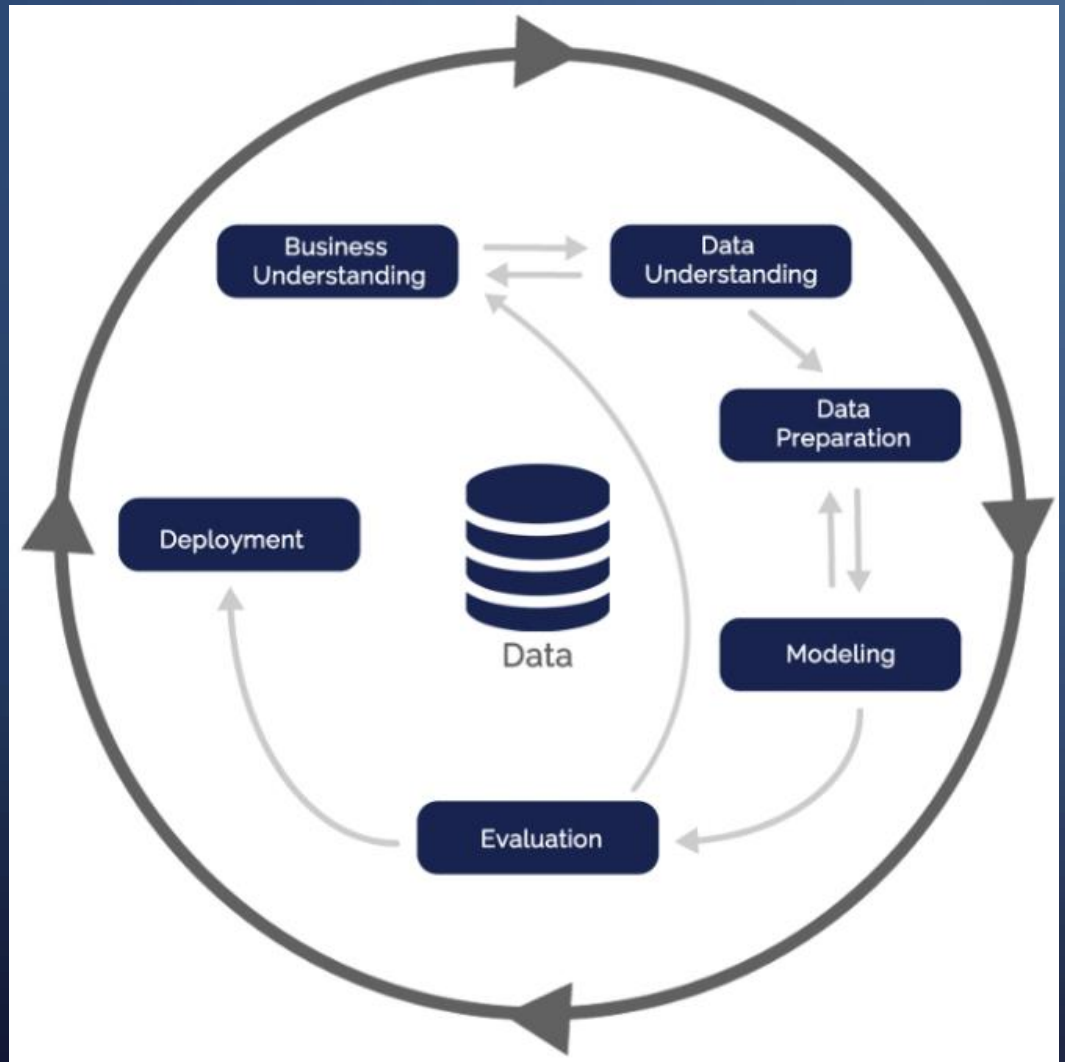
Location optimization for establishing a new Chinese restaurant in Vancouver city areas

