Market Segmentation Analysis using Machine Learning and Geo-Location data

Identifying the best location to open a new restaurant
Calgary, Alberta, Canada

By:

Mahmood Khordoo,

March 2019

1.Introduction

Manual site selection for business owners is still a very difficult problem. Finding the right location involves trying to quantify lots of variables that are hard to quantify and measure manually. In this project, we will try to subvert these obstacles by combining powerful data from multiple sources such as population, business license renewal, FourSquare location API with different machine learning techniques to model the problem and try to answer these question. The target city of the analysis is the City of Calgary located in Alberta, Canada. We believe that this approach can be used to make better decisions about where to open a new restaurant.

1.1 Bussiness Problem

As mentioned before, the goal here is to help a new business owner to find the best location for its restaurant in the City of Calgary. There are several factors that could influence business success. Our focus in this study is to answer the following questions as they could potentially help a business to succeed :

- Where should a new restaurant be located?
- What type of restaurant would be best in a given location?
- How will competition down the block or across town impact market share?

To summarize, We will use multiple data sources with different machine learning models (in particular K-mean clustering) to study the complex, dynamic, and often unobserved factors that would help to identify the best location for a new restaurant in the city.

1.2 City Info

Calgary, a cosmopolitan Alberta city with numerous skyscrapers, owes its rapid growth to its status as the center of Canada's oil industry. However, it's still steeped in the western culture that earned it the nickname "Cowtown," evident in the Calgary Stampede. It has a population of 1.2 million people. (2016 census).
Before we get the data and start exploring it, let's download all the dependencies that we will need.

2. Data Source

To build a powerful machine learning model, we needed powerful data. We will use the FourSquare API location data along with the different data sources from the web to identify the market segmentation in each part of the city. Understanding of restaurant similarity and local food preferences are amongst the factors that we consider to be important in our site selection hypotheticals.

2.1 Main data sources

- Neighbourhood boundaries in GeoJSON format(City of Calgary Open Data)
- Calgary Business Licenses Renewal (source)
- Calgary Neighourboohd Crime data (source)

- Calgary Census data-count of dwelling units and population (source)
- Neighourboohd Groups based on the Postal Codes (source)
- Foursquare Location API (source)

We will use these data along with the explore function of the FourSquare API to get the most common venue categories in each neighborhood, and then we will use this feature to group the neighborhoods into clusters. We can then suggest the location for each business based on the locations with the most similar business. We are also considering the distance to industrial zones, population density, crimes rate and the number of renewed business licenses in our analysis.
We believe this approach can be used not only to make better decisions about where to put restaurants but could also help to identify where to put other types of businesses like general Oil and Gas or Technology offices as well. The details of the approach and analysis will be presented and discussed in the methodology section.

2.2 Libraries and APIs

We will use the location data from the Foursquare Location API to explore neighborhoods in Calgary.
We will also use these library convert addresses into their equivalent latitude and longitude values.
Here is a summary of location data libraries and APIs :

- FourSquare API
- Geopy
- Geopandas

2.2 Downloading the data

In order to explore the different neighborhood in Calgary we use two sources of the data:

- Neighourhoods based on the PostalCode (Fiding Neighbourhood Names)
- Neighbourhoods based on the City classification boundaries (Finding boundaries)

This dataset for neoighbourhood names exists for free. We use the data provided in this http://totallycalgary.ca/loc/ to map the postal code to the neighbourhood name and then will map them to the neighbourhood boundaries.

2.2.1 Downloading the PostaCode page

We use the request library to download a page containing the postal codes and neighbourthoods of Calgary.

1.2 Parsing the HTML Table

Here we use the Beautiful Package to parse the content of the HTML options in the page and extract the required Postal code and their corresponding neighbourhood information.

	postalcode	neighbourhood
0	T1Y	Horizon,Monterey Park,Pineridge,Rundle,Sunridg...
1	T2A	Abbeydale,Albert Park/Radisson Heights,Applewo...
2	T2B	Dover,Erin Woods,South Foothills,Southview,Val...
3	T2C	Quarry Park,Riverbend,Shepard Industrial,Starf...
4	T2E	Bridgeland/Riverside,Calgary International Air...
5	T2G	Alyth/Bonnybrook,Downtown East Village,Golden ...
6	T2H	Burns Industrial,East Fairview Industrial,Fair...
7	T2J	Acadia,Bonavista Downs,Deer Ridge,Deer Run,Dia...

As you can see, the city of Calgary can be divided into 33 Neighbourhoods based o the postal codes.

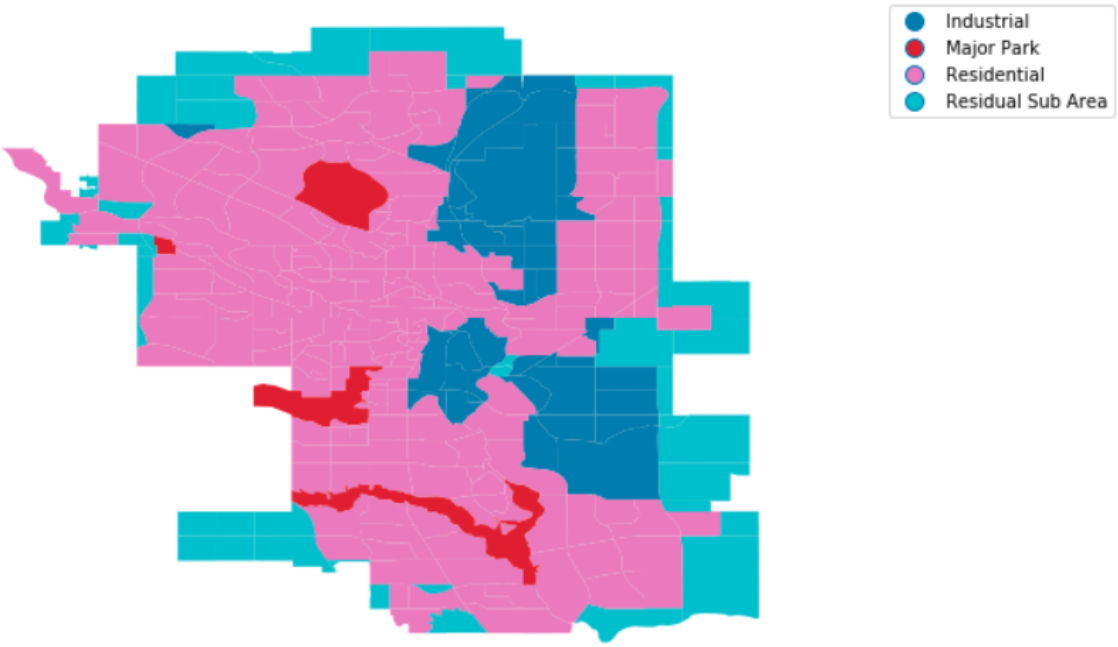
2.2.1 Downloading the Neighbourhood boundaries.

The city of Calgary Open data provides the boundaries of the neighbourhoods along with some information about them in a GeoJSON file. We have already downloaded the file.
Let's load the data using the geopandas library and display a few rows.

Lets see how many zones we have in the City Calgary based on this data:

	class
Residential	212
Residual Sub Area	48
Industrial	42
Major Park	4

As we can see the city is divided into four major zones two Residential zones along with Industrial and Major Parks. In general, we have 260 residential neighborhoods in Calgary.
Let's plot these zones to have a better understanding of their location.



