IBM Data Science Capstone: Week 5:

Jiawen Wang

August 10, 2019

Introduction

Canada has always been known for be a “melting pot” for culture. As of 2016, the population of immigrants in Canada sits at just above 7.5 million, making up almost 22% of the country’s population. With its people being so diverse, so too is its food.

There are two main audiences that this data would be for:

1. **If you’d like to open an restaurant**

Say you’d like to open a new restaurant and you’re looking for areas to set up your new shop. Maybe you’d like to open an Italian restaurant, or a Sandwich shop, or a Vietnamese restaurant. You’re thinking of a certain price point -- maybe you’re a stickler for fine dining or maybe you want to open a family-owned shop with cheap eats. Based on your requirements, the following analysis will pinpoint the cities in Canada that have proven to be the most popular foods at that price point.

2. **If you’re a new immigrant to Canada**

Though Canada is known to be diverse, some places are just more diverse than others. For example, almost half of Toronto’s population is made up of immigrants. Maybe you want somewhere to feel at home, and you want somewhere with a wide variety of different cultural foods. The clustering will help you do that, and also find an intersection between kinds of food, as well as the diversity of the population.

Data

A few datasets will be used, as follows:

For diversity measures amongst population:

- From Statistics Canada (2016 Census)
    - Immigration Data

        - This details the number of immigrants in Canada, in each of its major cities, distribution of immigratnts in Canada, and proportion compared to the rest of the population. It also reports these numbers for immigrants between 2011 and 2016 only.
    - Population Data
        - Only one column is needed: the population of each city. This will be used with the visible minorities data to figure out the proportion of Visible Minorities in each city
    - Visible Minorities
        - Statistics Canada defines a visible minority as: 'Visible minority' refers to whether a person belongs to a visible minority group as defined by the Employment Equity Act and, if so, the visible minority group to which the person belongs. The Employment Equity Act defines visible minorities as 'persons, other than Aboriginal peoples, who are non-Caucasian in race or non-white in colour'. The visible minority population consists mainly of the following groups: South Asian, Chinese, Black, Filipino, Latin American, Arab, Southeast Asian, West Asian, Korean and Japanese.

For coordinates I used a dataset from Simplemaps.com: Coordinates of major cities in Canada

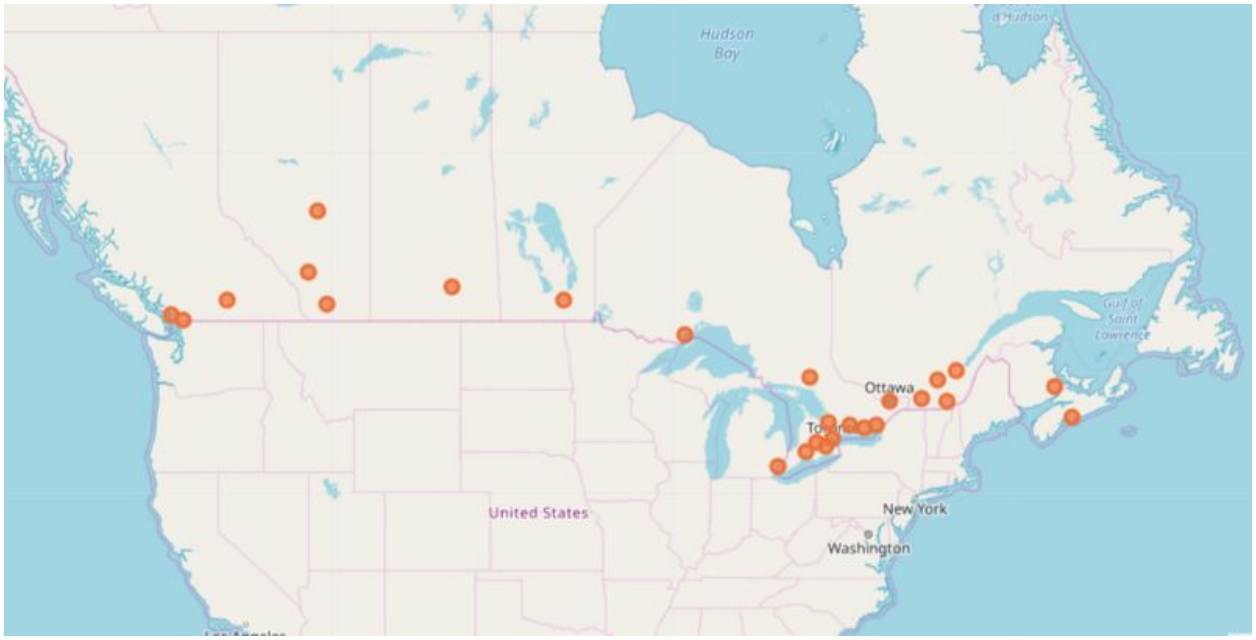
For venue information, I used the Foursquare Places API. In particular, I used the explor endpoint for each city to find that 50 closest “food places”. I excluded venues that are categorized as Cafes, since I would consider them a different kind of venue all on its own, and I excluded ‘Restaurants’ since this was too broad a term and probably arises from bad venue tags.

