Analysis of population and Venue distribution in Hong Kong

A study for finding potential location for new business

Tse Man Kit Michael

1. Introduction

Hong Kong is a highly developed territory, which ranks fourth on the UN Human Development Index. With over 7.4 million people of various nationalities in a 1,104-square-kilometre (426 sq. mi) territory, Hong Kong is one of the most densely populated places in the world. The dense space also led to a developed transportation network with public transport rates exceeding 90%. Hong Kong is ranked third in the Global Financial Centre Index, behind New York City and London. As it is a very important international financial, business services and recognized centre in the world, talents, and investors from worldwide are being concentrated in this place.

From the businessman's point of view, which location is the best for the new business is the key question. A retail store should be opened at a location with a high population but less similar varieties in the surrounding; an office should be located somewhere easy to go (which is not a problem in HK) and surrounded by restaurant, but also with a reasonable price of rent. Keeping the above things in mind that it is very difficult for an individual to find such a place in such a big city and gather this much information.

This report will visualize the population and venue information through a map and conclude on some potential locations for different venues.

2. Data

2.1 Data Sources

For this report, I followed the region(neighbourhood) definition in Hong Kong population distribution data, separating Hong Kong into 376 neighbourhoods in 18 Districts. The sources of data are listed as follows:

- Hong Kong neighbourhoods' names list: defined in Hong Kong population distribution data
- Hong Kong population data: https://www.bycensus2016.gov.hk/en/bc-own_tbl.html
- Neighbourhoods' Geographic Coordinates: Nominatim API
- Venues data: Foursquare API
- District boundary geojson: https://data.gov.hk/en-data/dataset/hk-had-json1-hong-kong-administrative-boundaries/resource/5b64f6ae-8827-444c-a092-b227112ab3ab

The Hong Kong population data was downloaded from the Hong Kong government website, and the neighbourhoods' definition of this report followed which in this file. The geographic coordinates of each neighbourhood were found by Nominatim API. A Foursquare API GET request is sent in order to acquire the surrounding venues that are within a radius of 1000m.

2.2 Data cleaning

The csv file of population of each district was very clear and not need for cleaning.

However, the list of neighbourhoods of each district needs to be cleaned and modified. The principle of defining neighbourhoods in the raw data set is partially based on the population, which is not useful in this project. This lead to 3 problems: some of the neighbourhoods like 'Tsim Sha Trui East' & 'Tsim Sha Trui West' which are actually the same; some are too near to each other; some cannot find a single geographic coordinates.

So, first of all, for the names of neighbourhood having any of the words at the end: 'North, East, South, West, Upper, Lower, I, II' , I deleted that word.

Secondly, I manually checked every neighbourhood, if any of them are too near to each other or hard to find a geographic coordinates using Nominatim API, I manually assigned a location.
Thirdly, if any one of the neighbourhood cannot find the coordinates using Nominatim API, then the district coordinates will be used, this can ensure that every neighbourhood have coordinates being assigned.

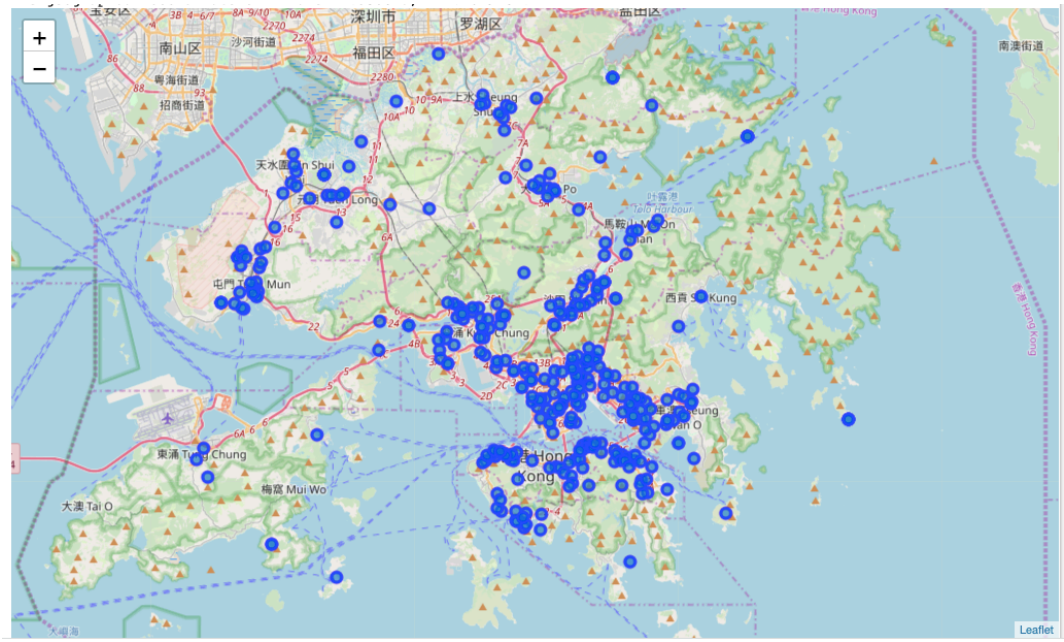
The categories of venues need to do modification manually, because some of the categories are not well enough, such as 'shops'(this is too general), or 'Food Court' (this should be classified as 'food').

