FINDING THE RIGHT NEIGHBORHOOD TO OPEN A NEW BANK IN ETOBICOKE

José G. Pérez

August, 2021

1. Introduction

1.1. Business understanding

An investor is interested in opening a new Bank in the city of Etobicoke, Toronto. This city is the fourth with the most neighborhoods, with **153.794** inhabitants and an average income per inhabitant of **$40.935** [1] The investor asked his new business department through his team of Data Scientists an analysis that allows generating recommendations of which is the ideal neighborhood to open to the new Bank based on three considerations: 1) Less competition 2) Greater amount of new clients, 3) Potential clients with better income. From these preliminary variables, the neighborhood where the new bank headquarters will be opened must be identified.

1.2. Analytic approach

To select the ideal neighborhood to open the new Bank, we will do an analysis of the demographic data of the Etobicoke neighborhoods, highlighting the number of population, per capita income and population growth. This will allow us to classify the most attractive neighborhoods considering these variables. From these variables we will use the **clustering machine learning algorithm** (**k-means**), prior to the application of the algorithm, the **standardization** of the variables will be carried out, so that they have the same scale. In this way we will group neighborhoods with common demographic characteristics. In this analysis we will incorporate the information from the most common offices around (**2000m**) of the Eobicoke neighborhoods, with emphasis on the Bank's offices. In this way, we will build a data set that will provide us with demographic data, information on bank offices and the results of the clustering analysis. We will support these analyzes with bank office location maps around neighborhoods acquired

from the **Foursquare API service**. All the analyzes described above will allow us to recommend the suitable vendor to open a new bank.

1.3. Data description

In order to make the necessary recommendations to the investor who is interested in opening a new Bank in the city of Etobicoke, the following data sources were used:

- Toronto demographics features [1].
- Principal 100 venues around 2000m of the Etobicoke neighborhoods using the Foursquare API services [2].

2. Methodology

The methodological approach used to find the solution to the problem posed was based on the procedure proposed in the course: *Data Science Methodology* [3].

2.1. Data collection

The main data source used for the analysis comes from the demographic data of the city of Toronto available on Wikipedia [1]. It's was collected by scrapping wiki URL using *Python Beautiful library*.

2.2. Data preparation

2.2.1. Explore dataset

The wiki URL scraped has information tables of several boroughs of the city of Toronto; we select the Etobicoke borough. The raw dataset has **15 columns and 13 rows**. The five first records of the raw dataset is shown in the Table 1; and mains stats of the raw dataset is shown in the Table 2.

Table 1. Raw dataset.

	Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Second most common language (after English) by name	Second most common language (after English) by percentage	Map
0	Alderwood	E	0211.00, 0212.00	11656	4.94	2360	-4.0	35239	8.8	8.5	Polish (6.2%)	06.2% Polish	NaN
1	Centennial	E	0236.01, 0236.02	12565	4.94	2544	0.5	34867	11.5	8.8	Polish (2.7%)	02.7% Polish	NaN
2	Clairville	E	0248.03, 0249.03	8506	6.71	1268	-3.3	26610	13.2	7.2	Punjabi (12.0%)	12.0% Punjabi	NaN
3	Eatonville	E	0213.00, 0221.02, 0222.01, 0222.02	19131	11.26	1699	4.3	36206	12.6	13.4	Serbian (3.2%)	03.2% Serbian	NaN
4	Humber Bay Shores	E	0200.00, 0201.00	10775	1.42	7588	-5.7	39186	15.7	31.6	Russian (5.2%)	05.2% Russian	NaN

Table 2. Mains stats of the raw dataset.

	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Map
count	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	0.0
mean	11709.800000	5.284000	3066.400000	0.646667	43175.333333	11.726667	14.693333	NaN
std	3845.808167	4.129361	1665.829429	6.461078	16503.088108	2.780151	7.835402	NaN
min	4674.000000	1.420000	421.000000	-7.400000	26610.000000	7.200000	5.800000	NaN
25%	9456.500000	2.815000	2304.500000	-4.850000	34141.000000	9.650000	8.650000	NaN
50%	10775.000000	4.130000	2652.000000	0.500000	37288.000000	11.800000	13.400000	NaN
75%	14325.500000	5.425000	3935.500000	4.000000	45290.500000	13.100000	20.000000	NaN
max	19131.000000	17.400000	7588.000000	15.400000	80618.000000	17.100000	31.600000	NaN

2.2.2. Data cleaning

The first step of the data cleaning was drop unimportant columns: 'FM', 'Census Tracts', '% Renters', 'Second most common language (after English) by name', 'Second most common language (after English) by percentage', 'Map'. The resulting dataset is shown in the Table 3.

Table 3. Raw dataset after column drop.

	Neighborhood	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %
0	Alderwood	11656	4.94	2360	-4.0	35239	8.8
1	Centennial	12565	4.94	2544	0.5	34867	11.5
2	Clairville	8506	6.71	1268	-3.3	26610	13.2
3	Eatonville	19131	11.26	1699	4.3	36206	12.6
4	Humber Bay Shores	10775	1.42	7588	-5.7	39186	15.7
5	Humber Heights	4674	1.69	2766	8.3	39738	10.1
6	Humberwood	7319	17.40	421	8.0	29576	7.9
7	Humber Valley Village	14453	5.45	2652	-0.1	80618	12.0
8	Islington – Six Points	16508	4.02	4106	3.7	43570	17.1
9	Kingsview Village	16254	4.05	4013	-6.2	32004	11.8
10	Long Branch	9625	2.22	4336	-7.4	37288	14.2
11	Markland Wood	10240	2.92	3507	1.0	51695	9.2
12	Mimico	14198	5.40	2629	15.4	47011	11.6
13	New Toronto	10455	2.71	3858	-5.8	33415	13.0
14	Princess Gardens	9288	4.13	2249	1.0	80607	7.2

The next step was to collect the longitude and latitude of each Etobicoke neighborhood using the geocoder from the Python library and subsequently integrate it into the working dataset. In Figure 1 we can see a map of the location of the neighborhood resulting after using geocoder, and in Table 4 we can see the working dataset with the data location of each neighborhood (longitude and latitude).

