Clustering Neighborhoods in Canada

Introduction

Canada is among the countries with a high immigration rates and in recent years it has become one of the most popular destination for educated professionals and skilled labor. In 2019 alone, Canada welcomed 341,000 immigrants and recently as per the new announcements it will be accepting over 400,000 immigrants per year between 2021-2023. This increase is to meet the economic growth targets and to replace the retiring workforce. Since Canada is a huge country standing at 9.985 Million Km^2, relocation for immigrants thus poses a huge challenge.

In this project, we will create a tool which will assist the immigrants in identifying different habitable and populated areas within Canada across all its provinces. In addition, we will segment the neighborhoods across different regions based on most common venues in the neighborhoods. This will enable the immigrants to quickly identify the regions and neighborhoods best suited for their purpose.

Data

To address the problem in question we will be using the location coordinates of major neighborhoods of Canada across all the provinces. This data can be downloaded for Canada from http://download.geonames.org/export/zip/. The data is in a tabular format so it can be read straight into pandas data frame. We will be using **Postal Code** (column 2), **Place Name** (column3), **Province Name** (column4), **Latitude** (column 10) and **Longitude** (column 11).

In addition, we will also be leveraging on Foursquare location data to pull in the most popular neighboring venues for each of the locations. We will achieve this through Foursquare APIs by passing the latitude and longitude location of each of the neighborhood into the ‘explore’ function below.

url= 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION, LATITUDE ,LONGITUDE, RADIUS, LIMIT)

Where:

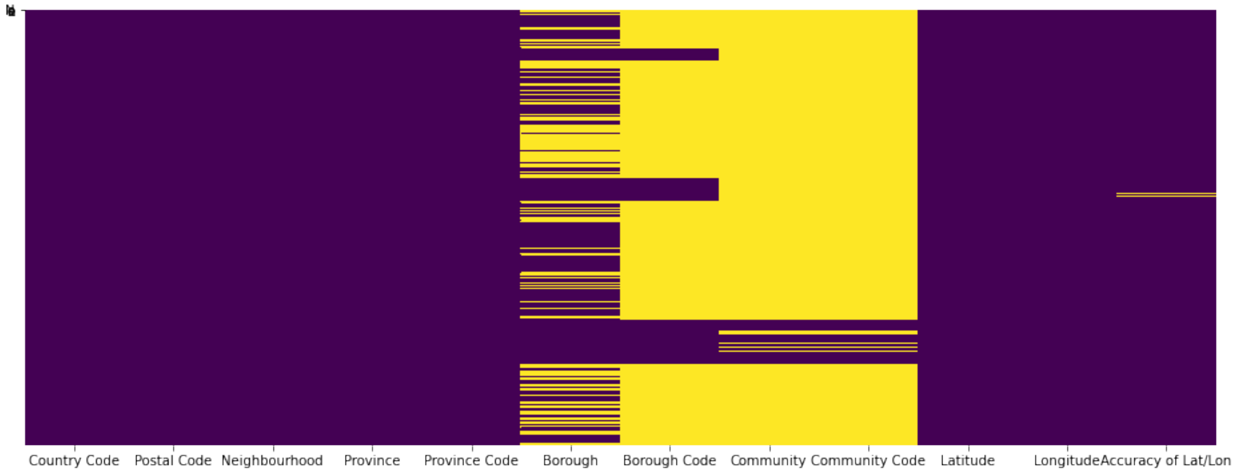
- CLIENT_ID and CLIENT_SECRET => Users credentials for Foursquare application
- LATITUDE, LONGITUDE => Location coordinates of the neighborhood
- RADIUS => Distance in meters within which to search for venues
- LIMIT => Maximum number of venue results to return

Data Wrangling and Cleaning

Clean dataset is a pre-requisite for any machine learning or data science problem. Therefore, after loading the data we will first try to clean the datasets. In doing so we will perform the following steps:

- Check for missing values
- Identify and drop irrelevant columns
- Drop rows with missing values

The visual below depicts the state of the missing data in our dataset. Here, the X-axis contains columns attributes, the Y-axis represents rows of the dataset and the missing data is colored in yellow.



