Segmenting and Clustering Calgary Neighbourhoods Based on Current Real Estate and Location Information

By Moriah Rempel

Introduction

Finding the perfect home can be a challenge as there are many different factors to consider when purchasing a home. This project aims to help buyers find their ideal neighbourhoods based on real estate data such as type of house and price as well as using neighbourhood location data to evaluate what kind of venues are close by.

The area of focus for this project will be in Calgary, AB although the methodology will be applicable to any city.

Data science will be implemented to segment and cluster neighbourhoods using information obtained from the real estate website realtor.ca and location information retrieved using four square. The K-means clustering technique will be used because the data can be separated into defined categories that can be represented numerically.

Buyers will be able to narrow down their neighbourhood selection to a cluster by looking at the common traits found within each cluster and selecting the cluster with the traits most desirable to them. This will save the buyer time and effort in their home buying search process as they can initially narrow down their neighbourhoods of interest.

Data

In order to investigate this problem the following data was needed:

1. Real estate data (coordinates, neighbourhood, building type, price, bedrooms, bathrooms, and size)
2. Calgary Neighbourhood list and coordinates
3. Neighbourhood venue information

For each point data was retrieved, cleaned/pre-processed and stored as a dataframe.

Acquisition and Pre-processing

1. Real Estate Dataframe

Real estate data was retrieved using the realtor.ca API. Information was returned for residential listings within Calgary. The search area was narrowed down by defining maximum and minimum latitude and longitude of Calgary boundaries and only returning listings defined as being in Calgary.

The API only allows 200 listings to be returned at a time so the code had to include a work around so that none of the listings would be missed. This work around first defined price intervals in a list at which the listings would be searched. If 200 results were returned this would indicate that there are potentially more listings within this price interval. The price interval would be divided in half and added back into the price interval list. This process would be repeated until all the results being returned had less than 200 results. To prevent double entries for listings, the information was stored in a dictionary. Since this section of code could take some time to run the dictionary was saved into a CSV so that it could be called upon later without needing to run this section of code again.

The information that was retrieved over the API was the listing address, coordinates, property type, building type, price, number of bathrooms, number of bedrooms, and interior size.

Neighbourhoods were listed within the address string and had to be extracted by searching for common surrounding characters.

The first 5 entries for resulting real estate dataframe can be seen in Table 1.

	Neighbourhood	lng	lat	property_type	building_type	price	bathrooms	bedrooms	InteriorSize
0	Red Carpet	-113.941113	51.040125	Single Family	Mobile Home	$19,000	1	2 + 0	715 sqft
1	Red Carpet	-113.937916	51.038809	Single Family	Mobile Home	$29,000	1	2 + 0	960 sqft
2	Greenwood/Greenbriar	-114.214703	51.089611	Single Family	Mobile Home	$32,500	1	3 + 0	1249 sqft
3	Bowness	-114.216046	51.089684	Single Family	Mobile Home	$38,500	1	2 + 0	983 sqft
4	Arbour Lake	-114.217323	51.127602	Single Family	Mobile Home	$39,900	2	2 + 0	892 sqft

Table 1. First 5 entries of the real estate dataframe created from data retrieved off realtor.ca.

2. Neighbourhood Coordinate Dataframe

A list of neighbourhoods was first scraped from Wikipedia. Anything that was not classified as residential was removed. This list was updated last on Wikipedia in 2012 so there were some changes that had to be made including replacing neighbourhood names with more commonly recognized ones and adding new community names to the list.

Google geocoder is no longer available to be used freely, it now requires payment for information to be retrieved. Using geolocator is unreliable in that it doesn't always return results and that it does not recognize the names of some neighbourhoods when searching for them. Neighbourhood locations could not be found using either of these methods so they were found by looping through the real estate data and averaging the coordinates of the listings within each neighbourhood. Neighhourhoods were removed removed from the neighbourhood list if they did not currently have any listings.

Once the neighbourhood dataframe was complete (see Table 2), this was used to clean up the data by cross referencing the neighbourhood dataframe with the real estate dataframe. If the neighbourhood in a listing in the real estate dataframe did not match any of neighbourhoods in the neighbourhood dataframe then the listing was removed.

