Profiling Major Cities Based on Venues Using Clustering

Rifat Jabbar

March 16, 2019

Introduction

In the Applied Data Science Capstone course, the k-means method was used to perform clustering on neighbourhoods in the city of Toronto, Ontario, and determine the defining characteristics of each cluster with respect to the types of venues that are present in those locations. In doing so, it was discovered that the k-means clustering method results in one cluster that is significantly larger than the rest. This study examined whether or not a similar outcome would be achieved for another major urban Canadian urban centre, and if so, whether or not the characteristics of its largest cluster would have similarities with Toronto's largest cluster. The data collection and clustering method for Toronto was replicated for the city of Montreal, Quebec, to perform the aforementioned comparison. The findings may be of interest to businesses who are interested in expanding their operations and opening establishments in either of these two cities, and would like to better understand the existing competitive environments.

Data

The following data was used in this study:

- Postal codes for Toronto with associated boroughs and neighbourhoods, obtained from Wikipedia[1].
- Postal codes for Montreal with associated neighbourhoods, obtained from Wikipedia[2].
- Geographic location data for postal codes in Toronto. This data was provided in a csv file by the instructors of the Capstone course[3].
- Geographic location data for postal codes in Montreal. As this data was not readily available, it was gathered manually from Google Maps, stored in a csv file, and hosted on Github[4].
- Foursquare location data for venues in Toronto and Montreal. This data includes the name, latitude, longitude, and category of each venue, and was obtained via the Foursquare API.

Methodology

Postal codes for both cities was scraped from the Wikipedia pages into Pandas dataframes. The formatting of the Toronto postal code data was conducive to scraping using the read_html method for Pandas. However, the same could not be performed on the Montreal postal code data as it was presented in a matrix, rather than a table, on the Wikipedia page. The BeautifulSoup HTML parser was used instead to gather this data. In both cases, data collection was selective and any postal codes that did not have pre-assigned neighbourhoods were not included. Following this, the geographic coordinates (latitude and longitude) of each postal code was added.

The Foursquare API was used to gather location data of venues that were near each neighbourhood, based on its postal code. This data was reduced to neighbourhoods that had ten or more venues, and then one-hot encoding was performed so that the top ten most common venues could be determined for each neighbourhood.

[1] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
[2] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_H
[3] https://cocl.us/Geospatial_data
[4] https://raw.githubusercontent.com/rjabbar/Capstone-project/master/Geospatial_Coordinates.csv

These neighbourhoods were then clustered using the k-means method. Prior to doing so, the elbow method was used to determine the optimum value for k. The largest cluster was identified and its most common venues were analyzed to determine its characteristics.

Results

Postal codes and their respective geographic coordinates were collected for 103 neighbourhoods in Toronto[5] and 121 neighbourhoods in Montreal. After gathering the location data of nearby venues and filtering for neighbourhoods that had at least ten venues, the venues dataframe for Toronto represented 62 neighbourhoods, and the venues dataframe for Montreal represented 55 neighbourhoods.

The elbow method was used to determine the optimum value for k to perform k-means clustering on the venues data. As shown in Figure 1, both Toronto and Montreal were clustered with k = 4.

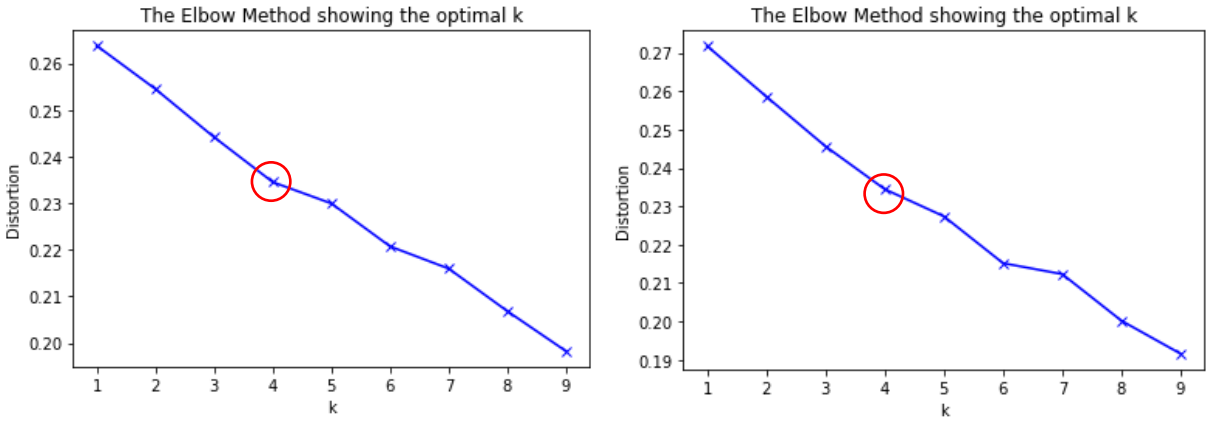

Figure 1: results of the elbow method for Toronto data (left) and Montreal data (right)

The resulting clusters were superimposed onto city maps using Folium, as shown in Figures 2 and 3. The first, second, and third-most common venues in the largest cluster of each city are shown in Figures 4 and 5.

