COMP4801
Final Year Project

Final Report

Author: Rishabh Jain (UID - 3035453608)

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

Supervisor: Prof. Cheng Reynold

Members:

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

18 April 2021

I

Abstract

With the emergence of the coronavirus, called COVID-19, the whole world has taken a strong hit in all facets of life. As a result, the economies have fallen globally, health conditions have deteriorated, death rates have risen and population lifestyles and productivity have crumbled. Thus, there is a significant need for the formulation of policies that alleviate the further effects of the pandemic. Hong Kong is a place with a well-connected network of public transport provided majorly by the MTR which may be one of the root causes of the aggravation of the COVID-19 spread. However, public transport is crucial in maintaining the population's lifestyle and productivity, and promoting mass transport to reduce pollution. Hence, a strict restriction on MTR-usage could cause more societal-damage than benefit. Hence, a balanced set of policies is required to tackle this issue sustainably. Accordingly, an in-depth understanding of the relationship between MTR passenger behaviour and COVID-19 spread could highlight major factors that result in this causality. The project, therefore, aims to define this relationship through a thorough analysis of available data regarding COVID-19 statistics and passenger trends, with a goal of making this analysis available to the government (and the public) to facilitate better decisions and policy formulation. This report presents Big Data, along with other methods, as an approach to formulate a solution to the abovementioned problem. It also explains how this solution could possibly play an instrumental role in reducing the spread of the virus while maintaining Hong Kong's operations. The methodologies, limitations, and current results have been discussed. Based on this information, further research and development will be conducted.

II

Acknowledgement

We would like to express our deepest and sincerest gratitude to our supervisor Dr. Reynold Cheng, Professor, Department of Computer Science, for guiding us and giving us the opportunity to work with him. His feedback, and presence through all the stages of the project has guided us towards the right path. Furthermore, we would like to extend our thanks to the Department of Computer Science, The University of Hong Kong, for equipping us with the required skills and resources.

Contents

List of Figures

List of Tables

Abbreviations

MTRC Mass Transit Railway Corporation

MTR Mass Transit Railway

COVID-19 Coronavirus Disease 2019

SAR Special Administrative Region

API Application Program Interface

ESRI Environmental Systems Research Institute

SQL Structured Query Language

1 Introduction

1.1 Coronavirus Disease 2019

The global pandemic of Coronavirus Disease 2019 (COVID-19) has had an immense impact on lives all over the globe. While close physical vicinity is the major means of contact for the virus, it can easily spread through the air (via small respiratory droplets and aerosols) and through contaminated surfaces [1].

From Figure 1.1.1 [2], it can be inferred that despite being successful at controlling the virus in the past, Hong Kong saw successive waves of COVID-19 cases from July to August and November to February. Thus it is important to be prepared with contingencies and an in-depth knowledge of factors that may be the cause of another wave in the future.

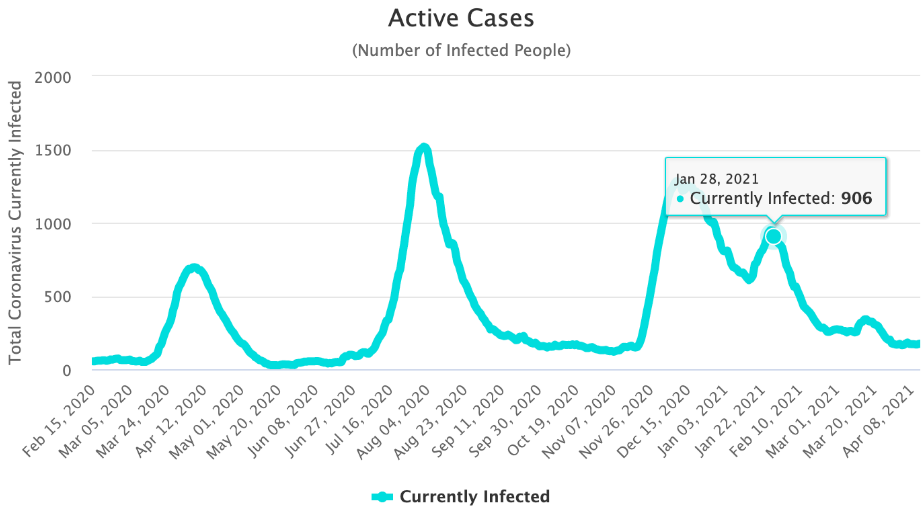


Figure 1.1.1 Graph representing Active COVID-19 cases in Hong Kong [7]

1.2 Mass Transit Railway Corporation

The Mass Transit Railway Corporation Limited (MTRC), in its 40 years of operation, has laid down a 262.6 km long railway network all over Hong Kong that covers all 18 districts in Hong Kong, carrying around 12 million people on an average weekday. In 2019, the MTR accounted for 47.9% of the total public transport boardings in Hong Kong [6]. As forecasted in Figure 1.2.1, it is highly likely that the MTR will continue to carry the majority of daily public transport passengers in 2021.

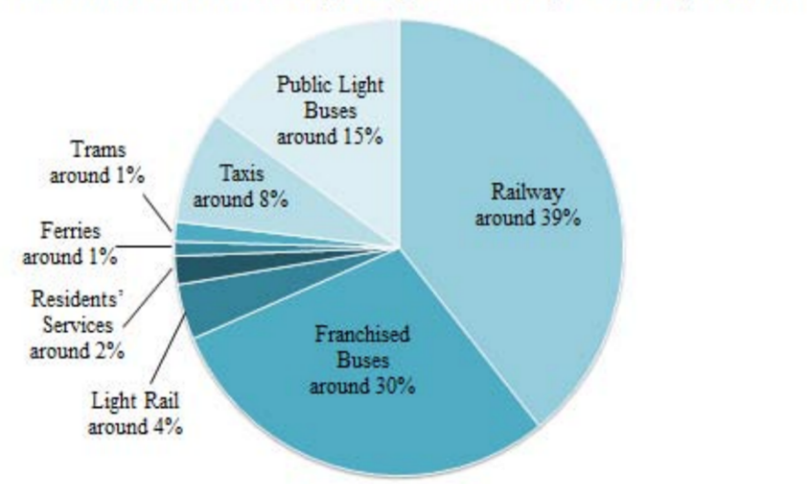


Figure 1.2.1: Forecasted distribution of public transport usage in 2021 [6]

From the abovementioned means of contact, it can be inferred that there would be a higher risk of virus transmission in crowded areas. Public areas that see a large volume of daily visitors may be considered as so-called "Hotspots" for the spread of the virus. Therefore, due to the high volume of passengers and lack of social distancing, the MTR may be deemed as a major COVID-19 Hotspot. This project, accordingly, aims to study the correlation between the development of the COVID-19 pandemic and MTR passenger travelling patterns in Hong Kong.

1.3 Goals and Objectives

The main objectives of this project are to collaborate with the MTRC, conduct in-depth research with the available data, and accordingly visualize and analyze the mutual correlation between the MTR-passenger travel patterns and the COVID-19 spread in Hong Kong. Additionally, the results of the research, analysis, and visualizations would be used to develop a single unified dynamic platform for a defined class of users, i.e. authorized HKU researchers and professors who have signed the confidentiality contract provided by the MTRC. Accordingly, the platform would allow these users to access useful analytical information and create visualizations. The project aims to utilize potential knowledge to enable the reduction of the risk of further spread of COVID-19 through the MTR mode of transport.

The initial goals included the pre-processing of MTRC-provided data and publicly available COVID-19 data, and creating relational database repositories to store the cleaned and formatted data. This is where the concept of Big Data would come in. The millions of transactions of MTR octopus card data and COVID-19 HK Government data were used to visualize, analyze, and accordingly extract useful insights from the trends.

The final deliverable of the project is aimed to be an efficient, dynamic, and widely scalable software. This software would enable users to view comprehensible data visualizations through geospatial and graph analysis on-demand. Apart from our insights and visualizations, the platform would aid other authorized researchers by providing them access to the two data repositories. Hence, the dynamic nature of the platform aims to fuel further future research into the field.

1.4 Project Contributions

Before the inception of this project, researchers and students utilized the MTRC-provided data and initial COVID-19 data to generate visualizations. However, failure to establish a good representation of the relationship between the two entities hindered their path to concrete conclusions. This failure might have risen from the previous lack of extensive COVID-19 data since the visualizations were created in the early stages of the pandemic. Additionally, the visualizations were previously based on traditional and inefficient tools such as MS Excel, which made them unsustainable in the long run.

Apart from the use of inefficient tools, the tedious means of manual extraction of data from DVDs (provided to one researcher and passed on to others) proves to be another source of hindrance to the efficiency of workflow and collaboration. The project's outcome would mitigate this hindrance by allowing authorized researchers to remotely connect, filter, access, and extract relevant data as per their requirements.

Thus, this project builds on the knowledge gained from prior work, improves on the methodologies, and adapts to more efficient technologies such as ESRI and Python.

1.5 Scope

To achieve the aforementioned goals in section 1.2, the project focuses on a three-part scope: (i) development of a structured SQL relational database for secure storage and maintenance of COVID-19 and confidential MTRC data, (ii) utilization of efficient and dynamic technologies such as ESRI and Python for trend analysis, and (iii) development of a visualization and analysis tool in the form of a web application for easy accessibility and utility for the researchers. The project aims to deliver a good big data solution by combining a data stream engine conducting fast complex analysis with a simple user interface for the supply of easy-to-understand information to the class of end-users mentioned in section 1.2.

1.6 Significance and Impact

Narrowing down the focus to Hong Kong SAR, the urgency of alleviating the effects of the pandemic calls for an in-depth analysis of the COVID-19 spread and its relationship with one of the most-used means of transportation, The Mass Transit Railway (MTR). The MTR is a crucial mode of public transportation in Hong Kong and may be considered a so-called “Hotspot” for the coronavirus spread. Therefore, studying the relationship between the MTR passenger traveling patterns and the spread of the virus could help in reducing the spread of the virus by increasing information available for researchers and the public. As a result, this may lead to better-suited transport policies that reduce the potential spread of the coronavirus while maintaining the accessibility and convenience of the MTR.

1.7 Outline

This report is organized into 8 chapters. The previous (first) chapter presented an overview of the COVID-19 pandemic and MTRC travel behavior relationship and introduced how technologies can be used to study that relationship. The second chapter will provide a literature review for theories and concepts relevant to the scope of the project. The third chapter will provide more focus on the methods that will be involved in the project to achieve the goals mentioned in chapter one (Section 1.3). It also justifies the relevance of using the specific methodologies for the benefit of the project. Chapter four presents the progress of the project and discusses all results found during the project, while chapter five showcases the project schedule in a tabular form. Chapter six discusses all limitations and challenges encountered during the final stages. Chapter eight mentioned the possible future works that can be potentially implemented as a part of the project and finally, chapter eight concludes the report by giving a summary of the solution to the specified problem.

