
DOUBLETREE
TORONTO

Moving to Toronto

1. The “Business problem”
    1.1. Target audience- who may be interested?
2. Data Section:
    2.1. Describe Data requirements and Sources needed to solve the problem
3. Methodology section:
    3.1. Data processing
    3.2. Exploratory data analysis
    3.3. Machine learning algorithms used.
4. Results section:
    4.1. Discussion of the results and findings
5. Discussion section:
    5.1. Discussion of observations noted and any recommendations
6. 6. Conclusion section:
    6.1. Options and conclusions.

Moving to Toronto

1. Discussion and Background of the Business Problem:

An executive manager has been expatriated to Toronto City and has to look for a new family home in one of its neighbourhoods

The key **criteria** to take into consideration are:

Family needs

- **good rated elementary schools** for his young children
- the presence of plenty of **malls** for his wife.

Personal needs.

- the existence of gyms in the chosen neighbourhood

Criteria excluded:

- **criminality factor** due to Toronto has low crime rates
- **cost of rental housing** as it is included in the expatriation package

Moving to Toronto

1.1 Target Audience:

1. Investors who could benefit from the model to assess real estate investments
2. Commercial Real Estate Brokers encouraged to offer brokerage services
3. Big Corporations interested in relocation processes for their expats.
4. Public Administration to right size their infrastructures to attract foreign talent
5. Toronto residents keen on taking data driven decisions based on the model
6. Individuals in expatriation situation who may have to face a similar situation

Moving to Toronto

2. Data Preparation:

1. Packages, if missing

beautifulsoup4 to scrape websites

geopy to geocode web services and to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources

folium to visualize data that's been manipulated in Python on an interactive leaflet map.It enables both the binding of data to a map visualization as well as passing rich vector/raster/HTML visualizations as markers on the map

2. Libraries

numpy # to handle data in a vectorized manner
pandas # for data analysis
json # to handle JSON files
Nominatim # to convert an address into latitude and longitude values
requests # to handle requests
json_normalize # to tranform JSON file into a pandas dataframe
matplotlib and associated # to plotting graphsmodules
sklearn # to use machine learning k-means at clustering stage
folium # to map rendering
geocoder # to get coordinates

Moving to Toronto

2. Data Sources:

1. Wikipedia

   https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2. Foursquare API

   https://developer.foursquare.com/docs/resources/categories

3. GeoSpace data

   https://open.toronto.ca/dataset/community-council-boundaries/

4. Websites

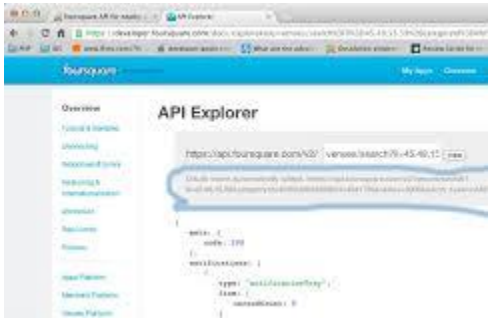
   ontario.compareschoolrankings.org

Moving to Toronto

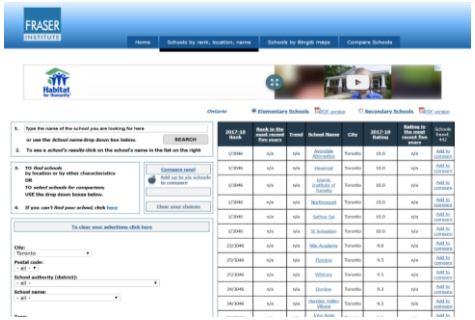
3. Methodology section:

3.1 Data Processing:

- Data wrangling process for Wikipedia source input
  - Quality rules applied



- API Foursquare
  - Extracting Venue Types information
    - Gym
    - Mall



- Website *ontario.compareschoolrankings.org*
  - building file with extracted information

