Medical institutions analysis in Singapore

Capstone Project
IBM Data Science Professional Certificate
Gaël Brunel

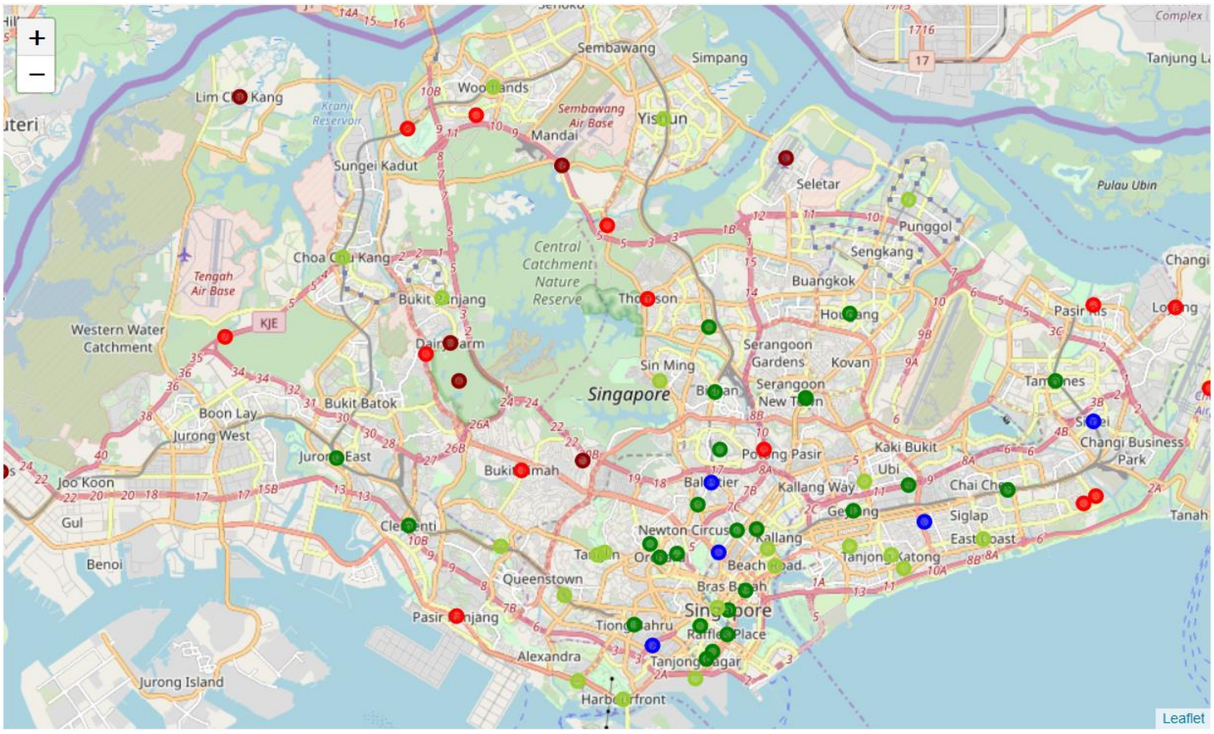


Table of Contents

Introduction

Singapore is an island city-state off southern Malaysia where approximately 6 million people are living in. It is one of the most important hubs of Southeast Asia, it possesses the world second busiest port in the world, and it is the place of choice of many multinationals who decide to establish in Southeast Asia.

I just spent one year in Singapore as part of my study and through this year I worked part time with an insurance company in Singapore. Their goal was to provide very personalized insurance for expats in Singapore, because healthcare is very expensive in Singapore, there is an important need for people to choose on what they want to be insured.

Moreover it is important for people to know if there is an easy access to health services from where they live, when an expat wants to settle in a new country we want to advise him where are the neighbourhoods that are close to medical centre, hospital, doctor's office, etc…

The problem I want to solve here is, how can I get an evaluation of the access to health services of each neighbourhood of Singapore. With this evaluation I will be able to tell if a neighbourhood is poorly provided or on the contrary has good access to any kind of health services.

