Evaluation of Calgary Communities and their Business Potentials Based on K-Means Method and Content Based Recommender System

by Siavash Fard

June 2020

Table of Contents

Abstract

Calgary is one of the fastest growing metropolitans in North America. Calgary small businesses account for 95 percent of all businesses in the city. Every year, an average of 4 to 5 percent of businesses decide to relocate because of operational costs or inadequate market research before opening their businesses. Better understanding of communities and their demographics, therefore, is needed to select the right community for small businesses.

In this report, community data from multiple sources have been collected, cleaned with data wrangling techniques. Communities were clustered into four clusters using K-Means algorithm with Python. Foursquare API has been used to evaluate each cluster and type of businesses in communities.

Last, a number of venues for one of the neighborhoods (Montgomery) has been recommended based on its clustered communities and using content-based recommender system. Similar approach can be taken to recommend other business opportunities in other communities in Calgary.

Introduction

Background

Calgary is one of the fastest growing metropolitans in North America mostly because of its oil and gas industry. The city has seen the second highest population growth in Canada and is the second youngest metropolitan in Canada. The city of Calgary has been suffering from economic downturn (mostly because of crude oil price crash) since 2015 but still with one of the highest unemployment rates in Canada, Alberta families earned an average of $72,700 in 2018, about $11,000 more than the national average of $61,400.

Calgary small business account for 95 percent of all businesses, the second highest number of small businesses per capita of the major cities in Canada. According to Statistics Canada, Retail Trade, and Accommodation and Food Services accounts for 8.5 and 5.6 percent of all small businesses in Calgary (Statistics Canada 2019). According to the City of Calgary, in 2019, there has been 1,587 new business licenses issued (a significant decrease from 7,085 in 2018). Every year, in average, between 4 to 5 per cent of businesses decide to relocate.

Motivation and Objective

Calgary has several popular neighborhoods with a high number of retail stores and eateries. Every year, number of new businesses start up in these neighborhoods to take advantage of already existing high demand; however, lack of market and neighborhood research would result in closures or relocation after a year or two. As a result, better research is needed to evaluate the different communities for their business opportunities.

In this report, communities across Calgary has been clustered into four different groups based on their demographic, income and average property assessment of the community. Moreover, potential business opportunities have been recommended for one of the communities based on their similarities in demographics, income, and average property values.

Methodology

Data Sources

Data for this report has been gathered from multiple sources. Demographic data for this report has been collected from the City of Calgary (https://data.calgary.ca/). Property assessment and values of more than 7 million properties have been collected (also from the City of Calgary) and averaged for each community.

The geojson file for community boundaries has been downloaded from https://data.calgary.ca/ and each community area has been calculated based on the polygon coordinates (to calculate population density for each community).

The median income for each neighborhood has been extracted from great-news.ca (the only available reference for income per neighborhood in Calgary).

All these datasets are combined in one data frame to be used for neighborhoods clustering and determine similar neighborhoods.

Finally, nearby venues for each neighborhood has been called from Foursquare website and sorted in order to determine the type of businesses and frequencies in each community.

Data Wrangling and Exploratory Data Analysis

Demographics data (Census_by_Community_2019.csv) for each community from the City of Calgary website contains various information such as age, gender, type of community (e.g. residential, industrial, etc.), community structure (e.g. developing, built out, etc.), residents count, dwellings, single family dwellings, dwelling types (apartment, etc.). Detailed description of data is given in Appendix.

The dwelling types have been removed from the dataset in order not to affect the communities' pattern. Also, some of these numbers contain outliers rendering these columns unreliable. In addition, there is many missing values in these columns which should not be included in modelling. Other data removed from the City of Calgary file are dog and cat counts, sector, converted dwellings, nursing homes, residents per dwelling types (e.g. apartment residents, etc.), number of residents per dwelling (contains missing data for some communities), gender considered as others (not male or female), etc.

Only Non 'Residential' Class and 'Employment' community structure (e.g. downtown area) have been filtered out. The ratios for each community have been calculated as below:

PRSCH_CHLD_RATIO = PRSCH_CHLD / RES_CNT

OWNSHIP_RATIO = OWNSHIP_CNT / DWELL_CNT

SING_FAMLY_RATIO = SING_FAMLY/ DWELL_CNT

Having examined the data, it was realized that some communities have SING_FAMLY_RATIO of 1.0 or 0.0 (e.g. BEL AIRE); therefore, these communities were removed from the final dataset.

Also, the ratios for both male and female for each age category have replaced the original counts.

Since each community has different area, the RES_CNT is not a great indicator of the neighborhood density. For example, some communities may have large population because of only large area and not just high population density (see Figure 1 and 2). This large population (as a result of large area) may falsely affect our understanding of population distribution and demands in Calgary. For example, Saddle Ridge has a very high residents count which one may conclude that there may be higher demands but because of its large area, it has relatively low population density.

As a result, area for each community has been calculated using the coordinates and the population density for each community has been calculated as below:

RES_CNT_THOUSANDS_PER_SQ_KM = RES_CNT/ AREA*10^3

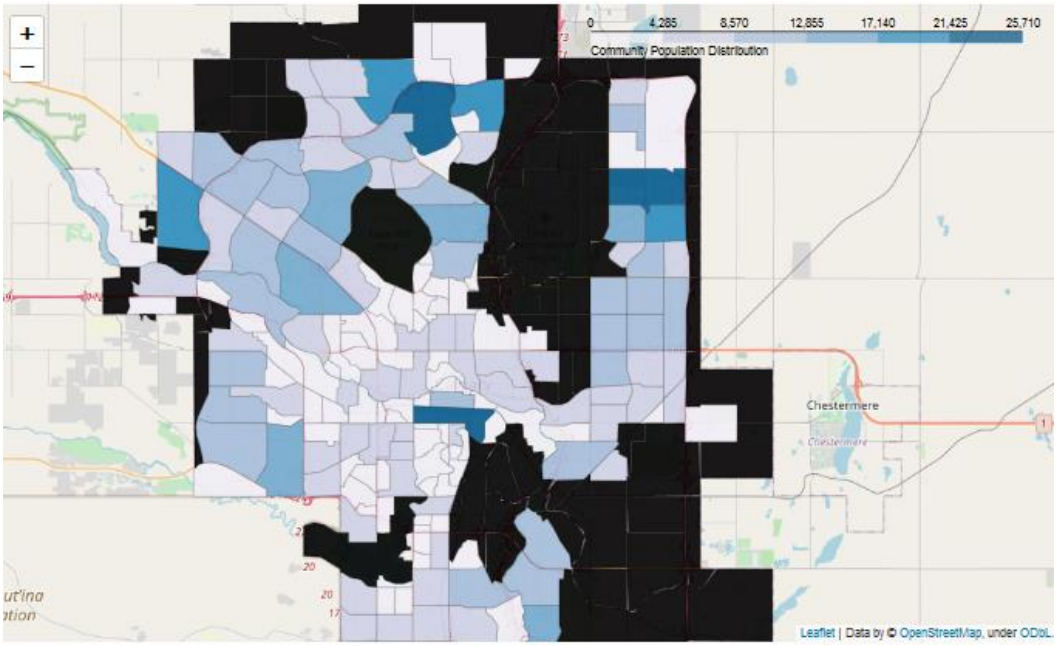

Figure 1 Residents Count for Each Neighborhood in Calgary. Note than black areas have been removed from the dataset.

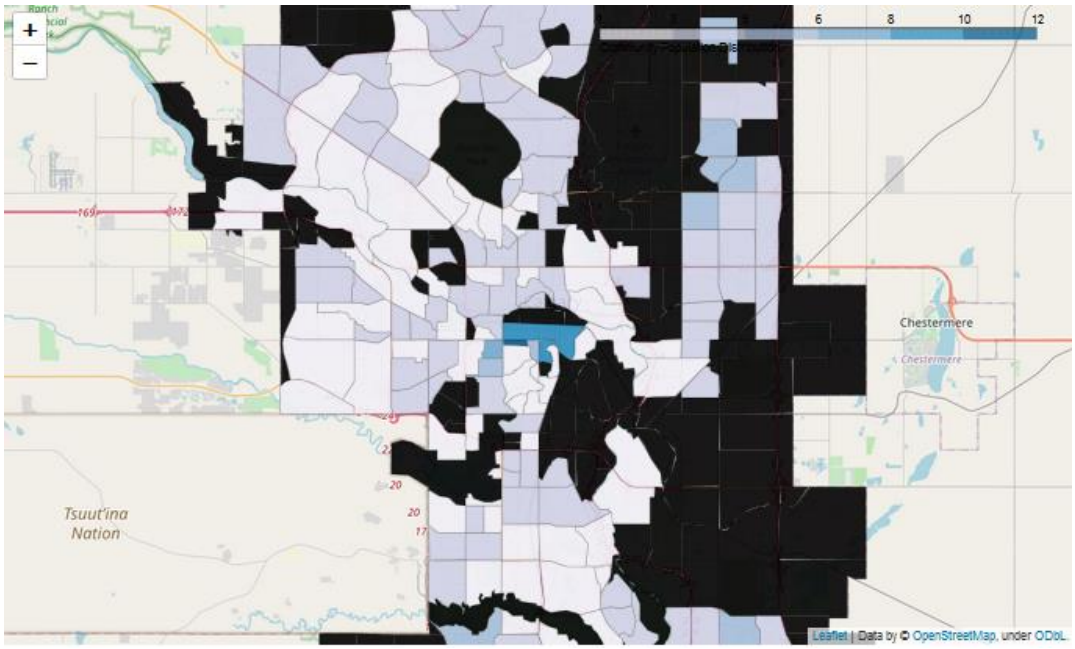

Figure 2 Population Density for Each Community (Thousands per Square Km). Note than black areas have been removed from the dataset.

Unfortunately, the author was not able to obtain any data for Calgarians income for 2019. The only available data was median income for 2014 (great-news.ca); even though it is preferred to gather information from official government sources, the income information is not typically available for communities. Still it is recommended to use this old dataset to characterize communities across Calgary.

