MGMT4000E Project Report
Project title: Epidemiological analysis of COVID-19 in HK

Group 3 - Stefani, Cynthia 20373350 | Sham, Cheuk Kiu 20354641 | Lau, Martin 20367870 | Pang, Ho NI Coco 20354043 | Rohatgi, Tanay 20380937
Advisor - Prof Yonghoon Lee

I. Introduction

Background - The COVID-19 pandemic was first confirmed to have spread to Hong Kong on 23 January 2020. Hong Kong was relatively unscathed by the first wave of the COVID-19 outbreak and had a flatter epidemic curve than most other places while being able to control the spread of the virus in the second wave. Compared to many other cities and countries, the control work of the HK government against COVID-19 was considered to be remarkable given its status as an international transport hub, proximity to China, and millions of mainland visitors annually.

Aim of the project - Our aim is to *investigate the spread, distribution, and possible clustering of COVID-19 in Hong Kong*. We conducted an in-depth analysis of cases to find patterns and to create a model that indicates a person's risk level. It is essential to get a better understanding of how the virus spread in Hong Kong as it is a serious health issue globally and the findings will pave the way to mitigate possible repercussions and maintain public health in the future.

II. Project Details

Steps of Research - We investigated the phenomenon and drew insights through analyzing the data of COVID-19 patients. Following is a breakdown of the approach we used:

1. Scraped and collated data about the first 937 cases from various sources such as HK govt dashboard, local websites and newspaper articles *(appendix)*
2. Parsed and plotted geographical locations of the patients on a map to figure out clusters, time-scale distribution and relational interdependence of cases
3. Conducted a thorough investigation on occupational data of the patients to figure out distribution and linkage of industries
4. Analyzed travel links between Hong Kong and other countries to understand relationship between countries and its effect on spread
5. Created a risk factor model, by using weighted scoring on the, to classify and figure out the risk confidence interval for a new individual

Hypothesis - There are three hypotheses for our research project:

1. **Geographical Density** - We hypothesize that the geographical density of the region is directly proportional to the spread of the virus. Geographical density supports the spread of the virus especially once a cluster is infected in or near a densely populated area. Our hypothesis aims to show that the infection rate for people living near the infected cluster will be higher as these people share common activities in the same area.
2. **Occupation** - We hypothesize that people working in certain industries are more prone to the virus. Singularly, individuals who work in the service industry are in frequent contact with and required to interact with more people on average. Our hypothesis aims to connect

certain occupations with a higher probability of meeting an infected cluster and being prone to infections.

3. **Travel History** - We hypothesize that the number of imported cases is positively correlated to countries/regions with most people travelling to and from HK. Our aim is to show the relationship between travel history and spread of the virus. Our findings will show that most imported cases were from countries that have a spatially or relationally "closer" relationship with HK.

III. Findings and Case Studies

Findings on Geographical Density - We observed that the second wave of the virus started in late February and reached its peak by late March (*fig 1*). The timing of the second wave corresponded with the return of Hong Kong students from abroad. To prevent rapid proliferation of the virus, the Hong Kong government began enacting strict control measures at HKIA such as prohibition of international transits, entry ban for all non-residents and 14-day quarantine order for returning residents. The effort shows success as the graph began to flatten by April.

As the cases spread throughout HK, our team has identified 4 main clusters/ places with most recorded cases through K-Means mapping (*fig 2)*. These clusters are centered at TST, HK Island, Tuen Mun, and Sha Tin (*fig 3)*. A possible explanation for this phenomenon is due to the densely populated areas becoming a catalyst of the spread of the virus. Most returning students come from families who can afford sending their kids abroad. More than 45% of the returning students reside on areas located on HK Island making it a central location with most discovered cases. It is essential to note that although most of the students were quarantined, the family members of the infected patients could be a leading cause of the spread due to these areas being densely populated. As a result, we believe that the virus infected significantly more people living in these 4 areas than anywhere else in Hong Kong.

Findings of Occupational Distribution - Out of 936 cases that we were analyzing, we could collect data regarding the occupation of 274 cases. We found information about 87 unique occupations and grouped these occupations by 12 different sectors. With regards to our second hypothesis about people working in certain industries being prone to the virus, following is deeper insight into the key takeaways from our findings.

The occupational distribution indicates an overwhelming 70.3% of cases can be classified as returning students followed by a mixture of occupations that all fall under 5% each (*fig 5*), which aligns with the second surge of COVID-19 that can be tied largely to imported cases. Removing students from the picture, the occupational distribution of COVID-19 cases is largely composed of occupations such as restaurant workers, taxi drivers, aviation crew, domestic helpers, and government sectors which amounts to a total of 53.1% (*fig 7)*. Given that these *service industries* are largely client facing, it supports our original hypothesis that those who frequently meet and interact with people would have a higher probability of being exposed to COVID-19.

Findings of Travel History - According to our findings, over 67.25% of the cases in Hong Kong were either imported or through close contact with imported cases (*fig 8*). This led us to further believe that Hong Kong was doing well in-terms of maintaining social distancing measures & reducing human activities locally. Upon further analyzing the travel information (*fig 9*) we had

about the infected cases, we found that over 48.4% cases were returning from the UK. This could be explained by a large number of HK residents having British citizenship and the exponential increase of cases in the region. We also found that an estimated 70% of returning students were from the UK. The US came in second, with over 10.2% cases being imported from the country. Our proposed explanation is the lack of early controls in the region and the strong trade relationship between US and HK. Overall, European countries showed a dominant amount of cases followed by neighboring Asian countries which have lower spatial distance with the region. The centrality of HK as a key market-player for global finance and also as a tourist destination could explain the reason for having a large amount of imported cases. This relates with the case for Singapore, a sister city which has found itself in a similar position.

