INTERCITY NEIGHBORHOOD CLUSTERING

Efren Mora

2020-09-01

Efren Mora
2020-09-01

1. Introduction

A small chain business owner stablished in the city of Toronto is looking to expand its business to another big city in Canada and is particularly interested in Montreal. The idea of expanding is new, and she wants to be thorough. She is quite familiar with the city of Toronto and its neighborhoods, but not so much with the city of Montreal. So, to be better oriented and get a better feel of this new city, she has asked for help identifying and mapping equivalent neighborhoods in these two cities and to carry a comparative analysis among these in terms of venue composition and census features.

Please note that the nature of the business was not specified. That is because the current analysis will aid business owners in different business categories make better-informed decisions moving forward. This study is meant to be carried out on the initial phases when considering a new location for your business. At a subsequent stage, a more business-specific analysis will have to be done.

2. Data Sources

Foursquare API:

- Main tool. It was used to virtually explore Toronto and Montreal Neighborhoods.
- When making 'GET Venues' requests to the Foursquare API, I used the 'search' endpoint to pick-up venues within specific category IDs.
- I used all the header (or main) category IDs from the Foursquare category tree. All other category IDs will ultimately fall underneath one these header category IDs. For example: A 'Japanese Restaurant' venue category will fall underneath the main venue category 'Food'.
    - Arts & Entertainment: 4d4b7104d754a06370d81259
    - College & University: 4d4b7105d754a06372d81259
    - Event: 4d4b7105d754a06373d81259
    - Food: 4d4b7105d754a06374d81259
    - Nightlife Spot: 4d4b7105d754a06376d81259
    - Outdoors & Recreation: 4d4b7105d754a06377d81259
    - Professional & Other Places: 4d4b7105d754a06375d81259
    - Residence: 4e67e38e036454776db1fb3a
    - Shop & Service: 4d4b7105d754a06378d81259
    - Travel & Transport: 4d4b7105d754a06379d81259
- For the complete list of Foursquare venue category IDs please visit https://developer.foursquare.com/docs/build-with-foursquare/categories/
- We deliberately searched for these venue categories in each neighborhood and gathered the data to create a features pool.

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

- To explore Toronto's and Montreal's neighborhoods, we first need a list of their names and addresses. In this case, the address to a Canadian neighborhood is the first 3 digits of a postal code, formally referred to as a forward sortation area or FSA for short.

Geocoder (Powered by Bing search engine):

- Because we are using Foursquare to explore the cities, and its API accept addresses in the form of latitude and longitude coordinates, FSAs need to be converted to geographical coordinates. Here is where Geocoder comes into play, facilitating this conversion.

FSA Boundary File from Census 2016: https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2016-eng.cfm

- To show the FSA neighborhoods boundaries superimposed on a map of the city of Toronto, and Montreal.

Population by FSA from Census 2016: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/comprehensive.cfm

- To calculate and include population density as a feature for neighborhood clustering.

Median total income of households per FSA from Census Profile, 2016 Census: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E

- To include median income by FSA as a feature for neighborhood clustering.
- The client would want to know the median income of families living in or near the neighborhoods that make up a particular cluster.

3. Methodology

The goal was to identify and group into clusters neighborhoods that share similar venues composition and certain census attributes such as population density and median income of households. By venues composition we mean the most frequent venue category types found in a neighborhood or neighborhood cluster.

Between the two cities, I had just over 200 neighborhoods, and on average 148 features per neighborhood were collected. Identifying which neighborhoods are similar and which ones are different by simply looking at the data or performing manual calculations can be a daunting task. Therefore, to accomplish this task, the unsupervised learning K-means clustering algorithm was used. This is a versatile algorithm that can be used to identify unknown groups in complex data sets.

In essence, I fed the algorithm an unlabeled dataset containing all the Toronto's and Montreal's neighborhoods and their respective features. The features were: the mean of the frequency of each venue category found in each neighborhood, and each neighborhood's normalized population density (pop per Km2), median income of households, and venues density (venues per km2). The elbow method helped identify the value for K, that is into how many groups or clusters our pool of neighborhoods can be divided. The neighborhoods that end up in a given cluster were similar to each other and dissimilar to neighborhoods in another cluster.

3.1 Data Cleaning

First, neighborhoods and their geographical location were identified.

Wikipedia hosts a couple of pages that have a list of the neighborhoods in each city and their respective postal codes.

- Beautiful Soup, a python library for pulling data out of HTML files, was used to scrape the webpages.
- The Toronto neighborhoods information on the Wikipedia page was already organized in the table format we wanted, postal code in one column and neighborhood on another. All was needed to convert the html table to a pandas dataframe and then remove rows were the value on the neighborhood column was 'Not assigned'.
- The Montreal neighborhoods information on the Wikipedia page was not in the format I wanted. The information was on a table but there were no defined column labels, and the data was not organized in any particular way. Each spot on the table had the postal code and its respective neighborhood in the same box. I did the following:
    - Read the html data with pandas, transformed it into an array and flatten all the data using the ravel function. Then, on a new dataframe, all the data was added to single column and parse each individual element and place it in a different column with its respective column label.
    - An important thing to point out, the data had no spaces between words, which also meant no space separating the postal code, Neighborhood name, and position (north, south, east, west).
    - The postal code was 3 characters long at the front, so was parsed using regular string manipulators. Each, the name of the neighborhood and the position started with capital letters, however some multi-word neighborhood names had hyphens in between. To parse through this, a regular expression (REGEX) function was created to add a space in between the concatenated string at the start of a capital letter, except when there was a hyphen preceding the capital letter.
- Geocoder, python library for handling geocoding JSON responses, was used to parse through geocoding results from Bing API geocoding requests, that ultimately allowed us to convert postal codes into geographical coordinates.

Second, obtaining census data

- Census files containing median income, land area and population per neighborhood were downloaded from the statistics Canada website, and uploaded to IBM cloud object storage for easy access from Jupyter notebook. Each of these 3 pieces of data came from separate files and only the rows and columns containing the desired information were extracted.
- Checked for data types and numeric fields were changed from string to float.
- Checked for and dropped rows containing null/NaN values. I started with 221 rows and end up with 212. Each row belonged to a different Forward Sortation Area (FSA). Usually each FSA correspond to a different neighborhood, with the exception of Downsview (M3N, M3M, M3L, M3K) and Don Mills (M3B, M3C) that expand to 3 and 2 different FSAs respectively.

