SIMILAR NEIGHBORHOODS IN DOWNTOWN TORONTO AND VILLE MARIE

Applied Data Science Capstone

Walter Dietrich

Agenda

- Introduction
- Data
- Methodology
- Results
- Discussion
- Conclusion

INTRODUCTION

Using Data Science to Guide Effective Marketing

Montreal

Important facts

- Second-most populous municipality in Canada
- 300 miles from Toronto, the most populous city in Canada
  - *Connected by the St. Lawrence River and Lake Ontario*
  - *Connected by Route 401*
- Divided into boroughs
- Government
  - *Mayor*
  - *City Council*
    - Democratically Elected
    - Mayor is “first among equals” on City Council
    - Executive Committee
    - Contains representatives from all of Montreal’s boroughs

Montreal would like to grow

This is the ”problem statement”.

- Executive committee and council want to increase growth rate
- Want to use Data Science to devise numbers-driven strategy for growth
- If a neighborhood in Toronto is like a neighborhood in Montreal, then businesses in that Toronto neighborhood are good candidates for inducements to open offices in Montreal
- Montreal's government would like start with a pilot project
  - *Assess viability*
  - *Find potential improvements*
  - *Future: expand scope*

Pilot Project

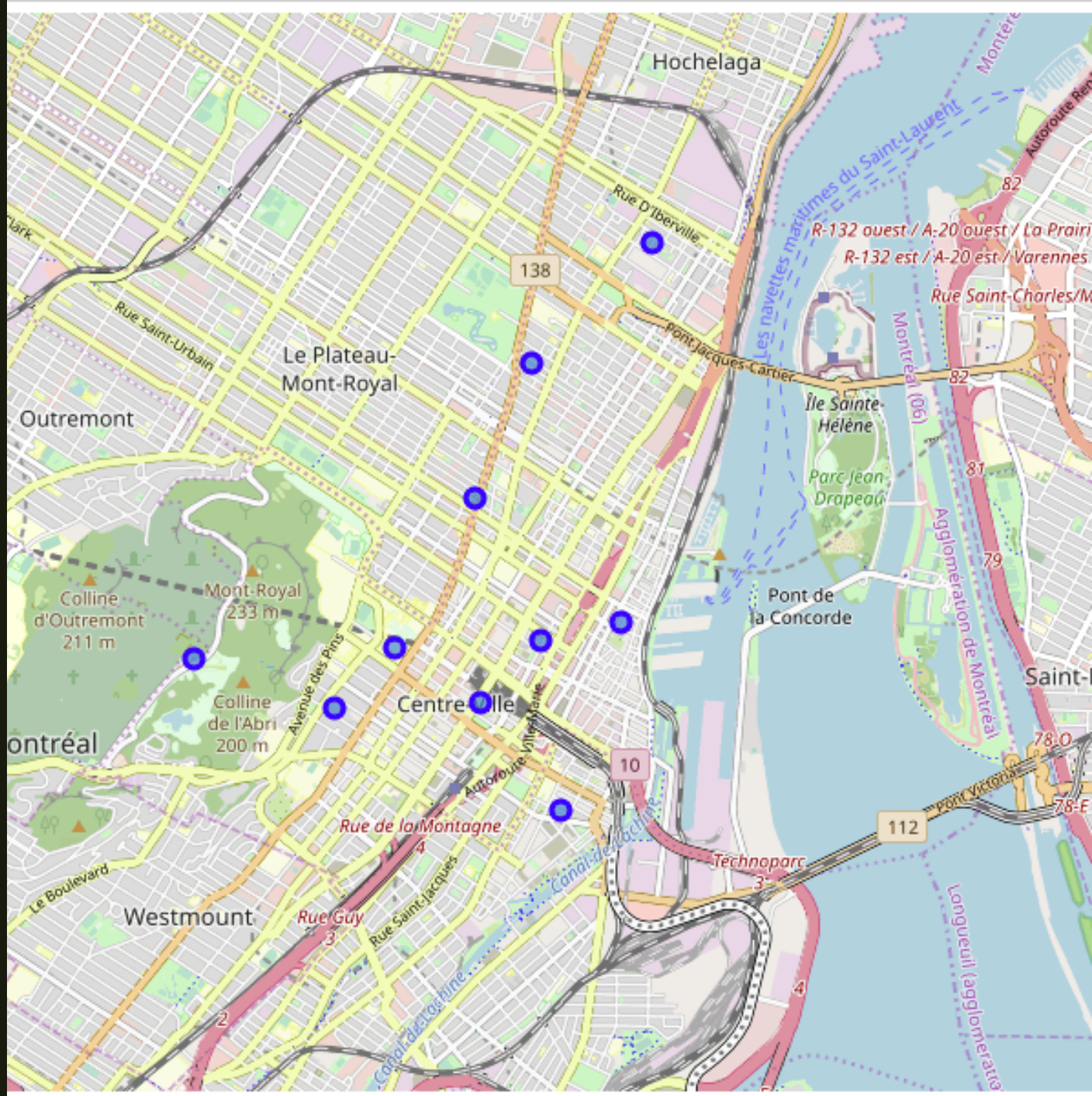
Use a pilot project to assess viability and find potential improvements

