Real State App – Helping hand

Jose Antonio del Río

August 05, 2021

1. Introduction

In this project, we want to provide real estate agencies with an objective criterion that allows them to identify neighborhoods with similar characteristics in two cities without having to know any of them. This will allow them to have satisfied customers, as they will move to an environment with the expected characteristics. One of the questions that we will be able to answer with our application is in which neighborhoods of Toronto I will have to move to have characteristics similar to Williamsbridge.

To achieve our goal we are going to use a concept called UQI (Urban Quality Index).Since the beginning of the 20th century, many public and private administrations have worked to measure the quality of cities, neighborhoods, etc. For this, many indicators / indices have been created that allow quantitatively comparing different administrative units with each other. These indicators have several dimensions, some objective such as economic, social, services and other subjective such as the perception that citizens have of their city, neighborhood, etc.

In our project, we are going to create three objective dimensions Social, Economic and Services, leaving for other projects the inclusion of a subjective dimension that can be obtained by processing the evaluations of the clients of the places.

Nor are we going to create a UQI, as is usually created, weighting the dimensions and getting a numerical value, but rather we will apply Machine Learning to the dimensions, which we will use as input variables of an unsupervised “Clustering” algorithm.

To weight each of the three dimensions in each neighborhood, we are going to measure the number of places in a radius of 500 meters above the point where it is geolocated, that is, the density. For example, for the Economic dimension we will count how many places of an economic type “Foursquare” returns us.

However "Foursquare" does not manage these variables, so we will have to perform a previous manual task of classifying each one of the "categories" uniquely. For example, "Music Scholl" is labeled as "Social", Internet Café as "Budget" and "Taxi" as "Service".

Therefore, once each neighborhood in New York has the three dimensions weighted, we proceed to classify them into 5 groups as indicated above.

Now we have a dataset with three characteristics and a class. Therefore, we can create a model that allows us to predict the class.

Therefore, if we take the neighborhoods of another city and calculate the dimensions for them, we can infer what type of neighborhood belongs to the five that we have, and ultimately we can indicate to a client in New York which neighborhoods in Toronto have characteristics similar to Williamsbridge, such as we said at the beginning.

1.1 Problem

Data that could help to compare neighborhoods in different cities without having to know the cities beforehand. This project aims to indicate which neighborhoods in Toronto are similar to those of New York, in order to recommend to a client the purchase of a home or premises.

1.2 Interest

Real estate agencies are very interested in having an objective value that allows determining the characteristics of a neighborhood and that is also comparable with any city in the world.

Additionally, this would apply to public administrations that need to compare administrative units with a view to making investments.

2. Data acquisition

Below we detail the different data sets that have been created in order to have a data set for the New York neighborhoods with the three dimensions indicated above, and thus be able to clustering them.

This set of classified data will in turn allow us to create a classification model that we will apply to the neighborhoods of the city of Toronto. For this last city, we will also have to create a dataset with the same three dimensions as New York.

2.1 New York Neighborhood.

In order to get the New York neighborhood dataset

1.Download json file from: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

2.Dataframe named "newyork_neighborhoods" with geolocated data is created

	Unnamed: 0	Borough	Neighborhood	Latitude	Longitude
0	0	Bronx	Wakefield	40.894705	-73.847201
1	1	Bronx	Co-op City	40.874294	-73.829939
2	2	Bronx	Eastchester	40.887556	-73.827806
3	3	Bronx	Fieldston	40.895437	-73.905643
4	4	Bronx	Riverdale	40.890834	-73.912585

2.2 Foursquare Categories

In order to get the Foursquare Categories dataset:

1.Foursquare API is called on to get all categories that provide, the response is stored in a dataframe
: https://api.foursquare.com/v2/venues/categories?&client_id=XXX&client_secret=YYY&v=20180605&m=foursquare

2. It is exported to an excel and classified manually, each category is assigned a dimension ECO (Economic), SOC (Social), SER (Service). Finally the excel is imported again and saved in a dataframe called "categories_pos".

	id_category	Venue Category	index
0	56aa371be4b08b9a8d5734db	Amphitheater	SER
1	4fceea171983d5d06c3e9823	Aquarium	SER
2	4bf58dd8d48988d1e1931735	Arcade	ECO
3	4bf58dd8d48988d1e2931735	Art Gallery	ECO
4	4bf58dd8d48988d1e4931735	Bowling Alley	ECO

