Capstone Project - The Battle of Neighborhoods

Danilo Evangelista (*https://github.com/danilodn/*)

1. Notes

Capstone project report for the “Applied Data Science Capstone” course from the “IBM Data Science Professional Certificate” on Coursera (*https://www.coursera.org/professional-certificates/ibm-data-science*).

2. Introduction/Business Problem

Stakeholders want to open a new coffee shop branch either on Vancouver or Toronto and need the subsidiary information on the city/competitors to decide about the expanding strategy.

Their intention is to open one Café shop right away and elaborate a long-term strategy to allocate several stores over the next years in these two cities.

The plan should consider the competitors’ location and neighborhood characteristics.

3. Dataset

The first set of data used was the list of boroughs and neighborhoods of Toronto and Vancouver, extracted from Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_V), as seen in the following images:

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Languages

List of postal codes of Canada: M

From Wikipedia, the free encyclopedia

This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website,[1] via its applications for such smartphones as the iPhone and BlackBerry,[2] and sells hard-copy directories and CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries.

Toronto - 103 FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes should start with M0. However, the postal code M0R 8T0 is assigned to an Amazon warehouse in Mississauga, suggesting that Canada Post may have reserved the M0 FSA for high volume addresses.

Postal Code	Borough	Neighborhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue, Humber Valley Village
M1B	Scarborough	Malvern, Rouge
M2B	Not assigned	Not assigned

Toronto’s list of Postal Cods, Boroughs and Neighborhoods

Wikipedia
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Languages
Français
Edit links

List of postal codes of Canada: V

From Wikipedia, the free encyclopedia

This is a list of postal codes in Canada where the first letter is V. Postal codes beginning with V are located within the Canadian province of British Columbia. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website,[1] via its mobile apps for such smartphones as the iPhone and BlackBerry,[2] and sells hard-copy directories and CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries.

British Columbia - 192 FSAs [edit]

Urban [edit]

V1A Kimberley	**V2A** Penticton	**V3A** Langley Township (Langley City)	**V4A** Surrey Southwest	**V5A** Burnaby (Government Road / Lake City / SFU / Burnaby Mountain)	**V6A** Vancouver (Strathcona / Chinatown / Downtown Eastside)	**V7A** Richmond South	**V8A** Powell River	**V9A** Victoria (Vic West / Esquimalt) Canadian Forces (MARPAC)
V1B Vernon East	**V2B** Kamloops Northwest	**V3B** Port Coquitlam Central	**V4B** White Rock	**V5B** Burnaby (Parkcrest-Aubrey / Ardingley-Sprott)	**V6B** Vancouver (NE Downtown / Gastown / Harbour Centre / International Village / Victory Square / Yaletown)	**V7B** Richmond (Sea Island / YVR)	**V8B** Squamish	**V9B** Victoria (West Highlands / North Langford / View Royal)
V1C Cranbrook	**V2C** Kamloops Central and Southeast	**V3C** Port Coquitlam South	**V4C** Delta Northeast	**V5C** Burnaby (Burnaby Heights / Willingdon Heights / West Central Valley)	**V6C** Vancouver (Waterfront / Coal Harbour / Canada Place)	**V7C** Richmond Northwest	**V8C** Kitimat	**V9C** Victoria (Colwood / South Langford / Metchosin)

Vancouver's list of Postal Cods, Boroughs and Neighborhoods

That way, after data preparation and cleaning we could get 29 different postal code neighborhoods from Vancouver and 102 from Toronto.

We also retrieved the list of venues for both cities from Foursquare API, each limited to 100 occurrences from 1000 meters of each neighborhood centroid, which resulted in 1772 different category venues for Vancouver and 4951 for Toronto.

3.1. Data Acquisition and cleaning

From Wikipedia we managed to get data through Pandas method "pd.read_html".

