IBM Data Science Capstone

Recommending best city in Scotland for Clothing business by a foreign investor

IREKE, Ukiwo Ireke

Introduction/ Business Problem

A Foreign Investor (customer) wishes to invest and open a Clothing Store Business in one of the Canada's big cities. He has a concept but as a foreigner, new in the country with little idea about Canada's City structure and therefore requires help.

He already owns a middle range Clothing Store Chain in the United Kingdom and this will be the first Store opened in Canada, therefore, it should meet some criteria to present his brand correctly.

After a meeting with him, he defined his business aim and informed me about the criteria's like following, it should be;

1. Opened in one of the big Cities in Canada (Population over 100.000 and more).
2. Within the max. 15 minutes walking distance from the Geographical coordinates of the City Centre.
3. As far away from other Clothing Stores as possible.
4. As close as possible to Italian Restaurants, because his collections are mostly Italian designs and he thinks, the customers visiting the Italian restaurants can be more interested in store windows as walking.
5. As close as possible to Hotels, because guests of the in-city hotels are generally tended to buy clothes nearby.
6. After all He stated honestly that, He needs a city that he can pay less possible salaries as he aims to give 20 employee job.
7. He added also, a city with highest possible unemployment rate would be an advantage for him, as finding personal in a short time. Otherwise he could wait longer to complete all employee team.
8. Population of the city also counts as a positive measure too. (City should be as crowded as possible).

Furthermore, it is very important to interpret investor's desires and convert these sentences to the scientific statements. As such, sentences like "…max. 15 minutes walking distance." *will mean: with an average walking speed of 5 km/h pedestrian, 1250 meters from the Geocoordinates of that city center*. And it will be used in Foursquare Api call as Radius measure (R=1250). i.e.:

```
def getNearbyVenues(names, latitudes, longitudes, radius=1250):
```

Data Description

To perform this analysis, we will need the following data:

1. General Venues data of all Major Cities in Canada.
2. Some Socioeconomic information such as Population, Average Income /person, Average Unemployment Rate and Area in km2 of that city.
3. Top venues in that city.

These data were collated from Wikipedia and Statistics Canada, the Canadian government official statistics website https://www.statcan.gc.ca.

After cleaning and preparing the data, we defined the following master DataFrame:

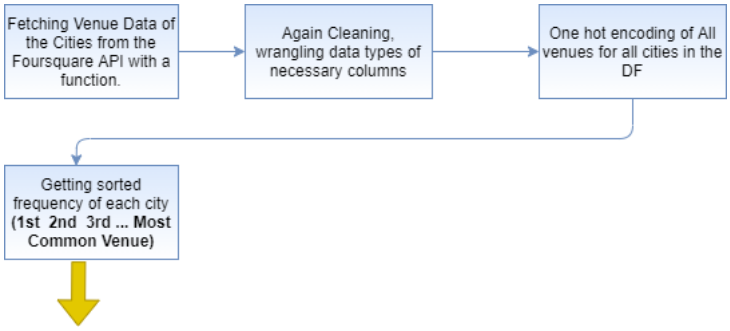
```
3 dfinal.head(20)
```

Out[98]:

	City	Province	Population 2016	Area(km2, 2011)	Population/km2	City_Lat	City_Long	Avg Income/ person CAD	Average Unemployment rate
0	Toronto, Ontario	Ontario	2731571	630.2	4334	43.653482	-79.383935	36600	10.6
1	Ottawa, Ontario	Ontario	934243	2790.2	334	45.421106	-75.690308	36600	10.6
2	Mississauga, Ontario	Ontario	721599	292.4	2467	43.589623	-79.644388	36600	10.6
3	Brampton, Ontario	Ontario	593638	266.3	2229	43.685815	-79.759934	36600	10.6
4	Hamilton, Ontario	Ontario	536917	1117.2	480	43.256080	-79.872858	36600	10.6
5	London, Ontario	Ontario	383822	420.6	912	42.983675	-81.249607	36600	10.6
6	Markham, Ontario	Ontario	328966	212.6	1547	43.856371	-79.337682	36600	10.6
7	Vaughan, Ontario	Ontario	306233	273.5	1119	43.794154	-79.526802	36600	10.6
8	Kitchener, Ontario	Ontario	233222	136.8	1704	43.451291	-80.492782	36600	10.6
9	Windsor, Ontario	Ontario	217188	146.3	1484	42.317099	-83.035343	36600	10.6
10	Richmond Hill, Ontario	Ontario	195022	101.0	1930	43.880078	-79.439392	36600	10.6
11	Oakville, Ontario	Ontario	193832	138.9	1395	43.447436	-79.666672	36600	10.6
12	Burlington, Ontario	Ontario	183314	185.7	987	43.324892	-79.796684	36600	10.6
13	Greater Sudbury, Ontario	Ontario	161531	3227.4	50	46.492720	-80.991211	36600	10.6
14	Oshawa, Ontario	Ontario	159458	145.7	1094	43.897556	-78.863532	36600	10.6
15	Barrie, Ontario	Ontario	141434	77.4	1827	44.389311	-79.690174	36600	10.6
16	St. Catharines, Ontario	Ontario	133113	96.1	1385	43.157981	-79.244100	36600	10.6
17	Guelph, Ontario	Ontario	131794	87.2	1511	43.546052	-80.249328	36600	10.6