Yan Houng

Story posted on Medium

Data Science Methodology

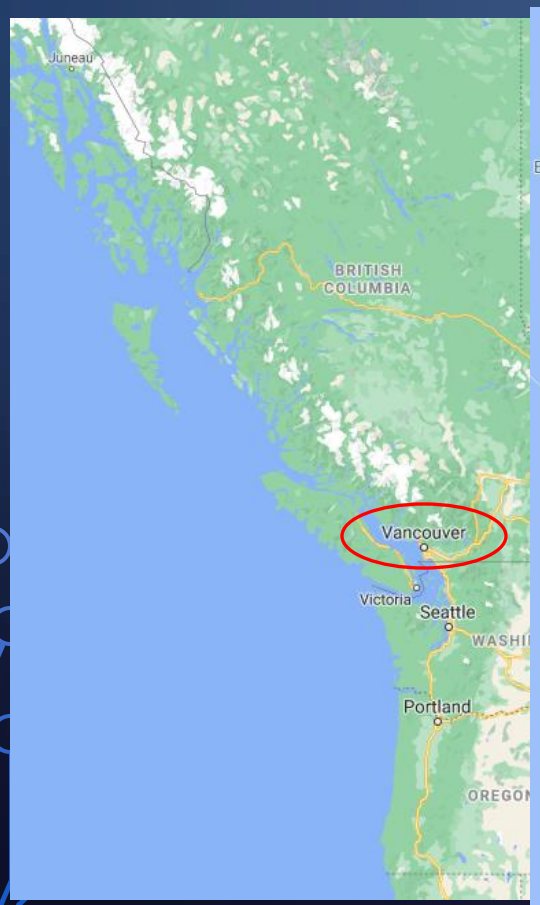


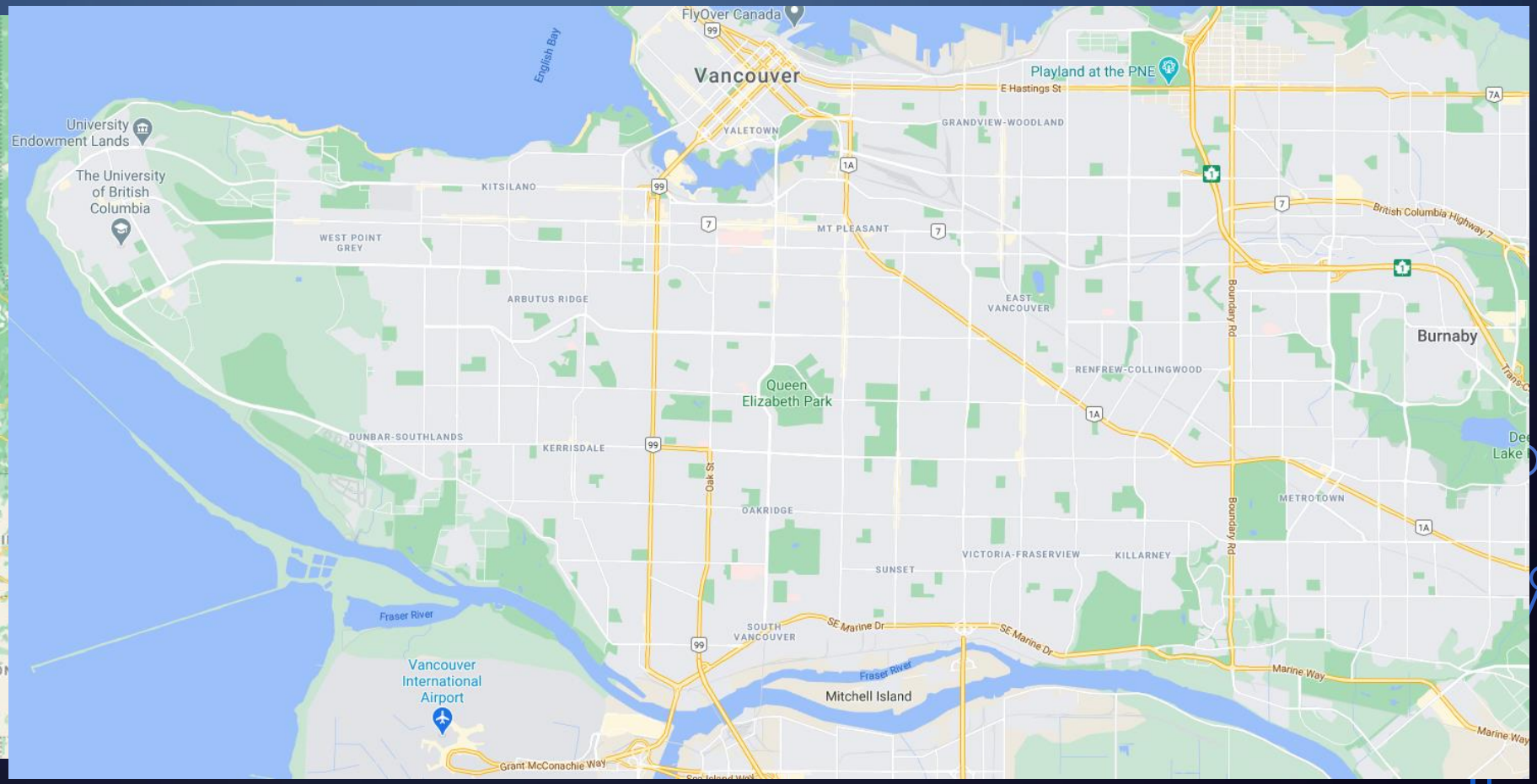
CRISP-DM (1996)
CRoss-Industry Standard Process for Data Mining

1) Business Understanding
   - Understand the business problem
2) Data Understanding
   - Study of the data collected
3) Data Preparation
   - Data pre-processing
4) Modelling
   - Build machine learning models
5) Evaluation
   - Evaluate the models built
6) Deployment
   - Implementation into actual usage

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.
- 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.





Business Problem :

- There are 167180 Chinese people staying in Vancouver city.
- 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.
- Need to scan through the neighborhood of Vancouver CSD to identify establishing a Chinese restaurant in which area will higher business opportunities and lesser competition.

Target audience :

- Investors who want to establish a new Chinese restaurant in Vancouver CSD.

Data Sources :

A. City of Vancouver census local area profiles 2016
   https://opendata.vancouver.ca/explore/dataset/census-local-area-profiles-2016/information/
   Access of the data in Vancouver city

B. Geocoder (Python library)
   To retrieve the latitude and longitude for each neighborhood in Vancouver city

C. Foursquare API
   To explore the famous venues exists in the neighborhoods of Vancouver CSD

Data Pre-processing :

I. Data Cleaning

- The data collected usually consists of many errors such as NaN (Not a Number – missing value), incorrect value (typo mistake during data entry) and etc.
- The data may contains too many irrelevant information.

II. Using Foursquare API (Application Programming Interface)

- Foursquare is a social networking service application for smartphones.
- Foursquare is to help in discovering and sharing information about businesses and attractions around a location.
- Foursquare API allowed us to extract the data in exploring the businesses and attractions around the location that we set.

