HOW TO MOVE TO A NEW CITY?

Aakash Vasudevan

1.0 Abstract:

This project aims to find neighborhoods in a city that closely match a desirable set of characteristics. To illustrate the methodology, the characteristics considered are the median age, income, population density and the proximity to certain types of venues like bars, restaurants and grocery store. The location data was obtained from the Foursquare database using their Places API. The median age, income and population density were obtained from reputable websites using web-scraping techniques. A feature matrix containing all these characteristics and their relative weights was assembled for all the neighborhoods in the new city. Then the K-means clustering algorithm was employed to find the neighborhoods that best matches our desired set of characteristics. The results were validated with Google search and domain knowledge of the author.

2.0 Introduction:

Moving cities is a major life change for most people. Leaving behind their life, community and relationships built over many years can seem overwhelming. The inertia is compounded by the uncertainty in the new environment and consequently, the time investment required for planning and acclimatization.

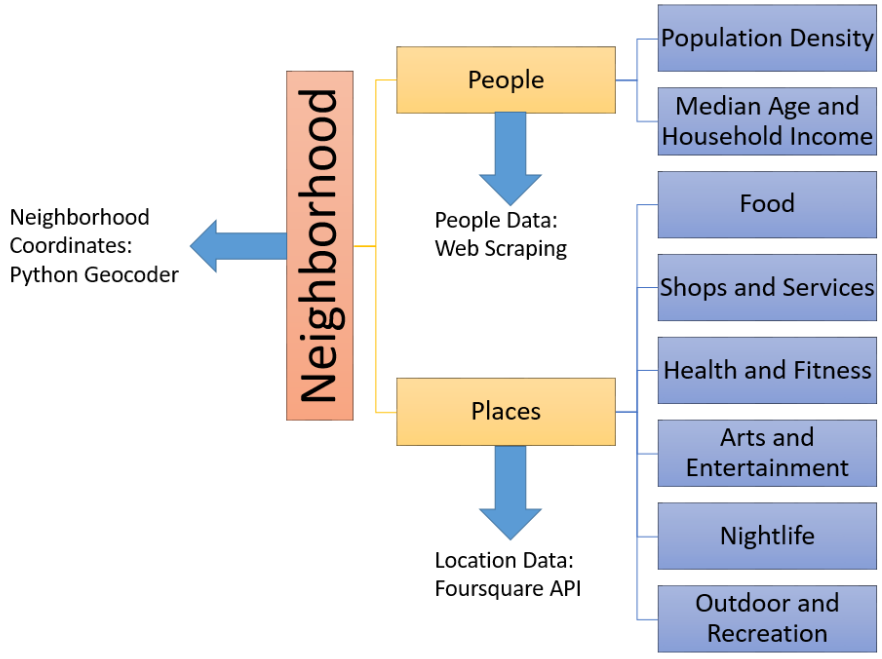
Given the current neighborhood of a person looking to relocate, can we find the most “similar” neighborhoods in the new city? This project aims to employ location intelligence and clustering to answer this question. After all, moving to a place with all the same restaurants, amenities, shops, parks…etc. would have a similar “vibe” that we are used to in our home city. The more familiarity we can manufacture, the less daunting the move becomes…

As a case study for the project, we will consider Alex, who is a recent engineering graduate from the University of Alberta living in the Strathcona neighborhood in Edmonton Alberta. Alex has accepted a job offer from a firm based out of Calgary, Alberta and is looking to relocate to a similar neighborhood in Calgary as Strathcona.

3.0 Dataset and Sources

The core idea is to represent neighborhoods in terms of their defining features in a high dimensional feature space and then cluster datapoints that are close together. The features that capture the essence of a neighborhood is an intricate and subtle problem in itself. For this exercise, we will broadly classify the features into those that represent the "people" in the neighborhood and those that represent the "places" in the neighborhood.

Here is a flowchart showing the breakdown of the feature set and their source:



It must be noted that the choice of features to represent neighborhoods is subjective and must be tailored to the end user (Alex in our case). The above features are by no means *the* unique combination to represent a neighborhood but merely provide a good starting point to illustrate the methodology used in this project.