- Use one borough from Montreal and one borough from Toronto
  - *Montreal borough: Ville Marie*
  - *Toronto borough: Downtown Toronto*
- Ville Marie
  - *Has more neighborhoods than any other Montreal borough*
  - *Includes downtown neighborhoods*
  - *Sometimes written as Ville-Marie*
- Downtown Toronto
  - *Includes downtown neighborhoods*
- Use clustering to find neighborhoods in Montreal that are similar to counterparts in Toronto

Ville Marie Postal Code Map

Each blue circle represents the geographic coordinates of one Ville Marie 3-character postal code prefix.

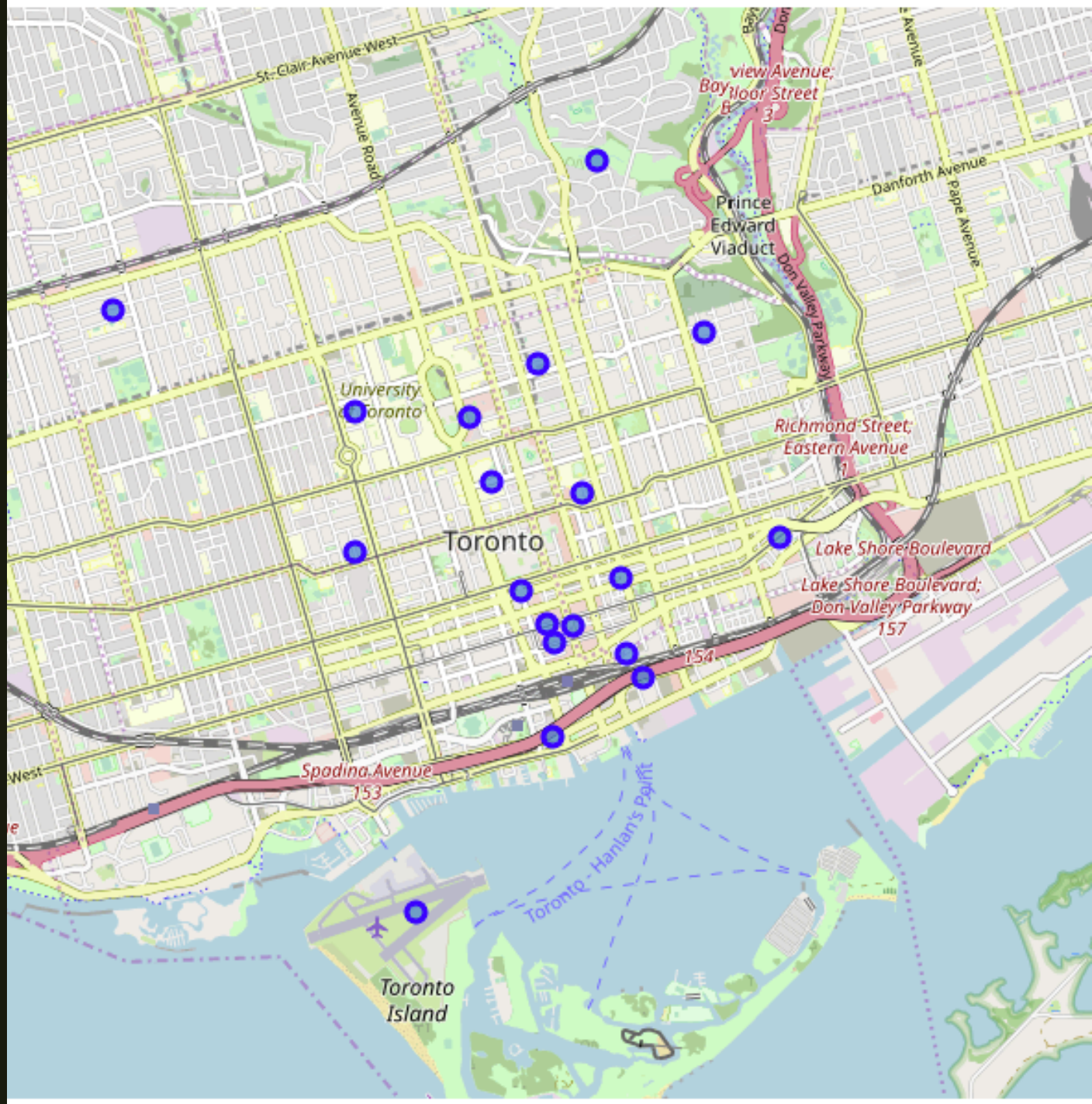
Each postal code prefix is associated with one or more Ville Marie neighborhoods.