Figure 1. Etobicoke neighborhoods before data cleaning

Figure 1. Neighborhood location before data cleaning.

Table 4. Working dataset integrated with data location.

	Neighborhood	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	Latitude	Longitude
0	Alderwood	11656	4.94	2360	-4.0	35239	8.8	43.604960	-79.541160
2	Clairville	8506	6.71	1268	-3.3	26610	13.2	43.748030	-79.631220
3	Eatonville	19131	11.26	1699	4.3	36206	12.6	43.689581	-79.494751
4	Humber Bay Shores	10775	1.42	7588	-5.7	39186	15.7	43.626860	-79.476710
5	Humber Heights	4674	1.69	2766	8.3	39738	10.1	43.652247	-79.486697
6	Humberwood	7319	17.40	421	8.0	29576	7.9	43.725165	-79.621555
7	Humber Valley Village	14453	5.45	2652	-0.1	80618	12.0	43.641466	-79.492537
8	Islington – Six Points	16508	4.02	4106	3.7	43570	17.1	43.634870	-79.530520
9	Kingsview Village	16254	4.05	4013	-6.2	32004	11.8	43.702510	-79.572090
10	Long Branch	9625	2.22	4336	-7.4	37288	14.2	43.593540	-79.532750
11	Markland Wood	10240	2.92	3507	1.0	51695	9.2	43.633910	-79.569480
12	Mimico	14198	5.40	2629	15.4	47011	11.6	43.617290	-79.498850
13	New Toronto	10455	2.71	3858	-5.8	33415	13.0	43.601430	-79.509250

As you can see in Figure 1 there are two neighborhoods far from the Etobicoke region, therefore we proceed to delete them as part of the ongoing data cleansing process. The resulting map is shown in the Figure 2.

Figure 2. Etobicoke neighborhoods after data cleaning

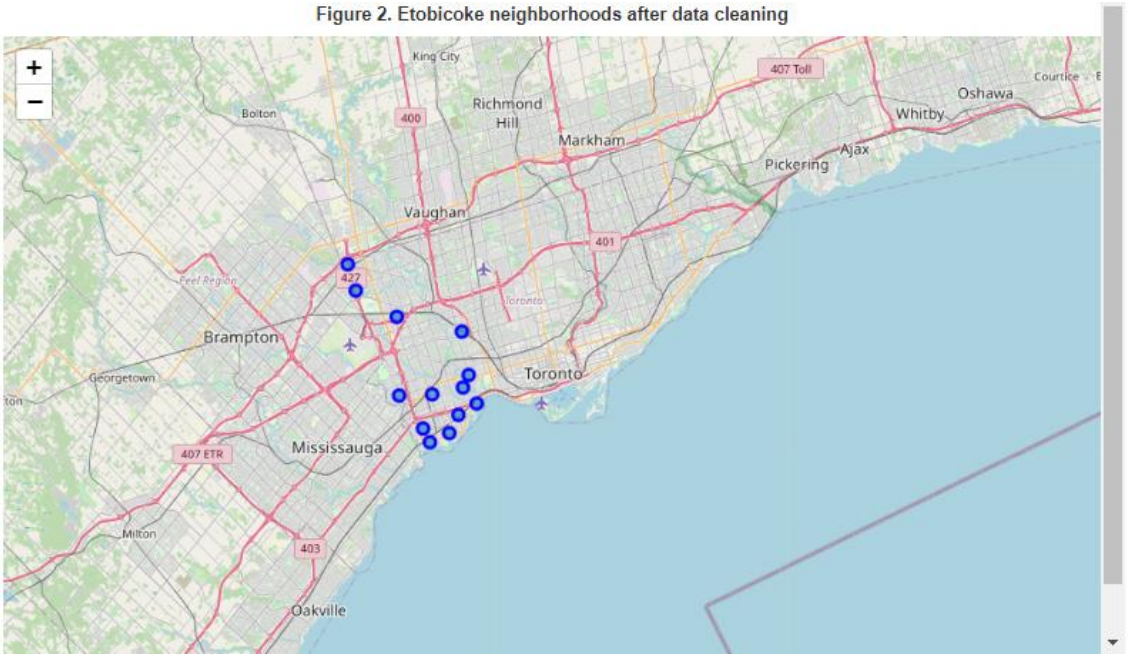

Figure 2. Etobicoke neighborhoods after data cleaning.

2.3. Data Integration

Using the Foursquare API services, we collected the top **100 venues** in each Etobicoke neighborhood within a **2000m radius**. Then, using python code (*onehot, groupby and others*) we obtained the **10 most common venues** around in a radius of 2000m (see Table 5)

Table 5. 10 most common venues around in a radius of 2000m of Etobicoke neighborhoods.

