Location optimization for establishing a new Chinese restaurant in Vancouver city areas.

1) Introduction :

Vancouver city is one of the big cities in Canada located in the west. Vancouver CSD (Census Subdivision), Vancouver city consists of 22 neighborhoods. Besides, Vancouver city is a multicultural city. It is formed by a mix of people who are of different races, having different religions, ethnicities, and cultural.

2) Business Problem :

There are 167180 Chinese people staying in Vancouver city. Thus, this makes establishing a Chinese restaurant a good choice of investment. As an investor, it is always important to find an optimal place to establish a Chinese restaurant. In this case, the neighborhood of Vancouver CSD need to be scanned through to identify establishing a Chinese restaurant in which area will higher business opportunities and lesser competition. This project is targeted to investors who would like to establish a new Chinese restaurant in Vancouver CSD.

3) Data Sources :

In order to establish a Chinese restaurant in Vancouver CSD, we will be scrapping and getting the data from the following:

A) City of Vancouver census local area profiles 2016:-

https://opendata.vancouver.ca/explore/dataset/census-local-area-profiles-2016/information/

Example of data:

A1) There are 22 neighborhoods in Vancouver CSD.

A2) There are 3045 Chinese people staying in the Arbutus-Ridge neighborhood.

B) Latitude and longitude of neighborhood in Vancouver city can be gotten through usage of python library Geocoder. Refer to the following link for more information:-

https://geocoder.readthedocs.io/

Example of data: The latitude and longitude of the Arbutus-Ridge neighborhood are 49.246305, -123.159636.

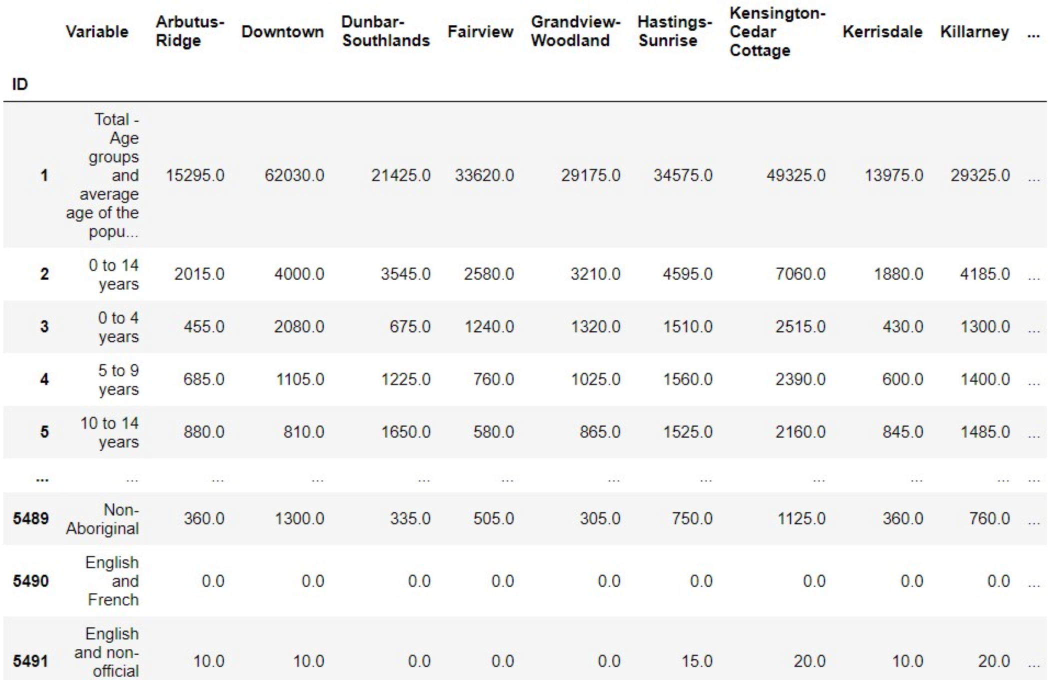
C) Foursquare API to explore the famous venues especially Chinese restaurants in the neighborhoods of Vancouver CSD. One will need to register her developer account from the following URL in order to access Foursquare API.

https://developer.foursquare.com/

Example of data: To explore the Arbutus-Ridge neighborhood and identify the number of Chinese restaurants.

4) Data Pre-Processing :

4.1) Data Wrangling / Data Cleaning :



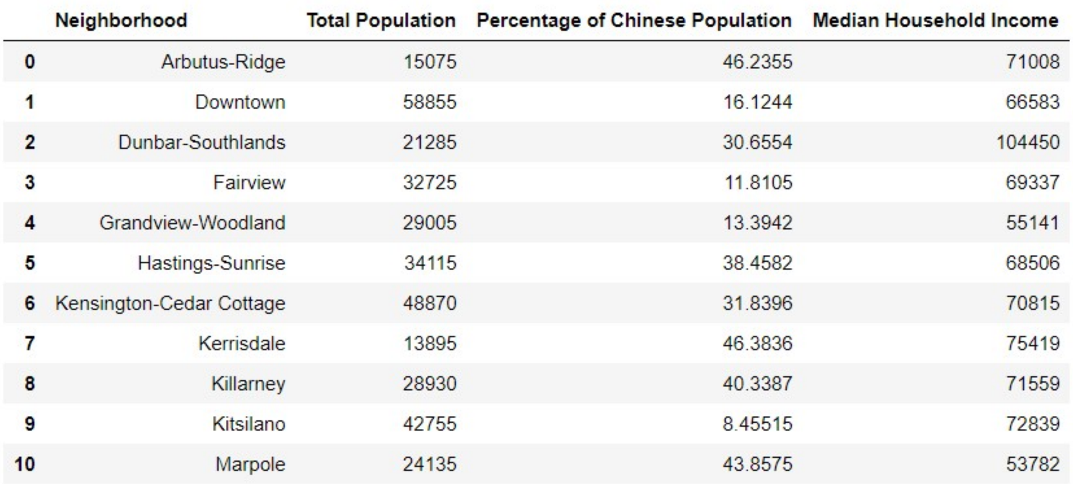
	Variable	Arbutus-Ridge	Downtown	Dunbar-Southlands	Fairview	Grandview-Woodland	Hastings-Sunrise	Kensington-Cedar Cottage	Kerrisdale	Killarney	...
ID											
1	Total - Age groups and average age of the popu...	15295.0	62030.0	21425.0	33620.0	29175.0	34575.0	49325.0	13975.0	29325.0	...
2	0 to 14 years	2015.0	4000.0	3545.0	2580.0	3210.0	4595.0	7060.0	1880.0	4185.0	...
3	0 to 4 years	455.0	2080.0	675.0	1240.0	1320.0	1510.0	2515.0	430.0	1300.0	...
4	5 to 9 years	685.0	1105.0	1225.0	760.0	1025.0	1560.0	2390.0	600.0	1400.0	...
5	10 to 14 years	880.0	810.0	1650.0	580.0	865.0	1525.0	2160.0	845.0	1485.0	...
...	...	...	...	...	...	...	...	...	...	...	...
5489	Non-Aboriginal	360.0	1300.0	335.0	505.0	305.0	750.0	1125.0	360.0	760.0	...
5490	English and French	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
5491	English and non-official	10.0	10.0	0.0	0.0	0.0	15.0	20.0	10.0	20.0	...

