Clustering Food Venues in Canadian Cities

Author: Lawrence Lee

Date: September, 2019

For IBM Data Science Professional Certificate

Table of Contents

Introduction

Background

Food venues are one of the most numerous and most diverse businesses in any major City. Following economic recessions, competition among these small businesses is fierce. Information on where different types of food restaurants are aggregated in major Canadian cities may help business owners understand both who their competition is but also where local population demographics may favor certain restaurant types.

Problem

How do geographic areas within a city and between cities compare, with regard to concentrations of different categories of food businesses?

Audience

This project is aimed at restauranteurs, realtors, small business owners, or even consumers, who might benefit from more insight into the distribution of different types of food businesses between postal code locations and cities.

Data

Canadian Postal Codes

I examined the three largest Canadian municipalities[1], using postal code lists collected on Wikipedia[2]. These tables were scraped using pandas read_html and reshaped as needed.

Foursquare Location Data

Foursquare's location dataset is one of the most comprehensive available. Built from 13+ billion crowd-sourced "check-ins" since 2009, the data is used by over 150,000 developers such as Apple, Samsung, Twitter, and Uber.[3] Foursquare's data is freely accessible via its Places API. The API's Search endpoint returns up to 50 venues within a radius of a location, including basic venue data such as name, categories, and location.

Example

Taking Toronto for example, I can call the Foursquare API's Search endpoint to obtain JSON containing the names and categories of three restaurants in Montreal, Quebec. I can then extract the relevant data from the JSON into a `DataFrame` like the one below.

Table 1

Sample dataframe of Foursquare venue data

Place	Place Latitude	Place Longitude	Name	Latitude	Longitude	Category
H0M	45.6986	-73.5025	Tenuta	45.694548	-73.509593	Italian Restaurant
H0M	45.6986	-73.5025	Dairy Queen Store	45.696467	-73.492754	Fast Food Restaurant
H0M	45.6986	-73.5025	Tim Hortons	45.690808	-73.496677	Coffee Shop

Note that the Foursquare geocoder can geocode using just the first three characters of the postal code (forward sortation area, or FSA). It can thus be used to obtain the coordinates of each FSA for further analysis.

Method

Overview

All data collection and analysis were performed in Jupyter Notebook running Python v.3.7.3.

After scraping 261 FSAs from Wikipedia, I used them to iteratively query the Foursquare Search endpoint to return 50 venues of the umbrella category 'Food' within 1 km of each postal code. This resulted in a dataset of 5798 venues distributed throughout each city, to represent the characteristics of each postal code area and each city.

I used Foursquare's categories hierarchy to further categorize each venue in broader categories. (E.g. 'Sushi Restaurant' may also be categorized as 'Japanese Restaurant' and 'Asian Restaurant'.)

For analysis, I used one-hot encoding on the categories, followed by k-means clustering machine learning to cluster the FSAs according to their characteristic food categories. The top categories per FSA, per city and per cluster were then determined, as well as the breakdown of clusters and categories between cities.

I then created map visuals of the postal codes, venues and clusters.

Modules

The python modules imported for use are listed below.

Table 2

Python Modules

Core Modules	Debugging Modules
numpy	time
pandas	dill
collections	
requests	
geopy.geocoders.Nominatim	
matplotlib	
sklearn.cluster.KMeans	
folium	

Postal Code Data

I scraped the following Wikipedia tables of postal codes in Canada:

1. Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. Montreal: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_H
3. Calgary: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T

Web scraping was achieved using:

```
html = requests.get(url).content

df = pd.read_html(html)
```

Tables were re-structured to obtain a list of 261 FSAs.

Inapplicable FSAs were excluded such as those designated ‘Not Assigned’. The letter ‘T’ FSAs include all of the province of Alberta. Cities other than Calgary were filtered out.

