Large Cities, Weather and Venues in Canadian Provinces

By: Heba Chehade

March 19, 2020

Background

Initially, my aim was to study how the top venues checked-into change as weather changes in two of the largest cities in Canada, Toronto and Vancouver, that are known to differ in average temperature and precipitation levels. Having learned more about the foursquare API, I came to understand that extracting venue information over time using the Foursquare API was not possible. Instead, I would have to extract venues over a year and that is not realistic for a Data Science Capstone project intended for a short-term course.

Introduction

This small study aims to uncover the top 100 venues in each of the largest cities in the ten Canadian provinces in order to undertake cluster analysis on the venue data to uncover insights that could serve entrepreneurs, economic analysts and marketing strategists who need to understand popular choices in each of the cities, how they differ and how to inform respective start-up and/or growth strategies.

Data

Cities - here are ten Canadian provinces, with a capital city in each. While capital cities are often the largest cities, it is not always the case. This study aims to uncover venue insights from the largest cities in each of the provinces to aid comparison. The data below was obtained from the latest data within Statistics Canada.

Table 1 – Canadian Provinces, Capital Cities, Largest Cities and Population

Province	**Capital City**	**Largest City**	**Population in Largest City (2019, *2017)**
Newfoundland and Labrador	St. John's		212,433
Prince Edward Island	Charlottetown		74,541*
Nova Scotia	Halifax		440,348
New Brunswick	Fredricton		60,879*)
Quebec	Quebec City	Montreal	4,319,000
Ontario	Toronto		6,472,000
Manitoba	Winnipeg		844,566
Saskatchewan	Regina	Saskatoon	330,674
Alberta	Edmonton	Calgary	1,515,000

British Columbia	Victoria	Vancouver	2,691,000

Weather - In addition to population figures, the study required data on weather. Canada is a large country and while most cities in Canada undergo similar weather patterns, they do slightly vary east to west particularly in precipitation levels. Environment Canada provides historical climate data on various cities in Canada and for this study, I extracted daily data between Jan 2014 until Dec 2018 to have a better picture of weather patterns over time. As some larger cities had more than one weather monitoring station, I used data from monitoring stations that were in close proximity to each of the largest cities above. Data was extracted in csv format for use in the analysis.

Venues - The other source of data used to capture venue information in each of the cities was foursquare. Using a developer account and respective API credentials, Python code was written within Jupyter notebook. I extracted the Top 100 venues, where possible, in each of the cities within a 10 km radius from the latitude and longitude of the cities. 10 km was chosen because it is, in my opinion, the furthest distance that can be travelled by foot.

Methodology

The following section outlines the thinking behind the Python code and what I did to extract data and ultimately carry out a cluster analysis on the data captured.

Weather Data

The first section of the code is about importing the necessary libraries and weather related csv files needed for the analysis. After checking the shape of each of the csv files to ensure that they data seems correct, I looked at the main headings to identify which columns are needed for the analysis which finally included:

- Latitude
- Longitude
- Date
- Year
- Month
- Day
- Mean Temperature
- Total Precipitation

After organizing the data, I cleaned up the data by looking for and managing missing values. If both Temperature and Precipitation data was missing, a row was dropped as it would be meaningless to include. If temperature was missing, the preceding temperature was used instead. I didn't find any total precipitation data missing on its own. I subsequently visualized mean temperatures and precipitation to understand weather patterns in each of the ten cities.

Venue Data

After setting locations and client ID information as extracted from my developer account, I ran code to request venue information in each of the cities on March 14th 2020. One of the key aspects is the latitude and longitude information for each of the cities and using foursquare, all the ones extracted looked comparable to what is publicly stated for each online except for St Johns which was not specifically identified in Foursquare. As a result, I manually fed in the latitude and longitude

information for St. John's only. Whereas I was able to extract the top 100 venues in each of the cities, it was not the case in Charlottetown and St. Johns where 90 and 91 venues were extracted respectively.

Exploratory Analysis

To carry out the analysis, I looked at the top 100 venues by category and ran a cumulative sum analysis on it to see if there was any point where majority of venues were captured by top limited amount. I also looked for duplicated in the same place using names. I then looked at distances and compared them to the latitude and longitude of the cities to understand dispersion of venues around the city center and drew Folium maps accordingly.

Cluster Analysis

With venue information on hand and clean, I ran the K-Means algorithm after converting into a vector to enable analysis which generated three clusters as depicted in the results section below. It is important to note that while the number of clusters is recommended to generally be between 3 and 5, I felt 3 was more appropriate given a sample of 10 cities for such a data on the venue data

Results

Extracted from the code used to extract and analyze data for this study.

