



COMP4801
Final Year Project

Final Report

**A Big-Data-Driven Approach for MTRC and
Coronavirus Analysis**

Supervisor:
Prof. Cheng Reynold

Members:
Ali, Marvin (3035361817)
Effendi, Janice Meita (3035492977) - Author
Jain, Rishabh (3035453608)
Nagra, Harsh (3035437707)
Widjaja, Marco Brian (3035493024)

18 April 2021

Abstract

The COVID-19 pandemic is caused by the spread of the coronavirus, which mainly occurs through physical contact. As the leading public transport in Hong Kong, the Mass Transit Railway (MTR) can be considered to be a hotspot for the coronavirus. With the prevalence of big data, this project aims to evaluate the behaviors of MTR passengers and how it has affected the development of COVID-19 in Hong Kong, as well as how the behaviors have changed in response to the latter. Preliminary data analyses have shown that they are correlated to one another, but more in-depth findings are required to establish what causal relationships are present. Through developing a unified data mining web application for these two data sources, this project will streamline the research procedure for professionals and propose preventative measures against the coronavirus for the population in Hong Kong.

Acknowledgments

First and foremost, I would like to express my gratitude towards Prof. Reynold Cheng, without whom the research involved in this final year project would not have been possible. I would like to thank him not only for his constant support but also for successfully relaying all our questions and concerns to the external parties (i.e. the MTRC) in order to ensure that the workflow of our implementation runs smoothly.

I would also like to acknowledge Shivansh Mittal, who ensures that our quality of work is optimal by providing valuable feedback and suggestions in every aspect of our work.

Furthermore, I have my teammates Marco Brian Widjaja, Marvin Ali, Harsh Nagra and Rishabh Jain to thank for their willingness to cooperate and handle any ad-hoc issues that we encounter.

Table of Contents

Abstract	1
Acknowledgments	2
Table of Contents	3
List of Figures	5
List of Tables	6
Abbreviations	7
1. Introduction	8
1.1 Motivation	8
1.2 Familiar strangers	9
1.3 Big data	10
1.3 Objectives	12
1.4 Project contribution	12
1.5 Outline	13
2. Methodology	15
2.1 Overview	15
2.2 Data pipeline	16
2.3 Database modeling	18
2.3 Mobility trend analysis	23
2.4 Geospatial analysis	24
2.5 Contact and behavior-based research	25
2.5.1 Someone Like You	25
2.5.2 Sensor Individuals	28
2.5 Interactive UI-based platform	30
2.6 Summary	34
3. Results	35
3.1 Overview	35
3.2 First wave of COVID-19 in Hong Kong	35
3.3 Second wave of COVID-19 in Hong Kong	36
3.3 MTR network route changes in COVID-19	40
3.4 UI-based web platform for MTR and COVID-19 data	42
3.4.1 Query	43
3.4.2 Visualization	48
3.4.3 Analysis	54
3.5 Challenges	57

3.5.1 Nature of Dataset	57
3.5.2 Computation Power	58
3.6 Future Improvements	59
3.6.1 Sensor Individuals	59
3.6.2 Real time location-based COVID-19 alerts	61
3.7 Summary	61
4. Conclusion	63
References	64

List of Figures

Figure 2.1: Project workflow	15
Figure 2.2: SSH protocol	17
Figure 2.3: Web Platform Layers of Security	18
Figure 2.4: Octopus card types	20
Figure 2.5: Raw COVID-19 cases data	20
Figure 2.6: Clean COVID-19 cases data	20
Figure 2.7: Database entity-relationship diagram	22
Figure 2.8: ‘Someone Like You’ analysis step-by-step	27
Figure 2.9: Sensor Individuals analysis step-by-step	29
Figure 2.10: Platform architecture	31
Figure 2.11: Technology stack	31
Figure 3.1: COVID-19 cases and MTR passenger count in February 2020 (by age group)	35
Figure 3.2: COVID-19 case distribution during Hong Kong’s second wave	37
Figure 3.3: MTR station density and COVID-19 hotspots in April 2020	38
Figure 3.4: Most popular MTR routes and COVID-19 hotspots in April 2020	39
Figure 3.5: Top 20 busiest MTR routes in January 2020	40
Figure 3.6: Top 20 busiest MTR routes in April 2020	41
Figure 3.7: Login page	42
Figure 3.8: Travel pattern query page	43
Figure 3.9: Passenger mobility query page	45
Figure 3.10: Station density query page	46
Figure 3.11: Raw data query page	47
Figure 3.12: Station density visualization page with station details	49
Figure 3.13: Station density visualization page with input form	50
Figure 3.14: Travel pattern density visualization page	51
Figure 3.15: Travel pattern density visualization page with input form	52
Figure 3.16: Passenger volume visualization page	53
Figure 3.17: Someone Like You analysis page	54
Figure 3.18: Sensor individuals analysis page	56
Figure 3.19: Sensor individuals algorithm proof-of-concept	60

List of Tables

Table 3.1: Travel pattern query sample output	44
Table 3.2: Passenger mobility query sample output	45
Table 3.3: Station density query sample output	46
Table 3.4: Raw data query sample output	48
Table 3.5: Someone Like You sample output	55
Table 3.6: Sensor individuals sample output	56

Abbreviations

MTRC

Mass Transit Railway Corporation

ESRI

Environmental Systems Research Institute

COVID-19

Coronavirus disease 2019

HKCHP

Hong Kong Centre for Health Protection

RDBMS

Relational database management system

SQL

Structured Query Language

SSH

Secure Shell

1. Introduction

1.1 Motivation

The mass transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is known to be the root cause of the COVID-19 pandemic. Though the emergence of the first case of COVID-19 was identified in Wuhan, China on 31 December 2019 (World Health Organization, 2020), Hong Kong confirmed its first identified cases not long after, on 23 January 2020 (Cheung, 2020). With the exponential growth in the number of infections, studying the dynamics of such a contagious disease - namely its possible routes of transmission and its socioeconomic consequences - becomes most essential.

La Rosa et al. (2020) identifies direct contact with respiratory droplets as the primary method of human transmission. SARS-CoV-2 viral particles are released as a by-product of coughing, sneezing, breathing, etc. Thus, in an effort to curb its rapid spread, social distancing measures are implemented, with the recommended distance declared to be 2 meters (6 feet) between each person (European Centre for Disease Prevention and Control, 2020). In addition to this, countries worldwide have adopted the lockdown policy, during which only essential services are allowed to operate, so as to advise local communities to stay home unless absolutely necessary. For instance, on 28 January 2020, employees in Hong Kong started to abide by the work-from-home policy (Cheung et al., 2020). Even with these regulations in place, however, it is almost impossible to abstain from the risk of exposure in commonly shared spaces such as public transportation.

Accounting for 47.4% market share of Hong Kong's public transportation in 2019 (MTR Corporation Limited., 2020), The Mass Transit Railway (MTR) is the leading transport arrangement in Hong Kong. Observing from a more general perspective, one may assume from the aforementioned statistics that the mobility of Hong Kong's population is reflected through MTR passengers' behaviors.

As a crucial mode of transport, with the added concern of the lack of social distancing in such an enclosed space, the MTR may be deemed as a coronavirus hot spot. This project aims to identify the correlation between MTR passengers' travel patterns and the development of the COVID-19 pandemic in Hong Kong, focusing predominantly on how these two factors have impacted one another.

1.2 Familiar strangers

The market of public transport passengers can be segregated into two extremes - those whose travel patterns exhibit high regularity versus those whose patterns are less predictable (Zhang et al., 2016). Given the consistencies in their individual journeys (i.e. MTR routes), the former is more likely to encounter 'familiar strangers' - individuals who have encountered one another multiple times but both parties have had little or no interaction (Zhou et al., 2020). However, a thorough comprehension of the concept of 'familiar strangers' poses various challenges within itself. One, although they have been presumed significant for 4 decades now (Liang, Li & Zhang, 2016), the defining metrics in identifying familiar strangers are far from concrete. Two,

discovering the implications of ‘familiar strangers’ involves looking at it from a multidimensional perspective.

Zhou et al. (2020) believe that, within the public health perspective, ‘familiar strangers’ could affect the transmission of infectious diseases within the general population. More specifically, in this case, MTR passengers who are ‘familiar strangers’ could bear catalytic consequences towards the coronavirus infection rate. Especially when direct physical contact comes into question, these individuals may unknowingly contract the virus through an infected ‘familiar stranger’. An example of this scenario would be, if two individuals were adhering to the same MTR route and consequently boarding the same MTR trains, they could be in close proximity for a considerable amount of time.

Previously, the significance of ‘familiar strangers’ greatly depended on hard evidence such as personal anecdotes or surveys (Zhou et al., 2020). However, with the increased relevancy and availability of big data analysis techniques, we can leverage these to depict a more accurate picture of ‘familiar stranger’ groups within the MTR passenger network.

1.3 Big data

The term big data refers to the usage of complex and large data sets, typically unstructured data, in order to pinpoint significant trends and characteristics that were not discovered previously (Boyd & Crawford, 2011). As compared to traditional methods where data is sampled into smaller volumes, big data involves extracting some value from the dataset, regardless of how

extensive the repository is. Through this, one is able to gain the advantage of obtaining additional insights. Consequently, this can lead to more informed future decisions and predictions.

In collaboration with the Hong Kong Mass Transit Railway Corporation (MTRC), this project intends to utilize MTR passenger traffic data from January 2020 to September 2020 to conduct data analyses and generate appropriate visualizations. The same data fields from the equivalent time period in 2019 may be used as the control set, so as to reflect previous regularities without factoring in the ongoing COVID-19 pandemic.

In a study conducted by Zhou et al. (2020), it was determined that the odds of encountering ‘familiar strangers’ is heavily influenced by the number of Beijing smartcard riders at any certain time frame. With both Beijing and Hong Kong being major metropolises, one can attempt to replicate the findings with the public transport network in Hong Kong by exploiting MTR check-in and check-out ticket transaction data.

Moreover, the COVID-19 data published by the Hong Kong Centre for Health Protection (HKCHP) will be used as an additional data source so as to better comprehend the societal implications. Through incorporating the 2 datasets, meaningful insights can be derived, especially from geospatial and temporal points of view.

1.3 Objectives

This project aims to implement a single unified platform that allows for efficient behavior analyses of MTR passengers in correlation with the pre-existing COVID-19 data. The scope of this project can be streamlined into several technical requirements.

Firstly, a centralized database will be created to store the MTR passenger data and COVID-19 data. Once this data repository is properly established, the next step involves uncovering the relationship between the two datasets and what exactly the correlation entails - for instance, quantifying the extent of a causal relationship, if it exists. At this stage, we hope to understand the impact of MTR passenger behavior on the development of the COVID-19 in Hong Kong, as well as how these behaviors have adapted in response to the outbreak.

Last but not least, we aim to develop a scalable system in the form of a web application. The application would serve as a querying and visualization tool, dynamically displaying relevant statistics or graphs in response to the users' data queries. In addition to this, we hope to integrate MTR and COVID-19 risk metrics as well; its details and use cases will be elaborated on in the later sections.

1.4 Project contribution

At present, previous researchers have attempted to use the data provided by the MTRC to generate visualizations but no key conclusions were made. These diagrams were also built on traditional data processing software such as Microsoft Excel, which might not be sustainable in

the long run due to the volume of the data available. This current workflow is also time-consuming and rather inefficient.

Upon successful implementation of the above requirements, the application will be accessible for interested researchers who have been provisioned with the proper access credentials. These research findings could serve as baseline information for professionals in the public health, public transportation, and/or government sectors. With the proper data pipeline, researchers will be able to gather the required findings in a more efficient manner. In addition to this, we plan to use the ArcGIS platform made available to us through our partner, the Environmental Systems Research Institute (ESRI), to project the data into its geospatial dimensions for more substantial analyses.

1.5 Outline

This report first covers the background theories and motivation behind the project. Following this general overview, the key objectives and significance of the project are properly identified. The second chapter explains the methodologies used to implement the system, including but not limited to, the technologies and data analysis methods specific to each technical requirement. Particularly, it begins by detailing the workflow and moves on to describe each phase in the implementation – building the data pipeline, mobility trend analysis, geospatial analysis and the mobile application development phase. The third chapter showcases the progress of the project thus far. Interim results, along with challenges and future steps (i.e. project schedule and

potential ideas) will be discussed. The final chapter concludes the report. It will reiterate the eventual goals and objectives of the project and highlight how we plan to accomplish that through the aforementioned methodologies.

2. Methodology

2.1 Overview

This chapter gives a detailed summary of how this project plans to leverage big data techniques in order to conduct MTR passenger behavior analysis in relation to COVID-19.

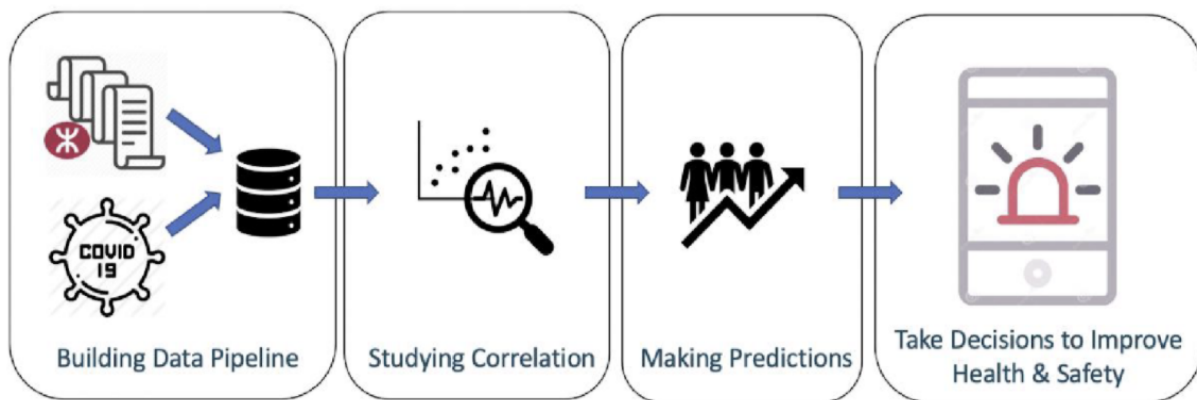


Figure 2.1 Project workflow

The implementation of the project adopts a number of steps in its workflow. Firstly, an appropriate data pipeline will be developed to handle the MTR passenger and COVID-19 data feed. Afterward, in order to study the correlation between the two datasets, the data analysis phase is further subdivided into mobility trend analysis, geospatial analysis and contact and behavior-based analysis. Once this is completed, the final phase involves compiling the findings and data reports from the analytics obtained into a single web application. Each of these phases will be further elaborated on in the coming subsections.

2.2 Data pipeline

In its raw form, the MTR passenger data supplied by the MTRC and the open-source COVID-19 data from HKCHP were both stored in CSV format. The former was entrusted to our team in the form of DVD copies, wherein any interested researcher would then be subjected to the overwhelming manual process of obtaining said DVDs and then copying the data onto their personal computers. It is crucial to note that data processing and retrieval is especially tedious with such high volumes of data. Hence, through migrating the data onto a MySQL database on our designated server, we not only alleviate this issue but also present enhancements to the current workflow.

