Calgary Coffee Shops Market Analysis

Raymond Mah
20 Jan 21

1. Introduction

1.1 Background

Coffee is a brewed drink that dates back to 15th century Arabia, where it was believed that coffee seeds were roasted and brewed for drinkers to stay awake during religious rituals. Coffee plants are now cultivated in over 70 countries. Canada is ranked 10th in the world in terms of coffee consumption on a per capita, per annum basis. Statistic Canada has reported that among those who drank coffee, consumption typically peaked at age 31 to 50, averaging 639 grams for men and 586 grams for women. To many Canadians a warm cup of coffee is an important part of their daily ritual.

The Canadian city of Calgary, Alberta (YYC) was selected as the basis for this market study on coffee shops. Calgary as of 2019 has a population of 1,285,711, which makes it the most populous city in the province of Alberta and the fourth largest census metropolitan area in Canada. The Economist Intelligence Unit ranked Calgary the most livable city in North America in both 2018 and 2019.

1.2 Problem

The city of Calgary as of 2019 has 200 plus neighborhoods, 42 industrial areas along with more communities slated for development as the city grows. For a new business owner who wishes to open a new coffee shop, choosing the best location to start the business would be a potential area of concern. Utilizing data science methodologies, this market analysis aims to provide clarity on which neighborhoods are best suited for opening a coffee shop.

1.3 Interest

Obviously, the most interested parties for the results of this market analysis is any coffee shop business owner who wishes to gain better insights and further their competitive advantage. Other interested parties may include general fans of coffee who are interested in all subjects concerning coffee.

2. Data acquisition and cleaning

2.1 Data Sources

In order to conduct this market study, I needed to collect demographic data on the various Calgary neighborhoods. The City of Calgary Open data bank provided the required data set of "Census by Communities 2019". This 2019 data set contained census data, which is an official count of dwelling units and its population for a given Calgary neighborhood. This data set is a substantial 306 rows by 142 column data set that details the population's age and gender distribution for all the listed Calgary neighborhoods.

To complement this data set, I also collected household income data for the various Calgary neighborhoods. I was able to retrieve this data from the Canadian Mortgage and Housing Corporation (CMHC). The CMHC amalgamated and reported the results from the Census of Canada and National Household Survey. This household income data set was available for the 2016 data and covers 180 neighborhoods. It reported the average & median after tax household income on a per neighborhood basis. It would have been ideal if similar data for 2019 was available. However, since only 2016 data was available it was deemed sufficient as the overall wealth of a given neighborhood typically does not fluctuate drastically.

To collect data on the existing coffee shops in Calgary, I utilized the Foursquare API to first search and collect data on all the nearby venues of a given neighborhood. This data set was then filtered specifically for listed "coffee shop" venues.

Finally, the geometry coordinates data of all the Calgary neighborhoods was collected from a Github repository (Calgary GIS). The .json file contained latitude longitude values that mapped out all the neighborhoods for the city of Calgary. This set of data allowed me to produce various choropleth maps for this market study.

2.2 Data Cleaning

The Calgary Census by Community was the first dataset that needed to be cleaned for python coding use. There were 306 rows of data that represented all the residential, industrial and residual sub areas of Calgary. Residual sub areas are uninhabited areas under development so for the purposes of this market analysis, any rows in the data set listed as a residual sub area was removed. Next any listed neighborhoods that did not have a "Number of Residents (RES_CNT)" count were also removed from the data set. Please note these particular neighborhoods may actually have residents. However, since the 2019 census was unable to collect data on them, there wouldn't be any of the required demographic data needed for this study.

For the 142 columns, the data set contained a wide assortment of information (i.e. number of preschool children, number of dwelling units, number of homeowners, etc.). For the purpose of this study, I only needed demographical data of the different neighborhoods. As such, any of the irrelevant data columns were removed from the python dataframe. The remaining data contain demographic information on the residents in a given neighborhood categorized by age group and gender. As shown in Table 1, sample demographic data for the first row would read as follows:

- Residential neighborhood of Legacy
  - RES_CNT = total number of residents in Legacy = 6420
  - MALE_CNT = total number of male residents in Legacy = 3125
  - MF_25_34 = number of male and female residents in Legacy aged 25 to 34 = 1518

Table 1: Data Cleaning Stage 1 - YYC Main

	CLASS	NAME	SECTOR	RES_CNT	MALE_CNT	FEMALE_CNT	MF_25_34	MF_35_44	MF_45_54	OTHER_25_34	OTHER_35_44	OTHER_4
0	Residential	LEGACY	SOUTH	6420	3125	3283	1518	1204	724	0	0	
1	Residential	HIGHLAND PARK	CENTRE	3838	1986	1846	777	798	485	0	0	
2	Residential	CORNERSTONE	NORTHEAST	2648	1409	1225	607	593	261	0	0	
3	Residential	MONTGOMERY	NORTHWEST	4515	2216	2291	1003	786	530	0	0	
4	Residential	TEMPLE	NORTHEAST	10977	5597	5357	1616	1727	1303	0	0	

Statistic Canada has reported, coffee consumption typically peaks at age 31 to 50. As such the target demographic of interest in Table 1 is essentially any residents that fall in that age range. The next subsequent step was to amalgamate the applicable data columns to create the new data column of "Number of Residents Age 25 to 54". This step in data cleaning can be observed in Table 2 below.

Table 2: Data Cleaning Stage 2 - YYC Main Demographics

	CLASS	NAME	SECTOR	RES_CNT	Number of Residents Age 25 to 54
0	Residential	LEGACY	SOUTH	6420	3446
1	Residential	HIGHLAND PARK	CENTRE	3838	2060
2	Residential	CORNERSTONE	NORTHEAST	2648	1461
3	Residential	MONTGOMERY	NORTHWEST	4515	2319
4	Residential	TEMPLE	NORTHEAST	10977	4646

At this stage, the main data frame shown in Table 2 contains data on 211 neighborhoods. Utilizing the python client GeoPy, the latitude and longitude values of each neighborhood were retrieved and added to the YYC Main data frame, as shown in Table 3.

Table 3: Data Cleaning Stage 3 - YYC Main Latitude Longitude

	CLASS	NAME	SECTOR	RES_CNT	Number of Residents Age 25 to 54	Latitude	Longitude
0	Residential	LEGACY	SOUTH	6420	3446	50.856893	-114.002560
1	Residential	HIGHLAND PARK	CENTRE	3838	2060	51.085355	-114.065809
2	Residential	CORNERSTONE	NORTHEAST	2648	1461	51.160280	-113.939608
3	Residential	MONTGOMERY	NORTHWEST	4515	2319	51.074802	-114.162474
4	Residential	TEMPLE	NORTHEAST	10977	4646	51.088424	-113.947877

Since the neighborhood latitude and longitude values were necessary to subsequently retrieve the nearby venue data, any neighborhood where the latitude and longitude values could not be retrieved was removed from the dataframe. This resulted in 210 remaining neighborhoods in the YYC main dataframe.

Using the Foursquare API, all venues located within the vicinity of the neighborhoods were retrieved. Within the python coding for the venue retrieval, the search radius from the neighborhood latitude longitude values was set at 850 meters. The average residential neighborhood in Calgary has an area of 2.325 km^2. As such the average neighborhood radius was calculated to be approximately 860 meters. Please note, the search radius may not provide 100% coverage of all the Calgary neighborhoods, as some neighborhoods are significantly larger than the majority. The radius used for search parameters may not be fully accurate for those select few neighborhoods, but the majority of yyc neighborhoods fall within the average values, so the radius value of 850 meters was deemed sufficient for this study.

The venue data retrieved was added to the main data frame as shown in Table 4 below.

Table 4: Data Cleaning Stage 4 - YYC Main Venues

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	LEGACY	50.856893	-114.002560	Shane Homes - Legacy	50.858113	-114.001418	Real Estate Office
1	LEGACY	50.856893	-114.002560	Domino's Pizza	50.862533	-114.009593	Pizza Place
2	LEGACY	50.856893	-114.002560	Bobby's Place	50.862768	-114.009751	Pub
3	LEGACY	50.856893	-114.002560	Sweet Cakes By Vernz	50.855133	-114.014450	Bakery
4	HIGHLAND PARK	51.085355	-114.065809	Citizen Brewing Company	51.083785	-114.058404	Brewery

Since this study was specific to Coffee Shop venues, all venues not listed as a coffee shop in the “Venue Category” column were removed from the data frame. This resulted in 235 Coffee Shops for the 210 neighborhoods used in the study.