Methodology

Some Exploratory Data Analysis

After preprocessing and compiling all the non-venue-related together to create the dataframe df_cities, I looked at few metrics.

First, here are the cities that I'll be working with:

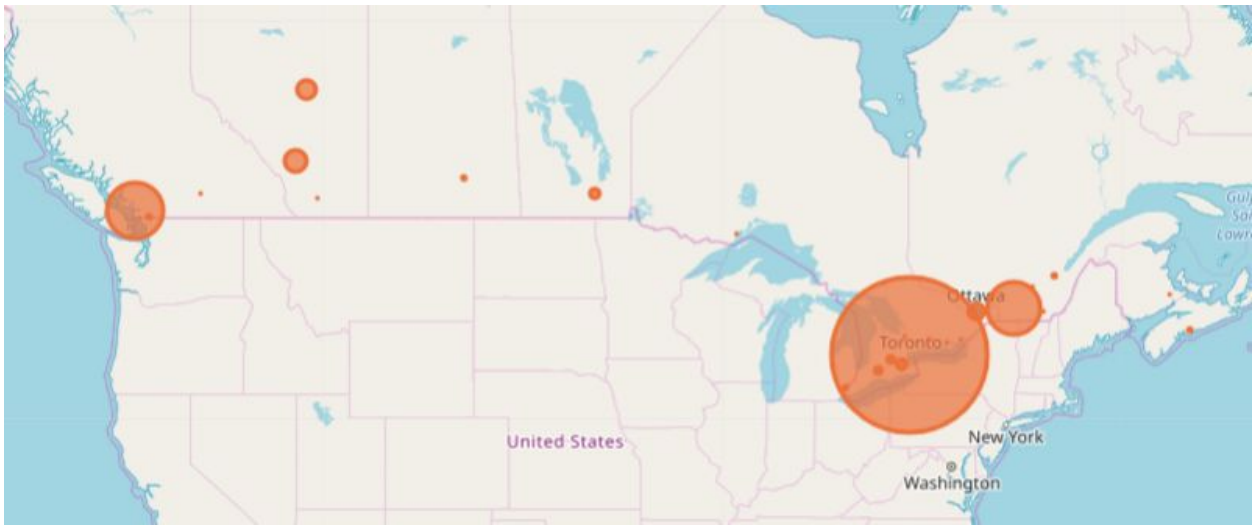


In alphabetical order, these are: Abbotsford, Barrie, Belleville, Calgary, Edmonton, Halifax, Hamilton, Kelowna, Kingston, Kitchener, Lethbridge, London, Moncton, Montréal, Ottawa, Peterborough, Québec, Regina, Sherbrooke, Sudbury, Thunder Bay, Toronto, Trois-Rivières, Vancouver, Windsor

