Applied Data Science Capstone

Similar Neighborhoods in Downtown Toronto and Ville Marie

Walter Dietrich

February 9, 2020

Table of Contents

1 Introduction

The government of Montreal would like to increase the city's commercial growth rate. Montreal is the second-most populous municipality in Canada. Montreal is about 300 miles from Toronto, and it is connected to Toronto by the St. Lawrence River and Lake Ontario. Route 401 also connects the two cities. Montreal was the commercial capital of Canada until it was surpassed by Toronto in the 1970's.

Montreal is led by a mayor, who is "first among equals" in the Montreal city council. The city council is democratically elected. It has representative from all of Montreal's boroughs. (Montreal , like New York and Toronto, is subdivided into Boroughs.) Much of the council's power is centralized in the Executive Committee.

In order to increase Montreal's growth rate, the executive committee and the council would like to use data science in order to find businesses that could be enticed to open offices in Montreal. If a neighborhood (as designated by a postal code) in Toronto is similar to a neighborhood (as designated by a postal code) in Montreal, then businesses in the Toronto neighborhood are good candidates for enticements to open up offices in Montreal. Similarly, if a neighborhood in Toronto does not have a similar neighborhood in Montreal, businesses in that neighborhood are not good candidates for inducements by Montreal.

The city council would like to start with a pilot project to assess the viability of a project that groups neighborhoods this way. The pilot project would apply the process to the neighborhoods in one borough from each city.

In Montreal, the borough with the most neighborhoods is "Ville Marie", which includes the Downtown neighborhood and several other neighborhoods. In Toronto, the "Downtown" borough includes the downtown Toronto, so the Toronto Downtown borough is a good match for the Montreal Ville Marie borough.

By finding out which neighborhoods in Ville Marie are similar to neighborhoods in Downtown Toronto, the Montreal government can target potential business to open locations in Montreal by appealing to business' preference what they are accustomed to. If they are accustomed to a location in Toronto that has a lot of gyms and cafes, then they will probably be more comfortable in a location in Montreal that contains a lot of gyms and cafes. In addition, if a business has a successful location in Toronto neighborhood, then they are more likely to succeed if they open a location in a similar neighborhood in Montreal. On the other hand, if the business wants to open an office in a new neighborhood that is different, the Montreal government will have evidence for recommending different neighborhoods.

Note: In the first Wikipedia article I refer to below, the borough of Ville Marie is written as “Ville Marie” and as “Ville-Marie”, with the two-word version occurring slightly more often than the hyphenated version. In Google maps, it is frequently written to as Ville-Marie. The original name of Montreal was Ville-Marie (City of Mary). In modern times, “Ville Marie” and “Ville-Marie” are interchangeable.

1.1 Literature Review

Data for this report was drawn from several sources. Since this is not a formal literature review, I am not using standard academic bibliography conventions.

1.1.1 Text Descriptions of Montreal, Ville Marie, and Toronto

https://en.wikipedia.org/wiki/Montreal

The above page describes Montreal. It includes descriptions of the boroughs. It contains a list of the neighborhoods in the borough of Ville Marie.

https://en.wikipedia.org/wiki/Ville-Marie,_Montreal

The above page contains more information about Ville Marie. It also contains a list of the neighborhoods in the borough of Ville Marie.

http://www.vieux.montreal.qc.ca/histoire/eng/v_mara.htm

The above page has a brief history of Montreal, from its beginnings as Ville Marie.

https://en.wikipedia.org/wiki/Red-Light_District,_Montreal

The above page contains a description of the Red-Light District in Montreal. I’ve included it because it is in maps of Ville Marie, but it is just a small area. It isn’t a neighborhood.

https://en.wikipedia.org/wiki/Toronto

The above page contains information about Toronto.

1.1.2 Postal Code Information

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

The above page contains all of the postal codes that are used in Montreal. It also contains some descriptive information about the neighborhoods. Some of it is in French, which is the predominant language of Montreal. (Montreal is the second-largest French-speaking city in the world, right behind Paris.) The neighborhood information is complimentary to information in the Wikipedia Montreal article.

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

The above page contains information about Toronto postal codes, including neighborhood names.

http://www.strategiclists.com/wp/wp-content/uploads/2015/07/Canada-FSAS.pdf

The above document contains maps of the areas of all of the three-characters postal code prefixes in Canada. Page 43 contains postal codes prefixes in Montreal. This could be used as an alternate source of postal code data.

https://cocl.us/Geospatial_data

The link above refers to a CSV file that contains 3 columns. The first column contains 3-character postal code prefixes that start with M. The second and third columns contain decimal numbers. The second column contains the latitude of the postal code prefix in the first column. The third column contains the longitude of the postal code prefix in the first column.

1.1.3 Maps Related to Ville Marie and Maps Related to Postal Codes that Contain Land in Ville Marie

https://www.google.com/maps/place/Ville-Marie,+Montreal,+QC,+Canada/@45.5127587,-73.5961059,13z/data=!3m1!4b1!4m5!3m4!1s0x4cc91a4d31166b3d:0xe16252d7fe06209e!8m2!3d45.5087937!4d-73.5553019

The above page contains a map that shows the boundaries of Ville Marie.

https://www.google.com/maps/place/Montreal,+QC+H2L,+Canada/@45.5196257,-73.5705852,15z/data=!3m1!4b1!4m5!3m4!1s0x4cc91bb156926d11:0xcc9a6ca5eaf5dd80!8m2!3d45.522199!4d-73.5641471

The above map shows the area contained in the H2L postal code prefix. It shows that Gay Village is in H2L. It also shows that part of the Le Plateau-Mont-Royal borough is in H2L. Since the prior analysis of Toronto was based on postal codes, I am using postal codes for this analysis even though some postal codes for Ville Marie also include neighborhoods from neighboring boroughs. One reason for doing this because Ville Marie contains Downtown Montreal, and I want to compare Downtown Montreal with Downtown Toronto.

https://www.google.com/maps/place/Montreal,+QC+H2X,+Canada/@45.5124287,-73.5844915,14z/data=!3m1!4b1!4m5!3m4!1s0x4cc91a4b85d8bcad:0xe808bc984c0f9884!8m2!3d45.5132931!4d-73.5694014

The above map shows the area contained in the H2X postal code prefix. It shows that the Latin Quarter and Quartier Des Spectacles are in this area.

https://www.google.com/maps/place/Montreal,+QC+H2Y,+Canada/@45.5058532,-73.5643487,15z/data=!3m1!4b1!4m5!3m4!1s0x4cc91a57232611a5:0x1bb3ff57ab036bb4!8m2!3d45.5052277!4d-73.5557318

The above map shows the area contained in the H2Y postal code prefix. Visual inspection shows that this area is entirely contained in Ville Marie.