Table 5: Data Cleaning Stage 5: YYC Main Coffee Shops

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
13	MONTGOMERY	51.074802	-114.162474	Starbucks	51.074935	-114.166081	Coffee Shop
16	MONTGOMERY	51.074802	-114.162474	Tim Hortons	51.071732	-114.163839	Coffee Shop
40	WOODBINE	50.942554	-114.128853	Starbucks	50.940328	-114.119909	Coffee Shop
49	UNIVERSITY HEIGHTS	51.070740	-114.137231	Starbucks	51.075199	-114.137260	Coffee Shop
53	UNIVERSITY HEIGHTS	51.070740	-114.137231	Tim Hortons	51.069060	-114.128237	Coffee Shop

Subsequently, a new column was created to show the coffee shop count on a per neighborhood basis.

Table 6: Data Cleaning Stage 6: YYC Main Coffee Shop Count

	CLASS	NAME	SECTOR	RES_CNT	Number of Residents Age 25 to 54	Latitude	Longitude	Coffee Shop CNT
0	Residential	LEGACY	SOUTH	6420	3446	50.856893	-114.002560	0.0
1	Residential	HIGHLAND PARK	CENTRE	3838	2060	51.085355	-114.065809	0.0
2	Residential	CORNERSTONE	NORTHEAST	2648	1461	51.160280	-113.939608	0.0
3	Residential	MONTGOMERY	NORTHWEST	4515	2319	51.074802	-114.162474	2.0
4	Residential	TEMPLE	NORTHEAST	10977	4646	51.088424	-113.947877	0.0

The 2016 house income data can now be merged with the main data frame. Any neighborhoods that did not contain average or median household income data were also removed. This was necessary as any further analysis was conducted based on the given neighborhood's target demographic and wealth level.

Table 7: Final Data Cleaning Stage: YYC Main Final

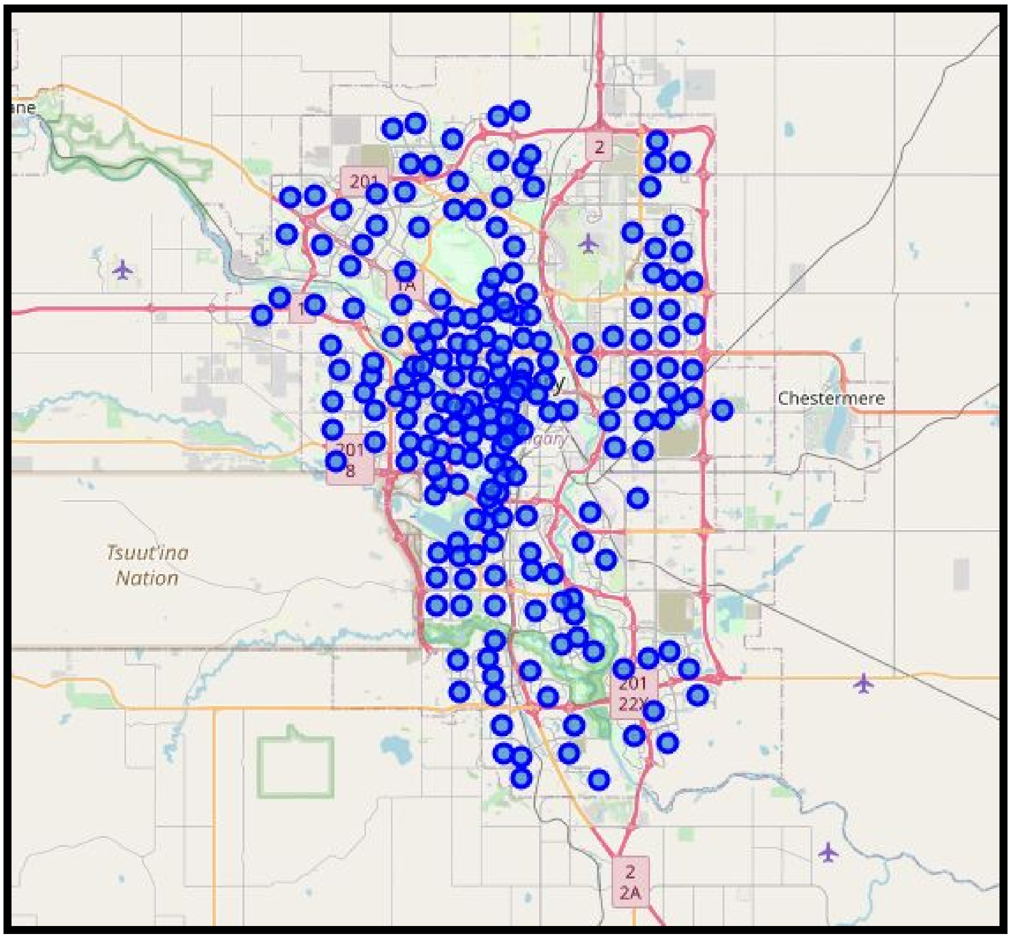
	CLASS	NAME	SECTOR	RES_CNT	Number of Residents Age 25 to 54	Latitude	Longitude	Coffee Shop CNT	Average Household Income After Taxes	Median Household Income After Taxes
0	Residential	LEGACY	SOUTH	6420	3446	50.856893	-114.002560	0.0	113658.0	95756.0
1	Residential	HIGHLAND PARK	CENTRE	3838	2060	51.085355	-114.065809	0.0	82832.0	64353.0
3	Residential	MONTGOMERY	NORTHWEST	4515	2319	51.074802	-114.162474	2.0	97159.0	67458.0
4	Residential	TEMPLE	NORTHEAST	10977	4646	51.088424	-113.947877	0.0	79742.0	72270.0
5	Residential	WOODBINE	SOUTH	8866	3316	50.942554	-114.128853	1.0	128229.0	94028.0

The final dataframe resulted in a data frame with 167 neighborhoods which will be used for further analysis. Due to the data limitations from the various data sources used, not every Calgary neighborhood is represented in this study. Despite this limitation, the final data frame of 167 neighborhoods is still considered sufficient for identifying neighborhoods with high potential for opening new coffee shops.

3. Methodology

I utilized python folium library to visualize geographical details of Calgary and all the listed neighborhoods under review for this study. As shown in Map 1, the map displays 210 neighborhoods which is consistent with data cleaning stage 3. This Map allows readers of this report to clearly visualize the distribution of neighborhoods for the city of Calgary. As it can be seen, there is a clustering of smaller neighborhoods (area wise) near the city center and comparatively larger neighborhoods in the outer suburban areas.

Map 1: The City of Calgary Map (210 Neighborhoods)