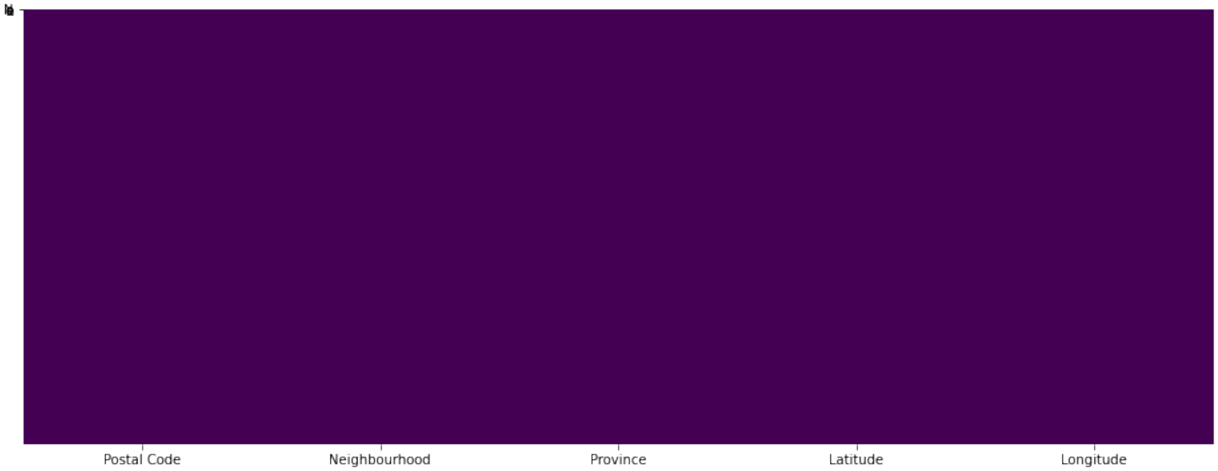
Missing values in Original dataset (size 1656,12)

It can be noticed that some columns such as **Borough**, **Borough Code**, **Community** and **Community Code** have significant rows with missing data and since none of these columns contains information relevant for this project we will drop these columns all together.

In addition, **Country Code** is not required since entire dataset is for Canada. **Province Code** is also redundant as we already have a column of **Province** name. Finally, as we are performing a holistic analysis for entire Canada we also do not need **Accuracy of Lat/Lon** column to differentiate between precise and approximated locations.

After, identifying these columns, in the next step we will drop these columns from dataset. We will also drop entire rows if any attribute value is missing, so that we get a cleaned dataset before developing our machine learning model.

The plot below illustrates final clean dataset, and it can be seen that there are no missing values in the dataset anymore.