	Postal Code	Borough	Neighborhood	Latitude	Longitude	City	Province	Land Area (km2)	Population	Median total income of households ($)
0	M4V	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park	43.685692	-79.402321	Toronto	ON	2.87	18241.0	85487.0
1	M4N	Central Toronto	Lawrence Park	43.728981	-79.391731	Toronto	ON	6.52	15330.0	137758.0
2	M4T	Central Toronto	Moore Park, Summerhill East	43.690659	-79.383560	Toronto	ON	2.18	10463.0	96571.0
3	M4S	Central Toronto	Davisville	43.703400	-79.385963	Toronto	ON	2.77	26506.0	75520.0
4	M5N	Central Toronto	Roselawn	43.711941	-79.419121	Toronto	ON	2.77	16610.0	94106.0

Figure 1. Sample of the *'metroToMo_data'* dataframe that includes neighborhood information for the city of Toronto and Montreal.

Third, collecting information about the venues in each neighborhood.

- Foursquare API was used to search each neighborhood for all types of venues that fall under different categories.
- Foursquare venue categories are organized in a tree shape, where main categories branch out into a number of sub-categories. For the complete list of Foursquare venue category IDs please visit https://developer.foursquare.com/docs/build-with-foursquare/categories/
- Search criteria included a fixed radius of 800m from the epicenter of each neighborhood, and the below main venue category ID's.

Main Category	Category ID
Arts & Entertainment	4d4b7104d754a06370d81259
College & University	4d4b7105d754a06372d81259
Event	4d4b7105d754a06373d81259
Food	4d4b7105d754a06374d81259
Nightlife Spot	4d4b7105d754a06376d81259
Outdoors & Recreation	4d4b7105d754a06377d81259
Professional & Other Places	4d4b7105d754a06375d81259
Residence	4e67e38e036454776db1fb3a
Shop & Service	4d4b7105d754a06378d81259
Travel & Transport	4d4b7105d754a06379d812

Table 1. Foursquare main venue categories and their respective category ID.

- In some cases, the search radius extended beyond the borders of its own neighborhood, which led for some venues to be found by 2 or more search requests, creating duplicates.

- Dealing with these duplicates was not as straight forward as you may think. The challenge was to know to which neighborhood the duplicate venue belonged to.
- First, the *drop_duplicates* function was used. This function keeps either the 1st or the last of the duplicates encountered when scanning the dataframe in question. The one kept may or may not be the one assigned to the correct neighborhood.
- The Foursquare response data included several pieces of data that helps assign the venue to the correct neighborhood. For example, it included the postal code of the venue, which its first 3 digits represent the neighborhood it belonged to. However, the feature was not found consistently on record for each venue, making it unreliable.
- Among all the features present in the JSON Foursquare response, we found that the latitude and longitude of the venue was the most reliable, it was present for all venues. A function was created to loop through each neighborhood and use its latitude and longitude to reverse geocode its postal code using the geocoder library for Python, powered by the Bing search engine.
- Once we had the first 3 digits of the postal code, which represents the neighborhood it belongs to, we did a left merge with a dataframe that had both Postal Code and Neighborhood name, to bring the name of the Neighborhood column to the dataframe containing all the venues found for each neighborhood.
- The method described above helped assign the correct postal code to 6,807 mislocated venues.

3.2 Exploratory Data Analysis

3.2.1 Demographic Variables Distribution

We have 116 observations (neighborhoods) for the city of Montreal and 96 for the city of Toronto. Montreal's neighborhoods cover a slightly bigger area than Toronto's, with 775.75 km2 versus 663.36 km2 respectively. However, Toronto's population (2,732,094) is larger than Montreal's (2,365,019) by 15% (367,075 people). This results in a higher population density for Toronto (4,118 people per km2) versus Montreal (3,048 people per km2).

The area sizes of Toronto's neighborhoods sit on a slightly tighter range than the area sizes of Montreal's neighborhoods (Figure 1). For Toronto, the neighborhood areas range from 0.30 km2 to 14 km2 with a few outliers extending all the way up to 32.73 km2. For Montreal, the neighborhood areas range from 0.33 km2 to 18.3 km2 with a couple of outliers extending all the way up to 32.13 km2. For Toronto, the interquartile range sits between 3.88 km2 and 8.67 km, while for Montreal's it sits between 2.81 km2 and 9.24 km2.

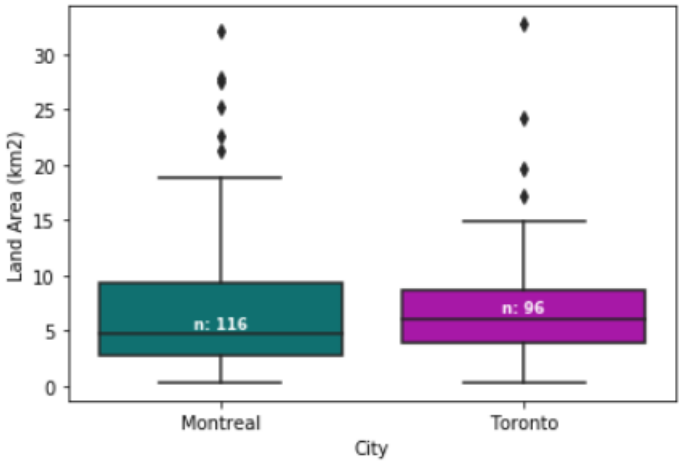


Figure 2. Land area (km2) per neighborhood for the city of Montreal and the city of Toronto.

By looking at figure 3, is visually clear that a great number of Toronto's neighborhoods are more populous than those of Montreal. Approximately, 50% of Toronto's neighborhoods have 25,000 people or more compared to 35% for Montreal's. The top 25% of Toronto's neighborhoods extends all the way to 65,000 people per neighborhood, excluding an outlier of 75,897; while Montreal's reaches a maximum of 48,556 people per neighborhood.

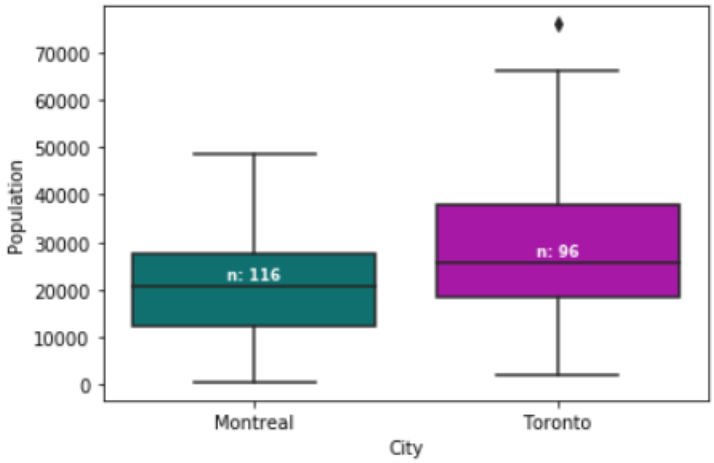


Figure 3. Population distribution for the city of Montreal and the city of Toronto.