City of Vancouver census local area profiles 2016 data

The data downloaded for City of Vancouver census local area profiles 2016 consists of all the Vancouver city data such as people from different age groups, household income, total visible minorities like Chinese population and etc for all the local areas (a.k.a. neighborhoods) in Vancouver CSD. Therefore the data to be focused on need to be filtered out. The data to be observed are the following :

- The 22 neighborhoods in Vancouver CSD
- The total population in each neighborhood
- The Chinese visible minority in each neighborhood
- The household income for each neighborhood (mean value and median value)

Thus, the ID number of the above information and located the information are identified in the data downloaded. Then, the data can be extracted out and put into a data frame shown below :

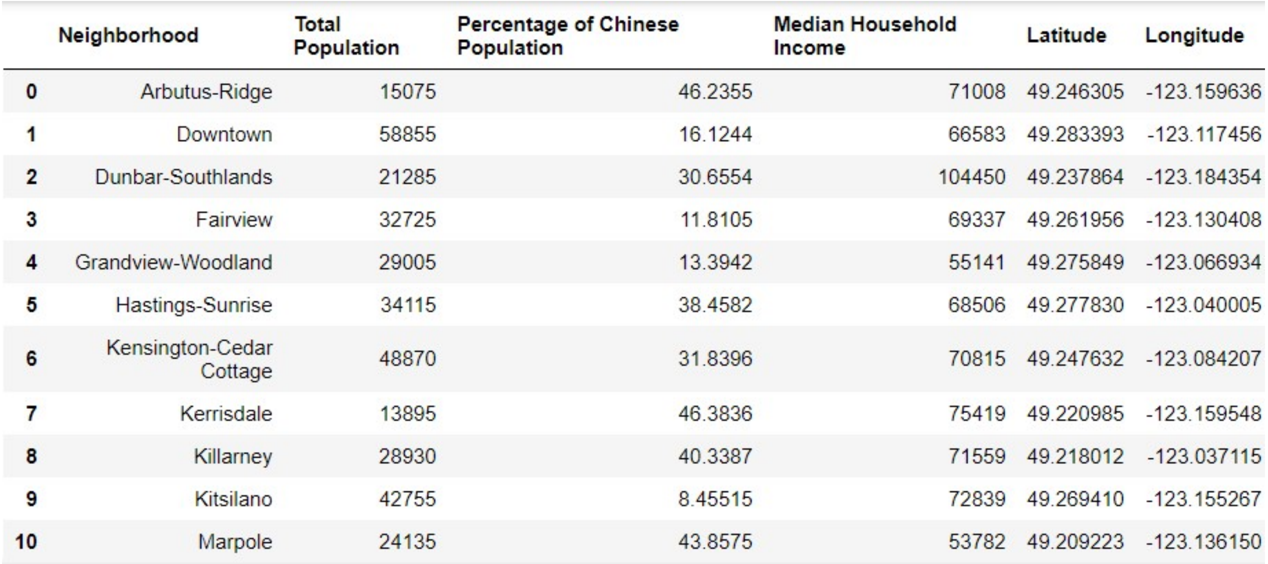


	Neighborhood	Total Population	Percentage of Chinese Population	Median Household Income
0	Arbutus-Ridge	15075	46.2355	71008
1	Downtown	58855	16.1244	66583
2	Dunbar-Southlands	21285	30.6554	104450
3	Fairview	32725	11.8105	69337
4	Grandview-Woodland	29005	13.3942	55141
5	Hastings-Sunrise	34115	38.4582	68506
6	Kensington-Cedar Cottage	48870	31.8396	70815
7	Kerrisdale	13895	46.3836	75419
8	Killarney	28930	40.3387	71559
9	Kitsilano	42755	8.45515	72839
10	Marpole	24135	43.8575	53782

The data frame showing all the data to be focused on (Showing only 10 of the neighborhoods)

A side note for the household income data. Median household income is taken into consideration in this case as the mean household income is higher than median household income. This means that the household income data is skewed. Taking median instead of mean value will be more appropriate.

In order to get the latitude and longitude information, geopy.geocoders.Nominatim package is used. After that, the data frame is updated with the latitude and longitude information.



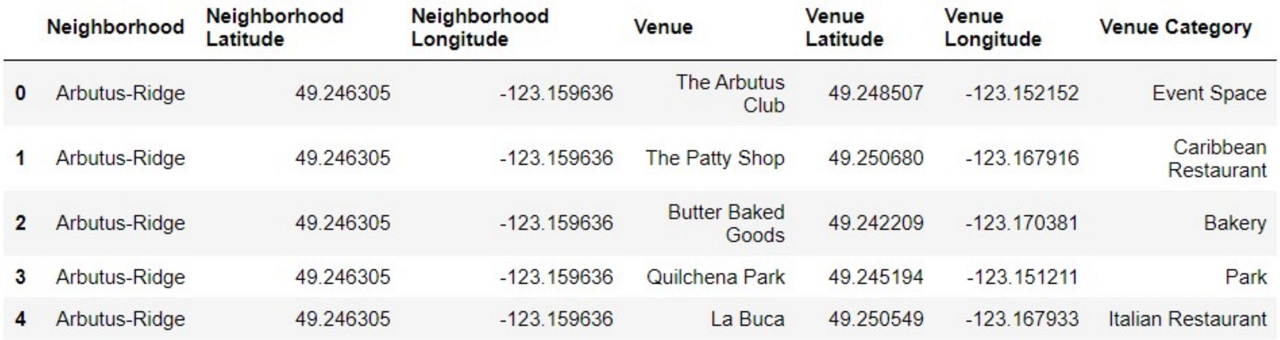
	Neighborhood	Total Population	Percentage of Chinese Population	Median Household Income	Latitude	Longitude
0	Arbutus-Ridge	15075	46.2355	71008	49.246305	-123.159636
1	Downtown	58855	16.1244	66583	49.283393	-123.117456
2	Dunbar-Southlands	21285	30.6554	104450	49.237864	-123.184354
3	Fairview	32725	11.8105	69337	49.261956	-123.130408
4	Grandview-Woodland	29005	13.3942	55141	49.275849	-123.066934
5	Hastings-Sunrise	34115	38.4582	68506	49.277830	-123.040005
6	Kensington-Cedar Cottage	48870	31.8396	70815	49.247632	-123.084207
7	Kerrisdale	13895	46.3836	75419	49.220985	-123.159548
8	Killarney	28930	40.3387	71559	49.218012	-123.037115
9	Kitsilano	42755	8.45515	72839	49.269410	-123.155267
10	Marpole	24135	43.8575	53782	49.209223	-123.136150

The data frame after added in latitude and longitude (Showing only 10 of the neighborhoods)

4,2) Using Foursquare application programming interface (API) :

