COMP 4801 – Final Individual Report

Final Year Project

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

Supervisors:

Professor Reynold Cheng

Shivansh Mittal

Group Members:

Ali, Marvin (3035361817) – Author

Effendi, Janice Meita (3035492977)

Jain, Rishabh (3035453608)

Nagra, Harsh (3035437707)

Widjaja, Marco Brian (3035493024)

Abstract

Hong Kong is one of the places in the world that were impacted by the COVID-19 pandemic. To suppress the spread of the disease, the Government of Hong Kong Special Administrative Region enforce several regulations including social distancing measures. As a result of these measures, the mobility of Hong Kong people was reduced dramatically, which can be observed through the decrease of MTR ridership. The purpose of this project is to create a platform where users can run queries of both MTR passenger data and COVID-19 cases, generate visualization, and run advanced contact-based research in order to see the impact of COVID-19 on the change of MTR patronage. Observing the change in MTR patronage can be a representation of change in the overall mobility of Hong Kong people as MTR is the main mode of transportation in Hong Kong. We aim to make a platform that can be used by government officials and academist to generate a meaningful visualization which can aid their research or regulations planning conveniently. Our findings found that there is a correlation between the volume of passengers and the emergence of COVID-19 cases in a particular area. Besides, we also observed that the measure taken by the government was effective to suppress the mobility of Hong Kong people as there is a huge decrease in patronage after several social distancing measures have been enforced.

Acknowledgement

We would like to express our gratitude to our supervisors Prof. Reynold Cheng from The Department of Computer Science, The University of Hong Kong for allowing us to participate in this project. We thankful to our Research Assistant Shivansh Mittal from the Department of Computer Science, The University of Hong Kong for their guidance to us on doing this project. We are grateful for the MTR Corporation for providing us with the data that we are using for the analysis. We also would extend our gratitude towards Esri Corporation for the ArcGIS pro software license sponsorship.

Table of Contents

List of Figures

List of Tables

Abbreviations

MTRC	Mass Transit Railway Corporation
ESRI	Environmental Systems Research Institute
COVID-19	Coronavirus disease 2019
HKCHP	Hong Kong Centre for Health Protection
RDBMS	Relational database management system
SQL	Structured Query Language
SSH	Secure Shell
HKU	The University of Hong Kong
ORM	Object Relation Mapping

1. Introduction

1.1.Background

The outbreak of COVID-19, which is caused by the SARS-CoV-2 virus, was first identified in Wuhan, China on 31st December 2019 (World Health Organization, 2020). On 23rd January 2020, Hong Kong confirmed the emergence of the first COVID-19 cases. As the virus is highly contagious, the government of Hong Kong SAR has enforced social distancing measures, such as limiting the number of people in a social gathering to suppress the spread of the virus. As a result, there was a drop in the mobility of Hong Kong people after this measure was enforced. One of the indicators is the change in the number of public transportation trips taken by Hong Kong residents.

In 2019, 47.4% of people in Hong Kong used the Mass Transit Railway (MTR) as their primary mode of transportation (MTR Corporation Limited, 2020). Thus, by analyzing the changes in the mobility of MTR passenger, we will be able to get the representation of the change of mobility of the Hong Kong population after the social distancing measures has been enforced. The University of Hong Kong has signed an agreement with the MTR Corporation to collaborate to develop visualization and data analysis of the MTR traffic.

1.2.Familiar Stranger

Familiar Stranger" refers to a stranger that someone recognizes due to regular meetings in a commonplace such as MTR station, yet both of these people never have any interaction. The concept of "Familiar Stranger" had been around for a while, yet it was hard to identify due to the lack of technology such as geolocation to support it. Previous research identified the phenomenon through qualitative measures such as survey and personal anecdotes (Zhang et. al., 2016).

There was research regarding the phenomenon of "Familiar Stranger," (Zhou et. al., 2020) which related to our analysis. Zhou et. al. (2020) has conducted in Beijing to determine the correlation between familiar stranger phenomena and human activity patterns. Zhou pointed out that there is a high positive correlation between the occurrence of "familiar stranger" and the number of smartcard riders in a given time frame. Due to the similarity between the Hong Kong and Beijing mass transportation system, there is a high chance that the research can also be replicated in Hong Kong to study the MTR passenger behavior. In addition, Zhou et al. (2020) also believes that in the public health field, identifying the 'familiar stranger' phenomenon can be used to track transmission of infection disease. Thus, in this project we will identify the "familiar stranger" phenomenon help us to identify any possible transmission of COVID-19.

1.3.Project Aims

This project aims to provide a platform that able to create a detailed visualization and analysis of the impact of the COVID-19 pandemic on the change of MTR passenger traffic. The target users of this project are divided into two different categories. The first category consists of users from the University of Hong Kong such as researchers and professors that have been authorized to access the MTR passenger data. The second category consists of government officials such as the Transport Department of Hong Kong and the Health Department of Hong Kong. Users in both categories will be required to sign a confidentiality agreement provided by the MTRC. The end goal of this system is to help these users to understand more about the correlation between the COVID-19 pandemic and the mobility of people by providing access to different analytical information and visualization without any sensitive information. This eventually will enable the user to understand the effectiveness of the current regulations to curb

the spread of COVID-19 and improve these regulations based on the feedback shown by the analysis that we have provided.

We also aimed to develop an analytical solution in the form of mobile or web applications to help academics and government officials replicate the analytical process in an efficient, effective, and convenient way.

1.4. Scope

The scope of this project can be divided into four different parts:

1. Develop a relational database management system to store and maintain both the MTR passenger data provided by the MTRC and the COVID-19 cases data
2. Conduct trends analysis by utilizing different technologies such as ArcGIS and Python
3. Develop a new system that enables users to receive the data more efficiently and able to replace the current data distribution system that still utilizes DVDs
4. Create a web application that works as a visualization tool

1.5.Significance and Impact

The system will enable both researchers and the government in Hong Kong SAR to conduct an in-depth analysis of the correlation between the spread of COVID-19 with the passenger mobility of the Hong Kong population overall through observing the MTR passenger mobility. As MTR is the most used public transportation mode in Hong Kong and one of the possible “Hotspot” for the COVID-19. Understanding the relationship between MTR passenger's mobility and the spread of the coronavirus will enable government officials to establish transport policies that able to reduce the suppress the spread of the COVID-19 while at the same time still maintaining the convenience of MTR services.

1.6.Outline of Report

This report is divided into 4 sections. Section 1 covered the brief information of both COVID-19 pandemic and changes in MTR passenger travel behavior, and the significance and impact of this project, background of the project, and existing research that is used to support this project. Section 2 will mainly be consisted of the methodologies of our project. In addition, justification of the decision to use a certain methodology and details of the implementation will also be covered in this section. Section 3 will cover the results obtained from this project. Section 5 will cover future implementation that will be taken and conclusion of this report

2. Methodologies

2.1.Overview

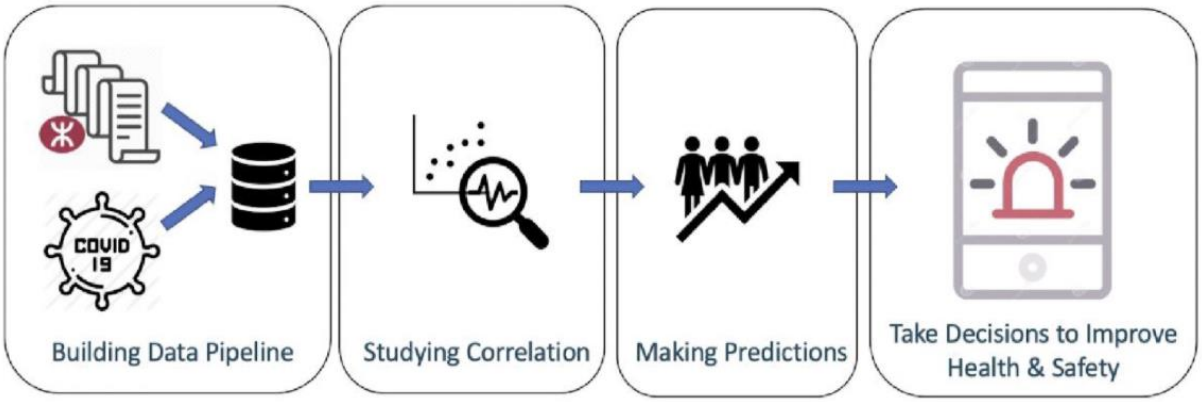

Figure 2.1. Project workflow

Figure 1 summarized the overall workflow of the project implementation. First, a data pipeline will be built to handle both the MTR passenger and COVID-19 data. Next, mobility trend analysis and geospatial analysis will be conducted to understand the correlation between the MTR passenger data and COVID-19 data. After this analysis has been completed, the result will be used to make some predictions on passenger behaviors and further compiled into a single application.

2.2.Data pipeline

2.2.1. Raw Data

Both MTR passenger data and COVID-19 cases data are needed to find the correlation between the spread of the COVID-19 disease and the mobility of Hong Kong people. The following subsections will cover the detailed information of MTR passenger raw data and COVID-19 cases raw data.

2.2.1.1.MTR Passenger Data

Through the agreement between The University of Hong Kong and MTR Corporation, passenger travel history from January 2019 until April 2020 can be retrieved for research purposes. The data was provided in CSV (comma-separated value) files. Data provided includes the ID of each card or ticket used in the transaction, entry points and exit points in the form of MTR station, timestamps of entry and exit, type of transaction (octopus or single ticket journey), and octopus type for the transaction using Octopus card.

2.2.1.2.COVID-19 Cases Data

The Department of Health of Hong Kong provided a database related to COVID-19 cases in Hong Kong, which can be accessed through https://data.gov.hk/en-data/dataset/hk-dh-chpsebcddr-novel-infectious-agent. From there, we decided to use the following database:

- Details of probable of confirmed COVID-19 cases in Hong Kong

- Buildings (residentials / non-residentials) in which probable/confirmed cases have resided in the past 14 days in Hong Kong

- Latest situation of reported cases of COVID19 in Hong Kong

2.2.2. Data Preprocessing

Preprocessing of the raw data is needed to standardize the format of the data within the database. Thus, data cleaning will be performed on both MTR passenger data and COVID-19 cases data. Data cleaning is the process of fixing or removing data that is incorrect, formatted wrongly, corrupted, or duplicated from a raw dataset (Tableau). Data cleaning can be done in several ways. Removing all the duplicated data from the dataset is one of the methods of data cleaning. Fixing structural error such as inconsistencies within data formatting is also one of the data cleaning methods. The following subsubsections will explain the data cleaning process of both MTR passenger raw data and COVID-19 cases raw data.

2.2.2.1.COVID-19 data

Our findings discovered that the data retrieved from the Department of Health were stored in an inconsistent format and had a lot of missing values. For example, the data field for "related case" in the building database, which displayed the case number related to a patient, was stored in several formats (e.g., "Case 14", "14", "Correlated to case 14"). In order to be able to perform a thorough analysis, we need to streamline all of this value. Thus, we perform string manipulation in R and Python with the pandas and NumPy library.

In addition, there are also several missing values in all the databases, predominantly the date column. As the percentage of missing data values was relatively high, especially in the building database, we conclude that it is not possible for us to delete the data. Thus, we perform an additional measure to find an appropriate date to fill the missing values. As the data was served in a single CSV for each day, we use the date where the database is published to fill the missing data values.

New fields were also added to the COVID-19 confirmed case data to facilitate better data analysis. These fields are “Asymptomatic” and “Age Group”. In some of our raw data, “Asymptomatic” are reported within the ‘Date of Onset’ field. The “Age Group” are added in order to enable an easier integration with the MTR passenger data. The age group classification will follow Octopus card’ age group classification (Child for cases with age 3 – 11, Adult for cases with age 12 – 64, and Elder for cases with age 65 or above) (Octopus).

2.2.2.2.MTR data

The purpose of data cleaning the MTRC data is to remove the data with missing values. In some of the transaction details data, we cannot find which exit or entry point of a passenger. As the frequency of this missing data was not a lot; removing them will not affect the integrity of the MTR data, we simply decided just to remove the data.

2.2.2.3.MTR data & COVID-19 data

As a geospatial visualization will be conducted for both the MTR passenger data & COVID-19 cases data. The address in both MTR & COVID-19 database needs to be converted into geolocation (in terms of longitude and latitude). This conversion is conducted by utilizing the geolocation feature in ArcGIS.

2.2.3. Database Modelling

To store the huge volume of the data (over 500 million MTR data and over 30 thousand COVID-19 data), we provision a database to store all the data. This database also enables users to retrieve the data concurrently in an efficient, effective, and convenient manner.

The database is stored in a form of a relational database as it enables us to perform queries to support our data mining process. This database is hosted on a private server owned by the Department of Computer Science of The University of Hong Kong. Both the MTR passenger data and the COVID-19 cases data are stored in a structured way. Thus, a Structured Query Language (SQL) type database is chosen to host this data. MySQL is a relational database management system (RDBMS) based on Structured Query Language (SQL). MySQL is used due to its open-source nature, enabling everyone in a team to connect to the database without paying additional fees.

Figure 2.2. illustrate all the tables within the MTR passenger database and COVID-19 cases database and the entity relationship between each table. Table 2.1 and Table 2.2 contain the explanation of the MTR tables and COVID-19 tables respectively.