Case Study: Fook Wai Ching She - FWCS is a Buddhist preaching hall/temple located at North Point. It is led by one person who has 3 different roles: monk, director, and head of operations. FWCS had believers/guests coming to the temple daily to carry out different activities ranging from volunteering, cleaning the temple, hearing Buddhist sermons and providing donations to the temple. These activities continued to be resumed normally even after the pandemic began. As a result, 19 out of 60 cases found in North Point can be traced back to the temple. After the Department of Health inspected the place in late February, the monk tested positive for the virus leading to a temporary closure of the temple to prevent further spread of the virus.

Fig 10 explains how occupation and level of centrality affects the spread of the virus relating to the temple. Patient 102 has the highest level of centrality at the temple, and this fact can be attributed to the multiplexity of his roles. Due to these roles, he interacted with different people daily ranging from followers, potential donors or potential believers introduced by current followers. His status as a super spreader can be attributed to his high level of centrality within his network group and the nature of his job which requires him to meet people directly (face-to-face).

The case also shows the association that homophily has to the spread of the virus. If a member of a homophilous network is infected, members of that network group will be exposed to higher risk of being infected. Patient 92 and 98 are not connected by any family ties but they knew each other and became friends through their shared activities of attending FWCS services frequently. Both patients tested positive for the virus the same day they visited the temple, meaning that either one of them could have infected the other during their visit to the temple or that both of them could have been infected by the super-spreader patient 102.

A significant number of infected returning students live in North Point along with the clusters from patient 92 and 98 living in the area. This makes the people living in North Point have even higher chances of coming into contact with the infected patients as North Point is one of the most densely populated areas in Hong Kong. As a result, North Point became one of the virus hotspots located on Hong Kong Island. The example of Fook Wai Ching She explains how occupation, level of centrality, homophily, and geographical density affects the spread of the virus.

IV. Risk Factor Model

Using a decision tree-based approach, we implemented a simple risk factor model for predicting the likelihood of an individual getting infected by the virus. The model works by calculating a

score based on geographical, occupational, travel and symptomatic circumstances. It takes into account factors such as the heuristic distance from the geographical cluster centers, the closeness of the occupational industry to people in their daily interaction, the increased exposure related to travel or being in close contact with someone who recently travelled and finally the symptomatic nature of the person.

Based on the factors above, we do a simple computation that predicts a risk score for the individual. In our implementation, we calibrated a score range of 0-30% is low risk, 30-60% is medium risk and 60% + is high risk based on the factors that we discussed. It is important to clarify that this score is not a measure of if a person is infected or not but rather a risk-based classification based on exposure. It is entirely possible and plausible to have a person with 90% risk but no infection and 10% risk but positive infection. The reason we went with this approach is because the dataset is too small to create a machine learning based classifier due to risk of overfitting and the changing infection trends in Hong Kong. We hope that our model gives people a correct estimate of their risk levels and allows them to take proper precautions.

V. Limitations and Reflections

One of our main limitations was the availability of data due to the confidentiality associated with patient data. The information that was officially available from the government website was limited in scope and could not be used for occupational or travel analysis. Further, due to the time-constraint on our project, we could only factor in the first 936 cases out of the total 1000+ cases that have happened in Hong Kong. Our approach was to collect information from various tertiary sources such as local news sites and other portals. This does add issues about the purity of data, but it was a risk we were willing to take to reap the benefits of the epidemiological analysis.

Another concern we had about our project was about connecting our findings with the class concepts. As our initial focus was on finding relationships about the various factors which were not directly related to the class concepts, we had to take a holistic approach in finding out aspects that matched with concepts of centrality, homophily and multiplexity. We provided a case study about the FWCS temple to illustrate these concepts as we think that this case is best suited to demonstrate the aforementioned theories. However, the case study of FWCS alone does not fully reflect the current situation of the spread in Hong Kong since other external factors such as hygiene and social distancing practice played a part in influencing the virus outbreak.

After finishing our presentation, we received some valuable feedback from our fellow classmates and Prof which we have added here as reflections. Many classmates showed interest in our approach for the risk factor model and we realized through interactions that the model is biased if it assumes equal importance for all different aspects. Our findings did not take into account the relationships between different factors and the overwhelming importance of one over the other. For example, a person’s occupation might have a much higher weightage for risk as compared to his geographical location or the two might be positively correlated to increase his risk levels. Given more time, this interdependence of and the unequal weightage of factors would be something that we will be willing to explore more.

Another important point raised by Prof. Lee was about the relationship between occupation and certain geographical locations in Hong Kong. For example, do cases with occupations in the finance industry have stronger geographical ties to locations on the island side such as Central. We reflected that this would be an interesting aspect to explore in terms of social networks, but our implementation remained constrained by the limited information about occupational data that we could find and almost zero datasets existing publicly that could help connect these two factors. On a micro-scale, we did see a relationship pointing to more finance industry cases being on the South Island side but with the handful of cases, we couldn't generalize these patterns.