Data

In order to solve this problematic I used the data listed below:

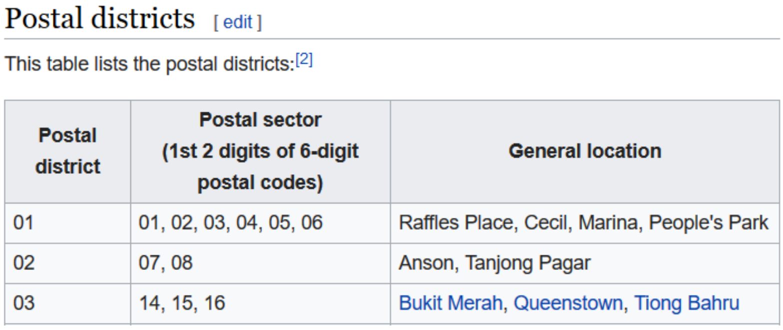
- First to get the name of every neighbourhood of Singapore, I went of the Wikipedia page listing all the postal code of Singapore, I saw that they provided a tab containing every postal district with their neighbourhoods in Singapore. https://en.wikipedia.org/wiki/Postal_codes_in_Singapore
- In order to get the geolocation of every neighbourhood of Singapore, I used the geocoder library
- I used the Foursquare API to get all the healthcare services around each neighbourhood

Methodology

1. Scrapping of the Wikipedia page containing all the post code of Singapore

In order to have an evaluation of the access to medical services of each neighbourhood of Singapore, I needed to have their name, and their localisation.

First, I went on the Wikipedia page listing all the postal code of Singapore, there I found a tab containing every postal district with their neighbourhoods in Singapore. I used BeautifulSoup4, a library that makes it easy to scrape information from web pages. From this library I obtained a parse tree where I retrieve only the tab with the name of each neighbourhood and their respective postal district code.

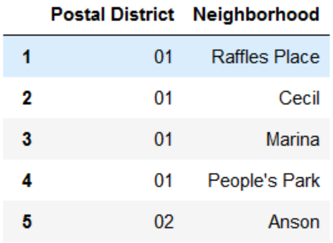


Postal districts [edit]

This table lists the postal districts:[2]

Postal district	Postal sector (1st 2 digits of 6-digit postal codes)	General location
01	01, 02, 03, 04, 05, 06	Raffles Place, Cecil, Marina, People's Park
02	07, 08	Anson, Tanjong Pagar
03	14, 15, 16	Bukit Merah, Queenstown, Tiong Bahru

Part of the tab from the Wikipedia page



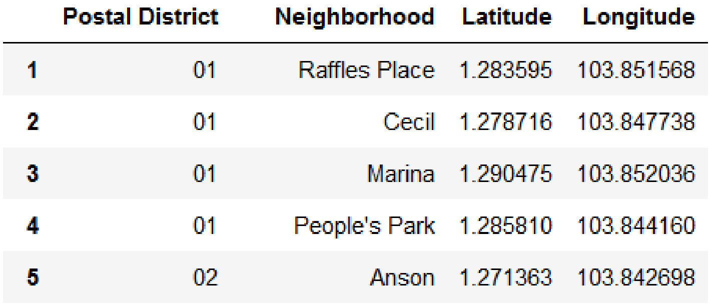
	Postal District	Neighborhood
1	01	Raffles Place
2	01	Cecil
3	01	Marina
4	01	People's Park
5	02	Anson

Datframe obtained after scrapping

2. Visualize Neighbourhood of Singapore on a map

After obtaining the name of the different neighbourhoods of Singapore, I wanted to be able to visualise all these neighbourhoods. To achieve this, I needed to have the coordinates of every neighbourhoods and then plot these coordinates on a map of Singapore.