Looking at figure 4, we can see that not only the inter quartile range but also the min-max range for the median income of households for neighborhoods in the city of Montreal is tighter than the ones for the city of Toronto's. This means that there is a smaller difference in median income of households among neighborhoods in the city of Toronto's than among neighborhoods in the city of Montreal. Additionally, Montreal is home to some poorer neighborhoods. The minimum of the median income distribution for neighborhoods in the city of Montreal starts in the 24,000s while for the ones in the city of Toronto starts in the 40,000s.

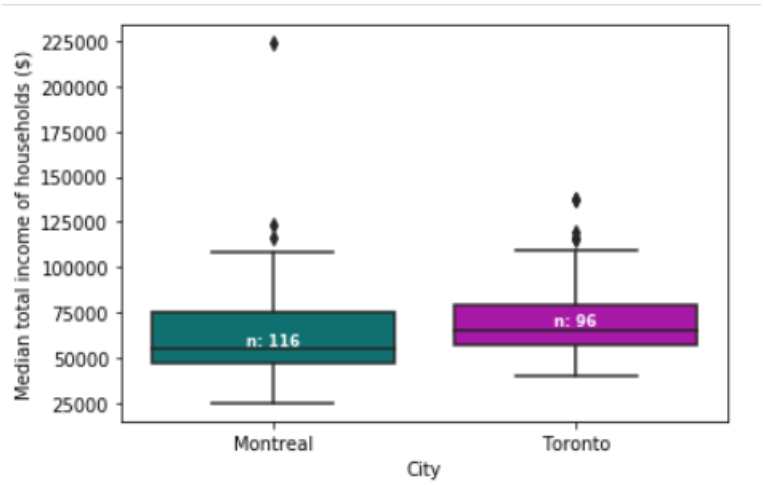
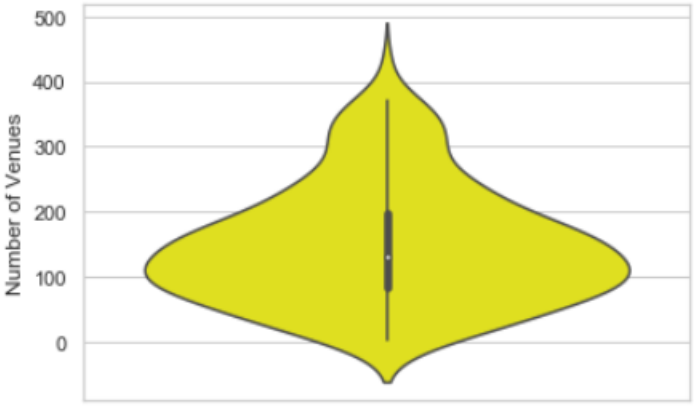


Figure 4. Median income per neighborhood for the city of Montreal and the city of Toronto.

3.2.2 Venues Distribution

A Foursquare venue search was performed on 212 FSA (or neighborhoods). The search returned 31,035 venues across 619 venue categories. Is worth mentioning that the number of venues found are not distributed evenly among neighborhoods. After removing neighborhoods in which less than 5 venues were found, 209 neighborhoods and 31,018 venues remained to be analyzed. As shown in figure 5 below, the median number of venues per neighborhood is 132. The distribution ranges from 5 to 390 venues per neighborhood, with an outlier of 427. The top half exhibits a wider range of number of venues per neighborhood (132-390) compared to the lower half (5-132), meaning that the top half neighborhoods have bigger differences in the number of venues per neighborhood compared to the lower half.



	Number of Venues per Neigh
count	209.000000
mean	148.464115
std	88.460008
min	5.000000
25%	84.000000
50%	132.000000
75%	195.000000
max	427.000000

Figure 5. Number of venues per neighborhood

3.3 Machine Learning – K-means Clustering Analysis

3.2.3 Feature transformation

Before we can feed our data to the algorithm, we need to transform and normalize our features.

3.2.3.1 One hot encoding

K means clustering algorithms cannot work directly with categorical data. Therefore, we needed to encode the *venue categories* to a numerical form. This was done through a procedure called One Hot Encoding, which requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector, that is all zero values except for the index of the integer, which is marked with a 1.

We invoked '*pd.get_dummies'* function and included as a parameter the column holding the venue category values. The output after running this function looked like the dataframe below.

Dataframe shape: (31030, 621)

	Neighbourhood	Postal Code	ATM	Accessories Store	Acupuncturist	Adult Boutique	Adult Education Center	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Gate	Airport Lounge	Airport Terminal	Alternative Healer	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Apres Ski Bar
0	Saint-Laurent Central	H4R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 6. Partial snapshot of the one hot encoded dataframe.

Each row represents one of the 31,018 venues found. The first column holds the name of the neighborhood in which the venue was found. Each of the remainder columns represent one of the 612 unique venue categories. Then, for a particular row, there is a value of 1 under the category name of the venue found, and a value of zero under the rest.

3.2.3.2 Normalization

Now that the categorical values are mapped into integer values, we consolidated the data in the one hot encoded dataframe to show the mean of the frequency of each venue category on a single row per neighborhood. See figure 7.

Additionally, to bring the census data variables to the same scale (0 to1) of the venue category variables, we used the Min Max Scaler from the Scikit Learn python library. Finally, our census variables ended up with smaller standard deviations, which can suppress the effect of outliers. See figure 7.

Shape of the dataframe: (209, 624)

	Postal Code	Neighbourhood	Land Area (km2)	Population	Median total income of households ($)	ATM	Accessories Store	Acupuncturist	Adult Boutique	Adult Education Center	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Gate	Airport Lounge	Airport Terminal	Alternative Healer	American Restaurant	Amphitheater
0	H1A	Pointe-aux-Trembles	0.399322	0.424174	0.189783	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	H1B	Montreal East	0.358002	0.260164	0.143670	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	H1C	Rivière-des-Prairies Northeast	0.533148	0.187398	0.312173	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	H1E	Rivière-des-Prairies Southwest	0.360469	0.555637	0.181758	0.012658	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	H1G	Montréal-Nord North	0.234659	0.637084	0.090255	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.007576	0.0	0.0

Figure 7. Snapshot of the features data after having consolidated and normalized its features.

3.2.3.3 Optimization

To help us select the optimal number of clusters to be set when executing the K-means clustering algorithm, we used the Elbow Method. This method fitted the K-means clustering model with a range of values for K. When the line chart resembles an arm, the “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. In this case we determined that 4 is an appropriate value for K.