3. Methodology

3.1 Exploratory Data Analysis

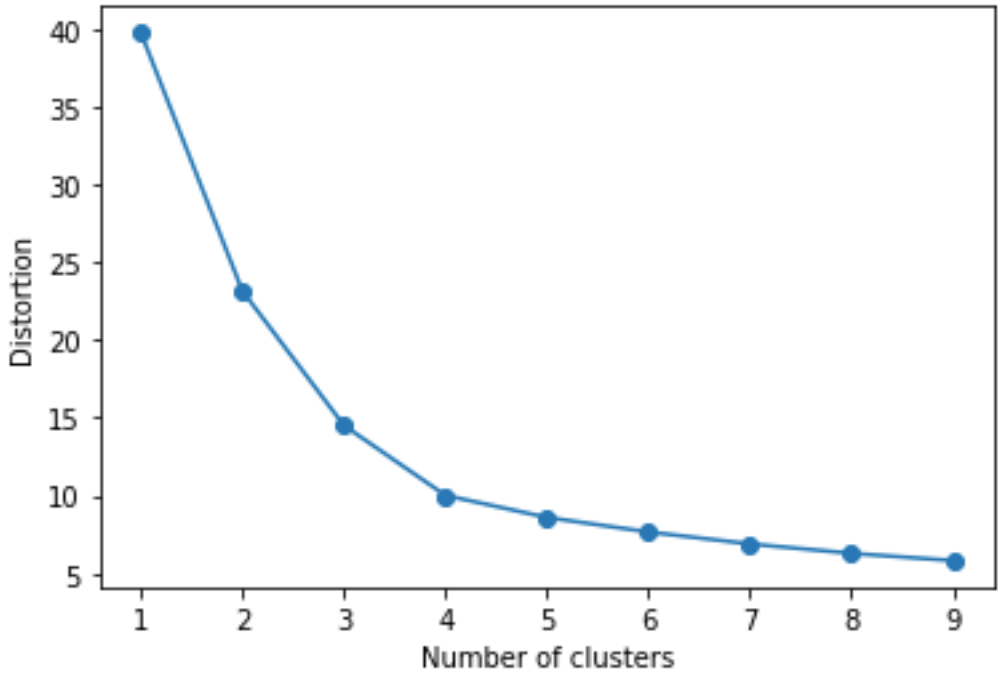
As the geojson file contain only 18 districts' boundaries are contained, I cannot draw all 430 neighbourhoods' regions on the map. So, assuming the population of each neighbourhood in the same district are uniform which is a good assumption since the neighbourhoods were originally defined in that way, the population of 18 districts is shown as a choropleth map and the neighbourhoods are represented as points with colours. The colours represent the cluster of the neighbourhood with respect to the others.

As discussed in 2.2 Data cleaning, some of the raw neighbourhoods share the same name; some are too near to each other; some cannot find a single geographic coordinates. Some of the categories of venues are not well enough, such as 'shops'(this is too general), or 'Food Court' (this should be classified as 'food'). So Data cleaning is done.
Data cleaning quality can be verified by data visualization. As in the figure, all main neighbourhoods (which means having a relatively high numbers of varieties) are covered. This can ensure that the varieties distribution in the later discussion is representable.



