FINDING THE BEST COMMUNITIES FOR NEWCOMERS TO CALGARY, ALBERTA, CANADA

IBM DATA SCIENCE PROFESSIONAL CERTIFICATION, COURSERA

APPLIED DATA SCIENCE CAPSTONE PROJECT

ARPITA CHAKRAWARTI

DECEMBER 6, 2020

1.0 INTRODUCTION

1.1 BACKGROUND

Calgary, AB is a beautiful city in western Canada very well known as the centre of Canada's oil industry. Per the fact sheet provided here,

- Calgary is currently the fourth largest city in Canada with the highest immigrant population next to Toronto, Vancouver and Montreal.
- By 2020, Calgary's total immigrant population is estimated to reach almost half a million.
- The Philippines, India, and China continue to be the lead source countries for immigrants to Calgary.

A very large percentage of the immigrant population belongs to working ages. It is often a struggle for new immigrants and newcomers to find the right neighbourhoods and communities, which can provide the right amenities, safety, affordable housing and good schools. It is therefore advantageous for newcomers to be aware of some general information about the various residential locations of this city.

1.2 PROBLEM

Data that might help determine which neighborhoods or communities may be preferable might include house prices, crime rates, population, popular amenities or venues. This project aims to find groups of communities which are similar in nature, based on this data.

1.3 INTEREST

The following may be interested in the analysis –

- New immigrants and newcomers to Canada
- The city of Calgary
- And others who may be interested in conducting a similar analysis of other cities

2.0 DATA ACQUISITION AND CLEANING

2.1 DATA SOURCES

All data for Calgary city, its crimes and property assessments can be found here, here and here . Data for venues can be found using the foursquare.com Places API – described here.

2.1.1 COMMUNITIES OF CALGARY

The data provided by the city of Calgary consists of 309 communities consisting of residential, industrial, residual sub area and major parks. For this exercise, only the 218 residential communities will be considered. This also consists of the latitude and longitude of the centre point of each community. Boundaries of communities are also provided by the city of Calgary which can be utilized for visual representation of data.

2.1.2 ANNUAL CRIME STATISTICS PER COMMUNITY

This data consists of total number of crimes and total population for each community for several years, starting 2012 until 2019. Only the 2019 statistics will be used for analysis.

2.1.3 2020 ASSESSED PROPERTY VALUES

This data consists of assessed value of all properties found within this community. The median of the assessed valuations by community will be calculated and used for further assessment.

2.1.4 SCHOOLS

The data provided by the city of Calgary for a complete list of schools and locations, including private and public schools can be found here. This data does not include the names of the communities the schools are located in. In order to compensate for this information, I have manually collected this information by using this website provided by the Calgary Board of Education. I have been able to map 248 Calgary Public Schools to the communities. Private schools have not been mapped.

2.2 DATA CLEANSING

Data downloaded from multiple sources have been combined into a single table with the communities as the key. Only residential communities have been considered in the analysis.

Some data cleansing has been required from each of the files, to retain the most recent data and to include only the communities of interest.

Features selected for the analysis include

- Total crimes per capita by community
- Total population by community
- Median value of assessed property value by community
- Counts of venues by a set of selected categories

