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.