VI. Conclusion

Our team is proud to acknowledge that we have successfully accomplished the aims that we set out to work on. Our epidemiological analysis of COVID-19 establishes some key findings related to geography, occupation and travel that could be insightful for health officials to strongly factor in future outbreaks or next waves. We provide quantitative proof to these discussions about the influence of centrality, homophily and multiplexity to the spread of the disease. Not only will it be essential for policy makers to set the correct precedent, it would also assist individuals to increase awareness about the spread and reduce exposure to the virus. In a post-COVID world, these preventive measures are likely to stay in place for a long time and our project is a small step in that direction to understand what the social network data of COVID-19 patients tell us about the virus and it's spread.

Appendix I: List of resources used for data collection and collation

1. Hong Kong Government's official website for information about COVID-19 - https://www.coronavirus.gov.hk/eng/index.html#Resource_Centre
2. Hong Kong Government's Statistics department dataset for COVID-19 cases - https://data.gov.hk/en-data/dataset/hk-dh-chpsebcddr-novel-infectious-agent
3. Local Newspaper with detailed information about individual COVID-19 cases - https://wars.vote4.hk/en/cases

Appendix II: Codebase

GitHub Link - https://github.com/trohatgi12/Epidemiological_Analysis_COVID-19

API Key for Google Geocode[1] - AIzaSyAhhWzdBeS-eG33_mU_S4qjFzIbVW83Mrc

Reference:

1. Python documentation for Google Geocoder https://python-googlegeocoder.readthedocs.io/en/latest/
2. Towards Data Science article on scrapping geographical coordinates from address data https://towardsdatascience.com/easy-steps-to-plot-geographic-data-on-a-map-python-11217859a2db

[1] This is required to run the Risk Factor Model, please do not share or upload to any open-source platform

