My Capstone Project: Neighborhood Data Analysis of Alberta(AB), Canada

1. Introduction

1.1 Background

Alberta is a province of Canada. With an estimated population of 4,067,175 people as of the 2016 census, it is Canada's fourth-most populous province and the most populous of Canada's three prairie provinces. Its area is about 660,000 square kilometers (250,000 sq mi). Alberta and Saskatchewan were formerly districts of the Northwest Territories until they were established as provinces on September 1, 1905.

Tourist destinations in the province include Banff, Canmore, Drumheller, Jasper, Sylvan Lake, and Lake Louise.[1]

Alberta is a city with a high population and population density. From the Real Estate investor’s point of view, we want to invest in such a business where the competition is moderate and footfalls will be high. Keeping the above things in mind it is very difficult for an individual to find such a place in such a big city and gather this much information.

1.2 Problem

The purpose of this Project is to help people in exploring better facilities around their neighborhood. It will help people making smart and efficient decision on selecting great neighborhood out of numbers of other neighborhoods in Scarborough, Toranto.

Lots of people are migrating to various states of Canada and needed lots of research for good housing prices and reputated schools for their children. This project is for those people who are looking for better neighborhoods. For ease of accessing to Cafe, School, Super market, medical shops, grocery shops, mall, theatre, hospital, like minded people, etc.

This Project aim to create an analysis of features for a people migrating to Scarborough to search a best neighborhood as a comparative analysis between neighborhoods. The features include median housing price and better school according to ratings, crime rates of that particular area, road connectivity, weather conditions, good management for emergency, water resources both freash and waste water and excrement conveyed in sewers and recreational facilities.

It will help people to get awareness of the area and neighborhood before moving to a new city, state, country or place for their work or to start a new fresh life.

1.3 Interest

I believe this is a relevant challenge with valid questions for anyone who wants to set up his/her business. The same methodology can be applied in accordance with demands as applicable. This case is also applicable to anyone interested in exploring starting or locating a new business in any city. Lastly, it can also serve as a good practical exercise to develop Data Science skills.

2. Data Section

2.1 Data sources

For the Alberta neighborhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighborhoods in Alberta. We'll be required to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas dataframe so that it is in a structured format.

- List of postal codes of Canada: T with their geographic co-ordinates (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T)[2]

2.2 Description of the Data

The dataframe will consist of five columns:

- PostalCode
- Borough
- Neighborhood
- Latitude
- Longitude

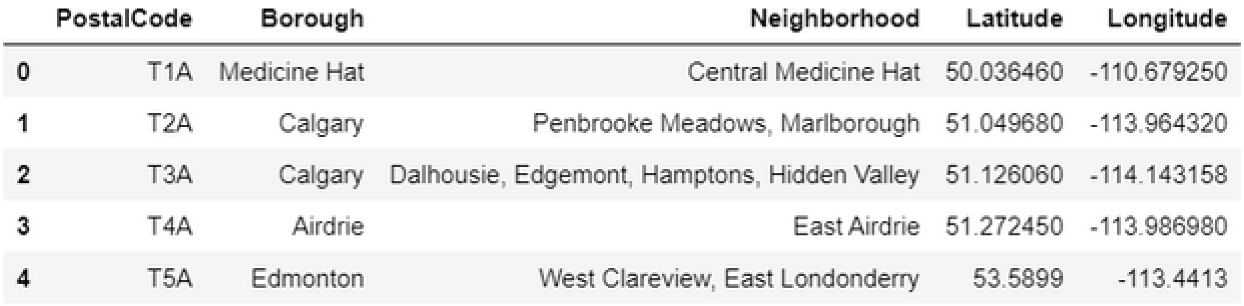
- Foursquare API: This project would use Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.[3]
- Work Flow: Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

2.3 Libraries used

- Pandas: For creating and manipulating dataframes.
- Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.
- Scikit Learn: For importing k-means clustering.
- JSON: Library to handle JSON files.
- XML: To separate data from presentation and XML stores data in plain text format.
- Geocoder: To retrieve Location Data.
- Beautiful Soup and Requests: To scrap and library to handle http requests.
- Matplotlib: Python Plotting Module.

