Table of contents

The Battle of Neighborhoods

-Where to start a Sushi restaurant in Calgary, Alberta?

1. Introduction

Calgary is a city in the western Canadian province of Alberta and it is situated about 80 km east of the Canadian Rockies. Calgary has a population of 1.3 million, making it Alberta's biggest city and the second largest in western Canada after only Vancouver.

As a multi-cultural city, Calgary has numerous cuisines from all over the world. Personally, I have been in Calgary for 15 years and love to try different cuisines and experience different flavours. Among them, I love the Sushi best.

Thus, the goal of this project is to study the neighborhoods in Calgary and determine the right places to start a Sushi restaurant. The project can be useful to some business owners who are looking to open up a new Sushi restaurant in Calgary.

To solve this problem, we need to acquire enough demographic data and food venue data of each neighborhood and analyze these data using some machine learning algorithm to provide recommendations for interested business owners and entrepreneurs.

2. Data Acquisition

We will need to collect the community demographic data from the web and food venue data from foursquare API.

2.1 Neighborhood demographics

To collect the data, we will start with the community demographics data. I will use two data sources to get these data. Pandas web scraper library will be used to get these publicly available data from the web.

1. Neighborhood and area from Wikipedia

   It seems not easy to get the community population density data after many unsuccessful tries and I decided to calculate it using the population and the area of a community. The below link provides the community and its area($km^{2)}$ in Calgary:

   https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary

	Community	Area
0	Abbeydale	1.7
1	Acadia	3.9
2	Albert Park / Radisson Heights	2.5
3	Altadore	2.9
4	Alyth/Bonnybrook	3.8
5	Applewood Park	1.6
6	Arbour Lake	4.4
7	Aspen Woods	3.8
8	Auburn Bay	4.5
9	Aurora Business Park	2.4

Fig 1

2. Neighborhood and demographics data from great-news.ca

Great-news.ca has a fairly complete dataset for Calgary community demographics data. The link here: https://great-news.ca/demographics/

	Community	Median Household Income	Population	Area	PopulationDensity
0	Abbeydale	55345.0	6071	1.7	3571.176471
1	Acadia	46089.0	10969	3.9	2812.564103
2	Albert Park / Radisson Heights	38019.0	6529	2.5	2611.600000
3	Altadore	53786.0	9518	2.9	3282.068966
4	Applewood Park	65724.0	6864	1.6	4290.000000
5	Arbour Lake	70590.0	10987	4.4	2497.045455
6	Aspen Woods	133939.0	7496	3.8	1972.631579
7	Auburn Bay	84350.0	11127	4.5	2472.666667
8	Banff Trail	49996.0	4204	1.5	2802.666667
9	Bankview	32474.0	5416	0.7	7737.142857

Fig 2

3. Geo data from Geocoder

We also need coordinates for each neighborhood to explore its venues using foursquare API. We will use geocoder library to get the coordinates for each neighborhood.

	Community	Median Household Income	Population	Area	PopulationDensity	Latitude	Longitude
0	Abbeydale	55345.0	6071	1.7	3571.176471	51.05976	-113.92546
1	Acadia	46089.0	10969	3.9	2812.564103	50.97227	-114.05882
2	Albert Park / Radisson Heights	38019.0	6529	2.5	2611.600000	51.04200	-113.99683
3	Altadore	53786.0	9518	2.9	3282.068966	51.01601	-114.10558
4	Applewood Park	65724.0	6864	1.6	4290.000000	51.04544	-113.92513
5	Arbour Lake	70590.0	10987	4.4	2497.045455	51.13364	-114.20307
6	Aspen Woods	133939.0	7496	3.8	1972.631579	51.04519	-114.21160
7	Auburn Bay	84350.0	11127	4.5	2472.666667	50.88976	-113.96397
8	Banff Trail	49996.0	4204	1.5	2802.666667	51.07472	-114.11297
9	Bankview	32474.0	5416	0.7	7737.142857	51.03412	-114.10044

Fig 3

2.2 Food venues in neighborhood

With the community dataset, we are ready to explore all the venues of a neighborhood using foursquare API.