2 Literature Review

When it comes to studying the pattern of contact between passengers within a public mode of transport, as some research papers suggest, the concept of Familiar Strangers comes into play. Familiar Strangers are urban visitors and residents who, despite being mutually unacquainted, encounter each other at particular locations during their daily routines [9]. The figure below depicts the co-presences of metro riders in Beijing during the weekdays, portraying a higher concentration of familiar strangers in certain lines due to similar routes during peak work hours.

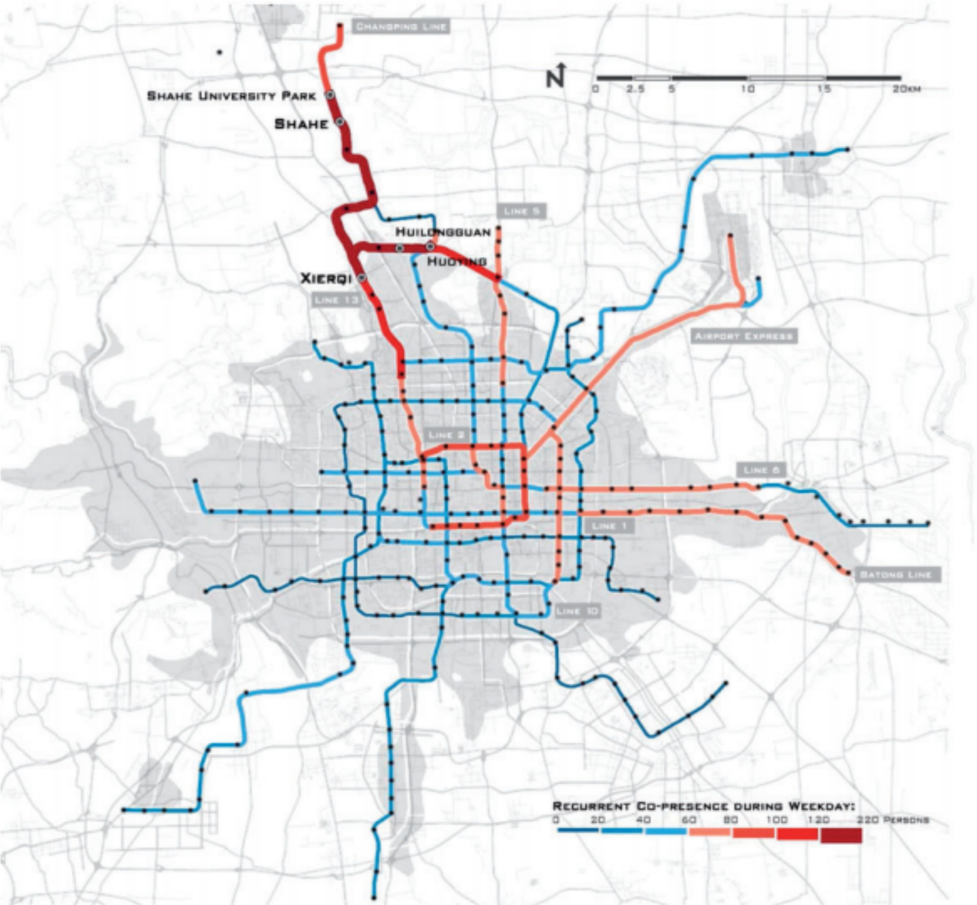


Figure 2.1: Co-presences of Metro Riders on Weekdays in Beijing [10]

From a public health standpoint, according to Zhou et al. (2017), a familiar stranger might play a role in affecting the spread of infectious diseases within a population. Bearing in mind the prevalence of consistent and routine travel patterns among MTR passengers within the Hong Kong population, the passengers might unknowingly contract the virus from an infected familiar stranger. For instance, a simple and most likely case of 'familiar strangers' would be two employees working in the area, with similar work hours, who stick to the same routes, board the same trains, and as a result, may remain in proximity for a considerable amount of time. Thus, the formation of ‘familiar strangers’ might act as a catalyst contributing to the spread of the coronavirus.

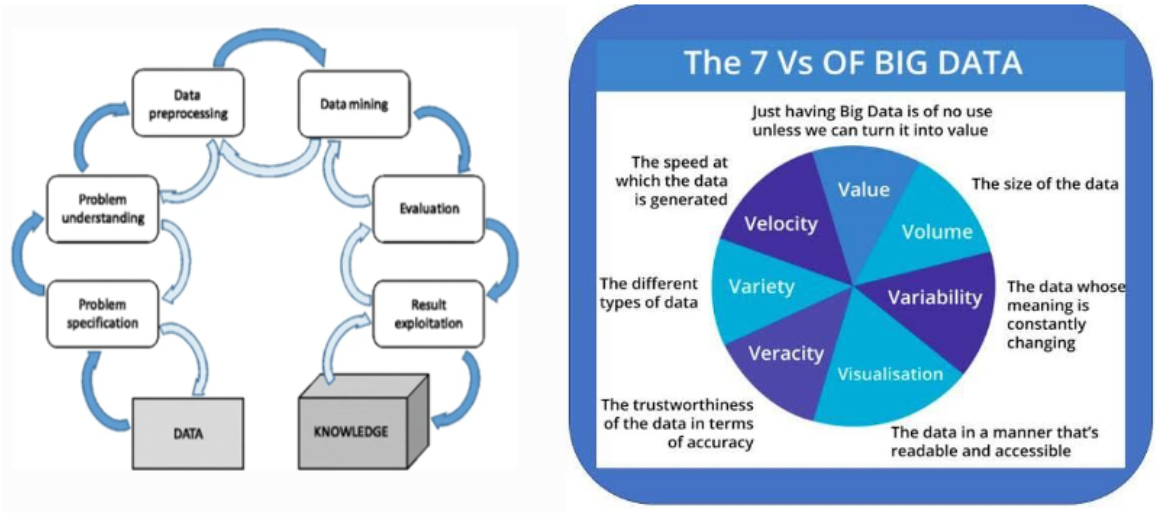
While the phenomenon of ‘familiar strangers’ has proved to be significant for many decades, the thorough comprehension and identification of familiar strangers bring about various challenges: (i) the understanding of a set of defining metrics required to identify familiar strangers is far from concrete; (ii) on identification, determining the consequential implications of familiar strangers requires a multi-dimensional perspective [7].

3 Methodology

This chapter discusses the methodology, technologies, and technical concepts that will be adopted into the project. These include details on the big data approach, database development, mobility trend analysis, and geo-spatial analysis. Each section justifies the technologies used along with some of the expected results.

3.1 Big Data Approach

The big data approach has been adopted for this project. The term big data encompasses but is not limited to data analysis, information extraction, or management of diverse collections of data with immensely complex, large, and varied structures to pinpoint trends and characteristics that cannot be found through traditional data management techniques. As portrayed in Figure 3.1.1, the process involved in the Big Data approach covers all stages - ranging from the detailed understanding of the problem at hand to the effective utilization of data [3].



Data preprocessing
Data mining
Problem understanding
Evaluation
Problem specification
Result exploitation
DATA
KNOWLEDGE
The 7 Vs OF BIG DATA
Just having Big Data is of no use unless we can turn it into value
The speed at which the data is generated
The size of the data
Value
Velocity
Volume
The different types of data
Variety
Variability
The data whose meaning is constantly changing
Veracity
Visualisation
The trustworthiness of the data in terms of accuracy
The data in a manner that's readable and accessible

Figure 3.1.1: Data handling process in the Big Data Approach

Looking at the ‘familiar stranger’ phenomenon from a big data perspective has led to a more technical approach for uncovering significant results. Previously, familiar strangers were identified through a tedious process of mass surveys and personal anecdotes. However, analysis conducted using a big data approach maps out a more accurate and extensive picture of familiar strangers within the Hong Kong MTR network.

3.2 Database

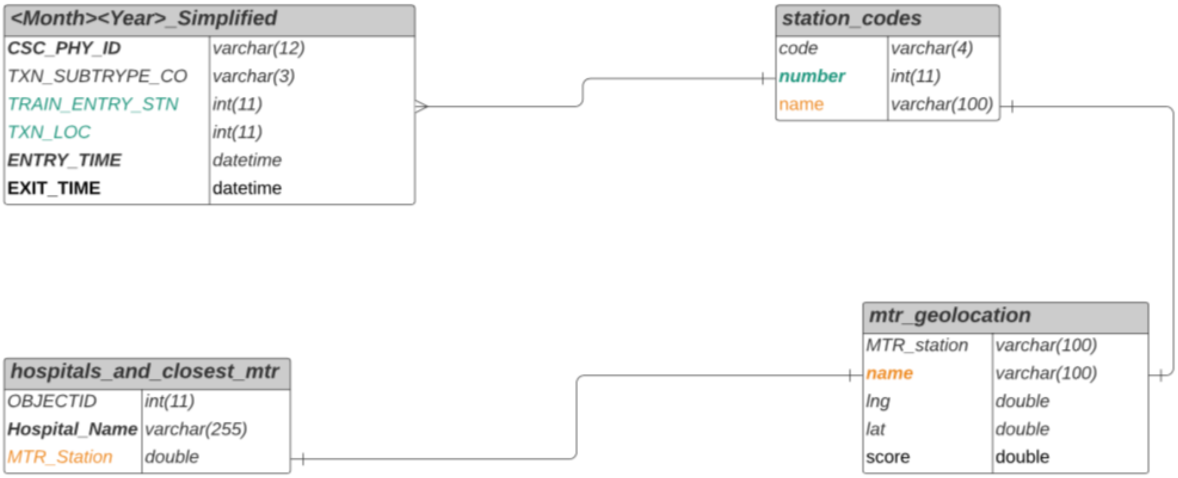
The database development process refers to the identification, specification, and creation of the schema that describes the organization of data in a table. The project will include a relational database storing MTRC-provided passenger data and the COVID-19 data provided by the Hong Kong Centre for Health Protection. This is a crucial stage in the development life-cycle as it would enable seamless and efficient data retrieval.

3.2.1 Preprocessing

The seamless and efficient usage of data in later stages of development is highly dependent on the initial stage of data pre-processing. Data Preprocessing refers to the conversion of raw data from diverse sources into relevant information [8]. The presence of clean data that contains strictly relevant information is instrumental for the success of Big Data. The team has successfully collaborated with the MTRC and accumulated around 32 million transactions of octopus data from the periods of January 2019 to August 2019 and January 2020 to August 2020. This data, in addition to the government-provided COVID-19-cases data repository, proves to be immensely vast and consequently accentuates the need to focus more on data pre-processing.

Overall, in the Database Development phase, the team aims to create relational database tables with concrete relationships and are devoid of redundancies. Thus, there would be a need for extensive standardization of data into consistent formats.

Third-party APIs, such as Google's Geocoding API, would be useful for adding some useful information and removing redundant attributes. Thereafter, the tables will be exported into the SQL server.



