2019

Neighborhood scouting for real estate

Elie MAZE

July 25, 2019

Table of Contents

Introduction

The North American Privilege Estate company (NAPE), a real estate company, is specialized in the sale of high-standing properties in the United States and Canada.

This company is aimed at a very demanding clientele. It always operates in the same way; the main steps are as follows:

- define the customer's environmental needs in order to select and acquire the most appropriate land,
- define the architecture of the future property,
- etc.

Problem

Recently, a customer has used NAPE services. He wants to acquire a property in the North of the United States or in the southeastern Canada. After several interviews with NAPE, 4 cities were selected:

- Toronto,
- San-Francisco,
- New-York,
- Chicago.

It has passed on all of the client's criteria, the objective is to find the most suitable land to satisfy its client's needs:

- the crime rate must be as low as possible,
- the land must be close to the greatest number of services and medical facilities,
- the property should be as close as possible to any transport venue,
- In addition, the client likes to be entertained and to watch sports: he wants to enjoy a wide variety of restaurants, be able to go out at night and be as close as possible to stadiums and sports complexes, monuments and cultural sites,
- finally, he likes to swim and run so the housing must be near a pool or green areas.

The future buyer is an antique collector. The proximity of an antique store would be an asset. Moreover, as he possesses a valuable collection, he would like to move into an area where the number of burglaries is as low as possible.

As the clientele of this company has a very high purchasing power, the price of land or properties is not considered.

The highest priority is defined by the criteria mentioned first (crime rate, transport, restaurants...). The second priority concerns the criteria relating to his hobby, namely antiques.

Interest

The company uses different specialists to complete all phases of the project. As mentioned earlier, one of them is to buy the land that best suits the needs of the client. As it operates over a very large geographic extent, it is not able to explore all terrains and find the one that best meets the needs of the future buyer in a reasonable time.

That’s why NAPE has called on us to complete this phase of the project. The use of data science will help NAPE to find out the top-5 neighborhoods that are the best compromise

regarding the client's criteria and in a relative short time. ANPE will then investigate in deeper this small area in order to determine the best terrain.

Data

As criminality and certain kind of venues are part of the criteria, we need to collect the crime rate and the venues categories for each neighborhood of the 4 cities. We will use the Foursquare API to identify venues within a given radius from the centroid of each neighborhood.

Because we do not have access to ready-to-use data, we need to grab data from different sources.

We will collect the crimes that were reported in each city in 2018. As the crimes are not always grouped by neighborhood, we will count the number of crimes occurring within a given radius from the centroid of each neighborhood (very time consuming). And fortunately, the latitudes and longitudes are available for each dataset.

Data source

First of all, we need to collect all the neighborhoods of the 4 cities and the geographical locations. The list of the neighborhoods can be found on Wikipedia and the locations are retrieved using the module Geopy. Except for New-York City and Toronto since we get data from the resources available for the notebooks of Coursera.

Most crime rate can be fetched from Kaggle. For Toronto, we find data on the Toronto police's data portal.

Data	City	Source	Details
Neighborhood list	NYC	JSON file	Coursera notebook
Neighborhood location	NYC	JSON file	Coursera notebook
Venues names & categories	NYC	Foursquare	
Crime rate	NYC	Kaggle	Link
Neighborhood list	Toronto	Wikipedia	Link
Neighborhood location	Toronto	CSV file	Coursera notebook
Venues names & categories	Toronto	Foursquare	
Crime rate	Toronto	Toronto police's data portal	Link
Neighborhood list	Chicago	Wikipedia	Link
Neighborhood location	Chicago	Geopy	
Venues names & categories	Chicago	Foursquare	
Crime rate	Chicago	Kaggle	Link
Neighborhood list	San-Francisco	Wikipedia	Link
Neighborhood location	San-Francisco	Geopy	
Venues names & categories	San-Francisco	Foursquare	
Crime rate	San-Francisco	Kaggle	Link

Data preparation

Collecting and cleaning

Data is stored in dataframes. At the beginning, there is one dataframe per city so we have 4 dataframes for the neighborhood list, 4 for the crimes and 4 for the venues.

Neighborhoods

After collecting the neighborhoods and locations, missing values are dropped. Missing values are:

- neighborhood names containing NaN or the mention "unassigned",
- no match found for the neighborhood address using Geopy, so the latitude and longitude are NaN.

Crimes

The dataframes containing the crimes data have also missing values. Sometimes the date of occurrence is missing or inconsistent (example: 03-18-1019 instead of 03-18-2019?). Rows with such anomalies are removed from the dataframe.

Then, the crimes data are filtered, we keep only crimes that occurred in 2018.

Finally we generate 2 dataframes:

- One unfiltered dataframe that contains all crimes occurring in 2018,
- The second filtered so that it contains only the burglaries. It will be treated later as it is a part of the second priority.

Venues

We use the Foursquare API to get the top 100 common venues within a radius of 500 meters from the neighborhood locations. We obtain a dataframe with the name and the category of the venues found for each neighborhood. This dataframe will be filtered later to extract the antique shops. Again, it will be processed in a second time as it is the second priority for the customer.

Data processing

Crimes

For each neighborhood, we compute the distance between the location of the neighborhood and the locations of all the crimes that occurred in the city in 2018. Then we filter the list to keep only distances less than or equal to 500 meters. Eventually we count the number of crimes. Finally, we get a single dataframe per city that contains the name of the neighborhoods with the latitudes and longitudes and the number of crimes that occurred at a maximum distance of 500 meters from each neighborhood position (centroid).

	Neighborhood	Latitude	Longitude	Crimes
0	Alamo Square	37.776357	-122.434694	1845
1	Anza Vista	37.780836	-122.443149	1230
3	Balboa Park	37.724949	-122.444805	580
5	Bayview	37.728889	-122.392500	1277
6	Belden Place	37.791744	-122.403886	5032

Figure 1: Preview of the San-Francisco dataframe.

Venues

As criteria are mainly based on some specific categories of venues, we need to explore all the categories and try to build a mapping in order to gather the categories per group (or final category). We extract the venue categories from the dataframe of each city and remove the duplicates. Then we manually analyze them to normalize the categories we are interested in.

