SOPHISTICATED URBAN PLANNING WITH MACHINE LEARNING

The Battle of Neighborhoods

Capstone Project – IBM Data Science Professional Certificate

Yeji Soh

Contents

1. Introduction

Urban Planning has been given a big importance for any city to have orderly development in urban and suburb areas. It includes such techniques as predicting population growth, zoning, geographic mapping/analysis, and identifying the way that land has been used. **[1]**

Clustering, as one of the most popular unsupervised Machine Learning concepts, is widely used to divide data points into a certain number of groups in the way that data points in the same groups, i.e. clusters, share more similar attributes compared to others in different groups.**[2]**

In the sense that urban planning requires various segmenting studies by the extent of development, clustering will be a good method to be used in terms of designing and planning neighborhoods.

This article is prepared to show how machine learning can be used to plan the development of cities or communities, considering different factors, for example, the number of residents, the number of dwellings, how close each community is to other developed communities, what types of amenities each community has and so on.

This article is focused on 'Developing' communities of city of Calgary, since these are the communities where city of Calgary officially considers having potentials to grow and are currently "developing". This study can be used for

- urban planners, by suggesting how to have impartial development for each neighborhood,
- potential business owners, by suggesting where to avoid, e.g. supersaturated area, and where to target to open their businesses.

2. Data

The official website of City of Calgary kindly made their census data**[3]** open to public. This census data includes the number of dwelling units and population for each unit, where it belongs to a certain sector (e.g. Centre, East, West, North, South, NE, NW, SE, SW) and is categorized as either 'built-out', 'developing', 'non-residential' or 'N/A' depending on the yearly development. Also, a class code represents the type of areas, which is categorized by four: Residential(1), Industrial(2), Major Park(3), and Residual Sub Area(4). Note that residual sub areas were deleted in this study since they don't contain any data.

To start with, the API call was made to access to the most recent census data . A 'requests' module of Python is used and the data is received in json format. Not all the columns of data are necessary, so data cleaning should be carried out in here to have only essential data in a data frame.

Next, using the name of neighborhood, the latitude and longitude for each neighborhood will be obtained with geocoder package in Python. These geographical data will be used in making an API call to Four Square to get nearby venues to understand what kind of business is needed for potential developing neighborhoods. For the data analysis, a variety of data science techniques are to be utilized, such as data wrangling, exploratory data analysis and K-Means Clustering.

2.1 Census Data

Census data called from API service of city of Calgary is shown in ***Figure 1***. In total, there are 306 neighborhoods, however, after excluding neighborhoods belonging to residual sub areas, it becomes 258 neighborhoods. Also, city of Calgary has 37 “Developing”, 167 “Built-out” and 54 “Non-Residential” or “N/A” neighborhoods. The number of keys, corresponding to columns in data frames, are 142 in total, and out of 142, only 6 key information was extracted to be used in the analysis, which are

- **Class Code**: 1=Residential, 2=Industrial, 3=Major Park, 4=Residual Sub Area (deleted)
- **Neighborhood**: Full name of the community district approved by City Council
- **Sector**: Planning sector polygon where the community is located: Centre, East, West, North, Northeast, Northwest, South, Southeast, Southwest
- **SRG (Suburban Residential Growth)**: shows the yearly development capacity or housing supply: BUILT-OUT, DEVELOPING, NON RESIDENTIAL, and N/A
- **Total Residents**: Total number of residents living in the community
- **Total Dwellings**: Total number of dwellings in the community

```
[{'class': 'Residential',
  'class_code': '1.0',
  'comm_code': 'LEG',
  'name': 'LEGACY',
  'sector': 'SOUTH',
  'srg': 'DEVELOPING',
  'comm_structure': 'BUILDING OUT',
  'cnss_yr': '2019',
  'res_cnt': '6420.0',
  'dwell_cnt': '2766.0',
  'prsch_chld': '850.0',
  'elect_cnt': '0.0',
  'emplyd_cnt': '0.0',
  'ownshp_cnt': '1826.0',
  'dog_cnt': '0.0',
  'cat_cnt': '0.0',
  'pub_sch': '1071.0',
  'sep_sch': '506.0',
  'pubsep_sch': '175.0',
  'other_sch': '107.0',
```


Figure 1 Raw Census Data in a json format

Census data merged with geographical information of each neighborhood, i.e. latitude and longitude, is shown in ***Figure 2***. Please note that some of communities, in case they were too close to other communities enough to be included there, were not updated yet on geocoder package, which resulted in n/a values for their latitudes and longitudes. For further data analytical purposes, those neighborhoods with n/a values of geographical information were excluded; 258 number of neighborhoods were changed to 247.