Figure 1: Major Cities in Canada Socioeconomic data used for this project.

Geo-coordinates of districts were obtained with the help of the geocoder tool in the notebook. Then top venues data in each City obtained from Foursquare API.




```
22 City_venues_sorted
```

Out[35]:

	City	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	Abbotsford, British Columbia	Coffee Shop	Sushi Restaurant	Pharmacy	Bank	Sandwich Place	Fast Food Restaurant	Restaurant	Gym / Fitness Center	Hotel	Gas Station
1	Airdrie, Alberta	Gas Station	Coffee Shop	Fast Food Restaurant	Sushi Restaurant	Pizza Place	Gym	Diner	Supermarket	Restaurant	Café
2	Ajax, Ontario	Coffee Shop	Fast Food Restaurant	Breakfast Spot	Pizza Place	Grocery Store	Sandwich Place	Mexican Restaurant	Bar	Bank	Ice Cream Shop
3	Aurora, Ontario	Coffee Shop	Sandwich Place	Bank	Japanese Restaurant	Thai Restaurant	Gas Station	Bakery	Pizza Place	Skating Rink	Café
4	Barrie, Ontario	Coffee Shop	Park	Pub	Vegetarian / Vegan Restaurant	Restaurant	Bar	Diner	Pizza Place	Pharmacy	Thai Restaurant
5	Belleville, Ontario	Harbor / Marina	Pharmacy	Pizza Place	Beer Store	Sushi Restaurant	Bank	Coffee Shop	Restaurant	Gas Station	Bike Shop
6	Blainville, Quebec	Fast Food Restaurant	Pharmacy	Pet Store	Soccer Field	Grocery Store	Train Station	Thai Restaurant	Coffee Shop	Sporting Goods Shop	Stables
7	Brampton, Ontario	Coffee Shop	Park	Pizza Place	Pub	Italian Restaurant	Sandwich Place	Bowling Alley	Gastropub	Bank	Baseball Field
8	Brantford, Ontario	Coffee Shop	Café	Grocery Store	Sandwich Place	Beer Store	Restaurant	Fast Food Restaurant	Asian Restaurant	Casino	Liquor Store
9	Brossard, Quebec	Fast Food Restaurant	Chinese Restaurant	Bank	Hotel	Grocery Store	Bakery	Sushi Restaurant	Supermarket	Liquor Store	Pharmacy

Figure 2: Cities Venue Data

Methodology and Analysis

After cleaning and preparing the data we then begin Analyzing the Investor Requirements (where Solving the Business Problem Begins):

Beginning from this section, we will only deal with the features (variables) those only interested the customer.

This means presenting to the investor a Clusterization analyse / map, with Greek Restaurants / Sushi Restaurants / Supermarkets / Gas Stations / Lakes… are in it , is irrelevant to his business problem, and even such an approach guides us to wrong results .

In order to understand investor's requirements, we will introduce a new concept called Customer ***Wishes_Matrix***. It is a features weight matrix (array) to express / to get investor desires in a scientific way.

Every feature that investor prerequisites from us, will be weighted to a scale from 1 to 10 in a manner of importance:

- **Opened in one of the big Cities in Germany (Population over 100,000 and more).** This is already satisfied because we have fetched cities only with population >100k
- **Within the max. 15 minutes walking distance from the Geographical coordinates of the City Centre.** This request is converted to a variable to use in Foursquare API call (Radius = 1250 meters in search).
- **As far away from other Clothing Stores as possible.** This request is very important for him and weighted as 9 points over 10 points. **(0.9)**
- **As close as possible to Italian Restaurant.** This request is somehow second degree and weighted as 5 points over 10 points. **(0.5)**
- **As close as possible to Hotels**. This request is second degree in importance but still weighted as 7 point over 10 points. **(0.7)**
- **Cities that statistically less possible salaries are paid**. For the investors salary issues are always important. So, it is weighted as 8 point. **(0.8)**
- **Cities with highest possible unemployment rate.** This request is also important but not more than salaries. So, gets 7 points over 10 pts. **(0.7)**
- **Population of the city also counts as a positive measure too. (City should be as crowded as possible).** Population value is directly related to the number of potential customers therefore it will be weighted as 8 points over 10. **(0.8)**