4. Results

As mentioned earlier, the neighborhoods were divided into 4 groups (given a K of 4). The clusters are color coded and you can visualize their resulting cluster distribution plotted on the maps below (figure 9) (figure 10).

For reference:
Cluster 1 = **Green**
Cluster 2 = **Blue**
Cluster 3 = **Red**
Cluster 4 = **Purple**

Metropolitan Toronto

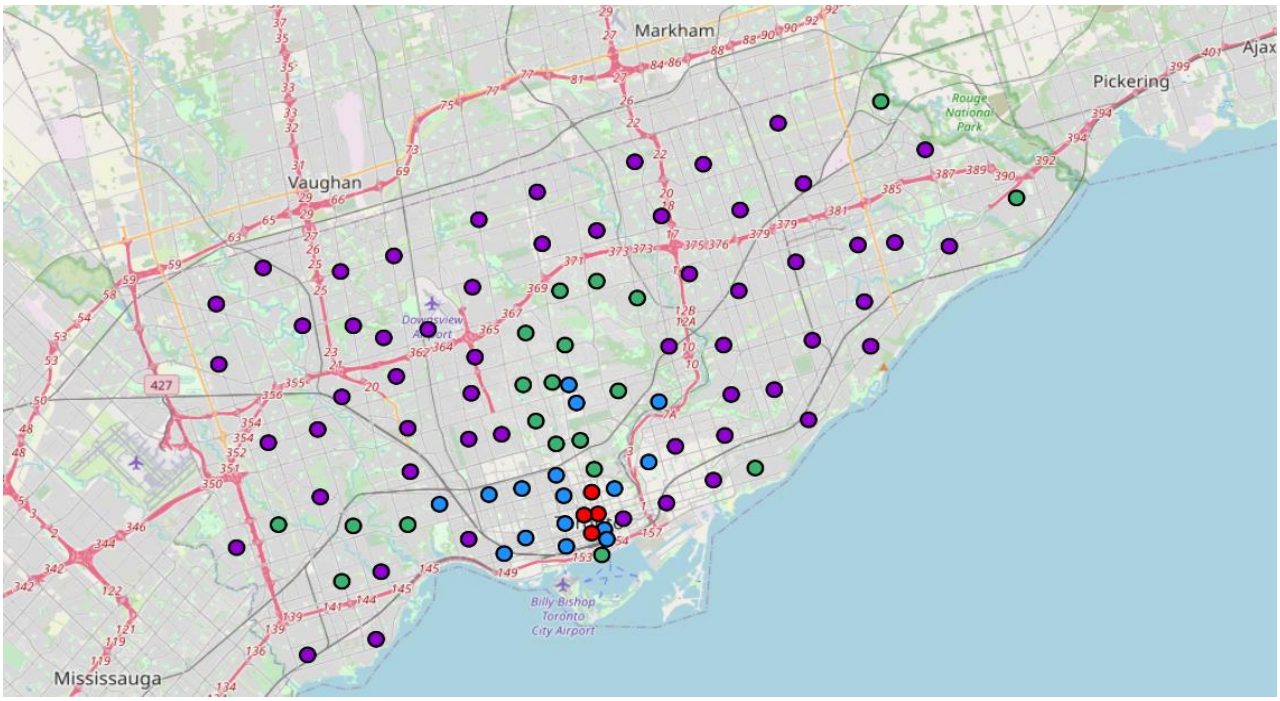


Figure 9. Neighborhood clusters superimposed on top of the city of Toronto map.

Metropolitan Montreal

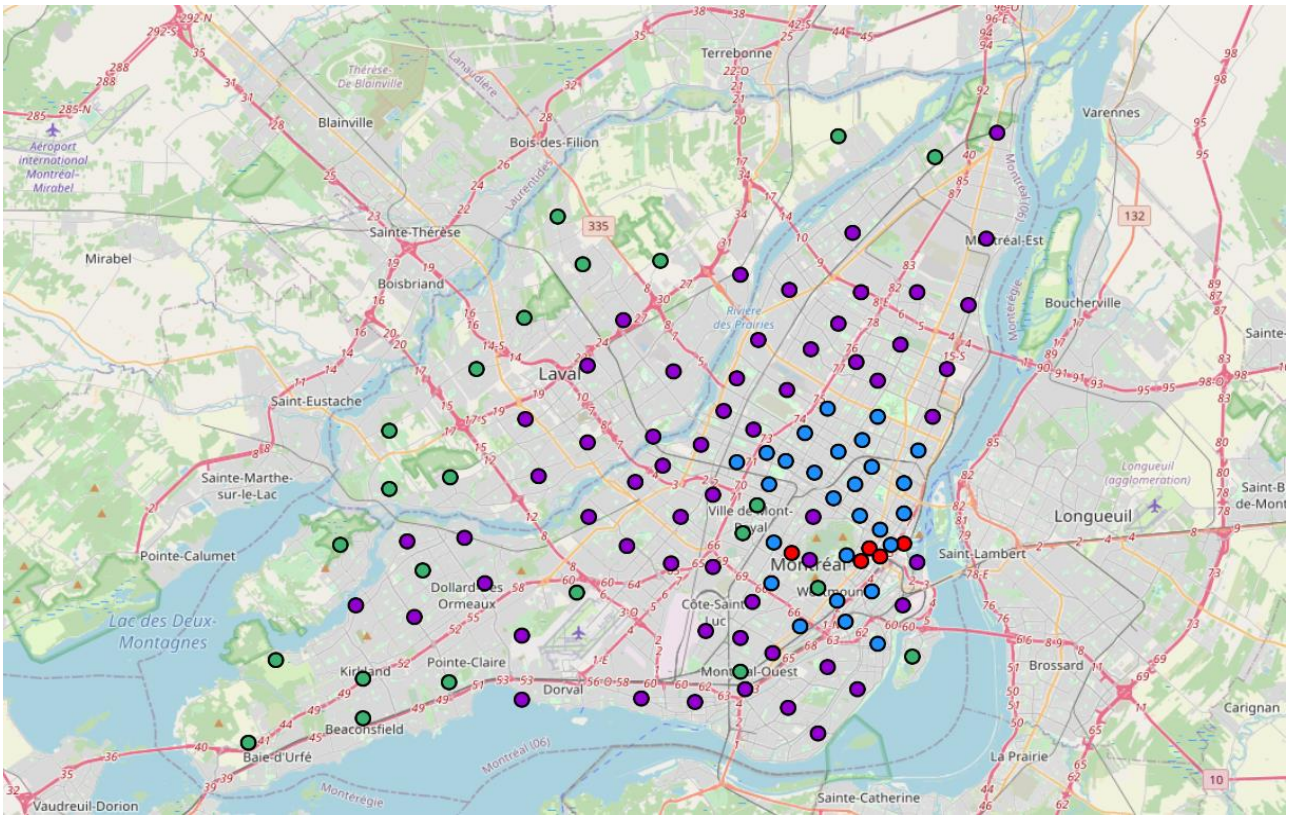


Figure 10. Neighborhood clusters superimposed on top of the city of Montreal map.