2.3 New York Venues

1. for each neighborhoods stored in the "newyork_neighborhoods" data frame, a search is performed to find the 500 nearest locations and stored in the "newyork_venues" data frame.
2. Finally we combine the dataframe "" to obtain the dataframe "result" where we include the dimension.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bronx	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Bronx	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Bronx	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Bronx	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
4	Bronx	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

	Unnamed: 0	Borough	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Category	Venue Latitude	Venue Longitude	index
0	0	Bronx	NewYork	Wakefield	40.894705	-73.847201	Lollipops Gelato	Dessert Shop	40.894123	-73.845892	ECO
1	1	Bronx	NewYork	Wakefield	40.894705	-73.847201	Rite Aid	Pharmacy	40.896649	-73.844846	ECO
2	2	Bronx	NewYork	Wakefield	40.894705	-73.847201	Walgreens	Pharmacy	40.896528	-73.844700	ECO
3	3	Bronx	NewYork	Wakefield	40.894705	-73.847201	Carvel Ice Cream	Ice Cream Shop	40.890487	-73.848568	ECO
4	4	Bronx	NewYork	Wakefield	40.894705	-73.847201	Dunkin'	Donut Shop	40.890459	-73.849089	ECO

2.4 New York Venues Grouped

The Venues are grouped by neighborhoods, making a sum in the corresponding dimension. The dataframe that we are going to use to perform the classtering will be named "newyork_grouped".

	Unnamed: 0	Neighborhood	ECO	SER	SOC
0	0	Allerton	25	4	0
1	1	Annadale	11	1	0
2	2	Arden Heights	3	1	0
3	3	Arlington	3	3	0
4	4	Arrochar	19	4	0
5	5	Arverne	14	7	0
6	6	Astoria	48	0	2
7	7	Astoria Heights	7	4	1
8	8	Auburndale	18	2	0
9	9	Bath Beach	47	3	0

2.5 Toronto Neighborhood

In order to get the Toronto neighborhood dataframe:

1. WebScraping is made on:
   "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"

2.Get Geospacial information from "https://cocl.us/Geospatial_data"

3.Create a Toronto_neighborhoods data frame joining both

	Unnamed: 0	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	0	M3A	North York	Parkwoods	43.753259	-79.329656
1	1	M4A	North York	Victoria Village	43.725882	-79.315572
2	2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494

2.6 Toronto Venues Grouped

The same procedure is carried out as in New York. This data frame will be used as input to predict what type of neighborhood you belong to in the model created with New York neighborhoods.

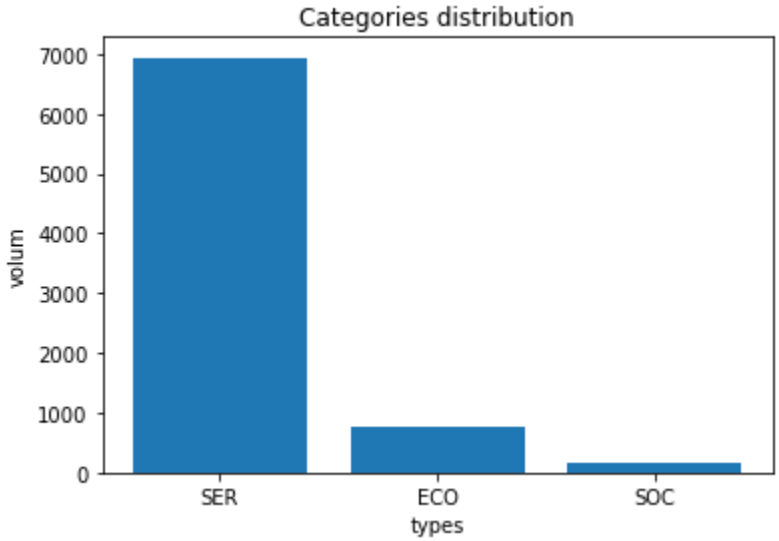
	Unnamed: 0	Neighborhood	ECO	SER	SOC
0	0	Agincourt	4	0	0
1	1	Alderwood, Long Branch	7	1	1
2	2	Bathurst Manor, Wilson Heights, Downsview North	22	1	0
3	3	Bayview Village	4	0	0
4	4	Bedford Park, Lawrence Manor East	24	0	0

3. Exploratory Data Analysis

In this section, we are only going to review the data of the categories provided by Foursquare, the rest of the data will be analyzed in the rest of the sections in order to contextualize them.