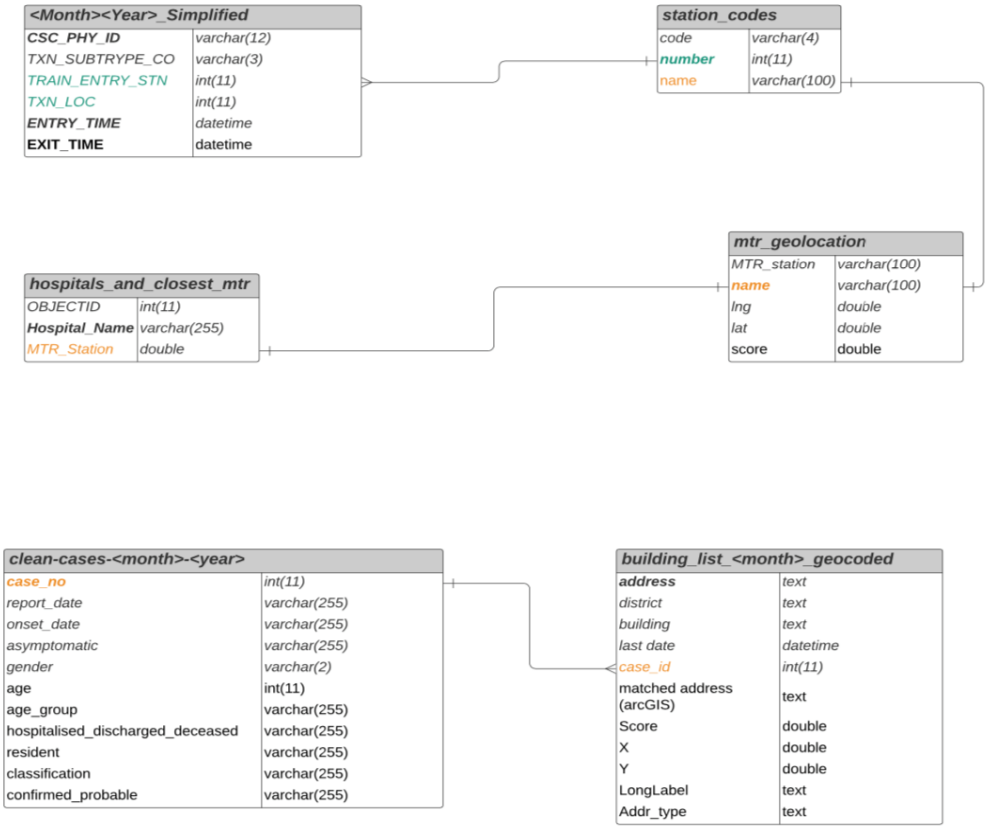

Figure 2.2. Database Entity Relationship Diagram

Table	Content
<Month><Year>_Simplified	MTR passenger transaction data; January – April 2020 and January – April 2019
Station_codes	Station codes as commissioned by the MTRC and its corresponding station name
Mtr_geolocation	Geolocation (longitude and latitude) of each MTR station
Hospitals_and_closest_mtr	Mapping of COVID-19 hospitals to closest MTR station

Table 2.1. Content of MTR Table

Table	Content
Clean-cases-<month>-<year>	List of COVID-19 cases in Hong Kong
Building_list_<month>_geocoded	Information of building that probable and confirmed COVID-19 cases has visited

Table 2.2. Content of COVID-19 Table

2.3.Mobility Trend Analysis

Mobility trend analysis is used to determine the impact of MTR passenger travel behavior on the change of several COVID-19 cases, and vice versa. This analysis is done with the aid of Python – utilizing several libraries such as Pandas, Matplotlib, and Plotly. To point out the mobility trend in general, the variation of MTR ridership from January to September 2020 is examined. In addition, and January to September 2019 will be used as a baseline to compare the impact of COVID-19 to the changes in passenger mobility. In order to take sociodemographic into considerations, changes in the passenger mobility trend in each Octopus card (i.e., infant, student, adult, and elderly) category are observed.

The analysis will also compare the differences between passenger travel behavior during the weekdays and weekends on a given hour of the day. The change of daily travel patterns will also be observed especially by comparing the travel patterns before and after a key date (i.e.

Chinese New Year Holidays, enforcement of Work From Home policy). The result will enable a better understanding of the impact of government policies in respect to curbing the COVID-19 spread to the change of human behavior in terms of passenger mobility

2.4.Geo-spatial analysis

Geo-spatial analysis is used to provide a visualization for certain factors that hard to understand in the data analysis stage. For example, the changes of both densities of passengers that enter or leaving each MTR station and travel density between two different MTR stations can be visualized better using Geo-spatial analysis. With respect to COVID-19 cases data, geo-spatial analysis can be used to point out the "hot-spots" area of COVID-19 confirmed cases in Hong Kong.

The result of geospatial visualization for both MTR passenger data and COVID-19 cases data will be combined in order to enable us to observe the relationship between the change sof MTR passengers travel behavior and the corresponding presence of confirmed COVID-19 cases in the respective area. As a prerequisite to generate a geo-spatial visualization, all the locations will be geocoded from addresses into longitude and latitude.

2.5.Contact and Behavior based research

2.5.1. Someone Like You

Two different riders who share the same trip trajectories (i.e. enter and exit the same MTR station at the roughly similar period of time) can be regarded as a pair of "someone like you". The objectives of this research are to identify MTR lines with highest number of "Someone Like You" occurrences and identifying the underlying spatiotemporal pattern of those that are regarded as "someone like you" to each other. Understanding the different number of "Someone Like You"

will enables government to change the scheduling of MTR routes to reduce the occurrence of "Someone Like You" which able to serve as an additional measures to prevent the spread of COVID-19. 'Someone Like You' analysis, will also enables us to discover a particular time and station pair that has the highest number of people travelling at the same time.

Here are the steps taken to process the MTR passenger travel history data in order to count the number of "someone like you":

1. First, create all the pairs of different MTR stations. Thus, if there are n number of station in the whole MTR network, there will be n * (n-1) pairs in total.

   (i.e. Station: A,B,C; Station Pairs: AB,AC,BC,BA,CB,CA)

2. For each day, the trips will be grouped according to their corresponding station pairs.

   Example:

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

Table 2.3 Sample of Raw Data for Someone Like You

   Trip ID 1,3,4 will be grouped with station pair AB and Trip 2,5 will be grouped with station pair AC

3. For every station pair, calculate how many of these trips have approximately similar entry times. This is done by, first, dividing one full day into 48 time periods with each period last around 30 minutes. Then, count the number of trips which has the same entry and exit station belongs to each of the period. The trip categorized into a same period if the entry

time are within the same time period. (*this 30 minute time interval can be varied to different minute intervals)

Example:

Trip ID	Entry time
1	8:10
3	8:20
4	9:20
7	9:35
8	9:55

Table 2.4. Sample trip from A to B on a given day

Time Period	Trip Count
8 - 8:30	2
9:00 – 9:30	1
9:35 – 9:55	2

Table 2.5. Sample trip from A to B that have been categorized to each time periods

4. For every station pair, accumulate the trip number for each time period that has more than 1 count. The total number will be the amount of "someone like you" in a day.
5. For every station pairs, calculate the average number of "someone like you" on weekends and weekdays of each week.

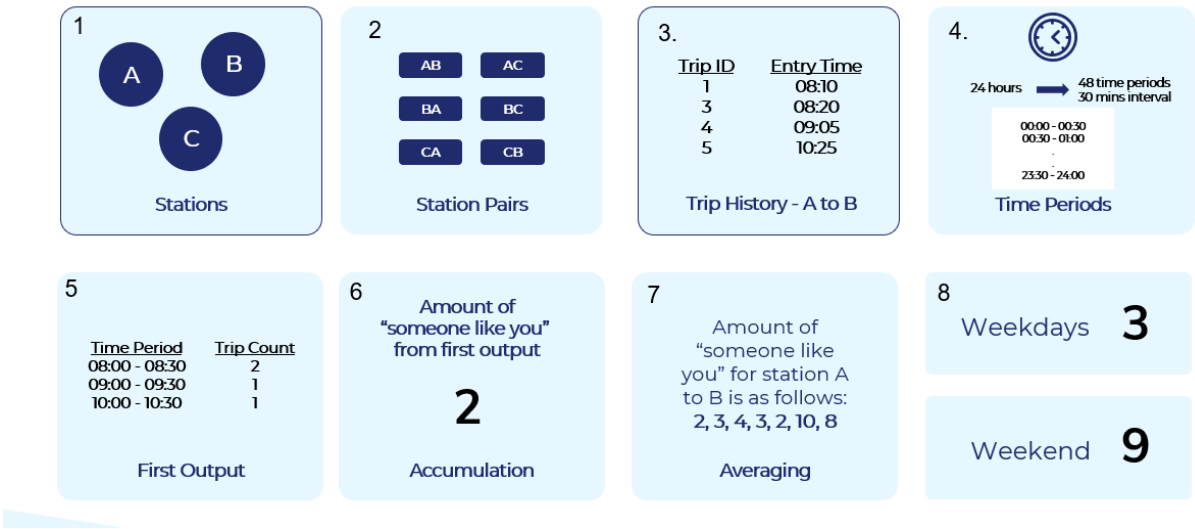

Figure 2.3. Overall process of "someone like you" analysis

Here are the assumptions used in this analysis:

1. People can be regarded as someone like you with each other even if they do not ride on the same MTR carriage.
2. The entry and exit time in the dataset are considered as entry and exit time from the train rather than entry and exit time from the station.
3. Each pair that is regarded as "someone like you" with each other are assumed to take the exact same route (i.e. if there are more than one station that allow them to switch lines, we would assume that they would switch at the same station)

2.5.2. Sensor individuals

Sensor individuals are a person who has a potential to have the most physical contact with other riders. In this project, identifying sensor individuals will enables us to identify any potential super spreader of COVID-19. The objectives of this analysis are to provide a means to visualize

spatiotemporal pattern of the copresence phenomenon within the MTR passenger and finding a probable super spreader. The objectives of this analysis are to provide a means to visualize spatiotemporal pattern of the copresence phenomenon within the MTR passenger and finding a probable super spreader. As we can see in table 2.6**.,** existing dataset of MTR passenger travel history only include the entry and exit points and their timestamps respectively, which makes it hard to identify the copresence phenomenon unless they are travelling from the same origin to same destination as can be seen in figure 2.4.

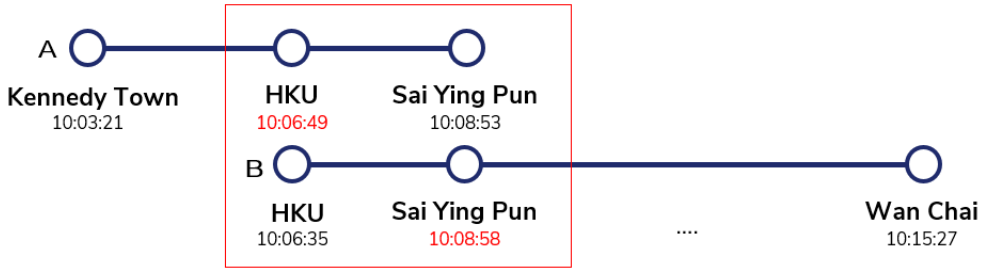

Figure 2.4. Copresence phenomenon illustration

Octopus ID	Entry Station	Exit Station	Entry Time	Exit Time
12345	Kennedy Town	Sai Ying Pun	10:03:21	10:08:53
67890	HKU	Wan Chai	10:06:35	10:15:27

Table 2.6. Existing Dataset of MTR passenger Travel History

Thus, every origin-destination information of each trip will be converted into a real path that contains all the intermediary stations that passed through the journey and its corresponding timestamp. Here are the steps taken to process the MTR passenger travel history data for sensor individual tasks:

1. Collect the time taken from one station to the next station. The data are collected manually utilizing Google Maps API. In graph theory, this time taken between 2 stations is referred as the geodesic distance of value 1 between two vertices (stations).

Example: [(A,B),(C,D),(E,F)] are neighboring station pairs.

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

Table 2.7. Sample of MTR Trip Information

2. Reconstruct a weighted undirected graph to represent the MTR train network. Python – **etwork** package is utilized to create this graph. It takes the csv file above and reconstruct a weighted undirected graph with the time durations between each station as the weight.
3. The package will implement Djikstra's shortest path algorithm to find the shortest path between a source station and the destination station.
4. After finding all of the intermediate stations, we will then give timestamps for entry time and exit time for each sub-trip.

   Here is the comparison of data before and after being processed in "sensor individuals":

   Input:

Octopus ID	Start	End	Entry Time	Exit Time
1234	Tsim Sha Tsui	Wan Chai	13:17	13:24

Table 2.8. Sample to be processed in Sensor Individual

Output:

Octopus ID	Start	End	Entry Time	Exit Time
1234	Tsim Sha Tsui	Admiralty	13:17	13:21
1234	Admiralty	Wan Chai	13:21	13:24

Table 2.9. Sample that has been processed in Sensor Individual

Below are the assumptions used in this analysis:

1. The MTR lines are indistinguishable. The graph modeled the MTR network as one big line with several different branches. Thus, trips from a station to the other station that can be done through more than one line (i.e. Central to Admiralty can be done in either Tseun Wan Line or Island Line) are considered as identical.
2. The entry and exit time in the dataset are considered as entry and exit time from the train rather than entry and exit time from the station.
3. The model does not take into considerations the time needed to switch train between different MTR lines.

To compensate the difference between the expected time calculated in our model and the actual time taken by a passenger that is caused by our assumption, we divide the amount of the difference by the number of sub-trip and add this calculated value to each sub-trip.

2.6.Web-app based platform

This section will cover the technologies stack used in the platform and its implementation.

2.6.1. Technology Stack

The front-end application of the platform is developed using React.js, which is a JavaScript library that is used to build user interface (Reactjs). React.js is used due to its high component

reusability and application state management capabilities. The front-end application is styled using CSS with Bootstrap framework.

The back-end application of the platform is developed in python using Django framework, which is a python web application framework that enable development of secure and maintainable websites. Django is used due to the built-in admin system that it provides which enables us to manage users of the platform. This functionality is essentials to handle user authorization to the system. Django also features a built in Object Relation Mapper (ORM) which enables the backend service to establish a communication with the SQL database through an interface, thus enabling the system to perform SQL queries using programmatic methods instead of SQL queries.

The back-end application and front-end application communicate with each other through REST APIs. The Django framework enables us to configure REST APIs through its MVT (Model-View-Template) design pattern (JavaTPoint). In the MVT design pattern, Model helps to handle the database. It works as a data access layer that handles the data. The Template works as a presentation layer that handles the User Interface part. While View initiates an interaction with the model to carry the data and render a template. In Django, raw HTTP request that is received from the front-end will be processed and converted into a pythonic HTTP request object which enables the back-end system to process the metadata of the raw request.

Figure 2.5. Technology Stack

2.6.2. Platform Architecture

Our database where MTR passenger data and COVID-19 confirmed cases data is stored are hosted in HKU HinCare server which is located within HKU intranet network. In order to access this server, user will need to establish an SSH-connection with HKU CS gatekeeper. Thus, as a pre-requisite, user will need to have a HKU CS intranet account. After the connection with the gatekeeper has been established, user will need to make another SSH tunnel to access the HinCare server. Thus, a multi-hop SSH tunnel which consist of 2 different SSH connections are required to enable user to connect to the HinCare server. This added complexity of access to the server is enforced to improve the security of the server as the data that MTR data that was stored in the database are confidential.

In order to be able to connect to the server in a convenient manner, while at the same time still able to maintain a secure connection, our team decided to establish a double port forwarding to connect to the HinCare database. The SSH port forwarding will be established between the HinCare server and HKU CS gateway server (port 8000 to 8080); and between HKU CS gatekeeper and FYP Virtual Machine (port 8080 to 8080). Therefore, enabling a direct access through fyp.cs.hku.hk:8080. As the front-end application of the platform is located in the FYP virtual machine server, it only required to be deployed on port 80 of the machine. Figure 2.6. the connection process between both back-end application, front-end application, and the established port forwarding.