[5] It should be noted that this did not include surrounding cities in the Greater Toronto Area, such as Markham, Vaughan, Brampton, or Mississauga.

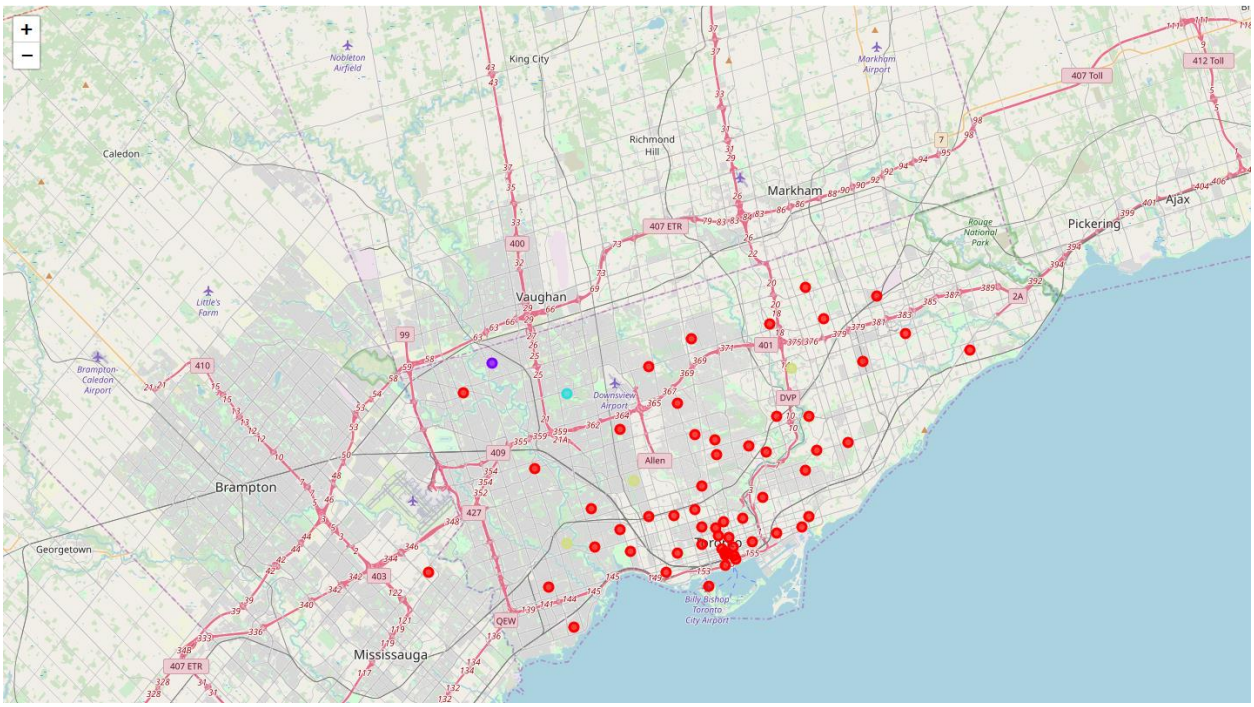

Figure 2: clusters in Toronto

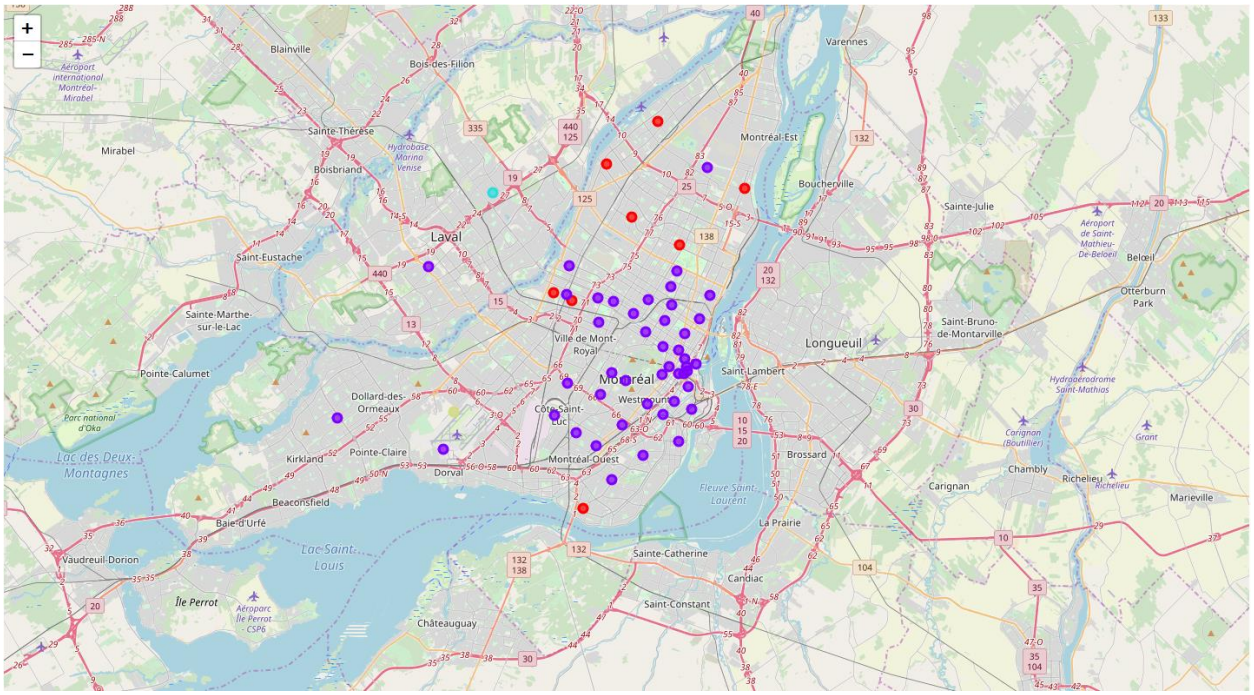

Figure 3: clusters in Montreal

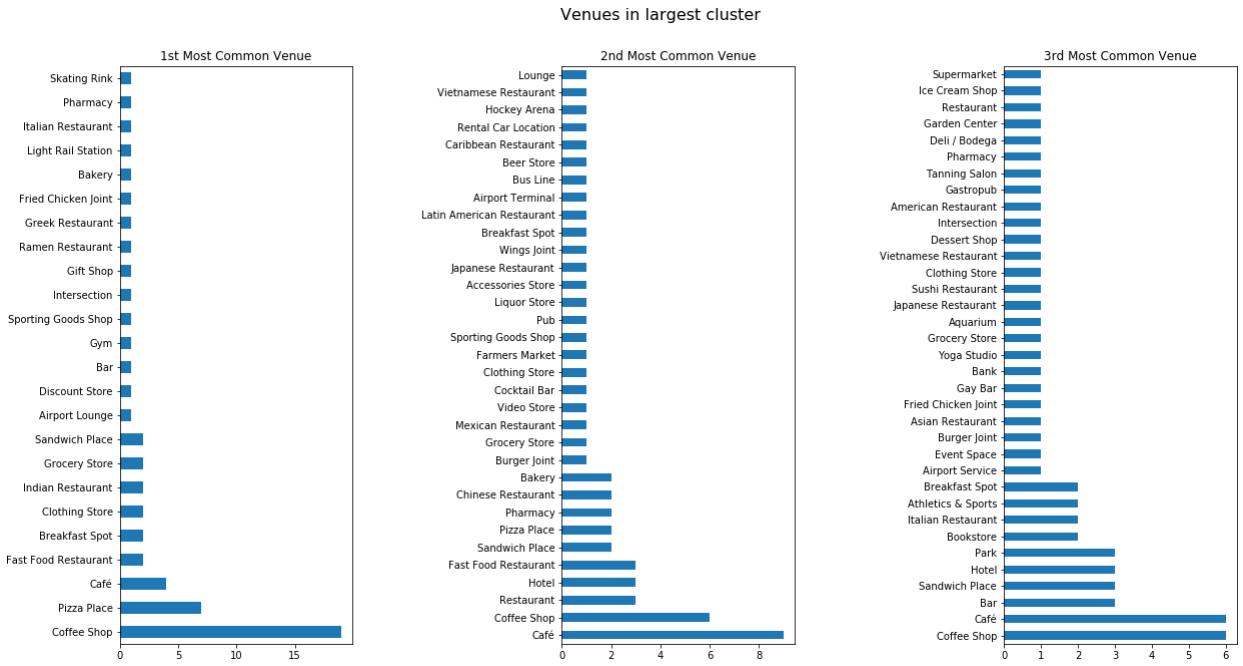

Figure 4: most common venues in Toronto's largest cluster

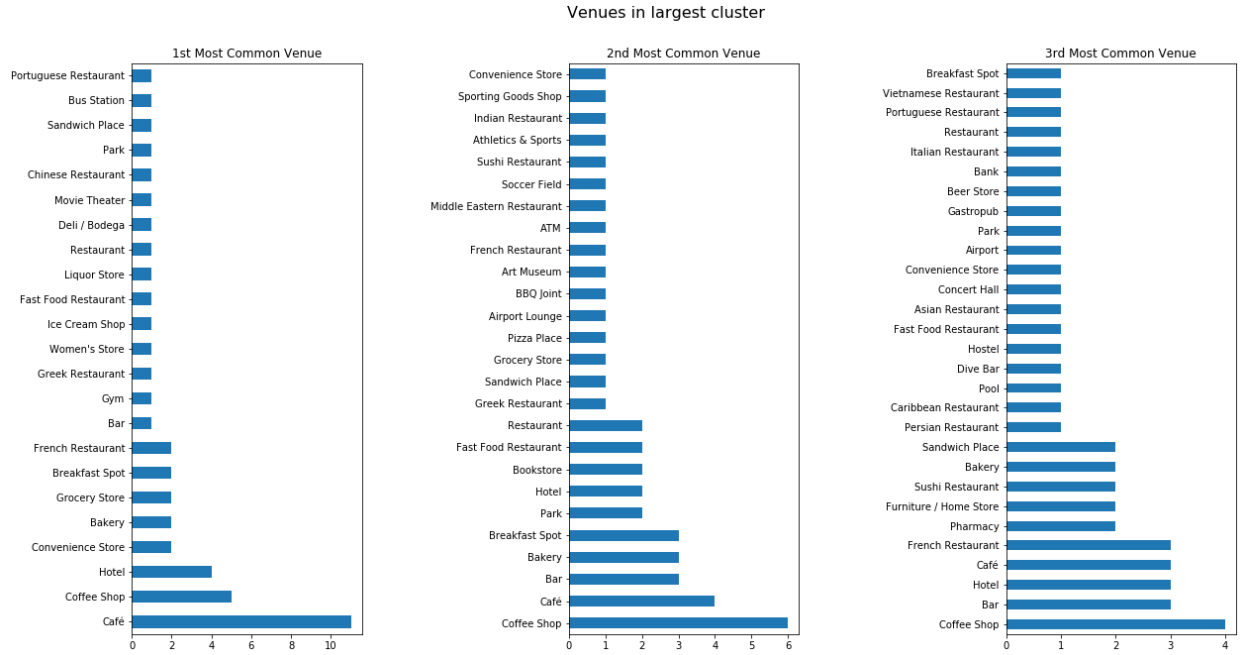

Figure 5: most common venues in Montreal's largest cluster

Discussion

When the two largest clusters are compared, it is evident that there are several similarities between the two cities, despite being in different provinces with different cultural contexts. Coffee shops and cafés are the most common venues in these clusters, along with bars, hotels, and various types of restaurants. This is to be expected of a large urban centre in North America, and therefore it can be assumed that this clustering method would likely produce similar results for other major cities as well. The largest cluster can be called “the urban centre cluster” and its constituents will likely be very similar for all major North American cities, with minor variations.

Beyond the similarities, the actual quantities of each type of venue would be of interest to different businesses. For example, a proprietor looking to open a single café or coffee shop in Toronto would be faced with stiff competition, and would either have to consider an alternate location, or to develop a unique value proposition that would entice customers despite the abundance of alternative venues nearby.

The code that was written for this study could be improved by including the ability to visualize the number of venues of a specific category (such as coffee shops) in all of the neighbourhoods to aid in the selection of a new location. Another improvement would be a standard function for scraping postal code data from the web, and the use of the Geocoder package to generate geographic coordinates from postal codes.

Conclusion

This study compared the largest cluster in Toronto with the largest cluster of another major Canadian city, Montreal. The study demonstrated that the data collection and clustering method can be replicated for other major cities and will likely result in the largest cluster having very similar characteristics with other major cities, with respect to the venues in the cluster. Businesses that are interested in learning about the competitive landscape in a particular urban centre can use this method to better understand the environment before making decisions to launch operations in that location.