Montreal or Toronto: the Battle of the Neighborhoods

How similar or dissimilar are the neighborhoods?

1. Introduction

Canada is accepting new immigrants at a far higher rate than the US [insert citation]. When applying for permanent residence or citizenship, applicants may face difficulties when comparing Canada's metropolitan cities. Moving across the world to a new country is already challenging. This project aims to capture the essence of the two cities so much that it would be possible for newcomers to Canada to choose a place to live based on the lifestyle.

This project will focus on three types of potential newcomers and advise where to move or not move based on their characteristics. First is the couple who either is looking to start a family or already has young kids. This family will not want to live close to the nightlife scene but would appreciate parks and playgrounds. The second newcomer is a young professional who's not a fan of their cooking; they want plenty of restaurant choices around them. The last newcomer is very active, basketball, tennis, soccer, hockey, you name they'll play it. They'd prefer sports fields and courts and gyms. Similar to the young family, they also aren't interested in the nightlife scene.

To capture each neighborhood's essence, the unsupervised machine learning algorithm k-means clustering was used, along with two values for k, five, and ten. The results will be calculated based on the most effective value.

2. Data

2.1 Datasets

Since the project focuses on the two cities' neighborhoods, the datasets weren't complicated or full of details. They included the postal code and neighborhood provided by Wikipedia pages here and here. Furthermore, the Toronto longitude and latitude were available by the CSV file provided by IBM here. While a defined function inserted the Montreal longitude and latitude via Geocoder.

2.2 Data Cleaning

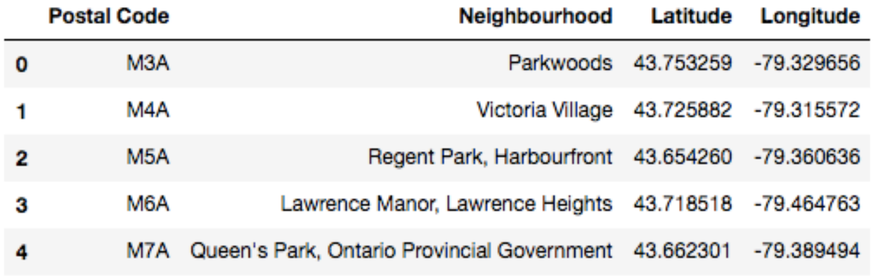
For this project, the data cleaning wasn't labor-intensive. Since the Toronto dataset was scraped on Wikipedia, it was necessary to drop the Boroughs column and anything labeled 'Not assigned.' Then reset the index and merge the longitude and latitude CSV file to the scraped HTML dataset.

The Montreal dataset had less cleaning as the data was entered manually. It was less labor-intensive to enter the data manually then to scrap the web page since the way the information was stored on Wikipedia was not clear when imported to a data frame. This allowed for manual removal of the neighborhoods labeled 'Not assigned.' Furthermore, postal codes beginning with H7 were also skipped, because that's the island of Laval. The first postal code and neighborhood on the Wikipedia page are H0H reserved for Santa Clause, which was skipped over. The next postal code and neighborhood is H0M for Akwesasne Region, which is on the border of Ontario, Quebec, and New York state, nowhere near Montreal.

The Montreal Wikipedia page also didn't include the longitude and latitude like the Toronto Wikipedia page. Therefore, it was necessary to define a function using Geocoder that would calculate the longitude and latitude for each neighborhood based on the postal code.

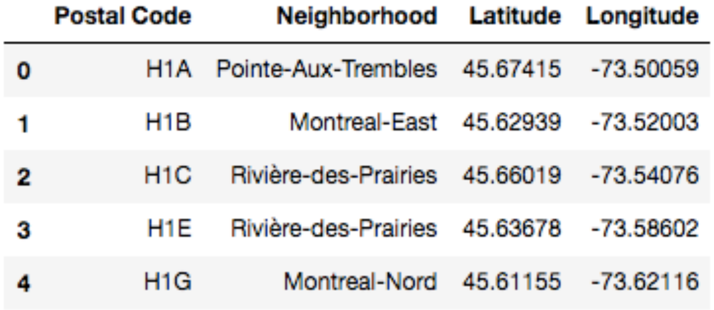
The difference between the two datasets is that Toronto has multiple neighborhoods per postal code, whereas Montreal has one neighborhood per postal code.