We found more than 500 unique venue categories that we will group into 8 final categories. For example Indian restaurant, French restaurant, Sandwich Place will correspond to the category "Food". The final categories are:

Final category	Description
Medical Care	All venue categories related to health (medical center, hospital, drugstore, doctor's office…).
Transport	All the transport venues: train station, bus, metro… We only consider the intra-city transport, so boat/ferry and planes are ignored.
Food	The venues were the customer can eat (various kinds of restaurants, BBQ places, Hot dog/Hamburger places, etc.).
Nature-Sport	This category gather all the venues related to green areas (outdoors, parks…) or the swimming pools.
Culture	The cultural venues, it could be monuments, museums, comedy clubs, Opera…
Entertainment	The venues where the customer will attend sport games (stadiums …).
Services	This category gather the most common services like the post office, bank or insurance.
Other	All other venue categories.

Note: The final category "Other" is useless but we kept it to understand the distribution of categories.

Feature selection

Some criteria are related to the proximity of certain places or the variety of the venue categories, so we will compute the average distance between the neighborhoods and certain venues or count the different venue categories, according to the final category.

The steps are:

1. Group the rows of the dataframe by neighborhood and:
   - Count the distinct venue categories when venues have the final category "Food",
   - Count the distinct venue categories when venues have the final category "Night-Entertainment".

   At the end of this step, we have 2 features: food and night-entertainment.

2. Compute the average distance between each neighborhood location and the venues locations belonging to a specific final category in order to create the following features:
   - Service_distance,
   - Stadium_distance,
   - Medical_distance,
   - Culture_distance,
   - Nature-Sport_dist.

3. Compute the minimum distance between each neighborhood location and the venues related to transportation.

4. Append the crimes to get the 9th feature.

Eventually, we have 4 dataframes (4 cities) with the following columns:

Feature name	Description
Neighborhood	Name of the neighborhood.
Crimes	Number of crimes occurring within a radius of 500m from the neighborhood centroid.
Food	The number of distinct kind of places to eat.
Night-Entertainment	The number of distinct kind of places to have fun during the night
Transport_distance	The average distance between the neighborhood centroid and all the intra-city transportation.
Service_distance	The average distance between the neighborhood centroid and all the services in the city.
Stadium_distance	The average distance between the neighborhood centroid and all the services in the city.
Medical_distance	The average distance between the neighborhood centroid and all the services in the city.
Culture_distance	The average distance between the neighborhood centroid and all the services in the city.
Nature-Sport_dist	The average distance between the neighborhood centroid and all the services in the city.

Then the 4 dataframes are merged into a single one.

	Food	Night-Entertainment	Neighborhood	Latitude	Longitude	Crimes	\
0	4	0	Wakefield	40.894705	-73.847201	91	
1	5	1	Co-op City	40.874294	-73.829939	75	
2	6	3	Eastchester	40.887556	-73.827806	60	
3	0	0	Fieldston	40.895437	-73.905643	7	
4	1	0	Riverdale	40.890834	-73.912585	24	

	Transport_distance	Service_distance	Stadium_distance	Medical_distance	\
0	24242.34	22450.49	26122.66	21794.76	
1	22961.58	21052.71	24889.13	20506.10	
2	24145.45	22298.05	26088.19	21745.23	
3	23653.48	21944.76	25205.44	20983.67	
4	23244.81	21548.39	24742.11	20525.05	

	Culture_distance	Nature-Sport_dist	City
0	22135.47	24207.80	New-York City
1	21077.66	22959.86	New-York City
2	22328.80	24209.00	New-York City
3	20362.15	23033.64	New-York City
4	19744.63	22509.59	New-York City

Figure 2: overview of the dataframe with the 9 features plus the neighborhoods information.

Methodology

Humans cannot explore manually a dozen variables for hundreds of neighborhoods in order to find the top-5.

The approach here is to transcribe the customer's criteria as features for all neighborhoods (already done) and apply clustering to split the neighborhoods into groups.

The second step consists in comparing the clusters and identify the group of neighborhoods that best fits the needs of the customer then we should have a smaller number of neighborhoods to analyze.

The last step is to compute the features related to the criteria defined by the second priority of the client: the number of burglaries and the distance to the closest antique shop. The selection of the top-5 neighborhoods will be based on these features.

Feature analysis and processing

Figure 3 shows the distribution of each feature. Our first observations are:

(1) We observe that the distributions of the 9 features look like log-normal distributions. ML model works better with normal distributions, furthermore Gaussian distributions are more appropriate for the next function which will try to help us to select a good number of clusters. So, we will apply a "np.log1p" transformation to all features in order to get 'gaussian-like' distributions.

(2) The value ranges of the 9 features are very different. The selection of the best neighborhoods is based on these features. As we want to consider all of them with the same importance and because the comparison of these features in a graph is not comfortable when the value range are very different, we will scale data between 0 and 1 with the MinMaxScaler of Scikit-Learn.

