COMP 4801 – Final Year Project

FYP – Individual Final Report

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

Supervisor: Prof. Cheng Reynold

Prepared By:

Nagra, Harsh (305437707)

Group members:

Ali, Marvin (3035361817)

Effendi, Janice Meita (3035492977)

Jain, Rishabh (3035453608)

Widjaja, Marco Brian (3035493024)

Abstract

The importance of big data is significantly increasing in today's world as it provides us with the opportunity to understand complex scenarios which were not possible to understand until now. Analysing the passenger behaviour in the MTR (Mass Railway Transit) provides us with the opportunity to uncover hidden patterns in passenger travel behaviour with respect to various events that would have had a direct impact on the passengers everyday life. This project aims to correlate the impact of COVID-19 on the busiest mode of transportation in Hong Kong, i.e. the MTR, by using the data provided by the Mass Transit Railway Corporation and Centre of Public Health Hong Kong. In addition to analysing this data, the project aims to provide other researchers with a platform to access and understand this data, and finally share their research results, therefore, providing them with a medium to collaborate. This platform would ensure the full utilization of the MTR data.

Acknowledgement

Firstly, I would like to express my gratitude towards Department of Computer Science, The University of Hong Kong for incorporating the Final Year Project into our curriculum to help us get hands on experience with real world application of our learnings.

Secondly, I would also like to thank our supervisor, Prof. Cheng Reynold, and our project RA, Shivansh Mittal, for providing me with the opportunity to work with him on an extremely interesting project which would not only equip me with a lot of knowledge but also present me with challenges which will help me grow to pursue my passion in the field of Big Data.

Table of Contents

List of Figures

List of Tables

Abbreviations

CHP	-	Centre for Health Protection
COVID-19	-	Coronavirus 2019
MTRC	-	Mass Transit Railway Corporation
HKU	-	The University of Hong Kong
MTR	-	Mass Transit Railway Corporation
DVD	-	Digital Versatile Disk
RDBMS	-	Relational Database Management System
UI	-	User Interface
CSV	-	Comma Separated Values
SSH	-	Secure Shell Protocol
CS	-	Computer Science
ESRI	-	Environmental Systems Research Institute
ArcGIS	-	Aeronautical Reconnaissance Coverage Geographic Information System
OD	-	Origin Destination
JS	-	JavaScript
JWT	-	JSON Web Token
JSON	-	JavaScript Object Notation

1 Introduction

1.1 Background and Motivation

The global pandemic of the COVID-19 started in December 2019 in the city of Wuhan in China (World Health Organization, 2020). Since then our lives have been majorly impacted in every aspect from getting a meal at a restaurant to travelling the world. COVID-19 is as contagious as normal flu but has is about 10 times deadlier and nobody has immunity against it (UCI Health, 2020).

Hong Kong is one of the cities which has successfully contained the outbreak of the virus despite going through four waves starting in January 2020 until recently in March 2021. These four waves are represented in Figure 1.1.1 (Worldometer, 2021) below.

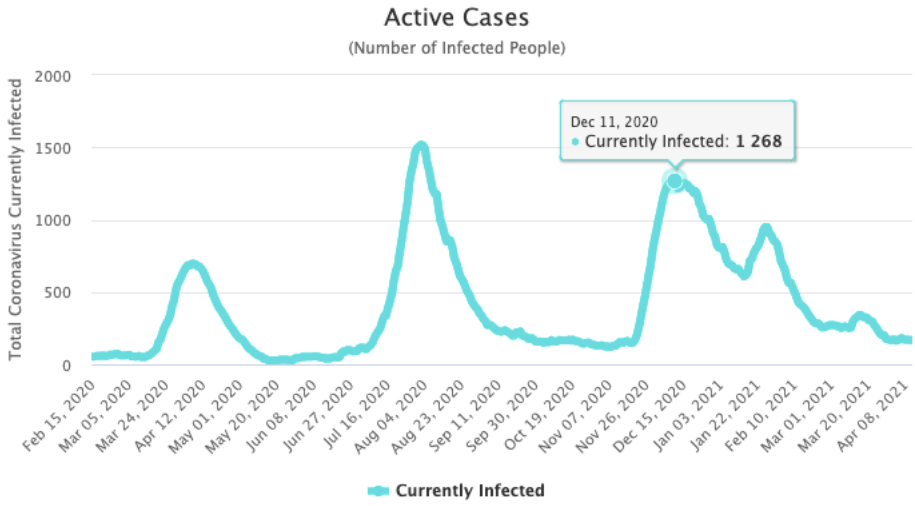

Figure 1.1.1 Active Cases vs Date in Hong Kong

To prevent the spread of the virus many countries imposed social distancing measures such as work from home, less number of people in public gatherings, quarantine, travel ban, etc. to ensure the safety of the public. Hong Kong rapidly responded to the possibility of the spread of the virus in the city and imposed these measures as early as 28th of January 2020 (Centre for Health Protection, 2020) and since then the city has not gone back to the normal routine. Even with these social distancing measures in place, it has been impossible to abstain from the risk of spread of the this virus in commonly shared places such as public transportation.

A study showed that different public transportation accounted for about 90% of passenger trips in Hong Kong every day which is the highest among the 27 major cities in the world (Public Transport Strategy Study, 2017). Among this population, in 2019, 47.4% of people used the MTR, making it the leading transportation medium in the city (2019 Annual Report of the MTR Corporation Limited, 2020). Figure 1.1.2 below displays the distribution of public shares in Hong Kong as of September 2020.

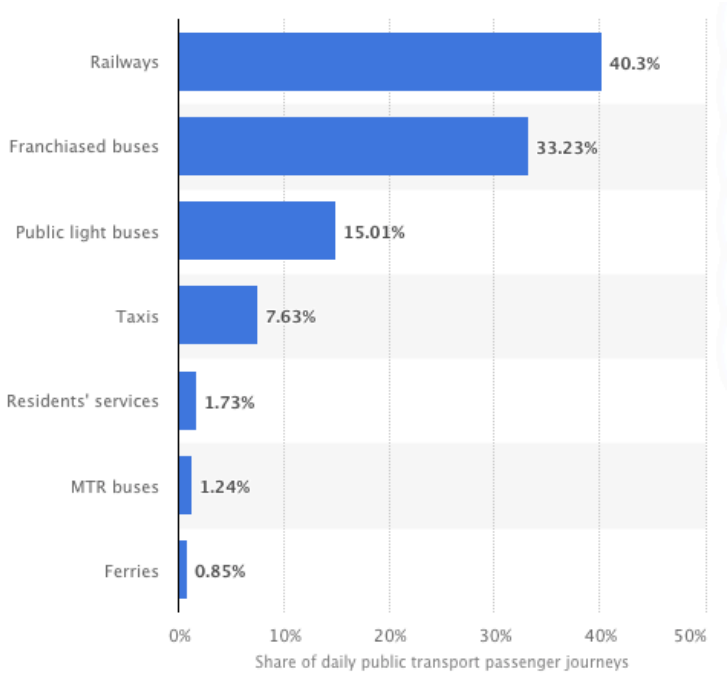

Figure 1.1.2 Avg. Daily Public Transport Journeys in Hong Kong (September 2020)

Now, after understanding the importance of public transportation and social distancing measures to prevent the spread of COVID-19 in Hong Kong, as displayed in Figure 1.1.3 it is very important for the Health Authorities to only impose measures that minimise the health risks but also at the same time ensure that the economy stays afloat. For example, the suspension of public transport like the MTR could have major economic impacts on the city.

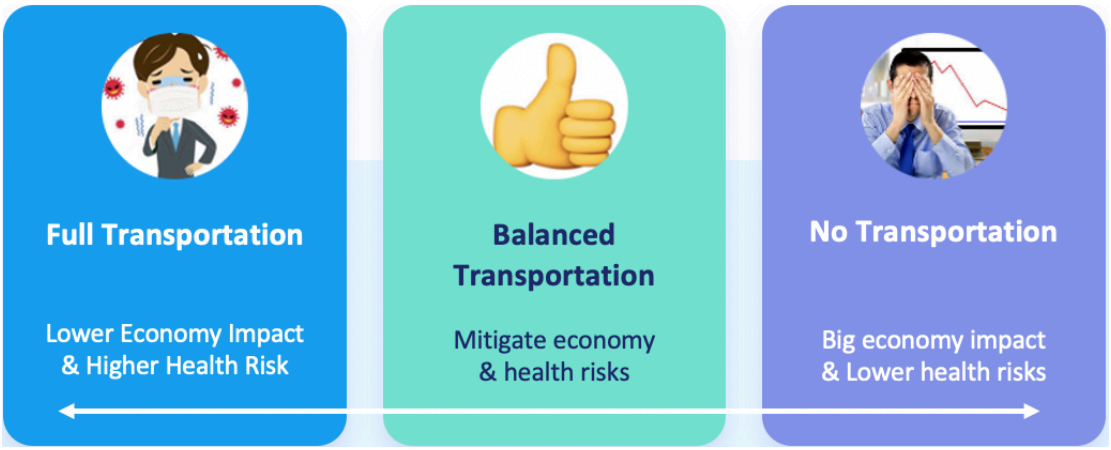

Figure 1.1.3 Balance between health risks and economic impact

1.2 Big Data

The complex and large datasets which are typically unstructured and can be used to analyse significant trends and characteristics are known as Big Data (Boyd, 2011). About $2x10^{18}$ bytes of data is produced every day across all industries in the world (Connall, 2021). These numbers represent the growing significance of data in every industry and also demonstrates the importance of leveraging the statistics that could be gathered from it, as companies like Netflix were able to influence 80% of content viewed by the subscribers due to accurate data insights (Connall, 2021). Whereas, 73.4% of companies still report the adoption of Big Data as a challenge (Connall, 2021).