Since some geo-coordinates of neighbourhoods are replaced by a more general geo-coordinate, and the distribution of varieties of each neighbourhoods are very divergence, a larger radius (r = 1000) for each neighbourhood is used.
This may lead to the same venue is double counted by two nearby neighbourhoods. However, normalized mean value of frequency of venues categories is used and hence the effect on double counting is minimized.

The neighbourhoods are clustered into ‘K’ groups in order to compare between neighbourhoods. The value ‘K’ is chosen by elbow method. The plot of sum of square distance v.s. K is shown below, and the elbow is at k = 4.



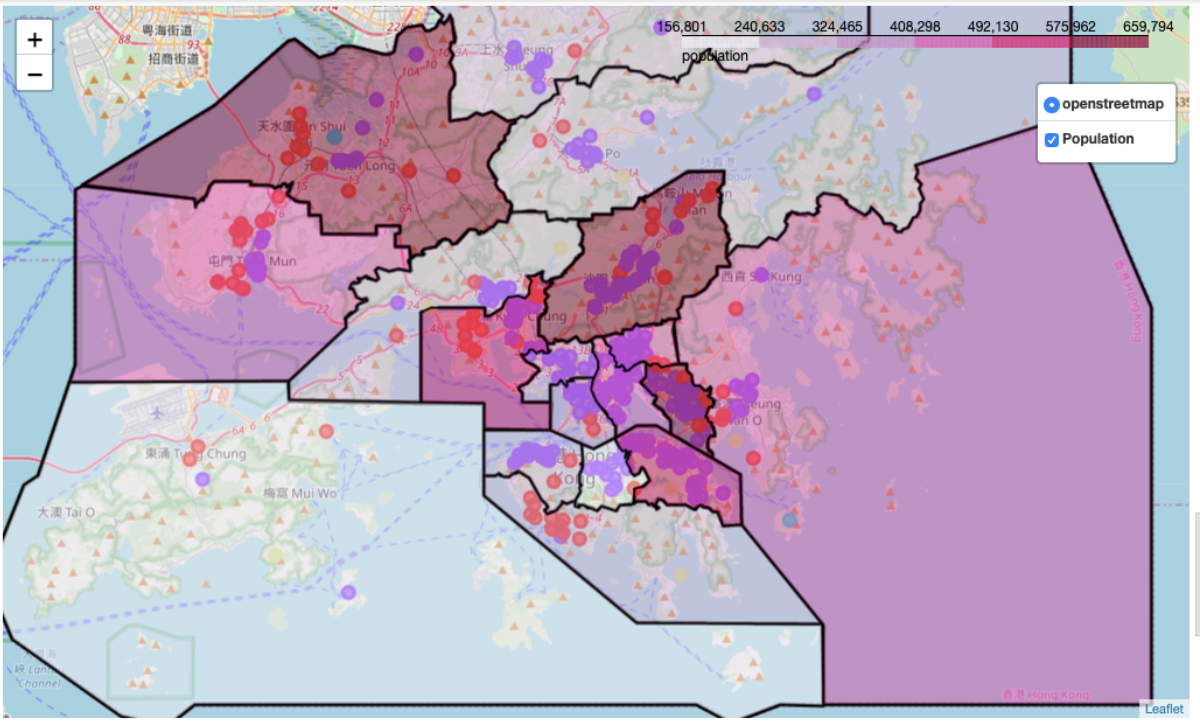
After examining each cluster, the categories of each labels are as following:
0: city with more hotels
1: city with more shopping mall
2: in between countryside city, with more hotels, café, farmer markets
3: countryside with mostly park

4. Results

The highest population is Sha Tin, second is Kwun Tong, third is Yuen Long.

