Capstone Project – Calgary Chinese Restaurant location selection

Introduction: Business Problem

Calgary is a city in the western Canadian province of Alberta, with a population of 1,285,711 in 2019, making it Alberta's largest city. It is the third-largest municipality in Canada (after Toronto and Montreal), and the largest in western Canada. Calgary has also always been one of the favorite cities for immigration from all over the world. Especially in these years, there are more and more Chinese immigrants moving to Calgary as their home. In 2016, 8.3% of the population in Calgary are Chinese immigrants.

As you can see from the numbers, with that many Chinese immigrants in Calgary, opening a Chinese restaurant in the right neighborhood in Calgary could be a good business. There are some very good Chinese restaurants with great success in Calgary already. As an investor who is looking to open a new Chinese restaurant, choosing the right location or neighborhood is essential. When we think of it by the investor, we expect from them to prefer the neighborhoods where there is relatively lower real estate cost, higher population and lower density of restaurants. We are also particularly interested in areas with no Chinese restaurants in vicinity. We would also prefer locations as close to city center as possible.

By utilizing our Data Science power, we could potentially help the investors/stakeholders to find some promising locations for opening a Chinese restaurant, so that they are open too more customers and facing less competition.

Data

In order to solve this business problem, we can search for the valuable data from multiple sources, then use our technical data analysis skills to process and clean the data, make them ready for deeper analysis during the later stage:

1. I searched "List of neighborhoods in Calgary" on Wikipedia, 'https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary'. In the Pycharm IDE, I was able to use "html" python library to read the table off the webpage into "pandas" data frame for processing and cleaning.
2. I collected some information about "median property price" in each neighborhood in Calgary from 'https://www.avenuecalgary.com/calgarys-best-neighbourhoods-2019/the-full-list/' webpage by using python library "BeautifulSoup" and "pandas".
3. By using "geopy" library in python, I was able to find all the latitudes and longitudes coordinates for each given neighborhood including the Calgary City Center.
4. After long period of data processing and cleaning, I was able to make a excel data file called "Final_data.xlsx", which will be used in my "Jupyter Notebook" for my Capstone project.
5. In my project, in order to accurately calculate distances on the map, I need to create grid of locations in Cartesian 2D coordinate system which allows me to calculate distances in meters

(not in latitude/longitude degrees). Then I used “pyproj” library to get the UTM Cartesian coordinates “X” and “Y” for each neighborhood.

6. I used my “Foursquare API” to get the restaurant venues and Chinese restaurant venues in each given neighborhood in Calgary.
7. For data visualization, I used python “folium’ library. Also, in order to create Calgary Choropleth Map, I downloaded the Calgary Neighborhoods Boundaries Geojson File from the Calgary City webpage ‘https://data.calgary.ca/Base-Maps/Community-Boundaries/ab7m-fwn6’.

Methodology

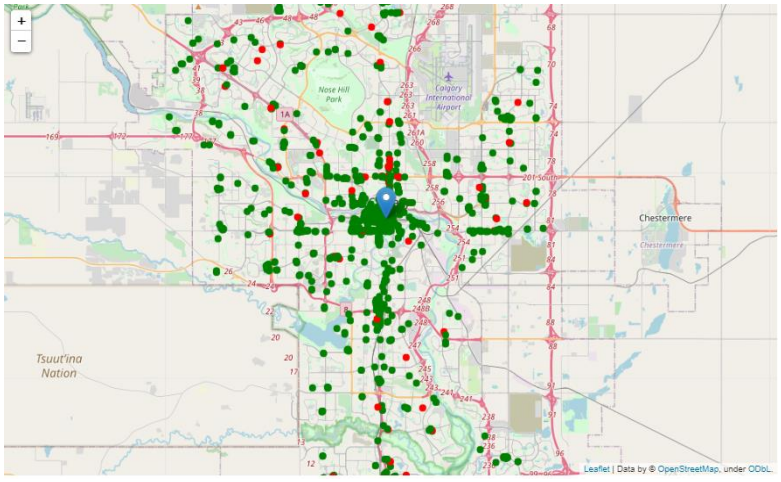
In this project we will direct our efforts on detecting areas of Calgary that have low restaurant density, particularly those with low number of Chinese restaurants. Also, we are looking for areas that have high population density with low property value (low rental fee). We can solve this business problem in three steps:

In the first step, I collected all the required data from different sources, then I used python’s libraries like pandas, numpy, geopy, requests, html and pyproj to build the original data excel file called “Final_data.xlsx”. Then, after some data cleaning and preparing, I was able to build a data frame called “df_calgary” for further analysis. This data frame contains all the Calgary communities with their population, area(km2), population density, latitude, longitude, property median price (CAD), UTM Cartesian coordinates and calculated distance to Calgary city center(m).

```
df_calgary.head()
```