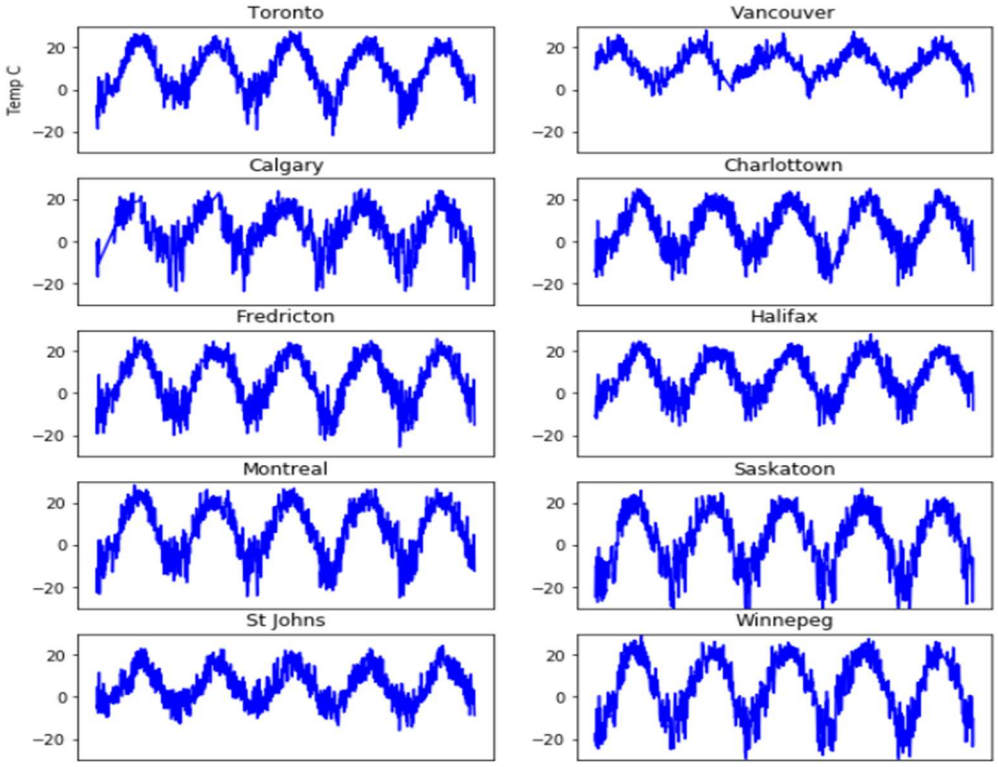
Weather Patterns per City

City	***Mean Temperature(C)***	***Min Temperature (C)***	***Max Temperature (C)***	***Mean Precipitation (mm)***	***Max* Precipitation (mm)***
Toronto	*9*	*-22*	*28*	*2*	*72*
Vancouver	*12*	*-4*	*28*	*5*	*100*
Calgary	*6*	*-24*	*25*	*1*	*41*
Charlottetown	*6*	*-21*	*25*	*3*	*93*
Fredricton	*6*	*-26*	*27*	*3*	*137*
Halifax	*8*	*-16*	*28*	*4*	*115*
Montreal	*7*	*-25*	*28*	*3*	*71*
Saskatoon	*3*	*-32*	*27*	*1*	*54*
St John's	*5*	*-16*	*24*	*4*	*72*
Winnipeg	*5*	*-29*	*29*	*1*	*49*

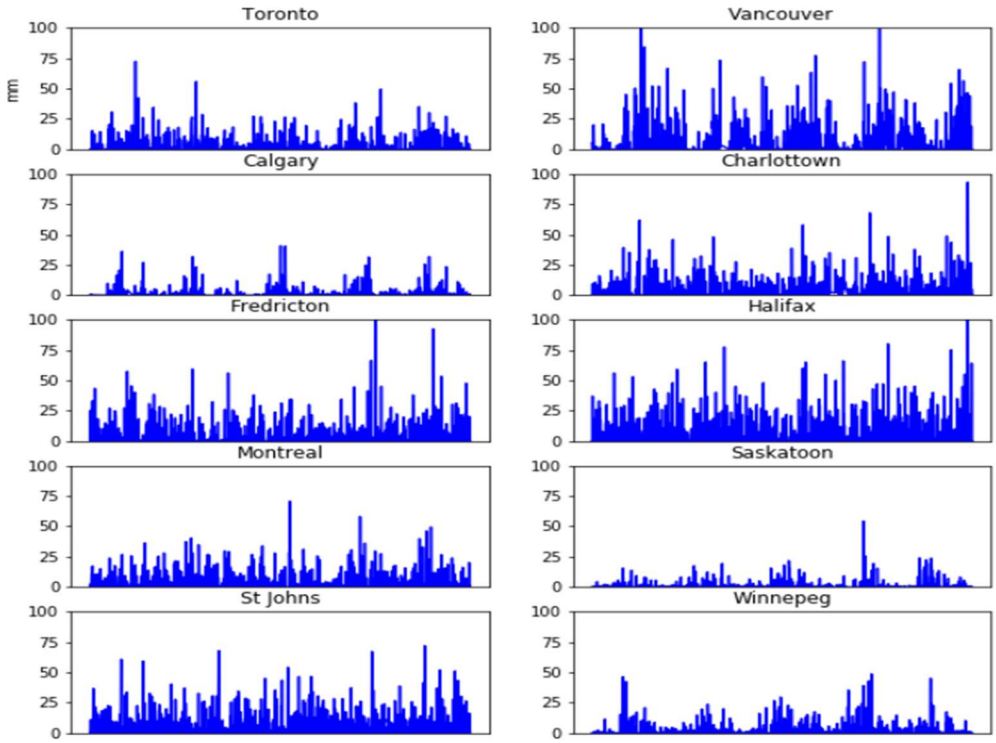
**Minimum precipitation is 0 mm for all*

	Lowest
	Highest

Weather 2014-2018



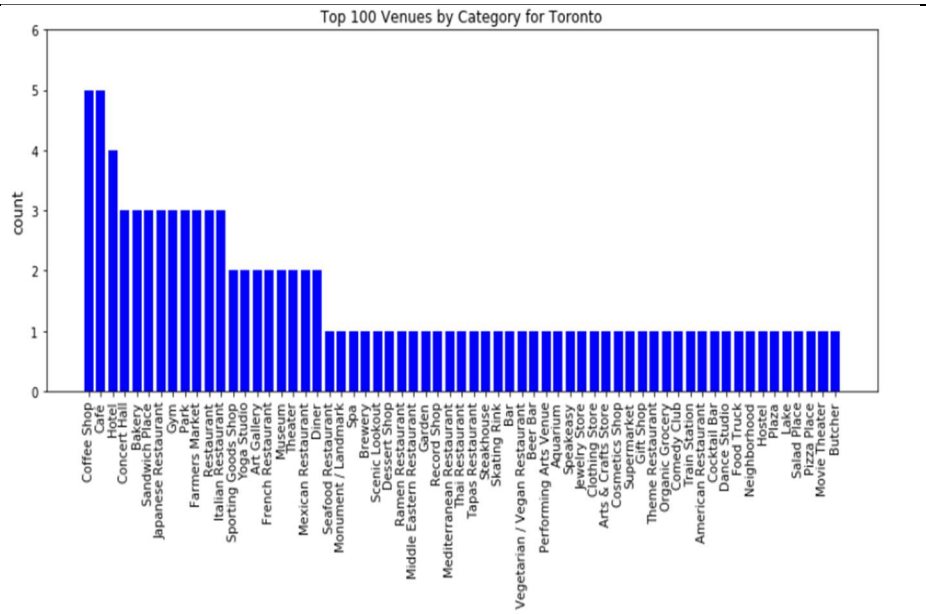
Precipitation 2014 - 2018



Top 100 Venues per City

Using the Foursquare API, the following summarises outputs from the python code depicting the top 100 venues (except for Charlottetown and St. John's) checked into on March 14, 2020, 10 km from what is considered the respective center of each of the cities.