Class Code	Neighborhood	Sector	SRG	Total Residents	Total Dwellings	Latitude	Longitude
1.0	LEGACY	SOUTH	DEVELOPING	6420.0	2766.0	50.856893	-114.002560
1.0	HIGHLAND PARK	CENTRE	BUILT-OUT	3838.0	2277.0	51.085355	-114.065809
1.0	CORNERSTONE	NORTHEAST	DEVELOPING	2648.0	1285.0	51.160280	-113.939608
1.0	MONTGOMERY	NORTHWEST	BUILT-OUT	4515.0	2013.0	51.074802	-114.162474
1.0	TEMPLE	NORTHEAST	BUILT-OUT	10977.0	3733.0	51.088424	-113.947877
...	...	...	...	...	...	...	...
2.0	FRANKLIN	NORTHEAST	N/A	0.0	3.0	51.047119	-113.994615
2.0	STONEGATE LANDING	NORTHEAST	N/A	0.0	0.0	51.163521	-113.985042
1.0	CAPITOL HILL	CENTRE	BUILT-OUT	4744.0	2440.0	51.071100	-114.101286
1.0	HIDDEN VALLEY	NORTH	BUILT-OUT	11566.0	3880.0	51.151085	-114.112672
1.0	RIVERBEND	SOUTHEAST	BUILT-OUT	9244.0	3474.0	50.974229	-114.014999

Figure 2 Census data with geographical information of each neighborhood

2.2 Amenities/Facilities Data

To be able to suggest what kind of amenities or facilities developing communities are lack of, it is necessary to figure out first what types of amenities they have and how many they have currently. *Error! Reference source not found.* illustrates the total number of amenities in developing neighborhoods. ***Figure 4*** shows top 20 built-out areas ordered by the total number of amenities. Here, it can be concluded that for developing areas, it is meaningless to say what is lacking or needs to be brought into the communities, because except Aspen Woods and Mckenzie Towne, most of the communities have around 10 or less than 10 amenities. Even for Aspen Woods and Mckenzie Towne, their numbers of amenities are not big figures compared to the ones of built-out areas. Considering developing neighborhoods are insufficient of all types of amenities overall, let's have a look at where these areas are located in the next section.

Neighborhood	Total Amenities
ASPEN WOODS	28
MCKENZIE TOWNE	17
HASKAYNE	11
CHAPARRAL	6
COUGAR RIDGE	6
WEST SPRINGS	5
CRANSTON	5
TARADALE	4
PANORAMA HILLS	4
SAGE HILL	4
COPPERFIELD	4
EVERGREEN	3
HOMESTEAD	3
AUBURN BAY	3
CITYSCAPE	3
SETON	2
NOLAN HILL	2
REDSTONE	2
SADDLE RIDGE	2
SKYVIEW RANCH	1
SHERWOOD	1
BELMONT	1
NEW BRIGHTON	1
ROCKY RIDGE	1
MAHOGANY	1
LEGACY	1
CORNERSTONE	1
YORKVILLE	1

Figure 3 Total number of amenities in developing neighborhoods

Neighborhood	Total Amenities
BELTLINE	56
LOWER MOUNT ROYAL	51
DOWNTOWN COMMERCIAL CORE	49
EAU CLAIRE	46
MISSION	37
CLIFF BUNGALOW	36
MEADOWLARK PARK	28
COUNTRY HILLS VILLAGE	27
BRIDGELAND/RIVERSIDE	22
DOWNTOWN EAST VILLAGE	21
TUXEDO PARK	21
HILLHURST	17
KINGSLAND	17
BANFF TRAIL	17
HOUNSFIELD HEIGHTS/BRIAR HILL	16
RIVERBEND	15
LAKE BONAVISTA	15
CRESCENT HEIGHTS	15
PARKHILL	15
INGLEWOOD	13