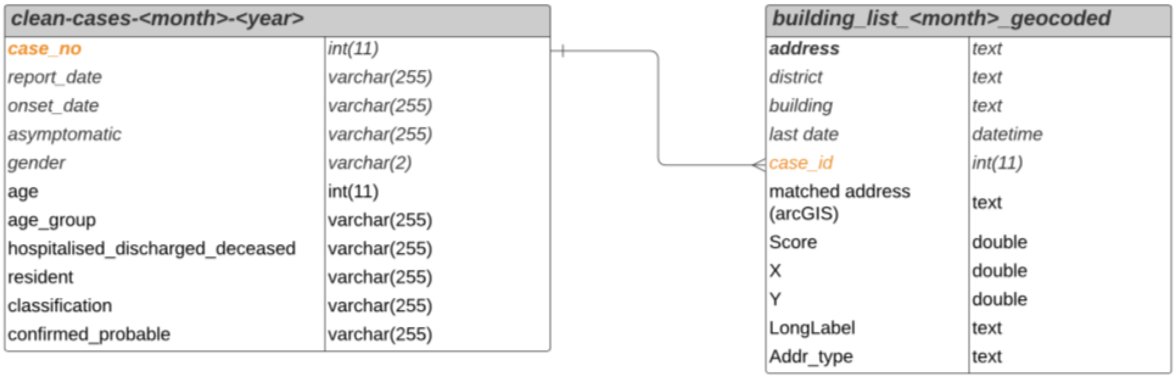
<Month><Year>_Simplified	
CSC_PHY_ID	varchar(12)
TXN_SUBTRYPE_CO	varchar(3)
TRAIN_ENTRY_STN	int(11)
TXN_LOC	int(11)
ENTRY_TIME	datetime
EXIT_TIME	datetime

station_codes	
code	varchar(4)
number	int(11)
name	varchar(100)

hospitals_and_closest_mtr	
OBJECTID	int(11)
Hospital_Name	varchar(255)
MTR_Station	double

mtr_geolocation	
MTR_station	varchar(100)
name	varchar(100)
lng	double
lat	double
score	double

Figure 3.2.1.1: Entity Relationship Diagram for MTRC octopus transaction data



clean-cases-<month>-<year>	
case_no	int(11)
report_date	varchar(255)
onset_date	varchar(255)
asymptomatic	varchar(255)
gender	varchar(2)
age	int(11)
age_group	varchar(255)
hospitalised_discharged_deceased	varchar(255)
resident	varchar(255)
classification	varchar(255)
confirmed_probable	varchar(255)

building_list_<month>_geocoded	
address	text
district	text
building	text
last date	datetime
case_id	int(11)
matched address (arcGIS)	text
Score	double
X	double
Y	double
LongLabel	text
Addr_type	text

Figure 3.2.1.2: Entity Relationship Diagram for Govt. COVID-19 data

Throughout the multiple cycles of development, testing, and requirement updates, the data repository underwent multiple alterations in order to create a database catering to the efficient retrieval requirements of the project. The diagrams above represent the schemas of the finalized entity tables in the relational database along with the concrete relationships between the entities.

3.2.2 Secure Shell (SSH) Tunnelling

As mentioned in section 1.3, the data procured from the MTRC is accompanied by a confidentiality agreement. Thus, the sensitivity of the data calls for its storage on a secure server. To prevent the data from being compromised, the SQL database would be deployed on the HINcare Server on top of the private HKU CS server. Accordingly, the system employs Secure Shell (SSH) protocol for establishing secure tunnels to the data source

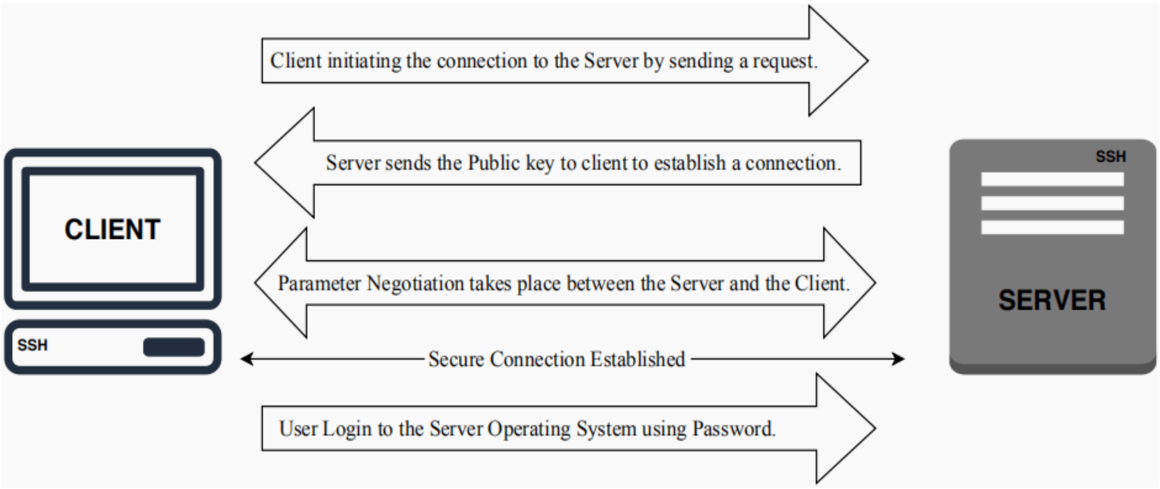

Client initiating the connection to the Server by sending a request.
Server sends the Public key to client to establish a connection.
Parameter Negotiation takes place between the Server and the Client.
Secure Connection Established
User Login to the Server Operating System using Password.
CLIENT
SSH
SSH
SERVER

Figure 3.2.2.1: Secure Shell (SSH) Protocol [5]

The server running the project's platform (the client) connects to the secure database via SSH channels opened only in response to the authentication of the admin credentials (Figure 3.2.2.1). Since the data is stored on the HINCare server on top of the CS server, the data security is enforced by using a multi-hop SSH tunnel to establish a connection to the database: (i) first, the connection to the HKU CS server is established using 'cs.hku.hk' credentials, then (ii) a connection to the database server is set up using unique HINCare server credentials.

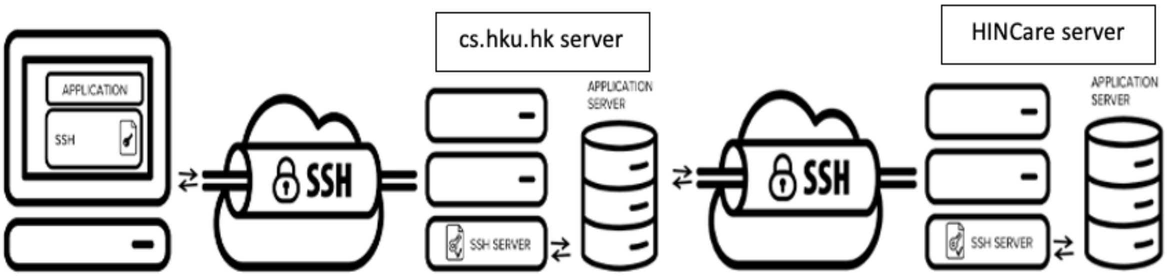


Figure 3.2.2.2: Multi-hop tunnel to access HINCare server [4]

3.3 Mobility Trend Analysis

The project will include a detailed analysis of the changes in mobility trends in Hong Kong with the spread of COVID-19. The passenger travel data provided by the MTRC portrays transportation trends in pre-COVID-19 Hong Kong (2019) and COVID-19-affected Hong Kong (2020). Thus, creating graphic visualization for monthly trends would enable the team to examine the influence of MTR ridership behavior on the fluctuating severity of the pandemic and vice versa.

In addition to the monthly reports, the analysis can be narrowed down to daily passenger travel patterns, wherein there would be an increased focus on certain key events - such as the Wuhan Lockdown, the initial outbreak, and the onset of the second, third, and fourth waves of widespread COVID-19 cases – and key dates – Chinese New Year holidays, Work-From-Home policy periods. The daily passenger travel data, with the entry station and exit station information, also enables us to formulate the most efficient travel routes. Subsequently, this formulation makes it possible to approximate the number of **familiar strangers**.

Sociodemographic factors can be taken into account while exploring the mobility trend in further detail. The MTRC-provided passenger travel data categorizes the octopus users based on their age group (Octopus type: Infant, Student, Adult, and Elderly). Thus, the mobility trend can be inspected independently for different age groups. As mentioned before, the elderly population is more susceptible to severe effects of COVID-19. Thus, analyzing based on age would be beneficial in an attempt to protect the at-risk demographic.

The project will also allow mobility analysis through the study of certain metrics – Passenger Mobility, Station Density, and Travel Pattern. Since the raw data procured from the MTRC merely consists of octopus transactions depicting entry and exit time along with entry and exit stations, filtering and aggregation (i.e. COUNT) using SQL queries would be required to retrieve relevant data that depicts the volumes of passengers present in a particular station or taking a particular train route, within a given time frame.

The abovementioned analyses will pave the way for a better understanding of human (passenger) behavior and how it has been affected by the pandemic. However, while the analyses will be beneficial in supporting our hypothesis, merely observing the trends might not be sufficient for the analytics encompassed by the project. Thus, a study of the data based on geolocation would be of great importance.

3.4 Geo-spatial Analysis using ArcGIS (ESRI)

Certain factors can be better observed during the data analysis stage through geospatial visualization. These include the density of passengers visiting each station and the travel density between entry and exit stations in the Hong Kong MTR network. Such visualization would indicate the level of crowding, not only in MTR stations but also within the moving trains. The geospatial mapping of COVID-19 case density may pinpoint the COVID-19 "hotspots" all over Hong Kong.

The project's aim is to combine the geo-spatial representation of MTR passenger data and COVID-19 case data with the purpose of confirming the hypothesized relationship between the two. In order to visualize the vast amount of data in a geospatial space, there is a need for geocoding addresses into longitude and latitude coordinates.

ESRI's Geometric Information System software, called ArcGIS, is widely used for the storage, analysis, and visualization of geometric information, and is capable of handling a large volume of complex and dynamic MTR and COVID-19 data. ArcGIS also provides some APIs that are capable of geo-coding addresses as required.

3.5 Contact and Behaviour Based Research

After the procurement, processing, and visualization of the MTR data, Contact and Behavior-based research will come into play to realize the underlying patterns of MTR passengers' travel behavior. This stage would be highly crucial in the analysis part of our platform since it would give insights into how the virus might spread within the MTR setting. The platform would allow the user to conduct two different types of contact and behavior research: (i) Someone like you and (ii) sensor individuals. The data repository resulted from the processes in section 3.2 would be used to conduct this research. Moreover, python libraries such as pandas and NumPy would be used for the analysis.

3.5.1 Someone Like You

The term “someone like you” refers to a pair of individual riders who exhibit co-presence for at least one trip in a day. In other words, two riders who follow the same trajectories such that they enter the same MTR station and exit the same MTR station within the same period would be considered as “someone like you” for each other. Thus, the presence of “someone like you” could be the key to identifying points of contact for the COVID-19 virus. As a result, the project would include this research to identify MTR routes with the maximum number of “someone like you” pairs and attempt to lay out the resulting spatiotemporal pattern.