Downtown Toronto Postal Code Centers

Each blue circle represents the geographic coordinates of one Downtown Toronto 3-character postal code prefix.

Each postal code prefix is associated with one or more Downtown Toronto neighborhoods.



Literature Review

Components of the Literature Review

- See the report for the details of the literature review
  1. *Text Descriptions of Montreal, Ville Marie, and Toronto*
  2. *Postal Code Information*
  3. *Maps Related to Ville Marie and Maps Related to Postal Codes that Contain Land in Ville Marie*
  4. *Algorithms*

Acknowledgements

- The author wishes to thank the following people and organizations

  - *The mayor of Montreal*
  - *The Montreal City Council*

  - *The authors of the Wikipedia pages about Toronto, Montreal, Ville Marie, and Canadian H and M postal codes*
  - *The people responsible for Scikit.learn classification code*

  - *The authors of the Coursera Applied Data Science courses*
  - *The Coursera students who reviewed my submissions*

DATA

Acquiring and Cleansing the Data that was used in this project

Overview

- This project depends on neighborhoods and the venues that are in the neighborhoods
- It builds on a previous project that used neighborhoods, 3-character postal code prefixes, the geographical coordinates of the postal code prefixes, and venues that are within a 500-meter radius of the postal code prefixes
- This project uses the 3-character postal code prefixes (commonly referred to as ”postal codes”), the neighborhoods that are in those postal codes, and the venues that are close to those postal codes’ coordinates

Descriptive Names

The descriptive names are the names of the neighborhoods

- Toronto
  - *Names and postal codes came from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M*
  - *Names and postal codes were scraped using Python*
  - *See https://github.com/WallyNY/Coursera_Capstone/blob/master/Capstone_week_3_part_1.ipynb*
- Montreal
  - *Names and postal codes came from sources listed in report*
  - *I built a CSV file manually from the sources mentioned*

Assigning Names to Postal Codes

- Each postal code specifies a geographic area
- Each neighborhood specifies a geographic area
- The borders of neighborhoods and postal codes are not necessarily aligned, so a neighborhood might have addresses that are in 2 different postal codes
- Postal codes can contain more than 1 neighborhood
- For each postal code, there is a list of neighborhoods that have addresses within the postal codes
  - *The lists appear in the following pages*

Coordinates

Coordinates of postal codes are essential

- Toronto
  - *Coordinates of postal codes came from CSV file at https://cocl.us/Geospatial_data*
  - *Link was provided in Coursera Applied Data Science Capstone course*
- Montreal
  - *Coordinates of Ville Marie postal codes came CSV file*
    - CSV file contained Google search results

Venues

- Venue information came from FourSquare
- For each postal code, found top venues within 500 meters of postal code's coordinates
  - *Up to 100 venues*
  - *Got each venue's Category*
  - *Categories were nouns like Café, Coffee Shop, Gym, Hotel, and Vietnamese Restaurant*
  - *Found the frequencies of each category in each postal code*
  - *Examples follow*

Sample venue category frequencies in Downtown Toronto, for top categories

Postal Code M4Y

---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

Postal Code M5C

----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

Sample venue category frequencies in Ville Marie, for top categories

Postal Code H3B

----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

Postal Code H2Z

----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

Data Cleaning

- See report for information about data cleaning

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
- See the following for examples

Sample Features

Neighborhood	American Restaurant	Antique Shop	Vietnamese Restaurant	Yoga Studio
Adelaide, King, Richmond	0.020	0.000	0.000	0.000
Berczy Park	0.000	0.000	0.000	0.000
CN Tower, Bathurst Quay, Island airport, Harbo...	0.000	0.000	0.000	0.000
Cabbagetown, St. James Town	0.000	0.000	0.000	0.000
Central Bay Street	0.012	0.000	0.000	0.012

