Clustering The COVID-19 Spread Groups based on Hospital Availability in The Nearby Area 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 project aims to cluster the spread groups of COVID-19 in South Korea and compare it with the number of hospital availability in the nearby area. This ratio could be a basic formula for other countries in order to win the fight against COVID-19.

1.3. Interest

Obviously, the government of other countries which affected with this outbreak would be interested to know what is the ideal ratio of the number of people infected and the number of hospital available 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. 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, cluster them, and find nearby hospitals.

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 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 that it can be plotted on the map.