Data Pre-processing :

I. Data Cleaning

ID	Variable	Arbutus-Ridge	Downtown	Dunbar-Southlands	Fairview	Grandview-Woodland	Hastings-Sunrise	Kensington-Cedar Cottage	Kerrisdale	Killarney	...
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	...

22 local areas + CSD total & CMA total

5491 different data categories

City of Vancouver census local area profiles 2016

Note : CMA = Census Metropolitan Area

Data Pre-processing :

I. Data Cleaning

We are focusing on the following data :

1. The 22 neighbourhoods in Vancouver city
2. The total population in each neighbourhood
3. The Chinese visible minority in each neighbourhood
4. The household income for each neighbourhood

	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

Data Pre-processing :

I. Data Cleaning

Using geopy.geocoder.Nominatim package in Python, we are able to get the latitude and longitude information for each neighbourhood in Vancouver CSD.

	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

Data Pre-processing :

II. Using Foursquare API

To extract the special venues around the neighbourhoods of Vancouver city using the latitude and longitude data we have gotten.

Radius of exploration = 1.5 km

Maximum limit of venues = 100

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
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

	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

Some venues explored in Arbutus-Ridge neighbourhood

Data Pre-processing :

II. Using Foursquare API

Summarised the venue category and sorted out for different types of restaurant.

	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

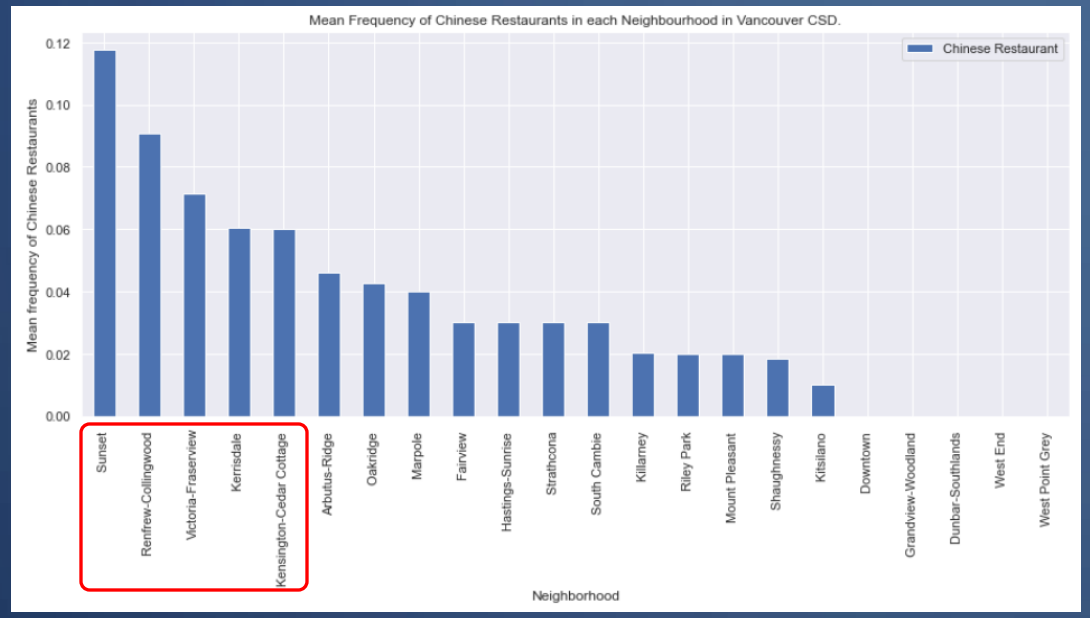
Total 43 types of restaurant in Vancouver

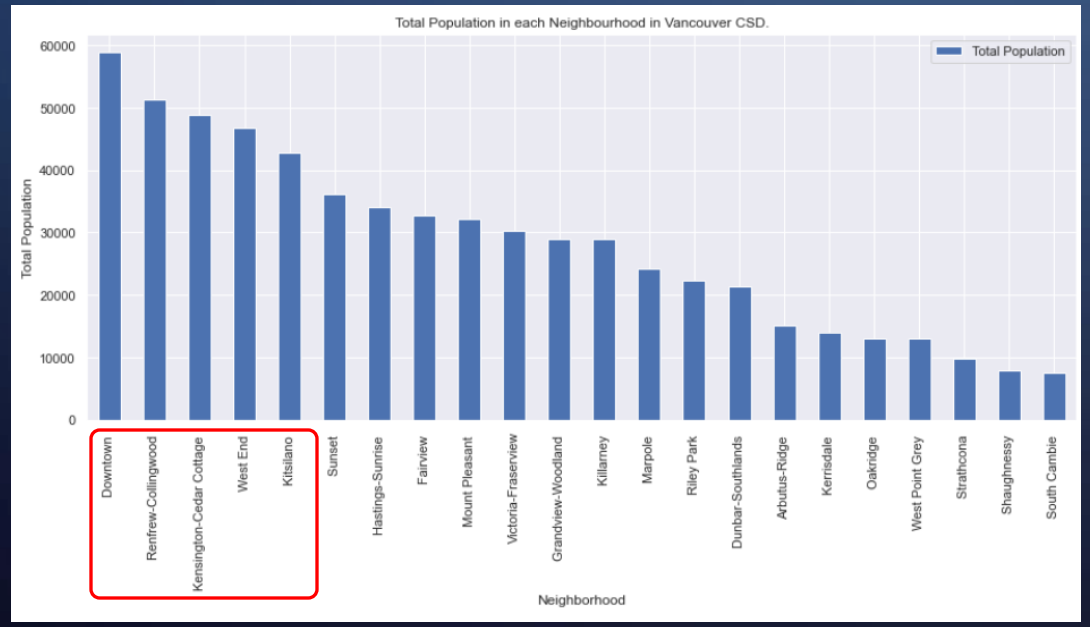
Combined 3 restaurant data and calculate frequency of occurrence of Chinese Restaurant.