2.4 Dataframes

- Dataframe of Postal Codes, with theri co-ordinates



	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	T1A	Medicine Hat	Central Medicine Hat	50.036460	-110.679250
1	T2A	Calgary	Penbrooke Meadows, Marlborough	51.049680	-113.964320
2	T3A	Calgary	Dalhousie, Edgemont, Hamptons, Hidden Valley	51.126060	-114.143158
3	T4A	Airdrie	East Airdrie	51.272450	-113.986980
4	T5A	Edmonton	West Clareview, East Londonderry	53.5899	-113.4413

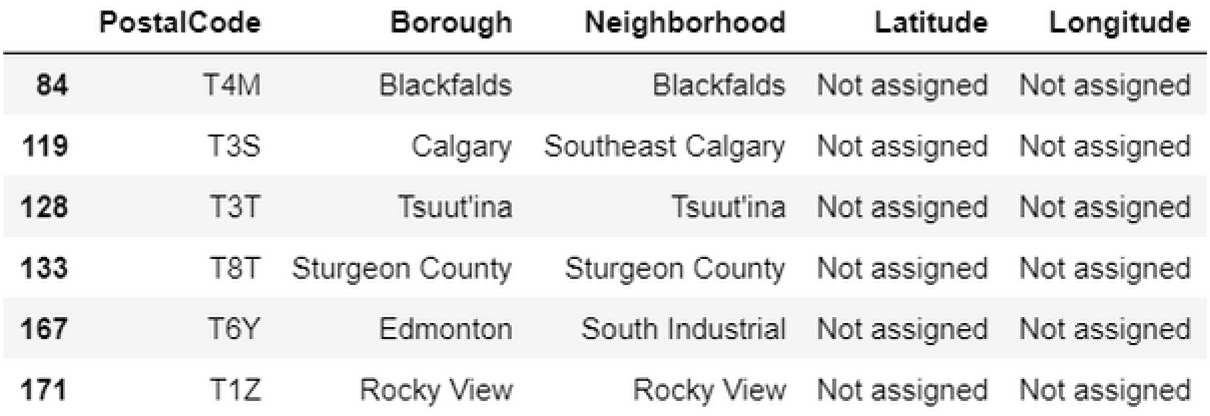
3.Methodology

3.1 Data Preprocessing

The dataframe contains uncleaned data for now. We can see that some values are valued as **Not assigned**. We process this dataframe by applying the below steps.

- Ignore cells with a borough that is **Not assigned**.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.
- More than one neighborhood can exist in one postal code area.

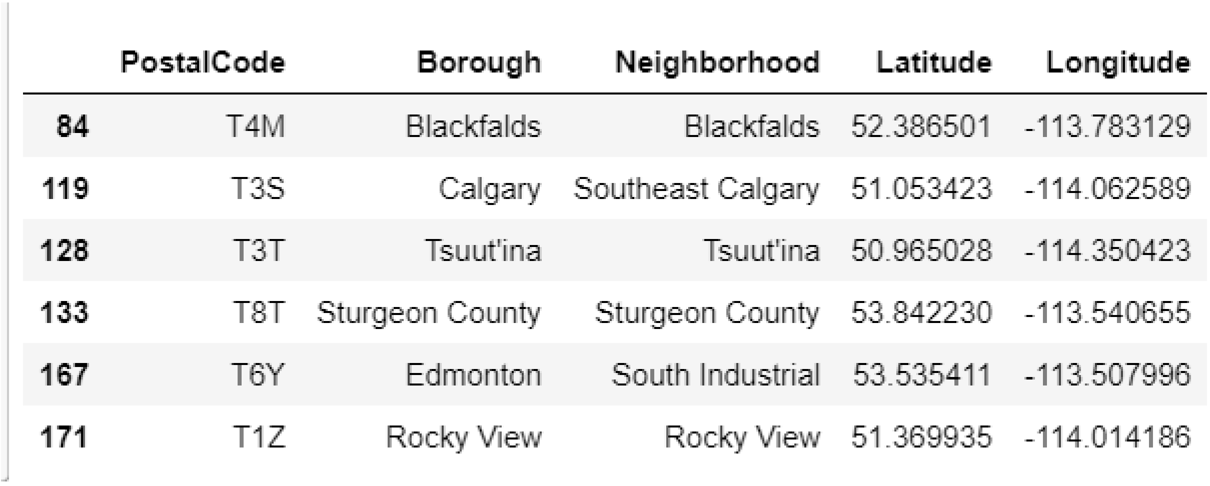
Also there are some values mising for co-ordinates of some postal codes.