3.2.a) Creating A Weight ("Wishes Matrix"):

Re W C1

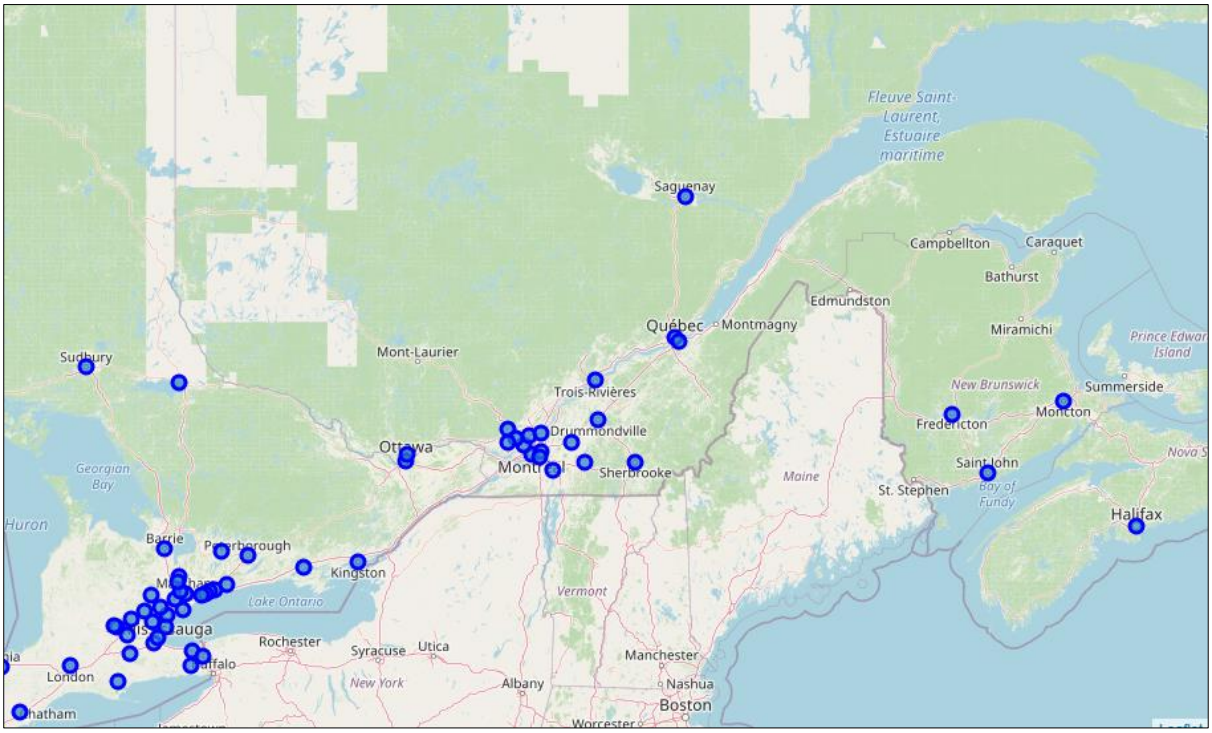
$$\begin{Bmatrix} Request\ 3 & 0{,}9 \\ Request\ 4 & 0{,}5 \\ Request\ 5 & 0{,}7 \\ Request\ 6 & 0{,}8 \\ Request\ 7 & 0{,}7 \\ Request\ 8 & 0{,}8 \end{Bmatrix} \times \begin{Bmatrix} -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ 1 \end{Bmatrix} = \text{Wishes Matrix}$$

Correlation bit: (-1 for negatively correlated variables and +1 is for positive correlation) For example increase on value of the Request 3 variable (Amount of Clothing stores in the area) will be added as punishment to end Data Frame, as it is not desired. But increase on population is positively counts (+) to end Data frame, as it is an advantage)

```
7 Wishes_Matrix
```

	Requirement	weight
0	Clothing Stores	-0.9
1	Italian Restaurants	0.5
2	Hotels	0.7
3	Income /Person CAD	-0.8
4	Avg unemployment rate	0.7
5	Population	0.8

Firstly, it's is quite useful to visualize the cities to be considered to have a view of the location spread.