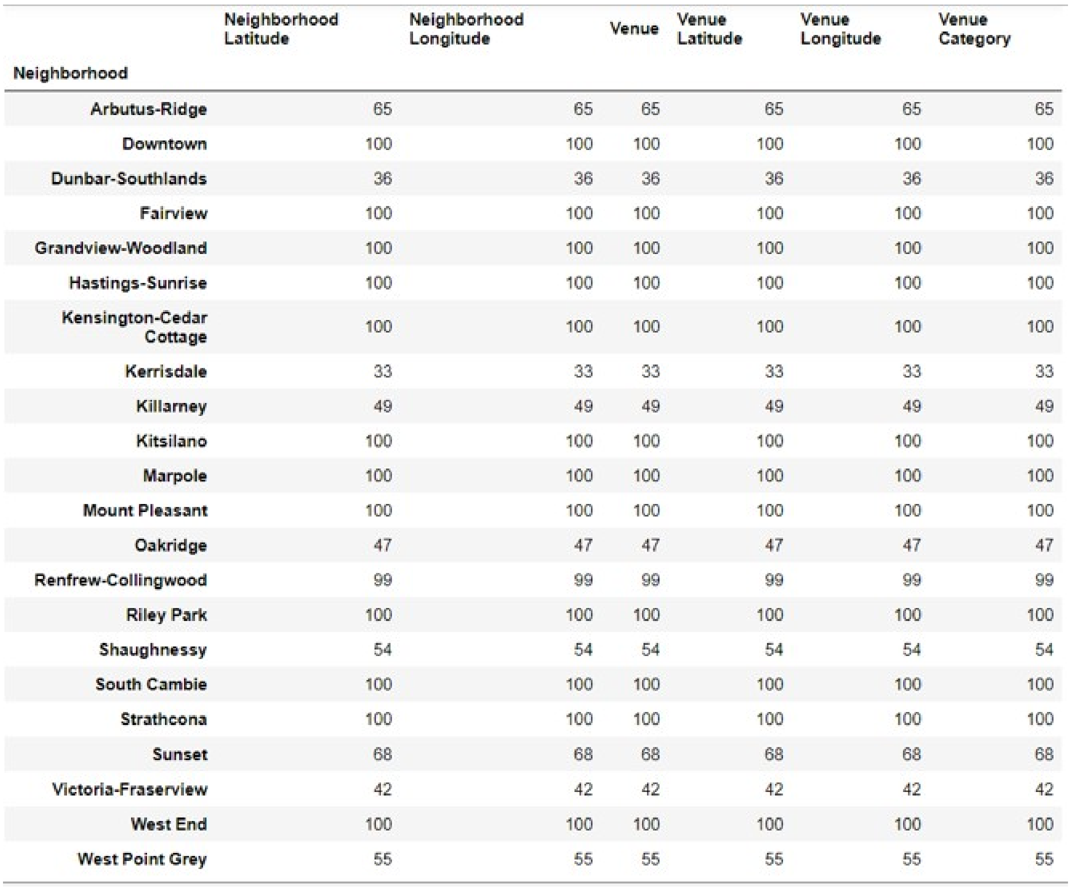
There is another important data that we need to gather and that data is about the number of Chinese restaurant in each neighborhood of Vancouver CSD. In order to get this data, the explore function in the Foursquare API will need to be used.

By using the explore function in Foursquare API, each neighborhood in Vancouver CSD is screened for the popular spots around the latitude and longitude information. In this case, the radius has been set to 1.5 km while the maximum limit of popular spots is 100. The 2 tables below show 1) the first 5 popular venues explored around the Arbutus-Ridge neighborhood and 2) the total number of popular venues explored around the neighborhoods in Vancouver CSD.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arbutus-Ridge	49.246305	-123.159636	The Arbutus Club	49.248507	-123.152152	Event Space
1	Arbutus-Ridge	49.246305	-123.159636	The Patty Shop	49.250680	-123.167916	Caribbean Restaurant
2	Arbutus-Ridge	49.246305	-123.159636	Butter Baked Goods	49.242209	-123.170381	Bakery
3	Arbutus-Ridge	49.246305	-123.159636	Quilchena Park	49.245194	-123.151211	Park
4	Arbutus-Ridge	49.246305	-123.159636	La Buca	49.250549	-123.167933	Italian Restaurant

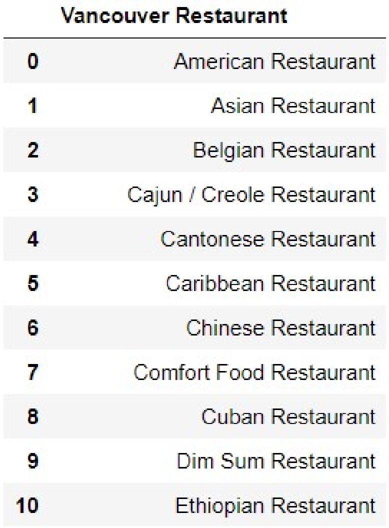
the first 5 popular venues explored around the Arbutus-Ridge neighborhood



	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Arbutus-Ridge	65	65	65	65	65	65
Downtown	100	100	100	100	100	100
Dunbar-Southlands	36	36	36	36	36	36
Fairview	100	100	100	100	100	100
Grandview-Woodland	100	100	100	100	100	100
Hastings-Sunrise	100	100	100	100	100	100
Kensington-Cedar Cottage	100	100	100	100	100	100
Kerrisdale	33	33	33	33	33	33
Killarney	49	49	49	49	49	49
Kitsilano	100	100	100	100	100	100
Marpole	100	100	100	100	100	100
Mount Pleasant	100	100	100	100	100	100
Oakridge	47	47	47	47	47	47
Renfrew-Collingwood	99	99	99	99	99	99
Riley Park	100	100	100	100	100	100
Shaughnessy	54	54	54	54	54	54
South Cambie	100	100	100	100	100	100
Strathcona	100	100	100	100	100	100
Sunset	68	68	68	68	68	68
Victoria-Fraserview	42	42	42	42	42	42
West End	100	100	100	100	100	100
West Point Grey	55	55	55	55	55	55

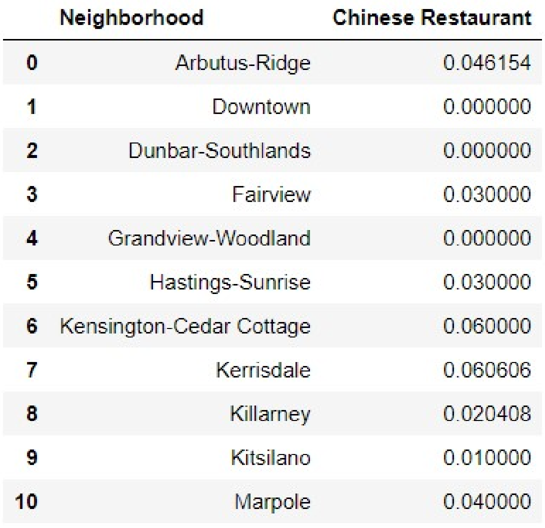
the total number of popular venues explored around the neighborhoods in Vancouver city

With the data above, we are able to identify all types of restaurants located in the Vancouver city area. The first 10 types of restaurants in Vancouver city are shown below follows alphabetical order:



	Vancouver Restaurant
0	American Restaurant
1	Asian Restaurant
2	Belgian Restaurant
3	Cajun / Creole Restaurant
4	Cantonese Restaurant
5	Caribbean Restaurant
6	Chinese Restaurant
7	Comfort Food Restaurant
8	Cuban Restaurant
9	Dim Sum Restaurant
10	Ethiopian Restaurant

Types of restaurant in Vancouver CSD



	Neighborhood	Chinese Restaurant
0	Arbutus-Ridge	0.046154
1	Downtown	0.000000
2	Dunbar-Southlands	0.000000
3	Fairview	0.030000
4	Grandview-Woodland	0.000000
5	Hastings-Sunrise	0.030000
6	Kensington-Cedar Cottage	0.060000
7	Kerrisdale	0.060606
8	Killarney	0.020408
9	Kitsilano	0.010000
10	Marpole	0.040000