Since our goal is to identify a Sushi restaurant location, we will focus on the food venues within 1500m (a reasonable walking distance) of a community.

To set the section parameter for foursquare API, we will use keyword **section=food** and follow this:

url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&***section={}***&limit={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION, lat, lng, radius, ***SECTION***,LIMIT)

	Community	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbeydale	51.05976	-113.92546	Atlas Pizza and Sports Bar	51.052481	-113.941859	Pizza Place
1	Abbeydale	51.05976	-113.92546	A&W	51.068291	-113.933571	Fast Food Restaurant
2	Abbeydale	51.05976	-113.92546	Subway	51.059239	-113.934423	Sandwich Place
3	Abbeydale	51.05976	-113.92546	Subway	51.069623	-113.932907	Sandwich Place
4	Abbeydale	51.05976	-113.92546	Subway	51.052786	-113.942449	Sandwich Place

Fig 4

3.Methodology

In this section, we will discuss the details of methodology used in this project.

3.1 Exploratory Data Analysis

To better understand the collected data, various exploratory data analysis was carried out and visualized in different charts.

Fig 5 shows the food venue distribution per category in Calgary.

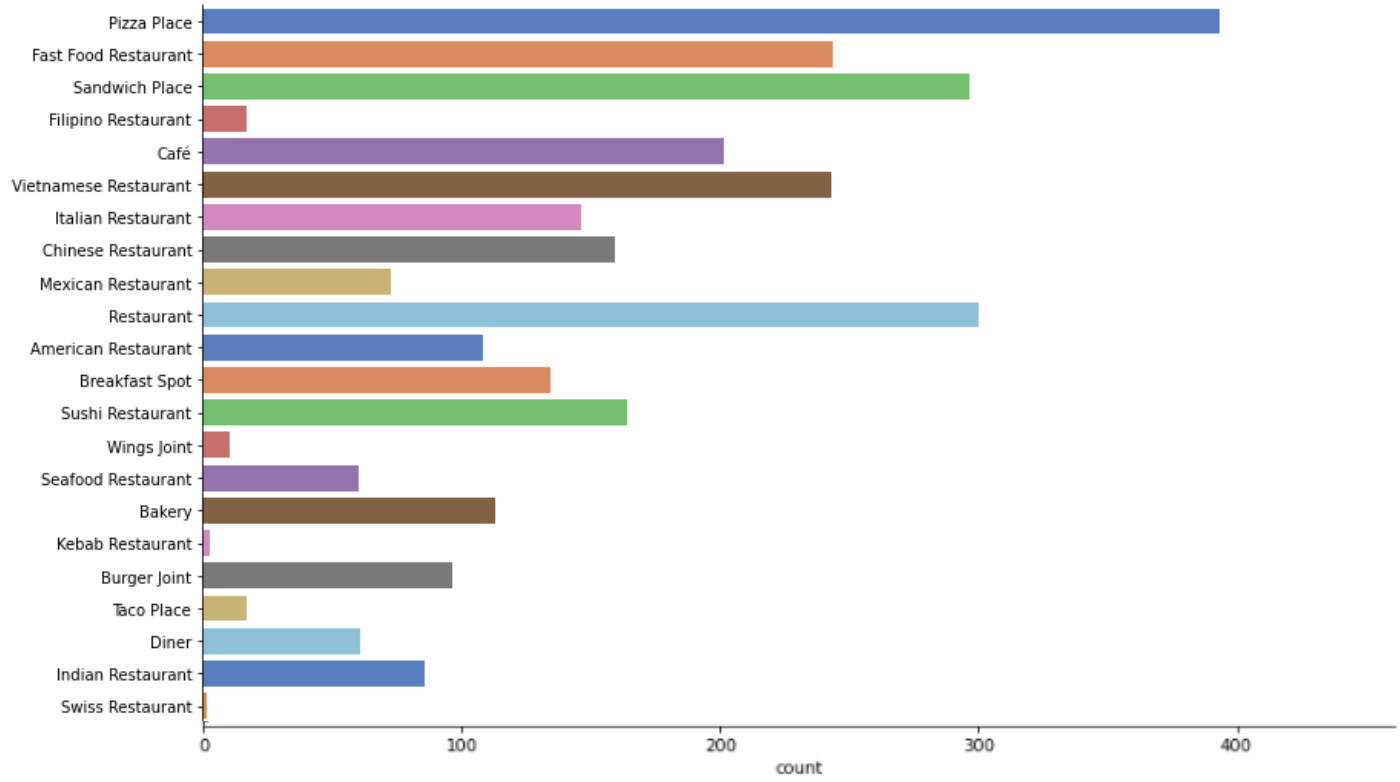


Fig 5

Figure 6 shows food venue distribution per community. We can clearly see that some community like Sunalta has much high number (indicating a fully developed mid-town community) compared to other community like Woodbine/Woodlands (rural communities).

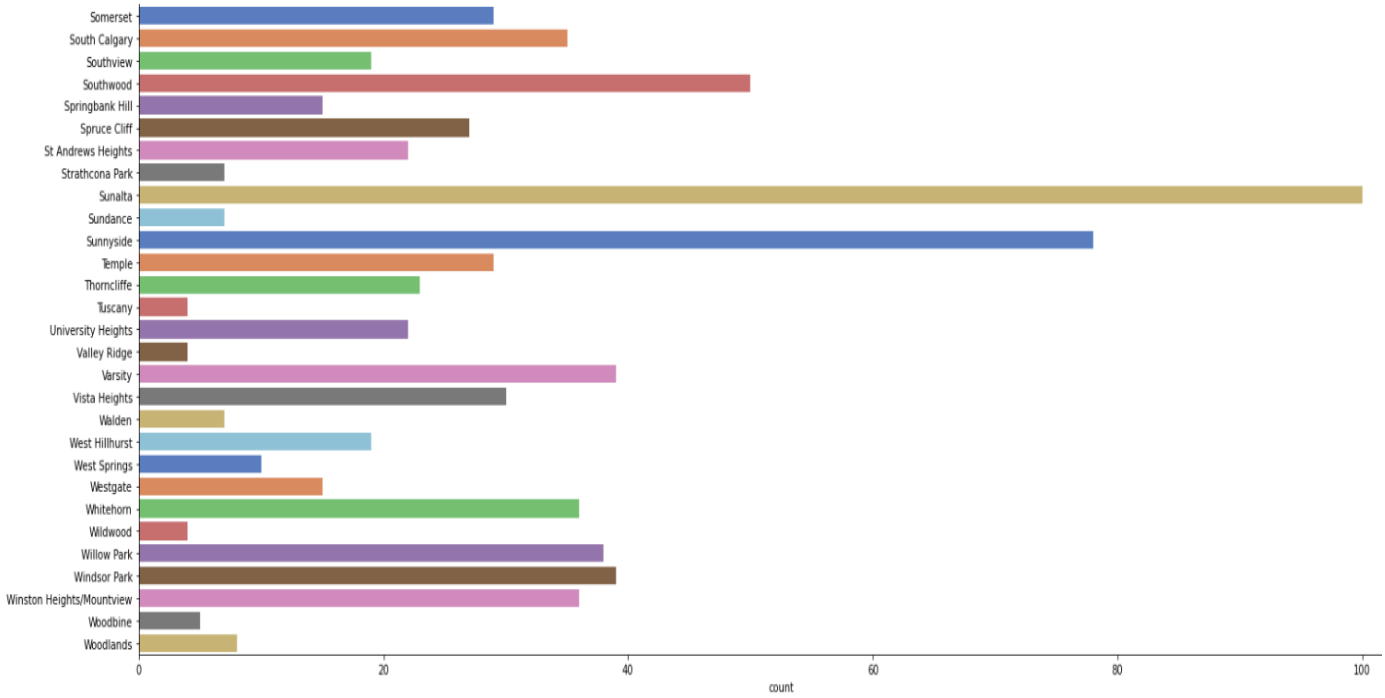


Fig 6

3.2 Feature Engineering

To identify the best location for a new Sushi restaurant, we need to analyze the data, such as, median household income, population density, total food venues (which can largely indicate if the neighborhood is fully developed, well developed, or underdeveloped). Since the competition to a Sushi restaurant is mostly from East Asian cuisines, we will also need to study the coverage of other East Asian restaurants, such as Chinese restaurants, Japanese restaurants, Dim Sum restaurants, and etc.