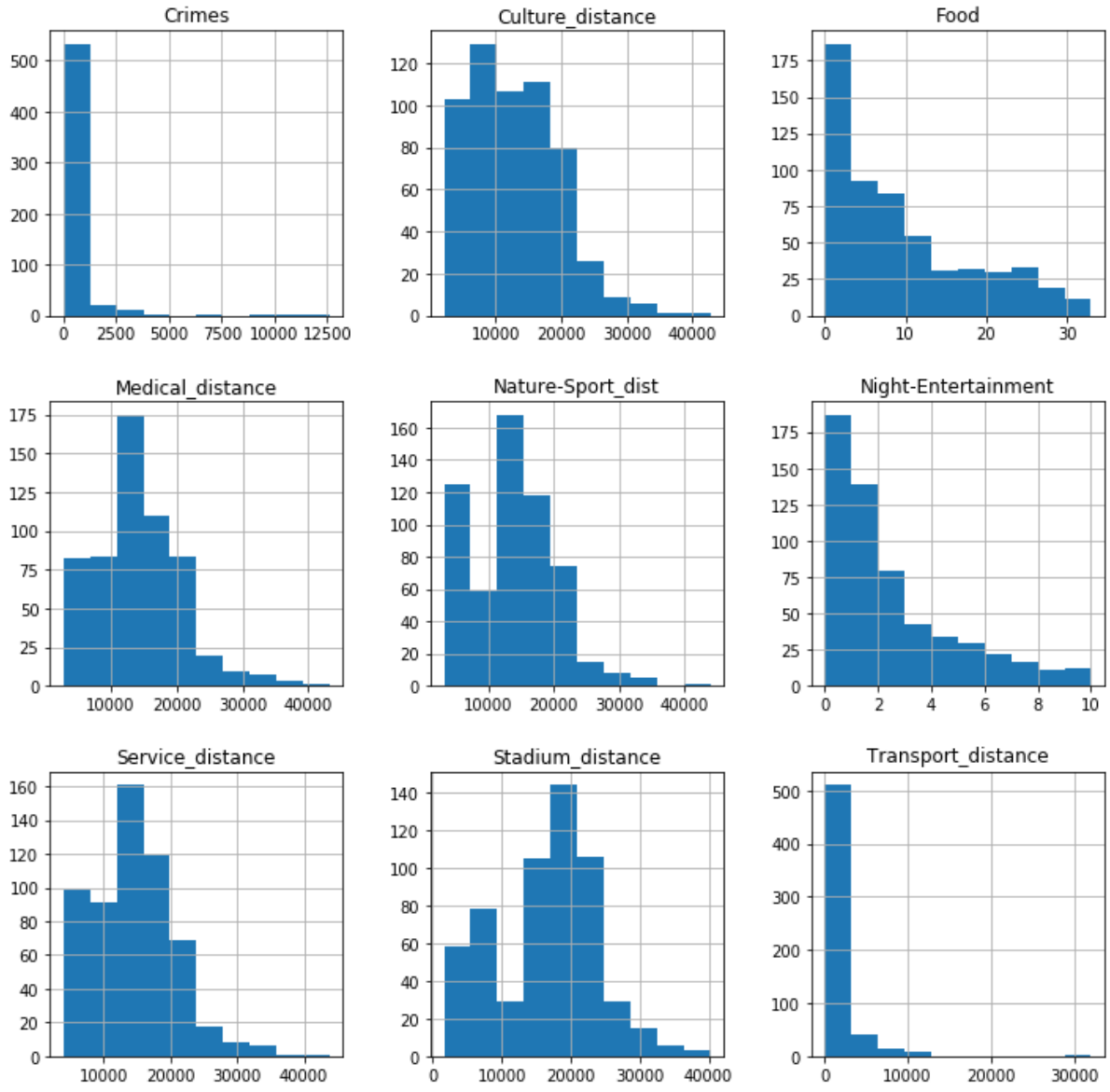

Figure 3: distribution of each features.

As we want to compare all criteria independently and give them the same importance, we scaled all the features between 0.0 and 1.0 so that they are in the same range.

To identify the best cluster, we must find the one that maximize the features:

- Food
- Night-Entertainment

And minimize the features:

- Crimes
- Medical / Nature-Sport / Service / Stadium / Transport / Culture distances

To make this operation easier, especially when representing features on a graph, we will apply the 1-x transformation (after the scaling) to the 7 features that we wanted to minimize. After this operation, we will just have to maximize all features.

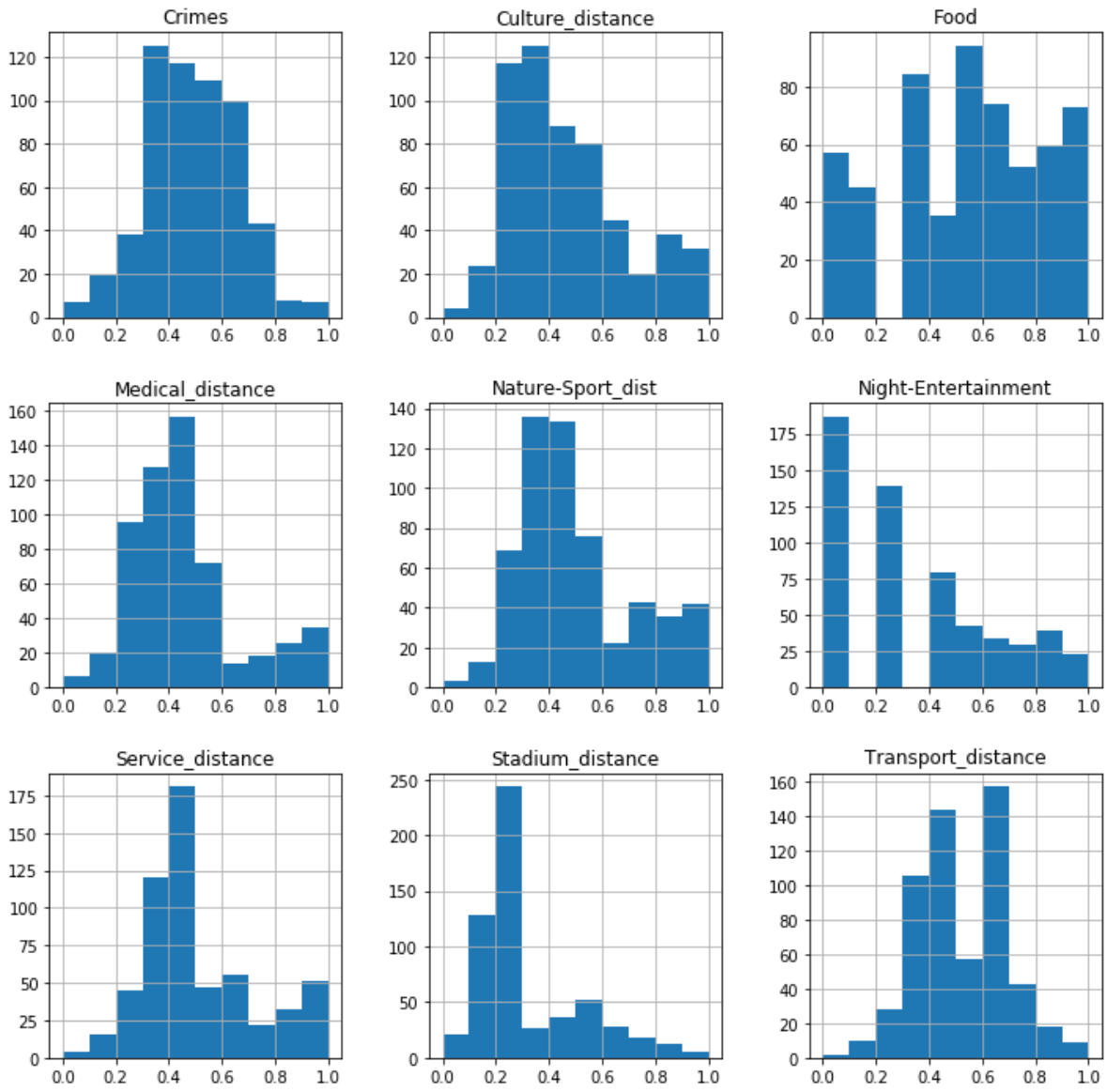

Figure 4: distribution of the features after the transformations.