Next, we will proceed with the following steps :

1. Getting only the necessary and related 'Venue' columns (features) to a new data frame called ***Customer_Venues*** using Foursquare API

```
5 Customer_Venues.head(10)
```

	City	Clothing Store	Italian Restaurant	Hotel
0	Abbotsford, British Columbia	0.0	0.00	0.03
1	Airdrie, Alberta	0.0	0.00	0.00
2	Ajax, Ontario	0.0	0.00	0.02
3	Aurora, Ontario	0.0	0.03	0.00
4	Barrie, Ontario	0.0	0.02	0.02
5	Belleville, Ontario	0.0	0.03	0.00
6	Blainville, Quebec	0.0	0.00	0.00
7	Brampton, Ontario	0.0	0.05	0.00
8	Brantford, Ontario	0.0	0.00	0.00
9	Brossard, Quebec	0.0	0.02	0.04

2. Again, getting only the necessary 'Social Data' columns (features) to a new data frame called ***Customer_Social***.

```
5 Customer_Social.head()
```

	City	Avg Income/ person CAD	Average Unemployment rate	Population 2016
0	Toronto, Ontario	36600	10.6	2731571
1	Ottawa, Ontario	36600	10.6	934243
2	Mississauga, Ontario	36600	10.6	721599
3	Brampton, Ontario	36600	10.6	593638
4	Hamilton, Ontario	36600	10.6	536917

3. Merging these two newly created Data Frames based on Key 'City'. Calling New DF as ***Customer_Merged***.

```
3 Customer_Merged.head(10)
```

	City	Clothing Store	Italian Restaurant	Hotel	Avg Income/ person CAD	Average Unemployment rate	Population 2016
0	Abbotsford, British Columbia	0.0	0.00	0.03	36700	10.7	141397
1	Airdrie, Alberta	0.0	0.00	0.00	40800	11.8	61581
2	Ajax, Ontario	0.0	0.00	0.02	36600	10.6	119677
3	Aurora, Ontario	0.0	0.03	0.00	36600	10.6	55445
4	Barrie, Ontario	0.0	0.02	0.02	36600	10.6	141434
5	Belleville, Ontario	0.0	0.03	0.00	36600	10.6	50716
6	Blainville, Quebec	0.0	0.00	0.00	34700	8.7	56863
7	Brampton, Ontario	0.0	0.05	0.00	36600	10.6	593638
8	Brantford, Ontario	0.0	0.00	0.00	36600	10.6	97496
9	Brossard, Quebec	0.0	0.02	0.04	34700	8.7	85721

4. Applying a Normalization to the ***Customer_Merged*** Data Frame.

```
2 Customer_Merged.head()
```

City	Clothing Store	Italian Restaurant	Hotel	Avg Income/ person CAD	Average Unemployment rate	Population 2016
Abbotsford, British Columbia	0.0	0.00	0.03	36700	10.7	141397
Airdrie, Alberta	0.0	0.00	0.00	40800	11.8	61581
Ajax, Ontario	0.0	0.00	0.02	36600	10.6	119677
Aurora, Ontario	0.0	0.03	0.00	36600	10.6	55445
Barrie, Ontario	0.0	0.02	0.02	36600	10.6	141434

We then apply a Normalization to the Customer_Merged Data Frame to bring our interested features to a comparable grade between 0 and 1 for scalability. This is because population scale is in millions way higher than frequency of stores and may introduce bias to our analysis. As an example, the highest Population value will be converted to 1.0, and at the same time highest frequency of the Hotel also will be converted to 1.This will ensure that each feature (each column) have the same WEIGHT (same importance) as the others.

```
2 Customer_Merged.head(10)
```

City	Clothing Store	Italian Restaurant	Hotel	Avg Income/ person CAD	Average Unemployment rate	Population 2016
Abbotsford, British Columbia	0.0	0.000000	0.12	0.899510	0.816794	0.051764
Airdrie, Alberta	0.0	0.000000	0.00	1.000000	0.900763	0.022544
Ajax, Ontario	0.0	0.000000	0.08	0.897059	0.809160	0.043813
Aurora, Ontario	0.0	0.230769	0.00	0.897059	0.809160	0.020298
Barrie, Ontario	0.0	0.153846	0.08	0.897059	0.809160	0.051778
Belleville, Ontario	0.0	0.230769	0.00	0.897059	0.809160	0.018567
Blainville, Quebec	0.0	0.000000	0.00	0.850490	0.664122	0.020817
Brampton, Ontario	0.0	0.384615	0.00	0.897059	0.809160	0.217325
Brantford, Ontario	0.0	0.000000	0.00	0.897059	0.809160	0.035692
Brossard, Quebec	0.0	0.153846	0.16	0.850490	0.664122	0.031382