In collaboration with the Hong Kong Mass Transit Railway Corporation (MTRC), our project aims to leverage the daily transaction data provided by the MTRC during the period January to August 2020. Each month of data comprises about 116 million rows and the size is about 13.6 GB on average. Figure 2.1.1 below represents the structure of the data provided by the MTRC under a confidentiality agreement between HKU and MTRC. The fields in this data include entry station, exit station, entry time, exit time, card type and the customer id. For confidential purposes, the customer id was masked by the MTRC before providing us with the data. This data was provided to us in DVDs, therefore, we believe MTRC just like many other firms has the access to Big Data but does not have a proper distribution or management system for the data.

Passenger ID	Card Type	Entry Station	Exit Station	Entry Time	Exit Time
49564026493	Adult	Central	Sham Shui Po	2020-04-17 08:14:33	2020-04-17 08:41:21
29564820568	Student	Kennedy Town	HKU	2020-04-17 17:07:46	2020-04-17 17:12:53
...	...	...	...	...	...

Figure 1.2.1 Columns of Big Data retrieved from MTRC

1.3 Literature Review

The research paper *"Who are My Familiar Strangers? Revealing Hidden Friend Relations and Common Interests from Smart Card Data" and "Understanding metropolitan patterns of daily encounters"* uncovers the concept of '*familiar strangers'* in a city of 3.02 million and 5 million people, respectively, by using a similar dataset like ours, i.e. the travel smart card transaction data (Zhang, 2016). Familiar Strangers refer to strangers that frequently meet in daily life and have common interests (Zhang, 2016). These research papers leverage high-level mathematical knowledge of the authors and are working on a smaller dataset and have access to more computing power. Whereas, the mathematical knowledge involved in this research would be beyond our scope.

'Someone like you' is a similar concept to *'Familiar Strangers'* but does not cover as many details. The term *'Someone like you'* is defined as the co-presence of smartcard holders who are travelling on similar trajectories i.e. from Station A to Station B for about X1 to X2 times (Jiangping & Yang, 2018). The research paper *"Someone like you: Visualising co-presences of metro riders in Beijing"* visualises the co-presence of metro riders to show and quantify patterns in the city of Beijing with a population of 22 million (Jiangping & Yang, 2018). Figure 1.3.1 displays the results from the research mentioned above. These results do not need as much mathematical understanding, therefore, are in the scope of our knowledge.

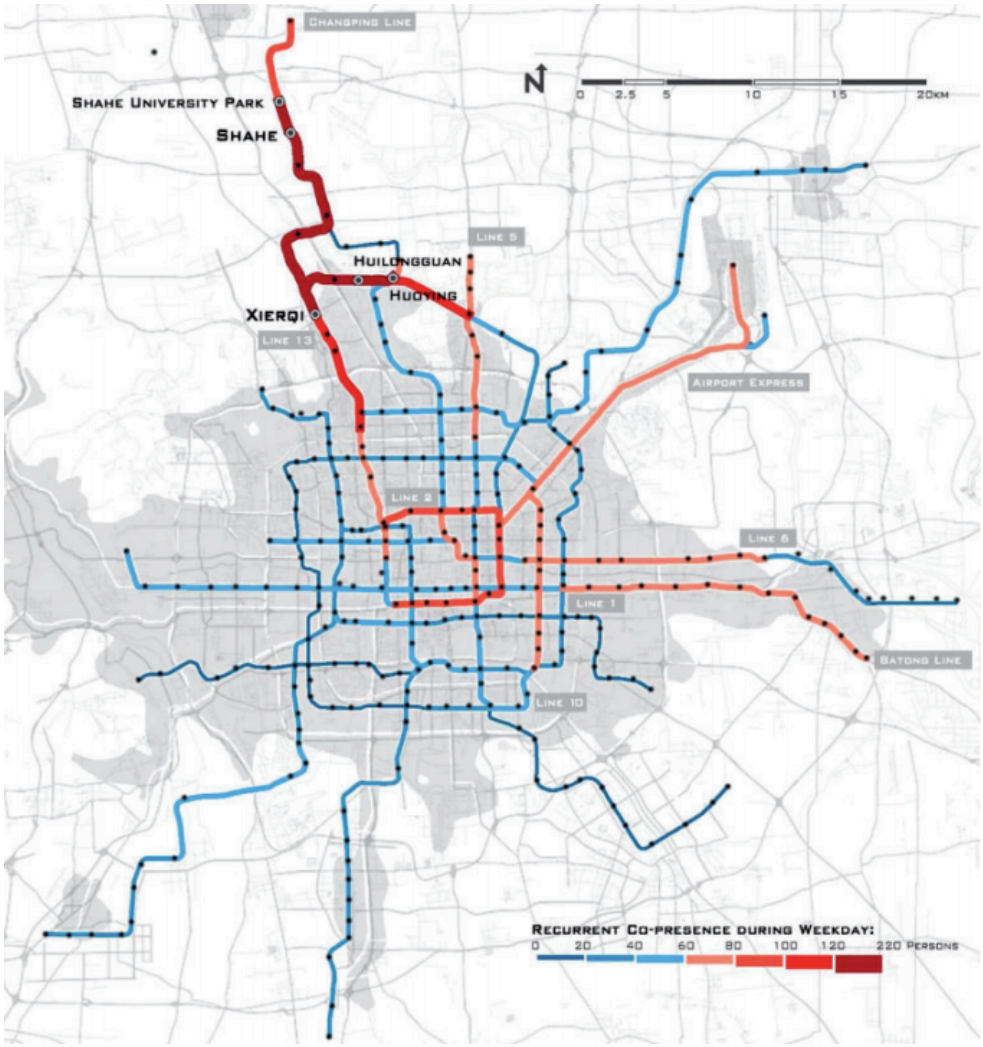

Figure 1.3.1 Co-presence of metro rides on weekdays in Beijing (10-16 August 2015)

Another famous research topic with smartcard transaction data is *'Sensor Individuals'*. This topic focuses on finding people who potentially have the most physical contact with other riders at their respective stations and hopefully at the MTR carriage level. These people could act as super-spreaders if infected by COVID-19 or any other contagious disease.

1.4 Objectives

This project aims at building a centralized repository. In addition to the data provided by the MTR, the centralised repository will also contain COVID-19 cases data provided by the CHP, Hong Kong. This centralized repository aims to contain only pre-processed data that can be directly accessed by any authorised application.

We have utilized this repository in two major ways i.e. analysing it to deduce a relation between COVID-19 and MTR, if any, and building a platform to maximise the utilisation of this platform using various tools.

The platform provides the users with three features. First, a querying tool that lets the users retrieve the data using various filters and methods. Second, a visualization tool to analyse the centralized repository geospatially on the map of Hong Kong. Finally, provide the result of the various analysis conducted by us.

Figure 1.4.1 Platform Objectives

1.5 Scope

To achieve the aforementioned goals in section 1.3, the project makes use of an RDBMS such as MySQL to securely store the data repository i.e. the MTR data and relate it to the COVID-19 data. The project also uses dynamic and scalable technologies such as React, Python Django, Python Pandas, ArcGIS, etc. to build the platform and conduct various analysis. This project aims at delivering a scalable, efficient and user-friendly solution to the current bottlenecks slowing down the research and analysis which could be conducted on this dataset.

1.6 Significance and Impact

The users of the platform mentioned above are other researchers within HKU. This querying tool aims to eliminate the usage of DVDs to distribute the data, as using DVDs is inefficient as well as the hardware required to access DVDs is not available in new systems. The visualizations will help these researchers realise the significance and potential of the data. And finally, the analysis will provide crucial results which can directly be used by them to draw further conclusions.

1.7 Outline

Chapter one provides us with an introduction to the project and covers the brief background and motivation. Following the introduction, the second chapter explains and justifies the key methodologies used in the project and provides other technical implementation details of these methods including the procedure behind the analysis as well as platform development. Chapter three covers the results from the methodologies covered in chapter two and also provides a discussion after each result to summarize the result. Further, chapter four provides us with an overview of the challenges and limitations faced during the implementation of the project. Chapter five provides us with a brief timeline of the project and key events during this period. Finally, chapter six provides us with possible future implementations to scale the project further and chapter seven concludes the learnings and outcomes from the whole project.

2 Methodology

2.1 Overview

This chapter provides a brief summary of the project plans and the techniques used to leverage the big data provided by the MTRC. The project has been implemented using the workflow described in the Figure 2.1.1.

