IBM Data Science
Applied Data Science Capstone

Calgary's Battle of Neighborhoods

OPENING NEW JUICE SHOP IN CALGARY

BY
TADEPALLI SARADA KIRANMAYEE

22 NOV 2020

1. Introduction

Calgary is one of the largest city in the western Canadian province of Alberta. It is 299 km (186 mi) south of the Alberat's capital- Edmonton

The city had a population of 1,285,711 in 2019. It is considered as one of the most-populous city in western Canada.It is the fourth-largest census metropolitan area (CMA) and second-largest in western Canada (after Vancouver). On October 2020, the 11 new communities are proposed for Calgary's outskirts. This provides a wide scope of business.

1.1 Business Problems

1. Finding out common interest of the people of the city by analyzing the most common places visited?
2. Understanding the top businesses in the current communities and setting up the new business in the new extended community?

1.2 Targeted Audience

When new communities are created then both the government officials and the businesses will be looking for the welfare and the convenience for the people dwelling in these areas. This analysis will help in urban planning for the government and business expansion plans for companies.

2. Data

Data that will be used to solve the above business problems are collected from various publicly-available sources.

1. The list of Planning Areas in Singapore are collected from a Wikipedia article entitled "List of neighbourhoods in Calgary" from https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary. This dataset lists the names of all Communities (Neighbourhood), Sector, areas (in km2), and population details etc.
2. The coordinates of each community is extracted from "https://data.calgary.ca/resource/j9ps-fyst.json". The coordinates obtained for each community will be used to plot map markers on the map for visualization.
3. The Foursquare API is used to extract the top 100 venues in each community within 500 radius. The data from the Foursquare API are extracted for relevant information, and descriptive statistics is performed for further analysis.

3. Methodology

3.1 Data Acquisition and Wrangling

The project uses different datasets that are publicly available online. The acquisition and wrangling of data is performed to retrieve the following datasets:

1. Calgary Communities
2. Each Community coordinates
3. Top venues in each Community

3.1.1 Data acquisition for Calgary Communities (Neighbourhood)

With the help of the Pandas Library the tables from the web are converted into a data frame. pandas.read_html() method helped to extract the following dataframe from the Wikipedia article - https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary:

Name[10]	Quadrant	Sector[11]	Ward[12]	Type[11]	2012 Population Rank	Population (2012)[10]	Population (2011)[10]	% change	Dwellings (2012)[10]	Area (km²)[11]	Population density
Abbeydale	NE/SE	Northeast	10	Residential	82	5,917	5,700	3.8	2,023	1.7	3,480.6
Acadia	SE	South	9	Residential	27	10,705	10,615	0.8	5,053	3.9	2,744.9
Albert Park/Radisson Heights	SE	East	10	Residential	75	6,234	6,217	0.3	2,709	2.5	2,493.6
Altadore	SW	Centre	11	Residential	39	9,116	8,907	2.3	4,486	2.9	3,143.4
Alyth/Bonnybrook	SE	Centre	9	Industrial	208	16	17	−5.9	14	3.8	4.2
Applewood Park	SE/NE	East	10	Residential	69	6,498	6,404	1.5	2,215	1.6	4,061.3
Arbour Lake	NW	Northwest	2	Residential	26	10,836	10,762	0.7	3,918	4.4	2,462.7
Aspen Woods	SW	West	6	Residential	92	5,271	4,469	17.9	2,281	3.8	1,387.1

Table 1: Communities in Calgary from Wikipedia

This dataframe provided Name and Type of Community required for analysis and other columns are unnecessary for the analysis and these columns are dropped. The “Name[10]” and the “Type[11]” columns are renamed as “Neighbourhood” and “Type”.

	Neighbourhood	Type
0	Abbeydale	Residential
1	Acadia	Residential
2	Albert Park/Radisson Heights	Residential
3	Altadore	Residential
5	Applewood Park	Residential

Table 2: Communities in Calgary after Data Wrangling

3.1.2 Data acquisition for Calgary Community coordinates

"The City of Calgary's Open Data Portal" provided json data for the coordinates of the communities in a dataset called "Wards and Communities Data Lens ". Pandas help in reading the json file with read_json(). The data frame derived is as follows.