In the case of Vancouver, the way the original table was organized, made the DataFrame to be a little messy and obliged us to do some wrangling, as we can see below:

```
In [237]: link_V = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_V"
          df_V = pd.read_html(link_V)[0]
          df_V.head()
```

Out[237]:

	0	1	2	3	4	5	
0	V1AKimberley	V2APenticton	V3ALangley Township(Langley City)	V4ASurreySouthwest	V5ABurnaby(Government Road / Lake City / SFU /...	V6AVancouver(Strathcona / Chinatown / Downtown...	V7ARichmor
1	V1BVernonEast	V2BKamloopsNorthwest	V3BPort CoquitlamCentral	V4BWhite Rock	V5BBurnaby(Parkcrest-Aubrey / Ardingley-Sprott)	V6BVancouver(NE Downtown / Gastown / Harbour C...	V7BRichmor Island / YVR
2	V1CCranbrook	V2CKamloopsCentral and Southeast	V3CPort CoquitlamSouth	V4CDeltaNortheast	V5CBurnaby(Burnaby Heights / Willingdon Height...	V6CVancouver(Waterfront / Coal Harbour / Canad...	V7CRichmor
3	V1ESalmon Arm	V2EKamloopsSouth and West	V3ECoquitlamNorth	V4EDeltaEast	V5EBurnaby(Lakeview-Mayfield / Richmond Park /...	V6EVancouver(SE West End / Davie Village)	V7ERichmor

After that, we used Pgeocode to get the coordinates for each postal code/neighborhood, resulting in the following DataFrame for each city:

[201]:

	PostalCode	Neighborhood	latitude	longitude
0	M3A	North York (York Heights / Victoria Village / ...	43.7545	-79.3300
1	M4A	North York (Sweeney Park / Wigmore Park)	43.7276	-79.3148
2	M5A	Downtown Toronto (Regent Park / Port of Toronto)	43.6555	-79.3626
3	M6A	North York (Lawrence Manor / Lawrence Heights)	43.7223	-79.4504
4	M7A	Queen's Park Ontario Provincial Government	43.6641	-79.3889

Then we run Foursquare API, getting the top 100 venues from 1000 meters from each postal code/neighborhood centroid coordinate, along with its name, category, and coordinates, resulting in another DataFrame for each city:

[178]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North York (York Heights / Victoria Village / ...	43.7545	-79.33	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	North York (York Heights / Victoria Village / ...	43.7545	-79.33	Brookbanks Park	43.751976	-79.332140	Park
2	North York (York Heights / Victoria Village / ...	43.7545	-79.33	Tim Hortons	43.760668	-79.326368	Café
3	North York (York Heights / Victoria Village / ...	43.7545	-79.33	A&W	43.760643	-79.326865	Fast Food Restaurant
4	North York (York Heights / Victoria Village / ...	43.7545	-79.33	Shoppers Drug Mart	43.760857	-79.324961	Pharmacy

As our focus was Coffee Shops and Cafés, we created one more DataFrame with only these venue types.

The Toronto DataFrames resulted in 4951 rows (all venues) and 601 rows (coffee shops and cafés) and the Vancouver resulted in 1779 rows (all venues) and 162 rows (coffee shops and cafés).

```
[202]: print(toronto_venues.shape)
       print(toronto_coffee.shape)

(4951, 7)
(601, 7)
```

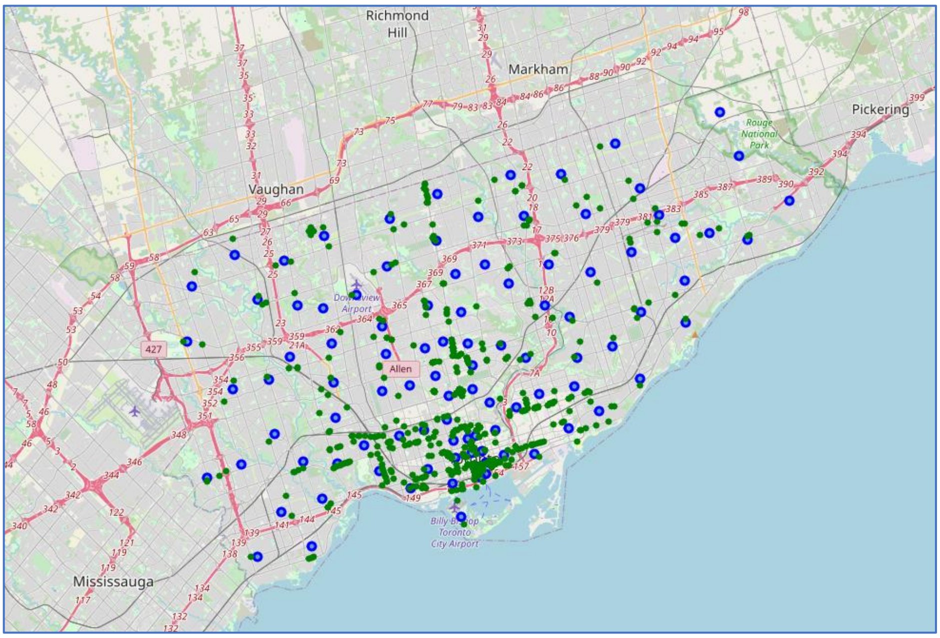
```
[203]: print(vancouver_venues.shape)
       print(vancouver_coffee.shape)

(1779, 7)
(162, 7)
```

