Clustering The Hospital to Treat COVID-19's Patients
Case Study : South Korea

Raden Artha Alam Pribadi

June 1st, 2020

1. Introduction

1.1. Background

Recently, there has been an outbreak that turn our world upside down. COVID-19 (Corona Virus Disease 2019) is an infectious disease caused by a newly discovered coronavirus which has infected millions people around the world. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019. COVID-19 is now a pandemic affecting many countries globally. The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. You can be infected by breathing in the virus if you are within close proximity of someone who has COVID-19, or by touching a contaminated surface and then your eyes, nose or mouth.

South Korea is one of the most successful countries in dealing with this virus. Excellent medical infrastructure is one of the contributing factor which leads to success. It has been recorded that South Korea has 11,541 people infected and only 272 deaths occurred. Meanwhile, 10,446 people has recovered from this disease (update June 1st, 2020). Since mid-April until mid-May 2020, South Korea managed to reduce the COVID-19 growth rate to an average of 10 new cases per day.

1.2. Problem

One of the major and most impactful medical infrastructure is hospital. This study aims to cluster the hospital which available to treat COVID-19's patients according to the severity of their symptoms. This cluster can ease the burden of medical workers or local governments when distributing COVID-19's patients. Local governments can also focused the placement of the medical workers in hospital which handle high risk patients.

1.3. Interest

Obviously, the government of other countries which affected with this outbreak would be interested to know how to build the cluster in order to win the fight against COVID-19.

2. Data Wrangling

2.1. Data sources

The dataset of COVID-19 infection cases in South Korea can be found in Kaggle released by KCDC (Korea Centers for Disease Control and Prevention). This dataset already classified the cases based on their spread groups (for example, church, gym, apartment, etc) and also provide it with their latitude and longitude. Foursquare API will be used to plot these spread groups on the map, find nearby hospitals, and cluster them.

2.2. Data cleaning

There are a lot of missing values in the latitude and longitude features inside the dataset. The coordinate feature represent the exact location of each spread group feature (the group which suspected has accelerated the spread of the virus). Therefore, the coordinate feature will not have any value if the spread group feature can not be identified ("etc") or indicate the incoming virus from overseas ("overseas inflow"). The total row of coordinate feature that has missing value is 30%. I decided to only use data that has latitude and longitude values so it can be plotted on the map. Also, it means that this study only highlights the local transmission inside South Korea. Finally, I dropped the unnecessary columns that irrelevant to this study.

	province	city	infection_case	confirmed	latitude	longitude
0	Seoul	Yongsan-gu	Itaewon Clubs	133.0	37.538621	126.992652
1	Seoul	Guro-gu	Guro-gu Call Center	99.0	37.508163	126.884387
2	Seoul	Dongdaemun-gu	Dongan Church	20.0	37.592888	127.056766
3	Seoul	Guro-gu	Manmin Central Church	41.0	37.481059	126.894343
4	Seoul	Eunpyeong-gu	Eunpyeong St. Mary's Hospital	14.0	37.63369	126.9165

Figure 1. Top 5 rows of final data

3. Methodology

3.1. Exploratory data analysis

From the dataset, we can find what province that has the most COVID-19 cases. I grouped the data by province and found that Daegu is the province that has the most COVID-19 cases with the total case of 4971 or nearly halves the total case of COVID-19 in South Korea. Then, I visualized the distribution in a bar chart using Seaborn library.