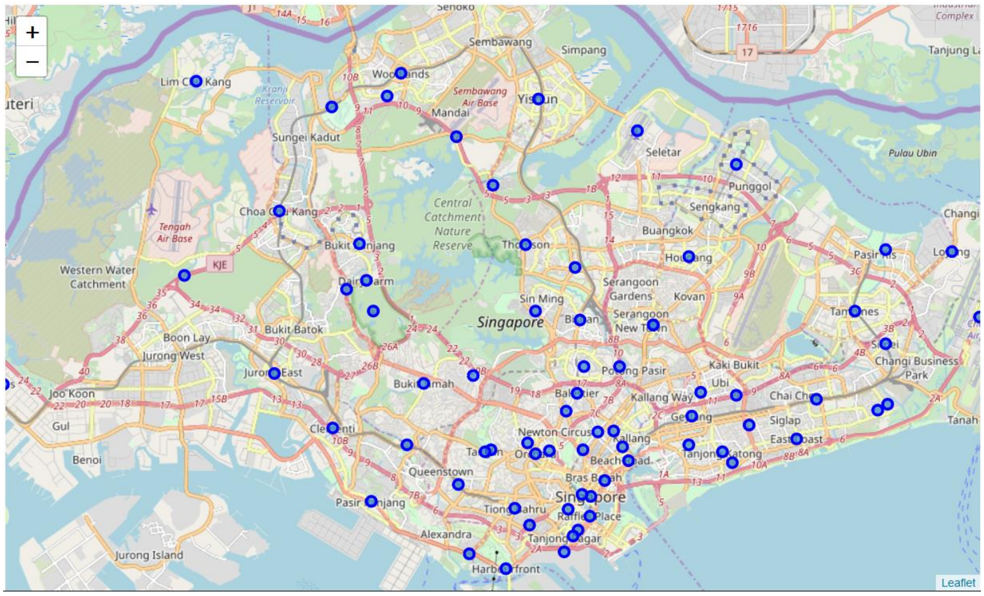
I used Geopy, a python client making easier to use some popular geocoding web services. For this project I used Nominatim, a geolocation service like Google Maps, but free. I execute a research on each neighbourhood and then add the obtained coordinates to my initial tab



	Postal District	Neighborhood	Latitude	Longitude
1	01	Raffles Place	1.283595	103.851568
2	01	Cecil	1.278716	103.847738
3	01	Marina	1.290475	103.852036
4	01	People's Park	1.285810	103.844160
5	02	Anson	1.271363	103.842698

To visualize Singapore on a map, I used the Python Folium library. It permitted me to show a map at a precise location and to add markers on it, I centred the map on Singapore and add a marker for every neighbourhoods at the correct geolocation that I obtained from Geopy.

Here the result:



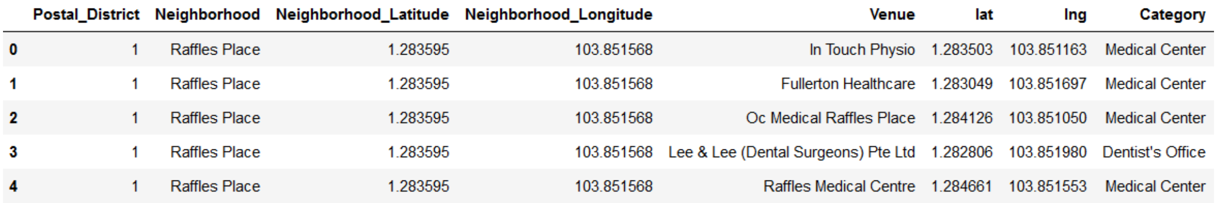
3. Exploring medical venues around neighbourhoods using Foursquare API

Now that we have all the location of the Neighbourhood of Singapore, we can start using Foursquare API to retrieve all the medical institutions around each neighbourhood.

We are only interested in medical services, so I restricted the research to only this kind of venue, for this I use a specific Id that we can find on the Foursquare website corresponding to 'Medical center'. (https://developer.foursquare.com/docs/resources/categories)

I designed the limit to 100 venues and the radius to 500 meters for each neighbourhood from their given latitude and longitude information.