Note: There are 26 neighborhoods and more than 100 venue categories, so the feature matrix has 26 rows and more than one hundred columns

METHODOLOGY

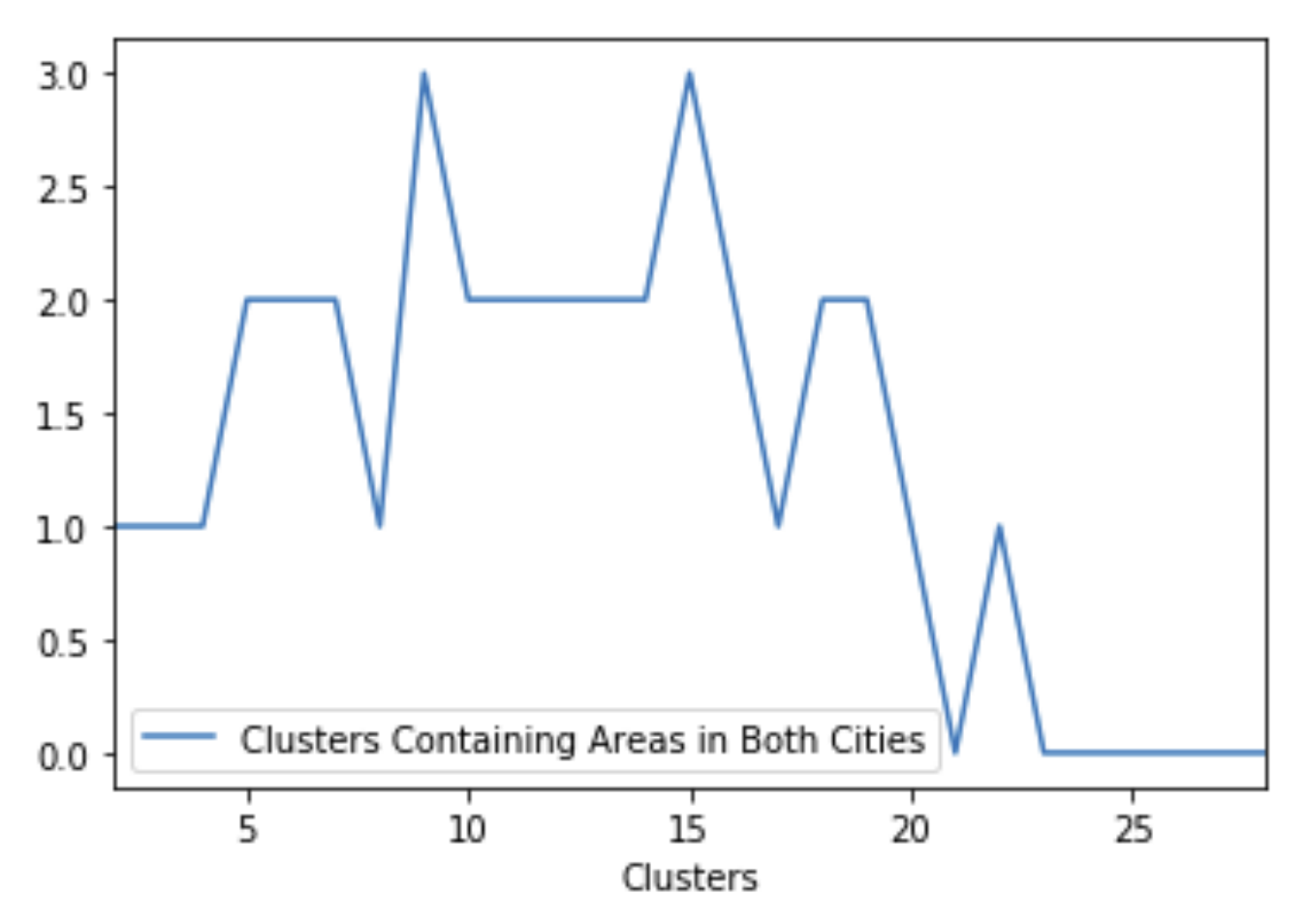
A scientific approach to finding matching neighborhoods in the two cities

K-means

- This project uses k-means clustering to find groups of postal codes that contain similar sets of venues
- K-means clustering: for a given value of *k*, k-means clustering finds *k* clusters, with each cluster containing one or more postal codes
- *K* is an input to k-means clustering, so this project had to find a good value for k
- See the report for an different way of finding matching neighborhoods