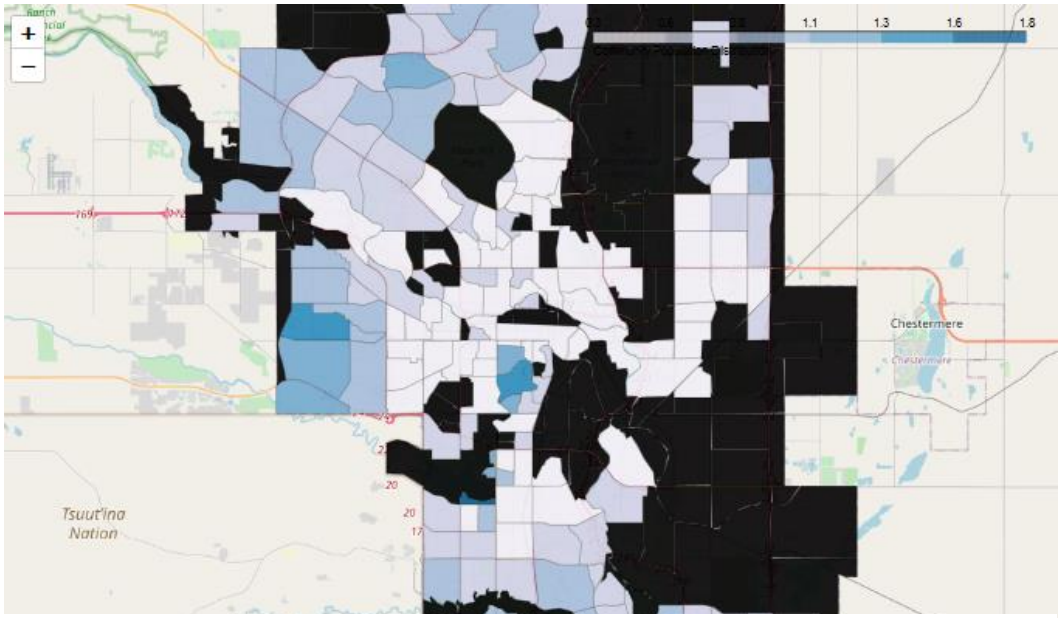

Figure 3 Median Income Distribution for Each Community in Calgary in 2014 (Hundred Thousands of Dollars). Note than black areas have been removed from the dataset.

Finally, property values of more than 7 million unique addresses have been collected and averaged for each community. This size of this file was extremely large, so the data was imported in chunks and grouped and averaged for each chunk. The code for this simple operation has not been included in the final report due to long processing time.

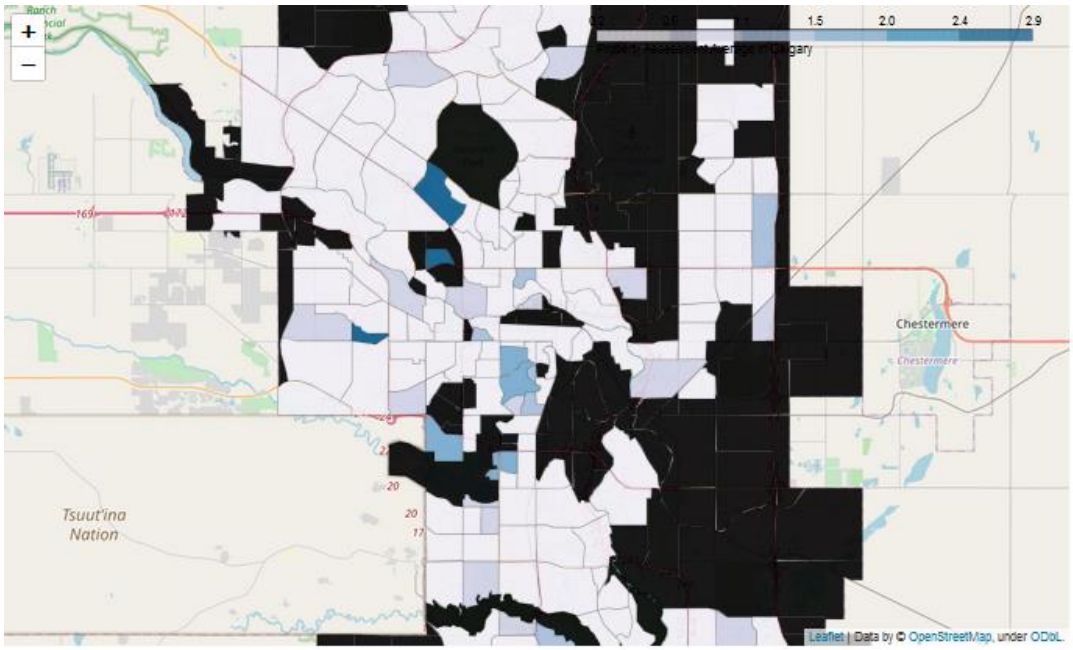

Figure 4 Property values for each community for Calgary in 2020 (million dollars). Note than black areas have been removed from the dataset.

After merging the relevant data together and replacing all counts with ratios, the final dataset (yyc_community) is comprised of 167 rows and 29 columns. However, LATITUDE and LONGITUDE should not be included in the clustering algorithm; otherwise, the location of the communities would also affect the clustering. So, the dataset input for the modelling has 167 rows and 27 columns.

Understanding the Communities

Above choropleth maps can shed some light about the city and its communities. As shown in Figure 2, the population density is slightly higher in the city center (e.g. Bankview). Income distribution in Figure 3 shows that except three communities (Upper Mount Royal, Britannia, and Elbow Park) in the inner city, outer city communities have relatively higher income than city center or inner-city neighborhoods. A lot of Calgarians with higher income would prefer to reside in the western outer city communities away from the city center traffic and closer to Rocky Mountains. Last, there are two communities which have the highest average property assessments but do not appear to be having high incomes. These two communities (i.e. Brentwood and University Heights) have higher values as they are closer to the University of Calgary.

Communities Clustering

K-Means Algorithm

K-Means algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group.** It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way K-Means algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters is not changing.

- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The approach k-Means follows to solve the problem is called **Expectation-Maximization**. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster.

The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} ||x^i - \mu_k||^2$$

where w_{ik}=1 for data point xi if it belongs to cluster *k*; otherwise, w_{ik}=0. Also, μ_k is the centroid of xi's cluster.

It's a minimization problem of two parts. We first minimize J w.r.t. wik and treat μ_k fixed. Then we minimize J w.r.t. μ_k and treat w_{ik} fixed. Technically speaking, we differentiate J w.r.t. w_{ik} first and update cluster assignments (*E-step*). Then we differentiate J w.r.t. , μ_k and recompute the centroids after the cluster assignments from previous step (*M-step*).

Implementation

Communities clustering with K-Means have been implemented using sklearn. Before implementing the K-Means method, the data was normalized using StandardScaler.