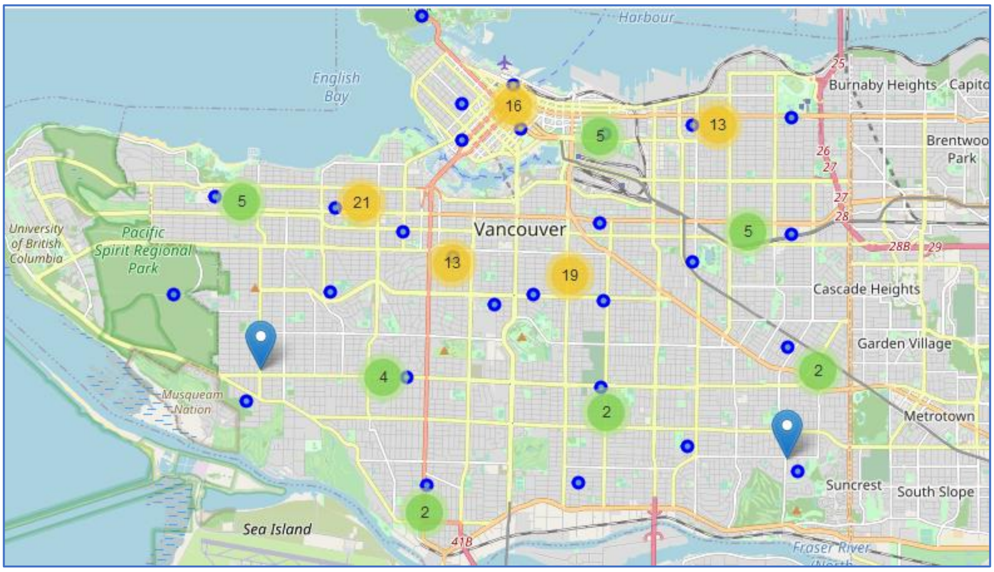
4. Methodology

With the DataFrames ready, first we used Python Folium to plot the neighborhoods in a map, to get a perspective of the cities and data.

Then we managed to superimpose the coffee shops and cafés in the same maps, to get a grip of the distribution of these venues over the cities, as we can see:



Also, we created a dynamic map using MarkerCluster, from Folium, which expands the venues when we zoom in the map and concentrate it when we zoom out, showing the number of venues per region. For example:



4.1. Clustering

We used the K-means clustering Classification algorithm, done with Python scikit-learn, to classify the neighborhoods in clusters of similarity or dissimilarity, to understand their characteristics, based on the venues populated there.

