B. Santos

City/neighborhood similarity

July 7, 2020

Introduction

Last year, a poll found that US citizens desire to relocate to Canada[1] does hit a new record. 16% of US citizens said that they would like to permanently move to another country. Another poll [2] studies the worldwide sentiment to migrate to another country, and the favorites are the USA, Canada, Germany, France, and Australia.

Right now, we are suffering the slashes of a pandemic, COVID-19. Admittedly, this will redefine the migratory movements for individuals and company relocations around the world. Now, we can use the time to redefine this exhausting and determining kinds of projects.

We are not interested in the factors that drive people to migrate, which is a complex subject [3]. We want to help people to decide the city of destination based on the city of origin.

In this brief study, we would like to compare cities or neighborhoods, using an arbitrary set of criteria to give a rough idea of similarity between towns. For instance, we can use

the foursquare API to classify one location given a set of features, and add other ones like services and cost of living.

The question that we will try to answer is, Does a city and neighborhood in another country is suitable for a particular lifestyle?

Analytic approach

First, we need to establish the country and city of origin. Second, we define the target country, i.e., the destination. With this information, we will use ideas from recommender systems to find similarities between the cities.

K-means will be used to cluster cities in the country of origin, U.S.A., and the destination country, Canada. We will study the similarities between these clusters. We will report the destination cities that are similar to the origin cities.

We propose the following procedure,

- We will focus on the destination cities which have a nearby airport.
- We will obtain the cost of living in the cities.

Libraries and tools

1. Pandas and numpy for data manipulation.

2. BeautifulSoup and requests for web scrapping.
3. Foursquare API, bing! API, and geocoder. For venues and geolocation.
4. Folium for maps.
5. Matplotlib for visualization.
6. Dotenv for api keys manipulation.
7. Scikit-learn for k-means clustering.

Data sources

1. **Foursquare:** We will characterize the city/neighborhood main venues and services by using the foursquare API.

2. **Wikipedia/Web:** We will collect other features by web scraping, such as services available, airports, etc. Wikipedia will be a resource.

3. **expatistan.com:** The main feature will be the cost of living, which will narrow the cities of destination to compare. The site expastitan.com offers such a service.

Data features and requirements

1. Features for city similarity. At this scale we will consider the following features:
    a. Cost of living
    b. Population
    c. International airport
    d. Parks
    e. Official languages

    f. Languages
    g. Public transportation
2. Features for neighborhood clustering:
    a. Popular venues (see Future endeavours)
3. Features for neighborhood similarity
    a. Cluster to cluster (see Future endeavours)
    b. Nearest hospital (see Future endeavours)
    c. Nearest school (see Future endeavours)

Data understanding. Exploratory data analysis.

We chose the USA as the country of origin and Canada as destination.

We found 114 international airports in the USA, and 18 in Canada. In the table below, we show the international airports of Canada.

City	Airport	IATA Code
Calgary	Calgary International Airport	YYC
Edmonton	Edmonton International Airport	YEG
Whitehorse	Erik Nielsen Whitehorse International Airport	YXY
Gander	Gander International Airport	YQX

Halifax	Halifax Stanfield International Airport	YHZ
Hamilton	John C. Munro Hamilton International Airport	YHM
Kelowna	Kelowna International Airport	YLK
London	London International Airport	YXU
Moncton	Greater Moncton International Airport	YQM
Montreal	Montreal-Pierre Elliott Trudeau International Airport	YUL
Ottawa	Ottawa Macdonald-Cartier International Airport	YOW
QuebecCity	Quebec/Jean Lesage International Airport	YQB
Regina	Regina International Airport	YQR
Saskatoon	Saskatoon John G. Diefenbaker International Airport	YXE
St.John's	St. John's International Airport	YYT
Mississauga/Toronto	Toronto Pearson International Airport	YYZ
Vancouver	Vancouver International Airport	YVR

Victoria	Victoria International Airport	YYJ
Winnipeg	Winnipeg James Armstrong Richardson International Airport	YWG

Scrapping expatistan.com gives the following for the most expensive cities in North America,

Latitude	Longitude	City	State	Score	Population	Index	Country
32.293	-64.782	Hamilton		0.00548	2000	294	BM
37.3928	-122.042	Mountain View	California	0.00421	74066	259	US
40.7143	-74.006	New York City		0.00410	8008278	256	US
37.7793	-122.419	San Francisco	California	0.00362	808976	244	US
40.7114	-74.0647	Jersey City	New Jersey	0.00311	247000	232	US

Next we show, a table for the U.S. cities with the scores:

City	TransportScaledScore	ParksScaledScore	CostScaledScore	HealthScaledScore	WealthScaledScore
Akron	0.5520833333	0.6243567753	0.614383204	0.3896457766	0.2097902098
Albany	0.84375	0.6243567753	0.614383204	0.7103931491	0.5454545455
Albuquerque	0.5104166667	0.8080808081	0.9642154825	0.2031919035	0.013986014
Anchorage	0.3645833333	0.6161616162	0.4964561483	0.3281432464	0
Appleton	0.40625	0.6243567753	0.614383204	0.281432464	0.0909090909