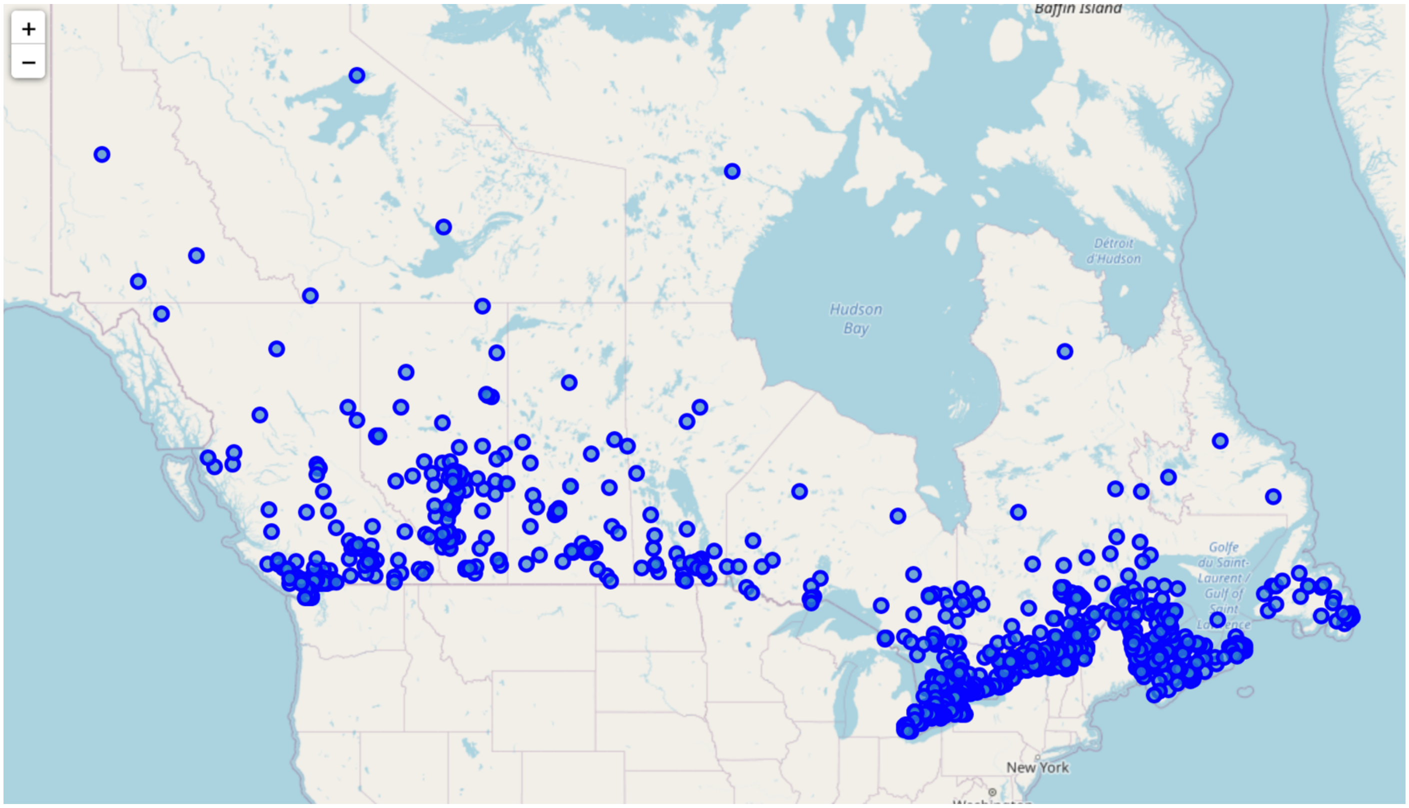
Cleaned dataset (size 1656,6)

Doing a quick check on the size of dataset we find that raw data had a size of 1656 rows and 12 columns and the cleaned dataset is of size 1656 rows and 6 columns. So basically, just by dropping irrelevant columns we are able to clean the dataset, without compromising on deleting any row.

Exploratory Data Analysis

Once, the data has been cleaned we will now perform some exploratory data analysis on data to understand it better. For this we will be using Folium visualization library.