Finding the best value for k

- Horizontal axis: number of clusters (*k*)
- Vertical axis: number of clusters containing postal codes in both cities
- Goal: most clusters containing postal codes from **both** cities
- As k goes up, the average number of postal codes per cluster goes down
- Goal: highest number of postal codes per cluster, subject to previous goal



Finding the best value for *k*

- When *k* is 9 or or 15, number of clusters containing postal codes from both cities is 3.
- When *k* is 15, the average number of postal codes per cluster is smaller than when *k* is 9
- Therefore, use *k* = 9
  - 9 clusters

Results of Clustering: First Cluster

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

Results of Clustering: Second Cluster

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

Results of Clustering: Third Cluster

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

Results of Clustering: Remaining Clusters

- Remaining clusters had one postal code per cluster
- See report for details
- See discussion section for interesting observation about 2 clusters

DISCUSSION

What makes 2 postal codes similar?

What characterizes each cluster?

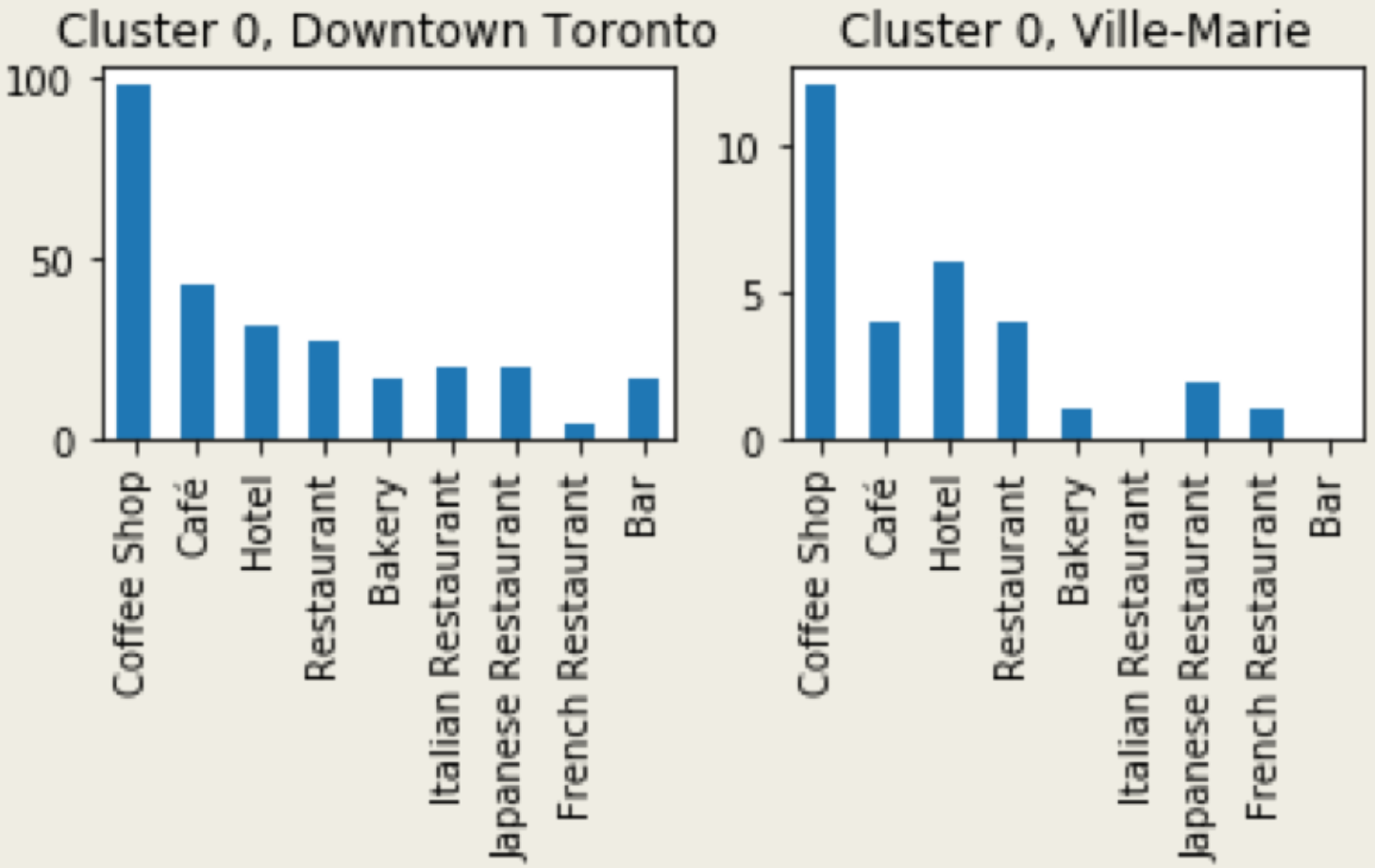
- Clustering is based on frequencies of venues
- Want to show the counts of the most common categories of venues in the different clusters
- For singleton clusters, show the counts of the venue categories