3.0 METHODOLOGY

3.1 FINDING THE COMMUNITIES TO EXPLORE

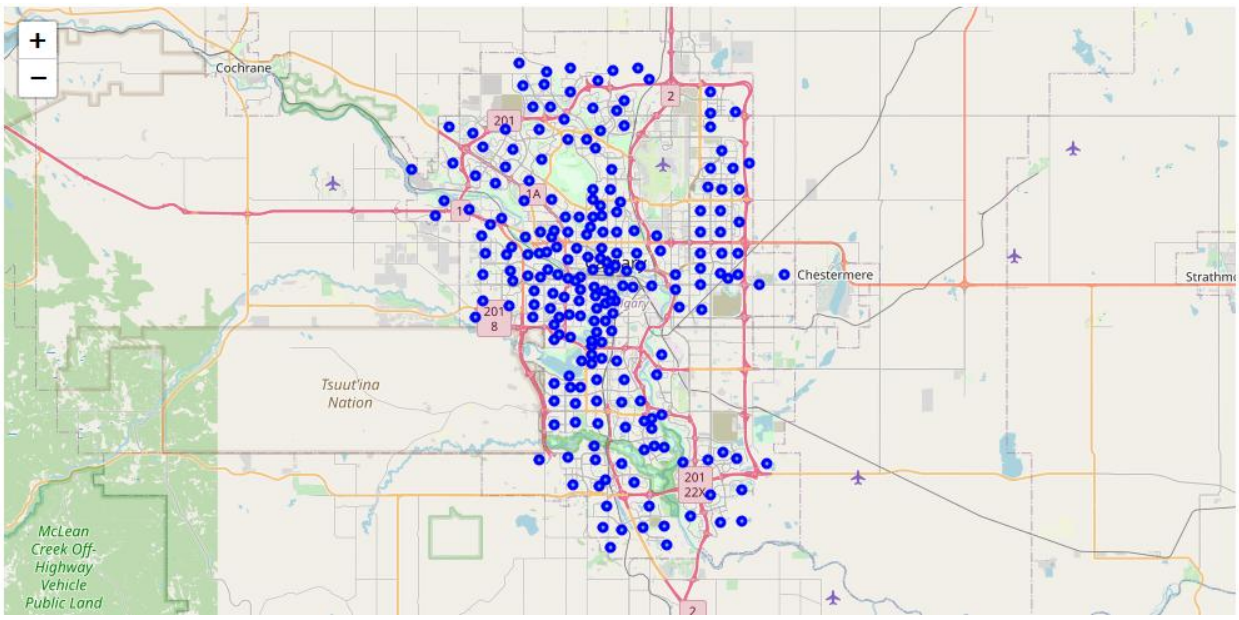
There are several communities in Calgary – 309 in all, as obtained from the open data provided.

	CLASS	CLASS_CODE	COMM_CODE	NAME	SECTOR	SRG	COMM_STRUCTURE	longitude	latitude	location
0	Residential	1	THS	TWINHILLS	EAST	DEVELOPING	BUILDING OUT	-113.877110	51.045111	(51.045111353378694, -113.87710975220665)
1	Residential	1	WIL	WILLOW PARK	SOUTH	BUILT-OUT	1960s/1970s	-114.056204	50.956623	(50.95662292848714, -114.05620363150967)
2	Residual Sub Area	4	05D	05D	NORTHEAST	NaN	UNDEVELOPED	-113.958662	51.179598	(51.17959764644064, -113.95866183876556)
3	Industrial	2	ST4	STONEY 4	NORTHEAST	NaN	EMPLOYMENT	-114.002762	51.176204	(51.17620448693238, -114.00276157771617)
4	Residential	1	PKH	PARKHILL	CENTRE	BUILT-OUT	1950s	-114.065552	51.018181	(51.01818071993347, -114.06555236114401)

Filtering out all but the Residential, left behind 218 communities loaded into a Pandas dataframe :

	CLASS	CLASS_CODE	COMM_CODE	NAME	SECTOR	SRG	COMM_STRUCTURE	longitude	latitude	location
0	Residential	1	THS	TWINHILLS	EAST	DEVELOPING	BUILDING OUT	-113.877110	51.045111	(51.045111353378694, -113.87710975220665)
1	Residential	1	WIL	WILLOW PARK	SOUTH	BUILT-OUT	1960s/1970s	-114.056204	50.956623	(50.95662292848714, -114.05620363150967)
4	Residential	1	PKH	PARKHILL	CENTRE	BUILT-OUT	1950s	-114.065552	51.018181	(51.01818071993347, -114.06555236114401)
5	Residential	1	PAT	PATTERSON	WEST	BUILT-OUT	1980s/1990s	-114.177047	51.063838	(51.06383775082155, -114.17704650860274)
6	Residential	1	RCK	ROSSCARROCK	WEST	BUILT-OUT	1950s	-114.145495	51.043280	(51.04328023810093, -114.14549516107789)

