THE BEST NEIGHBOURHOOD TO START A SUSHI RESTAURANT IN CALGARY

BATTLE OF THE NEIGHBOURHOOD | CAPSTONE PROJECT

ABSTRACT

THE BATTLE OF THE NEIGHBOURHOOD ANALYSIS AS PART OF THE ADVANCED DATA SCIENCE CAPSTONE PROJECT. IN THIS ANALYSIS, WE ANALYSE THE NEIGHBOURHOODS IN THE CITY OF CALGARY TO FIND THE BEST ONES FOR STARTING A SUSHI RESTAURANT

RAHUL T N U NAIR

ADVANCED DATA SCIENCE CAPSTONE PROJECT

Table of contents

The Battle of Neighborhoods

IDENTIFY THE BEST NEIGHBOURHOOD, IN THE CITY OF CALGARY, TO START A JAPANESE SUSHI RESTAURANT

1. Introduction

The Economist Intelligence Unit ranked Calgary the most livable city in North America in both 2018 and 2019. Calgary has been a top 5 contender for this title for the last ten years. Calgary was also ranked the best city in the world for drivers in 2019.

Calgary's economy includes energy, financial services, film and television, transportation and logistics, technology, manufacturing, aerospace, health and wellness, retail, and tourism sectors. The Calgary Metropolitan Region is home to Canada's second-highest number of corporate head offices among the country's 800 largest corporations. In 2015 Calgary had the highest number of millionaires per capita of any major Canadian city. In 1988 it became the first Canadian city to host the Winter Olympic Games.

Given the city's diverse socio-economic and demographic nature, starting a restaurant here might be a proposition that several business owners might consider. And what if, It is a Japanese Sushi restaurant?

Keep this hypothetical customer job in mind. The objective of the project is to determine the right places to start a Sushi restaurant.

The outcomes of this project will serve as a useful analysis for business owners seeking to open up a new Sushi restaurant in Calgary.

Problem Which Tried to Solve:
The major purpose of this project is to suggest the best neighbourhood in the city of Calgary to start a Japanese Sushi Restaurant

2. Data Acquisition

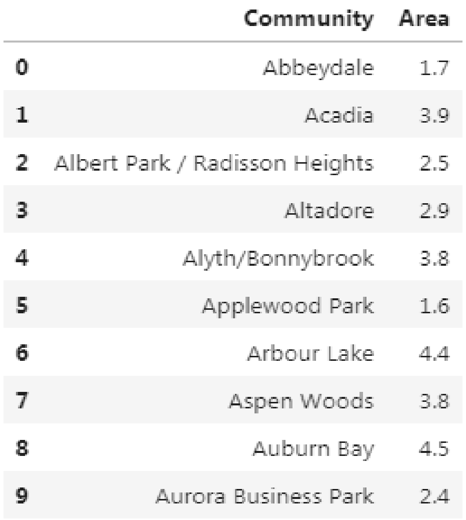
The analysis will require collecting the community demographic data from the foursquare API from the web and food venue data.