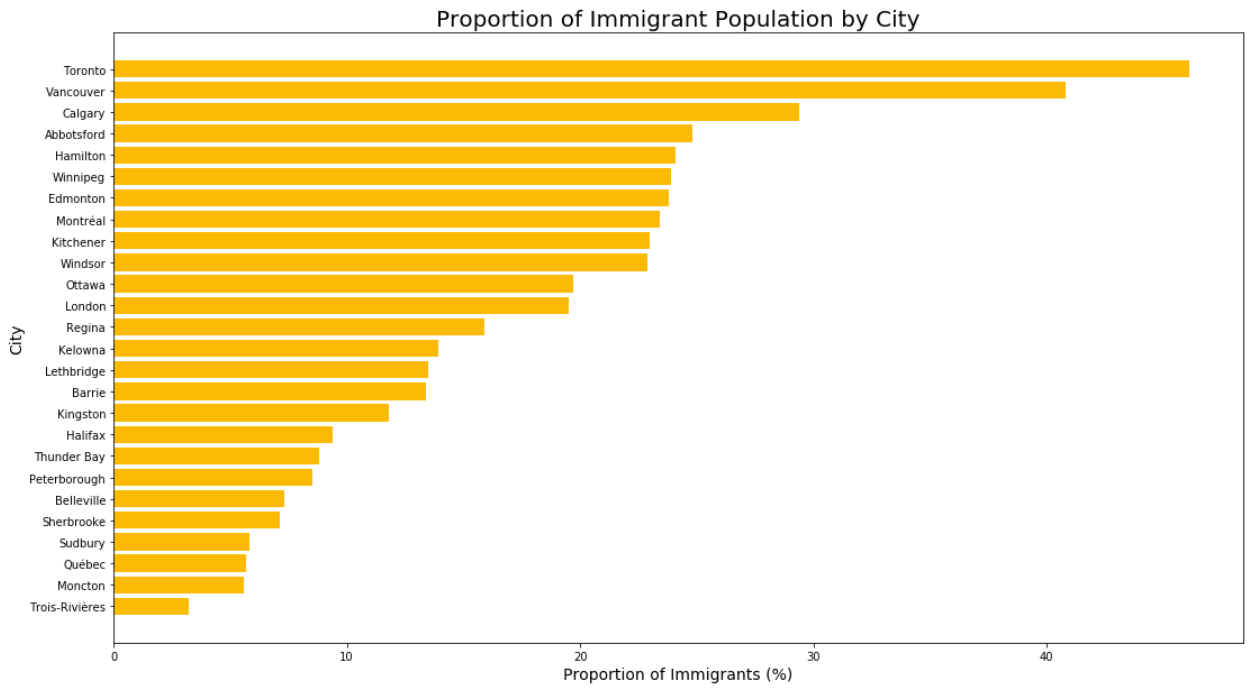
The first few rows of the main compiled dataframe that I used looks like the following, where proportion is in percent.

City	MedianIncome	VisibleMinorities	lat	lng	ImmigrationProportion	MinorityProportion
Regina	84447	30965	50.45	-104.62	15.9	17.58
Edmonton	94447	279280	53.55	-113.50	23.8	26.40
Calgary	99583	355315	51.08	-114.08	29.4	32.01
Lethbridge	75452	9440	49.70	-112.83	13.5	13.37
Kelowna	71127	11890	49.90	-119.48	13.9	9.50

