Final Year Project

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

Supervisors:

Professor Reynold Cheng

Shivansh Mittal

Group Members:

Marco Brian Widjaja (3035493024)

Janice Meita (3935492977)

Marvin Ali (3035361817)

Jain, Rishabh (3035453608)

Nagra, Harsh (3035437707)

Date: 17/April/2021

Abstract

Hong Kong was one of the first places that were hit by COVID19 pandemic. The Hong Kong government took swift actions to respond to the outbreak and therefore had been coping relatively well with the pandemic. Government strategies include the restriction of social gatherings, the closing of borders to mitigate the spread of the disease and travel restrictions on public transportation systems. During this crucial time, professionals from the field of epidemiology / the public sector require to perform an enormous amount of data analysis to gain a better understanding of the pandemic. These data analytics tasks usually involve a lot of manual work and can become extremely time-consuming. Our project aims to retrieve actionable insights from the pandemic data enhanced with the aid of MTR transaction data. We plan to create a software system in the form of a user-friendly application that allows professionals and researchers from the field of epidemiology, or in the public sector to find informative insights that help them to create better strategies to combat COVID19 and to measure the effectiveness of those strategies. Our software will allow professionals to retrieve faster insights on-demand and with much less work than previously needed. We are using industry-standard data science tools and languages such as Python, R, and their open-source libraries for data processing related tasks. Additionally, we are using ArcGIS Pro software from Esri for spatial data analytics and MySQL databases for the storage and retrieval of data. For our final product deliverable (public software) we will be using the latest web framework technologies available such as React.JS and Django (Python).

Acknowledgements

We would like to give our greatest acknowledgements to our supervisors Professor Reynold Cheng and Research Assistant Shivansh Mittal for their support, guidance and advice throughout the course of our project. We are thankful to the MTR Corporation for being very collaborative in providing the data for analysis. We are also thankful to Esri Corporation for providing us the licenses to work with ArcGIS Pro software.

Table of Contents

List Of Figures

Figure 16 – “Sensor Individuals” analysis page

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

HKCHP – Hong Kong Centre Of Health Protection

HKU – University Of Hong Kong

SQL - Structured Query Language

SSH – Secure Shell

ORM – Object Relation Mapping

Introduction

1.1 Background

Earlier this year in 2020, the world was faced with a global pandemic, the SARS-CoV-2 virus. Hong Kong was no exception and had encountered a few waves of the virus, though it has performed fairly well in coping with the spread of the virus as compared to other countries. The government had implemented multiple social distancing strategies such as quarantines, travel restrictions, and limiting social gatherings as a method to mitigate the spread of the virus. One of the important measurement indexes of social distancing is through the behaviour of local transportation amongst the people in the region.

The Mass Transit Railway (MTR) is the main mode of public transport in Hong Kong and accounts for 41% of all public transport trips made in a single day [1] which made it a good starting point for observing transportation trends. Through our collaboration with MTR Corporation, we were able to extract passenger travel data across age groups to understand the impacts and the effectiveness of government strategies on the movement of people. This type of research was previously not possible due to the lack of large-scale data on local transportation.

1.2 Project Aims

This project aims to perform analytics on the pandemic data in Hong Kong and enhance it with the aid of the MTR passenger travel data that has been provided by the MTRC. The project aims to gain important insights that could be used to guide future strategies in combatting the pandemic. We believe that with better insights, we could devise optimal strategies for the best returns. The concern that the main mode of transportation (MTR) and its usual crowdedness could breed for potential coronavirus hotspots increases our motivation to discover the correlation of COVID-19 spread with the use of MTR. We would like to understand how these two factors interact with one another and to understand how the behaviour of travel changed over the course of the pandemic, and to measure the effectiveness of government strategies.

The final deliverable of this project is to build a publicly available web-based platform capable of generating dynamic visualizations and analysis of the MTR and COVID19 data. The platform will be specifically catered towards 2 types of users. Firstly, authorized researchers and professors who have signed the confidentiality agreement contract from the MTRC. Secondly, officials from the public sector (health, transport sector) and the general public who have signed the confidentiality agreement contract from the MTRC. Our project aims to replicate the analytic process through a much more streamlined and user-friendly interface so that the professionals and researchers can use our software to create the analysis themselves on-demand. Previously, researchers had to perform time consuming tasks for pre processing data, and visualization and analytics tasks. Researchers previously used traditional methods of diagram visualization using Microsoft Excel which is not sustainable with the inflow of huge datasets. Our aim is to help these researchers to save time by providing a one-stop platform for all their data needs.

Because of the multidisciplinary nature and scope of this research topic, our hope is that with this web platform available, it will act as an auxiliary support in further advancing research interest and funding on this topic in the future.

1.3 Previous work and research

Familiar Strangers

Our project would like to advance further on previous works that had been done on the concept of “Familiar Strangers” into our analysis. This is the idea that there are people who you do not know (strangers) but they share a high degree of commonality (same occupation, same interests) which may result in two individuals who do not necessarily know one another but have close proximity and encounter at high frequencies [2]. The idea of “Familiar Strangers” had been existent for a long time but previously had not been easily measured or quantified due to the lack of geolocation technologies. Previous methods of identifying “Familiar Strangers” tend to be qualitative as only conducted through personal anecdotes and surveys of describing people they often see in public places [2].

There have been previous studies on understanding the human contact network in metropolitan areas through the use of Smart Card data (Smart Card used in public transportation such as bus and metro stations), in which many of them analysed the existence of “Familiar Strangers” [3]. We would like to take this research paper as an inspiration to perform similar analysis on our MTR data. Specifically, by knowing the exact time period of MTR travel of numerous passengers simultaneously we could infer and deduce the existence of such “Familiar Strangers” in a more quantitative manner.

We want to explore further the idea of “Familiar Strangers” because they may affect how diseases spread in the community [4]. In our case of MTR travel, people who tend to have similar starting travel time and travel destinations may be subjected to prolonged proximity with one another for a considerable amount of time. In this period of time of contact, there could be a probability of contacting the virus from the infected “familiar stranger”. Thus, knowing how to mitigate such risks could deem beneficial.

Furthermore, through the understanding of complex human network, there has also been research to show that we can strategize a more effective distribution of vaccines [5]. This could be further explored in the future when vaccination schemes have commenced. Thus, a simple strategy could be to perform vaccinations starting on places with the most “familiar strangers”.

1.4 Project Scope

This section briefly discusses the scope of the project.

1. Develop a secure SQL relational database for storing confidential MTR data and COVID19 data.
2. Replace the current data distribution methods (currently through physical DVDs) by providing a readily accessible platform for downloading the dataset. [This functionality is catered towards the academic researchers]
3. Utilize ArcGIS and Python for the MTR analysis
4. Develop dynamic visualization tools available in the web platform.

1.5 Report Outline

After discussion on the background, motivation and aims of this project, we continue onwards on the second section where the discussion will be focused on the methodologies performed in this project. In depth details relating to the tasks in the data pipeline (collection, processing, and analysis), mobility trend analysis, geospatial analysis and web platform development will be thoroughly elaborated. In section 3, the discussion will be focused on the results that we have obtained from our research and their implications for our project, we also will discuss functionality details of the web platform. Section 4 concerns the challenges that were faced during execution, the limitations of this project, and future developments that could be made. In section 5, we placed our closing remarks on the project.

2. Methodologies

Section Overview

This section discusses the methodologies of the data pipeline in chronological order. Starting from how data sources are collected, moving on to the methods for data processing, and finally how the processed data is stored and accessed for analytics. Furthermore, this section describes the different analytical methods that we utilized for the MTR data, such as station density, travel pattern, passenger volume/mobility, 'someone like you' and sensor individuals. Finally, the section ends with a detailed elaboration on the development of the web platform.

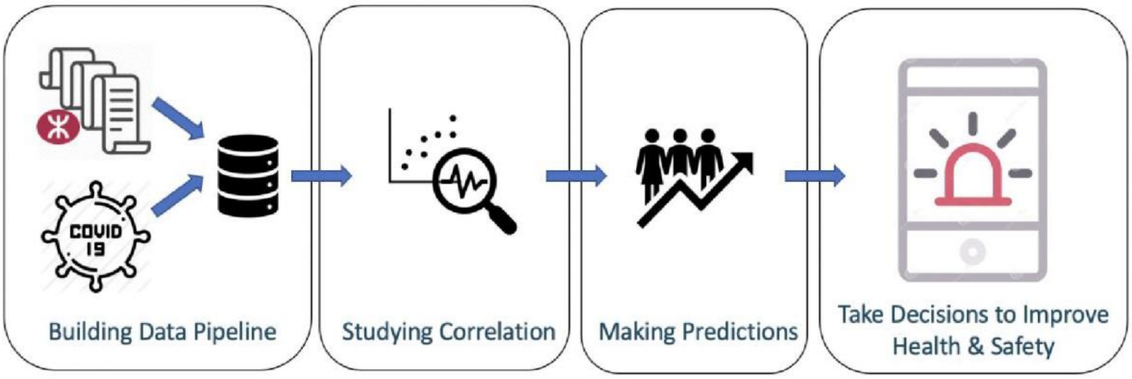


Figure 1 Step-by-step workflow

2.1 Data Sources

MTR Data

Upon a signed agreement for collaboration between HKU and MTR Corporation, we were able to obtain detailed passenger transactions data based of the Octopus card/ticket usage.

Each data row includes the card/ticket ID, entry station, exit station, entry and exit time, ticket type (octopus or single ride ticket), and the card type (adult, child, student, senior, others). The transaction data currently range from 1 January 2020 to 30 September 2020, we also have the same data from the same period from last year 1 January 2019 to 30 September 2019.

COVID19 Data

The COVID19 data was collected from https://data.gov.hk/. The website contains several trusted COVID19 datasets:

- Details of probable/confirmed cases of COVID-19 infection in Hong Kong
- Residential buildings which probable/confirmed cases have resided in the past 14 days, or non-residential building with 2 or more probable/confirmed cases in the past 14 days
- Latest situation of reported cases of COVID19 in Hong Kong

2.2 Data Cleansing & Processing

Cleaning & Processing MTR Data

Around 1 million records of MTR passenger transactions contained NULL values for the exit station. These data records are removed because they are erroneous and do not provide value for the dataset.

Cleaning & Processing COVID19 Data

The data from https://data.gov.hk/ is in csv file format and occasionally contains unstructured, inconsistent or missing data. For example, sometimes the dataset can have missing date fields, or spacings across text. We also want to have a consistent domain constraint on some of the data fields therefore this data cleansing is necessary. Basic data cleansing scripting was done with Python and R with the help of open source libraries (pandas, Numpy). We used Python and R because these programming languages are very powerful and intuitive when it comes to performing tasks related to string and file manipulation, they are the industry standard tool for data science related tasks. After data cleansing we obtained these data fields:

- Case ID (Unique identifier for a patient)
- Date Reported
- Date of Onset
- Asymptomatic (Yes, No)
- Gender
- Age, Age group
- Status (Recovered, Removed)
- Classification (Imported case, Locally transmitted)

We continued with processing the cleaned COVID19 data to create additional statistics from the raw data. Our processed data now includes:

- Cumulative cases

- Cases per day

Date	Total Cases	Total deaths	Total discharged	Critical Condition	New Cases	Active Cases	New Deaths	New Discharged

Processing Geographical Data

We used an external geographical information software, ArcGIS, to assist us in finding the geographical locations of MTR stations and COVID19 patients. Initially, addresses were in the form of text and we used the ArcGIS geocoding API service to convert the addresses into geographical coordinates, the longitude (X) , and latitude (Y). These geolocations are essential to perform spatial analysis and visualization in the ArcGIS software.

Additionally, we also used Google Maps API when ArcGIS failed to give an accurate result.

2.3 Database Modelling

The need to develop a database arises from the challenge to surmount the huge amount of data that we need to analyse and store. The MTR data has more than 500 million transactions (over 60 GB of data) , and the COVID19 dataset has thousands of records of patients. The database will allow us to have convenient and efficient access for information retrieval.

We have chosen to use a relational database schema to store our data because then we will be able to perform powerful and complex queries for data mining applications. With a relational database we will be able to retrieve derived information such as the volume of changes in MTR passengers, as well as analysing the different groups of passengers across varying timeframes. Thus is the reason why we chose an SQL database as compared to a NoSQL database, which will not allow us to perform the same set of complex queries.

Another reason for not using NoSQL databases is because the data will be stored in an unstructured format where no schema is enforced, but we would want to work with enforced schema-structured data because missing data fields can affect our results, thus our option to choose a SQL database is more suitable for our application.

Currently we are using a MySQL relational database that is running on a private HKU server (the server is called HinCare). We designed a database schema for storing the above data. Our database currently contains the following tables and schema:

MTR Data

- Passenger Transaction from January to April 2019/20

Octopus ID (anonymous)	Card Type	Entry Station	Exit Station	Entry Time	Exit Time

- MTR Geolocation

Station Name	Longitude (X)	Latitude (Y)

COVID19 Data

- COVID19 cases, this table contains all the details of each particular COVID19 case

Case ID	Reported Date	Date of onset	Asymptomatic (y/n)	Gender	Age	Age group	Status (Hospitalised / Discharge/ Deceased/ Pending admission)	Classification

- Building List (Jan – April 2020), this table contains all the buildings of probable and confirmed cases

Address	District	Building	Last Date	Case ID	Longitude	Latitude	Address type (residential/non-residential)

- Hospitals and Closest MTR station, this table contains the hospitals where COVID19 patients were admitted and the closest MTR station in its proximity

Hospital Name	MTR station

The database would serve as a source of information for us to perform our analysis and is the fundamental underlying infrastructure of our software. In the next phase during development of the application, the backend system would interact with this database to fetch, store and perform data queries. This database is crucial for us to retrieve the findings in the next section, in particular they are the source for the data of the ArcGIS visualization graphs and the graphs generated from Python.

2.4 Analysis

2.4.1 Mobility Trend Analysis

Mobility trend analysis is the study of the general movement and travel trend of the public. This project will include an analysis of the changes in travel decisions and the public's behaviour towards the usage of MTR during the development and spread of COVID19. This project is keen on keying into an attempt at understanding the influence of the COVID19 pandemic on public MTR behaviour and vice versa. Thus, our attempt is to specifically visualize the changes in MTR ridership during key events such as the start of the outbreak, Wuhan lockdown, and periods of COVID19 waves and compare with how it was before the pandemic started. A key metric that we use for measuring mobility trend is by using **passenger volume**, defined as the number of passengers riding the MTR at a given time period. Furthermore, we can investigate mobility changes in specific demographic groups as the MTR data is labelled with the specific card type (Adult, Elderly, Child). Understanding how different demographic groups act can be beneficial, especially when concerning high-risk groups such as the elderly. For example, we can examine whether the elderly are practicing social distancing and staying at home through activity in the elderly card type. Analysis will also include key comparisons between travel behaviour during weekends as compared to weekdays on specific hours of the day. For this example, we can strive to have a deeper understanding of how effective Hong Kong's work from home policy is. Thus, we can derive whether there is a significant change in the inflow and outflow of daily MTR passengers during working hours. This type of analysis will help us better understand the behaviour of the Hong Kong population with respect to new government policies enacted due to the COVID19 pandemic.