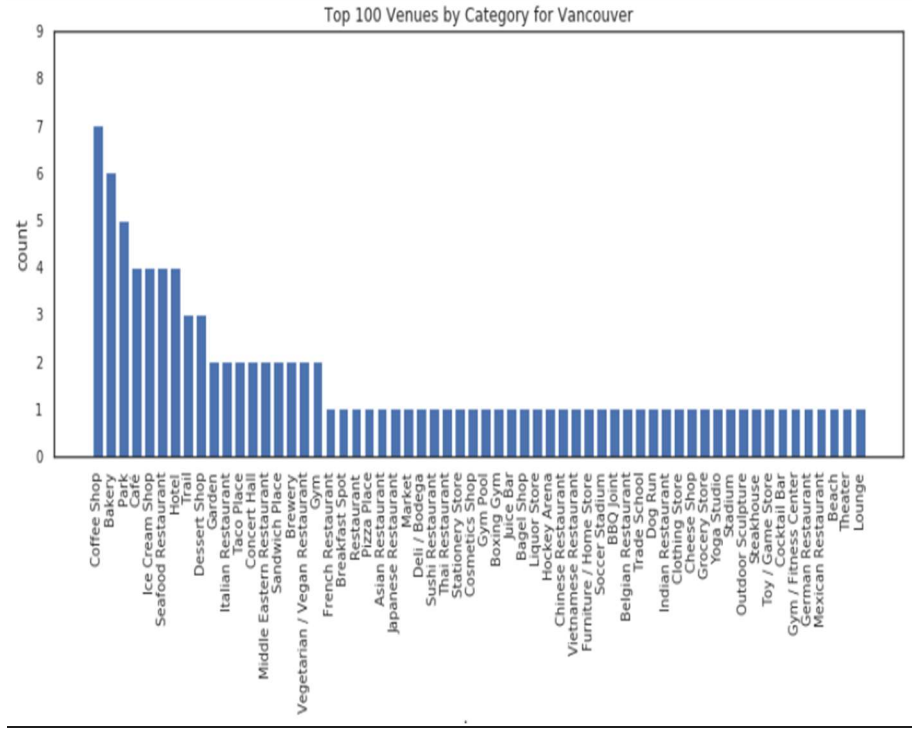
Toronto



Unique categories = 63

The top 20 categories account for approximately 50% of all check-ins and only two specific places had two check-ins each – Jimmy's Coffee and Banh Mi Boys. Based on distances from the city center coordinates, the average distance of a venue was 1210 m from the city center with the closest being at 174 m and furthest at 2093 m.

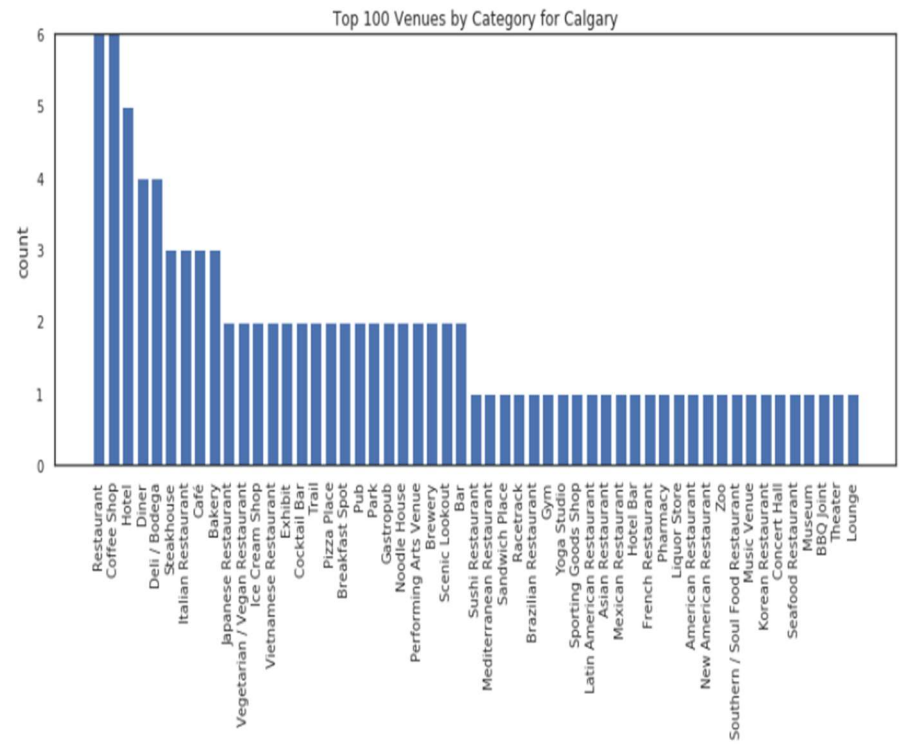
Vancouver



Unique categories = 60

The top 18 categories account for approximately 50% of all check-ins and similar to Toronto, only two specific places had two check-ins each – Matchstick Coffee Roastery and Earnest Ice Cream. Based on distances from the city center coordinates, the average distance of a venue was 1826 m from the city center with the closest at 307 m and furthest at 2726 m.

