IBM Data Science Capstone Project

Battle of Neighborhoods in the Province of Alberta, Canada

**Investigation of the best neighborhood for establishing new branches of Asian Chain Restaurant**

Spring 2021

Contents

[Introduction and Business Problem 2](#_Toc64791931)

[Target Audience 2](#_Toc64791932)

[Data Overview 2](#_Toc64791933)

[List of Neighborhoods 3](#_Toc64791934)

# Introduction and Business Problem

Nowadays every business owner has a dream of expanding his/her business globally. The business analysts helping these companies start with the largest cities in the country of destination then as the business starts its profitability they go after smaller cities in the same country. What they do is that they fist look for the neighborhoods of the target city. Then analyze each neighborhood to find out where is the best spot for them to establish new branches of their business.

I own an Asian chain restaurant that has been around in several major cities across the glob with more than 1000 branches. The company wants to expand its branches to more major cities in the world. Their plan is for the first phase to go for cities with a population of above one million and in the next phases focus on smaller cities. The project manager asked me to investigate cities with a population of above one million in the province of Alberta, Canada to find out where are the best neighborhoods to establish their new branches.

In the province of Alberta, there exist two cities with the criteria of having population above one million. These two cities are Edmonton and Calgary. Thus, in this project I will be using Foursquare location data to find out which neighborhoods are the best choice in Edmonton and Calgary to establish new branches.

# Target Audience

The audience of this project is the Asian chain restaurant who is looking for expanding its branches in major cities in the province of Alberta, Canada.

# Data Overview

To investigate the neighborhoods in two major cities of the province of Alberta Canada, we need to acquire six sets of data:

1. List of the neighborhoods in Edmonton
   1. This set of data must be retrieved from Wikipedia webpage.
2. Geographical location of each neighborhoods in Edmonton
   1. I use Georecorder package in python to export this set of data.
3. Venue data including the existing venues in each neighborhood in Edmonton
   1. I use Foursquare to retrieve venue data.
4. List of the neighborhoods in Calgary
   1. This set of data must be retrieved from Wikipedia webpage.
5. Geographical location of each neighborhoods in Calgary
   1. I use Georecorder package in python to export this set of data.
6. Venue data including the existing venues in each neighborhood in Calgary
   1. I use Foursquare to retrieve venue data.

## List of Neighborhoods

To have the list of the neighborhoods I have used Wikipedia webpage as a source. I first found the postal codes of neighborhoods in Alberta which is starting with T. Then searched for the list of the postal codes in Canada within the wiki webpage and as shown in Figure 1, I retrieved the list for both Edmonton and Calgary neighborhoods.

Table

Description automatically generated

Figure 1: A screenshot of list of neighborhoods in different cities in the province of Alberta, Canada.

Since data was in a suitable format for this project, I have scrapped it, cleaned Not Assigned neighborhoods, and exported them to two separate data frames one for Edmonton and the other for Calgary.

Graphical user interface, table

Description automatically generated

Figure 2: Edmonton and Calgary neighborhood data frames after retrieving and scrapping

Using the Geocoder package, I have added geospatial coordinates of centroids of each neighborhood to the Edmonton and Calgary neighborhood data frames. Figure 3 shows an example of it for Edmonton.

Graphical user interface, text, application, email

Description automatically generated

Figure 3: Adding geospatial coordinates to the data frame using Geocoder package

Then after, I used Foursquare website to export venues in each neighborhood in Edmonton and Calgary. Figure 4 shows the list of venues by neighborhoods in the city of Edmonton.

Table

Description automatically generated

Figure 4: Venue data retrieved and stored into a data frame from the Foursquare webpage for the City of Edmonton.

# Methodology

## Cleaning the data

The data collected from the Wikipedia website had some Borough and Neighborhoods where there were no information about their other field. For instance, in the data frame that was created for the Edmonton neighborhoods, there were some neighborhoods where their Boroughs were listed as “Not Assigned”. The same was true for some of the data in the Calgary data frame. Thus what I did was:

1. Cleaned the rows where the Borough column in that row was “Not Assigned”.
2. Cleaned the rows where the Neighborhood for that row was “Not Assigned”.
3. Combined all the rows with the same postal codes into one group of Neighborhoods.

Then after that I used Geocoder to assign latitude and longitude of centroid of each area with the same postal code to each row of the data frame for both Edmonton and Calgary. Figure 5 shows that for Calgary.

Graphical user interface, text, application, email

Description automatically generated

Figure : Calgary data frame after performing the cleaning process.