Showing 309 out of 309 rows

CLASS	CLASS_CODE	COMM_CODE	NAME	SECTOR	SRG	COMM_STRUCTURE	longitude	latitude	location
Residential	1	WOO	WOODLANDS	SOUTH	BUILT-OUT	1980s/1990s	-114.10633945400146	50.94287588247354	(50.94287588247354°, -114.10633945400146°)
Residential	1	THO	THORNCLIFFE	NORTH	BUILT-OUT	1950s	-114.06877697721175	51.103099775816034	(51.103099775816034°, -114.06877697721175°)
Residential	1	UOC	UNIVERSITY OF CALGARY	NORTHWEST	BUILT-OUT	OTHER	-114.1299167839167	51.07501240644861	(51.07501240644861°, -114.1299167839167°)
Residential	1	QPK	QUEENS PARK VILLAGE	CENTRE	BUILT-OUT	1950s	-114.07760880424992	51.08514667909161	(51.08514667909161°, -114.07760880424992°)
Residential	1	BDO	BONAVISTA DOWNS	SOUTH	BUILT-OUT	1960s/1970s	-114.03139762075935	50.943700716542395	(50.943700716542395°, -114.03139762075935°
Residential	1	WIN	WINSTON HEIGHTS/MOUNTVIEW	CENTRE	BUILT-OUT	1950s	-114.04185722666399	51.075323847960604	(51.075323847960604°, -114.04185722666399°
Residential	1	EAG	EAGLE RIDGE	SOUTH	BUILT-OUT	1960s/1970s	-114.09930184675723	50.98508133313613	(50.98508133313613°, -114.09930184675723°)
Residential	1	MIS	MISSION	CENTRE	BUILT-OUT	INNER CITY	-114.06632169398661	51.031090323799795	(51.031090323799795°, -114.06632169398661°
Residential	1	HUN	HUNTINGTON HILLS	NORTH	BUILT-OUT	1960s/1970s	-114.0667011472206	51.117583908420535	(51.117583908420535°, -114.0667011472206°)
Residential	1	CHW	CHARLESWOOD	NORTHWEST	BUILT-OUT	1950s	-114.11786352623943	51.08456161142416	(51.08456161142416°, -114.11786352623943°)
Residential	1	SAN	SANDSTONE VALLEY	NORTH	BUILT-OUT	1980s/1990s	-114.09495605533706	51.13773675598465	(51.13773675598465°, -114.09495605533706°)
Residential	1	HAY	HAYSBORO	SOUTH	BUILT-OUT	1950s	-114.08335093454583	50.97221692664152	(50.97221692664152°, -114.08335093454583°)

Table 3: Dataframe from Wards and Communities Data Lens

This dataframe provided Name and Coordinates for Calgary Community required for analysis and other columns are unnecessary for the analysis and these columns are dropped. The "name", "latitude" and "longitude" were renamed as "Neighbourhood", "Latitude", "Longitude".

	Neighbourhood	Longitude	Latitude
0	YORKVILLE	-114.076648	50.870403
1	WOLF WILLOW	-114.008637	50.870724
2	WEST SPRINGS	-114.206168	51.059732
3	WOODLANDS	-114.106339	50.942876
4	WINDSOR PARK	-114.083550	51.005040

Table 4: Dataframe after Data Wrangling

3.1.3 Display all the Neighbourhood on the Map

Using a geocoding service called geopyand the Nominatim geocoder is instantiated. Nominatim geocoder uses OpenStreetMap data, an open data and free to use. The coordinates of Calgary is obtained, and used as the starting coordinates for map visualisation.

The geographical coordinate of Calgary are 51.0534234, -114.0625892

With the help of folium library all the communities are plotted on the map starting with the Calgary geographical coordinates. The resultant map is shown below:

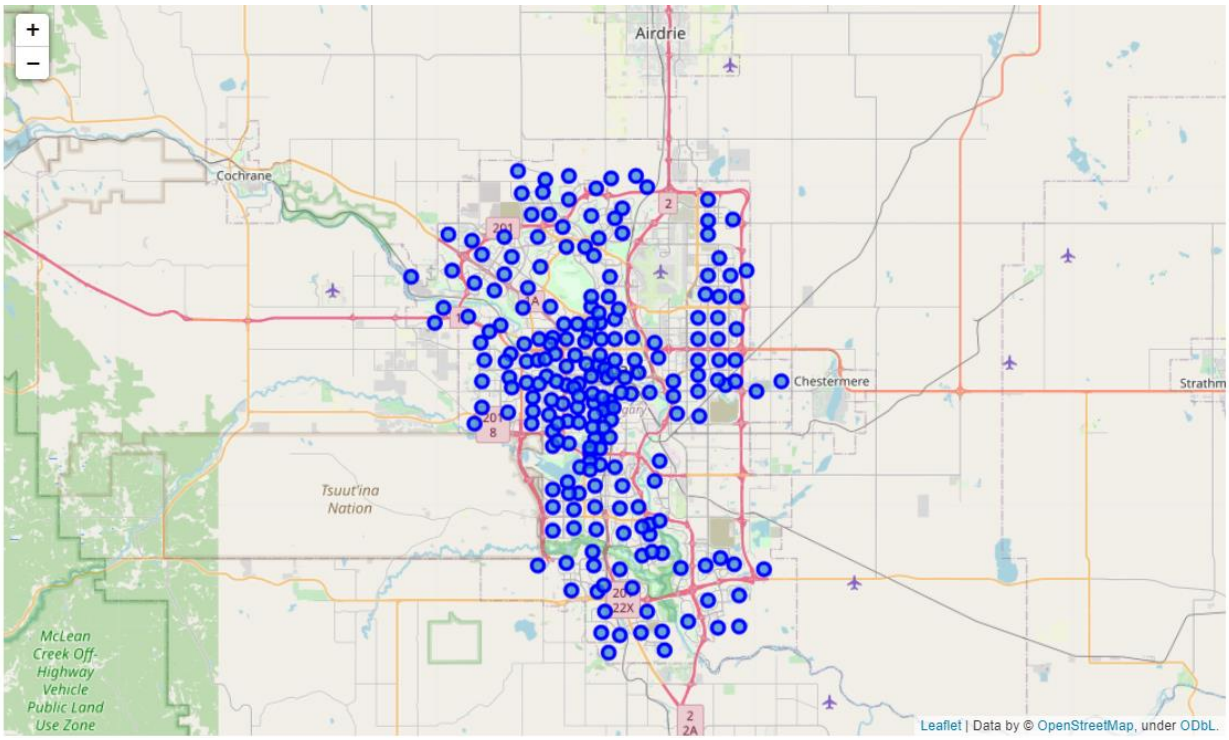

Figure 1: Map showing all Communities in Singapore.

3.1.4 Data acquisition and wrangling on top venues in each Community from FourSquare API

A Foursquare developer account is created and credentials are acquired. The Foursquare API is used to get the top 100 venues in each community, with search radius for each community. API calls are made to Foursquare by passing the coordinates and search radius of each community using loops. The venue data is returned from FourSquare in JSON format. Using the json() function, all the names, latitude, longitudes, and categories, are extracted and dataframe is created. There are 226 unique categories. The dataframe is as follows:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Langitude	Venue Longitude	Venue Category
0	YORKVILLE	50.870403	-114.076648	Mattamy Homes - Yorkville	50.869997	-114.070936	Home Service
1	WEST SPRINGS	51.059732	-114.206168	My Favourite Ice Cream Shop	51.061317	-114.209994	Ice Cream Shop
2	WEST SPRINGS	51.059732	-114.206168	Fergus & Bix Restaurant and Beer Market	51.059953	-114.212111	Restaurant
3	WEST SPRINGS	51.059732	-114.206168	Starbucks	51.060318	-114.212768	Coffee Shop
4	WEST SPRINGS	51.059732	-114.206168	Boston Pizza	51.062335	-114.210770	Pizza Place

Table 5: Venues from all Calgary's Communities

3.2 Exploratory Analysis and Normalizing Data

The count of the “Venue category” for each and every group is derived and displayed

	Neighbourhood	Venue Category
0	ABBEYDALE	4
1	ACADIA	2
2	ALBERT PARK/RADISSON HEIGHTS	4
3	ALTADORE	4
4	APPLEWOOD PARK	5

Table 5: Count of Venues in each Community

For further analysis, one-hot encoding is performed on “Venue Category” using pandas.get_dummies() method on the dataframe in Table 6. The categorical variables are converted into dummy/indicator variables. The resulting dataframe contains 227 columns and 186 rows. These dummy variables are grouped by each community.