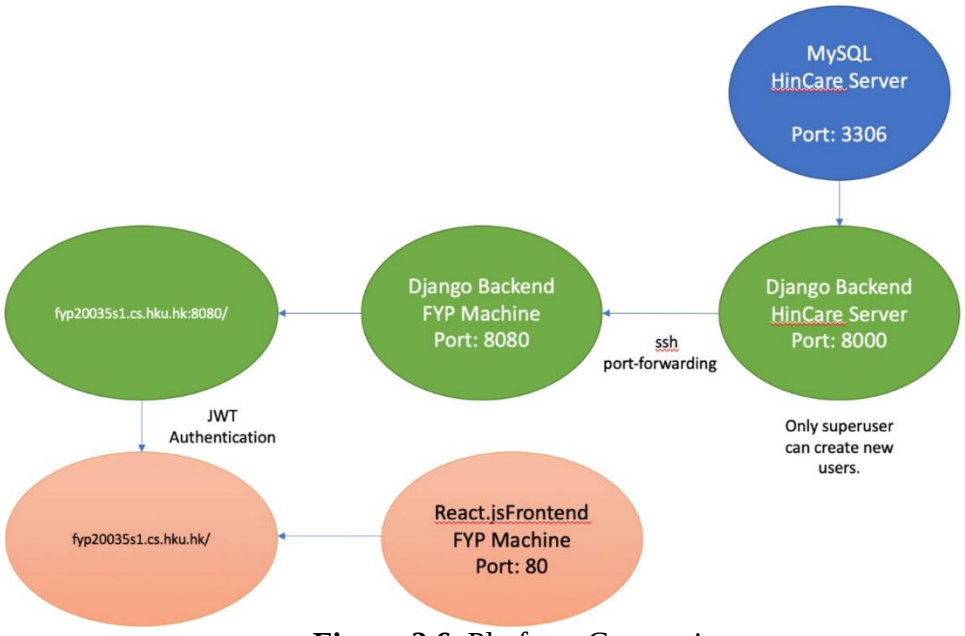

Figure 2.6. Platform Connection

The architecture described in Figure 2.6. also solve another problem. User are no longer required to have a HKU CS account in order to access the platform, as the architecture enables the management of authentication to be delegated to Django authentication service. The authentication system is powered by the JWT (JSON Web Token) which supported by rest_framework_simplejwt library. Once user are able to log in to the platform, the Django service will generate a JWT and

send it to the client side. The front-end application of the client will utilize this token whenever they do a REST APIs call to the back-end application. Without this token, the REST API call cannot be authorized. Thus, there will be no response returned from the back-end. The token will expired after a period of timeout, which will make the client to be logged out automatically from the platform.

2.6.3. Visualization feature

The system will enable user to retrieve both geospatial and non-geospatial visualization. The system' geospatial visualization platform is developed with the aid of the ArcGIS technology. Utilizing the JavaScript ArcGIS API, the system is able to render ArcGIS visualizations into our web-based applications which enables the user to have an access the ArcGIS visualizations tool right from the browser. In order to enable the creation of the visualizations on the front-end side, the arcgis-js-api module will need to be installed beforehand. The visualization code and logic are done through the front-end React application.

Here are the steps taken to embed the ArcGIS visualizations into our web application platform:

1. Fetching the specific data required for a visualization from the backend service utilizing REST API calls. (i.e.: geolocation data (longitude, latitude))
2. The data fetched from the backend will be converted into an ArcGIS Graphic objects. A collection of Graphic objects is converted into ArcGIS FeatureLayer object. (FeatureLayer will be set as a dataset and a Graphic object will be set as a row in the dataset).
3. A new ArcGIS Map object – which is baseline map that is used for visualization – will be rendered.

4. Configuring the data source of this map as the ArcGIS FeatureLayer object that we generated.
5. Modify the rendered visualization by specifying different configurations / variables when the Graphic and FeatureLayer object is created to obtain different visualization functions.

By following his procedure, we will be able to create various visualizations on the platform

2.7.Summary

This chapter covered the workflow of the required implementation in order to fulfil the project' technical requirement. Each section covers the justification of all engineering decisions such as choosing SQL databases, and uses of Python and ArcGIS for mobility trend analysis, visualization, and behavior-based research. Last subsection covered all the technical requirements of the platform.

3. Results

This section will cover the result of this project. The main result of this project is a platform that will enable user to get insights from both COVID-19 confirmed cases data and MTR passenger data. In addition, some of our findings regarding on the correlation between COVID-19 cases and MTR passenger mobility during the January – April 2020 period will also be covered.

3.1.The platform

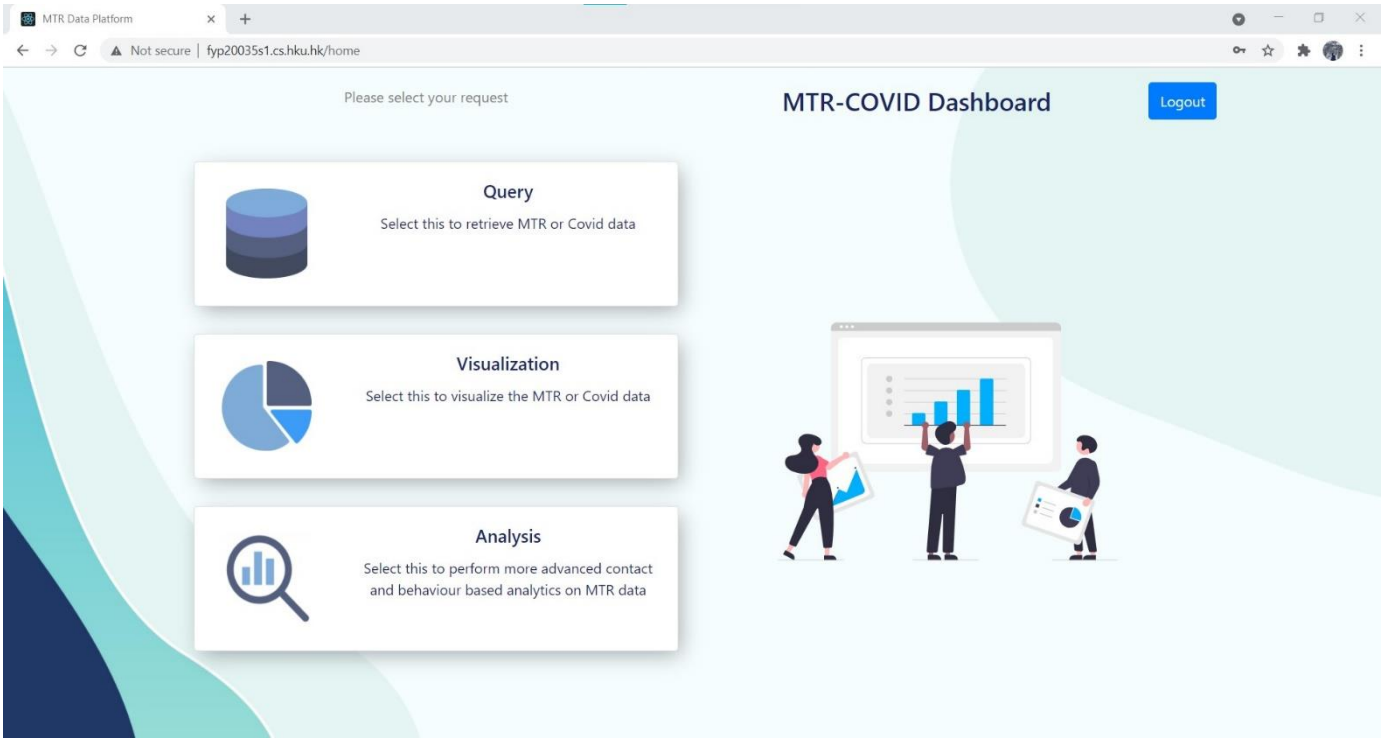

Figure 3.1. Home Screen Page

Figure 3.1. displayed the platform main menu which cover all the main functionalities of the platform. There are 3 major functionalities of the platform:

1. Query – User will be able to retrieve both COVID-19 confirmed cases data and MTR passenger data

2. Visualization – User will be able to get a visualization of the queried data that they retrieve from the platform
3. Analysis – User will be able to receive advanced contact and behavior-based analytics on the MTR passenger data.

3.1.1. Log-in feature

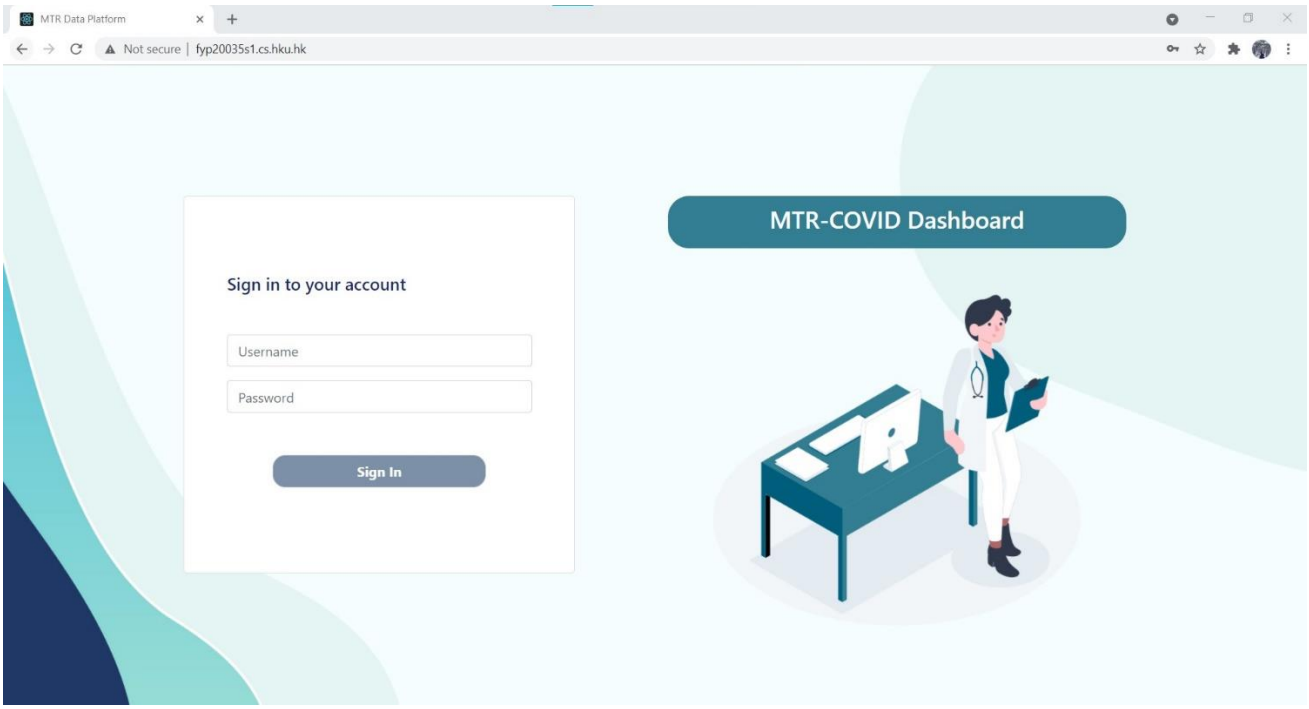

Figure 3.2. Log in Page

User will need to log in to the system by providing the username and password that is provided to them before having access to all the feature that was mentioned in section 3.1. Log in feature will enable the system to authenticate the user access to the platform and set a token for the user session. This token is used by the server as an identification of the user in the server side. This token will be expired after it is not used in a prolonged period of time for security considerations.

3.1.2. Querying feature

3.1.2.1. COVID-19 cases data query

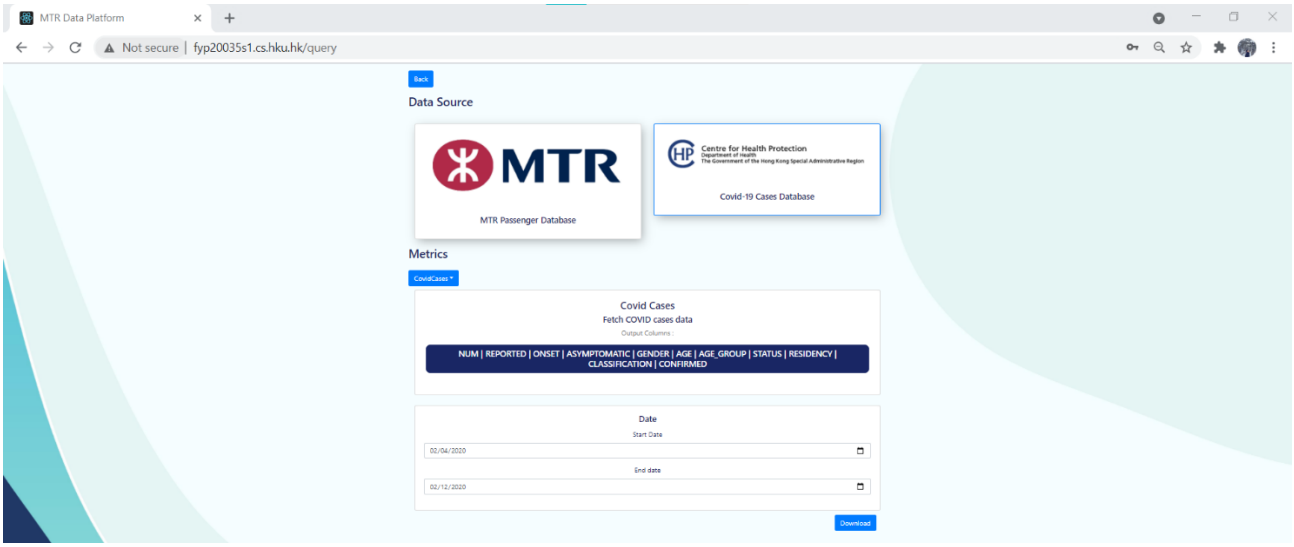

Figure 3.3. COVID-19 cases data query page

User will be able to get the COVID-19 data from the time period that they desired by inputting the "start date" and "end date" in the "start date" and "end date" field respectively. After giving the specified date, the system will return a CSV file with the following content:

NUM	REPORTED	ONSET	ASYMPTOMATIC	GENDER	AGE	AGE_GROUP	STATUS	RESIDENCY	CLASSIFICATION	CONFIRMED
14	2/1/2020	1/23/2020	NO	M	80	Elder	Discharged	HK resident	Imported case	Confirmed

Table 3.1. Sample output of COVID cases query

3.1.2.2.MTR passenger data query

For the MTR passenger data, user will be able to get the raw data that was commissioned by the MTR corporation and the data that has been preprocessed according to user needs. There are three different custom query that we provide: station density, passenger mobility, and travel pattern.