Here a part of the table containing all the venues retrieved, corresponding to 2119 venues in all Singapore:

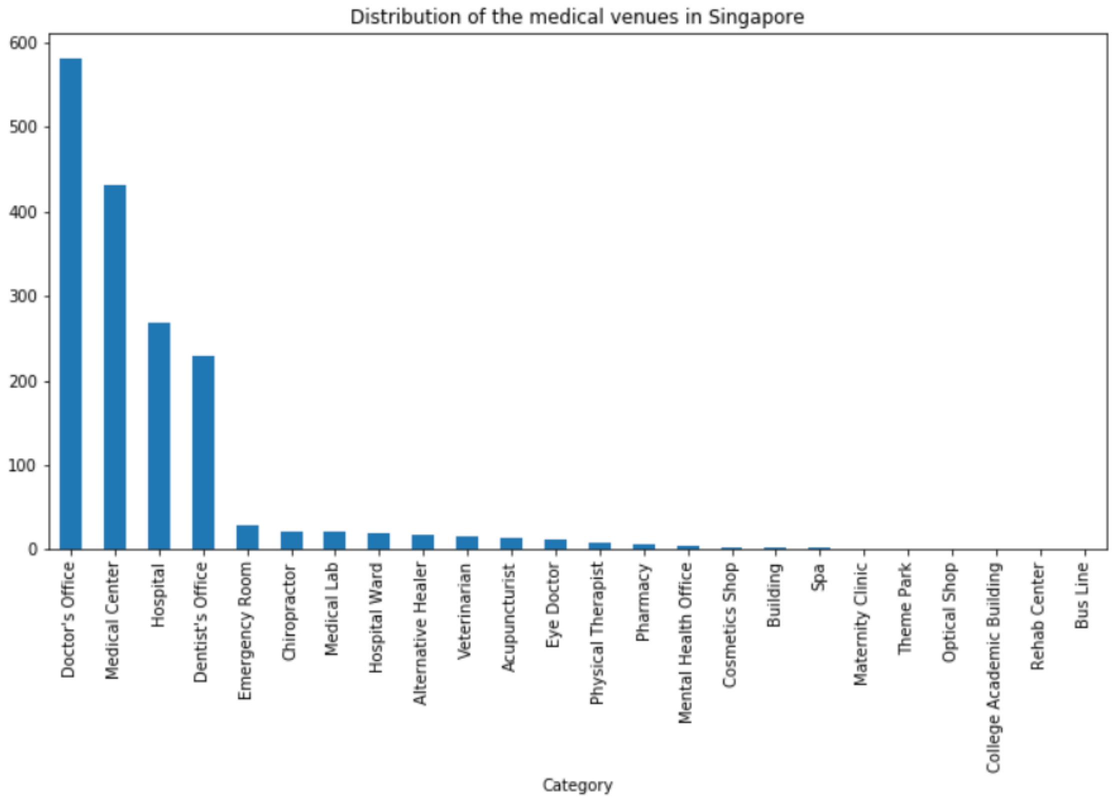


	Postal_District	Neighborhood	Neighborhood_Latitude	Neighborhood_Longitude	Venue	lat	lng	Category
0	1	Raffles Place	1.283595	103.851568	In Touch Physio	1.283503	103.851163	Medical Center
1	1	Raffles Place	1.283595	103.851568	Fullerton Healthcare	1.283049	103.851697	Medical Center
2	1	Raffles Place	1.283595	103.851568	Oc Medical Raffles Place	1.284126	103.851050	Medical Center
3	1	Raffles Place	1.283595	103.851568	Lee & Lee (Dental Surgeons) Pte Ltd	1.282806	103.851980	Dentist's Office
4	1	Raffles Place	1.283595	103.851568	Raffles Medical Centre	1.284661	103.851553	Medical Center