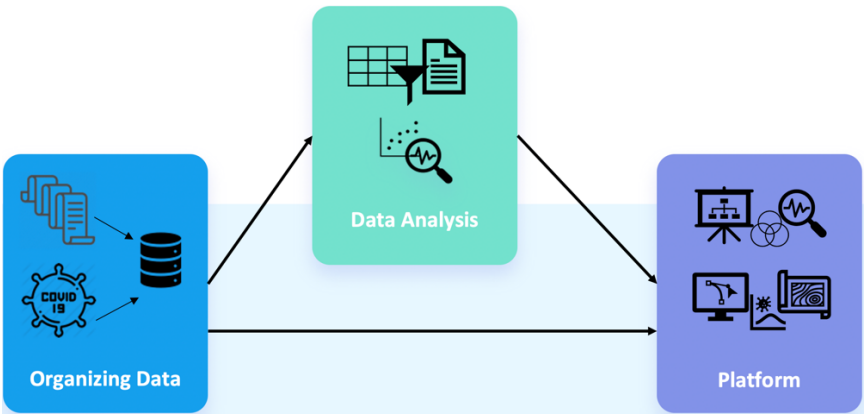

Figure 2.1.1 Project Workflow

The first step and most important step involves the creation of a secure data pipeline to provide access to the pre-processed MTR and COVID-19 data. This pre-processed data must process contain information to closely relate the MTR passenger flow with the geolocations of the spread of COVID-19. Therefore, making it easier for the users to study and relate this data geospatially.

Once the data pipeline is set up, the analysis of these two data sets is a crucial part of the project. The analysis is further divided into two sub-techniques, which are mobility trend analysis and geospatial analysis. The platform provides easier access to the data repository by connecting it to a UI using the data pipeline and also provide the results of our analysis to other users.

2.2 Database Development

The database development process involves three major tasks, i.e. identification, specification and creation of the schema that describes the organization of data in a table. The data retrieved from both MTRC (confidential) and CHP (open-source) was stored in CSV format. To facilitate the processing and analysis of the data, these CSV files were migrated onto a secure MySQL database.

However, before the migration of this data onto MySQL database, the data needs to be pre-processed. The data received from the MTR consisted of two rows for a single trip. The first row accounted only for the entry of the passenger into the MTR and had the same station code for both entry and exit station and the entry timestamp. Whereas, the second row had the actual entry and exit station but not the time stamp for the entry station. Therefore, pre-processing required connecting these two rows to form one row with all the information. Additionally, pre-processing helped us remove any inconsistencies and also account for the missing transactions in the dataset. After pre-processing, the data can be used seamlessly and efficiently without displaying any unexpected behaviour in later stages. Whereas, the COVID-19 data contains a large amount of human error as these CSVs/excel sheets are filling in manually. Ensuring that the data repository consists of clean and pre-processed data only is extremely important to ensure consistent behaviour from data in the later stages.

MySQL is a relational database management system (RDBMS) and is operated with the help of Structured Query Language (SQL). A relation database helps us identify and access data in relation to another piece of data in the database (What is a relational database?, 2020). Figure 2.2.1 represents our MTR database schema. *'<Month><Year>_Simplified'* represents the naming convention of our primary tables containing MTR transaction data. For example, the transaction data for April 2020 would be stored in Apr20_Simplified table. The other tables are *station_codes*, *mtr_geolocation*, and *hospitals_and_closest_mtr*. It is important to build a relational database as it would help us relate these tables. Whereas, a non-relational database (NoSQL) would not relate them as the structure of the database is not enforced and dynamic. A SQL database rather emphasizes a user-defined structure to regulate the relationship between different data fields (What is MySQL?, 2020). Therefore, relation database ensures data consistency such that the required fields are accounted for, there is no

duplication within unique fields and these fields are also guarded for maintaining user defined data type.

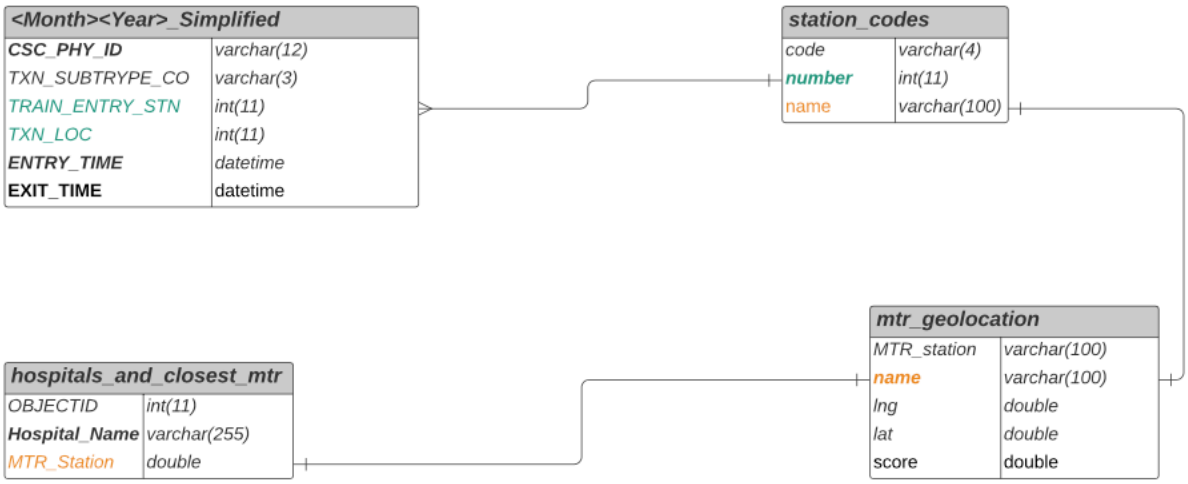

Figure 2.2.1 MTR Data Schema

The second part of our database contains COVID-19 data, which has been retrieved from CHP open-source website and pre-processed before migration (Figure 2.2.2). The table *'clean-cases-<month><year>'* contains details about every patient such as case_no (primary key), report_date, age, etc. and the table *'building list_<month>_geocoded'* contains the information about the building names related to every *case_no* in the former table. In addition to the address of the building, and case id, we have also added fields containing the geolocation (longitude, latitude, etc) of these buildings with the assumption that they will be helpful in the future to conduct a geospatial analysis. These geocoded locations were retrieved using Google Geocoding API and ArcGIS API.

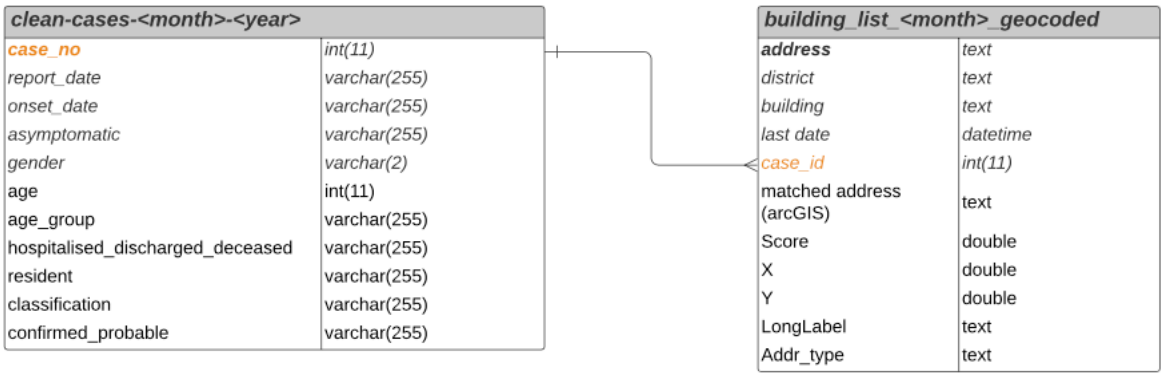

Figure 2.2.2 COVID-19 Data Schema

Building a scalable data repository is important as we will be receiving data from the MTR for the months after August 2020 and COVID-19 data is also updated on regular bases. Using a relational database and the abovementioned schema guarantees the scalability of our data repository. In addition to the scalability, the security of the repository is important as it contains confidential MTR data. In order to preserve data confidentiality, we have hosted the data on a private server named *"HinCare"* which can only be accessed using an HKU CS server. Therefore, to access this data we need to create an SSH tunnel to the HKU CS machine, and another tunnel from the HKU CS machine to the HinCare machine as displayed in Figure 2.2.3.

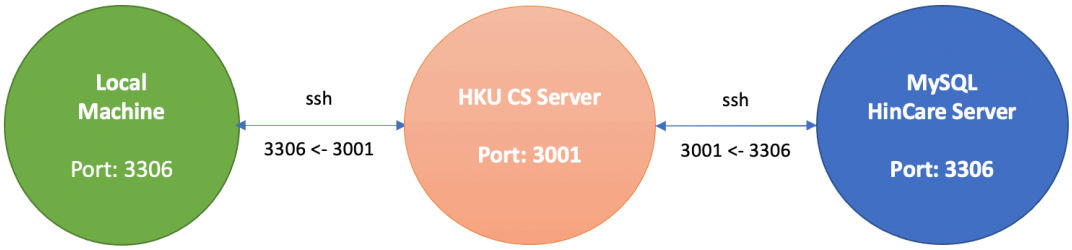

Figure 2.2.3 SSH Port-Forwarding (Tunnel)

2.3 Analysis

This subsection will provide a brief overview of data analysis techniques used to analyse the MTR and COVID-19 data repository.