Figure 4 Top 20 Built-out neighborhoods by the number of amenities

3. Data Analysis

Figure 5 displays 247 neighborhoods of Calgary: blue for Built-Out, Red for Developing, and Green for Other areas. It is observed that some developing areas are close to a number of built-out areas, while others are not. Especially 'Haskayne'(Developing) is all surrounded by built-out neighborhoods. It can be a logical conclusion that easy access to built-out areas would be one of the key factors for the independent development of developing neighborhoods. So, another data to be added for further analysis is how many nearby(within 5km) built-out areas each developing neighborhood has. There is a 'distance' method under geopy module in Python that allows to measure a distance with latitudes and longitudes. The number of built-out neighborhoods nearby(within 5km) for each developing community, is found in ***Figure 6***.

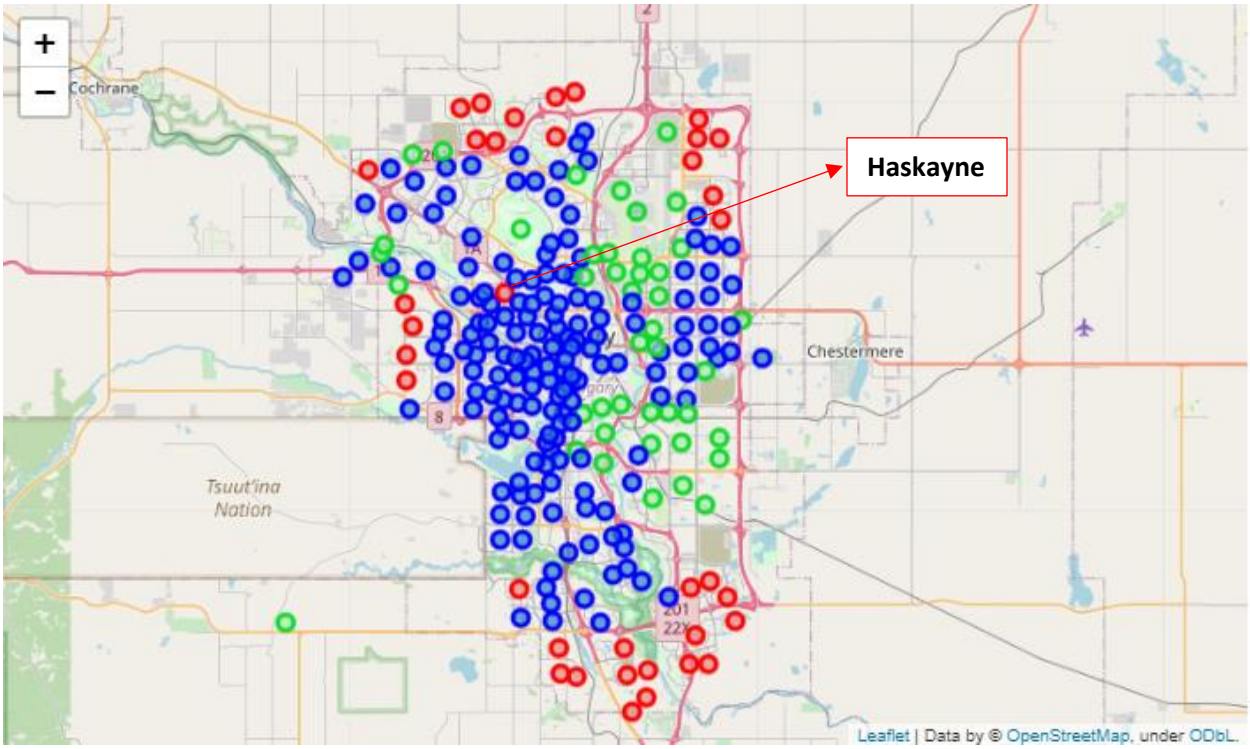

Figure 5 Map of Built-Out, Developing, and Other Neighborhoods of Calgary

4. Data Analysis with Anomaly Detection – K-Means Clustering

K-Means Clustering is a type of partitioning clustering, as one of unsupervised algorithms, where it divides the data into certain k number of non-overlapping subsets or clusters without any explicit labels. As a result, objects within the same cluster shows high similarities to each other, while objects across different clusters are very different.**[4]**