MySQL is a relational database management system (RDBMS) that operates using Structured Query Language (SQL). SQL is the language used for accessing and manipulating the data. Using SQL, otherwise known as the relational database model, is preferable in this case as opposed to using a NoSQL database. With the latter, the structure of the database (i.e. the database schema) is dynamic and not enforced. As opposed to this, however, an SQL database emphasizes a user-defined set of rules which then regulates the relationships between the different data fields (“What is MySQL?”, 2020). This ensures data consistency such that required fields are accounted for and no duplicates are present within the data. With our end goal of achieving a scalable system, it is imperative that the MTR data and COVID-19 data maintain consistency despite undergoing regular updates. In addition to this, the relational database model

is favored in situations where complex queries and data reports need to be generated (“What is a relational database?”, 2020).

Furthermore, in order to preserve data confidentiality as agreed upon with the MTRC, we made the conscious decision of using a designated private server instead of a public database server. We will employ an SSH-based protocol for security measures as well.

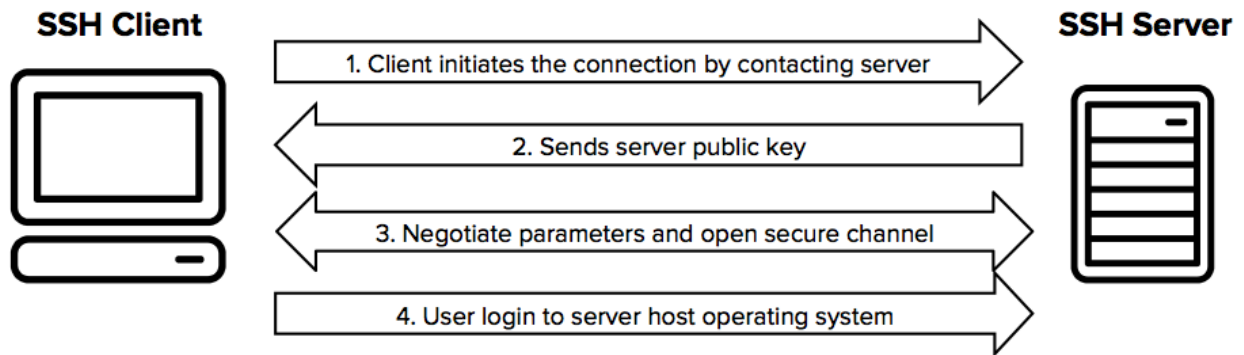


Figure 2.2 SSH protocol. From *SSH (Secure Shell)*, 2020, <https://www.ssh.com/ssh/>.

The client user connects to the database server via SSH, wherein a secure channel is opened in response to the authentication of the user’s credentials (Figure 2.2). For an added layer of security, a multi-hop SSH tunnel is used to access the server containing our database - a user accesses the first server using cs.hku.hk credentials, after which the user uses credentials unique to our team to access the second server where the database resides.

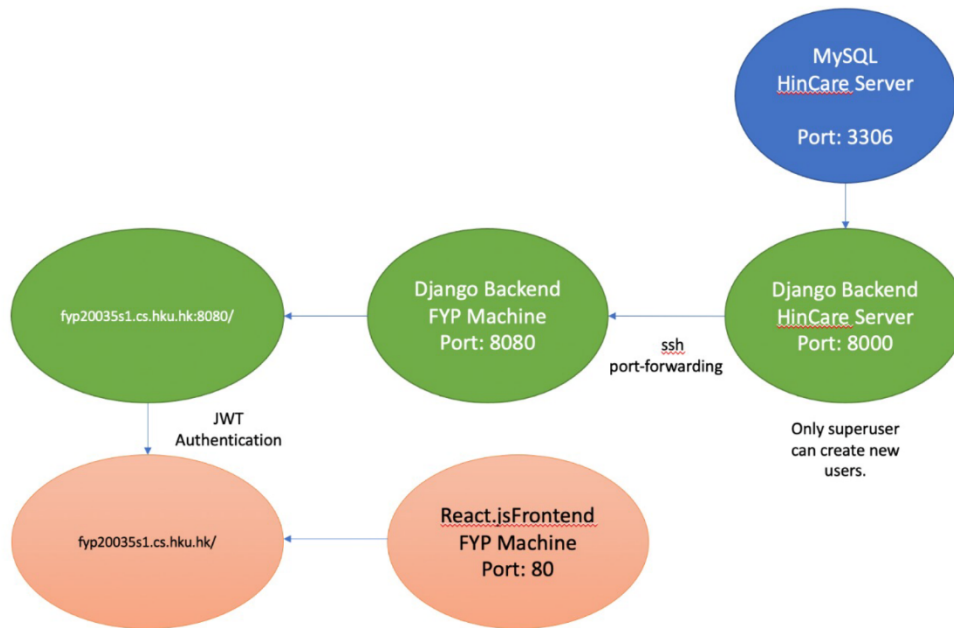


Figure 2.3 Web Platform Layers of Security

2.3 Database modeling

Prior to following through with data migration, however, the data needs to be preprocessed. This problem arises largely due to several data format inconsistencies and missing data fields. For instance, a null exit station field is observed for approximately 1 million MTR passenger data records - these records are deemed not valuable for our research purpose, hence introducing the need to eliminate them. Another prime example is the inconsistent wordings present among the COVID-19 data fields, its effect most prominent in the fields determining case classification throughout the different months.

In the interest of facilitating the course of our research and its further implications, new fields were added onto the COVID-19 cases table - namely the 'Asymptomatic' and 'Age Group' categorizations. 'Asymptomatic' was previously reported under the 'Date of Onset' field, however, we found that it would be more coherent if separated into its own field. Several assumptions were made in accomplishing this:

- a) Given that we are working with historical data dating back to several months ago, a patient previously listed as 'Asymptomatic' could not suddenly develop symptoms; and conversely,
- b) If a patient is listed under 'Unknown' or 'Pending', we determine that they were established to be asymptomatic.

'Age group' was added in accordance with the MTR Octopus smart card age divisions, as dictated in the figure below under the eligibility section (Figure 2.4).




Type	Deposit	Initial stored value*	Handling charge	Fare concession eligibility
Child 	HK\$50	HK\$20	-	Age 3-11
Adult 	HK\$50	HK\$100	-	No concession
Elder 	HK\$50	HK\$20	—	Age 65 or above

Figure 2.4 Octopus card types. From *Octopus*, 2021,

<https://www.octopus.com.hk/en/consumer/octopus-cards/products/on-loan/standard.html>

Figure 2.5 and Figure 2.6 show an excerpt of the clean data in comparison with the raw data obtained from HKCHP’s API.

2558	26/7/2020	Unknown	M		46		No admission	Unknown	Imported case	Confirmed
2559	26/7/2020	22/7/2020	M		90		Discharged	HK Resident	Local case	Confirmed
2560	26/7/2020	23/7/2020	M		41		Discharged	HK Resident	Local case	Confirmed
2561	26/7/2020	24/7/2020	M		19		Discharged	HK Resident	Epidemiologist	Confirmed
2562	26/7/2020	21/7/2020	M		52		Discharged	HK Resident	Epidemiologist	Confirmed
2563	26/7/2020	Asymptomatic	F		43		Discharged	HK Resident	Imported case	Confirmed

Figure 2.5 Raw COVID-19 cases data

2558	26/7/2020		NO	M	46	Adult	No admission	Non-HK resident	Imported case	Confirmed
2559	26/7/2020	22/7/2020	NO	M	90	Elder	Discharged	HK resident	Local case	Confirmed
2560	26/7/2020	23/7/2020	NO	M	41	Adult	Discharged	HK resident	Local case	Confirmed
2561	26/7/2020	24/7/2020	NO	M	19	Adult	Discharged	HK resident	Linked with I	Confirmed
2562	26/7/2020	21/7/2020	NO	M	52	Adult	Discharged	HK resident	Linked with I	Confirmed
2563	26/7/2020		YES	F	43	Adult	Discharged	HK resident	Imported case	Confirmed

Figure 2.6 Clean COVID-19 cases data

Majority of this data cleaning process - for both the MTR data and the COVID-19 data - was performed in R.

The end product of this stage is to have both the MTR database and the COVID-19 database on the server. The former will consist of individual trip transactions, specifically the entry and exit MTR station, the passenger's identification number, and their Octopus card type. The latter will cover the details of confirmed COVID-19 cases from January to September 2020, the daily cumulative number of cases, and the residential address of each confirmed case. The database schema is designed such that minimum latency is preserved; it comprises only relevant information. The entity relationships have been defined as seen in figure 2.7 below.

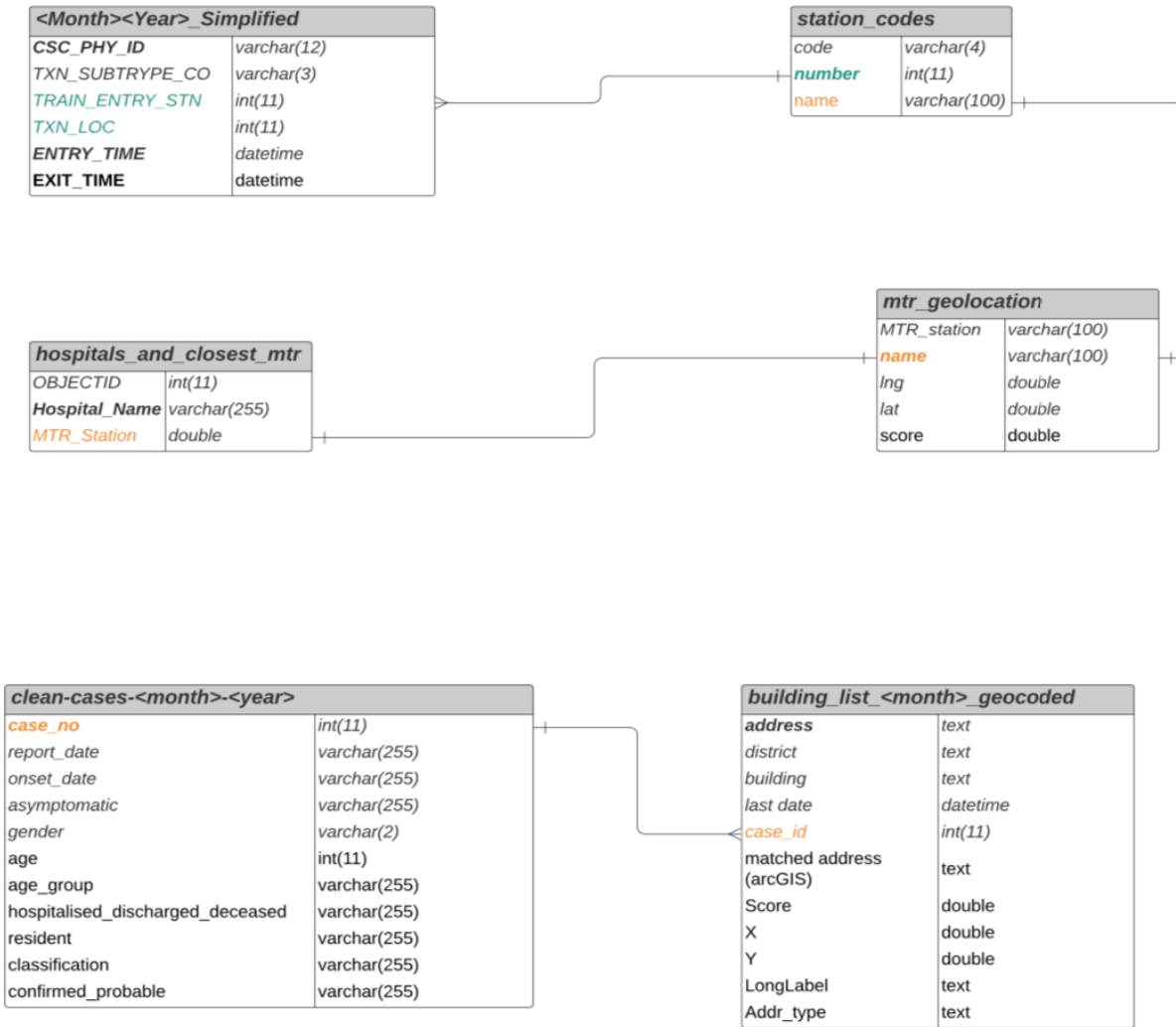


Figure 2.7 Database entity-relationship diagram

As seen in Figure 2.7, <Month><Year>_Simplified holds the MTR passenger transaction data from January - April 2020 and January - April 2019, with 2019 data being retained as a control set if necessary. Station_codes tracks the station code keys and its corresponding names, while the geolocation data, specifying a particular station’s longitude and latitude, is maintained in the mtr_geolocation table. As a supplementary component, we have mapped the COVID-19 hospitals to its closest MTR stations in hospitals_and_closest_mtr. As for the COVID-19 data,

we organize the clean cases by month in *clean-cases-<month>-<year>*.

Building_list_<month>_geocoded contains information on the buildings where these probable and confirmed cases reside.

2.3 Mobility trend analysis

Mobility trend analysis is a necessary preliminary step in our research to determine the correlation between the MTR dataset and the COVID-19 dataset. It will be conducted as an attempt to detect the influence of MTR passenger behavior on the fluctuations in the number of COVID-19 cases, and vice versa. This will be done on Python - using libraries such as Pandas, Matplotlib, Plotly, etc. To capture the general mobility trend, we will look into the variations in MTR ridership within the period of January to September 2020, as well as the same period in 2019 so as to provide us with the baseline statistics before the occurrence of the COVID-19 pandemic. To explore the mobility trend in further detail, sociodemographic factors can be taken into account. For this we can take advantage of the data field specifying Octopus card types - Adult, Elderly, and Child Octopus holders can be classified into their respective age groups.

In conjunction with monthly data reports, we will observe daily travel patterns as well. However, before this can be fulfilled, we will proceed by narrowing down the key dates that we plan to look into in greater detail. These include but are not limited to, Chinese New Year holidays 2020, the enforcing of the work-from-home policy, and the curfew policy.

In order to draw a side-by-side comparison of both datasets, we will also group the COVID-19 cases by age demographic (i.e. Adult, Elderly, and Child), and calculate the cumulative number of cases for the key dates identified. Although these analyses will be beneficial towards our correlation hypothesis, observing the general trend will not contribute much to the predictive analytics aspect of this project. Hence, this prompts us to study the locality of MTR passengers and the confirmed COVID-19 cases.

2.4 Geospatial analysis

Through conducting geospatial analyses, we aim to expedite the process of identifying hotspots for COVID-19 confirmed cases and the corresponding MTR stations. In this phase, we will primarily be utilizing the ArcGIS software provided by ESRI. It is more than capable of processing our highly complex and substantial dataset (approximately 1 billion ticket transactions per month) to generate the geospatial visualizations we require.

We will use the MTR data to look into the density of passengers in each MTR station as well as the travel density between stations in the MTR network map. In order to integrate the COVID-19 confirmed cases into the same spatial dimension, we will first retrieve the geocode of the cases' locations (i.e. their residential building) using the geolocation API available in the ArcGIS software. Then, using the Google Maps API, we attempt to match the building to the closest MTR station. As a final step, we will project the density of confirmed COVID-19 cases in the vicinity of each MTR station onto the same map.

2.5 Contact and behavior-based research

Another category of analysis worth exploring can be identified as contact and behavior-based research. This comes in alignment with our primary goal of adopting the big data analysis approach - particularly, we are attempting to derive several insights from massive volumes of MTR passenger data. We hypothesize that, especially with regards to the development of the COVID-19 pandemic, identifying, followed by interpreting passengers' travel patterns and behavior could lead to relevant observations. We have included two metrics belonging to the contact and behavior-based research category as part of our project scope - namely the identification of 'someone like you' and sensor individuals. These areas of analysis, which principally hedge around the 'familiar stranger' theory previously elaborated on in the introduction, could have had a significant effect on the spread pattern of COVID-19.