Appendix III: Charts, Tables and Figures

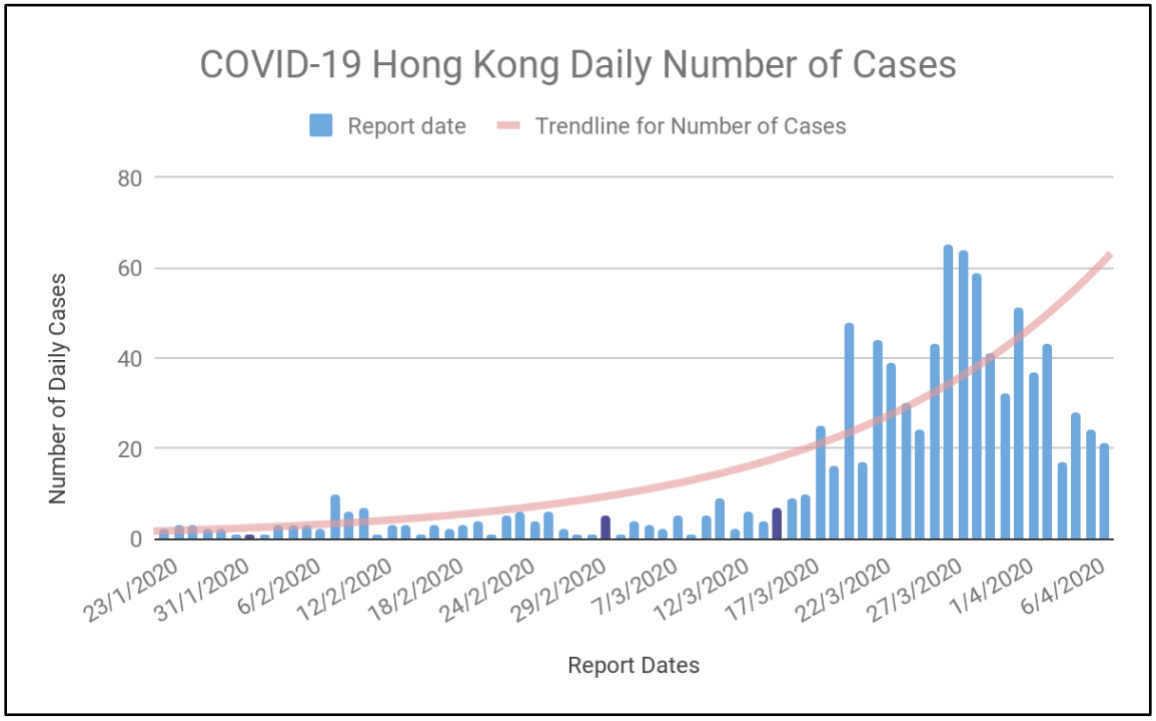


fig 1: Number of COVID-19 cases in Hong Kong on a timescale

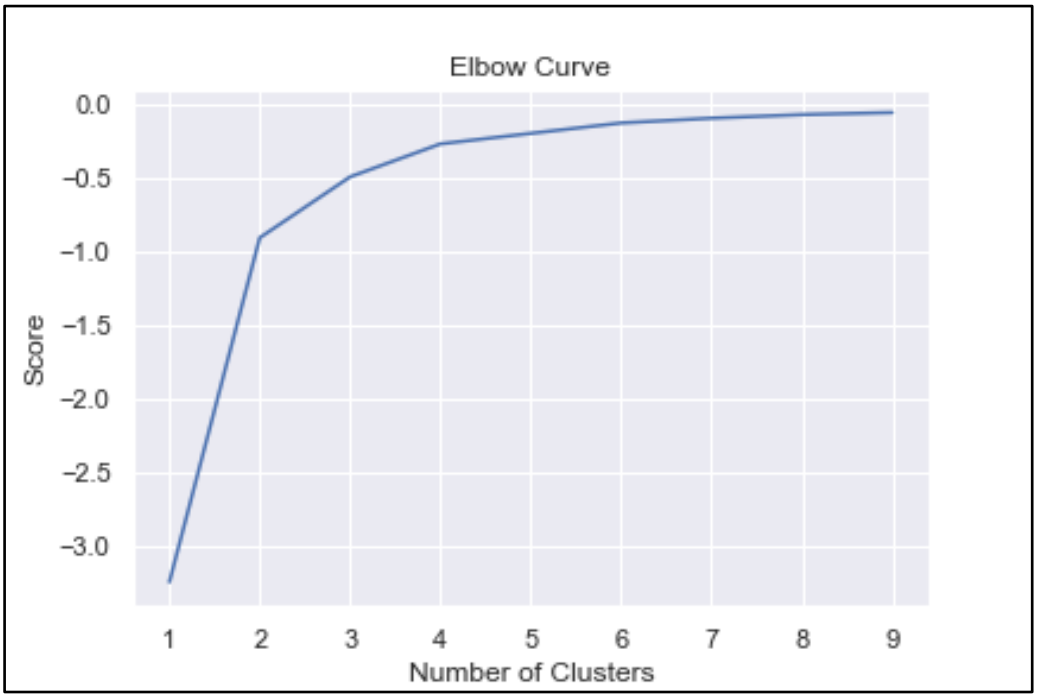


fig 2: Mapping the number of clusters based on K-Means