The same for Canada cities,

City	TransportScaledScore	ParksScaledScore	CostScaledScore	HealthScaledScore	WealthScaledScore
Calgary	0.8761329709	0.6153846154	0.6388161002	0.7000258004	0.4001562093
Edmonton	0.6888216457	0.6923076923	0.845613343	0.7000258004	0.4001562093
Gander	0.5203929576	0.6009615385	0.6942606072	0.9060324646	0.0351470971
Halifax	0.3323262993	0.9230769231	1	1	0.0481645405
Hamilton	0.640483321	0.4615384615	0.6942606072	0.518364851	1

Results and conclusions

From these scores we can compute the similarity between each city of origin and each city of destination. Then we kept the 3 most similar cities for each U.S. city. For the first part we neglect the languages.

Origin	Destination
Akron	['Montreal', 'Edmonton', 'Winnipeg']
Albany	['Calgary', 'Montreal', 'Edmonton']
Albuquerque	['Whitehorse', 'Saskatoon', 'Regina']
Anchorage	['Moncton', 'Halifax', 'Winnipeg']
Appleton	['Saskatoon', 'Montreal', 'Winnipeg']
Atlanta	['Saskatoon', 'Winnipeg', 'Regina']
Atlantic City	['Montreal', 'Calgary', 'Edmonton']
Austin	['Hamilton', 'Ottawa', 'London']
Baltimore	['Victoria', 'Calgary', 'Winnipeg']
Bangor	['Halifax', 'Gander', 'Moncton']
Birmingham	['Moncton', 'Whitehorse', 'Halifax']
Boise	['Whitehorse', 'Saskatoon', 'Regina']
Boston	['Victoria', 'Vancouver', 'Calgary']
Bradenton	['Montreal', 'Edmonton', 'Saskatoon']
Buffalo	['Victoria', 'Calgary', 'Montreal']
Charleston	['Whitehorse', 'Saskatoon', 'Regina']
Charlotte	['Saskatoon', 'Montreal', 'Winnipeg']
Chicago	['Victoria', 'Regina', 'Winnipeg']
Cleveland	['Montreal', 'Saskatoon', 'Regina']

Columbus	['Montreal', 'Edmonton', 'Calgary']
Covington	['Saskatoon', 'Regina', 'Whitehorse']
Dallas	['Montreal', 'Ottawa', 'Hamilton']
Dayton	['Montreal', 'Winnipeg', 'Edmonton']
Denver	['Regina', 'Winnipeg', 'Victoria']
Des Moines	['Saskatoon', 'Whitehorse', 'Regina']
Detroit	['Montreal', 'Saskatoon', 'Edmonton']
Durham	['Montreal', 'Winnipeg', 'Edmonton']
Fairbanks	['Saskatoon', 'Whitehorse', 'Montreal']
Fort Lauderdale	['Montreal', 'Winnipeg', 'Edmonton']
Fort Myers	['Montreal', 'Edmonton', 'Saskatoon']
Fresno	['Hamilton', 'London', 'Ottawa']
Grand Rapids	['Montreal', 'Edmonton', 'Winnipeg']
Green Bay	['Saskatoon', 'Winnipeg', 'Montreal']
Greensboro	['Montreal', 'Edmonton', 'Winnipeg']
Hartford	['Victoria', 'Winnipeg', 'Calgary']
Hilo	['Halifax', 'Moncton', 'Gander']
Honolulu	['Victoria', 'Winnipeg', 'Calgary']
Houston	['Hamilton', 'Montreal', 'London']
Huntsville	['Whitehorse', 'Saskatoon', 'Moncton']
Indianapolis	['Whitehorse', 'Saskatoon', 'Regina']
Jacksonville	['Quebec', 'Kelowna', 'Montreal']
Juneau	['Moncton', 'Halifax', 'Edmonton']
Kailua	['Moncton', 'Halifax', 'Gander']
Kansas City	['Montreal', 'Edmonton', 'Saskatoon']
Ketchikan	['Saskatoon', 'Montreal', 'Whitehorse']
Key West	['Montreal', 'Edmonton', 'Saskatoon']