3.1 Category Distribution

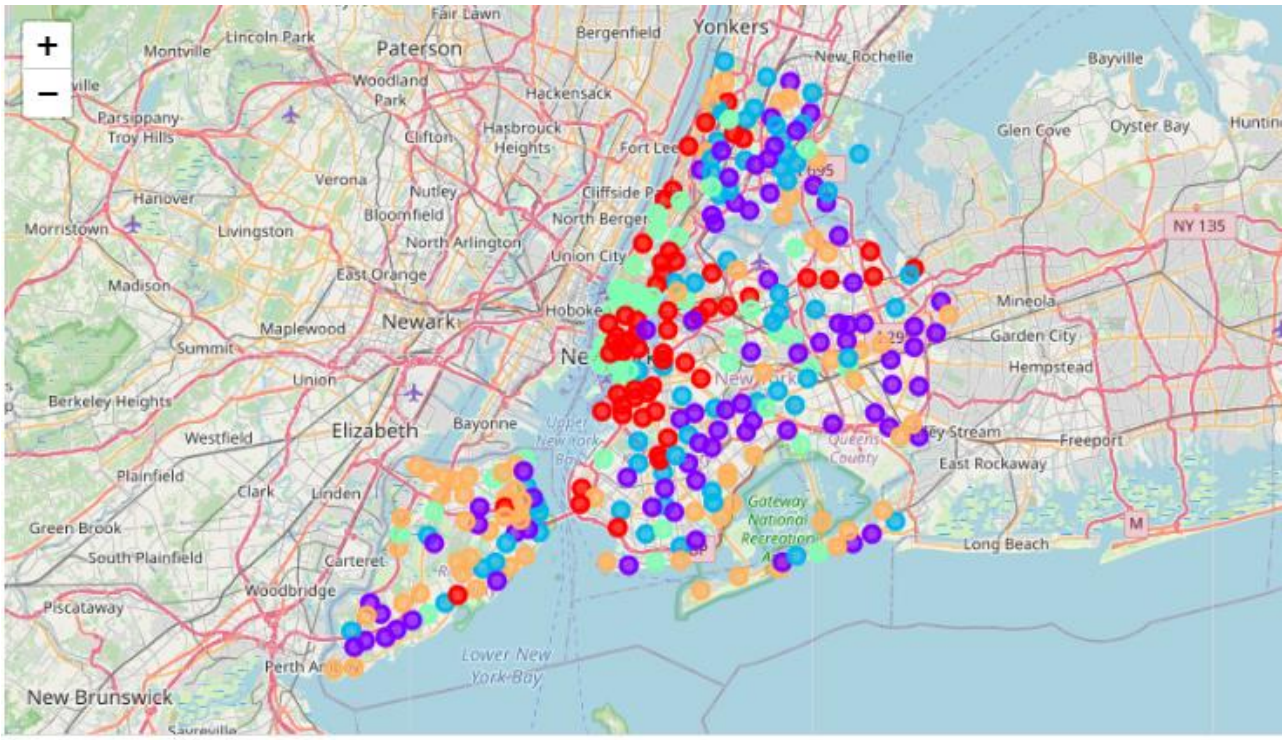
As you can see, most of the categorizations of the Foursquare places are of an economic type, while the size of the request and services is a minority. This is because it focuses on proximity businesses and leaves public services that have a better level of detail in the background.



4. Clustering Algorithm

K means that the algorithm has been selected to classify the New York neighborhoods. The value of K selected is "5", since it is a manageable and distinguishable value in the displays.

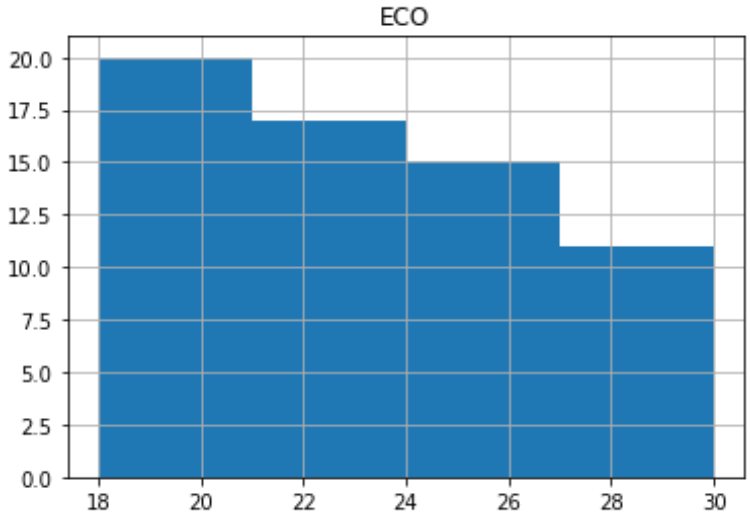
In the following map you can see how the neighborhoods of New York are classified.