	Name	population	Area	Populationdensity	Latitude	Longitude	Median_price	X	Y	Distance_to_center
0	ABBEYDALE	5917	1.7	3480.6	51.058836	-113.929413	305000	715167.687248	5.660854e+06	9708.906971
1	ACADIA	10705	3.9	2744.9	50.968655	-114.055587	385000	706728.683961	5.650467e+06	8426.114943
2	ALBERT PARK/RADISSON HEIGHTS	6234	2.5	2493.6	51.044845	-113.990195	318000	710972.897862	5.659123e+06	5308.353624
3	ALTADORE	9116	2.9	3143.4	51.015104	-114.100756	738000	703354.667651	5.655505e+06	4050.413819
4	APPLEWOOD PARK	6498	1.6	4061.3	51.044658	-113.928931	362000	715267.254465	5.659279e+06	9605.117760

Then, I used Foursquare API to get info on regular restaurants and Chinese restaurants in each neighborhood. By utilizing python’s folium library, I was able to show these restaurants on the map (green dots are the regular restaurants, red dots are the Chinese restaurants):



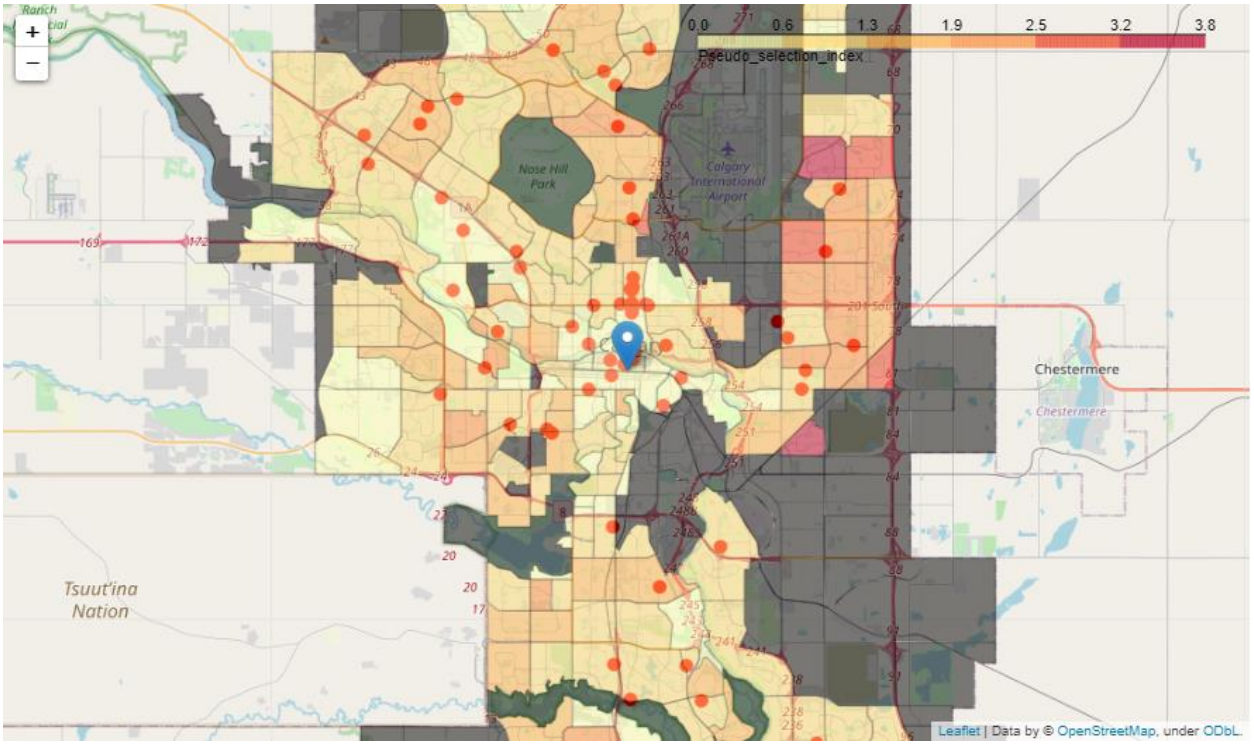
In the second step, my goal was to find the “region of interest”, a specific area that is generally a better choice for opening a regular restaurant.