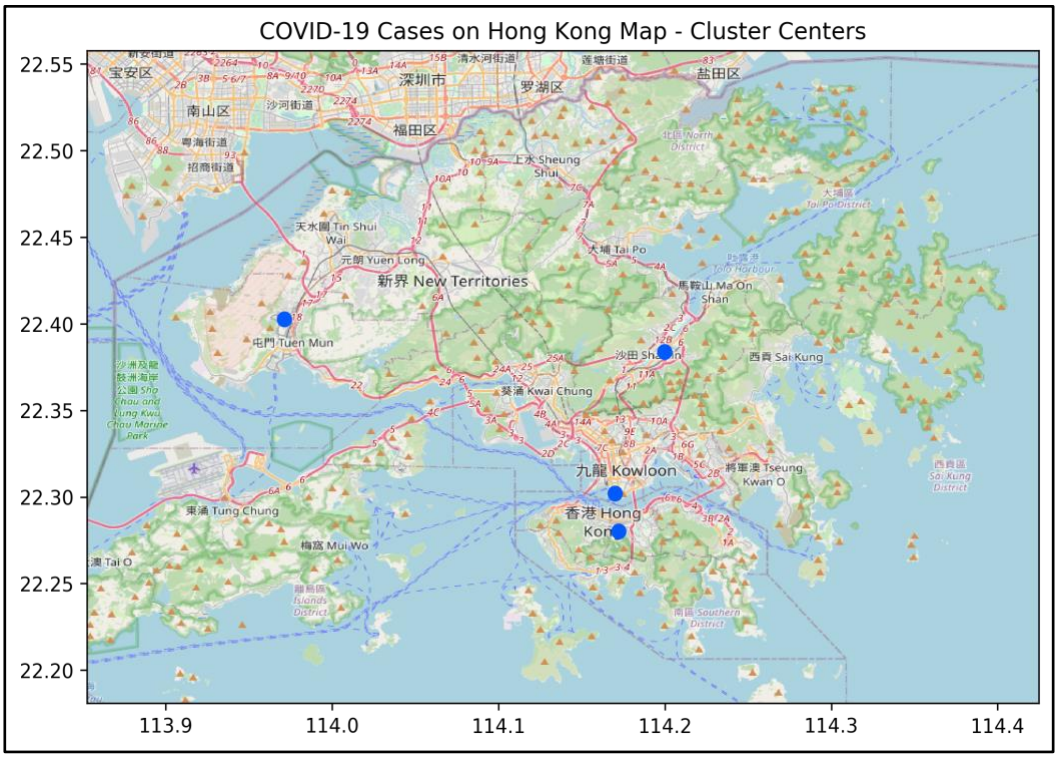


fig 3: Centers of the 4 main COVID-19 clusters in Hong Kong

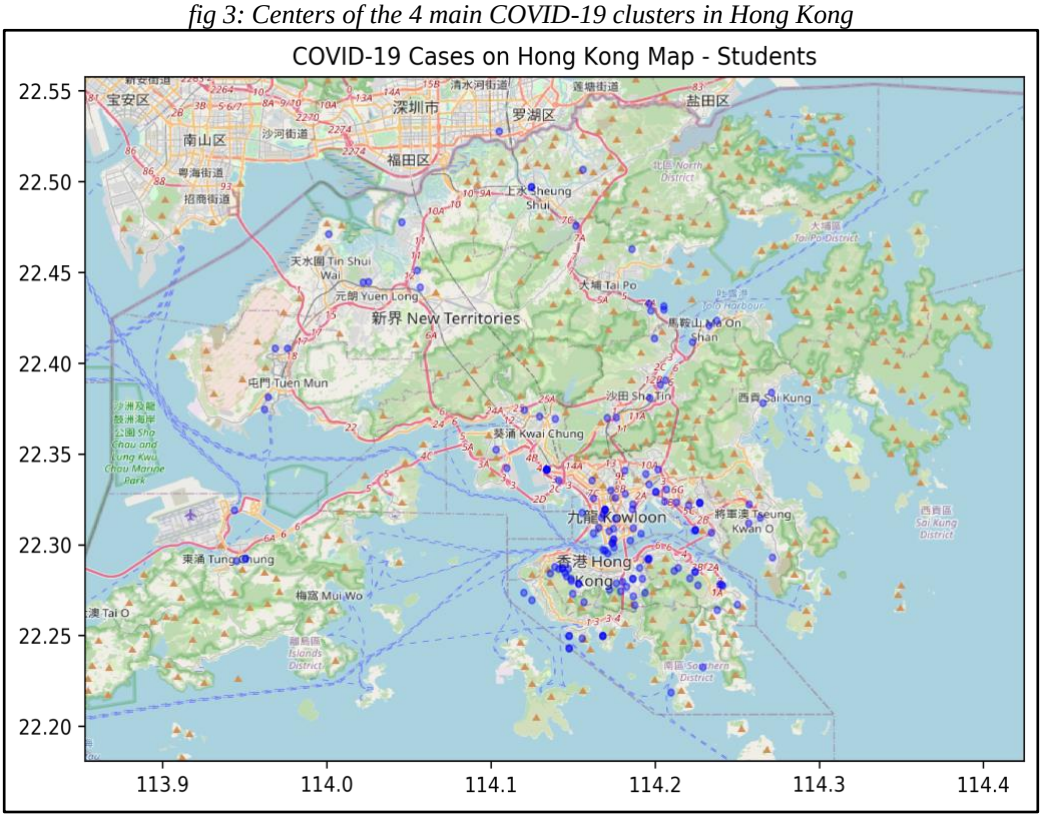


fig 4: COVID-19 cases by returning students in Hong Kong

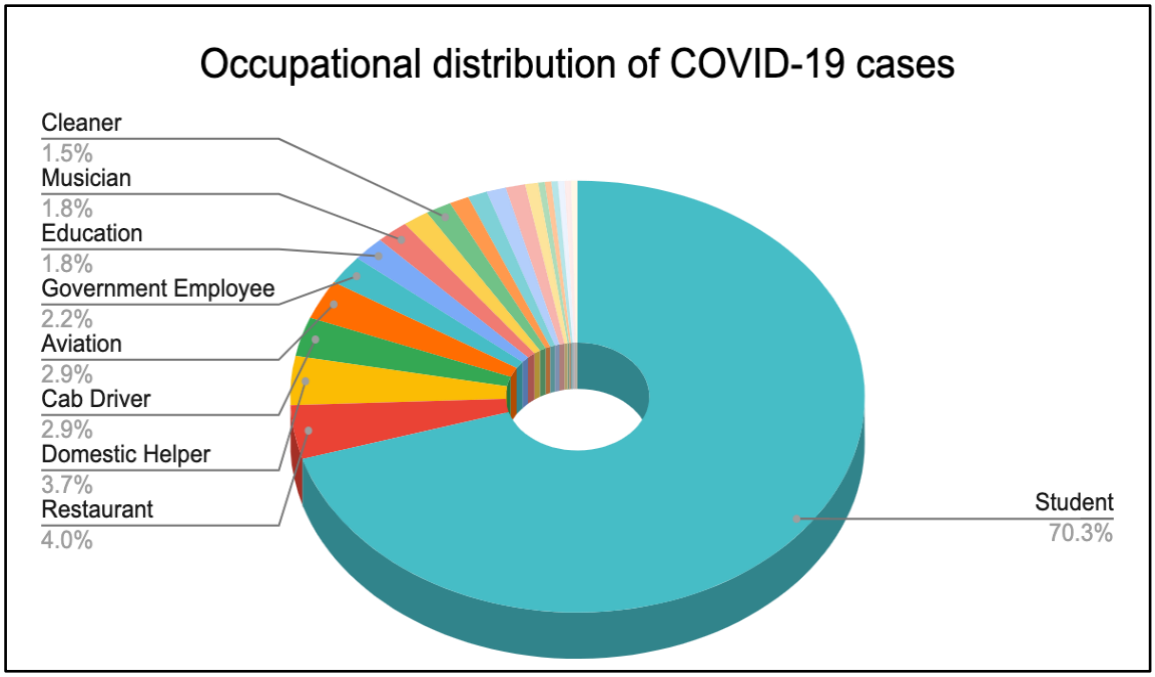


fig 5: Occupational distribution of COVID-19 cases in Hong Kong