Evaluation Using Elbow Method

Contrary to supervised learning where we have the ground truth to evaluate the model's performance, clustering analysis does not have a solid evaluation metric that we can use to evaluate the outcome of different clustering algorithms. Moreover, since K-Means requires *k* as an input and doesn't learn it from data, there is no right answer in terms of the number of clusters that we should have in any problem.

Elbow method gives us an idea on what a good *k* number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids.

Figure 5 shows the SSE for different cluster numbers. Based on the elbow method, cluster number of three (3) is recommended for K-Means modelling. However, cluster number of four (4) has been used to better differentiate between the communities.

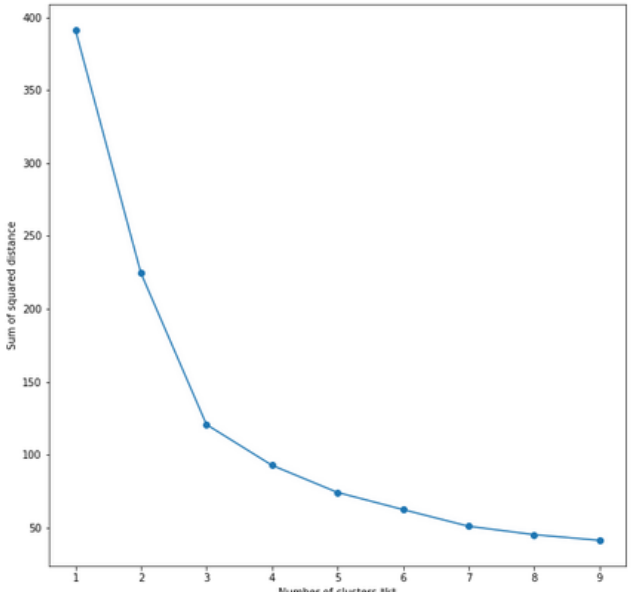

Figure 5 Sum of Squared Distance (SSE) for different cluster numbers. Based on the Elbow method, three clusters are recommended for the K-Means algorithm

Foursquare API

To gather retail stores and small businesses information, Foursquare API was called; the data gathered from Foursquare then was sorted and top three venues were tabled.

Foursquare API premium call for Cluster 3 was made and ratings were extracted for content-based recommender system.

Content Based Recommender System

A content-based recommender works with data that the user provides, either explicitly (rating) or implicitly (clicking on a link). Based on that data, a user profile is generated, which is then used to make suggestions to the user. As the user provides more inputs or takes actions on the recommendations, the engine becomes more and more accurate.

A content-based recommender system was used to suggest venues based on the venue category and ratings; even though a collaborative recommender system is preferred, not enough data was available for this algorithm because of lack of available ratings for most venues in these communities as well as Foursquare premium calls limit.

Results and Discussion

Based on the K-Means method, Calgary communities were grouped into four (4) clusters. Table 1 shows some of the results from the K-Means algorithm.

Table 1 Mean values for four (4) communities clusters in Calgary

Clus_km	RES_CNT_THOUSANDS_PER_SQ_KM	MEDIAN_INCOME_HUNDRED_THOUSAND	OWNSHIP_RATIO	SING_FAMILY_RATIO	ASSESSED_VALUE	PRSCH_CHLD_RATIO
0	4.107248	0.589976	0.561534	0.490928	0.380450	0.079689
1	2.752917	0.672690	0.662595	0.612507	0.635986	0.066085
2	9.251569	0.347463	0.229917	0.045762	0.329398	0.020671
3	1.672295	0.820945	0.726011	0.709260	0.671622	0.055932

Considering the number of columns in this table, the communities can be characterized as:

Cluster 0: Communities with relatively high population density, medium income level who own relatively more affordable properties; examples are South Calgary, Killarney and Crescent Heights

Cluster 1: Communities with higher income and relatively higher property values and older demographics than Cluster 0. A large majority of Calgary communities fall into this cluster.

Cluster 2: Communities with low income in highly populated communities. This cluster may be mostly younger adults with lower ownership ratio (most likely renting). Only a few communities fall into this category; examples are Mission, Beltline, Bankview and Lower Mount Royal.

Cluster 3: Highest paid communities with lowest population density and high ownership ratios and highest property values with older demographics; examples are Upper Mount Royal, Elbow Park, Inglewood, Southview, Sprinbank Hill.

