Assessing corona virus risks in Hong Kong constituencies

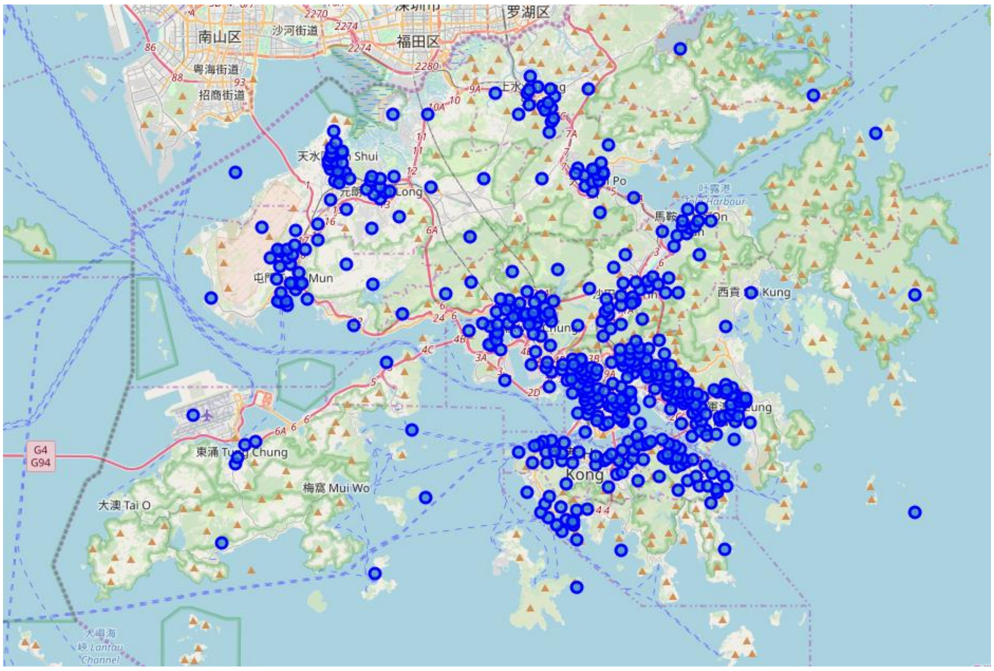
1 Introduction/Business problem

Due to the recent corona-virus outbreak in China, local government takes unprecedented steps to curb the ailment possible spread. It is clear that the virus creeps beyond mainland China. Thus, nearby areas, such as Hong Kong, Taiwan, South Korea, and Japan. In my study I want to focus on Hong Kong to arrange all constituency areas into clusters. Local government can develop different measures to be taken in constituency clusters to prevent epidemic. I expect that I can find a distinct groups of constituency areas based on their demographics. Several tiers of measures would help to save local authorities some funds and improve measure efficiencies.

2 Data

We are using open census data sets from Hong Kong government, available here. This dataset includes information about constituency areas population: total number, age and gender distribution.

Another data set that we would use is the information about the geography of constituency areas: location and area. The set is available on github here (GeoJSON).



3 Methodology

Let’s assume that main factors that correlate with higher mortality rate from the virus are:

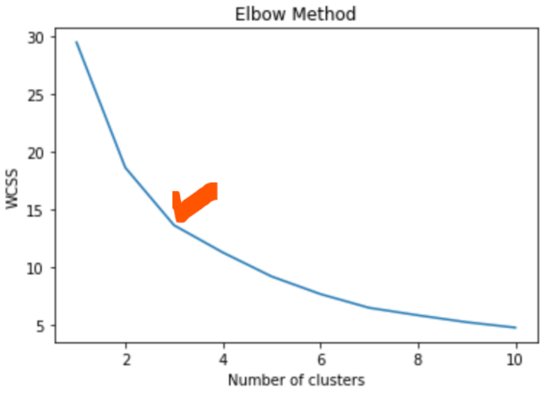
- Age (it was repeatedly reported that older people are more likely to develop severe conditions)
- Gender (also, we know that males are more likely die from the disease)
- Density (naturally, the higher density, the higher the speed of virus propagation)

From the first dataset, I will calculate the share of ‘old people’ (older than 50 years old) and the share of male population in the area. Also, by merging the first and the second dataset, we would be able to calculate the density of the area.

I also do some data cleansing (there are some NaN values in the merged dataset).

After calculating the required parameters (share of old people, share of males, density) I perform K-means clustering algorithm to form clusters. I choose this algorithm as it is a form of unsupervised learning that just fits my scenario.