I used “MinMaxScalor” method to normalize some data features in the data frame, and then I used these normalized features to create a “Pseudo_Selection_Index “= A1 * 'Population_density' / (A2 * 'Location_restaurants_number' + A3 * 'Property_Median_value' + A4 * 'distance_to_center'). Due to the lack of restaurants’ revenues information, I was not able to build a multi-linear regression model. In order to continue the analysis, I chose some reasonable values for the weight parameters for each feature: A1 for Populationdensity is 1, A2 for Property_Median_price is 0.2, A3 for Distance_to_center is 0.2, A4 for Restaurants_in_area is 1. A new normalized data frame called “df_new” was built.

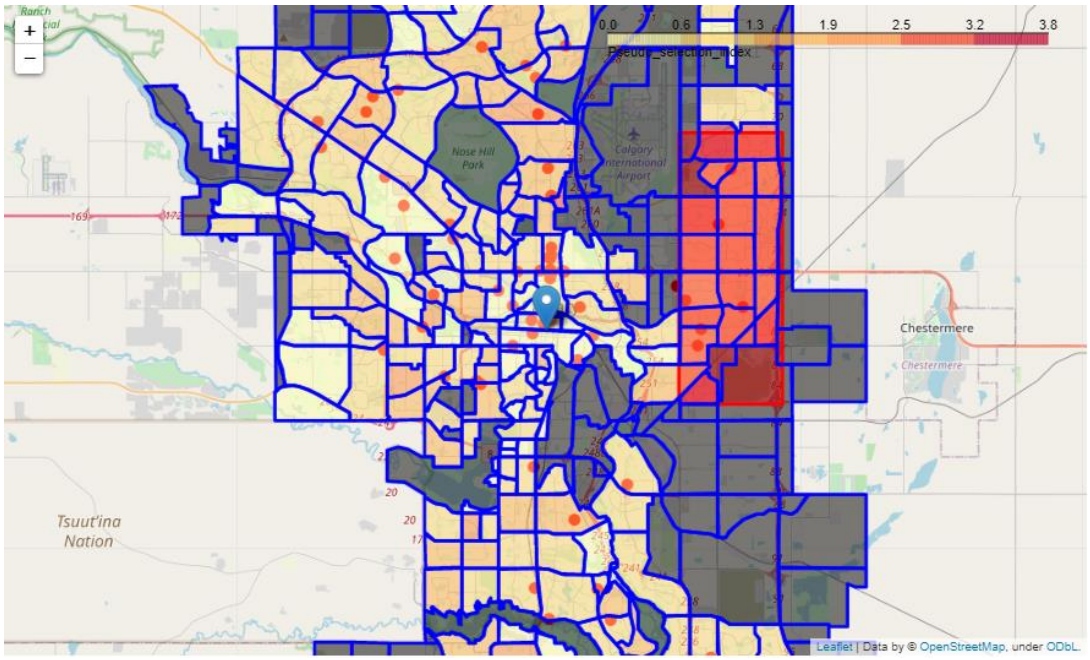
```
df_new.head()
```

	Populationdensity	Median_price	Distance_to_center	Restaurants_in_area	Name	Pseudo_selection_index
59	0.395618	0.167084	0.355554	0.000000	ERIN WOODS	3.784822
97	0.467843	0.188986	0.546357	0.000000	MARTINDALE	3.181120
168	0.539251	0.204005	0.592577	0.020619	TARADALE	2.996924
70	0.360712	0.108260	0.271922	0.061856	GREENVIEW	2.615901
183	0.419209	0.192741	0.405602	0.041237	WHITEHORN	2.605312