5. Multiplying the Wishes_Matrix (already same features) with the Customer_Merged Data Frame.

In this section we have assigned the weight column of the ***Wishes_Matrix*** to a NumPy array object ***Ar1***. And applying an arithmetic multiplication of 6 features to the Customers normalized data frame (keeping the features order same as the wishes matrix).

```
1 Weighted_Customer = Customer_Merged * Ar1
2
3 Weighted_Customer.head(5)
```

City	Clothing Store	Italian Restaurant	Hotel	Avg Income/ person CAD	Average Unemployment rate	Population 2016
Abbotsford, British Columbia	-0.0	0.000000	0.084	-0.719608	0.571756	0.041411
Airdrie, Alberta	-0.0	0.000000	0.000	-0.800000	0.630534	0.018035
Ajax, Ontario	-0.0	0.000000	0.056	-0.717647	0.566412	0.035050
Aurora, Ontario	-0.0	0.115385	0.000	-0.717647	0.566412	0.016238
Barrie, Ontario	-0.0	0.076923	0.056	-0.717647	0.566412	0.041422

6. Obtaining a ***Weighted_Customer*** Data Frame at hand , creating a ***Total_score*** Column on it.

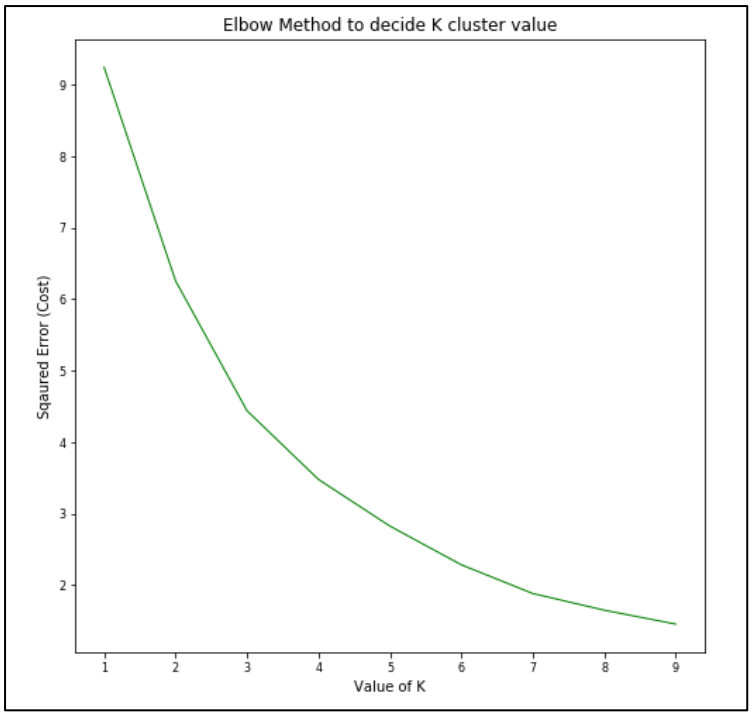
Descending sorting of all cities by their Total_Scores, is also our ultimate results to offer investor (customer). Toronto seems to be best optimal place to open a clothing store , in respect to Customers wishes. Following it the Cities : Toronto --> Vaughan --> Calgary .are the closer candidates.

```
1 Weighted_Customer.sort_values(by='Total_Score', ascending=False).head(10)
```