The log1p transformation do not give us the expected results for all features, but at least it does the job for some of them like Crimes or Culture_distance.

Clustering and cluster selection

We use the K-means algorithm to apply clustering. Then will compare the average characteristics of each cluster to determine the best group.

Clustering

We did the clustering with a number of clusters (k) equals to 3. Then we repeated the operation for different values of k: [3-11].

By clustering neighborhoods, we expect to get distinct distributions for each feature. For example: if we set k=3, we want 3 well separated distributions for each feature.

To select the best value of k, we chose a "naïve" approach that consists in:

- Maximizing the delta between the averages of the distributions of the clusters to have well separated distributions.
- Minimizing the standard deviation (these are approximately gaussian distributions) to have less spread distributions.

We compute indicators on the 9 features for each value of k and we represent the indicators as curves in a radar chart (see figure 5).

Then we identified the curve which maximize all indicators. The best compromise was found by calculating the area under each curve and by selecting the one which have the highest area.

We chose k=4 and restarted the clustering. The neighborhoods were split into 4 groups.

We then draw the distribution of the 9 features for each value of k, i.e. we have 4 curves per feature (figure 7).

The use of 4 groups improve a little bit the clustering. The clustering is not perfect but the distributions are quite well separated (see graph Nature-Sport_distance) even if we observed some overlapping (Culture_distance, group 1 & 3) or bi-modal (Night-Entertainment, group 1 & 2) distributions.

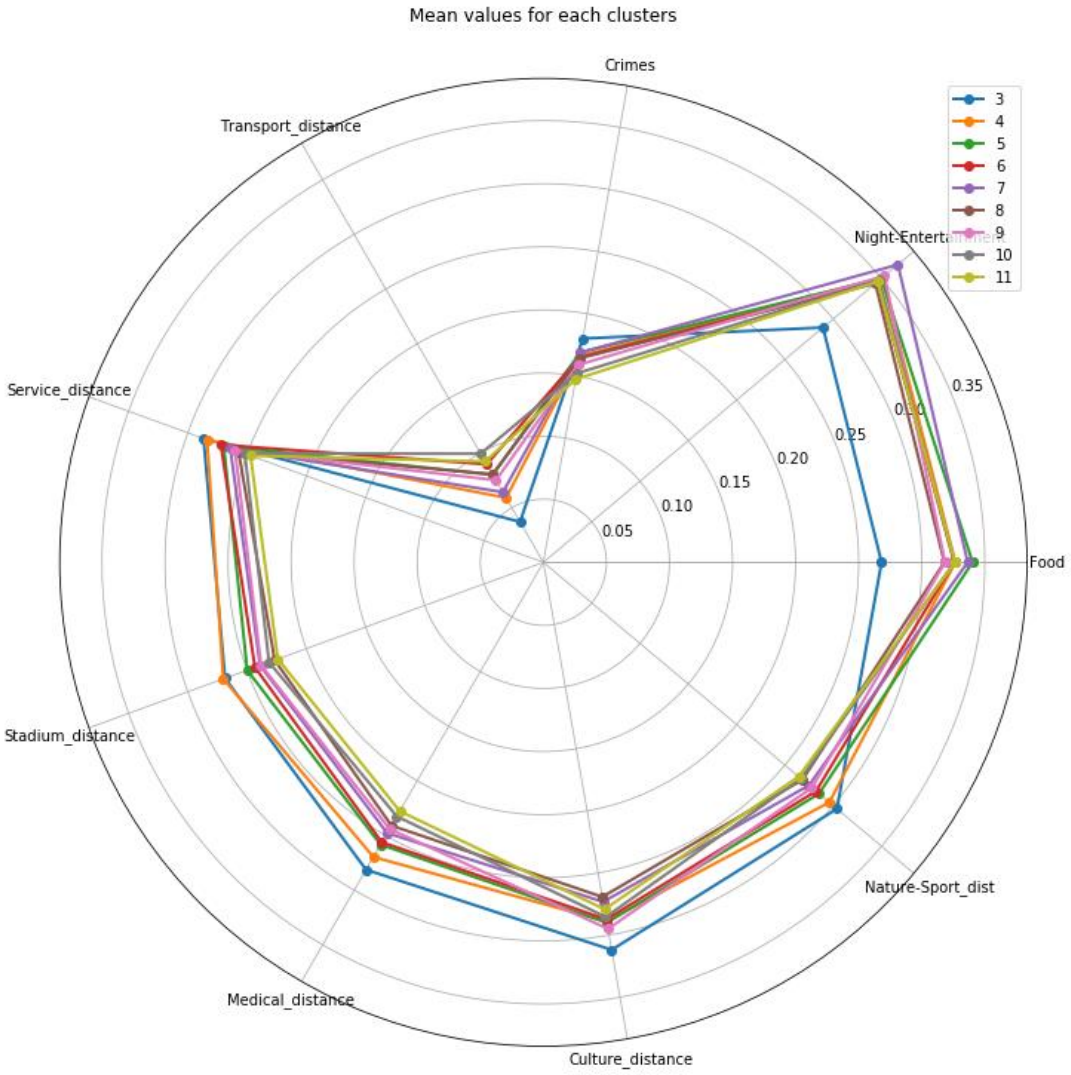

Figure 5: Each curve represents the indicators computed for each feature for a given k.

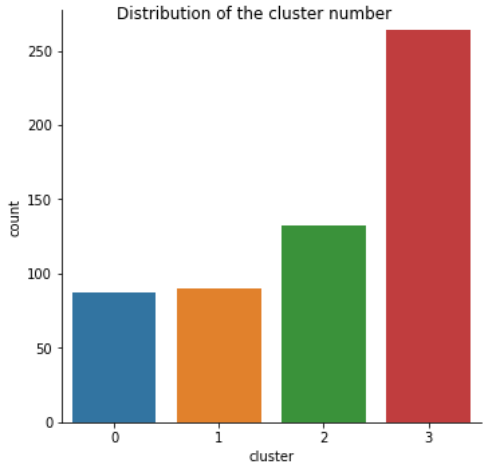

Figure 6: Distribution of the clusters.