Mean frequency of Chinese restaurant

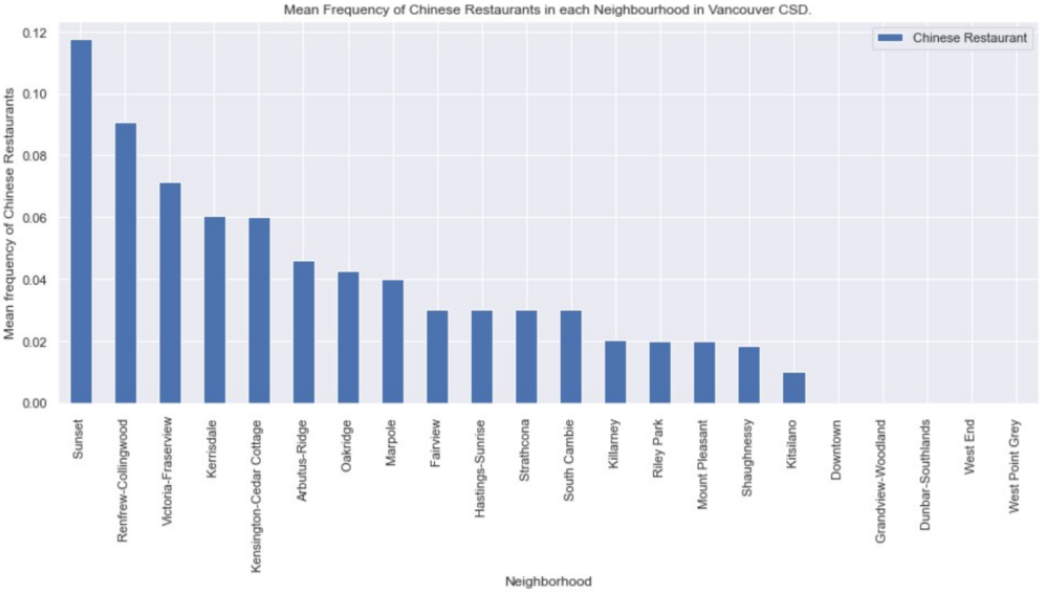
After going through the 43 types of restaurants in Vancouver, it is understandable that Cantonese Restaurant, Chinese Restaurant, and lastly Dim Sums Restaurant can be grouped together as Chinese Restaurant. Therefore, the mean frequency of Chinese restaurants is tabulated according to different neighborhoods shown in the table above.

5) Data Visualization :

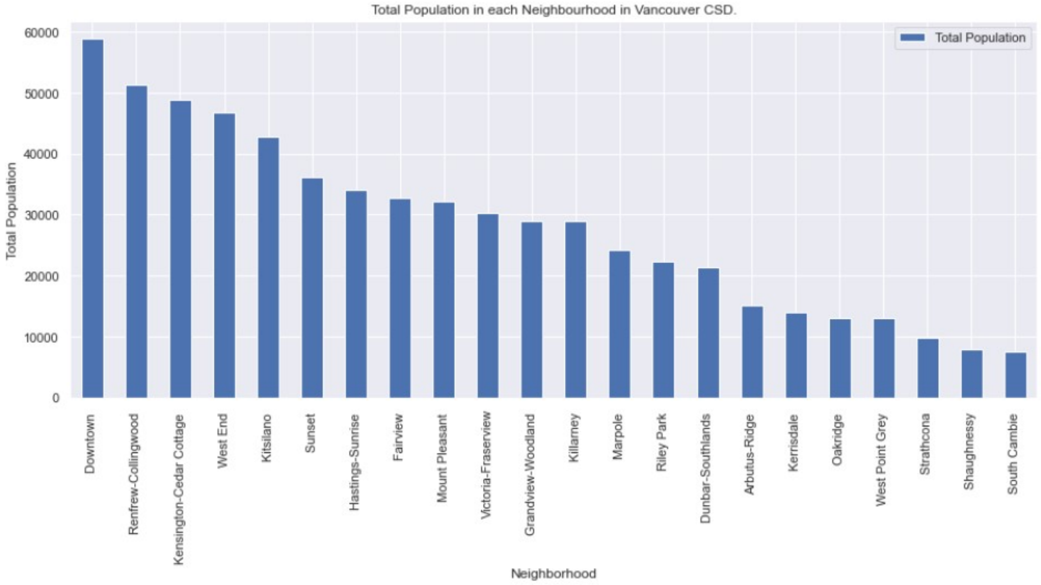
Data visualization is a graphical representation of information and data. By using chart or graphs, data visualization enable the user to easily understand the trends and patterns of the data and the information behinds.

5.1) Bar charts :

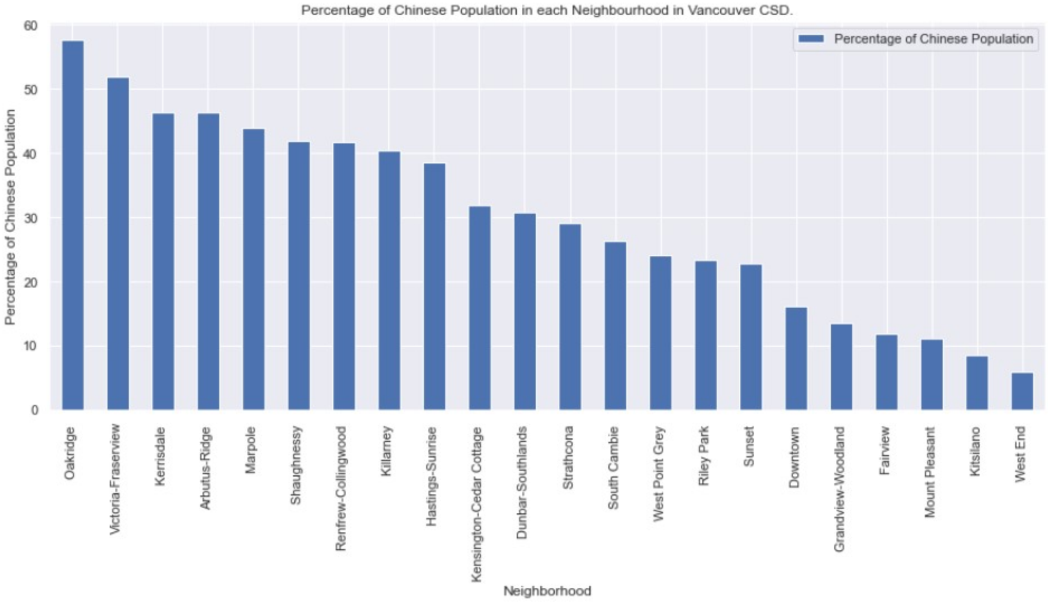
Let visualize the independent variables.



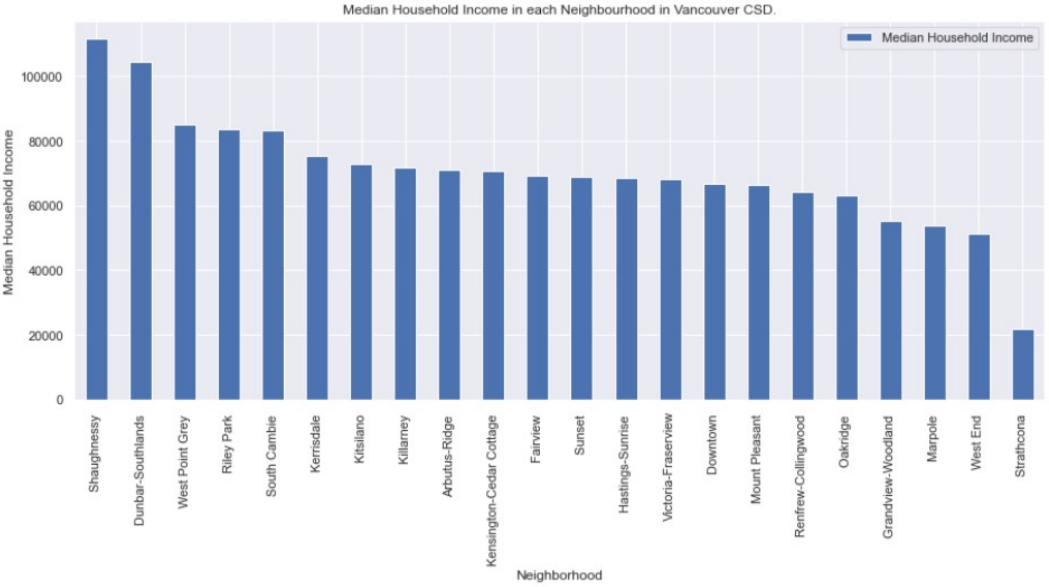
The top 5 neighborhoods which have the highest mean frequency of Chinese Restaurants are :
1) Sunset; 2) Renfrew-Collingwood; 3) Victoria-Fraserview; 4) Kerrisdale; 5) Kensington-Cedar Cottage
These neighborhoods have higher competition for Chinese restaurants. Thus, these neighborhoods will have lower priority when we are deciding the neighborhood to establish a new Chinese restaurant.