The 9 most frequently occurring venue categories:

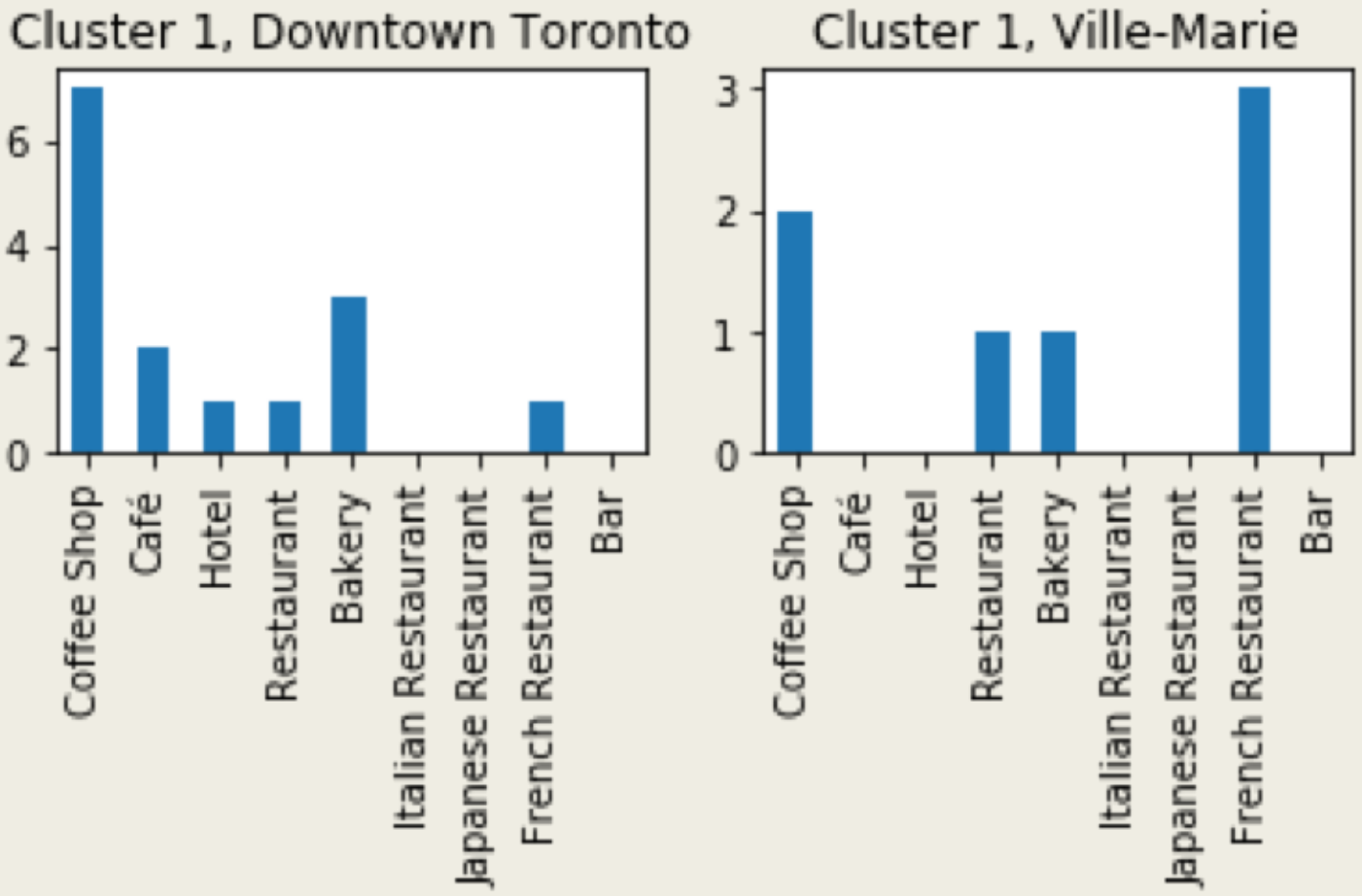
Category	Count
Coffee-Shop	166
Café	99
Hotel	74
Restaurant	54
Bakery	46
Italian-Restaurant	39
Japanese-Restaurant	36
French-Restaurant	33
Bar	32