https://www.google.com/maps/place/Montreal,+QC+H2Z,+Canada/@45.5053444,-73.5664642,16z/data=!3m1!4b1!4m5!3m4!1s0x4cc91a5073be9d71:0x78f81b587eb5ccfe!8m2!3d45.5039113!4d-73.56321

The above map shows the area contained in the H2Z postal code prefix. Visual inspection shows that this area is entirely contained in Ville Marie.

https://www.google.com/maps/place/Montreal,+QC+H3A,+Canada/@45.5057312,-73.5851572,15z/data=!3m1!4b1!4m5!3m4!1s0x4cc91a478d7d0b9d:0x1926a7f8491bef0!8m2!3d45.5035028!4d-73.5768503

The above map shows the area contained in the H3A postal code prefix. Careful visual inspection shows that it is contained in Ville Marie.

https://www.google.com/maps/place/Montreal,+QC+H3B,+Canada/@45.5011524,-73.5776599,15z/data=!4m5!3m4!1s0x4cc91a449d667617:0x43dba850448c97f1!8m2!3d45.4999144!4d-73.568918

The above map shows the area contained in the H3B postal code prefix. It contains a small area clearly inside of Ville Marie.

https://www.google.com/maps/place/Montreal,+QC+H3C,+Canada/@45.5023151,-73.5815692,13z/data=!3m1!4b1!4m5!3m4!1s0x4cc91af750b653c5:0x22f1d6a9e7709636!8m2!3d45.4927605!4d-73.5614161

The above map shows the area contained in the H3C postal code prefix. It contains the southeast portion of Ville Marie as well as Saint Helen's Island and Notre-Dame Island. Most of these areas are part of Ville Marie. Although this area extends beyond the boundaries of Ville Marie, most of the area in this postal code prefix is in Ville Marie.

https://www.google.com/maps/place/Montreal,+QC+H3G,+Canada/@45.4992627,-73.5897737,15z/data=!3m1!4b1!4m5!3m4!1s0x4cc91a400aac7669:0xdfdfa1a5fddbe76f!8m2!3d45.499505!4d-73.582556

The above map shows the area contained in the H3G postal code prefix. It is contained within the Ville Marie boundaries.

https://www.google.com/maps/place/Montreal,+QC+H3H,+Canada/@45.5007627,-73.6110584,14z/data=!3m1!4b1!4m5!3m4!1s0x4cc91a3d4cb16127:0xae3b8217ecd04d10!8m2!3d45.5026595!4d-73.5957654

The above map shows the area contained in the H3H postal code prefix. Except for part of the Notre Dame des Neiges Cemetery, all of this area is in Ville Marie.

https://www.cimetierenotredamedesneiges.ca/en/burial-location

The above contains a map of the Notre Dame des Neiges Cemetery.

https://www.google.com/maps/place/Cit%C3%A9+du+Havre,+Montreal,+QC+H3C+3R4,+Canada/@45.4922451,-73.5611646,14z/data=!3m1!4b1!4m5!3m4!1s0x4cc91a92b6a7ca67:0x98893b8c6ef71512!8m2!3d45.492247!4d-73.543655

The above map shows the location of Cité du Havre, and also shows that the postal code of Cité du Havre starts with H3C.

https://www.google.com/maps/place/Complexe+Le+Gleneagles/@45.4951439,-73.5965341,17z/data=!3m1!4b1!4m5!3m4!1s0x4cc91a10cd7ec55d:0x6d7594c72c8f65c5!8m2!3d45.4951439!4d-73.5943401

The above map shows the location of Îlot-Trafalgar-Gleneagles, and shows that the postal code of this location is H3H.

https://www.google.com/maps/place/Montreal,+QC+H2K,+Canada/@45.5298121,-73.5652603,15z/data=!3m1!4b1!4m5!3m4!1s0x4cc91bbe9e7b629b:0xc8cba6ffa90d0989!8m2!3d45.5301959!4d-73.5527213

The above map shows the area contained in the postal codes that starts with H3H. It shows that Sainte-Marie is in H3H.

1.1.4 Algorithms

https://en.wikipedia.org/wiki/K-means_clustering

The above describes k-means clustering

https://scikit-learn.org/stable/modules/clustering.html#k-means

The above has documentation for the k-means implementation I am using.

https://en.wikipedia.org/wiki/Curse_of_dimensionality

The above describes the general problems of using data in high-dimensional spaces. It also mentions the problem of high dimensionality in clustering. The Scikit-learn documentation also mentions this, and provides a potential solution.

http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PTXT&s1=6,282,318.PN.&OS=PN/6,282,318&RS=PN/6,282,318

The above contains a patent that combines pattern matching with optimization. It could be used to find the optimal matching of neighborhoods, if one wanted a one-to-one matching between neighborhoods in Downtown Toronto and Ville Marie.

1.2 Acknowledgements

I wish to thank the authors of the Coursera Applied Data Science courses, the authors of the Wikipedia pages I used, and the Coursera students who reviewed my submissions.

2 Data

This project depends on neighborhood names and the venues that are in the neighborhoods. To find the venues that are in neighborhoods, I need the coordinates of the neighborhoods. To find the coordinates of the neighborhoods, I use 3-character prefixes of Canadian Postal Codes. Each 3-character postal code prefix contains 0 or more neighborhoods. If a postal code contains no neighborhoods, it is not interesting. In the Descriptive Names section, I talk about acquiring the information about postal codes and neighborhoods. In the Coordinates section, I talk about getting the coordinates of the postal codes. In the Venues section, I talk about getting the venue information using the coordinates.

2.1 Descriptive Names

The names of the neighborhoods in each city are acquired from the data sources described in the Literature review. This subchapter describes the data acquisition in more detail.

2.1.1 Toronto

The Wikipedia page that has postal codes that start with M contains the postal codes and neighborhood names of neighborhoods in Toronto. I scrape the web page to create a dataframe that contains postal codes and the neighborhoods that are in the postal codes. A partial listing of the table is in
https://github.com/WallyNY/Coursera_Capstone/blob/master/Capstone_week_3_part_1.ipynb