	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

Data Visualization:

I. Bar Chart



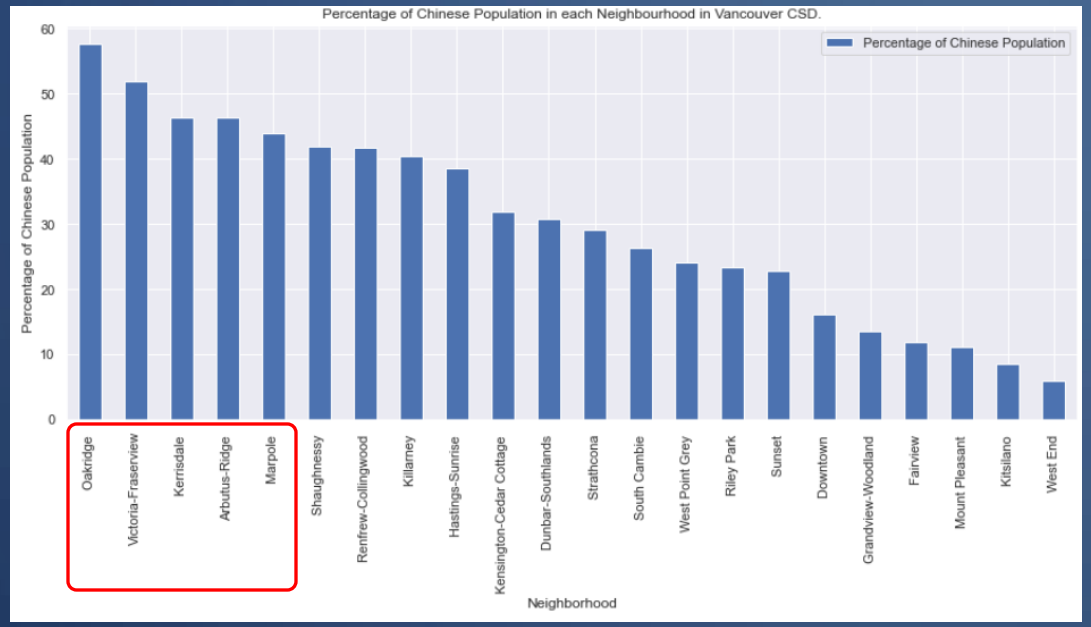


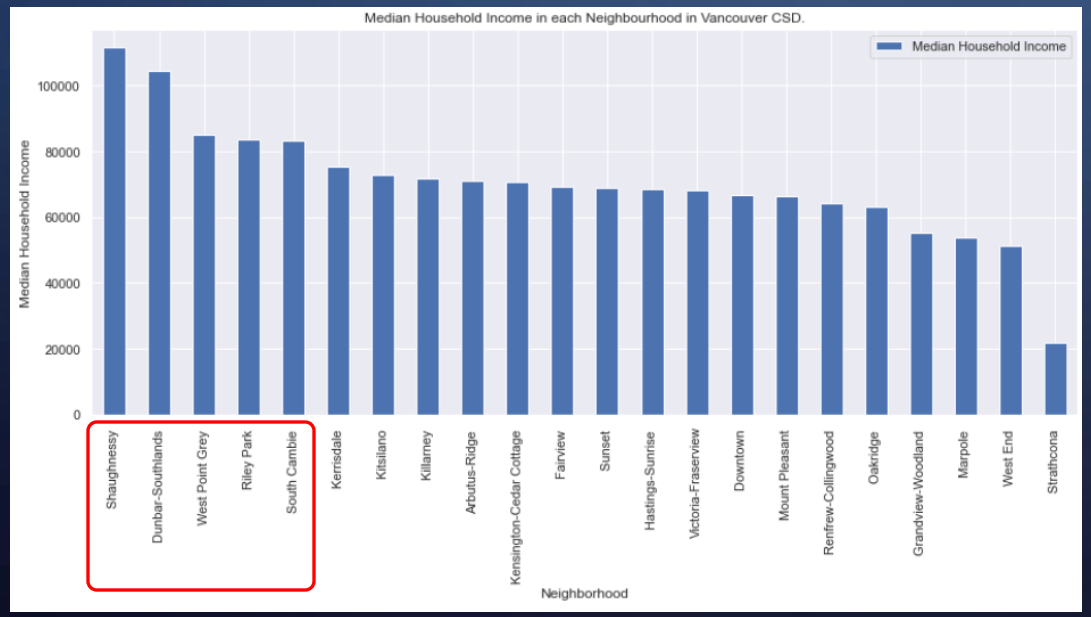
- High competition for Chinese restaurants → lower priority area to choose

- Higher population will bring more business to a restaurant

Data Visualization:

I. Bar Chart





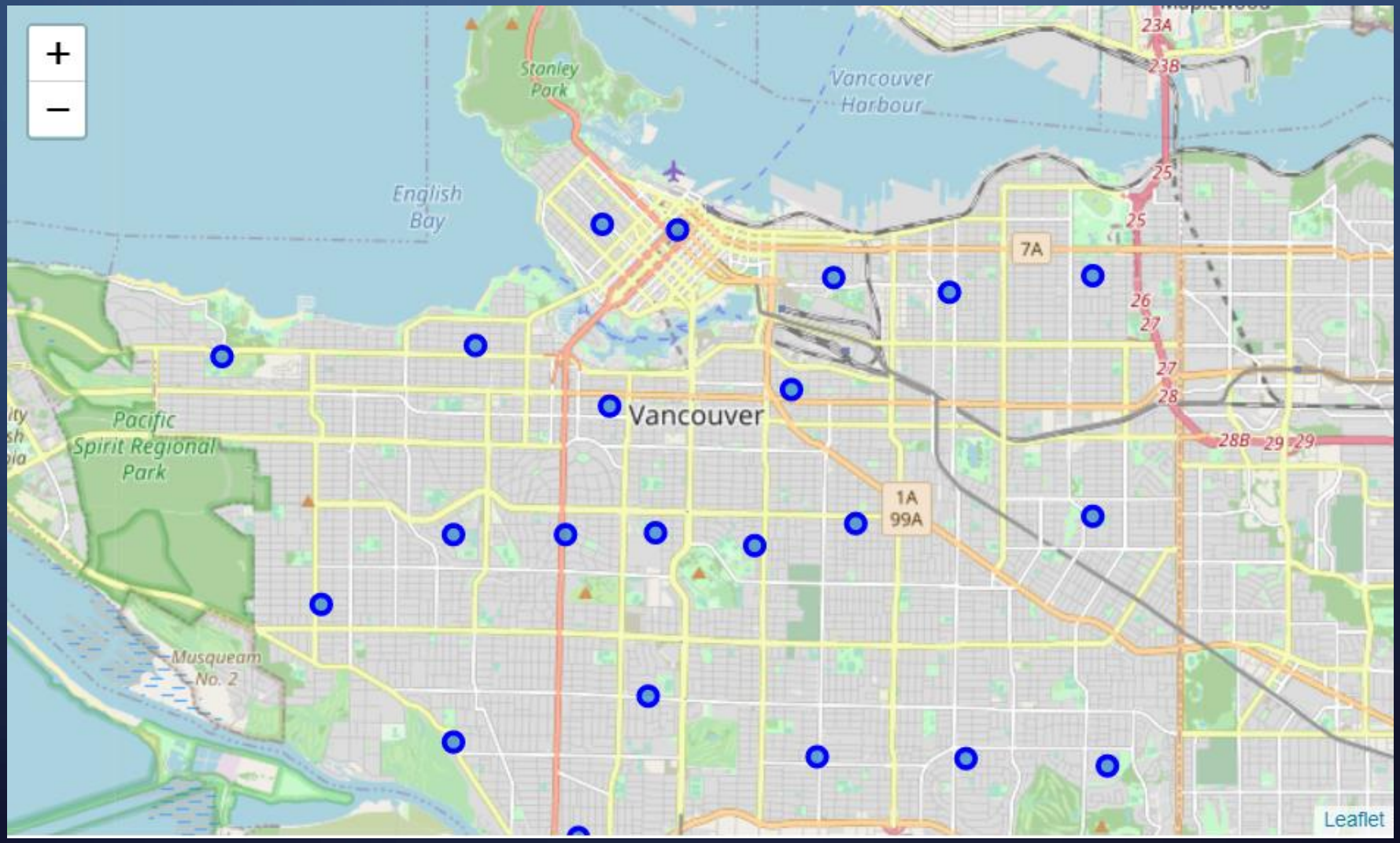
- Chinese people will prefer Chinese foods
- Higher Chinese population → more business

- Higher household income → higher spending power, more likely to have meal in restaurant

Data Visualization:

II. Map visualization

Using folium package, the Vancouver city map can be generated.



All 22 neighbourhoods in Vancouver CSD.

Machine Learning Modelling – Clustering problem:

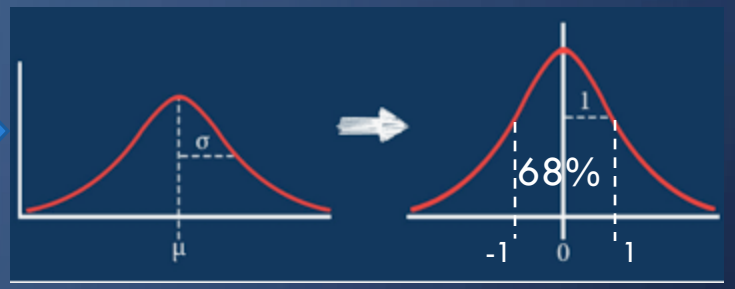
I. Feature scaling for all the features

The features to be used for clustering model are :

- Total population
- Percentage of Chinese population
- Household income
- Frequency of Chinese restaurant