2.5.1 Someone Like You

We define *someone like you* as any rider who shares a route or trajectory within the same time frame with another rider. Given the shared location and overlap in time frame, these groups of passengers are likely to have encountered one another. For instance, any two riders with corresponding entry and exit stations can be assumed to have shared a trip if they were travelling simultaneously. A real-life illustration of this would be, those whose workplaces are located close to the same MTR stop, presumably commute to work at around the same time in the mornings. By shifting the perspective to the entirety of the MTR network route, we fulfill the

goal of distinguishing those corridors (i.e., MTR routes) with the largest number of *someone like you*'s.

The data processing involves a number of steps; the pseudocode is detailed below as follows:

1. Create all possible station pairs within the MTR network. All the permutations will be considered - that is, n stations will result in $n*(n-1)$ station pairs.
2. Divide a whole day's period into 48 time frames, each with a half-hour timespan. This parameter could be fine-tuned accordingly. Decreasing it would narrow down groups of passengers, and vice versa.
3. For each day of MTR passenger data, all trips are grouped in accordance with the predefined station pairs in step 1.
4. These trips are further grouped into the time frames in step 2. For this step, we capture the *entry_time* for each trip and associate them with their respective 30-minute time frame clusters.
5. For each of the station pairs we collected, we accumulate the number of trips daily. Specifically, this involves totaling the number of trips in the 48 time frame groupings.
6. Count the total number of *someone like you*'s for each week for each station pair, taking note of the start date.
7. The result is captured as the average amount of *someone like you*'s throughout weekdays and weekends for each station pair. This data is saved in a CSV file containing start station, end station, week start date, amount of *someone like you*'s on weekdays, amount of *someone like you*'s on weekends.

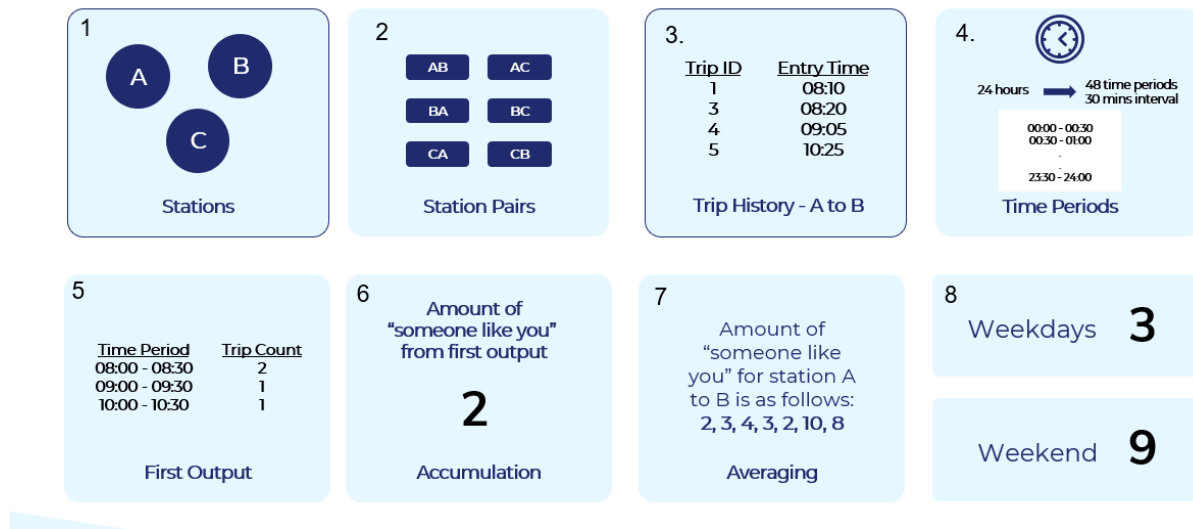


Figure 2.8 ‘Someone Like You’ analysis step-by-step

Figure 2.8 provides an illustration of the procedure, alongside example station codes and pairs. Following through with the above methodology, however, requires some assumptions. Firstly, we conduct the analysis under the assumption that, if an individual is a *someone like you* to another, they do not necessarily have to be sitting in the same MTR carriage. Secondly, given an occurrence wherein the entry station and the exit station are the same between two individuals, it is assumed that their journeys’ routes correspond to one another as well. We note that this might not always be the case: a passenger might interchange lines at a different station but still reach the same destination. However, it is reasonable to conclude that tolerating these outliers and edge cases would not have a considerable effect on our resulting data reports.

It is believed that the *someone like you* study could alleviate the issue of COVID-19 spread. These specific routes have been pinpointed to have the most number of *someone like you*’s.

Although still in its early research stages, a possible future use case could involve using these routes to influence the scheduling of MTR routes, hence serving as an additional preventative measure against the virus.

2.5.2 Sensor Individuals

Sensor individuals can be examined at a station level; they are defined as MTR passengers who bear the highest risk of physical contact with others. As opposed to the *someone like you* analysis wherein the passenger's routes play less of a role in the results (only the starting point and destination is taken into consideration), classifying such sensor individuals prompts us to study the possibility of encountering other individuals from station to station. This alludes to one of the common themes in the 'familiar stranger' theory - the co-presence phenomena, which recognizes an individual's spatio-temporal pattern in their respective journeys, and then proceeds to ask the question of whether or not overlaps exist between multiple patterns. Considering the context of the development of COVID-19, to determine those sensor individuals could allow us to identify specifically which individual is believed to be a super spreader in the public transport system.

As the database currently only holds records of entry and exit stations and their respective timestamps, this data needs to be modified such that we can properly single out each station involved in that particular route. The objective is to transform the origin-destination station pairs into a route detailing the station codes encountered by the rider on that journey and its estimated timestamps, closely following the pseudocode below:

1. A weighted undirected graph is initialized using the *networkx* package in Python.

2. The entirety of the MTR station network is modeled in this graph, with each node representing a single MTR station, and the weight defined to be the travel time between the MTR stations.
3. We take each origin-destination pair from the MTR passenger data provided, and determine the shortest path from *start_station* to *end_station* by implementing a modified version of Dijkstra's algorithm.
4. The timestamp for each layover station is determined by adding the start time to the each weight of the graph (i.e., the travel time between stations). However, in order to compensate for the discrepancies in the calculated time derived from the model and the actual time taken, this difference is divided by the total number of stations involved in the trip, and added back into the time taken to travel between the layover stations.

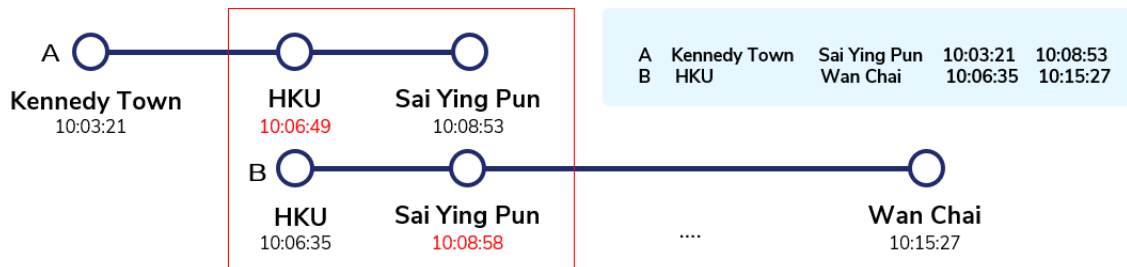


Figure 2.9 Sensor Individuals analysis step-by-step

Figure 2.9 serves as an illustration for the sensor individuals analysis approach. As with the *someone like you* study, we needed to establish several assumptions with the above methodology. Because the graph was modeled in such a way where its judgment is constructed mainly from the travel time between stations, it is impossible to distinguish between the different MTR lines

whilst coming up with the results. Hence, we make the assumption that trips that could involve multiple route possibilities, are considered to have taken the same identical route. Another assumption is that, in the event that a journey involves different MTR lines, the interchange time to switch between these lines is not taken into consideration when determining the individual timestamps.

The sensor individuals study serves as a starting platform for a multitude of additional research topics. A specific use case that concerns COVID-19, for instance, is the ability to identify a possible super spreader through locating the individual possessing the highest co-presence with regards to the other passengers in the public transport system in Hong Kong.

2.5 Interactive UI-based platform

The web-based UI platform aims to fulfill several technical requirements by supporting queries, visualizations and advanced analysis options. Prior to its development stages, constructing a robust system architecture is of utmost importance when it comes to big data solutions; data security, appropriate schemas and a suitable but scalable technology stack are incorporated within the design decisions.

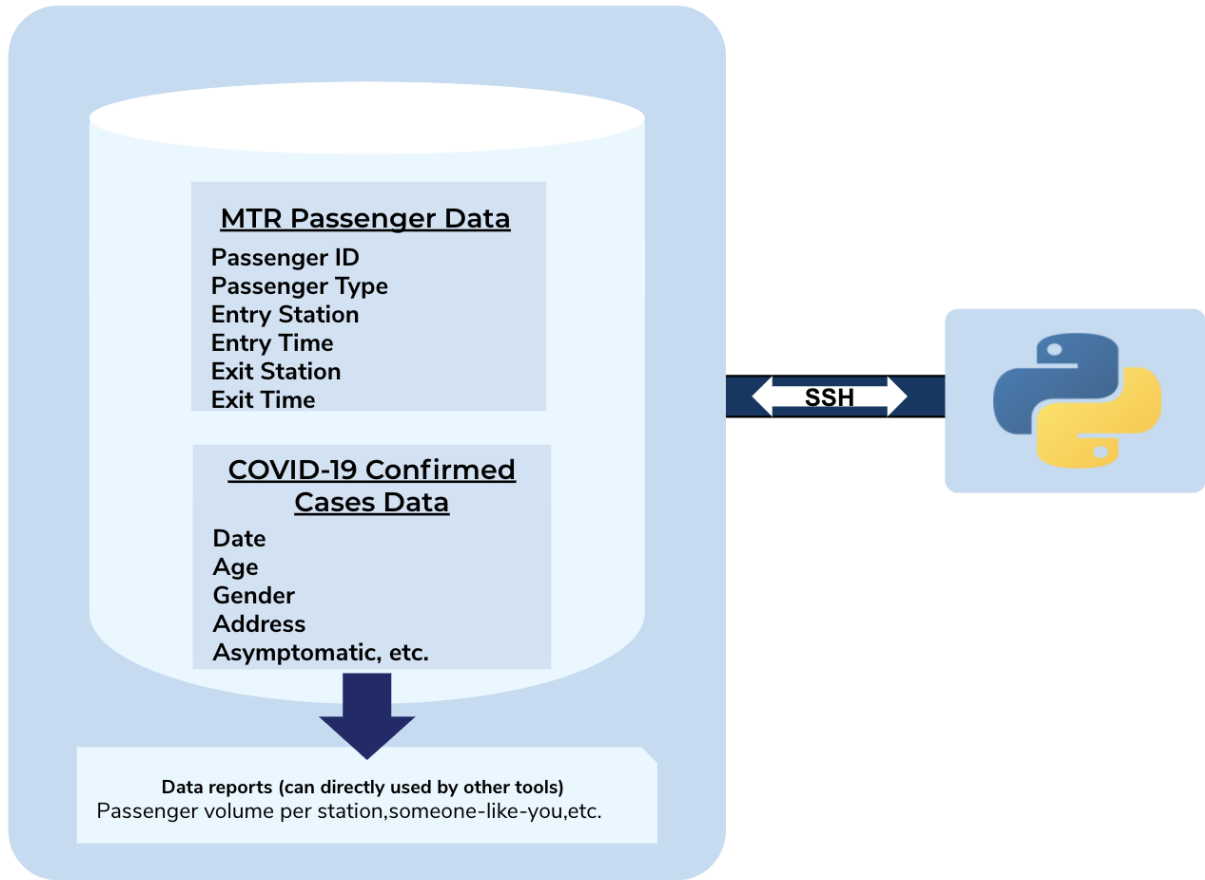


Figure 2.10 Platform architecture

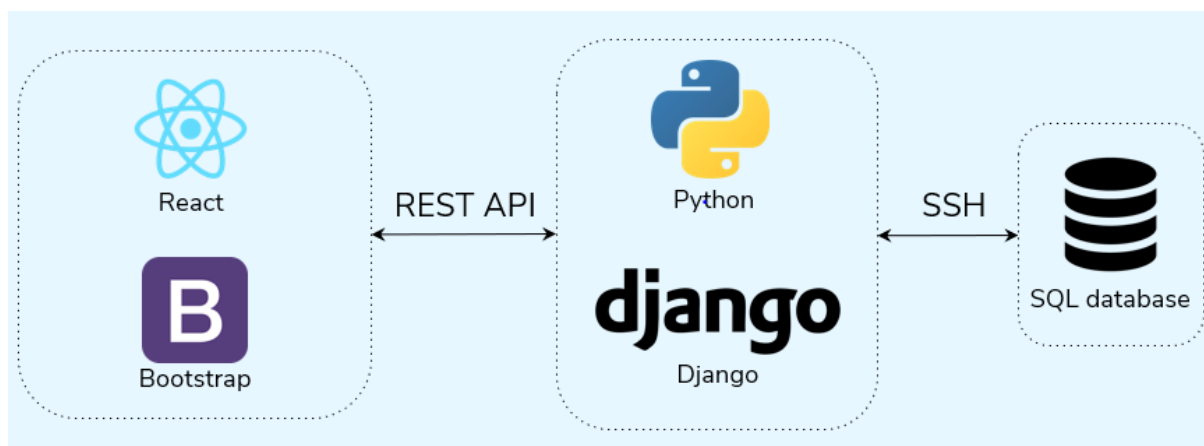


Figure 2.11 Technology stack

As detailed in figure 2.10, the data layer is held within the HKU CS Hincare server, only accessible via SSH to those authorized. The server layer hosts the querying mechanism, utilizing Python in conjunction with Django as its web framework. On the client layer shown in figure 2.11, the UI is rendered by React, a JavaScript library, in conjunction with Bootstrap so as to present an aesthetic front. React allows for UI components to be reused, which is especially necessary with the multitude of queries present within the system, and when we bring the issue of scaling into consideration. The visualization aspect of the system, with ArcGIS as its core backbone technology, is supported on the client layer as well.

The platform provides users with the means to query for MTR passenger data in a simplistic manner, in turn obtaining downloadable CSV files. Likewise, this functionality will be implemented for the COVID-19 data as well, so as to significantly expedite the process of data retrieval, in accordance with the user's needs. With the data pipeline in place, simply filling out several parameters on the user's end should be sufficient. The specific queries supported will be elaborated on in the results section.

Django employs a *Model-View-Template* design pattern - the team defines the API endpoints in *Views* and the *Models* operate by consuming this logic and integrating directly with the database objects. This query functionality is made possible through REST API calls to the backend server, which handles these requests accordingly. Django streamlines this process through its built-in support for passing state using request and response objects. Once the backend receives a request (i.e., a REST API call is sent to the running backend), Django encapsulates the request metadata into a `HttpRequest` object. This `HttpRequest` object handles the request type by directly querying

into the database defined and returning a QuerySet object, which is then parsed into columns stored in a CSV file. On the client side, the user receives this compiled CSV file containing the data requested.