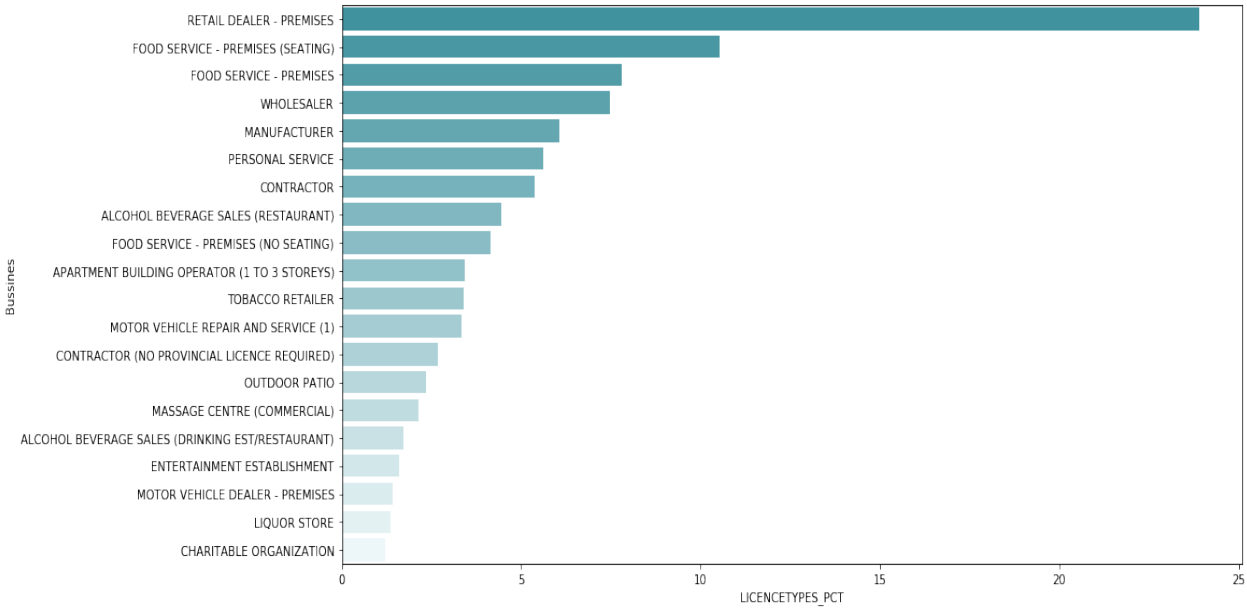
Neighbourdoods Boundaries:
Now, Let's take a look at different neighbourhood boundaries

2.2.3 Commercial License Renewal

Here we are going to load the commercial renewal license data.This would show the main area of the city which has the highest number of businesses. Lets load and take a look at the data.

Lets see how many different type of businesses do we have in the data:

	LICENCETYPES	LICENCETYPES_PCT
RETAIL DEALER - PREMISES	6491	23.899116
FOOD SERVICE - PREMISES (SEATING)	2864	10.544919
FOOD SERVICE - PREMISES	2120	7.805596
WHOLESALER	2032	7.481591
MANUFACTURER	1647	6.064065
PERSONAL SERVICE	1529	5.629602
CONTRACTOR	1463	5.386598
ALCOHOL BEVERAGE SALES (RESTAURANT)	1207	4.444035
FOOD SERVICE - PREMISES (NO SEATING)	1123	4.134757
APARTMENT BUILDING OPERATOR (1 TO 3 STOREYS)	927	3.413108



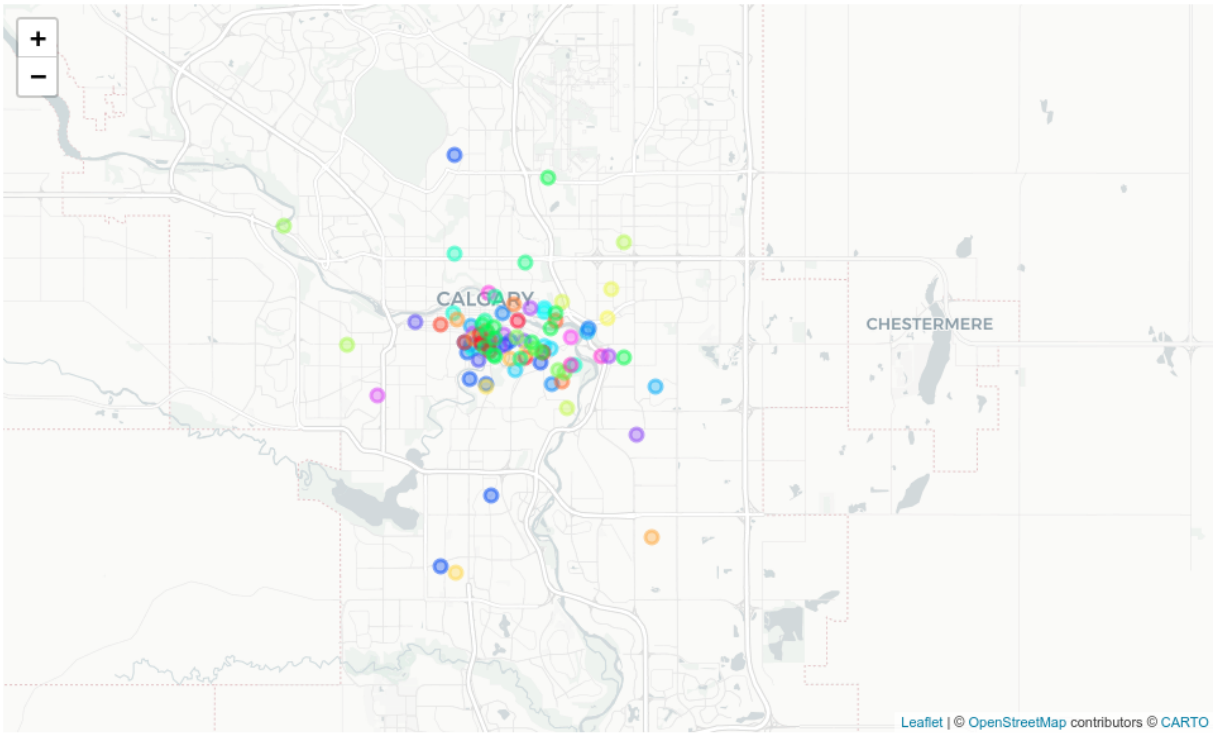
As we can see restaurant business(FOOD SERVICE(SEATING)) is in fact the second most popular business in Calgary.We will use these information to find their clusters and This would greatly help us to identify the target location for our restaurant.