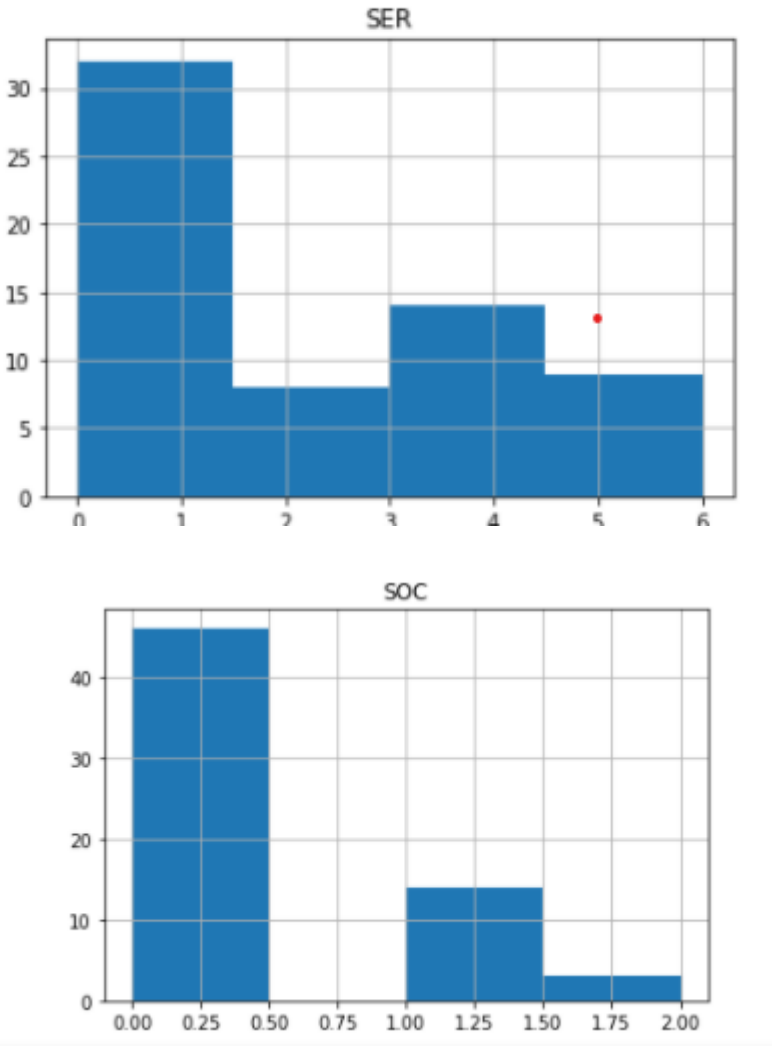


Below we show the detailed distributions of the 5 clusters in the three dimensions.

4.1 Cluster 0

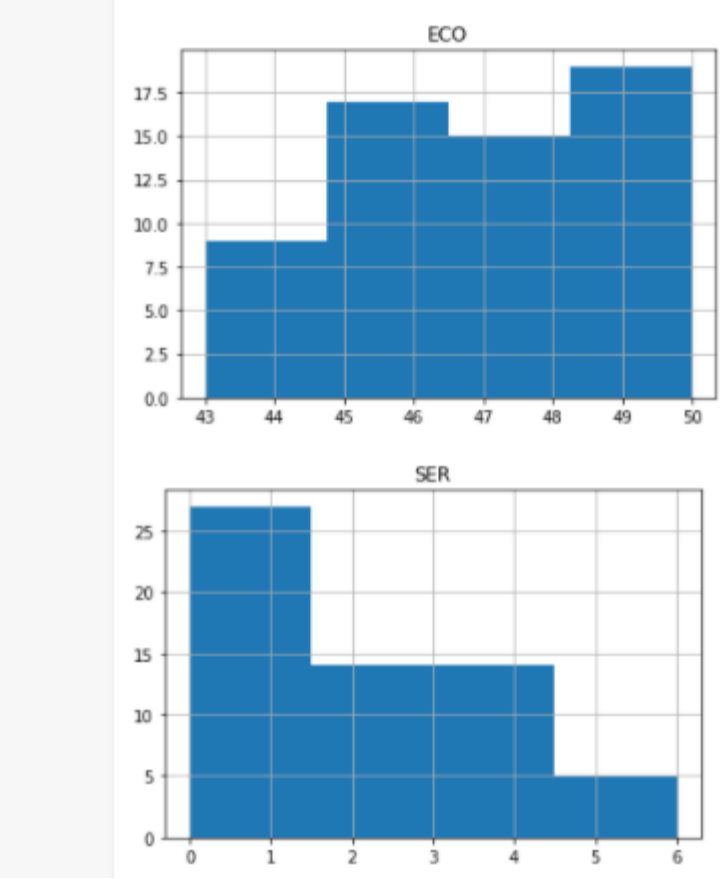
The neighborhoods that belong to "cluster 0" are neighborhoods with sufficient public and private services with acceptable infrastructures and enough shops at street level. If we qualify, it would be in the middle position "3".