Use these categories to show how clusters differ from each other

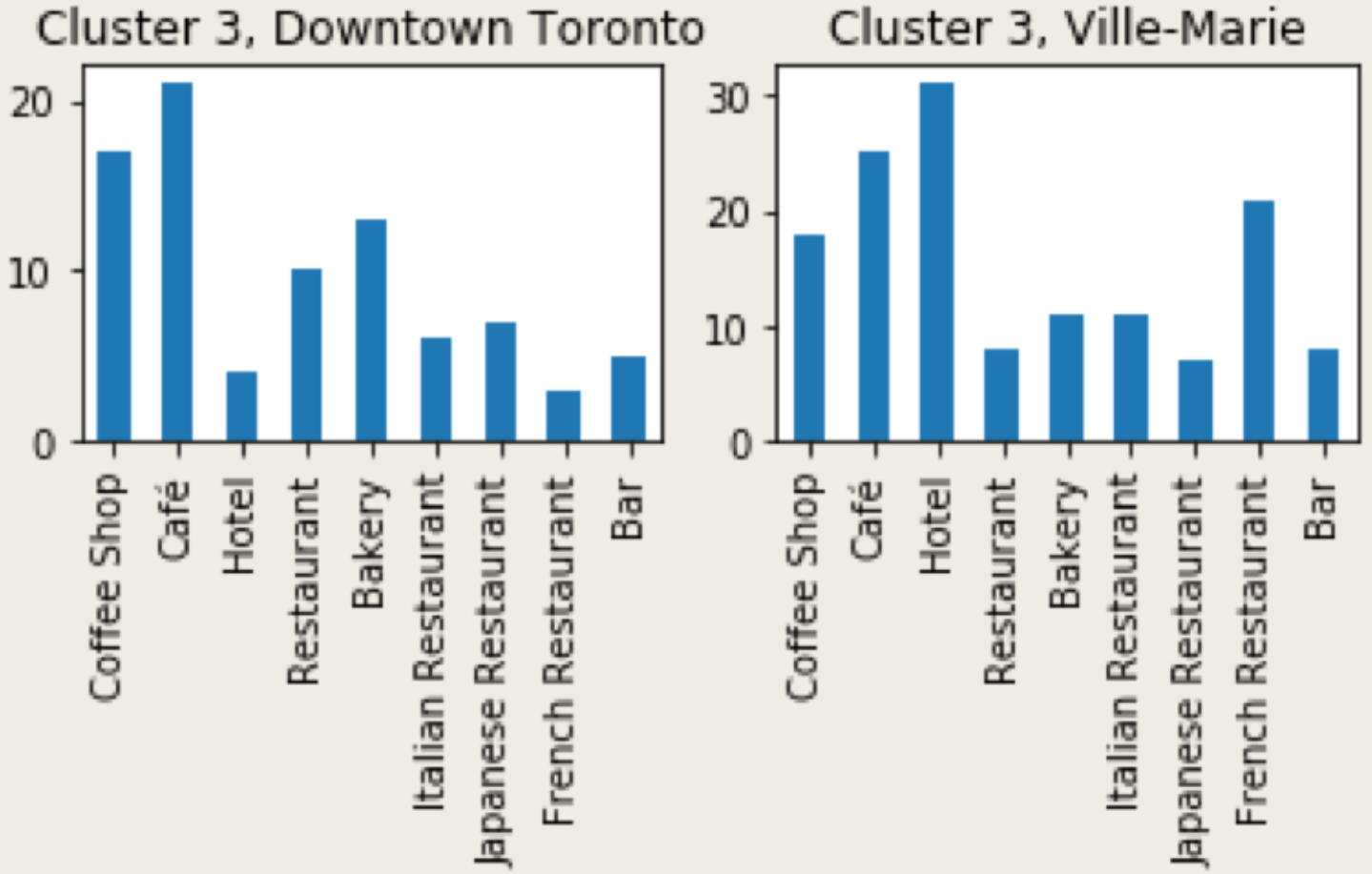
First cluster



Second Cluster



Third Cluster



Fourth Cluster

Borough: Downtown Toronto

Neighborhood(s): Rosedale

This postal code only contains 3 venue categories.

Venue Category	Count
Park	2
Playground	1
Trail	1

Fifth Cluster

Borough: Ville Marie

Neighborhood(s): Downtown Montreal Southwest, Shaughnessy Village, part of Mount Royal Park

This postal code only contains 4 venue categories.