4. Analysis of each Neighbourhood of Singapore

a. Exploratory Data Analysis

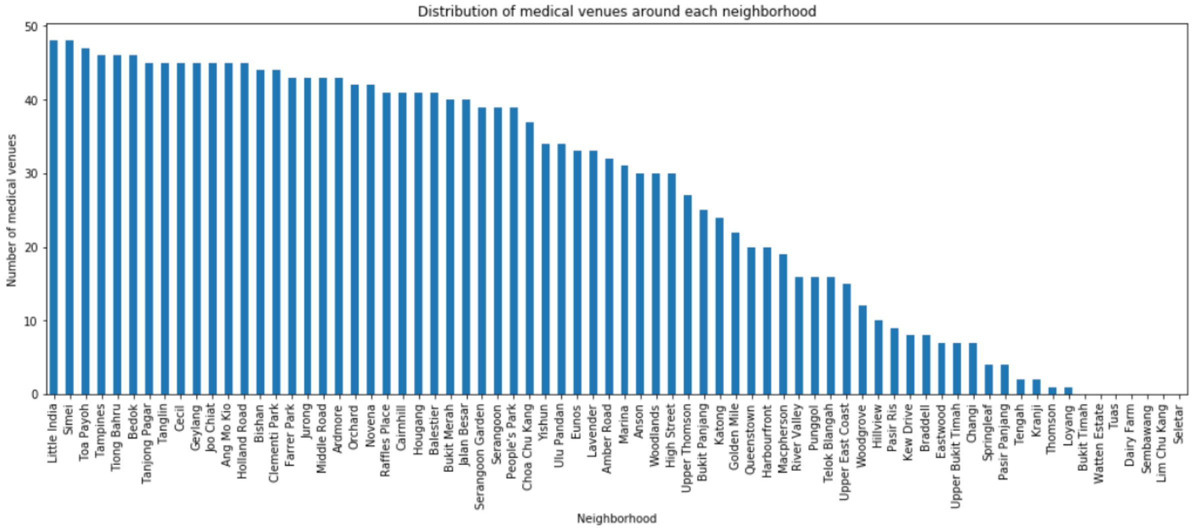
I started some Exploratory Data Analysis in order to have an overview and a first interpretation of the data I retrieved.



First, I visualized the distribution of the medical venues in Singapore, I saw that a lot of medical venues were poorly present in Singapore, like Maternity Clinic or Eye doctor. Moreover, Foursquare retrieved me some venues that are not at all medical venues, like a ‘Bus Line’ or a ‘Building’.

To clean the data, I removed all the categories that were not relevant, and I keep only the five venues that was the most present in Singapore, corresponding to 91.2% of the data.

Then I observed the distribution of medical venues around each neighbourhood:



We could see that there was an important difference between the neighbourhood, a lot was surrounded by more than 40 medical institutions which is very good, but on the other hand, a lot were poorly surrounded, some neighbourhoods didn’t even have any medical facilities around them.

b. Neighbourhoods clustering with unsupervised machine learning

My goal here is to be able to say if a neighbourhood is well surrounded by medical facilities or not, to answer this problematic, just looking at the distribution table of the medical venues around each neighbourhood is not enough.

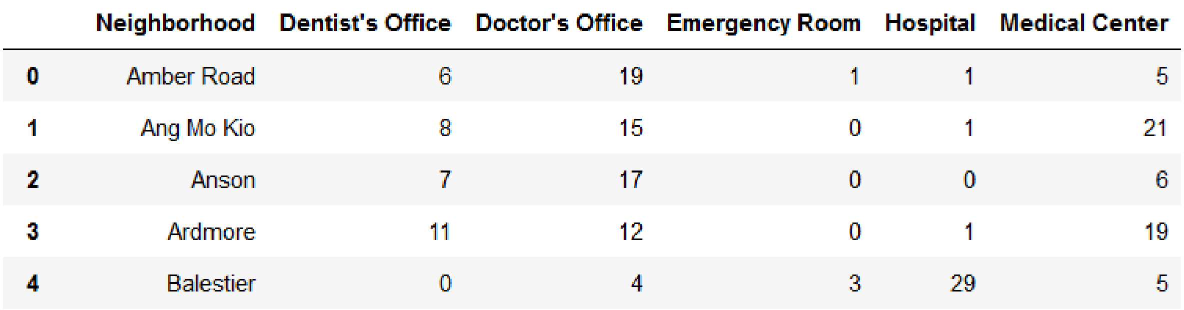
Because there is different kind of venues, and some of them are ‘more important’ than the others, it is normal to do a deeper analysis in order to evaluate the neighbourhoods. For example, there is much less hospitals than doctor’s Office, we can have 3 hospitals and 2 Doctor’s Office around us rather than 10 Doctor’s office, and both should be considered as well surrounded even if one has half of the number of venues of the other one because hospitals provide much more medical services than a Doctor’s Office.