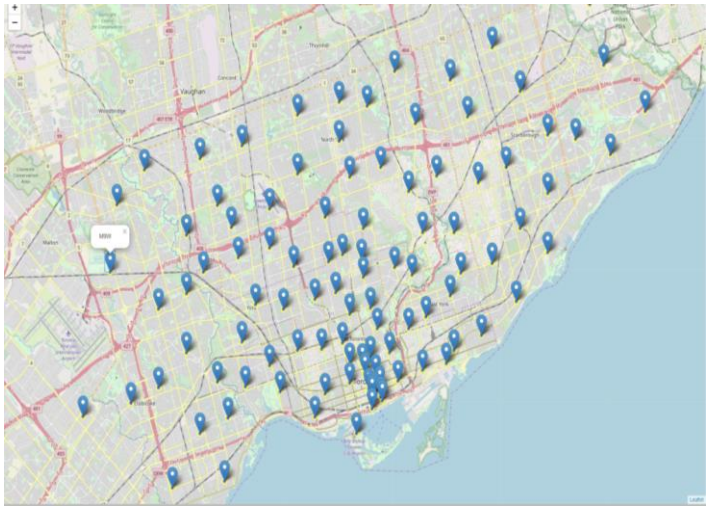
Microsoft Excel csv Worksheet

Moving to Toronto

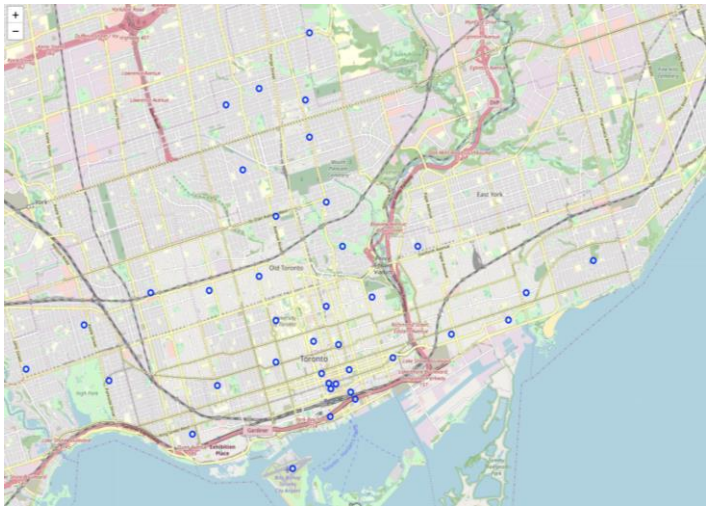
3. Methodology section:

3.1 Data Processing:

FROM



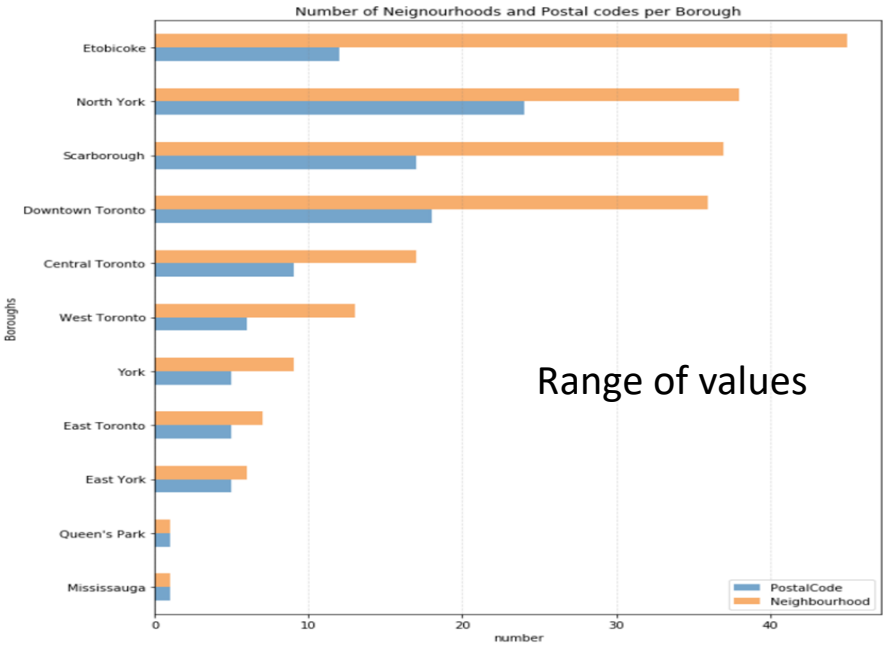
TO



Appying business decisión: Only part of Greater Toronto area

Moving to Toronto

3. Exploratory Analysis:



Range of values

List of values

	neighborhood	id	name	categories
0	Adelaide,King,Richmond	4	4	4
1	Berczy Park	1	1	1
2	Brockton,Exhibition Place,Parkdale Village	1	1	1
3	Central Bay Street	5	5	5
4	Church and Wellesley	11	11	11
5	Commerce Court,Victoria Hotel	2	2	2
6	Davisville North	2	2	2
7	Design Exchange,Toronto Dominion Centre	5	5	5
8	Dovercourt Village,Dufferin	2	2	2
9	First Canadian Place,Underground city	4	4	4
10	Forest Hill North,Forest Hill West	1	1	1