In addition to this, to ensure data confidentiality (especially applicable to data provisioned by the MTRC), a security layer is implemented onto our platform through the use of JSON web tokens. We used the *rest_framework_simplejwt* module to accomplish this; verification of a user's login credentials is subject to the following details:

1. Upon login, a request is sent to the backend server, which uses the user's credentials to generate a JSON web token pair - the access token and the refresh token. The token expiration time is currently set to a maximum of 20 minutes, after which the user has to repeat the login step to obtain a new JSON web token pair.
2. To maintain access control, any subsequent requests sent to the backend server requires for the JSON access token to be verified in order to prove authentication of the credentials provided by the user.

ArcGIS is integrated onto the platform through the JavaScript API with credentials provisioned by ESRI. This API specifically serves the function of rendering such visualizations for web-based platforms. The module *arcgis-js-api* supports handling of user input in order to dynamically generate the ArcGIS visualizations, thus satisfying the user experience requirement of the platform. As opposed to the query mechanism which is run on the Python Django backend and subsequently consumed by the frontend through REST API calls, the ArcGIS visualizations are rendered directly on the frontend side due to its fairly lightweight and straightforward nature.

After the user specifies the required data for a preferred visualization type, a set number of steps are executed synchronously. Firstly, similar to the aforementioned queries, the data is fetched through REST API calls to the running backend service. This could include geolocation data indicating longitude and latitude of either the MTR stations, COVID-19 cases, buildings and more. The response is then used to create the ArcGIS Graphic objects, which are then combined into a single dataset contained within the ArcGIS FeatureLayer object. The FeatureLayer object acts as the data source for the ArcGIS Map object - a component that presents the baseline map for visualizations when rendered. Different visualizations are generated through altering the configurations and parameters for the Graphic and the FeatureLayer objects, which are customizable depending on the function of that particular metric.

2.6 Summary

This chapter proposed the workflow of implementation required in order to fulfill the technical requirements of the project. The engineering decisions were justified in each section – namely the choice of an SQL database, and the use of Python and ArcGIS for mobility trend analysis and geospatial analysis, and contact and behavior-based research. The technical requirements of the final product, the web application, is detailed in the last subsection. This includes details on our system architecture, technology stack and the interactions that occur between each layer.

3. Results

3.1 Overview

This chapter includes details of our preliminary findings for the first and second wave of COVID-19 in Hong Kong, features of our UI-based web platform, challenges we encountered and concludes with future improvements that can be made to said platform.

3.2 First wave of COVID-19 in Hong Kong

Hong Kong’s first wave of COVID-19 started with the first confirmed case on 23 January 2020 and continued on until early March. Our preliminary findings reveal a significant decrease of 44% in MTR travel within Hong Kong during the first wave. Interestingly, this occurrence was recorded during the Chinese New Year holiday, wherein travel is generally popular within Hong Kong as well as between Hong Kong and its neighboring cities.

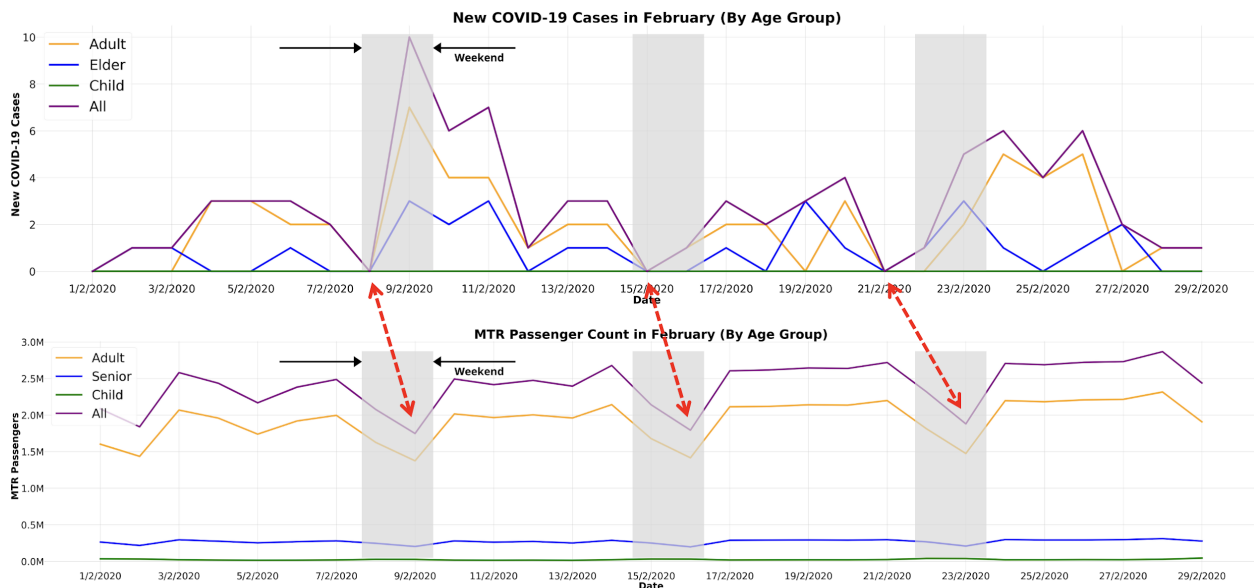


Figure 3.1 COVID-19 cases and MTR passenger count in February 2020 (by age group)

Figure 3.1 showcases a piece of evidence of the correlation between COVID-19 cases and the MTR passenger count. In this figure, the MTR passengers and the COVID-19 cases in February are grouped by age demographic - Adult, Elder, and Child, and the weekends are highlighted in the grey bars. There seems to be a general trend in the decrease of MTR passengers on weekends. However, through this side-by-side comparison, we can observe that the two graphs decline at similar time periods, implying that the decrease in MTR passengers and the decrease in COVID-19 cases are correlated with one another.

3.3 Second wave of COVID-19 in Hong Kong

Hong Kong faced a second wave of COVID-19 infections around mid-March 2020, where the new confirmed cases started to comprise mostly of imported cases (Lew, 2020). The second wave continued on throughout April.

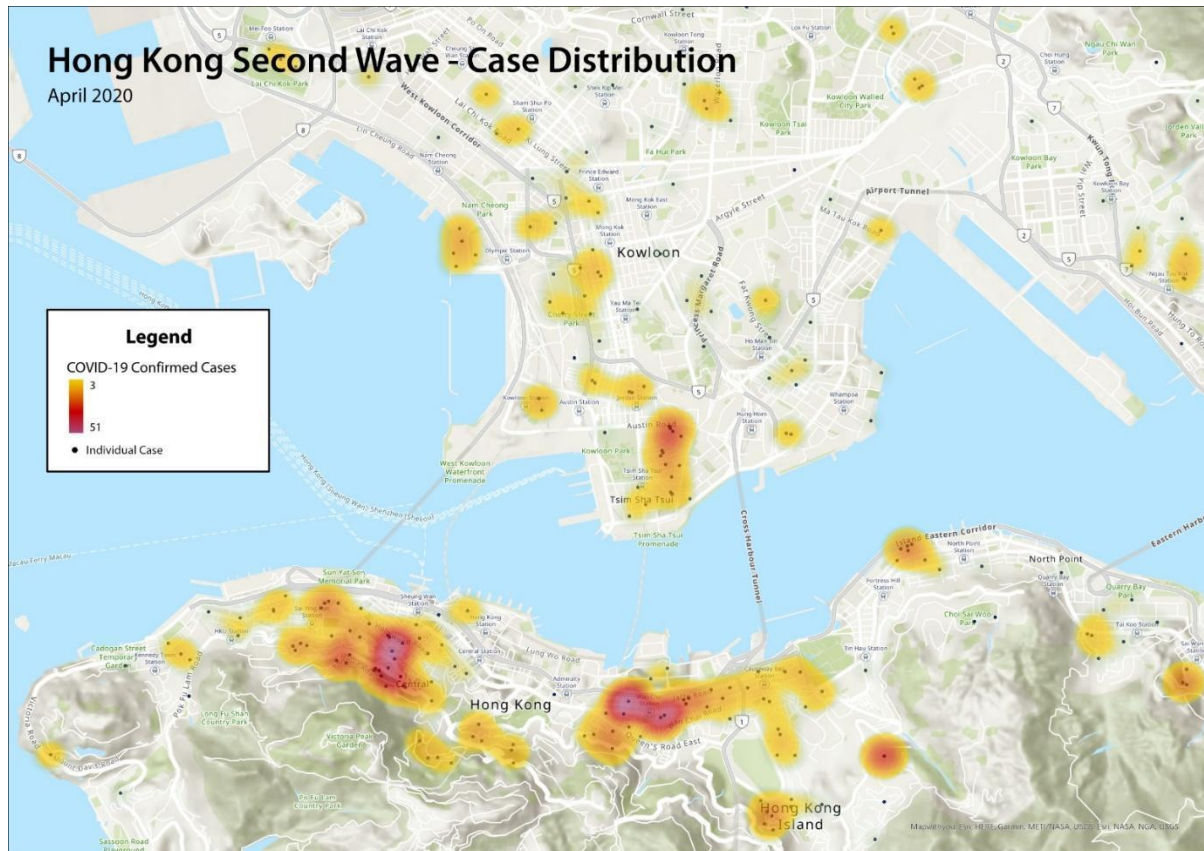


Figure 3.2 COVID-19 case distribution during Hong Kong’s second wave

Figure 3.2 pinpoints the individual COVID-19 cases scattered throughout Hong Kong in April 2020. As an attempt to isolate those areas most affected by the pandemic, we used the heat map feature in ArcGIS to highlight those confirmed cases.

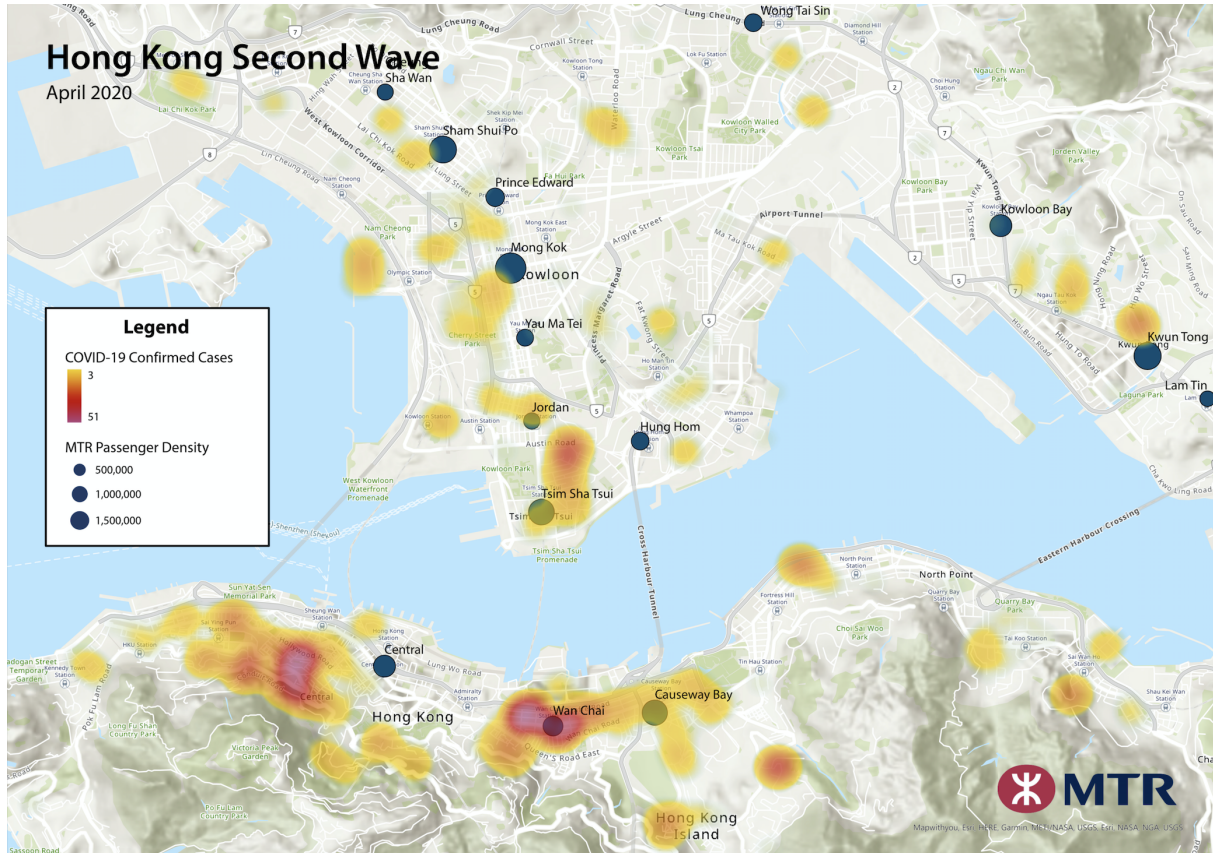


Figure 3.3 MTR station density and COVID-19 hotspots in April 2020

Figure 3.3 illustrates the density of each MTR station - in other words, how crowded each MTR station is - juxtaposed against the diagram in Figure 3.3. The most prominent finding from this figure can be inferred from Tsim Sha Tsui and Wan Chai. With the large MTR passenger density in those two areas, it is no surprise that they are also observed to be coronavirus hotspots, as seen from the red heatmap spots around them.

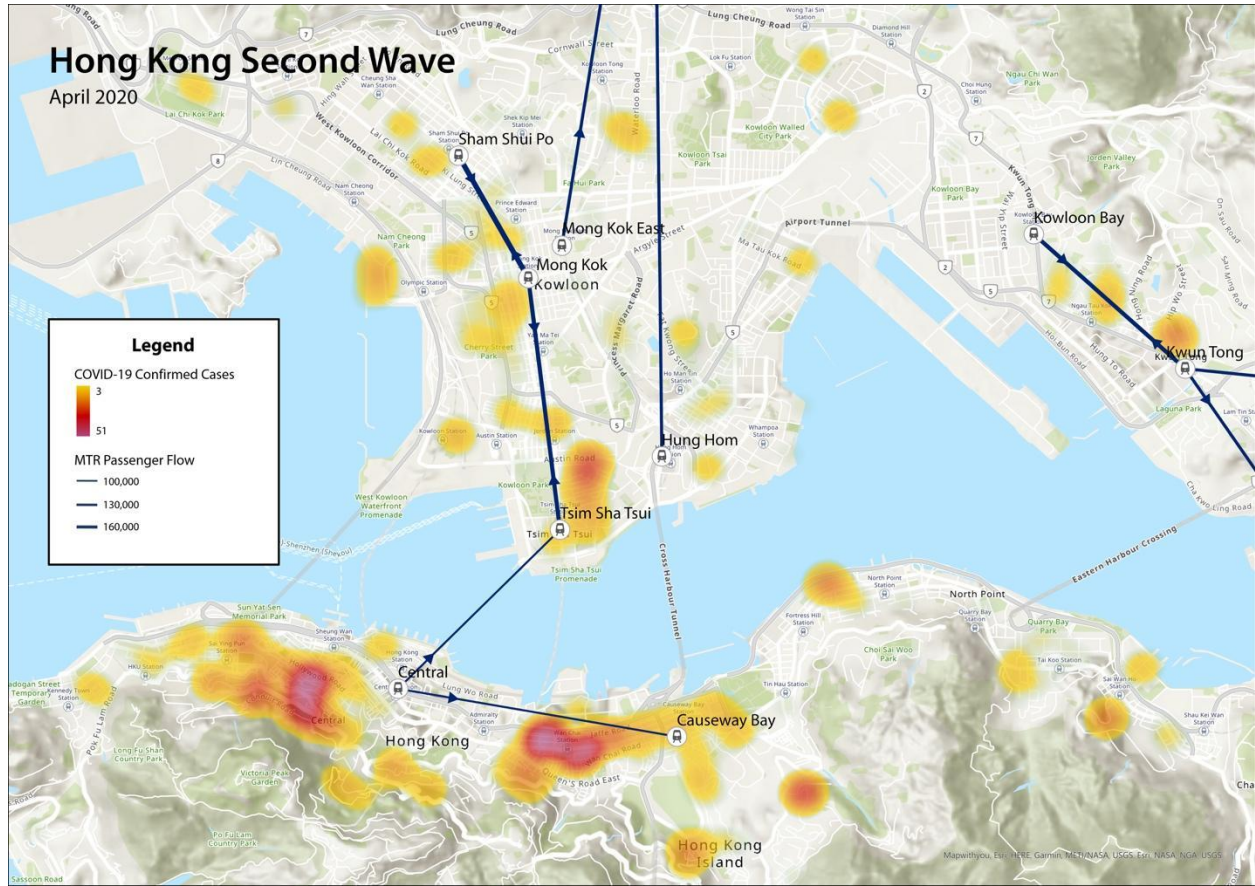


Figure 3.4 Most popular MTR routes and COVID-19 hotspots in April 2020

Figure 3.4 displays MTR passenger flow along popular MTR travel routes, as well as the number of COVID-19 confirmed cases in Hong Kong during the month of April 2020. From this figure, we can observe a trend wherein confirmed cases tend to spread across some of these routes. This effect is most prominent between the following MTR stations: Central and Causeway Bay, Sham Shui Po and Mong Kok, Mong Kok and Tsim Sha Tsui, Kowloon Bay and Kwun Tong. The above findings further emphasize on how MTR travel patterns have contributed to coronavirus spread.