2.1 Neighbourhood demographics

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

1. Neighbourhood and area from Wikipedia

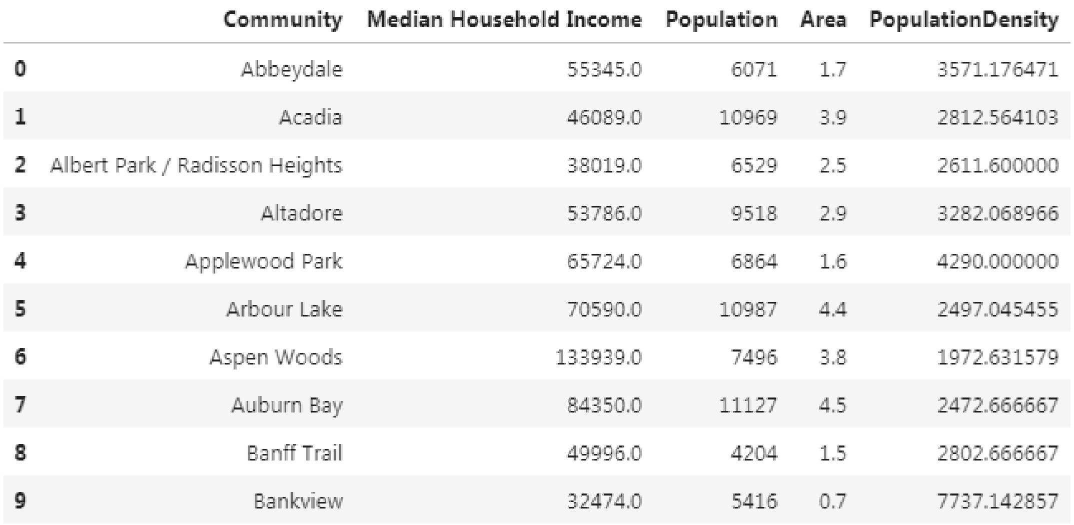
We used the link below to scrap the community and its area(km$^{2)}$ in Calgary: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary



	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

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

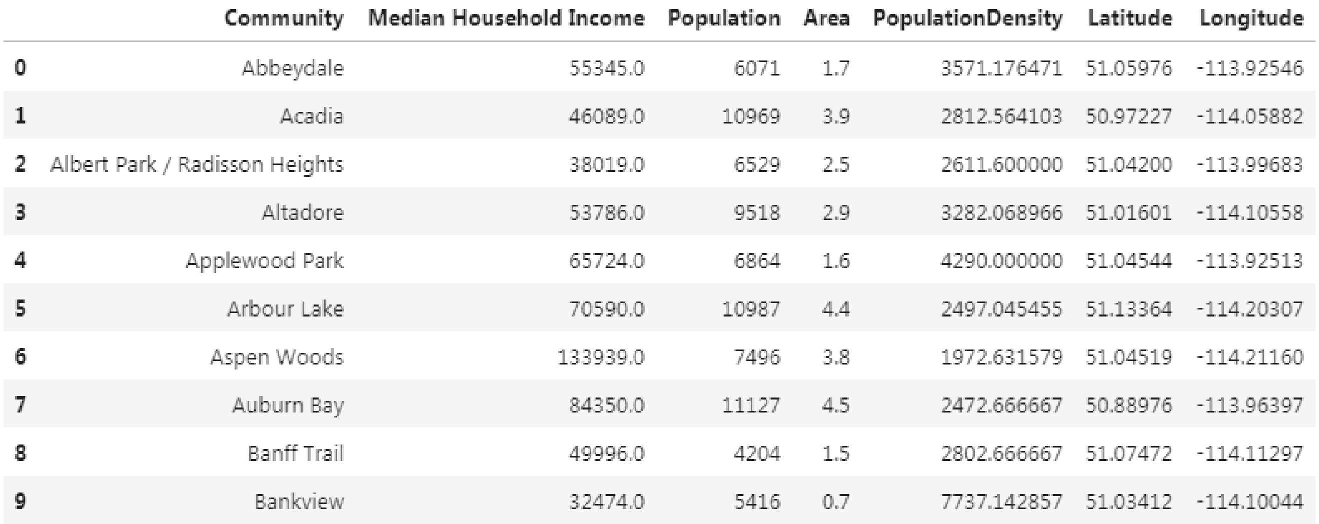
We used the following link to scrap Calgary's community demographics data. https://great-news.ca/demographics/



	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

3. Geodata from Geocoder

Using FourSquare API, we will capture the coordinates for each neighbourhood to explore its venues. Geocoder library was used to get the coordinates for each neighbourhood.