Datasets for features under the "People" category can be obtained through a combination of Wikipedia and census data available online. Data for the "Places" category can be retrieved from the Foursquare database through their Places API. Finally, the latitude and longitude coordinates of the candidate neighborhoods can be obtained using the Python Geopy library.

3.1 Features in “People Category”

The features under the "people" category are Population Density and the Median Age and Household Income.

The Population Density by neighborhoods can be web scraped off the Wikipedia page: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary

The next data we need is the median income and age for each neighborhood in Calgary. Conveniently, this website publishes both these data. We will follow the same procedure as before to scrape and store in a data frame.

Here is the final data frame showing the population density, median age and income for each neighborhood in Calgary:

:

	Neighborhood	Population Density	Median Age	Median Income
0	Abbeydale	3,480.6	34	$55,345
1	Acadia	2,744.9	42	$46,089
2	Altadore	3,143.4	37	$53,786
3	Applewood Park	4,061.3	33	$65,724
4	Arbour Lake	2,462.7	41	$70,590
...	...	...	...	...
170	Willow Park	1,537.9	45	$63,588
171	Windsor Park	3,173.8	37	$39,425
172	Winston Heights/Mountview	1,297	42	$41,065
173	Woodbine	2,853.4	42	$83,844
174	Woodlands	2,214.6	40	$71,234

175 rows × 4 columns

3.2 Latitude and Longitude Coordinates

As mentioned before, the geopy library in Python can be used to obtain the location coordinates of each neighborhood. We will pass each neighborhood in a loop to the geolocator object and store the results back in the input data frame.

Here is the data frame with the location coordinates added:

	Neighborhood	Latitude	Longitude	Median Age	Median Income	Population Density
0	Abbeydale	51.058836	-113.929413	34	$55,345	3,480.6
1	Acadia	50.968655	-114.055587	42	$46,089	2,744.9
2	Altadore	51.015104	-114.100756	37	$53,786	3,143.4
3	Applewood Park	51.044658	-113.928931	33	$65,724	4,061.3
4	Arbour Lake	51.136786	-114.202355	41	$70,590	2,462.7

3.3 Features in “Places” Category

All the features under the "Places" category can be obtained using the **Foursquare Places API**. We will use the *venues/explore* endpoint to obtain a json file with all the venues under each category in the vicinity of each neighborhood.

The json file can then be processed to extract and add all the venues as columns to our input data frame. Here is the result of this exercise:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbeydale	51.058836	-113.929413	Subway	51.059239	-113.934423	Sandwich Place
1	Abbeydale	51.058836	-113.929413	Mac's	51.059376	-113.934425	Convenience Store
2	Abbeydale	51.058836	-113.929413	roadside pub	51.059277	-113.934529	Wings Joint
3	Acadia	50.968655	-114.055587	Bow Valley Insurance	50.967936	-114.051084	Insurance Office
4	Acadia	50.968655	-114.055587	Highwest Electric Ltd	50.965847	-114.057257	Construction & Landscaping
...	...	...	...	...	...	...	...
1047	Winston Heights/Mountview	51.072303	-114.047588	Mount View School Age and Family Care Center	51.069705	-114.051977	College Classroom
1048	Woodlands	50.942435	-114.109359	3 Crowns	50.940765	-114.109430	Pub
1049	Woodlands	50.942435	-114.109359	Russian Store	50.941063	-114.109452	Food & Drink Shop
1050	Woodlands	50.942435	-114.109359	Woodpark Liquor	50.941202	-114.109502	Liquor Store
1051	Woodlands	50.942435	-114.109359	Mac's	50.940876	-114.109816	Food & Drink Shop

1052 rows × 7 columns

```
...
There are 196 uniques categories.
```

As we can see, the data frame contains all the venues and venue categories for each neighborhood. There are a total of 196 unique categories returned from the Foursquare API.