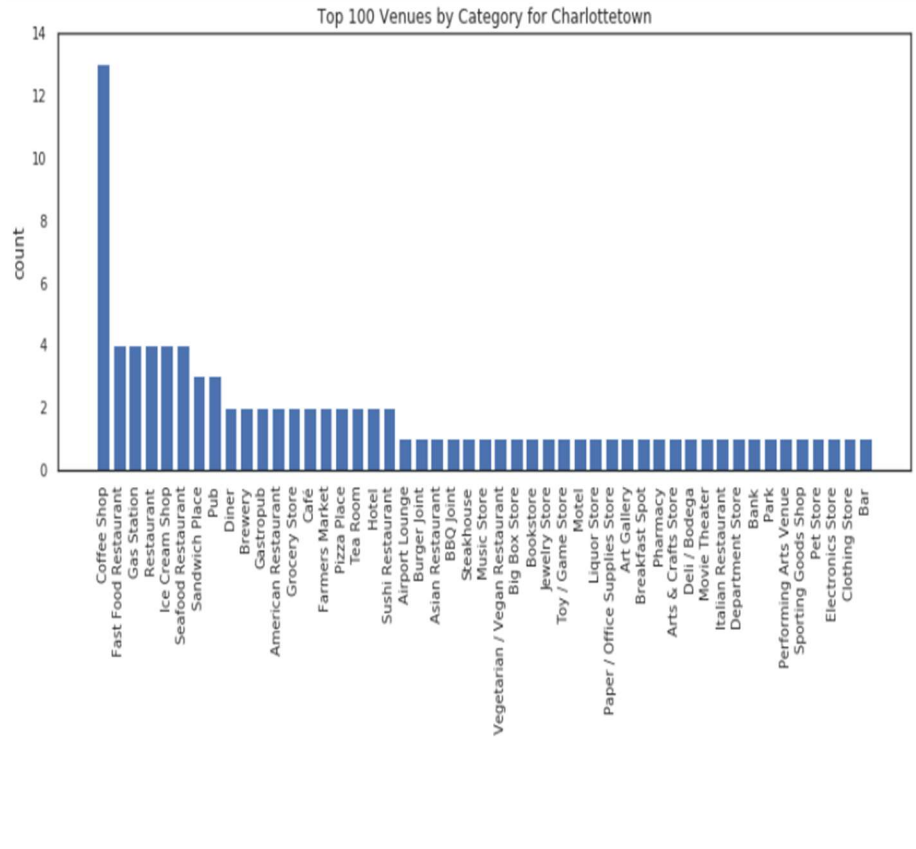
Calgary



Unique categories = 53

The top 12 categories account for approximately 50% of all check-ins and only one place had two check-ins – Over Easy Breakfast. Based on distances from the city center coordinates, the average distance of a venue was 1352 m from the city center with the closest check-in at 289 m and furthest at 2810 m.

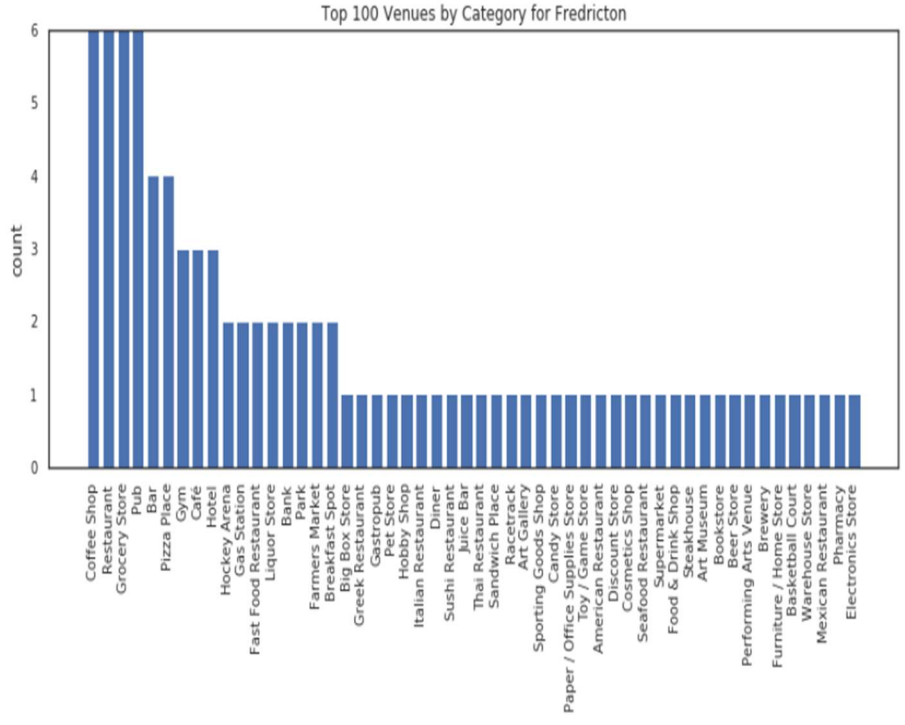
Charlottetown



Unique categories = 49

There were only 91 venues extracted for Charlottetown out of which the top 8 categories account for approximately 50% of all check-ins. As opposed to Toronto, Vancouver and Calgary, there were more places with multiple check-ins; Tim Hortons with 6, Starbucks with 3, Petro-Canada with 3 and Subway with 2. Based on distances from the city center coordinates, the average distance of a venue was 4895 m with the closest at 524 m and furthest at 9917 m.