To extract all the venue related data, we will do a one-hot encoding on the venue category once we got all the food venues for each community, Then we can easily get the occurrence frequency of a food venue category in a community by getting the group mean of each food venue category for that community.

After that, we can easily to extract the occurrence frequency of any cuisine in a community.

Fig 7 shows that our feature set is:

Median Household Income	PopulationDensity	venueCount	Chinese Restaurant	Japanese Restaurant	Sushi Restaurant	Dim Sum Restaurant

	Community	Median Household Income	PopulationDensity	venueCount	Chinese Restaurant	Japanese Restaurant	Sushi Restaurant	Dim Sum Restaurant
0	Abbeydale	55345.0	3571.176471	12.0	0.083333	0.0	0.00	0.00
1	Acadia	46089.0	2812.564103	50.0	0.020000	0.0	0.04	0.02
2	Albert Park / Radisson Heights	38019.0	2611.600000	27.0	0.000000	0.0	0.00	0.00
3	Altadore	53786.0	3282.068966	20.0	0.050000	0.0	0.15	0.00
4	Applewood Park	65724.0	4290.000000	8.0	0.000000	0.0	0.00	0.00

Fig 7

3.3 K-means clustering

1. Data normalization with max-min scaler

In data analytics, we usually need to do a feature rescaling to make sure features are on almost the same scale before we feed the data to the machine-learning algorithms. In this project, I will use max/min scaler to achieve this.

$$featureDF\_scaled = \frac{featureDF - featureDF.min()}{featureDF.max() - featureDF.min()}$$

	Median Household Income	PopulationDensity	venueCount	Chinese Restaurant	Japanese Restaurant	Sushi Restaurant	Dim Sum Restaurant
0	0.161939	0.319253	0.12	0.208333	0.0	0.00	0.00
1	0.101824	0.250325	0.50	0.050000	0.0	0.16	0.36
2	0.049412	0.232065	0.27	0.000000	0.0	0.00	0.00
3	0.151813	0.292985	0.20	0.125000	0.0	0.60	0.00
4	0.229347	0.384566	0.08	0.000000	0.0	0.00	0.00

Fig 8

2. K-Means clustering

K-means clustering is an unsupervised machine learning algorithm that aims to partition N instances into k clusters in which each instance belongs to the cluster with the nearest mean. After careful observation, I came to a conclusion that K-Means clustering fits very well with my project goal.

3. Find K with Elbow method

A fundamental step for K-means clustering is to determine the optimal number of clusters into which the data may be clustered. We will use the Elbow method determine this optimal value of k. We will use 2 different measurements to calculate the elbow number.

Distortion: It is calculated as the average of the squared distances from the cluster centers of the respective clusters.

Inertia: It is the sum of squared distances of samples to their closest cluster center.

We iterate the values of k from 1 to 12 and calculate the distortion and inertia for each value of k. See the below graphs.