1) I use elbow method to identify the best K. From the analysis, it looks like 3 is the option to choose.

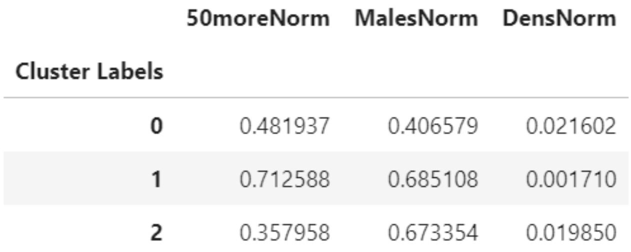


We can see that the optimal k parameter value is 3

2) I use 3 as a parameter to form three clusters.

4 Results

Let’s first analyze what clusters are formed using correlation parameters.



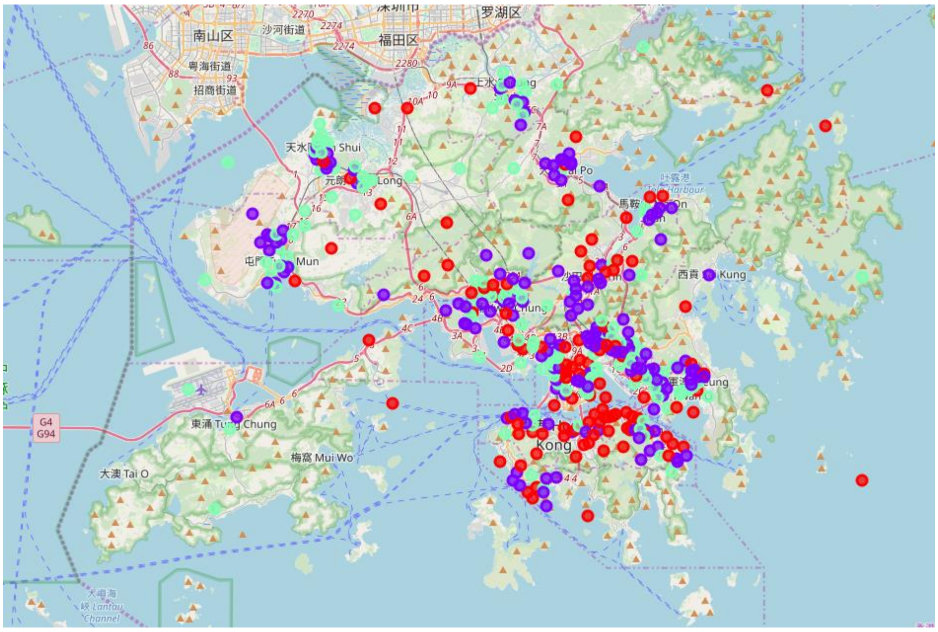
Cluster Labels	50moreNorm	MalesNorm	DensNorm
0	0.481937	0.406579	0.021602
1	0.712588	0.685108	0.001710
2	0.357958	0.673354	0.019850

From these clusters one can see that:

0 - Average age/prevalent women's population/average density - low risk

1 - Older age/prevalent males' population/low density - high risk, as people with the symptoms of flu have less access to doctor. That means even if the density is not high, potentially the mortality rate can be higher.

2 - Younger age/prevalent males' population/average density - low risk



On the map: red label - cluster 0, purple - cluster 1, green - cluster 2

5 Discussion

From the analysis we can conclude that some constituencies in Hong Kong do have higher risk of virus spread. Even though, I supposed that higher density of population is a risk factor, one can think that extremely low density is a risk as well. People who are living in remote areas are more likely to see a doctor later and more rarely compared to their city center peers. Simply, people in such areas have less access to hospitals. We also know that the later medical measures are taken, the less is the risk of death.

Based on this information, local authorities in such areas (cluster 1) can implement specific steps:

- Deploy medical patrols in remote locations
- Proactively visit such constituencies to check the situation
- Use social media and operator messaging to establish two way communication for the dweller of such areas

There are many limitations of such analysis. The problem would need more details investigation to define most impactful factors (distance from hospitals, distance from places where the contagious decease has been detected, the number of public places around) and more specific recommendations on the matter.

6 Conclusion

In this report, I did an attempt to evaluate the risk of virus spread in Hong Kong constituencies based on demographics factors. I used clustering to form three tiers of areas and developed a set of recommendations on one specific cluster of constituencies – scarcely populated zones with primarily older population of males.