2.3.1 Mobility Trend Analysis

Mobility trends refer to the general popular movement patterns of the public. Having access to MTR transaction data, conducting such an analysis would understand the data better and also retrieve high level insights. In addition to January – August 2020 MTR transport data, the MTRC provided us with 2019 MTR transaction data for the same months to help us monitor the aggravation of COVID-19 by comparing the transactions statistics for both years. In addition, it will also help us eliminate any yearly trends and not relate them to COVID-19 spread in the city. For example, MTR transactions significantly reduce every year during Chinese New Year due to holidays. Our key focus would be on major events such as

the Wuhan Lockdown, the first COVID-19 case in Hong Kong, and the onset of other waves in the city.

To effectively understand this analysis, we will be plotting these statistics to create graphic visualizations. These visualizations would help the general audience to grasp the trends in an efficient manner.

Further, the analysis can be broken down to target a particular group of MTR passengers based on card type, i.e. adult, child, senior and students. This analysis would reveal how these group of passengers were affected due to the pandemic. To draw a side-by-side comparison, the COVID-19 cases will be subdivided into the same four categories.

We expect to see correlations between the data sets, even though, it might not be possible to pin-point the reason behind the correlation or guarantee the factors leading to this correlation. As the two datasets do not have any relation connecting them each other.

2.3.2 Metrics

We will also be running SQL queries over the MTR data to create three analysis metrics. The three main analysis metrics that we will be using to visualize the MTR data with the help of ArcGIS are Travel Pattern, Passenger Mobility and Station Density.

Travel Pattern analysis takes MTR passenger routes into account (Start Station, Destination Station, Date and Volume of Passengers), this analysis will help us analyse the MTR passenger flow better and also visualise it for easier interpretation. Passenger Mobility analysis aims to capture the general trend in number of MTR passengers in a given time frame (Date and Volume of Passengers). Finally, Station Density analysis would illustrate the passenger density at each station i.e. how crowded every station during a given time frame (MTR station, Date and Volume). The tables for all the respective analysis techniques were created in the Database using SQL queries.

We believe the results from these metrics will come in handy for other researchers to analyse the data without having to process the data themselves, therefore, saving time.

2.3.3 Geospatial Analysis

The limitation of our data repository pushed our team to find other methods to correlate the datasets. Having access to both geocodes of MTR stations and COVID-19 cases, we decided to conduct a geospatial analysis on these datasets.

There are not many open-source sophisticated software available in the market to conduct a geospatial analysis, therefore, ESRI provided us access to their ArcGIS software. ArcGIS is a popular software for providing geographic information system to work with maps and geographic information (Figure 2.3.1). ArcGIS also works well with large amount of complex data, therefore, making it the perfect platform for our geospatial analysis.

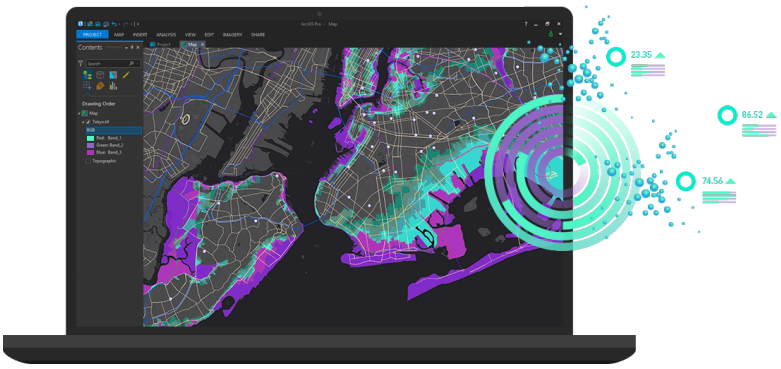

Figure 2.3.1 Sample image ArcGIS software

Using ArcGIS we aim to visualize the combination of COVID-19 dataset, i.e. the COVID-19 cases table alongside the building details with geolocations. The resulting geographic visualizations would represent the heat maps of COVID-19 spread in Hong Kong.

Our final aim is to combine the analysis on the two datasets mentioned above i.e. relating COVID-19 heat map to the three different analysis metrics (mentioned in the previous section) on MTR dataset over a given time frame. We hope this would provide us a holistic view of how these travel patterns, station densities and passenger mobility patterns were affected.

2.4 Contact and Behaviour based research

The contact and behaviour based research will only focus on the MTR dataset. The purpose behind this research is to find relations within the MTR passengers. For instance, this research will help us project the impact of one passenger's journey onto the other passengers, i.e. if the former is found to be COVID-19 positive, is it possible to track/predict the impact on the latter. To perform this analysis our team focused on the following two techniques.

2.4.1 Someone like you

As mentioned in section 1.3, *'Someone like you'* is defined as co-presence of smartcard holders who are travelling on a similar trajectories i.e. from Station A to Station B for about X1 to X2 times (Jiangping & Yang, 2018). For example, two riders who share a ride between Kennedy Town and Mong Kok around the same time of the day at least once a day over a given period of time will form a *'someone like you'* pair.

The process to find these *'someone like you'* pairs involves 5 steps, which are displayed with an example in Figure 2.4.1 and Figure 2.4.2. First, is the formation of station pairs. Our database contains station codes for 120 MTR stations, i.e. a total of 14,280 (120 times 119) station pairs. Second, step involves querying and the trips from each station pair together. Third, we need to divide the day into equal time intervals such that we can count the trips sharing approximately the same entry and exit time for each of the station pairs mentioned above.

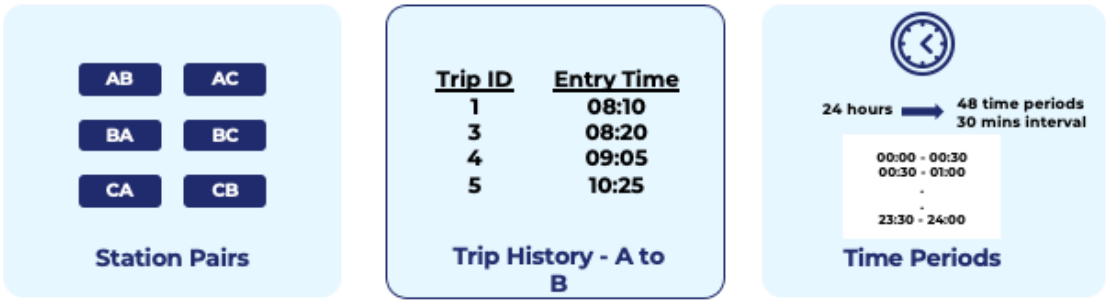

Figure 2.4.1 Step 1, 2 & 3 of *'Someone like you'* analysis

The penultimate step requires calculating the total number of trip shares throughout the day for every station pair. The fifth and the final step averages the total number of trip shares for all days in a selected time period for every station pair. The analysis would be

conducted separately on weekdays and weekends (or holidays) as the trip patterns between the two could vary to a great extent.

Figure 2.4.2 Step 4 & 5 of *'Someone like you'* analysis

In conclusion, *'someone like you'* analysis would provide us with the statistics for passengers with the same travel pattern, therefore, correlating them. By making changes in the analysis steps above, we could also find the *'someone like you'* individuals for a particular passenger i.e. the result would contain customer ids. This result would not be useful in our scenario as these id's are masked, therefore, not having any real life application.

Here are a few assumptions used in *'someone like you'* analysis:

1. Passengers will be regarded as *'someone like you'* even if they do not travel in the same carriage.
2. Every passenger with the same entry and exit station around the same time is assumed to have taken the same route, even if there are more than one possible ways to reach the destination from the source station.
3. The entry and exit time are assumed to the entry and exit time into the MTR carriage and not the entry and exit time of the paid area of the station.

2.4.2 Sensor Individuals

'Sensor Individuals' as a topic is not as explored as *'Someone like you'*, therefore, there are not many previous research papers available publicly. The details regarding this method have been provided by the supervisor in form of a summary.

As mentioned in section 1.3, *'Sensor Individuals'* are people who potentially have the most physical contact with other riders at their respective stations and hopefully at the MTR carriage level. Therefore, this analysis is not as simple as *'Someone like you'* as it requires the

details of the path followed by the passenger along with the timestamps, whereas, the information in our dataset was enough to perform *'Someone like you'* analysis.

The procedure to create *'Sensor Individuals'* requires involves three steps. The first and most important step is to compute the shortest path between two stations, i.e. calculating the shortest path by O-D pairs. Second, we need to pre-process the data to convert it from the form (Start Station, Exit Station) to (Start Station, .. stations in between .., Exit Station). In addition to adding the stations in between, we also need to add a time stamp for the time the passenger will be at each of these stations as explained in Figure 2.4.3. To calculate this path and the timestamps we created a weighted undirected graph utilizing the library networkx provided by python where each node was a station and the travel time between each station was the weight. After the graph was ready, we used Dijkstra's algorithm to find the shortest path between the two stations.