Folium maps were used to plot the centres of the communities on a map:



3.2 VENUES IN THE COMMUNITIES

As required by the Applied Data Science Capstone (IBM Data Science Professional Certificate on COURSERA), I used the FOURSQUARE Places API to explore venues in the communities. I used a limit of 100 venues per search and in a radius of 1km from the center of each community. This returned 4465 venues for 218 communities. This included 273 unique categories.

This data was reduced to only 13 categories which were found to be the highest in number and are likely the venue categories which are the most important in daily life.

The following depicts categories of interest by community:

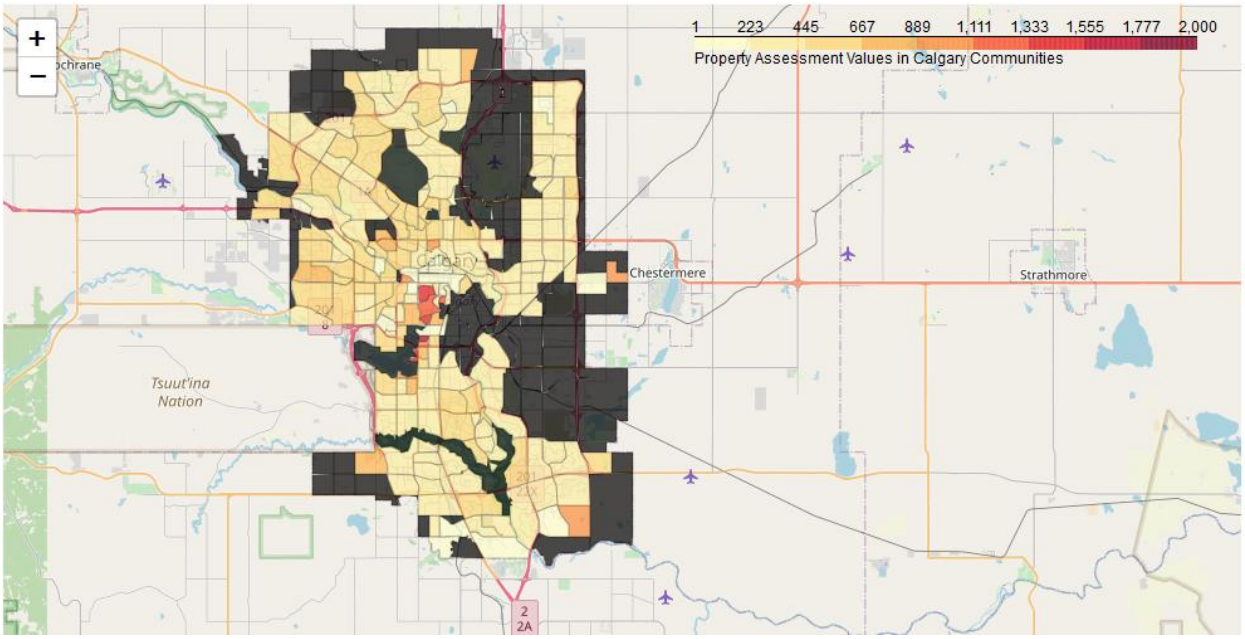
NAME	Bank	Café	Coffee Shop	Convenience Store	Fast Food Restaurant	Gas Station	Grocery Store	Gym / Fitness Center	Park	Pharmacy	Pizza Place	Playground	Real Estate Office
ABBEYDALE	0	1	0	1	0	0	0	0	0	0	0	0	0
ACADIA	0	0	1	1	0	0	0	1	0	1	0	0	0
ALBERT PARK/RADISSON HEIGHTS	1	0	0	1	1	0	1	0	0	1	2	0	0
ALPINE PARK	0	0	0	0	0	0	0	0	0	0	0	0	0
ALTADORE	0	0	1	0	0	0	0	0	1	0	0	0	0

3.3 2020 ASSESSED PROPERTY VALUES

This is a large data set which provides assessment values for every property in the 218 residential communities. This data was cleaned to exclude all properties which are not residential. Finally the median value by community was calculated to be used further as a feature in this study.