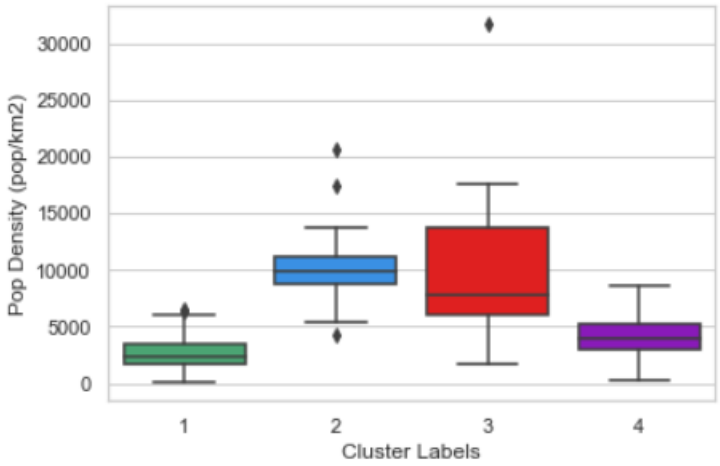


Figure 11. Comparing the distribution of *population density* data among the four clusters.

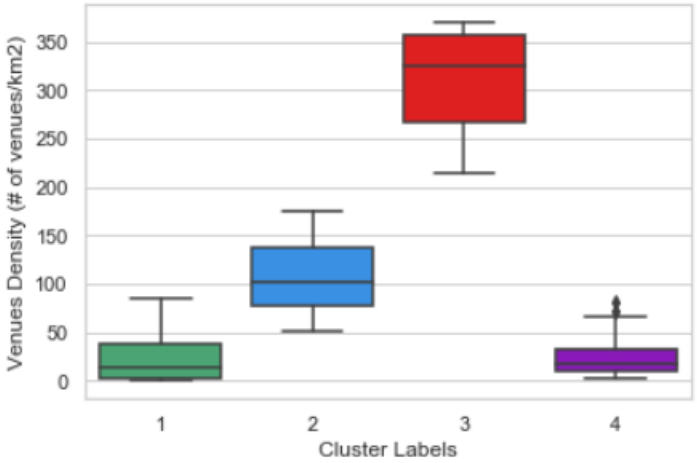


Figure 12. Comparing the distribution of *venues density* data among the four clusters.

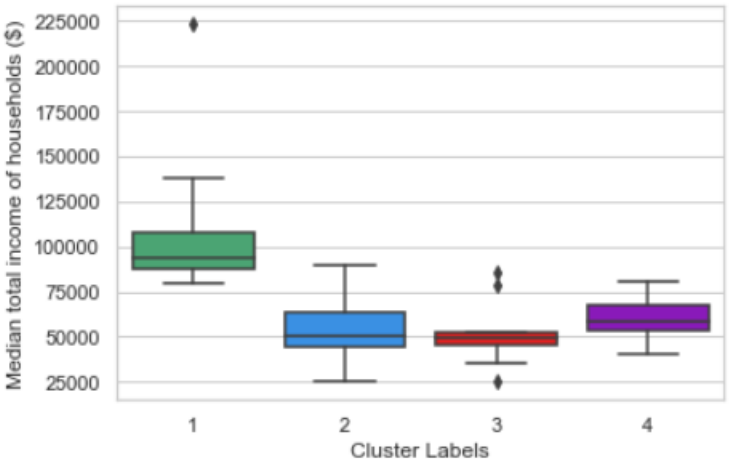


Figure 13. Comparing the distribution of *Median Total Income of Households* data among the four clusters.

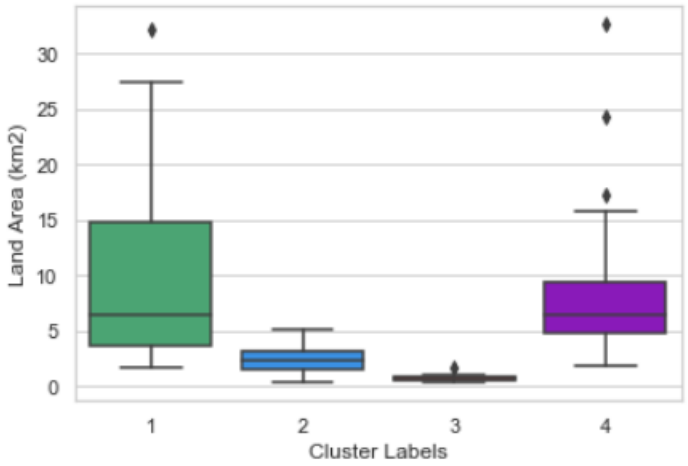


Figure 14. Comparing the distribution of *Land area (km2)* data among the four clusters.

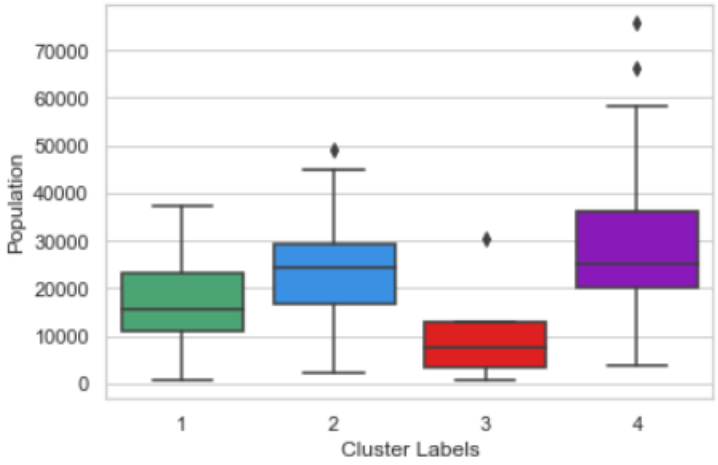


Figure 15. Comparing the distribution of the *population* data among the four clusters.

Cluster 1

Cluster 1 is color coded "**green**". Neighborhoods in this cluster can be visualized in figure 9 and 10. Here we have 43 neighborhoods, 20 from Toronto and 23 from Montreal. Its top venue categories (see figure 15) are offices, parks, residential Buildings (apartment and condominiums), bakeries, Italian restaurants, bus lines, gyms and playgrounds, and more.