	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	Alderwood	Coffee Shop	Café	Fast Food Restaurant	Department Store	Bakery	Pizza Place	Electronics Store	Restaurant	Seafood Restaurant	Burger Joint
1	Clairville	Intersection	Coffee Shop	Sporting Goods Shop	Home Service	Indian Restaurant	Flower Shop	Music Venue	Department Store	Cemetery	Campground
2	Eatonville	Coffee Shop	Pizza Place	Convenience Store	Gas Station	Sandwich Place	Grocery Store	Park	Restaurant	Train Station	Fast Food Restaurant
3	Humber Bay Shores	Park	Italian Restaurant	Bank	Coffee Shop	Indian Restaurant	Sushi Restaurant	Grocery Store	Pharmacy	Pizza Place	Liquor Store
4	Humber Heights	Bakery	Coffee Shop	Café	Italian Restaurant	Sushi Restaurant	Park	Pub	Breakfast Spot	Bar	Pizza Place
5	Humber Valley Village	Coffee Shop	Italian Restaurant	Sushi Restaurant	Bakery	Pizza Place	Pub	Bank	Café	Thai Restaurant	Breakfast Spot
6	Humberwood	Coffee Shop	Restaurant	Sandwich Place	Fast Food Restaurant	Ice Cream Shop	Pizza Place	Bank	Café	Gas Station	Campground
7	Islington – Six Points	Coffee Shop	Bank	Fast Food Restaurant	Sushi Restaurant	Grocery Store	Restaurant	Sandwich Place	Thai Restaurant	Bakery	Burger Joint
8	Kingsview Village	Coffee Shop	Hotel	Fast Food Restaurant	Pizza Place	Gas Station	Sandwich Place	Restaurant	Bank	Asian Restaurant	Rental Car Location
9	Long Branch	Coffee Shop	Park	Pizza Place	Sandwich Place	Bank	Pub	Café	Burger Joint	South American Restaurant	Burrito Place
10	Markland Wood	Coffee Shop	Park	Sandwich Place	Sporting Goods Shop	Café	Beer Store	Restaurant	Discount Store	Gym	Breakfast Spot
11	Mimico	Restaurant	Coffee Shop	Sushi Restaurant	Pizza Place	Bank	Park	Italian Restaurant	Liquor Store	Grocery Store	Ice Cream Shop
12	New Toronto	Park	Coffee Shop	Skating Rink	Brewery	Pizza Place	Fast Food Restaurant	Café	Bakery	Sandwich Place	Liquor Store

To identify the neighborhoods that do not have banks or have fewer banks around a radius of **2000 meters**, a table was generated to count how many banks there are around each neighborhood (see Table 6)

Table 6. Number of Banks around 2000 meters of each neighborhood.

Neighborhood	Number of Banks
Alderwood	2
Clairville	0
Eatonville	0
Humber Bay Shores	5
Humber Heights	2
Humber Valley Village	4
Humberwood	2
Islington – Six Points	6
Kingsview Village	2
Long Branch	2
Markland Wood	2
Mimico	5
New Toronto	0

The integrated dataset used to the analysis result of the integration of the information of Table 4, Table 5 and Table 6. The resulting dataset is shown in the Table 7.

Table 7. Integrated dataset to made the data analysis.

	Neighborhood	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	Latitude	Longitude	Number of Banks	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Alderwood	11656	4.94	2360	-4.0	35239	8.8	43.604960	-79.541160	2	Coffee Shop	Café	Fast Food Restaurant
1	Clairville	8506	6.71	1268	-3.3	26610	13.2	43.748030	-79.631220	0	Intersection	Coffee Shop	Sporting Goods Shop
2	Eatonville	19131	11.26	1699	4.3	36206	12.6	43.689581	-79.494751	0	Coffee Shop	Pizza Place	Convenience Store
3	Humber Bay Shores	10775	1.42	7588	-5.7	39186	15.7	43.626860	-79.476710	5	Park	Italian Restaurant	Bank
4	Humber Heights	4674	1.69	2766	8.3	39738	10.1	43.652247	-79.486697	2	Bakery	Coffee Shop	Café
5	Humberwood	7319	17.40	421	8.0	29576	7.9	43.725165	-79.621555	2	Coffee Shop	Restaurant	Sandwich Place
6	Humber Valley Village	14453	5.45	2652	-0.1	80618	12.0	43.641466	-79.492537	4	Coffee Shop	Italian Restaurant	Sushi Restaurant
7	Islington – Six Points	16508	4.02	4106	3.7	43570	17.1	43.634870	-79.530520	6	Coffee Shop	Bank	Fast Food Restaurant
8	Kingsview Village	16254	4.05	4013	-6.2	32004	11.8	43.702510	-79.572090	2	Coffee Shop	Hotel	Fast Food Restaurant
9	Long Branch	9625	2.22	4336	-7.4	37288	14.2	43.593540	-79.532750	2	Coffee Shop	Park	Pizza Place
10	Markland Wood	10240	2.92	3507	1.0	51695	9.2	43.633910	-79.569480	2	Coffee Shop	Park	Sandwich Place
11	Mimico	14198	5.40	2629	15.4	47011	11.6	43.617290	-79.498850	5	Restaurant	Coffee Shop	Sushi Restaurant
12	New Toronto	10455	2.71	3858	-5.8	33415	13.0	43.601430	-79.509250	0	Park	Coffee Shop	Skating Rink

2.4. Data understanding

In the Figure 3 we can see an integrated barplot with the population, average income, population density and % change of population (since 2001), there are the main demographics features of the Etobicoke neighborhoods.