2.4.2 Geo Spatial Analysis using ArcGIS (Esri)

This project utilizes the geographical visualization rendering technology of the ArcGIS platform to display the MTR and COVID data on a geographical map. One of the visualizations created out of the MTR data is **station density**. The station density of a station is defined to be the **average number/volume of inflow and outflow of passengers for a station in a given time period**. Another is the **travel pattern** visualisation, which is defined to be **the number of passengers travelling from one entry station to another exit station in a given time period.** These 2 metrics allow us to identify the level of crowding in specific MTR stations as well as in moving trains. With these metrics we can visualize them into the ArcGIS geographical map, for **station density** it will be a point-based graphic that scales in size with the increased/decreased amount of passenger volume. For **travel pattern,** the visualization will be a point-to-point line based graphic from one station to another station and the width of each line is scaled to the amount of passenger volume (The next section will have example figures for illustration).

In addition to these prior MTR visualizations is to overlay them with COVID19 data. In an attempt to confirm the hypothesized relationship between the two data, we will plot the geographical spots of the COVID19 cases on the map and in particular show heatmaps that may signify the presence of a COVID hotspot / cluster. Thus, a more holistic view can be derived by looking at the busiest travel regions in Hong Kong as well as looking into locations where COVID19 hotspots are located.

2.4.3 Contact and Behaviour Based Research

Contact and behaviour-based research will be conducted using MTR data. This research is based upon the idea of the "Familiar stranger" that was discussed in the introduction section. The main objective of this research is to understand the underlying patterns and behaviour of MTR passengers, to understand their behaviour in a more individual user basis and on approximating passengers' contact with other passengers.

Our platform will include a feature that enables the user to do two types of contact and behaviour research: **"someone like you"** and **sensor individuals**.

2.4.3.1 Someone Like You

"Someone like you" is defined as two different riders who simultaneously share trajectories for at least one trip in a day. Two riders who enter the same MTR station and leave the same MTR station at approximately the same period of time can be regarded as sharing a trip and therefore can be regarded as "someone like you" to each other. The goal is to identify MTR routes that have the highest count of 'someone like you' and identify the spatiotemporal pattern of those "Someone Like You".

In order to make the "Someone Like You" analysis there are assumptions that had to be made and these are the following assumptions:

1. People can be regarded as someone like you with each other even if they do not ride on the same MTR carriage.
2. Each pair of individuals that is considered as "someone like you" with each other are assumed to take identical MTR route / path (example if there are more than one station

that allow them to switch lines, we would assume that they would switch at the same station)

These are the steps to process the MTR transactions data and find passengers "someone like you":

1. Firstly, all the possible station pairs in the MTR network are grouped together. (If there are n stations, then there will be (n*n-1) station pairs altogether).

   Example: Stations A, B, C

   Station pairs: AB, AC, BA, BC, CA, CB

2. For each day, we then will group riders' trips according to the corresponding station pairs.

   Example:

Trip ID	Entry station	Exit station
1	A	B
2	B	C
3	A	B
4	B	C
5	A	B

   Trip ID 1, 3, 5 should be in a group with station pair A to B.

   Trip ID 2,4 should be in a group with station pair B to C

3. In each station pair group, we then count how many of these trips have approximately similar entry times. This is done by dividing one full day into 48 time periods of length 30 minutes and we define the similar entry time to be trips that are 30 minutes* away from one another. (*this 30 minute time interval can be varied to different minute intervals)

   Example:

Trip ID	Entry_station	Exit_station	Entry time
1	A	B	8:10
2	B	C	9:30
3	A	B	8:20
4	B	C	9:05
5	A	B	10:25

From this table we can group Trip 1,3 together and Trip ID 2,4 together.

Time Period	Trip Count
8 - 8:30	2
9 – 9:30	2
10 – 10:30	1

4. For each station pair, count the amount of “someone like you”, defined as those with Trip Count greater or equal to 2.
5. For each of the station pairs, count the average amount of “someone like you” on days (weekends, weekdays) in each week.

The end result is that the users will obtain a CSV file that contains information of the average number of “someone like you” on weekdays and weekends in all the possible stations pairs.

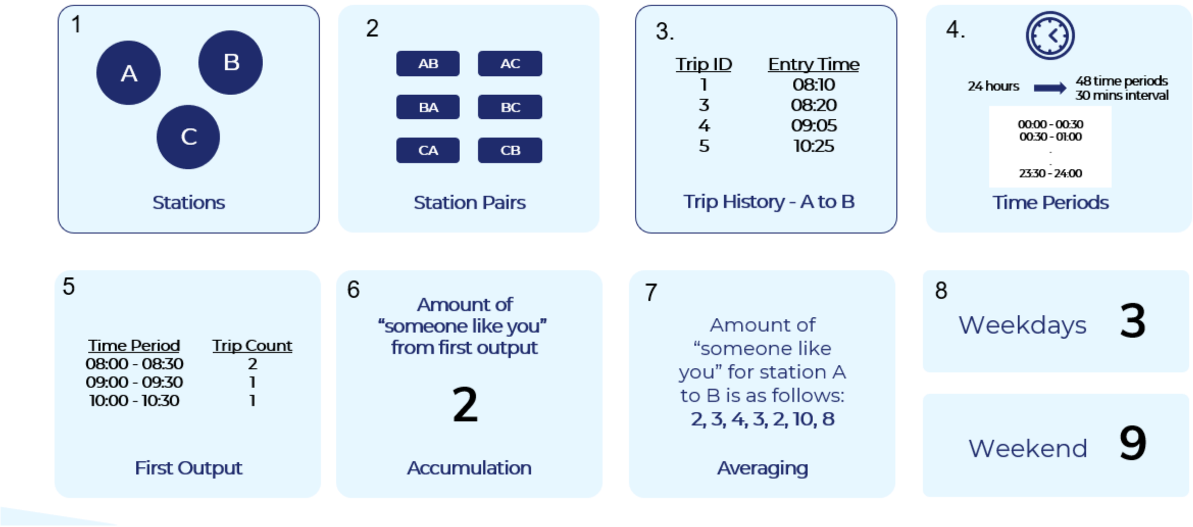


Figure 2 Step-by-step someone like you illustration

With the 'someone like you' analysis, we can discover which particular time and station pair has the most people travelling at the same time. We can use this information to potentially alert people on a busy schedule time and might want to mitigate risks of catching COVID19 from meeting many "someone like you" passengers.

2.4.3.2 Sensor Individuals

Sensor individuals are passengers that potentially have the most physical contact with other passengers at the station or MTR carriage level. These individuals under the COVID-19 situation could potentially become a super spreader, thus being able to identify these individuals can be beneficial. This analysis is aimed to have a better picture of the co-presence amongst MTR passengers. The current MTR transaction data format will require some further data pre-processing before we can perform this "Sensor Individual" task. Currently each MTR transaction only contains the rider's entry and exit station for each trip. Using this data only, we will not have the full path of the rider's journey and therefore intermediate stations that are travelled are not recorded. Thus, a complicated process needs to be done to convert the start and destination into a full journey with intermediate stations and timestamps. For this task we used a package called **networkx** to simulate the whole MTR rail system as a weighted undirected graph.

These are steps to process the data for Sensor Individuals task:

1. The first step we did was to collect the time it takes to get from one station to the next direct station (ex Kennedy Town and HKU is a direct station). In terms of graph theory, this is the geodesic distance of value 1 between two vertices (stations). Collect the time taken to travel for each station to its neighbouring stations. This data is collected manually through using the Google Maps API.

   Ex: (A,B) (D,E) (F,G) are neighbouring station pairs. The following is an example of the data format in CSV.

Station 1	Station 2	Time Duration (mins)
A	B	2
D	E	3
F	G	1

2. Using the **networkx** package, we take the csv file above and the package will reconstruct the following data into a weighted undirected graph which simulates the MTR rail line system.
3. The package allows to find the shortest path between one station to the another station by implementing Djikstra’s shortest path algorithm. Thus, we can implement intermediate timestamps and station journey and find the complete path of the journey with only start and end station as input.

The input is a journey containing only starting and end stations, the output of this analysis will be a CSV file containing the details of the full travel journey for each person including each intermediary station and timestamps.

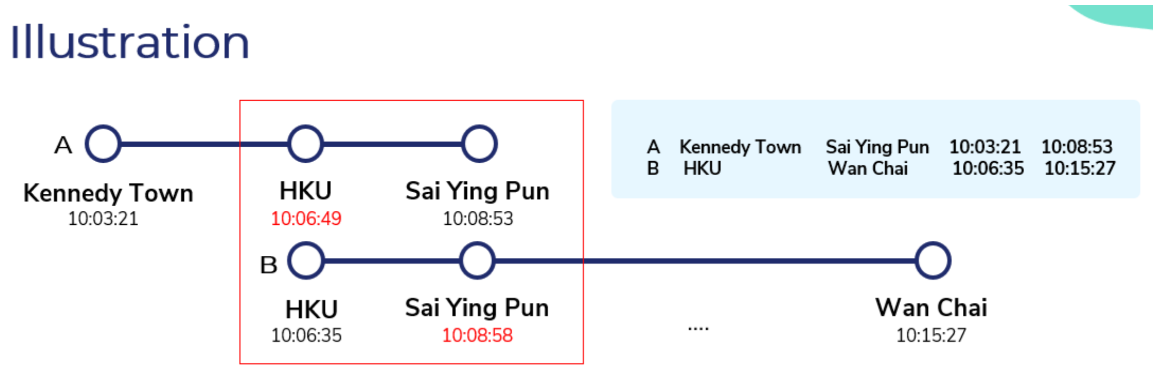
Example (columns are not exactly how it is in code just for easier demonstration):

Input:

Octopus ID	Start	End	Entry Time	Exit Time
1234	Kennedy Town	Sai Ying Pun	2:40	2:45

Output:

Sequence	Octopus ID	Start	End	Entry Time	Exit Time
0	1234	Kennedy Town	HKU	2:40	2:43
1	1234	HKU	Sai Ying Pun	2:40	2:43



A	Kennedy Town	Sai Ying Pun	10:03:21	10:08:53
B	HKU	Wan Chai	10:06:35	10:15:27

Figure 3 Sensor Individuals

Finally, after having multiple full journey trips from multiple people, we then can try to find these sensor individuals by looking at the intersection of timestamps. We want to be able to find the pairs of people who have intersecting journeys and keep that record pair and the journey where they intersected. Therefore, the final result will look like:

Passenger 1 (ID)	Passenger 2 (ID)	Journey Intersection (Station Code)
1234	1983	[12, 32, 33]
9420	8392	[1, 4]

With this record we could look for the Passengers with the greatest number of pairs (potential sensor individuals), we can also search for individual passengers by their ID and find all the other passengers that are in the same path as them.

These are the assumptions made during the analysis of "Sensor Individuals":

1. Trips that can be done through different MTR lines are considered identical. (Ex from Central to Admiralty a person can take either the Tsuen Wan Line or the Island Line, but in our case, we consider them to be just one line since we do not have information of passengers interchanging stations).
2. The intermediary entry and exit times are considered to be the time when people enter / exit the train carriage and not the station.
3. The model does not take into the consideration of the time needed to switch between train lines or the time from Octopus transaction to the point at the train. Therefore we calculate the difference between the expected arrival time in our model and the actual time taken by a passenger for travel. The result is the time difference person not travelling inside the train. We take this time difference and divide by the number of intermediary trips (stations travelled) and add this extra delay time to each intermediary trips.

2.5 Web Platform

This section will discuss the technologies used for the development of the web application. This section also discusses the implementation of such technologies.

2.5.1 Technology Stack

Our web platform makes use of the latest modern open-source web technologies for development. Our front-end application is built using React.js, a modern JavaScript library for building Single Page applications. Our team decided to use React.js for its reusable components and application state management capabilities, which helps increase code-readability and development efficiency. The front-end application is styled with the Bootstrap CSS framework.

The backend application (server) is developed with Django, a python-based web framework for building secure and scalable web applications. We opt to use Django for the backend service because of its wide range of functionalities and flexibility. Our team specifically chose Django for its built-in secure authentication and authorization system which is essential for our application. Django provides a built-in admin system which allows easy management of users and superusers using the platform. This use-case is essential for handling the different users that will be interacting with our platform. Django also provides a built in ORM (Object Relation Mapper) which allows the backend service to communicate via an interface with the SQL database. The Django's ORM allows us to perform SQL queries through programmatic methods instead of using raw SQL queries. The Django ORM allows us to easily transform SQL data into workable pythonic objects that we can interact with in the backend service.

The communication between the front-end React application and the backend Django application is done through REST APIs. Django allows the team to easily set up REST APIs through their **Model-View-Template** design pattern. This allows us to create API URLs **Views** and integrate easily to the **Models** for database access. Django allows the processing of HTTP requests in a simpler manner by having built-in support that automatically converts raw HTTP requests into pythonic **HTTPRequest** objects which contains the metadata of the raw request.

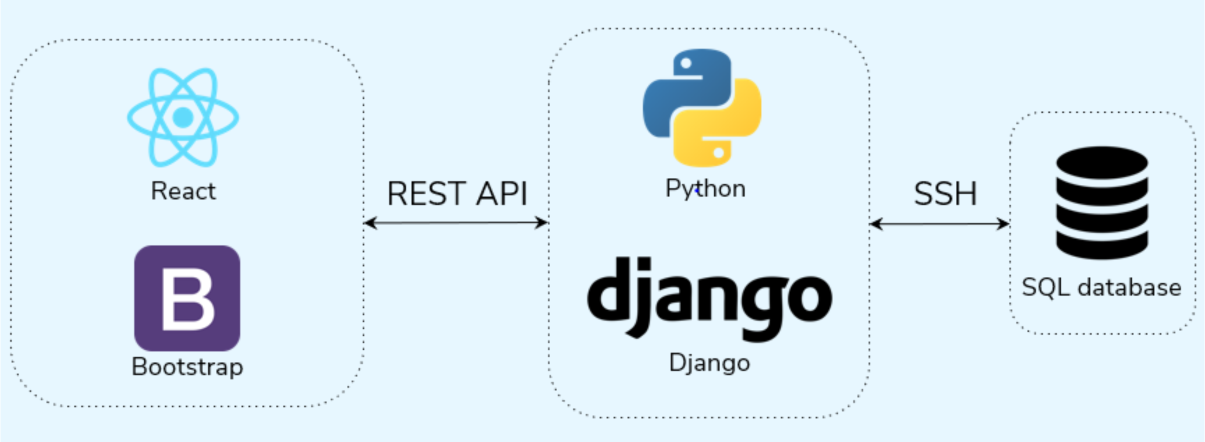


Figure 1.1 Tech Stack

2.5.2 Platform Architecture

The diagram shows the full picture of where each process of the technology stack is deployed. The complexity of the architecture will be discussed, reasoned, and elaborated in this section.

Our database of MTR data and COVID19 data is stored in the HKU HinCare server which is within the HKU intranet network. To access this HinCare server, the user has to login to the HKU CS gateway network through SSH (secure shell) terminal. In order to access the HKU CS gateway network, the user is required to own a HKU Computer Science intranet account.

Once the user has access to the HKU CS gateway network, another SSH is required to access the HinCare server which contains the MTR data. Thus 2 SSH connections are required to access the HinCare server. The added complexity of this method is required because the confidentiality agreement of the MTR data does not allow us to share the data to the public. This requires storing the data privately in-house and not on cloud public services domain.

This added complexity requires multiple workarounds during the development of the platform, especially from the backend development side. The problem was to figure out how to get the data available to the public domain for authorized users in a secure manner. Our team came up with the solution of using double port forwarding to be able to get connection from the private HinCare server to the hosting for FYP Virtual Machine. Initially the Django backend application is run in the HinCare server, in which it has direct access to the private MTR database. The first SSH port forwarding is done from the HinCare server to the HKU CS gateway server (from port 8000 to 8080). Second SSH port forwarding is done from the HKU CS gateway to the FYP Virtual Machine (from port 8080 to 8080). Thus, the backend will be accessible through the fyp.cs.hku.hk domain in port 8080.