The neighborhoods that have the highest population are :
1) Downtown; 2) Renfrew-Collingwood; 3) Kensington-Cedar Cottage; 4) West End; 5) Kitsilano
These neighborhoods having high population which means it will bring in more business for a Restaurant.



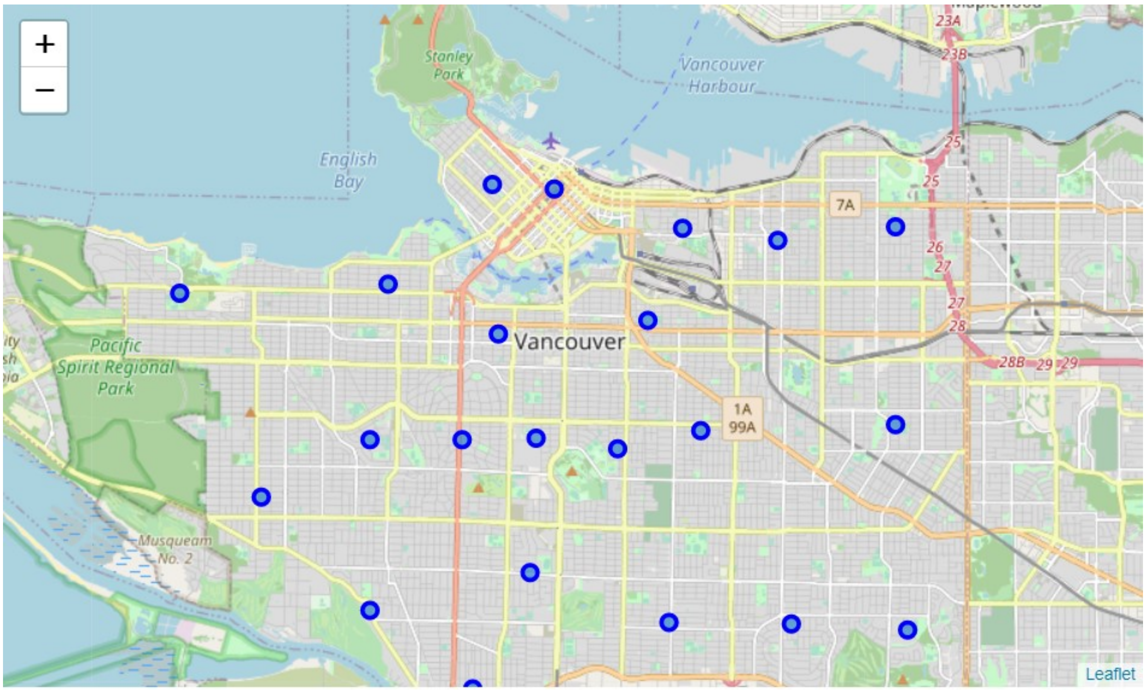
The top 5 neighborhoods that have a high percentage of Chinese people are :
1) Oakridge; 2) Victoria-Fraserview; 3) Kerrisdale; 4) Arbutus-Ridge; 5) Marpole
These neighborhoods have a high percentage of Chinese people staying. And it means that the investor might get more business since Chinese people will usually prefer Chinese foods more.



The top 5 neighborhoods which have the highest mean frequency of Chinese Restaurants are :
1) Shaughnessy; 2) Dunbar-Southlands; 3) West Point Grey; 4) Riley Park; 5) South Cambie
The people who stay in the above neighborhoods have higher household incomes. The people in these neighborhoods will have higher spending power.

5.2) Map visualization using Folium package :

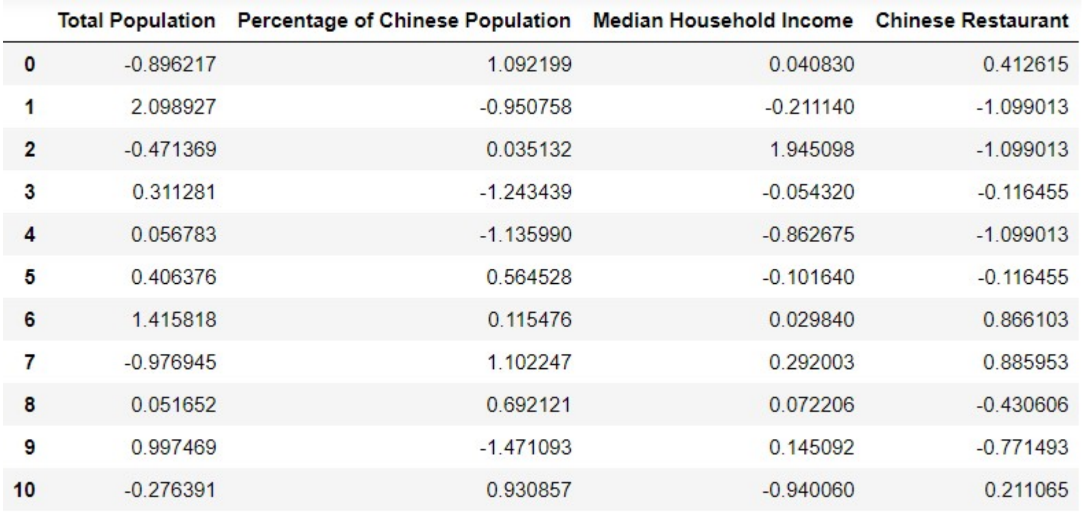
By using the Folium package, the Vancouver map can be generated with the latitude and longitude of Vancouver city (49.26038, -123.11336). The map below showing all the 22 neighborhoods in Vancouver CSD.



6) Clustering Vancouver Neighborhoods :

6.1) Feature Scaling for independent variables :

First of all, the independent variables need to be defined. The independent variables (features) used in the clustering model are: 1) Total population of each neighborhood; 2) Percentage of Chinese People in each neighborhood; 3) Median household income for different neighborhood; 4) Mean frequency of Chinese Restaurant in each neighborhood. Next, these independent variables need to be undergone feature scaling (a.k.a standardization). Feature scaling is used to standardize all the different independent variables to a particular range and it is crucial to the clustering algorithm, K-means used later. With feature scaling, the K-mean algorithm will not be biased toward any one feature.



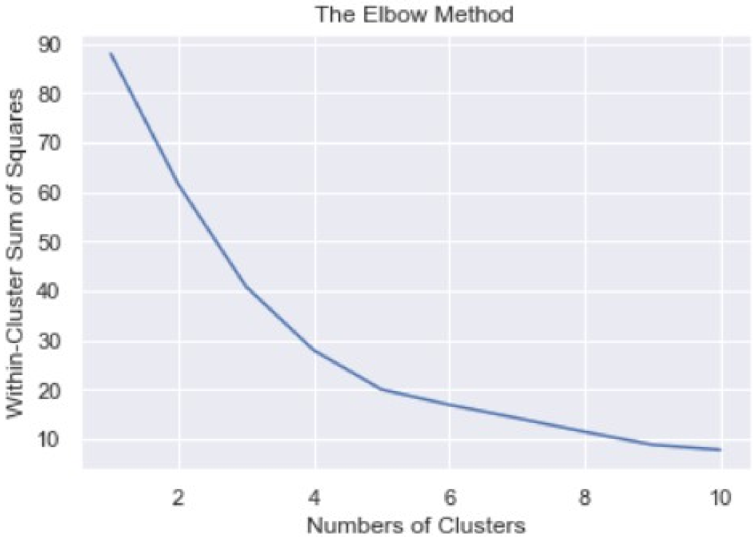
	Total Population	Percentage of Chinese Population	Median Household Income	Chinese Restaurant
0	-0.896217	1.092199	0.040830	0.412615
1	2.098927	-0.950758	-0.211140	-1.099013
2	-0.471369	0.035132	1.945098	-1.099013
3	0.311281	-1.243439	-0.054320	-0.116455
4	0.056783	-1.135990	-0.862675	-1.099013
5	0.406376	0.564528	-0.101640	-0.116455
6	1.415818	0.115476	0.029840	0.866103
7	-0.976945	1.102247	0.292003	0.885953
8	0.051652	0.692121	0.072206	-0.430606
9	0.997469	-1.471093	0.145092	-0.771493
10	-0.276391	0.930857	-0.940060	0.211065

All features after feature scaling

6.2) K-means clustering algorithm :

In this project, K-means algorithm is selected as the unsupervised machining learning algorithm as it is quite a simple algorithm to be implemented. One of the disadvantages of using K-means algorithm is that the number of clusters, K value needs to be chosen correctly in order to get a good result.

