Vancouver Neighbourhood Analysis

Instruction

This project is to analyze the venues and crime rates of each neighbourhood in metro Vancouver, which is known as greater Vancouver. The City of Vancouver is located on the western half of the Burrad Peninsula. It is one of the most ethnically and linguistically diverse cities in Canada. The city is divided into 24 neighbourhoods (according to my analysis on the crime data of Vancouver). Each neighbourhood has many kinds of venue. [1] Mastering neighbourhood venue information and crime status may be vital to those people who consider selecting an area for living, staying or even opening a business in metro Vancouver.

Data Preparation

The City of Vancouver official website[2] provided crime data from year of 2003 to April 2019. I have extracted data of 2003 to 2018 for data integrity purposes. There are 24 unique neighbourhoods in the dataset. I calculated average crime records of each neighbourhood, presented on a bar plot. I used names of the neighbourhoods to locate related coordinates. However, there are very limited free API to get accurate coordinates. So I have tried to use the Geocoder Python package instead. The problem with the package was that not to get the complete geographical coordinates all the time. For solving this problem, I had to use a loop to continuously get the results, which could be very time consuming and it ended up with a dead loop. Therefore, I have decided to manually find the coordinates of each neighbourhood. [3]

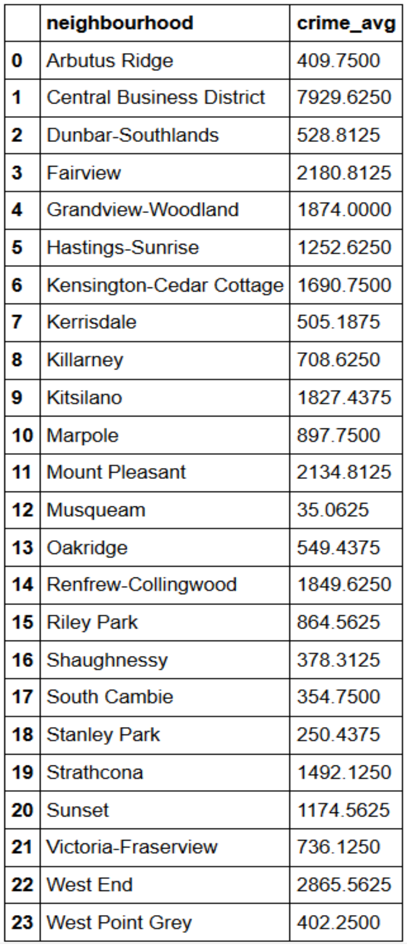
Forsquare API is location-based services. I used the coordinates to get venue information in each neighbourhood, and analyzed that venue information by sorting the top 10 common venues, and clustering the neighbourhoods.

Methodology

1. Organize the crime table. Get a neighbourhood data frame associated with average crime. (Figure 1)
2. Manually enter latitude and longitude values into the neighbourhood data frame. (Figure 2)
3. Present the neighbourhood data frame with average crime into a bar plot. (Figure 3)

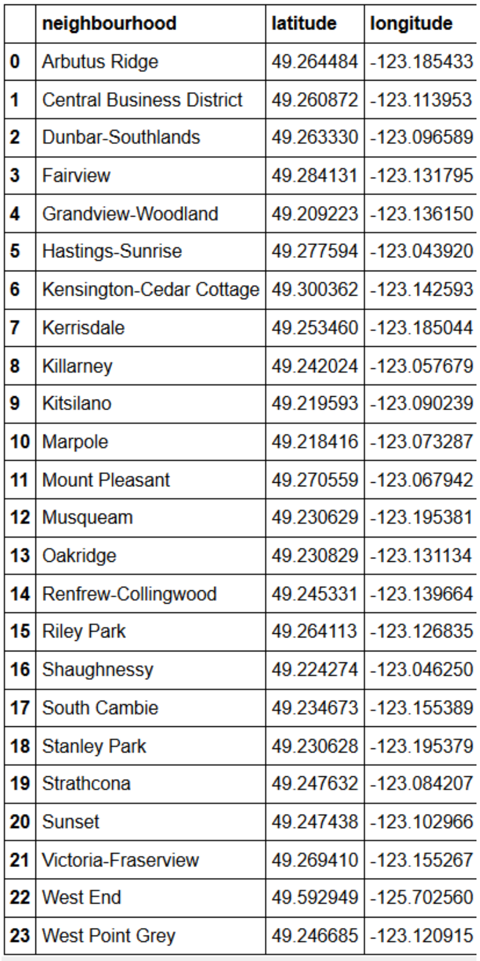
4. Create a map of Metro Vancouver by using python folium library. (Figure 4)
5. Mark the neighborhoods on the map of Metro Vancouver. (Figure 5)
6. Get nearby venues in each neighbourhood by calling Forsquare API. Requested json result shows there are five neighbourhoods have zero venues. In fact, there should have been many different venues. I am unsure at this point why I didn't get any venue information of those five neighbourhoods. Perhaps free calls don't get sound results? (Figure 6)
7. K-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data.[4] Cluster venues and visualize on the map by applying k-mean clustering. (Figure 7)
8. Finally, I have used two approaches to determine optimal k: exam clusters one by one, and elbow method (Figure 8). In this case, the elbow method may not be the best approach to find the optimal k as the line is almost straight.