The deployment of the front-end application is much simpler as only required deploying on port 80 in the FYP virtual machine server.

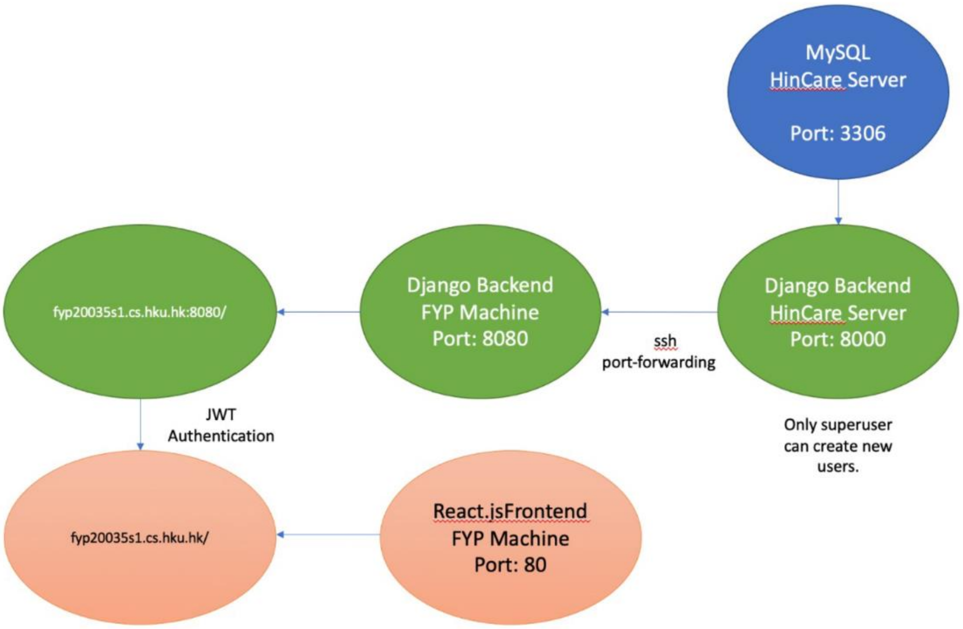


Figure 4.2 Platform Architecture

The architecture above solved an important problem. Previously, to access the MTR data people would need to own a HKU CS account. With this architecture we are now capable of creating an authentication service with Django, which is able to handle user authentication. Our Django service acts as a middleman to access the MTR data. Our Django application running inside the private HinCare server has access grant to the MTR database. Thus, Django acts as a user authentication management service for handling and storing user credentials and passwords. Users may get access to data through Django as long as they have the valid authentication credentials. Therefore, the authorization flow for getting authentication permissions would be:

1. Researcher requests/apply for platform access.
2. If the applicant is eligible then the applicant will require to sign the confidentiality agreement with MTR.
3. Once approved, a new user account will be created in the platform (Django) for that user

4. With the new user account, the user will be able to access the platform and be able to retrieve the MTR data.
5. New users do not need a HKU account anymore to be able to access our platform. Meaning that researchers from other faculties may collaborate without the need to make HKU CS account for them.

The authentication system is powered by JWT (JSON Web Token) by using a python library called **rest_framework_simplejwt**. Once a user successfully logs in to the platform, a JSON Web Token will be generated from the backend Django service and the token is sent to the client side. The client side will use this token when calling REST APIs to the backend, without this token the REST API call will not be authorized, and no response will be generated back. Currently, this token expiry date is set to a 20 minute timeout, the user will be logged out of the platform and will have to re-login for new token credentials. The token expiry time is set for increased security.

2.5.3 JavaScript ArcGIS API (Front-End)

The platform's visualization capability is dependent on the ArcGIS technology. ESRI provides a JavaScript ArcGIS API which allows to render ArcGIS visualizations onto web-based applications. This allows the creation of interactive user experience with dynamic ArcGIS visualizations right from the browser. Initially, the **arcgis-js-api** module will need to be installed before the visualizations can be created on the frontend. All of the visualization code and logic is done through the front-end React application. The process of creating ArcGIS visualizations into the web application is the following:

1. Specific data for a particular visualization is fetched from the backend service through REST API calls. [Exampe: geolocation data (longitude, latitude)]

2. Data from the backend is converted into ArcGIS ***Graphic*** objects. A collection of ***Graphic*** objects is converted into ArcGIS ***FeatureLayer*** object. (Imagine a ***FeatureLayer*** to be a dataset and a ***Graphic*** object to be a row in that dataset).
3. Render a new **ArcGIS Map** object (this is the baseline map for visualization) and set the data source of this map to be the ArcGIS ***FeatureLayer*** object that we created.
4. By specifying different configurations / variables during the creation of the ***Graphic*** and ***FeatureLayer*** object we can alter the rendered visualizations for different visualization functions.

Thus, by using this procedure, we can create multiple various visualizations on the platform.

2.5.4 Backend Optimization

Initially our platform receives queries requests and performs the computation only after receiving a request to generate the query results. This was the method for generating queries for the MTR metrics such as ‘Station Density’, ‘Travel Pattern’, ‘Passenger Volume’. Our team realized that this ad-hoc approach can quickly become an overhead as the number of requests increases, and the system could quickly freeze trying to perform too many heavy computations simultaneously. Initially, by using the ad-hoc approach of computing results at request time took the machine around 2 minutes on average* to generate results for each request. Therefore, to increase the performance of our platform, we have precomputed the results in the background and stored them into tables for direct querying. This significantly improved the response time for each API calls into milliseconds.

*The initial request time for each metrics is: Passenger Volume (2 minutes 13 seconds), Station Density (2 minutes 2 seconds), Travel Pattern (3 minutes 33 seconds).

Our team also utilizes database indexing to increase the performance of SELECT operations on the tables.

3. Results

Section Overview

This section discusses the results that are obtained in consequential from the previous steps. The results were created from taking the previously processed data and using software tools such as Python, R and ArcGIS to create the graphs.

COVID19 New Cases (categorized Age group)

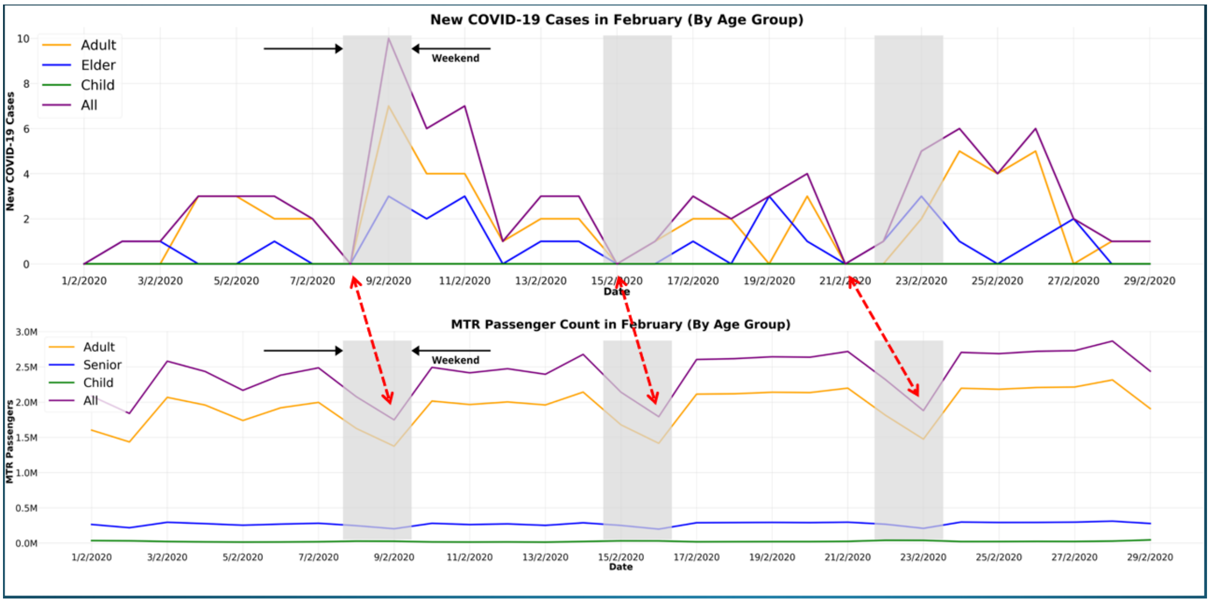


Figure 5 Graphs representing new covid19 cases and the MTR passenger travel categorized by age group

There are two graphs above that describes the relationship between the COVID19 cases and MTR Passengers over time. The top graph shows the daily cases of COVID19 during February 2020 (first wave) categorized by the age group (Adult, Elderly, Child) and the bottom graph shows the MTR passengers categorized by age group in the same month. Our study found a correlation between the new cases of COVID19 with the passenger flow in

MTR. The emphasized greyed area shows that the moment where MTR passenger flow increases, the cases of COVID19 increases as well, especially in the adult age group. There is almost no MTR activity for the children age group most likely due to the closure of schools, thus we also see that there are no child related cases of COVID19.

Spatial Visualizations

By using the ArcGIS software, we were able to have a better visualisation and understanding of our datasets. Here are some of the visualisations that we have obtained from using ArcGIS.

Busiest MTR routes

Here below we are showing the top 20 busiest MTR routes during January and April 2020 in Figure 2.1 and 2.2.

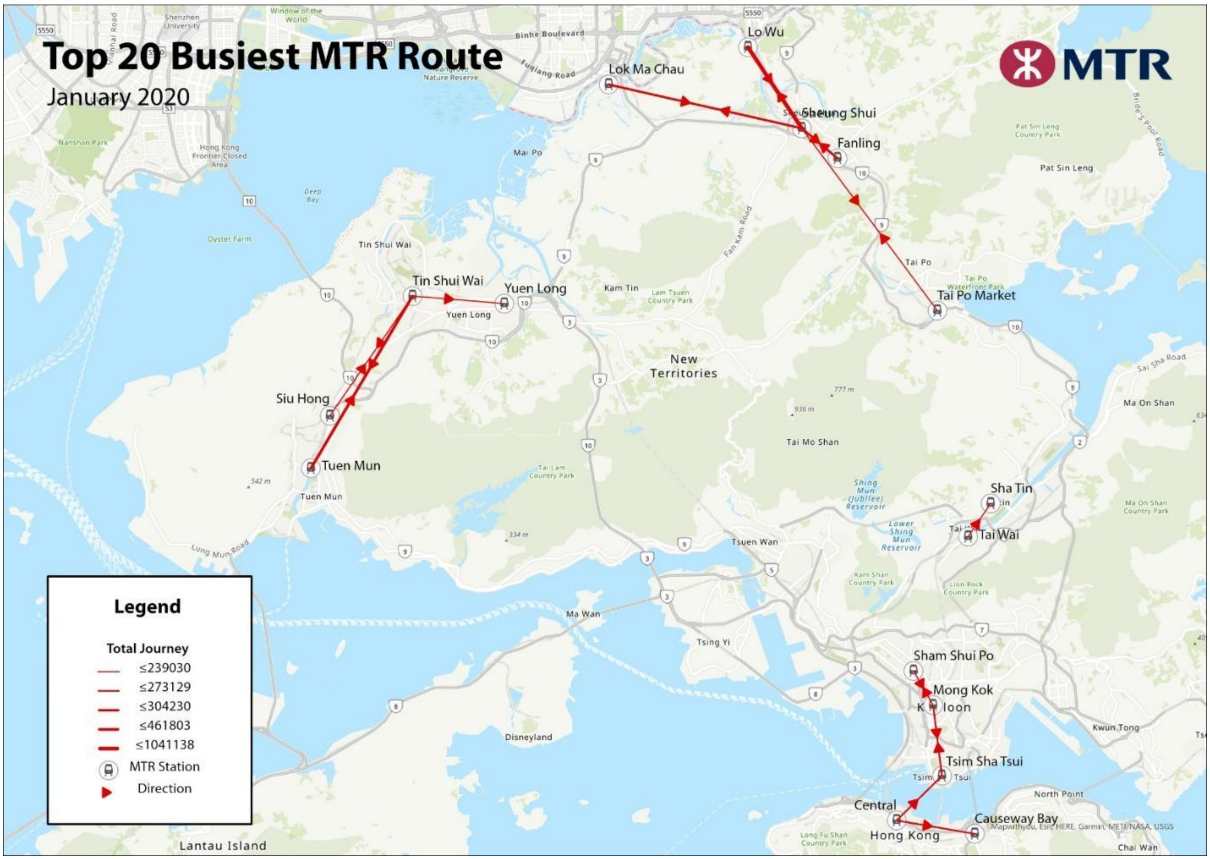


Figure 6.1 Top 20 Busiest MTR Routes January 2020

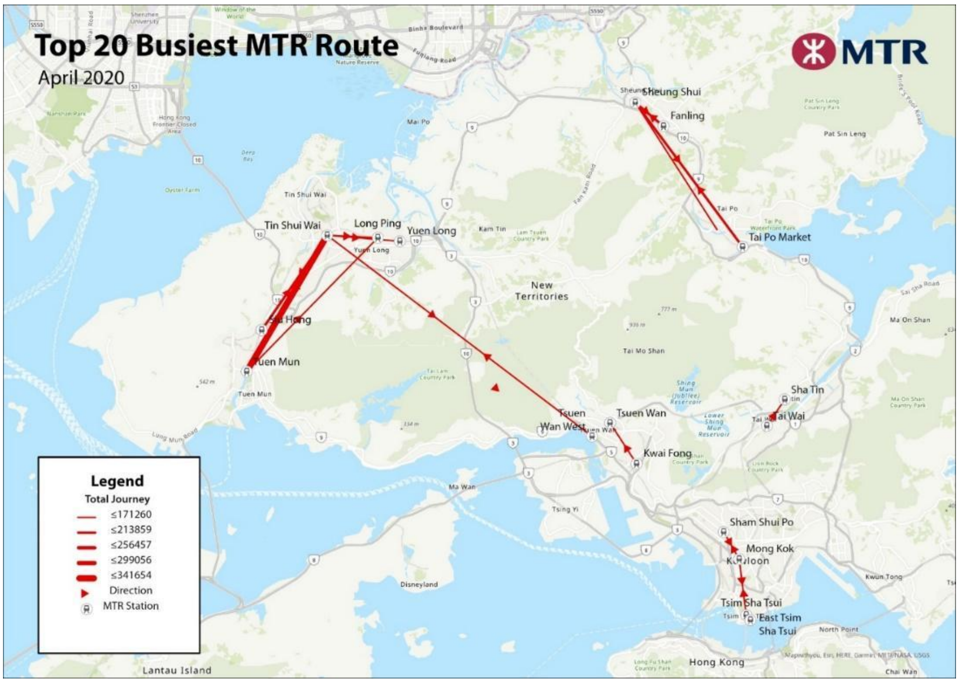


Figure 6.2 Top 20 Busiest MTR Routes during April 2020

The graphs above show the changes in the trend of MTR passengers' movement throughout the months of January and April (February and March Omitted). There is a change in the busiest MTR routes ever since the pandemic began. The most significant change is that the volume of people travelling in the busiest route dropped by 67%. In January the busiest route had 1,041,138 passengers and by April it dropped to 341,654 passengers.