To conduct the abovementioned “someone like you” research, certain operations are done on the MTRC octopus transaction data. All stations are uniquely paired with each other and each day is divided into 48 time periods (each with a length of 30 minutes). Thus, all trips throughout a day are grouped based on station pairs and periods. Accordingly, the number of trips is accumulated for each day. Finally, the resulting numbers would be averaged for weekdays and weekends separately, since it has been observed that the two have vastly different patterns (weekdays more organized while weekends more spread and random). The resulting data would be a CSV file depicting the number of mutual “someone like you” passengers for each station pair separately for weekdays and weekends. Moreover, visualization of spatiotemporal patterns of “someone like you” would be created.

3.5.2 Sensor Individuals

Sensor Individuals refer to passengers who potentially have come into the most physical contact with other passengers at the station level. From a health perspective, the identification of Sensor Individuals can lead to the detection of a COVID-19 “super spreader”. By conducting research and analysis on this phenomenon, the project aims to map out the spatiotemporal pattern of the MTR passengers’ co-presence, and accordingly attempt to pinpoint the existence of potential super spreaders within the MTR network.

The identification of sensor individuals would require some hefty operations on the pre-existing database. The existing database only consists of entry and exit stations, and not the stations encountered by a passenger in transit. Thus, it would not be possible to identify contact between passengers sharing a partial path (or a sub-path) within their transit routes. Hence, a single journey would be broken down into sub-journeys including each MTR station encountered within a journey, along with timestamps. The result would exhibit the details of the travel journey of each passenger in a CSV format.

3.6 Web Application

The results from the abovementioned methodologies would be made available to the target class of users (as mentioned in section 1.3) through a platform developed in the form of a **web application**. The user-friendly design of the web application will fulfill the three technical requirements mentioned below.

Firstly, the web application will provide users with a means to query both the MTR and COVID-19 databases in a simple manner. With the database and SSH tunnels in place (as mentioned in section 3.2.2), users will have the ability to fill in parameters (period, location, etc.) for efficient data retrieval. Additionally, for research purposes, the platform would allow users access to some metrics (travel pattern, station density, or passenger mobility) that potentially illustrate underlying behavioral patterns of MTR passengers. The abovementioned data requests would retrieve clean data formatted into CSV files, allowing users to utilize the data for their research purposes.

Secondly, results of the geospatial analysis (mentioned in section 3.4) would be exhibited through the platform's visualization feature, allowing users to generate trend maps with an optional time-period filter. The **ArcGIS JavaScript API** will be utilized to render a full-scale map component on the platform, allowing users to visualize and traverse through the markings made as per the MTR transit patterns and COVID-19 reported-case locations.

Finally, the analysis section of the platform would provide users, mainly researchers, access to contact and behavior-based research. This feature would play the crucial role of serving data portraying “someone like you” and “sensor individuals” analyses, and thus, pave the way to further insights and collaboration with other researchers.

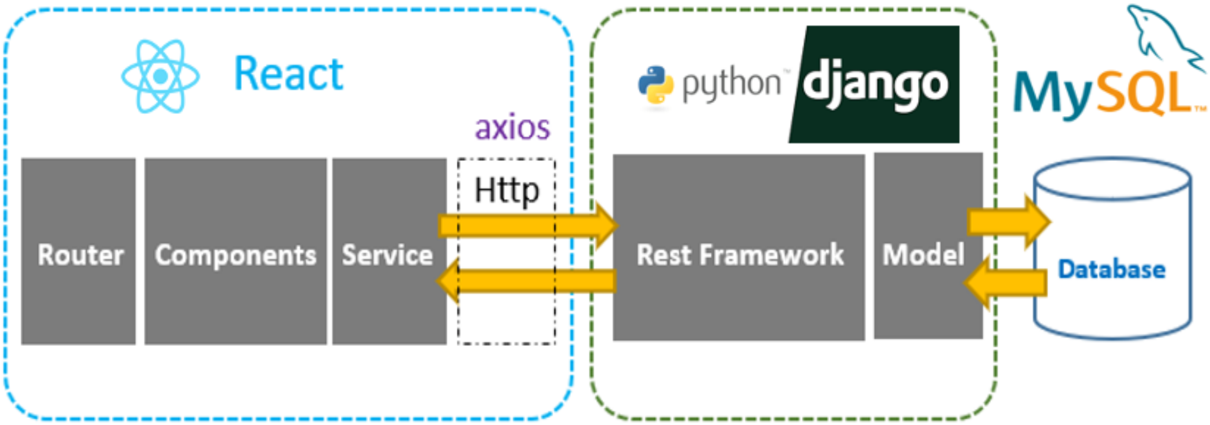


Figure 3.6.1: Web Application Technology Stack and Architecture

3.7 Summary

The abovementioned methodologies are crucial to the progress of the project and are highly dependent on each other. The database development stage acts as the backbone of all other methodologies since it produces clean data to be used in all stages. The geo-spatial analysis, mobility trend analysis and Contact and Behaviour research go hand in hand as they facilitate data interpretation and give actual meaning to the random numbers we get from the databases. Overall, all the outputs from all stages are exploited in the Big Data approach to devise crucial analytical knowledge (last stage in Figure 3.1.1) which would be utilized for further research and made available to the end-users in the form of a web application.

4 Results and Discussion

This chapter reports the project's preliminary findings, discusses the project's schedule and gives a description of the final deliverables. It then moves on to point out the immediate and long-term plans that would be fuelled from the current outputs.

4.1 Preliminary Results

Amidst the course of the fluctuations in the severity of the COVID-19 pandemic, Hong Kong encountered four major waves of rising infections. The preliminary research included in-depth **mobility and geo-spatial analysis** of the first two waves of COVID-19 cases and corresponding MTR ridership within Hong Kong. The main objective of this stage was to support the main hypothesis - the presence of a strong correlation between the MTR ridership behavior and pandemic's severity,

The onset of the first wave of COVID-19 commenced with the first confirmed case on 23rd January 2020 and continued on until the end of March.

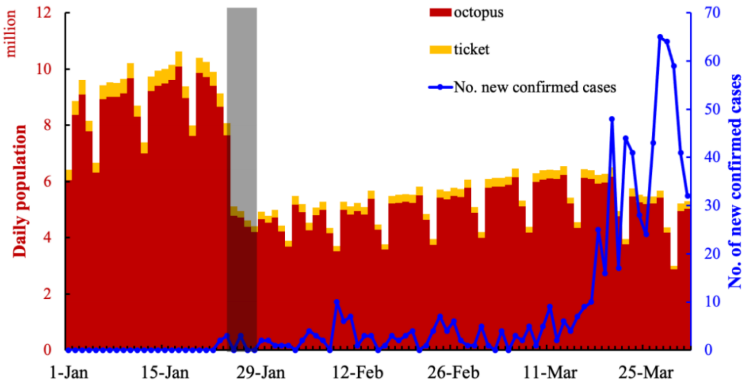


Figure 4.1.1: MTR ridership during the first COVID-19 wave

The onset of the first wave of COVID-19 commenced with the first confirmed case on 23rd January 2020 and continued until the end of March. A careful study of MTR ridership trends and COVID-19 cases pinpointed a significant decrease of about 44% in MTR usage within Hong Kong, exhibited by the grey marked area in figure 4.1.1. It is noteworthy that this occurrence was observed despite the ongoing Chinese New Year Holidays - a time of peak travel and exploration. This indicated a rise in the population's awareness of the COVID-19 pandemic.

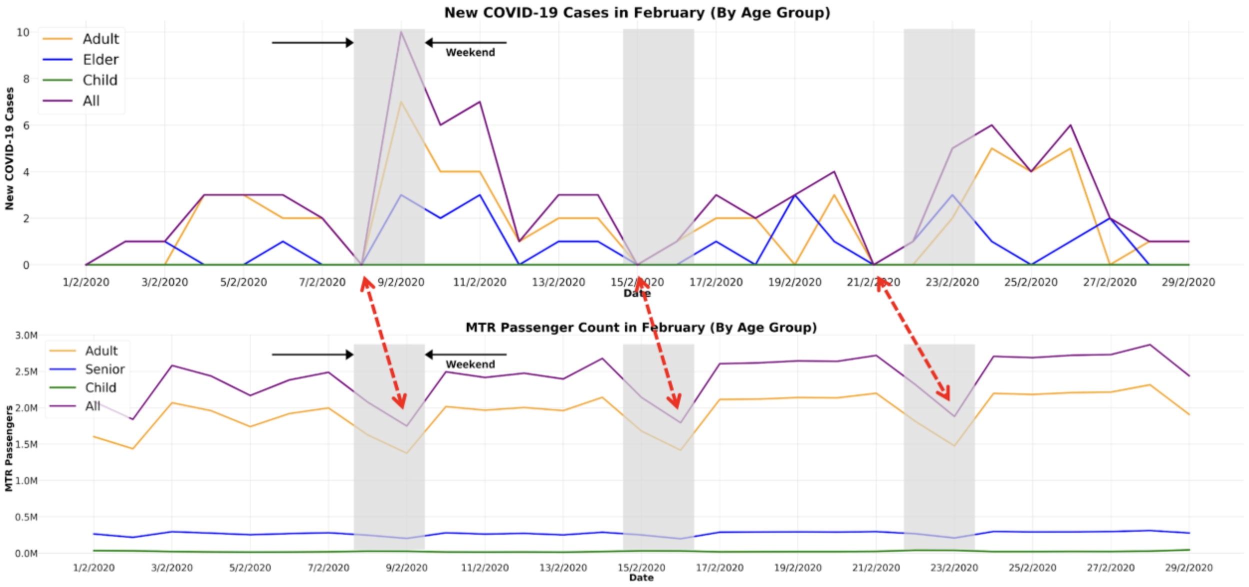


Figure 4.1.2: New COVID-19 cases & MTR Passenger Density vs Date (February 2020)

Figure 4.1.2 displays the results of the Mobility Trend Analysis (section 3.3). It is observed that during February 2020 the MTR passenger density and the number of new COVID-19 cases simultaneously exhibited a general downward trend during the weekends in comparison to the fluctuating trend during the weekdays. The similar trend of the two metrics during similar periods proves to be a good demonstration of the correlation between them, and further supports the hypothesis.

Subsequently, the preliminary stage also included a Geo-Spatial Trend analysis conducted during the second wave of COVID-19 infections in Hong Kong (from March to April 2020).