	Neighbourhood	Accessories Store	American Restaurant	Arcade	Argentinian Restaurant	Art Gallery	& Crafts Store	Asian Restaurant	Astrologer	Athletics & Sports	...	Vegetarian / Vegan Restaurant	Video Game Store	Vietnames Restaurar
0	ABBEYDALE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
1	ACADIA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
2	ALBERT PARK/RADISSON HEIGHTS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
3	ALTADORE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
4	APPLEWOOD PARK	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
181	WINDSOR PARK	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
182	WINSTON HEIGHTS/MOUNTVIEW	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
183	WOODBINE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
184	WOODLANDS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
185	YORKVILLE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.

186 rows × 227 columns

Table 6: One-hot Encoding

Top 10 venues of each community is calculated with the frequencies.

```
--------ABBEYDALE---------
                      Venue  freq
0                Wings Joint  0.25
1  Health & Beauty Service  0.25
2            Sandwich Place  0.25
3         Convenience Store  0.25
4         Accessories Store  0.00
5    Performing Arts Venue  0.00
6                 Nightclub  0.00
7              Noodle House  0.00
8     Other Great Outdoors  0.00
9         Other Repair Shop  0.00
```

Table 7: Top 10 venues of each community

3.3.1 K-Means clustering algorithm

K-Means clustering is an unsupervised learning algorithm. It analyses the clusters of communities with similarities. The Silhouette score will help to determine the number of clusters.

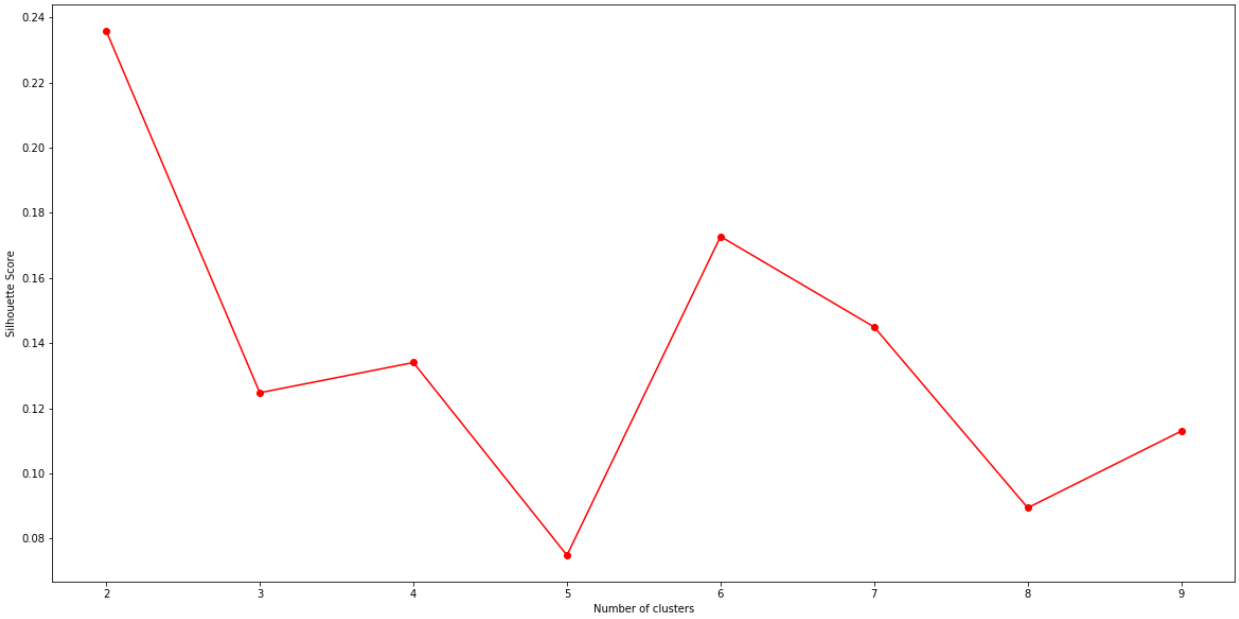

Figure 2: Silhouette score for the number of clusters

The K-Mean Clusters are created and all the clusters are visualized on the map as shown in figure below.