Venue Category	Count
Bus Stop	1
Historic Site	1
Lake	1
Mountain	1

- This cluster is in Montreal
- The previous cluster is in Toronto
- Although the clustering algorithm did not match the neighborhoods in this cluster and the previous cluster, they are both characterized by outdoor venues, so they are similar.
- **Future project:** explore the use of higher-level venue categories, like “outdoor”, “restaurant”, and “store”, and cluster based on higher-level categories.
  - Additional benefit: Using higher-level categories would avoid the curse of dimensionality

Sixth Cluster

Borough: Downtown Toronto

Neighborhood(s): Queen's Park

This postal code contains more than 20 venue categories. Although the venues are dominated by coffee shops, the remaining venues are in categories that aren't even in the top 9 categories.

Venue Category	Count
Coffee Shop	11
Gym	2
Park	2
Arts & Crafts Store	1
Bar	1
Beer Bar	1
Burger Joint	1
Burrito Place	1
Café	1
Chinese Restaurant	1
College Auditorium	1
Creperie	1
Diner	1
Fast Food Restaurant	1
Fried Chicken Joint	1
Hobby Shop	1
Italian Restaurant	1
Mexican Restaurant	1
Music Venue	1
Nightclub	1
Portuguese Restaurant	1
Sandwich Place	1
Smoothie Shop	1
Sushi Restaurant	1
Theater	1

Seventh Cluster

Borough: Downtown Toronto

Neighborhood(s): CN Tower, Bathurst Quay, Island airport, Harbourfront West, King and Spadina, Railway Lands, South Niagara

This cluster is dominated by venues found at airports. Ville Marie doesn't have an airport, so that fact that this cluster is in only one borough isn't surprising. Marketing to businesses in this postal code would probably be best when Montreal's airport is in the scope of a (future) project.

Venue Category	Count
Airport Service	3
Airport Lounge	2
Airport Terminal	2
Airport	1
Airport Food Court	1
Bar	1
Boat or Ferry	1
Boutique	1
Harbor / Marina	1
Plane	1
Rental Car Location	1
Sculpture Garden	1

Eighth Cluster

Borough: Ville Marie

Neighborhood(s): Centre-Sud South, Gay Village

No coffee shops. No cafes. The result might be different if coffee shops and breakfast spots were in the same higher-level group

Venue Category	Count
Breakfast Spot	2
Restaurant	2
Sushi Restaurant	2
Asian Restaurant	1
Beer Bar	1
Bike Rental / Bike Share	1
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
Supermarket	1
Thai Restaurant	1

Ninth Cluster

Borough: Downtown Toronto

Neighborhood(s): Christie

Let's look at the top 9 categories: 3 Cafes. Only 1 coffee shop. No hotels. Only 1 "Restaurant". No Bakeries. Only 1 Italian Restaurant. No Japanese Restaurants. No French Restaurants. No Bars. No gyms.

This isn't like the first cluster because there are no coffee shops. It isn't like the second cluster because there are no coffee shops and no French restaurants. It isn't like the 3rd cluster because it does not have many venues in the top 9 categories.

Venue Category	Count
Café	3
Grocery Store	3
Park	2
Athletics & Sports	1
Baby Store	1
Candy Store	1
Coffee Shop	1
Convenience Store	1
Diner	1
Italian Restaurant	1
Nightclub	1
Restaurant	1

CONCLUSION

Results

- k-means clustering 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.
- Use of k-means to cluster neighborhoods in the two cities is very useful for finding neighborhoods in Ville Marie that are like neighborhoods in Downtown Toronto.

Opportunity

- Discussion section shows that there is room for improvement
  - *Categories from FourSquare are generally very detailed*
- Algorithm failed to match two postal codes even though they contained a preponderance of outdoor venues, because they contained different **categories** of outdoor venues.
- Results could be improved by adding higher-level categories to the feature matrix, and either removing the corresponding low-level categories or leaving them in.
  - *Removing the low-level categories would reduce the dimensionality, and this is generally beneficial*

Summary

- This shows that matching up neighborhoods in Downtown Toronto with neighborhoods in Ville Marie using k-means clustering gives **excellent results**
- Montreal could use these results to design targeted campaigns to attract more business to Montreal
- Future project could expand the scope to all of Montreal instead of just one borough in Montreal
  - *Also expand scope to include all of Toronto*
  - *Could also expand scope to target other Cities both inside and outside of Canada*

THANK YOU

MERCI