	PostalCode	Borough	Neighborhood	Latitude	Longitude
84	T4M	Blackfalds	Blackfalds	Not assigned	Not assigned
119	T3S	Calgary	Southeast Calgary	Not assigned	Not assigned
128	T3T	Tsuut'ina	Tsuut'ina	Not assigned	Not assigned
133	T8T	Sturgeon County	Sturgeon County	Not assigned	Not assigned
167	T6Y	Edmonton	South Industrial	Not assigned	Not assigned
171	T1Z	Rocky View	Rocky View	Not assigned	Not assigned

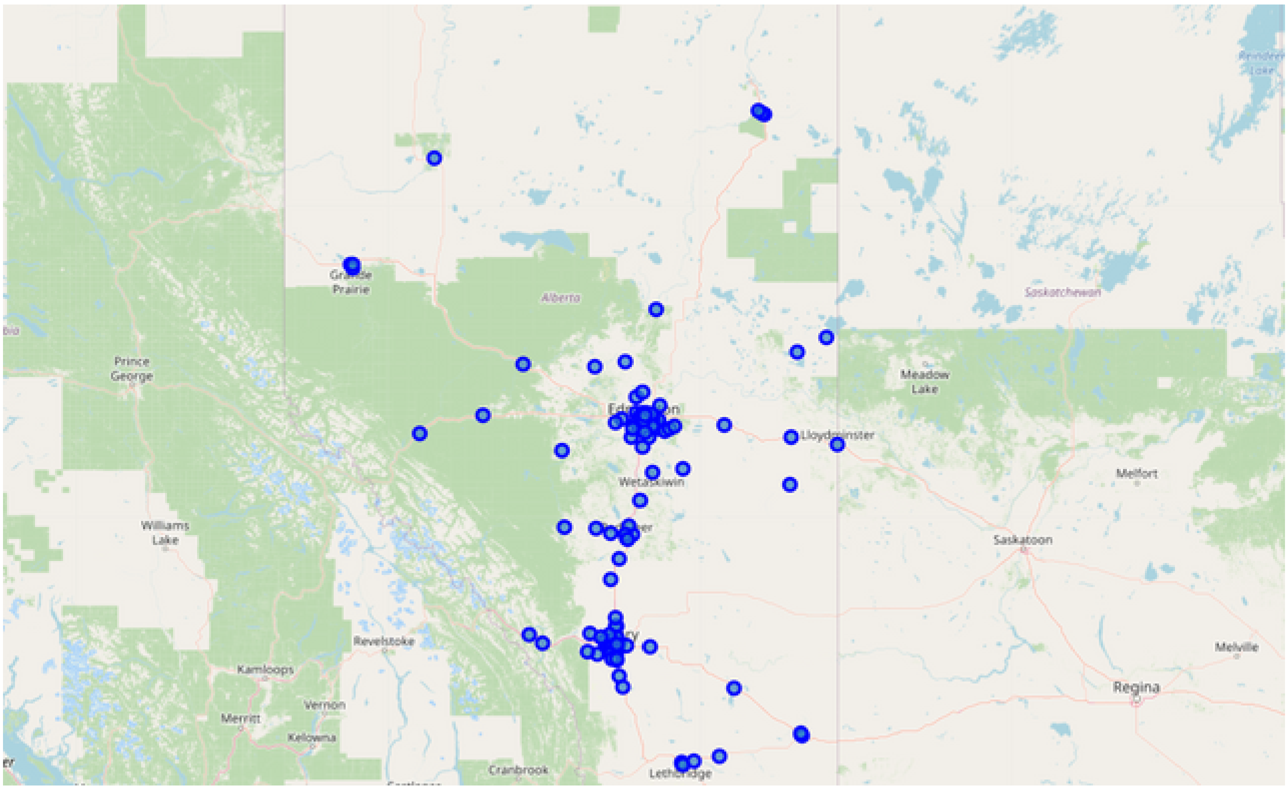
For this I'm using geocoder python packge to get the latitude and longitude of all the Borough. We’ll iterate through the spliced data frame and then we will join this dataframe with the original dataframe.