	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

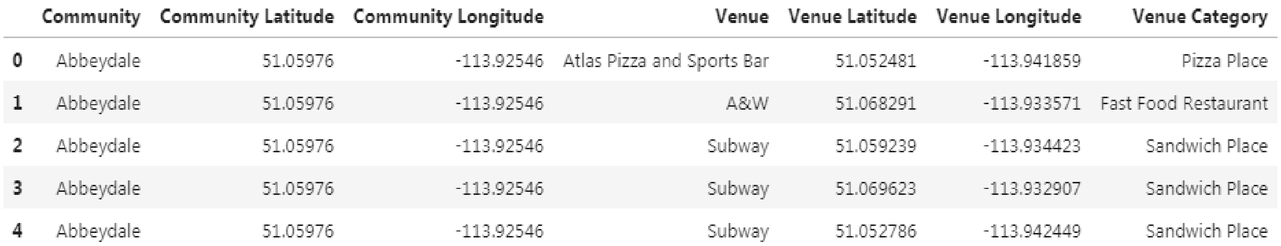
2.2 Food venues in the neighbourhood

Given we now have the community dataset, we will explore all the neighbourhood venues using the foursquare API.

Our objective is to identify the best neighbourhoods for starting a sushi restaurant. We will focus on the food venues within1500m (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)



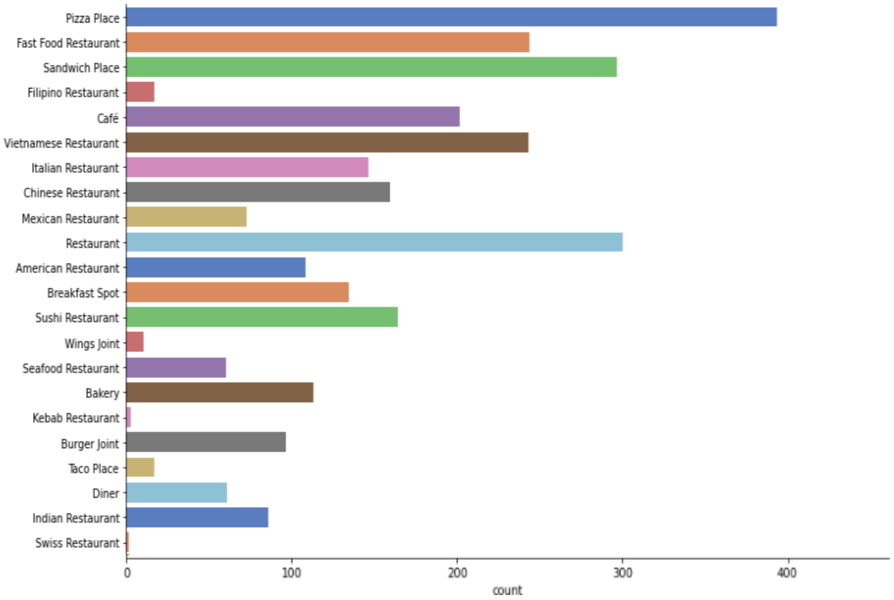
	Community	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbeydale	51.05976	-113.92546	Atlas Pizza and Sports Bar	51.052481	-113.941859	Pizza Place
1	Abbeydale	51.05976	-113.92546	A&W	51.068291	-113.933571	Fast Food Restaurant
2	Abbeydale	51.05976	-113.92546	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