Distribution of the Bussiness
Lets use folium library to plot the distribution of different type of businesses on the City map

	LICENCETYPES	longitude	latitude	Count	LICENCETYPES_CAT
0	ADVERTISER CANVASSER OR DISTRIBUTOR	-114.044790	51.045828	1.0	0
1	MOTOR VEHICLE REPAIR AND SERVICE (MOBILE WASH)	-114.007455	51.041693	1.0	59
2	PERSONAL SERVICE (FITNESS CONDITIONING)	-114.061819	51.024043	1.0	68
3	PERSONAL SERVICE	-114.072682	51.035016	1.0	67
4	PAYDAY LENDER (GRANDFATHERED)	-114.048975	51.039014	1.0	66

In order to plot the data based on their categories we need to convert the license type to numerical values.We use Pandas Categorical utility

Plotting Top 20 Business distributions

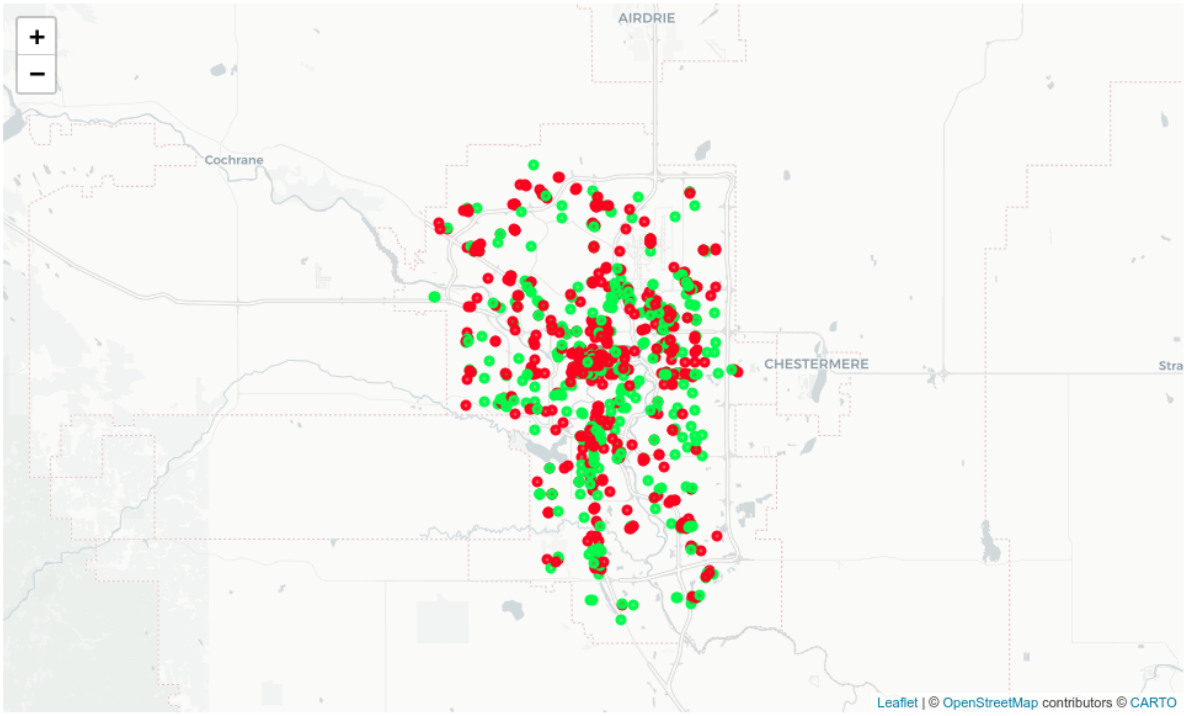


Plotting Food Services

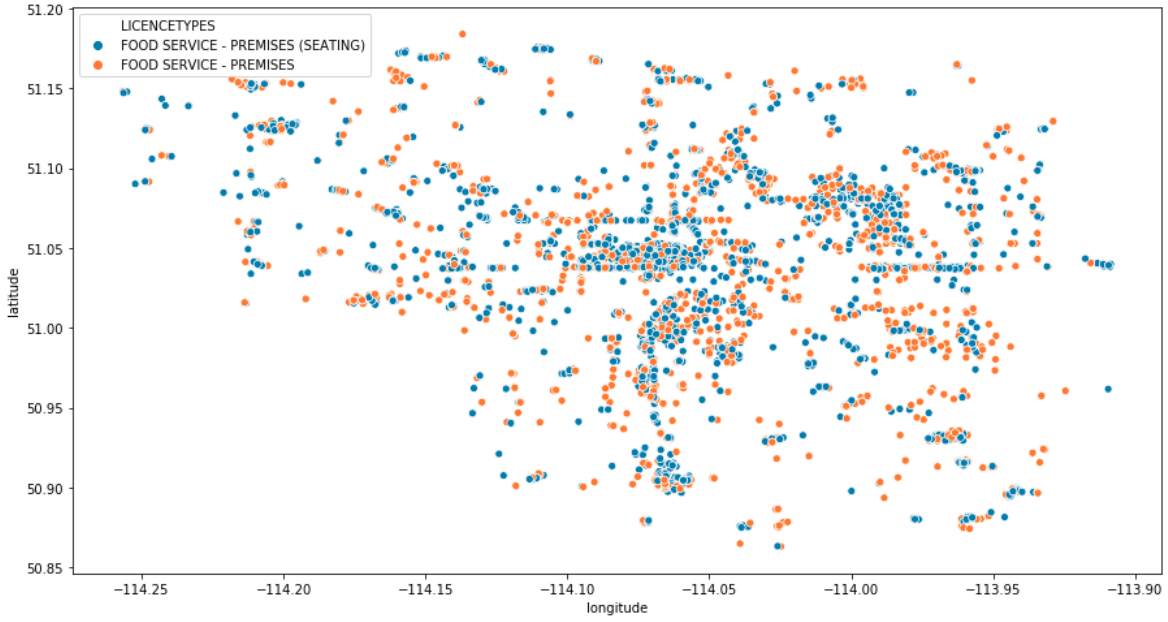
In the next plot we show the mean center for food businesses.This includes the following categories:

- Food Service Premises (Seating)

- Food Service Premises



In the above map the red dots shows the FOOD SERVICE - PREMISES (SEATING) and the green shown the Food Service Premises categories respectively. As you can see we can not find a very good pattern in the data visually.Lets use the K-mean algorithm to find the cluster center for each category:



As you can see the data are highly scattered across the city and it is highly unlikely that the K-Mean algorithm can separate these two classes from each other.

4. Cluster Centers for Food Businesses

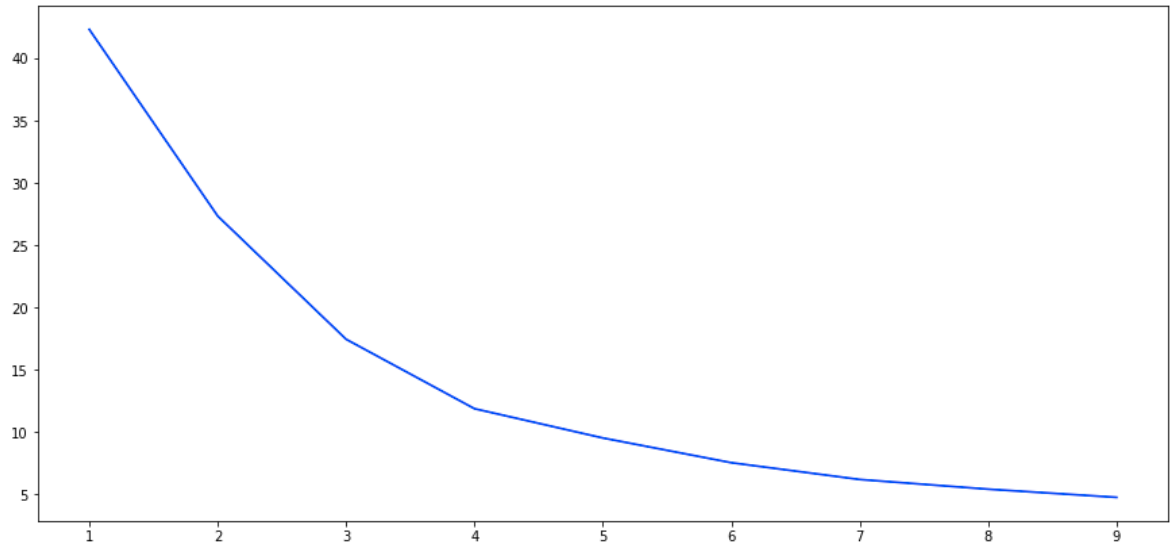
4.1 K Mean clustering

In the above map the red dots shows the FOOD SERVICE - PREMISES (SEATING) and the green shown the Food Service Premises categories respectively. As you can see we can not find a very good pattern in the data visually.Lets use the K-mean algorithm to find the cluster center for each category:

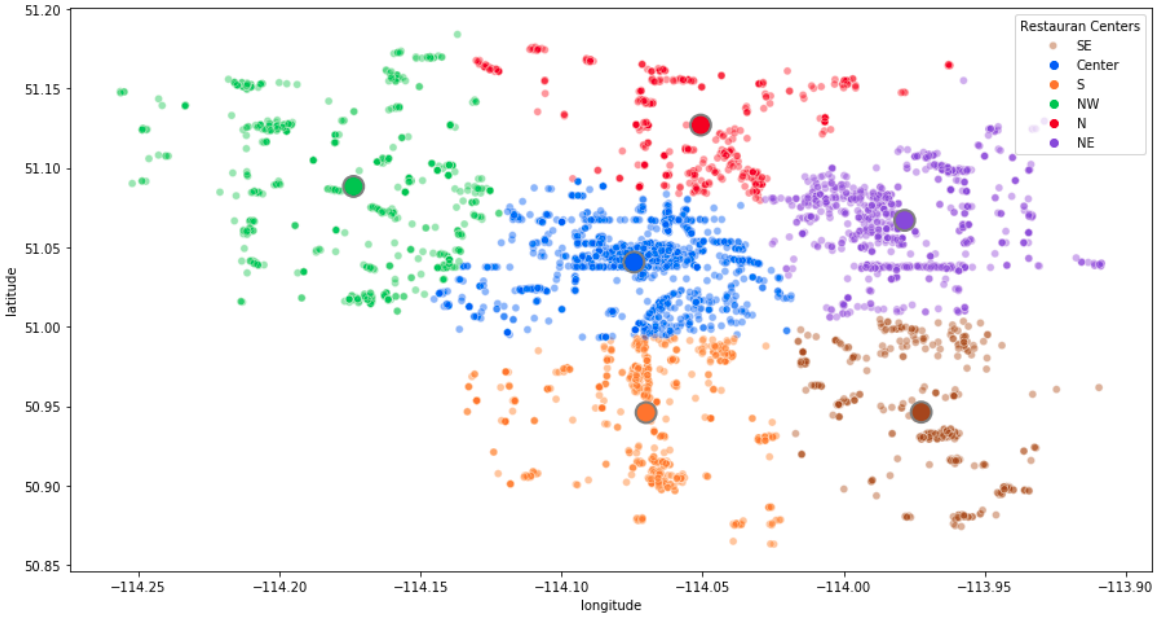
	latitude	longitude
11	50.879272	-113.959724
12	51.037467	-113.980436
14	51.023429	-114.107628
21	51.060466	-114.136802
25	51.066458	-114.063494

4.2 Optimum number of Clusters

We use the elbow method to find the optimum number of clusters in the data. Here we plot the sum of squared distances to the closest centroid for all observations vs the number of clusters.The optimum number of cluster seems to be 5