Heatmap Graphs of COVID case distribution (Second Wave)

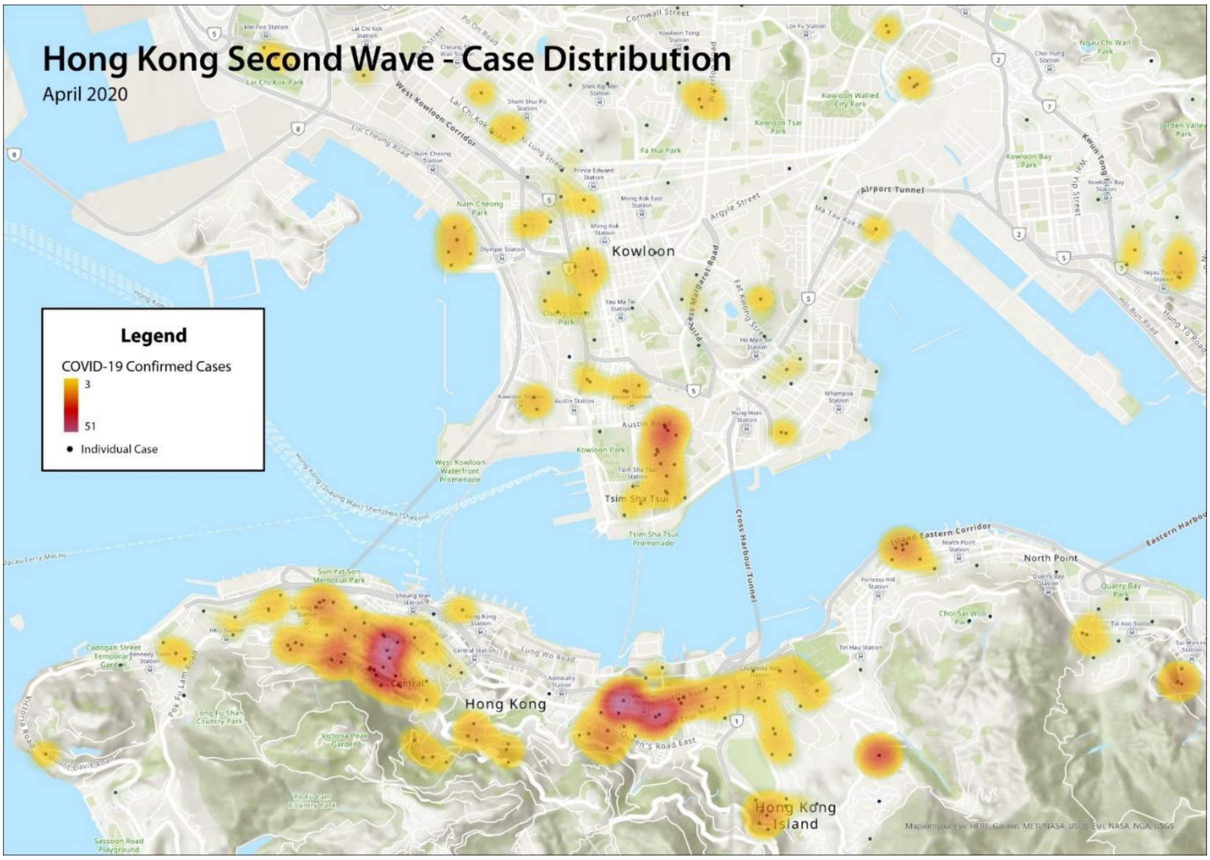


Figure 7.1 April 2020 COVID19 Heatmap

The graph above shows a COVID heatmap distribution to know where the virus hotspots were located across the region. In the example above is the graph during April 2020.

Furthermore, we performed an overlay on the graph of MTR incoming volume of passengers to the heatmap graph. We took the top MTR stations with the greatest number of incoming volume of passengers in April 2020. We retrieved the following results shown in Figure 3.2:

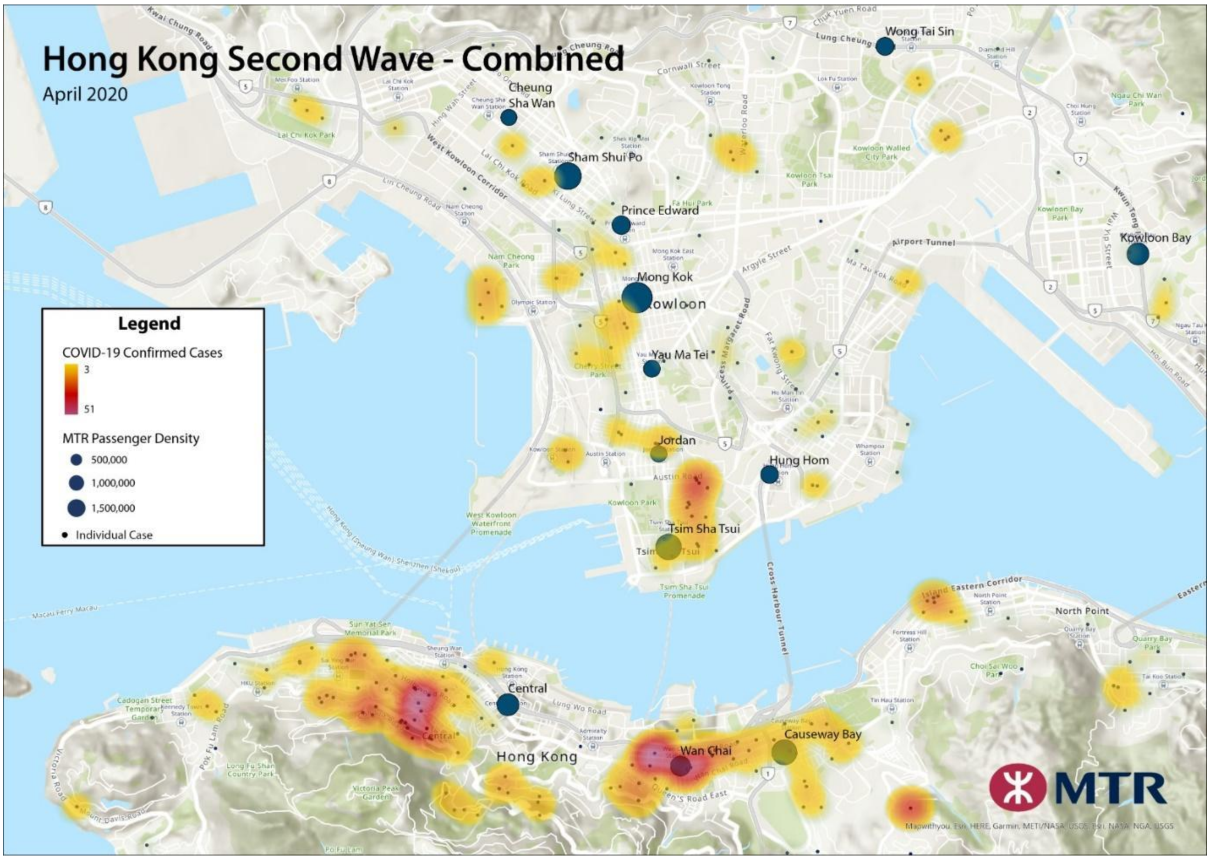


Figure 7.2 April 2020 heatmap with MTR volume density

The blue circles represent the volume of passengers at a busy MTR station. The bigger the circle meant that there were more passenger volume. It was observed that there was a correlation between the areas of covid hotspot and busy MTR stations. The places with the busiest MTR activity are closer to the covid hotspot areas.

Web Platform User Flow & Functionality Details

This section details on the user experience of the web platform. The team had successfully deployed a working platform that is available at http://fyp20035s1.cs.hku.hk/

Login Page

The user passes the correct authentication credentials (username and password) that has been created by the admin.

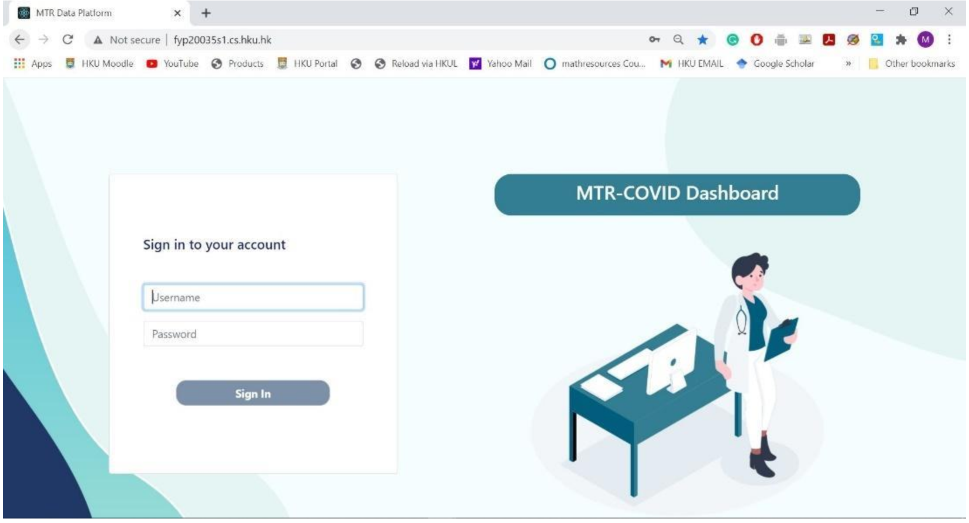
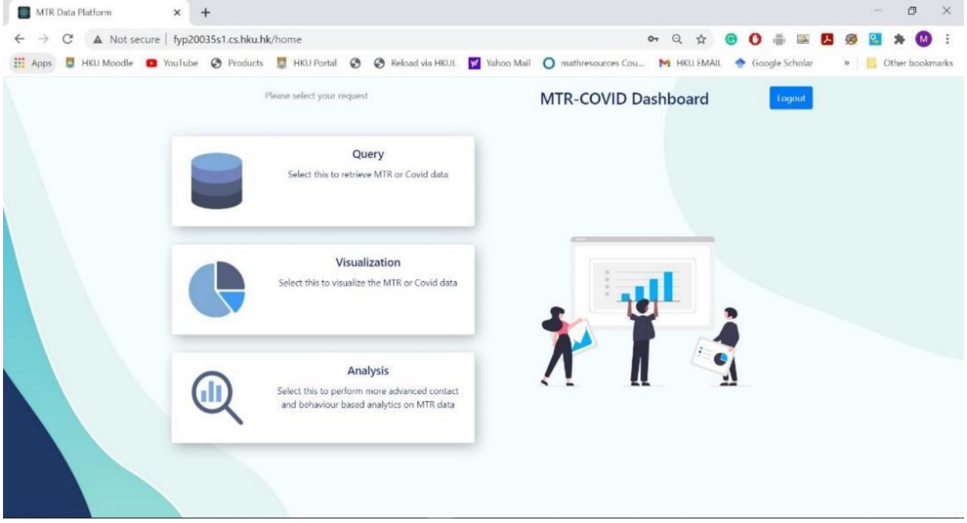


Figure 8 Log In Page

Home Page

After the authenticated user signs into the platform. They will be redirected right into the home page which allows a selection of 3 different platform features. Querying raw data, performing dynamic visualizations, and "Contact and behaviour-based" analysis.



Select this to retrieve MTR or Covid data

Select this to visualize the MTR or Covid data

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

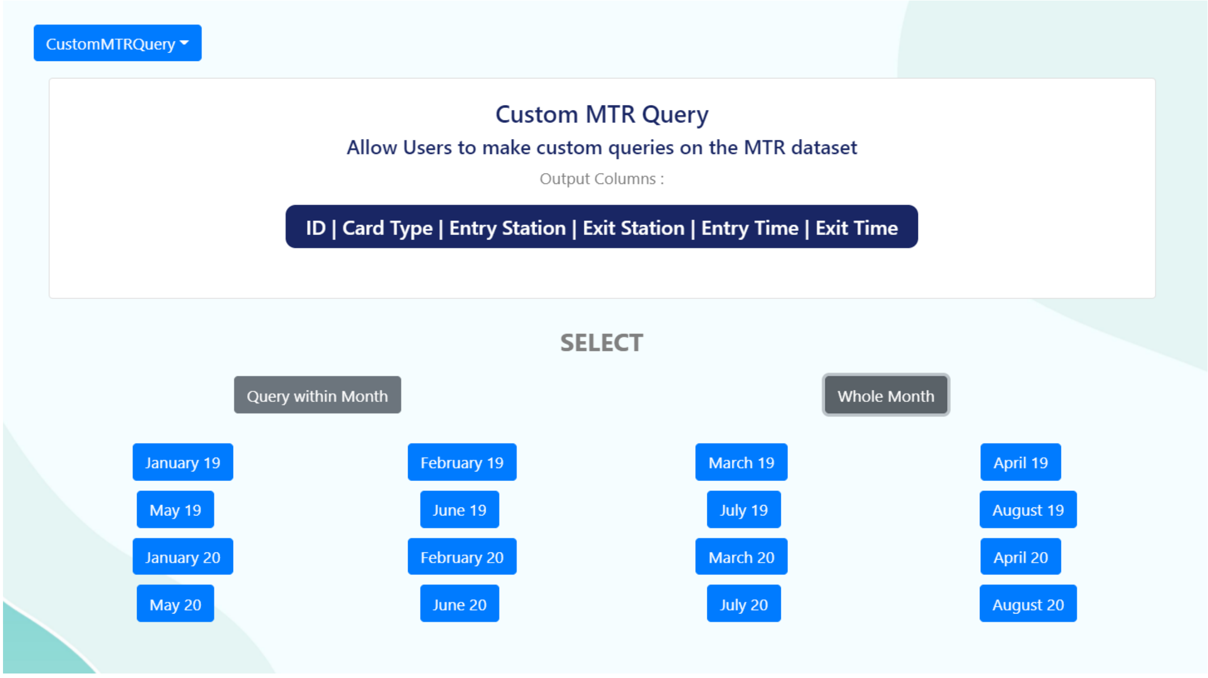
Figure 9 Home Page

Query Page

The query page allows the users to download data based on the given metric choices. Users can choose amongst available MTR queries which includes Custom MTR query, Passenger Volume, Station Density, Travel Pattern queries. Query page also has the option to fetch preprocessed COVID19 dataset.