Below are the forementioned figures:



	neighbourhood	crime_avg
0	Arbutus Ridge	409.7500
1	Central Business District	7929.6250
2	Dunbar-Southlands	528.8125
3	Fairview	2180.8125
4	Grandview-Woodland	1874.0000
5	Hastings-Sunrise	1252.6250
6	Kensington-Cedar Cottage	1690.7500
7	Kerrisdale	505.1875
8	Killarney	708.6250
9	Kitsilano	1827.4375
10	Marpole	897.7500
11	Mount Pleasant	2134.8125
12	Musqueam	35.0625
13	Oakridge	549.4375
14	Renfrew-Collingwood	1849.6250
15	Riley Park	864.5625
16	Shaughnessy	378.3125
17	South Cambie	354.7500
18	Stanley Park	250.4375
19	Strathcona	1492.1250
20	Sunset	1174.5625
21	Victoria-Fraserview	736.1250
22	West End	2865.5625
23	West Point Grey	402.2500

Figure 2 Average Number of Crime



	neighbourhood	latitude	longitude
0	Arbutus Ridge	49.264484	-123.185433
1	Central Business District	49.260872	-123.113953
2	Dunbar-Southlands	49.263330	-123.096589
3	Fairview	49.284131	-123.131795
4	Grandview-Woodland	49.209223	-123.136150
5	Hastings-Sunrise	49.277594	-123.043920
6	Kensington-Cedar Cottage	49.300362	-123.142593
7	Kerrisdale	49.253460	-123.185044
8	Killarney	49.242024	-123.057679
9	Kitsilano	49.219593	-123.090239
10	Marpole	49.218416	-123.073287
11	Mount Pleasant	49.270559	-123.067942
12	Musqueam	49.230629	-123.195381
13	Oakridge	49.230829	-123.131134
14	Renfrew-Collingwood	49.245331	-123.139664
15	Riley Park	49.264113	-123.126835
16	Shaughnessy	49.224274	-123.046250
17	South Cambie	49.234673	-123.155389
18	Stanley Park	49.230628	-123.195379
19	Strathcona	49.247632	-123.084207
20	Sunset	49.247438	-123.102966
21	Victoria-Fraserview	49.269410	-123.155267
22	West End	49.592949	-125.702560
23	West Point Grey	49.246685	-123.120915