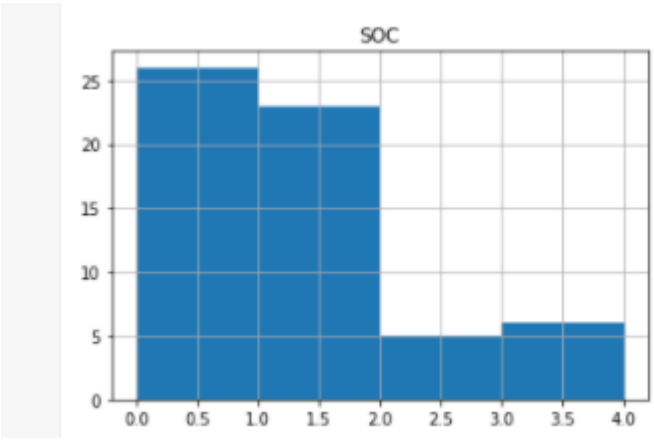




4.2 Cluster 1

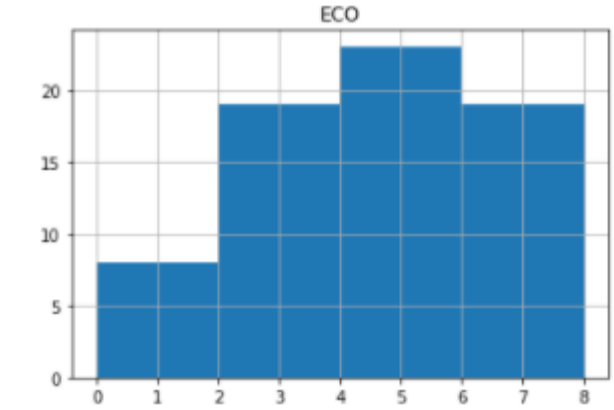
The neighborhoods that belong to "cluster 1" are neighborhoods with better public and private services with excellent infrastructures and many shops at street level. If we qualify, it would be in a high position "1".

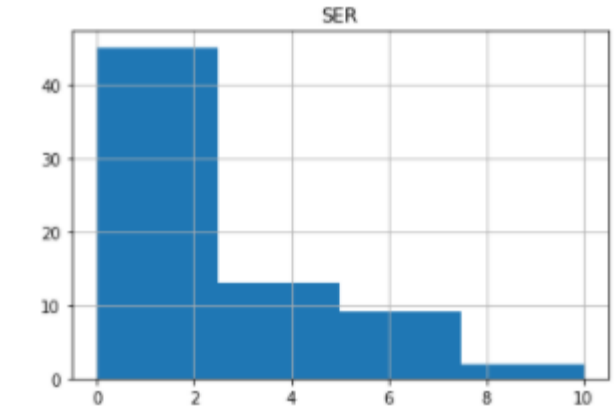


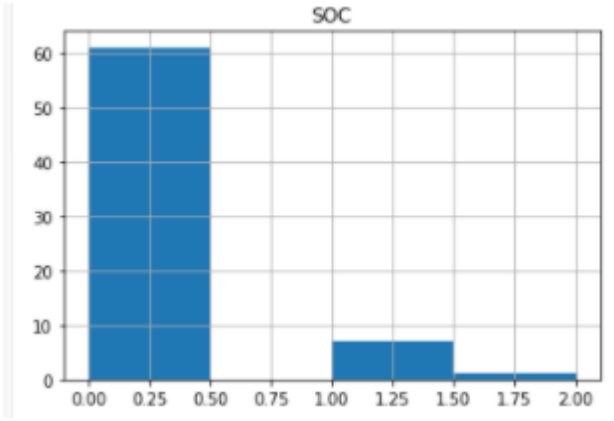


4.3 Cluster 2

The neighborhoods that belong to "cluster 2" are neighborhoods with few public and private services with little infrastructure and few businesses at street level. If we classify, it would be in position 5, that is, the worst of all.