To cluster developing neighborhoods by similarities, it is critical to set up the criteria that would like to be compared. In this study, they would be

- Total Residents,
- Total number of Amenities,
- Number of Built-Out Neighborhoods nearby (within 5km)

Since these three criteria use different scale or range of data, it is important to normalize data first so that data can be set at the similar level of scale. ***Figure 6*** shows raw data set before normalization, while ***Figure 7*** represents the data after normalization.

Now, what is the proper number of clusters? Here, "Elbow Method" can be used. The elbow method could operate k-mean clustering on the dataset for certain range of k values and RSS (Residual Sum of Square) is calculated for each k value. When RRS values are plotted with each k value, the line chart would look like an arm, and here "elbow" (the inflection point) can be chosen visually.**[5]** ***Figure 8*** describes how RSS value changes with k value, and it can be easily observed that the optimal k number in this study is three.

Now that every data point, i.e. every developing neighborhood, is assigned certain cluster number, the average values for each cluster can be known in ***Table 1***. However, by looking at the number of built-out areas close-by, they present high similarity across the clusters, which should not be in proper clustering analysis. It is a good time to figure out possible anomalies in the dataset.

Figure 10 displays how data points are distributed, and it can be observed clearly that we have three anomalies (red circles in the figure) for each criterion: total residents, total amenities, and number of close built-out areas. Note that purple is for Cluster 0, green is for Cluster 1, and yellow is for Cluster 2. The anomalies for Cluster 0 and Cluster 1 are differentiated by having much higher number of total amenities, whereas the one for Cluster 2 is distinguished by having higher figure on the number of close built-out areas. Now, it is known that Aspen Woods, Mckenzie Towne, and Haskayne are the anomalies and they need to be excluded in clustering study, because they are showing their own characteristics. (***Figure 11***, ***Figure 12*** , and ***Figure 13***)

To see if it is still valid to use three clusters after anomalies are removed, elbow method is re-used to the dataset excluding three anomalies. The result was still the same showing three clusters are optimal for this study(***Figure 9***). "Corrected" cluster-representative values are shown in ***Table 2***. It is normal for some data points to change their belonged cluster one to another, however, in this study, it was checked that all the data points stayed in the same cluster even after getting rid of anomalies.

	Neighborhood	Total Residents	Total Amenities	No of BuiltOut < 5km
0	LEGACY	6420	1.0	0
1	CORNERSTONE	2648	1.0	1
2	ASPEN WOODS	9446	26.0	13
3	COUGAR RIDGE	6997	6.0	15
4	AUBURN BAY	17607	4.0	2
5	BELMONT	86	1.0	5
6	TARADALE	19026	5.0	8
7	ROCKY RIDGE	8398	1.0	7
8	SHERWOOD	6246	NaN	7
9	REDSTONE	5848	2.0	0
10	SKYVIEW RANCH	11707	NaN	1
11	SADDLE RIDGE	22321	2.0	6
12	SILVERADO	7655	1.0	7
13	WALDEN	6228	NaN	1
14	COPPERFIELD	13823	5.0	1
15	YORKVILLE	14	1.0	5
16	PINE CREEK	14	NaN	0
17	CHAPARRAL	12654	7.0	7
18	WOLF WILLOW	0	NaN	2
19	KINCORA	6889	NaN	8
20	CITYSCAPE	3104	2.0	2
21	LIVINGSTON	1477	NaN	4
22	HASKAYNE	0	10.0	40
23	CRANSTON	19884	4.0	1
24	SETON	1134	2.0	1
25	MAHOGANY	11784	1.0	1
26	HOMESTEAD	0	4.0	0
27	SAGE HILL	7924	3.0	3
28	NOLAN HILL	7505	2.0	3
29	EVANSTON	17685	NaN	7
30	CARRINGTON	572	NaN	5
31	WEST SPRINGS	10758	5.0	20
32	MCKENZIE TOWNE	18283	21.0	5
33	NEW BRIGHTON	13103	1.0	2
34	PANORAMA HILLS	25710	1.0	9
35	EVERGREEN	21500	6.0	13
36	SPRINGBANK HILL	9943	NaN	12