3.3 MTR network route changes in COVID-19

The figures created below are meant to visualize the busiest MTR routes, narrowed down to the top 20 preferred journeys taken by passengers in January 2020 (Figure 3.5) and April 2020 (Figure 3.6). These months are selected specifically for the purpose of evaluating the travel pattern changes from when the pandemic issue came into existence, to when the second wave emerged.

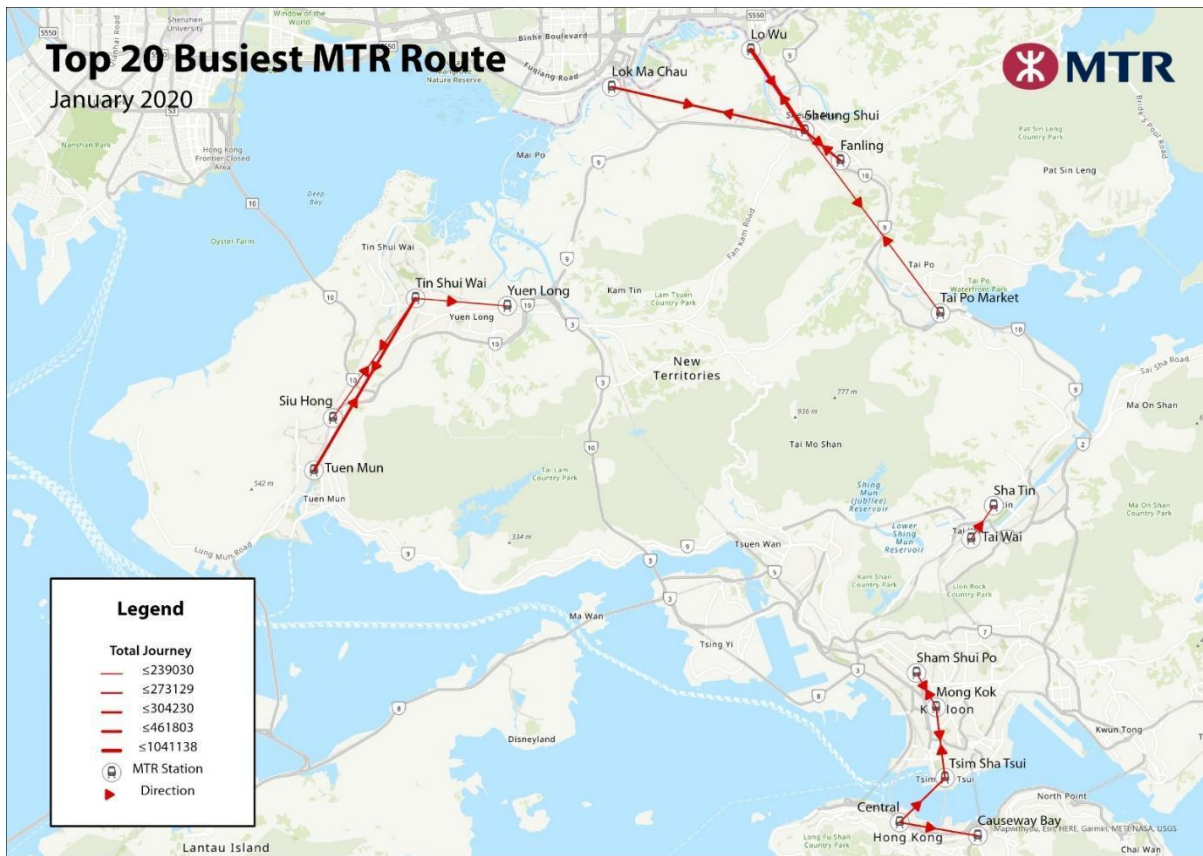


Figure 3.5 Top 20 busiest MTR routes in January 2020

In January 2020, as seen in figure 3.5 above, the busiest route hosted approximately 1 million passengers.

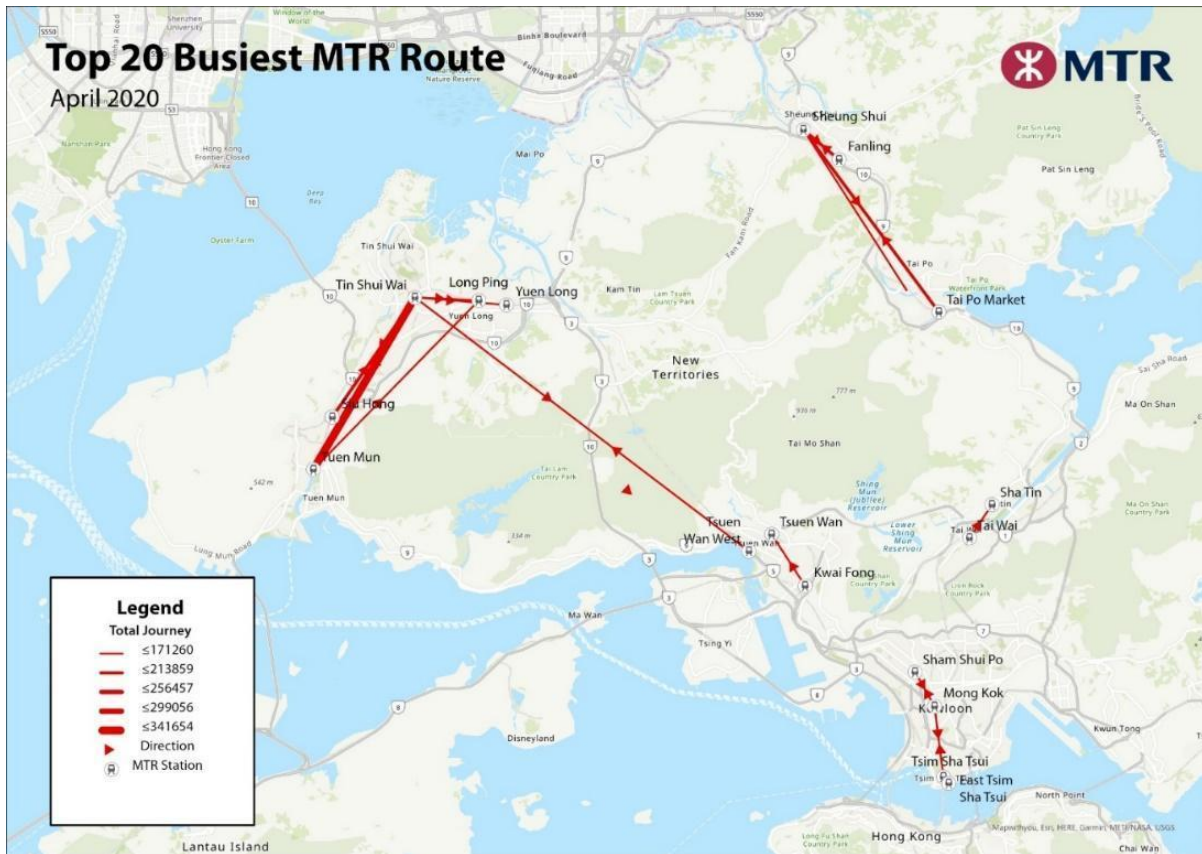


Figure 3.6 Top 20 busiest MTR routes in April 2020

In April 2020, however, Figure 3.6 uncovers that the number of passengers on the busiest route has dropped to approximately 300,000. A significant drop of 67% is observed in the passenger volume. We accredited this change to the pandemic's severe consequences on public transport.

3.4 UI-based web platform for MTR and COVID-19 data

The team has successfully deployed our web platform on our designated FYP machine server. The FYP machine exposes the URL <http://fyp20035s1.cs.hku.hk/> for client usage. However, the backend server, which is hosted on a separate port, only grants authorized users access to its endpoints. The following subsections describe the features available for use upon successful login and validation of credentials. The login page is shown in Figure 3.7 below.

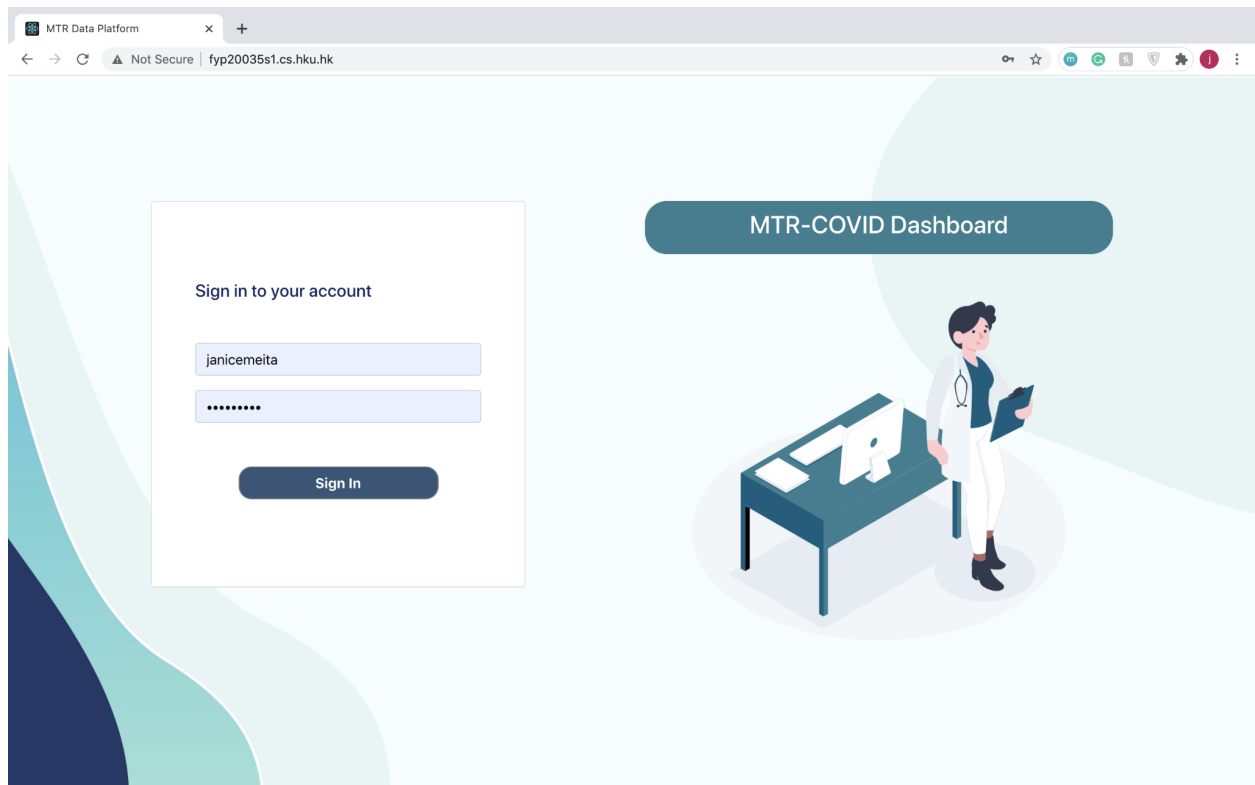


Figure 3.7 Login page

3.4.1 Query

Querying is made accessible through our platform - users have the privilege of obtaining raw MTR data without having to acquire the original DVDs commissioned by the MTRC.

Furthermore, this data has been enhanced through our pre-processing steps, lest to assure that the data obtained is in accordance with the user's needs. There are several query metrics supported by our platform at present.

Firstly, the **Travel Pattern** metric takes into account the MTR travel routes as a method to understand the MTR passenger flow better. The output columns are: Origin, Destination, Date, and Volume.

The screenshot shows a web browser window with the URL `fyp20035s1.cs.hku.hk/query`. The page features two database selection buttons at the top: "MTR Passenger Database" and "Covid-19 Cases Database". Below these is a "Metrics" section with a dropdown menu currently set to "TravelPattern". The main content area is titled "Travel Pattern" and includes a descriptive text: "Travel pattern analysis takes into account the MTR travel routes as a method to understand the MTR passenger flow better." Below this text, it lists "Output Columns : Origin | Destination | Date | Volume". The interface includes two input panels: "Date" with "Start Date" and "End Date" fields (both showing "dd/mm/yyyy" and a calendar icon) and "Station" with "Starting Station" and "Destination Station" dropdown menus (both showing "Select..."). A "Download" button is located at the bottom right of the form.

Figure 3.8 Travel pattern query page

Origin	Destination	Date	Volume
1	2	1-2-2020	1000

Table 3.1 Travel pattern query sample output

The output is sampled in Table 3.1 above. This signifies that, on 1-2-2020, there are 1000 passengers traveling from origin MTR station with code 1 to destination MTR station with code 2.

In addition to travel pattern analysis, **Passenger Mobility** analysis can be used to capture the general trend in the number of MTR passengers in a given time frame. This query comes with the option to filter by Octopus card type (Adult, Child, Senior, Student) as well, an option which might be beneficial in demographic-centered studies. The output columns are: Date and Volume.