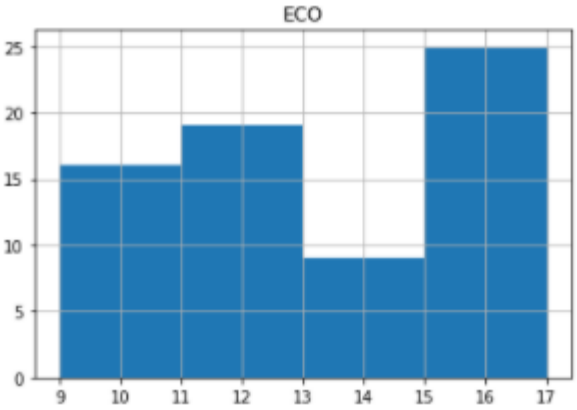


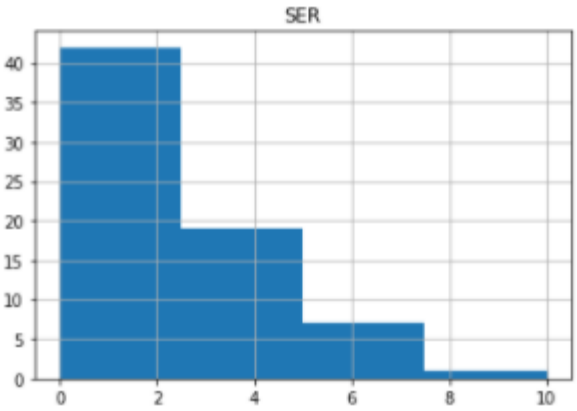




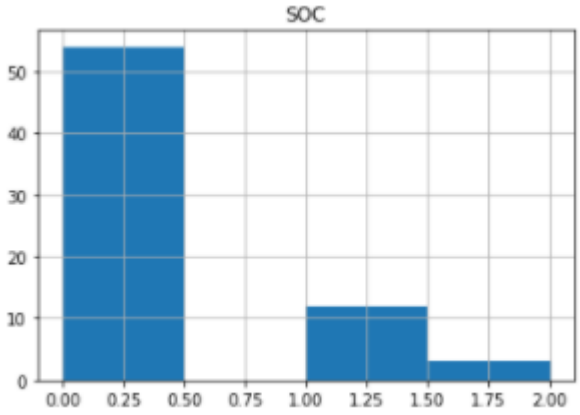
4.4 Cluster 3

The neighborhoods that belong to "cluster 3" are neighborhoods with few public and private services with little infrastructure and few shops at street level. If we qualify, it would be in position 4.