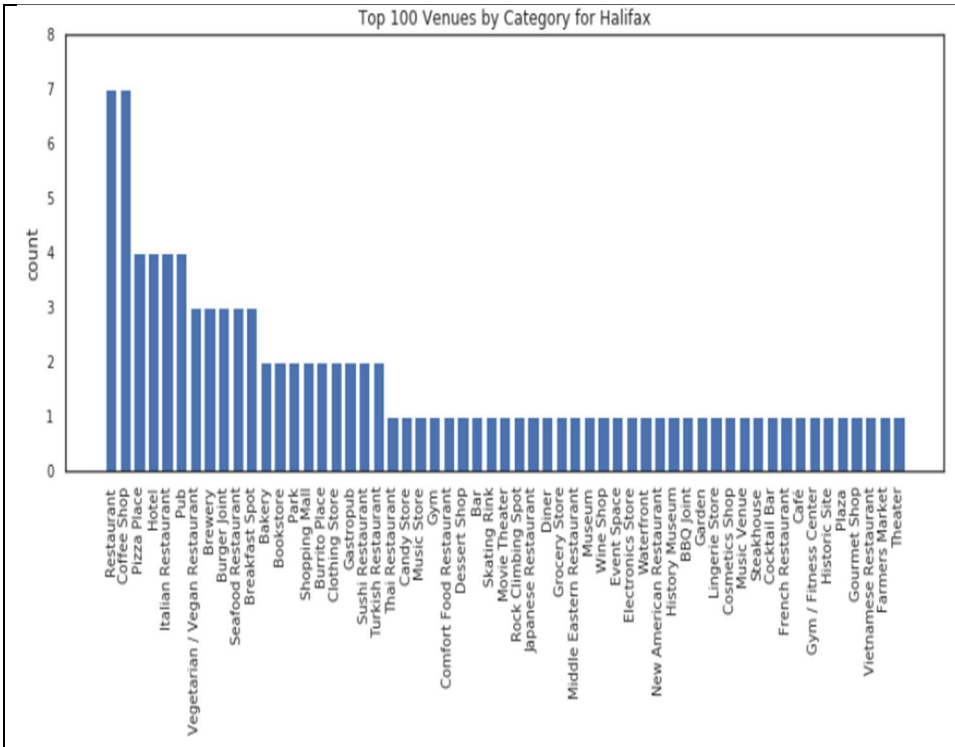
Fredricton



Unique categories = 52

The top 9 categories account for approximately 50% of all check-ins with multiple check-ins at Tim Hortons (4) and each of Second Cup, Sobeys, Circle K, TD Canada Trust, Alcool Superstore, Goodlife Fitness, Atlantics Superstore having 2. Based on distances from the city center coordinates, the average distance of a venue was 2207 m with the closest at 275 m and furthest at 9618 m.

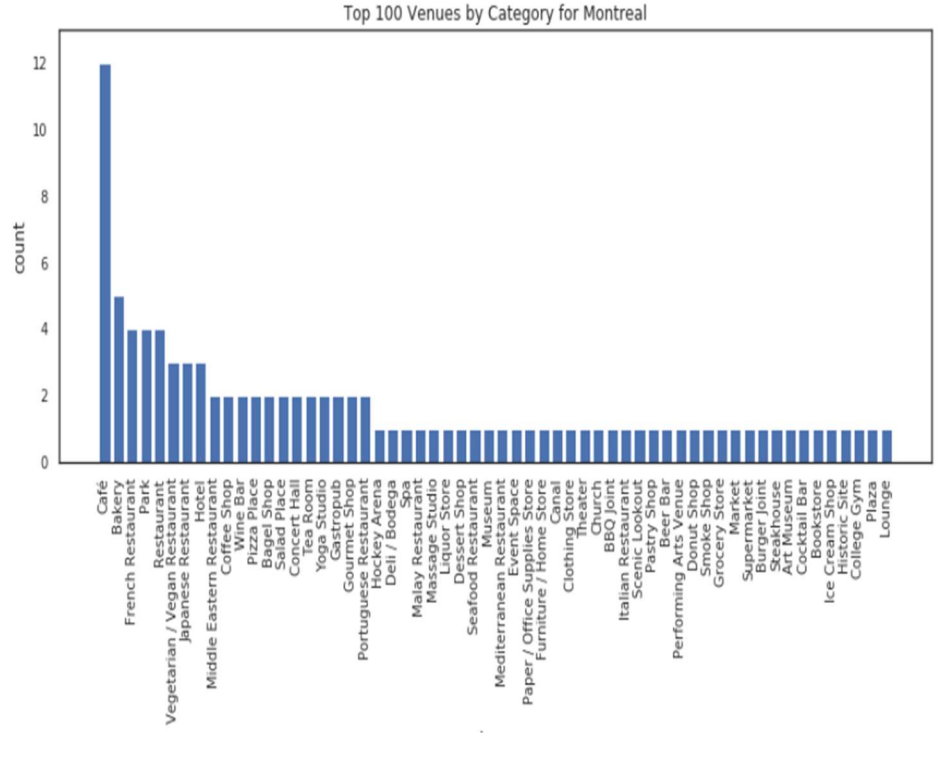
Halifax



Unique categories = 57

The top 11 categories account for approximately 50% of all check-ins and like Toronto, Vancouver and Calgary, there was only one place with 2 check-ins, Burrito Jax. Based on distances from the city center coordinates, the average distance of a venue was 1326 m with the closest at 127 m and furthest at 3598 m.