3.1.2.2.1. Station Density

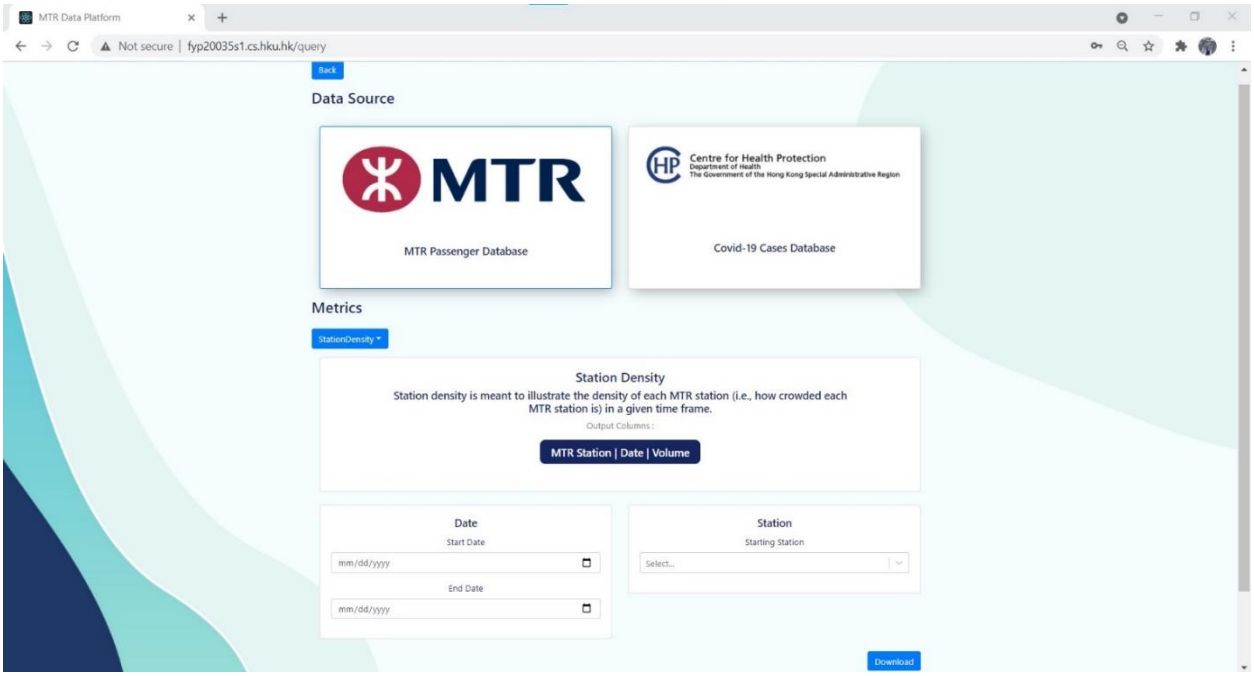

Figure 3.4. Station Density Page

Station density is used to illustrate the denseness of each MTR station in a given day. User will be able to further filter the result that they need. First, user will be able to get the data from the time period that they desired by inputting the "start date" and "end date" in their respective input field. In addition, they can also filter the list of station that they need by inputting the station that they want to observe in the "starting station" input field. After clicking the "download" button, user will be able to download a CSV file with the following content:

MTR Station	Date	Volume
1	02-02-2020	15000

Table 3.2. Sample output of station density

The example can be interpreted as following, “there are 15000 passenger that went out from MTR station with MTR station code “1” on 2-2-2020”. The station code is commissioned by the MTR corporation.

3.1.2.2.2. Travel Pattern

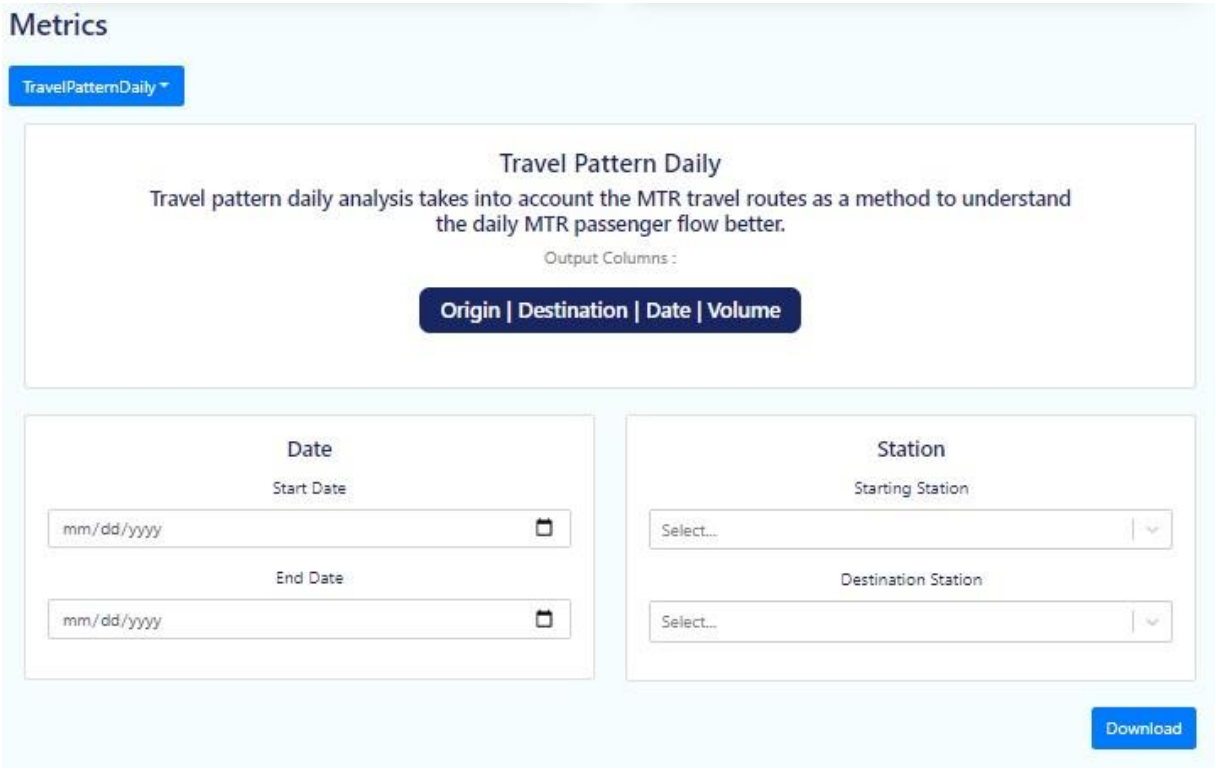

Figure 3.5. Daily Travel Pattern Page

Travel pattern is used to illustrate the number of passengers from one MTR station to the other in a day. Use will be able to filter the period of time of data that they need by specifying the start date in the “start date” input field and the end date in the “end date” input field. In addition, user can also get the data from a specific station pair by inputting the “starting station” and “destination station” field (Figure 3.5.). If user want to get hourly travel pattern data, user will need to choose a different metrics type which is called “Travel Pattern by Hour” (Figure 3.6.).

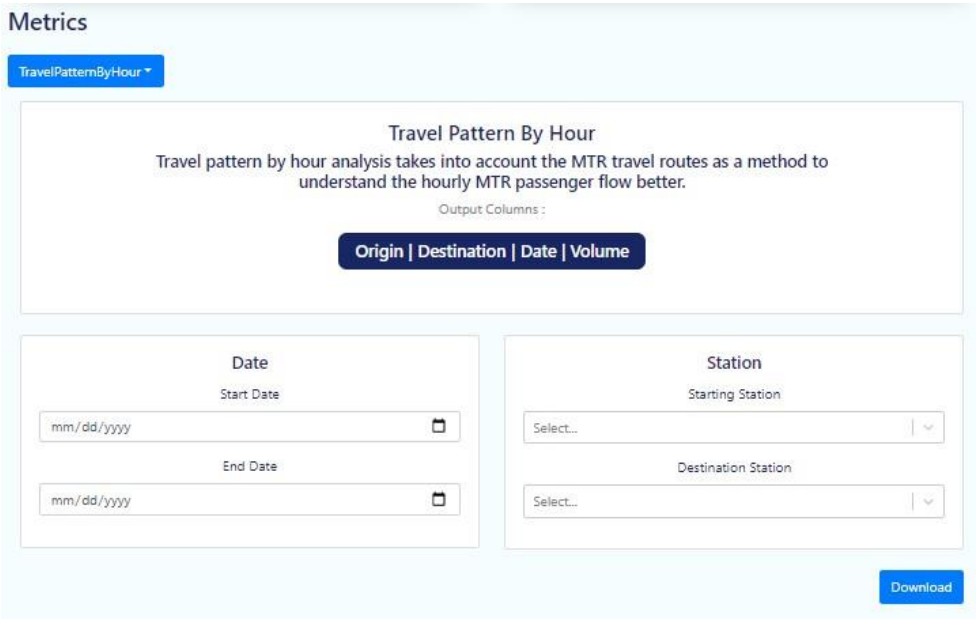

Figure 3.6. Hourly Travel Pattern Page

Both “TravelPattern” and “TravelPatternByHour” query will then return a CSV file with the following content:

Origin	Destination	Date	Volume
3	7	2-2-2020	7000

Table 3.3. Sample output of Travel Pattern and Travel Pattern by Hour

The example in Table 3.3 means that there are 7000 people travelling from MTR station with code “3” to MTR station with code “7” on 2-2-2020.

3.1.2.2.3. P*assenger mobility*

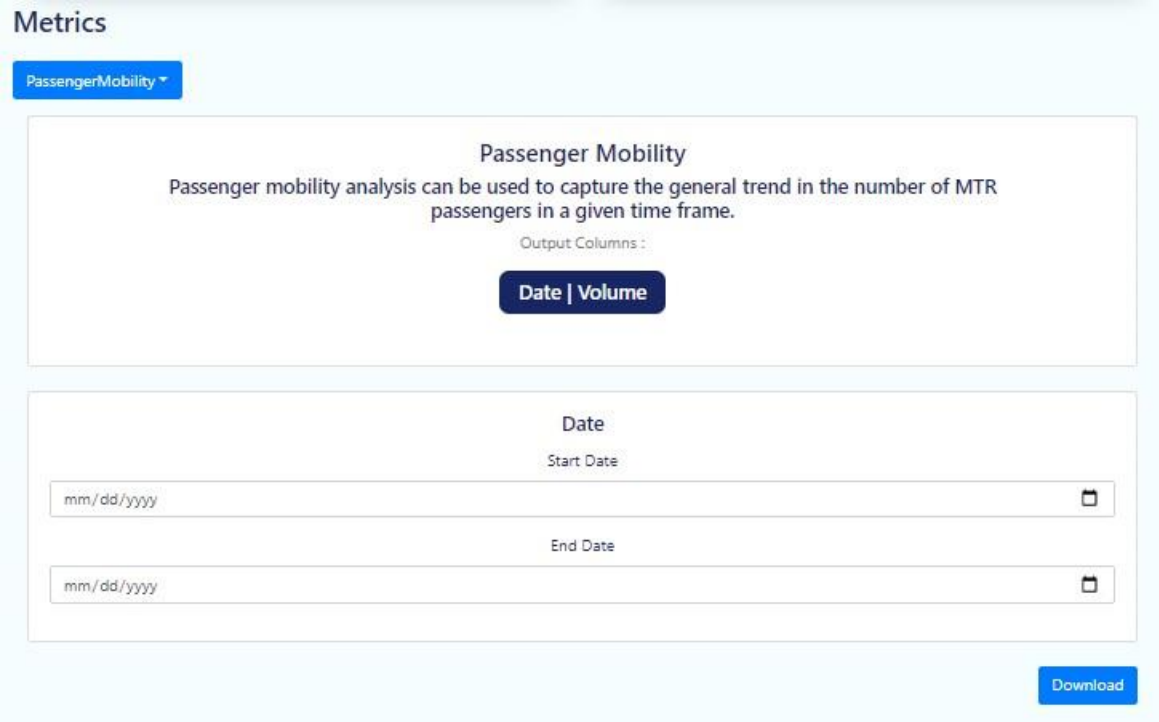

Figure 3.7. Passenger Mobility Page

Passenger mobility is used to observe the general trend in number of MTR passengers in a day. User will be able to get the dataset from a certain period of time by inputting the start date and the end date that they desired in the "start date" and "end date" input field respectively (Figure 3.7.). Table 3.4. illustrated the sample result of the passenger mobility query. The example in table 3.4. means that there are 12000 people taking the MTR on 2-2-2020.

Date	Volume
2-2-2020	12000

Table 3.4. Sample output of Passenger Mobility

Should the user need to get the result based on the octopus card type, they will need to choose another metric type ("PassengerMobilityByCard") – Figure 3.8. Table 3.5. illustrated the sample result of the passenger mobility by card query. The example means that there are 80000 number of Adults taking the MTR on 2-2-2020.