	Neighbourhood	Latitude	Longitude
0	Abbeydale	51.059	-113.926
1	Acadia	50.9735	-114.061
2	Albert Park/Radisson Heights	51.0404	-113.993
3	Altadore	51.0168	-114.104
4	Applewood Park	51.0422	-113.93

Table 2. First 5 entries of the Neighbourhood dataframe created from a Calgary neighbourhood list found on Wikipedia and coordinates averaged from listings in the real estate table.

3. Neighbourhood Venue Dataframe

The venue dataframe (Table 3) was acquired using the Four Square API. Venues within 750 m (a maximum of 100 were returned for each neighbourhood) were found for each neighbourhood based on the coordinates in the neighbourhood dataframe.

The venue information retrieved using the API was the neighbourhood, neighbourhood coordinates, venue name, venue coordinates, and venue category.

This dataframe was saved to the CSV as Four Square has a maximum amount of data retrieval per day.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbeydale	51.059006	-113.925914	Subway	51.059239	-113.934423	Sandwich Place
1	Abbeydale	51.059006	-113.925914	Vanity Fitness	51.062982	-113.926238	Health & Beauty Service
2	Abbeydale	51.059006	-113.925914	Mac's	51.059376	-113.934425	Convenience Store
3	Abbeydale	51.059006	-113.925914	roadside pub	51.059277	-113.934529	Wings Joint
4	Abbeydale	51.059006	-113.925914	Magic Touch Stone Ltd	51.065379	-113.927304	Construction & Landscaping

Table 3. First 5 entries for the venue datframe which lists venues within 750 m of each neighbourhood in the neighbourhood dataframe.

Methodology

Once the data was collected and stored in dataframes these were used to perform K-means clustering to create clusters of neighbourhoods with similar real estate attributes and venue locations.

K-means Clustering

K-means clustering is performed by defining a set number of clusters and then randomly distributing the set number of cluster centroids throughout the data. Data is assigned to the nearest cluster centroid, the average of the data within the cluster is found and the centroid is

moved to the average. The process of assigning data to the nearest cluster and then averaging to find the new centroid location is iterated until there is no change in the centroid location.

Neighbourhood Analysis

In order to visualize the neighbourhoods and the coordinates found for the neighbourhoods the information was input into folium to create a map of Calgary and the neighbourhoods in the neighbourhood dataframe (Figure 1). This represents all the neighbourhoods in Calgary that currently have listings.