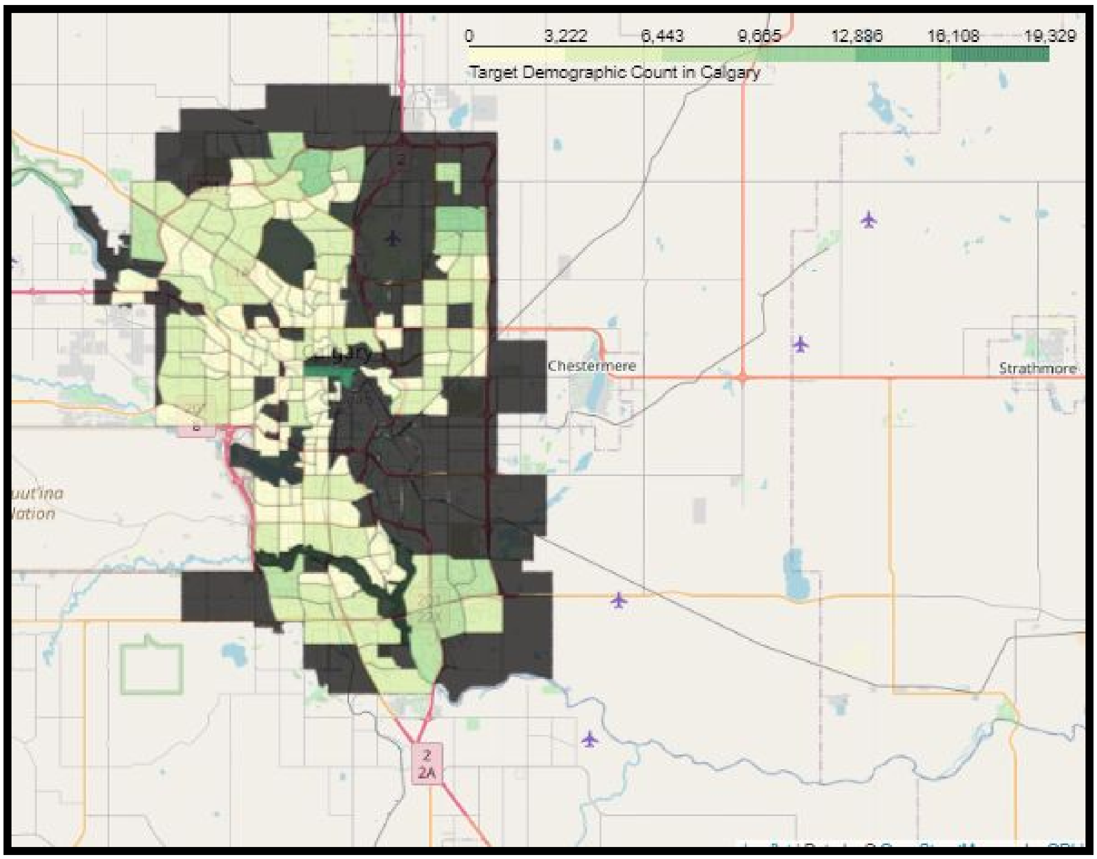


3.1 Choropleth Maps

After the final stages of data cleaning were completed, various choropleth maps (i.e. heat maps) were generated via the python folium library to review market conditions in Calgary. The following maps utilized the YYC Main final data frame with data on 167 neighborhoods. I created choropleth maps to review the distribution of the target demographic, current number of coffee shops and average & median after tax household income for a given neighborhood.

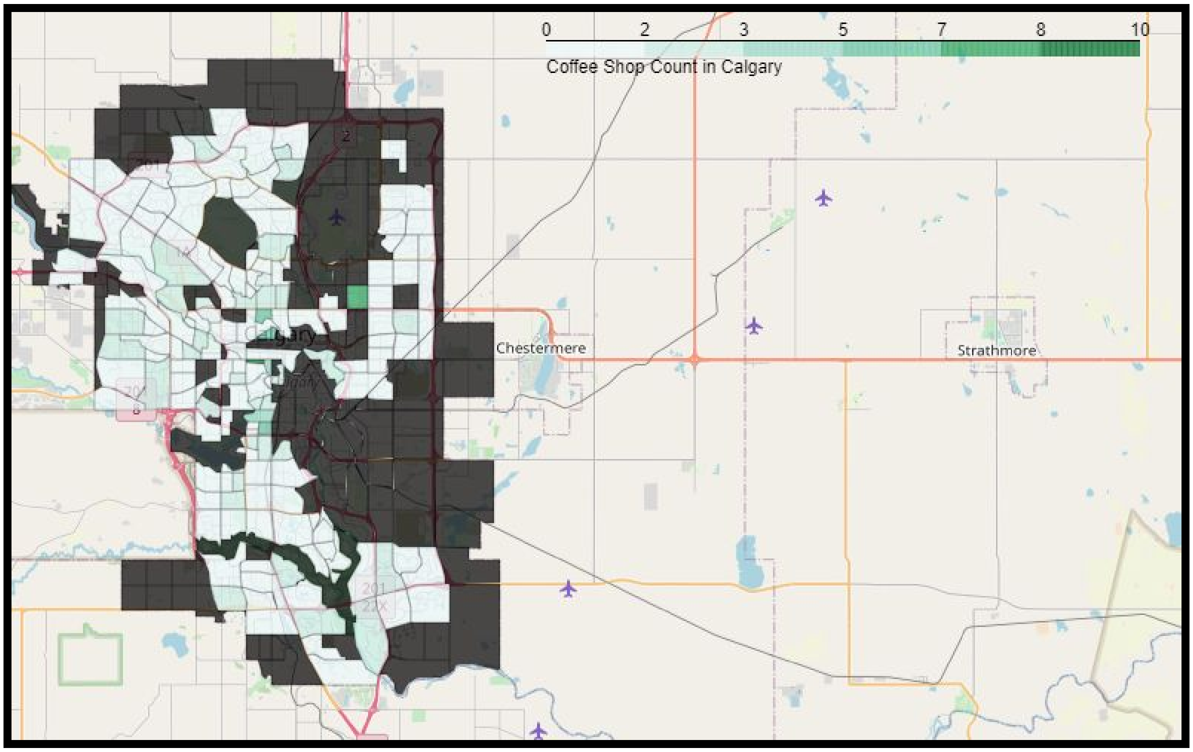
As reported by Statistics Canada, peak coffee consumption is around age 31 to 50. For this study the target demographic group is set at residents age 25 to 54 due to how the raw Census data was structured. Map 2 highlights the distribution of the target demographic for the 167 neighborhoods reviewed. An example of an insight that can be extracted from Map 2 is that potential/existing coffee shop owners can see areas with low count of the target demographic group.

Map 2: YYC Target Demographic Choropleth Map

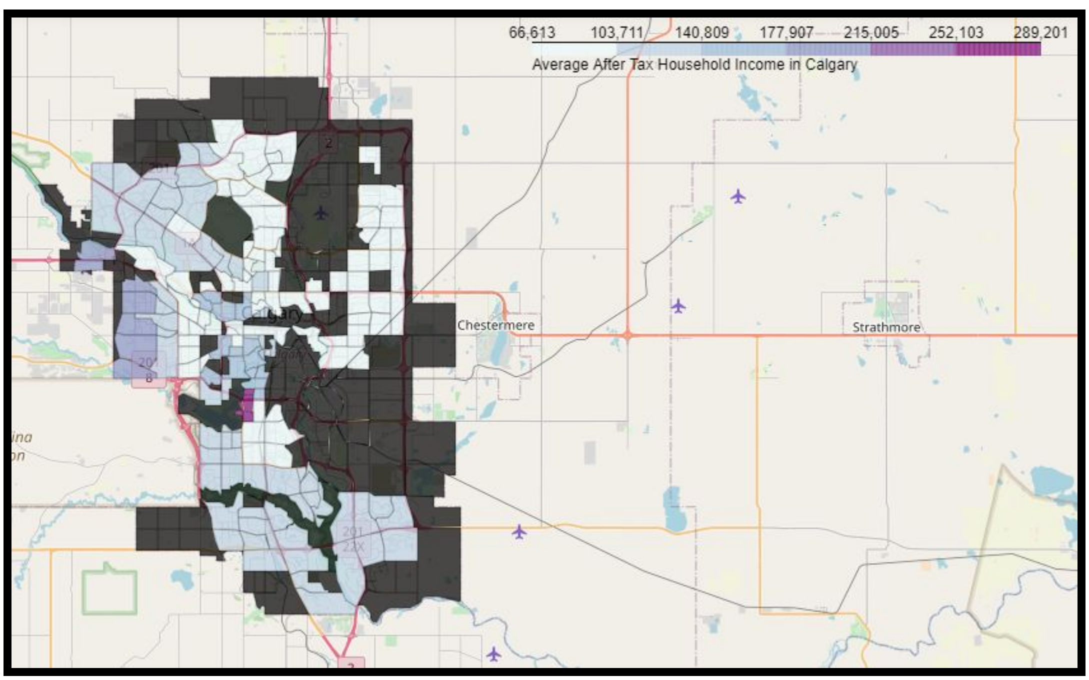


If those areas already have established coffee shops then those areas are potentially overly saturated with competition. Visually, Map 2 results can be quickly compared with Map 3 results to see the level of market competition in a given neighborhood. Of course it is important to keep in mind that this does not paint the full picture. As certain neighborhoods may have a strong industrial core where a lot of workers go to as a place of employment. So although the target demographic count may be low (i.e. only counts residents) and the number of coffee shops is relatively high in the neighborhood, a coffee shop there may still be able to sustain its business because of the amount of workers there.