2.1.2 Montreal

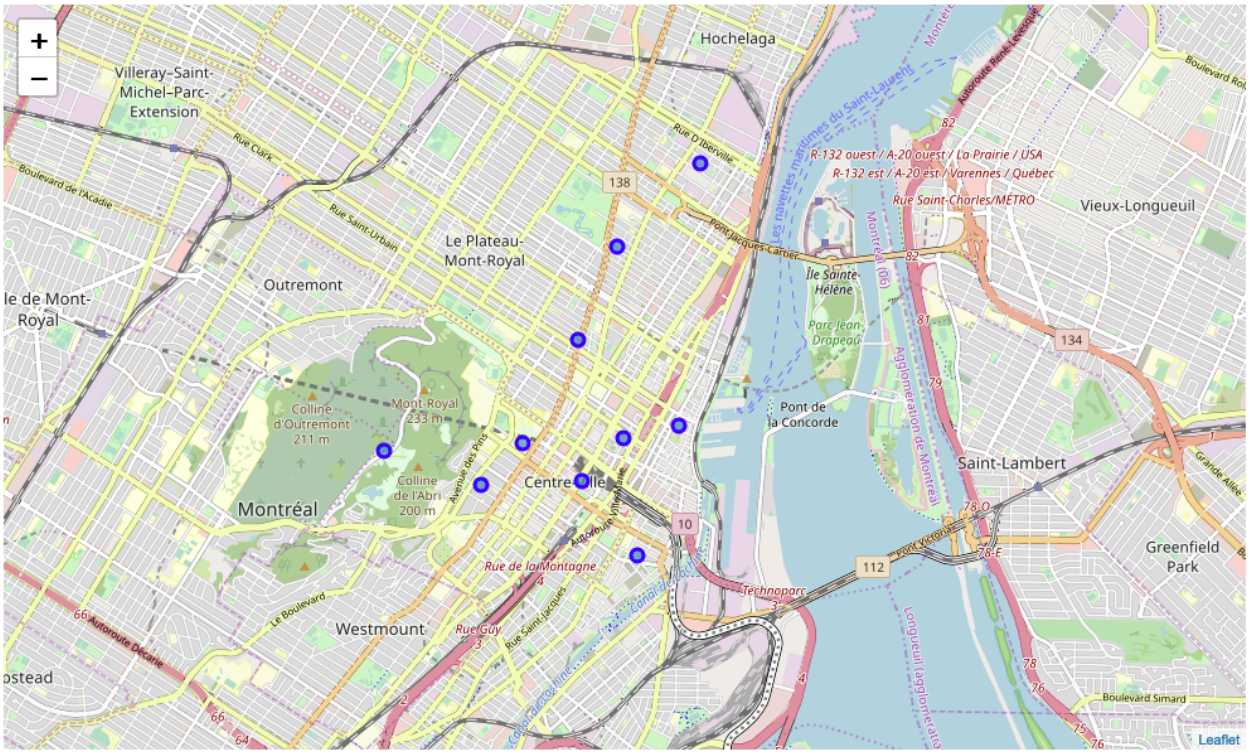
By using the information in the Introduction as well as information contained in the Literature Review section, one can make a table containing all of the postal code prefixes that have land in Ville Marie. The following table contains the postal code prefixes that contain neighborhoods in Ville Marie, the descriptions of the postal code prefixes' areas from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_H, and names of other neighborhoods that are in those postal code prefixes based on Google Maps and the Wikipedia pages about Montreal and Ville Marie. The "Neighborhood" column contains the descriptions from the Wikipedia page for postal codes. The Additional Neighborhoods column contains information gleaned from the other sources.

PostalCode	Neighborhood	AdditionalNeighborhoods
H2K	Centre-Sud North	Sainte-Marie
H2L	Centre-Sud South	Gay Village
H2X	Plateau Mont-Royal Southeast	Quartier Des Spectacles
H2Y	Old Montreal	Quartier International De Montréal
H2Z	Downtown Montreal Northeast	
H3A	Downtown Montreal North (McGill University)	Golden Square Mile
H3B	Downtown Montreal East	
H3C	Griffintown (Includes Île Notre-Dame & Île Sainte-Hélène) (Université de Montréal)	Cité du Multimédia, Saint Helen's Island, Notre Dame Island
H3G	Downtown Montreal Southeast (Concordia University)	
H3H	Downtown Montreal Southwest	Shaughnessy Village, part of Mount Royal Park

2.2 Neighborhood Maps

2.2.1 Toronto

2.2.2 Montreal



2.3 Coordinates

I need the coordinates or each 3-character postal code prefix in order to find the top 100 venues within 500 meters of them.

2.3.1 Toronto

I try to use geocoder to get the coordinates of the 3-character postal code prefixes, but every attempt results in the following:

```
<[REQUEST_DENIED] Google - Geocode [empty]>
```

Therefore, I use the CSV file containing postal codes prefixes and coordinates that is mentioned in the literature review.

After selecting only the postal codes in Downtown Toronto, this is the set of postal codes and neighborhood names I am using:

PostalCode	Borough	Neighborhood	Latitude	Longitude
M4W	Downtown Toronto	Rosedale	43.679563	-79.377529
M4X	Downtown Toronto	Cabbagetown, St. James Town	43.667967	-79.367675
M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
M5B	Downtown Toronto	Ryerson, Garden District	43.657162	-79.378937
M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306
M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
M5H	Downtown Toronto	Adelaide, King, Richmond	43.650571	-79.384568
M5J	Downtown Toronto	Harbourfront East, Toronto Islands, Union Station	43.640816	-79.381752
M5K	Downtown Toronto	Design Exchange, Toronto Dominion Centre	43.647177	-79.381576
M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817
M5S	Downtown Toronto	Harbord, University of Toronto	43.662696	-79.400049
M5T	Downtown Toronto	Chinatown, Grange Park, Kensington Market	43.653206	-79.400049
M5V	Downtown Toronto	CN Tower, Bathurst Quay, Island airport, Harbo...	43.628947	-79.394420
M5W	Downtown Toronto	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846
M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.382280
M6G	Downtown Toronto	Christie	43.669542	-79.422564
M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494

2.3.2 Montreal

The coordinates of the postal code prefixes were critical for this study. I got them by googling "coordinates of XXX" where XXX was a postal code. If there were no legal or practical prohibitions, I think I could have scraped the coordinates from the result page in Python. However, the previous labs provided CSV files with the location data, so I decided not to try to scrape Google pages in my Python code. I copied the coordinates from the Google results and put them into a CSV. My Python code converts the strings of characters from Google into numerical coordinates that can be used with FourSquare and Folium.

The following shows the text that is in the CSV file that my code reads in.