In terms of population density, neighborhoods in cluster 1 ranks 1st (from least to most dense) compared to the other 3 clusters (see figure 11 and 14). The median number of people per square kilometer is 2,381. Neighborhoods in this cluster rank 2nd in venues density, having a median of 14.6 venues per square kilometer and a maximum of 84. Additionally, these neighborhoods rank 1st (from highest lowest) in terms of median total income of households ($). The median figure is $94,000 with a minimum of $79,000, a maximum of $137,000 and outliers up to $23,000.

	Median total income of households ($)	Pop Density (pop/km2)	Venues Density (# of venues/km2)
count	43.000000	43.000000	43.000000
mean	100572.627907	2789.596744	21.417886
std	24106.093101	1686.539180	21.624835
min	79117.000000	26.370000	0.155618
25%	87762.500000	1677.615000	3.188865
50%	94106.000000	2381.420000	14.579439
75%	107840.000000	3490.070000	39.313116
max	223744.000000	6454.640000	84.403670

Figure 14. Cluster 1 numerical features distribution.

Figure 15. Word cloud chart based on the venue categories frequencies of the neighborhoods in cluster 1.

Cluster 2

Cluster 2 is color coded "blue". Neighborhoods in this cluster can be visualized in figure 9 and 10. Here we have 43 neighborhoods, 16 from Toronto and 27 from Montreal. Its top venue categories (see figure 17) are entertainment venues, bus lines, light rail and metro stations, residential Buildings (apartment and condominiums), bars, cafés and restaurants, art galleries, and more.

In terms of population density, neighborhoods in cluster 2 are the most dense ranking 4th (from least to most dense) compared to the other 3 clusters (see figure 11 and 14). The median number of people per square kilometer is 9,883 (see figure 16). That is 4 times the population density of cluster 1, which ranked 1st.

Neighborhoods in this cluster rank 3rd in venues density (from least to most dense). These have a median value of 101.5 venues per square kilometer and range from 51 to 175. That value is almost 7 times larger than the one from cluster 1, which ranked 1st (see figure 11).

Additionally, when it comes to median total income of households ($), neighborhoods in this cluster rank 3rd (from highest to lowest). The median figure is $50,500 and the distribution ranges from $25,500 to $90,000. This is a wide distribution, considering that within it fits the median income range for clusters 3 (red) and 4 (purple). Cluster 2 (blue) ranked 3rd and not 2nd (cluster 4 purple) because the percentage of neighborhoods doing worse than those of cluster number 4 is greater than the percentage of neighborhoods doing better than those of cluster 4.

	Median total income of households ($)	Pop Density (pop/km2)	Venues Density (# of venues/km2)
count	43.000000	43.000000	43.000000
mean	53730.906977	10008.642558	106.342645
std	14024.169102	2922.274296	35.591591
min	25517.000000	4205.450000	51.369863
25%	44310.000000	8722.270000	76.904275
50%	50556.000000	9883.700000	101.533742
75%	63272.000000	11164.600000	137.164074
max	89969.000000	20615.840000	175.141243

Figure 16. Cluster 2 numerical features distribution.

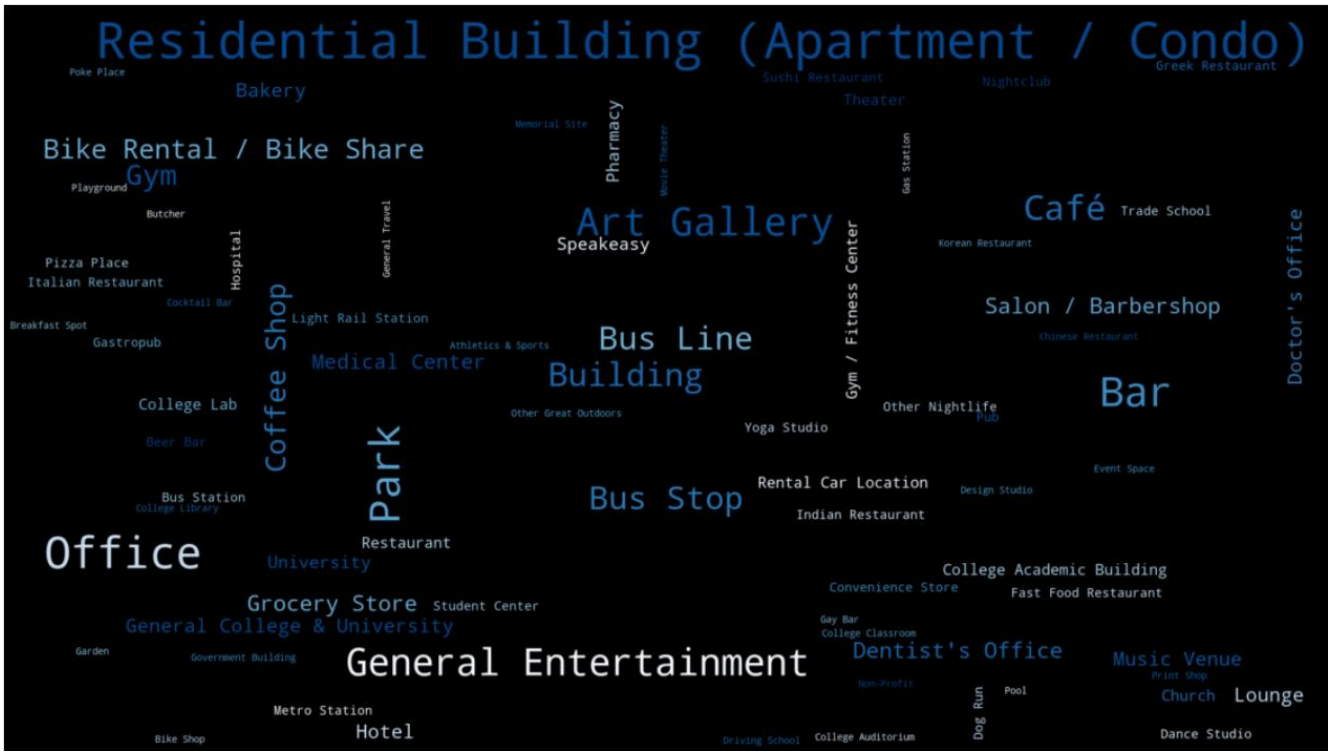


Figure 17. Word cloud chart based on the venue categories frequencies of the neighborhoods in cluster 2.

Cluster 3

Cluster 3 is color coded "red". Neighborhoods in this cluster can be visualized in figure 9 and 10. Here we have only 9 neighborhoods, 4 from Toronto and 5 from Montreal. Its top venue categories (see figure 19) are entertainment venues, residential Buildings (apartment and condominiums), offices, pubs, cafés and coffee shops, university and college buildings, shopping malls, historic sites and more.