## Data Exploration

After cleaning the data frame, I used Folium to create a map of Edmonton and Calgary with the neighborhoods centroid being highlighted using circles. In Figure 6 you can see how I used Folium to show the neighborhoods on the map.

Map

Description automatically generated

Figure : Neighborhoods centroid geospatial locations for Edmonton and Calgary

After visualizing the neighborhoods for both Edmonton and Calgary, I used the Foursquare webpage to extract all the required information for venues existing in each neighborhood. Figure 7 and Figure 8 represent first five rows of the extracted information about venues in each neighborhood in Edmonton and Calgary, respectively that are attached to the data frame I already created using wiki page.

Graphical user interface, application

Description automatically generated

Figure : Information extracted from Foursquare for venues in Edmonton area.

Graphical user interface, application

Description automatically generated

Figure : Information extracted from Foursquare for venues in Calgary area.

## Machine Learning

I implemented One hot encoding technique to transfer all the categorical data in the data frame into numerical data. For each specific neighborhood in each city, each venue turned into a frequency and the table got updated based on that as shown in Figure 9.

Calendar

Description automatically generated

Figure : Using One hot encoding technique to transfer all the categorical variables into numerical.

Then I grouped the data by averaging on the frequency of occurrence of each venue category as shown in Figure 10.

Table

Description automatically generated

Figure : grouped data over each venue category for each neighborhood.

Then top ten most common venues at each neighborhood was extracted and listed in a data frame as shown in Figure 11.

Graphical user interface, table

Description automatically generated with medium confidence

Figure : Top 10 most common venues in each neighborhood.

## K-Means Clustering

In this part of the research, I implemented K-Means clustering algorithm to cluster similar neighborhoods in a cluster. To do so, I needed to find the best K for the data of Edmonton and Calgary. So, I used the Elbow Point technique to find the best K for each data frame of Edmonton and Calgary. Figure 12 shows the Elbow Point analysis of the K.

Chart, line chart

Description automatically generated

Figure : Finding the bet K for K-means clustering.

After finding the best K for each data set, for the next step I created a new data frame for each city that includes Neighborhoods, frequency of Asian restaurants in that neighborhood, and the cluster label for that neighborhood. Figure 13 shows the new data frame for both Edmonton and Calgary.

Graphical user interface, application

Description automatically generated

Figure : New data frame created with the cluster label for each neighborhood.

I created a new table by merging tables in Figure 13 with the venue data for each neighborhood to be able to start analyzing possibility of establishing new Asian restaurant in each neighborhood for the two major cities in the province of Alberta, Canada. Then I used Folium to create a map with a circle at the centroid of each neighborhood color coded based on the cluster label for that specific cluster. Figure 14 shows it for both cities.

Map

Description automatically generated

Figure : Edmonton and Calgary maps using Folium and after performing K-Means clustering. Cluster 1 is Red, 2 is Purple, 3 is Turquoise, and 4 is Dark Khaki.

## Data Analysis

Based on my machine learning work I have a total of four cluster for Edmonton and three cluster for Calgary. The first step here is to see how many neighborhood I have in each cluster and what is the frequency of Asian restaurant in each cluster. Using Matplotlib bar graph, I plotted number of neighborhoods and frequency of Asian restaurant in each cluster.

Figure 15 shows how many neighborhood I have in Edmonton and Calgary.

Chart

Description automatically generated

Figure : Number of neighborhoods in each cluster for Edmonton and Calgary

As per shown by Figure 15, in Edmonton there are 32 neighborhoods in cluster 1, and one neighborhood in each cluster from cluster 2 to cluster 4. The same figure shows that in Calgary we have 29 neighborhoods in cluster one, one neighborhood in cluster 2 and 3, and 3 neighborhood in cluster 4.

Figure 16 shows how what is the frequency of Asian restaurant in each cluster for Edmonton and Calgary.