Also we should change the datatypes of **Latitude** and **Longitude** columns to *Float*



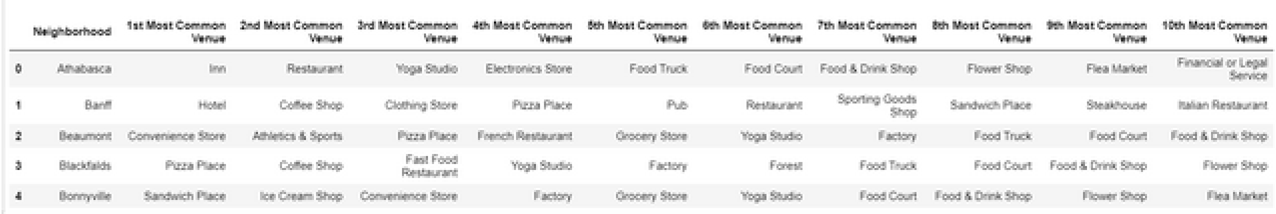
	PostalCode	Borough	Neighborhood	Latitude	Longitude
84	T4M	Blackfalds	Blackfalds	52.386501	-113.783129
119	T3S	Calgary	Southeast Calgary	51.053423	-114.062589
128	T3T	Tsuut'ina	Tsuut'ina	50.965028	-114.350423
133	T8T	Sturgeon County	Sturgeon County	53.842230	-113.540655
167	T6Y	Edmonton	South Industrial	53.535411	-113.507996
171	T1Z	Rocky View	Rocky View	51.369935	-114.014186

I used python folium library to visualize geographic details of Alberta and its boroughs and I created a map of Alberta with boroughs superimposed on top. I used latitude and longitude values to get the visual as below:



I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as 100 venue and the radius 500 meter for each borough from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

Finally by using the Foursquare API in conjunction with the created datasets, a table of most common visited venues in Alberta neighborhoods is generated.