In terms of population density, neighborhoods in cluster 3 rank 3rd (from least to most dense) compared to the other 3 clusters (see figure 11 and 14). The median number of people per square kilometer is 7,815 (see figure 18) and it ranges from 1,760 to 17,500 with an outlier in the 30,000s. Its wide range is nearly 2 times that of the other 3 neighborhoods.

The venue density of these neighborhoods is unparallel compared to the other 3 clusters (see figure 12). Unlike in the other metrics where a given cluster show an overlap in relation to any of the other 3, here, the entire range is outside of the others. Neighborhoods in this cluster rank 4th in venues density (from least to most dense). These have a median value of 326 venues per square kilometer and range from 215 to 370. For comparison, the median of the second most dense cluster is 100 venues per squared kilometer and ranges from 50 to 175.

Additionally, when it comes to median total income of households ($), neighborhoods in this cluster rank 4th (from highest to lowest). The median figure is $49,700. Its narrow distribution ranges from $24,800 to

$51,000 with 2 outliers in the high 70,000s. Even though cluster 2 (blue) and 4 (purple) have similar median values than cluster number 3 (red), the entire distribution of cluster 3 (red) fits below the median values of the former and the latter, explaining its ranking.

	Median total income of households ($)	Pop Density (pop/km2)	Venues Density (# of venues/km2)
count	9.000000	9.000000	9.000000
mean	52083.888889	10971.454444	302.846206
std	19359.384174	9182.059047	61.466528
min	24864.000000	1760.610000	214.634146
25%	44949.000000	6067.800000	267.213115
50%	49696.000000	7815.240000	325.581395
75%	52167.000000	13808.200000	357.575758
max	86101.000000	31741.670000	370.666667

Figure 18. Cluster 3 numerical features distribution.

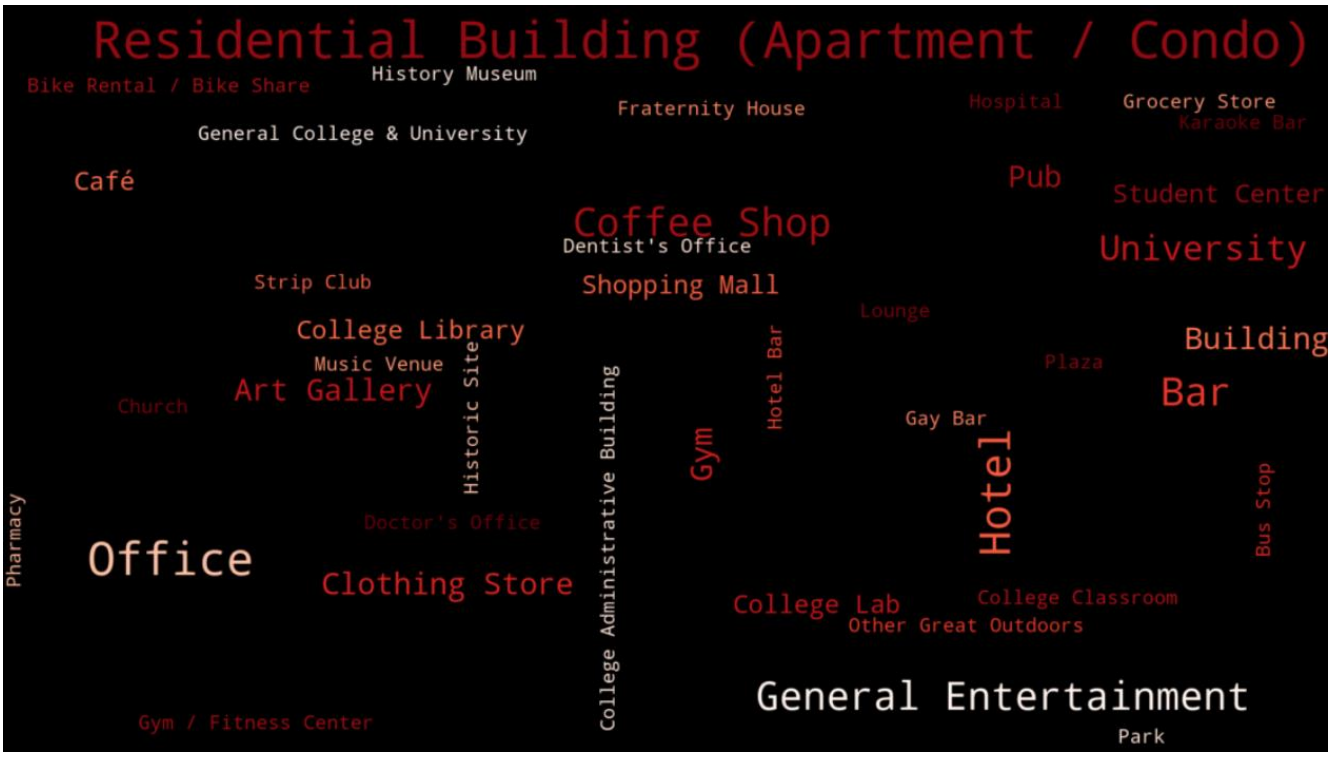


Figure 19. Word cloud chart based on the venue categories frequencies of the neighborhoods in cluster 3.

Cluster 4

Cluster 4 is color coded "purple". Neighborhoods in this cluster can be visualized in figure 9 and 10. This is the biggest cluster featuring 114 neighborhoods, 56 from Toronto and 58 from Montreal. Its top venue categories (see figure 21) are residential buildings, dentists, pharmacies, doctor's offices and other healthcare venues, educational centers, assorted ethnic restaurants, convenience stores, and more.

In terms of population density, neighborhoods in cluster 4 rank 2nd (from least to most dense) to those of the other 3 clusters (see figure 11 and 14). Its values are a slight notch up to those of cluster 1. The median number of people per square kilometer is 3,995 (see figure 20) and it ranges from 275 to 8,552.

Neighborhoods in this cluster rank 1st in venues density (from least to most dense). Cluster 4 have a median value of 17 venues per square kilometer and range from 2 to 75. For comparison.

Additionally, when it comes to median total income of households ($), neighborhoods in this cluster rank 2nd (from highest to lowest). The cluster median figure is $58,980 and its distribution ranges from $40,730 to $81,014. Cluster 2 (blue) ranked 3rd and not 2nd (cluster 4 purple) because the percentage of neighborhoods doing worse than those of cluster number 4 is greater than the percentage of neighborhoods doing better than those of cluster 4.