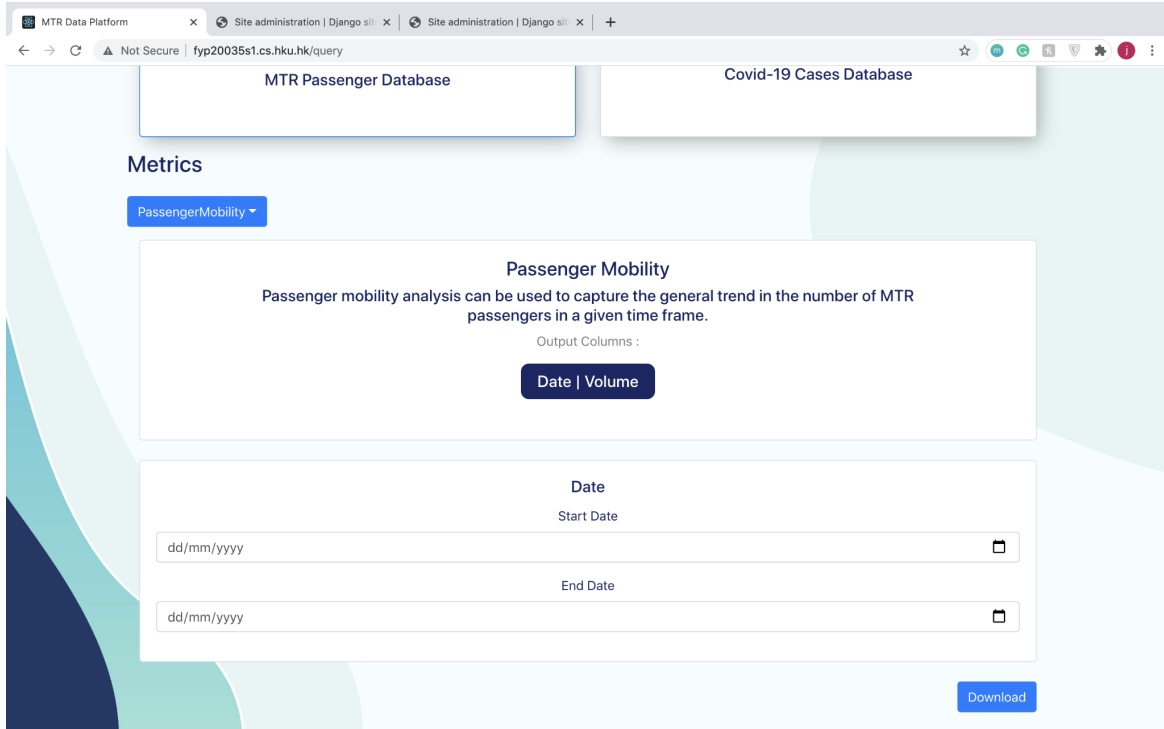


Figure 3.9 Passenger mobility query page

Date	Volume
1-1-2020	10000

Table 3.2 Passenger mobility query sample output

The sample output shown in Table 3.2 indicates that on 1-1-2020, there were a total of 10000 passengers using the MTR as a form of public transport. The output for card type-specific queries is similar to the above, except it is filtered by the chosen demographic grouping.

Next, **Station Density** is meant to illustrate the density of each MTR station (i.e., how crowded each MTR station is) within a given time frame. The user could choose to query for daily station

density or hourly station density. The output columns are: MTR Station Code, Date and Volume.

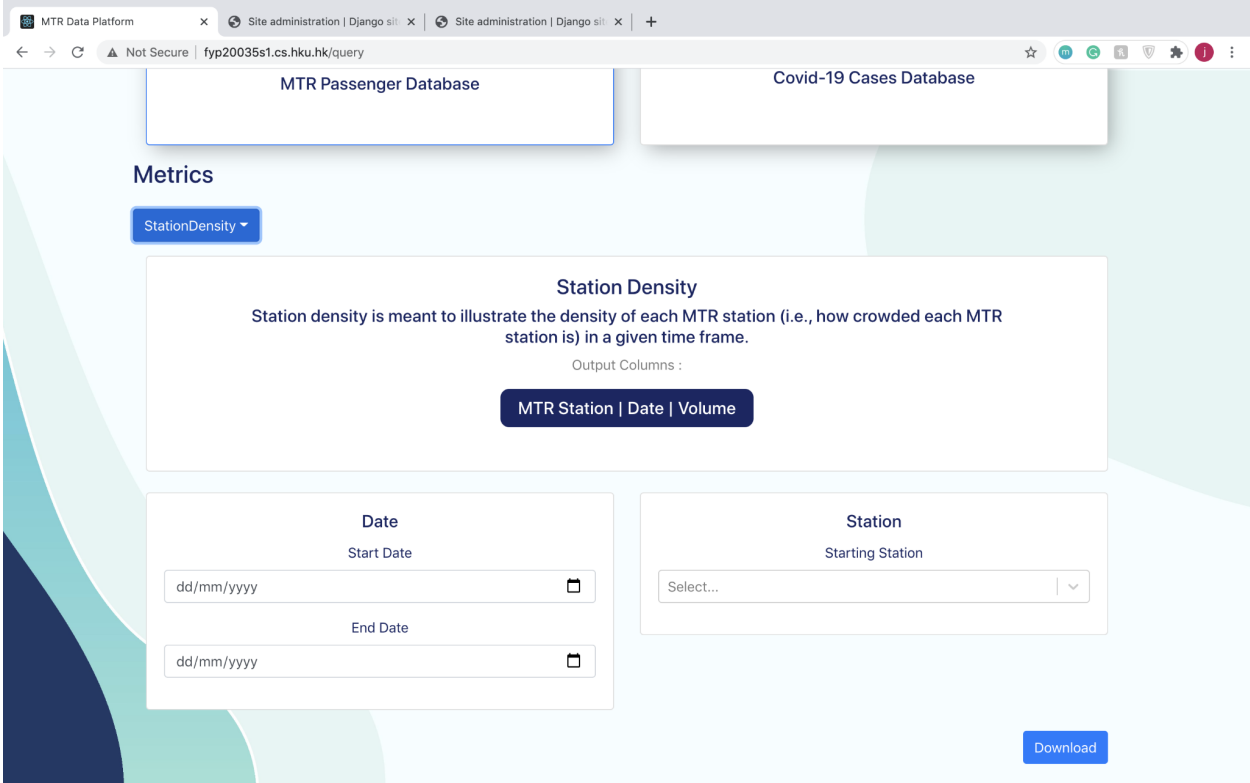


Figure 3.10 Station density query page

MTR Station	Date	Volume
1	01-01-2020	100000

Table 3.3 Station density query sample output

Table 3.3 again shows the sample output: MTR station of code 1 was crowded with 10000 passengers on 01-01-2020.

The platform supports **Raw Data Queries**: querying with no specifications (i.e., detailed columns from the MTR database) as well, which fulfills the use case wherein the user requires raw data. The output columns are: ID (with anonymity preserved), Card Type, Entry Station, Exit Station, Entry Time, Exit Time. However, given the timeout constraints with querying large volumes of raw data, we engineered the decision to separate raw data queries into queries within the same month, and queries specifying an entire month's worth of data in order to optimize the process.

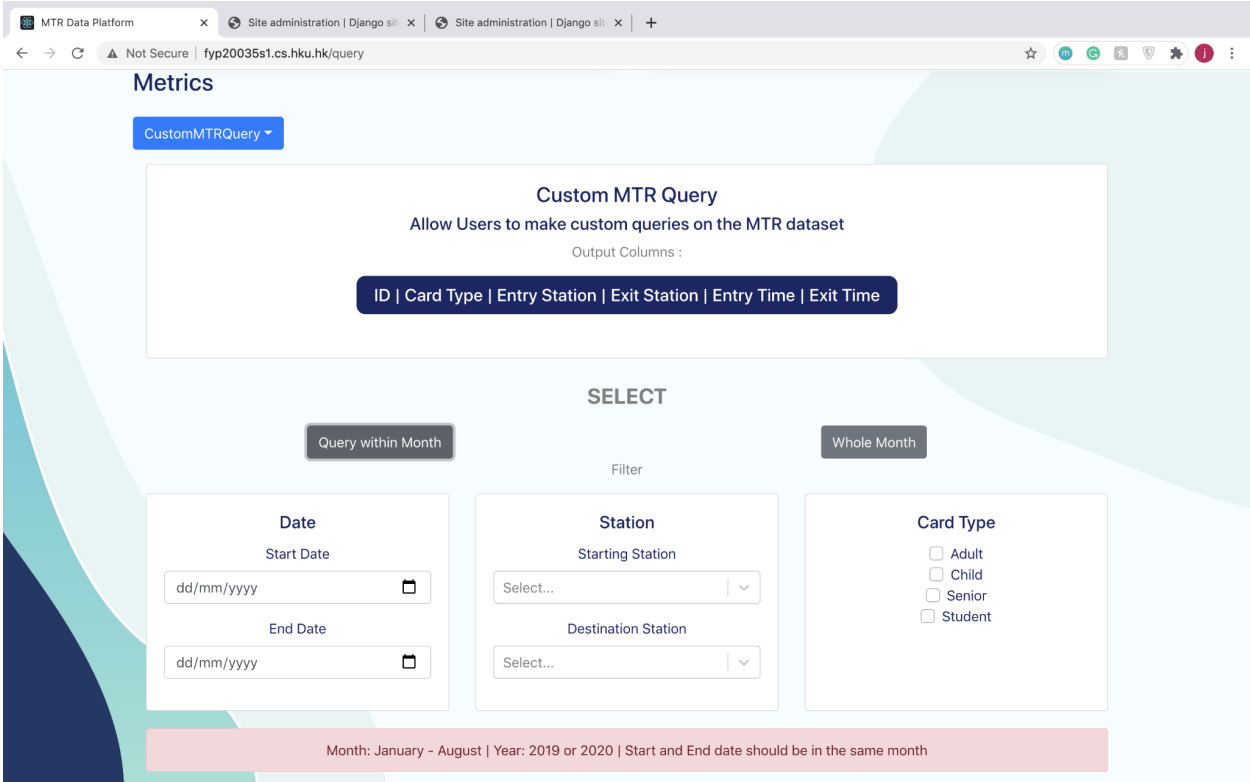


Figure 3.11 Raw data query page

ID	Card Type	Entry Station	Exit Station	Entry Time	Exit Time
12345	SEN	1	12	2020-02-01 06:00:00	2020-02-01 06:20:00

Table 3.4 Raw data query sample output

A row of the raw MTR data is sampled in Table 3.4. It contains information on a passenger with ID 12345, card type SEN (senior) who entered MTR station with code 1 at the timestamp of 2020-02-01 06:00:00 and exited station 12 at the timestamp of 2020-02-01 06:20:00.

3.4.2 Visualization

The visualization page features relevant graphs as well as geospatial visualizations for both the MTR passenger data and the COVID-19 data.

One of the visualization options currently supported is **Station Density**. The primary objective of these diagrams is to infer the crowdedness of MTR stations located in Hong Kong. The user's input is customized by date range, and number of stations which is dynamically rendered to correspond to the range slider input. This constructs an interactive map visualization of the station density, wherein the circles shown (as in Figure 3.12) are proportionate to the passenger volume in each station. If desired, the user can bring attention to a particular station as seen in Figure 3.12; the pop-up holds the details for the station selected. As a supplementary feature, the platform also recognizes that the user might need to visualize the COVID-19 cases data on the same map. This is supported by overlaying said data onto the station density map (option 'with

COVID data shown in Figure 3.13), so as to take into account the correlation between the two datasets, in an effort to further understand the sequence of events.

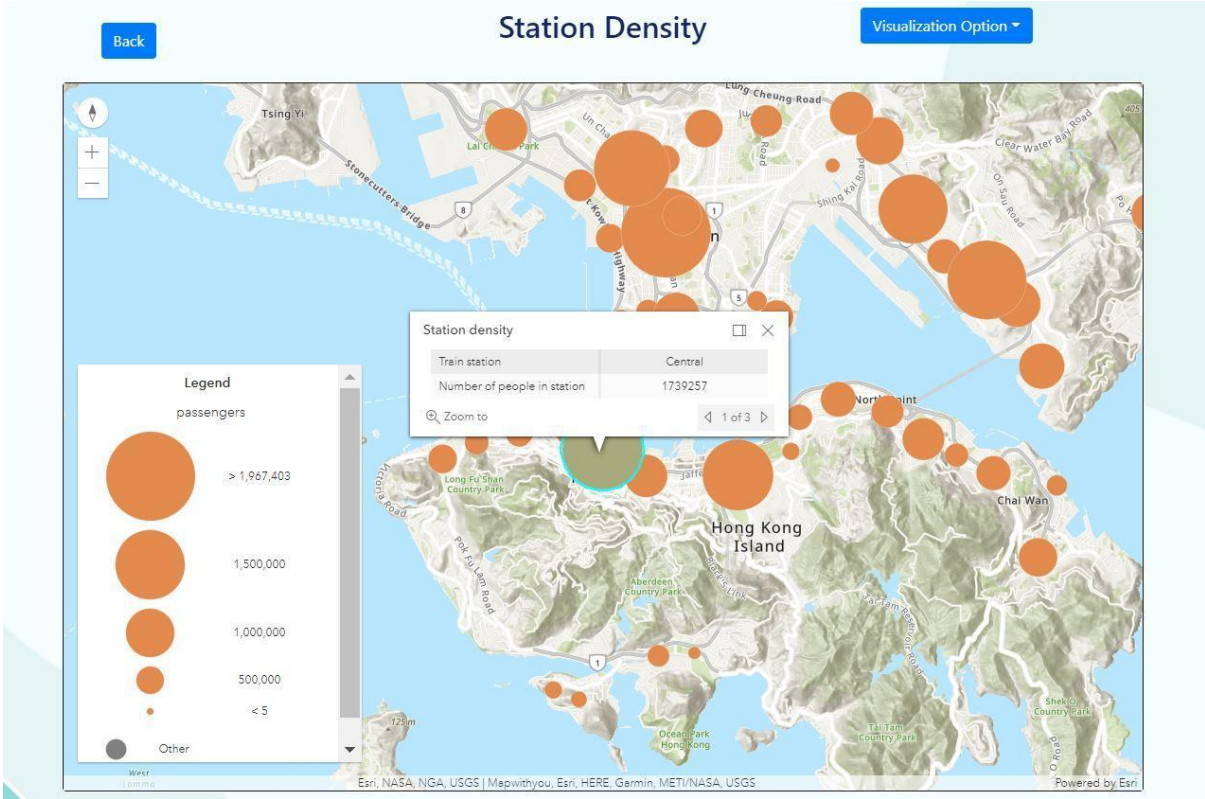


Figure 3.12 Station density visualization page with station details

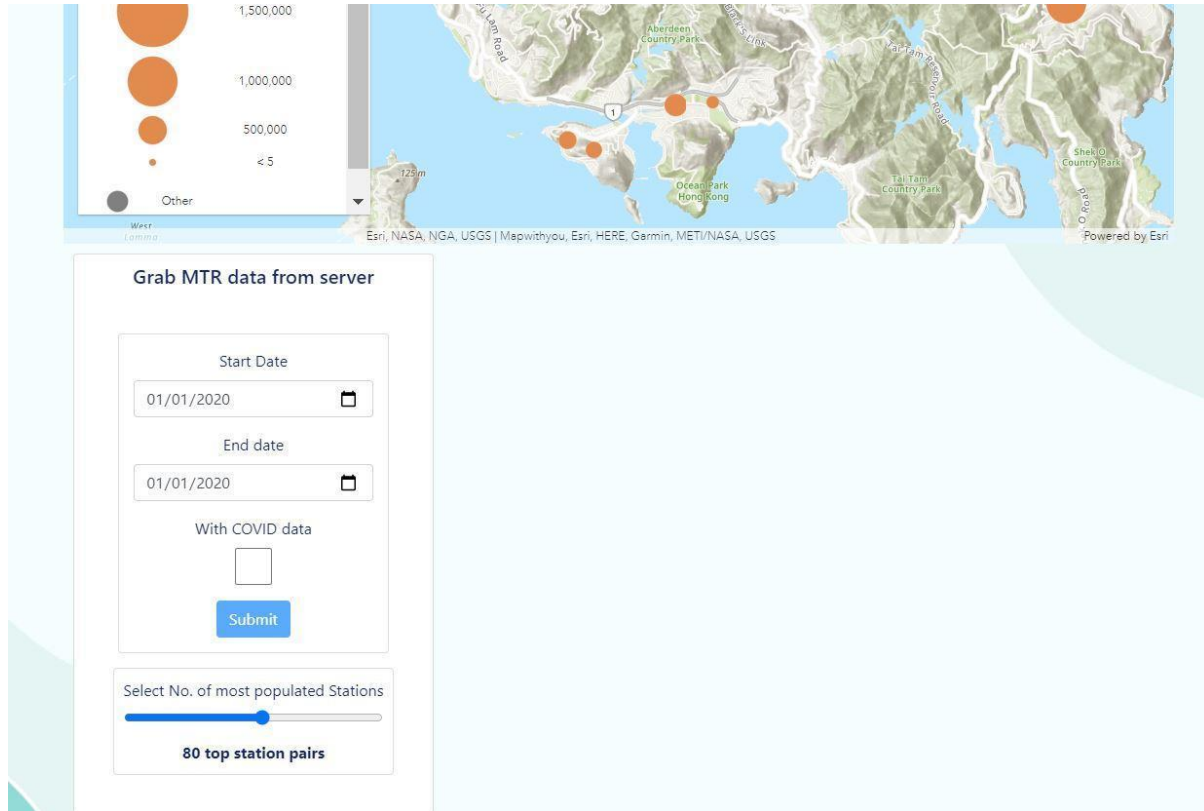


Figure 3.13 Station density visualization page with input form

Another visualization option a user can choose to explore is **Travel Pattern**. Similar to station density, the user customizes their desired visualization through specifying the date ranges as seen in Figure 3.14. However, this option delves into the specific routes - that is, travel volume trends between station pairs - instead of focusing on each MTR station. The range slider may be used to indicate the top N preferred routes, sorted by passenger volume. As with station density, COVID-19 confirmed cases data could be overlaid with this visualization as well.