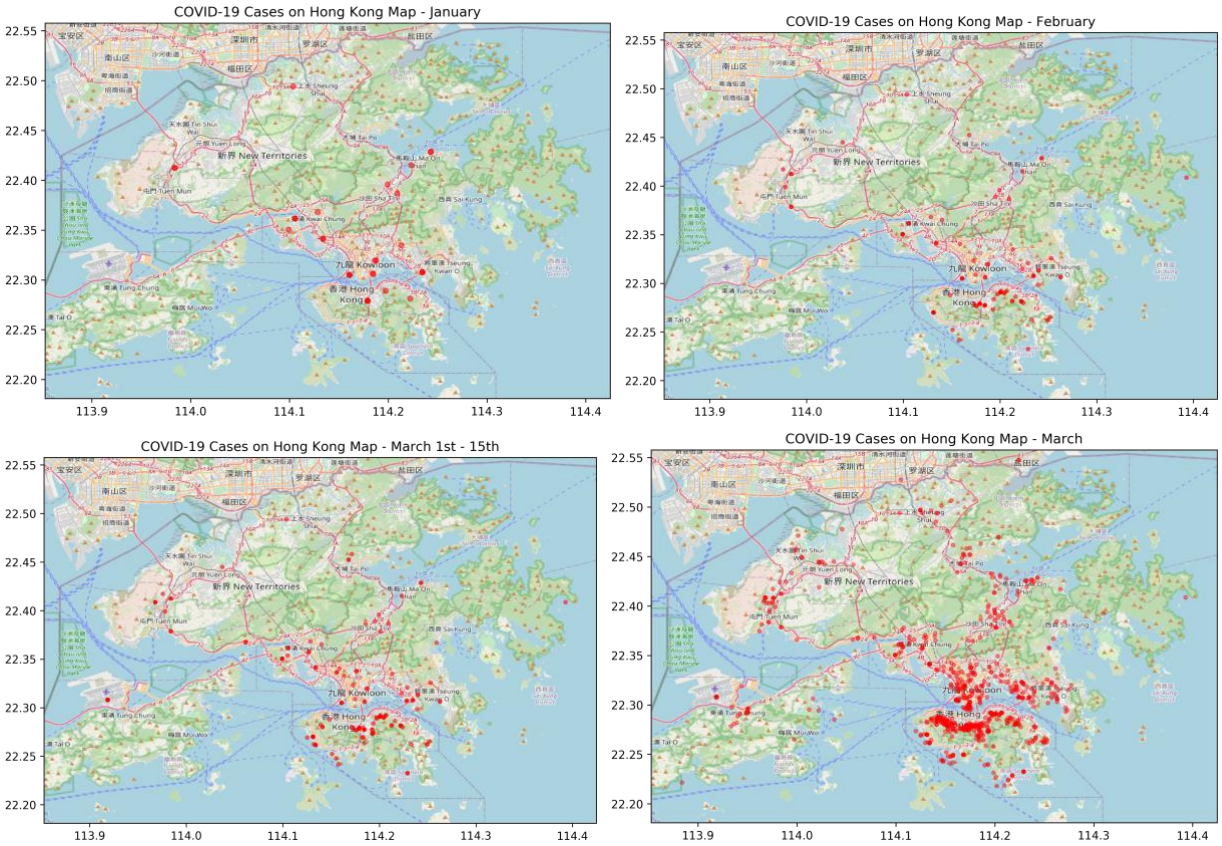


fig 6: Month-by-month spread of COVID-19 cases in Hong Kong

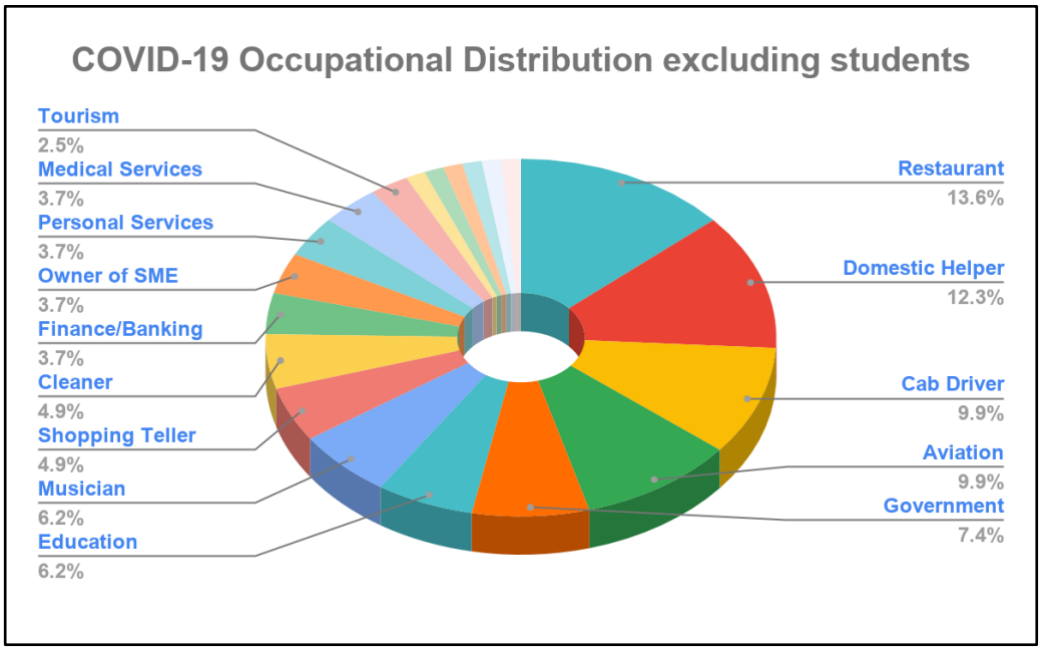


fig 7: Occupation distribution of COVID-19 cases in Hong Kong (excluding students)