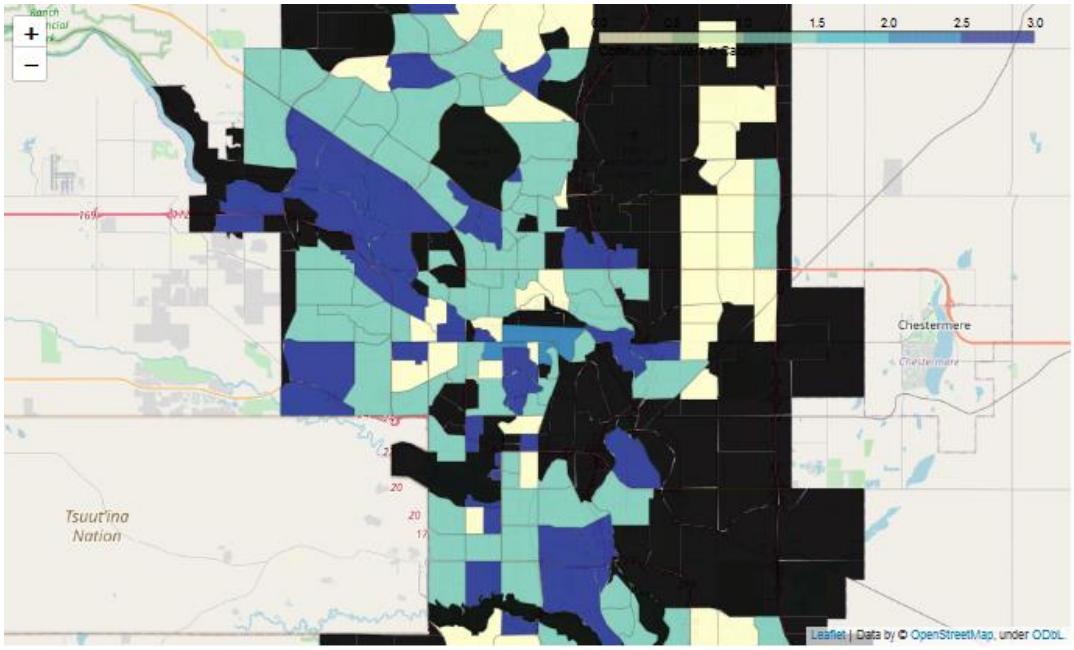

Figure 6 Community clustering of Calgary from K-Means method.

Table 2 Foursquare API results for communities in Calgary

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	ABBEYDALE	Wings Joint	Convenience Store	Sandwich Place	Health & Beauty Service	Yoga Studio
1	ACADIA	Pool	Gym / Fitness Center	Yoga Studio	Flower Shop	Garden Center
2	ALTADORE	Pub	Ice Cream Shop	Dog Run	Yoga Studio	Flower Shop
3	APPLEWOOD PARK	Liquor Store	Park	Trail	Home Service	Construction & Landscaping
4	ARBOUR LAKE	Residential Building (Apartment / Condo)	Lake	Grocery Store	Yoga Studio	Flower Shop

Foursquare API data was filtered to only Cluster 3 communities to better evaluate the nearby venues. There are approximately 40 communities in this cluster with 97 available ratings. Content-based recommender system was used on “Montgomery” and its profile was generated using venue categories.

Table 3 Venue ratings and category for Montgomery

	Neighborhood	Venue id	Venue Name	Venue Rating	Venue Category
0	MONTGOMERY	4c62047ceb82d13a969604d6	NOtaBLE	8.3	Restaurant
1	MONTGOMERY	4bb3ec87737d76b061e83a7c	Shouldice Athletic Park	8	Park
2	MONTGOMERY	55351231498ee5b605d43cb9	Tim Hortons	6.3	Coffee Shop
3	MONTGOMERY	4d2624b8d2668cfa2a14c5db	Subway	6.2	Sandwich Place
4	MONTGOMERY	4dcb5515c65bccd86744de6d	Pizza Hut	6.3	Pizza Place
5	MONTGOMERY	4dae25c55da3cca6f0a39f60	KFC	6.2	Fast Food Restaurant
6	MONTGOMERY	4bf4a947370e76b04796bd4a	7-Eleven	6	Convenience Store

The community profile was calculated using the Venue Rating and the binarized category of each Venue Category. The weighted average of the top five similar communities are given in Table 4.

Table 4 Weighted average of top five communities suggested for Montgomery

Neighborhoods	Weighted Average
BAYVIEW	0.084567
WILDWOOD	0.084567
BRITANNIA	0.053277
LAKE BONAVISTA	0.043793
EAGLE RIDGE	0.043693

Table 5 Venue ratings of suggested communities for Montgomery

	Neighborhood	Venue id	Venue Name	Venue Rating	Venue Category
8	WILDWOOD	4bba279598c7ef3be3c43202	Edworthy Park	9.3	Park
9	WILDWOOD	4b3ffecbf964a520e2b325e3	Edworthy Dog Park	7.5	Dog Run
10	LAKE BONAVISTA	5325e880498eb79a55daadb7	The Lake House	7.4	Restaurant
11	LAKE BONAVISTA	4b7a26c3f964a520b0242fe3	Brewsters Lake Bonavista	7.1	Brewery
12	LAKE BONAVISTA	4d6d1a14cf7e41bd2cad8285	TD Canada Trust	6.5	Bank
13	LAKE BONAVISTA	4ba3cba4f964a520406038e3	Subway	6.4	Sandwich Place
14	LAKE BONAVISTA	4c9bdfe90e9bb1f7e7c3ce5f	Shoppers Drug Mart	6.2	Pharmacy
15	LAKE BONAVISTA	4bd4f1d429eb9c7460b592e1	Safeway Bonavista Shopping Plaza	6.1	Grocery Store
16	LAKE BONAVISTA	4b5b257af964a5206ce628e3	Lake Bonavista Promenade	5.6	Shopping Mall
24	BRITANNIA	54457979498e5391918f06c4	Village Ice Cream	8.5	Ice Cream Shop
25	BRITANNIA	5a679a8a59c42311fd6bec88	Monogram Coffee	8.3	Coffee Shop
26	BRITANNIA	4b771183f964a520007a2ee3	Sunterra Market	8	Food & Drink Shop
27	BRITANNIA	4ba12ae3f964a520419e37e3	Starbucks	7.1	Coffee Shop
28	BRITANNIA	4c08f9a6a1b32d7fbdcd96f0	RBC Royal Bank	6.5	Bank
68	EAGLE RIDGE	4b0586e9f964a520c17422e3	Heritage Park Historical Village	8.5	History Museum
69	EAGLE RIDGE	4ba50faef964a5206cd738e3	Railroad Cafe	6.8	Sandwich Place
70	EAGLE RIDGE	4e8673f7d66a9b178e910fed	Petro-Canada	6.2	Gas Station
83	BAYVIEW	4b0586e9f964a520cb7422e3	South Glenmore Park	9.3	Park
84	BAYVIEW	4bc930f03740b71375705e65	Petro-Canada	6.3	Gas Station

As discussed earlier, content-based recommender system is not preferred given few available ratings and how the recommendation system works.