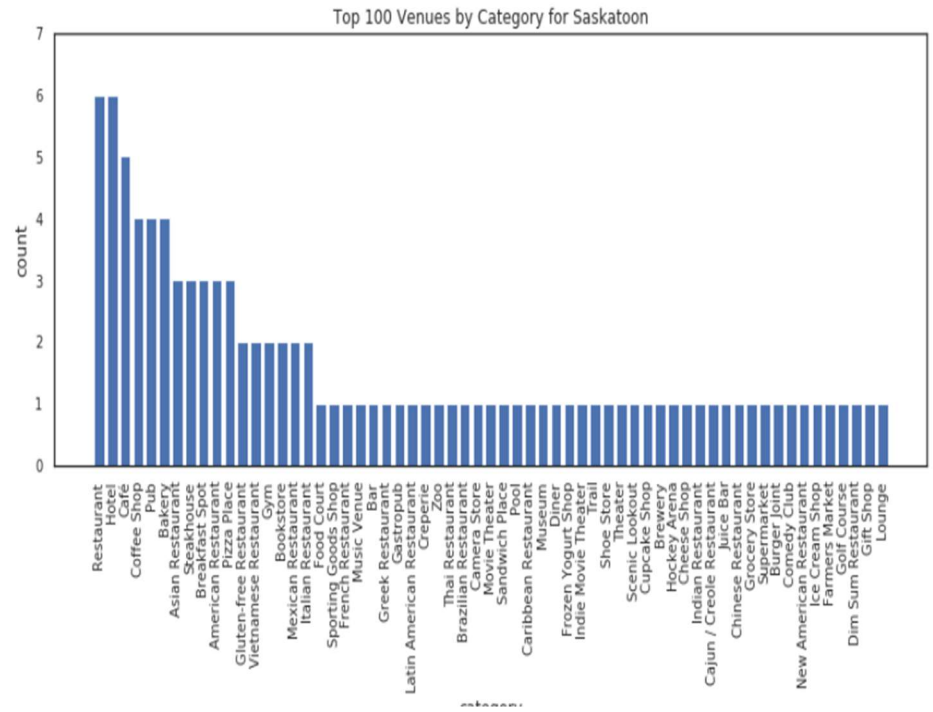
Montreal



Unique categories = 58

The top 20 categories account for approximately 50% of all check-ins and like Toronto, Vancouver and Calgary, there was only two places with 2 check-ins, Mandy's and MELK Bar a Café. Based on distances from the city center coordinates, the average distance of a venue was 2951 m with the closest at 1249 m and furthest at 3675 m.

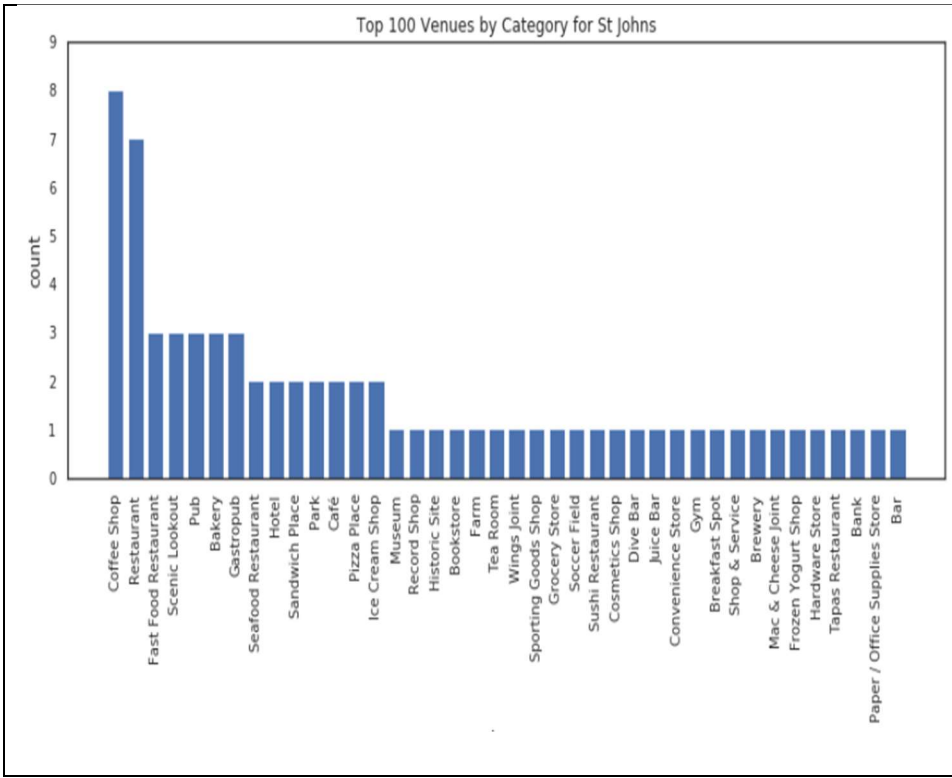
Saskatoon



Unique categories = 61

The top 15 categories account for approximately 50% of all check-ins and there was only one place with 2 check-ins, Starbucks. Based on distances from the city center coordinates, the average distance of a venue was 2204 m with the closest at 247 m and furthest at 6556 m.

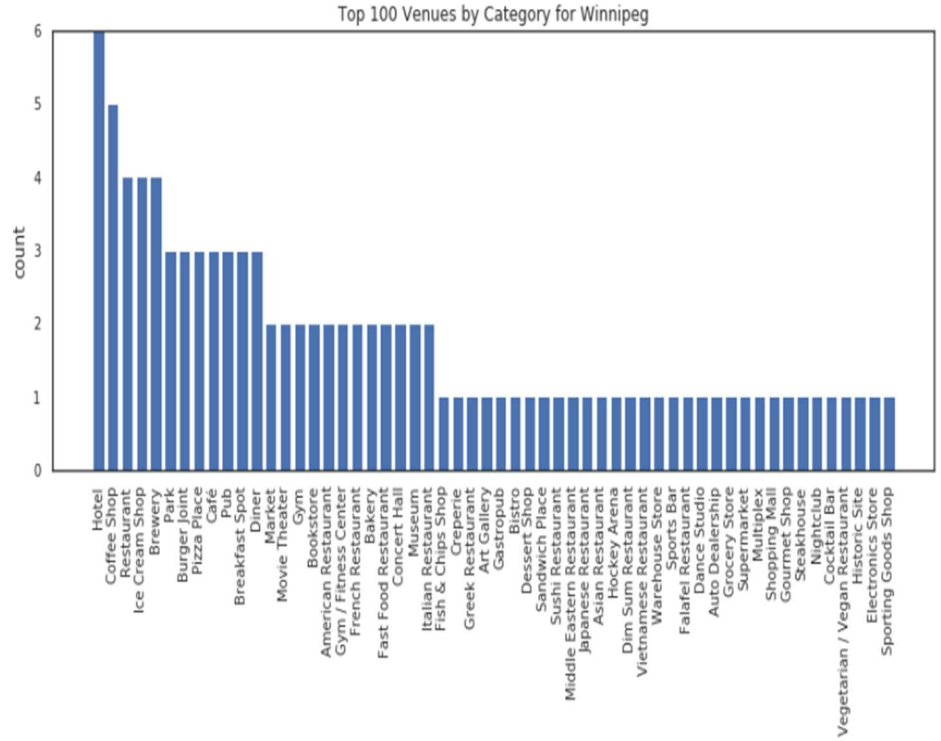
St. John's



Unique categories = 40

Like Charlottetown, there were less than 100 venues extracted, namely, 91. The top 14 categories account for approximately 50% of all check-ins with only one place with 5 check-ins, Tim Hortons. Based on distances from the city center coordinates, the average distance of a venue was 4092 m with the closest at 162 m and furthest at 9943 m.