NAME	ASSESSED_VALUE
ABBEYDALE	289000.0
ACADIA	386000.0
ALBERT PARK/RADISSON HEIGHTS	299000.0
ALPINE PARK	700500.0
ALTADORE	721000.0

The Choropleth map of the Property Assessment Values is depicted below which can provide a better idea of the distribution of property values in the city –

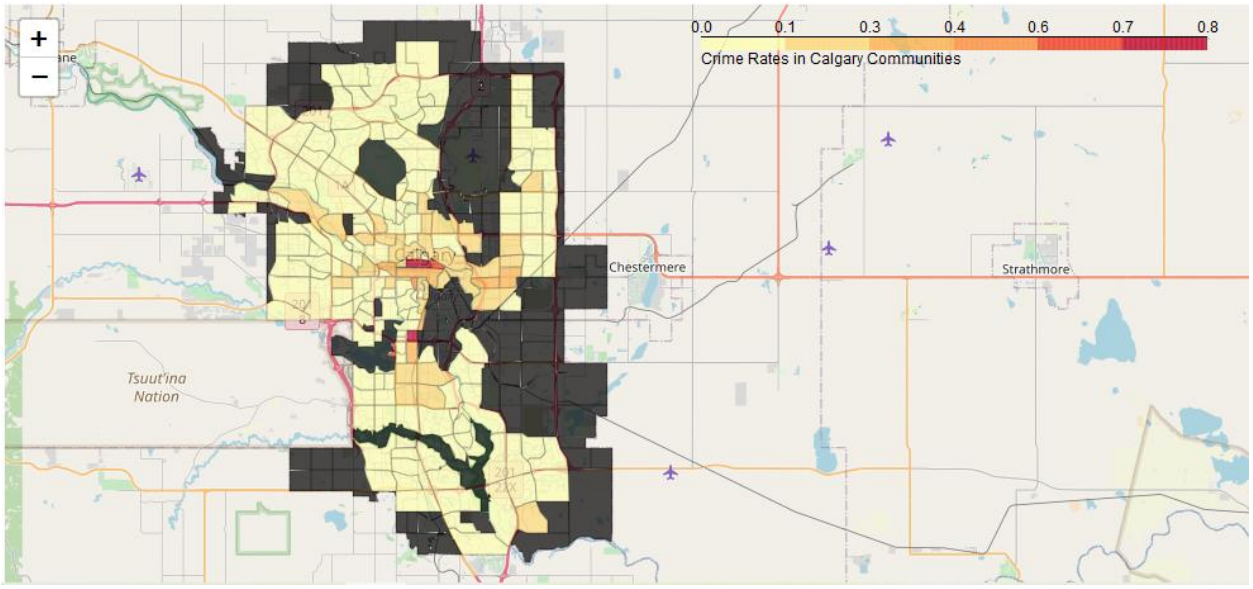


3.4 TOTAL CRIMES BY COMMUNITY

The Calgary city open data already contained the total number of crimes by community alongside the estimated population in each community. The data was cleaned to only evaluate the most recent numbers i.e. for the year 2019. For each community, the per capita number of crimes was calculated and carried forward as one of the features.