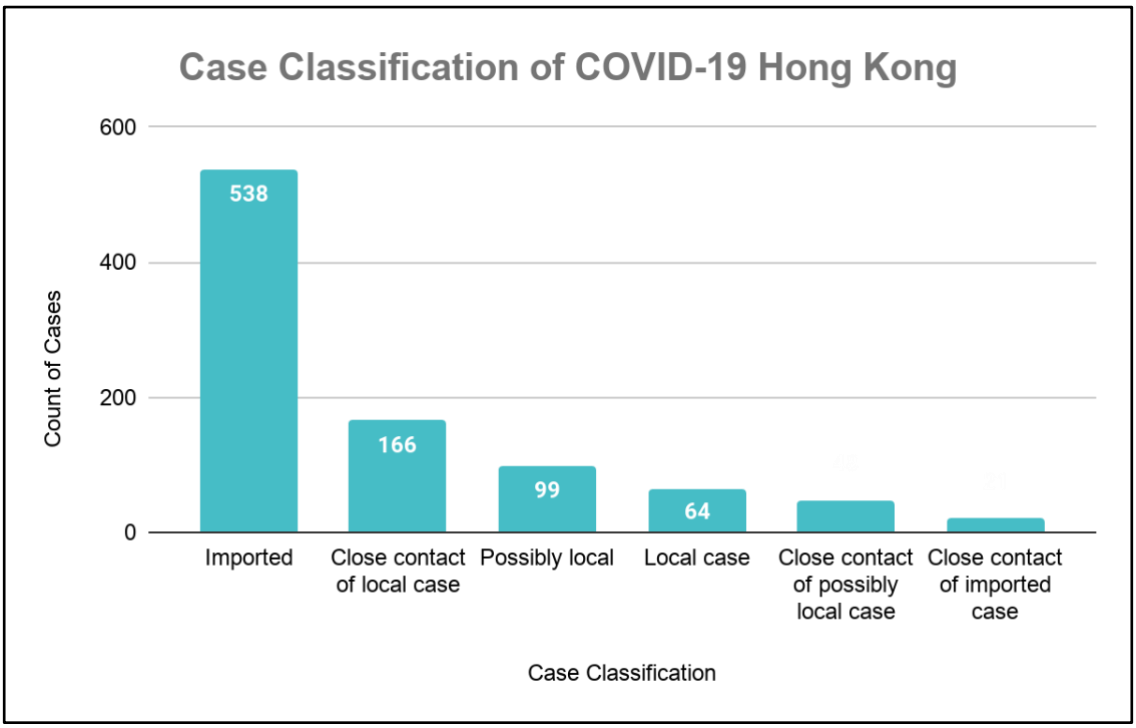


fig 8: Case Classification of COVID-19 cases in Hong Kong

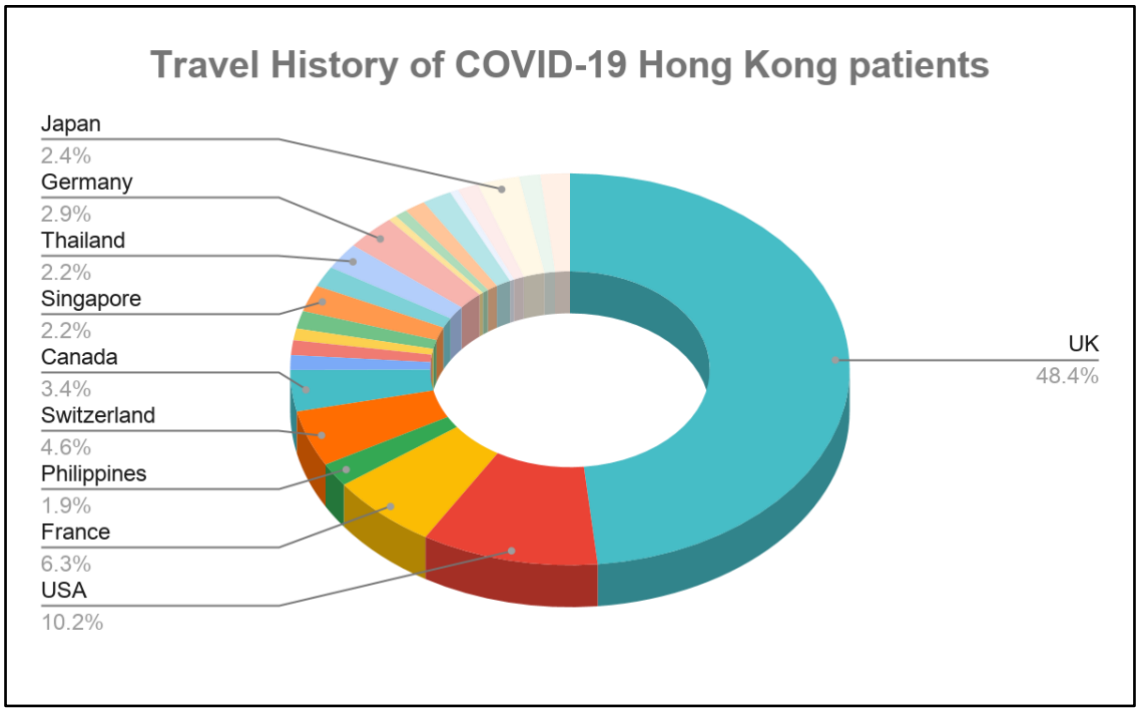
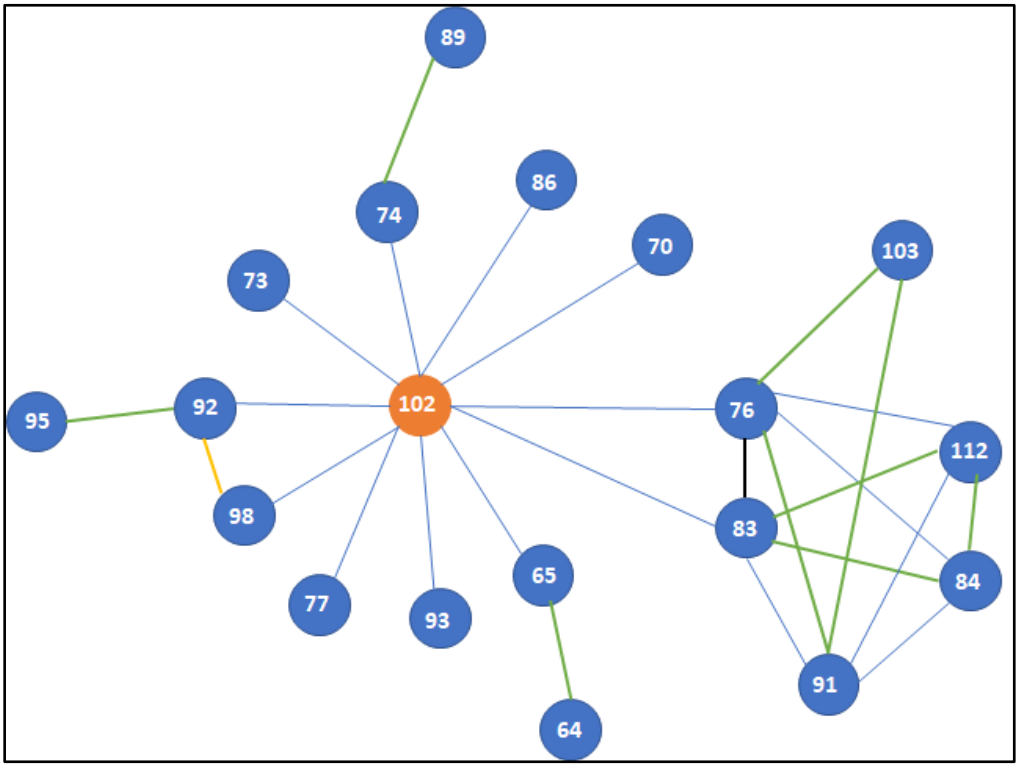