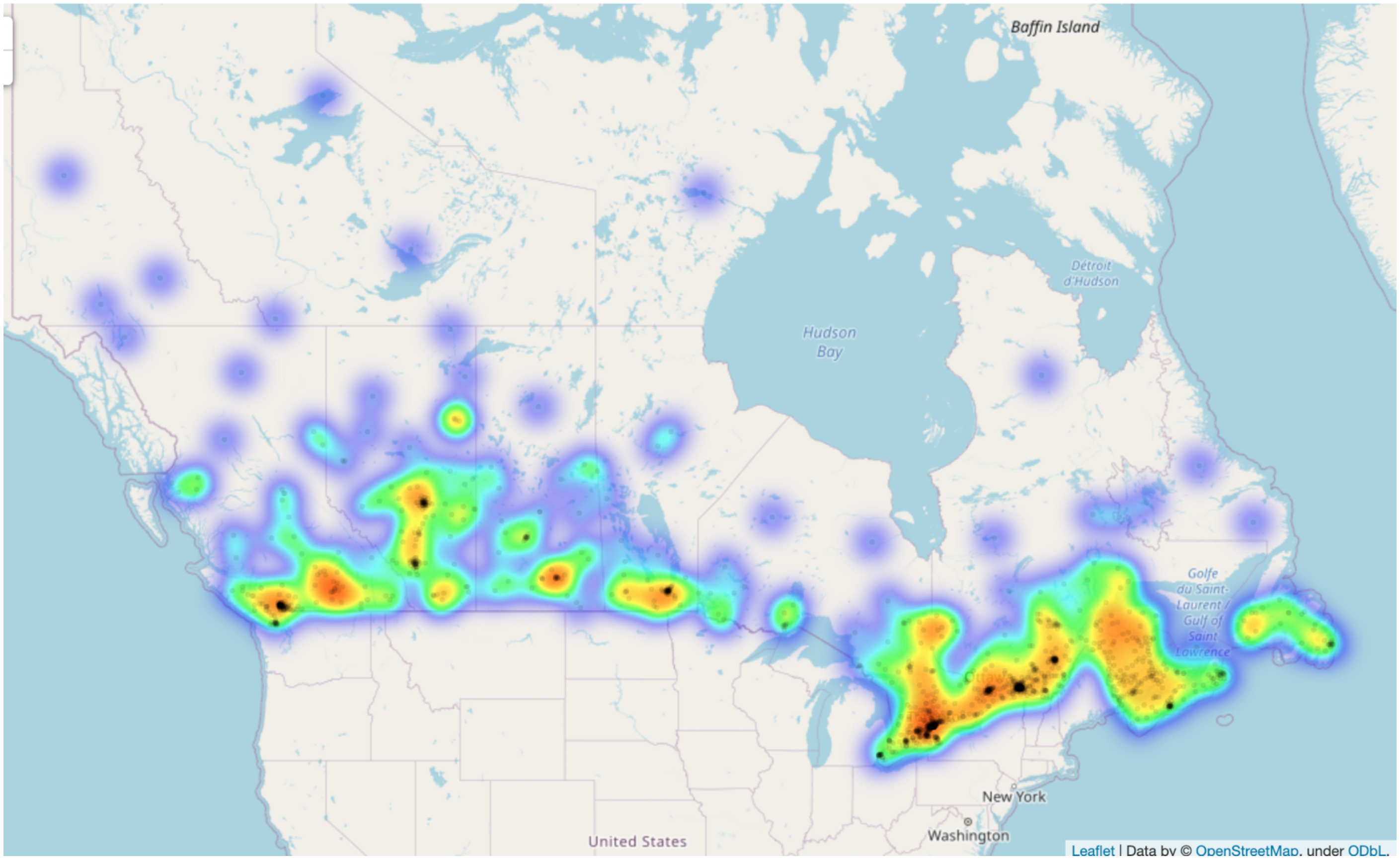
We will start by plotting all the neighborhoods in Canada on a map as shown below.



Neighborhoods plotted as points on map

The map above shows all the neighborhoods scattered across Canada. We can easily see that most of these neighborhoods are towards the South of Canada, as we move up towards the North they tend to get sparsely populated.

The scatter plot above may not be the best representation of the neighborhoods density so let's create a heat map of all the neighborhoods to easily visualize the information.


Baffin Island
Détroit d'Hudson
Hudson Bay
Golfe du Saint-Laurent / Gulf of Saint Lawrence
New York
Washington
United States
Leaflet | Data by © OpenStreetMap, under ODbL.

Heatmap showing density of populated neighborhoods plotted

From this heat map it gets very clear now that most of the popular neighborhoods are populated closer to the US border. From the density of black point on the map you could easily identify the major cities form West to East such as Vancouver, Calgary, Edmonton, Toronto, Ottawa, Montreal, Halifax. Combining the data science methodologies along with visualizations we are now able to understand popular neighborhoods across Canada. This kind of information is very useful for a new immigrant who needs to make a choice of which region to move to.