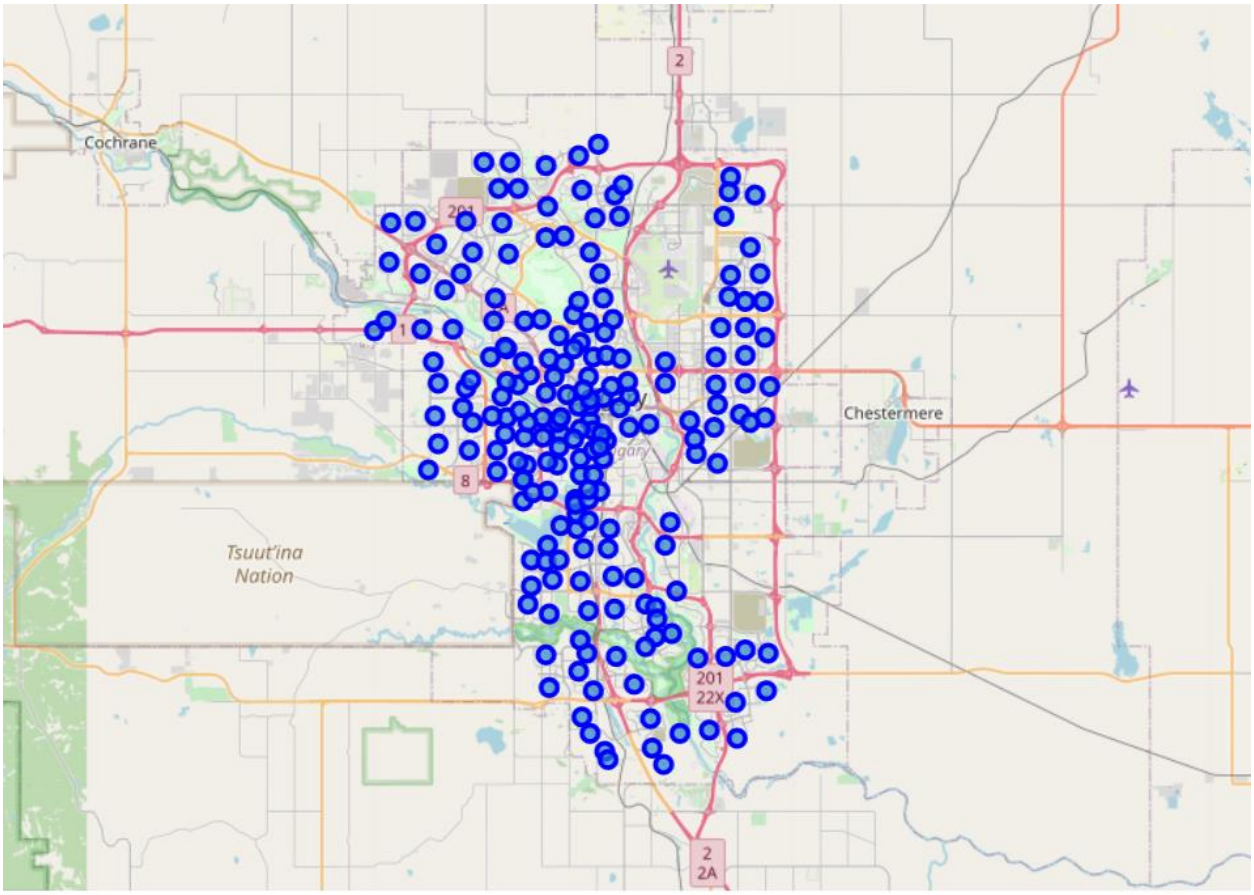

Figure 1. Map of Calgary. Each blue dot represents a neighbourhood with residential listings. Blue dot coordinates were calculated by averaging listing coordinates within the respective neighbourhood.

Neighbourhood Venue Analysis

The data that was used in K-means clustering for the neighbourhood venue data was the top 5 venue categories and the number of venues for each neighbourhood.

To find the top 5 categories, a dummy dataframe was created for the venue category where each venue category was turned in to a column and a value of 0 representing that the venue is not present or a value of 1 representing the venue is present are placed as values. This was then grouped by neighbourhood so that there was only one row per neighbourhood. Grouping was

performed by averaging the dummy entries of each venue within the neighbourhood. The top 5 venues for each neighbourhood were found by sorting the values for each venue category and selecting the category columns with the 5 highest values.

For each neighbourhood the number of venues were counted using the groupby and count functions. Then the venue count for each neighbourhood was sorted into more generalized groups labeled as >50 venues, 25-50 venues, and >100 venues. This was then merged with the top 5 categories for each neighbourhoods to produce the dataframe seen in Table 4.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Most Common Venue Count
0	Abbeydale	Wings Joint	Health & Beauty Service	Sandwich Place	Convenience Store	Construction & Landscaping	<25
1	Acadia	Sandwich Place	Pub	Fast Food Restaurant	Coffee Shop	Yoga Studio	<25
2	Albert Park/Radisson Heights	Grocery Store	Indian Restaurant	Restaurant	Fried Chicken Joint	Fast Food Restaurant	<25
3	Altadore	Liquor Store	Coffee Shop	Pub	Brewery	Greek Restaurant	<25
4	Applewood Park	Home Service	Coffee Shop	Park	Liquor Store	Food & Drink Shop	<25

Table 4. First 5 entries of the top 5 venues and the venue count for each neighbourhood.

K-means clustering was then implemented using the top 5 venue categories and the venue count for each neighbourhood to see what the results would look like when only clustering using the venue data. The neighbourhoods were divided into 5 clusters. Figure 2 was created using folium to plot each neighbourhood as a colour corresponding to a cluster.