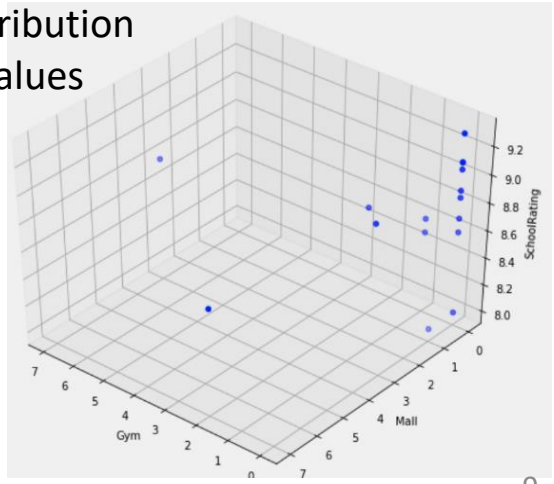
```
print(crit_1.shape)
crit_1.head()
```

(261, 16)

Level of detail

	address	categories	cc	city	country	crossStreet	distance	formattedAddress	id	labeledLatLngs	lat	lng
0	1048 Broadview Ave.	Gym / Fitness Center	CA	Toronto	Canada	NaN	679	[1048 Broadview Ave., Toronto ON M4K 2R8, Canada]	539a6357498ef1b1c9b888e7	[{'label': 'display', 'lat': 43.684524, 'lng':...	43.684524	-79.357102
0	Carlaw Ave	Gym / Fitness Center	CA	Toronto	Canada	Carlaw & Dundas	432	[Carlaw Ave (Carlaw & Dundas), Toronto Ont, Ca...	52c9880b498edcce9881e7b0	[{'label': 'display', 'lat': 43.66341163749049,...	43.663412	-79.341104
1	233 Carlaw Ave.	Gym	CA	Toronto	Canada	btwn Dundas St & Queen St. E	382	[233 Carlaw Ave. (btwn Dundas St & Queen St. E...	4cdf14f2f8a4a1434ae3dbbc	[{'label': 'display', 'lat': 43.66293178257155,...	43.662932	-79.340321
0	140 Erskine	Gym	CA	Toronto	Canada	NaN	271	[140 Erskine, Toronto ON, Canada]	4da89b34a86e771ea70e84c1	[{'label': 'display', 'lat': 43.71312601210131,...	43.713126	-79.393537
1	900 Mount Pleasant Road	Gym / Fitness Center	CA	Toronto	Canada	NaN	174	[900 Mount Pleasant Road, Toronto ON M4P 3J9, ...	4c3f2724db3b1b8d635e6695	[{'label': 'display', 'lat': 43.71167058860572,...	43.711671	-79.391767