fig 9: Travel History of COVID-19 cases in Hong Kong



Green Line: Live together
Yellow Line: friends sharing the same routine
Black Line: family sharing the same routine

fig 10: Fook Wai Ching She temple case and the related clusters associated

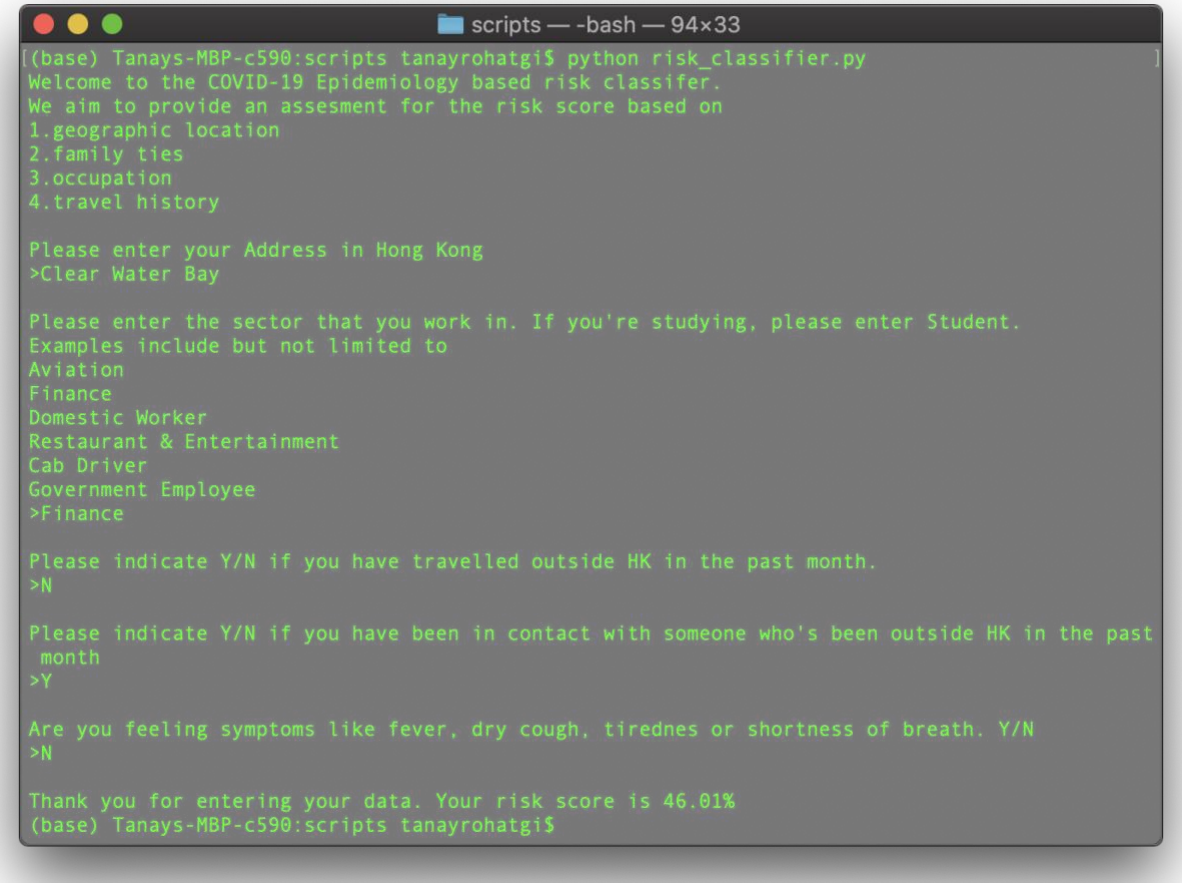


fig 11: Risk factor model sample run with {G: Clear Water Bay, O: Finance, T: N-Y, S: N}