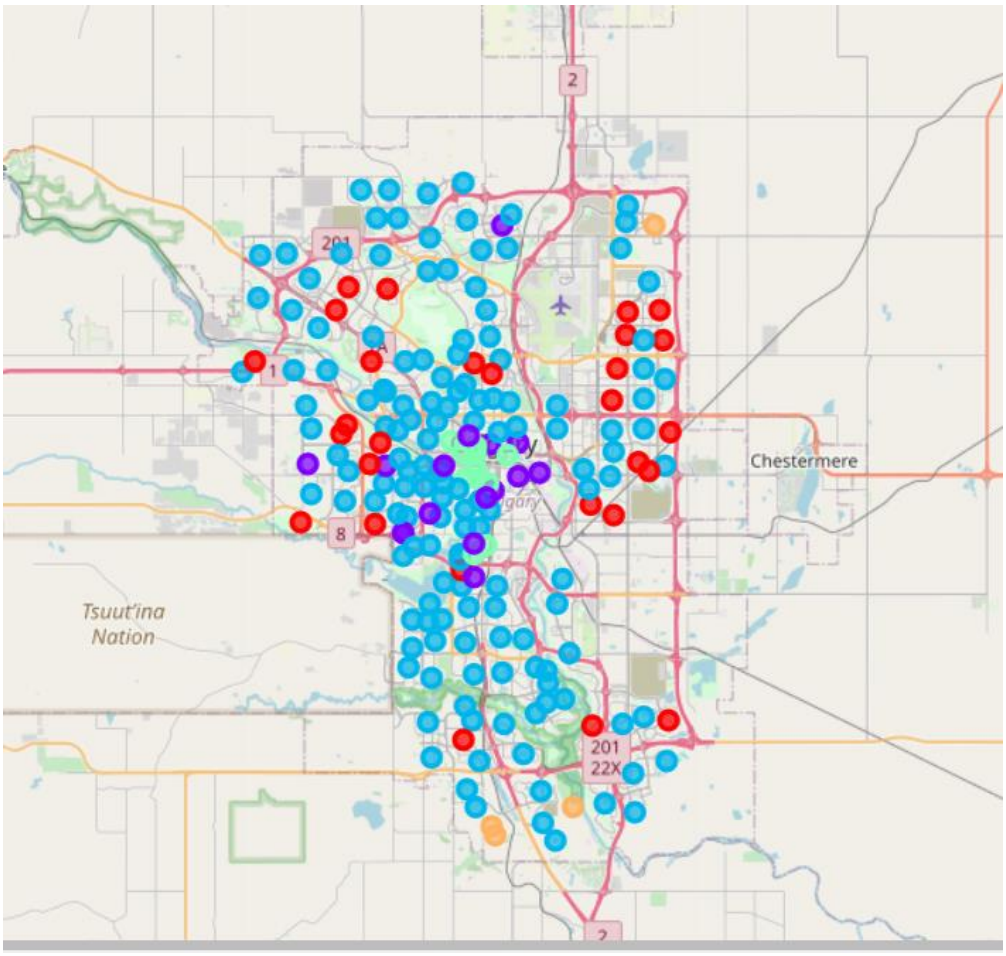

Figure 2. Calgary neighbourhoods clustered based on top 5 venues and venue count. Cluster 0: red, Cluster 1: purple, Cluster 2: blue, Cluster 3: green, Cluster 4: orange.

Real Estate Attribute Analysis

The data that was used for k-means clustering performed on the real estate data were from the dataframe columns: building type, price, number of bathrooms, number of bedrooms, and interior size. For each of these attributes the top 2 were found.

Since there was a wide range prices, these were broken down into the following price bracket for use in clustering: <$250,000, $250,000-$500,000, $500,000, $750,000-$1,000,000, $1,000,0000-$2,000,000, $2,000,0000-$3,000,000, and $3,000,000+. The interior size was treated similarly and broken into brackets of: <500 sqft, 500-1000 sqft, 1000-2000 sqft, and 2000+ sqft.

A loop ran each real estate attribute so that it created a dummy dataframe expressing values as 0 or 1 depending on their presence, sorted the dummy dataframes by neighbourhood and averaged the dummy values, and then found the top 2 values for each attribute. These top 2 values were then all merged into a single table to produce Table 5.