In order to find the good K value, within cluster sum of square (WCSS) will need to be minimized. In this case, the elbow method is applied. The best result will be the lower number of cluster, K with the lower WCSS. Therefore, K-means algorithm is applied to standardized features shown in part 6.1 with different number of clusters, K. A graph can be plotted to show how the WCSS changes with the number of clusters, K.

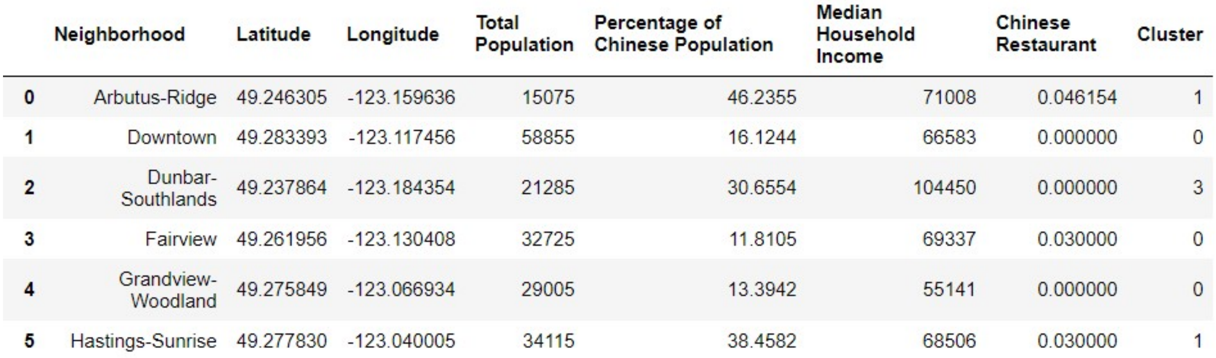


The elbow method to find the best number of clusters

From the graph plotted, it can be clearly seen that the best K value, the number of clusters is 5. It is because the WCSS dropped dramatically from K value = 1 to K value = 5 and the drop is reduced significantly after K value = 5. Therefore, K value = 5 is the number of clusters to be chosen to be used with K-means algorithm.

7) Result :

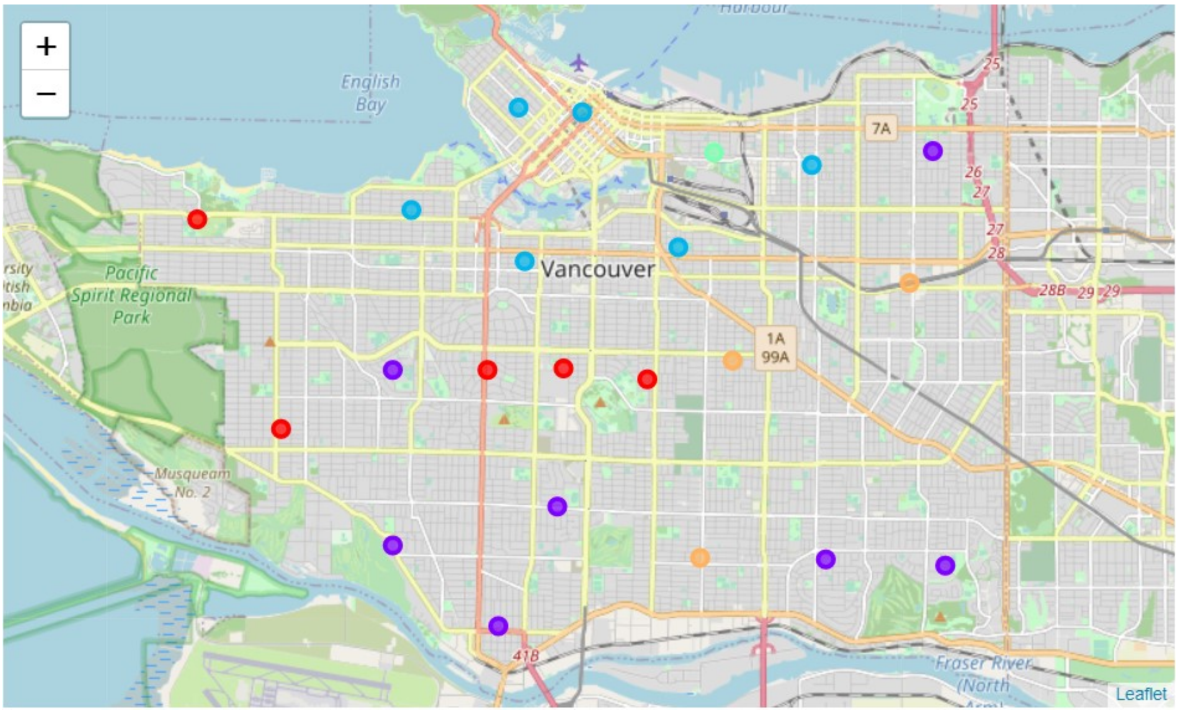
The standardized features undergone K-means algorithm with K-value = 5 and the 22 neighborhoods in Vancouver CSD will be separated into 5 cluster groups. These cluster labels are then inserted into the data frame shown below.



	Neighborhood	Latitude	Longitude	Total Population	Percentage of Chinese Population	Median Household Income	Chinese Restaurant	Cluster
0	Arbutus-Ridge	49.246305	-123.159636	15075	46.2355	71008	0.046154	1
1	Downtown	49.283393	-123.117456	58855	16.1244	66583	0.000000	0
2	Dunbar-Southlands	49.237864	-123.184354	21285	30.6554	104450	0.000000	3
3	Fairview	49.261956	-123.130408	32725	11.8105	69337	0.030000	0
4	Grandview-Woodland	49.275849	-123.066934	29005	13.3942	55141	0.000000	0
5	Hastings-Sunrise	49.277830	-123.040005	34115	38.4582	68506	0.030000	1

The data frame after included the cluster label for each neighborhood

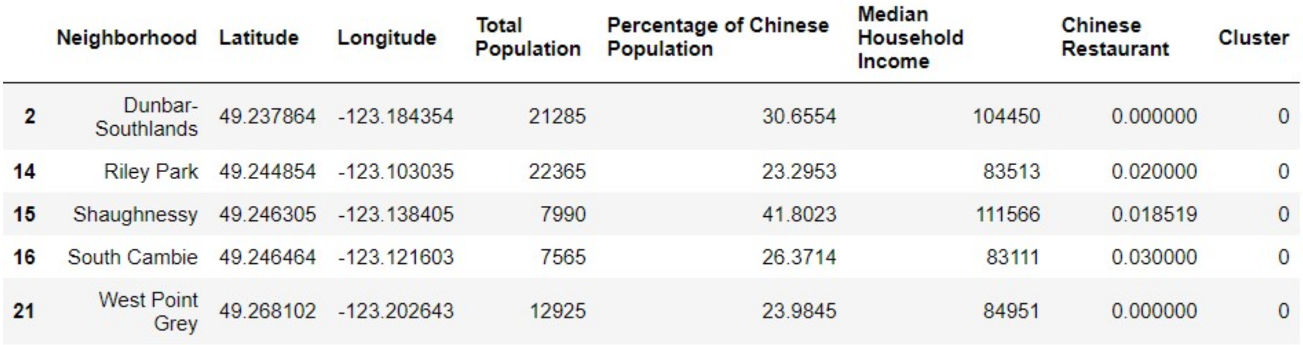
By using Folium, a map showing Vancouver city can be generated with different cluster indicated by different colors.