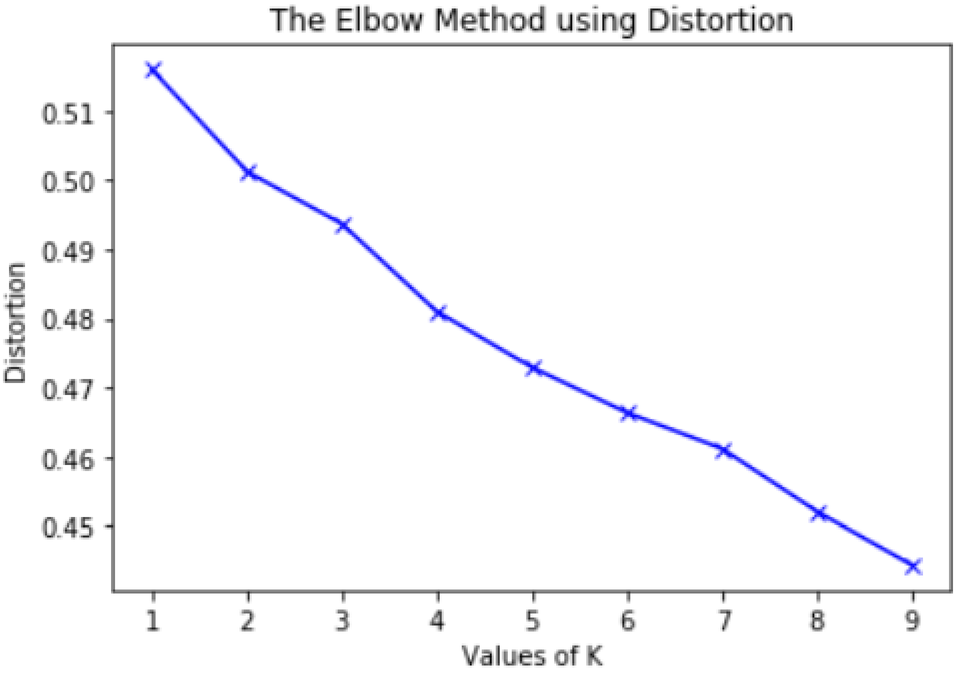


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Athabasca	Inn	Restaurant	Yoga Studio	Electronics Store	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market	Financial or Legal Service
1	Banff	Hotel	Coffee Shop	Clothing Store	Pizza Place	Pub	Restaurant	Sporting Goods Shop	Sandwich Place	Steakhouse	Italian Restaurant
2	Beaumont	Convenience Store	Athletics & Sports	Pizza Place	French Restaurant	Grocery Store	Yoga Studio	Factory	Food Truck	Food Court	Food & Drink Shop
3	Blackfalds	Pizza Place	Coffee Shop	Fast Food Restaurant	Yoga Studio	Factory	Forest	Food Truck	Food Court	Food & Drink Shop	Flower Shop
4	Bonnyville	Sandwich Place	Ice Cream Shop	Convenience Store	Factory	Grocery Store	Yoga Studio	Food Court	Food & Drink Shop	Flower Shop	Flea Market

3.2 Machine Learning

We have some common venue categories in boroughs. In this reason I used unsupervised learning K-means algorithm to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning.