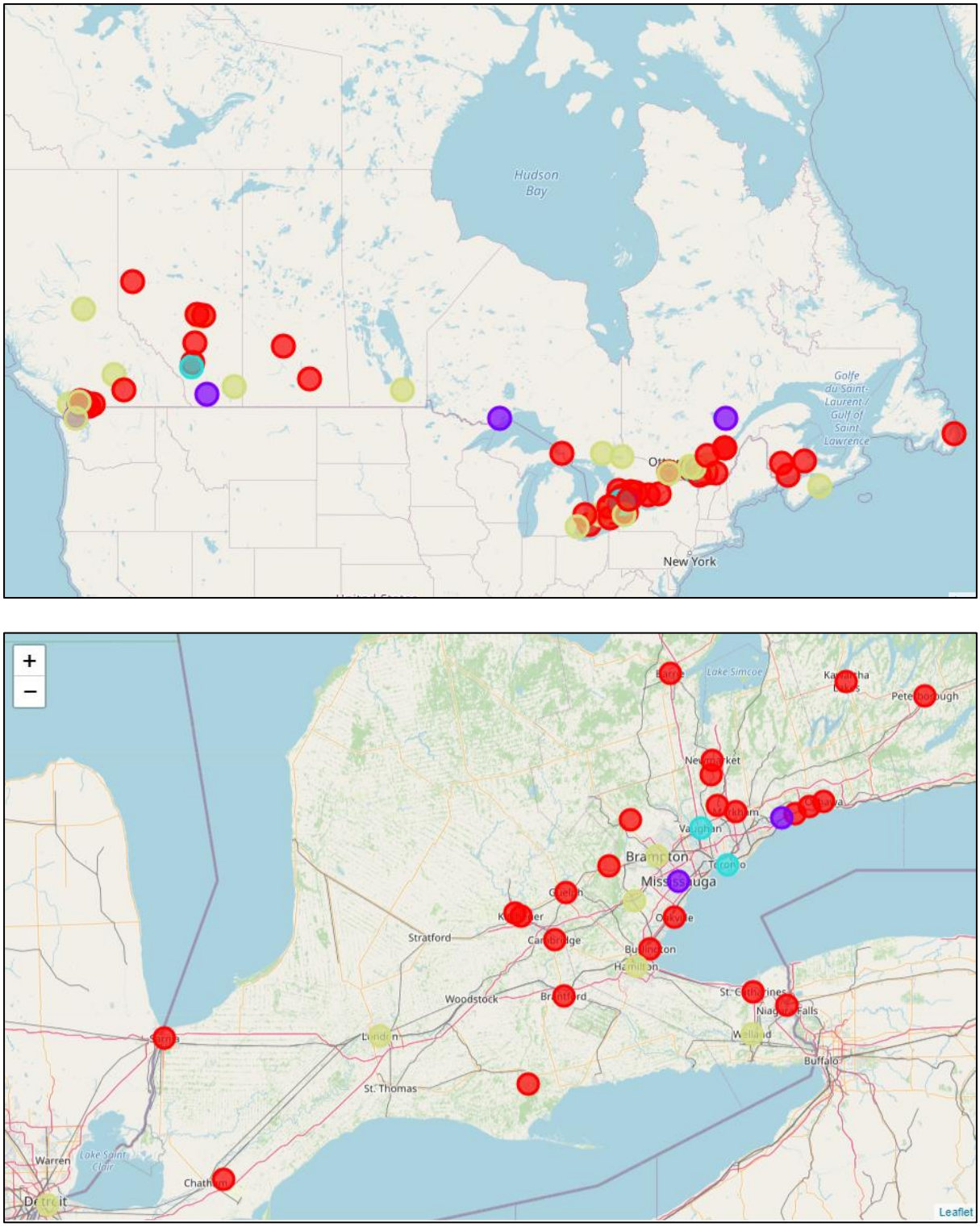
City	Clothing Store	Italian Restaurant	Hotel	Avg Income/ person CAD	Average Unemployment rate	Population 2016	Total_Score
Toronto, Ontario	-0.075	0.076923	0.112	-0.717647	0.566412	0.800000	0.762688
Vaughan, Ontario	-0.000	0.000000	0.700	-0.717647	0.566412	0.089687	0.638452
Calgary, Alberta	-0.000	0.153846	0.196	-0.800000	0.630534	0.362933	0.543313
Montreal, Quebec	-0.000	0.076923	0.084	-0.680392	0.464885	0.499257	0.444673
Windsor, Ontario	-0.000	0.423077	0.084	-0.717647	0.566412	0.063608	0.419450
Edmonton, Alberta	-0.000	0.230769	0.056	-0.800000	0.630534	0.273116	0.390420
Saint-Hyacinthe, Quebec	-0.000	0.500000	0.000	-0.680392	0.464885	0.016298	0.300791
Saint-Jerome, Quebec	-0.000	0.384615	0.084	-0.680392	0.464885	0.021774	0.274883
Ottawa, Ontario	-0.075	0.038462	0.140	-0.717647	0.566412	0.273613	0.225840
Brampton, Ontario	-0.000	0.192308	0.000	-0.717647	0.566412	0.173860	0.214933

7. **Cluster Analysis:** To identify groups (clusters) with similar characteristics, the unsupervised learning method to our data, namely K-Means algorithm, was applied to our data**.** Making a City clusterization and visualization for the ***Weighted_Customer*** DF.

To identify the optimal number of clusters, the Elbow method is used. It can be seen from the graph that ***four*** clusters are the best choice. Furthermore, this elbow point is confirmed programmatically with the *KneeLocator* of the *kneed* python package. As such k is set as 4.