Distribution of values



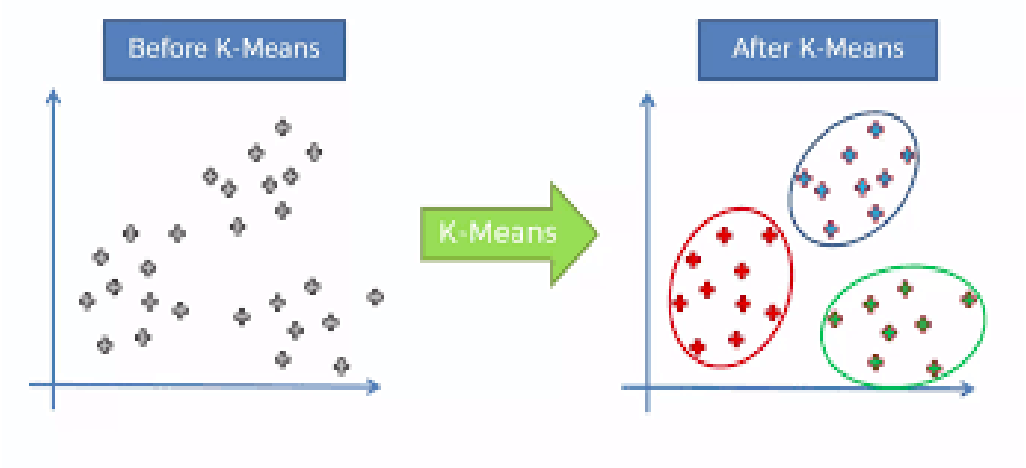
Moving to Toronto

3. Exploratory Analysis:

3.3. Machine learning algorithms used:

Using scikits learn libray

We apply k-means clustering :
a method of vector quantization, aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.



We apply K means as it is one of the algorithms that can be used for items segmentation

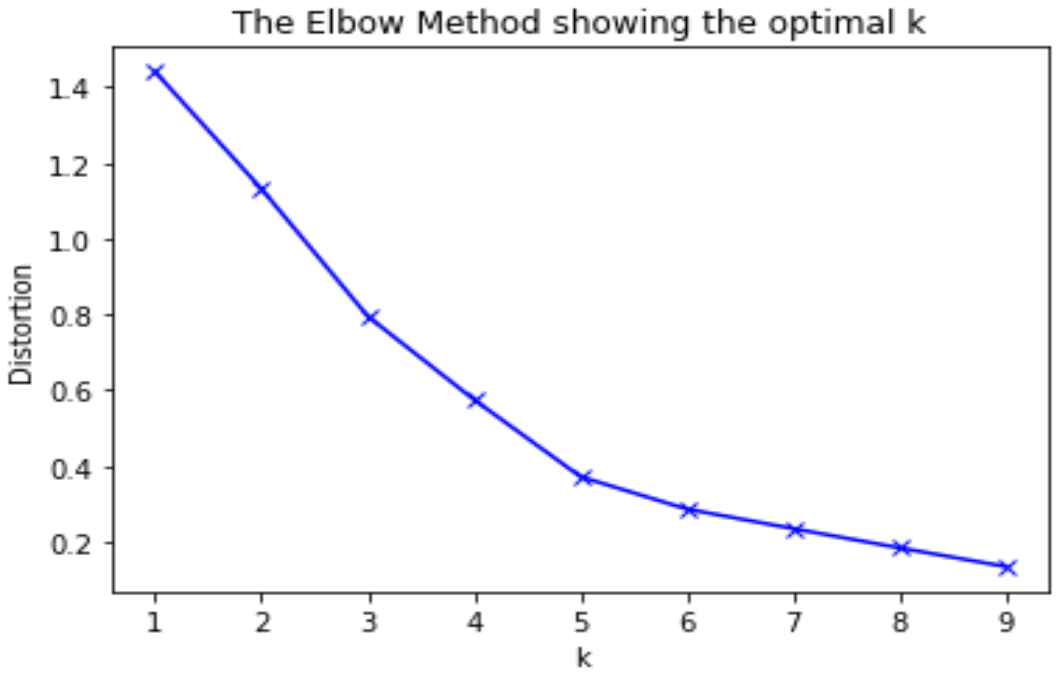
K-Means can group data only unsupervised based on the similarity of items to each other.