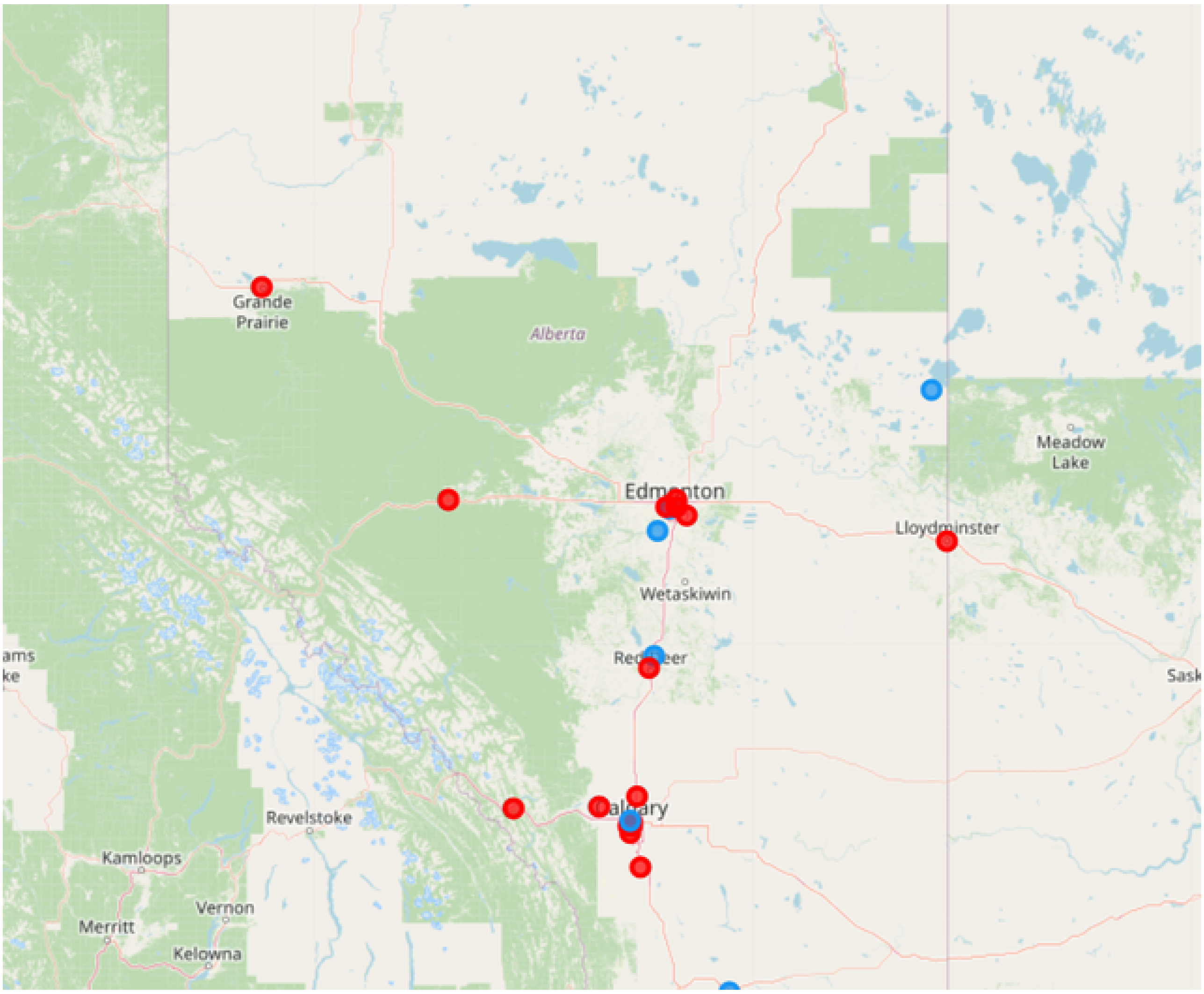
First, I will run K-Means to cluster the boroughs into 6 clusters because when I analyze the K-Means with elbow method it ensured me the 6 degree for optimum k of the K-Means.



4.Visualisation

First I visualize the cluster and you can see the clustered map below:

- Red(cluster 0)
- Blue(cluster 1)



5.Result

Most of the restro are concentrated in the Edmonton & Calgary neighborhood of Alberta province of Canada, with the highest number in cluster 0. On the other hand, cluster 1 has only one resto in the neighborhoods. This represents a great opportunity and high potential areas to open new Fast Food Restro as there is very little to no competition from existing restro. Meanwhile, restro in cluster 0 are likely suffering from intense competition due to oversupply and high concentration.

6.Discussion section

This also shows that the oversupply of restro mostly happened in the central area of the city, with the suburb area still have very few fast food restro. Therefore, this project recommends property developers to capitalize on these findings to open new restro in neighborhoods in cluster 1 with little to no competition.

Lastly, property developers are advised to avoid neighborhoods in cluster 0 which already have high concentration of restro and suffering from intense competition.

7.Conclusion

As people are turning to big cities to start a business or work. For this reason, people can easily interpret where to start a new restro. Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.

A.References

1.https://en.wikipedia.org/wiki/Alberta (https://en.wikipedia.org/wiki/Alberta)

2.https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T)

3.https://developer.foursquare.com/ (https://developer.foursquare.com/)