Figure 6 Final Dataset to be used in Clustering Analysis

```
array([[-0.36199724, -0.44089372, -0.8238747 ],
       [-0.88363089, -0.44089372, -0.68778825],
       [ 0.05647132,  4.19350045,  0.94524909],
       [-0.28220334,  0.48598512,  1.21742198],
       [ 1.18506415,  0.11523358, -0.5517018 ],
       [-1.23793242, -0.44089372, -0.14344247],
       [ 1.38129907,  0.30060935,  0.26481687],
       [-0.08845765, -0.44089372,  0.12873042],
       [-0.38605988, -0.62626948,  0.12873042],
       [-0.44109969, -0.25551795, -0.8238747 ],
       [ 0.36914726, -0.62626948, -0.68778825],
       [ 1.83696791, -0.25551795, -0.00735602],
       [-0.19120786, -0.44089372,  0.12873042],
       [-0.38854911, -0.62626948, -0.68778825],
       [ 0.66177101,  0.30060935, -0.68778825],
       [-1.24788937, -0.44089372, -0.14344247],
       [-1.24788937, -0.62626948, -0.8238747 ],
       [ 0.50010884,  0.67136089,  0.12873042],
       [-1.24982545, -0.62626948, -0.5517018 ],
       [-0.29713876, -0.62626948,  0.26481687],
       [-0.82057019, -0.25551795, -0.5517018 ],
       [-1.04556964, -0.62626948, -0.27952891],
       [-1.24982545,  1.22748819,  4.61958311],
       [ 1.49995275,  0.11523358, -0.68778825],
       [-1.09300346, -0.25551795, -0.68778825],
       [ 0.37979567, -0.44089372, -0.68778825],
       [-1.24982545,  0.11523358, -0.8238747 ],
       [-0.15400758, -0.07014218, -0.41561536],
       [-0.21195151, -0.25551795, -0.41561536],
       [ 1.19585084, -0.62626948,  0.12873042],
       [-1.170723  , -0.62626948, -0.14344247],
       [ 0.2379091 ,  0.30060935,  1.89785421],
       [ 1.27854886,  3.26662162, -0.14344247],
       [ 0.56220149, -0.44089372, -0.5517018 ],
       [ 2.3056361 , -0.44089372,  0.40090331],
       [ 1.723431  ,  0.48598512,  0.94524909],
       [ 0.12520194, -0.62626948,  0.80916265]])
```

Figure 7 Normalized Dataset

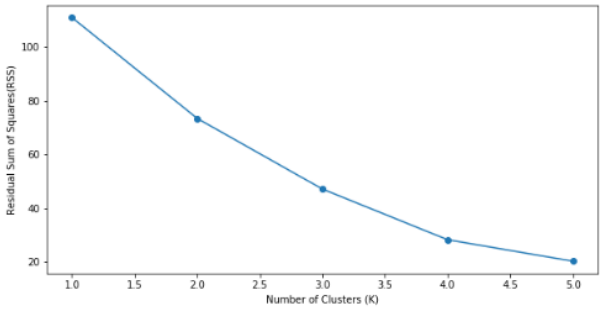

Figure 8 RSS vs K Before Anomaly Detection

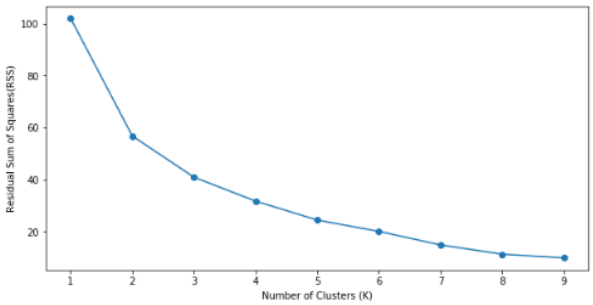

Figure 9 RSS vs K After Anomaly Detection

Table 1 Representative average values for each Cluster before anomaly detection

Cluster No.	Total Residents	Total Amenities	No of Built-Out < 5km
0	9073.778	3.389	6.000
1	20252.000	5.375	6.375
2	822.638	1.909	5.909