Knoxville	['Saskatoon', 'Whitehorse', 'Winnipeg']
Lakeland	['Montreal', 'Edmonton', 'Moncton']
Lansing	['Montreal', 'Edmonton', 'Calgary']
Las Vegas	['Whitehorse', 'Saskatoon', 'Regina']
Little Rock	['Whitehorse', 'Saskatoon', 'Regina']
Los Angeles	['Ottawa', 'Toronto', 'Hamilton']
Louisville	['Whitehorse', 'Saskatoon', 'Quebec']
Melbourne	['Montreal', 'Edmonton', 'Moncton']
Memphis	['Kelowna', 'Montreal', 'Whitehorse']
Mesa	['Moncton', 'Halifax', 'Whitehorse']
Miami	['Regina', 'Victoria', 'Winnipeg']
Midland	['Hamilton', 'London', 'Ottawa']
Milwaukee	['Saskatoon', 'Winnipeg', 'Regina']
Myrtle Beach	['Whitehorse', 'Moncton', 'Saskatoon']
Nashville	['Whitehorse', 'Saskatoon', 'Montreal']
New Orleans	['Saskatoon', 'Regina', 'Whitehorse']
New York City	['Vancouver', 'Toronto', 'Victoria']
Newark	['Victoria', 'Regina', 'Montreal']
Newburgh	['Kelowna', 'Edmonton', 'Halifax']
Norfolk	['Saskatoon', 'Montreal', 'Regina']
Oakland	['Ottawa', 'Toronto', 'Hamilton']
Oklahoma City	['Whitehorse', 'Kelowna', 'Saskatoon']
Omaha	['Whitehorse', 'Saskatoon', 'Regina']
Ontario	['Hamilton', 'Ottawa', 'London']
Orange County	['Hamilton', 'Ottawa', 'London']
Orlando	['Montreal', 'Edmonton', 'Calgary']
Palm Springs	['Ottawa', 'Hamilton', 'London']

Panama City	['Montreal', 'Edmonton', 'Moncton']
Pensacola	['Montreal', 'Edmonton', 'Saskatoon']
Philadelphia	['Moncton', 'Halifax', 'Edmonton']
Phoenix	['Montreal', 'Saskatoon', 'Whitehorse']
Pittsburgh	['Victoria', 'Montreal', 'Edmonton']
Portland	['Victoria', 'Regina', 'Calgary']
Portland-Maine	['Gander', 'Victoria', 'Edmonton']
Providence	['Gander', 'Halifax', 'Edmonton']
Racine	['Saskatoon', 'Regina', 'Montreal']
Reno	['Whitehorse', 'Saskatoon', 'Regina']
Richmond	['Saskatoon', 'Regina', 'Montreal']
Rochester	['Edmonton', 'Kelowna', 'Montreal']
Rochester-Maine	['Moncton', 'Halifax', 'Gander']
Rockford	['Edmonton', 'Montreal', 'Calgary']
Sacramento	['Ottawa', 'Hamilton', 'London']
Salt Lake City	['Regina', 'Saskatoon', 'Winnipeg']
San Antonio	['London', 'Hamilton', 'Kelowna']
San Bernardino	['Ottawa', 'Hamilton', 'London']
San Diego	['Ottawa', 'Hamilton', 'London']
San Francisco	['Toronto', 'Ottawa', 'Hamilton']
San Jose	['Ottawa', 'Hamilton', 'London']
Sanford	['Montreal', 'Edmonton', 'Saskatoon']
Savannah	['Saskatoon', 'Whitehorse', 'Regina']
Scranton	['Edmonton', 'Montreal', 'Calgary']
Seattle	['Regina', 'Victoria', 'Winnipeg']
Snohomish County	['Montreal', 'Edmonton', 'Winnipeg']

Spokane	['Montreal', 'Winnipeg', 'Edmonton']
St. Louis	['Moncton', 'Halifax', 'Whitehorse']
St. Paul	['Victoria', 'Moncton', 'Regina']
St. Petersburg	['Montreal', 'Edmonton', 'Winnipeg']
Syracuse	['Edmonton', 'Calgary', 'Montreal']
Tallahassee	['Montreal', 'Edmonton', 'Calgary']
Tampa	['Montreal', 'Edmonton', 'Calgary']
Washington	['Regina', 'Winnipeg', 'Saskatoon']
West Palm Beach	['Moncton', 'Edmonton', 'Montreal']
Wilmington	['Montreal', 'Edmonton', 'Saskatoon']

For the second part we added new features. These features correspond to languages spoken at home. The city similarity study gives:

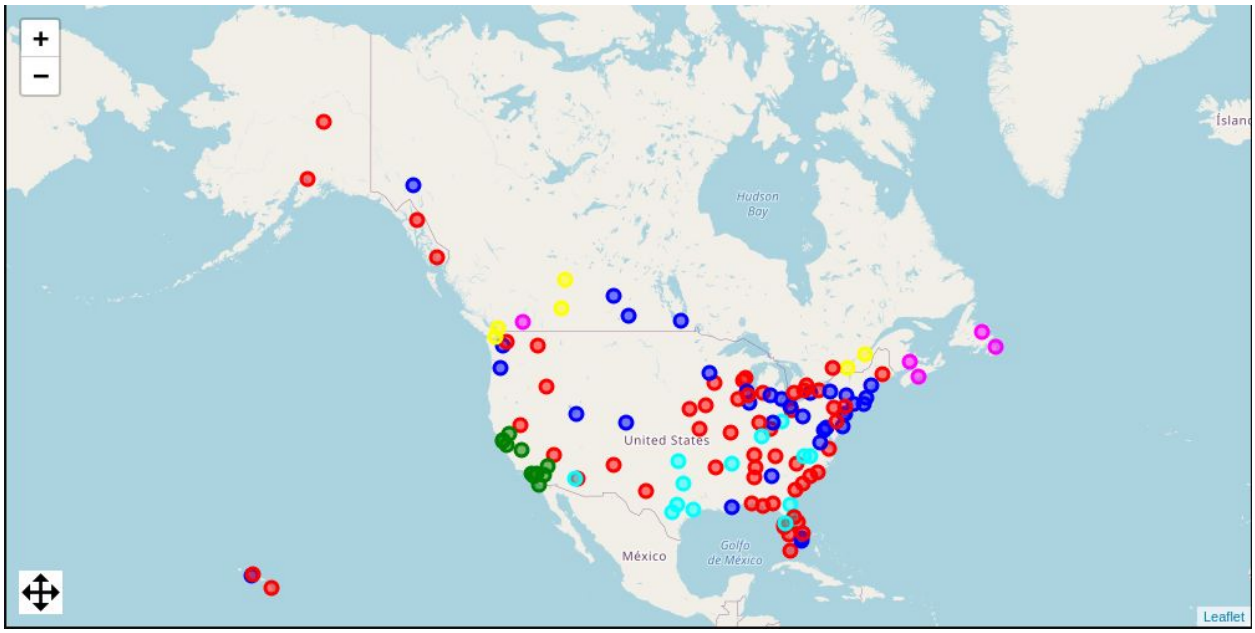
Origin	OCluster	Destination 1	DCluster 1	Destination 2	DCluster 2	Destination 3	DCluster 3
New York City	0	Toronto	2	Hamilton	1	Montreal	1
St. Paul	1	Ottawa	1	Kelowna	3	Victoria	3
Richmond	1	Calgary	0	Ottawa	1	Edmonton	0
Boston	1	Toronto	2	Ottawa	1	Edmonton	0
Baltimore	1	Ottawa	1	Edmonton	0	Toronto	2
San Diego	1	Vancouver	0	Calgary	0	Toronto	2
Washington	1	Ottawa	1	Calgary	0	Edmonton	0
Spokane	1	Edmonton	0	Vancouver	0	Winnipeg	0
Omaha	2	Saskatoon	0	Edmonton	0	Kelowna	3

Las Vegas	2	Edmonton	0	Kelowna	3	Ottawa	1
Albuquerque	2	Kelowna	3	Saskatoon	0	Whitehorse	3
Charlotte	2	Edmonton	0	Ottawa	1	Calgary	0
Huntsville	2	Moncton	3	London	1	Saskatoon	0
Oklahoma City	2	Edmonton	0	Halifax	3	Calgary	0
Philadelphia	2	Edmonton	0	London	1	Toronto	2
Providence	2	Toronto	2	Ottawa	1	Montreal	1
Charleston	2	Kelowna	3	Saskatoon	0	Regina	0
Memphis	2	London	1	Edmonton	0	Hamilton	1
Houston	2	Edmonton	0	Calgary	0	Ottawa	1
Salt Lake City	2	Kelowna	3	Victoria	3	Vancouver	0
Columbus	2	Kelowna	3	Hamilton	1	Toronto	2
Portland	2	Kelowna	3	Edmonton	0	Victoria	3
Detroit	2	Toronto	2	Ottawa	1	London	1
Bangor	2	Quebec	3	Ottawa	1	Montreal	1
Anchorage	2	Victoria	3	Whitehorse	3	Kelowna	3
Phoenix	2	Ottawa	1	Kelowna	3	Edmonton	0
Little Rock	2	Calgary	0	Ottawa	1	Edmonton	0
Denver	2	Calgary	0	Kelowna	3	Edmonton	0
Tampa	2	Kelowna	3	Victoria	3	Whitehorse	3
Atlanta	2	Edmonton	0	Ottawa	1	Calgary	0
St. Louis	2	Ottawa	1	Calgary	0	Edmonton	0
Honolulu	2	Vancouver	0	Whitehorse	3	Winnipeg	0
Indianapolis	2	Kelowna	3	Hamilton	1	Saskatoon	0
Des Moines	2	Kelowna	3	Calgary	0	Saskatoon	0
Covington	2	Kelowna	3	Edmonton	0	Saskatoon	0
New Orleans	2	Quebec	3	Montreal	1	Ottawa	1