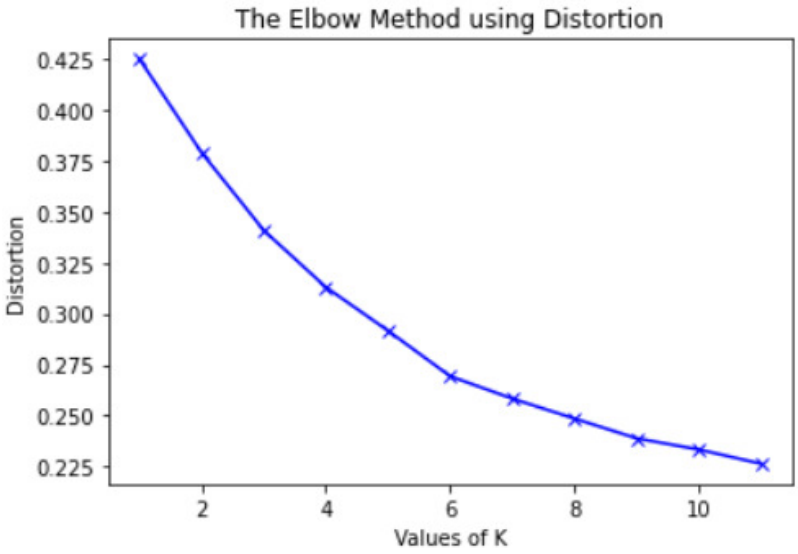


Fig 9

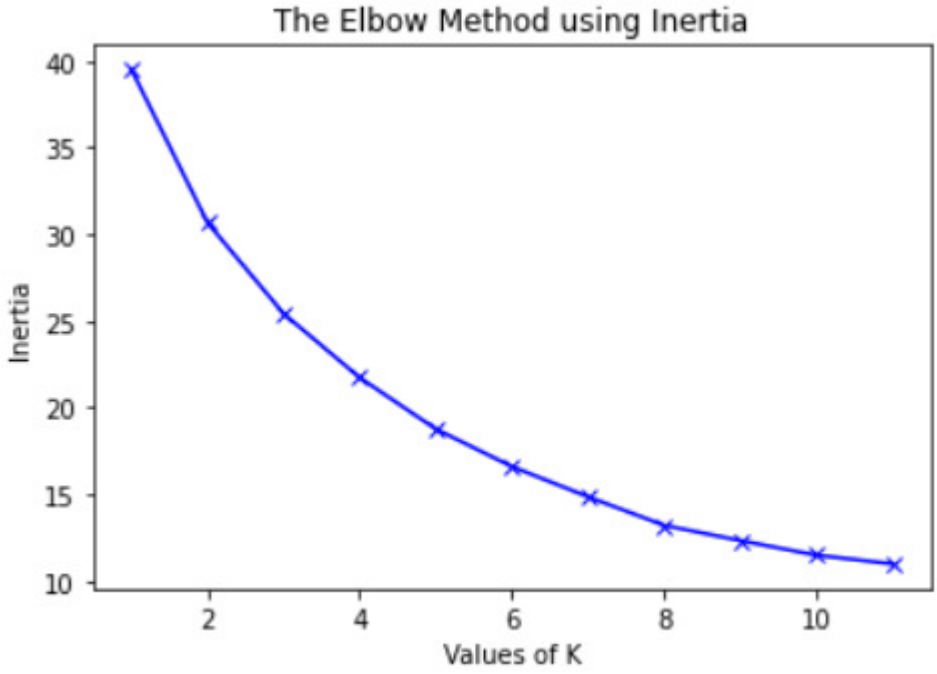


Fig 10

To determine the optimal number of clusters, we have to select the value of k at the "elbow", where the point after which the distortion/inertia start decreasing in a linear fashion. Thus, for the given data, we can conclude that the optimal number of clusters for the data is 5 or 6. However, it is not very easy to explain the clustering results if we use k = 6, so I decided to keep it at k = 5.

4. Clustering

The code to implement the K-means clustering is rather simple as below.

```
# set number of clusters
kclusters = 5

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(calRestDF_scaled)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([0, 0, 0, 1, 0, 1, 2, 2, 0, 4])
```

4. Results

In this section, we will present the clustering result in 3 different forms. First let's check how many neighborhoods are distributed in each cluster. As we can see, cluster 0, cluster 1 accounts for most of the neighborhoods in Calgary. Cluster 2, 3, and 4 each owns fairly small number of communities.