3. Methodology

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

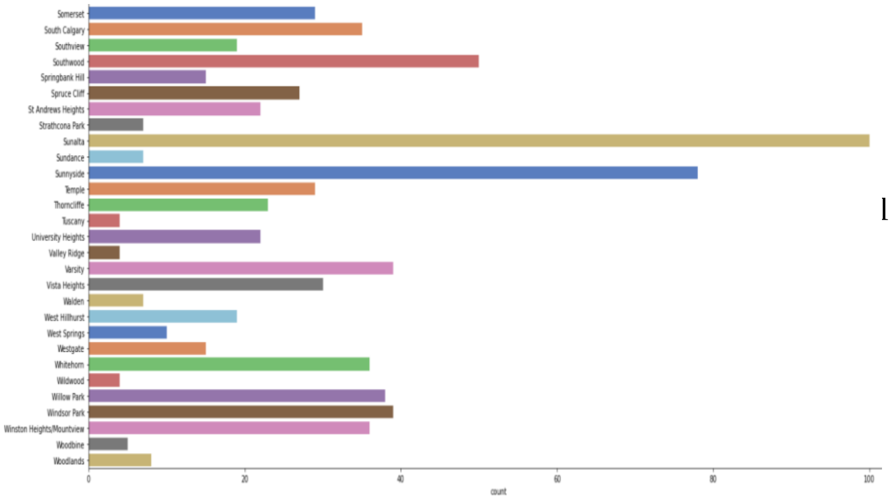
3.1 Exploratory Data Analysis

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



The figure to the right shows the food venue distribution per category in Calgary.

We can observe that there is a good supply of Sushi Restaurants in the market with a count of ~200, and it is comparable to the number of Chinese restaurants in the city



The figure to the right shows the food venue distribution per community.

We can 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).

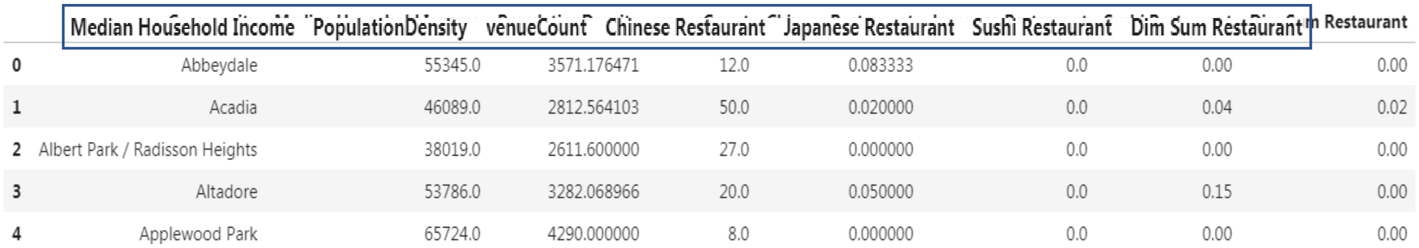
3.2 Feature Engineering

To identify the best location for a new Sushi restaurant, we will be analyzing additional data points including: medianhousehold income, population density, total food venues (which can largely indicate if the neighbourhood 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 East Asian restaurants as a category which includes Chinese restaurants, Japanese Restaurants, Dim Sum restaurants, etc.

Once we got all the food venues for each community, we will do a one-hot encoding on the venue category to extract all the venue-related data. 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.

After that, we can easily extract the occurrence frequency of any cuisine in a community.
Fig 7 shows that our feature set is:



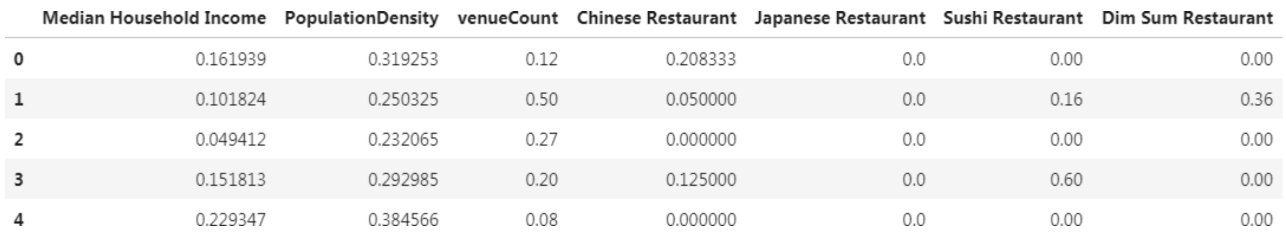
	Median Household Income	PopulationDensity	venueCount	Chinese Restaurant	Japanese Restaurant	Sushi Restaurant	Dim Sum Restaurant	n Restaurant
0	Abbeydale	55345.0	3571.176471	12.0	0.083333	0.0	0.00	0.00
1	Acadia	46089.0	2812.564103	50.0	0.020000	0.0	0.04	0.02
2	Albert Park / Radisson Heights	38019.0	2611.600000	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