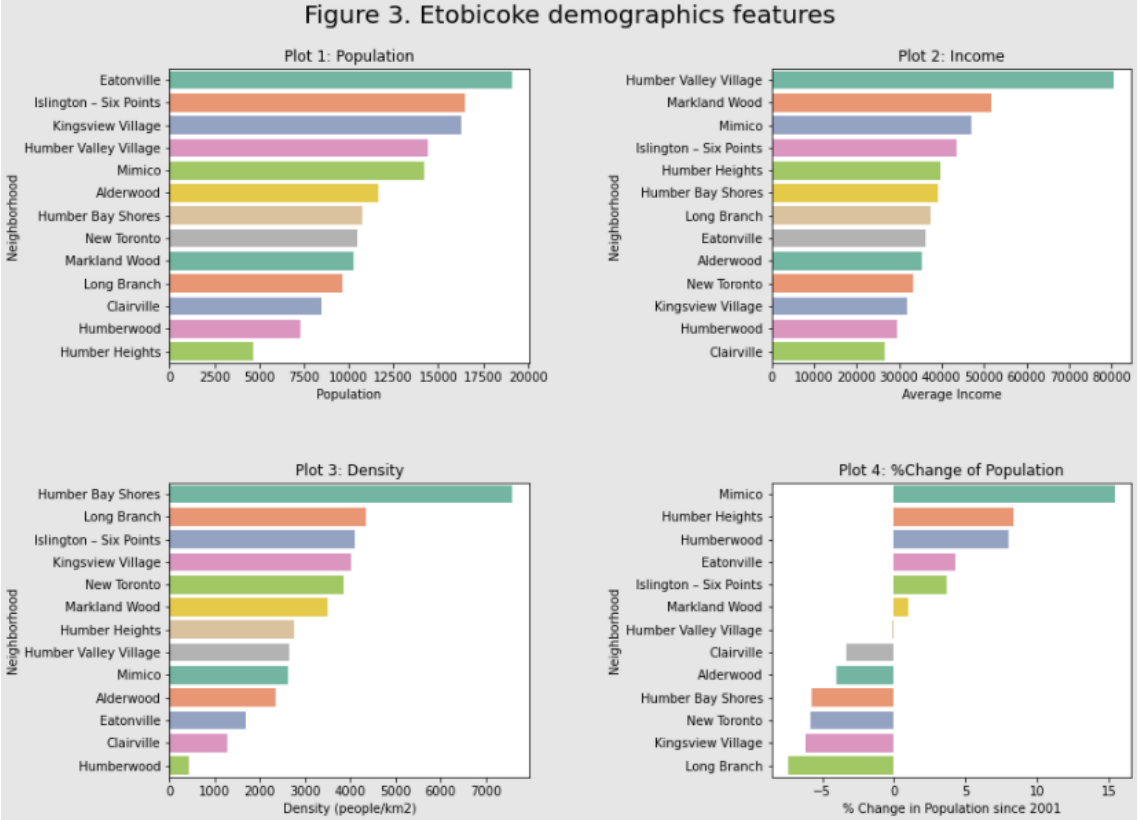

Figure 3. Main demographics features of the Etobicoke neighborhoods.

Figure 3 shown that the neighborhood with more population is Eatonville, the neighborhood with the highest average income is Mimico, the Humber Bay Shores neighborhood has the highest population density. These features will help us, join the Bank's venues information, to figure out what is the best neighborhood to open a new bank. That analysis is in the next modeling section.

2.5. Modelling

2.5.1. Machine learning algorithm

We will use the **clustering analysis** to group the neighborhoods with similar demographic characteristics (see Table 7), the clustering algorithm to use is **k-means**.

Because the demographic characteristics shown in Table 7 have different scales, before performing the cluster analysis we standardized these characteristics ('Population', 'Density (people/km2)','Average Income','% Change in Population since 2001','Number of Banks') using the Sckitlearn StandarScaler MinMax library (Table 8).

Table 8. Main demographics features standardized.

```
array([[0.48294944, 0.27054556, 0.15977263, 0.14912281, 0.33333333],
       [0.26506191, 0.11818055, 0.        , 0.17982456, 0.        ],
       [1.        , 0.17831729, 0.17767738, 0.51315789, 0.        ],
       [0.4220101 , 1.        , 0.23285439, 0.0745614 , 0.83333333],
       [0.        , 0.32719408, 0.2430751 , 0.68859649, 0.33333333],
       [0.18295635, 0.        , 0.05491779, 0.6754386 , 0.33333333],
       [0.67641973, 0.31128785, 1.        , 0.32017544, 0.66666667],
       [0.8185654 , 0.51416213, 0.31402755, 0.48684211, 1.        ],
       [0.80099606, 0.50118599, 0.09987409, 0.05263158, 0.33333333],
       [0.34246386, 0.54625366, 0.19771145, 0.        , 0.33333333],
       [0.3850038 , 0.43058462, 0.46446823, 0.36842105, 0.33333333],
       [0.65878121, 0.30807869, 0.37774033, 1.        , 0.83333333],
       [0.39987549, 0.47955909, 0.12599985, 0.07017544, 0.        ]])
```

2.6. Model evaluation

The most widely used technique to evaluate which is the ideal cluster number for the analyzed data set is the elbow plot, according to the elbow plot, the **ideal cluster number is 7** (Figure 4).