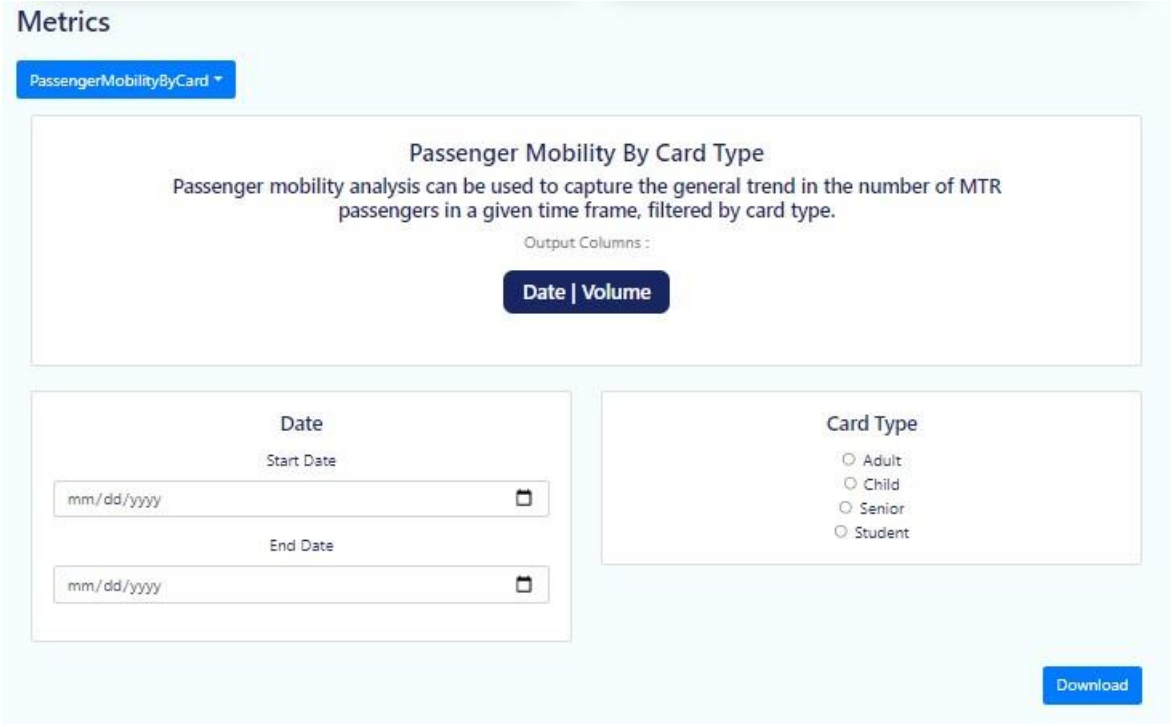

Figure 3.8. Passenger Mobility by Card Type Page

Date	Volume	Card Type
2-2-2020	8000	ADL

Table 3.5. Sample Output of Passenger Mobility By Card Type

3.1.2.2.4. Raw Data Query

Raw data query is used to get the dataset that was given by the MTRC. User can filter which data they want to retrieve. Yet, due to the limitations of our system, user can only get a filtered raw data if the data requested are within the same month. Thus, user will need to select the "query within month" options (Figure 3.9)

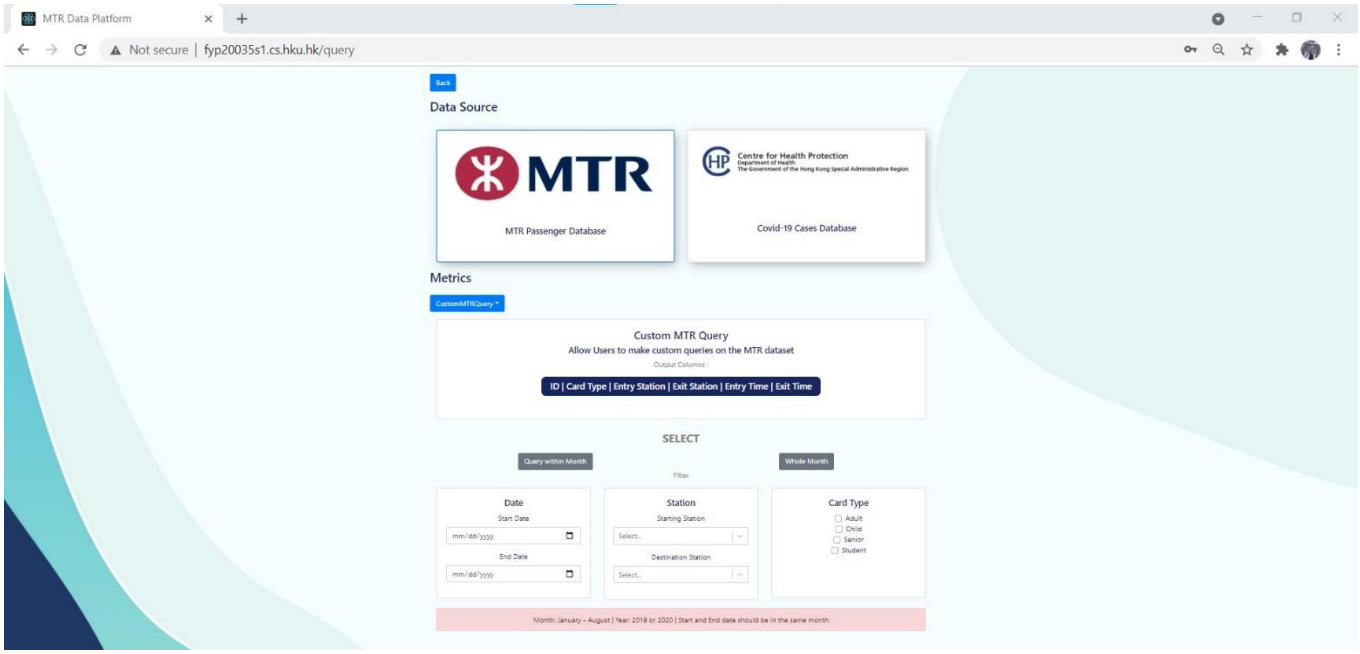

Figure 3.9. Custom MTR Query Page

The user can specify the period of time of data that they needed by entering the start date and end date of their desired period in the “start date” and “end date” input field. User will also be able to filter which station that they want to include in their dataset by specifying either or both starting and destination station in “starting station” and “destination station” input field respectively. In addition, user can specify which data that they need according to the Octopus card types by selecting the card type that they need in the card type checkboard field.

User can also download the entire raw data in a given month by selecting the “whole month” option. In this menu, user can specify which data that they need according to the month and year of the data

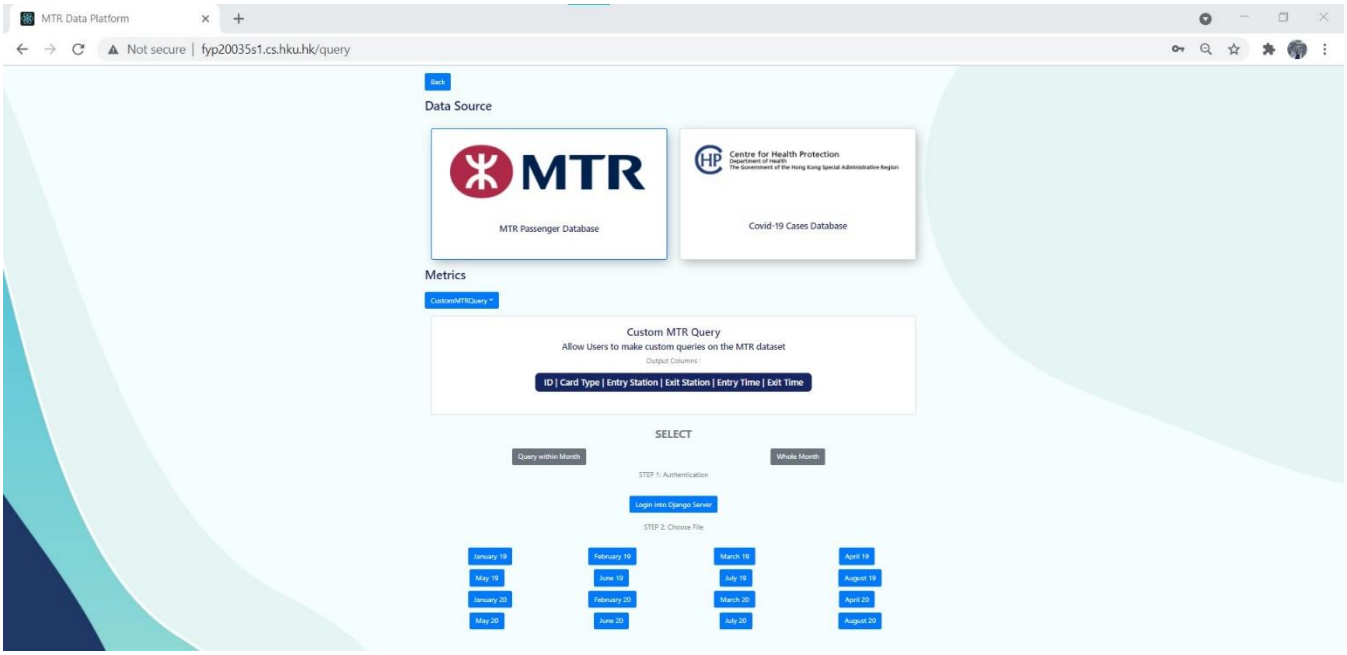

Figure 3.10. Custom MTR Query Page (Whole Month Data)

Both of the options will return the same CSV file. Table 3.6. illustrate an example of CSV file returned from this query

ID	Card Type	Entry Station	Exit Station	Entry Time	Exit Time
12345	ADL	1	2	12:00	12:03

Table 3.6. Sample output of Raw Data Query

The data can be interpreted as “Passenger with ID 12345, who has a ADL card, take a trip from MTR station with code 1 to MTR station with code 2, the passenger enter the station at 12:00 and exit the station at 12:03”.

3.1.3. Visualization Feature

The visualization feature will enable user to make both geospatial and non-geospatial visualization for both MTR passenger data and COVID-19 confirmed cases data. A geospatial visualization of both Travel Pattern and Station Density will be generated. In addition, the geospatial visualization of the COVID-19 confirmed cases which visualized the distribution of COVID-19 cases can be overlayed over the geospatial visualization of MTR passenger. A non-geospatial visualization in the form of a line chart will be generated for Passenger Mobility.

User will first need to choose the visualization that they need. Then, for both Travel Pattern, Station Density, and Passenger Volume the user just need to specify the date where they want their data from. Further personalization can be done for all the visualizations. For, Station Density visualization, user can specify the number of most populated station to be displayed as observed in figure 3.13. For the travel pattern, user can specify the number of most populated station pairs that want to be displayed (Figure 3.11.). Both of these visualization, there is an option to overlay the COVID-19 cases data (Figure 3.12 and figure 3.14). For passenger volume, user can specify which type of passenger that want to be visualized (Figure 3.15).