For this reason, I used unsupervised learning, and more precisely, the **K-means algorithm** to cluster the neighbourhoods.

i. Data Pre-processing

To use the **K-means algorithm**, I needed to first do some Data Pre-processing. I transformed my data in order to only keep the information needed for the algorithm, I only keep the category of every medical venues around each neighbourhood. Then transform this categorical data in numerical data, sum all the venues from the same category in the same neighbourhood to have the frequency of each venues.

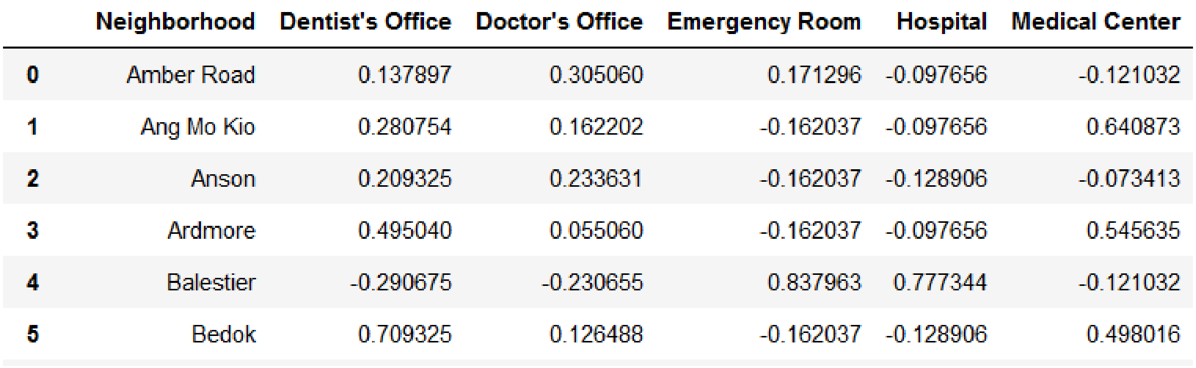
I also added the different neighbourhood that were not present in the tab because they were not surrounded by any medical venues, so I manually put their different venue frequencies at 0.



	Neighborhood	Dentist's Office	Doctor's Office	Emergency Room	Hospital	Medical Center
0	Amber Road	6	19	1	1	5
1	Ang Mo Kio	8	15	0	1	21
2	Anson	7	17	0	0	6
3	Ardmore	11	12	0	1	19
4	Balestier	0	4	3	29	5

Then I did **Data Normalization,** it is a very important pre-processing step, used to rescale values to fit in a specific range to assure better convergence. If we don't do this then some of the features (those with higher frequency than the others like Doctor's Office) will be weighted more in the cost function. The data normalization makes all features weighted equally.

I obtained this final tab that could be used for the K-means algorithm:

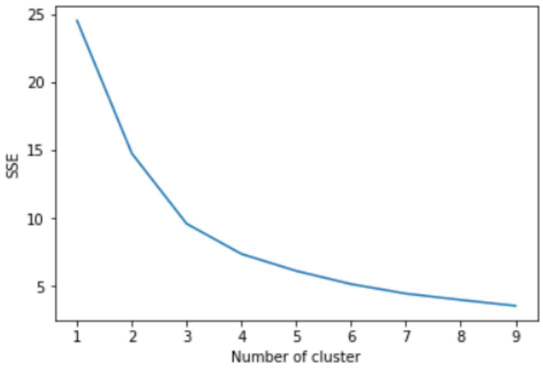


	Neighborhood	Dentist's Office	Doctor's Office	Emergency Room	Hospital	Medical Center
0	Amber Road	0.137897	0.305060	0.171296	-0.097656	-0.121032
1	Ang Mo Kio	0.280754	0.162202	-0.162037	-0.097656	0.640873
2	Anson	0.209325	0.233631	-0.162037	-0.128906	-0.073413
3	Ardmore	0.495040	0.055060	-0.162037	-0.097656	0.545635
4	Balestier	-0.290675	-0.230655	0.837963	0.777344	-0.121032
5	Bedok	0.709325	0.126488	-0.162037	-0.128906	0.498016

ii. K-means algorithm

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. It allows us to segment the data into K cluster, with K a number we choose. Each cluster containing very similar data, and very different data compare to the other clusters. This way we can separate the neighbourhood between well surrounded or not well surrounded.