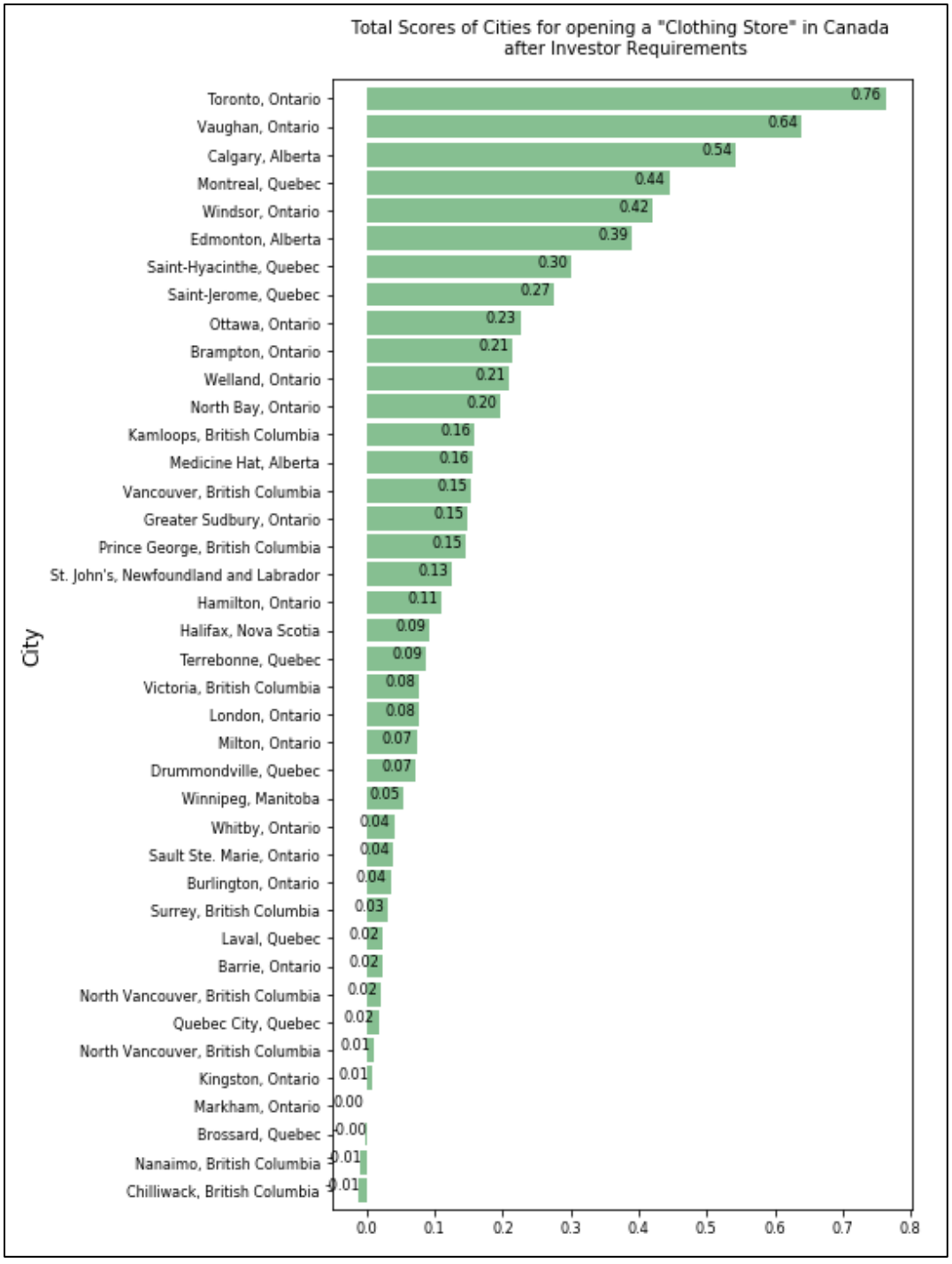
Visualizing the clustering results based on the total score of venues in each city across Canada, the following maps were created. The first showing the distribution of clusters across the entire nation and the second map taking a closer look at nearby cities.