Foursquare Data

I used a function to iteratively call the Foursquare API with a search of the 50 closest venues within a radius 1km of each postal code. The Foursquare Search endpoint returns the nearest venues to the point of origin, matching the search criteria: the ‘Food’ top level category id. For simplicity I only included the primary category of each venue, as some venues have multiple categorizations, but this is rare.

The following data elements were extracted from the JSON into a Pandas DataFrame:

```
response['geocode']['feature']['geometry']['center']['lat']

response['geocode']['feature']['geometry']['center']['lng']

For v in requests.get(url).json()['response']['venues']:

       v['id'],

       v['name'],

       v['location']['lat'],

       v['location']['lng'],

       v['categories'][0]['name'], # primary category only

       v['categories'][0]['id']
```

Similarly, the Foursquare categories hierarchy was extracted from JSON by looping through each hierarchy level to create a DataFrame with additional rows for each parent category, as below.

Table 3

Sample pivoted category hierarchy. Note the last two items correspond to the same child category id.

Category Id	Categories
4d4b7105d754a06374d81259	Food
503288ae91d4c4b30a586d67	Afghan Restaurant
4bf58dd8d48988d1c8941735	African Restaurant
4bf58dd8d48988d10a941735	African Restaurant
4bf58dd8d48988d10a941735	Ethiopian Restaurant

The ‘Coffee Shop’ and ‘Café’ categories were excluded as they dominated the data at 19% of total venues. The remaining categories were merged to the original venues data.

Analysis

Categories of each venue were one-hot encoded to produce a wide DataFrame counting the applicable categories for each venue.

```
pd.get_dummies(df[['Categories']])
```

After dropping venues with category mismatches, the final count of unique venues was 5798.

The encoded data was then grouped by FSA, calculating the mean, as the proportion of nearby venues that matched each category.

```
df.groupby('Place').mean()
```

KMeans Clustering

These values were used for KMeans Cluster analysis. Clustering is a machine learning technique to segment data points into mutually exclusive clusters. It is an unsupervised technique, meaning pre-existing labels are not provided, rather the inferences are made based on data similarities. K-means is a partition-based clustering algorithm which produces sphere-like clusters and is relatively efficient for medium and large datasets.

```
KMeans( n_clusters=5).fit( df.iloc[:,1:]
```

The cluster labels were then merged with the previous data to generate several summary tables breaking down the clusters, cities and FSAs with their most frequent food categories, as below.

Table 4

Sample of FSAs with location, cluster number and most frequent foods

City	Place	Place Latitude	Place Longitude	Cluster	1st Most Frequent	2nd Most Frequent	3rd Most Frequent	4th Most Frequent	5th Most Frequent
Montreal, Quebec	H1Y	45.5486	-73.5788	0	Fast Food Restaurant	Asian Restaurant	Breakfast Spot	Italian Restaurant	Pizza Place
Toronto, Ontario	M9R	43.6898	-79.5582	0	Pizza Place	Asian Restaurant	Sandwich Place	Chinese Restaurant	American Restaurant
Montreal, Quebec	H0M	45.6986	-73.5025	0	Fast Food Restaurant	Italian Restaurant	Xinjiang Restaurant	Falafel Restaurant	Food Court
Montreal, Quebec	H1E	45.6342	-73.5842	0	Italian Restaurant	Fast Food Restaurant	Pizza Place	Restaurant	Diner
Montreal, Quebec	H1G	45.6109	-73.6211	0	Fast Food Restaurant	Asian Restaurant	Restaurant	Sandwich Place	Breakfast Spot
…									

Mapping

The geopy nominatim module was used to obtain coordinates of each city.

```
Nominatim(user_agent="explorer") .geocode(city)
```

Maps were generated using the Folium module. By looping through each city, the coordinates were used to centre the map and corresponding dataframes used to build map markers. In this way I generated maps Toronto, Montreal and Calgary display postal codes alone, venues alone, and postal codes coded by KMeans Cluster.

```
folium.Map(location=[c_lat, c_lng])
```

```
folium.CircleMarker([p_lat, p_lng]) ).add_to(map)
```