3.3 K-means clustering

1. Data normalization with max-min scaler

In this analysis, we will be using the max/min scaler to ensure all the features are on almost the same scale before we feed the data to the machine-learning algorithms.

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



	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

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. K-Means clustering fits very well with the overall projectgoal.

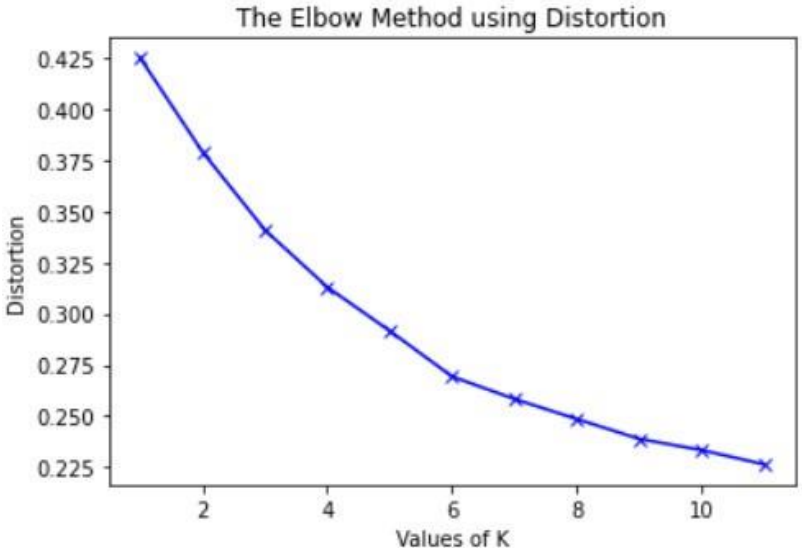
3. Find K with Elbow method

We have used the Elbow method to determine this optimal value of, k. and we have used two2 different measurements to calculate the elbow number.

Distortion: the average of the squared distances from the cluster centres of therespective clusters.

Inertia: the sum of squared distances of samples to their closest cluster centre

Values o k iterated from 1 to 12, and distortion and inertia calculated for each value of k, presented in the chart below.