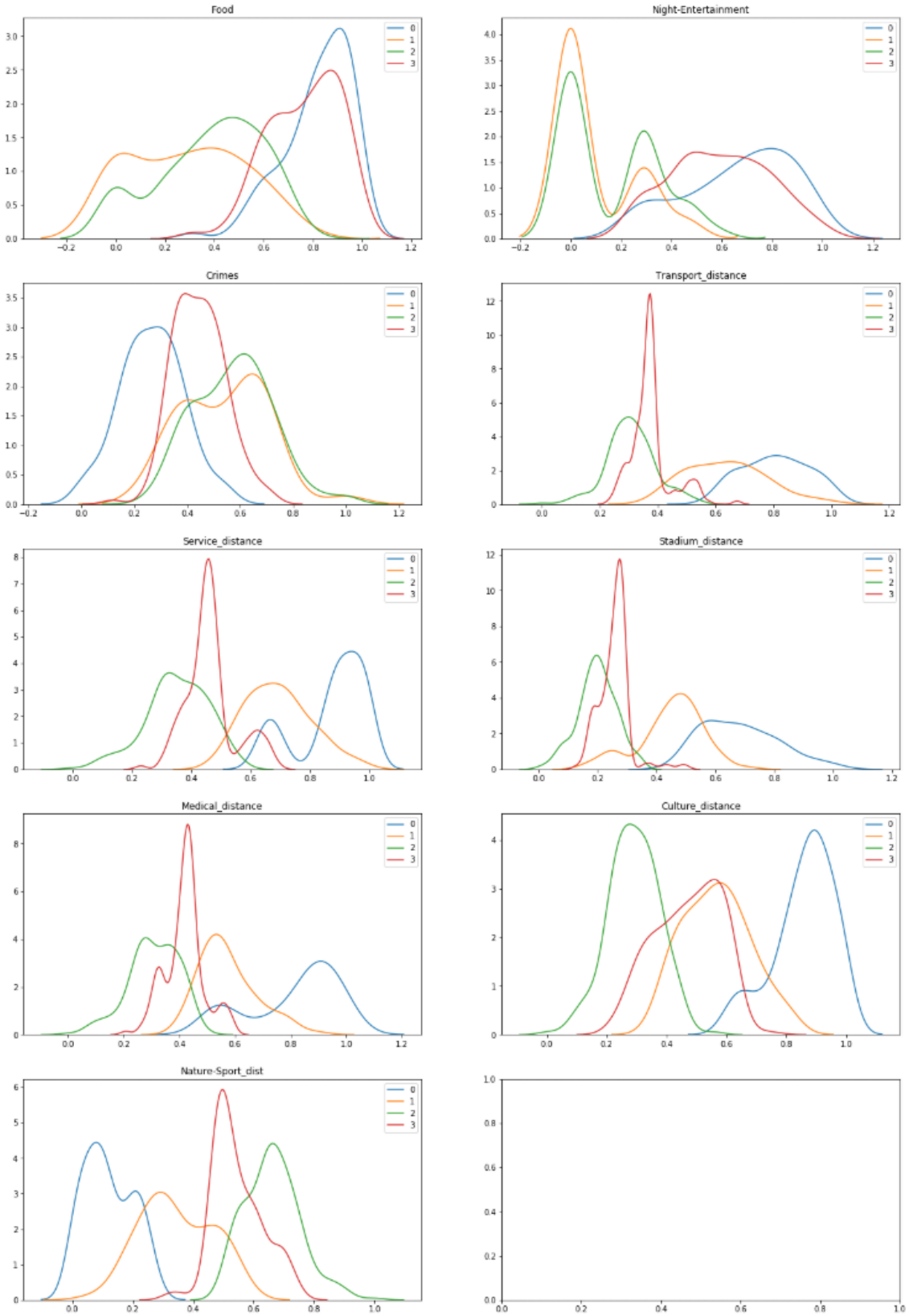

Figure 7: Distribution of the 9 features. Each curve is related to a cluster [0-3].

Cluster selection

As the best cluster should maximize each feature, we would like to identify a cluster whose distributions are always on the right of each graph.

Of course, we observe that is not possible, the ideal which would satisfy all the criteria does not exist but at least we can identify the best compromise.

It seems the blue curve (group 0) have distributions that are the closest to one 7 times. Not so bad, but let's try with radar chart method to confirm this 1st observation.

Figure 8 represents for each cluster the average values computed on each feature.

Figure 8: mean values computed on the 9 features for each cluster.

We tried to identify the best cluster by computing the area under the curves. With this approach and considering our observations of the spider chart:

- The best cluster seems to be group 0 (blue curve). The disadvantage is that the crime rate is more important in the neighborhoods of cluster 0.
- The worst group (according to the client's criteria) is cluster 3 (red curve).

We decide to select the group 0 as the best cluster.

The next figure shows the distribution of the clusters for the 4 cities. Our first observation is that cluster 0 is only present in Toronto and San-Francisco. So, our first conclusion is to exclude NYC and Chicago.