Custom MTR query

The custom MTR query allows users to filter MTR data for specific dates, card types, and stations. The feature also allows for whole month dataset querying (getting the whole dataset for a particular month)



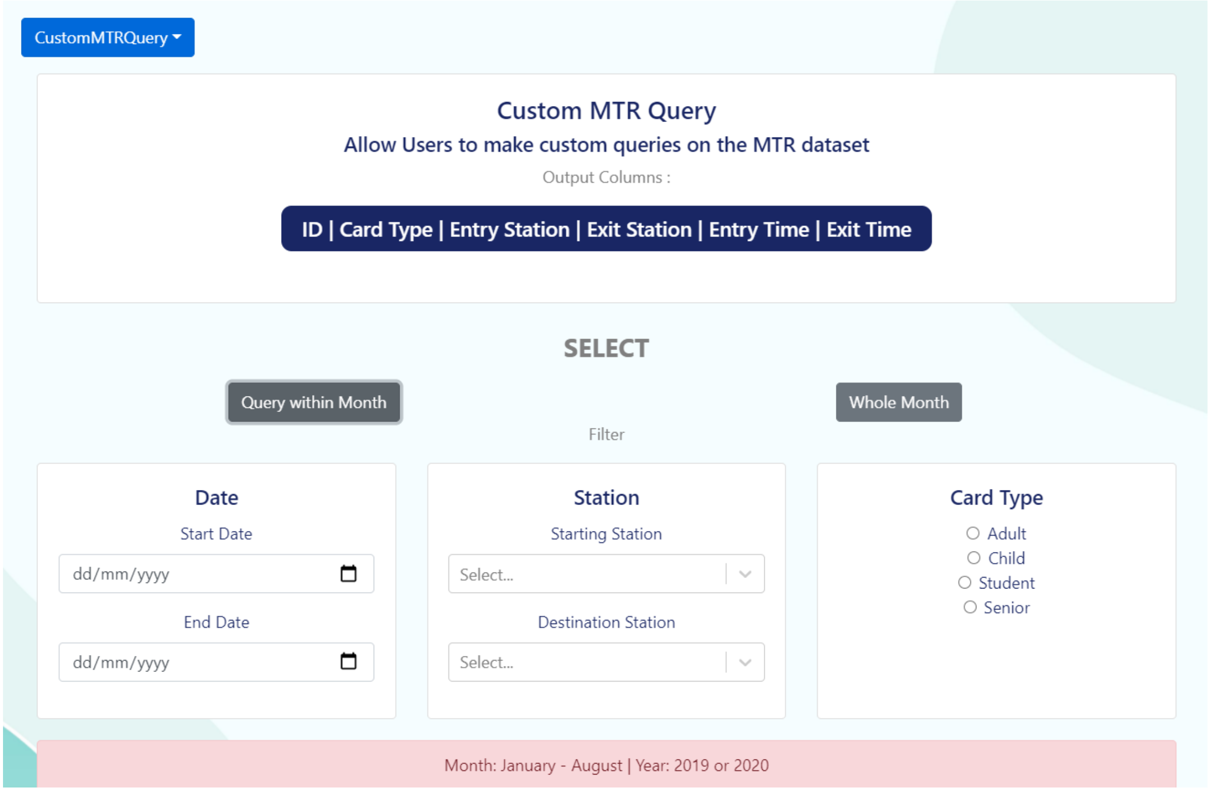
Custom MTR Query

Allow Users to make custom queries on the MTR dataset

Output Columns :

ID | Card Type | Entry Station | Exit Station | Entry Time | Exit Time

Figure 10.1 Whole month query



Custom MTR Query

Allow Users to make custom queries on the MTR dataset

Output Columns :

ID | Card Type | Entry Station | Exit Station | Entry Time | Exit Time

Date

Start Date

End Date

Station

Starting Station

Destination Station

Card Type

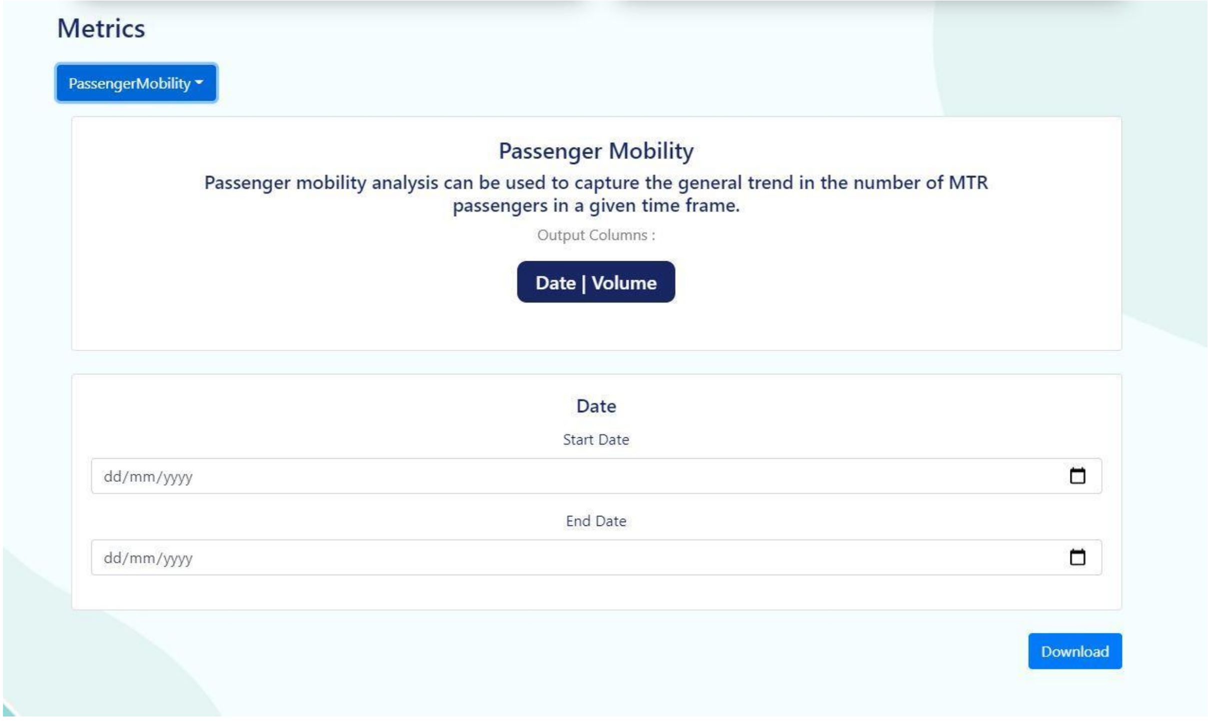
○ Adult
○ Child
○ Student
○ Senior

Month: January - August | Year: 2019 or 2020

Figure 10.2 Custom MTR query filtering

Passenger Volume/Mobility Query

The passenger mobility query page fetches passenger volume aggregated across all card types. The passenger mobility query fetches passenger volume grouped according to card type (Adult, Senior, Child, Student).



Passenger Mobility

Passenger mobility analysis can be used to capture the general trend in the number of MTR passengers in a given time frame.

Output Columns :

Date

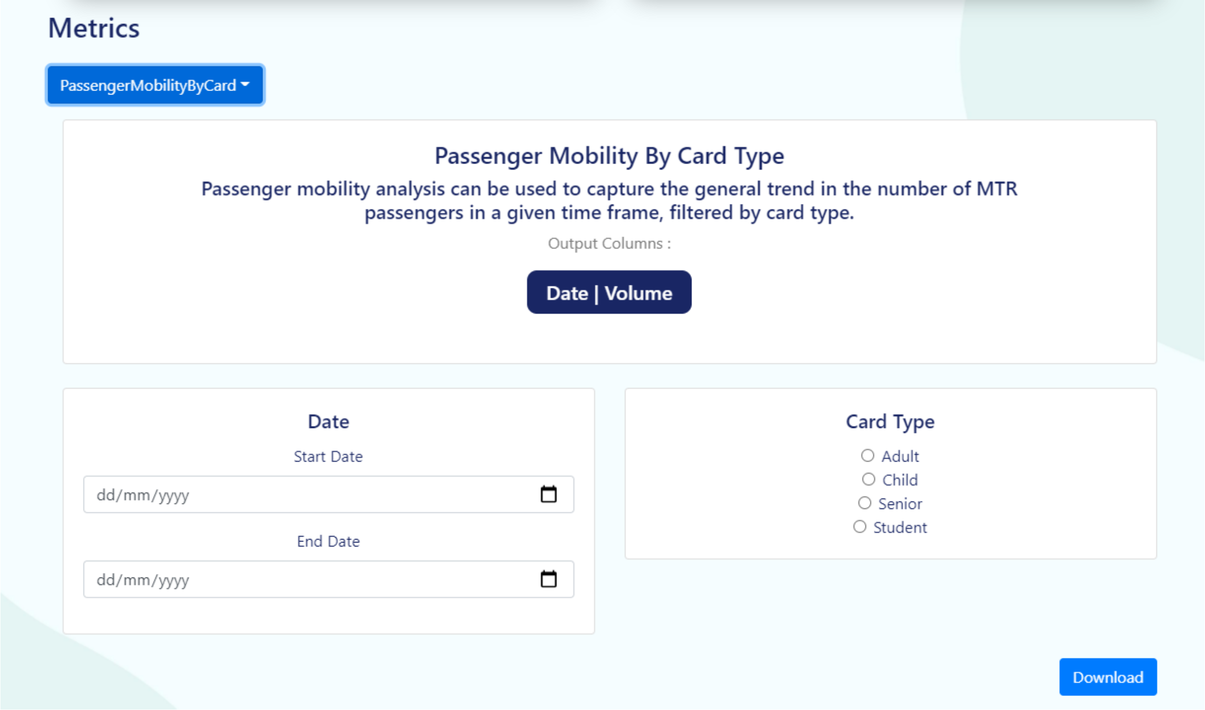
Start Date

dd/mm/yyyy

End Date

dd/mm/yyyy

Figure 11.1 Passenger Volume Query



Passenger Mobility By Card Type

Passenger mobility analysis can be used to capture the general trend in the number of MTR passengers in a given time frame, filtered by card type.

Output Columns :

Date

Start Date

dd/mm/yyyy

End Date

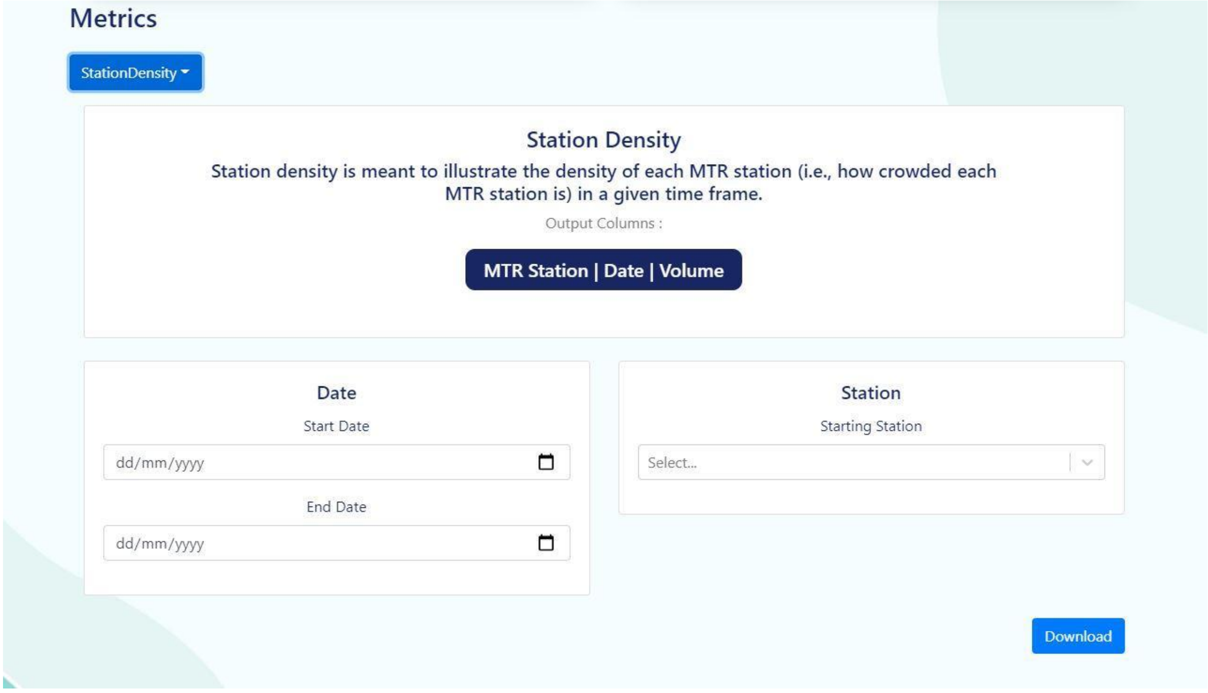
dd/mm/yyyy

Card Type

○ Adult
○ Child
○ Senior
○ Student

Figure 11.1.1 Passenger Volume Query by Card Type

Station Density Query



Metrics

Station Density

Station density is meant to illustrate the density of each MTR station (i.e., how crowded each MTR station is) in a given time frame.

Output Columns :

Date

Start Date

dd/mm/yyyy

End Date

dd/mm/yyyy

Station

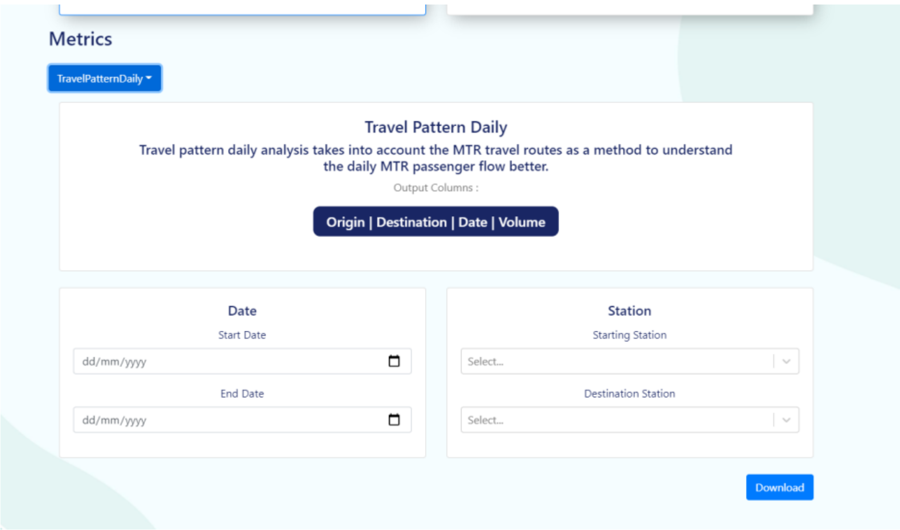
Starting Station

Select...

Figure 11.2 Station Density Query

Travel Pattern Query

Users have the option to query for travel pattern by daily volume or hourly volume. With the hourly volume user get to have a better picture of intra-day passenger movements, while with the daily volume the user will have insights in a much macro scale.



Metrics

Travel Pattern Daily

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

Output Columns :

Date

Start Date

dd/mm/yyyy

End Date

dd/mm/yyyy

Station

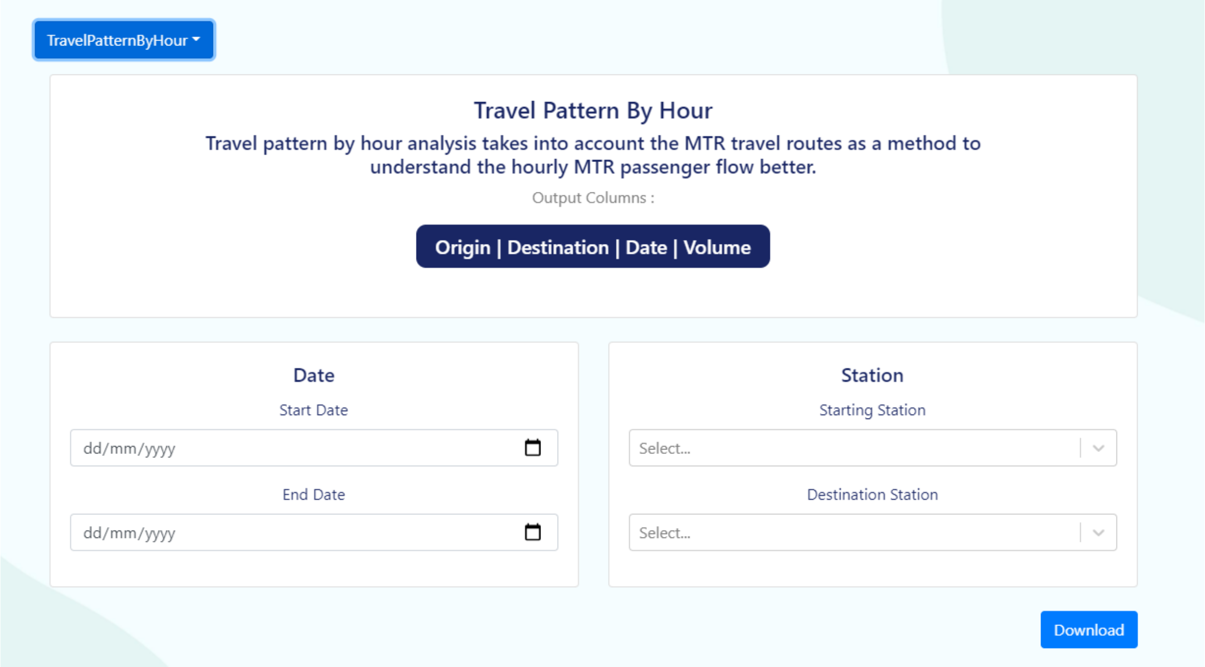
Starting Station

Select...