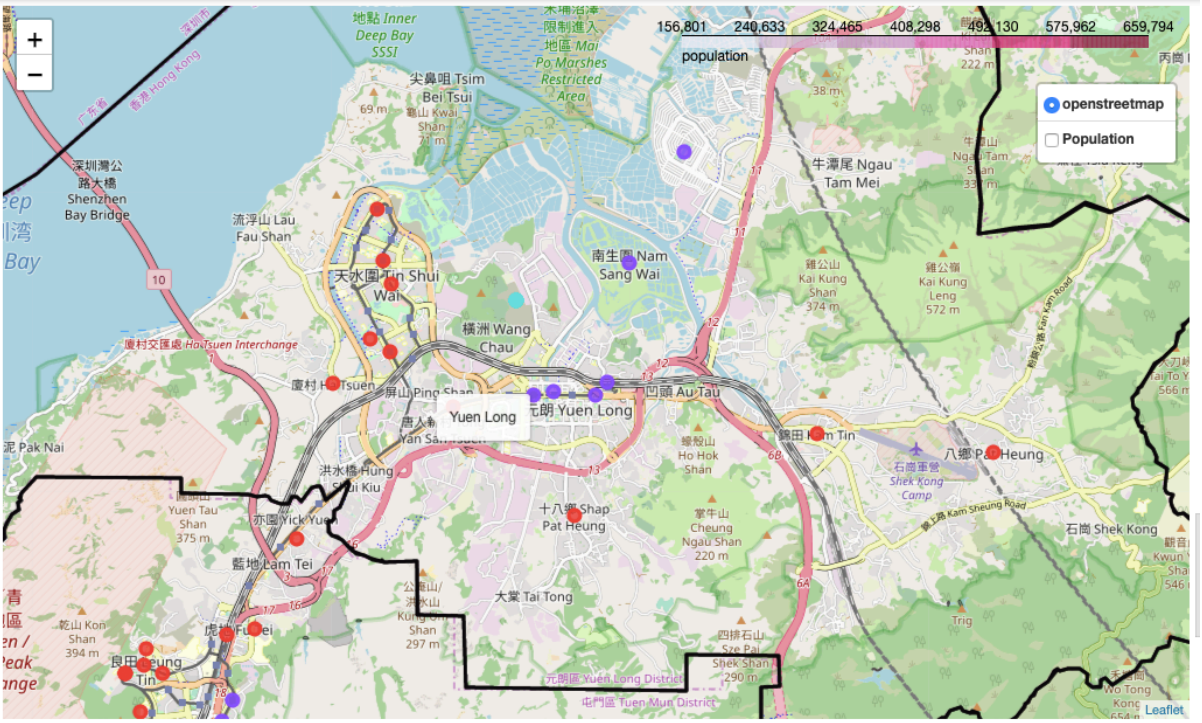
	neighborhood	population
15	Sha Tin	659794
8	Kwun Tong	648541
12	Yuen Long	614178
2	Eastern	555034
9	Kwai Tsing	520572
11	Tuen Mun	489299
16	Sai Kung	461864
7	Wong Tai Sin	425235
6	Kowloon City	418732
5	Sham Shui Po	405869
4	Yau Tsim Mong	342970
10	Tsuen Wan	318916
13	North	315270
14	Tai Po	303926
3	Southern	274994
0	Central and Western	243266
1	Wan Chai	180123
17	Islands	156801

Below is the final map showing population as choropleth, neighbourhood as points where label 0: red, label 1: purple, label 2: blue, label 3: green.



As shown in the map, Yuen Long District is high in population but fewer shops. Wang Chau is even in between countryside and city which have not many

restaurants.



5. Discussion

The definition of neighbourhoods based on the gov population result is too fine, which may lead to failure in generating the geographic coordinate through geocoding API. That's why large work on data cleaning was done in this project. This could be improved if a geocoding API with more accurate and detailed information about HK is found, such as Google Maps Geocoding API (but it is not free to use).

The choropleth map can be finer defined, as there is an official source about the neighbourhood's boundaries, but takes time to turn it into suitable format.

Note that the population data are showing where Hong Kong people live, but most of the HK people are working far away from their home. In other words, for another kind of business, such as fitness studio which needs a high stream of people, may need to consider data other than population in districts. Such as the rate of public transport usage in districts which can simulate the stream rate of people.

6. Conclusion

According to the final map shown in 4.Results section, Yuen Long may have a high potential in further development and hence it is a good location to start a residential related business, such as restaurants or stores.

As a result, there is an increasing trend of people coming into HK. For this reason, people can achieve better outcomes through their access to platforms where such information is provided.
Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.