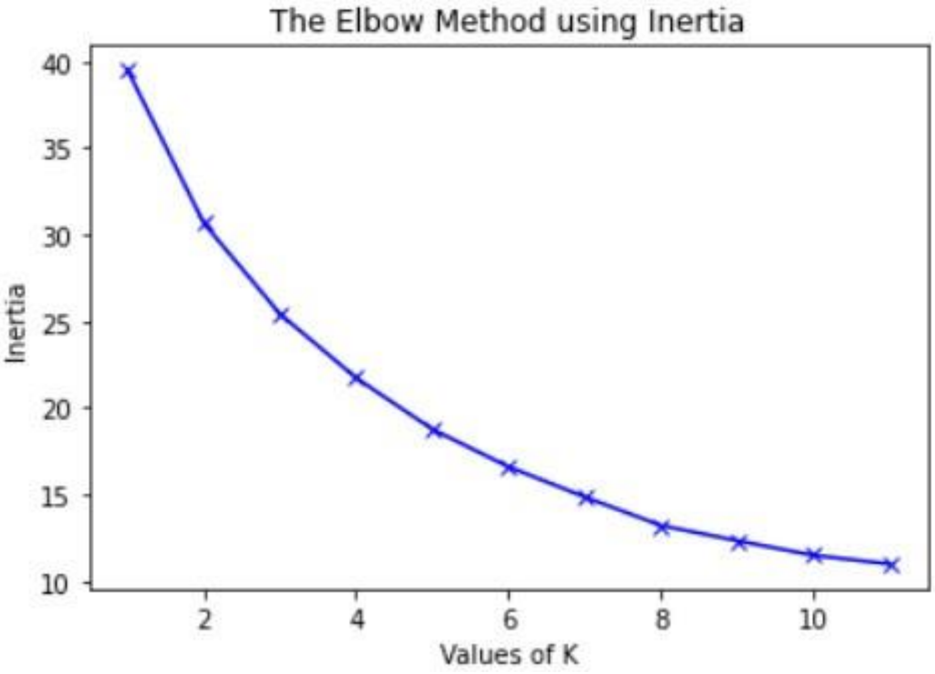
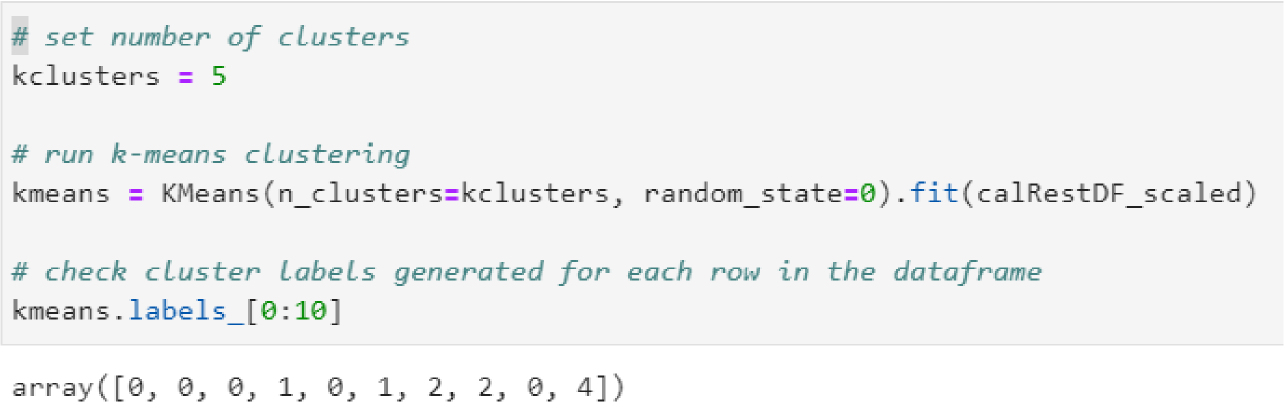


Fig 10

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 line linearly or the given data, the optimal number of clusters for the data is 5 or 6. However, to keep the analysis less complex, the analysis uses k = 5.