To display the large amount of venue markers, the FastMarkerCluster plugin was used.

```
map.add_child(FastMarkerCluster(marker_data.values.tolist()))
```

Color coding was assigned using the Matplotlib qualitative colormap ‘Set1’.

```
matplotlib.cm.Set1(np.linspace(0, 1, 9))
```

A legend for the colormap was also added in html.

```
'<table><tr><td>Cluster {}</td><td><i class="fa fa-circle"
style="color:{}"></i></td>'.format(item, color)
```

```
map.get_root().html.add_child(folium.Element(legend_html))
```

Results

Reviewing the 3 city maps of the 261 FSAs, below, it is obvious that the distribution of FSAs is less dense in Calgary.

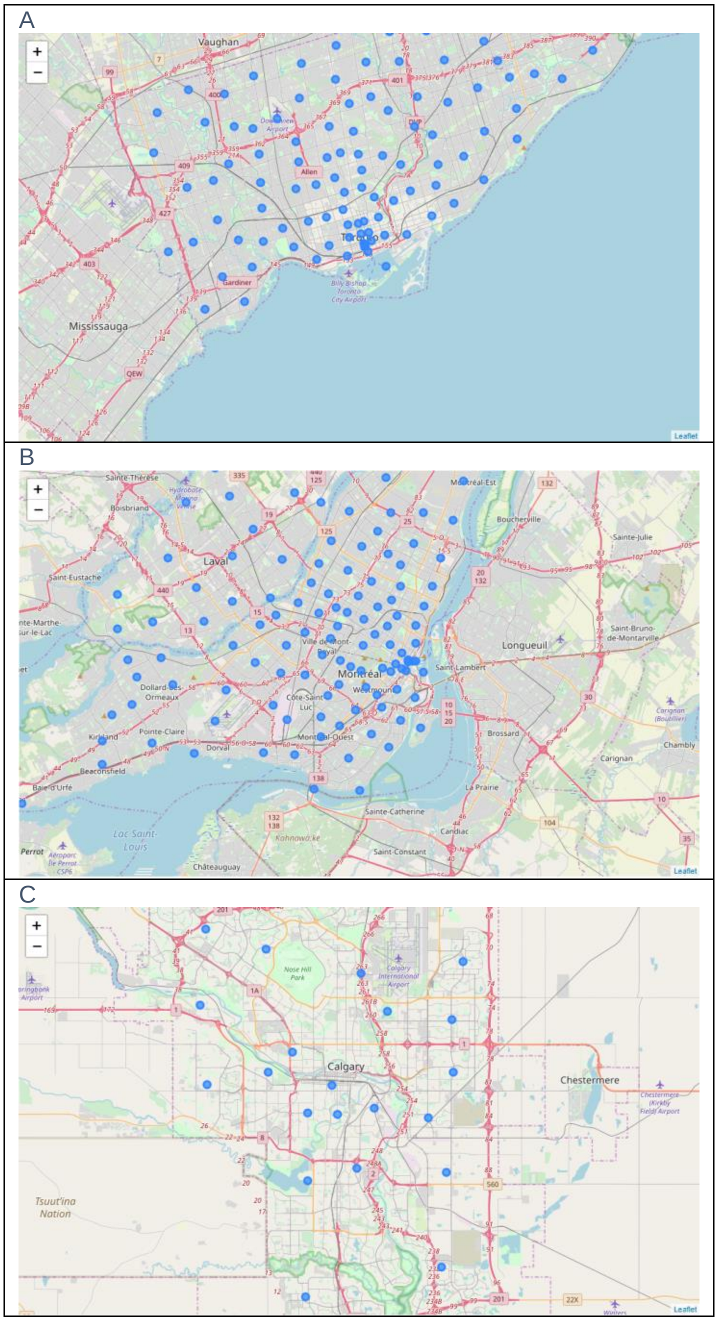

Figure 1. Maps of FSAs in A) Toronto, B) Montreal and C) Calgary.

