Canadian City Dissimilarity

Brett Wiens – October 10, 2019

Coursera – IBM Data Science Using Python – Capstone Project

# **Problem/Introduction**

As Canada heads towards an election, it is often described as a very regional defined country. The country is easily divided into distinct areas. The Maritimes, Quebec, Ontario, the prairie provinces, and British Columbia. At the same time, despite being the second largest country in the world by area, it is an urbanized country with more than 80% of the population living in cities. Contrasting with the perception of a fragmented region-based country, cities are considered sterile and uniform. Which is it? If we compare amenities available to each of the major cities in Canada, will we determine that all cities are the same, or are there clusters of similar amenities in each FSA (forward sortation area) in each city that differentiates them from each other. This assignment will attempt to use Python, foursquare, and a variety of cartographic and analytics approaches to answer that question.

# **Data**

All postal codes (FSA) and a somewhat meaty list of community names are available through Wikipedia. Frustratingly, Toronto (the sample demonstrated in the course material) is unique in its structure and format. Fortunately, every other province is consistent. Once the method for scraping was figured, it could be replicated.

- All FSA data is being scraped from Wikipedia and processed using BeautifulSoup.

- Geographic information for FSAs can be obtained from Statistics Canada, the shapefile was downloaded from:

https://www150.statcan.gc.ca/n1/en/catalogue/92-179-X

- The FSA centroids were downloaded from Geogratis: <http://geogratis.gc.ca/services/geolocation/en/locate?q>

# **Methodology**

All FSAs that correspond to one of Canada's major cities will be scraped from the Wikipedia site. Conveniently, all major cities have some of the neighbourhoods in parentheses after the name. For example, "Calgary (Briar Hill, Capitol Hill, Hillhurst)." This makes for a convenient way to identify them without manually delineating each necessary FSA. For this study, I will use Vancouver, Calgary, Edmonton, Winnipeg, Toronto, and Montreal. A few of the other major cities (Ottawa and Quebec City) do not have the neighbourhood shortcut, so I've left them out of scope.

The FSA data will be combined with amenity data from foursquare for each FSA and further combined with the geographic centroids of each FSA courtesy of Geogratis, and the polygon shapefile from statistics Canada. There is an abundance of sociodemographic data available from statistics Canada as well, but we will leave this out. The amenity data will be what we use to determine the similarity or dissimilarity of various FSA in Canada's big cities.

The ten most common amenity types will be determined for each FSA, these will then be compared using K-means clustering to determine which communities are most similar or dissimilar, and whether different cities show different characteristics. This will be visualized cartographically using geopandas. If different cities show a greater variety of amenity distributions than those within the same city, it will be determined that there are marked regional differences. If clustering is approximately the same between cities as within cities, then it will be determined that cities are in fact bland uniform melting pots.

# **Exploratory Analysis**

The first stage of the exploratory analysis is just to validate that the FSAs that I've collected do correspond to the cities they ought to. We would expect to see a cluster of red markers on this map showing Vancouver, Calgary, Edmonton, Winnipeg, Toronto, and Montreal. Figure 1 - Validating Location of Cities Scraped shows that as we expected, the six cities do correspond to the locations we would expect. It is hardly a beautiful figure, but it does the trick.

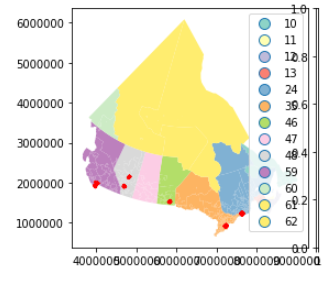
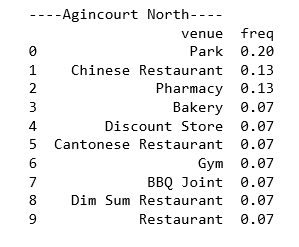
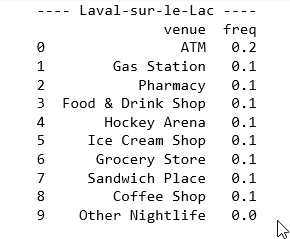


Figure 1 - Validating Location of Cities Scraped

Now that we have all the amenity information around each FSA, the next step is to validate that the information is meaningful... Did the data pull collect information about each community, are the counts different from community to community, and are there enough diversity of amenities that there will be meaningful distinction? A quick check reveals that there are 14,057 records that have been pulled from foursquare, and there are 418 unique categories. I've demonstrated what those venues are in the following box for thoroughness sake.