Figure 1: Toronto Dataset



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

Figure 2: Montreal Dataset



	Postal Code	Neighborhood	Latitude	Longitude
0	H1A	Pointe-Aux-Trembles	45.67415	-73.50059
1	H1B	Montreal-East	45.62939	-73.52003
2	H1C	Rivière-des-Prairies	45.66019	-73.54076
3	H1E	Rivière-des-Prairies	45.63678	-73.58602
4	H1G	Montreal-Nord	45.61155	-73.62116

2.3 Foursquare

To capture the essence of each neighborhood, it's essential to focus on the most popular venues. For example, if there's a lot of bars and clubs in one neighborhood, it may be loud at night. Moreover, if there's a lot of playgrounds, it may be a more family-friendly neighborhood.

There are a few parameters that need to be established when using Foursquare API; client ID and secret, limit, version, latitude, longitude, and radius. Of course, to access the Foursquare location data, one must have an account, here is where you'll find your client ID and secret. The limit parameter is defined as how many venues you'd like the API to return. In the test and k-means clustering, the limit was set as 100 venues. The version parameter is the date. Capturing the neighborhoods' essence, it was critical to focus on before COVID-19, instead of taking the time when the project started. The parameters of longitude and latitude are defined in the data frames based on the postal codes. The radius is how many meters from the defined latitude and longitude the Foursquare API will search for venues. This project focused on a 500-meter radius since the longitude and latitude are the centers of the towns.

3. K-Means Clustering

K-means clustering is a type of portioning clustering. K-means can group data based on their similarities or dissimilarities into non-overlapping segments. Objects within the same cluster are similar, whereas objects across different clusters are dissimilar. Deciding the value for k is considered a hard problem because there's no clear-cut answer. In this case, we tested the values of k as five and ten. Comparing the results from k =5 and k = 10, it was clear that k =10 was more accurate as the clusters are more homogenous.

3.1 Toronto Clusters

First, we tested k = 5, which resulted in figure 3. As visualized, the map of Toronto is dominated by the second cluster or, in this case, the purple markers. When k = 10 was tested next, the dominant cluster didn't change much. However, the second most predominant cluster, the red markers, was broken down into various clusters, such as light blue and orange markers.

Moving from k = 5 to k =10 was expected to modify the dominant cluster much more than it did. This indicates that the dominant cluster is more homogenous than initially thought. Furthermore, it was expected that the clusters would be along the lines of or geography. However, the further you get from downtown Toronto, the neighborhood's essence does not change as much as initially thought. This proves that Toronto neighborhoods' core isn't based on location at all.

Figure 3: Toronto Map When k = 5

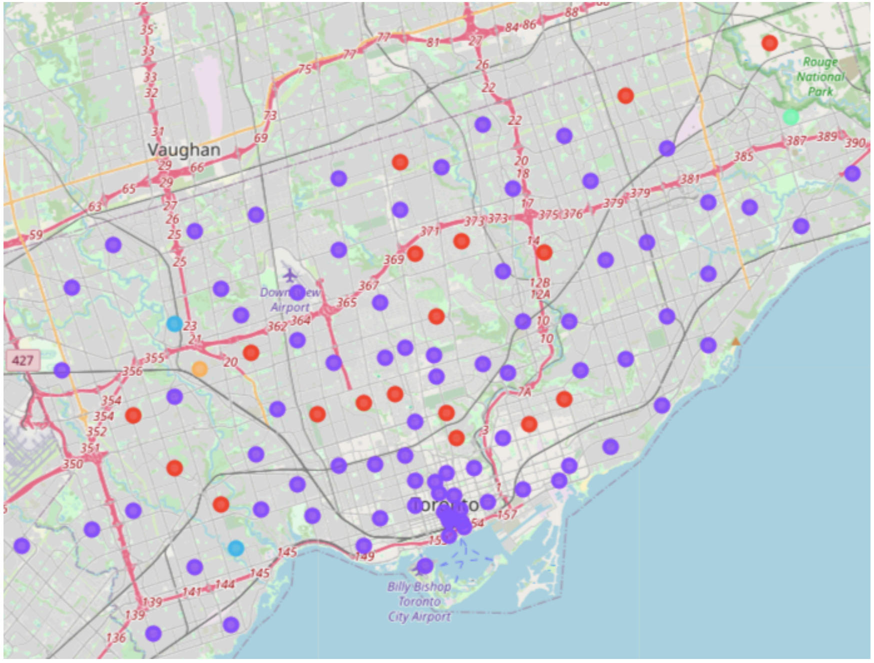
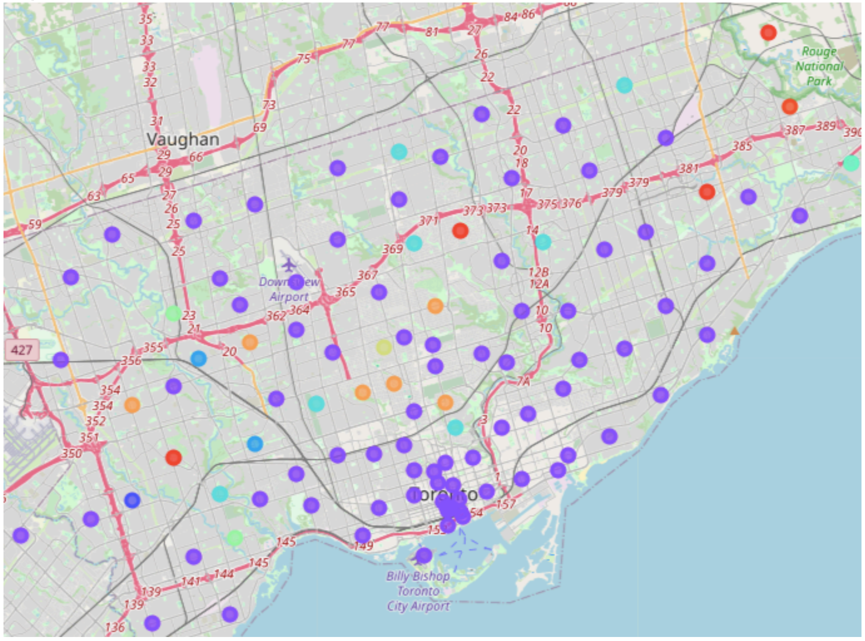