Table 2 Representative average values for each Cluster after anomaly detection

Cluster No.	Total Residents	Total Amenities	No of Built-Out < 5km
0	9051.882	2.058	5.588
1	20533.286	3.143	6.571
2	904.900	1.100	2.500

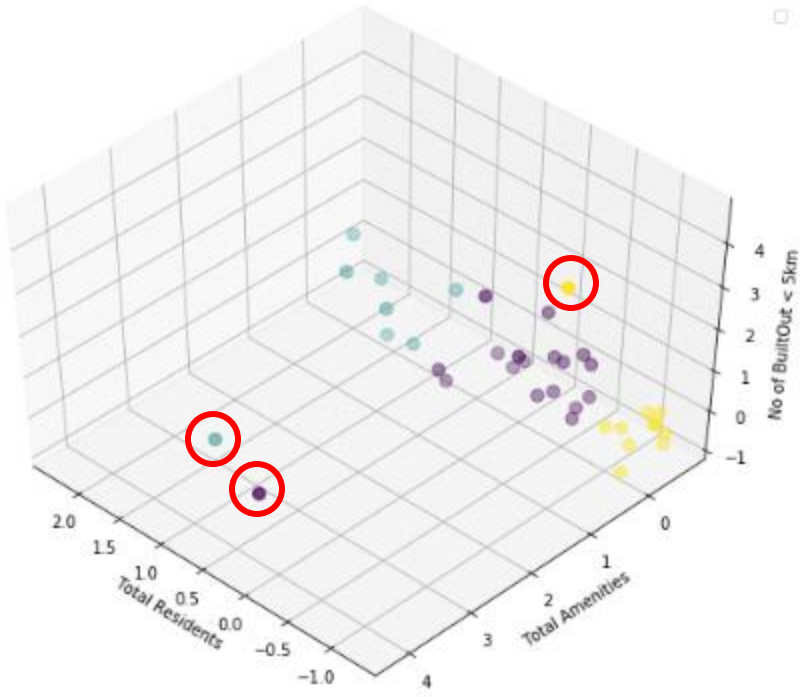

Figure 10 Plot of dataset (Purple: Cluster 0, Green: Cluster 1, Yellow: Cluster 2)

Neighborhood	Total Residents	Total Amenities	No of BuiltOut < 5km	Cluster No	Latitude	Longitude
ASPEN WOODS	9446	26.0	13	0	51.043119	-114.210185
CHAPARRAL	12654	7.0	7	0	50.883594	-114.021265
COUGAR RIDGE	6997	6.0	15	0	51.070710	-114.210968
COPPERFIELD	13823	5.0	1	0	50.912048	-113.932098
WEST SPRINGS	10758	5.0	20	0	51.058822	-114.204254
SAGE HILL	7924	3.0	3	0	51.178975	-114.145868
REDSTONE	5848	2.0	0	0	51.170807	-113.957483
NOLAN HILL	7505	2.0	3	0	51.177160	-114.163274
NEW BRIGHTON	13103	1.0	2	0	50.920726	-113.947085
MAHOGANY	11784	1.0	1	0	50.898702	-113.925905
LEGACY	6420	1.0	0	0	50.856893	-114.002560
SILVERADO	7655	1.0	7	0	50.884366	-114.078035
ROCKY RIDGE	8398	1.0	7	0	51.143274	-114.242722
KINCORA	6889	0.0	8	0	51.158447	-114.133026
WALDEN	6228	0.0	1	0	50.869568	-114.018688
SKYVIEW RANCH	11707	0.0	1	0	51.160534	-113.958135
SHERWOOD	6246	0.0	7	0	51.159564	-114.148693
SPRINGBANK HILL	9943	0.0	12	0	51.028926	-114.209793

Figure 11 Developing Neighborhoods in Cluster 0 (ASPEN WOODS: Anomaly)