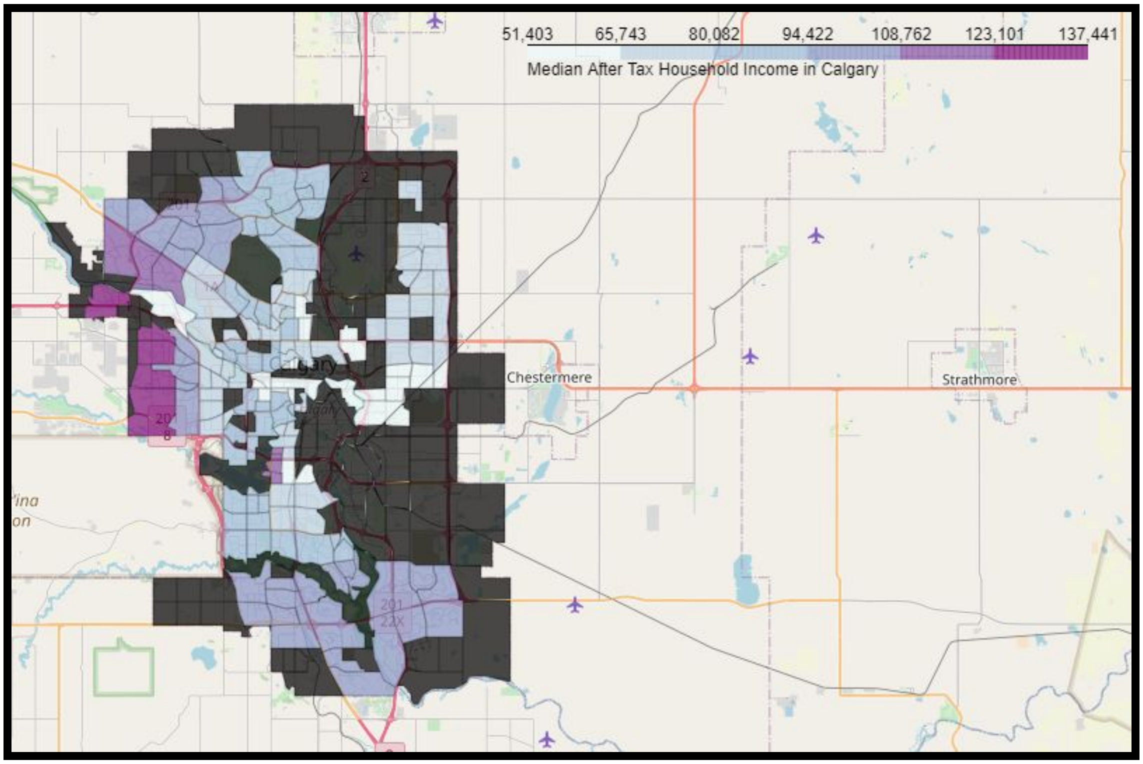
Map 3: Existing Coffee Shop Count Choropleth Map



Map 4: Average After Tax Household Income YYC - Choropleth Map



Map 5: Median After Tax Household Income YYC - Choropleth Map



The results from Map 4 can be reviewed in conjunction with Map 5 to review the overall wealth distribution for the neighborhoods in Calgary. It was interesting to note that there are some neighborhoods that had high average household income but only moderate median household income. This may imply that for those areas, the high average may be skewed by a small number of residents with extremely high household income that would raise the average value, but would not affect the median value. From this insight, further analysis should prioritize the median household income value over the average value. As the median value would better represent the overall wealth of a given neighborhood.

3.2 Exploratory Data Analysis

The following methodology section will discuss and describe any exploratory data analysis that was completed. As such, this section will cover the machine learnings performed. For the purposes of this Coffee Shop market study, the machine learning modeling method of K-means clustering was used to segment all the neighborhoods in YYC Main Final data frame.

3.2.1 Machine Learning Modeling

To reiterate the original objective of this market study, I want to provide clarity on which neighborhoods are best suited for opening a coffee shop. In order to achieve that, I needed to segment all the neighborhoods under review into distinguishable clusters. I would need to find commonality amongst the neighborhoods based on the attribute factors of target demographic count, average after tax household income and median after tax household income.

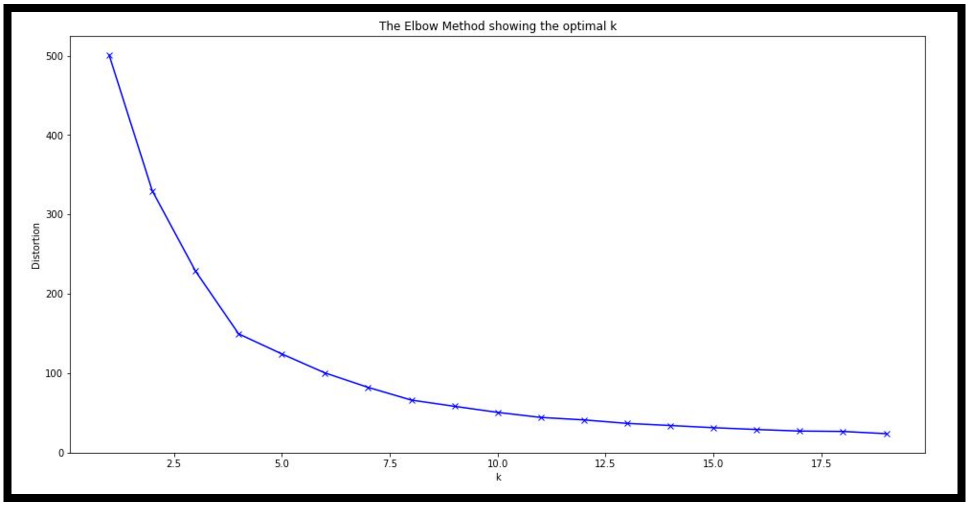
I was able to achieve this via the machine learning modelling method of k-means clustering. This method utilizes an iterative refinement approach and is an unsupervised machine learning algorithm that groups data into k number of clusters. As shown in Table 8, those 3 data columns were pulled from the YYC Main Final data frame for machine learning modelling.

Table 8: K-Means Clustering Variables

	RES_CNT	Number of Residents Age 25 to 54	Average Household Income After Taxes	Median Household Income After Taxes
0	6420	3446	113658.0	95756.0
1	3838	2060	82832.0	64353.0
3	4515	2319	97159.0	67458.0
4	10977	4646	79742.0	72270.0
5	8866	3316	128229.0	94028.0

For K-means clustering, a limitation is that the number of clusters are user-defined and the algorithm will try to group the data even if this number is not optimal for the specific case. As such, the optimal k-value needs to be determined first. The Elbow method is a very popular technique for determining optimal k value. This methodology entails running k-means clustering for a range of clusters k (i.e. k = 1 to 15) and for each value, the sum of squared distances from each point to its assigned cluster centroid (distortions) is calculated. When the distortions are plotted and the plot looks like an arm then the “elbow”(i.e. the point of inflection on the curve) is the best value of k.

Figure 1: Determining Optimal K value



For this market study, optimal K was determined to be 4 as there were no more substantial drops in distortion past the K value of 4 (i.e. Curve starts flattening after K = 4). With completion of finding optimal K value, the YYC Main Final dataframe will segment the 167 neighborhoods into 4 distinctive clusters. For further mapping purposes, cluster labels were added for each neighborhood in the data frame as shown in Table 9.

