculation of the percentile 50% (10.88%) we can divide in a map the neighborhoods with more people living alone with less than 65 years (in green) and the others (in red).

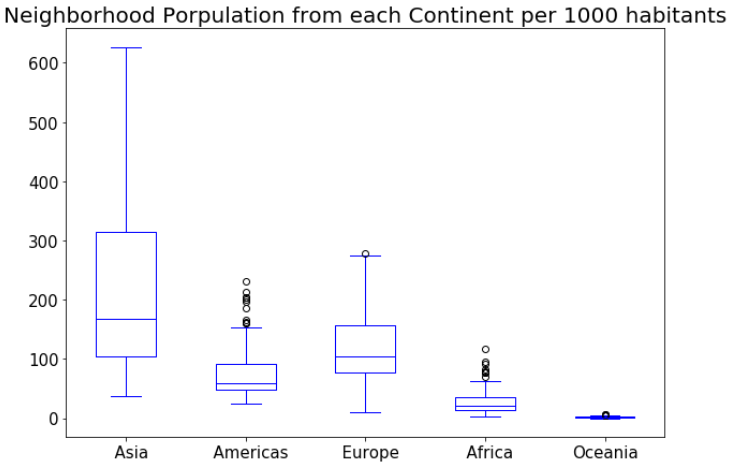


In this case, our subject is interested in the neighborhoods marked in green. Currently, his neighborhood is marked in red, with a percentage of 2.65% of people living alone with less than 65 years. Rest of the data in this category will be used after the filtering having in mind all the characteristics involved in the decision.

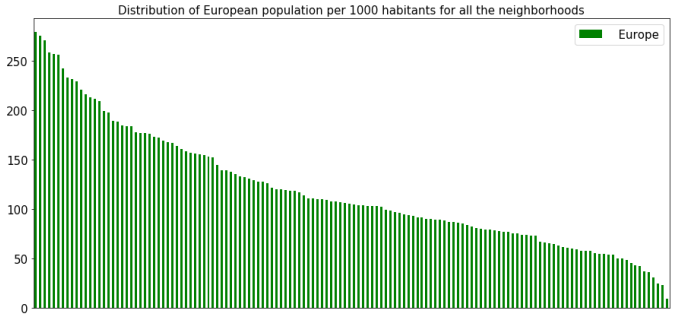
- Income of the population: the idea is displaying the most frequent category between three categories: income less than 19,999 (in red in map below), income between 20,000 and 49,999 (in yellow) and income more than 50,000 (in green). Although this parameter is not so important as the security of the neighborhood, it is interesting in order to understand the wealthy of each neighborhood. The idea is to maintain the same level as the current neighborhood (income between 20,000 and 49,999) or increase it



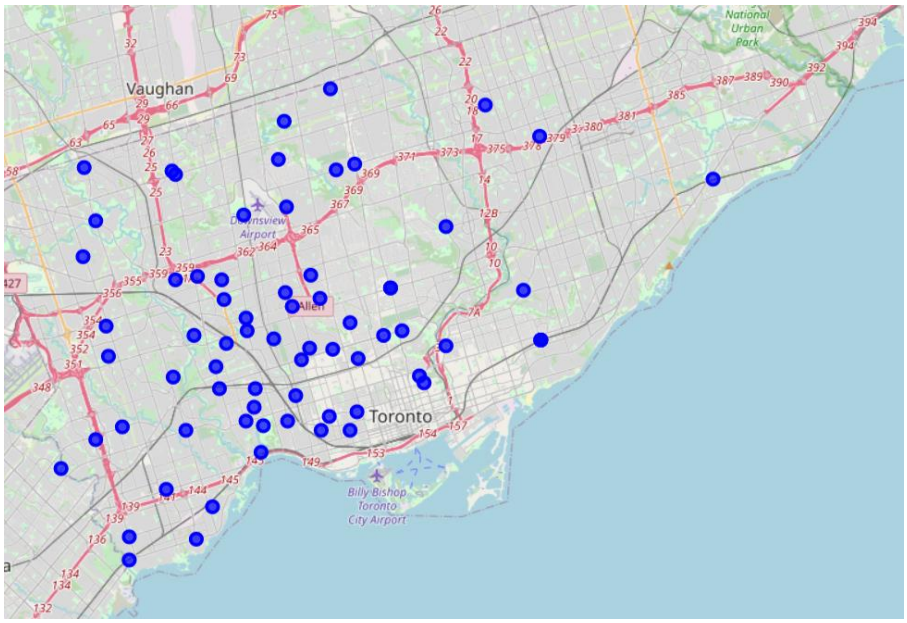
- Immigration: As our subject is European, we are going to understand how many European people is living in each neighborhood. As we can see in the box plot below, immigration from Asia is the most common followed by the immigration from Europe and there are not neighborhoods with isolated values.