Winnipeg

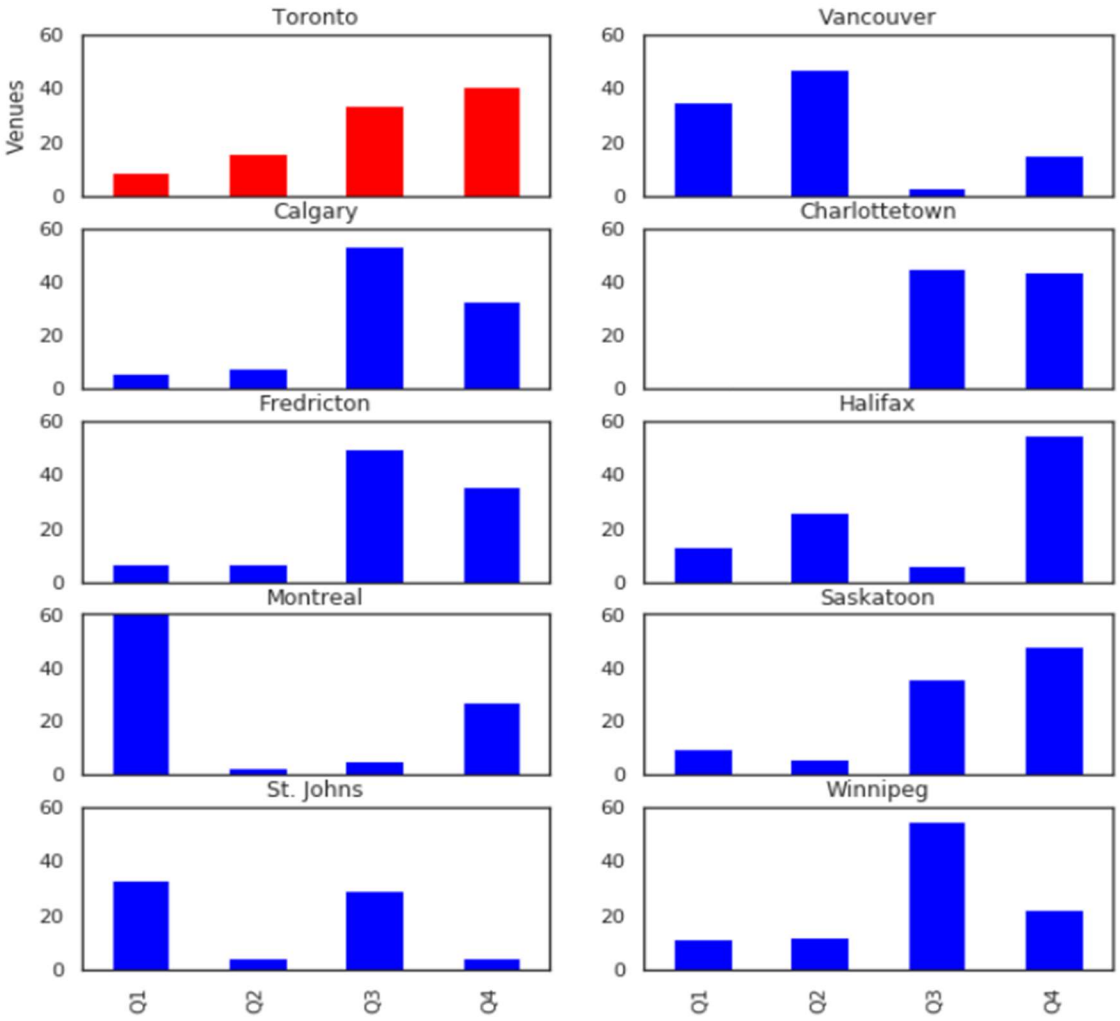


Unique categories = 56

The top 13 categories account for approximately 50% of all check-ins with no multiple check-ins noted. Based on distances from the city center coordinates, the average distance of a venue was 3164 m with the closest at 69 m and furthest at 8173 m.

Dispersion of Venues per City

Using the data above extracted from foursquare, location coordinates of each of the venues were compared to the respective coordinates of each of the cities to see the dispersion of the top 100 (exception of Charlottetown and St. John's) venues in each of the cities as compared to the respective city centers. Each of the quarters are defined clockwise starting from the midnight position and going all the way around. Note that the red color is only for illustrative purposes and does not signify anything unique about Toronto.



Clusters

Using the data extracted from four square, data was converted into vector format generating three clusters with the Top 10 most common venues and based on all the above, named the clusters according to that.