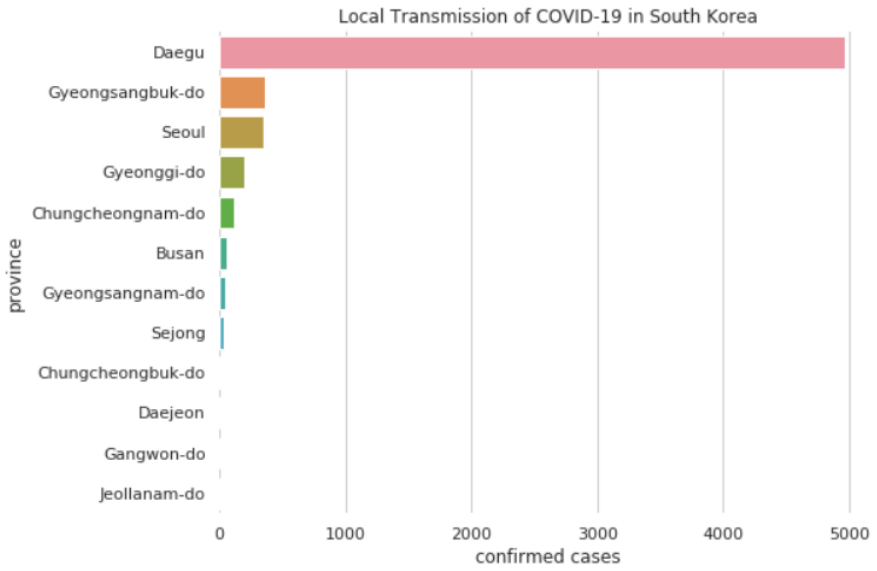

Figure 2. Local transmission of COVID-19 in South Korea

We can also derive what venue that accelerated the most of virus spread. From a quick look, I managed to find that church and hospital itself has been act as a medium to spread the virus further. In order to gain the exact number of COVID-19 cases classified by the venue, I calculated the confirmed cases by feature “infection_case” that contains string of “church” or “hospital”. Also, I visualized the result in pie chart using Seaborn library.

	spread_group	confirmed_cases	percentage
0	Church	4710	76.55
1	Hospital	674	10.95
2	Others	769	12.50

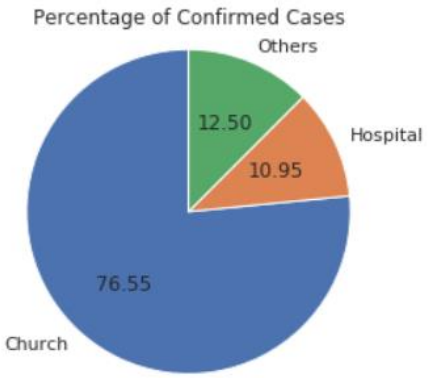

Figure 3. Total confirmed cases classified by venue

As we can see, Daegu province is the epicenter of COVID-19 cases in South Korea. Nearly halves (48% to be exact) of the COVID-19 total cases occurred in Daegu. Therefore, we will focused this study to identify and cluster the hospital around Daegu. I made a new dataframe that only consist data from Daegu and there are 5 spread groups identified with the highest number of confirmed cases come from Shincheonji Church.

	province	city	infection_case	confirmed	latitude	longitude
0	Daegu	Nam-gu	Shincheonji Church	4510	35.840080	128.566700
1	Daegu	Dalseong-gun	Second Mi-Ju Hospital	196	35.857375	128.466651
2	Daegu	Seo-gu	Hansarang Convalescent Hospital	128	35.885592	128.556649
3	Daegu	Dalseong-gun	Daesil Convalescent Hospital	100	35.857393	128.466653
4	Daegu	Dong-gu	Fatima Hospital	37	35.883950	128.624059

Figure 4. Spread groups and total confirmed cases in Daegu province

The spread groups above I plotted into a map using Folium library. Then, I determined the coordinate of Daegu as centroid of the map generated by Nominatim from Geopy library. Folium makes it easy to visualize data on an interactive leaflet map. Geopy is suitable for

Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.

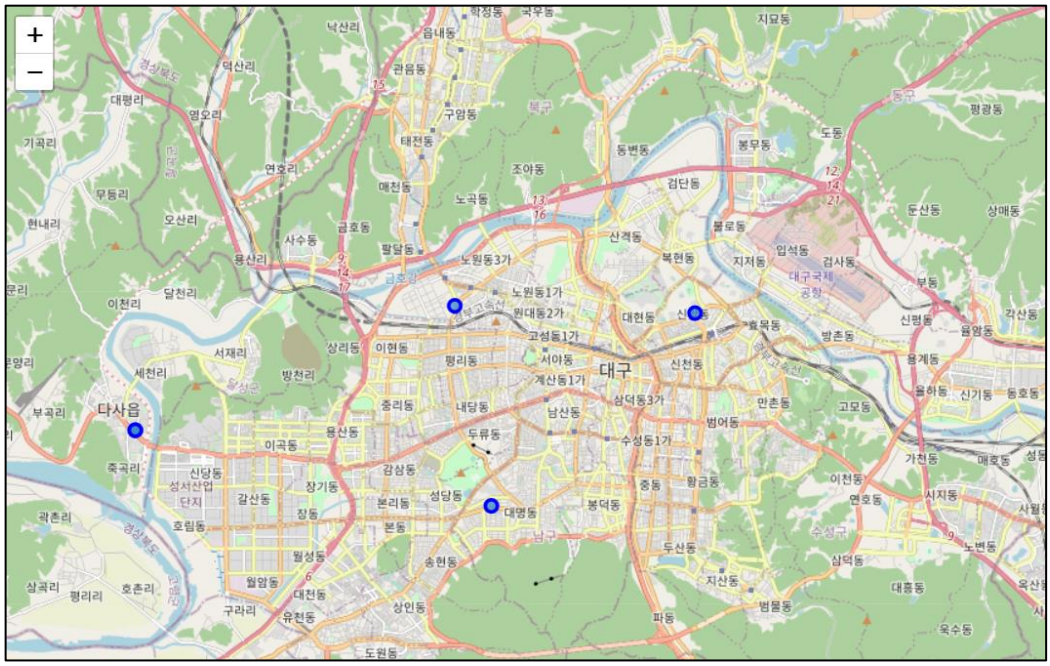

Figure 5. Daegu spread groups plotted

As we can see, the spread groups are located within the downtown of Daegu. The high risk area is located in the southern part of Daegu as 4510 people with COVID-19 live near Shincheonji Curch. After this, I need to find nearby hospital using Foursquare API.