Destination Station

Select...

Figure 11.3 Travel Pattern Query



Travel Pattern By Hour

Travel pattern by hour analysis takes into account the MTR travel routes as a method to understand the hourly MTR passenger flow better.

Output Columns :

Origin | Destination | Date | Volume

Date

Start Date: dd/mm/yyyy

End Date: dd/mm/yyyy

Station

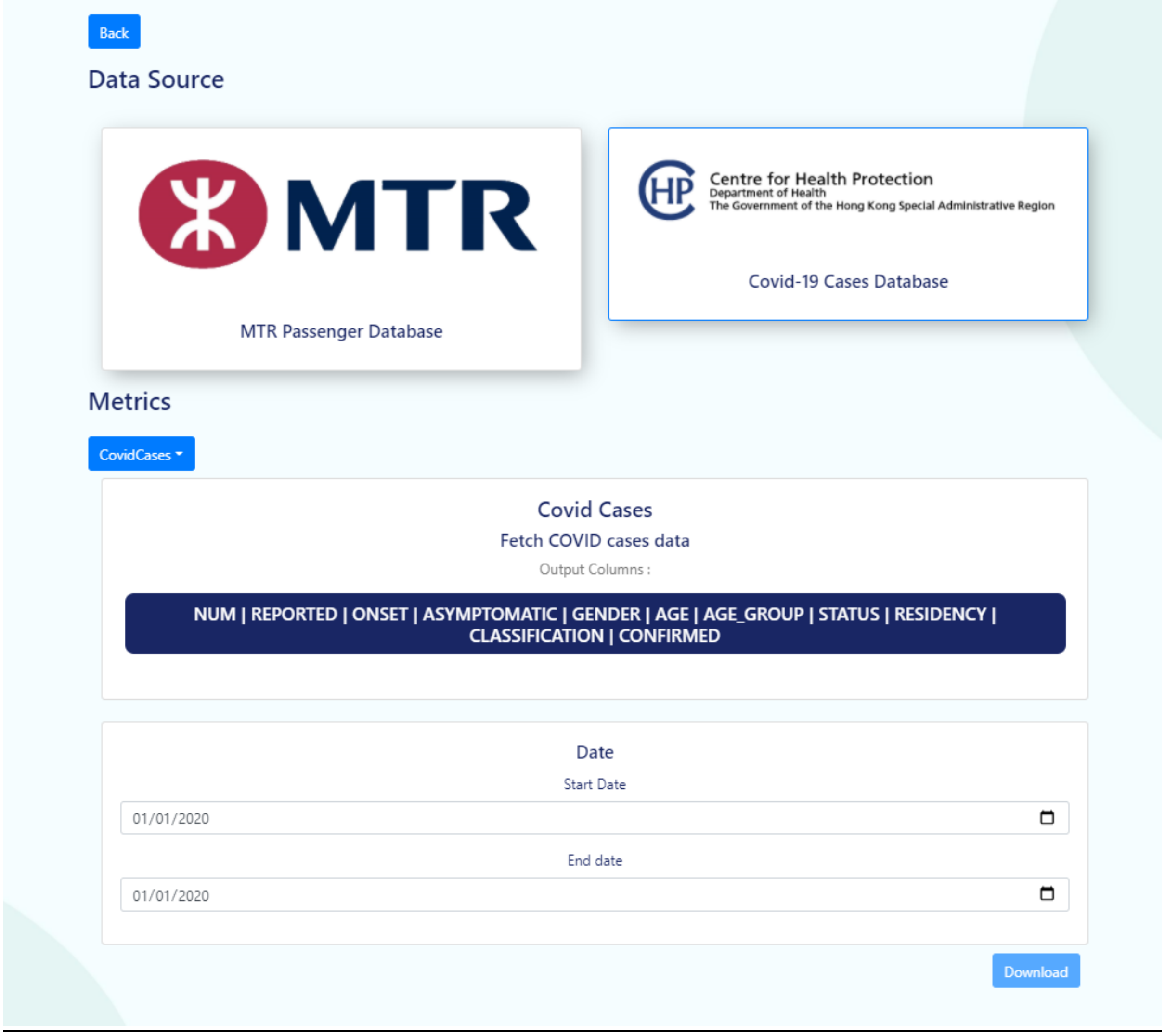
Starting Station: Select...

Destination Station: Select...

Figure 11.3.2 Travel Pattern by Hour Query

COVID19 Query

This query section allows the user to select a date range for fetching the pre-processed COVID19 dataset.


Back
MTR
Centre for Health Protection
Department of Health
The Government of the Hong Kong Special Administrative Region
CovidCases
Download

Data Source

MTR Passenger Database

Covid-19 Cases Database

Metrics

Covid Cases

Fetch COVID cases data

Output Columns :

NUM | REPORTED | ONSET | ASYMPTOMATIC | GENDER | AGE | AGE_GROUP | STATUS | RESIDENCY | CLASSIFICATION | CONFIRMED

Date

Start Date

01/01/2020

End date

01/01/2020

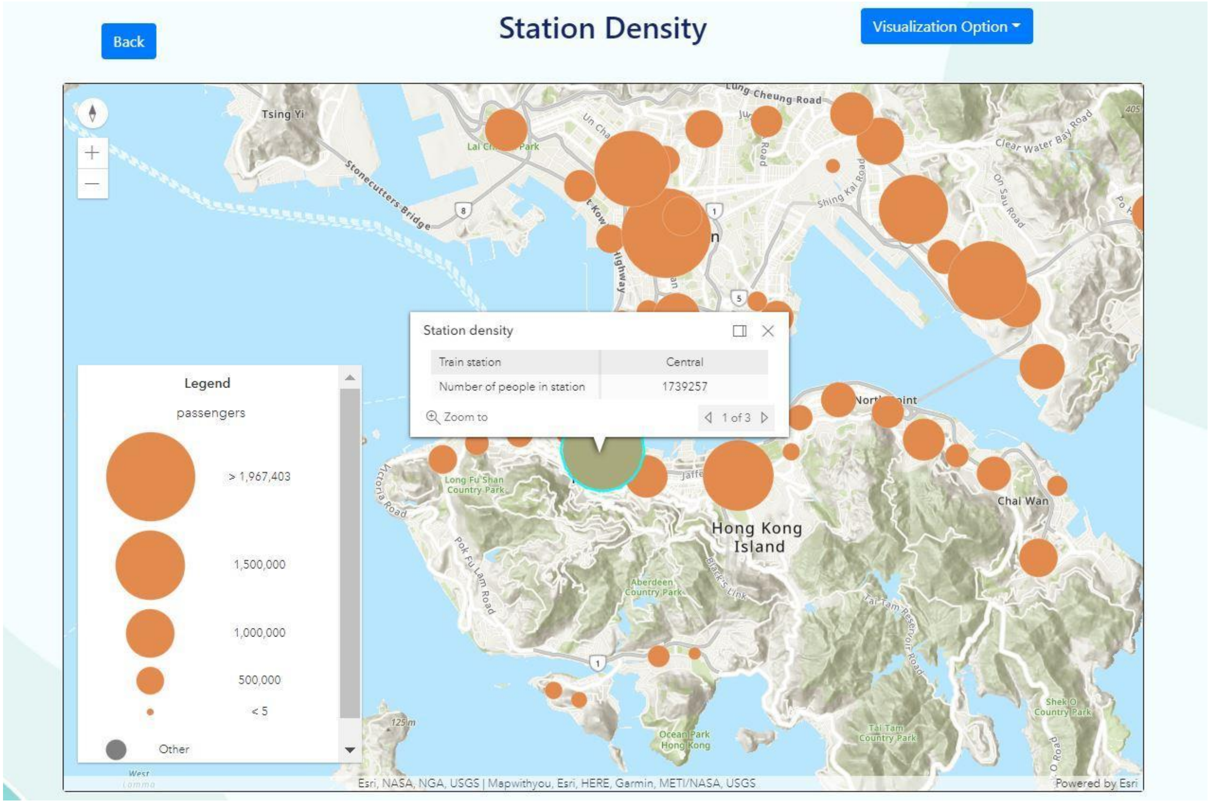
Figure 11.4 Covid19 Query

Visualization Page

The visualization page contains the dropdown ‘Visualization option’ menu which allows you to pick either one of the available visualizations.

Station Density Visualization

The user can use the form to select specific date ranges to observe the station density trend of each MTR station. A submission form is available for the user to select this date range option. The user then can use the range slider below the form to re-render the number of stations that want to be displayed. The map plotted is interactive and therefore the users can click on the figure to show a pop-up of the specific details of the station such as **station name** and **number of passengers**. The size of the circle is directly proportional to the volume of passengers in the station. With the given station density visualization, the user may be able to deduce levels of crowdedness of each station from this given information. The form also has a checkbox for selecting the option which allows users to overlay the visualization with COVID19 geolocation data, this allows the user to have a more holistic view of the events that unfold.

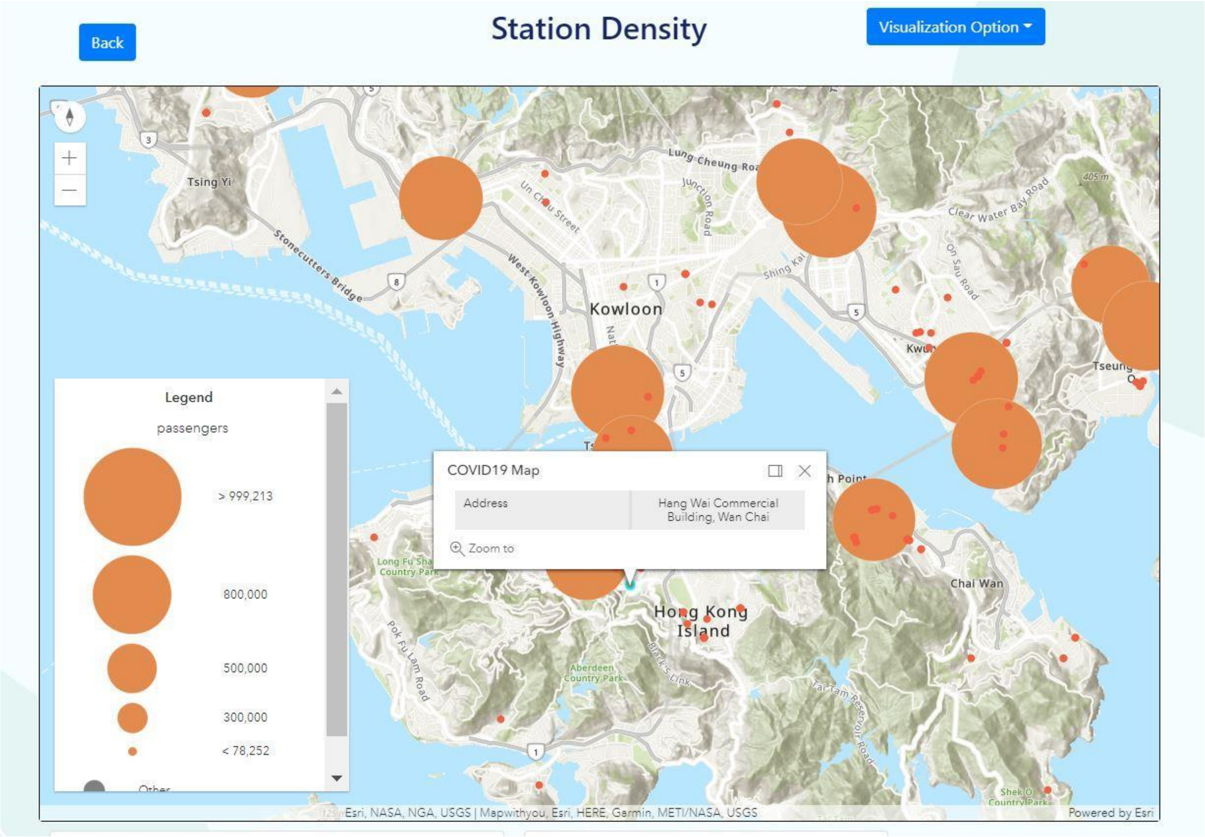


Station density

Train station	Central
Number of people in station	1739257

Zoom to 1 of 3

Figure 12.1 Station Density Visualization

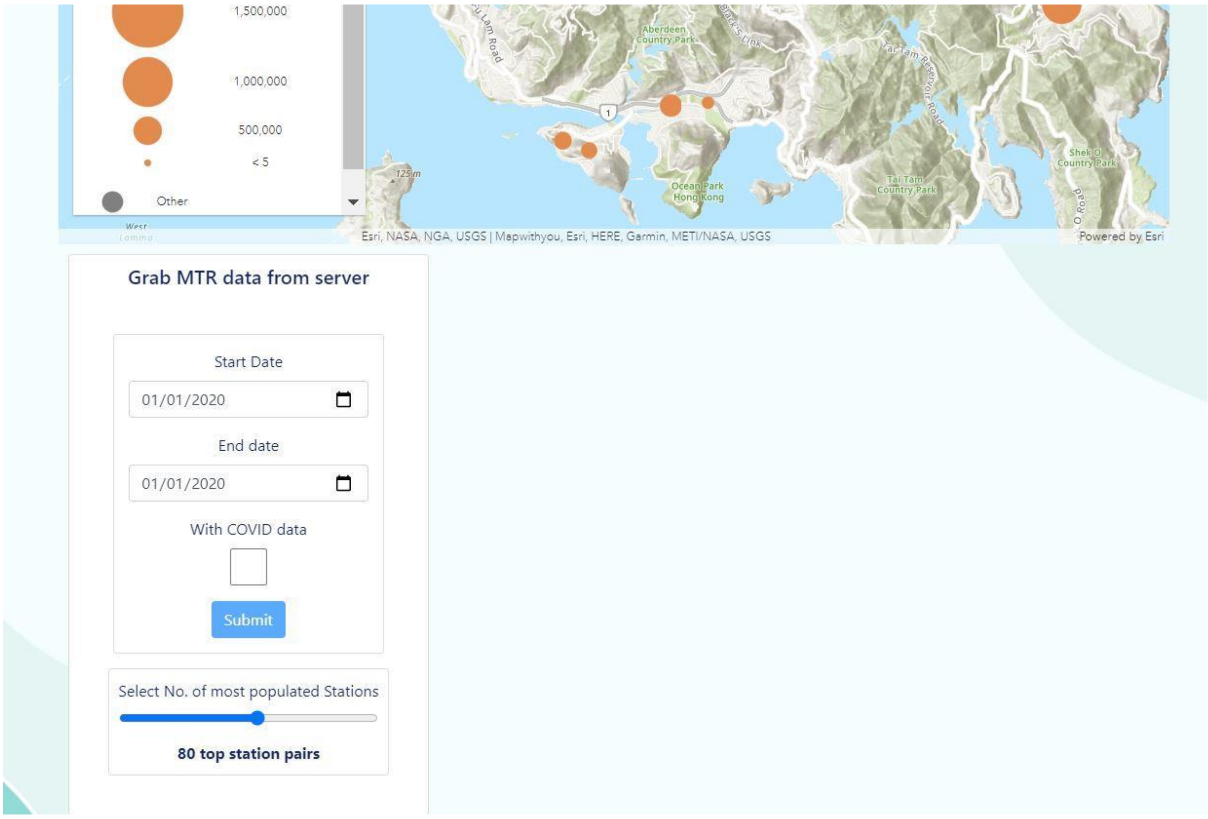


COVID19 Map

Address	Hang Wai Commercial Building, Wan Chai

Zoom to

Figure 12.2 Station Density with Covid Data



Grab MTR data from server

Start Date
01/01/2020

End date
01/01/2020

With COVID data
□

Select No. of most populated Stations
80 top station pairs

Figure 12.3 Form Input for Station Density Visualization

Travel Pattern Visualization

Functionality wise is similar to the station density feature. The user can use the form to select specific date ranges to observe the travel pattern trend of each MTR station pair. Again, users will have to submit the form for the visualizations to appear. The user then can use the range slider below the form to re-render the number of station pairs that want to be displayed, it will display the top station pairs with the highest count of travel pattern volume. Users can also interact and click on the figure to show a pop-up of the specific details of the station pair, showing details of passenger volume of that selected station pair. With this visualization, the user can investigate specific MTR station pairs and observe which MTR stations pairs are the busiest pairs. The user can repeatedly perform this visualization over different time periods and observe how the busy routes might change over time. The form also allows the user to opt for overlaying the visualization with COVID19 geolocation data. In figure 13.2, is a demo visualization for the month of March in 2020 overlayed with COVID19 geolocation data.