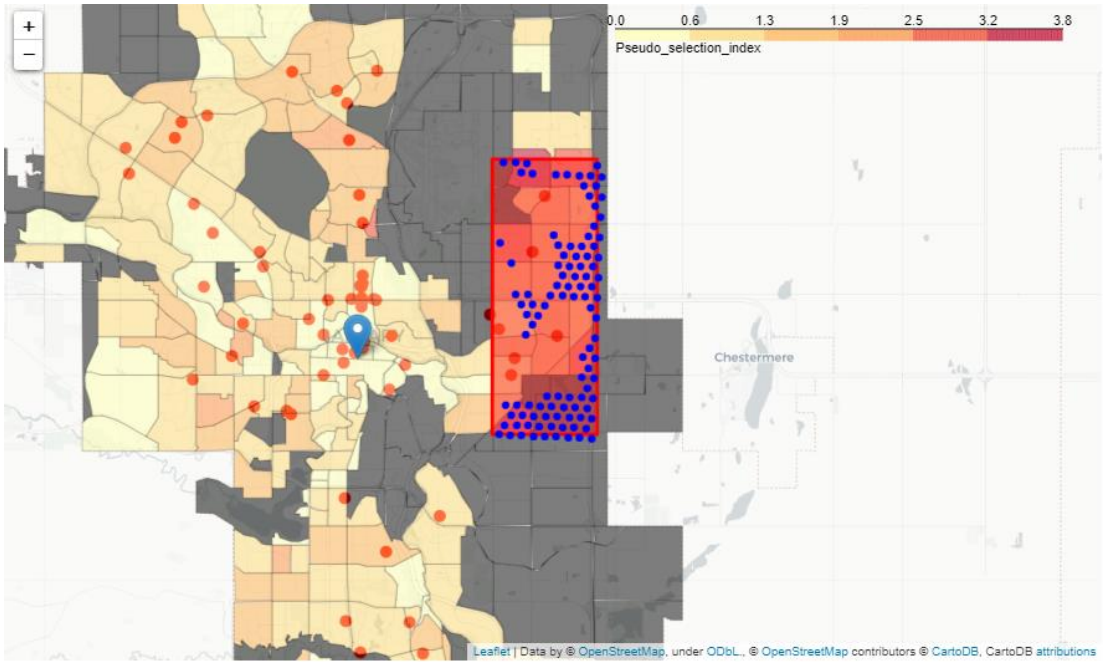
In order to plot the Chorolepth map of Calgary to show the value of 'Pseudo_selection_index' for each neighborhood, the Calgary Neighborhoods boundaries geojson file was downloaded from the "City of Calgary" Webpage, and by using this geojson file, I was able to build the wanted Chorolepth map to show the value of 'Pseudo_selection_index' for each neighborhood with the chinese restaurants marked as red dots.



In theory, the higher the “Pseudo_selection_index” value is, the better choice the Neighborhood is. As you can see from the map, the North-East part of Calgary is showing deeper red color on the map, which indicating those neighborhoods have higher “Pseudo_selection_index”, which means, in theory, they are better choices for opening a regular restaurant. Now, I have my “region of interest” (marked in the red rectangle):



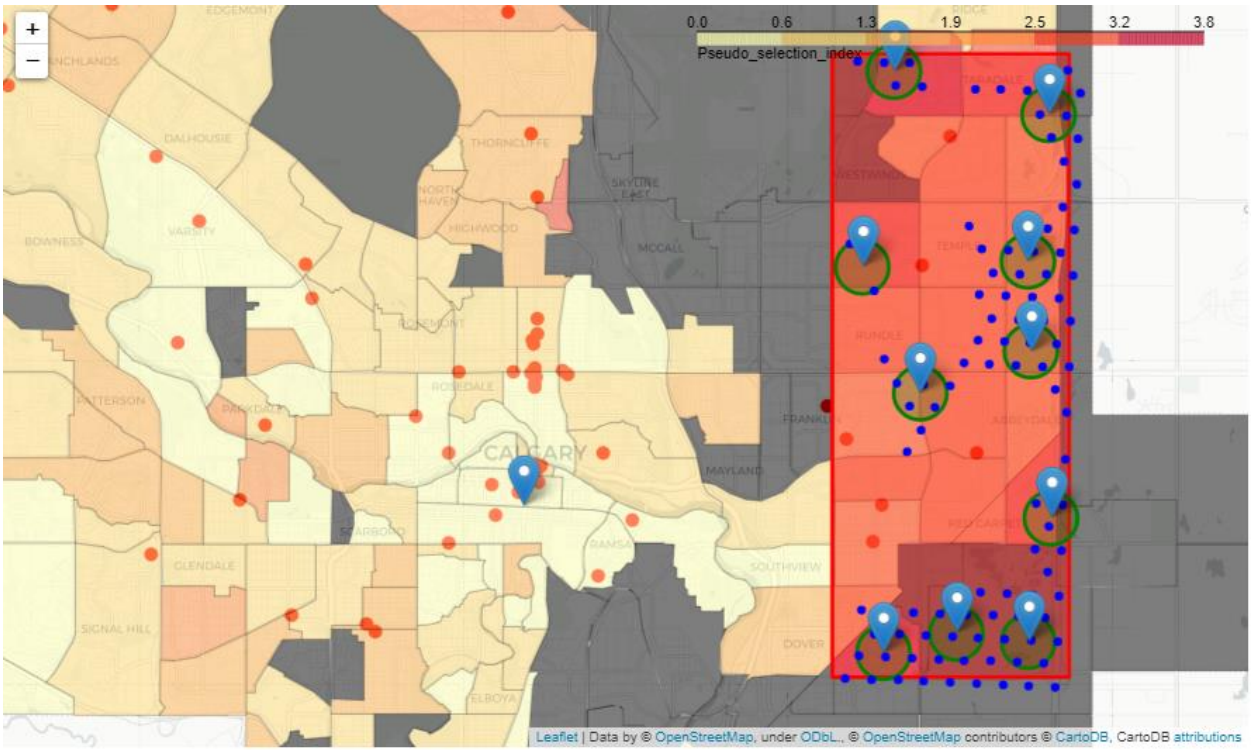
In the third step, we need to find the more detailed locations with no more than 1 restaurant around and especially no Chinese restaurant around within 1km to those locations.