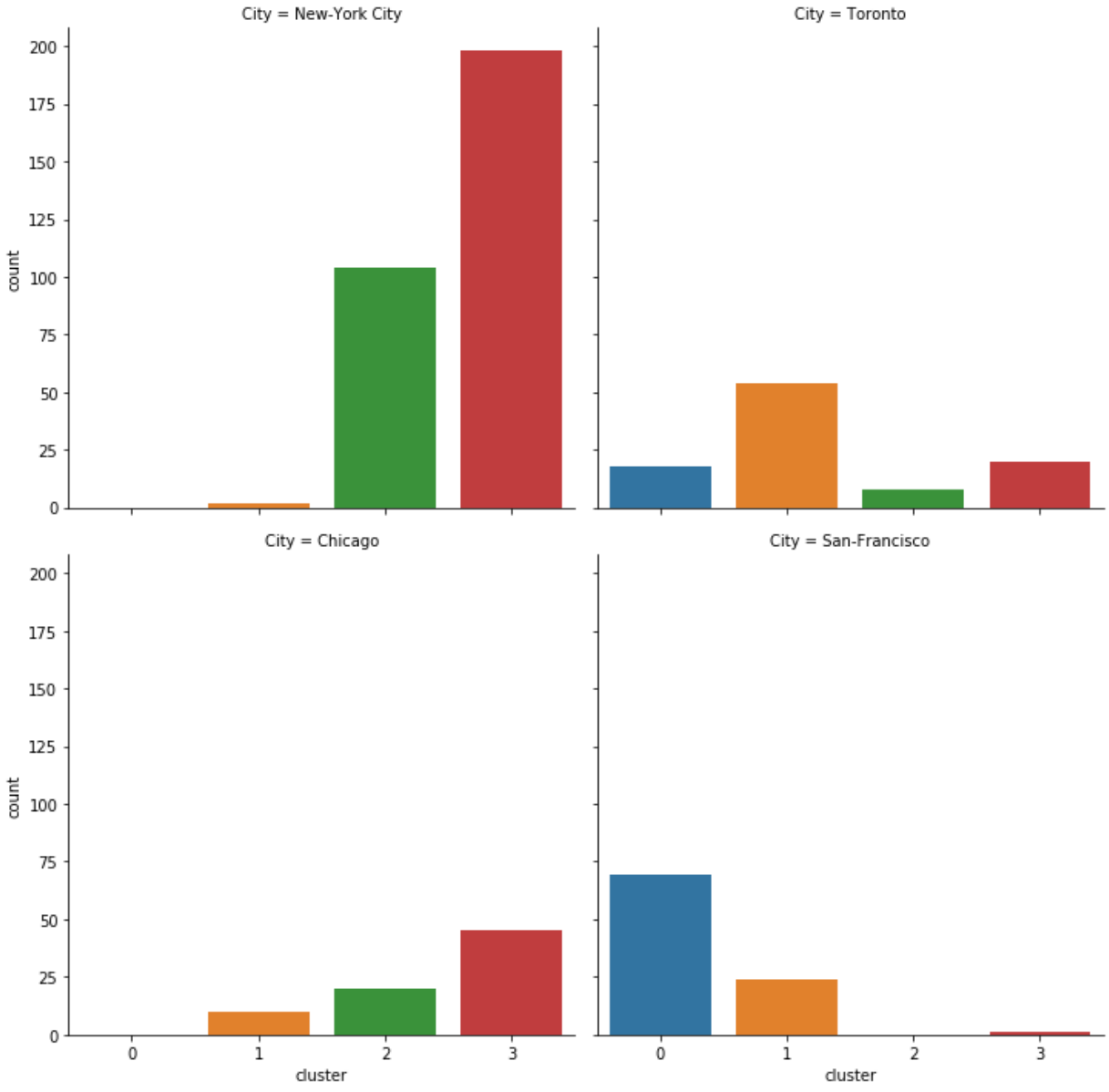

Figure 9: Distribution of the clusters per city.

The figure 10 shows the location of each neighborhoods in Toronto. The neighborhoods of cluster 0 seems to be located in down town. The neighborhoods of cluster 3 (the worst) are located on the outskirts of the city.

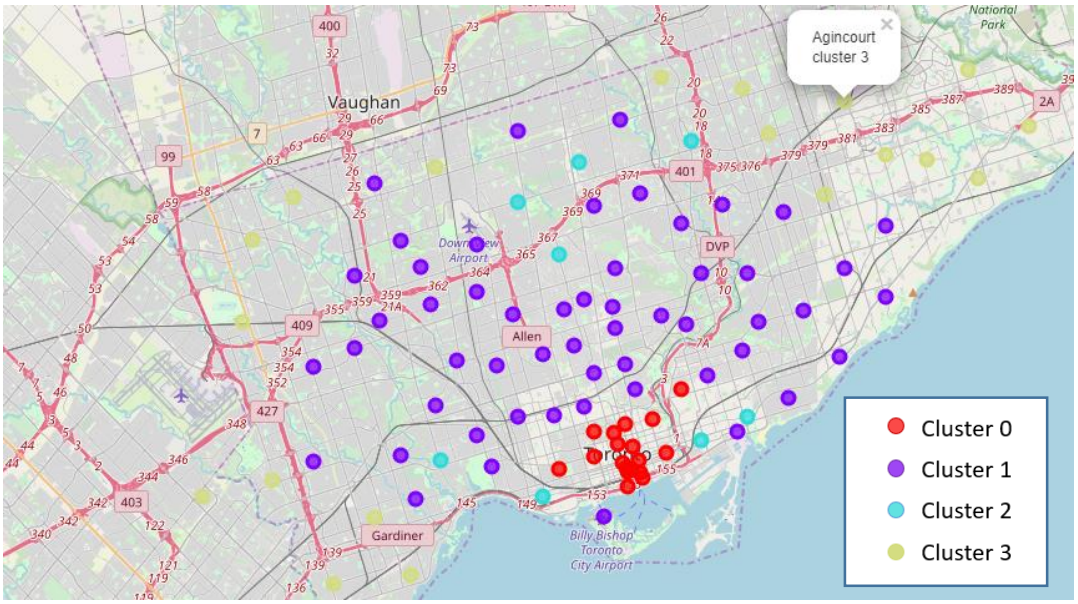

Figure 10: location of each neighborhood in Toronto.

In San-Francisco (figure 11), the group 0 covers a larger extent and there are only 2 groups (0 and 1).