3.2. Foursquare API

One of the methods to utilized spatial data is by accessing Foursquare API. Foursquare is a social location service that allows users to explore the world around them. Foursquare launched its API in November 2009, allowing application developers to extend the platform in interesting ways. Developers can build location management tools, custom search engines, and even games and other tools that interact with the Foursquare API.

I searched the hospital within 30 km radius from the Daegu spread groups. Why 30 km? I think it is maximum distance to transport patient without worsen their condition. The query result showed that there were 138 hospitals found. But, the result was mixed up because the

spread group is close to each other, so there were a lot of duplicate values. After I removed the duplicate values, there were 19 unique hospitals left in the dataset.

	Hospital	Hospital Latitude	Hospital Longitude
0	대구가톨릭대학교병원 (Catholic University of Daegu Hospi...	35.843541	128.568839
1	Hyosung Hospital (효성병원)	35.842730	128.612706
2	Sodaegu Hospital	35.869783	128.553078
3	Future Women Hospital	35.850686	128.537456
4	Kyungpook National University Hospital (경북대학교병원)	35.866562	128.604793
5	Goodyear Hospital	35.867560	128.621608
6	Keimyung University Dongsan Medical Center (계명...	35.868964	128.582900
7	Daegu Fatima Hospital (대구 파티마 병원)	35.883748	128.623681
8	Union Hospital (유니온병원)	35.872249	128.603750
9	Moon Medical Care Hospital (문요양병원)	35.915609	128.548146
10	Rosemary Women & Children's Hospital (로즈마리병원)	35.943261	128.557775
11	하양삼성병원 / Hayang Samsung Hospital	35.914852	128.825968
12	Gumi Gangdong Hospital (구미강동병원)	36.096262	128.423090
13	Kabul Hospital (갑을구미병원)	36.107246	128.402776
14	WELLKIDS Chilndren's Hospital (웰키즈아동병원)	36.106975	128.420729
15	Soon Chun Hyang University Hospital Gumi (순천향대...	36.102840	128.382613
16	Gumi Cha Hospital (구미차병원)	36.114023	128.340177
17	Hyeonggok Children's Hospital (형곡아동의원)	36.113870	128.338730
18	SCH Pediatric Hospital (순천향소아과의원)	36.121227	128.374188

Figure 6. List of nearby hospitals in Daegu province

3.3. K-Means clustering

K-means clustering is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

Next thing I do was divided those hospitals into 3 clusters. The first cluster is intended for preliminary checking related to COVID-19 and for patient who has mild symptoms. The second cluster is intended for patient who has intermediate symptoms without comorbid disease. The third cluster is intended for patient who has severe symptoms with comorbid

disease and in risk of dying. The clusters were automatically formed using the k-means algorithm. Finally, I plotted the clusters into a map.