To choose the correct number of K cluster to use I decided to use an evaluation method called the Elbow method. I plotted the distortion that we obtain after running the Kmeans algorithm with different number of clusters.

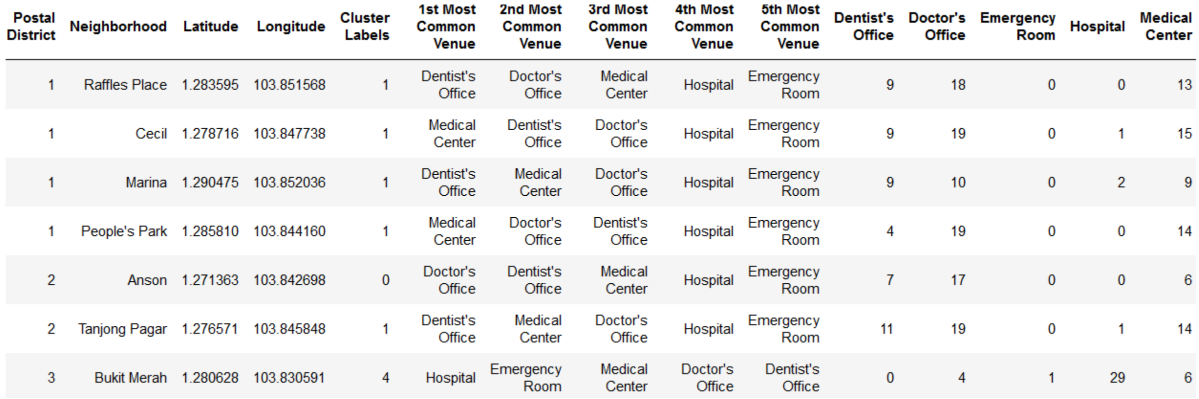


Unfortunately, as we can see on the graph, there is not a clear elbow point, where the distortion goes down rapidly until a certain point and then the distortion goes down very slowly. So, we can't really evaluate what would be the most appropriate number of clusters.

The evaluation being not relevant, I decided to run k-means to cluster the neighbourhood into 5 clusters. I choose to have a lot of clusters to obtain a better differentiation to know which area was poorly surrounded by medical venues. With 4 clusters some areas that I judge enough surrounded were in the same cluster than some areas very poorly surrounded by medical venues.