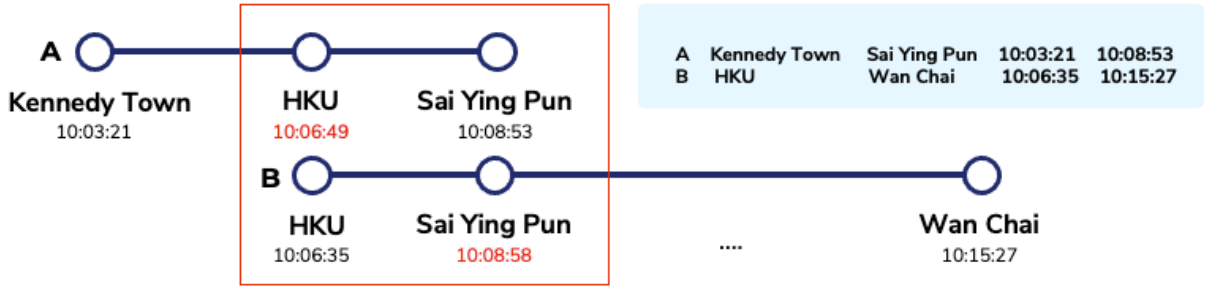

Figure 2.4.3 Full path taken by the passenger with timestamps (Step 1 & 2)

Finally, we need to uniformly distribute these passengers over *n* carriages, as we do not have a method to find out the actual carriage number of each passenger. For example, as shown in Figure 2.4.4 below, 18 passengers between the time 10:00 – 10:05 could be distributed into 8 carriages by randomly assigning 2 passengers in 6 cars each and 3 to two cards each.

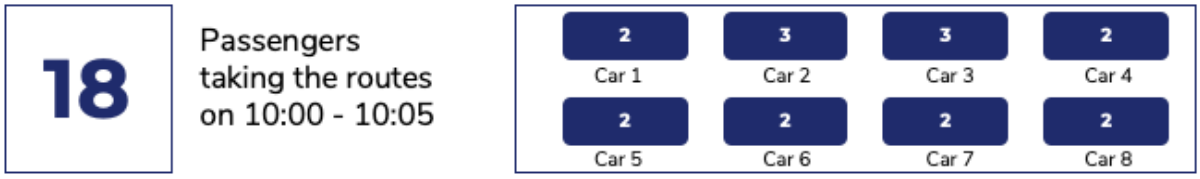

Figure 2.4.4 Passenger distribution in carriages

Once the data is pre-processed, the results would be in the format [A,(Kennedy Town, 10:03:21,2),(HKU,10:06:49,2),(Sai Ying Pun, 10:08:53,2)] where [Passenger ID, (Station Name, Time Stamp, Carriage n), (Station Name, Time Stamp, Carriage n), ...] represents the format of the data within the squared brackets.

Therefore, we could efficiently visualize the spatio-temporal pattern of co-presence phenomena of the MTR passengers. These visualizations would help us further single out the possibility of super-spreaders in the public transportation system.

Here are a few assumptions used in *'Sensor Individuals'* analysis:

1. The graph has been modelled with the assumption that the MTR network is one big line with several branches, therefore, trips from one station to another station from different lines are considered as identical (For example, Central to Admiralty could be done of two different lines but is considered as one).
2. The time taken to switch lines is not taken in consideration.
3. The entry and exit time are assumed to the entry and exit time into the MTR carriage and not the entry and exit time of the paid area of the station.
4. As the time calculated by our model is not the exactly same as the time taken by the passenger, we divide this time difference onto each sub-trip.

2.5 Platform

To provide access to the data as well as our findings, we decided to build a scalable and efficient full-stack web based platform. The process involves development of the platform, i.e. the frontend and the backend, and deployment of the platform.

2.5.1 Development

The frontend of the platform will be developed using React JS, a popular open-source JS framework made by Facebook. React JS allows users to develop UIs for web projects on a component basis. The reason behind choosing React JS was to ensure the scalability of the platform as well as leverage the component to write clean code. We are using react bootstrap for the styling of the User Interface. To integrate the frontend to the backend, we will utilize another JS library, known as Axios. The frontend will be retrieving MTR and COVID-19

data as well as our advanced analysis results from the backend. To retrieve the dynamic visualizations we will be leveraging the JavaScript API provided by ESRI's ArcGIS software (arcgis-js-api), the data required by this API is fetched from the backend.

The backend of the platform is developed using Django, another popular open-source python-based web library which is based on model-template-view architecture. Django was our top preference for backend development as we can leverage its model-template-view architecture which will closely connect our backend with the database and we can also utilize Django restful framework to create endpoints. In addition to the rest framework, we are using JWT authorisation to protect the endpoints. The JWT authorisation is implemented using Simple JWT library provided by Django rest framework.

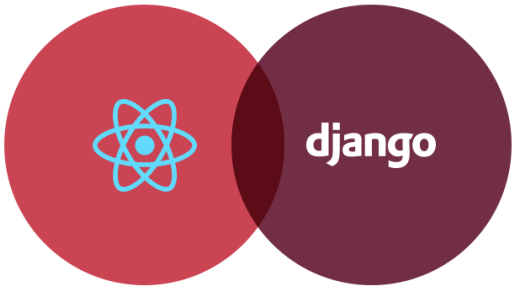

Figure 2.5.1 React JS & Django logos

2.5.2 Deployment

One of the major challenges of our project was working with a database hosted on a private server to protect the data. To ensure the security of the data, we hosted the backend on HinCare server as well. To deploy this backend to the public, we are using an SSH tunnel, this tunnel connects our FYP Machine to the HinCare server which can only be accessed through the HKU CS Server (the flow has been described in Figure 2.5.1 below). The backend can be accessed at http://fyp20035s1.cs.hku.hk:8080 (i.e. Port 8080 of the FYP machine). The endpoints provided by the our backend are protected by JWT authorisation or Django-Admin Portal default authorisation. To access these endpoints, the user will need an account which can only be created by our project supervisor (superuser) using the Django-Admin portal. Whereas, deploying the frontend was rather simple, as we had to run the React App at Port 80 of the FYP machine, i.e. the UI is available to the public at http://fyp20035s1.cs.hku.hk/ .

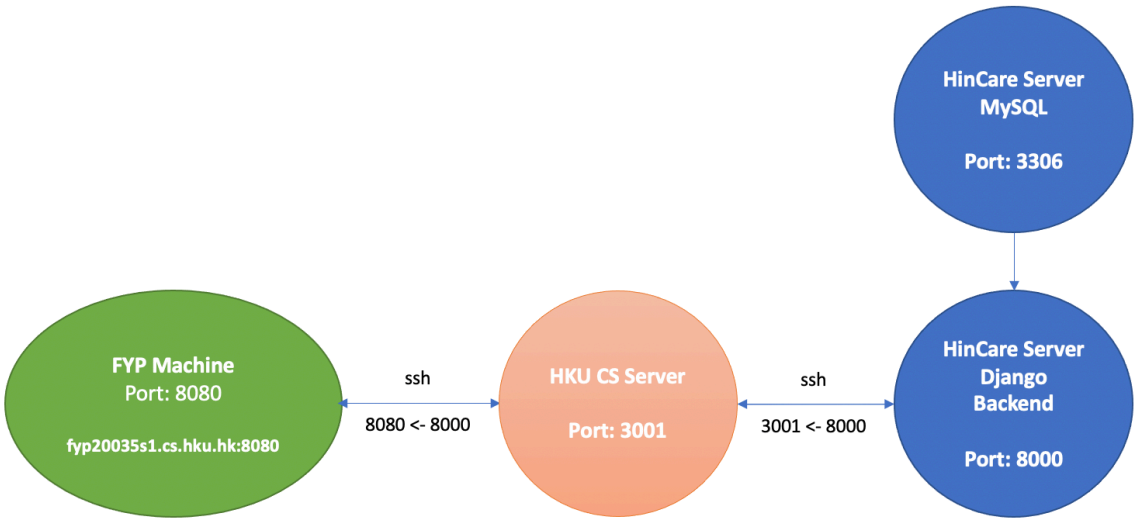

Figure 2.5.1 Django Backend SSH Tunnel

2.6 Summary

In conclusion, this chapter provided us with the workflow of our project. Following the described workflow was extremely crucial to the project due to the co-dependency of each task. Database development acts as the backbone of the project, and the analysis further plays an important role in the results presented by our platform. In addition to the workflow, it also covered the brief technical implementation and theoretical details about the project. The engineering decisions were justified in each section, i.e. the reason behind the choice of technologies used through the course of the project. The implementation challenges faced were also mentioned in the respective sections.

3 Results and Discussion

3.1 Overview

This chapter reports the final results and findings concluded by our project over the period of our final year. These results would follow a similar flow as the chapters above, starting with the initial insights from database development, followed by the analysis results and finally, the platform to provide access to the two. The chapter will also include discussions concluded from these results.