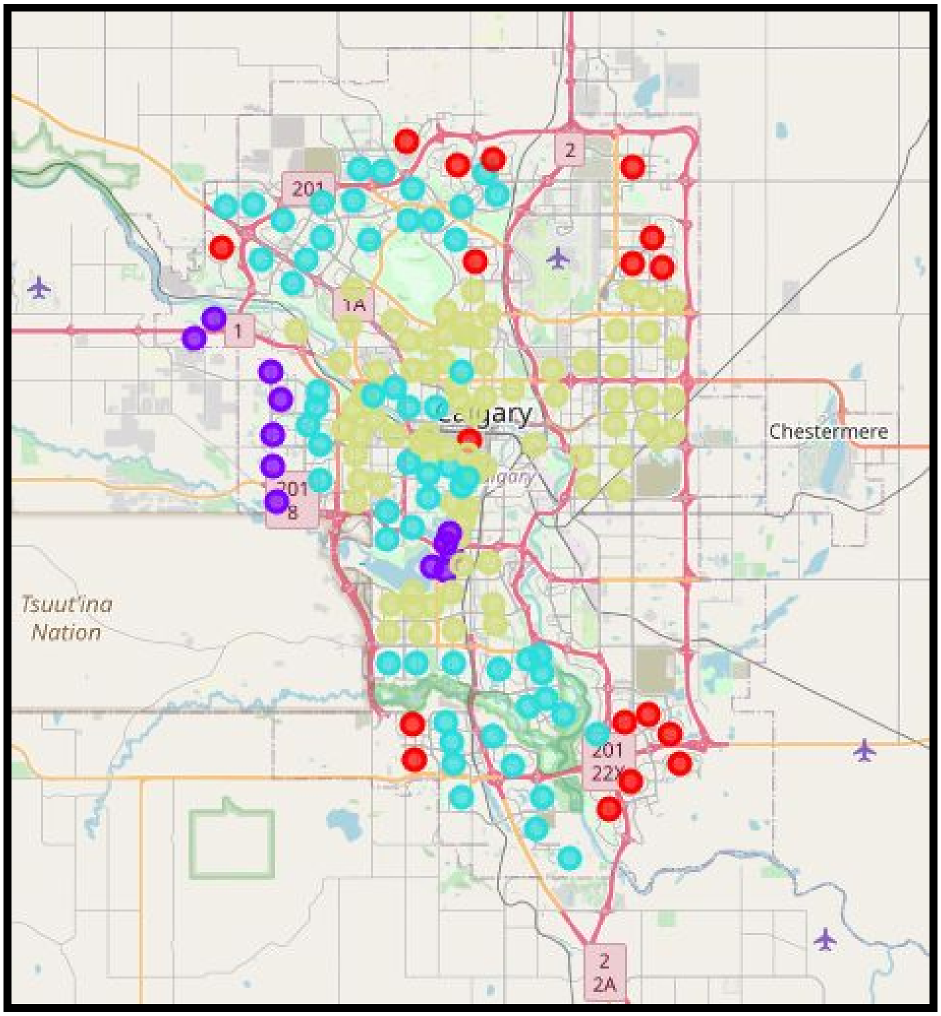
Table 9: YYC Main Final Data Frame - Cluster Label Column

	CLASS	NAME	SECTOR	RES_CNT	Number of Residents Age 25 to 54	Latitude	Longitude	Coffee Shop CNT	Average Household Income After Taxes	Median Household Income After Taxes	Clus_km
0	Residential	LEGACY	SOUTH	6420	3446	50.856893	-114.002560	0.0	113658.0	95756.0	2
1	Residential	HIGHLAND PARK	CENTRE	3838	2060	51.085355	-114.065809	0.0	82832.0	64353.0	3
3	Residential	MONTGOMERY	NORTHWEST	4515	2319	51.074802	-114.162474	2.0	97159.0	67458.0	3
4	Residential	TEMPLE	NORTHEAST	10977	4646	51.088424	-113.947877	0.0	79742.0	72270.0	3
5	Residential	WOODBINE	SOUTH	8866	3316	50.942554	-114.128853	1.0	128229.0	94028.0	2

4. Results

4.1 Final clustering results

Map 6: YYC Neighborhood Segmentation



Map 6 displays final cluster results for the 167 neighborhoods reviewed in this study. The neighborhoods are marked as follows:

- Red Neighborhoods = Cluster 1
- Purple Neighborhoods = Cluster 2
- Blue Neighborhoods = Cluster 3
- Yellow Neighborhoods = Cluster 4

4.2 Examine the Clusters

After completing clustering and segmentation of the 167 neighborhoods, the neighborhoods that already had existing coffee shops will not be reviewed further. This was completed to further refine the analysis parameters and identify the neighborhoods with the highest potential.

This decision to exclude those data values could easily be reversed if it was so desired. The python coding structure and the YYC Main Final data frame remains unchanged, and it would simply entail updating the decision parameter to include those neighborhoods. However, for all the following analysis and discussion, I only reference the 81 neighborhoods that do not have existing coffee shops listed.

Table 10: Neighborhood Count per Cluster

0 Coffee Shops Listed	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Neighborhoods	8	7	32	34

Table 11: Cluster Categorization Description

Description	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Target Demographic	High	Low	Moderate	Moderate
Average Household Income (After Tax)	Moderate	High	Moderate	Low
Median Household Income (After Tax)	Moderate	High	Moderate	Low

Table 12: Cluster Attribute Average Values

Average Values	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Target Demographic	8410	1186	2690	2609
Average Household Income (After Tax)	$103,661.38	$ 229,194.14	$ 127,925.03	$ 87,573.71
Median Household Income (After Tax)	$ 90,654.38	$ 126,497.43	$ 96,680.00	$ 69,628.53

As it is shown in Table 10 to 12, the machine learning clustering results found clear distinctions between the 4 clusters. So, preference for any particular cluster depends on the personal desire of the individual reader. For example, if there is a new coffee shop business owner who wishes to open a coffee shop that will be extremely high end. The beverages sold are high in price point and emphasizes customer customization. The coffee shop owner would like to brand their franchise on user experience and move away from the volume base model. With those considerations, it may be recommended that this coffee shop owner explore available locations in Cluster 2. Conversely, if another coffee shop owner wants to have moderate pricing but use a volume based sales model then Cluster 1 locations should be considered.