Figure 4: Toronto Map When k = 10



3.2 Montreal Clusters

Unlike Toronto, Montreal changed a lot when k = 10. The red marks in figure 5 split into three groups, red markers, purple markers, and light red markers as predicted Montreal's clusters did partition one geographic line to a certain extent. The west island is heterogeneous compared to the rest of the island. There are four unique clusters with only one neighborhood: Senneville, St. Anne-de-Bellevue, Beaconsfield, and Kirkland. Similarly, the east-end was different from the rest of the island. Moreover, the second cluster (purple markers) breaks up the ninth (light red markers), by cutting through the island entirely, proving that is some segmentation based on geographical lines.

Unlike Toronto, the map wasn't divided based on income. It was expected that high-income neighborhoods such as Westmount wouldn't be in a cluster with lower-income neighborhoods such as Montreal-Nord. This was true for the first cluster when k =10. However, in the most significant cluster, light red markers, high-income neighborhoods such as Mount-Royal and Outremont, are paired with low-income neighborhoods such as Montreal-Nord. Therefore, it can be concluded that income is not a determining factor for the clusters.

Figure 5: Montreal Map When k = 5

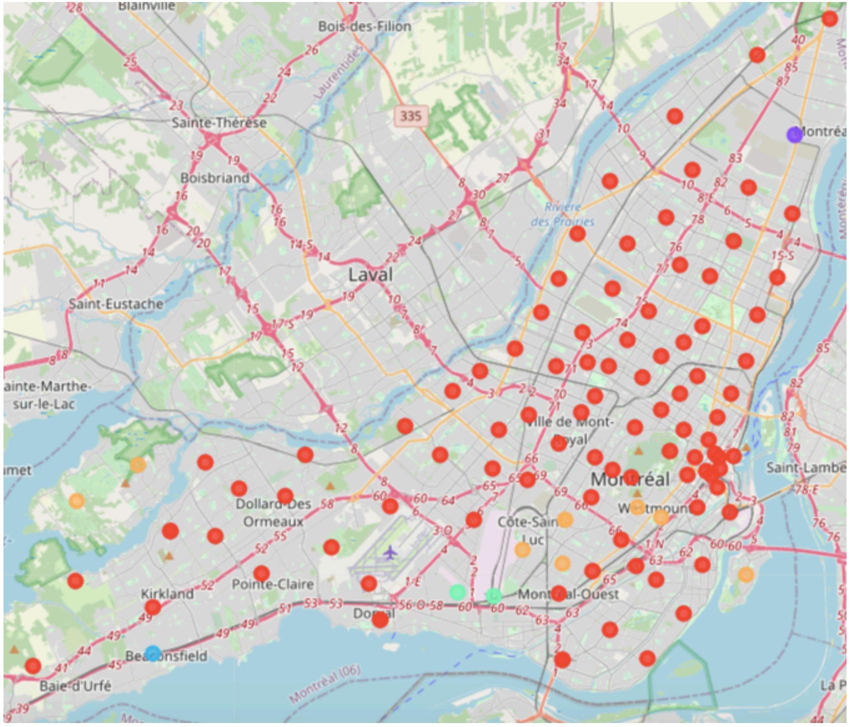
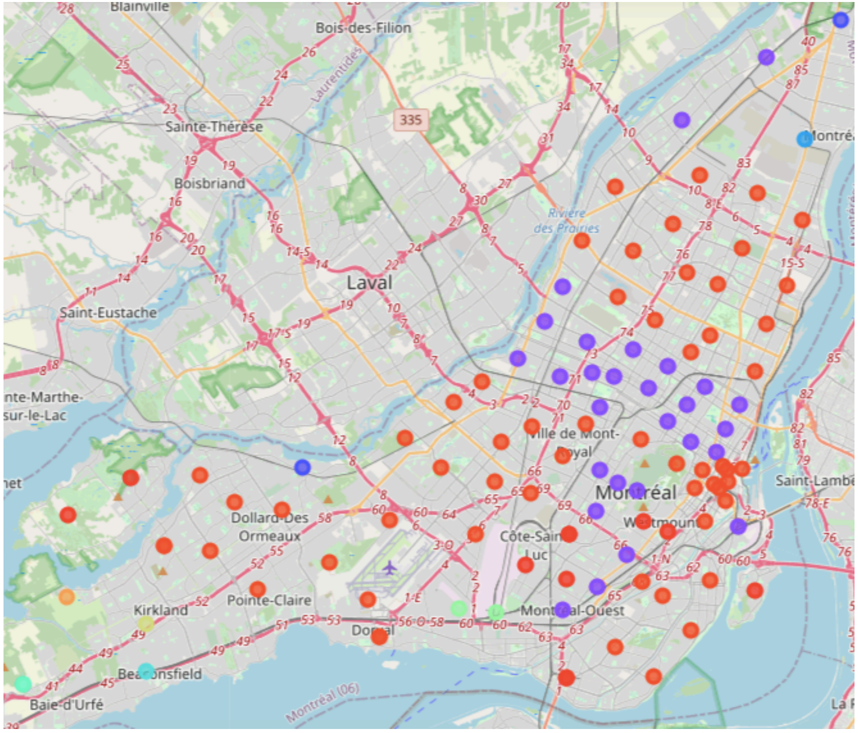