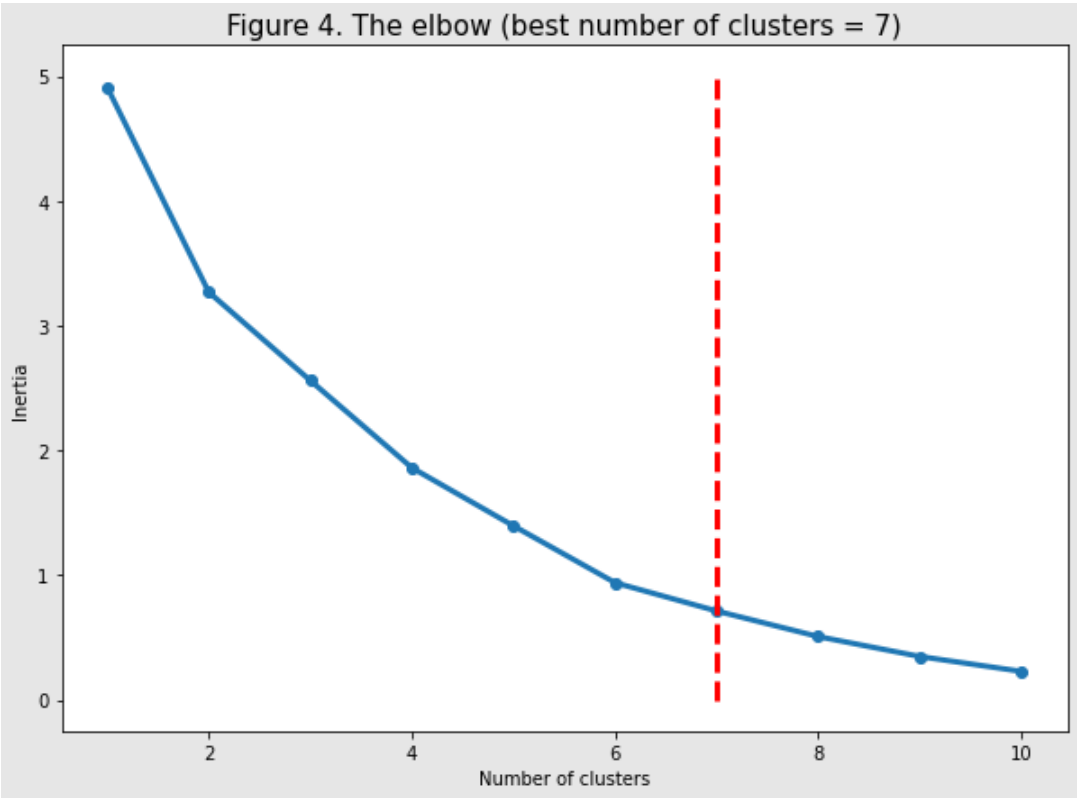

Figure 4. Cluster analysis elbow plot.

3. Results

3.1. Deployment

Once run the cluster algorithm k-mean on the standardized dataset we can see the results in Figure 5 (Income_scaler vs Density_Scaler). The graph shows the 7 clusters after running the k-means. The resulting cluster labels of the k-means process were added into a dataset, the results are shown in Figure 6 shows the cluster analysis on the Etobicoke neighborhoods, as we can see Figure 5 and Figure 6 are not very explicit alone; we must combine these results with the quantity venues bank located around the neighborhoods.

As we can see in Figure 7, Clairville and Eatonville neighborhoods don't have banks in a radius of 2000 m. Clairville belongs to cluster 0 and Eatonville belongs to cluster 5 (see Table 9). After we grouped de dataset by cluster (Table 10) we can confirm that around a radius of 2000 m there are not any banks venues.