4. Clustering

e code to implement the K-means clustering is 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]
```

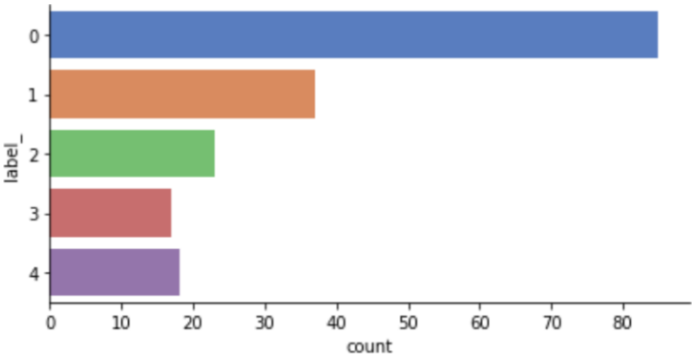
```
array([0, 0, 0, 1, 0, 1, 2, 2, 0, 4])
```

4. Results

This section presents the clustering result in 3 different forms.

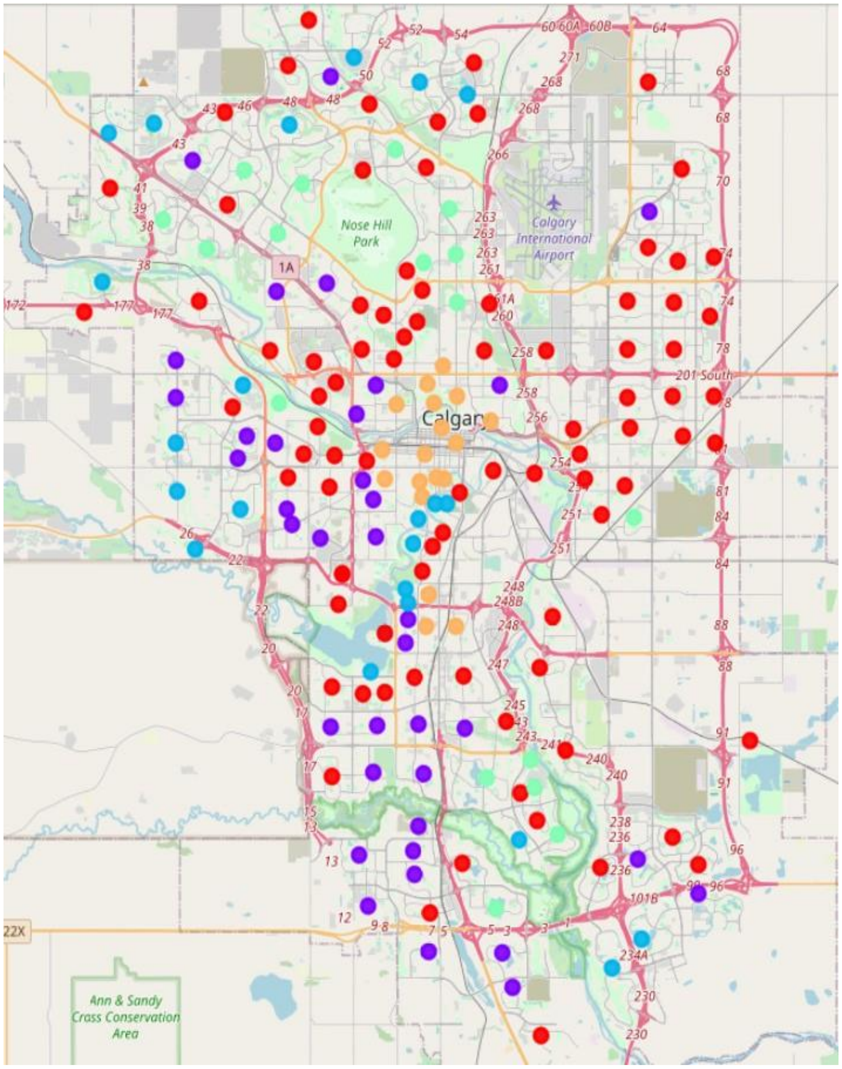
4.1 Neighborhood distribution cluster

First: we looked at how many neighbourhoods are distributed in each cluster. As we can see, cluster 0, cluster 1 accounts for most of the neighbourhoods in Calgary. Cluster 2, 3, and 4 each owns a fairly small number of communities.



4.2 Map the clustering result

Second: for clear visual analysis, the results are plotted on a map as shown below.

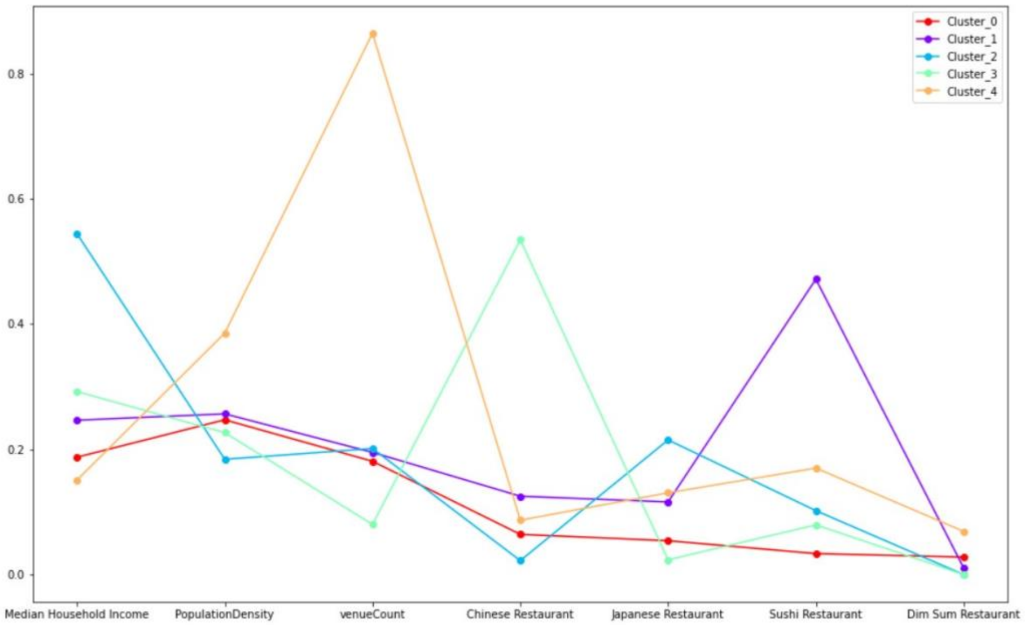


Nose Hill Park
Calgary International Airport
Calgary
1A
201 South
22X
Ann & Sandy Cross Conservation Area

4.3 Cluster line graphs

Third: To visualize the clustering result, we calculated the cluster mean on each measuring feature d visualize them in the graph, as shown in the right. We will discuss the graph details in the next section .

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



5. Discussion & recommendations

Looking at the live chart above, let's start to analyze each line graph:

	Family Income	Population Density	Food Venue Coverage	East Asian Restaurant overage
Cluster 0 (red line)	Low	Medium	Medium	low East Asian restaurant coverage
Cluster 1 (purple line)	Medium	Medium	Medium	fair Chinese/Japanese restaurant coverage, but highest Sushi restaurant coverage
Cluster 2 (blue line)	Highest	Lowest	Medium	highest Chinese restaurant coverage, but low Sushi restaurant coverage
Cluster 3 (green line)	High	Low	Lowest	highest Chinese restaurant coverage, but low Sushi restaurant coverage
Cluster 4 (orange line)	Lowest	Highest	highest	high East Asian restaurant coverage

The recommendation based on the analysis is as follows::

- Cluster 0: ***Not recommended*** to start a new Sushi restaurant in this cluster even if communities in thiscluster 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 havehighest 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 havehigh 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 thiscluster have lowest family income, highest population density and overall food venue coverage. Alow-end Sushi restaurant will thrive in this cluster which are mainly consisted of fully developed downtown or hub communities.

6. Conclusion

In this project,

- we collected and cleaned community demographic data and food venue data.
- We used Feature extraction 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.
- Finally, by analyzing the clustering result, we provided the following recommendations:
    1. **Cluster 2 and 3 are highly recommended for a high-end sushi restaurant**
    2. **Cluster 4 is recommended for a low-end fast- food sushi restaurant catering to Downtown/ CBD customers**

7. How can we further improve the analysis?

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. Using a paid account in the future, we could get more complete venue data which will help to provide a more accurate clustering result.

2. Data limit

We only considered three parameters in the analysis: family income, population density, overall food venue coverageand East Asian restaurant coverage in a community. Enriching this analysis with other parameters like age, median rent, proximity to C-train line and more factors, will further enrich the recommendation.