Folium generated map showing different neighborhoods belong to different clusters.

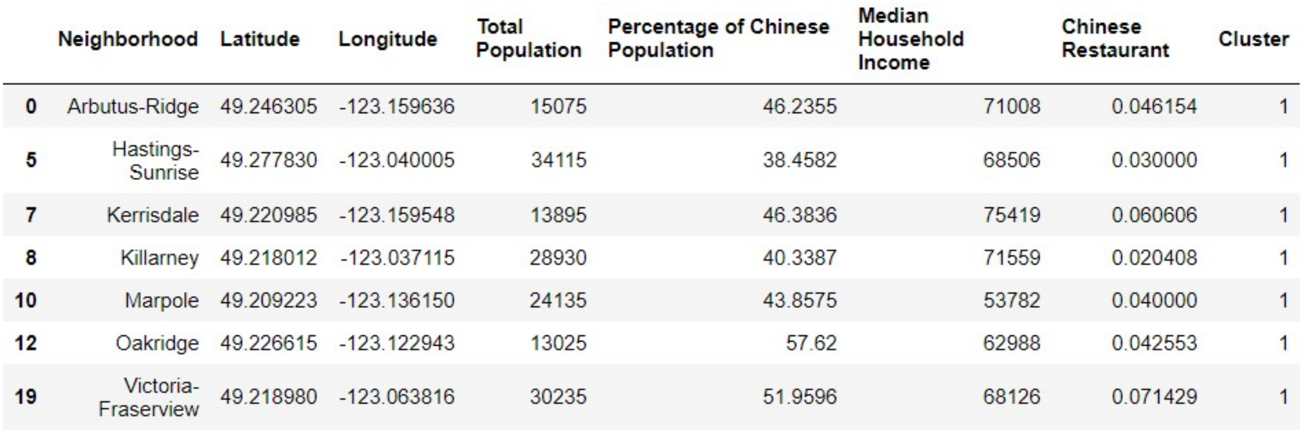
7.1) Clusters analysis :

Each cluster is being analyzed for the 4 features: total population and percentage of the Chinese population, median household income, and lastly mean frequency of Chinese restaurant. The median household income can be interpreted as the spending power of a household while the mean frequency of Chinese Restaurants represents the market competition that the investor is going to face.



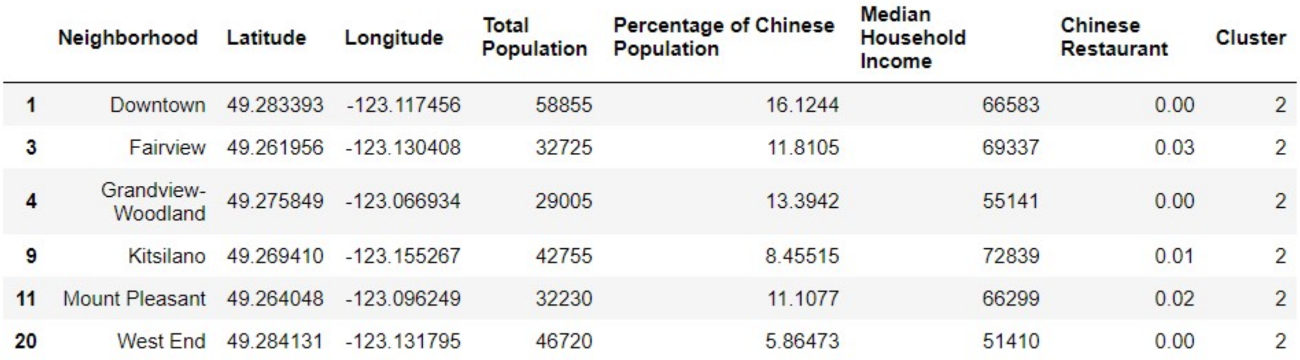
	Neighborhood	Latitude	Longitude	Total Population	Percentage of Chinese Population	Median Household Income	Chinese Restaurant	Cluster
2	Dunbar-Southlands	49.237864	-123.184354	21285	30.6554	104450	0.000000	0
14	Riley Park	49.244854	-123.103035	22365	23.2953	83513	0.020000	0
15	Shaughnessy	49.246305	-123.138405	7990	41.8023	111566	0.018519	0
16	South Cambie	49.246464	-123.121603	7565	26.3714	83111	0.030000	0
21	West Point Grey	49.268102	-123.202643	12925	23.9845	84951	0.000000	0

Cluster 0: Low Chinese population, high household income, and low mean frequency of Chinese restaurants.



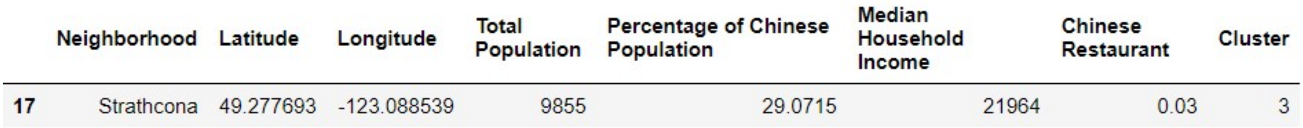
	Neighborhood	Latitude	Longitude	Total Population	Percentage of Chinese Population	Median Household Income	Chinese Restaurant	Cluster
0	Arbutus-Ridge	49.246305	-123.159636	15075	46.2355	71008	0.046154	1
5	Hastings-Sunrise	49.277830	-123.040005	34115	38.4582	68506	0.030000	1
7	Kerrisdale	49.220985	-123.159548	13895	46.3836	75419	0.060606	1
8	Killarney	49.218012	-123.037115	28930	40.3387	71559	0.020408	1
10	Marpole	49.209223	-123.136150	24135	43.8575	53782	0.040000	1
12	Oakridge	49.226615	-123.122943	13025	57.62	62988	0.042553	1
19	Victoria-Fraserview	49.218980	-123.063816	30235	51.9596	68126	0.071429	1

Cluster 1: Medium Chinese population, medium household income, and medium mean frequency of Chinese restaurants.



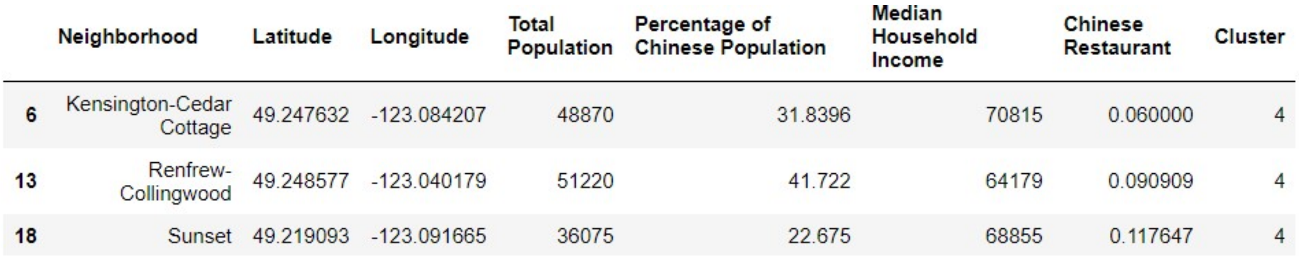
	Neighborhood	Latitude	Longitude	Total Population	Percentage of Chinese Population	Median Household Income	Chinese Restaurant	Cluster
1	Downtown	49.283393	-123.117456	58855	16.1244	66583	0.00	2
3	Fairview	49.261956	-123.130408	32725	11.8105	69337	0.03	2
4	Grandview-Woodland	49.275849	-123.066934	29005	13.3942	55141	0.00	2
9	Kitsilano	49.269410	-123.155267	42755	8.45515	72839	0.01	2
11	Mount Pleasant	49.264048	-123.096249	32230	11.1077	66299	0.02	2
20	West End	49.284131	-123.131795	46720	5.86473	51410	0.00	2

Cluster 2: Low Chinese population, medium household income, and low mean frequency of Chinese restaurants.