4.1 Neighborhood distributions per cluster

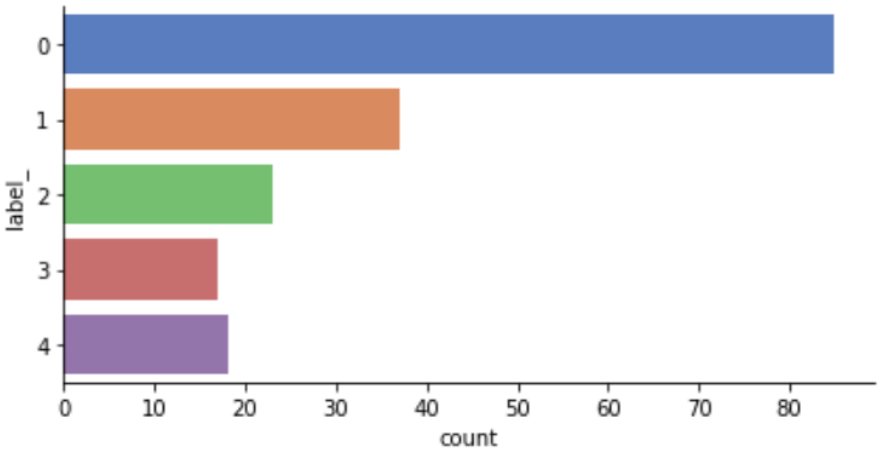


Fig 11

4.2 Map the clustering result

To help users fully grasp the clustering result, we can show the clustering result in a map as shown below.