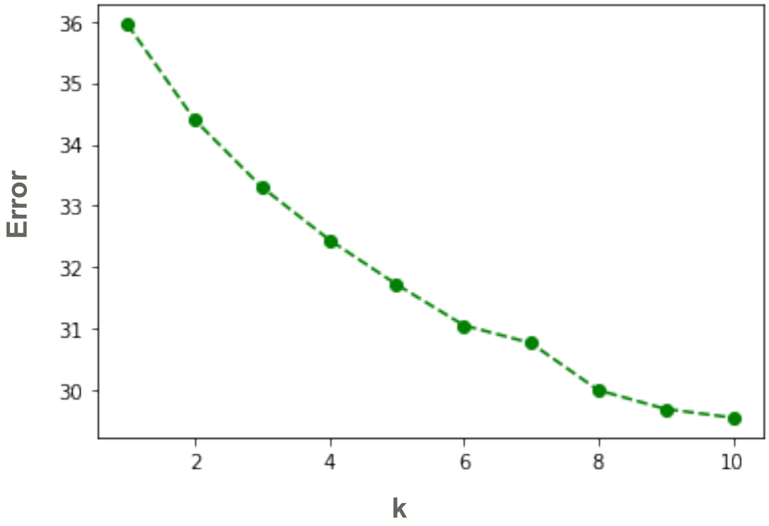
Next, we will find the most popular venues in each neighborhood and cluster them together to identify which neighborhoods across different cities in Canada are similar. Here we leverage the Foursquare APIs to extract information on popular venues around each neighborhood. Once we have this information we are all set to segment the data using an unsupervised machine learning model.

For this we will be using k-means clustering approach because of its simplicity and applicability for the problem in question. Also, we will be using the elbow method to identify optimum number of clusters to use. Following is the workflow used to run k-means clustering:

- Remove all neighborhoods which have less than 5 venues in vicinity. We are doing this to remove isolated neighborhoods and this way the clusters we create will be more meaningful for someone to examine and make relocation decisions
- One hot encode all categorical features
- Group rows per neighborhood by taking the mean of the frequency of occurrence of each venue category
- Filter top 5 venues for each neighborhood
- Run k-means clustering for different cluster values and pick the optimum using elbow method
- Run final clustering on the optimum cluster value picked from the elbow method.