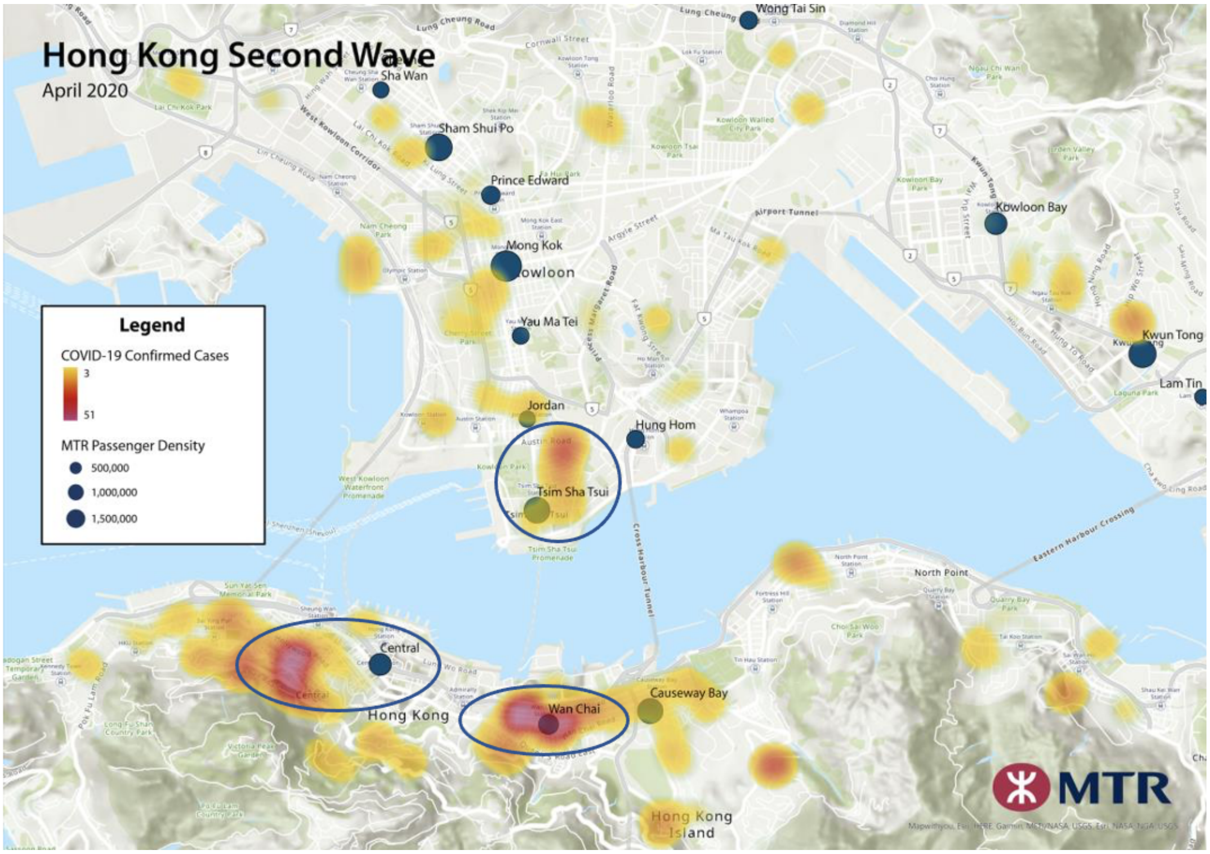


Figure 4.1.3: MTR Station Density and COVID-19 Hotspots (April, 2020)

Figure 4.1.3 comprises the passenger density of each MTR station (blue circles) along with the heatmap of the COVID-19 cases all over Hong Kong (yellow-ish/red spots). The most prominent discovery from the map can be observed around the regions of Wan Chai, Central, and Tsim Sha Tsui. Being among the MTR stations with large passenger density, it is no surprise that these stations come into proximity with some of the major coronavirus hotspots, observed as nearby red spots.

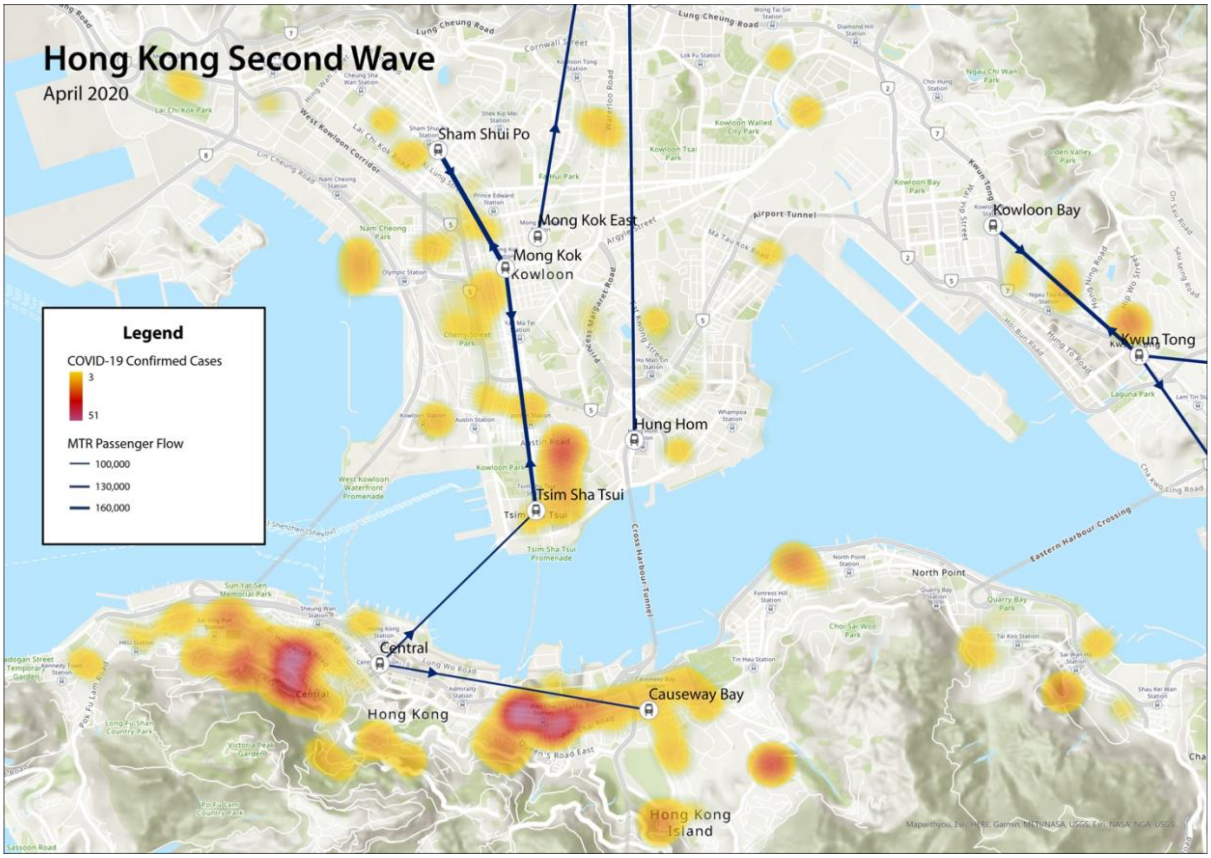


Figure 4.1.4: MTR Travel Pattern and COVID-19 Hotspots (April, 2020)

Figure 4.1.4 exhibits the MTR passenger travel pattern (blue lines) along with the heatmap of the COVID-19 cases all over Hong Kong (yellow-ish/red spots) during April 2020. The thickness of the blue lines in the map generated represents the trip frequency along each route. The figure shows a trend portraying the spread of confirmed cases across some paths. This effect can be most observed along the paths of Central - Causeway Bay, Sham Shui Po – Mong Kok, Mong Kok - Tsim Sha Tsui, and Kowloon Bay – Kwun Tong.

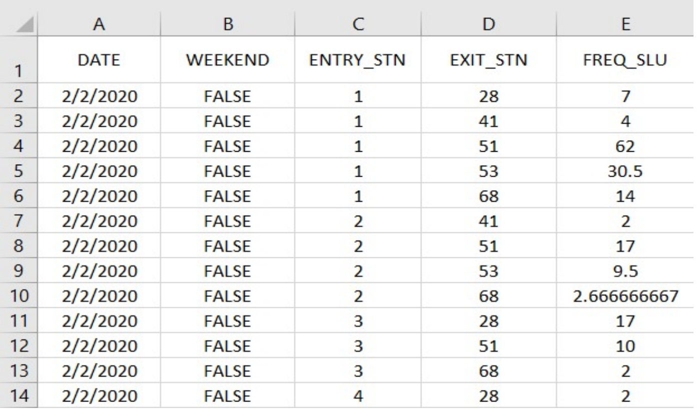
The mobility trend and geo-spatial trend analysis conducted during the preliminary stages successfully supported our hypothesis by bringing further focus towards the influence of MTR passenger travel patterns on the coronavirus spread, and vice versa.

4.2 Someone Like You and Sensor Individuals

After the preliminary stage, the analysis focus was diverted to the Contact and Behavior-based research as mentioned in section 3.5. While the previous stages of mobility and geospatial trend analysis gave results in the form of visualizations, the outcome of this stage was in the form of powerful data which could potentially enable experienced researchers to utilize complex visualization methods and operations to charge further research into the field.

Table 4.2.1 illustrates a sample result from a script prepared for the “someone like you” analysis. This script, written in python, resides in the Django backend and is utilized to process the MTRC-provided data stored on the SQL server. The input parameters to the algorithm comprise the station pairs to be filtered (optional) and the time period to be used as a window of identifying ‘someone like you’ passengers (required). The algorithm output columns are: (i) DATE (analysis date), (ii) WEEKEND (whether the date describes a weekend day - TRUE/FALSE), (iii) ENTRY_STN (passengers’ entry station), (iv) EXIT_STN (passengers' exit station), and finally (v) FREQ_SLU (number of ‘someone like you’ passengers). The last column ‘FREQ_SLU’ represents the calculated average number of ‘someone like you’ passengers along the path between the entry station and exit station on a particular date.

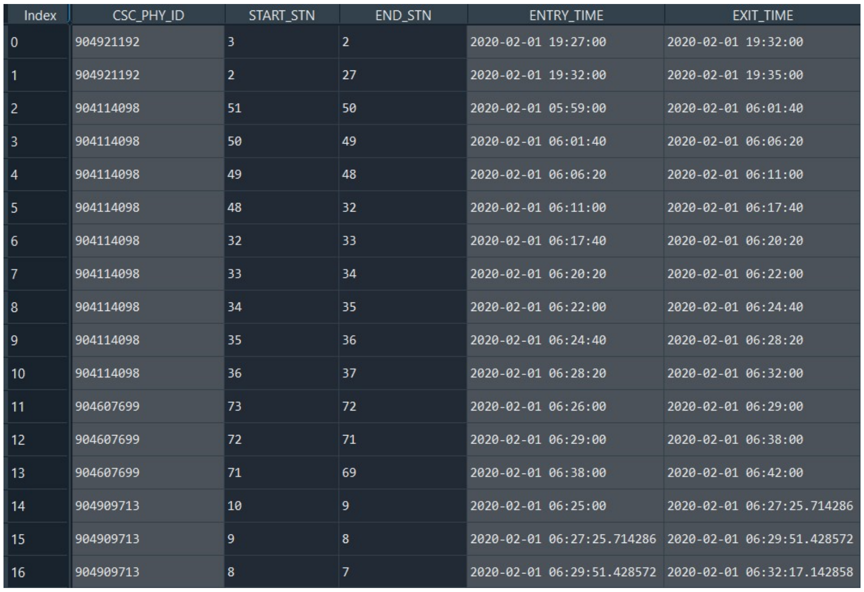
Table 4.2.1: Someone Like You Analysis sample result



	A	B	C	D	E
1	DATE	WEEKEND	ENTRY_STN	EXIT_STN	FREQ_SLU
2	2/2/2020	FALSE	1	28	7
3	2/2/2020	FALSE	1	41	4
4	2/2/2020	FALSE	1	51	62
5	2/2/2020	FALSE	1	53	30.5
6	2/2/2020	FALSE	1	68	14
7	2/2/2020	FALSE	2	41	2
8	2/2/2020	FALSE	2	51	17
9	2/2/2020	FALSE	2	53	9.5
10	2/2/2020	FALSE	2	68	2.666666667
11	2/2/2020	FALSE	3	28	17
12	2/2/2020	FALSE	3	51	10
13	2/2/2020	FALSE	3	68	2
14	2/2/2020	FALSE	4	28	2

This data (Table 4.2.1) can further be utilized to generate visualizations similar to the map depicting “Co-presences of Metro Riders on Weekdays in Beijing” (Figure 2.1) in section 2. Visualizing such data would enable easier detection of spatiotemporal patterns depicting periods and paths with the maximum number of ‘someone like you’ passengers that play a role in the spread of the virus. This query takes around 5-10 minutes if the station pair is not specified as a filter parameter.