3.2 Initial findings

After pre-processing the data received from the MTRC, we found 1 million incomplete transactions in the MTR dataset, representing, a person entered the station but never exited it (exit station had null value). We believe this must be due to technical errors faced by the passengers or due to other human errors such as two passengers exiting at the same time. After the pre-processing, in order to get an estimate of the total number of passengers and their distribution, we ran a few queries and the results of these queries are displayed below in Figure 3.2.1.

Table 3.2.1 MTR passenger distribution (January to March 2020)

Subtype	January	February	March	Jan - Mar
Adult	5,972,974	3,734,522	3,826,613	6,541,538
Child	513,763	159,080	190,632	559,352
Disabled	99,245	73,892	78,356	109,225
Senior Citizen	1,062,943	655,453	692,643	1,193,492
Student	358,217	239,646	251,516	375,937
All	8,047,967	4,905,746	5,082,557	8,810,951

3.3 Analysis findings

The analysis findings are distributed into two parts, as mentioned in section 2.3 (mobility trend analysis and geospatial analysis) and section 2.4 (someone like you and sensor individuals). For the mobility trend and geospatial analysis, we vastly focused on January to April 2020 data as it was conducted towards the start of our project.

3.3.1 Mobility Trend and Geospatial Analysis

To get started with the mobility trend analysis, we plotted the MTR using Python with the help of Pandas and Plotly. The result is displayed in Figure 3.2.2 below. We can see a clear shift in the pattern starting 21st January 2020 onwards, we believe this is a result of the regulations imposed by the Hong Kong government after the first case of COVID-19 was reported in the city.

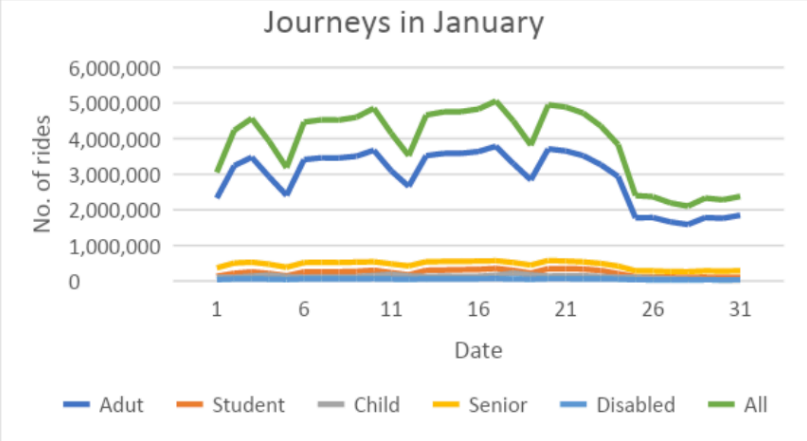

Figure 3.3.1 January 2020 MTR Data plotted using Python.

After realising the shift in trends around the same period as the first few COVID-19 cases in the city, we plotted another graph to compare the number of COVID-19 cases against the total number of MTR passengers. The result of the comparison is displayed in Figure 3.2.3 below. The grey area in this figure highlights the period in which the number of MTR passengers dropped by 44%, right after emergence of first few COVID-19 cases in the city.

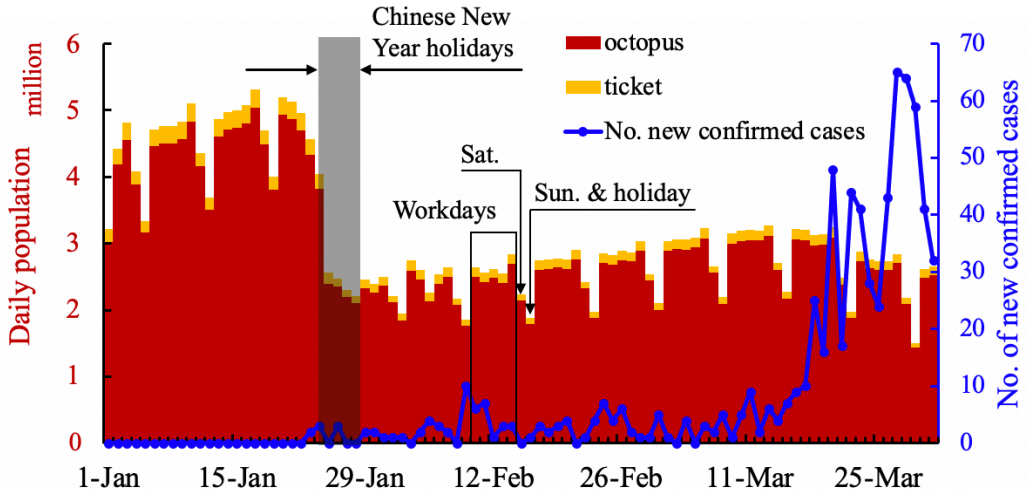

Figure 3.3.2 COVID-19 Cases and number of MTR passengers vs Day

To further analyse the relation between the two datasets, we generated another visualisation displayed in Figure 3.2.4 below. This visualisation focuses on different age groups of MTR passengers and the reported COVID-19 cases for the same age groups. The grey area represents a correlation between the two graphs, representing a drop in MTR passengers and COVID-19 cases during the weekend. We cannot pin point this relation to any specific reason as onset date of COVID-19 varies from person to person, therefore, eliminating the possibility of less cases due to less MTR passengers flow.

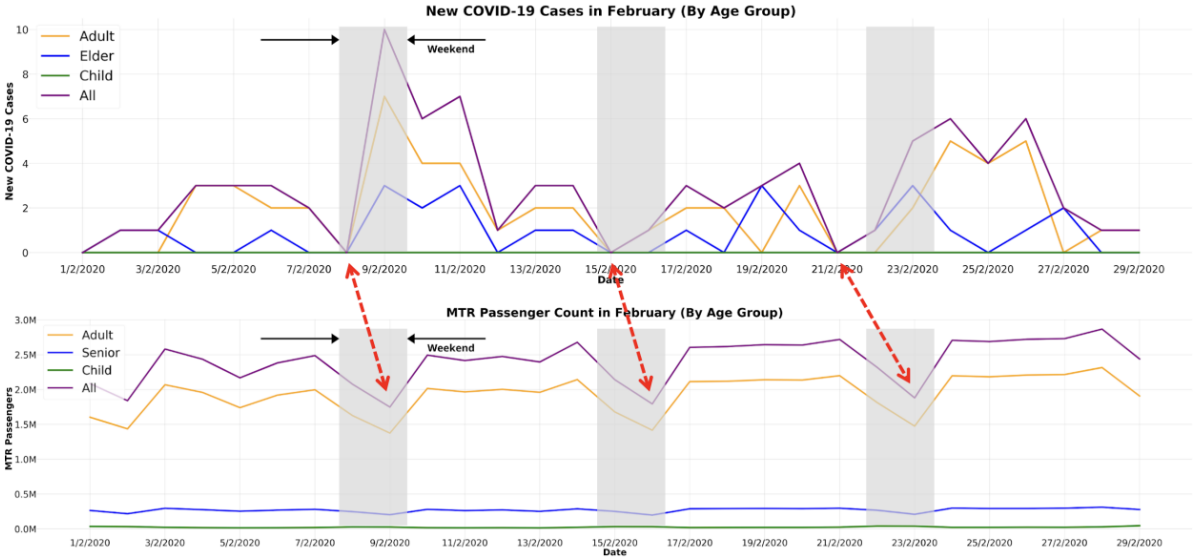

Figure 3.3.3 New COVID-19 cases and MTR passengers vs Date

Moving onto the Geospatial Analysis, we leveraged the capabilities of ArcGIS to visualize the MTR data and COVID-19 spread at the same time. Figure 3.3.4 below visualizes station densities of the most dense MTR stations and the COVID-19 spread using heat patterns. The station densities represent how crowded every station is and the heat patterns represent the affected buildings due to COVID-19 cases. The visualization does not indicate any clear pattern, but the stations Tsim Sha Tsui and Wan Chai have one of the highest station densities and also represent hot spots for COVID-19 spread in the month of April 2020. We believe the correlation is due to the high population density in the two areas, resulting in more number of MTR passengers and easier spread of COVID-19. There is no evidence blaming the spread on the MTR transportation system.