PostalCode	CoordsFromGoogle
H2K	45.5302° N, 73.5527° W
H2L	45.5222° N, 73.5641° W
H2X	45.5133° N, 73.5694° W
H2Y	45.5052° N, 73.5557° W
H2Z	45.5039° N, 73.5632° W
H3A	45.5035° N, 73.5769° W
H3B	45.4999° N, 73.5689° W
H3C	45.4928° N, 73.5614° W
H3G	45.4995° N, 73.5826° W
H3H	45.5027° N, 73.5958° W

This shows the parsed postal codes and the associated neighborhood names:

PostalCode	Neighborhood	Latitude	Longitude
H2K	Centre-Sud North, Sainte-Marie	45.5302	-73.5527
H2L	Centre-Sud South, Gay Village	45.5222	-73.5641
H2X	Plateau Mont-Royal Southeast, Quartier Des Spe...	45.5133	-73.5694
H2Y	Old Montreal, Quartier International De Montréal	45.5052	-73.5557
H2Z	Downtown Montreal Northeast	45.5039	-73.5632
H3A	Downtown Montreal North (McGill University), G...	45.5035	-73.5769
H3B	Downtown Montreal East	45.4999	-73.5689
H3C	Griffintown (Includes Île Notre-Dame & Île Sai...	45.4928	-73.5614
H3G	Downtown Montreal Southeast (Concordia Univers...	45.4995	-73.5826
H3H	Downtown Montreal Southwest, Shaughnessy Villa...	45.5027	-73.5958

2.4 Venues

In addition to the above, I use FourSquare to get information about venues near the postal code prefixes' locations. I search for the 100 top venues that are within a 500-meter radius of each postal code coordinates. I used the same technique for both Montreal and Toronto.

2.4.1 Toronto

The following shows the 5 most common categories of venues for each postal code in Downtown Toronto.

```
----Adelaide, King, Richmond----
         venue  freq
0  Coffee Shop  0.08
1   Steakhouse  0.04
2         Café  0.04
3          Bar  0.04
4       Bakery  0.03


----Berczy Park----
            venue  freq
0     Coffee Shop  0.07
1    Cocktail Bar  0.06
2  Farmers Market  0.04
3        Beer Bar  0.04
4            Café  0.04


----CN Tower, Bathurst Quay, Island airport, Harbourfront West, King and
Spadina, Railway Lands, South Niagara----
              venue  freq
0   Airport Service  0.19
1    Airport Lounge  0.12
2  Airport Terminal  0.12
3             Plane  0.06
4          Boutique  0.06


----Cabbagetown, St. James Town----
                venue  freq
0              Bakery  0.06
1         Coffee Shop  0.06
2         Pizza Place  0.04
3  Italian Restaurant  0.04
4              Market  0.04


----Central Bay Street----
                venue  freq
0         Coffee Shop  0.16
1                Café  0.05
2  Italian Restaurant  0.05
```

```
3        Burger Joint  0.04
4      Ice Cream Shop  0.04


----Chinatown, Grange Park, Kensington Market----
                           venue  freq
0                           Café  0.06
1          Vietnamese Restaurant  0.05
2  Vegetarian / Vegan Restaurant  0.05
3            Dumpling Restaurant  0.04
4             Chinese Restaurant  0.04


----Christie----
           venue  freq
0  Grocery Store  0.18
1           Café  0.18
2           Park  0.12
3    Candy Store  0.06
4     Restaurant  0.06


----Church and Wellesley----
                 venue  freq
0          Coffee Shop  0.08
1     Sushi Restaurant  0.05
2  Japanese Restaurant  0.05
3              Gay Bar  0.04
4           Restaurant  0.04


----Commerce Court, Victoria Hotel----
                venue  freq
0         Coffee Shop  0.11
1                Café  0.07
2               Hotel  0.06
3          Restaurant  0.05
4  Seafood Restaurant  0.03


----Design Exchange, Toronto Dominion Centre----
         venue  freq
0  Coffee Shop  0.14
1        Hotel  0.08
2         Café  0.07
3   Restaurant  0.04
4          Bar  0.04


----First Canadian Place, Underground city----
         venue  freq
0  Coffee Shop  0.12
1         Café  0.07
2   Steakhouse  0.04
3        Hotel  0.04
4   Restaurant  0.04
```

```
----Harbord, University of Toronto----
                venue  freq
0                Café  0.14
1          Restaurant  0.06
2   Italian Restaurant  0.06
3  Japanese Restaurant  0.06
4           Bookstore  0.06


----Harbourfront----
           venue  freq
0    Coffee Shop  0.16
1         Bakery  0.07
2           Park  0.07
3            Pub  0.07
4  Breakfast Spot  0.04


----Harbourfront East, Toronto Islands, Union Station----
               venue  freq
0        Coffee Shop  0.12
1           Aquarium  0.05
2  Italian Restaurant  0.04
3               Café  0.04
4              Hotel  0.04


----Queen's Park----
               venue  freq
0        Coffee Shop  0.29
1               Park  0.05
2                Gym  0.05
3        Yoga Studio  0.03
4  Chinese Restaurant  0.03


----Rosedale----
              venue  freq
0              Park  0.50
1        Playground  0.25
2             Trail  0.25
3  Afghan Restaurant  0.00
4       Music Venue  0.00


----Ryerson, Garden District----
                      venue  freq
0               Coffee Shop  0.10
1            Clothing Store  0.05
2            Cosmetics Shop  0.04
3                      Café  0.04
4  Middle Eastern Restaurant  0.03


----St. James Town----
           venue  freq
0           Café  0.06
1    Coffee Shop  0.06
```

```
2      Restaurant  0.05
3           Hotel  0.03
4  Cosmetics Shop  0.03


----Stn A PO Boxes 25 The Esplanade----
                venue  freq
0         Coffee Shop  0.12
1                Café  0.04
2   Seafood Restaurant  0.03
3  Japanese Restaurant  0.03
4         Cocktail Bar  0.03
```

2.4.2 Montreal

The following shows the 5 most common categories of venues for each postal code in Ville Marie:

```
----Centre-Sud North, Sainte-Marie----
                  venue  freq
0      French Restaurant  0.12
1                   Park  0.12
2            Coffee Shop  0.08
3  Performing Arts Venue  0.08
4         Sandwich Place  0.08


----Centre-Sud South, Gay Village----
              venue  freq
0        Restaurant  0.10
1  Sushi Restaurant  0.10
2    Breakfast Spot  0.10
3      Concert Hall  0.05
4    Hardware Store  0.05


----Downtown Montreal East----
            venue  freq
0     Coffee Shop  0.12
1           Hotel  0.06
2  Clothing Store  0.04
3            Café  0.04
4      Restaurant  0.04


----Downtown Montreal North (McGill University), Golden Square Mile----
            venue  freq
0           Hotel  0.09
1     Coffee Shop  0.07
2  Clothing Store  0.07
3  Sandwich Place  0.06
4             Gym  0.04


----Downtown Montreal Northeast----
                venue  freq
```

```
0               Hotel  0.08
1    Asian Restaurant  0.06
2   French Restaurant  0.06
3  Chinese Restaurant  0.06
4         Coffee Shop  0.05


----Downtown Montreal Southeast (Concordia University)----
         venue  freq
0   Art Museum  0.08
1        Hotel  0.08
2         Café  0.08
3  Coffee Shop  0.06
4 Burger Joint  0.04


----Downtown Montreal Southwest, Shaughnessy Village, part of Mount Royal Park-
---
             venue  freq
0    Historic Site  0.25
1         Mountain  0.25
2         Bus Stop  0.25
3             Lake  0.25
4 Arepa Restaurant  0.00


----Griffintown (Includes Île Notre-Dame & Île Sainte-Hélène) (Université de
Montréal), Cité du Multimédia, Saint Helen's Island, Notre Dame Island----
                   venue  freq
0                   Café  0.05
1               Pharmacy  0.05
2 Furniture / Home Store  0.05
3     Italian Restaurant  0.05
4                 Bakery  0.05


----Old Montreal, Quartier International De Montréal----
               venue  freq
0  French Restaurant  0.09
1              Hotel  0.08
2               Café  0.07
3         Steakhouse  0.04
4 Italian Restaurant  0.04


----Plateau Mont-Royal Southeast, Quartier Des Spectacles----
              venue  freq
0              Café  0.06
1 Indian Restaurant  0.05
2               Bar  0.05
3             Hotel  0.04
4  Sushi Restaurant  0.03
```