	Neighbourhood	1st Most Common building_type	2nd Most Common building_type	1st Most Common price	2nd Most Common price	1st Most Common bedrooms	2nd Most Common bedrooms	1st Most Common bathrooms	2nd Most Common bathrooms	1st Most Common InteriorSize	2nd Most Common InteriorSize
0	Abbeydale	House	Mobile Home	250,000–500,000	<=$250,000	3 bedrooms	4 bedrooms	2 bathrooms	3 bathrooms	1000-2000 sqft	500-1000 sqft
1	Acadia	House	Apartment	250,000–500,000	<=$250,000	4 bedrooms	3 bedrooms	2 bathrooms	3 bathrooms	1000-2000 sqft	500-1000 sqft
2	Albert Park/Radisson Heights	Apartment	Duplex	<=$250,000	250,000–500,000	2 bedrooms	3 bedrooms	2 bathrooms	4 bathrooms	500-1000 sqft	1000-2000 sqft
3	Altadore	House	Row / Townhouse	750,000–1,000,000	500,000–750,000	4 bedrooms	3 bedrooms	4 bathrooms	3 bathrooms	1000-2000 sqft	2000+ sqft
4	Applewood Park	House	Apartment	250,000–500,000	<=$250,000	3 bedrooms	4 bedrooms	2 bathrooms	3 bathrooms	1000-2000 sqft	500-1000 sqft

Table 5. First 5 entries of the top 2 values for the attributes building type, price, number of bedrooms, number of bathrooms, and interior size.

This dataframe was then used to perform K-means clustering to create 5 neighbourhood clusters. Figure 3 shows a map of neighbourhoods coloured to correspond with the assigned cluster.