We will now encode all the categories for each neighborhood using one-hot encoding. The result is a new data frame containing all the categories as separate columns with either a "1" indicating that that category is in proximity to the neighborhood or "0" otherwise as shown below.

	Neighborhood	Accessories Store	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bank	...	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio
0	Abbeydale	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	Abbeydale	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	Abbeydale	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
3	Acadia	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	Acadia	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 196 columns

Finally, the neighborhoods can be grouped by the mean of each venue category. This will give a feature matrix containing the relative occurrence of each venue for a given neighborhood. For example, in the snippet below, the category "Wings Joint" is weighted 33% amongst the venues that are in proximity to Abbeydale.

	Neighborhood	Accessories Store	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bank	...	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio
0	Abbeydale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.333333	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	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	Altadore	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	0.0	0.0	0.0	0.000000	0.0	0.0
3	Applewood Park	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Arbour Lake	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	0.0	0.0	0.0	0.000000	0.0	0.0

5 rows × 196 columns

3.4 Compile Final Dataset

We can compile the final data set by merging the above data frame with our data frame from section 3.2. Also, since we are intending to employ a clustering algorithm to determine "similar" neighborhoods to Alex's current neighborhood, we need to add "Strathcona" as an extra data point to the above dataset.

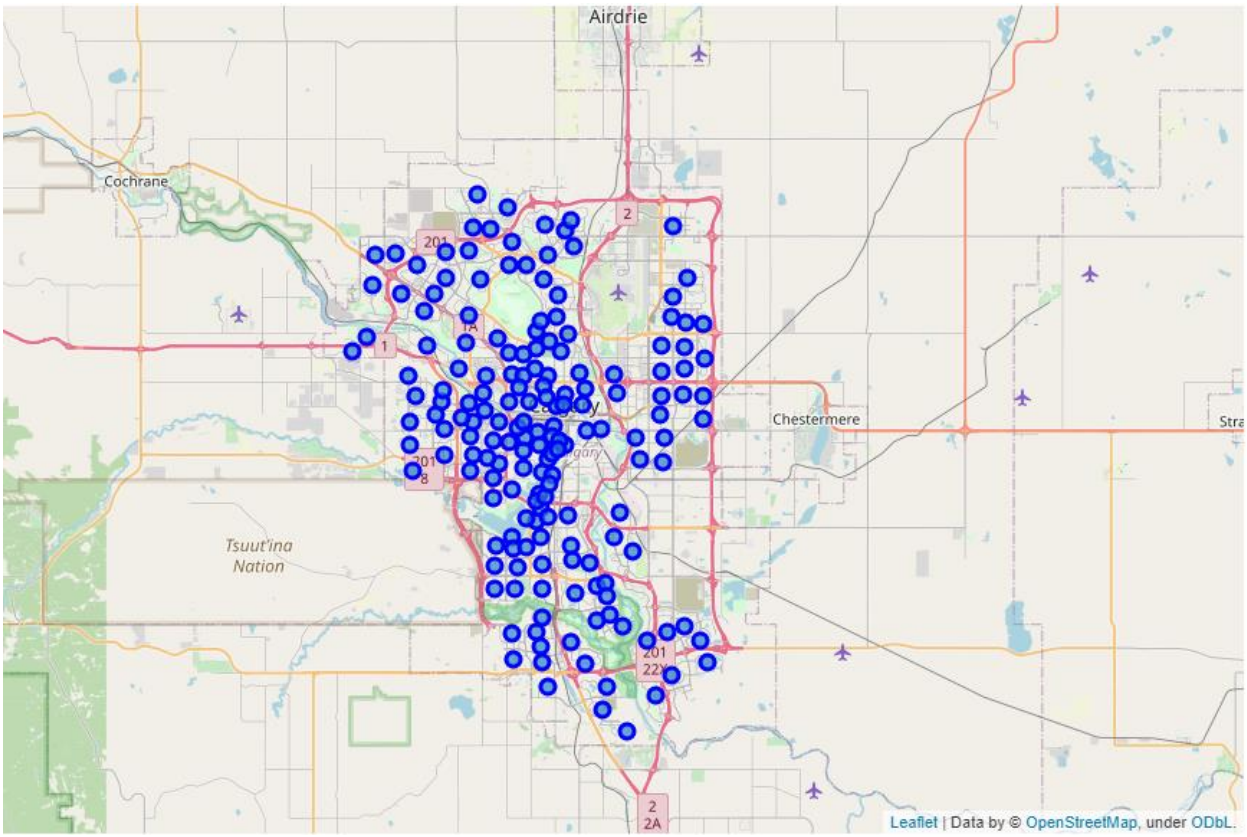
The resulting data frame is shown below:

	Neighborhood	Latitude	Longitude	Median Age	Median Income	Population Density	Accessories Store	American Restaurant	Art Gallery	Arts & Crafts Store	...	Vietnamese Restaurant	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	Farmers Market	General Entertainment	Jazz Club	Public Art
0	Abbeydale	51.058836	-113.929413	34.0	55345.0	3480.6	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.333333	0.0	0.0	0.000000	0.000000	0.000000	0.000000
1	Acadia	50.968655	-114.055587	42.0	46089.0	2744.9	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
2	Altadore	51.015104	-114.100756	37.0	53786.0	3143.4	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
3	Applewood Park	51.044658	-113.928931	33.0	65724.0	4061.3	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
4	Arbour Lake	51.136786	-114.202355	41.0	70590.0	2462.7	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
155	Willow Park	50.960293	-114.054645	45.0	63588.0	1537.9	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
156	Windsor Park	51.006165	-114.076187	37.0	39425.0	3173.8	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
157	Winston Heights/Mountview	51.072303	-114.047588	42.0	41065.0	1297.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
158	Woodlands	50.942435	-114.109359	40.0	71234.0	2214.6	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
159	Strathcona	53.522000	-113.492000	35.0	68403.0	5722.3	0.0	0.0	0.0	0.0	...	0.041667	0.0	0.0	0.000000	0.0	0.0	0.041667	0.041667	0.041667	0.041667

Finally, we see that the Median Age, Median Income and Population Density features are disproportionately greater than the other features. This might induce distortions in the K-means clusters as variables with large variance will tend to be more separated than variables with small variance. To overcome this, we will normalize the data in all the columns and use the resulting matrix for the clustering algorithm.

3.5 Exploratory Data Analysis

We have obtained all the candidate neighborhoods and their features in our final data set. However, it would be prudent to visualize the neighborhoods on a map to ensure that we have neighborhoods uniformly distributed throughout the city. If the neighborhoods are all crowded in a small region, the resulting analysis and results may not be useful.



As we can see, we have a reasonably good spread of neighborhoods throughout the city with a few odd sparse and dense locations. Depending on the use case, it might be necessary to revisit the list of

neighborhoods and perhaps add some more to the list. However, for our purposes, the above samples are sufficient.

Another useful check would be to visualize the venue categories for each neighborhood in a more intuitive structure than the relative occurrence matrix we saw earlier. We will sort the venue categories for each neighborhood by the 1st, 2nd, 3rd, etc. most common venues in proximity to that neighborhood.

	Neighborhood	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	Yoga Studio	Food Court
1	Acadia	Insurance Office	Construction & Landscaping	Yoga Studio	Food Service	Golf Course
2	Altadore	Dog Run	Coffee Shop	Massage Studio	Greek Restaurant	Gourmet Shop
3	Applewood Park	Park	Liquor Store	Home Service	Yoga Studio	Food Service
4	Arbour Lake	Bus Station	Moving Target	Grocery Store	Lake	Residential Building (Apartment / Condo)