Different mean, different variance

Standardization



	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

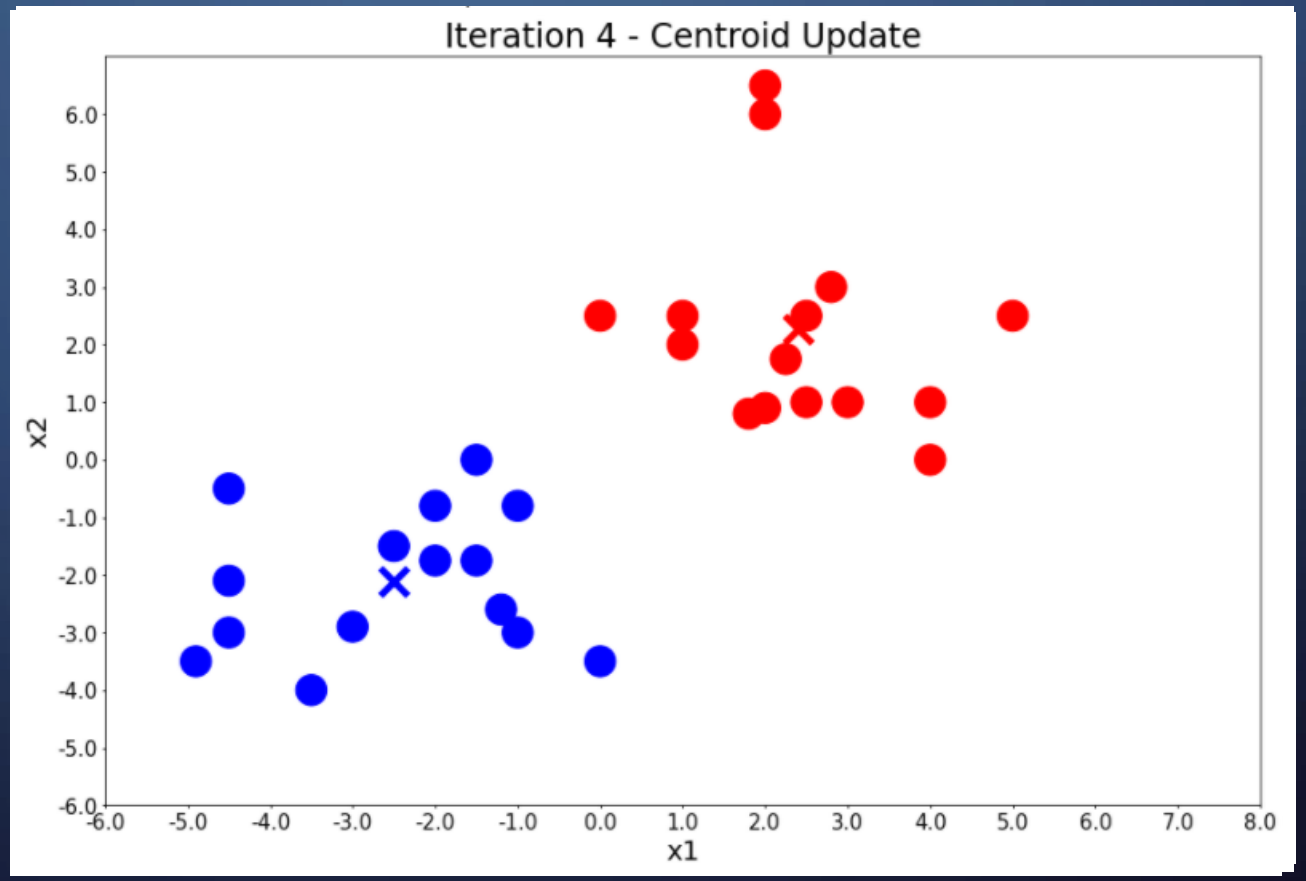
Machine Learning Modelling – Clustering problem:

II. K-means clustering algorithm

- One of the simplest unsupervised machine learning algorithm
- Need to specify the number of clusters, K

Example:

1) Random centroids initiation
2) For each point, compare the distance to different centroid
3) Assign point to closest centroid
4) Update centroid position
5) Repeat steps 2 – 4 until all iterations completed



Machine Learning Modelling – Clustering problem:

II. K-means clustering algorithm

- How to find the best number of clusters, K ?