Cluster A: The Small City, Limited Options, Spread out from Center 0 – 10 km

Cluster Labels	City	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	Charlottetown	Coffee Shop	Fast Food Restaurant	Gas Station	Seafood Restaurant	Restaurant	Ice Cream Shop	Sandwich Place	Pub	Grocery Store	Brewery
0	Fredricton	Coffee Shop	Restaurant	Grocery Store	Pub	Bar	Pizza Place	Gym	Hotel	Café	Liquor Store
0	St Johns	Coffee Shop	Restaurant	Bakery	Pub	Fast Food Restaurant	Gastropub	Scenic Lookout	Café	Ice Cream Shop	Park

Cluster B: The Cosmopolitan City, Many Options, Close to Center 0 – 4 km

Cluster Labels	City	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
1	Toronto	Coffee Shop	Café	Hotel	Italian Restaurant	Japanese Restaurant	Gym	Bakery	Farmers Market	Park	Concert Hall
1	Vancouver	Coffee Shop	Bakery	Park	Hotel	Seafood Restaurant	Ice Cream Shop	Café	Trail	Dessert Shop	Garden
1	Montreal	Café	Bakery	French Restaurant	Restaurant	Park	Vegetarian / Vegan Restaurant	Hotel	Japanese Restaurant	Yoga Studio	Pizza Place

Cluster C: The Medium City with Visitors, Some Options, Between 3 – 8 km from Center

Cluster Labels	City	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
2	Calgary	Restaurant	Coffee Shop	Hotel	Deli / Bodega	Diner	Steakhouse	Café	Bakery	Italian Restaurant	Exhibit
2	Halifax	Coffee Shop	Restaurant	Hotel	Pizza Place	Pub	Italian Restaurant	Brewery	Seafood Restaurant	Burger Joint	Breakfast Spot
2	Saskatoon	Restaurant	Hotel	Café	Coffee Shop	Bakery	Pub	Breakfast Spot	Pizza Place	American Restaurant	Steakhouse
2	Winnipeg	Hotel	Coffee Shop	Ice Cream Shop	Restaurant	Brewery	Breakfast Spot	Diner	Pub	Pizza Place	Park

Discussion

The cluster analysis uses venue data to identify logical clusters that could be used to provide insight sought after in this and other studies. While population and weather data was captured for each of the cities, this is used to provide greater insight on the three clusters. This section focuses on explaining the three clusters and then subsequently commenting on the challenges faced.

Cluster A: The first cluster ended up being made up of small cities in that their respective populations don't exceed 250,000. Weather wise, these three cities are similar in terms of average temperature and precipitation. Both Charlottetown and St. John's had less than 100 venues. Within the top venues, the furthest venue in each of the cities was close to the 10 km radius set for the foursquare API ranging between 9.6 and 9.9 km. The three cities had locations with multiple check-ins with the highest between 4 and 6 check-ins. The two most common venue categories in all three were 'coffee shop' and 'restaurant' and the top specific venue was Tim Hortons.

Advice to stakeholders: In this cluster, one could look into covering spaces where venues are limited especially when it comes to the popular venues where multiple check-ins are noted. As seen from the dispersion of venues in these cities, one may expand services around a wider circle as opposed to just the city center as venues are widely spread and also cover for example the Western side in Charlottetown, Eastern side in Fredricton and South East / North West of St. John's.

Cluster B: The second cluster is almost the opposite of the first. They are primarily composed of the top three largest cities in all of Canada with populations over 2 million. The same couldn't be said about weather patterns as Vancouver was warmest with the highest average precipitation. The furthest venue in all three cities from the center of the respective city did not go beyond 3.7 km. Three of them did not have many locations with multiple check-ins signifying that the cities provide people with a highly diversified set of options for entertainment, food and services. 'Café',' Coffee Shop' and 'Bakery' were the top three categories.

Advice to stakeholders: In this cluster, one could look into unique ideas that could further diversify venues within the 2.5 km radius to the city center and likely successful in the top venue categories, e.g. Coffee Shops, Cafes and Bakeries. Based on the dispersion of venues, one may find opportunities to the East of Toronto's city center, West of Vancouver's city center and south of Montreal's city center.

Cluster C: This cluster comes somewhere in between both clusters above with four instead of three cities. With the exception of Calgary that has a population of 2.7 million, the other three cities have populations that range between 330,000 and 850,000, i.e. between small and large cities in Clusters A and B. Weather wise, all cities with the exception of Halifax have the same average precipitation of 1 mm and even temperatures vary between 3 and 8 C. The furthest venue in these cities, with the exception of Calgary (furthest venue at 2.8 km), is also somewhere between those in Clusters A and B ranging between 3.7 and 8.7 km. All cities in this cluster have limited, or none in the case of Winnipeg, multiple check-ins in the same venue and the top three categories include 'restaurant', 'coffee shop' and 'hotel'. That is why I called this cluster the in-between cluster. The issue with this cluster is Calgary where the population and dispersion of venues more closely relate to Cluster B. However, looking closely, Calgary is more comparable to this cluster in terms of venues and as a result was include din this cluster post the K-means cluster analysis.

Advice to stakeholders: In this cluster, one has more room for exploring new possible venues but also within a wider radius from the city center. Investors may look into what services can be provided close to or within hotels that are one of the top two most common venues in these cities with, as seen from the dispersion analysis, opportunities on the west side in each of the cities' city center..

Challenges

There were two key challenges that played a role in the outcome of the study and cluster analysis; location accuracy and data extracted from Foursquare by date / time.

As already mentioned, St. John's coordinates had to be manually entered in order to extract the top 100 venues. Having said that, what is defined as 'city center' in Foursquare is something that any researcher conducting a similar study needs to double check prior to accepting the data generated to ensure that data is representative of what a researcher is seeking.

The other is related to the fact that one could not extract venues using the foursquare API specifying date / time. This limits a researcher's ability to run the code multiple times on the same data unless

one ensures that the data is saved. This particularly became an issue when I would have liked to run the cluster analysis again on the same data setting the clusters to 4 and 5 to see if the results. I did not save the data because I did not expect the global pandemic to hit Canada this quickly which, as interesting as it is, is a key environmental factor that would shift data and make this study obsolete.

Conclusion

In conclusion, Foursquare data is very beneficial and can tell stakeholders a lot of information as long as one uses objective standards to set coordinates of cities and define the study well in terms of time as it might require a longer study than originally planned given limitations in the Foursquare API.

From an output perspective, it was very interesting to see how Canada's largest cities can be clustered and how one cannot base development and investment decisions just based on population and weather, but also each of the clusters depict different venue characteristics, dispersion around the center of each of the cluster cities and customer behavior.