	Median total income of households ($)	Pop Density (pop/km2)	Venues Density (# of venues/km2)
count	114.000000	114.000000	114.000000
mean	59564.719298	4190.326754	22.763636
std	9759.744277	1680.098752	16.535248
min	40730.000000	275.440000	2.175439
25%	53221.750000	2975.092500	10.764253
50%	58980.500000	3995.440000	17.202105
75%	67409.250000	5251.742500	33.131467
max	81014.000000	8552.190000	81.521739

Figure 20. Cluster 4 numerical features distribution.

Figure 21. Word cloud chart based on the venue categories frequencies of the neighborhoods in cluster 4.

	# of Neigh	Population Density	Venues Density	Median Household Income	Type and Location	Popular Venue Categories
Cluster 1	Total: 43 M.O.: 23 T.O.: 20	Low	Low	High	Residential and commercial areas. Mid-town	parks, residential buildings, offices, bakeries and Italian restaurants, bus lines, gyms and playgrounds, and more.
Cluster 2	Total: 43 M.O.: 27 T.O.: 16	Med-to-High	Med	Med-to-High	Residential and commercial areas. Downtown	entertainment venues, bus lines, light rail and metro stations, bars, cafés and restaurants, art galleries, apartments and condominiums, and more
Cluster 3	Total: 9 M.O.: 5 T.O.: 4	Low-to-high	High	Low-to-Med	Mostly commercial areas. Downtown	entertainment venues, apartment and condominiums, offices, pubs, cafés and coffee shops, university and college buildings, shopping malls, historic sites and more.
Cluster 4	Total: 114 M.O.: 58 T.O.: 56	Low-to-Med	Low	Med	Residential and some industrial areas. City Suburbs.	residential buildings, dentists, pharmacies, doctor's offices and other healthcare venues, educational centers, assorted ethnic restaurants, grocery stores, factories, and more

Table 2. Neighborhood clusters results summary.

5. Discussion

To best interpret the results, one must be familiar with at least one of the cities. This way you can uncover subtleties that may be present in one city. Then, results can be extrapolated or investigated to see if these are also true on the target city. For example, because I'm familiar with the city of Toronto, I can recognize how clusters 2 (blue) and 1 (green) are distributed across the city resembling the shape of the subway system. I can also notice how, in the same way that a street called Bloor does, some neighborhoods under cluster 2 (blue) form a straight line dividing downtown-Toronto from mid-town-Toronto. And, how neighborhoods in cluster number 3 (red) represent busy areas in downtown core, facing waterfront. All the above phenomena apply as well to neighborhood clusters in the city of Montreal.

However, even if you are not familiar with any of the cities in the study, the report on its own reveal a great deal of information. For instance, it seems that in cluster 1 (green) we are dealing with some distinguished neighborhoods. For starters, the median value for the median income of households for neighborhoods in this cluster are 59% higher than that of the next highest cluster. These neighborhoods are not too crowded in terms of both people (low population density), and commercial buildings (low venues density), are some of the biggest in terms of land area (km2), sit in close proximity to downtown, are big on parks and green areas. Additionally, they have a great assortment of commercial venues nearby such as restaurants, bakeries, gyms and playgrounds, general entertainment, and more. Here, pieces of real state should be quite expensive.

Neighborhoods in cluster 3 (red) are densely packed, mostly commercial areas downtown. Population is very low (see figure 15), but land areas (km2) are even smaller; That's what drives the relatively high population density in the area. But what is really filling up this neighborhoods are the high number of commercial venues, that paired with the floored land areas (km2), in comparison with the rest of the clusters, catapults the venues density (# of venues/km2) numbers to levels unattainable by the rest of the neighborhoods. In the word cloud on figure 19, you can see that most words are about the same font size, meaning that there is an even or well-balanced variety of venues, with no one venue category frequency sticking out.

Going outwards on any direction from neighborhoods in cluster 3 (red) (mostly commercial downtown areas), you will start seeing more residential areas, which indicates you have arrived at areas of cluster 2 (blue). These areas are still considered to be downtown; they are just the outer layer of downtown's core.

By looking at the higher population densities and the lower venue densities of neighborhoods in cluster 2 (blue) compared to the ones of cluster 3 (red), It seems like what is different between the two is that cluster 2 (blue) neighborhoods have traded commercial venues in favor for population. Looking deeper into cluster 2's numbers reveal that it is true that the population is higher, in fact its median is 3 times the one of cluster 3 (red) (see figure 15); although, the number of venues per neighborhood between the two is not much different. What is helping maintain their lower venues density figures are the slightly bigger

land areas of cluster 2 (blue). That extra land space is likely being used to accommodate the significantly higher number of residents with residential real estate in the form of condominiums, apartments, houses.

Lastly, by looking at the summary table above (table 2), you can tell how neighborhoods under cluster 4 (purple) represent mostly residential areas and a few industrial areas outside the city's core. We do see factories among the most frequent venue categories on the word cloud on figure 21, but for the most part we see a higher concentration of venues that cater to family households such as dentists, pharmacies, doctor's offices and other healthcare venues, educational centers, assorted ethnic restaurants, grocery stores, and more. The numerical features also corroborate what is expected out of suburban areas of the city, low to medium population density, low venues density and medium median household income.

6. Conclusion

In this project I used the k-means clustering algorithm. This method helped me identify and group into clusters neighborhoods, from the city of Toronto and the City of Montreal, that share similar venues composition and certain census attributes such as population density, and median income of households. I found that grouping neighborhoods into 4 clusters explains most of the variation observed in these two cities. Based on my findings, I categorized my clusters into cluster 1 (green) residential and commercial areas mid-town, cluster 2 (blue) residential and commercial areas downtown, cluster 3 (red) commercial area downtown, cluster 4 (purple) residential and some industrial areas city suburbs.

Finally, there are many ways in which I could have approach this problem. But my results show the power of thinking outside of the box with unsupervised methods like *k*-means.

7. Future Directions

This study is meant to be carried out on the initial phases when considering a new location for your business. At a subsequent stage, a more business-specific analysis will have to be done. This report, however, will help make better-informed decisions in the final stages of your search. For example, you want to expand your chain of vegan restaurants. After unfolding the more business-specific (vegan restaurants) analysis, you have located several pocket areas of low restaurant density (also areas of low vegan-restaurant density) across the target city (e.g. Montreal) that are at an adequate distance from the city center. With the results of my study on hand, stakeholders can now zero-in on that optimal location for your vegan restaurant. They can use this information to evaluate each of the tentative locations based on the attractiveness of the neighborhoods these are in; proximity to parks, water, or major roads; presence of venues that trigger the use of vegan restaurants such as a variety of entertainment venues, shopping malls, gyms, and more; real estate availability; social and economic dynamics of every neighborhood, which may include population density, median total income of households, and more.