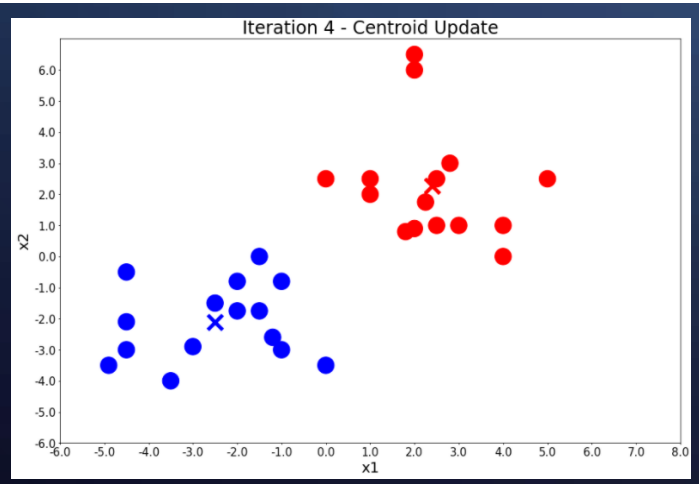
➡ The elbow method

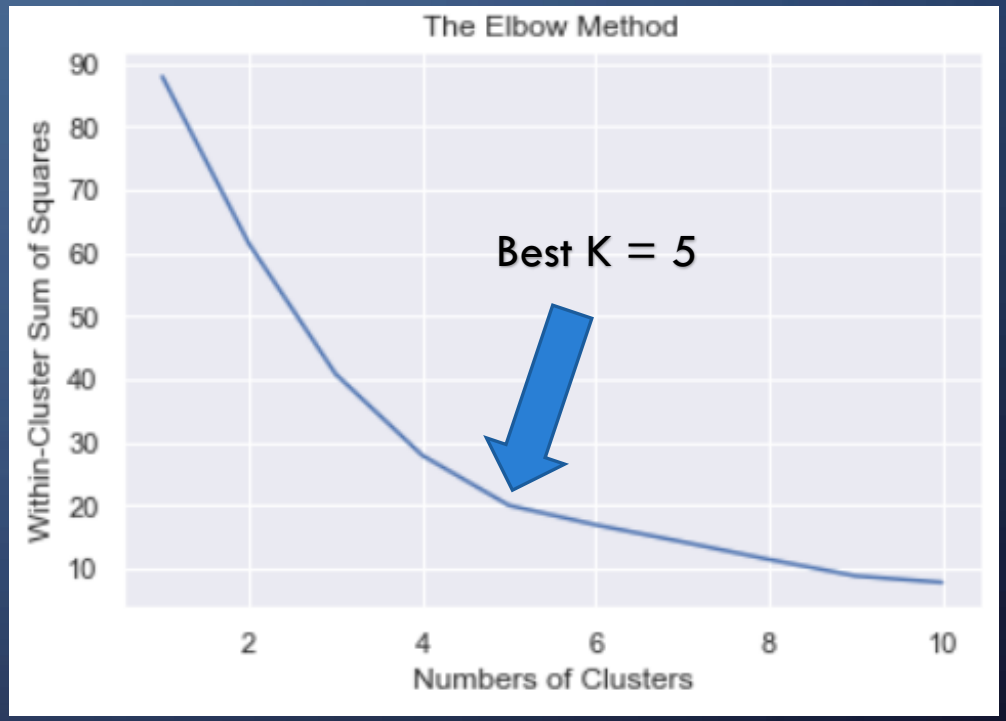
1. Decide a range of K
2. Run K-means algorithm
3. Calculate the WCSS (within-cluster sum of square)

$$\text{WCSS} = \sum_{C_k}^{C_n} \left(\sum_{d_i \text{ in } C_i}^{d_m} distance(d_i, C_k)^2 \right)$$

Where,

C is the cluster centroids and d is the data point in each Cluster.

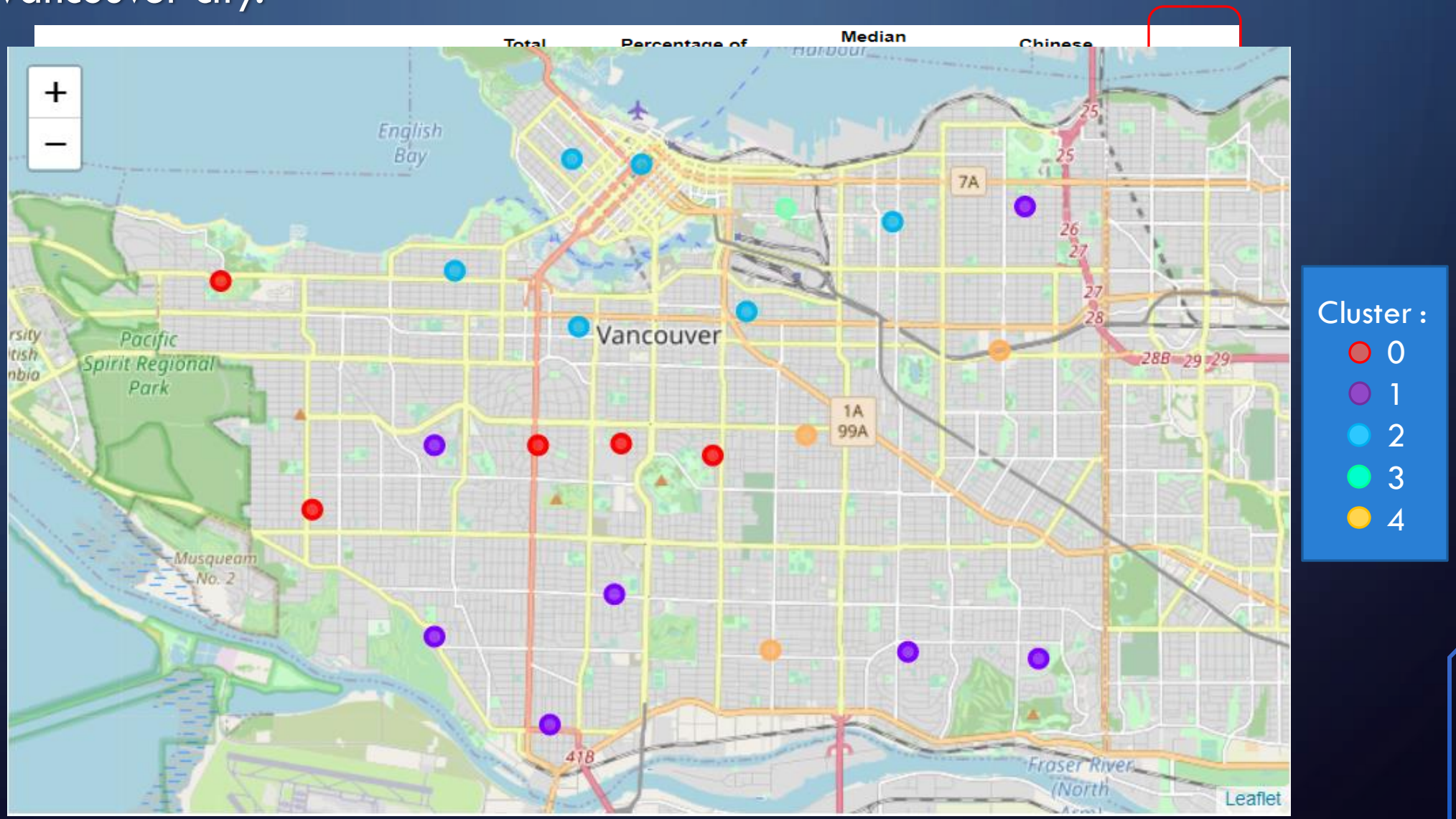




The best number of clusters, K is with a low WCSS and a low K value

Results :