4.2.1 Target Demographic Distribution

Figure 2: Cluster 1 Target Demographic Distribution

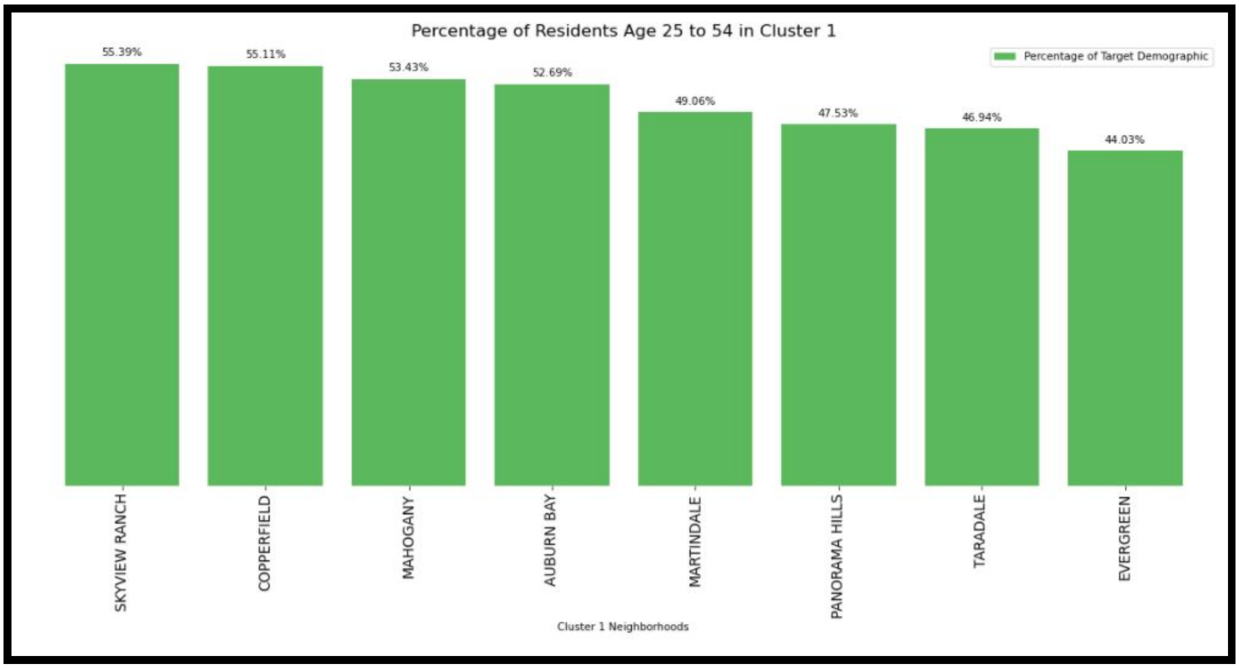


Figure 3: Cluster 2 Target Demographic Distribution

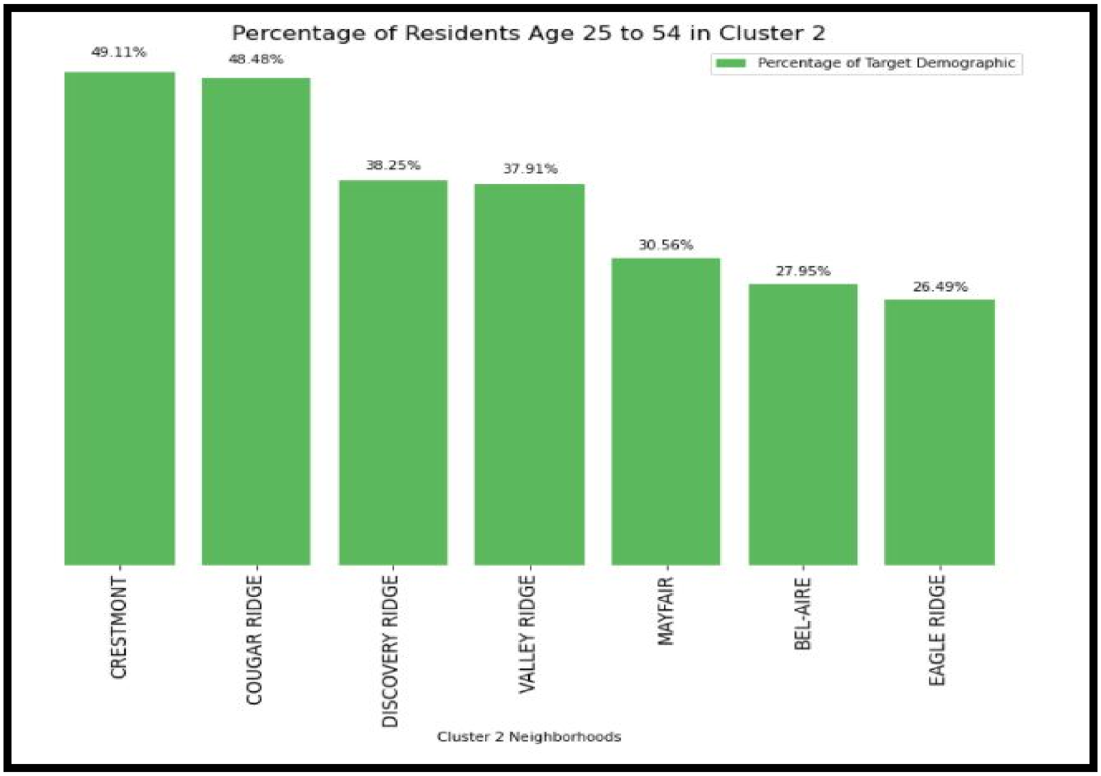


Figure 4: Cluster 3 Target Demographic Distribution

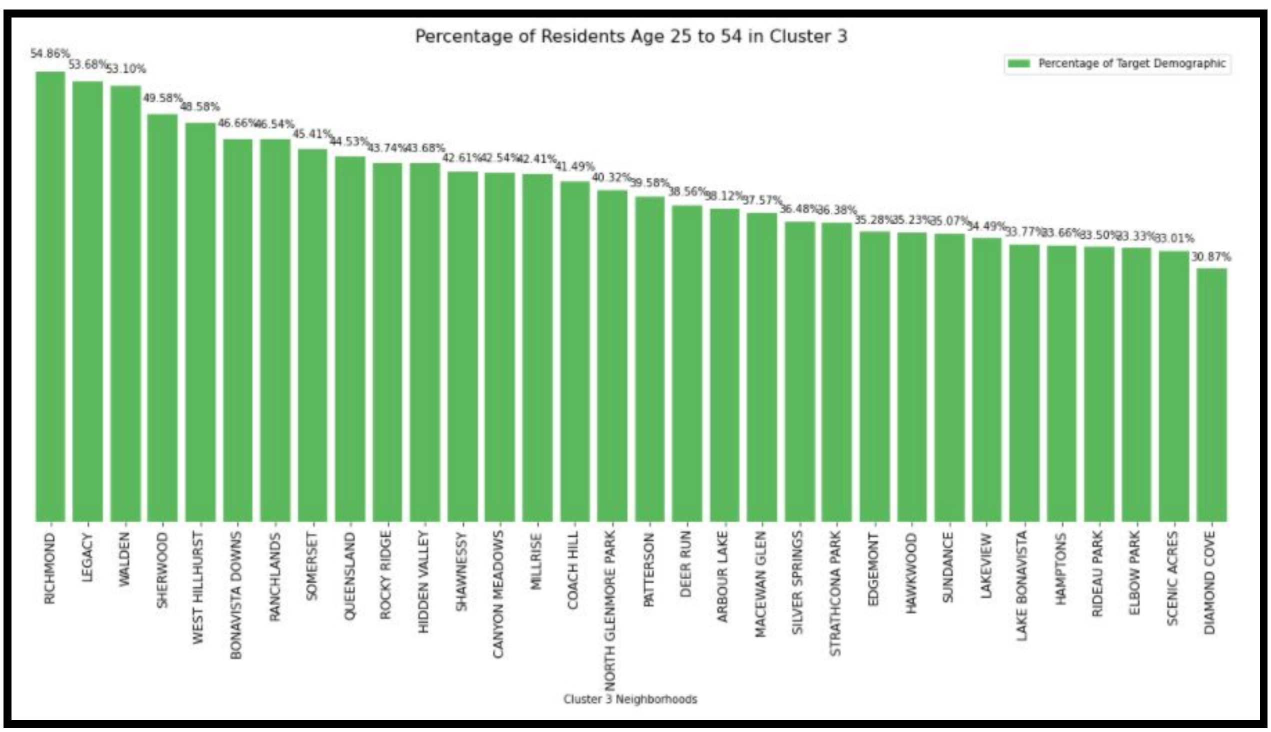


Figure 5: Cluster 4 Target Demographic Distribution

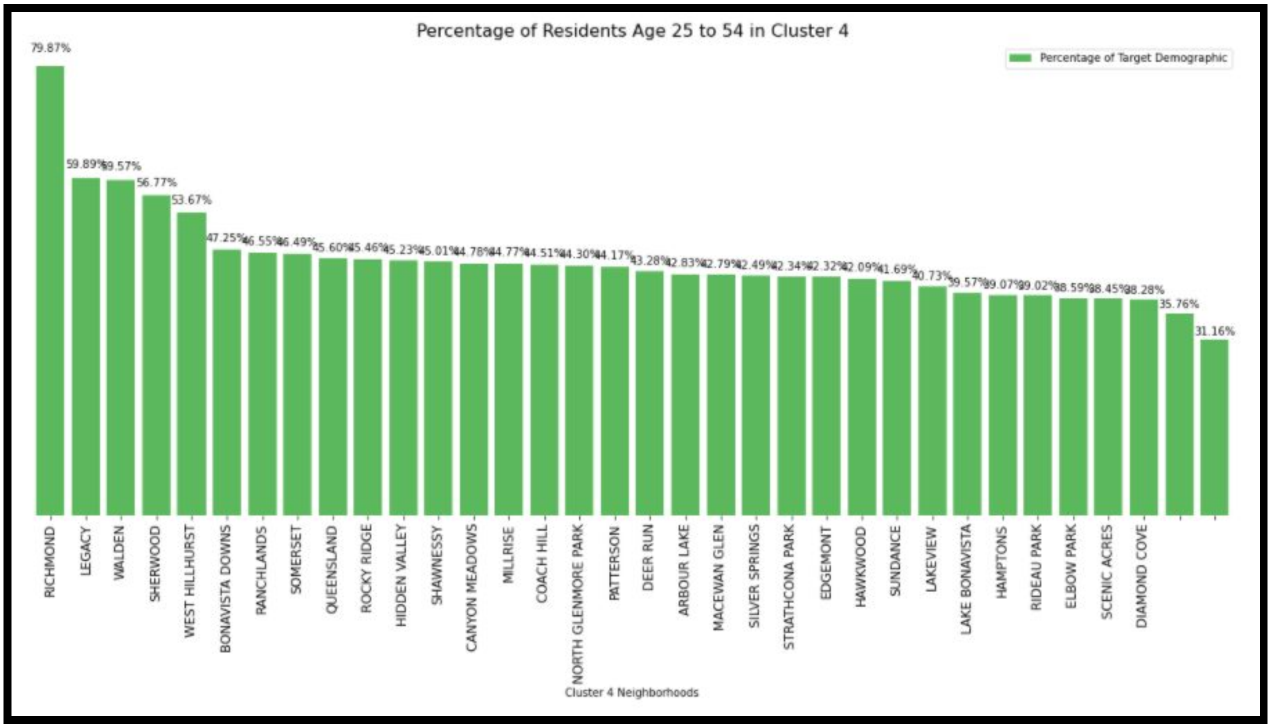


Figure 2 to 5 highlights the percentage of the target demographic in each of the neighborhoods on a per cluster basis. The percentages were sorted in descending order so readers can easily visualize the target demographic distribution of the neighborhoods in relations to each other. It is important to note that a neighborhood can have a lower target demographic percentage, but actually have an above average number of residents in the target demographic group (i.e. some neighborhoods have more total residents than others - potentially resulting in a smaller percentage value).

5. Discussion

From the above results, individual readers are already provided with lists of potential neighborhoods that may suit their needs. To further narrow down the choices, I would provide the top 2 or 3 choices of each cluster based on the following selection parameters:

- In order of Priority for Selection, must satisfy all 3 criteria