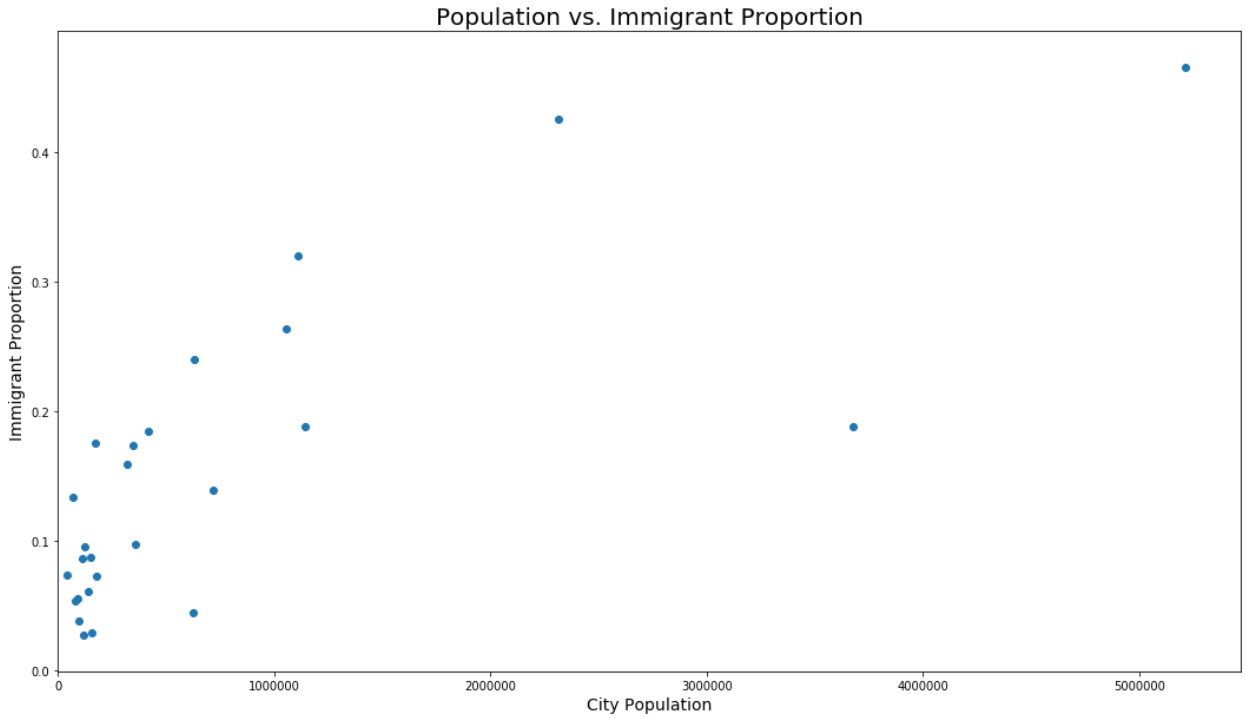
Just to put it into perspective, here are the same cities but with markers with radiuses proportionate to the city's population:



Some other metrics include the analysis of immigration population in each city, sorted in descending order:



In addition, here is a scatter plot that shows that immigrants are more likely to move to highly-populated cities in Canada.