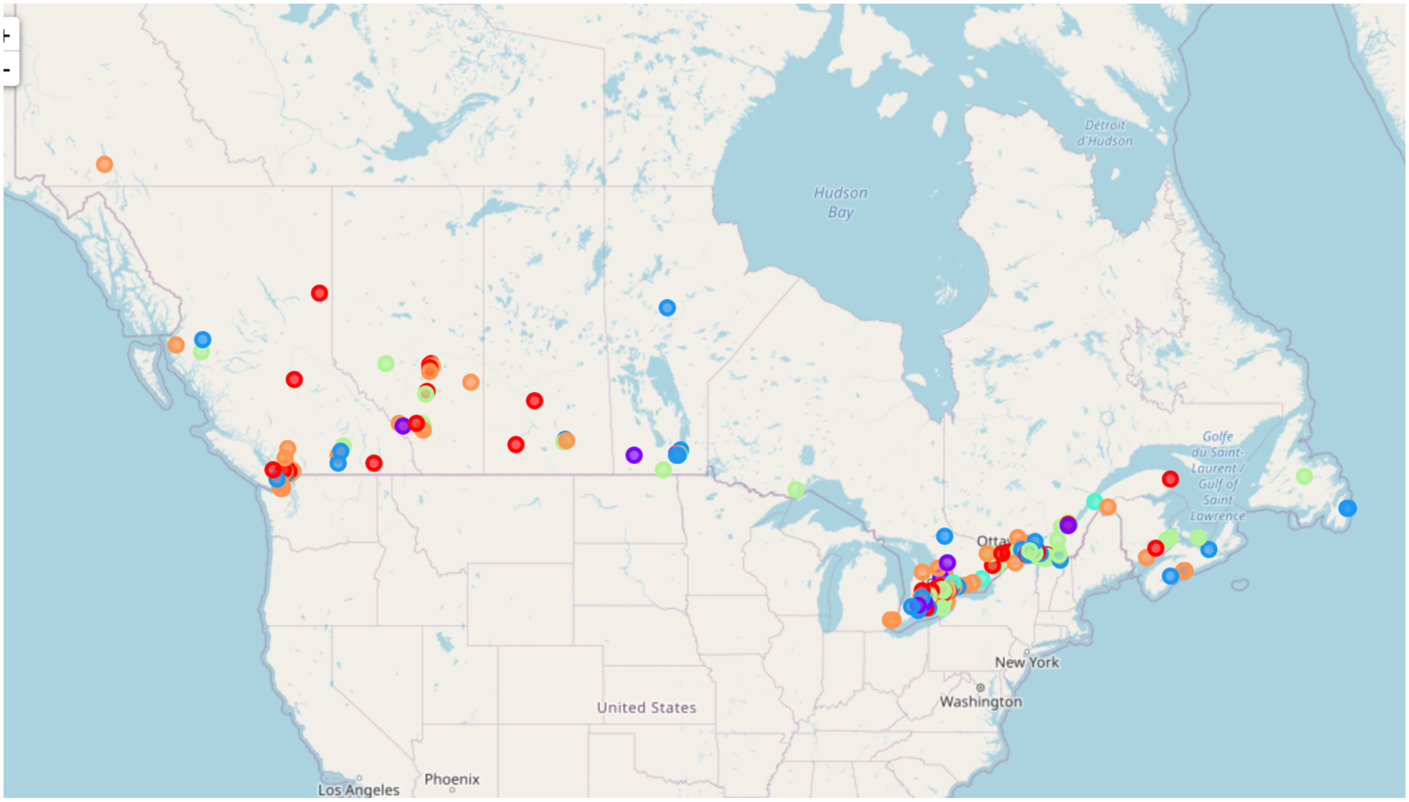
Results

The result of elbow method are shown below



Elbow Method to choose optimum k value

You can then see how the error decreases with an increasing number of clusters. However, each additional cluster provides a smaller net benefit. After k=6 the drop in error becomes small, so we pick 6 as our number of optimum clusters and re-run k-mean with it. Different clusters generate and their respective locations are shown below

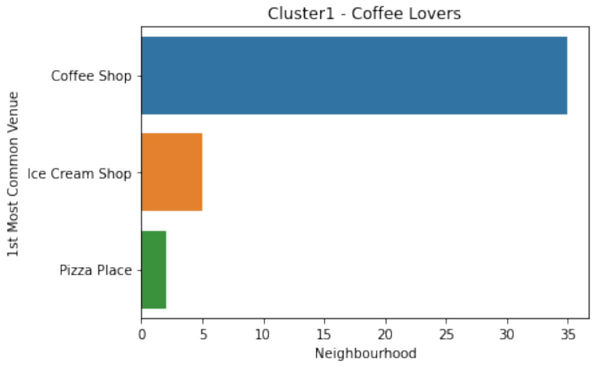


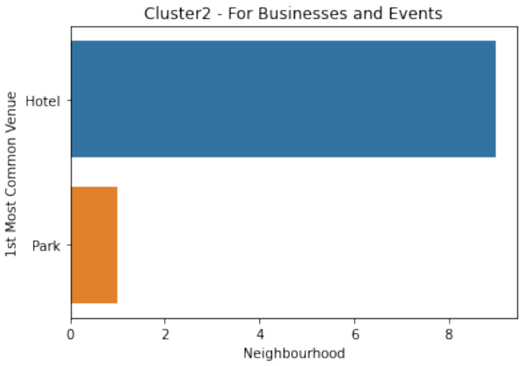
Clusters of popular neighborhoods in Canada