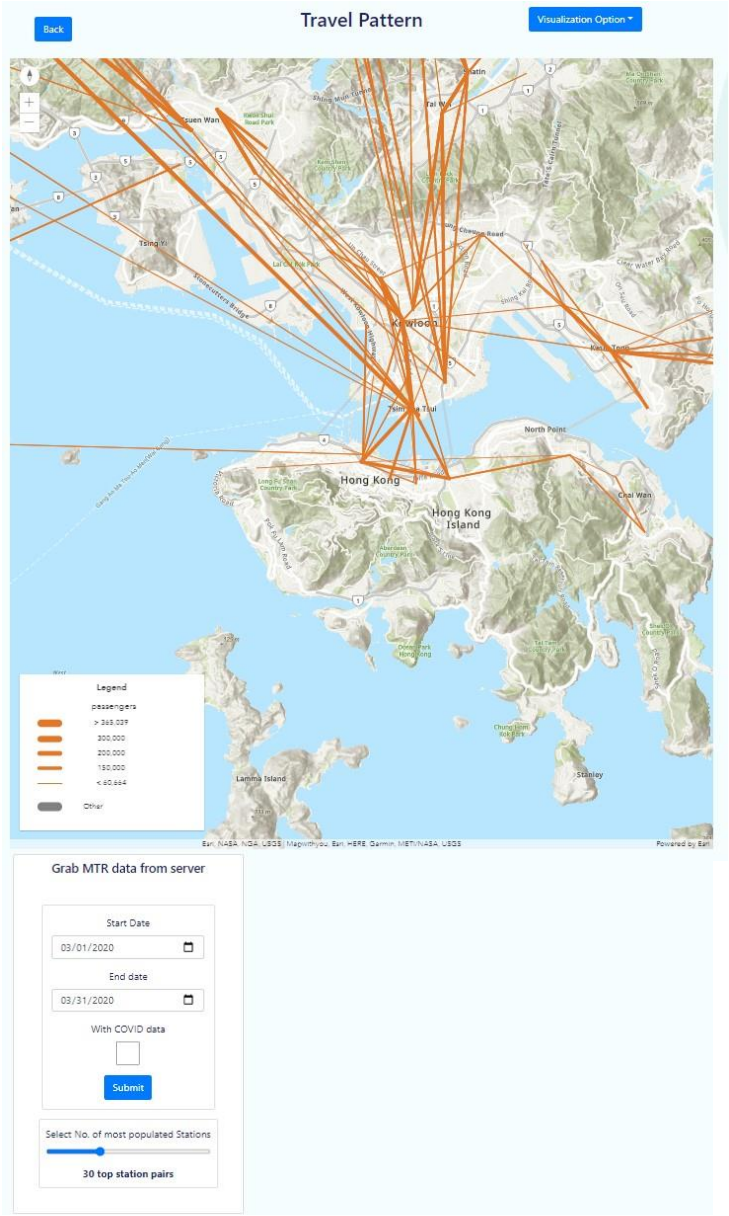

Figure 3.11. Travel Pattern Visualization

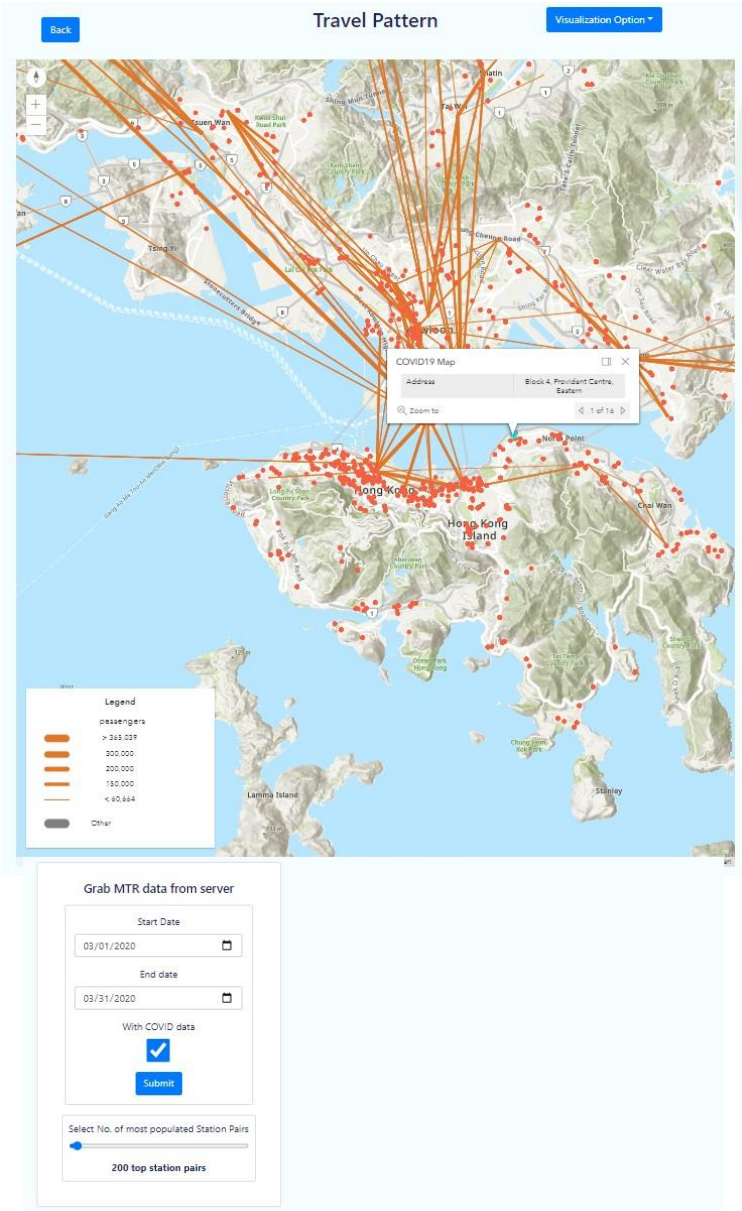

Figure 3.12. Travel Pattern and COVID – 19 Visualization

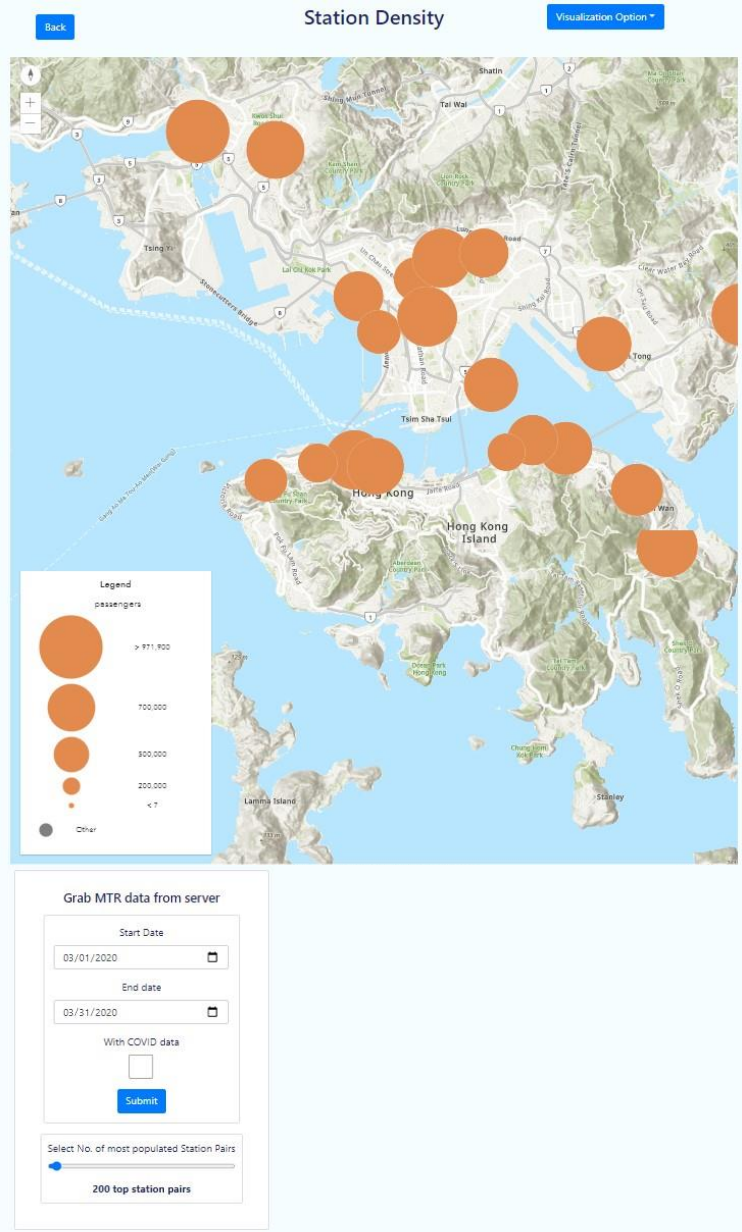

Figure 3.13. Station Density Visualization

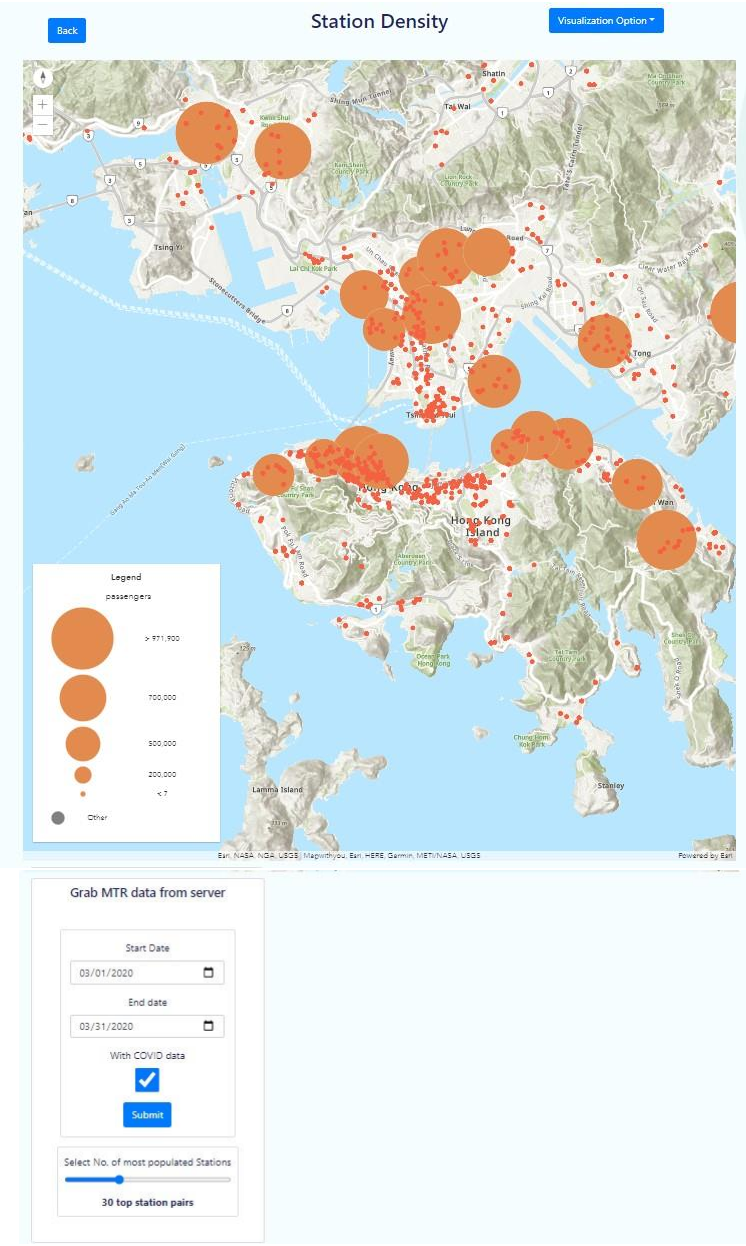

Figure 3.14. Station Density and COVID-19 Visualization

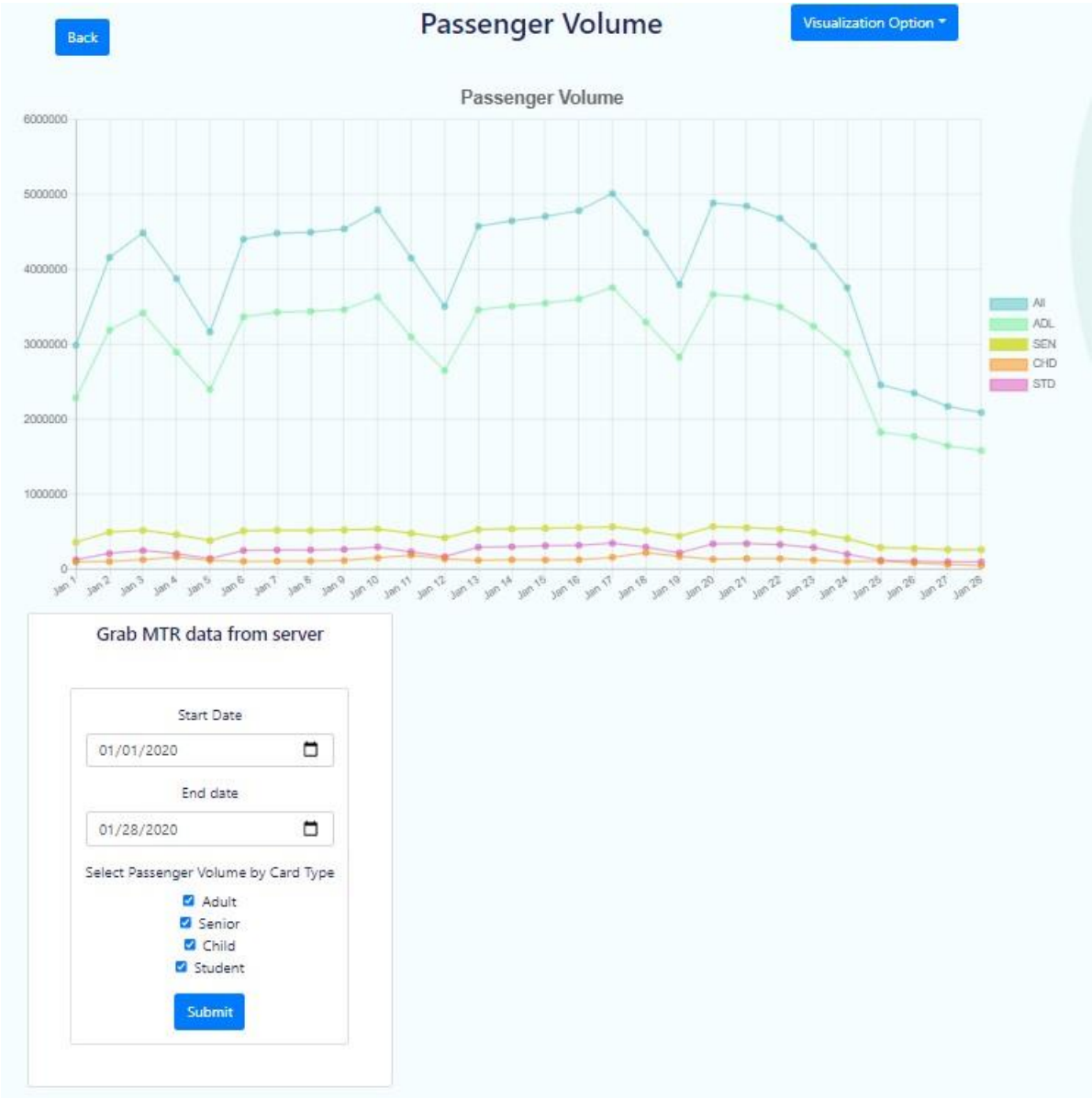

Figure 3.15. Passenger Volume Visualization

3.1.4. Analysis feature

This feature will enable user to run advanced contact and behavior-based analysis on the MTR passenger data. User will be able to run analysis on both “someone like you” and “sensor individuals”.

3.1.4.1.Someone like you

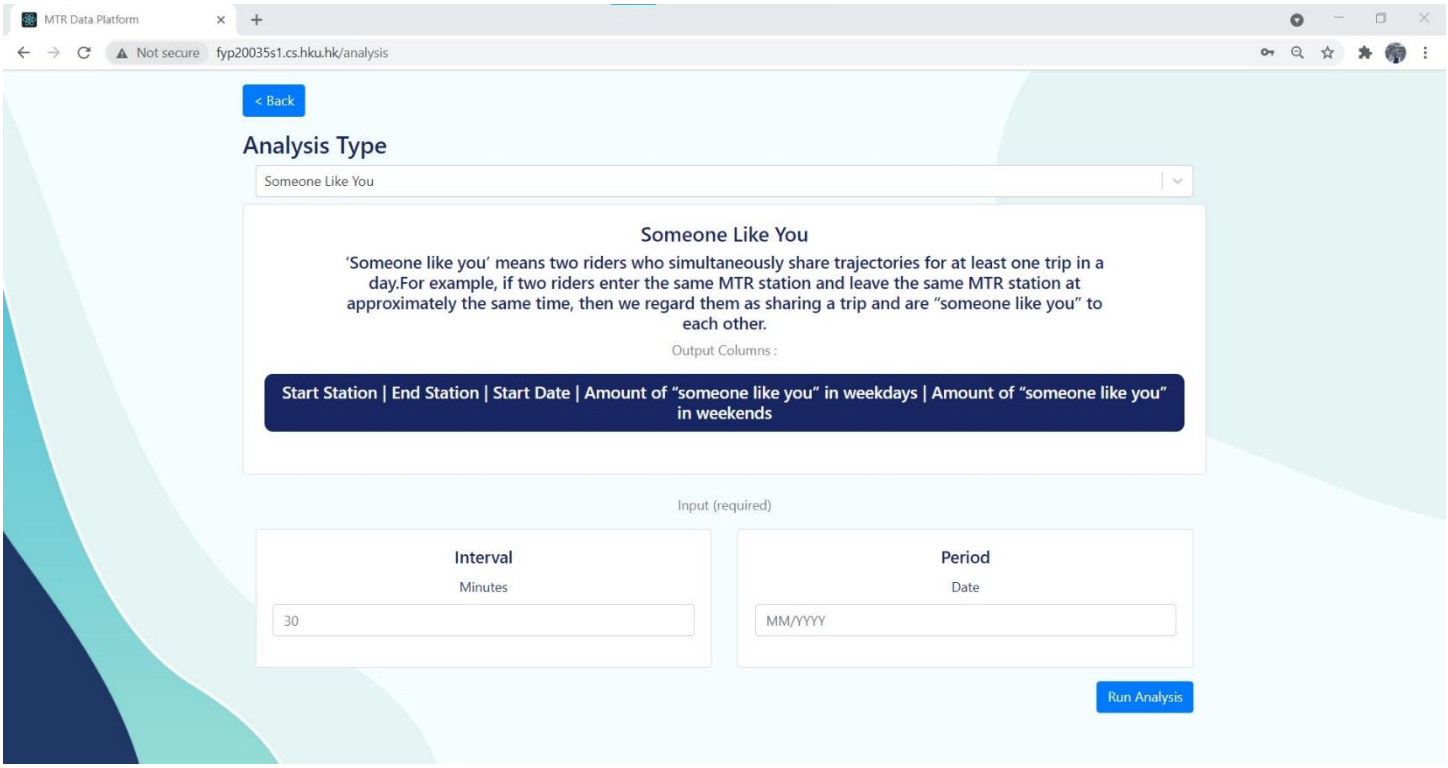

Figure 3.16. Someone Like You Page

As mentioned in the methodology section, a user can be regarded as a "someone like you" to other user if they are travelling from the same origin to the same destination in roughly same period of time. As seen in figure 3.16. In the platform, user need to specify how long the interval of a time period by changing the "interval" input field. User will also need to specify the month and year of data that need to be analyzed by changing the "date" input field. Table 3.7. contained the sample CSV file output of someone like you analysis

Date	Weekend	Entry_Stn	Ext_Stn	Freq_SLU
2-2-2020	TRUE	1	2	5

Table 3.7. Sample Output of Someone Like You

The data in table 3.7. can be interpreted as follow," on the day in the weekend of week that start on 2-2-2020 in average there are 5 people who are someone like you to each other on the route from MTR station with code 1 to MTR station with code 2".