Neighborhood	Total Residents	Total Amenities	No of BuiltOut < 5km	Cluster No	Latitude	Longitude
MCKENZIE TOWNE	18283	21.0	5	1	50.916499	-113.964353
EVERGREEN	21500	6.0	13	1	50.916379	-114.111578
TARADALE	19026	5.0	8	1	51.116704	-113.938464
AUBURN BAY	17607	4.0	2	1	50.890605	-113.959565
CRANSTON	19884	4.0	1	1	50.875210	-113.965956
SADDLE RIDGE	22321	2.0	6	1	51.129706	-113.944796
PANORAMA HILLS	25710	1.0	9	1	51.160946	-114.081322
EVANSTON	17685	0.0	7	1	51.171292	-114.116352

Figure 12 Developing Neighborhoods in Cluster 1 (MCKENZIE TOWNE: Anomaly)

Neighborhood	Total Residents	Total Amenities	No of BuiltOut < 5km	Cluster No	Latitude	Longitude
HASKAYNE	0	10.0	40	2	51.076889	-114.124009
BELMONT	86	1.0	5	2	50.868495	-114.062773
YORKVILLE	14	1.0	5	2	50.870531	-114.076523
CARRINGTON	572	0.0	5	2	51.183050	-114.080296
LIVINGSTON	1477	0.0	4	2	51.184978	-114.064808
WOLF WILLOW	0	0.0	2	2	50.871446	-114.001443
CITYSCAPE	3104	2.0	2	2	51.148549	-113.962668
CORNERSTONE	2648	1.0	1	2	51.160280	-113.939608
SETON	1134	2.0	1	2	50.874878	-113.948915
PINE CREEK	14	0.0	0	2	50.849703	-114.014795
HOMESTEAD	0	4.0	0	2	53.431253	-113.493237

Figure 13 Developing Neighborhoods in Cluster 2 (HASKYANE: Anomaly)

5. Results & Discussion

From the final result in ***Table 2***, all the three categories: total number of residents, total number of amenities, and the number of close(within 5km) built-out areas, are proportional to each other. In other words, the higher number of total residents is, the higher number of total amenities, as well as, the number of built-areas are. Therefore, it would be valid to determine these three clusters by the level of independency, so from now on, they are titled with “Low”, “Med”(for Medium), and “High”. ***Figure 14*** shows the map of developing neighborhoods marked in different colors by each cluster: pink for Cluster 0(Med), purple for Cluster 1(High), and orange for Cluster 2(Low), along with built-out neighborhoods marked in blue color.

Following suggestions can be made for each cluster for urban and regional planners to come up with different plans on city development, or for any entrepreneurs to initiate their future businesses.

Cluster 2 (Low)	If looking at the map, all the neighborhoods, belonging to this cluster, are located around the outer boundaries of Calgary. This cluster "Low" is rather

	close to cluster "Med" than cluster "High" or built-out areas. So, for the time being that cluster "Low" is catching up with its own development, it would be wiser to develop cluster "Med" areas first to have more number of amenities, which attracts more residents to come and live and makes "Med" to "High" level of development areas. This will naturally lead to following development of cluster "Low", with being located close to semi-cluster "High".
Cluster 0 (Med)	All the neighborhoods in cluster “Med” are located around the outer boundaries, just like cluster "Low", but with being closer to built-out districts. West Srpings, Cougar Ridge, and Springbank Hill are not recommended areas for low-risk seeking business persons to start their businesses. Although they show low number of amenities within their areas, since quite many number of built-out areas, over ten, are closeby, their new businesses should compete not only with the ones within the neighborhood, but also with the ones in those built-out areas. Therefore, low-risk-seeking business persons would rather like to start their businesses in the neighborhoods within this cluster, except West Srpings, Cougar Ridge, and Springbank Hill. However, it could also be interpreted that, these three areas are very "sexy" in terms of having possibilities to target people in built-out areas closeby, as well as people within the neighborhood, if the business is attractive enough to compete over the ones in built-out areas.
Cluster 1 (High)	All the neighborhoods in this cluster would be "sweetpies" for potential business persons, because each neighborhood already secures large enough number of residents(average 20,000) who could be potential customers, even if competition with businesses in close built-out areas is inevitable.
Anomalies:	
HASKAYNE	Since Haskayne is one of the University of Calgary’s school districts, this area has zero resident but high number of amenities that students can use. Also, it is surrounded by many built-out areas.
ASPEN WOODS	Aspen Woods has abnormally high number of amenities compared to any other developing areas.
MCKENZIE TOWNE	Mckenzie Towne has a similar characteristic with Aspen Woods in terms of having high number of amenities, but with double population than the one in Aspen Woods. It is expected to be categorized as a built-out area in near future, assuming more amenities and residents are coming in.