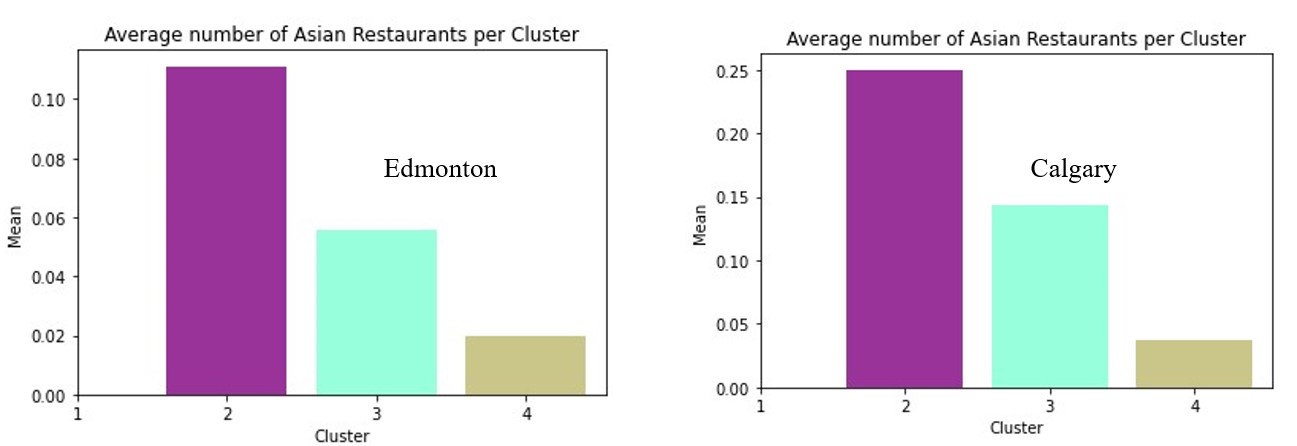


Figure : Frequency of Asian Restaurant in each cluster

As shown in Figure 16, cluster 1 for both Edmonton and Calgary has zero frequency for Asian restaurants. Cluster two for both cities has the highest frequency, and cluster 3 and cluster 4 are next clusters for both.

## Cluster Analysis

In this section I am analyzing each specific cluster to find out how it looks like in terms of Asian restaurant.

## Cluster 1

For both Edmonton and Calgary there is no Asian restaurant in cluster 1. Despite the fact of having the most neighborhoods in cluster 1 for Edmonton (32 neighborhood) and for Calgary (29 neighborhood), neighborhoods in this cluster do not have any Asian restaurant in them.

## Cluster 2

Cluster 2 for Edmonton and Calgary has only one neighborhood. This cluster also has one record of Asian restaurant as well. In Figure 17 we see that Edgemont City in the Calgary and Pinoy Grill in Edmonton are the only restaurants in cluster 2 for each city. Another important thing is that highest average of Asian restaurants is happening in cluster 2 for both of the cities with Edmonton (0.11) and Calgary (0.25).

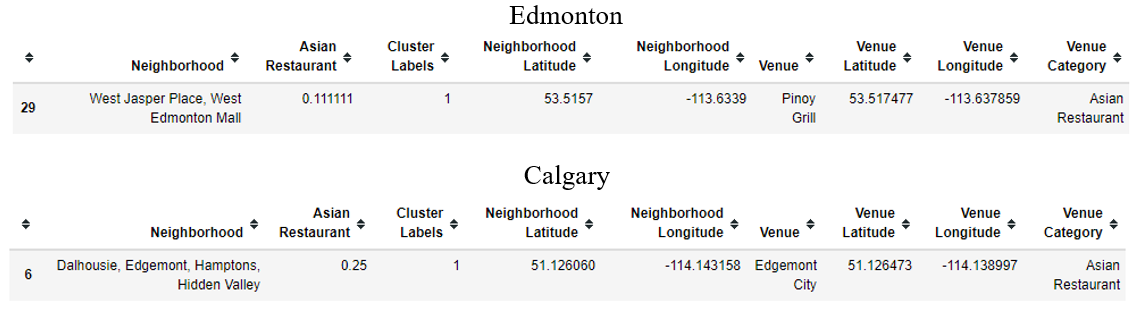


Figure : Number of restaurants in Cluster 2 for Edmonton and Calgary

## Cluster 3

As shown in Figure 18, in cluster 3 for Edmonton and Calgary there is only one Asian restaurant with an average of 0.056 in Edmonton and 0.14 in Calgary.