Figure 1 Neighbourhood Coordinates

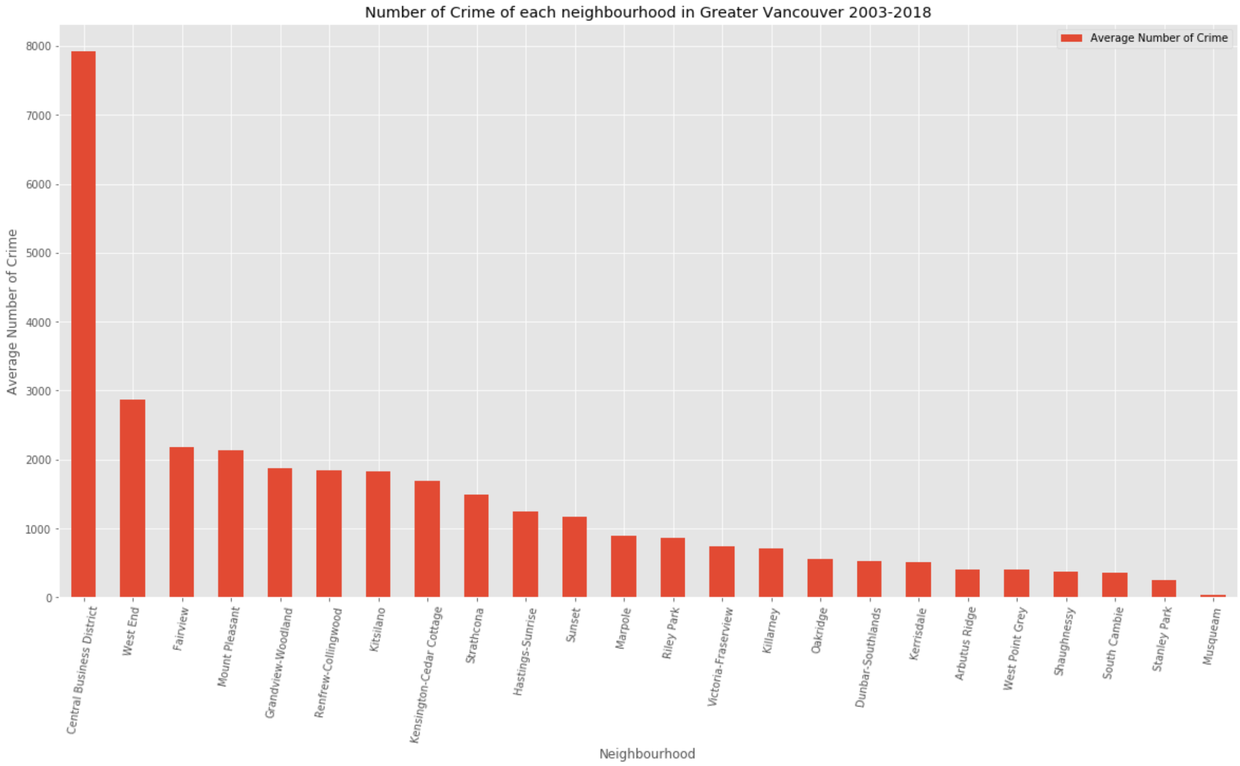


Figure 3 Average Crime of Neighbourhood

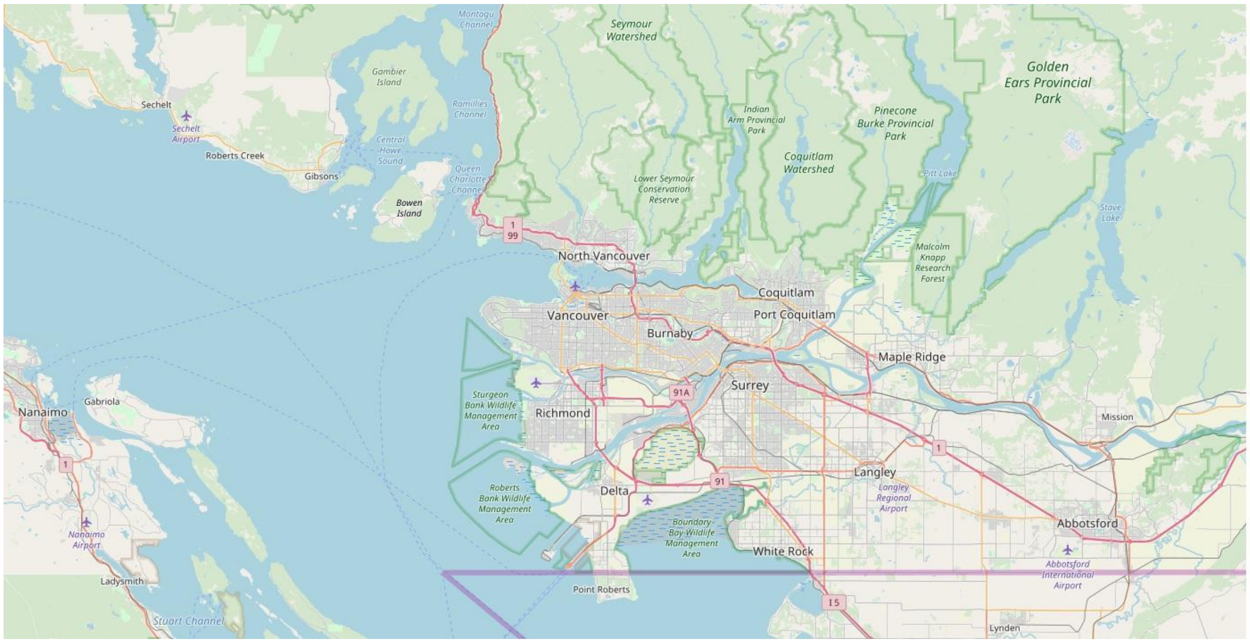


Figure 4 Vancouver Map

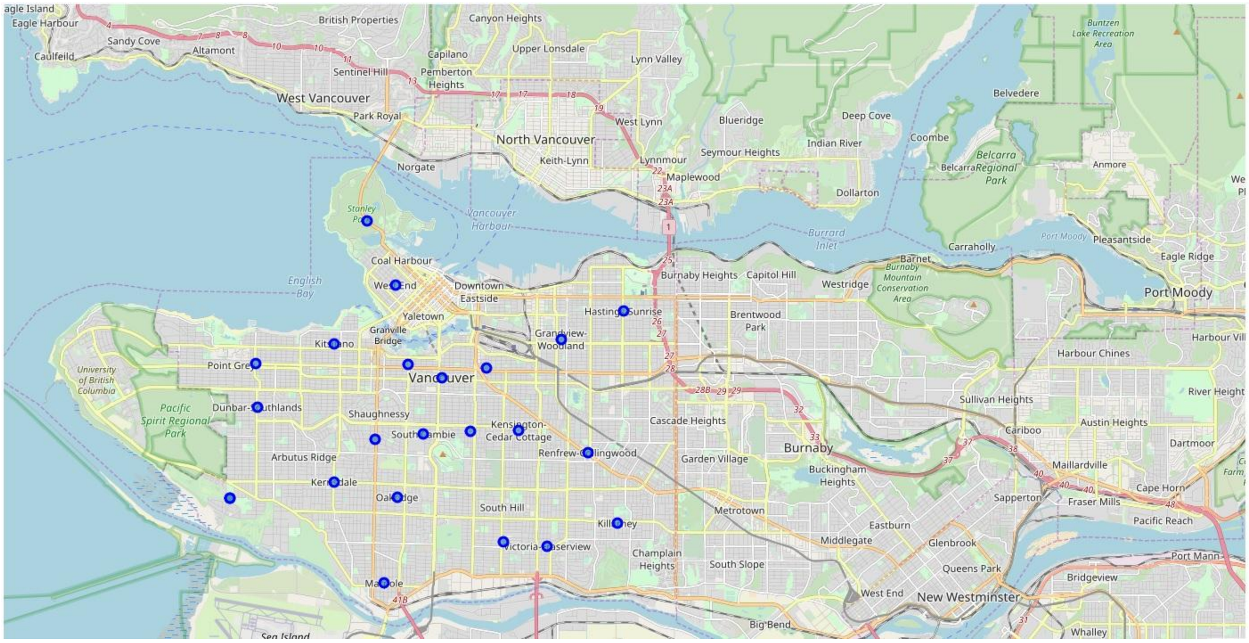


Figure 5 Vancouver Neighourhood

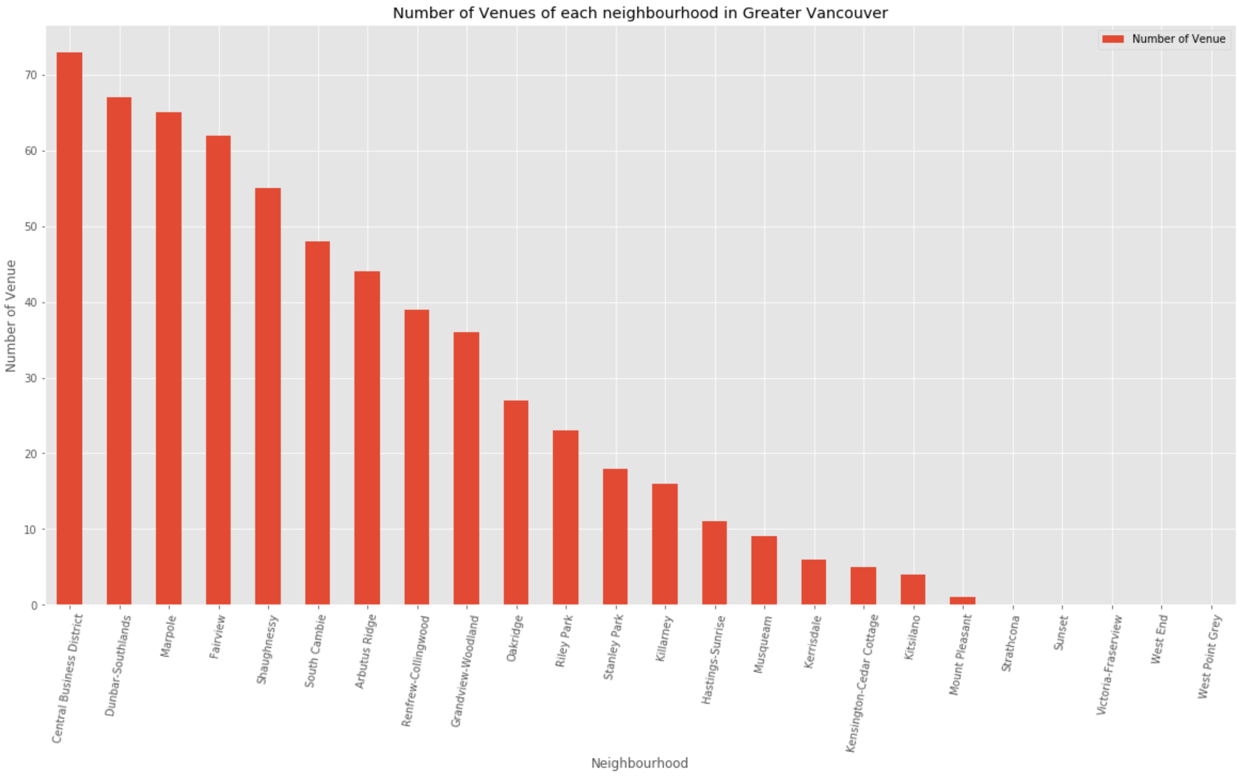


Figure 6 Venue

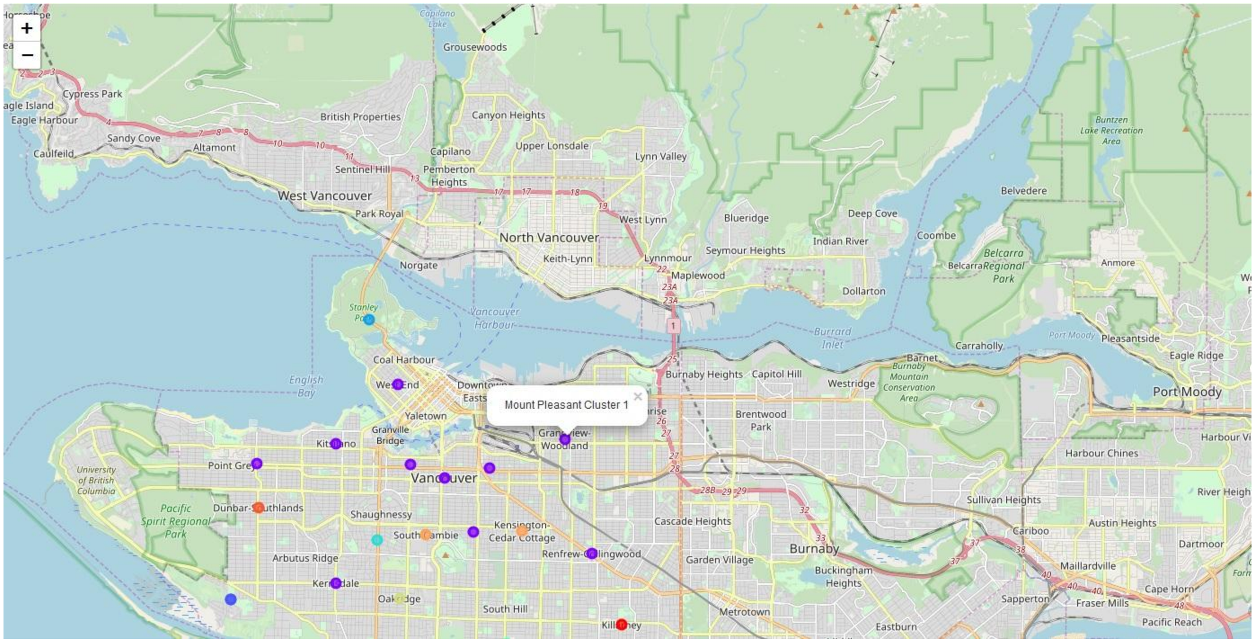


Figure 7 Neighbourhood Clustering