2.5 Data Cleaning

The data related to Toronto needs significant cleaning. Many postal codes have no borough information. These postal codes are removed from the data set. Some postal codes do not borough names but not neighborhood names. In this case, I assume that the neighborhood name is the same as the borough name. (All of the postal codes that are in Downtown Toronto are associated with neighborhoods, so this would only matter if the project is expanded to include other boroughs.)

The data related to Montreal also needs significant cleansing. Ville Marie only contains 10 postal codes, so I use a spreadsheet to parse and clean the Wikipedia Montreal Postal Code page. I then add neighborhoods that are in the Montreal and Ville Marie Wikipedia pages to the CSV file. This file contains cleaned data that can be processed by Python without further cleaning.

2.6 Assigning Neighborhoods to Postal Codes

After cleansing, the Toronto and Montreal data have one or more than one neighborhood per postal code. The neighborhoods for each postal code are combined into a comma separated string because the search for venues is based on the coordinates of each postal code rather than the coordinates of each neighborhood. (I could have used the coordinates or each neighborhood, but that would have been an expansion of the scope of the project. If the customers want a more detailed analysis, that would be a good path to pursue.)

2.7 Feature Selection

For each postal code, the 100 top venues within 500 meters are found using Foursquare. The category of each venue is also found using Foursquare. The number of venues in each category is counted. The features that are used with k-means are the frequencies of the venues in each category. Some postal codes have fewer than 100 venues, but this does not matter because the clustering is done based on the frequencies of the venues in each category, and even postal codes that have more than 100 venues do not have venues in every category. If a postal code does not have a venue in a given category, the frequency of the venues in that category is zero, whether the venue has 100 top venues or less than 100 top venues. The following shows all of the rows and some of the columns in the feature matrix, along with the corresponding neighborhoods.

	Neighborhood	American Restaurant	Antique Shop	Vietnamese Restaurant	Yoga Studio
0	Adelaide, King, Richmond	0.020000	0.000000	0.000000	0.000000

	Neighborhood	American Restaurant	Antique Shop	Vietnamese Restaurant	Yoga Studio
1	Berczy Park	0.000000	0.000000	0.000000	0.000000
2	CN Tower, Bathurst Quay, Island airport, Harbo...	0.000000	0.000000	0.000000	0.000000
3	Cabbagetown, St. James Town	0.000000	0.000000	0.000000	0.000000
4	Central Bay Street	0.012048	0.000000	0.000000	0.012048
5	Centre-Sud North, Sainte-Marie	0.000000	0.000000	0.040000	0.000000
6	Centre-Sud South, Gay Village	0.000000	0.000000	0.000000	0.000000
7	Chinatown, Grange Park, Kensington Market	0.000000	0.000000	0.053191	0.000000
8	Christie	0.000000	0.000000	0.000000	0.000000
9	Church and Wellesley	0.011905	0.000000	0.011905	0.023810
10	Commerce Court, Victoria Hotel	0.030000	0.000000	0.000000	0.000000
11	Design Exchange, Toronto Dominion Centre	0.030000	0.000000	0.000000	0.000000
12	Downtown Montreal East	0.000000	0.000000	0.000000	0.010000
13	Downtown Montreal North (McGill University), G...	0.000000	0.000000	0.000000	0.014706
14	Downtown Montreal Northeast	0.000000	0.000000	0.000000	0.010000
15	Downtown Montreal Southeast (Concordia Univers...	0.000000	0.000000	0.000000	0.020000
16	Downtown Montreal Southwest, Shaughnessy Villa...	0.000000	0.000000	0.000000	0.000000
17	First Canadian Place, Underground city	0.020000	0.000000	0.000000	0.000000
18	Griffintown (Includes Île Notre-Dame & Île Sai...	0.000000	0.000000	0.000000	0.000000
19	Harbord, University of Toronto	0.000000	0.000000	0.000000	0.027778
20	Harbourfront	0.000000	0.022222	0.000000	0.000000
21	Harbourfront East, Toronto Islands, Union Station	0.000000	0.000000	0.000000	0.000000
22	Old Montreal, Quartier International De Montréal	0.000000	0.000000	0.010000	0.010000
23	Plateau Mont-Royal Southeast, Quartier Des Spe...	0.000000	0.000000	0.010753	0.010753
24	Queen's Park	0.000000	0.000000	0.000000	0.026316
25	Rosedale	0.000000	0.000000	0.000000	0.000000

	Neighborhood	American Restaurant	Antique Shop	Vietnamese Restaurant	Yoga Studio
26	Ryerson, Garden District	0.010000	0.000000	0.010000	0.000000
27	St. James Town	0.010000	0.000000	0.000000	0.000000
28	Stn A PO Boxes 25 The Esplanade	0.010526	0.010526	0.000000	0.000000

3 Methodology

In this project, I use k-means clustering to find groups of postal code prefixes that are similar based on the most popular types of venues that are near the postal code prefixes locations. I experiment with different values for k because if k is too low, there won't be much differentiation between different postal codes, but if k is too large, each postal code would be in a cluster by itself. The best value of k will provide at least one cluster that contain postal codes from both cities, and it will provide clusters that are significantly different from other clusters.

Note: I use the word "postal code prefix" to mean either a 3-character string or to mean the area on a map that has addresses that use that 3-character prefix. I believe the reader will be able to differentiate between the two meanings.

Another way of doing this analysis would be to consider each postal code prefix in Montreal, and then find the postal code prefix in Toronto that is the most similar. The biggest potential problem with this is that it might find a match for a Montreal postal code prefix that was actually not very similar. This difficulty could be overcome by setting a minimum similarity score so that no match would be returned if the most similar neighborhood had a score that was too low. Another potential problem is that many postal codes in Montreal would match up with the same postal code in Toronto. If this happened, the pool of potential business to be recruited would be small because many other similar neighborhoods in Toronto would be left out. This difficulty could be overcome by finding the best N matches for each neighborhood in Montreal, but then we would have to find the optimal value of N. This could be an interesting way of solving the problem if using k-means clustering does not solve it.