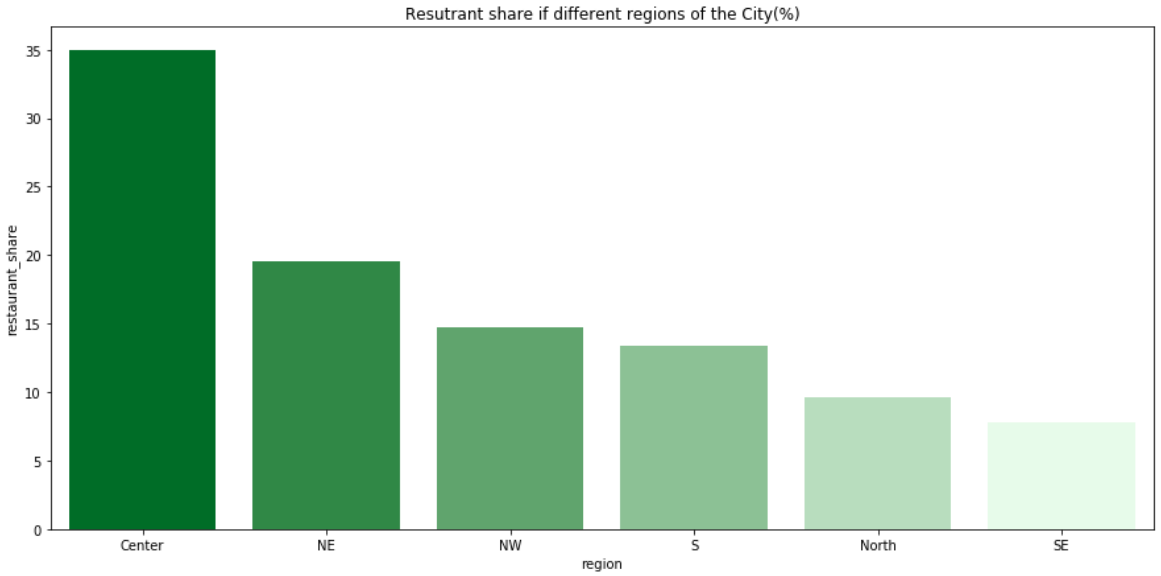


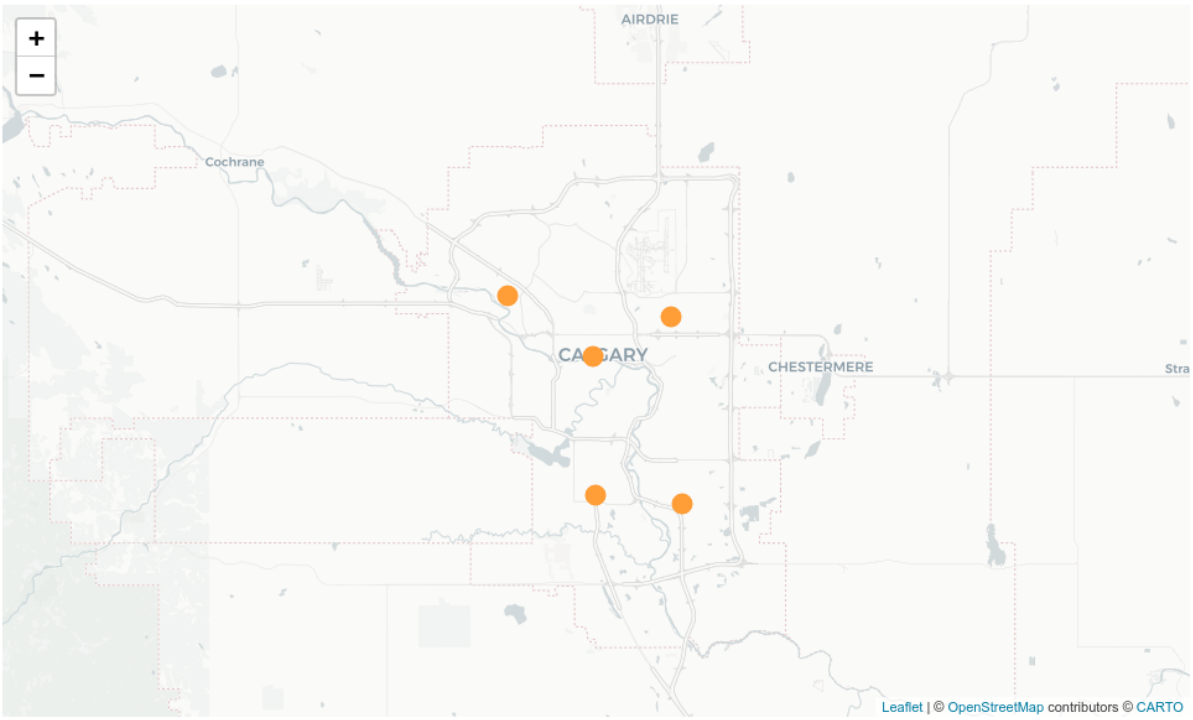
We can see that the optimum number of cluster is 6. Folium can not handle drawing a large amount of data.So Lets plot the centers on along with all the point on a Scatter plot:


Restauran Centers
SE
Center
S
NW
N
NE
51.20
51.15
51.10
51.05
51.00
50.95
50.90
50.85
latitude
−114.25
−114.20
−114.15
−114.10
−114.05
−114.00
−113.95
−113.90
longitude

4.3 Market Share per Region

Lets find the number of restaurants in each region to see which area is more popular:





4.2 Recommended location

The following coordinate shows the coordinates of the best location for opening a restaurant in Calgary :

	latitude	longitude	area
0	50.954995	-114.070993	S
1	51.051511	-114.073327	Centre
2	51.079709	-113.987043	NE
3	51.094039	-114.169078	NW
4	50.948672	-113.974758	SE

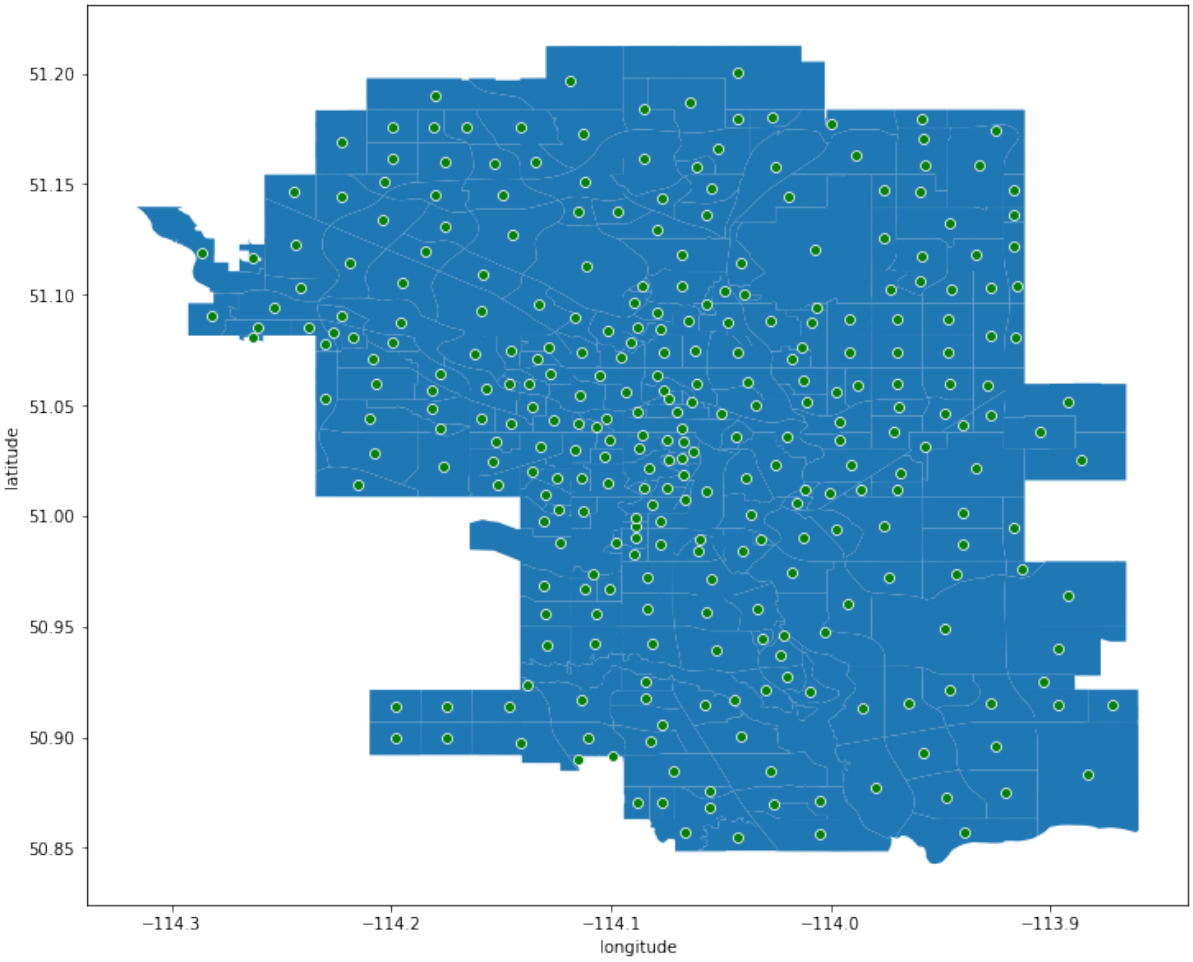
Best Locations:

- City Center:
- NW
- NE
- S
- SE

Let's create a new dataframe that includes the cluster as well as the business name.

2.2.3 Using four square API

In order to use the Foursquare API first we calculate the central location for each neighbourhood.