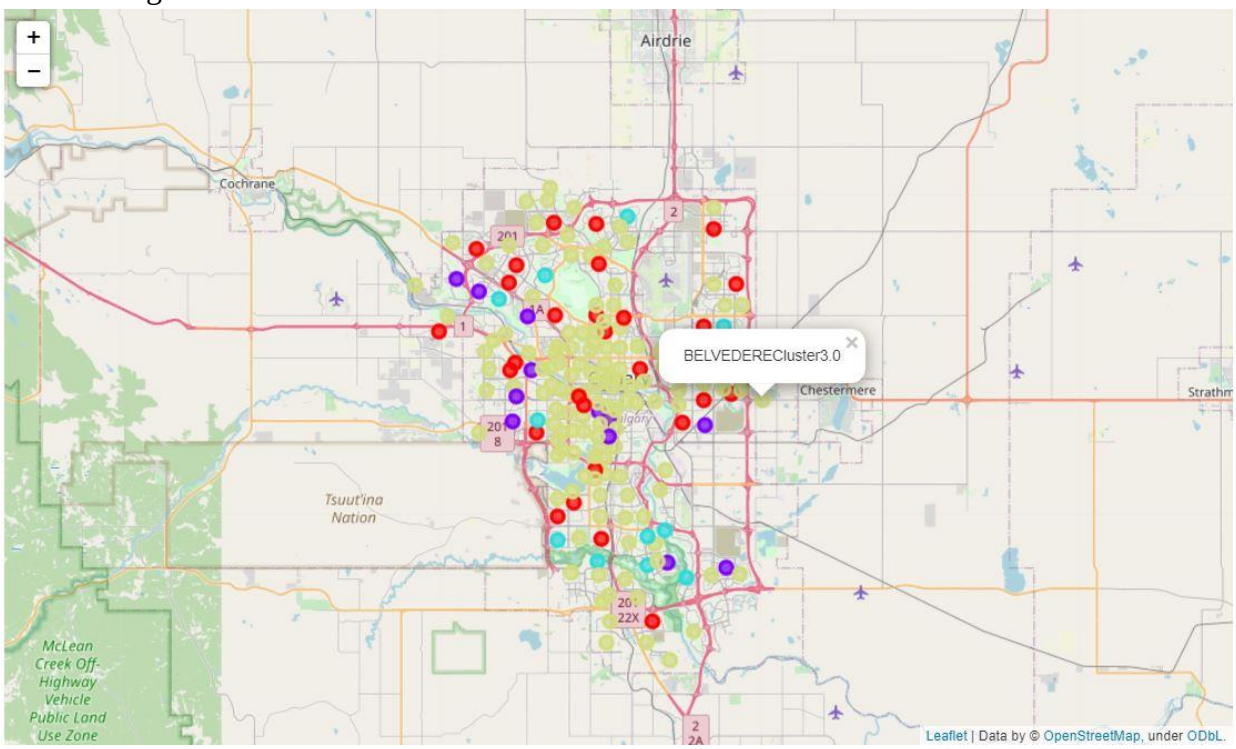

Figure 3: The K-mean clusters are visualized on the Calgary's map

4. Results

The clusters are visualized as a bar chart and the we can see that cluster 3 is biggest cluster and the top 5 venues are “Yoga Studio”, “Gift Shop”, “German Restaurant” , “Gastropub” , “Gas Station”.