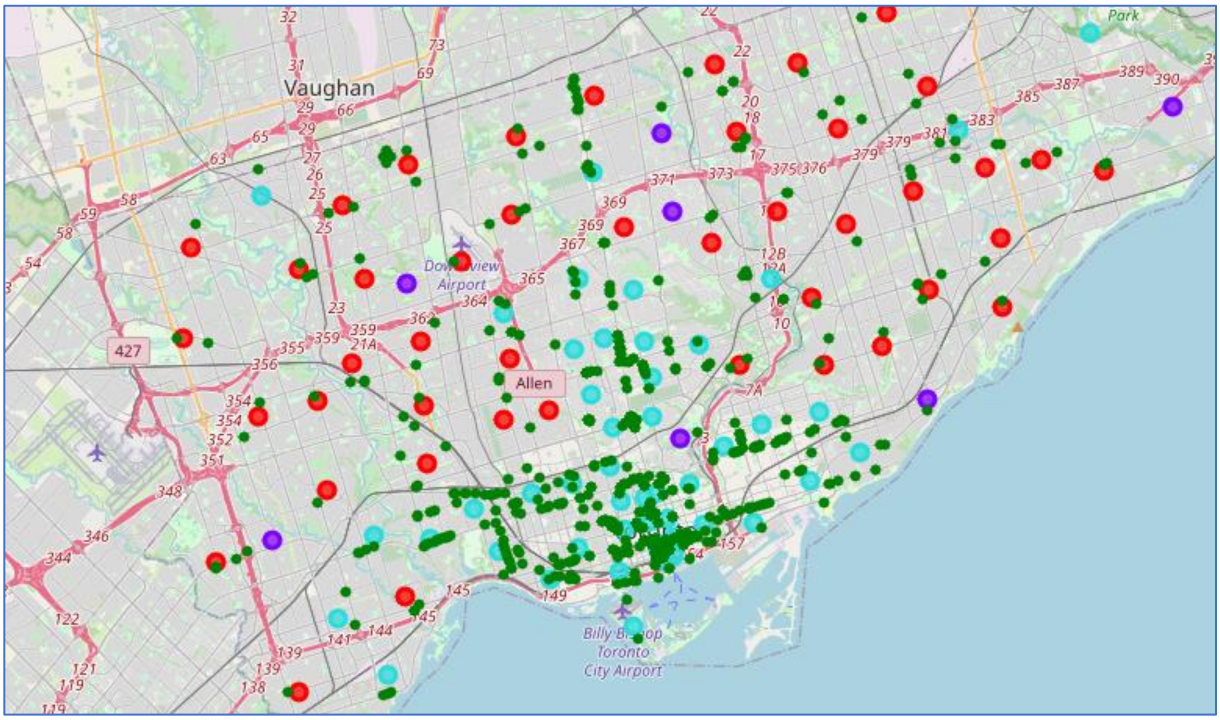
To do that, we used one-hot encoding to get dummy data and grouped neighborhoods, getting the top 10 venues for each neighborhood, for each city.

[124]:

	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	Central Toronto (Davisville North)	Coffee Shop	Italian Restaurant	Pizza Place	Café	Park	Sushi Restaurant	Restaurant	Yoga Studio	Diner	Pub
1	Central Toronto (Davisville)	Sushi Restaurant	Italian Restaurant	Pizza Place	Indian Restaurant	Café	Coffee Shop	Bakery	Sandwich Place	Bank	Bar
2	Central Toronto (Forest Hill North & West)	Park	Coffee Shop	Café	Bank	Sushi Restaurant	Italian Restaurant	Pharmacy	Gym / Fitness Center	Trail	Bagel Shop
3	Central Toronto (Lawrence Park East)	Sushi Restaurant	Italian Restaurant	Coffee Shop	Bakery	Bus Line	Café	Fast Food Restaurant	Pub	Bank	Asian Restaurant
4	Central Toronto (Moore Park / Summerhill East)	Coffee Shop	Sushi Restaurant	Italian Restaurant	Grocery Store	Thai Restaurant	Gym	Park	Gastropub	Spa	Bank

Trying some values for k, the best values were 3 for Toronto and 4 for Vancouver, grouping this way the neighborhoods according to its venue's categories.

Then, we created maps with the clustered neighborhoods and finally maps with the clusters and coffee/café's venues superimposed on it (green dots), which gave a better sense of all data together.