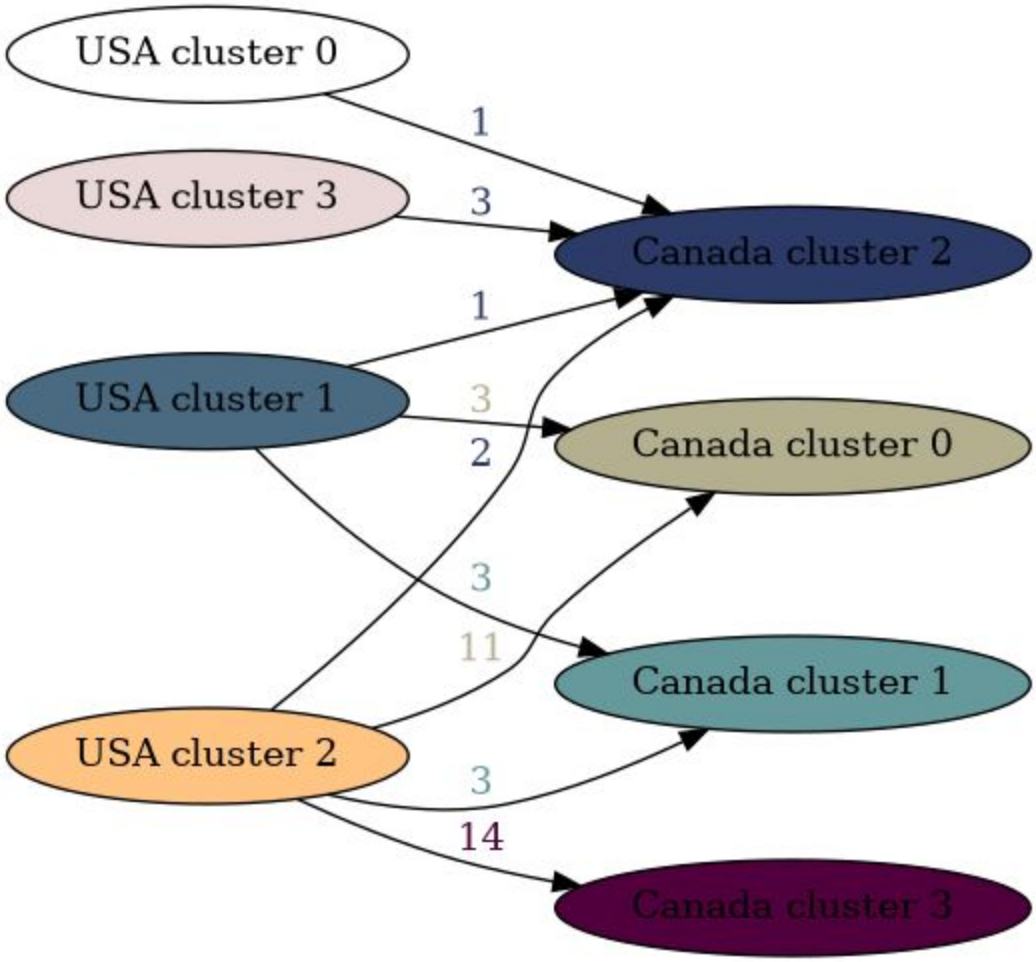
Boise	2	Vancouver	0	Edmonton	0	Saskatoon	0
Milwaukee	2	Kelowna	3	London	1	Victoria	3
Chicago	3	Toronto	2	Hamilton	1	Calgary	0
Hartford	3	Toronto	2	Hamilton	1	London	1
Newark	3	Toronto	2	Edmonton	0	Ottawa	1

U.S. and Canada cities clustering:

We used k-means clustering, to find clusters of cities in the U.S. and Canada:

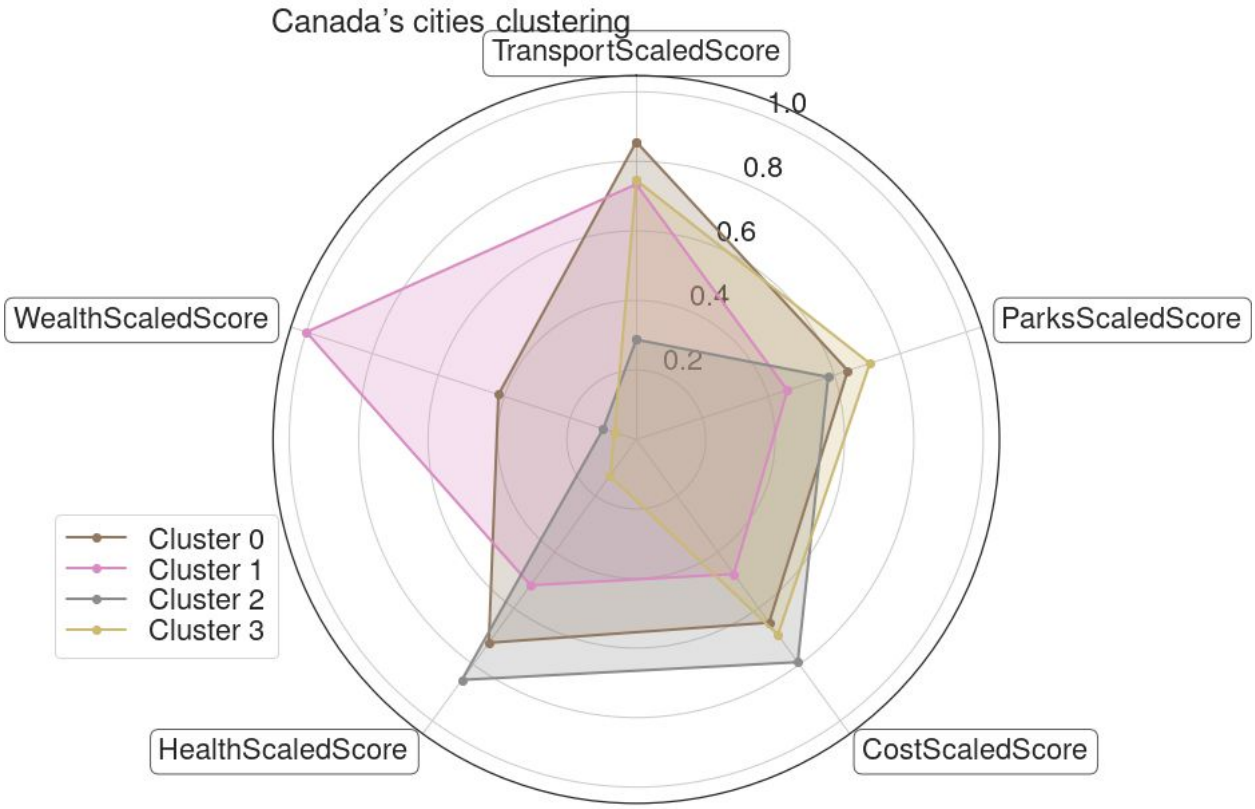


The relation between clusters (see table above) is shown in the next figure.



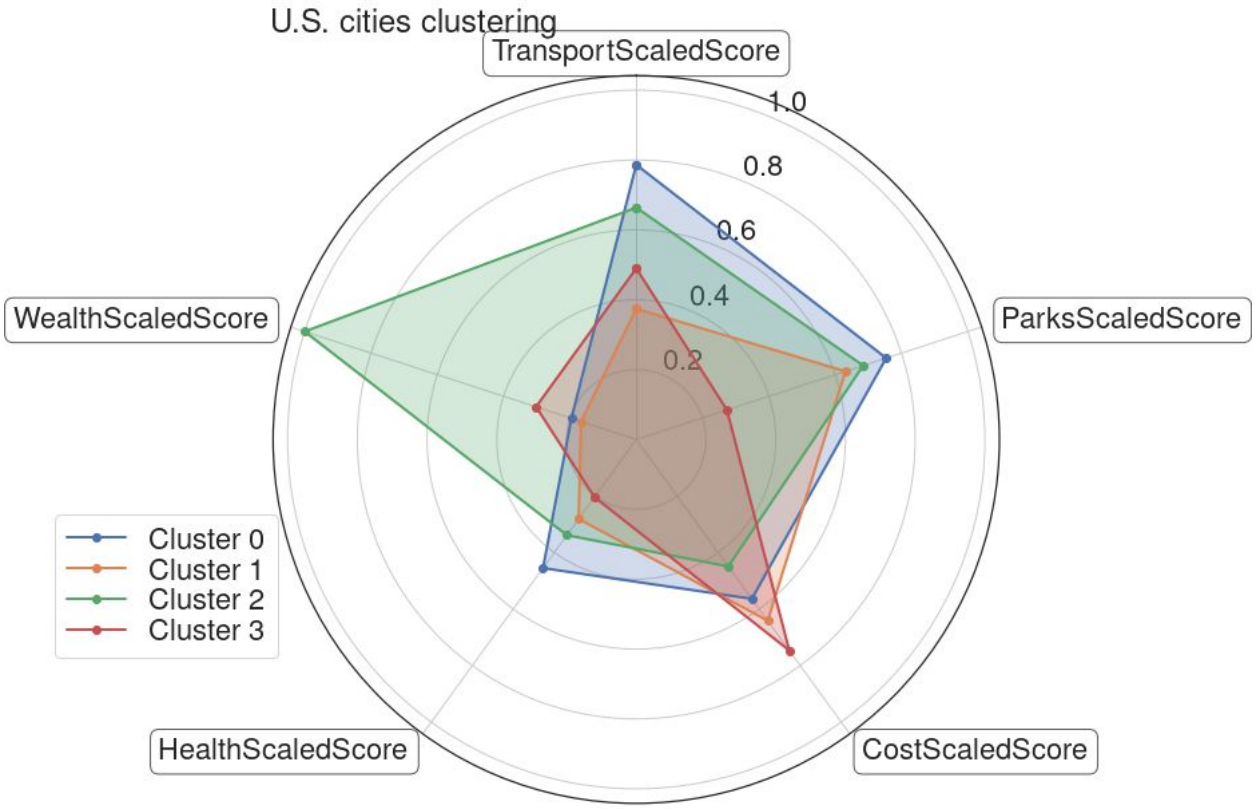
We clearly see that U.S. cities in cluster 1 and 2, are related to Canada clusters 1 and 3.