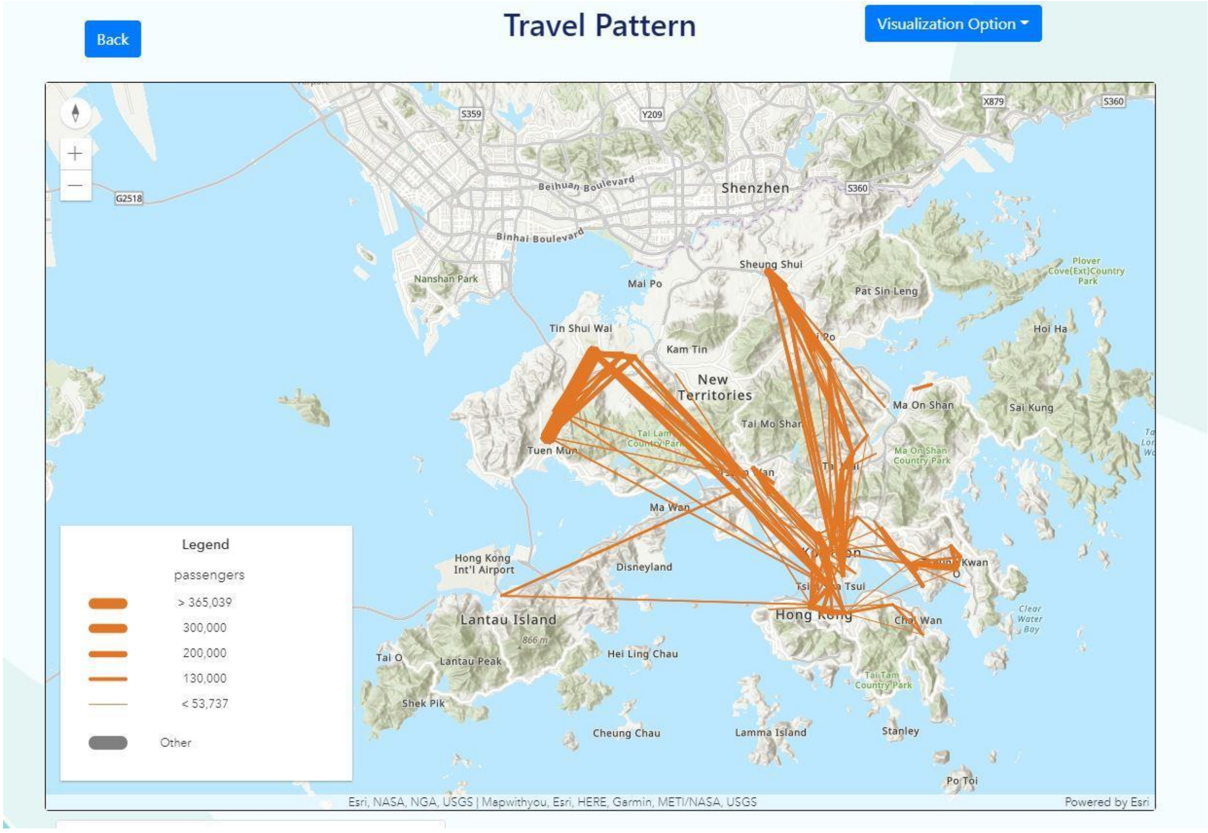
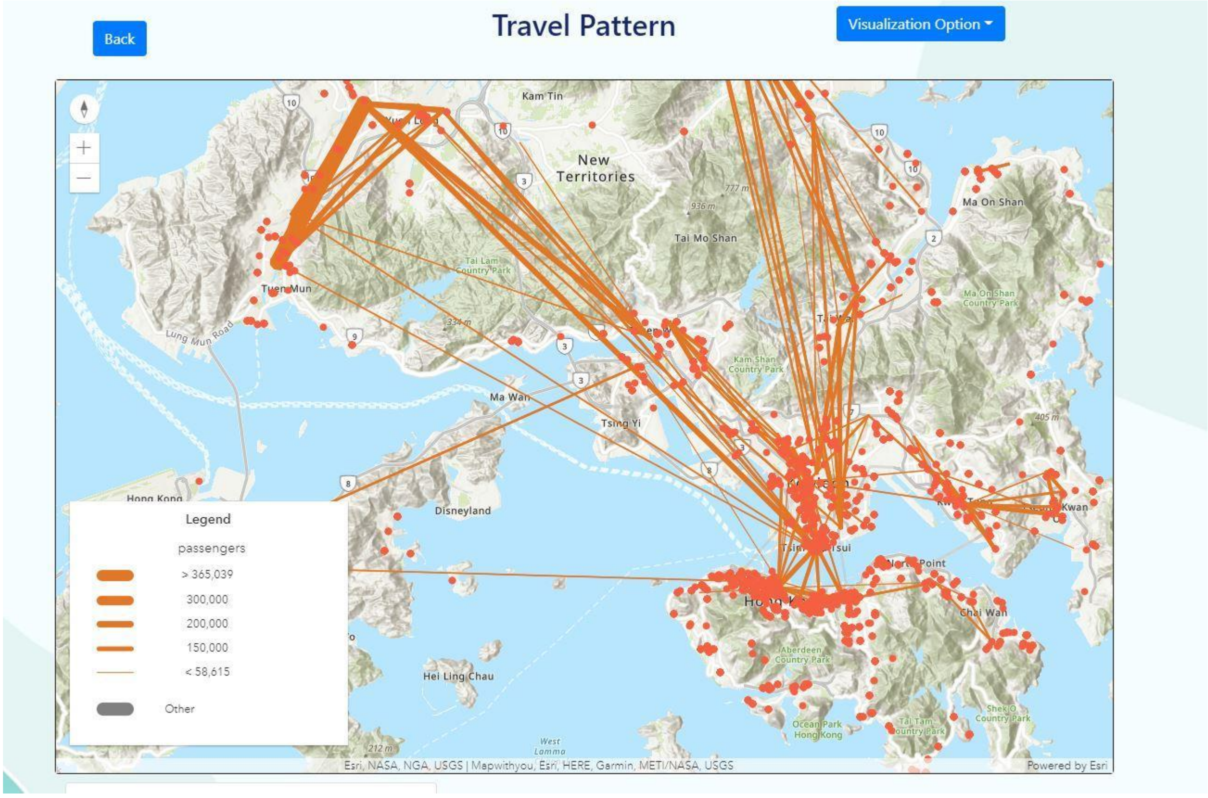


Figure 13.1 Travel Pattern Visualization

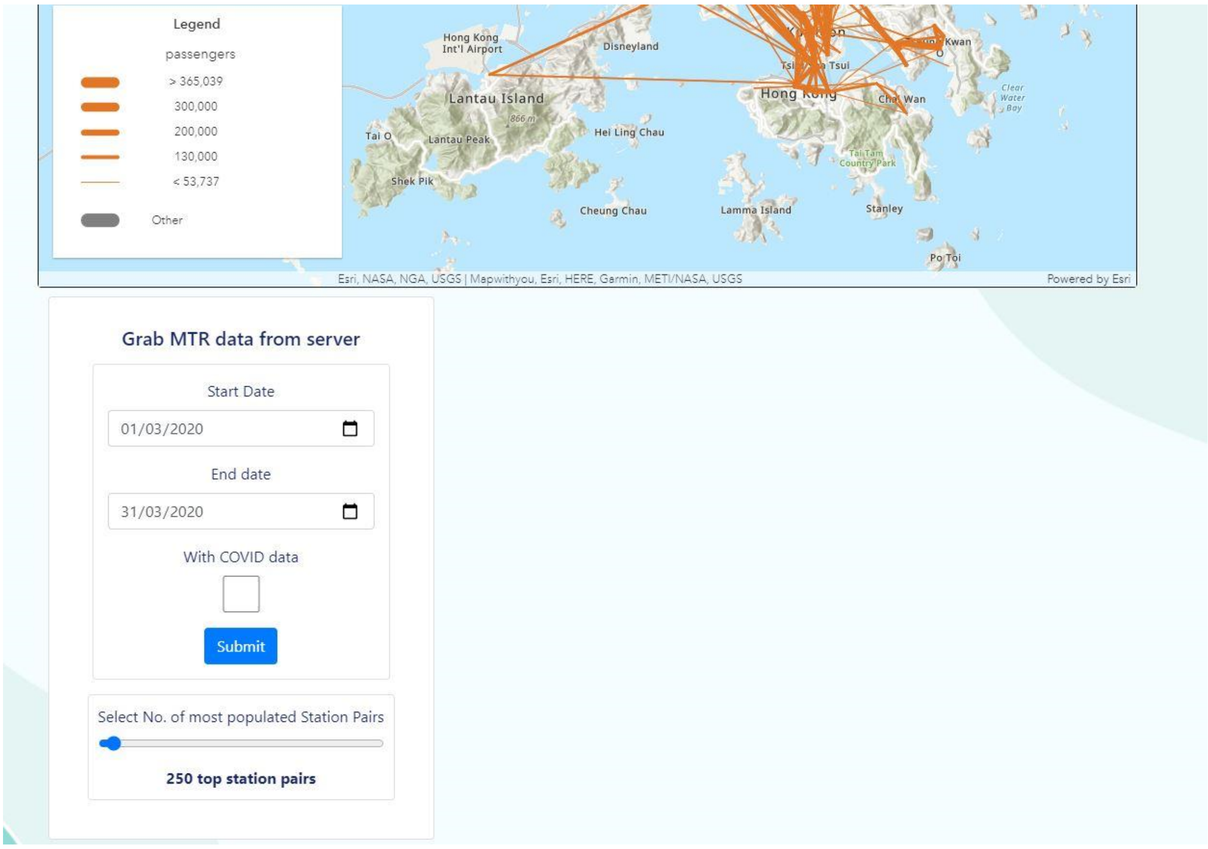


Legend

passengers

- > 365,039
- 300,000
- 200,000
- 150,000
- < 58,615
- Other

Figure 13.2 Travel Pattern with Covid19 data. March 2020



Legend

passengers

- > 365,039
- 300,000
- 200,000
- 130,000
- < 53,737
- Other

Grab MTR data from server

Start Date

01/03/2020

End date

31/03/2020

With COVID data

□

Submit

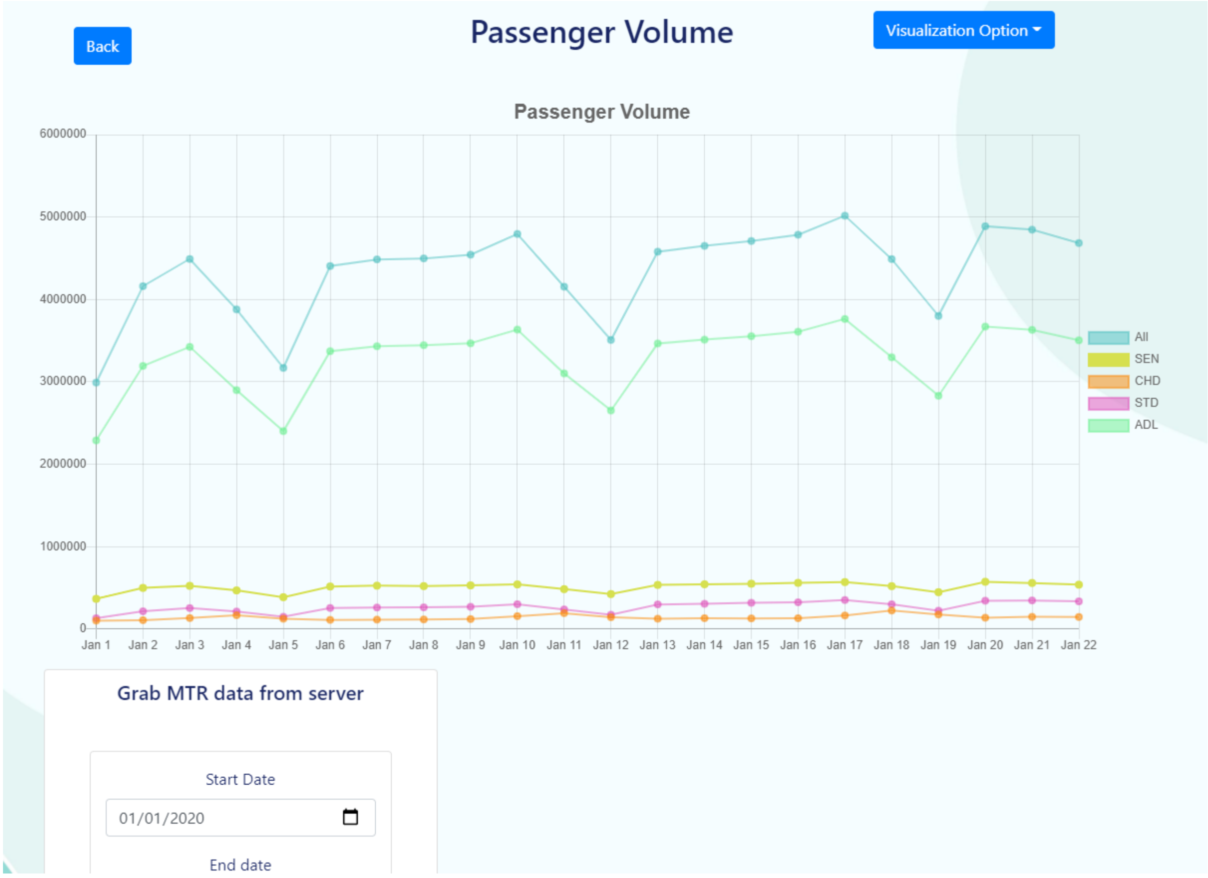
Select No. of most populated Station Pairs

250 top station pairs

Figure 13.3 Form Input for Travel Pattern

Passenger Volume Visualization

Users may use the form submission and select a date range they would like to observe. The result is an interactive Time Series graph of passenger volume over time. The user can use this graph to easily observe changes in passenger volume activity over a specified amount of time. There is an extra option to select the passenger volume separated by card types (Adult, Senior, Child, Student).