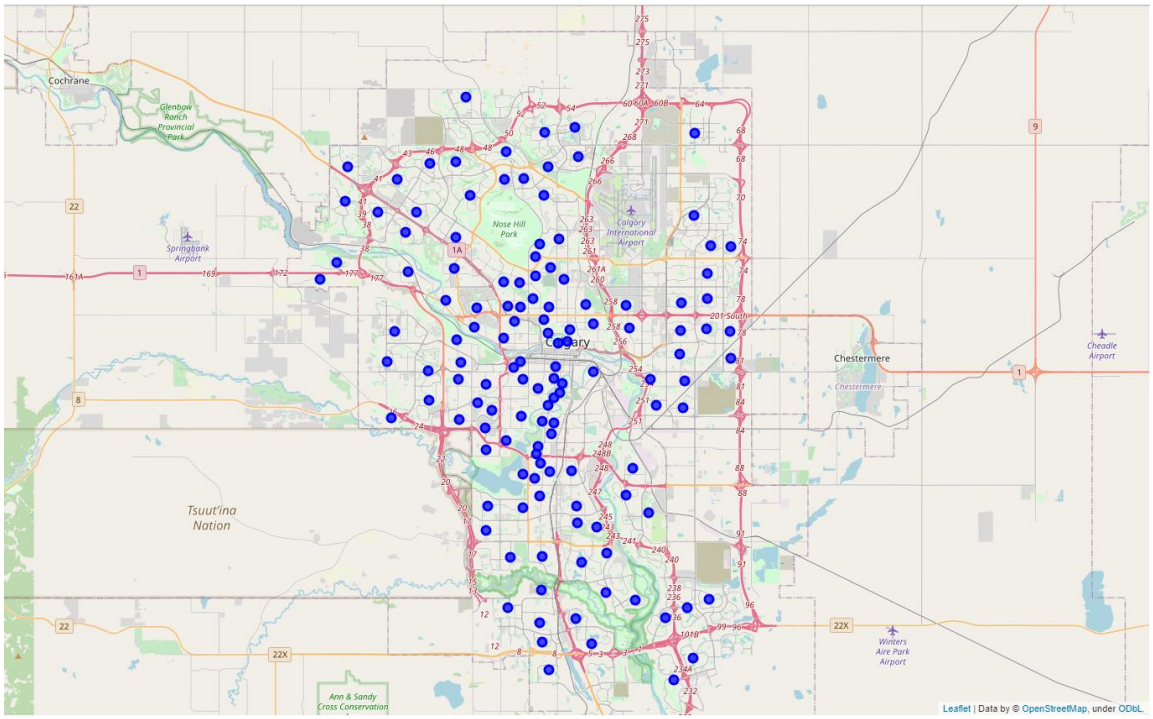
We see that the neighborhood Abbeydale has more Wing Joints than any other venue category. This makes sense from our earlier observation that the Wings Joint category was weighted pretty heavily for Abbeydale in the relative occurrence matrix.

4.0 Methodology

4.1 Naïve Clustering

As explained in the earlier sections, the general strategy is to apply the K-means clustering algorithm to segment and label neighborhoods in Calgary that are similar to Strathcona. We will first naively apply the clustering algorithm to our normalized dataset from section 3.4. When visualizing the clusters, we will note some interesting patterns that will suggest that the clusters are not quite what we intended.

Training and fitting the model using 10 clusters yields the following subset of neighborhoods with the same label as Strathcona:



Clearly, the K-means clustering seems to indicate that most neighborhoods in Calgary are pretty similar to the Strathcona neighborhood in Edmonton. However, intuitively, we know that this cannot be true. As a concrete example, Signal Hill located near the western boundary of the city has a reputation for being a quiet residential neighborhood. On the other hand, Strathcona is a neighborhood known for its busy pubs, restaurants and night clubs. And yet, they are both in the same cluster. How do we explain this discrepancy?

3.2 Curse of the Dimensions

As it turns out, clustering in higher dimensions is a bit tricky. Our traditional "distance" metric to measure similarity between two sample points tends to zero as the number of dimensions increase. That is, any two data points will have a distance value close to zero if the feature space dimensionality is very high. This can be verified to some extent by playing with the epsilon parameter in the DBSCAN algorithm and visualizing the effect (or lack thereof) on the labels.