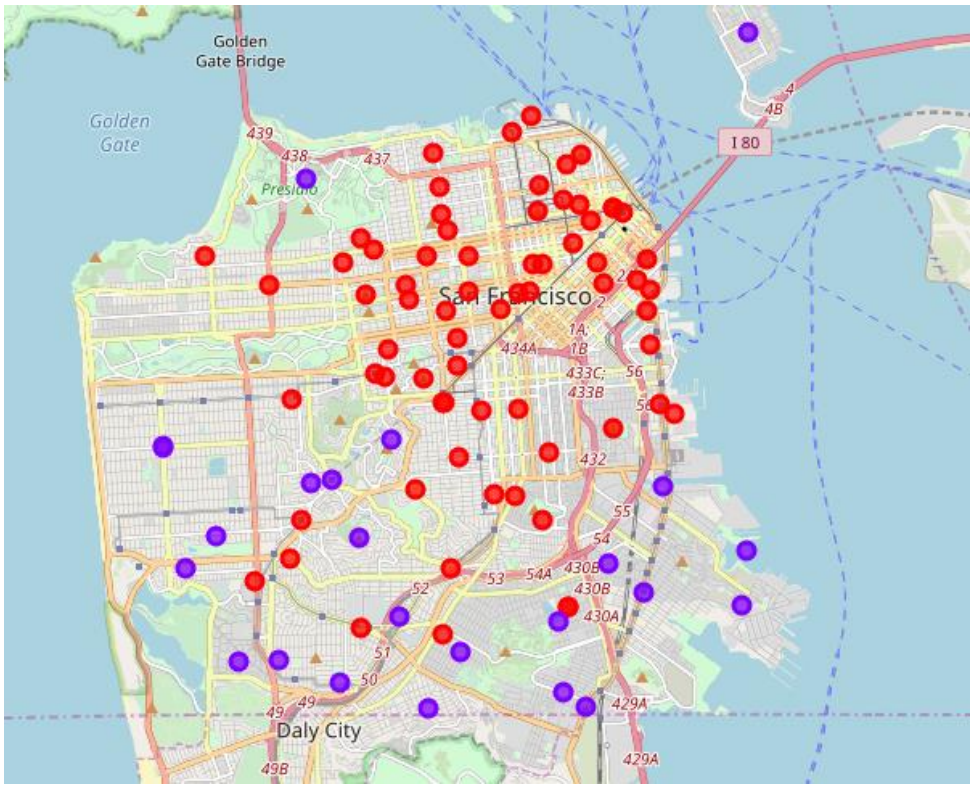

Figure 11: Location of the neighborhoods in San-Francisco.

Top-5 best neighborhoods selection

There are 87 neighborhoods in cluster 0. We had more than 500 neighborhoods before the clustering.

The objective is to identify the top-5 best neighborhoods. This selection will based on the second priority, i.e. the criteria related to antiquities and burglaries.

Firstly, we will compute the number of burglaries that occurred in 2018 within a radius <= 500m from the neighborhood center.

The burglaries lists (for each city) were already generated in Chapter "Data".

Secondly, we will filter the venues by Venue Category. We only keep "Antique shop". Then we compute the minimum distance between each neighborhood and all the antique. We perform this operation for each city.

Figure 12 represents a scatter plot, we plotted the minimum distance to antique shops versus the number of burglaries for each neighborhood. As we want the property to be as close as possible to an antique shop and to be in area where the number of burglaries is low: we must find 5 points when x and y are minimum, i.e. we have to look at the left bottom corner of the chart.

Regarding the graph, we identified 5 neighborhoods that we consider as the best (orange points). Note that there are 2 overlapping points.

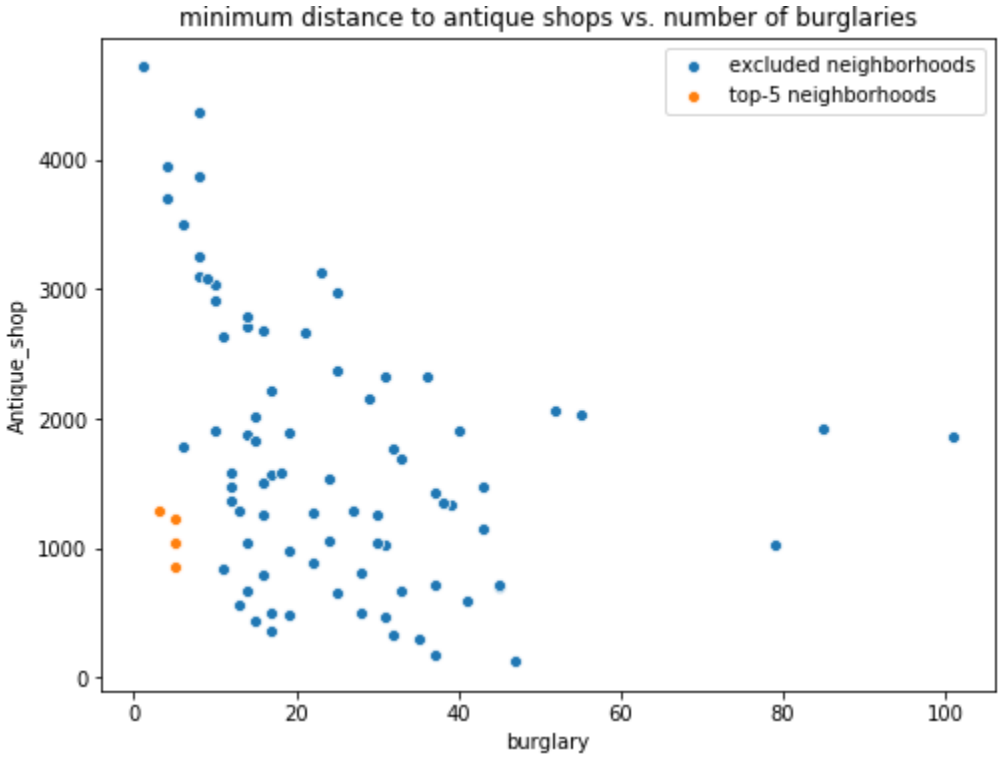

Figure 12: Antique shops vs. burglaries. Each point represents one of the 87 neighborhoods of group 0.

The 5 neighborhoods are listed on figure 13.

Results

Our analysis shows that the neighborhoods that satisfy as much as possible the criteria of the customer are located in San-Francisco or Toronto. We had 573 neighborhoods to evaluate and the clustering helped us to filter the neighborhoods and we got 87 potential best neighborhoods.

The second criteria of the future buyer allowed us to select 5 of the 87 neighborhoods.

	Neighborhood	Latitude	Longitude	City
372	First Canadian Place, Underground city	43.648429	-79.382280	Toronto
498	Embarcadero	37.792864	-122.396912	San-Francisco
502	Financial District	37.793647	-122.398938	San-Francisco
503	Financial District South	37.793647	-122.398938	San-Francisco
504	Fisherman's Wharf	37.809167	-122.416599	San-Francisco

Figure 13: The top-5 neighborhoods.