This trend is also evident in the distribution of the 5798 venues, below. Only the downtown core of Calgary has a density of food venues resembling most of Toronto and Montreal.

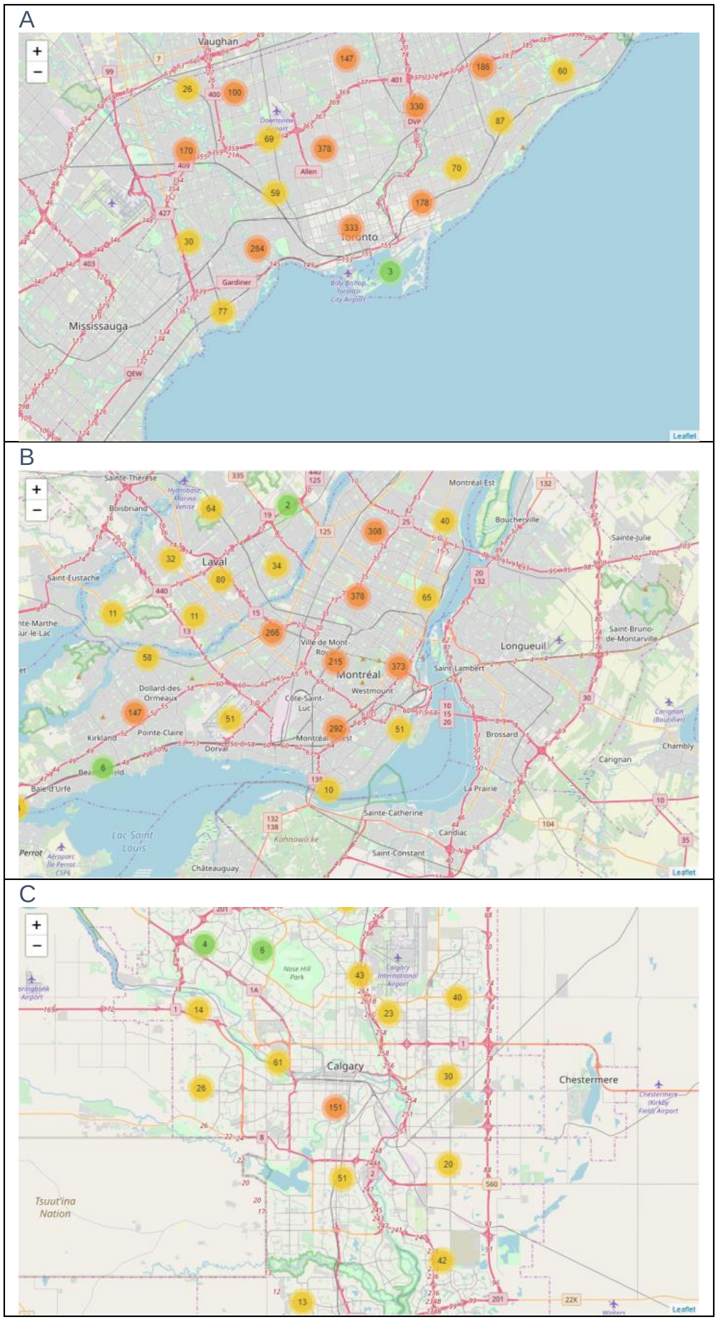

Figure 2. Maps of grouped venues in A) Toronto, B) Montreal and C) Calgary.

Reviewing the cluster data below, we can see that all five clusters had Asian restaurants as the most frequent food venue for the average FSA region. These are followed by pizza places, fast food and the overall restaurant category. However, once we get into the 4th and 5th most frequent categories differences in clusters emerge.

It's also notable that cluster 0 and 1 were mostly in Montreal (66%, 79% repectively) and infrequent in Toronto (13%, 7%). These two clusters had more dessert shops as the 5th and 4th most frequent food.