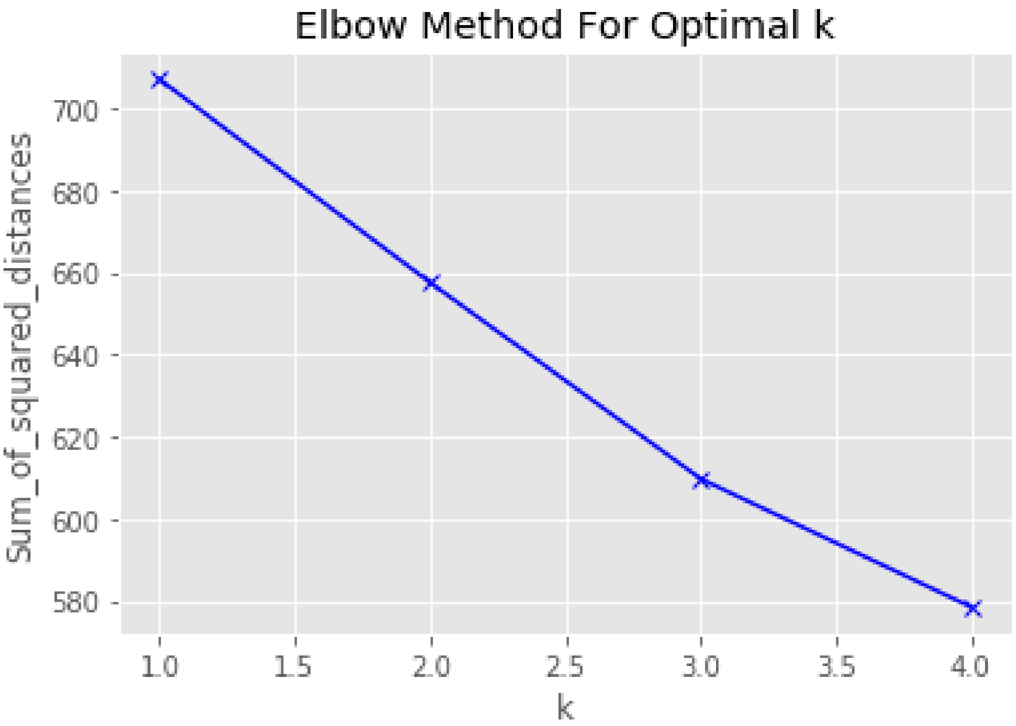


Figure 8 Optimal K

Discussion

Those five neighbourhoods without any venue information made me consider the accuracy of json result from Forsquare API by using free calls, which may impact later analysis on neighbourhood clustering. Further research is needed to explore more on requested json result via Forsquare API and possible impacts on neighbour clustering.

Conclusion

For this project, which way is the best to cluster neighbour venues depends on personal favorite. For instance, if someone would like to open a yoga studio in Vancouver, perhaps the preferred neighbourhood would be less competition, low crime rate, busy neighbourhood etc. Cluster 1, 4, 5 show yoga studio is one of top 5 most common venue in many areas. Cluster 2 shows yoga studio/gym or any other activity facilities less common but many dining options. Therefore, the optimal k would be 2 to cluster the venue.

References

[1] Python code is shared at my GitHub repository.

[2] The City of Vancouver website: https://vancouver.ca/

[3] https://www.latlong.net/

[4] Applied Data Science Capstone - k-means Clustering lab