Calculating Neighbourhood Radius

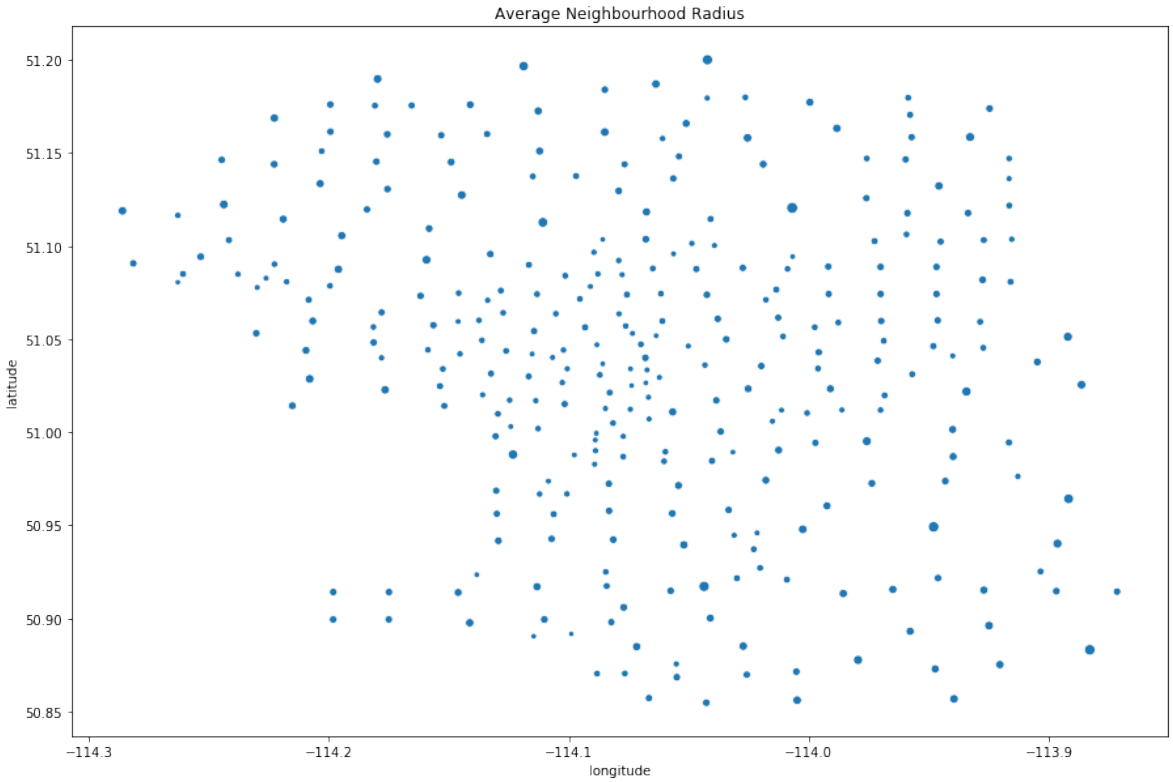
The next step is to find the radius of each neghoubordood.We Do this by first calculating the area of each neghoubordood and then use some arithemtic calculation to find teh radius of a circle that would fit whitin this area.

1.1 Convert to Cartesian system

We change the the projection to a Cartesian system (EPSG:3857, unit= m as in the answer of ResMar) so that the radius would be in meter unit.

2. Calculate the average radius for each neighbourhood:

As we can see from the above table the uni of geometry are now in meteres and therefroe we can go ahead and calculate the area of each district and then find the fitting circle in each neighbourhood



1. Search for a specific venue category:

Now, let's let's define a query to search for restaurants which are that is within the average radius of each neighbourhood.

To avoid fetching the data multiple time from the API ,we save the result into a data frame and load it in subsequent runs.Lets take a look at the data:

latitude	longitude	name	cat_name	cat_pluralName	cat_shortName
51.044450	-114.092754	Korean Village Restaurant	Korean Restaurant	Korean Restaurants	Korean
51.037457	-114.094972	Moti Mahal Restaurant	Indian Restaurant	Indian Restaurants	Indian
51.037839	-114.095126	Green Chili Indian Restaurant	Indian Restaurant	Indian Restaurants	Indian
51.043242	-114.092588	Christos Greek Restaurant	Greek Restaurant	Greek Restaurants	Greek
51.038006	-114.095573	The Rock Cafe and Restaurant	Modern European Restaurant	Modern European Restaurants	Modern European

Fiding missing values

Lets see if we have any missing values for the short Catgory names.This is important siince later on we want to Classify restautant based on this short name:

Property	Number of Missing Records
latitude	0
longitude	0
name	0
cat_name	108
cat_pluralName	108
cat_shortName	108

There are 108 restaurant which only have their name and their short name is missing.Lets take a look at one of these restaurants.

latitude	longitude	name	cat_name	cat_pluralName	cat_shortName
51.038208	-114.094518	Oshii Village Japanese restaurant	NaN	NaN	NaN
50.999720	-114.070960	Whiskey Restaurant - Saddledome	NaN	NaN	NaN
51.085968	-114.128802	Murder Restaurant	NaN	NaN	NaN
51.081580	-113.985390	Misai Japanese restaurant	NaN	NaN	NaN
51.081752	-114.000855	Basil Vietnamese Restaurant	NaN	NaN	NaN

Filling missing values

As we can see for some of these resturant ,we can identify its category from its name. For example Oshii Village Japanese restaurant doset have the short name but we can see that it is a Japanese restaurant. To do these first we create a list of all the unique short Category names in the data.Then we will see if any of these words are present in the name of the restaurant and then we assign the class to its short name. First lets create a list of all the unique short name for the category of restaurants:

```
['Korean', 'Greek', 'Modern European', 'Szechuan', 'Indian',
       'Italian', 'Pub', 'Vietnamese', 'Breakfast', 'Restaurant',
       'Fast Food', 'Vegetarian / Vegan', 'Beer Garden', 'Japanese',
       'Chinese', 'Diner', 'Seafood', 'American', 'Sushi', 'Bar',
       'Filipino', 'South American', 'Thai', 'Hookah Bar', 'Gastropub',
       'Ethiopian', 'Eastern European', 'Lounge', 'Nightclub', 'Pizza',
       'Café', 'Steakhouse', 'Cocktail', 'Dim Sum', 'Shop', 'Falafel',
       'Mediterranean', 'Noodles', 'Brewery', 'Asian', 'Pakistani',
       'Entertainment', 'Office', 'Deli / Bodega', 'Mexican', 'Park',
       'Wine Bar', 'New American', 'Persian', 'Hotel Bar', 'Caribbean',
       'Peruvian', 'Latin American', 'French', 'Ice Cream', 'Hunan',
       'Wings' ]
```

There are 57 unique categories for the restaurants.Now lets use these list to fill the missing values. Now,Lets apply these function to identify the missing restaurants classes in the whole dataset

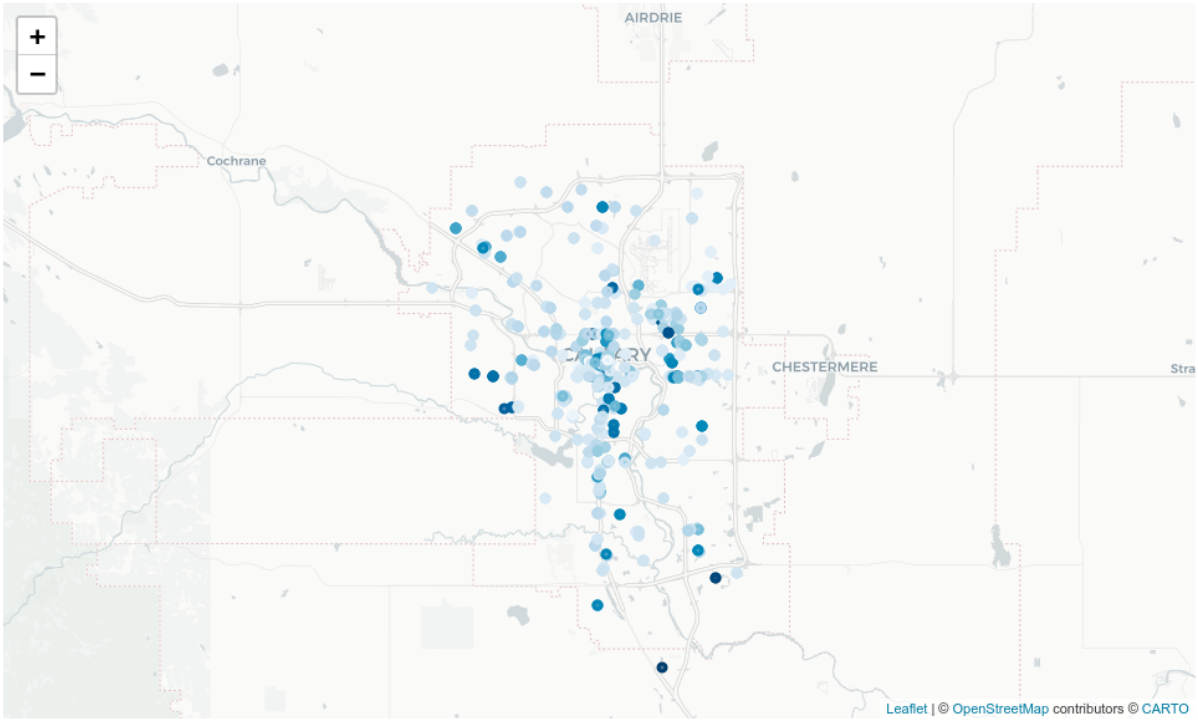
Great! We we able to correctly identify the category of 105 restaurant out of a total of 108 which is great.Lets see what are 3 missing restaurants that we were not able to identify. Now we only have 3 missing valuesLets take a look at those records.