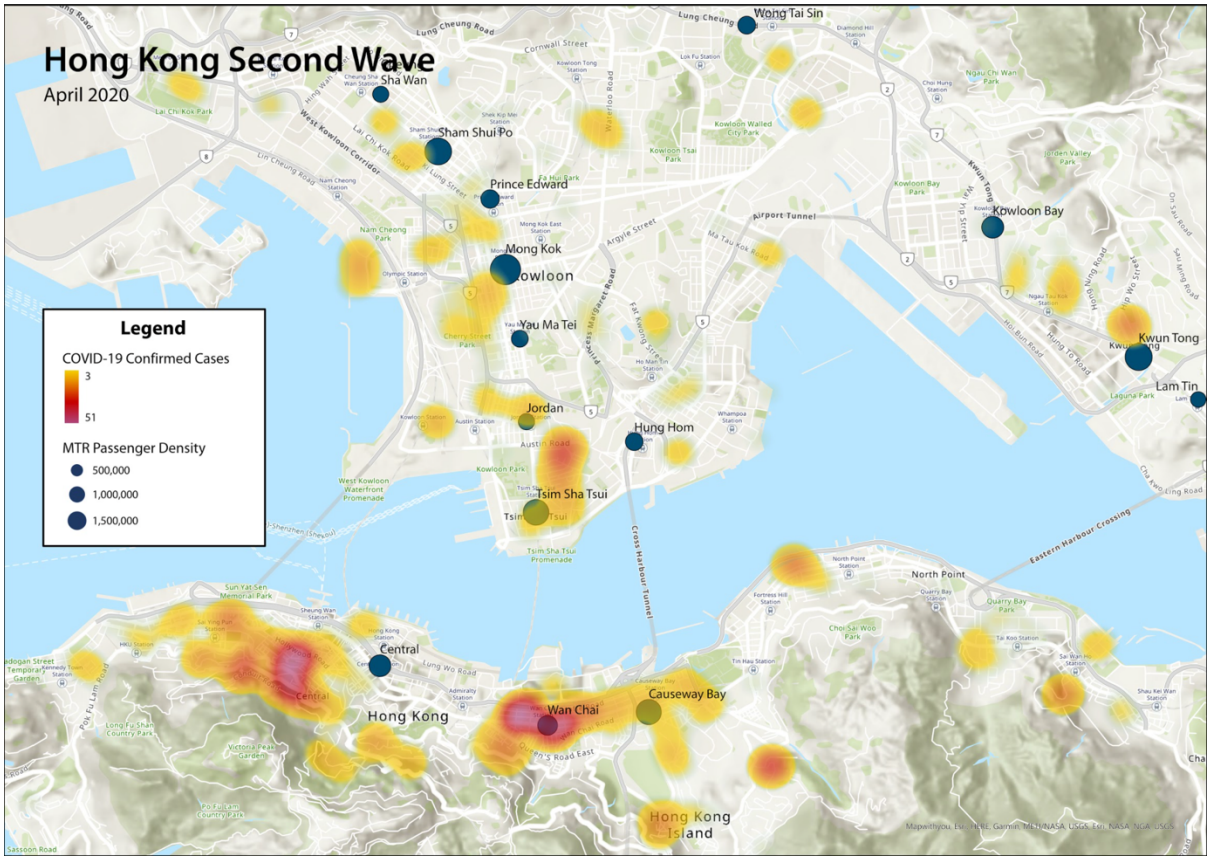

Figure 3.3.4 MTR station density and COVID-19 hotspots in April 2020

Another geospatial visualization was to analyse the most popular travel routes along with the COVID-19 confirmed cases heat map for the month of April 2020, as shown in Figure 3.3.5 below. From this figure, we observe a trend wherein confirmed cases tend to spread across some of the busiest routes. The effect is most prominent between the following MTR stations: Causeway Bay and Central, Sham Shui Po and Mong Kok, Mong Kok and

Tsim Sha Tsui, and Kowloon and Kwun Tong. We realise that these correlations are highly dependent on the local population of different areas which in turn, has a direct impact on total MTR passengers in and out from that area. These populated areas also account for the most COVID-19 cases during the initial few months of 2020.

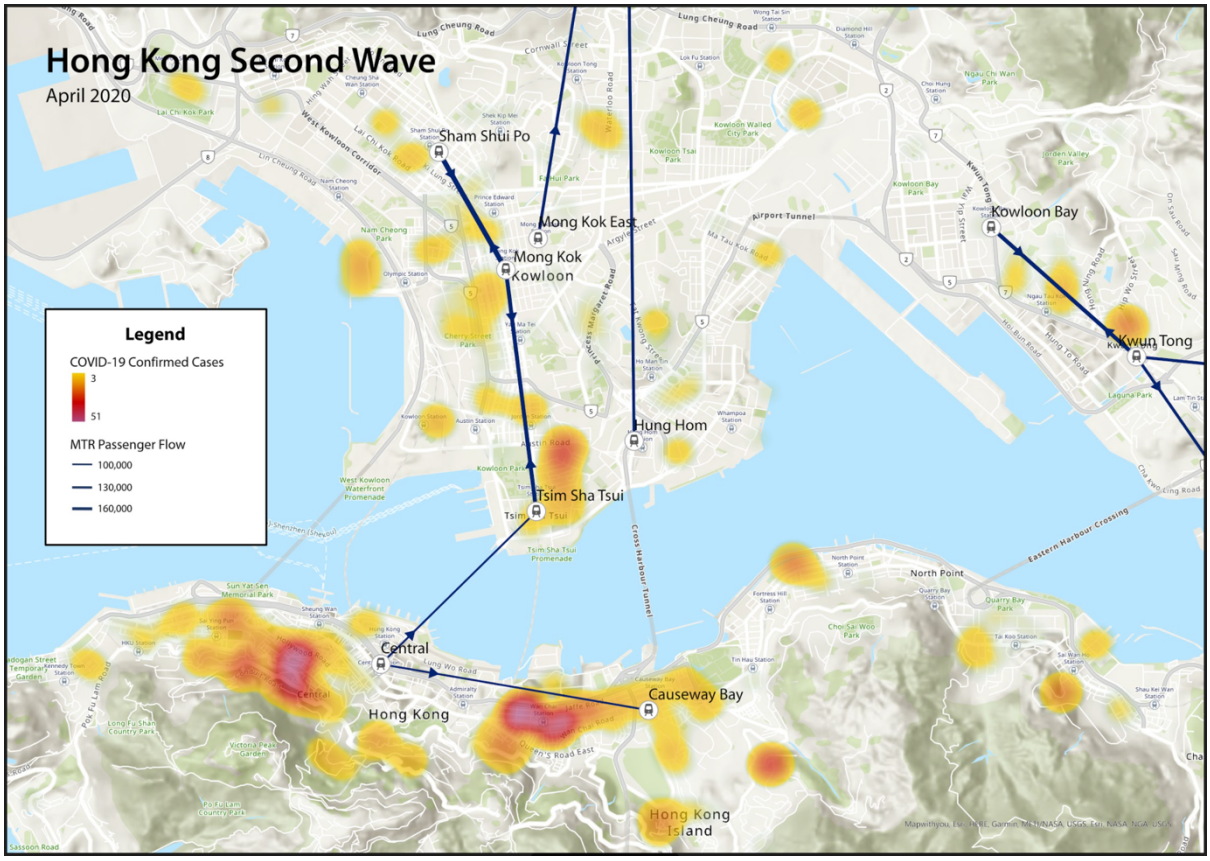

Figure 3.3.5 Popular MTR routes and COVID-19 hotspots in April 2020

3.3.2 Analysis Metrics

After scripting the SQL queries for each of the metrics, we created tables for them in our database hosted on our private server. These tables would provide easy and fast access to the results of these queries for any further analysis.

Figure 3.3.6 represents the first few rows from the result of passenger query. The two columns represent the date and the number of passengers on that day. On an average, passenger mobility query took about 2 minutes 13 seconds on one month of data, i.e. to run this query on 8 months of data could take about 20 minutes.

	ENTRY_TIME	VOLUME
▶	2018-12-31	14508
	2019-01-01	6932213
	2019-01-02	9284138
	2019-01-03	9530735
	2019-01-04	5086404
	2019-01-05	4506838

	VOLUME	ENTRY_TIME	CARD_TYPE
▶	5379980	2019-01-01	ADL
	263205	2019-01-01	CHD
	728657	2019-01-01	SEN
	2289422	2020-01-01	ADL
	102263	2020-01-01	CHD
	366405	2020-01-01	SEN

Figure 3.3.6 Passenger Volume daily and daily with card type results

Figure 3.3.7 represents the first few rows from the result of station density query. The three columns represent the date, station code and the number of passengers on that day in a particular station. On an average, station density query took about 2 minutes 2 seconds on one month of data, i.e. to run this query on 8 months of data could take about 16 minutes.

	ENTRY_TIME	ENTRY_STN	VOLUME
▶	2019-01-01	1	171993
	2019-01-02	1	258222
	2019-01-03	1	268814
	2019-01-04	1	143335
	2019-01-05	1	89790
	2019-01-06	1	87155

Figure 3.3.7 Station Density results

Figure 3.3.8 represents the first few rows from the result of travel pattern query. The three columns represent the date, entry station code, exit station code and the number of passengers on that day with the same entry and exit station. On an average, travel pattern query took about 3 minutes 33 seconds on one month of data, i.e. to run this query on 8 months of data could take about 28 minutes.

ENTRY_TIME	ENTRY_STN	EXIT_STN	VOLUME
2019-01-01	1	2	3434
2019-01-02	1	2	4763
2019-01-03	1	2	5898
2019-01-04	1	2	2965
2019-01-05	1	2	1411
2019-01-06	1	2	1880
2019-01-07	1	2	2663

Figure 3.3.8 Travel Pattern results

3.3.3 Someone like you and Sensor Individuals

After conducting the mobility trend and geospatial analysis, we moved onto the contact and behaviour based research. This results from this analysis were not in the form of visualizations but rather, in the form data which could further be utilized by other researchers to perform more detailed mathematical and complex visualization techniques.