Table 9

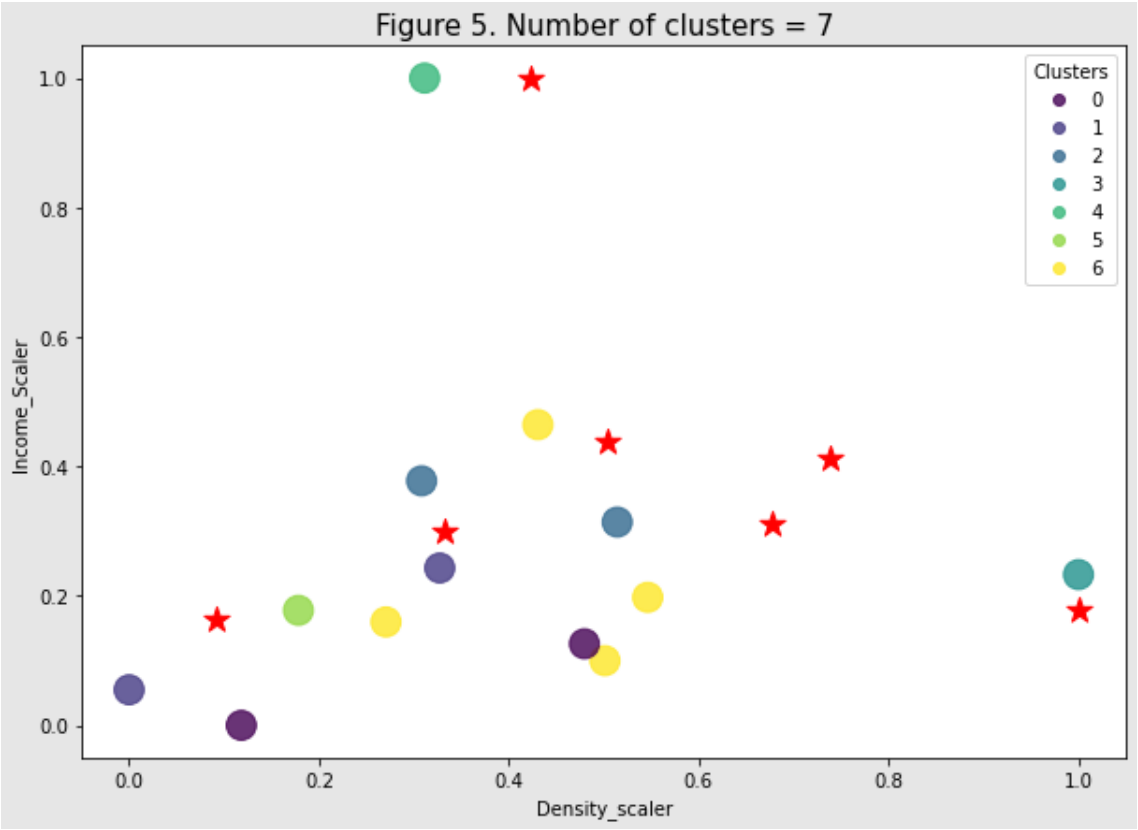

Figure 5. k-means Cluster analysis (Income_scaler vs Density_scaler)

Figure 6 shows the cluster analysis on the Etobicoke neighborhoods, as we can see Figure 5 and Figure 6 are not very explicit alone; we must combine these results with the quantity venues bank located around the neighborhoods.

As we can see in Figure 7, Clairville and Eatonville neighborhoods don't have banks in a radius of 2000 m. Clairville belongs to cluster 0 and Eatonville belongs to cluster 5 (see Table 9). After we grouped de dataset by cluster (Table 10) we can confirm that around a radius of 2000 m there are not any banks venues.

Table 9. Resulting dataset after adding the cluster labels.

	Neighborhood	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	Latitude	Longitude	ClusterLabels	...	1st Most Common Venue
0	Alderwood	11656	4.94	2360	-4.0	35239	8.8	43.604960	-79.541160	6	...	Coffee Shop
1	Clairville	8506	6.71	1268	-3.3	26610	13.2	43.748030	-79.631220	0	...	Intersection
2	Eatonville	19131	11.26	1699	4.3	36206	12.6	43.689581	-79.494751	5	...	Coffee Shop
3	Humber Bay Shores	10775	1.42	7588	-5.7	39186	15.7	43.626860	-79.476710	3	...	Park
4	Humber Heights	4674	1.69	2766	8.3	39738	10.1	43.652247	-79.486697	1	...	Bakery
5	Humberwood	7319	17.40	421	8.0	29576	7.9	43.725165	-79.621555	1	...	Coffee Shop
6	Humber Valley Village	14453	5.45	2652	-0.1	80618	12.0	43.641466	-79.492537	4	...	Coffee Shop
7	Islington – Six Points	16508	4.02	4106	3.7	43570	17.1	43.634870	-79.530520	2	...	Coffee Shop
8	Kingsview Village	16254	4.05	4013	-6.2	32004	11.8	43.702510	-79.572090	6	...	Coffee Shop
9	Long Branch	9625	2.22	4336	-7.4	37288	14.2	43.593540	-79.532750	6	...	Coffee Shop
10	Markland Wood	10240	2.92	3507	1.0	51695	9.2	43.633910	-79.569480	6	...	Coffee Shop
11	Mimico	14198	5.40	2629	15.4	47011	11.6	43.617290	-79.498850	2	...	Restaurant
12	New Toronto	10455	2.71	3858	-5.8	33415	13.0	43.601430	-79.509250	0	...	Park

Table 10. Dataset Groupby-means by cluster.

	ClusterLabels	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	Latitude	Longitude	Number of Banks
0	0	9480.50	4.7100	2563.0	-4.55	30012.5	13.10	43.674730	-79.570235	0.0
1	1	5996.50	9.5450	1593.5	8.15	34657.0	9.00	43.688706	-79.554126	2.0
2	2	15353.00	4.7100	3367.5	9.55	45290.5	14.35	43.626080	-79.514685	5.5
3	3	10775.00	1.4200	7588.0	-5.70	39186.0	15.70	43.626860	-79.476710	5.0
4	4	14453.00	5.4500	2652.0	-0.10	80618.0	12.00	43.641466	-79.492537	4.0
5	5	19131.00	11.2600	1699.0	4.30	36206.0	12.60	43.689581	-79.494751	0.0
6	6	11943.75	3.5325	3554.0	-4.15	39056.5	11.00	43.633730	-79.553870	2.0

We can affirm that clusters 0 and 5 meet the first conditions that we set ourselves when addressing the problem, and that is that we were going to choose neighborhoods where there were no nearby banks to avoid having competition from other banking agencies.

In Figure 8 we can see the whole of main demographics features grouped by clusters, and Figure 9 the same features only for our interest clusters, 0 and 5. As we can see cluster 5 has the highest values of demographics features in comparison to cluster 0.