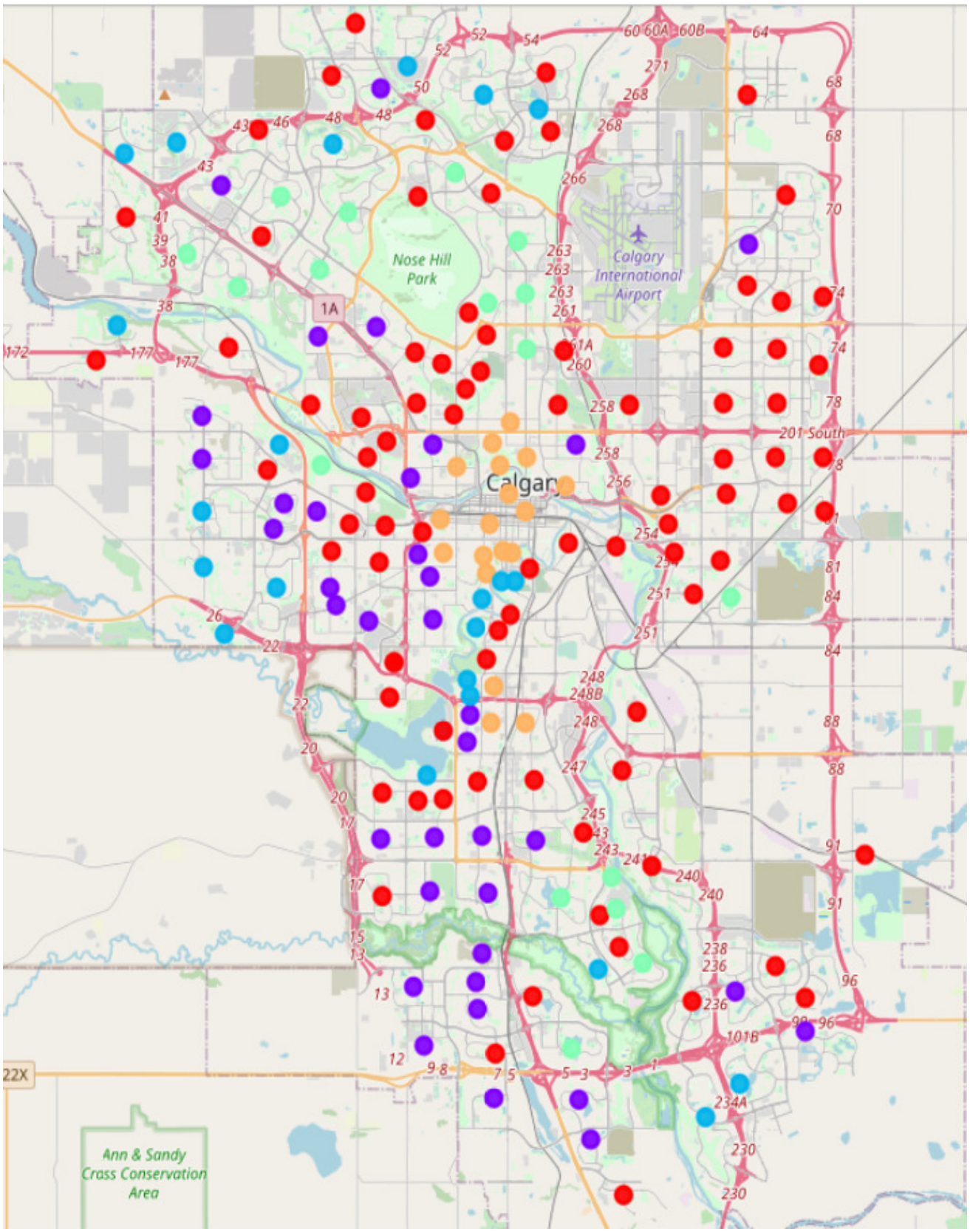
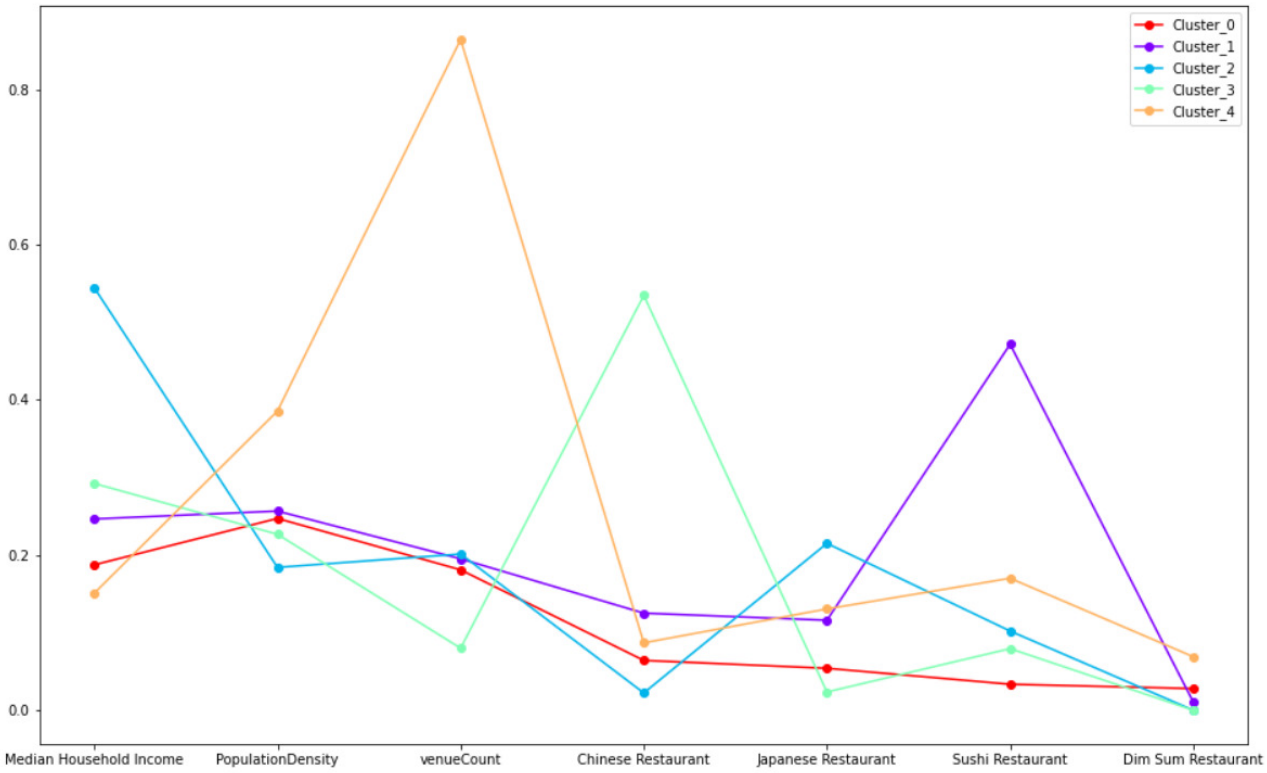