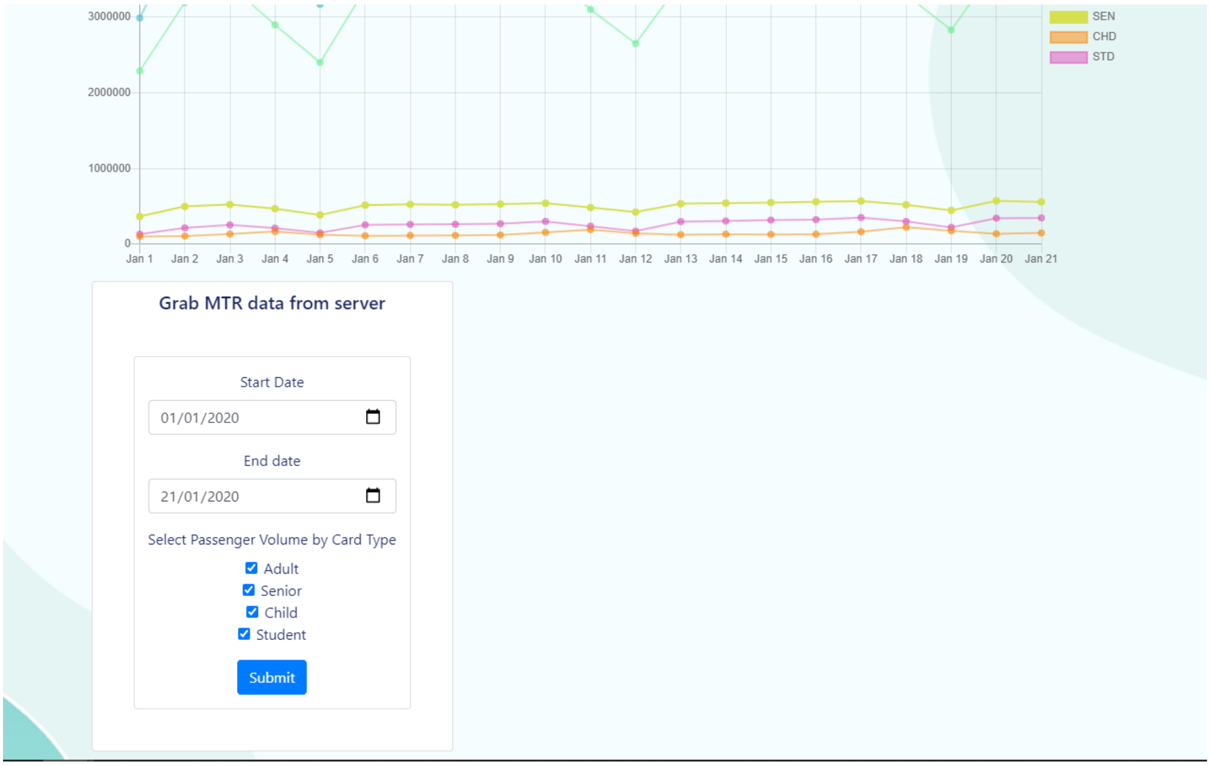
Grab MTR data from server

Start Date

01/01/2020

End date

Figure 14.1 Passenger Volume Visualization



Grab MTR data from server

Start Date

01/01/2020

End date

21/01/2020

Select Passenger Volume by Card Type

☑ Adult
☑ Senior
☑ Child
☑ Student

Figure 14.2 Form Submission for Passenger Volume

Analysis Page

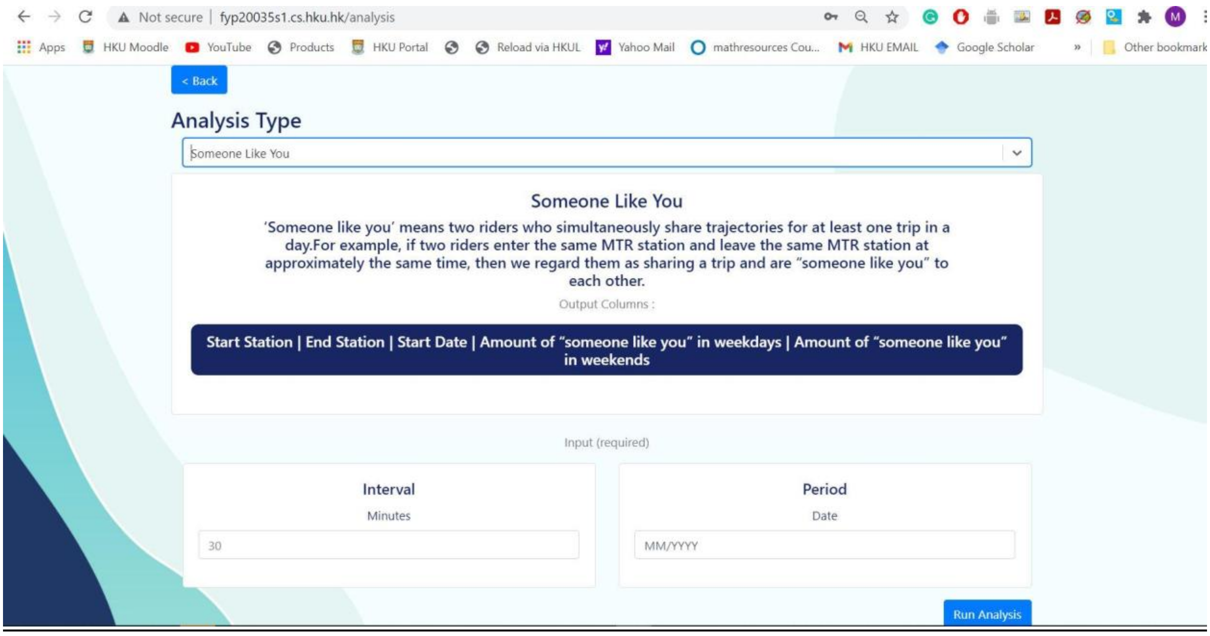
The analysis page allows the user to select from the dropdown component for either the “Someone Like You” or “Sensor Individuals” analysis.

Someone Like You

The input for the “Someone Like You” analysis requires two parameters:

First is the time interval in minutes, this is the time interval that we split for the whole day as explained in the methodology section. In which we consider people entering the same station who has entry time within the same time interval are considered to be “someone like you”. Users can opt to change the time interval into shorter or longer time spans according to research requirements.

Second parameter is the month of the year that we would like to analyse. The result of the analysis is the “someone like you” results for weekdays and weekends for each station pairs in the MTR.



Analysis Type

Someone Like You

Someone Like You

'Someone like you' means two riders who simultaneously share trajectories for at least one trip in 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 sharing a trip and are "someone like you" 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: MM/YYYY

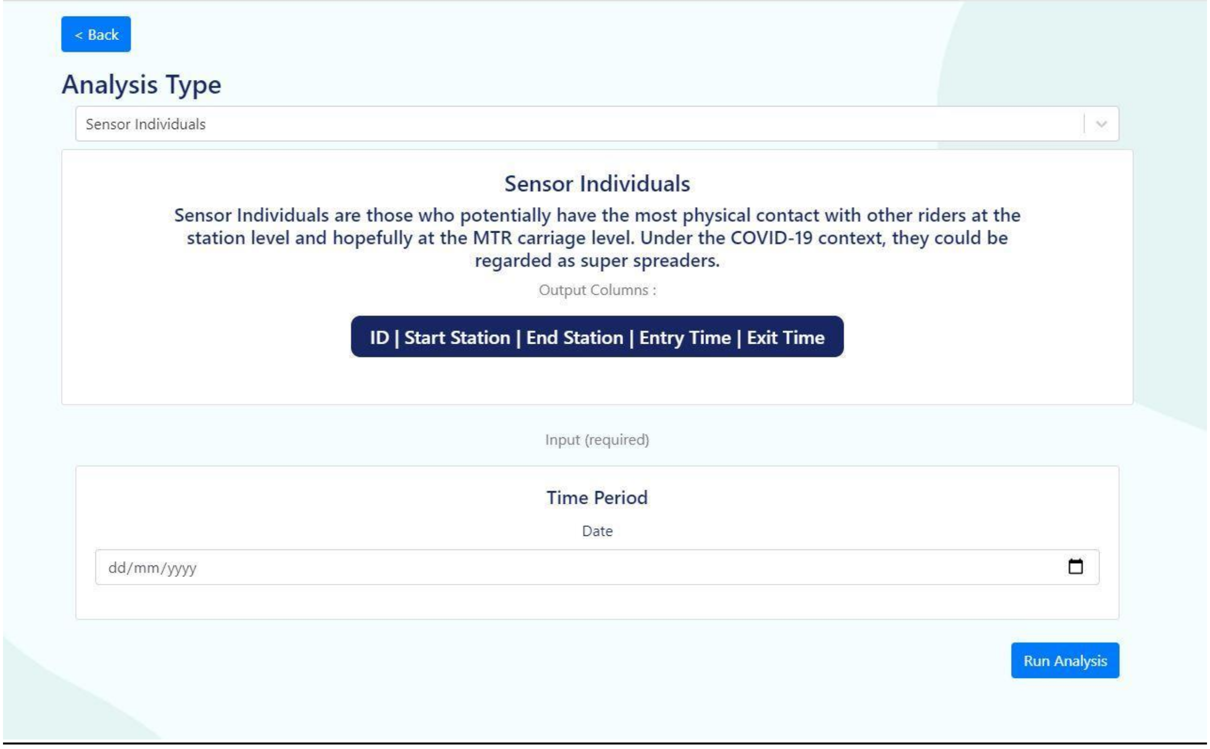
Figure 15 Someone Like You

This is a sample row from the result. The date column shows the starting date of that particular week. The rows below state that on the weekend of 1/3/2020 from station 1 to station 34 there is 4.5 the count of "Someone Like You" (FREQ_SLU), and on the weekdays of the same week there are on average 8.8 "Someone Like You".

Date	Weekend	Start_Stn	End_Stn	Freq_SLU
1/3/2020	TRUE	1	34	4.5
1/3/2020	FALSE	1	34	8.8

Sensor Individuals

The input parameter for the "Sensor Individual" analysis is a date. This will return the full journey of the passengers as modelled through the simulated MTR network (as defined in the "Sensor Individual" methodology section) in a particular day.



Sensor Individuals

Sensor Individuals are those who potentially have the most physical contact with other riders at the station level and hopefully at the MTR carriage level. Under the COVID-19 context, they could be regarded as super spreaders.

Output Columns :

ID | Start Station | End Station | Entry Time | Exit Time

Figure 16 Sensor Individuals

4. Limitations & Future Improvements

Challenges

The tasks of analysing the COVID19 and MTR dataset requires presumed knowledge and expertise on the field of epidemiology, transportation, and urban environments. Thus, analysing the dataset to in-depth length is a challenge for our team because of our background from Computer Science, which is an unrelated field. Therefore, our team's objective is to focus our efforts towards the software development aspect of this project and to improve functionalities upon requests from the academic researchers.

Limitations

Nature of Dataset

Our team was faced with a huge challenge for analysing the MTR dataset. One issue is that the nature of the dataset is quite difficult to work with in the sense that we do not have complete data points for a holistic analysis. For example, we cannot pin-point the exact geographic coordinates of where an individual went to. Our MTR data comprises only of starting and exit MTR stations. The best we could do is to estimate the journey for each passenger through intermediary stations such as in “Sensor Individuals” task. In reality, a person could be walking around in a particular MTR station and have complex interactions with the world, this of which our data cannot capture. Therefore, using our data could only be a rough estimate of general trends in the MTR system and almost impossible to accurately justify more personal interactions.

Confidentiality and Privacy

The signed agreement states that the MTR data is highly confidential and should not be shared to the public. This implies that we may not be able to use public cloud services such as Google Cloud Platform, Amazon Web Services to perform our data analysis or storing data using their services. The implication is that we will not be able to use state of the art cloud platforms to perform our tasks. Some advantages of using the public cloud services are that the application will be easily scalable, robust, and have much better performance than in-house servers. We may also be missing out from using their already available data science, machine learning tools to improve our results. Our approach will be to use open-source solutions that allows us to keep the data private by running the software privately.

Lack of Computing Power

Advanced analysis such as “Someone Like You” and “Sensor Individuals” take a considerable amount of runtime to perform. These tasks are computationally intensive and with the current server’s computational power, such a task is almost impractical to have fast results and performance.

Future Improvements

Real Time Platform

Currently the system works on historical data from previous months and therefore could only be used as retrospection of events that had already happened. If we were to make a fully dynamic application with the ability for real time alerts, we would need to make an application which has a steady stream pipeline of data from the MTRC and HKCHP. Consequently, this will require further and deeper collaboration with the MTRC and would also require more technical knowledge and computing power for such a data intensive and high performant application. Such a collaboration would be a huge project and would be involving multiple stakeholders (HKU, MTRC, HKCHP).

Integration with Contact Tracing Application

To improve the application and real-world use-case of our dataset, we can share our analysis results by future integration with Contract Tracing application. We can set up an API for getting “Sensor Individuals” / “Someone Like You” data which could be used to provide alerts to the contact tracing app. For example, a MTR route can be busy with many “Someone like you” individuals and therefore may be good to give a small precaution notification to the app. Integration with Contact Tracing application will help us to cover the data points that could not be captured by the MTR data (like previously mentioned) and help to give us a more holistic understanding of the movement of the population.

5. Conclusion

This paper had presented the background knowledge, statistics and insights on the COVID19 and MTR transactions data. Elaborate detail on the dataset was discussed as well as the pre-processing methods undertaken to prepare the data for storage and accessibility. Furthermore, the paper discussed about the methods and steps needed to perform the various trend analysis utilized from the given data. We have analysed through visual data the trends of local transport during the pandemic in Hong Kong and examined the correlation between MTR movements and COVID19.

Our team had also built a secure platform-independent web application as a one stop solution for academic researchers and professors for data retrieval, data visualization and data analysis. We went through the different features and use-cases of the platform and guide users on how to use them.

We hope that our platform will continue in aiding academic researchers on this topic and hope that our contribution can help in preventing and fighting against future pandemic outbreaks.

References

[1] (2020). Retrieved 27 October 2020, from https://www.legco.gov.hk/research-publications/english/1718issh07-mtr-train-service-performance-20171220-e.pdf

[2] Zhang, F., Jin, B., Ge, T., Ji, Q., & Cui, Y. (2016). Who are My Familiar Strangers? *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. doi:10.1145/2983323.2983804

[3,4] Sun, L., Axhausen, K., Lee, D., & Huang, X. (2013). Understanding metropolitan patterns of daily encounters. *Proceedings Of The National Academy Of Sciences*, *110*(34), 13774-13779. doi: 10.1073/pnas.1306440110

[5] Salathe, M., Kazandjieva, M., Lee, J., Levis, P., Feldman, M. and Jones, J., 2010. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51), pp.22020-22025.

[6] Developers.arcgis.com. (2021). https://developers.arcgis.com/javascript/latest/api-reference