Figure 6: Montreal Map When k = 10



4. Results

What neighborhoods recommendation for each newcomer? Let's remember their characteristics. First, the couple with a young family or looking to start a family soon. They want a family-friendly neighborhood with parks and playgrounds for their children. They should look into the cluster of the fifth and ninth cluster in Toronto. The fifth cluster includes the following neighborhoods; Parkwoods, Caledonia-Fairbanks, Willowdale, Newtonbrook, York Mills West, Milliken, Agincourt North, Steeles East, Rosedale, The Kingsway, Montgomery Road, and Old Mill North. The ninth cluster includes the following neighborhoods; Humewood-Cedarvale, North Park, Map Leaf Park, Upwood Park, Lawrence Park, Forest Hill North & West, Forest Hill Road Park, Kingsview Village, St. Phillips, Martin Grove, Moore Park, and Summerhill East.

The second newcomer is a young professional with a demanding job who doesn't like their cooking. They're looking for a neighborhood with many restaurant choices. They should choose a neighborhood within Toronto's second cluster, such as Glencairn, Bedford Park, and Lawrence Manor East, all of which have their top four most common venues like restaurants. The young professional would also be happy in one of Montreal's neighborhoods in the tenth cluster, such as Lasalle, Saint Laurent, Montreal-Nord, and Dollard-des-Ormeaux. All of which have at least three of their top four top venues as some type of restaurant.

The final newcomer is athletic. They're interested in sports courts and fields. They aren't interested in the nightlife scene. They should choose a neighborhood in Toronto's seventh cluster; Humberlea, Emery, Old Mill South, King's Mill Park, and Sunnylea, which have a baseball field as their most common venue. They also have options in Montreal's first cluster, such as Nun's Island with a tennis court or Ile-Bizard with a hockey arena.

5. Conclusion

Canada is accepting new immigrants at a far higher rate than the US [insert citation]. When applying for permanent residence or citizenship, applicants may face difficulties when comparing Canada's metropolitan cities. This project focuses on three types of potential newcomers and provides advice on where to move or not to move based on their characteristics; a young family, a young professional, and an athlete.

To capture the essence of each neighborhood in the Canadian cities, unsupervised machine learning algorithm k-means clustering was used. Two values for k were tested, five and ten. Ultimately k = 10 was the most effective.

To conclude, the young family of newcomers should choose a neighborhood within the Toronto fifth and ninth clusters, including neighborhoods such as Willowdale and Kingsway. The young profession who likes to eat out has the most choices in both cities as their options are in the largest cluster. Some of neighborhood options are Glencairn, Toronto, or Lasalle, Montreal. The last newcomer is an athlete. Their lifestyle would hit in Humberlea, Toronto, or Nun's Island, Montreal.