To better understand Montgomery community, K-Means method was used to cluster different communities based on their venue categories and their frequencies. Communities have been filtered to only include Cluster 3 (Clus-km == 3.0) and only clusters which include Montgomery (Cluster Labels == 1.0)

Table 6 Communities clustering based on venue type and frequencies

	NAME	Clus_km	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	MONTGOMERY	3	1.0	Japanese Restaurant	Restaurant	Food & Drink Shop	Park	Coffee Shop
12	LAKE BONAVISTA	3	1.0	Park	Bank	Skating Rink	Shopping Mall	Brewery
25	INGLEWOOD	3	1.0	Gas Station	Diner	Cycle Studio	Pool	Liquor Store
65	VALLEY RIDGE	3	1.0	Restaurant	Greek Restaurant	Convenience Store	Yoga Studio	Flower Shop
117	GLENDALE	3	1.0	Chinese Restaurant	Pizza Place	Yoga Studio	Flower Shop	Garden Center
152	PATTERSON	3	1.0	Health & Beauty Service	Vietnamese Restaurant	Trail	Pizza Place	Convenience Store
155	CHARLESWOOD	3	1.0	Building	Pharmacy	Coffee Shop	Gas Station	Garden Center
159	BAYVIEW	3	1.0	Gas Station	Grocery Store	Other Great Outdoors	Park	Gastropub

Before drawing any conclusion, Montgomery demographics may give a better understanding of the community.

Table 7 Montgomery demographics versus Cluster 3 and Calgary communities

	MONTGOMERY	Clus_km == 3	CALGARY
RES_CNT_THOUSANDS_PER_SQ_KM	1.52943	1.626422	2.964469
MALE_0_4	0.0676895	0.053354	0.057813
MALE_5_14	0.0902527	0.119206	0.120619
MALE_15_19	0.0415162	0.054457	0.051547
MALE_20_24	0.0694946	0.055260	0.058977
MALE_25_34	0.224278	0.118116	0.152069
MALE_35_44	0.17509	0.144355	0.164303
MALE_45_54	0.118682	0.139578	0.135988
MALE_55_64	0.11056	0.148259	0.130329
MALE_65_74	0.0636282	0.104378	0.082112
MALE_75	0.0388087	0.063037	0.046244
FEM_0_4	0.0597992	0.050360	0.054470
FEM_5_14	0.0811873	0.111781	0.112691
FEM_15_19	0.0336098	0.050291	0.047791
FEM_20_24	0.071148	0.047889	0.055698
FEM_25_34	0.220864	0.117001	0.153601
FEM_35_44	0.173723	0.146471	0.163017
FEM_45_54	0.116543	0.142868	0.134415
FEM_55_64	0.121344	0.150286	0.132074
FEM_65_74	0.0628546	0.107734	0.086863
FEM_75	0.0589262	0.075319	0.059382

ASSESSED_VALUE	0.424912	0.700592	0.573765
OWNSHIP_RATIO	0.510184	0.732554	0.643523
SING_FAMILY_RATIO	0.490313	0.717585	0.593847
PRSCH_CHLD_RATIO	0.0726467	0.053326	0.065743
MEDIAN_INCOME_HUNDRED_THOUSAND	0.42795	0.842413	0.683159

Montgomery has younger population with lower median income when compared to Cluster 3 communities. Assuming that we are not interested in businesses which already exist in the area, from Table 6, we can conclude that the following venues could be a match for Montgomery:

- Coffee shop (local or high end)
- Grocery store
- Brewery
- Gastropub

These venues also satisfy our conclusion from the demographics of Montgomery community (Table 7) which mostly consists of younger population with lower than average income.

Conclusion

In this report, communities across Calgary were clustered and visualized based on their demographics and venue categories. Calgary communities were clustered into four different groups based on their demographics, income level and average property assessments of communities.

- Cluster 0: Communities with relatively high population density, medium income level who own relatively more affordable properties; examples are South Calgary, Killarney and Crescent Heights
- Cluster 1: Communities with higher income and relatively higher property values and older demographics than Cluster 0. A large majority of Calgary communities fall into this cluster.
- Cluster 2: Communities with low income in highly populated communities. This cluster may be mostly younger adults with lower ownership ratio (most likely renting). Only a few communities fall into this category; examples are Mission, Beltline, Bankview and Lower Mount Royal.
- Cluster 3: Highest paid communities with lowest population density and high ownership ratios and highest property values with older demographics; examples are Upper Mount Royal, Elbow Park, Inglewood, Southview, Sprinbank Hill.

A number of venues for one of the neighborhoods (Montgomery) has been recommended based on its clustered communities and using content-based recommender system.

The same approach in this report can also be used to determine the similarity of two communities based on their demographics and venue category. This report can be improved by taking the locations and demographics of venue reviewers on Foursquare to understand the customers behavior in different communities.

Appendix

Description of the data columns and their types

CLASS	Industrial, Major Park Area, Residential, Residual Sub Area	Plain Text
CLASS_CODE	1 = Residential, 2 = Industrial, 3 = Major Park, 4 = Residual Sub Area	Number
COMM_CODE	A three-character code assigned to the Community District	Plain Text
NAME	Full name of the Community District as approved by City Council	Plain Text
SECTOR	Planning Sector Polygon the Community is Located in: Centre, East, West, North, Northeast, Northwest, South, Southeast, Southwest	Plain Text
SRG	Reflects the yearly development capacity or housing supply as outlined in the Suburban Residential Growth document, the valid values are: BUILT-OUT, DEVELOPING, NON-RESIDENTIAL, and N/A	Plain Text
COMM_STRUCTURE	Used to identify life-cycle patterns and to develop the demographic model of Calgary. Where a decade is listed at least 51% of a community's peak population must be in place by the end of the decade it is assigned to. Valid values are: 19502, 1960s/1970s, 1970s/1980s, 1980s/1990s, 1990s/2000s, 2000s, Building Out, Centre City, Employment, Inner City, Other, Undeveloped	Plain Text
CNSS_YR	Year census data was gathered	Number
FOIP_IND	Indicates results subject to Freedom of Information and Protection of Privacy Legislation. Freedom of Information and Protection of Privacy rules are applied to the data to ensure that no individual can be identified in any of the data released.	Plain Text
RES_CNT	Number of residents	Number
RES_CNT_THOUSANDS_PER_SQ_KM	Number of residents per square kilometer of the community	Number
DWELL_CNT	Number of dwellings	Number
PRSCH_CHLD	Number of preschool children	Number
PRSCH_CHLD_RATIO	Ratio of preschool children to total number of residents (RES_CNT)	Number
ELECT_CNT	Number of enumerated voters	Number