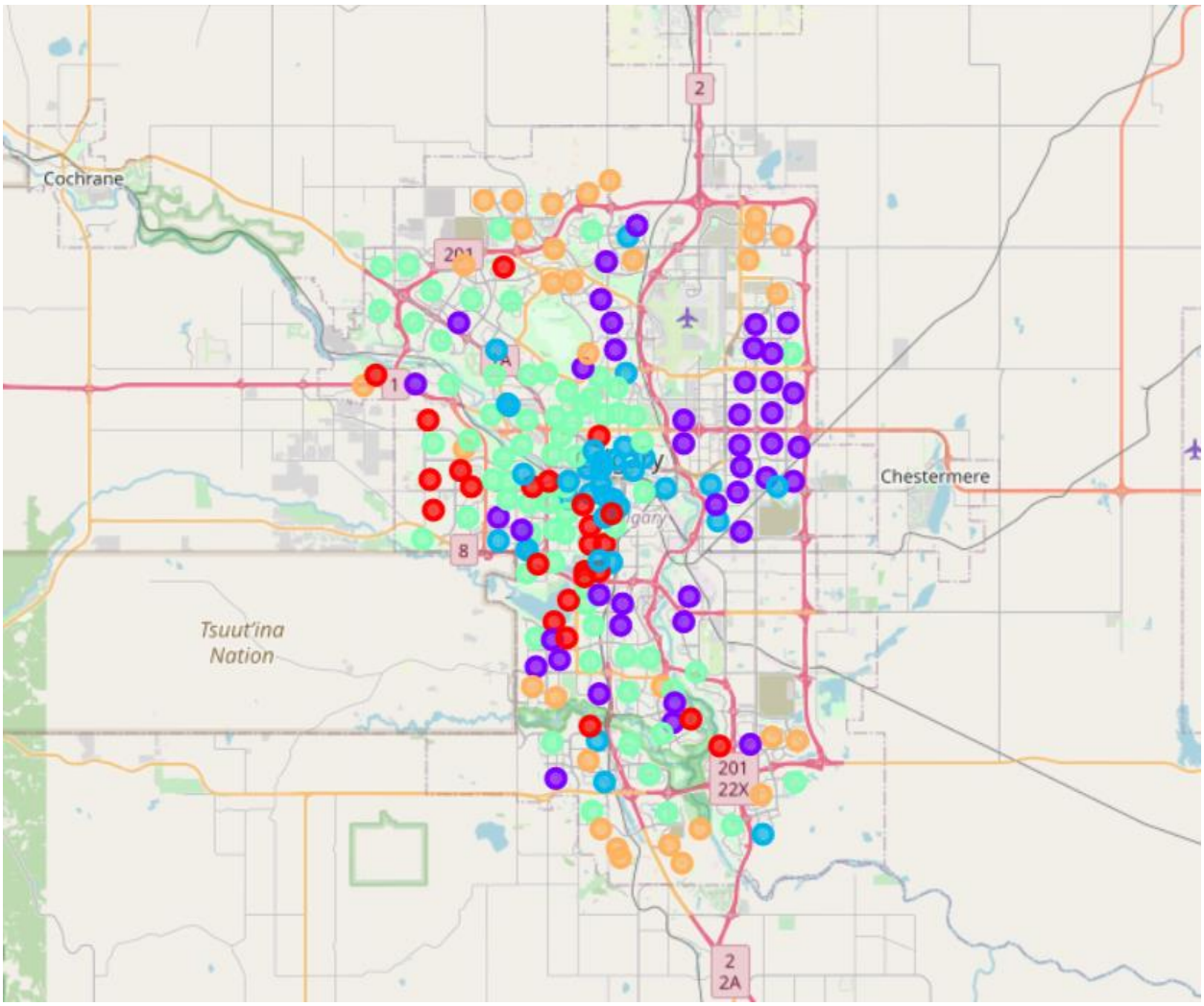

Figure 3. Calgary neighbourhoods clustered based on real estate attributes. Cluster 0=red, Cluster 1: purple, Cluster 2: blue, Cluster 3: green, Cluster 4: orange.

Combined Venue and Real Estate Analysis

The goal of this project was to cluster based on both the venue and real estate data. In order to do this the venue dataframe and the real estate dataframe were merged into a master dataframe. K-means clustering was then run on this master dataframe to create 5 clusters based the top 5 venues, the venue count, and the real estate attributes to create Figure 4.