The distribution of the European population for each neighborhood is shown below. It is shown in the graph below that the distribution is not constant, so the European population is more dense in some areas.



Having in mind this, a good approach is taking the percentile 50, displayed in map below for the chosen of the neighborhood.



4. Results

Merging all the data, the next conclusions can be reached:

- Cluster 0 of the categories venues is composed for the neighborhoods more similar to subject's neighborhood

- The neighborhoods with more crimes are specially in the center of Toronto, maybe due to the concentration of tourism and other commercial activities. Besides the northwest has a big concentration of crimes.

- There is more concentration of people living alone with age less than 65 years in the center and area close to the bay.

- There is a mix between different cluster per income along all the city, so the wealth is distributed

- The European population lives more in the west area of the city.

So, in conclusion, these are the neighborhoods with best fits for all the criteria: Briar Hill-Belgravia, Cabbagetown-South St. James Town, Dovercourt-Wallace Emerson-Junction, Dufferin Grove, Junction Area, Mimico (includes Humber Bay Shores), Oakwood Village, Palmerston-Little Italy, Playter Estates-Danforth, Rosedale-Moore Park, and Trinity-Bellwoods. These neighborhoods are displayed in map below.

And these are the data for each category:

- Crimes and population:

Neighborhood	Population	Crimes per 1000	Persons living alone (total)	Liv. alone age less than 65 (percentage)
Briar Hill-Belgravia	14257	12.765659	1725	8.066213
Cabbagetown-South St. James Town	11669	23.738110	3315	20.353072
Dovercourt-Wallace Emerson-Junction	36625	14.962457	4655	10.279863
Dufferin Grove	11785	15.528214	2020	13.831141
Junction Area	14366	12.529584	1985	11.067799
Mimico (includes Humber Bay Shores)	33964	14.662584	8205	18.578495
Oakwood Village	21210	13.342763	2650	7.190005
Palmerston-Little Italy	13826	13.742225	2230	12.765804
Playter Estates-Danforth	7804	17.811379	1455	13.582778
Rosedale-Moore Park	20923	13.191225	4170	11.614013
Trinity-Bellwoods	16556	17.516308	2160	10.660788

- Category Venues:

Neighborhood	Bakery	Bank	Gym	Ice Cream Shop	Italian Restaurant	Mexican Restaurant	Park	Pharmacy	Restaurant	Coffee	Food Store	Fast Food	Asiatic Restaurant
Briar Hill-Belgravia	0.142857	0.142857	0.000000	0.000000	0.000000	0.000000	0.142857	0.000000	0.000000	0.285714	0.000000	0.142857	0.000000
Cabbagetown-South St. James Town	0.061224	0.020408	0.000000	0.000000	0.040816	0.000000	0.020408	0.020408	0.061224	0.163265	0.040816	0.081633	0.081633
Dovercourt-Wallace Emerson-Junction	0.062500	0.000000	0.000000	0.000000	0.000000	0.000000	0.062500	0.062500	0.000000	0.375000	0.062500	0.000000	0.000000
Dufferin Grove	0.039216	0.000000	0.000000	0.000000	0.098039	0.039216	0.000000	0.000000	0.078431	0.196078	0.000000	0.019608	0.019608
Junction Area	0.037736	0.000000	0.018868	0.018868	0.075472	0.056604	0.000000	0.000000	0.000000	0.150943	0.056604	0.018868	0.094340
Mimico (includes Humber Bay Shores)	0.000000	0.142857	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.142857	0.428571	0.000000	0.000000
Oakwood Village	0.043478	0.000000	0.000000	0.000000	0.043478	0.086957	0.000000	0.000000	0.000000	0.086957	0.043478	0.130435	0.043478
Palmerston-Little Italy	0.014925	0.000000	0.000000	0.000000	0.059701	0.000000	0.000000	0.000000	0.014925	0.089552	0.000000	0.059701	0.044776
Playter Estates-Danforth	0.038462	0.038462	0.000000	0.000000	0.000000	0.038462	0.000000	0.076923	0.000000	0.153846	0.076923	0.038462	0.038462
Rosedale-Moore Park	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000
Trinity-Bellwoods	0.047619	0.000000	0.000000	0.023810	0.023810	0.000000	0.023810	0.000000	0.023810	0.071429	0.023810	0.023810	0.047619

- Income per ranges

Neighborhood	Below 19,999	Between 20,000 and 49,999	Over 50,000
Briar Hill-Belgravia	4495	5285	1990
Cabbagetown-South St. James Town	3175	3335	3470
Dovercourt-Wallace Emerson-Junction	11555	12305	6920
Dufferin Grove	3645	4155	2325
Junction Area	3760	4300	3505
Mimico (includes Humber Bay Shores)	8770	10445	9735
Oakwood Village	6545	6985	3560
Palmerston-Little Italy	4025	4750	3450
Playter Estates-Danforth	1910	2015	2370
Rosedale-Moore Park	4020	4425	8875
Trinity-Bellwoods	5185	5230	3740

- Immigrants per each 1000 habitants in the Neighborhood:

Neighborhood	Asia	Americas	Europe	Africa	Oceania
Briar Hill-Belgravia	234.972294	113.628393	233.218770	17.535246	0.000000
Cabbagetown-South St. James Town	126.831777	68.129231	110.549319	15.853972	1.285457
Dovercourt-Wallace Emerson-Junction	125.733788	69.624573	177.201365	17.883959	1.774744
Dufferin Grove	109.036911	73.398388	176.495545	13.576580	0.848536
Junction Area	114.158430	59.515523	129.124321	11.485452	2.088264
Mimico (includes Humber Bay Shores)	132.934872	63.743964	176.363208	20.610058	1.619362
Oakwood Village	139.321075	130.834512	172.324375	26.638378	0.471476
Palmerston-Little Italy	83.176624	48.821062	152.972660	10.125850	1.446550
Playter Estates-Danforth	63.429011	43.567401	107.637109	11.532547	3.844182
Rosedale-Moore Park	81.728242	54.963437	105.386417	17.922860	1.911772
Trinity-Bellwoods	145.566562	28.388500	167.310945	6.644117	0.906016

5. Discussion

Although more filters can be applied in order to reduce the neighborhoods to only a perfect one, it is better to leave some option to the subject just in case he prefers giving more importance to one category or another. Currently 11 neighborhoods of 140 have been selected following the criteria defined at the beginning of this document.

Other criteria can be used to obtain similar results, such as the range of ages per population or neighborhoods with population from a specific country, but this result is a good approach having in mind the expected result.

At future analysis, a clustering with other values could be defining, for example reducing the venue categories from foursquare and adding other kind of categories, such as criminality or immigration of the neighborhood. This could be done in further studies.

6. Conclusion

After this analysis, 11 potential neighborhoods have been found which fit the criteria defined by our subject:

- They have similar services as his current neighborhood
- The criminality is not high
- There are people living in the similar situation as him (alone with less than 65 years)
- The income of the population is similar to his income
- There is European population which could be good for his social life in case he is missing to talk with people with the same origin.

These neighborhoods are potentially and improvement of the current situation of the subject, having in mind that his current neighborhood doesn’t accomplish all the criteria defined.