Then, I used K-means Clustering method to cluster those locations to create 10 centers of 10 area containing all the good locations. These 10 centers and corresponding addresses will be the final result of our analysis.

	Cluster_Center_Longitude	Cluster_Center_Latitude	Location_Address
0	-113.968260	51.019234	Erin Meadow Close SE, Erin Woods, Calgary, Alb...
1	-113.929108	51.085908	Anaheim Place NE, Monterey Park, Calgary, Albe...
2	-113.922863	51.042168	Rotary Mattamy Greenway, Applewood Park, Calga...
3	-113.923510	51.110835	GW-189, East Calgary Greenway, Coral Springs, ...
4	-113.974029	51.084889	Whitlock Place NE, Whitehorn, Calgary, Alberta...
5	-113.965422	51.118163	Martin Crossing Drive NE, Martindale, Calgary,...
6	-113.928305	51.070620	California Boulevard NE, Monterey Park, Calgar...
7	-113.929033	51.021117	Rotary Mattamy Greenway, Applewood Park, Calga...
8	-113.958298	51.063448	Tim Hortons, 52 Street NE, Marlborough Park, C...
9	-113.948474	51.022504	Forest Lawn Industrial, Calgary, Alberta, T2B ...

These 10 locations/zones are showed on the map:



Results and Discussion

	Cluster_Center_Longitude	Cluster_Center_Latitude	Location_Address
0	-113.968260	51.019234	Erin Meadow Close SE, Erin Woods, Calgary, Alb...
1	-113.929108	51.085908	Anaheim Place NE, Monterey Park, Calgary, Albe...
2	-113.922863	51.042168	Rotary Mattamy Greenway, Applewood Park, Calga...
3	-113.923510	51.110835	GW-189, East Calgary Greenway, Coral Springs, ...
4	-113.974029	51.084889	Whitlock Place NE, Whitehorn, Calgary, Alberta...
5	-113.965422	51.118163	Martin Crossing Drive NE, Martindale, Calgary,...
6	-113.928305	51.070620	California Boulevard NE, Monterey Park, Calgar...
7	-113.929033	51.021117	Rotary Mattamy Greenway, Applewood Park, Calga...
8	-113.958298	51.063448	Tim Hortons, 52 Street NE, Marlborough Park, C...
9	-113.948474	51.022504	Forest Lawn Industrial, Calgary, Alberta, T2B ...

These locations have the following features: good population density; cheaper property value which means less rent; only 1 existing restaurant and no existing Chinese restaurants within 1km. Stakeholders should be able to use these 10 locations as a reference, with further investigation or personal preference, they should be able to find their optimum location for opening a Chinese restaurant.

Furthermore, a potential regression model could be built based on the dataset if we can find more data of annual revenues of some Calgary restaurants as our target variable, take other features in the data set as independent variables, then we could use Linear or non-linear regression to train the model to finalize the weight parameters for each independent variables. Due to the lack of information about the revenues of these restaurants, I can only give some reasonable random number for the weight parameters of 'Population_density', 'Location_restaurants_number', 'Property_Median_value' and 'distance_to_center' when I calculate the "Pseudo_Selection_Index". This part of the analysis could be optimized with more data on restaurants' revenues.

Conclusion

My analysis consists of three main steps: preparing and processing the data; finding the region of interest by utilizing the "Pseudo_Selection_Index"; further analysis with focus on region of interest. After these three steps, we are able to find these 10 locations/zones showing above in the data frame. Result of all this is 10 zones containing largest number of potential new restaurant locations based on number of and distance to existing venues - both restaurants in general and Chinese restaurants particularly. Meanwhile, the communities with good population density and cheaper property value (cheaper rent) were also taken into the consideration. Although I could build a more accurate model with more information, the stakeholder/Chinese restaurants investors should be able to use these 10 locations as a reference for further investigation in order to make a final decision.