	NAME	CRIME_COUNT_2019	RESIDENT_COUNT_2019	Year	CRIME_PER_CAPITA_2019
0	SUNALTA	1049	3268	2019	0.320991
7	MAPLE RIDGE	161	1916	2019	0.084029
17	NEW BRIGHTON	406	13103	2019	0.030985
22	WESTGATE	149	3202	2019	0.046533
29	NOLAN HILL	179	7505	2019	0.023851

The following Folium map in the Choropleth style, depicts Crime Rates in communities –



3.5 SCHOOLS IN THE COMMUNITY

Finally, after mapping all the public schools in Calgary School Division to their respective communities, barring a handful, the number of schools per community could be estimated.

NAME	NUM_SCHOOLS
ABBEYDALE	1
ACADIA	3
ALBERT PARK/RADISSON HEIGHTS	2
ALTADORE	3
ARBOUR LAKE	2

3.6 COMBINED FEATURE SET

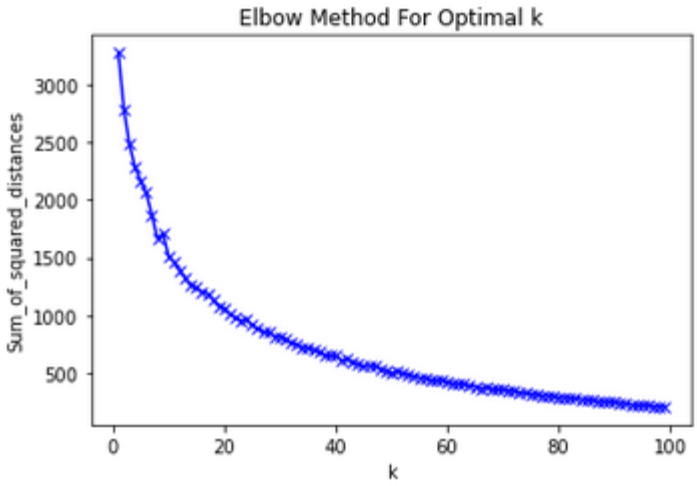
After combining the data for all the features using the names of the communities as the common field, the following feature set with 218 rows and 16 columns was created –

	Bank	Café	Coffee Shop	Convenience Store	Fast Food Restaurant	Gas Station	Grocery Store	Gym / Fitness Center	Park	Pharmacy	Pizza Place	Playground	Real Estate Office	ASSESSED_VALUE	CRIME_PER_C
0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	289000.0	
1	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	386000.0	
2	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	2.0	0.0	0.0	299000.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	700500.0	
4	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	721000.0	

3.7 K-MEANS CLUSTERING

K-Means Clustering will be used the technique to predict similar communities based on the values provided for the various features selected for the analysis.