Table 5

Clusters with FSAs per city and most frequent foods

Cluster	Calgary, Alberta	Montreal, Quebec	Toronto, Ontario	1st Most Frequent	2nd Most Frequent	3rd Most Frequent	4th Most Frequent	5th Most Frequent
0	7	21	4	Asian Restaurant	Pizza Place	Fast Food Restaurant	Restaurant	Dessert Shop
1	2	11	1	Asian Restaurant	Fast Food Restaurant	Pizza Place	Dessert Shop	Restaurant
2	7	45	61	Asian Restaurant	Pizza Place	Fast Food Restaurant	Restaurant	Bakery
3	2	2	4	Asian Restaurant	Fast Food Restaurant	Restaurant	Japanese Restaurant	Italian Restaurant
4	7	36	31	Asian Restaurant	Fast Food Restaurant	Restaurant	Pizza Place	Japanese Restaurant

When comparing cities, we can see that again, all three cities have Asian restaurant as the most frequent, followed by pizza place and fast food. The 5th most frequent shows that Vietnamese restaurants are common in Calgary, Dessert shops in Montreal and Japanese restaurants in Toronto.

We see again that Toronto has relatively few FSAs of cluster 0 and 1, but a high proportion of Cluster 2, which has more Bakeries.

Table 6

Cities with FSAs per cluster and most frequent foods

City	Clus0	Clus1	Clus2	Clus3	Clus4	1st Most Frequent	2nd Most Frequent	3rd Most Frequent	4th Most Frequent	5th Most Frequent
Calgary, Alberta	7	2	7	2	7	Asian Restaurant	Pizza Place	Fast Food Restaurant	Sandwich Place	Vietnamese Restaurant
Montreal, Quebec	21	11	45	2	36	Asian Restaurant	Fast Food Restaurant	Restaurant	Pizza Place	Dessert Shop
Toronto, Ontario	4	1	61	4	31	Asian Restaurant	Pizza Place	Fast Food Restaurant	Restaurant	Japanese Restaurant

Finally, reviewing the clustered map shows us that cluster 4, high in Japanese restaurants, is especially common in the downtown core in all 4 cities. Cluster 2, high in bakeries, is distributed throughout each city. Cluster 0, high in dessert shops, are more frequently on the outskirts of the city.