  - Above cluster average value (i.e. Table 12 results) of “After Tax Median Household Income”
  - Above cluster average value (i.e. Table 12 results) of “Number of residents Age 25 to 54”
  - Above cluster average value (i.e. Table 12 results) of “After Tax Average Household Income”

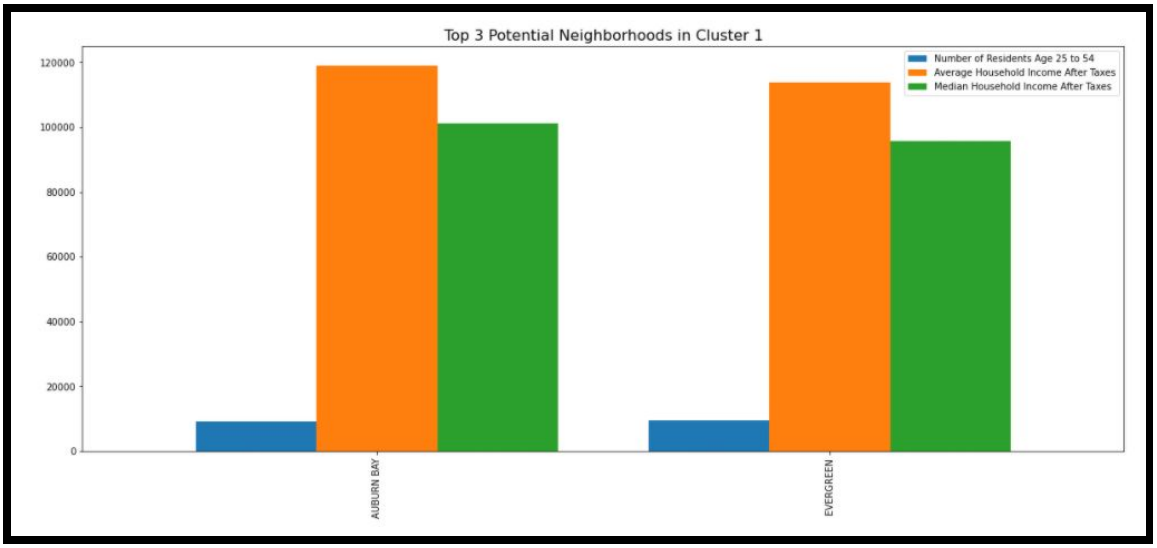
Cluster 1 Recommendations

Neighborhood in Cluster 1 must meet the following:
- "After Tax Median Household Income" above $90,654.00
- "Number of residents Age 25 to 54" above 8410
- "After Tax Average Household Income" above $103,661.00

Table 13: Cluster 1 Recommendations

NAME	Number of Residents Age 25 to 54	Average Household Income After Taxes	Median Household Income After Taxes
AUBURN BAY	9277	119066.0	101028.0
EVERGREEN	9467	113658.0	95756.0

Figure 6: Cluster 1 Recommendations Bar Chart



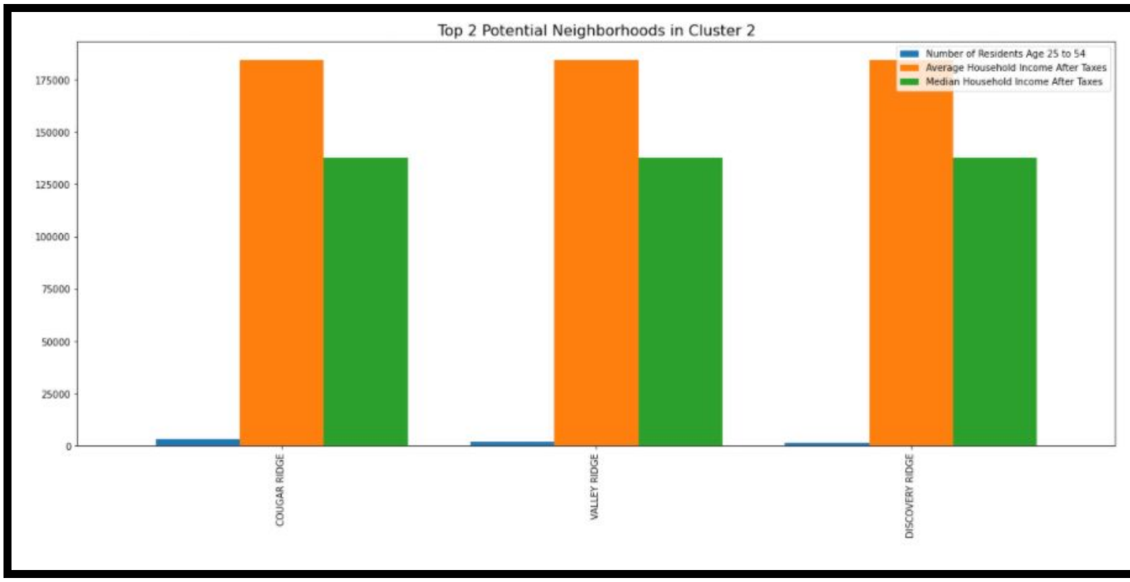
Cluster 2 Recommendations

Neighborhood in Cluster 2 must meet the following:
- "After Tax Median Household Income" above $126.497.00
- "Number of residents Age 25 to 54" above 1186
  - "After Tax Average Household Income" criteria was excluded for Cluster 2 as the average value of $229,194.00 was skewed high with cluster 2 only having 7 neighborhoods. If a neighborhood was extremely wealthy the average value is easily skewed.