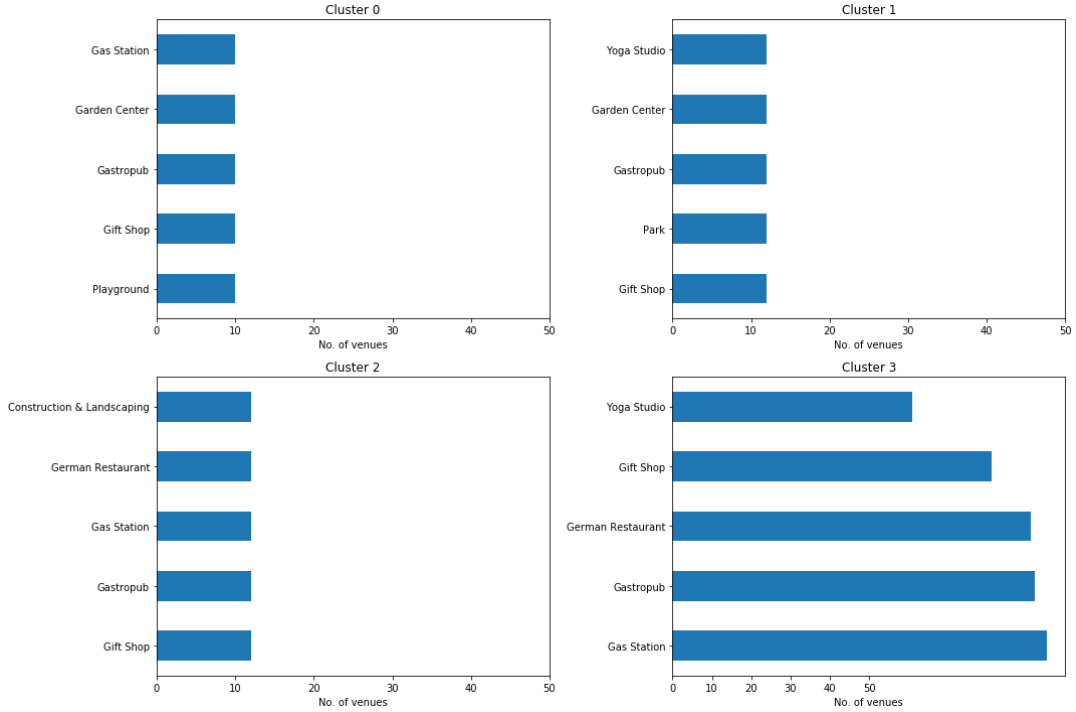

Figure 4: The bar chart for the top 5 venues in the clusters

4 Discussion

The above results show that the cluster 3 is the biggest cluster and the main interest of the Calgary population. The details provided by the analysis shows that the people give more importance to their health and best quality food because the most visited places “Yoga Studio” and the “Gastropub” most visited . In Figure 5, provided by the Calgary government showcasing the new proposed communities.

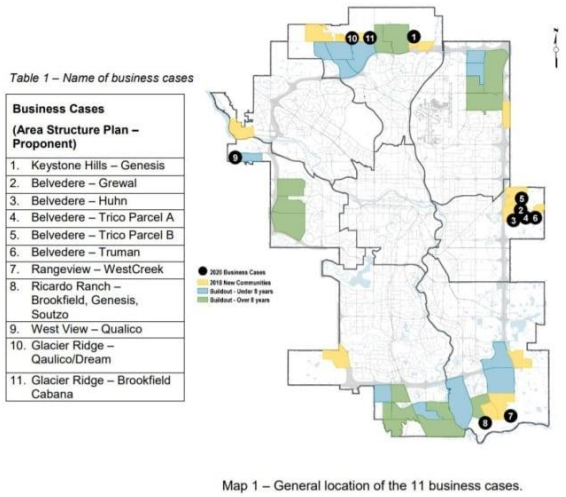

Table 1 – Name of business cases

Business Cases (Area Structure Plan – Proponent)
1. Keystone Hills – Genesis
2. Belvedere – Grewal
3. Belvedere – Huhn
4. Belvedere – Trico Parcel A
5. Belvedere – Trico Parcel B
6. Belvedere – Truman
7. Rangeview – WestCreek
8. Ricardo Ranch – Brookfield, Genesis, Soutzo
9. West View – Qualico
10. Glacier Ridge – Qaulico/Dream
11. Glacier Ridge – Brookfield Cabana

Map 1 – General location of the 11 business cases.

Figure 5: Proposed new communities of Calgary

According to the figure 5 we can infer that “Belvedere” is getting maximum new proposed communities. So, as per figure 4 “Belvedere” falls in the cluster 3 so this is the best place for opening the juice and health drink bar.

5. Conclusion

In this Capstone project, the data was extracted from different sources, coordinates for each community are visualized and, Foursquare API to get trending venues surrounding every area. Data is wrangled, formatted, and normalized so that further data analysis can be performed. Exploratory analysis and visualizations are done to gain a better understanding of the data. Finally, machine learning algorithms, i.e., K-Mean Clustering is used to cluster data. After the analyzing the clusters, the best place to open the juice and health bar is determined, i.e., cluster 3 near the new communities adjacent to “Belvedere”.