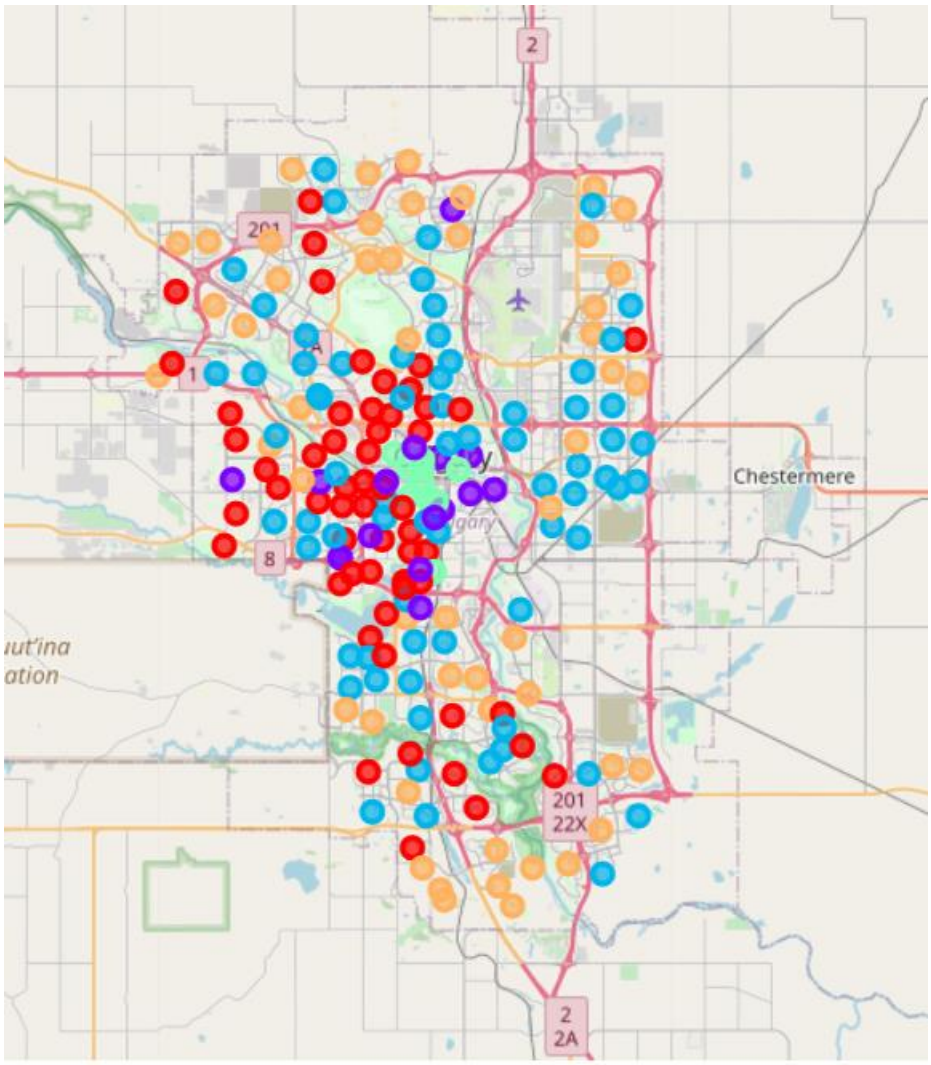

Figure 4. Calgary neighbourhoods clustered based on top 5 venues, venue count, and real estate attributes.

Results

There will not be results shared in this report for the neighbourhood clustering that resulted from the venue data or the real estate date alone as the goal of the project was to find the clusters resulting from the combination of this data.

The full dataframe tables for each resulting cluster can be found on https://github.com/moriahrempel/Data-Science-Capstone/blob/master/Data%20Science%20Capstone%20Project.ipynb .

To summarize the dataframe tables, the final resulting clusters produced from the combine data are:

Cluster 0

Neighbourhoods: Altadore, Banff Trail, Bayview, Bel-Aire, Britannia, Cambrian Heights, Capitol Hill, Currie Barracks, Charleswood, Christie Park, Collingwood, Coral Springs, Cougar Ridge, Deer, Run, Diamond Cove, Discovery Ridge, Eagle Ridge, Edgemont, Elbow Park, Elboya, Evergreen, Glendale, Hamptons, Highwood, Hounsfield Heights/Briar Hill, Killarney/Glengarry, Lake Bonavista, Lakeview, Mayfair, McKenzie Lake, Meadowlark Park, Midnapore, Mount Pleasant, North Glenmore Park, Parkdale, Pump Hill, Richmond, Rosedale, Scarboro, Sunalta West, Shagnappi, Shawnee Slopes, Sherwood, Silverado, Springbank Hill, Strathcona Park, Sundance, Tuscany, University Heights, Upper Mount Royal, Valley Ridge, West Hillhurst, West Springs, Wildwood, Winston Heights/Mountain View, Garrison Green

Summary: This cluster contains larger and more expensive houses or townhouses. Common prices are over 500,000 dollars and include neighbourhoods with average house prices over one million. The homes in this cluster are usually are over 1000 sqft with many over options 2000 sqft and have at least 4 bedrooms and at least 4 bathrooms. Looking at the map most of these neighbourhoods closer to the inner city. The neighbourhoods in this cluster will have lower accessibility to venues as the venue count is commonly <25.

Cluster 1

Neighbourhoods: Aspen Woods, Bridgeland/Riverside, Lincoln Park, Chinatown, Country Hills Village, Erlton, Inglewood, Kingsland Ramsay, Rosscarrock, Roxboro, Sunalta, Sunnyside, Windsor Park, Garrison Woods

Summary: This cluster contains mostly apartments and townhouses. Prices are commonly under 500,000 but there are a range of price options within this cluster. They are commonly under 2000 sqft with many options under 1000 sqft with less than 4 bedrooms and bathrooms. These neighbourhoods have a venue count of 25-50 meaning these neighbourhoods have good venue accessibility and by looking at the map are just outside the city core.

Cluster 2

Neighbourhoods: Abbeydale, Acadia, Albert Park/Radisson Heights, Applewood Park, Arbour Lake, Bankview, Beddington Heights, Bowness, Braeside, Brentwood, Bridlewood, Canyon Meadows, Cedarbrae, Country Hills, Crescent Heights, Dalhousie, Deer Ridge, Dover, Erin Woods, Falconridge, Forest Heights, Forest Lawn. Glamorgan, Glenbrook, Greenview, Greenwood/Greenbriar, Haysboro, Highland Park, Huntington Hills, Kelvin Grove, Kincora, Mahogany, Marlborough Park, Mayland Heights, McKenzie Towne, Millrise, North Haven, Oakridge, Ogden, Palliser, Parkhill, Parkland, Patterson, Penbrooke Meadows, Pineridge, Queensland, Ranchlands, Red Carpet, Renfrew, Rideau Park, Rosemont, Rundle, Rutland Park, Sage Hill, Seton, Signal Hill, Skyview Ranch, Somerset, South Calgary, Southwood, Spruce Cliff, Taradale, Throncliffe, Tuxedo Park, University of Calgary, Varsity, Vista Heights, Whitehorn, University District.