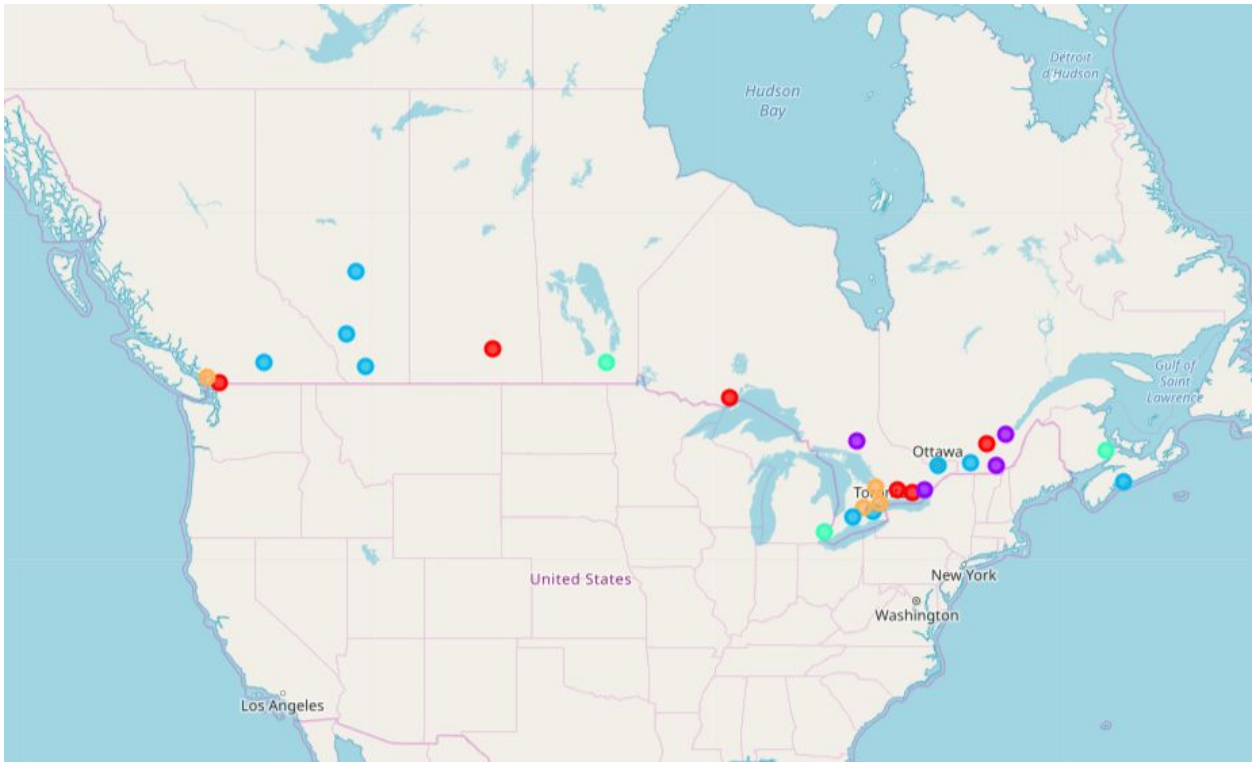


For the main component of the analysis, I used K-Means clustering, in particular k=5 from the sklearn package, to find both the most cities that are closest to each other in diversity as well as food diversity.

First, for the food, I compiled my venue frequencies via one-hot encoding. That is, I counted the types of venues for each city, and gave each category (e.g. Vietnamese Food, Sandwich Shop), and divided it by the total number of venues. I sorted them in order of frequency, and fed the frequency into the K-Means Model.

I used a similar tactic for diversity measurements, but instead using immigration proportions and visible minority proportions, but this time with a MinMax Scaler from sklearn.

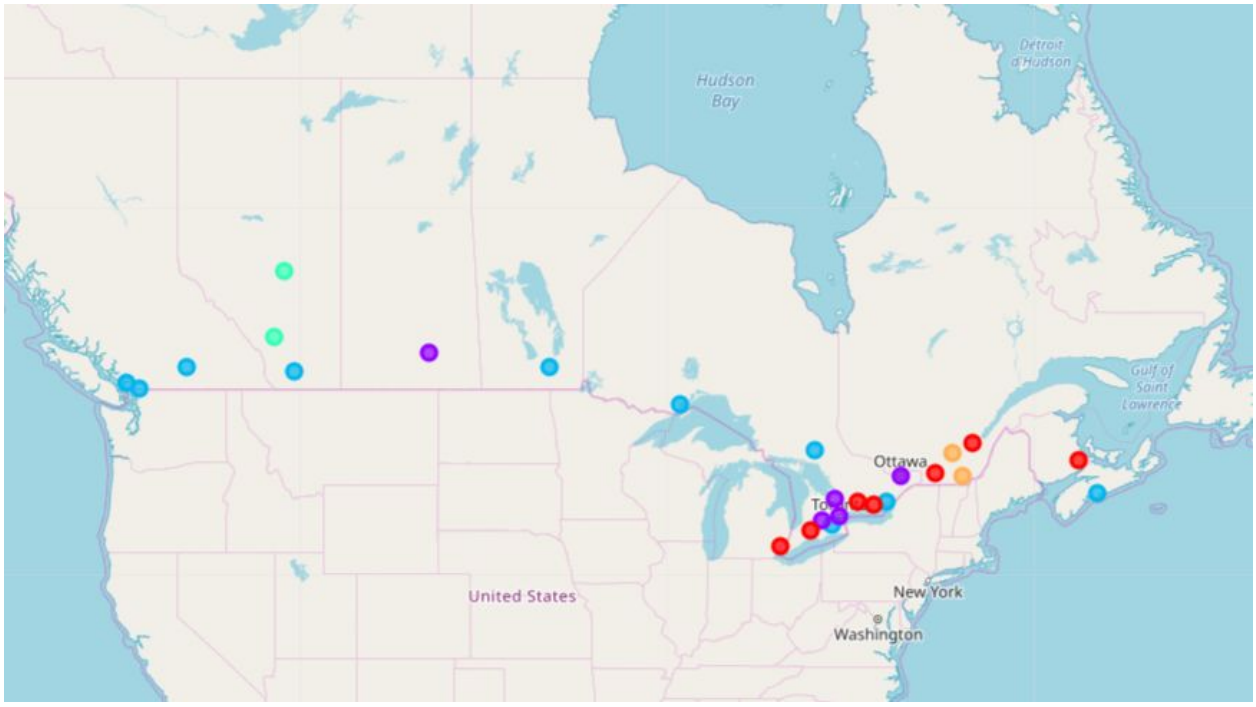
Results



Clustering 1:

- Cluster 0 (Red): This cluster consist of smaller cities whose main food places are Fast Food, Sandwich Shops, and Pizza Places. Examples include Abbotsford, Peterborough, Regina, and Thunder Bay. This might be due to the fact that these places are often mid-way points for people travelling around Canada.
- Cluster 1 (Purple): Though not very multicultural, this cluster contains slight more sit-down type food venues, such as Italian restaurants, pubs, and Breakfast spots. Members of this cluster include Sudbury and Kingston, with slightly higher populations that Cluster 0.
- Cluster 2 (Blue): Interestingly enough, these seen slightly localized despite not giving the algorithm any coordinate information. Here we see more multiculturalism, with more higher-end places like Steakhouses, and some niche foreign foods such as French resaurants and Middle-Eastern restaurants
- Cluster 3 (Light turquoise): Quite similar to Cluster 1, these are smaller cities with more sit-down restaurants, though slightly lacking in foreign foods

- Cluster 4 (Orange): Localized at and near Toronto, with the added city of Vancouver. Seems to have the most high-end areas with the most multicultural foods, with Italian, Indian, Thai, and Vietnamese foods seemingly the most popular.



Clustering 2:

A summary of the clusters is as follows

Cluster 0, in red, holds cities with average to high diversity, and relatively low median income. Cluster 1, in purple, include Ottawa, Barrie, and Toronto, with average median income and relatively high diversity. Cluster 2, in light blue, has a lower income than Cluster 1, but similar diversity, and includes Winnipeg, Kingst, Vancouver, and Sudbury. Cluster 3, in bright green, have a very high median household income with high diversity. Cluster 4 has the lowest income and the least diversity.

Results

With this info, prospective restaurant owners can decide where to set up shop so that they can leverage the popularity and the diversity of the surrounding area, and set price points that are appropriate for that area.

For example, if someone wishes to open a high-end Chinese Restaurant such as a fancy dimsum restaurant, a good area would be somewhere in Edmonton, where it is both diverse and where median income is high. We can see that Chinese restaurants are the 2nd most popular kind of food menu Edmonton.

Say that someone native to Quebec has an established sandwich chain, where prices are affordable and production costs are relatively low. Then that person may benefit from establishing a new branch in Trois-Rivieres, where median income is low, and sandwiches are the kinds of food that passerbyers and locals have come to expect.

The opposite is possible too. Perhaps someone wants to fill a niche in a certain area that has not yet been explored. Maybe that person can establish a Vietnamese restaurant in Abbotsford, where immigration and diversity is relatively high, but the number of Asian restaurants are low. They could be a leader in their own market.

Finally, imagine an immigrant from Thailand, who wishes to establish a new life in Canada. They wish to be surrounded by like-minded people and feel comfortable where they live, so they'd like to choose somewhere with a high proportion of immigrants and visible minority. They could go with somewhere in cluster 4 of the food places, such as Toronto or Vancouver, where diversity is high and so is Asian-style restaurants.

Conclusion

This was brief analysis of the cities around Canada based on diversity and types of food venues. Just visually, you can see the relationship between the two maps, with some clusters looking quite similar. It is also interesting that the clusters are somewhat localized, despite not having given the model any information about longitude or latitude.

Though I would have liked the two clusterings to be more obviously related, the parameters and independent variables can be tweaked in the future for more specific results. The Foursquare API is somewhat limited as well; I would also have liked to extend the analysis to other cities, or a larger sample of venues with more descriptive categories.

In the future, another good idea would be to focus on specific ethnicities, both in food and in people in different areas, so it could cater to more specific immigrants.

Further decision making can be made if we included factors like the average price range of all venues, density of population, or I could use info from places like UberEats and explore at-home dining.

Although the analysis is far from comprehensive, it can give us a good idea of what the different foods communities are in Canada, along with the demographics of the people who live there.