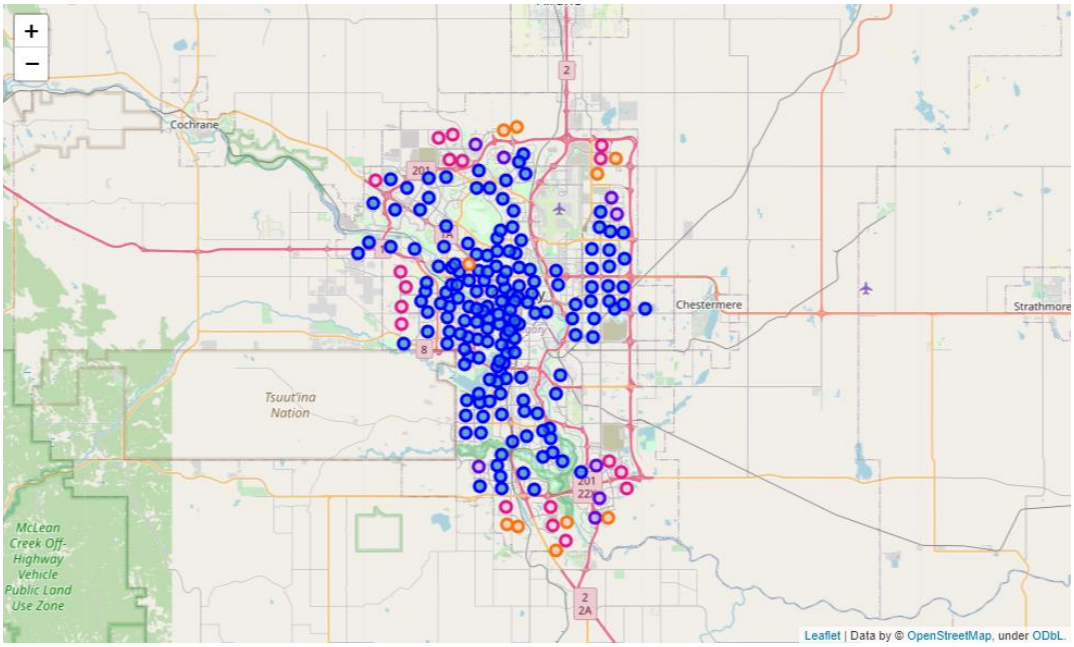

Figure 14 Map of Developing neighborhoods by cluster, and Built-Out Areas

6. Conclusions

To build up specific development plans for "developing" neighborhoods, since city of Calgary specifies them as the areas with high potentials to grow, machine learning, more specifically clustering analysis, was successfully used to divide 37 developing areas into 3 groups and suggest strategic urban plans for each one of them. 3 groups were categorized by the level of independency measured with total number of residents, total number of amenities, and number of "built-out" neighborhoods close to each developing one. Although there are many other various factors to be reflected in real-life studies, I believe this study is meaningful enough to consider the external factor: how close built-out areas can affect developing ones' independent developments, as well as internal factors, such as populations or the number of facilities. Also, the analytical model in this study has a high flexibility to be used in other big cities that would like to enhance their urban planning strategies.

7. Reference

[1] https://en.wikipedia.org/wiki/Urban_planning

[2] https://en.wikipedia.org/wiki/Cluster_analysis

[3] https://data.calgary.ca/Demographics/Census-by-Community-2019/rkfr-buzb

[4] 'Machine Learning with Python' course from IBM Data Science Professional Certificate

[5] https://www.scikit-yb.org/en/latest/api/cluster/elbow.html