Table 4.2.2: Sensor Individuals Analysis sample



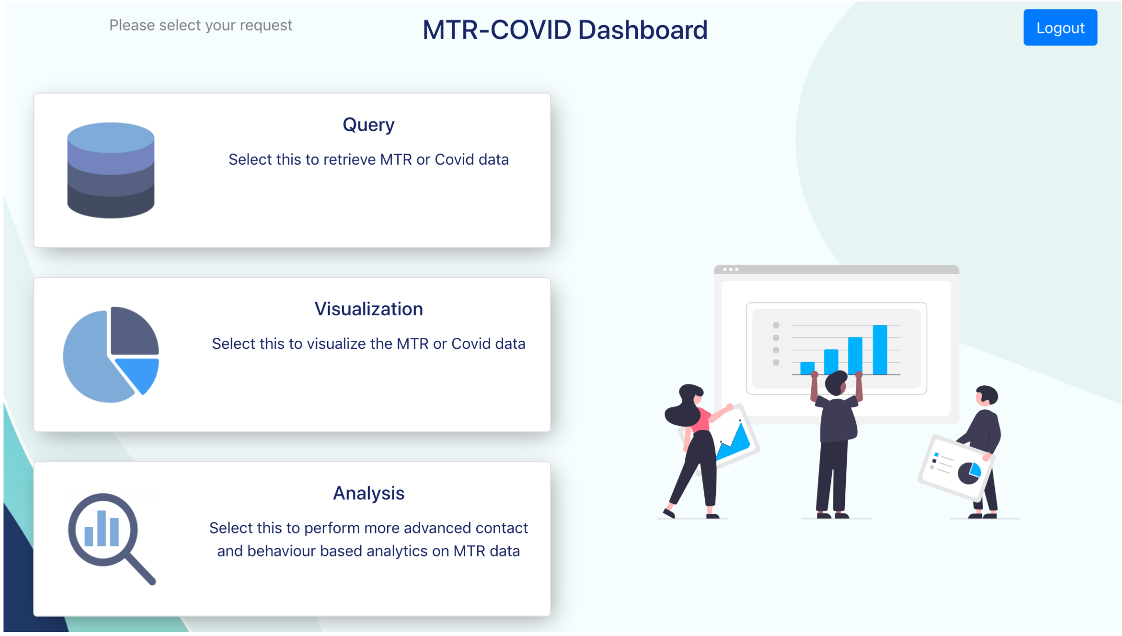
Index	CSC_PHY_ID	START_STN	END_STN	ENTRY_TIME	EXIT_TIME
0	904921192	3	2	2020-02-01 19:27:00	2020-02-01 19:32:00
1	904921192	2	27	2020-02-01 19:32:00	2020-02-01 19:35:00
2	904114098	51	50	2020-02-01 05:59:00	2020-02-01 06:01:40
3	904114098	50	49	2020-02-01 06:01:40	2020-02-01 06:06:20
4	904114098	49	48	2020-02-01 06:06:20	2020-02-01 06:11:00
5	904114098	48	32	2020-02-01 06:11:00	2020-02-01 06:17:40
6	904114098	32	33	2020-02-01 06:17:40	2020-02-01 06:20:20
7	904114098	33	34	2020-02-01 06:20:20	2020-02-01 06:22:00
8	904114098	34	35	2020-02-01 06:22:00	2020-02-01 06:24:40
9	904114098	35	36	2020-02-01 06:24:40	2020-02-01 06:28:20
10	904114098	36	37	2020-02-01 06:28:20	2020-02-01 06:32:00
11	904607699	73	72	2020-02-01 06:26:00	2020-02-01 06:29:00
12	904607699	72	71	2020-02-01 06:29:00	2020-02-01 06:38:00
13	904607699	71	69	2020-02-01 06:38:00	2020-02-01 06:42:00
14	904909713	10	9	2020-02-01 06:25:00	2020-02-01 06:27:25.714286
15	904909713	9	8	2020-02-01 06:27:25.714286	2020-02-01 06:29:51.428572
16	904909713	8	7	2020-02-01 06:29:51.428572	2020-02-01 06:32:17.142858

Table 4.2.2 above displays results from the script prepared to conduct the 'Sensor Individuals' analysis. The python script involved in this analysis proves to be more computationally heavy in comparison to the 'someone like you' script. Thus, it has a much longer running time. To shorten the running time and increase feasibility, a smaller sample dataset was used to test this algorithm. The algorithm breaks down a single user trip into multiple sub-journeys, depicting every station encountered during the trip along with the timestamp for each encounter. For instance, the first two rows, with the starting (START_STN) and ending (END_STN) station pairs (3, 2) and (2, 27) represent sub-journeys of a complete trip from station 3 to station 27 (with 2 as an intermediate station). Thus, the total number of sub-journeys depends on the number of stations encountered on the route. The result returned is in the form of a CSV file.

While the result does not identify sensor individuals in the dataset, it paves way to do so through visualisations and computations that can be generated in the future.

4.3 The Platform

This section discusses the final product of the project - the platform - which would allow end-users to access the result of all the abovementioned methodologies, research, and development. It will focus on the key features and functionalities provided by the platform.



Please select your request

MTR-COVID Dashboard

Logout

Query

Select this to retrieve MTR or Covid data

Visualization

Select this to visualize the MTR or Covid data

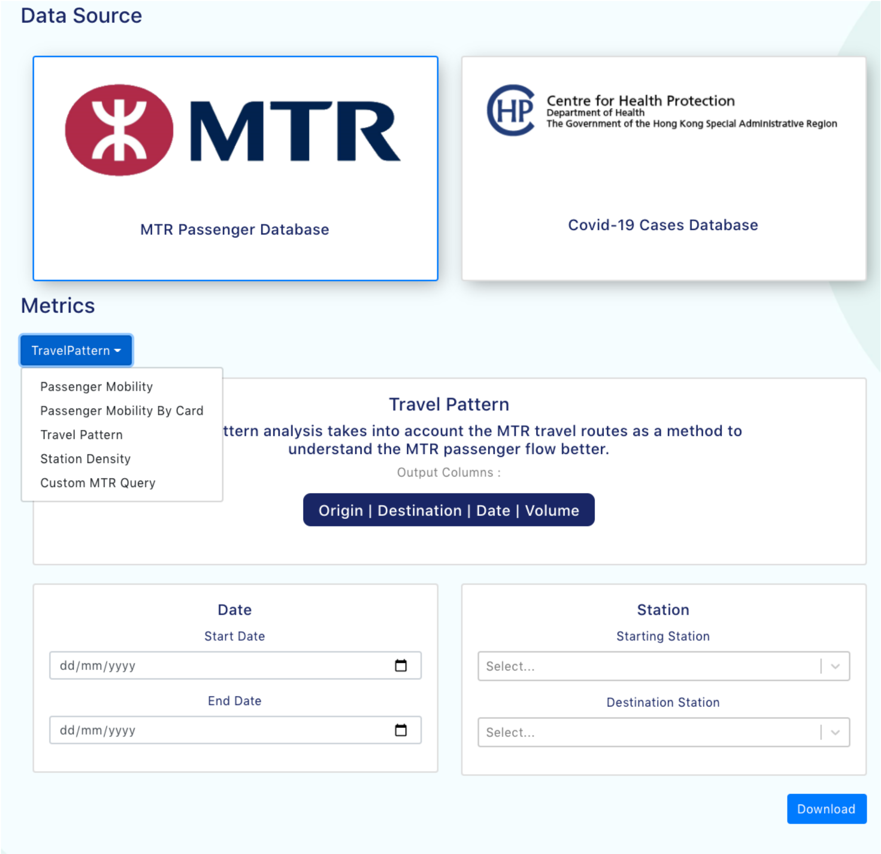
Analysis

Select this to perform more advanced contact and behaviour based analytics on MTR data

Figure 4.3.1: Functionalities of the platform

The platform provides three main functionalities: (i) query, (ii) visualisation, and (iii) analysis.

4.3.1 Query



Travel Pattern

ttern analysis takes into account the MTR travel routes as a method to understand the MTR passenger flow better.