3.1.4.2.Sensor individuals

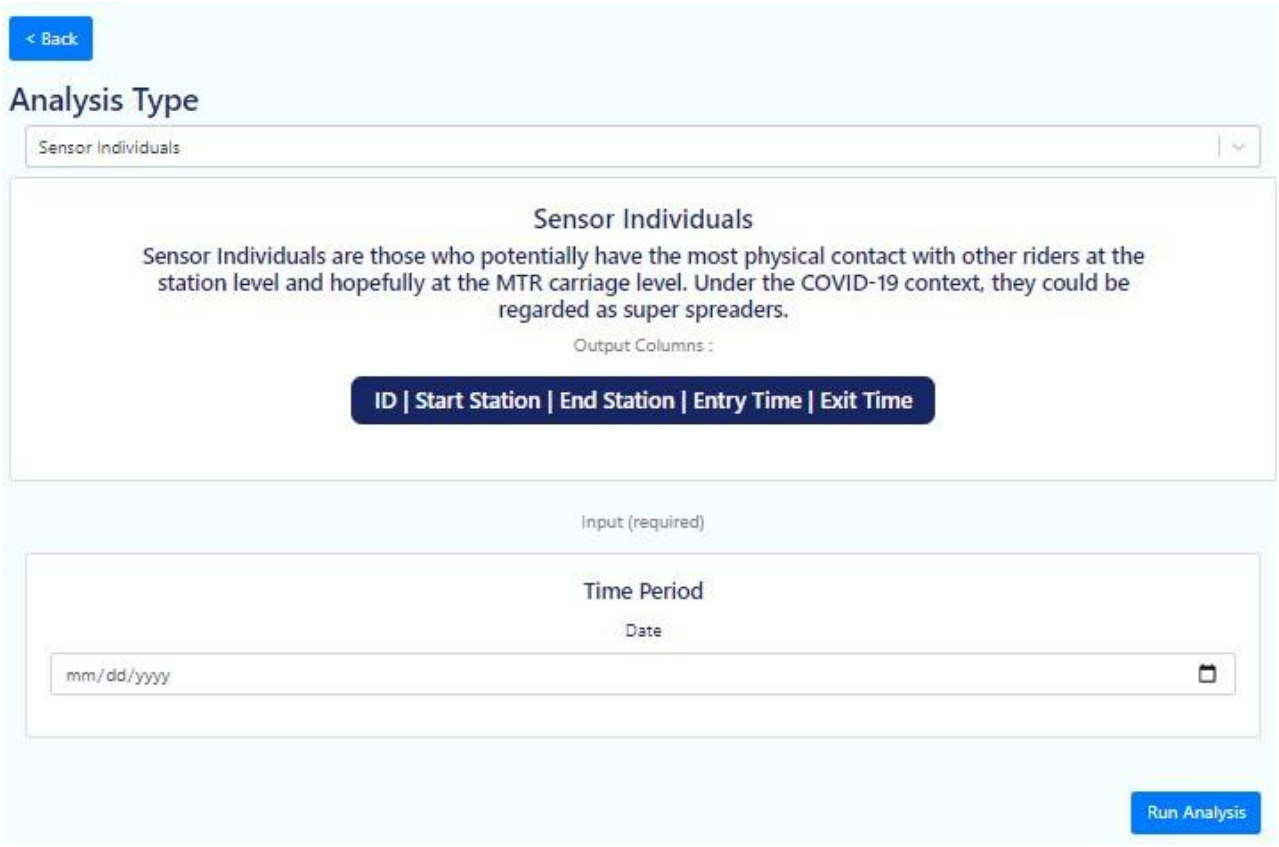

Figure 3.17. Sensor Individuals Page

For sensor individual's analysis, user can specify the start of time period where they want to get the data from by inputting it in the "Time Period" input field. Then, the system will only run the "sensor individuals" on data that has "Entry Time" within the 10 minutes time from the submitted start date. The result of this query can be seen in table 3.8.

Octopus ID	Start	End	Entry Time	Exit Time
1234	1	2	13:17	13:24

Table 3.8. Sample output of Sensor Individuals

3.2.Trend discovered

This section will cover the trend that is discovered regarding the correlation between the MTR passenger mobility and the spread of COVID-19.

3.2.1. Trend during the first wave of COVID-19

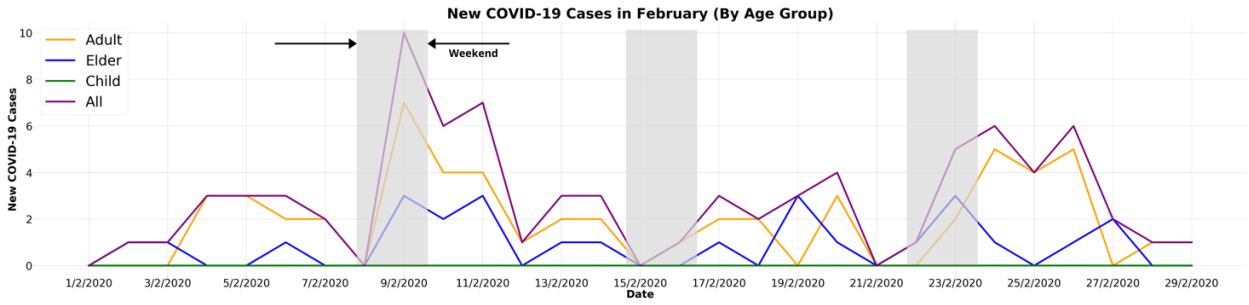

Figure 3.18. Changes in COVID-19 cases in February 2020 categorized by age group in February 2020 (during the first wave)

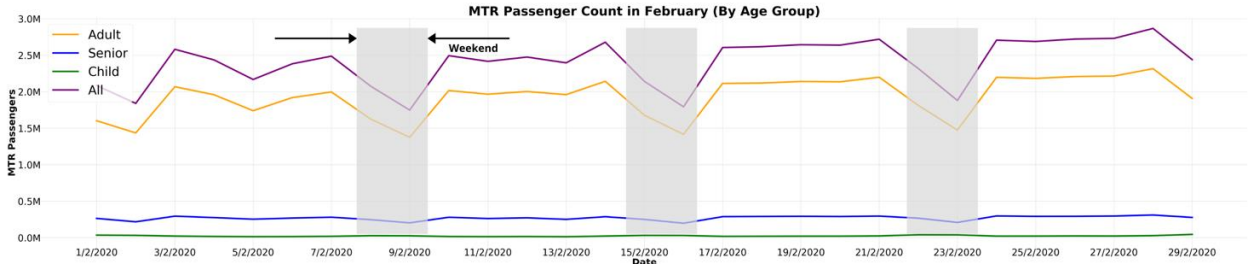

Figure 3.19. Changes in the number of MTR passengers categorized by age group in February 2020 (during the first wave)

Figure 3.18 represents the change in the number of new COVID-19 every day from 1 February 2020 until 29 February 2020, categorized by each age group. Figure 3.19 represents the change in the number of volumes of MTR passengers from 1 February 2020 until 29 February 2020, which is categorized by each age group. From those two graphs, there is a correlation between the changes in the number of new cases and the number of MTR passenger count. Noticed

in the adult group, whenever there is an increase in MTR passenger flow, the number of new COVID-19 cases also increases.

3.2.2. Trend during the second wave of COVID-19

Utilizing the ArcGIS software, several geospatial visualizations are generated to observe the trend within the second wave of COVID-19 pandemic in Hong Kong

3.2.2.1.Change in busiest MTR Route

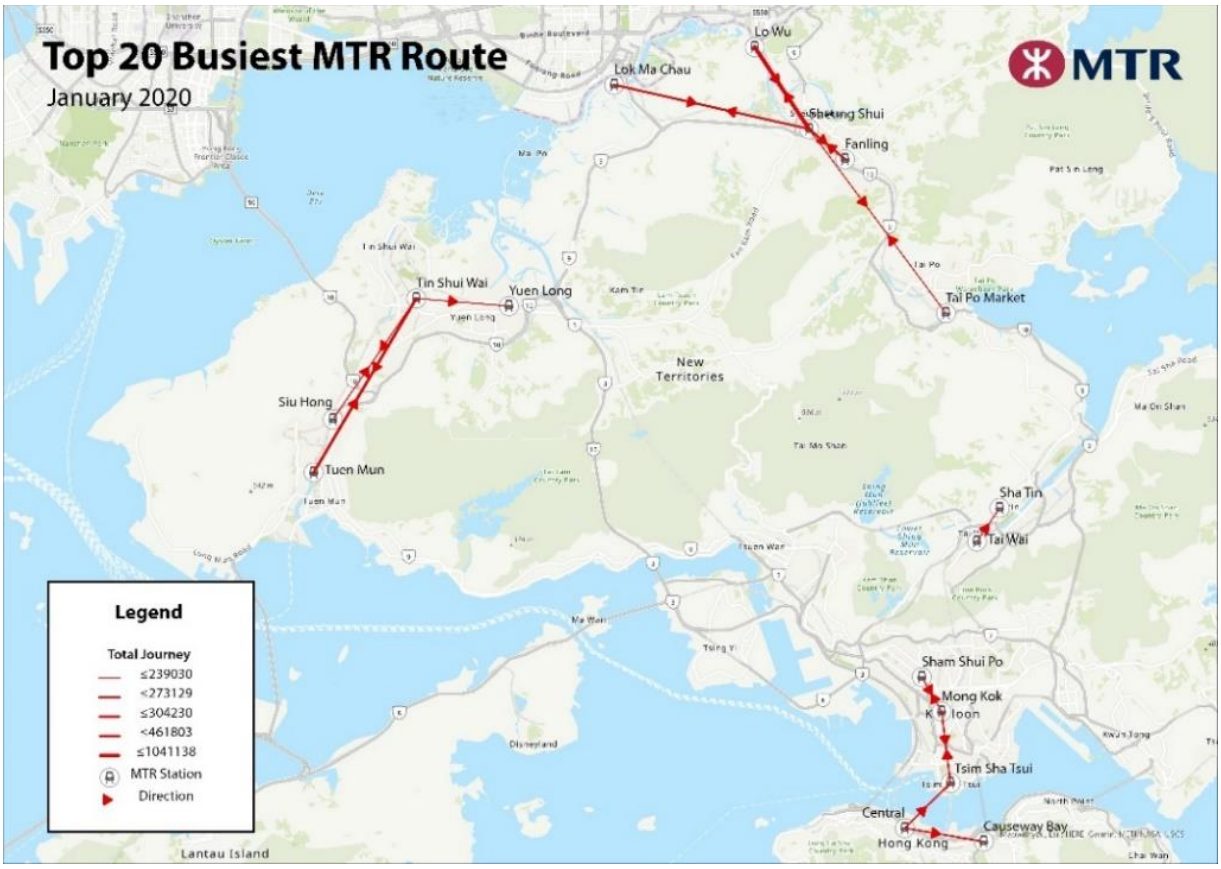

Figure 3.20. Top 20 Busiest MTR Route in January 2020

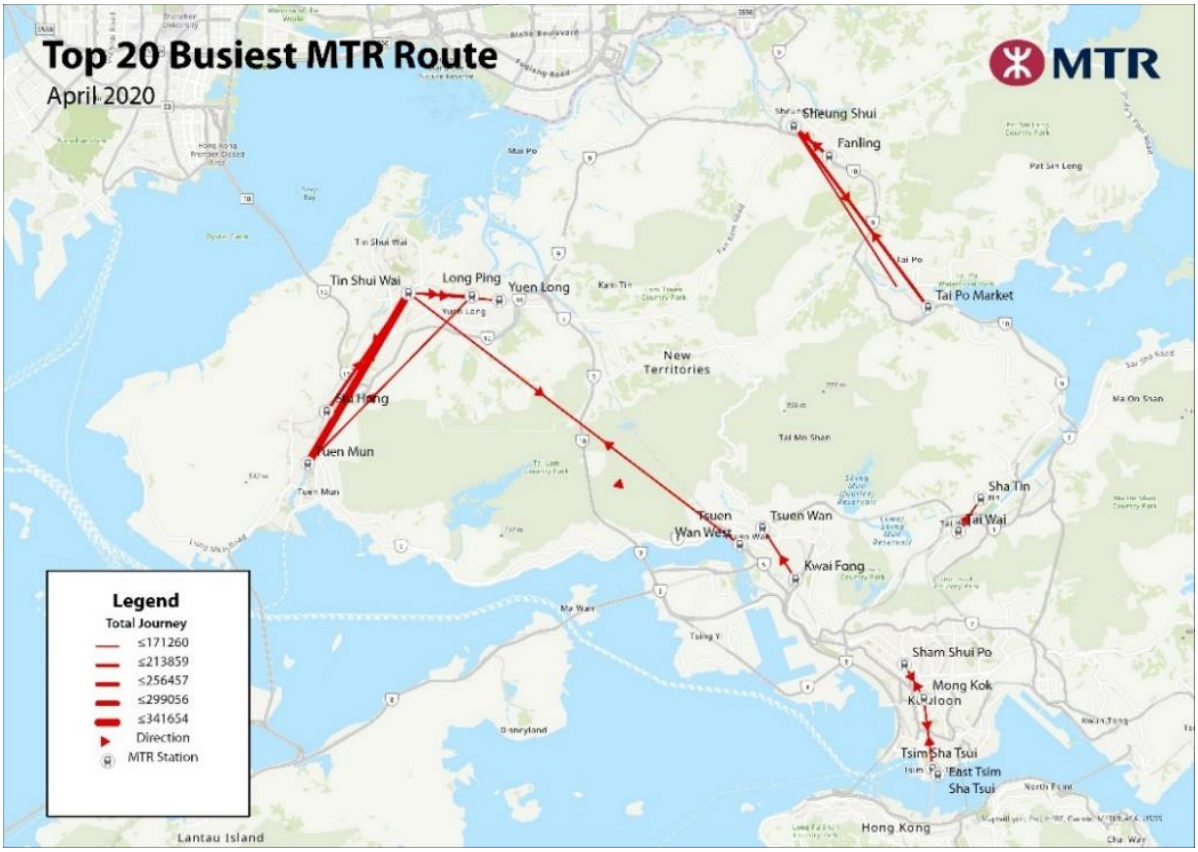

Figure 3.21. Top 20 Busiest MTR Routes in April 2020

Based on figures 3.20. and 3.21, a noticeable drop in terms of the total number of the journey taken from one station to the other can be observed. In January 2020, the maximum total journey is 1,041,138 from Sheng Shui station to Lo Wu station. However, in April 2020, the maximum total journey is only 341,654 from Tuen Mun station to Tin Shu Wai station. This drop can be attributed to the enforcement of social distancing measures. Moreover, since February 2020, the Hong Kong government has closed the Lo Wu border, which explained why there is not any traffic from Sheung Shui Station to Lo Wu Station/Lok Ma Chau Station and vice versa.