Nose Hill Park
Calgary International Airport
Calgary
Ann & Sandy Crass Conservation Area
1A
201 South
22X

Fig 12

4.3 Cluster line graphs

To visualize the clustering result, we can calculate the cluster mean on each measuring feature and show them in the graph below. We will discuss the graph details in the next section .



Cluster graphs show how each cluster performs on each measuring matrix on X-axis
Fig 13

5. Discussion & recommendations

Looking at the Fig 13, let's start to analyze each line graph:

Cluster 0(Red): This cluster represents the communities with low family income, medium population density, fair overall food venue coverage and very low East Asian restaurant coverage.

Cluster 1(Purple): This cluster represents the communities with medium family income, medium population density, medium overall food venue coverage, and with fair Chinese/Japanese restaurant coverage, but highest Sushi restaurant coverage.

Cluster 2(Blue): This cluster represents the communities with highest family income, lowest population density and fair overall food venue coverage, but with highest Japan restaurant coverage and low Sushi restaurant coverage.

Cluster 3(Green): This cluster represents the communities with high family income, lower density and lowest overall food venue coverage, highest Chinese restaurant coverage, but low Sushi restaurant coverage

Cluster 4(Orange): This cluster represents the communities with lowest family income, highest population density and overall food venue coverage, high East Asian restaurant coverage.

Base on the above clustering analysis, I can have the following ***recommendations:***

Cluster 0: ***Not recommended*** to start a new Sushi restaurant in this cluster even if communities in this cluster have a low East Asian restaurant coverage, due to its low family income, fair overall food venue coverage.

Cluster 1: ***Not recommended*** to start a new Sushi restaurant in this cluster due to its highest Sushi restaurant coverage (which means fierce competition for a new start-up), medium population density, and medium family income.

Cluster 2: ***Highly recommended for a high-end Sushi restaurant***. Communities in this cluster have highest family income, low Sushi restaurant coverage. However, the new Sushi restaurant will face some competition from Japanese restaurants in the community.

Cluster 3: ***Highly recommended for a high-end Sushi restaurant***. Communities in this cluster have high family income, low Sushi restaurant coverage. However, the new Sushi restaurant will face some competition from Chinese restaurants in the community.

Cluster 4: ***Highly recommended for low-end fast-food like Sushi restaurant***. Communities in this cluster have lowest family income, highest population density and overall food venue coverage. A low-end Sushi restaurant will thrive in this cluster which are mainly consisted of fully developed downtown or hub communities.

6. Conclusion

In this project, community demographic data and food venue data are collected and cleaned. Feature extraction was used to help finding the intrinsic features of a neighborhood from the food venue data.

The final feature dataset was normalized using max-min scaler and an unsupervised machine learning algorithm, K-means clustering, was employed to cluster the final dataset.

By analyzing the clustering result, we provided different recommendations for different business goals:

A high-end Sushi restaurant should be located in communities in cluster 2 or cluster 3;

A low-end Sushi restaurant should be started in communities in cluster 4.

7. Future Considerations

1. Foursquare API limit

   Foursquare API only provides up to 100 venues in a community, there are some cases in which a downtown/hub community could have more than 100 venues and we just lose all those venue data exceeding 100. With a paid account, I probably could get more complete venue data which will help to provide a more accurate clustering result.

2. Data limit

   In this project, I only considered family income, population density, overall food venue coverage and East Asian restaurant coverage in a community. There are some other factors, like median

age, median rent, proximity to C-train line and more factors, which could affect the final business decision. Given more time and more data, I could continue to improve the clustering model and provide better analysis and recommendations for the business owners.