Figure 6. Etobicoke Neighborhood Cluster Map

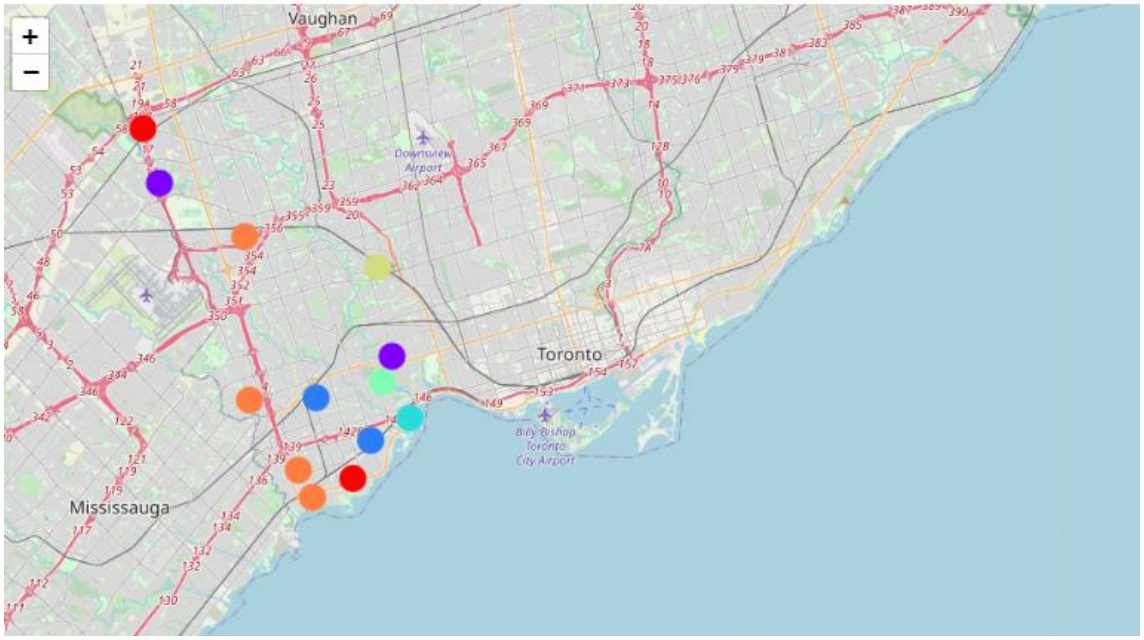

Figure 6. Cluster analysis on the map of Etobicoke neighborhoods.

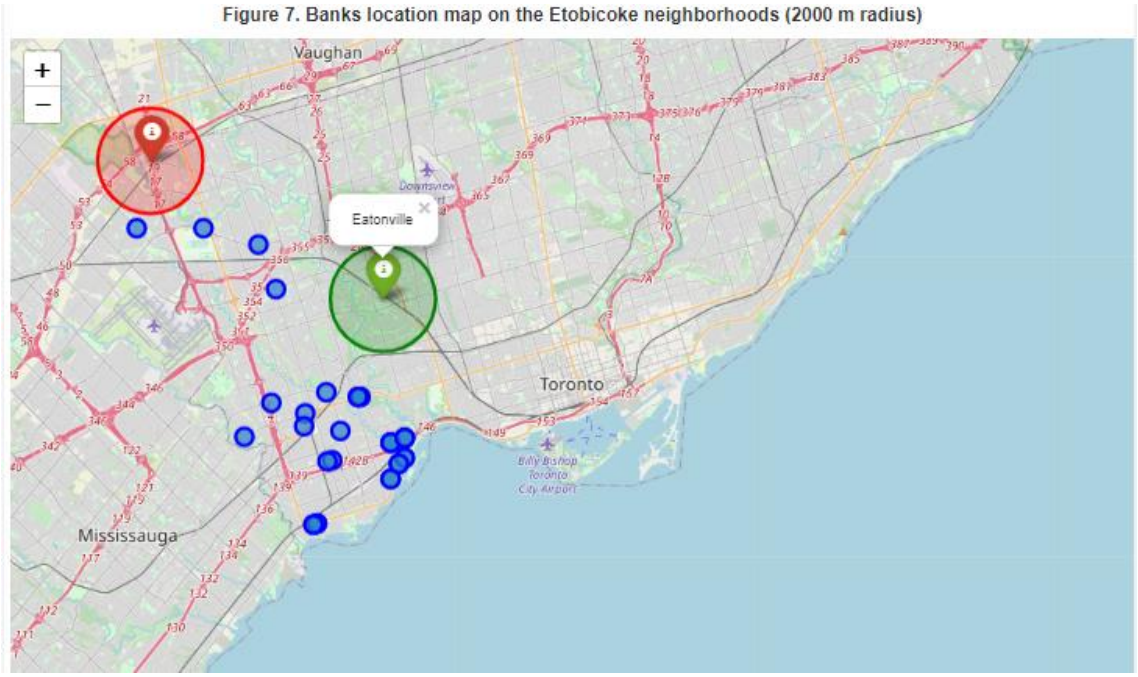

Figure 7. Banks location map on the Etobicoke neighborhoods (2000 m radius)