4 Results

This section explores the data in more details and looks at the results of the k-means clustering.

4.1 Data Exploration

The feature matrix is very sparse. In other words, most of its entries are zero. In order to get a better look at the data, I construct data frames that show the top categories of venues for each neighborhood.

	Neighborhood	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	Adelaide, King, Richmond	Coffee Shop	Steakhouse	Café	Bar	Hotel	Sushi Restaurant	Restaurant	Bakery	Thai Restaurant	Asian Restaurant
1	Berczy Park	Coffee Shop	Cocktail Bar	Seafood Restaurant	Steakhouse	Bakery	Café	Farmers Market	Cheese Shop	Beer Bar	Comfort Food Restaurant
2	CN Tower, Bathurst Quay, Island airport, Harbo...	Airport Service	Airport Lounge	Airport Terminal	Plane	Bar	Rental Car Location	Sculpture Garden	Boutique	Boat or Ferry	Airport
3	Cabbagetown, St. James Town	Bakery	Coffee Shop	Pizza Place	Restaurant	Café	Flower Shop	Pub	Italian Restaurant	Market	Pet Store
4	Central Bay Street	Coffee Shop	Italian Restaurant	Café	Sandwich Place	Burger Joint	Ice Cream Shop	Japanese Restaurant	Bar	Salad Place	Bubble Tea Shop

....

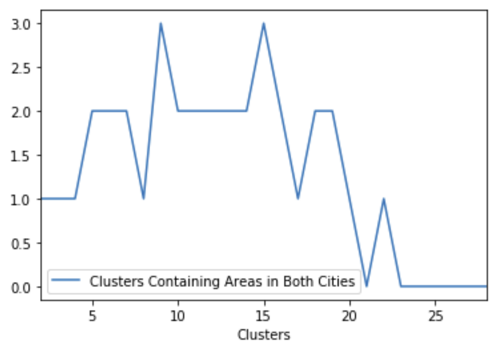
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
16	Downtown Montreal Southwest, Shaughnessy Villa...	Historic Site	Lake	Mountain	Bus Stop	Yoga Studio	Diner	Ethiopian Restaurant	Empanada Restaurant	Electronics Store	Eastern European Restaurant

...

	Neighborhood	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
24	Queen's Park	Coffee Shop	Gym	Park	Fried Chicken Joint	Portuguese Restaurant	Beer Bar	Smoothie Shop	Sandwich Place	Burger Joint	Burrito Place
25	Rosedale	Park	Trail	Playground	Donut Shop	Diner	Discount Store	Dive Bar	Dog Run	Doner Restaurant	Yoga Studio
26	Ryerson, Garden District	Coffee Shop	Clothing Store	Café	Cosmetics Shop	Middle Eastern Restaurant	Italian Restaurant	Fast Food Restaurant	Sporting Goods Shop	Pizza Place	Bubble Tea Shop
27	St. James Town	Coffee Shop	Café	Restaurant	Clothing Store	Beer Bar	Breakfast Spot	Bakery	Cocktail Bar	Hotel	Cosmetics Shop
28	Stn A PO Boxes 25 The Esplanade	Coffee Shop	Café	Cocktail Bar	Japanese Restaurant	Restaurant	Seafood Restaurant	Hotel	Beer Bar	Park	Bakery

4.2 Finding the Optimal Value for K in K-means Clustering

One of the challenges in using k-means is to find the best value for k. K is the number of clusters that will be found, and it is an input to the algorithm. For this analysis, I want at least two clusters, because if there is only one cluster, there is no differentiation between any of the neighborhoods. I only have 29 postal codes, so I can't have more than 29 clusters (unless a cluster is empty, which doesn't make sense). Since I want to map neighborhoods in Downtown Toronto to similar neighborhoods in Ville Marie, I want clusters that contain neighborhoods in both boroughs. Since I want to differentiate the different types of neighborhoods, more clusters is better than fewer clusters. I can run k-means with 28 different values of k, and then find out how many clusters contain neighborhoods in both boroughs for each k. The result is presented in the following graph:



This shows that there are 2 values of k that generate 3 clusters that containing neighborhoods from both boroughs, and that 3 is the maximum number of clusters that contain neighborhoods from both cities. The following table shows that k = 9 and k = 15 each result in 3 clusters containing areas in both boroughs.

K	Clusters Containing Areas in Both Cities
9	3
15	3

I choose to use 9 clusters for the rest of this analysis because using more clusters would result in more single-borough clusters. Since the goal is to find neighborhoods in Ville Marie that are similar to neighborhoods in Downtown Toronto, clusters containing only one borough are not useful.

4.3 Results of Clustering

Here are the boroughs and neighborhoods that are in the clusters when k = 9.

CLUSTER NUMBER: 0

	Borough	Neighborhood
9	Downtown Toronto	Church and Wellesley
26	Downtown Toronto	Ryerson, Garden District
4	Downtown Toronto	Central Bay Street
0	Downtown Toronto	Adelaide, King, Richmond
21	Downtown Toronto	Harbourfront East, Toronto Islands, Union Station
11	Downtown Toronto	Design Exchange, Toronto Dominion Centre
10	Downtown Toronto	Commerce Court, Victoria Hotel
28	Downtown Toronto	Stn A PO Boxes 25 The Esplanade
17	Downtown Toronto	First Canadian Place, Underground city
12	Ville-Marie	Downtown Montreal East

CLUSTER NUMBER: 1

	Borough	Neighborhood
20	Downtown Toronto	Harbourfront
5	Ville-Marie	Centre-Sud North, Sainte-Marie

CLUSTER NUMBER: 2

	Borough	Neighborhood
25	Downtown Toronto	Rosedale

CLUSTER NUMBER: 3

	Borough	Neighborhood
3	Downtown Toronto	Cabbagetown, St. James Town
27	Downtown Toronto	St. James Town
1	Downtown Toronto	Berczy Park
19	Downtown Toronto	Harbord, University of Toronto
7	Downtown Toronto	Chinatown, Grange Park, Kensington Market
23	Ville-Marie	Plateau Mont-Royal Southeast, Quartier Des Spe...
22	Ville-Marie	Old Montreal, Quartier International De Montréal
14	Ville-Marie	Downtown Montreal Northeast
13	Ville-Marie	Downtown Montreal North (McGill University), G...
18	Ville-Marie	Griffintown (Includes Île Notre-Dame & Île Sai...
15	Ville-Marie	Downtown Montreal Southeast (Concordia Univers...

CLUSTER NUMBER: 4

	Borough	Neighborhood
16	Ville-Marie	Downtown Montreal Southwest, Shaughnessy Villa...