	Unnamed: 0	latitude	longitude	name	cat_name	cat_pluralName	cat_shortName
712	12	51.037909	-114.072639	El Sombrero Restaurante Mexicano	NaN	NaN	None
1005	26	51.037909	-114.072639	El Sombrero Restaurante Mexicano	NaN	NaN	None
1060	11	51.037909	-114.072639	El Sombrero Restaurante Mexicano	NaN	NaN	None

We can see that these are the Mexican restaurants which belong to our South American category .So lets assign this category to these restaurants an by that we will not have any missing categories.

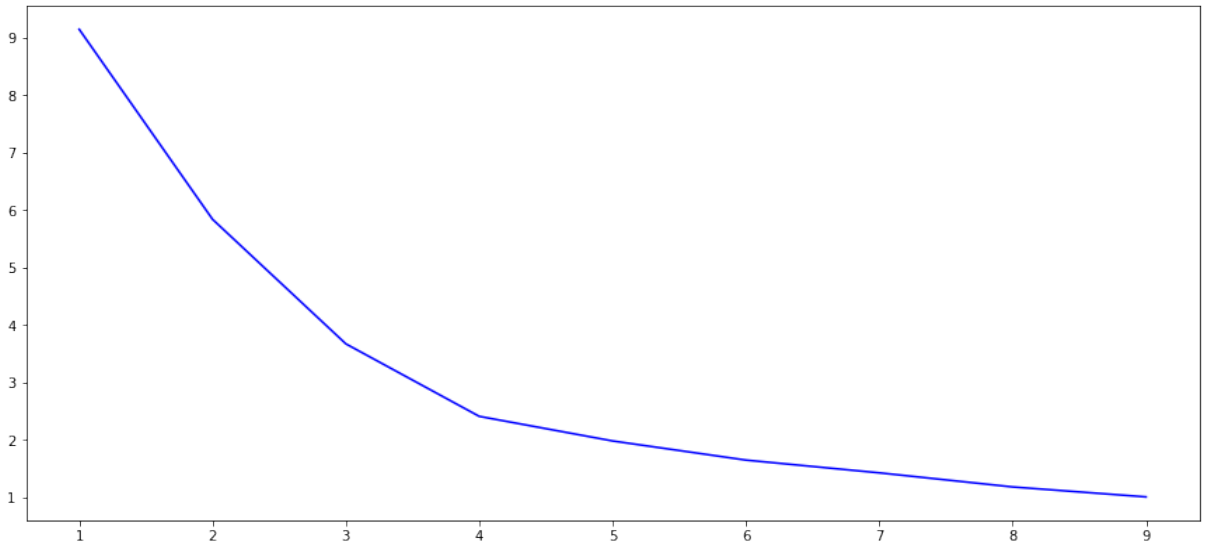
Know we ca see that we do not have any missing values for the cat_shortName of the resturants and therefore we can proceed to the analysis section.

Let's plot the data to see their distributino on the map

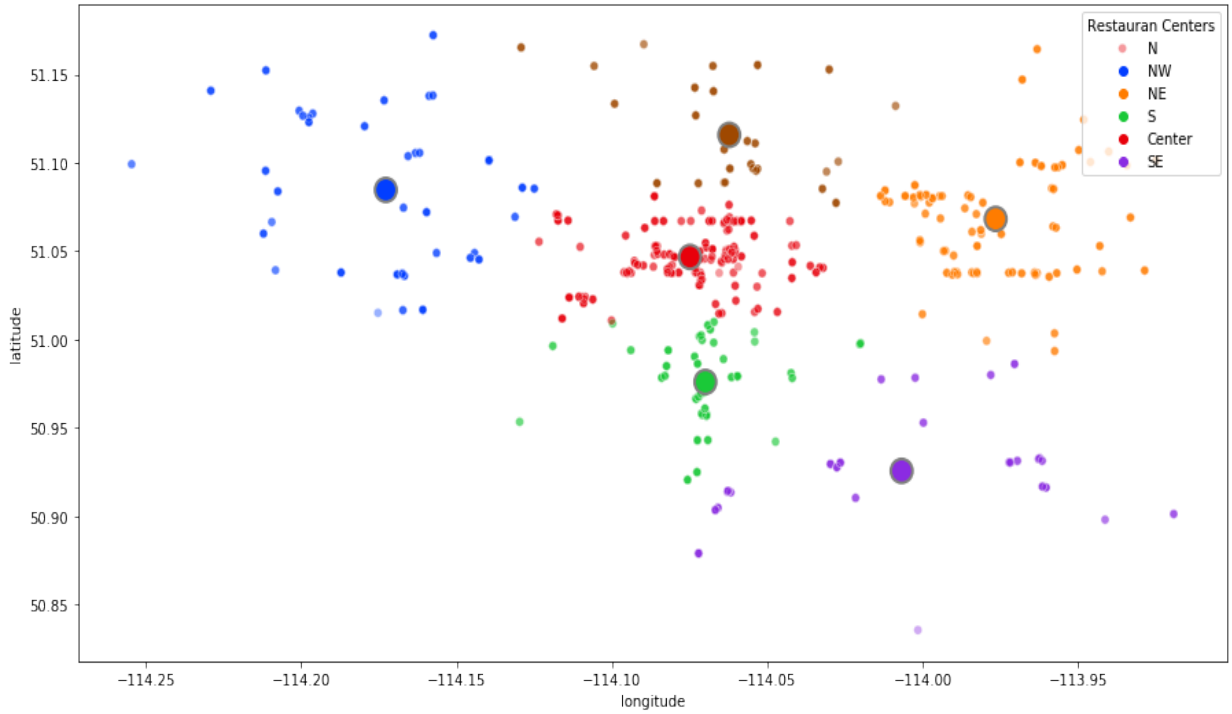


Clustering of the Restaurants

We use the same approch as the one we used befor.



From the above chart we can see that the optimum number of clusters is 6. This makes sense as it divides the city into six regions.