If we plot the scores in radar plots we have,



We can identify,

- Cluster 0: is the largest with the high transport score, and good parks.
- Cluster 1: is balanced.
- Cluster 2: is expensive with a good health score.
- Cluster 3: has the best park score but the lower cost of life and bad health.

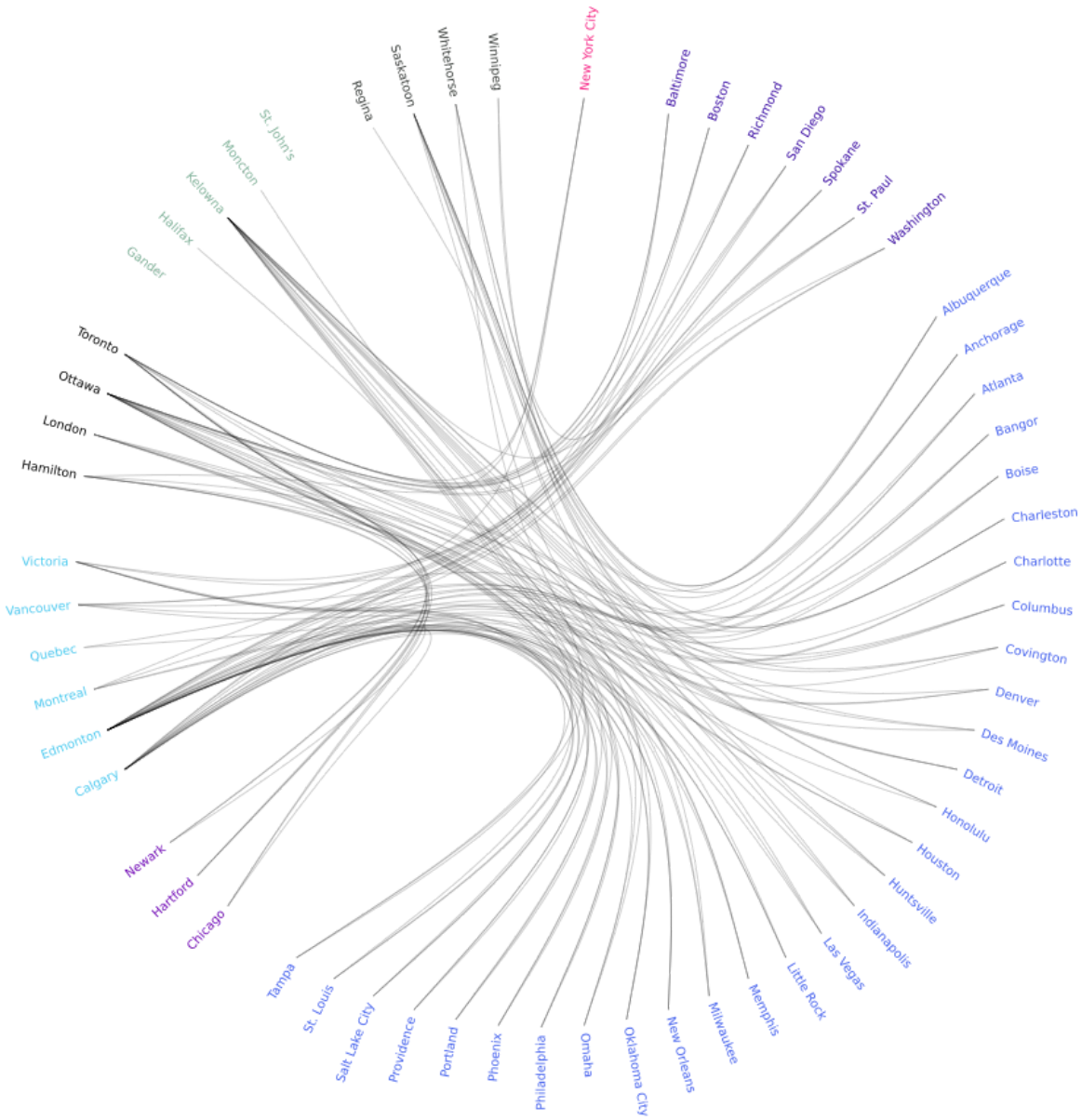


Similarly, for U.S. cities we have:

- Cluster 0: New York.
- Cluster 1: has the lower cost of life, bad health score and worst gdp score.
- Cluster 2: excellent wealth score and average in the other features.

The similarity between cluster 1 in the U.S. and cluster 1 in Canada is given by the fact that they have very balanced scores.

The relation between cities is explored in the following figure:



Future endeavours

When a couple of suitable destination cities are found, we could use k-means clustering to group neighborhoods with similar features, using the Foursquare API. Also, the same analysis can be performed in the community of origin. Finally, we can study the destination boroughs that are similar to the neighborhood of origin with all this data.