CLUSTER NUMBER: 5

	Borough	Neighborhood
24	Downtown Toronto	Queen's Park

CLUSTER NUMBER: 6

```
            Borough                                     Neighborhood
2  Downtown Toronto  CN Tower, Bathurst Quay, Island airport, Harbo...


CLUSTER NUMBER: 7

       Borough                 Neighborhood
6  Ville-Marie  Centre-Sud South, Gay Village


CLUSTER NUMBER: 8

            Borough Neighborhood
8  Downtown Toronto       Christie
```

One interesting observation is that the clusters that don't contain neighborhoods from both boroughs in fact don't even contain two neighborhoods from the **same** borough. This means that, according to k-means clustering, those neighborhoods are one-of-a-kind. Is that correct? See the discussion section for more.

5 Discussion

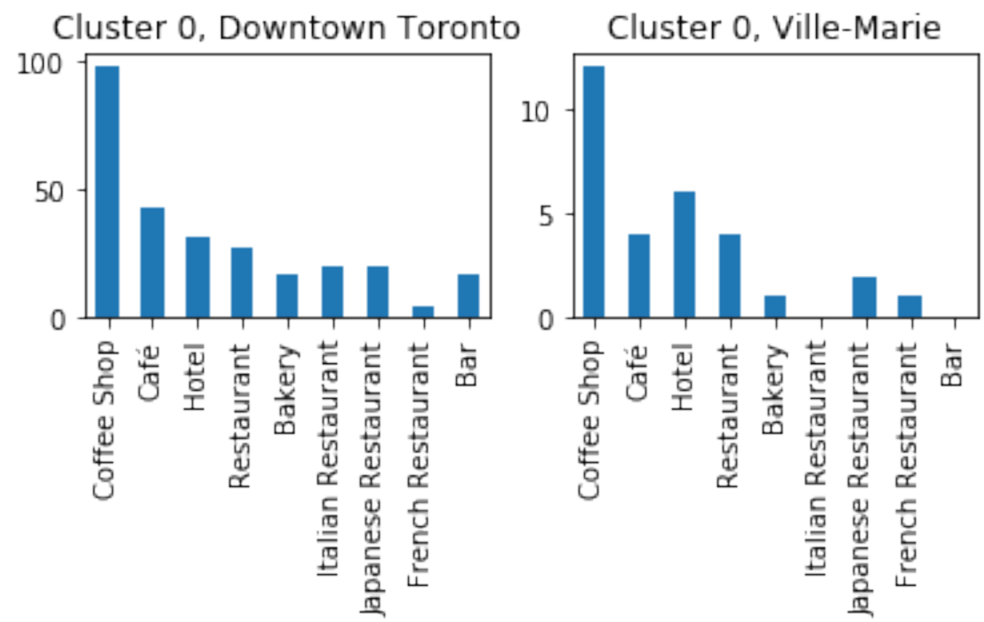
The previous chapter shows the results of the clustering. One might wonder whether these are good results. What makes the neighborhoods that are in the clusters similar? Why are some neighborhoods so different that they are not clustered with neighborhoods in the other borough? I decided to use graphs to show the similarities between the clusters contained neighborhoods from both cities, and tables to show why some clusters were unique. Remember that the clusters were chosen using the frequencies of the different categories of venues. Graphs containing the frequencies of 100 different types of venues are not easy to compare, so I only graph the frequencies of the 10 most popular types of venues. First, I find the most common categories of venues in all of the neighborhoods

Category	**Count**
Coffee-Shop	166
Café	99
Hotel	74
Restaurant	54
Bakery	46
Italian-Restaurant	39
Japanese-Restaurant	36
French-Restaurant	33
Bar	32
Gym	29

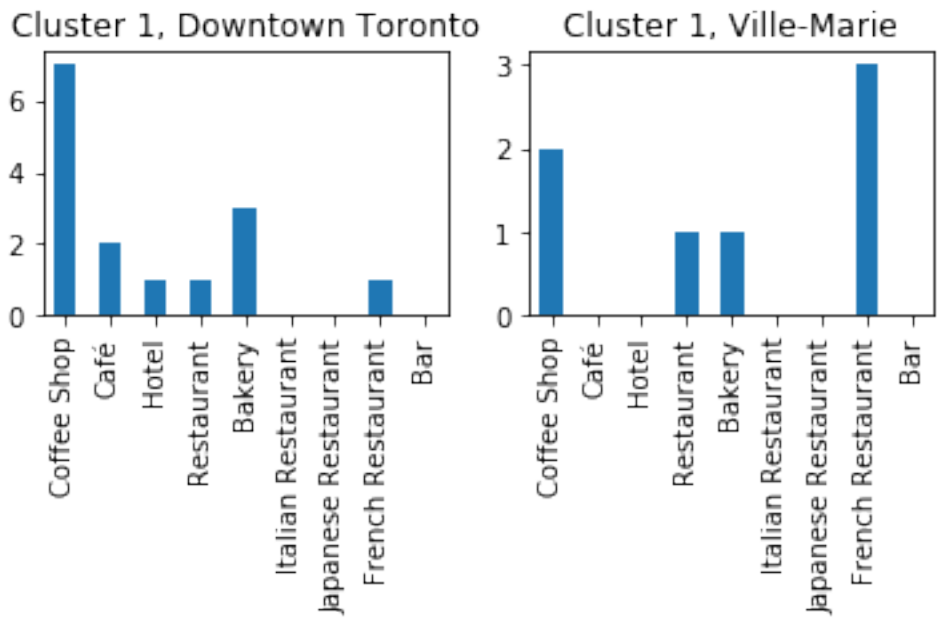
I choose to use the top 9 categories because the French Restaurant and Bar are almost tied for 8th and 9th place, but Gym is 10% less common than Bar.

For each cluster, I create either a pair of bar graphs showing the frequency of the 9 most-popular categories in each cluster, or a table showing the venue categories (referred to as venue types) that are in the neighborhoods in that cluster

5.1.1.1 Cluster 0, which where coffee shops, cafes, hotels, and restaurants are most popular, especially coffee shops



5.1.1.2 Cluster 1, which where coffee shops are popular, but bakeries are also popular, and hotels and cafes aren't so popular