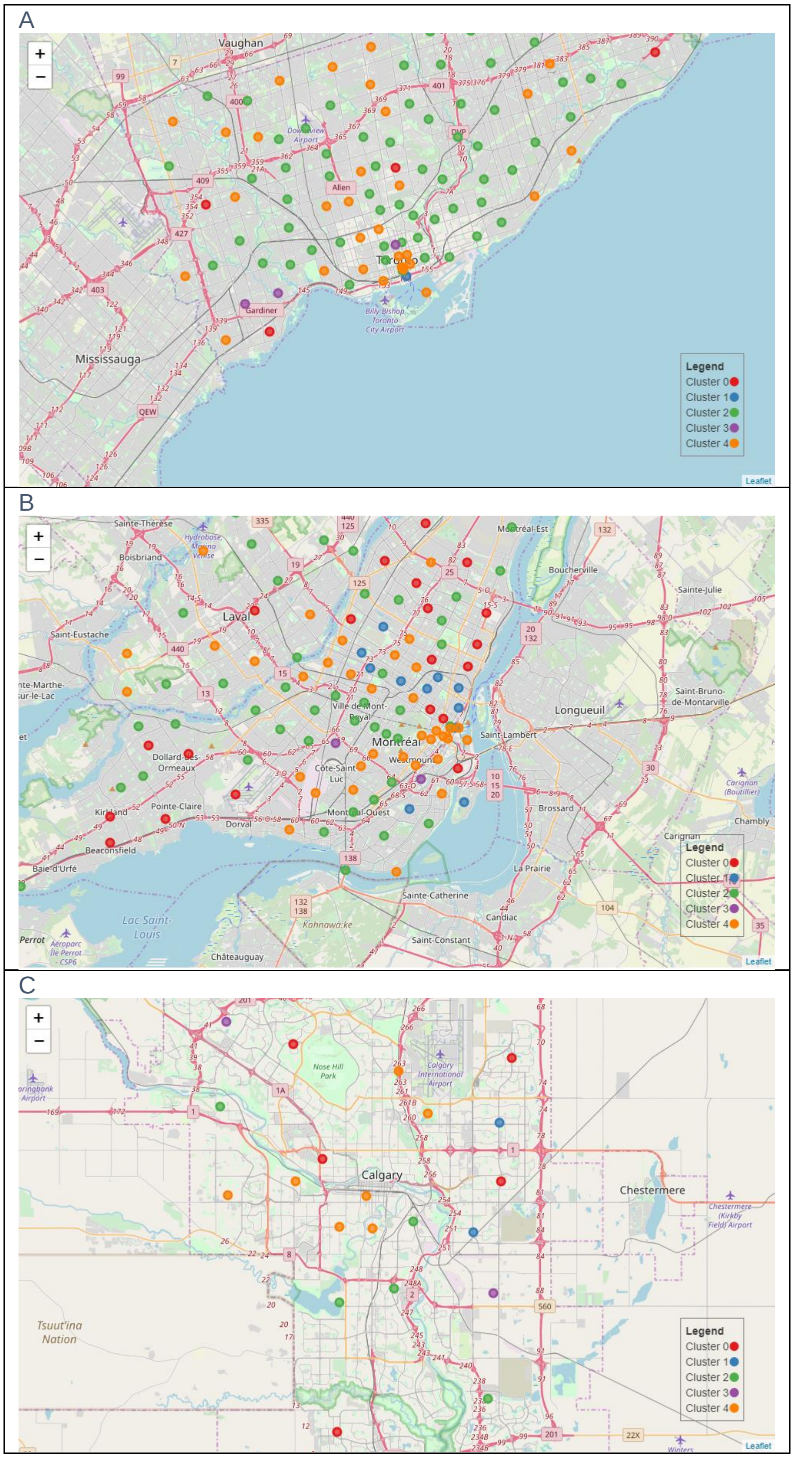

Figure 3. Maps of FSAs clustered by means of counts of nearby food category in A) Toronto, B) Montreal and C) Calgary.

Discussion

The low densities of FSAs in Calgary has some implications. First, it is likely due to Calgary being less densely populated, and having fewer FSAs spread out over a larger area, as shown below.

Table 7

Land area and population of Toronto, Montreal and Calgary per Wikipedia[1].

City	Land Area	Population (2016)
Toronto	630.2	2,731,571
Montreal	365.1	1,704,694
Calgary	825.3	1,239,220

Furthermore, only 25 of the 35 Calgary FSAs survived the data processing, which suggests that many Calgary suburbs may not have any food venues within 1km to analyse. As a result, our sample of FSA regions and associated venues may not be as accurate a representation of Calgary, compared to Toronto and Montreal. Future work would have to account for the disparity in FSA areas when comparing cities of different densities.

Asian restaurants, fast food and pizza were common across all clusters and cities. Asian restaurants however are a fairly broad umbrella category for many food subcategories.

Reviewing the clusters revealed some notable differences between Montreal and Toronto, despite the two cities being of similar size and density, and geographically close to each other. Clusters 0 and 1 were predominant in Montreal and revealed that Dessert shops may be relatively common in Montreal.

Each city also had a distinct 5th most frequent food venue near the FSA centers: Vietnamese in Calgary, dessert shops in Montreal, and Japanese in Toronto.

Conclusion

Despite the confound of differences in FSA density, concrete similarities and differences between clusters and cities emerged when analysing food venue categories near FSA centres.

This exploratory analysis paves the way for a closer look comparing Montreal and Toronto and which food businesses thrive in each of these two large Canadian cities.

References

1. https://en.wikipedia.org/wiki/List_of_the_100_largest_municipalities_in_Canada_by_population
2. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
3. https://foursquare.com/about