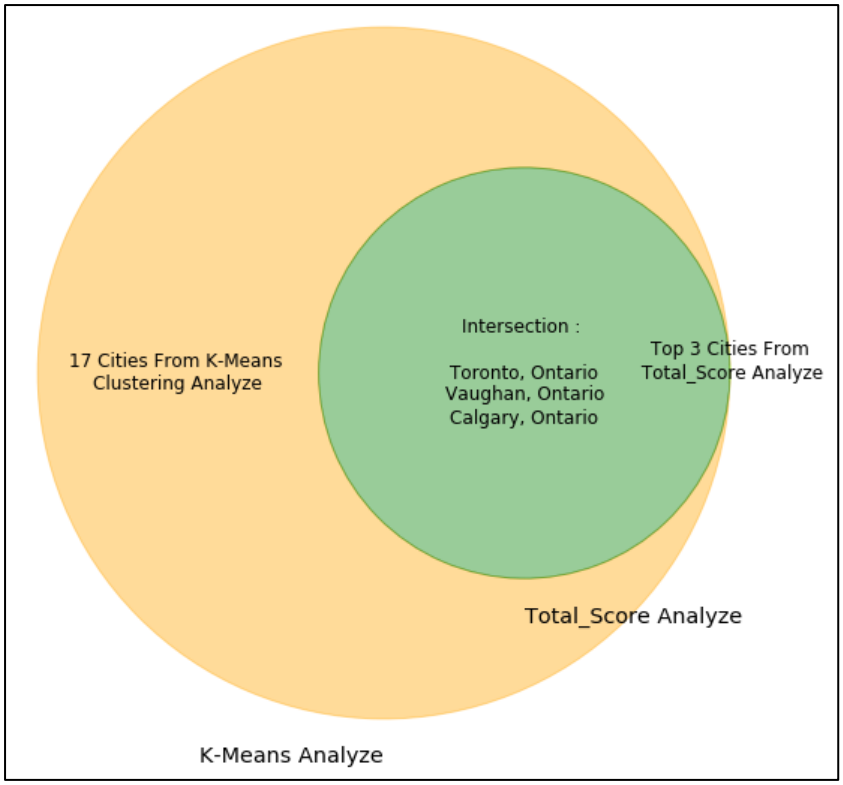
We can see that nearby cities did not belong to similar clusters. This was mainly due to clusters based on total score of the several requirements of the investor. Most cities may differ in population or other factors.

Results

From our Analysis, Toronto revealed to have the highest total score based on investors requirements. Thus obviously, the best location to suggest for investment as shown in the figure below.



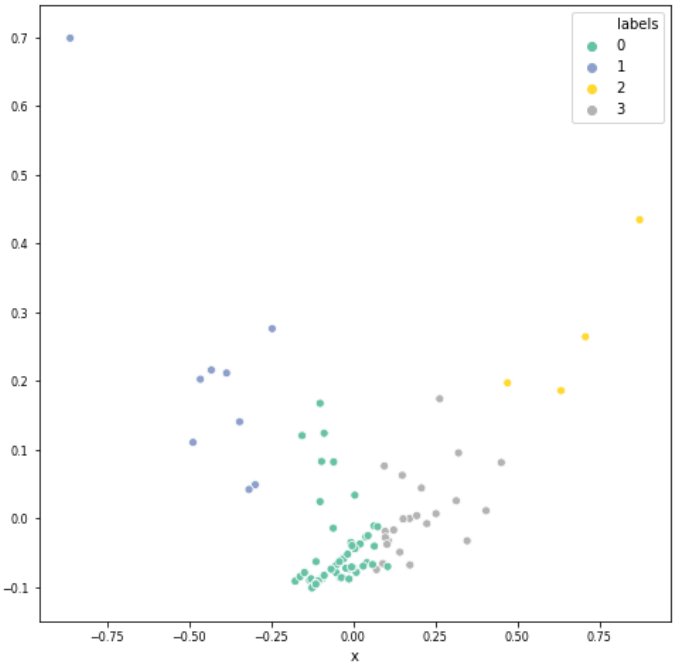
Summary of result on a Venn diagram showing best cities as intersection is shown below:



What could be done better?

Foursquare doesn't represent the full picture, since many venues are not on the list. For that reason, another map could be utilized such as Google map or Openstreet map.

Locations have too complex geometry, thus defining the closest venues within the certain radius brings additional error to our analysis.



Conclusion

To conclude, the basic data analysis was performed to identify the most optimal location for the clothing business in Canada. During the analysis, several important statistical features of the Venues in each City were explored and visualized. Furthermore, clustering helped to highlight the group of optimal areas.

Cluster analysis showed cluster 2 to consist mainly of the top major cities to be recommended to the investor. These cities also showed the highest total score above 50% with Montreal dropped since it scores is lower. Finally, ***Toronto, Vaughan and Calgary*** were chosen as the most attractive Cities for investment according to investors demands.