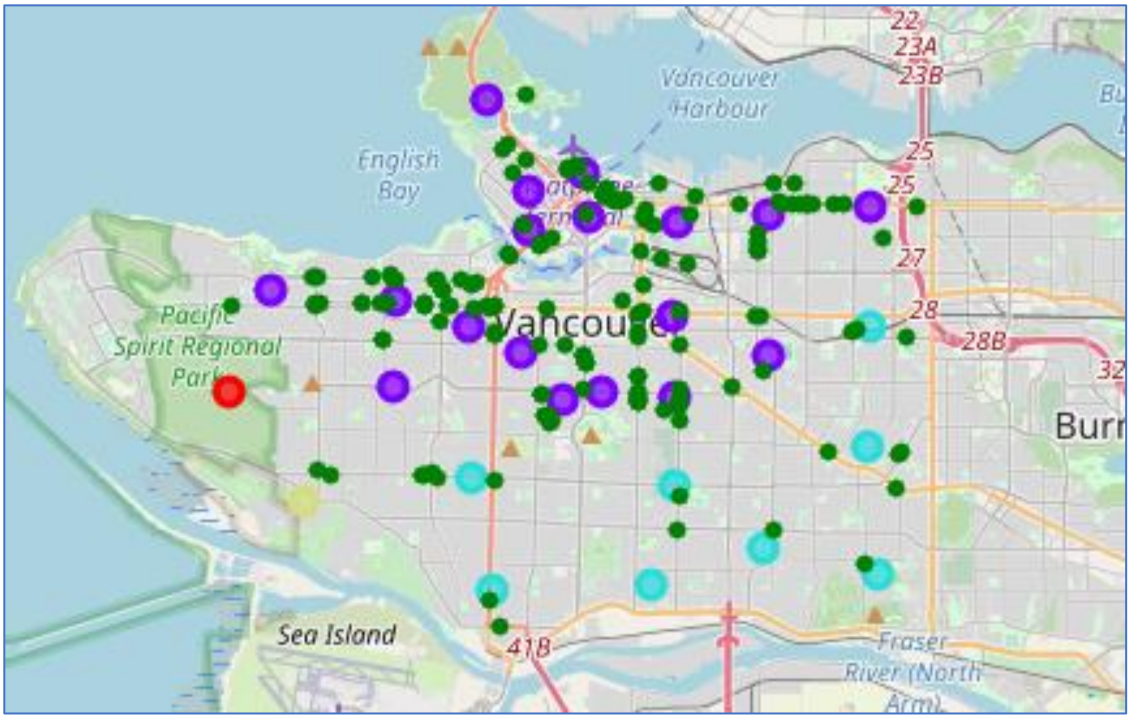


Resulting clusters for Toronto are:

- 0 (red): residence neighborhoods, predominantly with bus stops, parks, grocery stores, pharmacies, and banks, along a few restaurants, malls, and cafés.
- 1 (purple): a few sparse neighborhoods with distinct venues, as parks, zoo, soccer field, baseball field, golf course, theater, etc.
- 2 (cyan): neighbors with downtown characteristics with a lot of coffee shops/cafés, restaurants, and hotels.

Vancouver:

- 0 (red): one neighborhood situating the Pacific Spirit Regional Park.
- 1 (purple): neighbors with downtown characteristics with a lot of coffee shops/cafés, restaurants, and a few hotels (less than Toronto).
- 2 (cyan): residence neighborhoods, predominantly with bus stops, parks, restaurants, and other sparse venues.
- 3 (yellow): one neighborhood situating a golf course.



5. Results

From this Project we managed to understand Toronto and Vancouver neighborhood's characteristics, according to their venue vocation, and the distribution of coffee shops/cafés in the cities.

We classified satisfactorily the neighborhoods in clusters of similarity based on its venues and venue categories.

Also provided dynamic maps to the stakeholders, making the results visually appealing and easily understandable, making it very appropriate for the business decision making.

Must we say that this project can be used in the future too, updating the status of the neighborhoods and its venues at any time needed.

6. Discussion

The scope of this report could be extended by researching population of neighborhoods to be used together with the number of venues for each one.

It could have considered individual income from each city too, but it escapes the predetermined objective.

Also, we could have done some histograms, bar charts, area plots, etc. to help in the visualization of the information.

7. Conclusion

It is an open field and Foursquare API can be used for an infinite number of applications, since government policies to businesses decision to individual interests.

The bigger challenges are to get enough good data to use, and have the time to explore everything, but we must have in mind the objective, without escaping focus of the project.