The most meaningful fix in our case would be to reduce the feature space dimensionality. Let us arbitrarily pick the following features for the reduced input space and drop all the other venue categories:

1. Bar

2. Train Station

3. Cafe

3. Grocery Store

4. Indian Restaurant

In other words, our hypothetical end user (Alex) has decided that the above venues are his most important/desirable characteristics in a neighborhood. Here is the reduced data frame:

	Neighborhood	Latitude	Longitude	Median Age	Median Income	Population Density	Bar	Train Station	Café	Grocery Store	Indian Restaurant
0	Abbeydale	51.058836	-113.929413	34.0	55345.0	3480.6	0.000000	0.000000	0.000	0.000000	0.000000
1	Acadia	50.968655	-114.055587	42.0	46089.0	2744.9	0.000000	0.000000	0.000	0.000000	0.000000
2	Altadore	51.015104	-114.100756	37.0	53786.0	3143.4	0.000000	0.000000	0.000	0.000000	0.000000
3	Applewood Park	51.044658	-113.928931	33.0	65724.0	4061.3	0.000000	0.000000	0.000	0.000000	0.000000
4	Arbour Lake	51.136786	-114.202355	41.0	70590.0	2462.7	0.000000	0.000000	0.000	0.166667	0.000000
...	...	...	...	...	...	...	...	...	...	...	...
155	Willow Park	50.960293	-114.054645	45.0	63588.0	1537.9	0.000000	0.000000	0.000	0.200000	0.000000
156	Windsor Park	51.006165	-114.076187	37.0	39425.0	3173.8	0.000000	0.000000	0.000	0.000000	0.000000
157	Winston Heights/Mountview	51.072303	-114.047588	42.0	41065.0	1297.0	0.000000	0.000000	0.000	0.000000	0.000000
158	Woodlands	50.942435	-114.109359	40.0	71234.0	2214.6	0.000000	0.000000	0.000	0.000000	0.000000
159	Strathcona	53.522000	-113.492000	35.0	68403.0	5722.3	0.041667	0.041667	0.125	0.083333	0.041667

160 rows × 11 columns

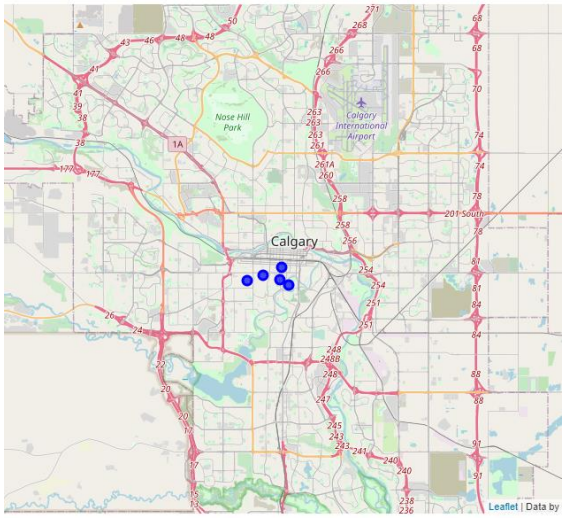
For same reasons as explained earlier, we will normalize all the columns before running the clustering algorithm.

4.0 Results

Running the K-means clustering algorithm on this lower dimensional feature space labels the following neighborhoods as similar to Strathcona:

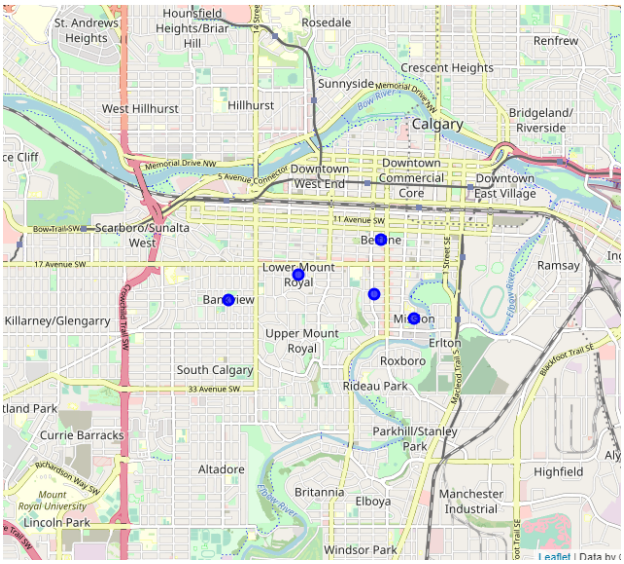
	Neighborhood	Latitude	Longitude	Median Age	Median Income	Population Density	Bar	Train Station	Café	Grocery Store	Indian Restaurant	k-Labels
8	Bankview	51.033887	-114.099518	32.0	32474.0	7458.6	0.000000	0.000000	0.000000	0.142857	0.000000	3
12	Beltline	51.040498	-114.072593	33.0	33901.0	6786.6	0.053571	0.000000	0.017857	0.000000	0.000000	3
29	Cliff Bungalow	51.034436	-114.073833	32.0	35576.0	4840.0	0.030303	0.000000	0.030303	0.030303	0.030303	3
78	Lower Mount Royal	51.036645	-114.087139	34.0	35570.0	10600.0	0.000000	0.000000	0.000000	0.020000	0.000000	3
91	Mission	51.031758	-114.066720	34.0	37040.0	8650.0	0.000000	0.000000	0.027778	0.027778	0.027778	3

Here is a map showing all these neighborhoods:



5.0 Discussion

Let us take a closer look at the five neighborhoods:



The neighborhoods classified to be similar to Strathcona are all in lively locations just outside the core Calgary downtown area. This is a promising start as Strathcona is known for its busy bars and night life. These neighborhoods all have lots of bars, cafes, grocery stores and Indian restaurants within walking distance. Moreover, since many companies have offices downtown, young professionals tend to stay in these neighborhoods, as evidenced by the median age. Finally, as someone that has stayed in both Strathcona and Lower Mount Royal neighborhoods, I can confirm that the ambience and "vibe" are indeed pretty similar. Therefore, we have reasonable validation for the model and can recommend the candidate neighborhoods to Alex.

The next step is to further assess these neighborhoods based on other preferences and come up with a ranking list. For example, Alex may prefer to be as close to work as possible to minimize travel time in the winter. Or his priority may be to minimize rental costs since a significant chunk of his salary must be allocated for student loan repayment. If there are multiple competing considerations, a weighted average metric could be employed to come up with a ranking.

6.0 Conclusion

The K-means clustering algorithm was successfully employed on the reduced feature data set to obtain a reasonable cluster of neighborhoods that are similar to Strathcona. Obviously, the methodology can be extended to any combination of cities and with more intricate feature selection, perhaps even across countries.

To avoid the dimensionality issue, it is recommended that the training set be restricted to 10 or less feature variables. This would involve some careful consideration in choosing the best subset of features that could have the most meaningful impact for the use case. As with most data science projects, some iteration might be required to arrive at a reasonable combination of features and clusters.

7.0 Dataset Sources and Links

Strathcona Edmonton - Population Density: https://en.wikipedia.org/wiki/Strathcona,_Edmonton

List of Neighborhoods in Calgary and Population Density:
https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary

Median Age and Household Income for Calgary Neighborhoods: https://great-news.ca/demographics/

Strathcona Edmonton - Median Age and Income:
https://public.tableau.com/profile/city.of.edmonton#!/vizhome/2019EdmontonMunicipalCensus/2019EdmontonMunicipalCensusNeighbourhood

Location Information - Foursquare API: https://developer.foursquare.com/docs/places-api/

Link to Jupyter Notebook on Github Repository:
https://nbviewer.jupyter.org/github/AakashVasudevan/Coursera_Capstone/blob/main/Battle%20of%20Neighborhoods.ipynb