Output Columns :

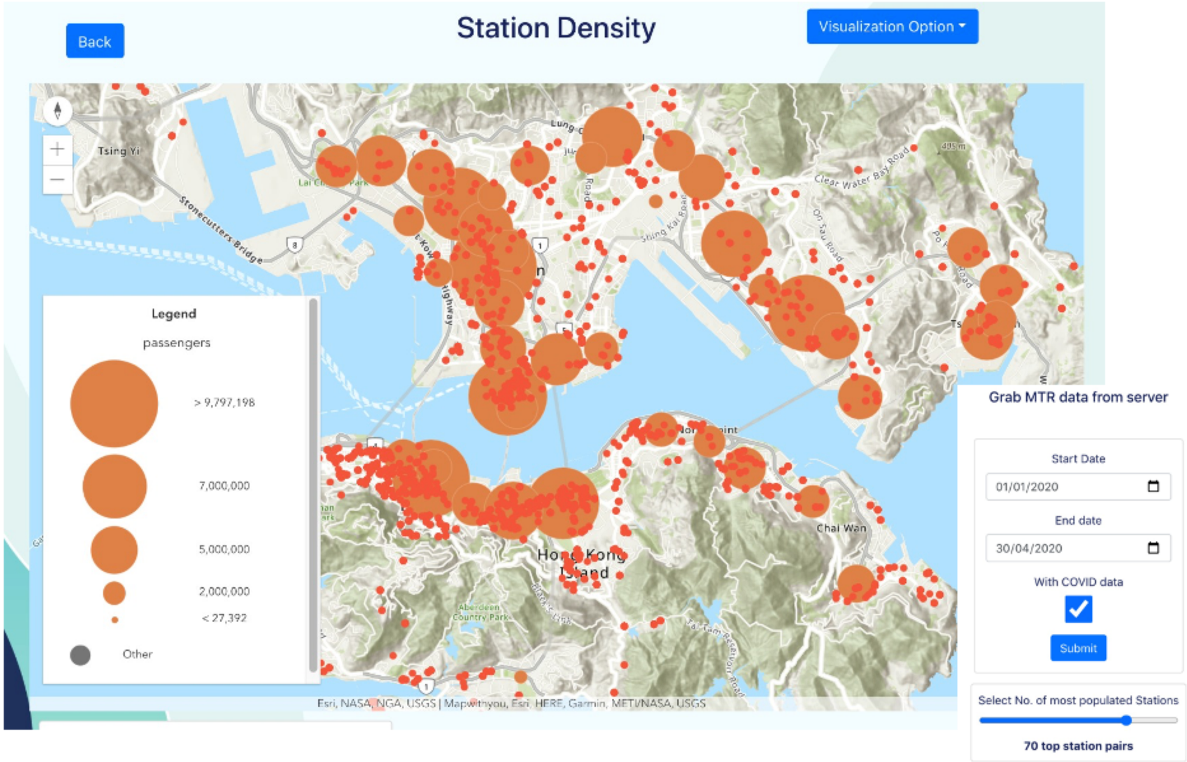
Origin | Destination | Date | Volume

Figure 4.3.1.1: Query functionality of the platform

The query feature allows users to fill in certain parameters (optional) and retrieve clean data in the form of CSV files as per their requirements. Apart from the custom MTR queries that allow original MTRC-provided data retrieval, there would also be access to certain **metrics** prepared by the team. These metrics include (i) Travel Pattern (daily/hourly MTR travel route frequency), (ii) Station Density (daily/hourly volume of passengers entering a station), (iii) Passenger Mobility (Daily volume of passengers across the MTR network). Additionally, the query feature can also be utilized to retrieve government-provided data on COVID-19 cases within Hong Kong.

4.3.2 Visualisation

This feature plays a crucial role in exhibiting dynamic visualizations that effectively support our hypothesis and possibly equip researchers for further research into the field. The interface allows the user to select a particular time period (start-date and end-date) and accordingly generate visualizations. The feature provides three types of visualization options to the user: (i) Travel Pattern, (ii) Station Density, and (iii) Passenger Volume.



Start Date
01/01/2020

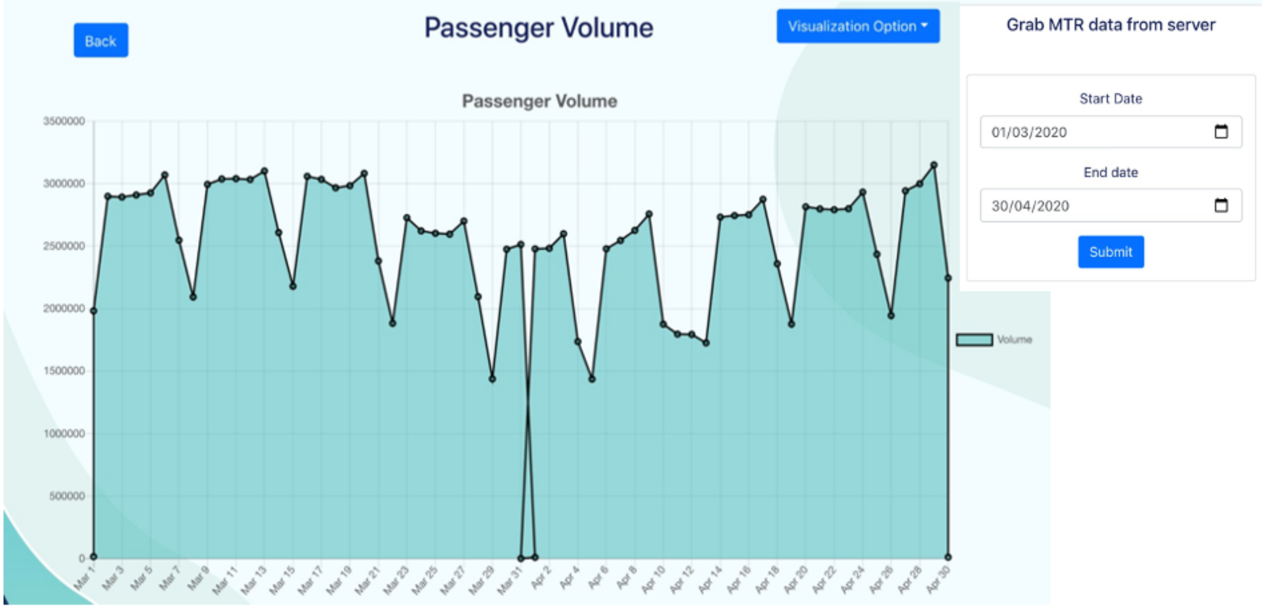
End date
30/04/2020

With COVID data ☑

Select No. of most populated Stations
70 top station pairs

Figure 4.3.2.1: Visualisation Feature - Geo-spatial Station Density (Jan - Apr 2020)

Figure 4.3.2.1 above maps out the COVID-19 case geolocations (red spots) and the MTR station density (orange circles) as per the user-entered time period filter (1st Jan 2020 to 30th Apr 2020). This visualization provides a better portrayal of the previously established direct correlation between station passenger density and the presence of COVID-19 hotspots, as mentioned in section 4.1 (Figure 4.1.3).

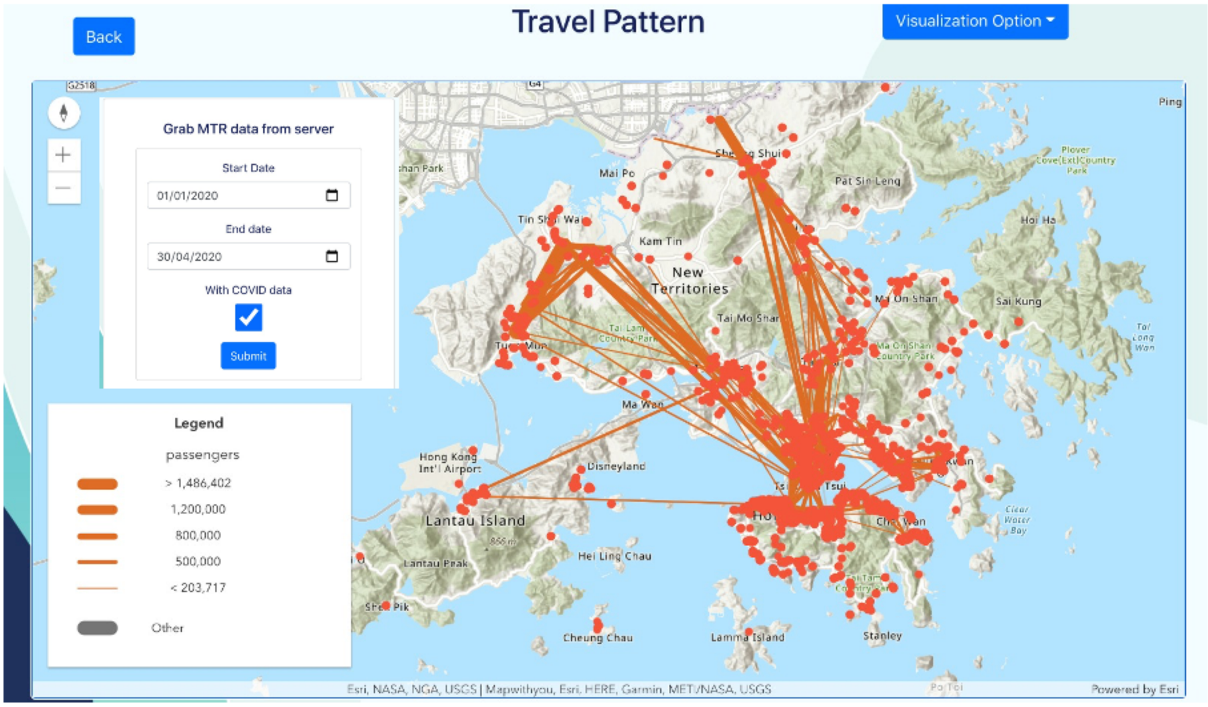


Start Date
01/03/2020

End date
30/04/2020

Figure 4.3.2.2: Visualisation Feature - Passenger Volume (Mar-Apr 2020)

Figure 4.3.2.2 above graphs out the daily passenger volume in the MTR trains as per the filter entered by the user (1st Mar 2020 to 30th Apr 2020). This graph effectively portrays the downward trend in passenger density during weekends as seen during the preliminary stage in Section 4.1 (Figure 4.1.2).



Travel Pattern

Grab MTR data from server

Start Date: 01/01/2020

End date: 30/04/2020

With COVID data ☑

Legend

passengers

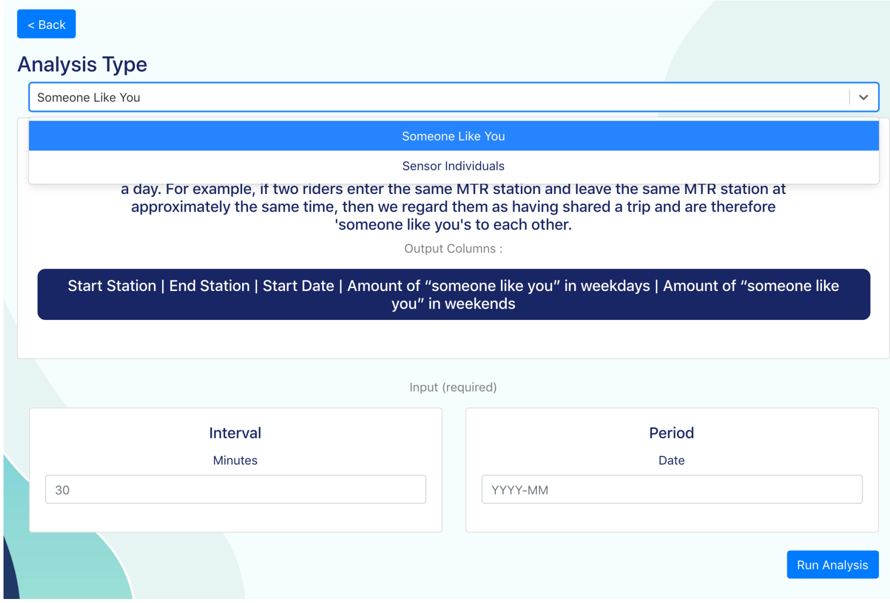
▬	> 1,486,402
▬	1,200,000
▬	800,000
▬	500,000
▬	< 203,717
▬	Other

Figure 4.3.2.3: Visualisation Feature - Geo-spatial Travel Pattern (Jan - Apr 2020)

Figure 4.3.2.3 above maps out the COVID-19 case geolocations (red spots) and the MTR travel route frequency (orange lines) as per the user-entered time period filter (1st Jan 2020 to 30th Apr 2020). This geospatial visualization exhibits the correlation between the frequently taken MTR routes and the distribution of coronavirus cases all over Hong Kong. Although previously established in Section 4.1 (Figure 4.1.4), the correlation is more effectively presented through the above visualization.