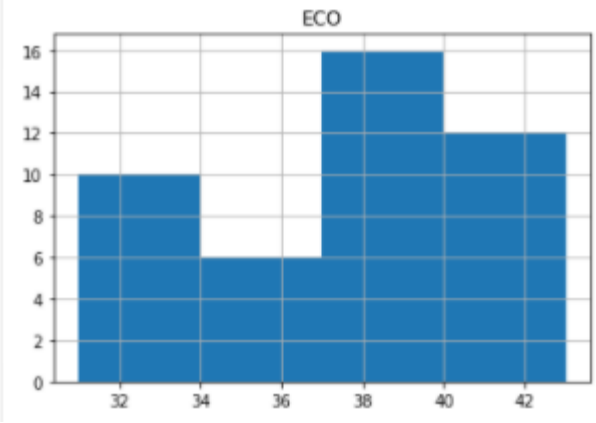


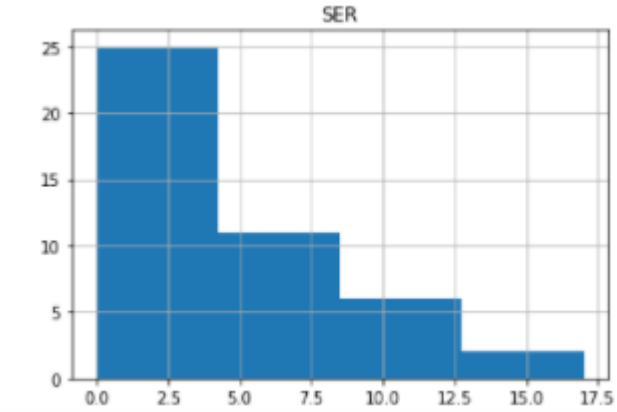
<

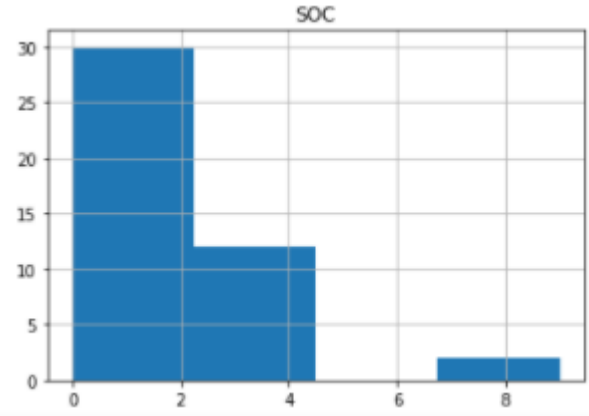


4.5 Cluster 4

The neighborhoods that belong to "cluster 4" are neighborhoods with high public and private services with abundant infrastructure and many shops at street level. If we qualify, it would be in a high position "2".



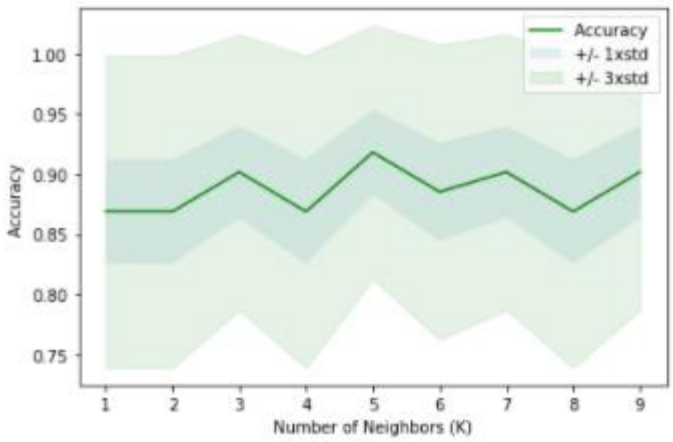




5. k-nearest neighbors algorithm

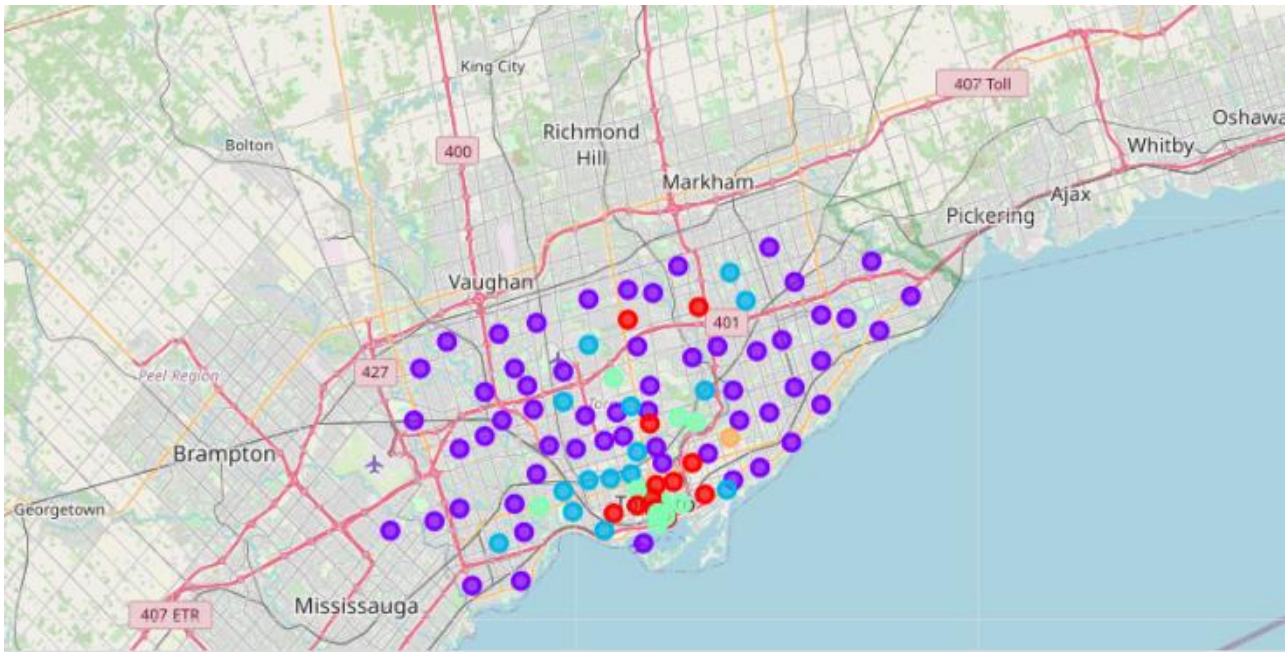
Once we have a set of classified data, what we have done is train a model using the "k-nearest neighbors" algorithm; this will allow us to identify similar neighborhoods in other cities. In our case, we are going to compare it with Toronto.