These 5 neighborhoods are located in very urbanized areas. The map (figure 14) is focused on the neighborhood "First Canadian Place, Underground city" in Toronto.

We can find a very important number of restaurants, bars, services and transports. There also many drugstores and one hospital (St Michael's Hospital) located to the northeastern.

In term of culture, there are many theaters and some museums like the Toronto Railway Museum located to the southwestern of the neighborhood.

The Scotiabank Arena is a complex located on the south of the city not far from the neighborhood. Many basketball games are played there.

In term of green areas, we find a few squares on the outskirts of the neighborhood.

Finally, there 2 antique shops located to the west and east of the neighborhood:

- Sunday Antique Market
- Toronto Antiques On King

Our choice seems coherent with the criteria of the future buyer.

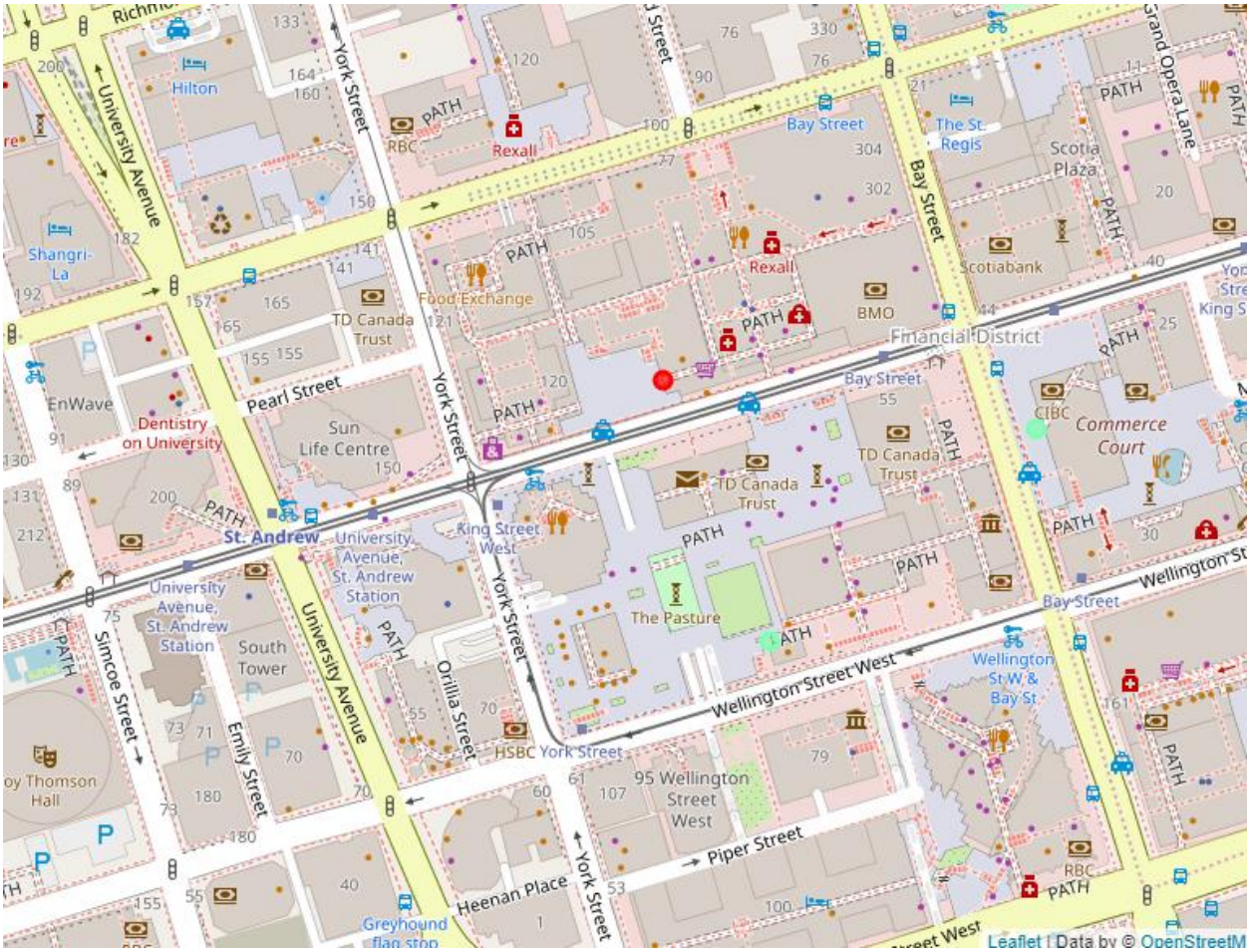

Figure 14: First Canadian Place, Underground City.

Discussion

The use of clustering is quite difficult to handle because we don't have labeled data to evaluate the results and it's quite difficult to estimate the best value of k. We use a "naïve" approach based on the characteristics of the distributions of each cluster but it's not necessary the best approach. In the future, it would be interesting to test clustering with a greater range of k values (we test k in [3:11]).

In order to improve the clustering, we should transform the feature crimes that is the number of crimes reported in 2018 within a radius <= 500m from the neighborhood center. The problem is that neighborhoods have not the same population density, so it make sense to divide the number of crimes by the population. The problem is obviously that censuses are not done every year and they are not set up in each city at the same time.

There are also multiple ways to generate features from the criteria of the customer. We decide to count the distinct types of venues for some group of categories of venues, and to compute the average distance with other types of venues. To go deeper, it will be interesting to try other features like the minimum distance, etc.

Conclusion

The objective of this project was to identify 5 neighborhoods for the company NAPE. This will help NAPE to focus on a few neighborhoods to find a land that should suits as much as possible the needs of a client who has planned to acquire a property soon.

The first instructions was to focus on only 4 cities: Chicago, NYC, Toronto and San-Francisco.

We used Foursquare and Geopy module to explore the venues nearby each neighborhood of these cities and compute some features, and then by clustering the neighborhoods we managed to determine a group of neighborhoods to investigate in deeper. We used the representation of the features in spider chart to identify the best group of neighborhoods that are the best compromise and selected the top-5 best neighborhoods according to the second criteria of the customer.

After a geographical analysis of these 5 neighborhoods, we identified on the map many venues that should meet the needs of the customer and that confirm that our choice is coherent.