4.3.3 Analysis

The platform's analysis feature allows users to generate relevant data for research and insight into the topics of 'someone like you' and 'sensor individuals'. The results of the two metrics have been explained in Section 4.2.



a day. For example, if two riders enter the same MTR station and leave the same MTR station at approximately the same time, then we regard them as having shared a trip and are therefore 'someone like you's to each other.

Output Columns :

Start Station | End Station | Start Date | Amount of "someone like you" in weekdays | Amount of "someone like you" in weekends

Input (required)

Interval

Minutes

30

Period

Date

YYYY-MM

Figure 4.3.3.1: Analysis Features – Someone Like You and Sensor Individuals

4.3.4 Security

The data provided by the MTRC, comprising of a large volume of octopus transaction information, is sensitive and bound by a confidentiality agreement. Thus, the data cannot be disclosed to the general public. Accordingly, the platform is made secure such that it can only be accessed by admin users who have signed the confidentiality agreement. This layer of security is enforced by JSON Web Token (JWT) creation and verification through credentials provided to users with data privileges. The platform's login page protects sensitive data from users without privileges.

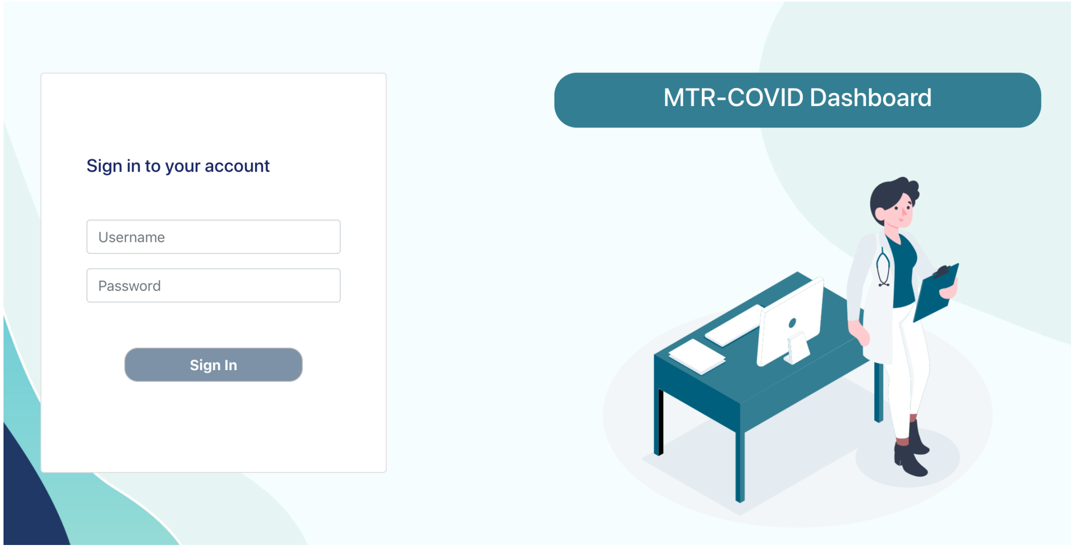


Figure 4.3.4.1: Platform Login Page (JWT functionality)

5 Project Schedule

Table 6.1 shows the project schedule along with the status of each of the tasks. The tasks till phase 3, including the final web application deliverable, have been completed. The team is currently commencing the final presentation preparation.

Table 6.1: Project Plan

Date	Tasks	Status
1 Sep 2020	Pre-processing of data: 1. COVID-19 data 2. MTRC data	Completed
1 Oct 2020	Development of databases	Completed
4 Oct 2020	**Deadline for deliverables of phase 1 (Inception):** **1. Detailed project plan** **2. Project website**	Completed
31 Oct 2020	Mobility trend analysis: 1. Basic visualizations and insights 2. Data streaming/mining	Completed
Mid - Nov 2020 (Date TBA)	Showcase of progress in the opening of The Tam Wing Fan Innovation Wing Phase 1.	Completed
3 Jan 2021	Geo-Spatial Analysis	Completed
11 – 15 Jan 2021	First presentation	Completed
24 Jan 2021	**Deadline for deliverables phase 2 (Elaboration)**	Completed

	1. Preliminary implementation **2. Detailed interim report**	
18 Apr 2021	**Deadline for deliverables phase 3 (Construction):** **1. Finalized tested implementation** **2. Web Application platform** **3. Final report**	Completed
19 – 23 Apr 2021	Final presentation	Not yet started
4 May 2021	Project exhibition	Not yet started

6 Limitations and Challenges

6.1 Dataset and Server Performance

While the large volume of data is exceptionally advantageous and is the key to the project's success, its vast and confidential nature brings about challenges to its efficiency.

Bound by the MTRC's confidentiality agreement, maintaining data security has been a priority throughout the project. Accordingly, the sensitive data resides on a private HINCare server that is accessed through multi-hop SSH logins using admin credentials (Figure 3.2.2.2). However, this extra layer of security leads to slow data retrieval speeds. While this efficiency has improved throughout the project, it still has drawbacks in cases where complex queries (with wide parameters) are generated, data retrieval tends to take a larger amount of time (5-10 minutes).

Additionally, the low processing power and computation capabilities of the FYP server leads to a slower performance of the complex analysis functionalities of the platform - 'someone like you' and 'sensor individuals'. As a result the data response in the two cases may take an immense amount of time (around 15 minutes).

The generation of spatiotemporal visualizations entails extensive complexities. This, in addition to the aforementioned inefficiencies, led to a failure to generate spatiotemporal patterns (Section 3.5.1) for the two analyses, which could have potentially led to the identification of **super spreaders**. These visualisations, while possible, were hindered due to the time constraints of the project.

6.2 COVID-19 and MTR Relationship

While the project has led to some promising results supporting the hypothesis that MTR ridership behavior has a strong influence on the coronavirus spread, it does not confirm the said hypothesis. This report features results of periods with a larger number of new COVID-19 cases, and thus, can portray the correlation between the two datasets. However, the Hong Kong government has been able to contain the spread of the virus, therefore, the infection patterns are inconsistent.

6.3 ArcGIS

ArcGIS, one of the most sophisticated tools used in the project, was utilized to create complex visualizations. Understanding the complexity of the software was among the main challenges of the project. Subsequently, realizing the full potential of the software would have required a lot more knowledge.

Additionally, the ArcGIS JavaScript API utilized in the frontend of the platform is not as developed as the software itself, and therefore, it lacks certain functionalities. Thus, the platform cannot create heatmaps for the COVID-19 hotspots similar to those exhibited in Figure 4.1.3.

7 Future Works

The vast amount of data stored in the project's repository has immense applications and can be harnessed to great extents. Therefore, since this project had a defined timeline, bound by certain time limitations, more functionalities can be implemented in the future. This chapter discusses the possible methodologies and features that can be added to the project in the future.

7.1 Real-time Data Streaming

Currently, the COVID-19 data is being manually fetched, processed, and added to the SQL database. However, the open HK government-provided APIs can be used to fetch real-time data. This can further enhance the dynamic nature of the platform.

The current process involves downloading raw data and pre-processing it using python scripts (involving ArcGIS APIs and Pandas library). However, it is possible to implement an automated process on the python-based Django backend that fetches the data, runs operations on it, and, when requested, sends it to the frontend per the parameters specified.

7.2 Complex Visualisations

As mentioned section 5.1 and 5.3, the current version of the platform does not support certain complex visualisations. These include the spatiotemporal visualisations for the two types of analyses - ‘someone like you’, and ‘sensor individuals’ - and the heatmaps for coronavirus hotspots in Hong Kong. For the former, the failure of implementation was mainly due to the time constraints of the project. Thus, if more resources and time are allocated to the feature, some powerful visualisations can be created. Additionally, since the failure of implementation of heatmaps was due to the lack of the feature in the ArcGIS API, some other API can be utilized for heatmap generation.

8 Conclusion

The report presents a trustworthy and resourceful Big Data Solution that serves as a versatile toolbox for research purposes. The solution has three parts: a Data-stream engine for an immense data repository, an exhibition of complex visualizations comprehensible to both experts and non-experts, and lastly, a web application with a user-friendly interface for the supply of knowledge gained from the project.

Despite the global emergence and distribution of vaccines, the Coronavirus Pandemic is far from over, with new cases are emerging daily. Thus, it is crucial that humanity bands together and utilizes collective knowledge to fight against the spread of the virus. Intending to contribute to this purpose, the project plays a role in sharing the knowledge and insights gained by the team. While the project fails to pinpoint strong analytical conclusions, it enables more knowledgeable researchers to do by providing multi-purpose tools through the platform's functionalities. Furthermore, the web application serves as a secure platform where HKU researchers can accumulate their collective findings, resulting in a higher possibility of uncovering promising results.

9 References

1. Appendices. (n.d.). Retrieved October 28, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/appendix.html
2. *China, Hong Kong SAR*. Worldometer. https://www.worldometers.info/coronavirus/country/china-hong-kong-sar/.
3. García, S., Ramírez-Gallego, S., Luengo, J. et al. Big data preprocessing: methods and prospects. Big Data Anal 1, 9 (2016).
4. How an SSH tunnel can bypass Firewalls, add encryption to application protocols, and help access services remotely. (n.d.). Retrieved April 15, 2021, from https://www.ssh.com/academy/ssh/tunneling
5. Ishitamaheshwari, Pasupuleti, C., M.H.M, M., & Rajaraman, J. (2020, December 12). Secure shell protocol (ssh protocol): Home page. Retrieved April 15, 2021, from https://iot4beginners.com/secure-shell-protocol-ssh-protocol/
6. KEEP MOVING - MTR. (n.d.). Retrieved October 28, 2020, from http://www.mtr.com.hk/archive/corporate/en/investor/annual2019/EMTRAR19.pdf
7. Liang, D., Li, X., & Zhang, Y. Q. (2016). Identifying familiar strangers in human encounter networks. Europhysics Letter, 116(1), 18006.
8. Miller, R. (n.d.). Data Preprocessing: What is it and why is important. *CEOWORLD Magazine*. doi:December 13, 2019

9. Zhou, J., Li, Y., & Yang, Y. (2017). Familiar strangers: Visualising potential metro encounters in Beijing. *Environment and Planning A: Economy and Space, 50*(2), 262-265. doi:10.1177/0308518x17745874

10. Zhou, J., Yang, Y., Li, Y., &amp; Maurer, V. (2018). Someone like You: VISUALISING co-presences of metro riders in Beijing. Environment and Planning A: Economy and Space, 50(4), 752-755. doi:10.1177/0308518x18774049