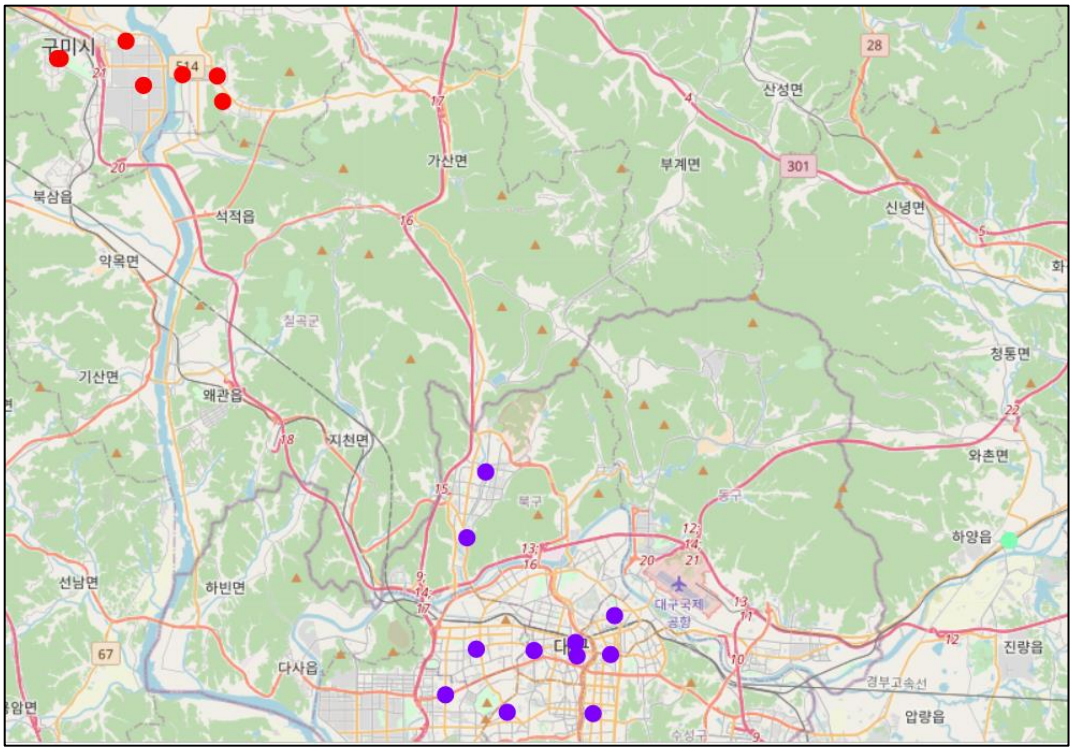

Figure 7. Cluster of hospital in Daegu province

The green point in the eastern part of Daegu is the first cluster for preliminary checking. The red point in the outskirt of Daegu is the second cluster for patient who has intermediate symptoms. The purple point in the downtown of Daegu is the third cluster for patient who has severe symptoms. Notice that there are considerable distance between each cluster. It is because the k-means algorithm is always try to maximize the inter-cluster distances. In real life, the distance could prevent the patient with different symptoms got mixed up and increase the local transmission of COVID-19.

4. Result

From this study, I managed to create 3 clusters of hospital located in the epicenter of COVID-19 outbreak in South Korea, which is Daegu province. Roughly, **there is 1 hospital for every 262 people** who got infected by COVID-19. The clusters are as follow:

First cluster (preliminary checking and mild symptoms)

	Hospital	Cluster Label	Hospital Latitude	Hospital Longitude
0	하양삼성병원 / Hayang Samsung Hospital	2	35.914852	128.825968

Second cluster (intermediate symptoms)

	Hospital	Cluster Label	Hospital Latitude	Hospital Longitude
0	Gumi Gangdong Hospital (구미강동병원)	0	36.096262	128.423090
1	Kabul Hospital (갑을구미병원)	0	36.107246	128.402776
2	WELLKIDS Chilndren's Hospital (웰키즈아동병원)	0	36.106975	128.420729
3	Soon Chun Hyang University Hospital Gumi (순천향대...	0	36.102840	128.382613
4	Gumi Cha Hospital (구미차병원)	0	36.114023	128.340177
5	Hyeonggok Children's Hospital (형곡아동의원)	0	36.113870	128.338730
6	SCH Pediatric Hospital (순천향소아과의원)	0	36.121227	128.374188

Third cluster (severe symptoms)

	Hospital	Cluster Label	Hospital Latitude	Hospital Longitude
0	대구가톨릭대학교병원 (Catholic University of Daegu Hospi...	1	35.843541	128.568839
1	Hyosung Hospital (효성병원)	1	35.842730	128.612706
2	Sodaegu Hospital	1	35.869783	128.553078
3	Future Women Hospital	1	35.850686	128.537456
4	Kyungpook National University Hospital (경북대학교병원)	1	35.866562	128.604793
5	Goodyear Hospital	1	35.867560	128.621608
6	Keimyung University Dongsan Medical Center (계명...	1	35.868964	128.582900
7	Daegu Fatima Hospital (대구 파티마 병원)	1	35.883748	128.623681
8	Union Hospital (유니온병원)	1	35.872249	128.603750
9	Moon Medical Care Hospital (문요양병원)	1	35.915609	128.548146
10	Rosemary Women & Children's Hospital (로즈마리병원)	1	35.943261	128.557775

5. Discussion

This study is quite simple. I only compared the availability of the hospitals with the number of COVID-19 confirmed cases and generated some clusters. In order to build a better model and better recommendation, I recommend to add some more data to the algorithm, such as the number of health worker, the number of medical equipment, demography of the patient, detail of the symptoms, the health status of each patient, and so on. In the meantime, I am quite satisfied with the outcome considering the lack of data that can be accessed.

6. Conclusions

In this study, I managed to develop a k-means algorithm to clusters hospital in Daegu province, South Korea. The tools that I used were completely free and open source. This model can be very useful in helping the health worker when transporting COVID-19 patients to the hospital that match with their symptoms severity.