3.2.2.2.Distribution of COVID-19 Cases

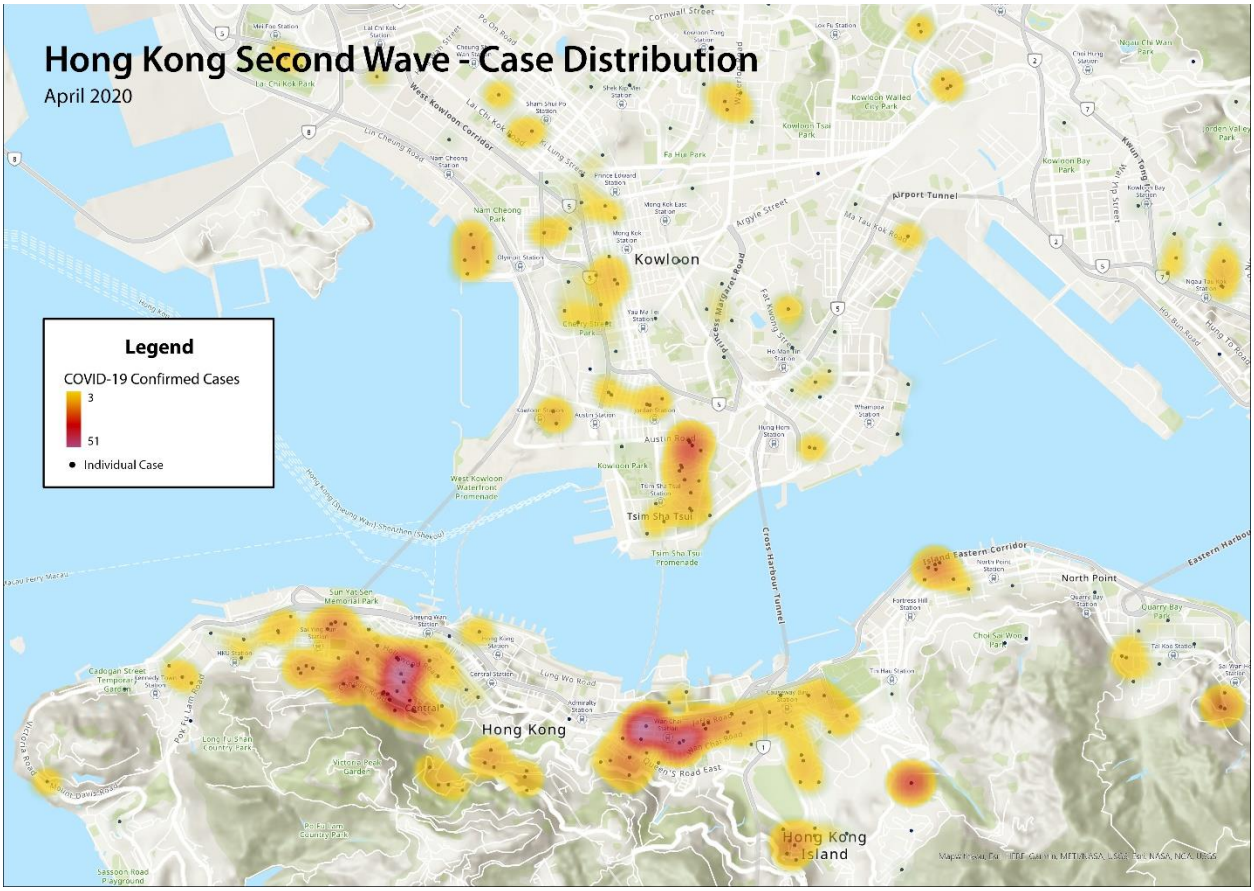

Figure 3.22 Case Distribution of COVID-19 in April 2020

Figure 3.22 covered the distribution of COVID-19 cases in Hong Kong during April 2020. A heatmap visualization is created to enable us to see the concentration of the location of COVID-19 cases easier. Color red indicated that there are a lot of COVID-19 cases in that area, while the color yellow indicated that there are not too many cases in that area.

3.2.2.3.Correlation between the distribution of COVID-19 and MTR traffic

To observe the correlation between COVID-19 cases and MTR traffic, a visualization of both COVID-19 case distribution and volume of the incoming passenger to a specific MTR station

is created. The objective of this observation is to determine whether a higher volume of MTR passengers corresponds to a higher number of COVID-19 cases in an area.

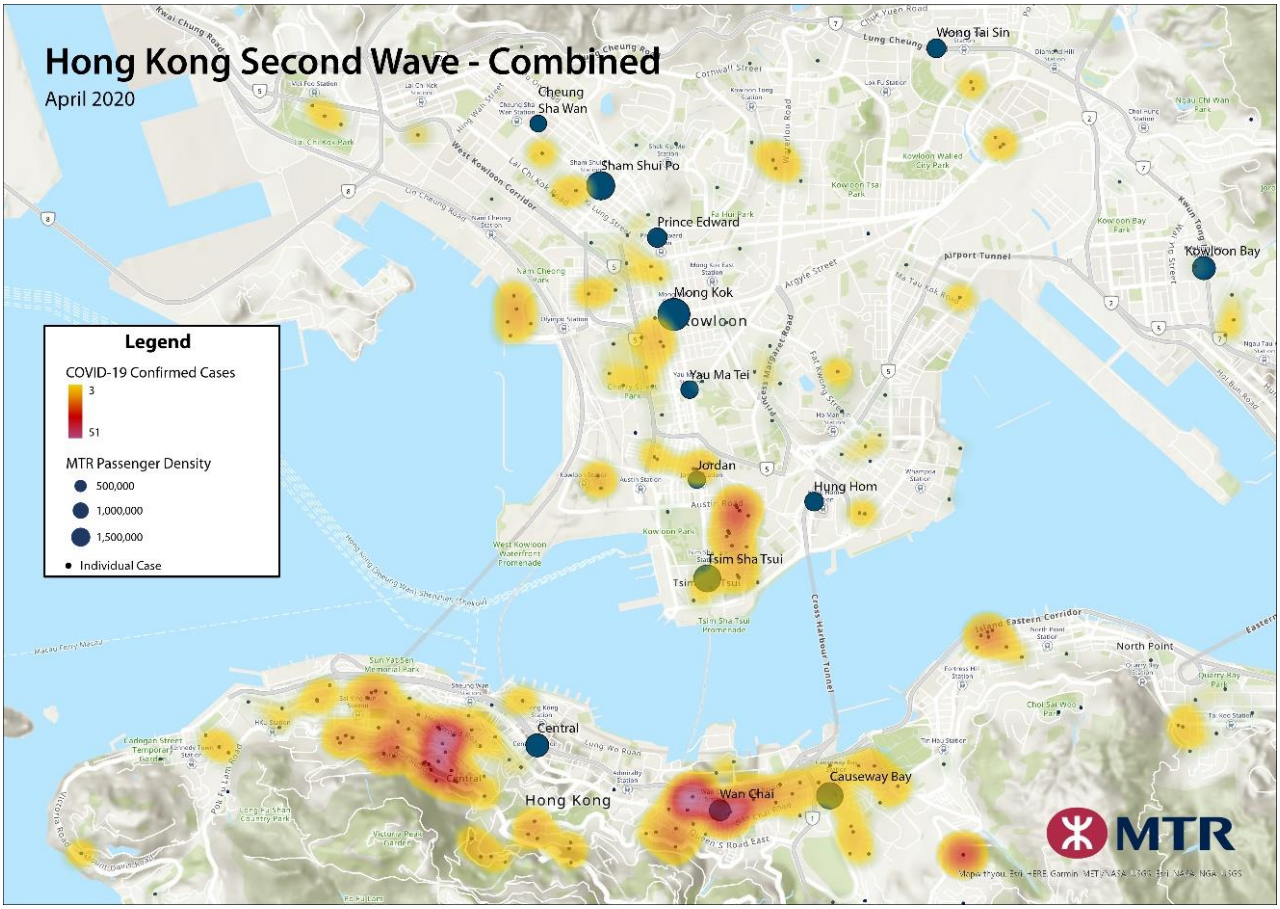

Figure 3.23. Combined Case Distribution Heatmap and MTR Incoming Passenger Volume

In figure 3.23., the blue circle's size represents the volume of the incoming passenger to a certain MTR station. In this graph, only the top 20 stations with the highest incoming volume of the passenger are displayed. It can be observed that around the area where there are a vast number of COVID-19 cases; we can see a total of at least 500,000 MTR incoming passengers in that station

4. Limitations and Future Development

4.1.Limitations

4.1.1. Nature of Dataset

While working on the MTR passenger data, a numerous number of assumptions need to be made, especially on the contact and behavior-based research as the MTR data that only consist of origin and destination MTR station for each passenger, which makes it hard to pin-point the exact geographic location that each passenger visited. The only possible tracking that can be conducted is to identify intermediary stations that a passenger passed in a trip, which has been implemented in "Sensor Individuals". In the reality, a person may stop in intermediary station for a while, which is not captured in our data. Thus, the data provided by the MTR Corporation are more suitable for capturing the general trends rather than accurately pinpoint personal interactions.

4.1.2. Computation Power

The huge volumes of the MTR data have a great effect on the computation power of the server. As a result, there will be overhead whenever there is more than one request at the same time. Thus, to relieve this issue, several optimization methods has been implemented, such as:

- Providing a precomputed result of queries that take a long processing time. Thus, if user run this query, we will send them directly the precomputed result
- Storing all the data that have been requested previously by user. Thus, if there is a subsequent request that use the same parameter, our platform can transfer these results rather than running these queries which significantly reduce the overhead.
- Provide indexing to improve performance on queries request

Using this approach, the processing time was able to be reduced significantly. Initially, Passenger Mobility, Station Density, Travel Pattern, and Someone Like You analysis took 2 minutes 13 second, 2 minutes 2 seconds, 3 minutes 33 seconds, and 5 minutes 4 seconds of processing time, respectively. Using this approach, the processing time are reduced into a matter of milliseconds.

The computation power can be further improved by hosting both our platform and database in public cloud services such as Amazon Web Services or, Google Cloud Platform Yet it is not advisable due to the confidentiality of the MTR passenger data.

4.2.Future Improvements

This section will cover the brief description of the application that we plan to develop in the future.

4.2.1. Sensor individuals

In current implementation, sensor individuals only return a CSV containing the detailed path that is taken in each trip. There is a potential to use this data to locate COVID-19 super spreader. The team has developed a search algorithm where the user able to specify an Octopus ID and the date and will get the passenger detailed of anyone that the person encountered during their trip.

As observed in figure 4.1., passenger with ID 905294529 have encountered passenger with ID 904802245 at two separate occasions. However, there still further enhancements needed ad the complexity of current algiruthm is $O(n^2)$. As our dataset consisted of millions of data, it will take a long time to complete the query.

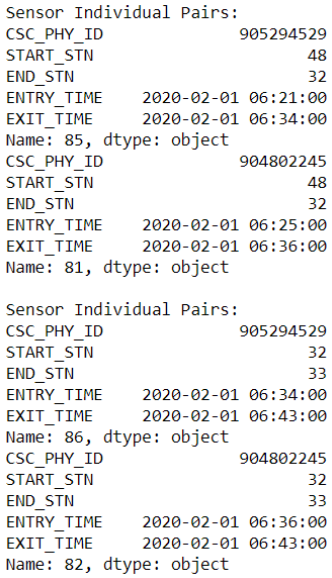

Figure 4.1. Sample of future implementation of sensor individual output

4.2.2. Real Time platform

In the current system, the platform only works with past data from previous months. Thus, it can only be used as a observation of past events. Should there a possibility to get a steady and real-time data from both the MTR Corporation and HKCHP, the system will be able to be developed into a fully-dynamic application with ability to give a real time alerts. Yet, in order to enable this functionality, there should be a new agreement established with the MTR Corporation, understanding of complex data analytics techniques, and high computing resources.

5. Conclusion

The report has covered in detail the background of this project including hypotheses on correlation between the spread of COVID-19 and MTR passenger mobility, methodology taken to complete the project according to the requirements, presented the result which proves the hypotheses, and suggesting potential areas of future research and implementation. The main objective of this project is to provide a platform where users especially researchers or government officials will be able to query data, retrieve visualization, and conduct advance analysis of both COVID-19 data and MTR passenger data in a convenient manner. The information retrieved from the platform can be used as a consideration by government officials and researchers to mitigate the risk of spreading COVID-19 within the MTR routes and station. The final goal of this project is to support the research on “Familiar Strangers” theory and the Big Data field.

References

JavaTPoint. (n.d.). Django MVT. Retrieved April 18, 2021, from https://www.javatpoint.com/django-mvt

MTR Corporation Limited. (2020). 2019 Annual Report of the MTR Corporation Limited. Retrieved from https://www.mtr.com.hk/archive/corporate/en/investor/annual2019/EMTRAR19.pdf

MTR and HKU sign MoU on railway operation big data analysis - Press Releases - Media - HKU. (n.d.). Retrieved October 23, 2020, from https://www.hku.hk/press/press-releases/detail/18893.html

Octopus (n.d.). Standard octopus. Retrieved October 30, 2020, from https://www.octopus.com.hk/en/consumer/octopus-cards/products/on-loan/standard.html

Reactjs (n.d.). Tutorial: Intro to react. Retrieved April 16, 2021, from https://reactjs.org/tutorial/tutorial.html

Tableau (n.d.). Data cleaning: The benefits and steps to creating and using clean data. Retrieved April 18, 2021, from https://www.tableau.com/learn/articles/what-is-data-cleaning

World Health Organization. (2020, January 12). *Novel Coronavirus – China.* Retrieved from https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/

Zhang, F., Jin, B., Ge, T., Ji, Q., & Cui, Y. (2016). Who are My Familiar Strangers? Proceedings of the 25th ACM International on Conference on Information and Knowledge

Management. doi:10.1145/2983323.2983804

Zhou, J., Yang, Y., Ma, H., & Li, Y. (2020). “Familiar strangers” in the big data era: An exploratory study of Beijing metro encounters. *Cities, 97,* 102495. doi:10.1016/j.cities.2019.102495