Table 14: Cluster 2 Recommendations

NAME	Number of Residents Age 25 to 54	Average Household Income After Taxes	Median Household Income After Taxes
COUGAR RIDGE	3392	184189.0	137441.0
VALLEY RIDGE	2117	184189.0	137441.0
DISCOVERY RIDGE	1637	184189.0	137441.0

Figure 7: Cluster 2 Recommendations Bar Chart



Cluster 3 Recommendations

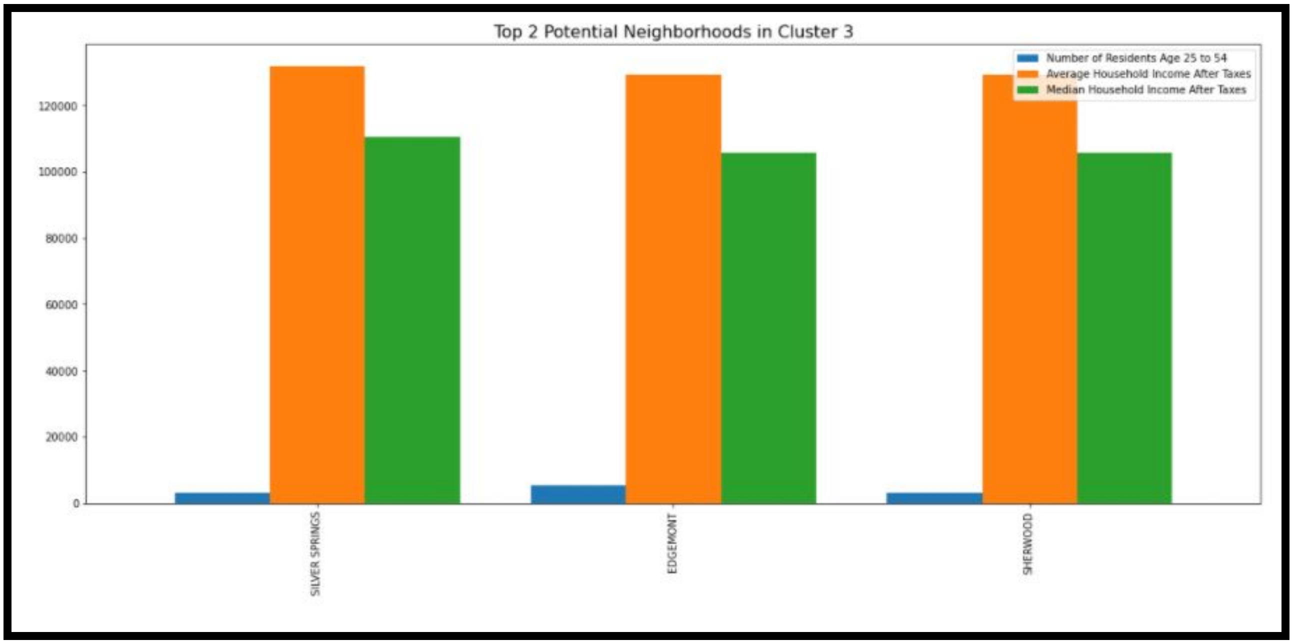
Neighborhood in Cluster 3 must meet the following:

- “After Tax Median Household Income” above $96,680.00
- “Number of residents Age 25 to 54” above 2690
- “After Tax Average Household Income” above $127,925.00

Table 14: Cluster 3 Recommendations

NAME	Number of Residents Age 25 to 54	Average Household Income After Taxes	Median Household Income After Taxes
SILVER SPRINGS	3166	132035.0	110614.0
EDGEMONT	5432	129218.0	105839.0
SHERWOOD	3097	129218.0	105839.0

Figure 8: Cluster 3 Recommendations Bar Chart



Cluster 4 Recommendations

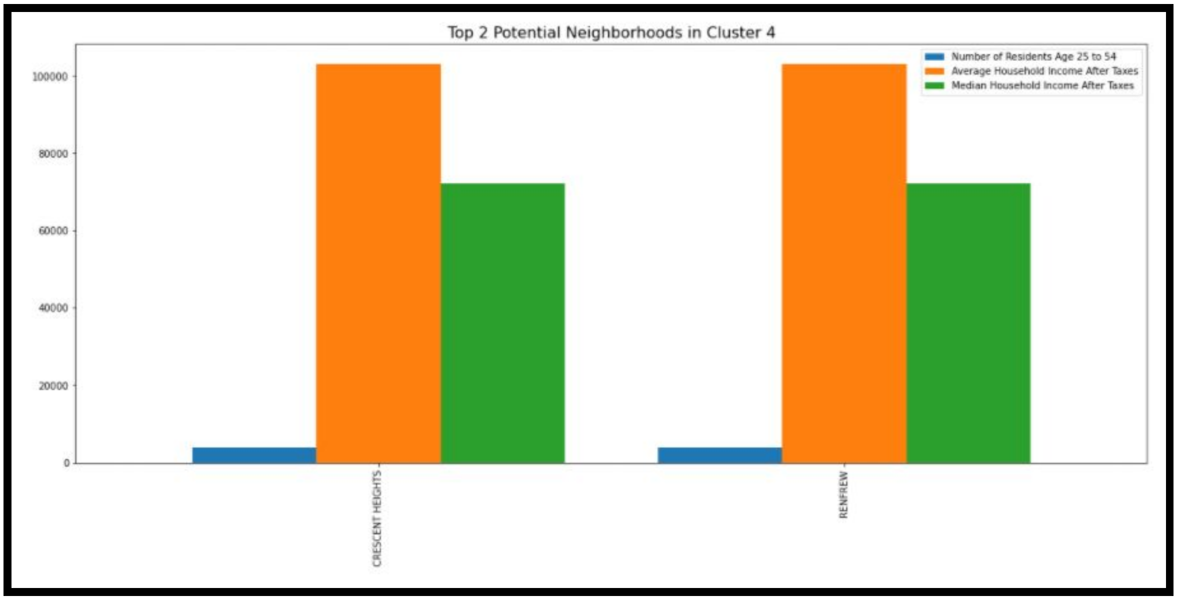
Neighborhood in Cluster 4 must meet the following:

- “After Tax Median Household Income” above $69,628.00
- “Number of residents Age 25 to 54” above 2609
- “After Tax Average Household Income” above $87,573.00

Table 15: Cluster 4 Recommendations

NAME	Number of Residents Age 25 to 54	Average Household Income After Taxes	Median Household Income After Taxes
CRESCENT HEIGHTS	3965	103136.0	72113.0
RENFREW	3921	103136.0	72113.0

Figure 8: Cluster 4 Recommendations Bar Chart



Recommendation Summary

In Table 16, the recommendation summary indicates the top choices within each cluster based on the selection criteria used above.

Table 16: Neighborhood Recommendations per Cluster

Recommendations	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1st Choice	Auburn Bay	Cougar Ridge	Silver Springs	Crescent Heights
2nd Choice	Evergreen	Valley Ridge	Edgemont	Renfrew
3rd Choice		Discovery Ridge	Sherwood	

Of course these are based on selection parameters that I determined to be reasonable. Adjustments to the parameters can easily be made to have either more tolerance or restrictions. As there is such a level of variability, very different approaches could be used in clustering and classification studies such as this one.

Future Direction

Looking beyond this market study, if the required data is available the following are potential areas of interest worth exploring:

- Complete the exact same market study while factoring consideration on venues that are in indirect competition to coffee shops such as venues listed as Cafes from the Foursquare API
- Complete the final recommendations section of this report with the inclusion of the neighborhoods with existing coffee shops and compare it to current recommendations list
- Explore potential inclusion of more data attributes for the main dataframe
    - Analyze the number of people who work in a given neighborhood as a place of employment
    - Review legislative commercial zoning considerations on a per neighborhood basis to indicate neighborhoods that actually permit the opening of a new coffee shop
    - Review commercial real estate data to include the most current average price of commercial real estate in a given neighborhood

These are all potential areas of interest that would add value and insights to this market study of Calgary Coffee shops. If reliable data can be found for those items, then this market study could be expanded on and the python dataframe built can provide even greater insights.

6. Conclusion

Calgary is a growing city with an ever expanding list of neighborhoods. For new coffee shop owners, selecting the best location may be a daunting task when there are so many locations to choose from. To restate the aim of this market study, the objective was to utilize data science methodologies to provide clarity on selecting the most ideal neighborhoods for opening a new coffee shop.

I was able to achieve this objective through the process of data cleaning raw data from multiple sources and subsequent utilization of machine learning methodology (i.e. k means clustering) to properly segment the Calgary neighborhoods. Market segmentation was conducted based on the parameters of target demographic count, average after tax household income and median after tax household income. With sincerity, I hope this market study was found to be useful for the readers and that potential coffee shop owners can gain valuable insights from it.

All the best,

Raymond Mah

7. References

[1] The City of Calgary Open data bank

[2] Foursquare API

[3] Calgary GIS Github Repository

[4] Canada Mortgage and Housing Corporation