We propose the following procedure,

- Given a borough and city of origin, we will use postal code information to get the borough of the destination cities, given the similarities studied in this report.
- We could use bing to get the coordinates of the boroughs.
- We could use follium to display the map and the boroughs.
- We could use foursquare API to get the most popular venues in destination and origin cities.
- We could use k-means clustering to cluster the boroughs with common features.
- We could determine the similarity between the borough of origin and each cluster in the city of destination, using a recommendation engine.

Example of a case of study:

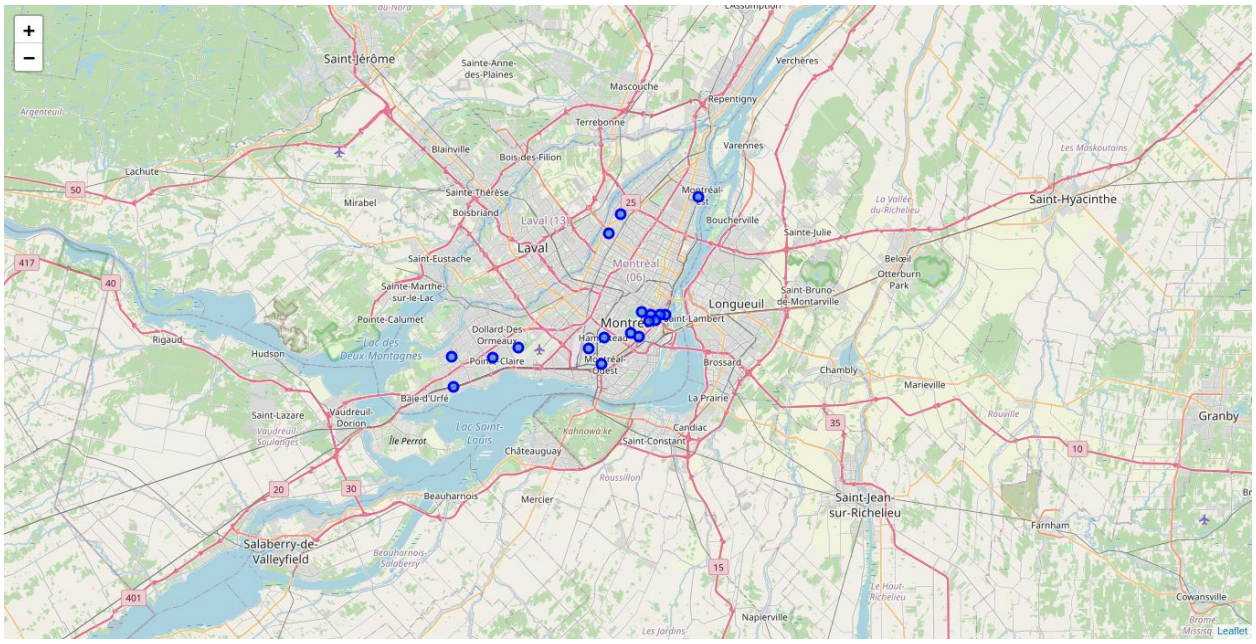
Suppose that the city of origin is Dallas and we found that a similar city to migrate is Montreal. We can compare the cities using the cost of living, expatistan gives: "*Montreal is 14% cheaper than Dallas, Texas. Jul 2020 Cost of Living*".

From the postal codes of boroughs in each city, we obtained the coordinates and popular venues. For instance, these are the boroughs of Montreal.

Borough	PostalCode	Latitude	Longitude
Montreal East	H1B	45.629	-73.505
Montreal North North	H1G	45.612	-73.621
Montreal North South	H1H	45.591	-73.639
Old Montreal	H2Y	45.506	-73.554
Downtown Montreal Northeast	H2Z	45.505	-73.562
Downtown Montreal North	H3A	45.505	-73.575
Downtown Montreal East	H3B	45.500	-73.569
Downtown Montreal Southeast	H3G	45.498	-73.579
Downtown Montreal South & West	H3H	45.508	-73.589
Hampstead	H3X	45.482	-73.646
Westmount West	H3Y	45.486	-73.606
Westmount East	H3Z	45.482	-73.593
Cote-Saint-Luc West	H4W	45.470	-73.669

Montreal West	H4X	45.453	-73.650
Kirkland	H9J	45.461	-73.874
Dorval Outskirts	H9P	45.471	-73.774
Pointe-Claire	H9R	45.460	-73.813
Beaconsfield	H9W	45.430	-73.873

By using Folium the boroughs can be located in a map, as shown below.



References

1. https://news.gallup.com/poll/245789/record-numbers-americans-leave.aspx?
2. https://news.gallup.com/poll/245255/750-million-worldwide-migrate.aspx
3. Francesco Castelli, Drivers of migration: why do people move?, Journal of Travel Medicine, Volume 25, Issue 1, 2018, tay040, https://doi.org/10.1093/jtm/tay040