EMPLYD_CNT	Employed persons include those 15 years of age and older who are employed full or part time. This includes those who are self employed, employed by others and persons who may not be working temporarily due to health, vacation, weather, labour disputes or other personal reasons such as bereavement.	Number
OWNSHP_CNT	Number of homeowners	Number
OWNSHP_RATIO	Ratio of homeowners to dwelling count (DWELL_CNT)	Number
DOG_CNT	Number of dogs	Number
CAT_CNT	Number of cats	Number
PUB_SCH	This represents dwellings that support the Public School system (Calgary Board of Education)	Number
SEP_SCH	This represents dwellings that support the Separate School system (Calgary Catholic School District)	Number
PUBSEP_SCH	This represents dwellings that support both the Public School system (Calgary Board of Education) and the Separate School system (Calgary Catholic School District)	Number
OTHER_SCH	This represents dwellings that support school systems other than Public or Separate.	Number
UNKNWN_SCH	This represents dwellings whose school support is unknown or undetermined.	Number
SING_FAMLY	Number of single family dwellings	Number
SING_FAMILY	Ratio of single family dwellings to total dwelling (DWELL_CNT) in the community	Number
DUPLEX	Number of duplexes	Number
MULTI_PLEX	Number of structures designed and built to contain at least three or more dwelling units on one or two levels.	Number
APARTMENT	Number of structures designed and built to contain at least three or more dwelling units on three or more levels.	Number
TOWN_HOUSE	A structure originally designed and built to contain three or more attached or semi-detached dwelling units.	Number
MANUF_HOME	A structure originally built to be movable, whethere it is now movable or on a permanent foundataion. Also referred to as a manufactured home.	Number
CONV_STRUC	The additional dwelling unit in a structure that contains more units than the building	Number

	was originally designed and built to contain. Also referred to as a converted structure.	
COMUNL_HSE	A structure that contains one dwelling unit, in which multiple individuals, are accommodated, who have separate sleeping facilities but share common cooking and/or bathroom facilities. Also referred to as a communal house.	Number
RES_COMM	A structure that is primarily commercial but which also contains one or two dwelling units.	Number
OTHER_RES	Any residential structure that contains a dwelling unit but does not fit the other structure types listed.	Number
NURSING_HM	A structure originally designed to contain one or more dwelling units with is designated as a nursing home, auxiliary hospital, care centre, etc.	Number
OTHER_INST	A structure where multiple residents are temporarily living and where the cooking is centrally provided for and which is not prepared by the residents, i.e hospice, jail, etc.	Number
HOTEL_CNT	A structure that provides lodging.	Number
OTHER_MISC	A structure that does not fit any of the other categories.	Number
APT_NO_RES	Number of apartment units that are used for non-residential purposes	Number
APT_OCCPD	Number of apartment units that are occupied	Number
APT_OWNED	Number of apartment units that are owned by the resident	Number
APT_PERSON	Total number of persons occupying apartment units	Number
APT_VACANT	Number of apartment units that are vacant	Number
APT_UC	Number of apartment units that are under construction	Number
APT_NA	Number of apartment units that are inactive	Number
CNV_NO_RES	Number of converted structure units that are used for non-residential purposes	Number
CNV_OCCPD	Number of converted structure units that are occupied	Number
CNV_OWNED	Number of converted structure units that are owned by the resident	Number
CNV_PERSON	Total number of persons occupying converted structure units	Number

CNV_VACANT	Number of converted structure units that are vacant	Number
CNV_UC	Number of converted structure units that are under construction	Number
CNV_NA	Number of converted structure units that are inactive	Number
DUP_NO_RES	Number of duplex units that are used for non-residential purposes	Number
DUP_OCCPD	Number of duplex units that are occupied	Number
DUP_OWNED	Number of duplex units that are owned by the resident	Number
DUP_PERSON	Total number of persons occupying duplex units	Number
DUP_VACANT	Number of duplex units that are vacant	Number
DUP_UC	Number of duplex units that are under construction	Number
DUP_NA	Number of duplex units that are inactive	Number
MFH_NO_RES	Number of manufactured home units that are used for non-residential purposes	Number
MFH_OCCPD	Number of manufactured home units that are occupied	Number
MFH_OWNED	Number of manufactured home units that are owned by the resident	Number
MFH_PERSON	Total number of persons occupying manufactured home units	Number
MFH_VACANT	Number of manufactured home units that are vacant	Number
MFH_UC	Number of manufactured home units that are under construction	Number
MFH_NA	Number of manufactured home units that are inactive	Number
MUL_NO_RES	Number of multiplex units that are used for non-residential purposes	Number
MUL_OCCPD	Number of multiplex units that are occupied	Number
MUL_OWNED	Number of multiplex units that are owned by the resident	Number
MUL_PERSON	Total number of persons occupying multiplex units	Number
MUL_VACANT	Number of multiplex units that are vacant	Number
MUL_UC	Number of multiplex units that are under construction	Number
MUL_NA	Number of multiplex units that are inactive	Number
OTH_NO_RES	Number of other structure units that are used for non-residential purposes	Number