Figure 2- Comparing the top 10 venues in Laval and Agincourt



The second stage of the analysis is just a review that the amenities make sense. The above table shows that the one-hot encoding worked, and the venues do vary from place to place. It is amusing that a waterfall is considered a venue. Figure 2- Comparing the top 10 venues in Laval and Agincourt sorts and lists the top 10 venues by community. I won't dig too deeply into this, but just looking between the first two (Laval and Agincourt) we can see that there are similarities and differences. The number 1 and 2 in Laval (ATM/Gas Station) aren't even in the top 10 for Agincourt, similarly, Agincourt's #1 and #2 (Park/Chinese Restaurant) don't show up in Laval's top 10. In fact, the only commonality between the two is the #3 (Pharmacy) otherwise the lists are mutually exclusive.



Figure 3- The most common venues in 10 FSAs (head)

Figure 3- The most common venues in 10 FSAs (head)is a really powerful table that shows the different venues for the first ten communities. What I find most fascinating about this is how unique they all are. We don't see the same two things in a row for any of the choices. Even in Ahuntsic, which is thrice represented here, the amenities are quite distinct. If we see regional trends, they will certainly have a high hurdle to clear.

There is good confidence that there will be meaningful results.

# **Results**



Figure 4- Canadian City FSA Clusters by City

The six plots in Figure 4- Canadian City FSA Clusters by City comprise the primary findings of this assignment. Six different maps have been created showing the cities and their cluster distributions.

- Vancouver is almost entirely assigned to Cluster 1

- Calgary is split between Clusters 2 and 6

- Edmonton is split between Clusters 1, 3, and 6.

- Winnipeg is mostly Cluster 2.

- Toronto is primarily Cluster 4, with some Clusters 1 and 5.

- Montreal is a patchwork of Cluster 2, 3, and 6.

Toronto is the only city with any of Cluster 4, and it is the dominant cluster. Vancouver is a nearly uniform Cluster 1, while Calgary, Edmonton, Winnipeg, and Montreal are more mixed.

# **Discussion**

The question presented asked if Canadian cities are homogenous entities, indistinguishable from one another based on their venue availability. The analysis suggests that this is not the case. While it is not determinable whether regional tendency or unique urban criteria describe it, it is clear that Canadian cities are unique and distinct from one another.

Vancouver and Toronto are quite different from one another, but within each of them, they are nearly uniform in terms of what services and amenities are available to the citizens. Montreal has a patchwork of various types of amenities that varies greatly from neighbourhood to neighbourhood. Calgary and Winnipeg share a similar distribution of amenities while Edmonton looks like it is partway between its prairie neighbours (Calgary and Winnipeg) and Vancouver to the West.

Referring to the expected regionality. If clusters were to be applied to the cities themselves, it is likely that the expected regions would appear if we were to create four clusters:

- Vancouver (West Coast)

- Calgary, Edmonton, Winnipeg (Prairies)

- Toronto

- Montreal

It is interesting to note that despite similar needs, and even similar ethnic, gender, and age characteristics, regional trends influence service availability in each city.

# **Conclusion**

Those services that are available to citizens of Canadian cities vary by region. Cities within the same region demonstrate more similar characteristics, and great consistency within themselves, while cities between region show different characteristics. A city may in itself be a pretty consistent, sterile environment, but travelling between Canadian cities will provide a variety of experiences. For politicians, it can probably be concluded that the message needs to be modified for each unique place.