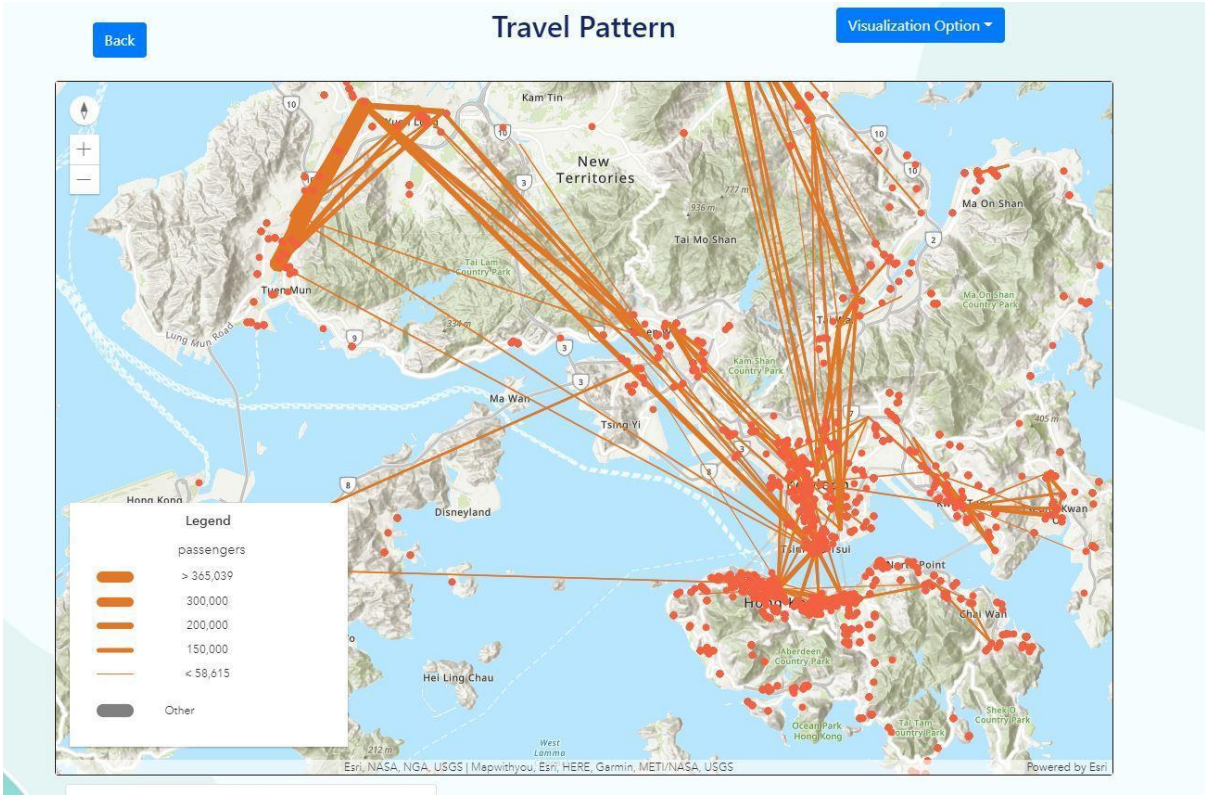


Figure 3.14 Travel pattern density visualization page

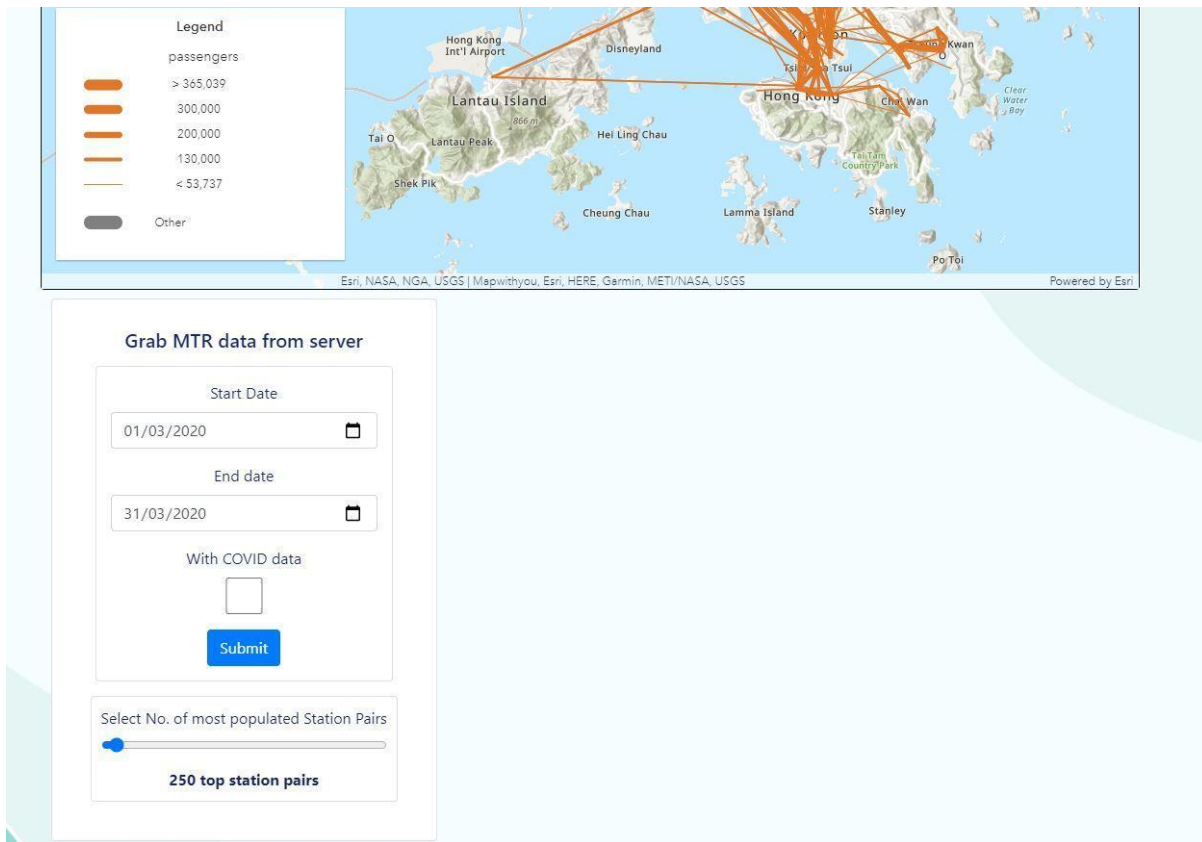


Figure 3.15 Travel pattern density visualization page with input form

Passenger volume captures the mobility trend in the form of a time series graph of passenger volume plotted against time. With this option, the trend in overall passenger activity can be gathered, especially through specifying key date ranges that are most affected by the pandemic. The graph is rendered upon successfully requesting the MTR passenger data from the server.

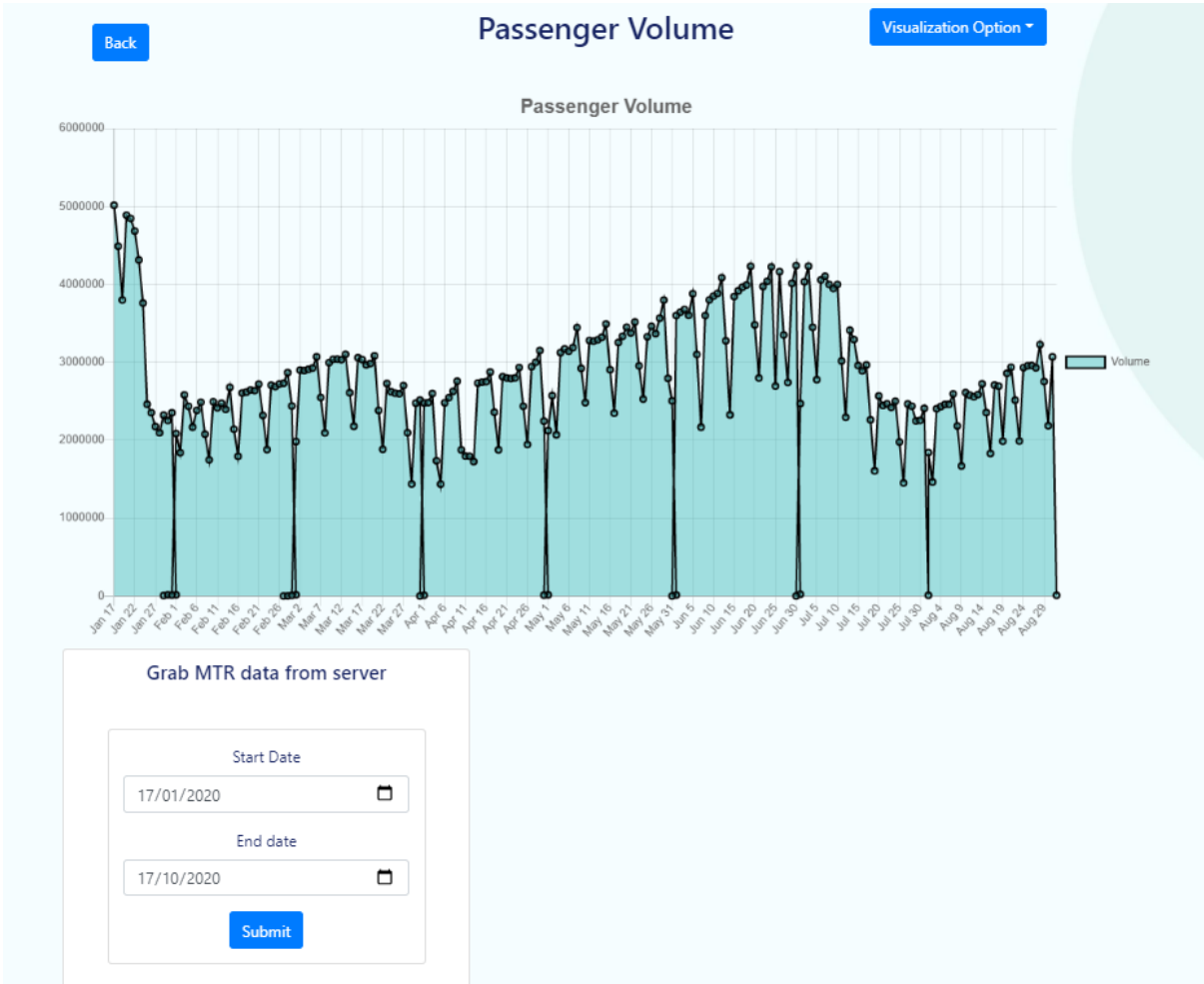


Figure 3.16 Passenger volume visualization page

3.4.3 Analysis

The advanced analysis feature summarizes the findings from the contact and behavior-based research as detailed in the methodology section.

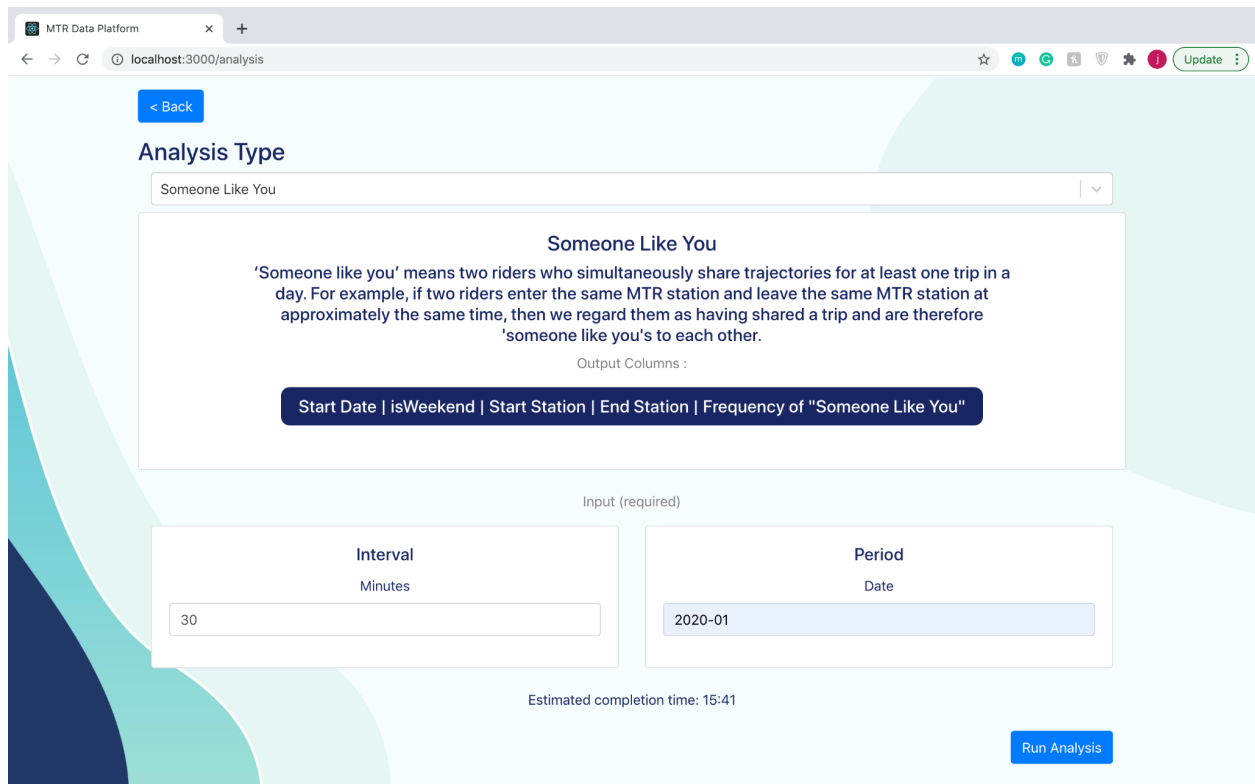


Figure 3.17 Someone Like You analysis page

Figure 3.17 shows the **Someone Like You** analysis page. Users are prompted to indicate the interval, which narrows down the timespan of ‘someone like you’ encounters. The default value is set to 30 minutes. However, as discussed in the methodologies we adopted, decreasing the parameter could lead to increased sensitivity. This is customizable depending on the user’s

research requirements - certain users might require investigation of those ‘someone like you’ encounters within a shorter timespan, while others might prefer dividing the day into fewer time span groupings to gain a more holistic overall perspective. In addition to this, a particular date needs to be chosen for this analysis to be conducted.