We have searched for the best Ks and in our case, it is "5" with accuracy “0.91”



Once we apply the algorithm to the Toronto data, we have classified them with the same criteria as in New York.

In the map below you can see how the neighborhoods of Toronto are classified.

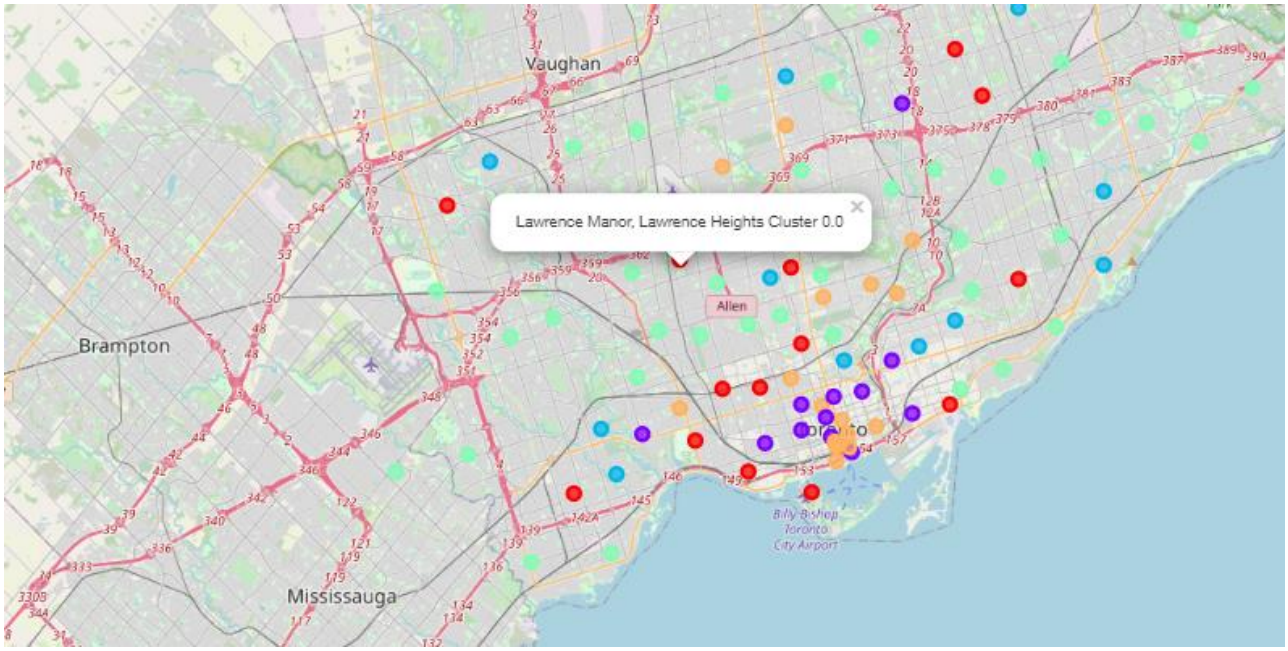


6. Conclusions

Our application helps us to homogenize the comparison of neighborhoods in different cities, both visually and with the two dataframes generated. (NewYork_print, Toronto_print).

It can be applied to many use cases. We are going to carry out a practical case so that it is better understood.

Client lives in the "North Riverdale" neighborhood and wants to see Toronto neighborhoods similar to this one. As it belongs to Cluster 0 we have to look for a neighborhood in Toronto inferenced with Cluster 0.


Vaughan
Lawrence Manor, Lawrence Heights Cluster 0.0
Allen
Brampton
Billy Bishop
Toronto
City Airport
Mississauga

7. Future directions

To improve the state of the art, progress could be made in several directions.

One of them would be to incorporate other data sources that allow us to add more dimensions, for example subjective, which are those that are based on the opinion that neighborhoods have about their neighborhood. Dimensions that take into account the social circumstances of the neighbors, unemployment, drugs, and so on could also be included.

Other algorithms could also be checked to see if they improve accuracy.

Finally, more cities could be included and thus expand the catalog of cities to be compared.