Moving to Toronto

3. Exploratory Analysis:

3.3. Machine learning algorithms used:

In order to determine the optimal k number, we'll apply **the elbow method**

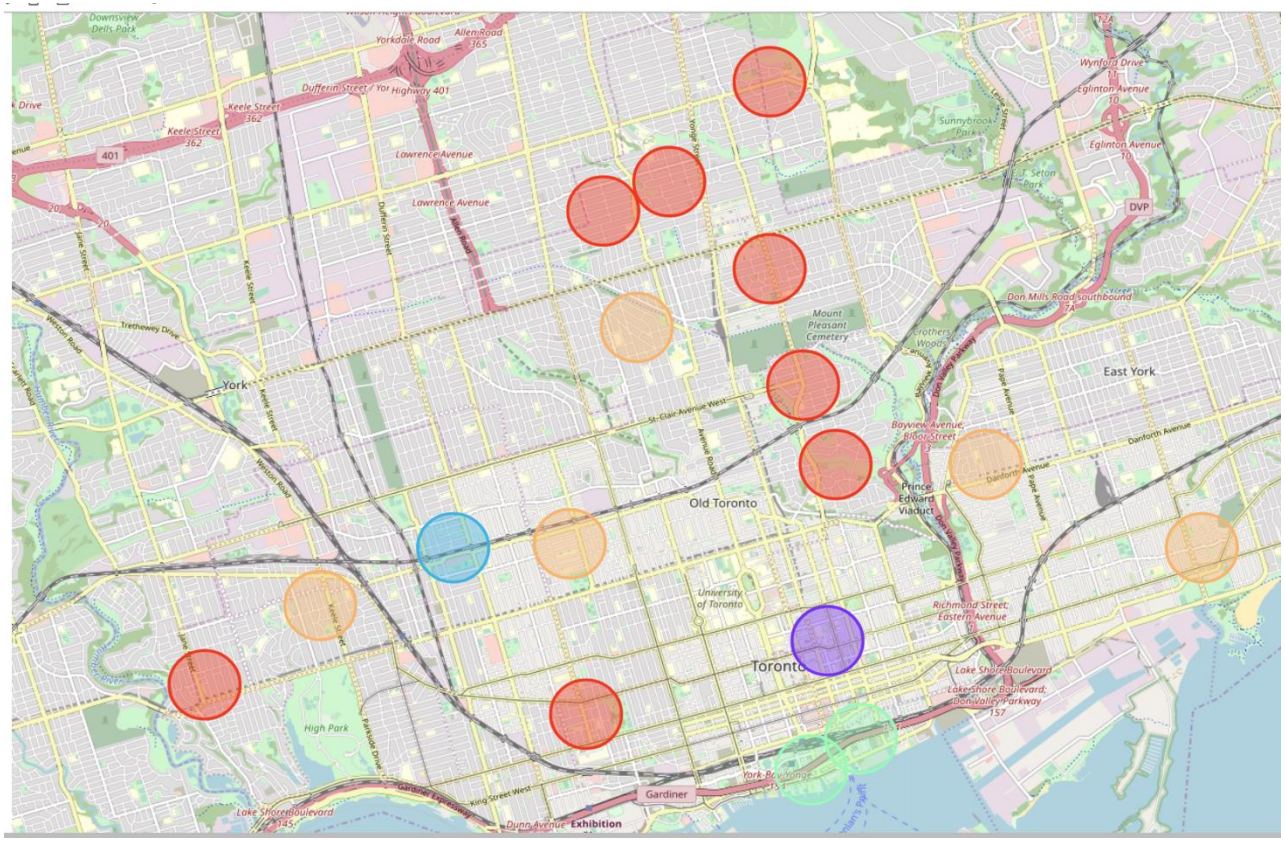


We will use k=5 to find the clusters

Moving to Toronto

4. Results:

5 clusters



Moving to Toronto

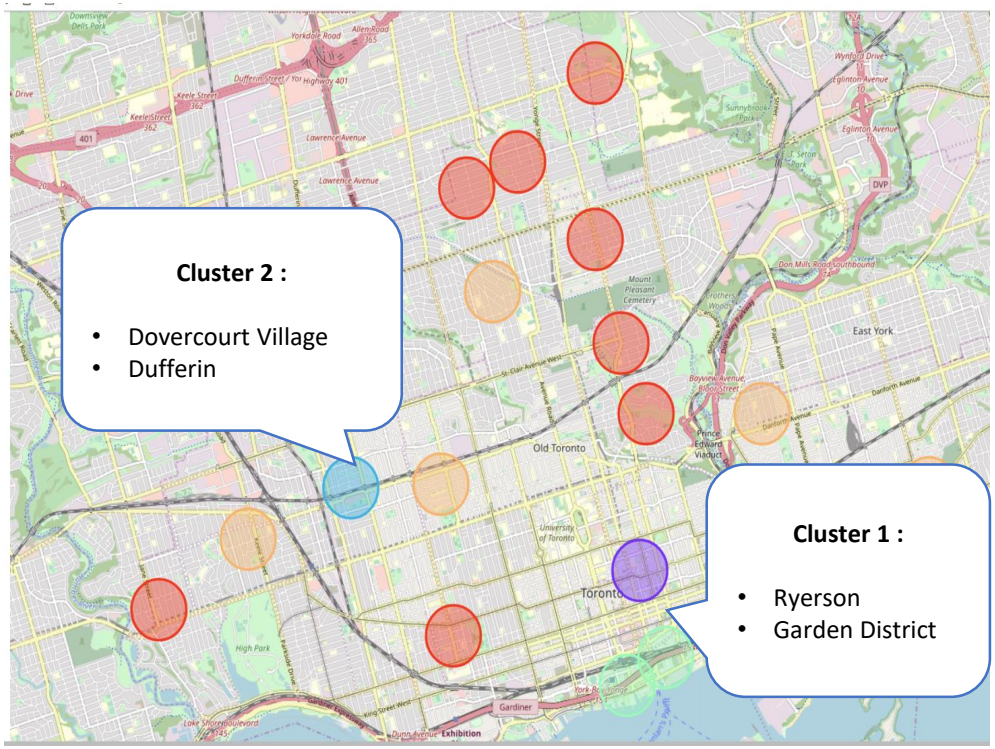
4. Results:

5 clusters :

Cluster 0 (red) have the best ranked schools but no scores in Gym nor Mall

Cluster 4 (orange) it's the group that scores either on Gym either on Malls but never on both

Cluster 3 (green) scores on both Gym and Mall but low



Cluster 1 (purple) and Cluster 2 (blue) offer above 8.7 in schools and higher scores in Gym and Mall than Cluster 3

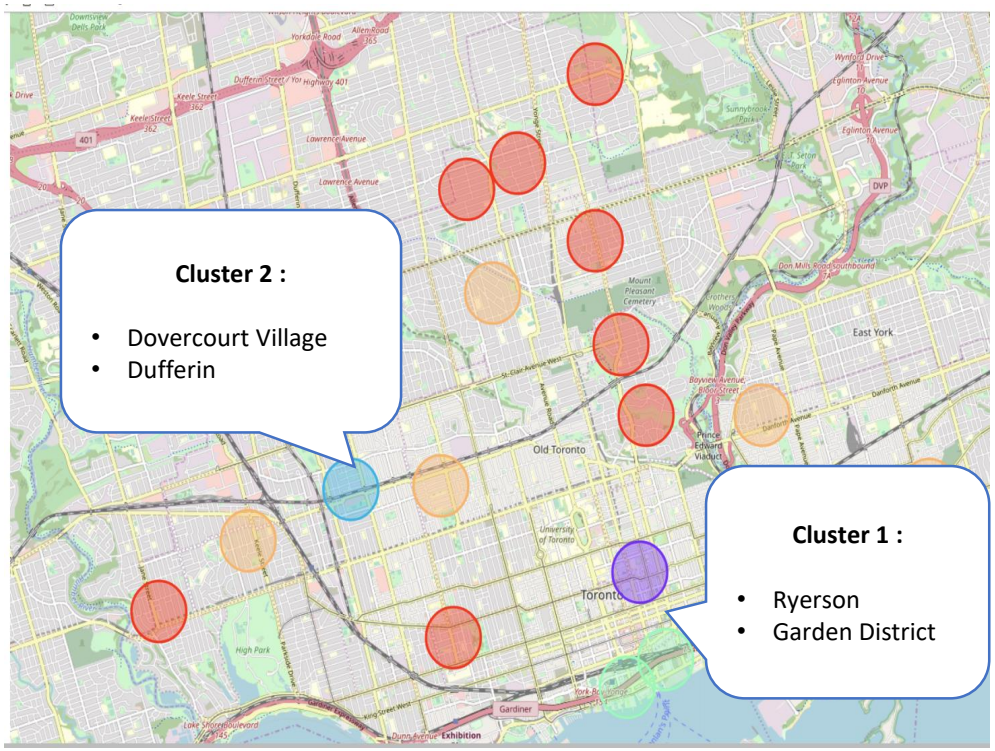
Cluster 1 scores more in Gym than Cluster 2 and vice versa

A trade-off decision

Moving to Toronto

6. Conclusion:

A choice among 4 neighbourhoods in 2 clusters



It's clear than many more criteria could be added to refine the decision adding more complexity

and a pitfall of this analysis could be the consideration of only one major area of Toronto, taking into account all the areas may be would have provided a more realistic picture.

Furthermore, this results also could potentially vary if we use some other clustering techniques like DBSCAN.

Thanks !