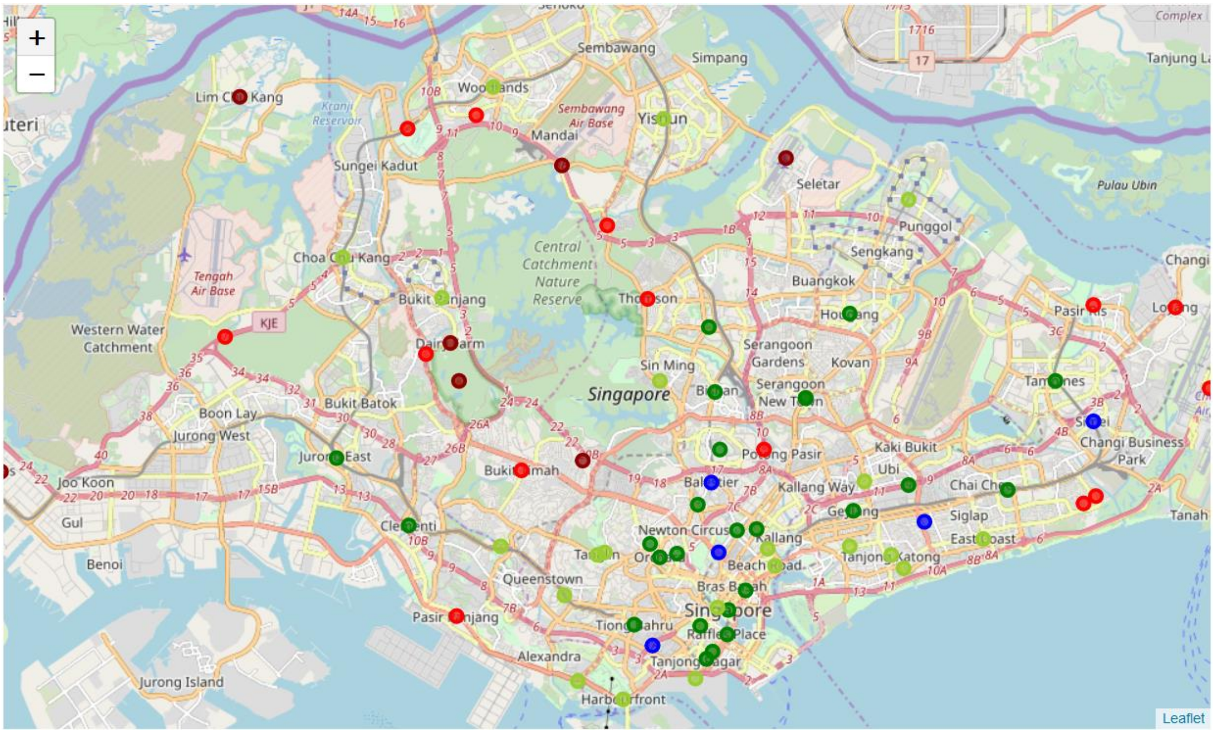
Results

As a result of the K-Means algorithm, each neighbourhood is classified under a certain cluster. I analysed by myself each of these cluster and was able to clearly differentiate which one was representing neighbourhood well surrounded by medical services or the opposite.



Postal District	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Dentist's Office	Doctor's Office	Emergency Room	Hospital	Medical Center
1	Raffles Place	1.283595	103.851568	1	Dentist's Office	Doctor's Office	Medical Center	Hospital	Emergency Room	9	18	0	0	13
1	Cecil	1.278716	103.847738	1	Medical Center	Dentist's Office	Doctor's Office	Hospital	Emergency Room	9	19	0	1	15
1	Marina	1.290475	103.852036	1	Dentist's Office	Medical Center	Doctor's Office	Hospital	Emergency Room	9	10	0	2	9
1	People's Park	1.285810	103.844160	1	Medical Center	Doctor's Office	Dentist's Office	Hospital	Emergency Room	4	19	0	0	14
2	Anson	1.271363	103.842698	0	Doctor's Office	Dentist's Office	Medical Center	Hospital	Emergency Room	7	17	0	0	6
2	Tanjong Pagar	1.276571	103.845848	1	Dentist's Office	Medical Center	Doctor's Office	Hospital	Emergency Room	11	19	0	1	14
3	Bukit Merah	1.280628	103.830591	4	Hospital	Emergency Room	Medical Center	Doctor's Office	Dentist's Office	0	4	1	29	6

Then I visualized the result on a map of Singapore showing the different neighbourhood with a specific colour depending on the classification of the neighbourhood.



Legend:

- Green: Very well surrounded by Dentist and doctor offices and medical centres
- LightGreen: Well surrounded by Dentist and doctor offices and medical centres
- Blue: Neighbourhood with lots of hospitals compare to the other area
- Red: Lack of medical services compare to the other area

- Darkred: Not a single medical facility is present around there

Discussion

Our analysis showed that Singapore is quite well deserved by medical institutions, with overall more than 2000 medical venues retrieved. And as we can see on the result map, we can really cluster Singapore into different zones just by looking at the segmentation that gave the medical venues. The city centre of Singapore is very well deserved, as opposite to the more external area.

We can easily guess a correlation on the activity and population of these neighbourhood just by looking at the result of my analysis on the medical venues.

We can really see here the efficiency of the K-means algorithm, it was able to clearly determine which neighbourhood have the most important Hospitals of Singapore (in blue), which neighbourhood we should or shouldn't go if we want to have easy access to healthcare.

Conclusion

As a result, this analysis was very useful for the insurance company I worked with, they were be able to advise new expats arriving in Singapore, which are the best place to live in, if they want to have an easy access to health or to an hospital.

It could be also useful for a local doctor who would like to know which place the best is to establish his office. For example, we can see easily see few neighbourhoods that are in the middle of an urban area with lots of habitation but with a very poor number of medical venues, like Upper Bukit Timah or Braddell.