Date	Weekend	Start_Stn	End_Stn	Freq_SLU
1/3/2020	TRUE	1	16	8.61

Table 3.5 Someone Like You sample output

A sample output is shown in Table 3.5. The date column shows the starting date of the week. The above signifies that there is an estimated 8.61 *someone like you*'s on the weekend of week 1/3/2020, entering the MTR from station code 1 and exiting from station code 16.

However, this advanced analysis is subject to a shortcoming - given the limitations of the server's computation power and the volume of data to be processed, running this would take up a considerable amount of time. In order to ensure that our users are aware of this, we have displayed the estimated completion time (set to 8 minutes after the request is sent) as a prior warning. In the case that the analysis ends up running longer than the expected time, users have the option to report this anomaly back to our team.

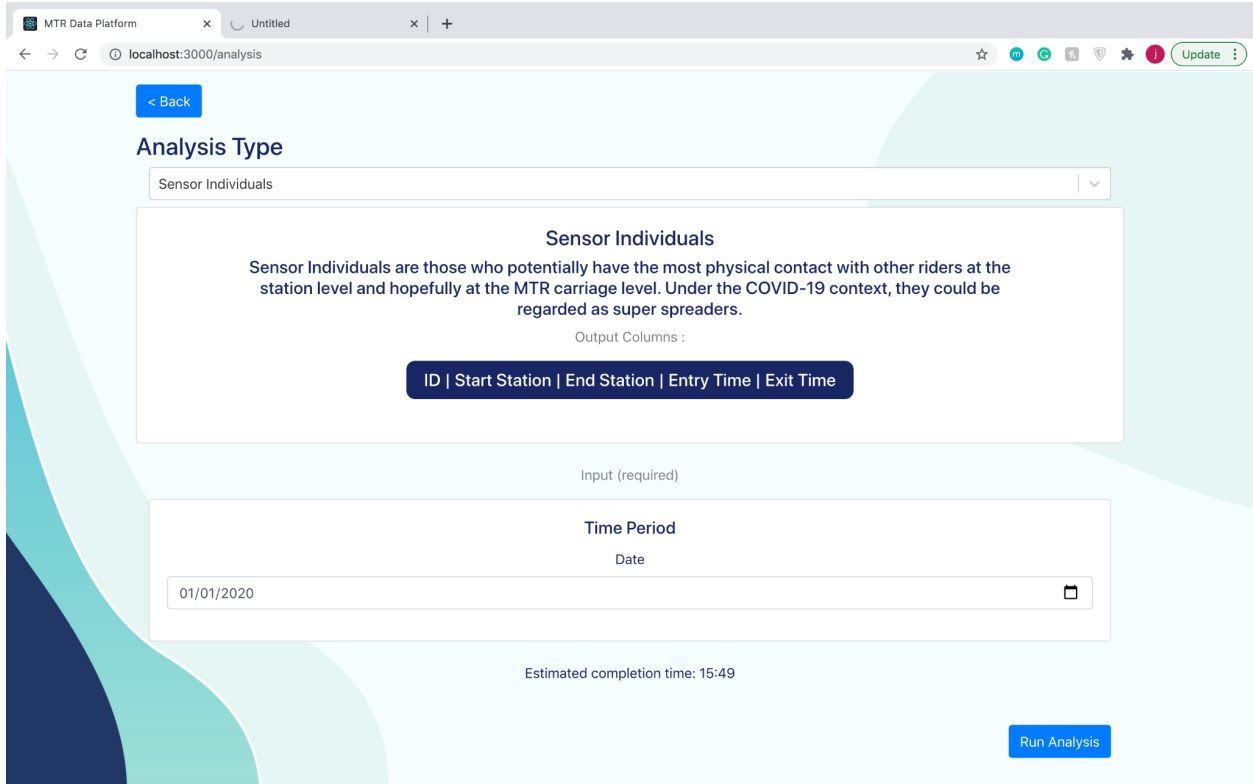


Figure 3.18 Sensor individuals analysis page

Figure 3.18 shows the **Sensor Individuals** analysis page. The required input for this analysis is the time period - date and time. The output lists down all the sub-paths (with their respective timestamps) throughout the journeys taken by all MTR passengers on the date and time specified. This data is available in the form of a preconfigured 10-minute time span; however, this parameter can be easily tweaked later on.

ID	Start_Stn	End_Stn	Entry_Time	Exit_Time
12345	1	3	1/1/2020 12:05:50 AM	1/1/2020 12:07:59 AM

Table 3.6 Sensor individuals sample output

A sample output of this analysis is shown in Table 3.6 above. This row signifies that a passenger with ID 12345 entered MTR station 1 and exited MTR station 2 at 12:05:50 AM and 12:07:59 AM respectively, on the date of 1/1/2020.

As with the *someone like you* analysis, given the limitations in computation power, the user is provided with an estimated completion time, which has been set to around 15 minutes following the request time.

3.5 Challenges

3.5.1 Nature of Dataset

In the process of working with the MTR dataset to complete several analyses, we were faced with the challenge of having to make several assumptions. This is especially prevalent in the contact and behavior analysis aspect of our project - since the dataset only includes the starting station and the exit station, we are unable to accurately pinpoint a passenger's route, therefore having to assume that all passengers with the same starting and ending points had to have taken the same route. However, this only serves as a best estimate, as this is not always the case in reality. In addition to this, the only timestamps specified within the dataset is the entry time and the exit time. Because of this, we again have to assume that within this time frame the passenger stays in the paid area, he follows through with his MTR journey without pursuing any other activities within the MTR paid area.

Another potential challenge will be establishing which key dates to further look into. There are several factors that play into influencing MTR passenger behavior - for instance, the enforcement of a new government policy could influence the volume of passengers. The team will need to keep a record of these regulations and their implementation dates to recognize the extent of their effects.

3.5.2 Computation Power

Given that we are working with large volumes of data, the performance of our system is greatly dependent on the computation power of the server executing all these tasks. This might result in overhead at times when there is a large number of requests. In order to alleviate this issue, the team has utilized several optimization methods, listed below:

1. The results of all data requests are stored by the system. This means that whenever another user requests for the same data with identical parameters to a previously completed request, the backend simply requests for this CSV result without running the methods again. This is acceptable since we are working with historical data for this research project, hence the output is not subject to change no matter when the request is made.
2. For queries made to the database, we use indexing on the tables to improve the performance of those SELECT commands.
3. For queries that we expect to take up a large amount of time (i.e., queries with no specific parameters asking for raw data), we performed precomputation of the results and stored them into separate tables. These tables support direct SQL queries. We judged the

improvement in response time through conducting manual testing: each API call takes milliseconds to complete, as opposed to the initial completion time of 2-3 minutes.

Although we note that computation power could be further improved by hosting our data on public cloud services such as Google Cloud Platform or Amazon Web Services, using these publicly available platforms to analyse and store data is not advisable. This is due to the confidentiality and privacy issue imposed by our team's agreement with the MTRC. Hence, scaling, computation power and robustness of the system still depends on the specifications of our in-house server.

3.6 Future Improvements

3.6.1 Sensor Individuals

Currently, the analysis concerning sensor individuals only produces a list of passengers' intermediate locations within the MTR network during the specified timestamps. This data has the potential of being the baseline for further use cases. This includes the aforementioned use case of pinpointing a COVID-19 super spreader. As a proof of concept, we have experimented with a search algorithm where a user is able to specify one Octopus ID (perhaps the potential super spreader concerned) and the date. The output is in the form of the passenger details for anyone who has been considered to be this person's sensor individuals.

```

sensor_individuals('905294529', '2020-02-01')

Sensor Individual Pairs:
CSC_PHY_ID          905294529
START_STN           48
END_STN             32
ENTRY_TIME          2020-02-01 06:21:00
EXIT_TIME           2020-02-01 06:34:00
Name: 85, dtype: object
CSC_PHY_ID          904802245
START_STN           48
END_STN             32
ENTRY_TIME          2020-02-01 06:25:00
EXIT_TIME           2020-02-01 06:36:00
Name: 81, dtype: object

Sensor Individual Pairs:
CSC_PHY_ID          905294529
START_STN           32
END_STN             33
ENTRY_TIME          2020-02-01 06:34:00
EXIT_TIME           2020-02-01 06:43:00
Name: 86, dtype: object
CSC_PHY_ID          904802245
START_STN           32
END_STN             33
ENTRY_TIME          2020-02-01 06:36:00
EXIT_TIME           2020-02-01 06:43:00
Name: 82, dtype: object

```

Figure 3.19 Sensor individuals algorithm proof-of-concept

Figure 3.19 shows an example run of the algorithm we have devised using a sample of our data. An individual with ID 905294529 is said to have encountered 904802245 at two separate time frames.

However, this is still experimental given that the complexity of the algorithm is $O(n^2)$. This implies that, when this is run on the entirety of our dataset and not just a sample as shown above, the analysis could end up taking a long time to complete. In the future, we will look to improve this and later implement this as part of the analysis aspect of our platform.

3.6.2 Real time location-based COVID-19 alerts

Given the amount of insights that could be derived from combining both the MTR dataset and the COVID-19 dataset, the team recognizes that the platform could be of use to alleviate risk of COVID-19 exposure especially in those MTR areas with extensive passenger volumes. However, the system currently only utilizes historically available information. Especially given the erratic development of the COVID-19 pandemic, using historical data is not sufficient to accurately devise a risk metric. Agencies (2020) mentioned that a certain mutation in the coronavirus could increase its infectious ability. In view of possible inaccuracies, we need to develop a data stream engine for real-time coronavirus data such that this is used in addition to the historical data to generate our predictions.

This data stream engine would need to be capable of asynchronously updating the database based on information provided by the MTRC and HKCHP. As our current agreement only provides us with access to a few months of historical MTR passenger data, the team does not have the appropriate resources to develop this at the current stage. However, the team is confident that with some improvements to the computing capabilities, the current system could be scaled to handle this use case. Some possibilities include using the data reports generated from the contact and behavior-based analysis to calculate a risk of exposure metric based on the number of ‘Sensor Individuals’ and/or ‘Someone Like You’ in close proximity to the user.

3.7 Summary

This chapter detailed the results and features of our platform, and outlined the potential challenges. The team has successfully rolled out the data pipeline and conducted several

preliminary data analyses. The latter has allowed us to identify significant relationships from the general mobility trend and from a geospatial point of view. In addition to this, the web platform implementing the query, visualization and analysis aspects detailed above, has been deployed onto the designated FYP server and can be readily utilized by any researcher or interested party who has been granted the appropriate access. This chapter also recognizes that there are several areas for future improvements to the system. The concluding chapter will summarize the findings so far and reiterate the key objectives of the project.

4. Conclusion

This report has detailed the background behind the hypothesis, outlined the methodology and described interim results proving the correlation hypothesis, before suggesting potential areas of further research. In addition to gaining a better understanding of the correlation between MTR passenger behavior and the development of COVID-19 in Hong Kong, this project aims to provide a platform where users, especially professionals in the research field, are able to explore visualizations in a simplistic, efficient manner. As a supplementary feature, the web application hopes to assist users in generating queries and advanced analysis for assessing COVID-19 risk based on passenger's behaviors. This information serves as the baseline for improvements to be made in multiple aspects - namely the avoidance of COVID-19 risk in MTR routes, and also specific MTR stations. The eventual goal is to contribute to future research regarding the 'familiar strangers' theory and the field of big data as a whole.

References

- Agencies. (2020, June 17). *Coronavirus mutation makes it even more infectious and is emerging as the dominant kind, research finds*. South China Morning Post. Retrieved from <https://www.scmp.com/lifestyle/health-wellness/article/3089361/coronavirus-mutation-makes-it-even-more-infectious-and>
- Boyd, D. & Crawford, K. (2011, September 21). Six Provocations for Big Data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. Retrieved from <https://ssrn.com/abstract=1926431>
- Cheung, E. (2020, January 22). *China coronavirus: death toll almost doubles in one day as Hong Kong reports its first two cases*. South China Morning Post. Retrieved from <https://www.scmp.com/news/hong-kong/health-environment/article/3047193/china-coronavirus-first-case-confirmed-hong-kong>
- Cheung, T., Lum, A., Cheung, E., & Sum, L. K. (2020, January 28). *China coronavirus: Hong Kong scrambles to roll out containment plan stopping short of total closure, with cuts on cross-border travel and reduced transport services with mainland*. South China Morning Post. Retrieved from <https://www.scmp.com/news/hong-kong/health-environment/article/3047907/china-coronavirus-hong-kong-government-deny-entry>
- European Centre for Disease Prevention and Control. (2020, June 30). *Transmission of*

COVID-19. Retrieved from

<https://www.ecdc.europa.eu/en/covid-19/latest-evidence/transmission>

La Rosa, G., Bonadonna, L., Lucentini, L., Kenmoe, S., & Suffredini, E. (2020). Coronavirus in water environments: Occurrence, persistence and concentration methods – A scoping review. *Water Research*, 179, 115899.

Lew, L. (2020, March 19). *As Beijing, Hong Kong face second coronavirus onslaught, quarantine gets serious*. South China Morning Post. Retrieved from <https://www.scmp.com/news/china/society/article/3075880/beijing-hong-kong-face-second-coronavirus-onslaught-quarantine>

Liang, D., Li, X., & Zhang, Y. Q. (2016). Identifying familiar strangers in human encounter networks. *Europhysics Letter*, 116(1), 18006.

MTR Corporation Limited. (2020). *2019 Annual Report of the MTR Corporation Limited*.

Retrieved from

<https://www.mtr.com.hk/archive/corporate/en/investor/annual2019/EMTRAR19.pdf>

SSH (Secure Shell). (2020). Retrieved 27 October 2020, from <https://www.ssh.com/ssh/>

What is a relational database?. (2020). Retrieved 27 October 2020, from

<https://www.oracle.com/hk/database/what-is-a-relational-database/>

What is MySQL?. (2020). Retrieved 27 October 2020, from

<https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html>

World Health Organization. (2020, January 12). *Novel Coronavirus – China*. Retrieved from

<https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/>

Zhang, F., Jin, B., Ge, T., Ji, Q., & Cui, Y. (2016). Who are my familiar strangers? Revealing hidden friend relations and common interests from smart card data. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 619-628.

Zhou, J., Yang, Y., Ma, H., & Li, Y. (2020). “Familiar strangers” in the big data era: An exploratory study of Beijing metro encounters. *Cities*, 97, 102495. doi:

10.1016/j.cities.2019.102495