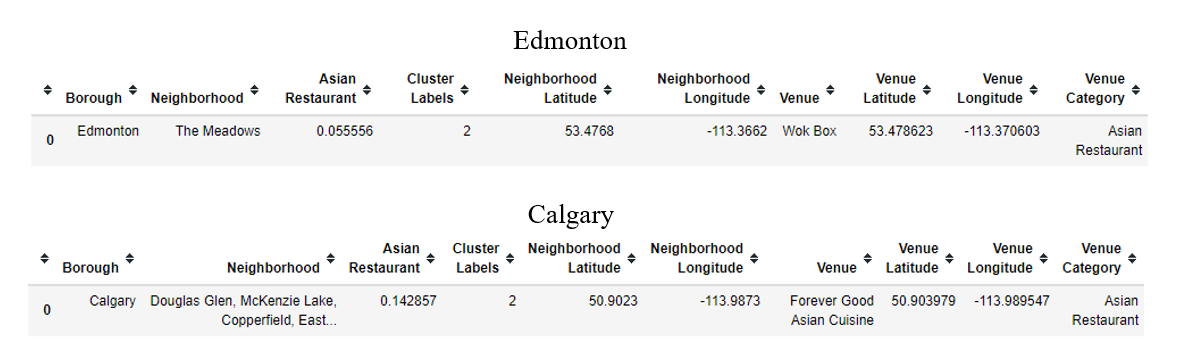


Figure : Number of restaurants in Cluster 3 for Edmonton and Calgary

## Cluster 4

As shown in Figure 19, in cluster 4 for Edmonton there are two Asian restaurant with an average of 0.02 and for Calgary there are four restaurant with an average of 0.03 to 0.048.

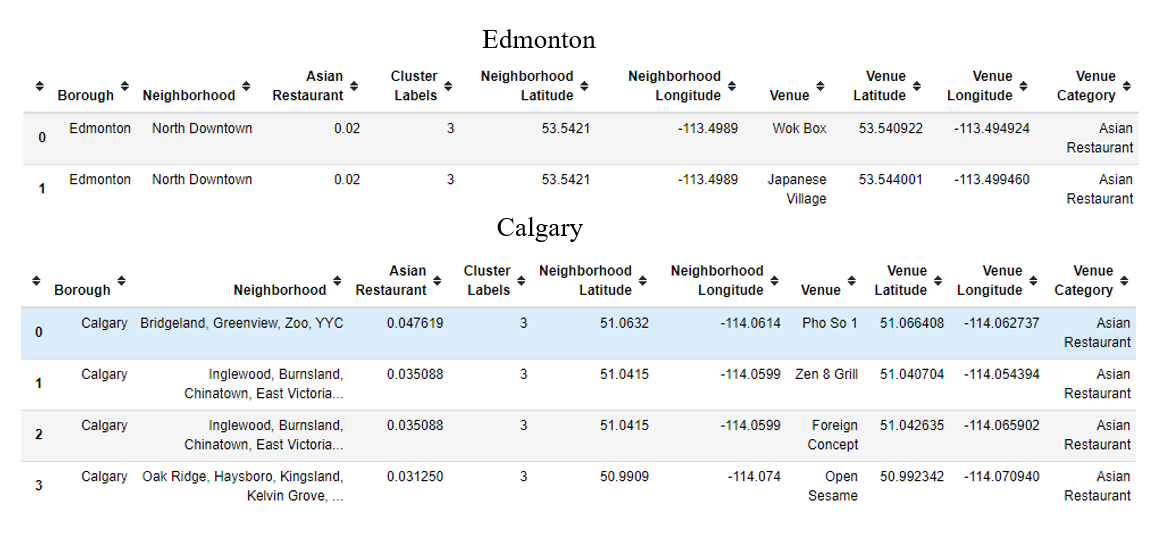


Figure : Number of restaurants in Cluster 4 for Edmonton and Calgary

Based on the cluster analysis, cluster 1 with highest number of neighborhoods and lowest average for Asian restaurant is the best cluster for both Edmonton and Calgary to establish a branch.

# Discussion

Most of the neighborhoods fall into cluster 1 in Edmonton and Calgary. However, there is no Asian restaurant in either cities in cluster 1.

Edmonton has 32 neighborhoods in cluster 1 with zero average of Asian restaurants in this cluster. Calgary has 29 neighborhoods in this cluster with no Asian restaurant.

Cluster 2, however, is mostly dense area with number of Asian restaurant. There is only one neighborhood in cluster 2 of Edmonton and Calgary but the average number of Asian restaurant is 0.1 for Edmonton and 0.24 for Calgary.

# Conclusion

In this project I used python programming language with its numerous libraries, Wikipedia, foursquare, ant etc. to examine where is the best neighborhood in the city of Edmonton and city of Calgary to establish new branches of the Asian chain restaurant.

For both of the cities, neighborhoods falling into cluster 1 are the best neighborhoods to look for a place to establish a new branch of our restaurant.