	Neighborhood	Latitude	Longitude	Total Population	Percentage of Chinese Population	Median Household Income	Chinese Restaurant	Cluster
17	Strathcona	49.277693	-123.088539	9855	29.0715	21964	0.03	3

Cluster 3: Low Chinese population, low household income, and low mean frequency of Chinese restaurants.

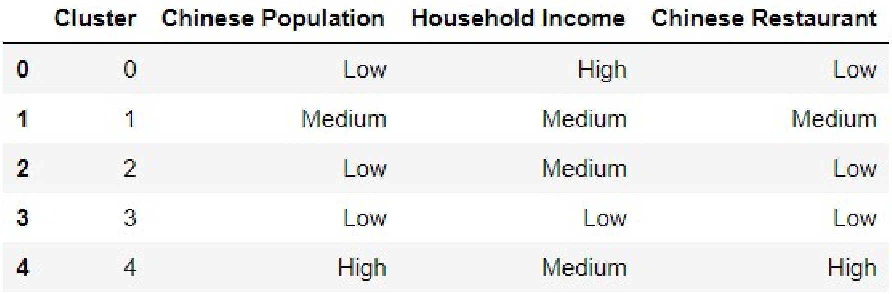


	Neighborhood	Latitude	Longitude	Total Population	Percentage of Chinese Population	Median Household Income	Chinese Restaurant	Cluster
6	Kensington-Cedar Cottage	49.247632	-123.084207	48870	31.8396	70815	0.060000	4
13	Renfrew-Collingwood	49.248577	-123.040179	51220	41.722	64179	0.090909	4
18	Sunset	49.219093	-123.091665	36075	22.675	68855	0.117647	4

Cluster 4: High Chinese population, medium household income, and high mean frequency of Chinese restaurants.

7.2) Results' Summary :

The whole project is to help investors in finding the optimum neighborhood for establishing a new Chinese restaurant in Vancouver city. With the unsupervised machine learning algorithm used, the K-means model, the neighborhoods of Vancouver CSD are grouped into 5 cluster groups. The analysis above is summarized in a table.

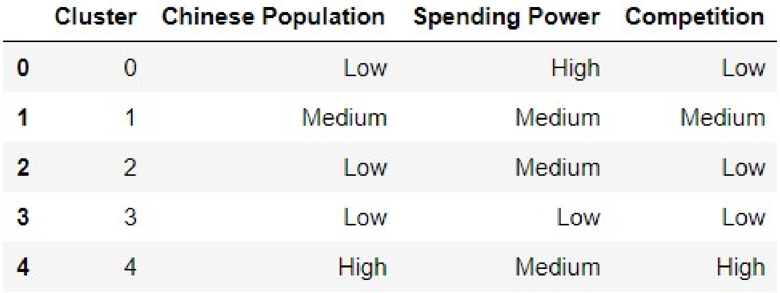


	Cluster	Chinese Population	Household Income	Chinese Restaurant
0	0	Low	High	Low
1	1	Medium	Medium	Medium
2	2	Low	Medium	Low
3	3	Low	Low	Low
4	4	High	Medium	High

The cluster analysis summary table

8) Discussion :

It is assumed that Chinese people will prefer Chinese food, thus they will go to Chinese restaurants more often. This means that the higher the Chinese population in the neighborhood, there will be more business for a Chinese Restaurant. Besides, the household income decides the spending power of a household. And lastly, the mean frequency of Chinese restaurant represents the market competition.



	Cluster	Chinese Population	Spending Power	Competition
0	0	Low	High	Low
1	1	Medium	Medium	Medium
2	2	Low	Medium	Low
3	3	Low	Low	Low
4	4	High	Medium	High

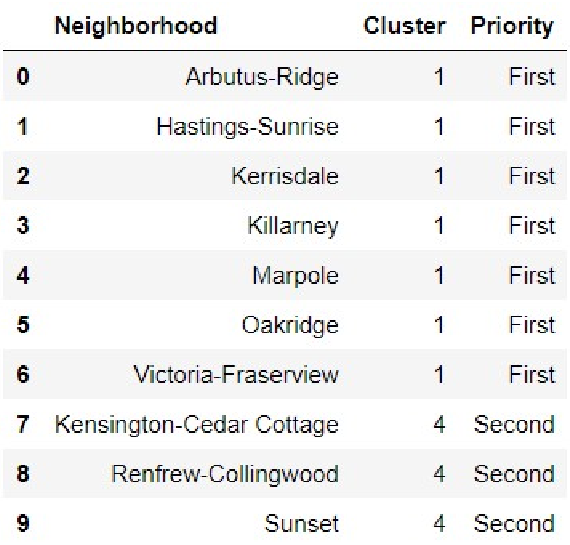
From the summary table above, the neighborhoods which have a low Chinese population can be ignored as it means there will be lesser business if the investor opened a Chinese restaurant in those neighborhoods. Then, the neighborhood with high and medium spending power should be prioritized. And lastly, the investor will focus on the neighborhoods with low market competition. However, the investor should also consider that lesser market competition might be due to lesser demand.

9) Conclusion :

As a result, neighborhoods in cluster group 1 are to be prioritized by the investor who is going to establish a new Chinese restaurant in Vancouver CSD. These neighborhoods are Arbutus-Ridge, Hastings-Sunrise, Kerrisdale, Killarney, Marpole, Oakridge, and Victoria-Fraserview. Establishing a new Chinese restaurant within these neighborhoods should bring in more income to the Chinese restaurant.

Besides cluster group 1, the investor might also consider the neighborhoods in cluster group 4 for opening a new Chinese restaurant. These neighborhoods are Kensington-Cedar Cottage, Renfrew-Collingwood, and Sunset. Although there are more Chinese restaurants in these neighborhoods of cluster group 4, the Chinese population there is higher as well. The higher Chinese population in these neighborhoods will increase the Chinese cuisines demand. Thus, establishing a new Chinese restaurant in neighborhoods of cluster group 4 can be one of the considerations for investors.

This information can be compiled into a table below. The priority means the priority of the neighborhoods for the investor to consider when the investor is establishing a new Chinese restaurant in Vancouver CSD.



	Neighborhood	Cluster	Priority
0	Arbutus-Ridge	1	First
1	Hastings-Sunrise	1	First
2	Kerrisdale	1	First
3	Killarney	1	First
4	Marpole	1	First
5	Oakridge	1	First
6	Victoria-Fraserview	1	First
7	Kensington-Cedar Cottage	4	Second
8	Renfrew-Collingwood	4	Second
9	Sunset	4	Second

10) Reference :

I) My Github repository for this project :- https://github.com/houng87/coursera-capstone-project

II) Census local area profiles 2016 of Vancouver CSD :- https://opendata.vancouver.ca/explore/dataset/census-local-area-profiles-2016/information/

III) Foursquare API :- https://developer.foursquare.com/