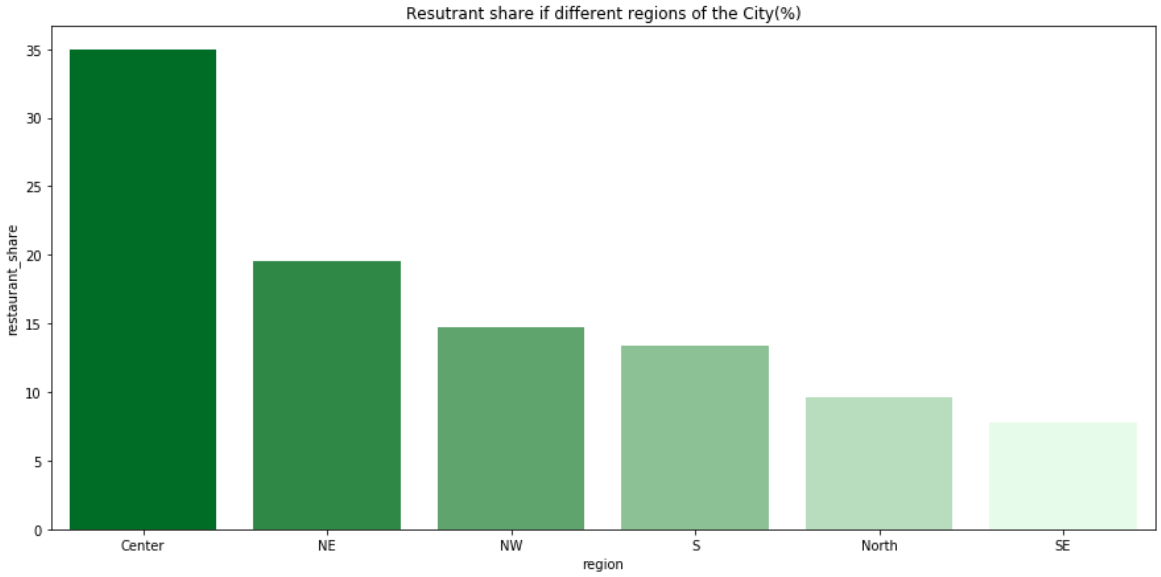


As we saw before with the license data ,again we see a similar pattern in the data and we can classify the data into five cluster.

Top locations

Lest group the data based on the cluster and find out which region has the highest concentration of the restaurants:

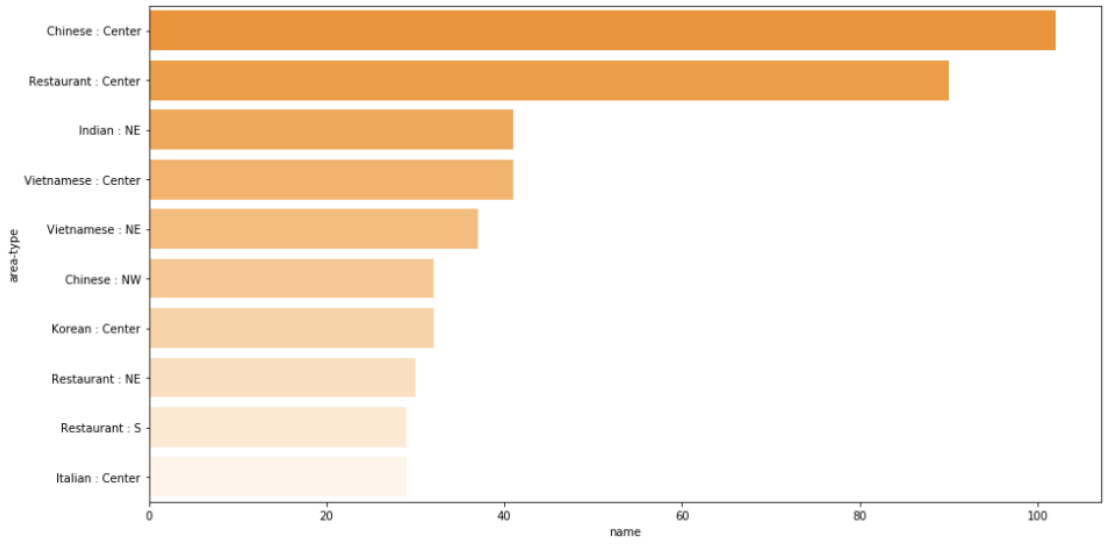
	region	restaurant_share
0	Center	42.172285
1	NE	22.921348
4	S	11.910112
2	NW	10.786517
3	North	7.640449
5	SE	4.569288



The results are in agreement with the previous analysis we obtain from the business license renewal data.As we can see the Center and NE areas are the top locations for opening a restaurants and SE is the least favorable place to open a restaurant business.This makes sense since the SE is an industrial area of the city.

Top restaurant types

Lets first take a look at the top 10 popular restaurants type the city along with their location. We can see the most popular type in Calgary is Chinese restaurant which are located in City center.This makes sense since the china town is also located in City center. The second most popular one is Indian which are located in NE and the third is Vietnamese which are located in Center and NE.



Restaurant popularity per region

In this part lets find out what are the most popular restaurant types in each area of the city

```
SE
[{'type': 'Restaurant', 'count': 12}, {'type': 'Fast Food', 'count': 9},
{'type': 'Chinese', 'count': 8}, {'type': 'Brewery', 'count': 8}, {'type':
'Breakfast', 'count': 7}, {'type': 'Vietnamese', 'count': 5}]
========================================

S
[{'type': 'Restaurant', 'count': 29}, {'type': 'Chinese', 'count': 20},
{'type': 'Fast Food', 'count': 19}, {'type': 'Vietnamese', 'count': 10},
{'type': 'Breakfast', 'count': 10}, {'type': 'Filipino', 'count': 9}]
========================================

North
[{'type': 'Chinese', 'count': 26}, {'type': 'Fast Food', 'count': 15},
{'type': 'Restaurant', 'count': 11}, {'type': 'Vietnamese', 'count': 11},
{'type': 'Japanese', 'count': 11}, {'type': 'Diner', 'count': 5}]
========================================

NW
[{'type': 'Chinese', 'count': 32}, {'type': 'Fast Food', 'count': 28},
{'type': 'Restaurant', 'count': 17}, {'type': 'Vietnamese', 'count': 9},
{'type': 'Seafood', 'count': 8}, {'type': 'Japanese', 'count': 8}]
```

```
========================================

NE
[{'type': 'Indian', 'count': 41}, {'type': 'Vietnamese', 'count': 37},
{'type': 'Restaurant', 'count': 30}, {'type': 'Fast Food', 'count': 27},
{'type': 'Breakfast', 'count': 17}, {'type': 'Chinese', 'count': 15}]
========================================

Center
[{'type': 'Chinese', 'count': 102}, {'type': 'Restaurant', 'count': 90},
{'type': 'Vietnamese', 'count': 41}, {'type': 'Korean', 'count': 32}, {'type':
'Italian', 'count': 29}, {'type': 'Indian', 'count': 26}]
========================================
```

Conclusions

There are 6 major areas in the City of Calgary that has the high number of restauranst.We performed the restaurant site selection analysis using two different data sources :

- Business License Renewal
- Four Square API
- We came with the consistent conclusions with both datasets showing the central part of the city has the highest number of concentration of the restaurants. These area is the most likely place to open a new shop. We also identified the popularity of each type of restaurants in different areas of the city .So we can select the restaurant location based on the type of the restaurant as well.
-