Summary: This cluster contains houses (although townhouses and apartments are options as well) under 500,000 with options under 250,000. Most homes are under 2000 sqft with 3-4 bedrooms and 2-3 bathrooms. Venue accessibility is lower with less than 25 nearby venues.

Cluster 3

Neighbourhoods: Beltline, Cliff Bungalow, Downtown Commercial Core, Downtown East Village, Downtown West End, Eau Claire, Hillhurst, Lower Mount Royal, Manchester, Mission

Summary: This cluster contains apartments and townhouses under 500,000 and less than 1000 sqft. They commonly have 1-2 bedrooms and 1-2 bathrooms. These neighbourhoods have access to over 50 venues and looking at the map they are located in the city core.

Cluster 4

Neighbourhoods: Auburn Bay, Bonavista Downs, Castleridge, Chaparral, Chinook Park, Citadel, Cityscape, Coach Hill, Copperfield, Coventry Hills, Cranston, Crestmont, Douglasdale/Glen, Evanston, Fairview, Harvest Hills, Hawkwoos, Hidden Valley, Legacy, MacEwan Glen, Maple Ridge, Marlborough, Martindale, Monterey Park, Montgomery, New Brighton, Nolan Hill, North Haven Upper, Panorama Hills, Redstone, Riverbend, Rocky Ridge, Royal Oak, Saddle Ridge, Sandstone Valley, Scenic Acres, Shawnessy, Silver Springs, Southview, Temple, Walden, Westgate, Willow Park, Woodbine, Woodlands, Yorkville, Carrington, Wolf Willow, Cornerstone, Belmont, Pine Creek.

Summary: This cluster contains houses and townhouses priced under 750,000 with options under 500,000. These homes have 3-4 bedrooms and 3-4 bathrooms. Homes in this cluster are commonly 1000-2000 sqft. Venue accessibility is lower with under 25 nearby venues. Looking at the map many of these neighbourhoods lie near the edges of the city

Discussion

From the traits noted for each cluster in the results, the five clusters can be used to describe 5 different types of buyers.

The first type of buyer would be associated with Cluster 0. They would be a wealthy buyer looking for a large house perhaps for a family unit larger than 2 people. Due to the low accessibility to venues they would most likely have to own a vehicle to get everywhere they desire.

The second type of buyer would be associated with Cluster 1. They would be less wealthy to average and probably a single person or a couple looking for something small like an apartment or townhouse. They would be looking for something with good accessibility to venues, having a vehicle would not be necessary most of the time, but would want to live outside of the city core.

The third type of buyer would be associated with Cluster 2. They would be less wealthy with a smaller budget and be looking for a smaller house. Due to the low venue accessibility they would most likely need to own a car.

The fourth type of buyer would be associated with Cluster 3. They would be of average wealth and would be looking for something small like an apartment or townhouse, most likely for a single person or couple. Owning a car would not be necessary as this cluster has the highest number of venues available. Buyers would be looking to live downtown and would most likely work downtown as well.

The fifth (last) type of buyer would be associated with Cluster 4. They would be of average wealth looking for a small to mid-sized house. These would most likely be buyers looking for a house for a family unit greater than 2 and would need a vehicle due to the low venue accessibility in these neighbourhoods.

Conclusion

Neighbourhoods were clustered into 5 groups based on real estate attributes and type and number of venues. The clusters contain similar traits and can help buyers narrow down the neighbourhoods they were searching for real estate in by matching them with a cluster. K-means clustering was deemed an applicable clustering method as it is a robust and quick clustering method to apply and the data was suitable to be used in a K-means algorithm.

Future directions for this project would be to determine which neighbourhoods are most expensive or most affordable based on having the same house attributes like size, number of bedrooms and number of bathrooms. Another factor that could be taken into consideration for this project is the accessibility to transit, this would assessed by looking at the number of bus stops or number of routes that pass through a neighbourhood.