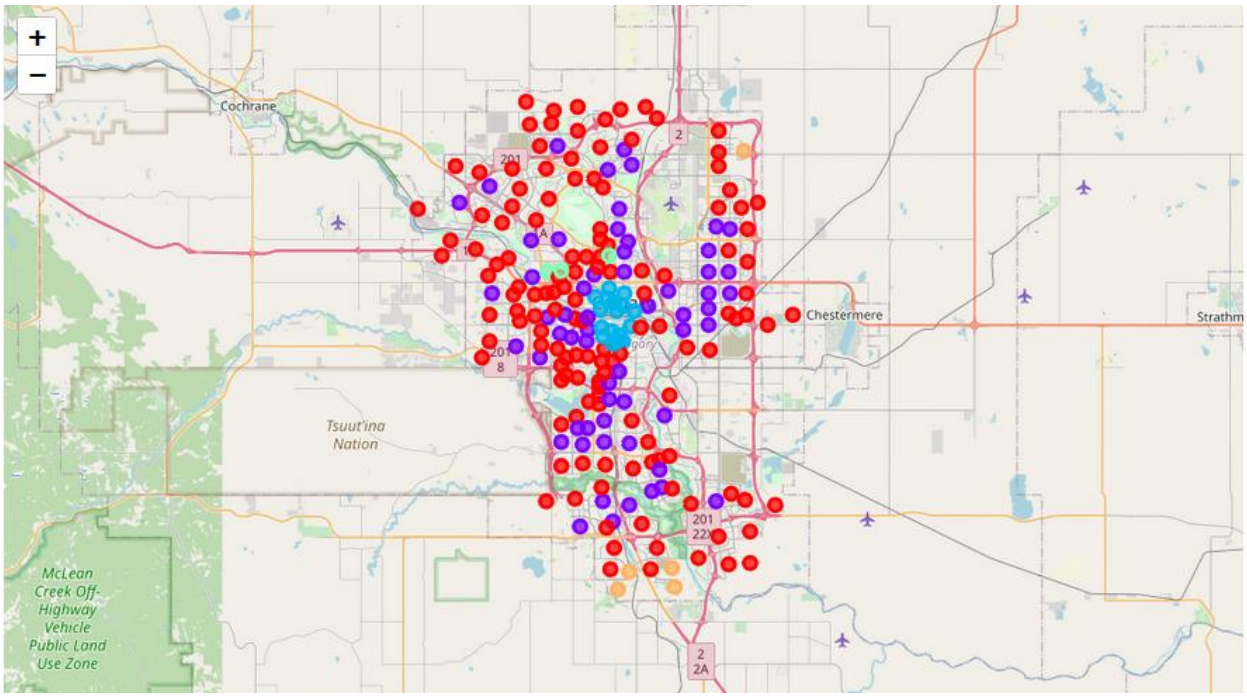
The elbow method for optimal K provided the following –



There is a sudden change in the sum of squared distances at the value of k=5 , the slope changes faster after that point so the value of 5 was chosen . Also, making an in-depth manual analysis of the changes in the feature values after the value of k=5 did not provide any additional insight which was another reason for continuing analysis with that value.

4.0 RESULTS

The following Folium map was created using the results of the K-Cluster analysis.



5 different types of communities were found –

CLUSTER 0 - RED CLUSTER POINTS

Characteristics:

- A large number of communities fall into this cluster - 134 out of 218.
- At least 1 or more schools,
- Low crime per capita,
- Most common venues can be found close by but seems to be less than other communities,
- Property rates higher than others

CLUSTER 1 - VIOLET CLUSTER POINTS

Characteristics:

- 59 out of 218 communities fall into this category, majority are built-out communities.
- 1 or more schools available
- Coffee shops, grocery stores, convenience stores, pizza places very commonly available in these communities.
- Property rates are low.
- Crime is higher than in other community clusters except for the city center

CLUSTER 2 - BLUE CLUSTER POINTS

Characteristics:

- 17 out of 218 communities fall into this category, all are built out and located in the center of the city
- Fewer schools available in these communities
- Plenty of cafes, parks and pizza places around.
- Property rates are the second highest as compared to the other clusters.
- Per capita Crime is highest here

CLUSTER 3 - GREEN CLUSTER POINTS

Characteristics:

- 3 out of 218 communities fall into this category, all around the university area.
- Very few schools available in these communities
- Some coffee shops are available.
- Property rates do not appear to be affordable.
- Per capita Crime is very low here

CLUSTER 4 - ORANGE CLUSTER POINTS

Characteristics:

- 5 out of 218 communities fall into this category, all around outskirts of the city and developing communities.
- There are no schools here
- There are no grocery stores or other useful venues around.
- Property rates are very affordable.
- Per capita Crime is close to none.

5.0 DISCUSSION

There is much scope to delve deeper into the areas of property pricing, crimes and schools. These have always been a high interest area and further features can be added to further enhance the quality of the analysis. Some of the features that could be included –

- Student results in schools
- Kinds of crime
- Property assessment using the mode values instead of median values

6.0 CONCLUSION

Further analysis of this nature could help city officials look at the overall living conditions within the city to further improve and ensure the safety and happiness of all citizens.