with K = 5, we are able to get the cluster group for each neighbourhoods in Vancouver city.



Results :

	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
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
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
17	Strathcona	49.277693	-123.088539	9855	29.0715	21964	0.03	3
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

Frequency of Chinese Restaurant

Household income

	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

Conclusion :

	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

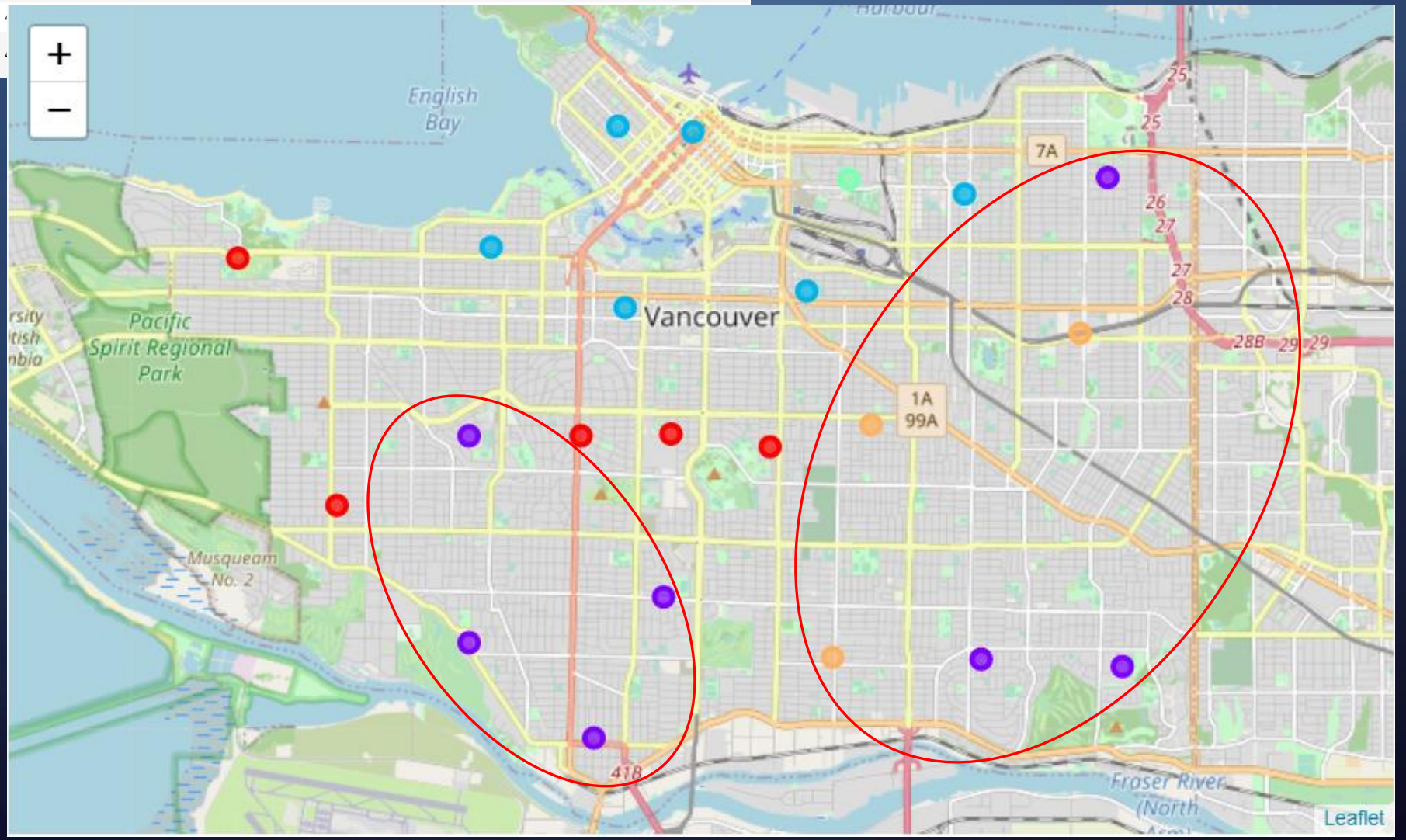
- Low Chinese population → less business for Chinese restaurant
- The best choice to establish a new Chinese restaurant : Cluster 1 neighbourhoods
  - Medium Chinese population with medium spending power and the Chinese restaurant competition there is not as competitive as Cluster 4.

- 2nd choice : Cluster 4 neighbourhoods
  - Although the Chinese restaurant competition here is very competitive, the Chinese population is also high at these neighbourhoods.
  - The high Chinese population can overcome the competition here.

Conclusion :

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							
19	Victoria-Fraserview							

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 :
- 0
- 1
- 2
- 3
- 4

Thank you