OTH_OCCPD	Number of other structure units that are occupied	Number
OTH_OWNED	Number of other structure units that are owned by the resident	Number
OTH_PERSON	Total number of persons occupying other structure units	Number
OTH_VACANT	Number of other structure units that are vacant	Number
OTH_UC	Number of other structure units that are under construction	Number
OTH_NA	Number of other structure units that are inactive	Number
TWN_NO_RES	Number of townhouse units that are used for non-residential purposes	Number
TWN_OCCPD	Number of townhouse units that are occupied	Number
TWN_OWNED	Number of townhouse units that are owned by the resident	Number
TWN_PERSON	Total number of persons occupying townhouse units	Number
TWN_VACANT	Number of townhouse units that are vacant	Number
TWN_UC	Number of townhouse units that are under construction	Number
TWN_NA	Number of townhouse units that are inactive	Number
SF_NO_RES	Number of single family units that are used for non-residential purposes	Number
SF_OCCPD	Number of single family units that are occupied	Number
SF_OWNED	Number of single family units that are owned by the resident	Number
SF_PERSON	Total number of persons occupying single family units	Number
SF_VACANT	Number of single family units that are vacant	Number
SF_UC	Number of single family units that are under construction	Number
SF_NA	Number of single family units that are inactive	Number
OTH_STRTY	Other Structure Type	Number
DWELSZ_1	Dwelling has 1 occupant	Number
DWELSZ_2	Dwelling has 2 occupants	Number
DWELSZ_3	Dwelling has 3 occupants	Number
DWELSZ_4_5	Dwelling has 4 or 5 occupants	Number
DWELSZ_6	Dwelling has 6 occupants	Number

MALE_CNT	Total number of male residents	Number
FEMALE_CNT	Total number of female residents	Number
MALE_0_4	Ratio of male residents aged 0 to 4 total number of male residents	Number
MALE_5_14	Ratio of male residents aged 5 to 14 total number of male residents	Number
MALE_15_19	Ratio of male residents aged 15 to 19 total number of male residents	Number
MALE_20_24	Ratio of male residents aged 20 to 24 total number of male residents	Number
MALE_25_34	Ratio of male residents aged 25 to 34 total number of male residents	Number
MALE_35_44	Ratio of male residents aged 35 to 44 total number of male residents	Number
MALE_45_54	Ratio of male residents aged 45 to 54 total number of male residents	Number
MALE_55_64	Ratio of male residents aged 55 to 64 total number of male residents	Number
MALE_65_74	Ratio of male residents aged 65 to 74 total number of male residents	Number
MALE_75	Ratio of male residents aged 75+ to total number of male residents	Number
FEM_0_4	Ratio of female residents aged 0 to 4 to total number of female residents	Number
FEM_5_14	Ratio of female residents aged 5 to 14 to total number of female residents	Number
FEM_15_19	Ratio of female residents aged 15 to 19 to total number of female residents	Number
FEM_20_24	Ratio of female residents aged 20 to 24 to total number of female residents	Number
FEM_25_34	Ratio of female residents aged 25 to 34 to total number of female residents	Number
FEM_35_44	Ratio of female residents aged 35 to 44 to total number of female residents	Number
FEM_45_54	Ratio of female residents aged 45 to 54 to total number of female residents	Number
FEM_55_64	Ratio of female residents aged 55 to 64 to total number of female residents	Number
FEM_65_74	Ratio of female residents aged 65 to 74 to total number of female residents	Number
FEM_75	Total number of female residents aged 75+	Number
MF_0_4	Total number of male and female residents aged 0 to 4	Number
MF_5_14	Total number of male and female residents aged 5 to 14	Number

MF_15_19	Total number of male and female residents aged 15 to 19	Number
MF_20_24	Total number of male and female residents aged 20 to 24	Number
MF_25_34	Total number of male and female residents aged 25 to 34	Number
MF_35_44	Total number of male and female residents aged 35 to 44	Number
MF_45_54	Total number of male and female residents aged 45 to 54	Number
MF_55_64	Total number of male and female residents aged 55 to 64	Number
MF_65_74	Total number of male and female residents aged 65 to 74	Number
MF_75	Total number of male and female residents aged 75+	Number
OTHER_CNT	Total number of other residents	Number
OTHER_0_4	Total number of other residents aged 0 to 4	Number
OTHER_5_14	Total number of other residents aged 5 to 14	Number
OTHER_15_19	Total number of other residents aged 15 to 19	Number
OTHER_20_24	Total number of other residents aged 20 to 24	Number
OTHER_25_34	Total number of other residents aged 25 to 34	Number
OTHER_35_44	Total number of other residents aged 35 to 44	Number
OTHER_45_54	Total number of other residents aged 45 to 54	Number
OTHER_55_64	Total number of other residents aged 55 to 64	Number
OTHER_65_74	Total number of other residents aged 65 to 74	Number
OTHER_75	Total number of other residents aged 75+	Number
ASSESSED_VALUE	Average property assessment of the community (million dollars)	Number
MEDIAN_INCOME_HUNDRED_THOUSAND	Median income of families in the community (hundred thousand dollars)	Number

References

1. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
2. https://www.analyticsvidhya.com/blog/2015/08/beginners-guide-learn-content-based-recommender-systems/
3. https://data.calgary.ca/Demographics/Census-by-Community-2019/rkfr-buzb
4. https://data.calgary.ca/dataset/2020-Assessed-Property-Values/qwrb-nw8u
5. https://data.calgary.ca/dataset/2020-Assessed-Property-Values/qwrb-nw8u