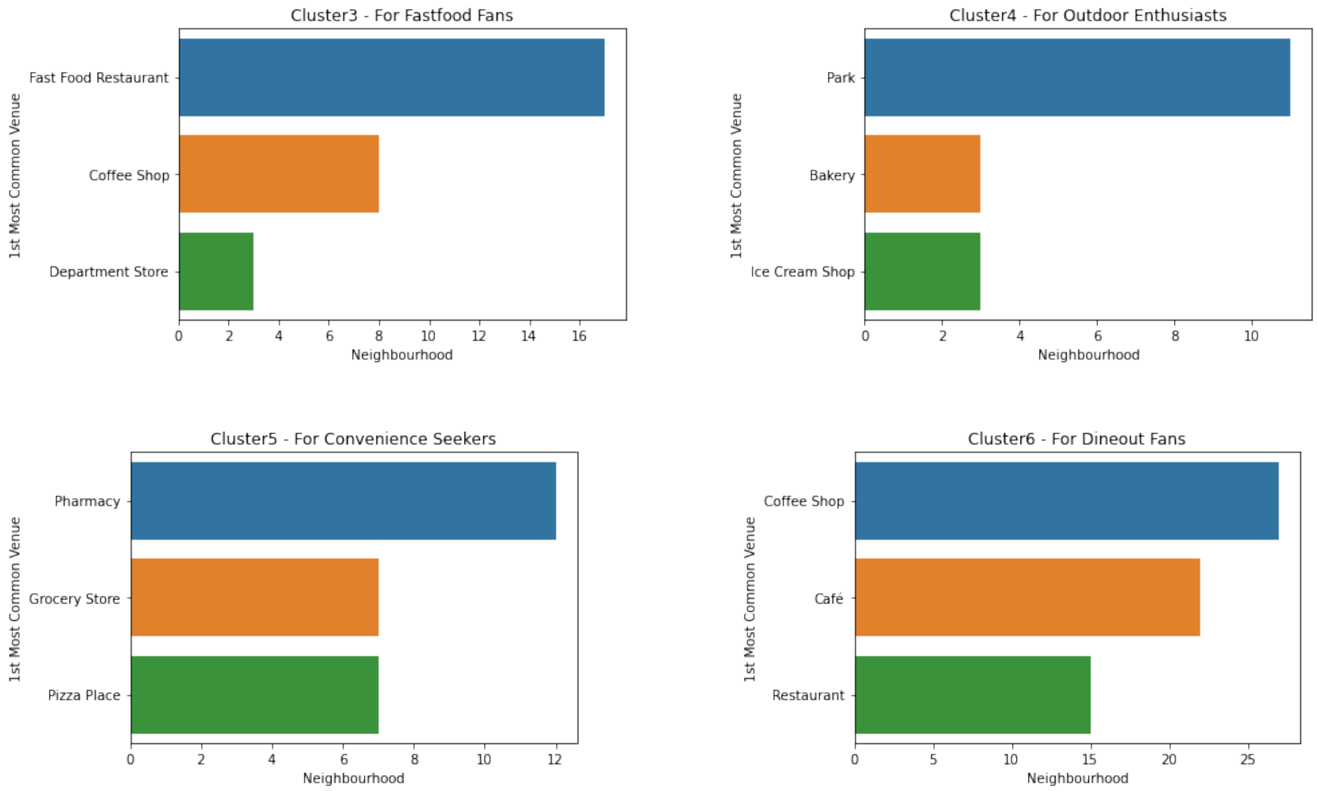
The map above shows the most popular neighborhoods in Canada colored by the cluster they belong to. It can be noticed that they most popular neighborhoods are in the major cities.

Discussion

Now let's explore and describe each of these clusters. To categorize the clusters let's look at the 1st most common venue in each and plot the results.







Top venues in each cluster

Based on these trends we can verbally describe clusters as follows:

- **Cluster 1:** Neighborhood popular with people who love coffee
- **Cluster 2:** Neighborhood with a lot of Hotels around, good for businesses and events
- **Cluster 3:** Neighborhood popular with people who prefer Fast food
- **Cluster 4:** Neighborhood popular with people who enjoy outdoors
- **Cluster 5:** Neighborhood with Pharmacies and Grocery store around for those who prefer convenience
- **Cluster 6:** Cluster 6 is somehow similar to Cluster 1, but also features Cafes and Restaurants

Conclusion

We saw how we were able to leverage huge amount of data and information to visualize and understand habitable and popular neighborhoods. This information is very useful for new immigrants when deciding which province or region to relocate to. Next, we segmented the neighborhoods into different clusters so that the immigrants can drill down into the region of their choice and choose the neighborhood which best suits their taste. We were able to identify 6 distinct neighborhood type and we also verbally named them based on the popular venues. With this application we are now able to facilitate anyone looking for a place to relocate.