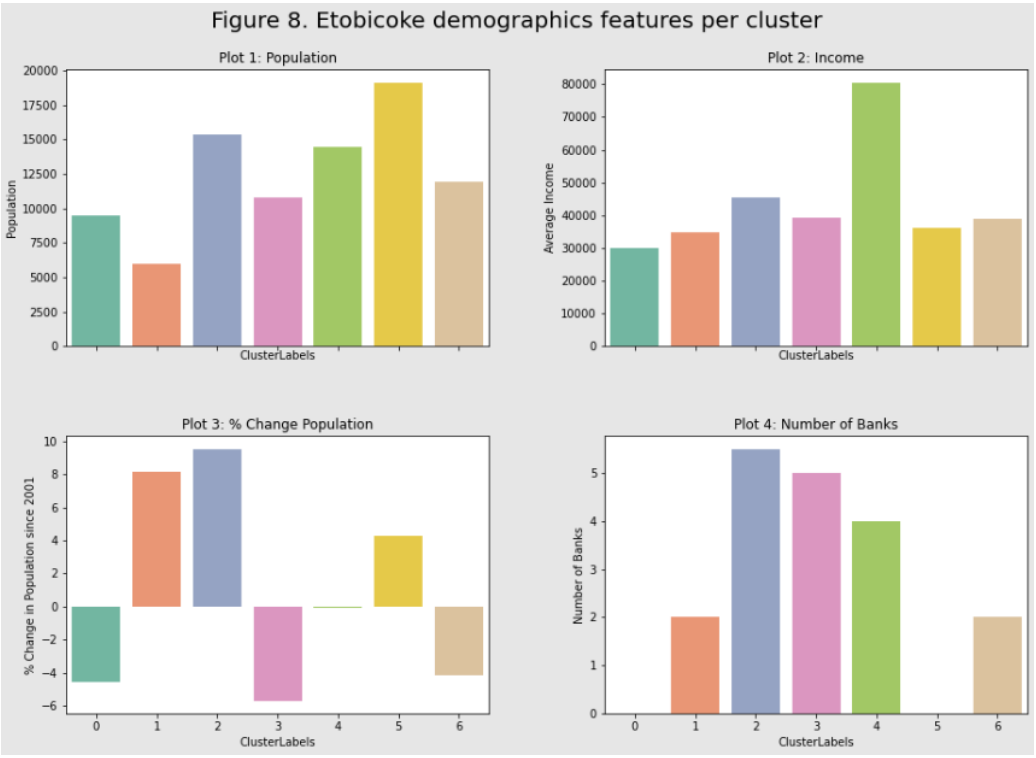

Figure 8. Main demographics features grouped by clusters (groupby.means)

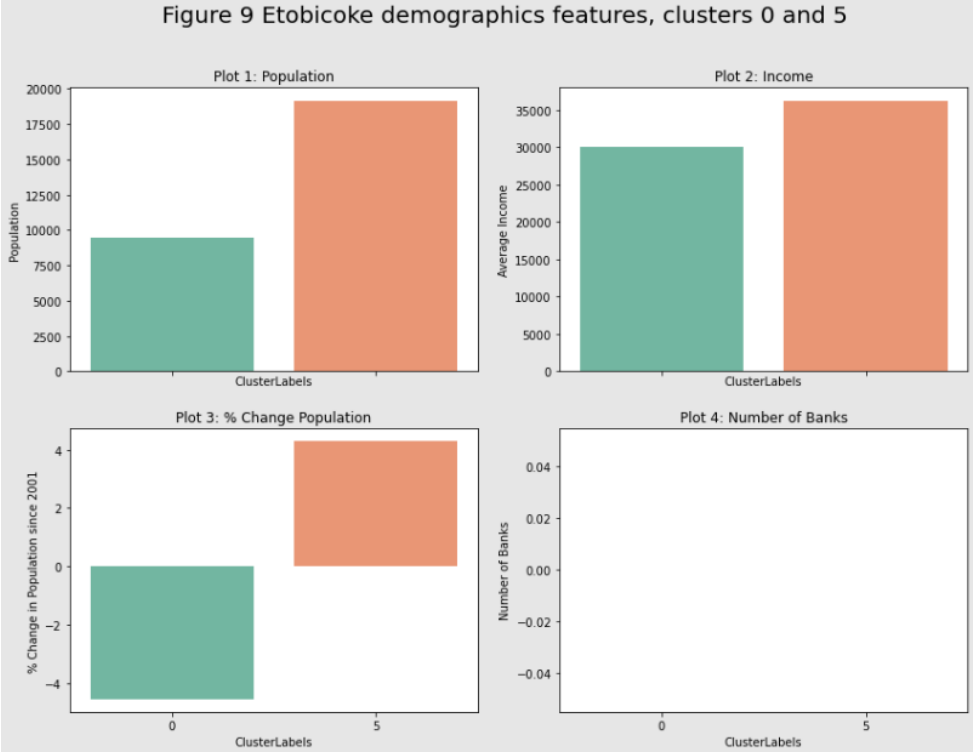

Figure 9. Main demographics features grouped by clusters (groupby.means), only cluster 0 and 5.

Finally, after filtering the data set only for clusters 0 and 5 (Table 11 and Table 12), we can see that the Eatonville neighborhood (**cluster 5**) is the one with the best demographic indicators and does not have banks within a radius of 2000 meters, therefore, taking these into account Indicators It is the right neighborhood to open a new bank.

Table 11. Dataset filter by cluster 0.

	Neighborhood	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	Latitude	Longitude	ClusterLabels
1	Clairville	8506	6.71	1268	-3.3	26610	13.2	43.74803	-79.63122	0
12	New Toronto	10455	2.71	3858	-5.8	33415	13.0	43.60143	-79.50925	0

Table 12. Dataset filter by cluster 5.

	Neighborhood	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	Latitude	Longitude	ClusterLabels
2	Eatonville	19131	11.26	1699	4.3	36206	12.6	43.689581	-79.494751	5

4. Conclusions

In cluster 5 there is only the neighborhood of Eatonville with a population of **19,131** inhabitants and an average income of **36,206**, it has also had a population growth of **4.3%** since 2001. In contrast, cluster 0 groups the neighborhoods of Clairville and New Toronto who are geographically distant (see Figure 7). In addition, both do not have a larger population or higher income than the Eatonville neighborhood, and additionally since 2001 their population has had a decrease of **4.55%.** So according to these analyzed variables **the new Bank office should be opened in the neighborhood of Eatonville, since there are no other banks around and it is the neighborhood with the best demographic indexes of the neighborhoods where there are no banks**.

5. Reference

[1] Toronto demographics (https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods).

[2] Foursquare API services (https://es.foursquare.com/developers/apps)

[3] Data science methodology (https://www.coursera.org/learn/data-science-methodology/home/welcome)