# Introduction

## Background

Toronto is a hot location, with growth accelerating every year. Many people want to move into and set up business in Toronto. As such this is a highly competitive landscape, with high rents and lots of competition. A client wants to open a series of coffee shops in the region. They have a fair amount of capital so are able to shop quite freely within the GTA area. Coffee shops are built on foot traffic, most business occurs in store as such getting the necessary foot traffic is essential for the success of the business.

## Problem

They don’t want to waste their money, they want a location that is not already saturated with coffee shops or cafes and will provide them with a good deal of foot traffic.

# Data

## Sources

The data used will include postal code data for boroughs and neigbourhoods, the co-ordinate information for the postal codes, income and tax amounts for each postal code, and Four Square venue information. In order to use this data it will need to be imported from various sources.

Postal code data for boroughs and neigbourhoods will be scraped from Wikipedia. This is not necessarily the best source of data for many applications as it is unverified, however for basic data such as postal codes I’d consider it accurate

Co-ordinate information will be extracted from Statistics Canada’s Forward Sortation Areas (FSA) boundary files. This is a reputable source with accurate location data.

Income and tax amounts were obtained from the Canada Revenue Agency at <https://www.canada.ca/en/revenue-agency/programs/about-canada-revenue-agency-cra/income-statistics-gst-hst-statistics/individual-tax-statistics-fsa/individual-tax-statistics-fsa-2017-edition-2015-tax-year.html>. This is a reputable data source with accurate economic data. It covers income tax amounts for 2015 so is slightly out of date. However, I believe it still provides a reasonable view of the current economic landscape.

Venue information in each FSA was be provided by the Four Square API.

## Cleaning

All this information will need to be combined into a single data frame in order to apply analysis. This is difficult as the sources are diverse.

For the postal code data for boroughs and neighbourhoods, I scraped the table on <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>. I was able to directly pull the data from the url using pandas. I then removed rows with an unassigned borough and checked for duplicate postal codes. No duplicate postal codes were present so I proceeded to the next data set.

The co-ordinate information for each of the postal codes could be found with Statistics Canada’s Forward Sortation Areas (FSA) boundary files, which is where I extracted the data and put into a csv file. This data set was very clean with all data present so I proceeded to merge it with the postal code data for boroughs and neighbourhoods.

With the co-ordinate data, I was then able to make calls to the Four Square API. For our purposes we were interested in the venues so that is the data that was called.

Income and tax amount data was delivered in the form of a CSV file, downloadable from the CRA website at the url mentioned above. This was a very clean data set with complete data for all the areas of interest. The only exception was Queen’s park, which reported no income tax as you would expect. The government doesn’t need to pay taxes to itself after all!

# Methodology

In order to address the clients problem, I performed clustering analysis. With this I was able to paint a semantic map of the types of businesses within an area. This allows us to identify potentially lucrative locations based on the amount of competition present and the likelihood of getting high foot traffic based on the number and types of venues within the FSA.

With the income data, we’re also able to see what the general economic state of the areas of interest are. The client will then be able to use this data when searching for places to rent in the identified areas. A lower ratio of price to reported income will likely indicate a better deal.

# Results

A picture containing text, map

Description automatically generatedUsing kmeans clustering, I was able to identify four qualitatively distinct regions as shown below. Cluster 1 is shown as purple, cluster 2 is shown as teal, cluster 3 is shown as yellow, and cluster 4 is shown as red.

# Discussion

Among all the clusters, cluster 4 appears to be the best bet. All regions within this cluster have significant recreational facilities (all have parks or playgrounds as the most common venue) with few food offerings (and no café’s or coffee shops in the top 10). There is also significant diversity in income among regions within this cluster with total income per capita ranging from 28 185 per year to 211 828 per year. This means there will be a broad range of places and populations to pick from. This income data will be used to further analyze real estate purchasing decisions.

Conclusion

I would recommend that we start looking for real estate in the 11 regions in cluster number 4. Once possible real estate opportunities are identified and the asking price is known, we can use the economic data here to identify the best price deals.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Borough** | **Total Population** | **Total Income per Cap** | **Net Income per Cap** | **Taxable Income per Cap** | **Cluster Labels** | **1st Most Common Venue** | **2nd Most Common Venue** | **3rd Most Common Venue** | **4th Most Common Venue** | **5th Most Common Venue** | **6th Most Common Venue** | **7th Most Common Venue** | **8th Most Common Venue** | **9th Most Common Venue** | **10th Most Common Venue** |
| North York | 26010 | 44054.98 | 40817.49 | 39324.14 | 3 | Park | Food & Drink Shop | Yoga Studio | Dance Studio | Eastern European Restaurant | Drugstore | Donut Shop | Dog Run | Distribution Center | Discount Store |
| North York | 22570 | 54777.94 | 50934.51 | 49339.3 | 3 | Park | Japanese Restaurant | Pub | Metro Station | Yoga Studio | Dance Studio | Donut Shop | Dog Run | Distribution Center | Discount Store |
| York | 30200 | 34659.74 | 32632.25 | 30979.5 | 3 | Park | Women's Store | Pool | Yoga Studio | Curling Ice | Drugstore | Donut Shop | Dog Run | Distribution Center | Discount Store |
| East York | 26970 | 50671.97 | 46201.45 | 44701.63 | 3 | Park | Convenience Store | Metro Station | Yoga Studio | Dance Studio | Drugstore | Donut Shop | Dog Run | Distribution Center | Discount Store |
| North York | 16150 | 34011.76 | 31900.06 | 30049.97 | 3 | Park | Construction & Landscaping | Bakery | Yoga Studio | Deli / Bodega | Eastern European Restaurant | Drugstore | Donut Shop | Dog Run | Distribution Center |
| North York | 28490 | 40288.1 | 37400.14 | 35991.65 | 3 | Park | Yoga Studio | Dance Studio | Eastern European Restaurant | Drugstore | Donut Shop | Dog Run | Distribution Center | Discount Store | Diner |
| Central Toronto | 12020 | 211828.8 | 199085.5 | 192124.7 | 3 | Park | Swim School | Bus Line | Yoga Studio | Dance Studio | Drugstore | Donut Shop | Dog Run | Distribution Center | Discount Store |
| York | 19260 | 33422.27 | 31354.36 | 29138.42 | 3 | Park | Convenience Store | Yoga Studio | Dance Studio | Eastern European Restaurant | Drugstore | Donut Shop | Dog Run | Distribution Center | Discount Store |
| North York | 6520 | 133867.6 | 125571.9 | 122879.3 | 3 | Park | Convenience Store | Yoga Studio | Dance Studio | Eastern European Restaurant | Drugstore | Donut Shop | Dog Run | Distribution Center | Discount Store |
| Scarborough | 46190 | 28185.36 | 26218.77 | 24482.46 | 3 | Playground | Park | Yoga Studio | Curling Ice | Drugstore | Donut Shop | Dog Run | Distribution Center | Discount Store | Diner |
| Downtown Toronto | 11600 | 202622.1 | 190372.7 | 184430.5 | 3 | Park | Playground | Trail | Yoga Studio | Curling Ice | Drugstore | Donut Shop | Dog Run | Distribution Center | Discount Store |