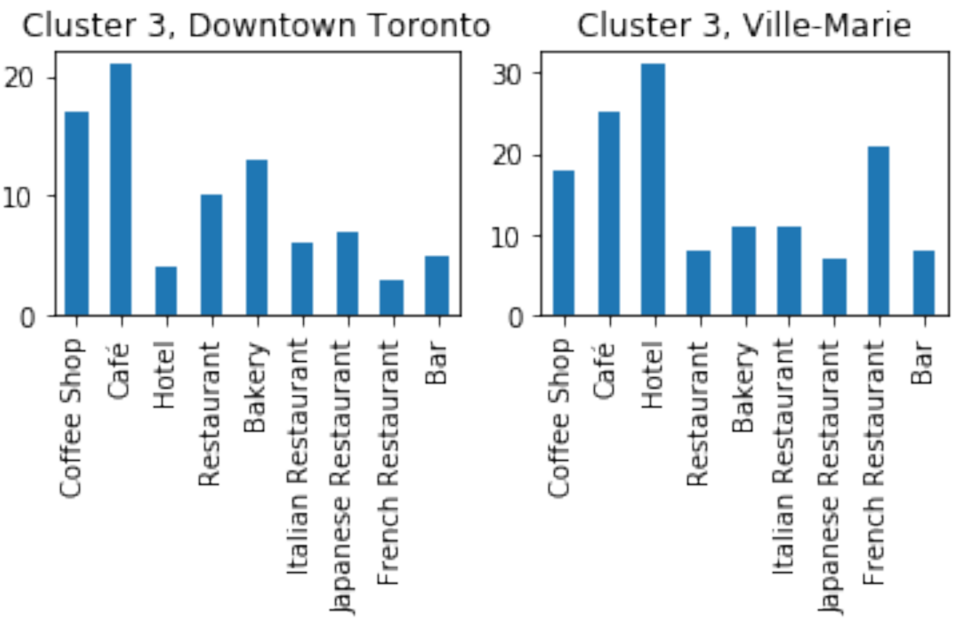


5.1.1.3 Cluster 2, where outdoor venues are popular, and which only has a presence in one city

```
Cluster 2, Downtown Toronto
Neighborhood group: Rosedale
        Venue Type                      Count
        Park                                2
        Playground                          1
        Trail                               1
```

5.1.1.4 Cluster 3, where cafes are more popular than coffee shops, and where all 9 of the most popular venues types are well represented

Diversity of venue types is valued here.



5.1.1.5 Cluster 4, where outdoor venues are popular, but not the same kind of outdoor venues as Cluster 2's venues

This cluster's neighborhood is in Ville Marie, whereas cluster 2's neighborhood is in Downtown Toronto. Although the clustering algorithm did not match these 2 venues, this discussion shows that they are both characterized by outdoor venue types, so they might be good matches for each other. In a future project, it would be worth exploring the use of higher-level venue categories, like "outdoor", "restaurant", and "store", and doing clustering based on those higher-level categories. Using higher-level features would also be a good way to avoid the "curse of dimensionality". For more about the curse of dimensionality, see the Wikipedia and Scikit-learn documentation that are referenced in the literature search.

```
Cluster 4, Ville-Marie
Neighborhood group: Downtown Montreal Southwest, Shaughnessy Village, part of
Mount Royal Park
        Venue Type                     Count
        Bus Stop                           1
        Historic Site                      1
        Lake                               1
        Mountain                           1
```

5.1.1.6 Cluster 5, which is only represented in one city, and has a disproportionate number of coffee shops

```
Cluster 5, Downtown Toronto
```

```
Neighborhood group: Queen's Park
        Venue Type                    Count
        Arts & Crafts Store               1
        Bar                               1
        Beer Bar                          1
        Burger Joint                      1
        Burrito Place                     1
        Café                              1
        Chinese Restaurant                1
        Coffee Shop                      11
        College Auditorium                1
        Creperie                          1
        Diner                             1
        Fast Food Restaurant              1
        Fried Chicken Joint               1
        Gym                               2
        Hobby Shop                        1
        Italian Restaurant                1
        Mexican Restaurant                1
        Music Venue                       1
        Nightclub                         1
        Park                              2
        Portuguese Restaurant             1
        Sandwich Place                    1
        Smoothie Shop                     1
        Sushi Restaurant                  1
        Theater                           1
        Yoga Studio                       1
```

5.1.1.7 Cluster 6, which is only represented in one city, and which contains venues popular at airports

Ville Marie doesn't have an airport, so that fact that this cluster is in only one city isn't surprising.

```
Cluster 6, Downtown Toronto
Neighborhood group: CN Tower, Bathurst Quay, Island airport, Harbourfront West,
King and Spadina, Railway Lands, South Niagara
        Venue Type                    Count
        Airport                           1
        Airport Food Court                1
        Airport Lounge                    2
        Airport Service                   3
        Airport Terminal                  2
        Bar                               1
        Boat or Ferry                     1
        Boutique                          1
        Harbor / Marina                   1
        Plane                             1
        Rental Car Location               1
        Sculpture Garden                  1
```

5.1.1.8 Cluster 7, which is only represented in one city, and doesn't contain coffee shops nor cafes

```
Cluster 7, Ville-Marie
Neighborhood group: Centre-Sud South, Gay Village
        Venue Type                     Count
        Asian Restaurant                   1
        Beer Bar                           1
        Bike Rental / Bike Share           1
        Breakfast Spot                     2
        Caribbean Restaurant               1
        Concert Hall                       1
        Farmers Market                     1
        Fast Food Restaurant               1
        Gym                                1
        Hardware Store                     1
        Hostel                             1
        Pharmacy                           1
        Poutine Place                      1
        Restaurant                         2
        Supermarket                        1
        Sushi Restaurant                   2
        Thai Restaurant                    1
```

5.1.1.9 Cluster 8, which is only represented in one city, and doesn't seem to have a lot of the most popular venues

```
Cluster 8, Downtown Toronto
Neighborhood group: Christie
        Venue Type                     Count
        Athletics & Sports                 1
        Baby Store                         1
        Café                               3
        Candy Store                        1
        Coffee Shop                        1
        Convenience Store                  1
        Diner                              1
        Grocery Store                      3
        Italian Restaurant                 1
        Nightclub                          1
        Park                               2
        Restaurant                         1
```

6 Conclusion

The k-means clustering algorithm finds 15 postal codes in Downtown Toronto that have good matches in Ville Marie. Further analysis finds one more postal code in Downtown Toronto that has a good match in Montreal. The use of k-means to cluster neighborhoods in the two cities is very useful for finding neighborhoods in Ville Marie that are similar to neighborhoods in Downtown Toronto.

The discussion section shows that there is room for improvement in the method. The categories that are drawn from FourSquare are generally very detailed. The algorithm failed to match two postal codes even though they contained a preponderance of outdoor venues, because they contained different categories of outdoor venues. The results of the model could be improved by adding higher-level categories to the feature matrix, and either removing the corresponding low-level categories or leaving them in. This is a worthwhile avenue for future exploration.

This report shows that matching up neighborhoods in Downtown Toronto with neighborhoods in Ville Marie using k-means clustering give excellent results. The government of Montreal could use these results to design targeted campaigns to attract more business to Montreal. This technique could be used to match more neighborhoods in Montreal with more neighborhoods in Toronto, expanding the scope of the study, and resulting in more opportunities for attracting business to Montreal.