Figure 3.3.6 below represents the sample output from Someone like you analysis. We have written an algorithm in Python to process this query with the help of SQL, and the same code is being used in our python-based backend. The result from this algorithm includes the following columns which are the *date* (date of the analysis), *weekend* (whether the date is a weekend or not), *entry_stn* (entry station of passengers), *exit_stn* (exit station of passengers) and *freq_slu* (total number of *'someone like you'*). The last column represents the result from the *'someone like you'* query, representing the total number of *'someone like you'* individuals from *entry_stn* to *exit_stn*. The input parameter for the algorithm is the period during which we are looking for *'someone like you'* individuals and optionally asks for a particular station pair, otherwise returns the results for all the station pairs. Since, we are working on a huge dataset these queries could end up taking 5-10 minutes if the station pair is not mentioned.

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

Figure 3.3.9 Someone like you sample output

These results could be used to create visualizations similar to the one's created by the paper *"Someone like you: Visualising co-presences of metro riders in Beijing"* to quantify these patterns on the map of Hong Kong. These results would inform us about the routes with the most *'someone like you'* individuals.

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

Figure 3.3.10 Sensor Individuals sample output

Figure 3.3.7 above represents our results from the *'Sensor Individuals'* analysis. This analysis algorithm was more complex than *'Someone like you'* and therefore, the running time for the algorithm was much more as well. To shorten the running time, we created a sub-dataset to run this algorithm. The algorithm retains the format of the input (i.e. the MTR data) and contains the following rows: index (serial no.), csc_phy_id (masked customer id), start_stn (entry station), end_stn (exit station), entry_time (entry time) and exit_time (exit time). The input row (904921192, 3, 27, 2020-02-01 19:27:00, 2020-02-01 19:35:00) was broken down into two sub-journeys, (904921192, 3, 2, 2020-02-01 19:27:00, 2020-02-01 19:32:00) and (904921192, 2, 27, 2020-02-01 19:32:00, 2020-02-01 19:35:00) as represented in the table below. The total number of sub-journeys will depend on number of stations the MTR goes through between the entry and exit station. Therefore, the output finally contains

the expanded input data as it breaks down each row into multiple rows to represent sub-journeys with the timestamps and this data is finally returned in the form of a CSV file by the algorithm. This algorithm was not integrated with our platform due to its poor performance on our large MTR dataset.

These results could be leveraged by the other researchers to efficiently visualize the spatio-temporal pattern of co-presence phenomena of the MTR passengers as mentioned in section 2.4.2.

3.4 Platform

The platform was successfully completed and deployed to the public at the address http://fyp20035s1.cs.hku.hk/ as shown in Figure 3.4.1. The platform is visible to the public but can only be accessed by users authorized by our supervisor.

Figure 3.4.1 Platform login page

The platform is divided into three main sub-sections as well, focusing on the three areas of development of our project as shown in Figure 3.4.2 below. First, focuses on providing users with access to our database with our query tool. Second, provides high level insights of the dataset using geospatial visualizations. Third, lets the user retrieve the results from our contact and behaviour based research algorithms.

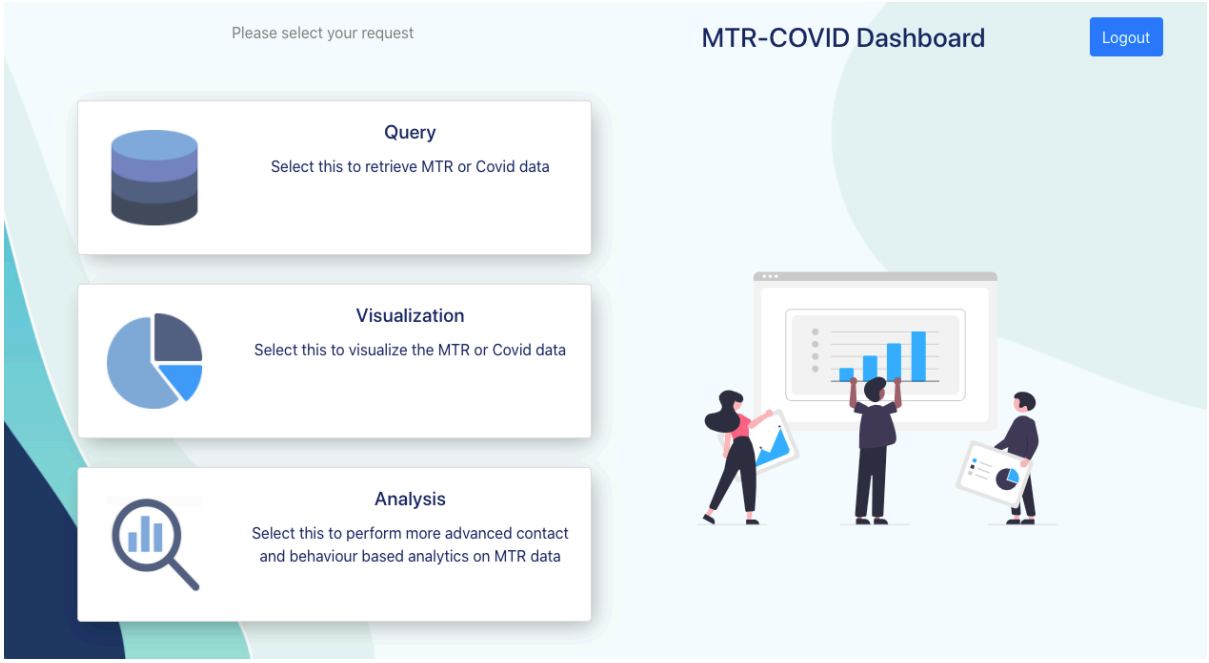

Figure 3.4.2 Platform home page

Figure 3.4.3 and Figure 3.4.4 represent our query tool to retrieve data from our database in the form of CSV files. For our selected partial query in Figure 3.4.3 the result has been downloaded in the form of a CSV file in the bottom left corner. The user can click on the metrics drop down menu and select one of the following options: Travel Pattern (daily or by hour), Station Density (daily or by hour), Passenger Mobility (with or without card type) and Custom MTR query. The first few options provide access to the three metrics discussed in section 2.3.2 and the last option provides direct access to the MTR data containing all the transaction data of every passenger (displayed in Figure 3.4.4). All the parameters are optional i.e. the user can use them partially or decide not to use them at all.

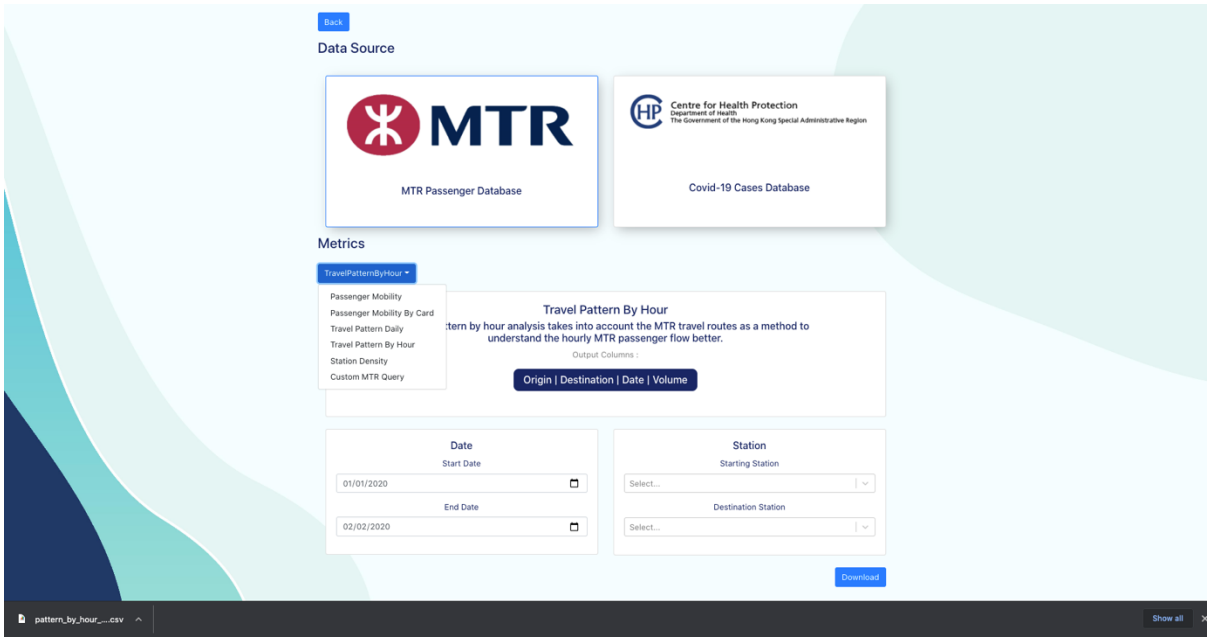

Figure 3.4.3 Platform – Travel Pattern query

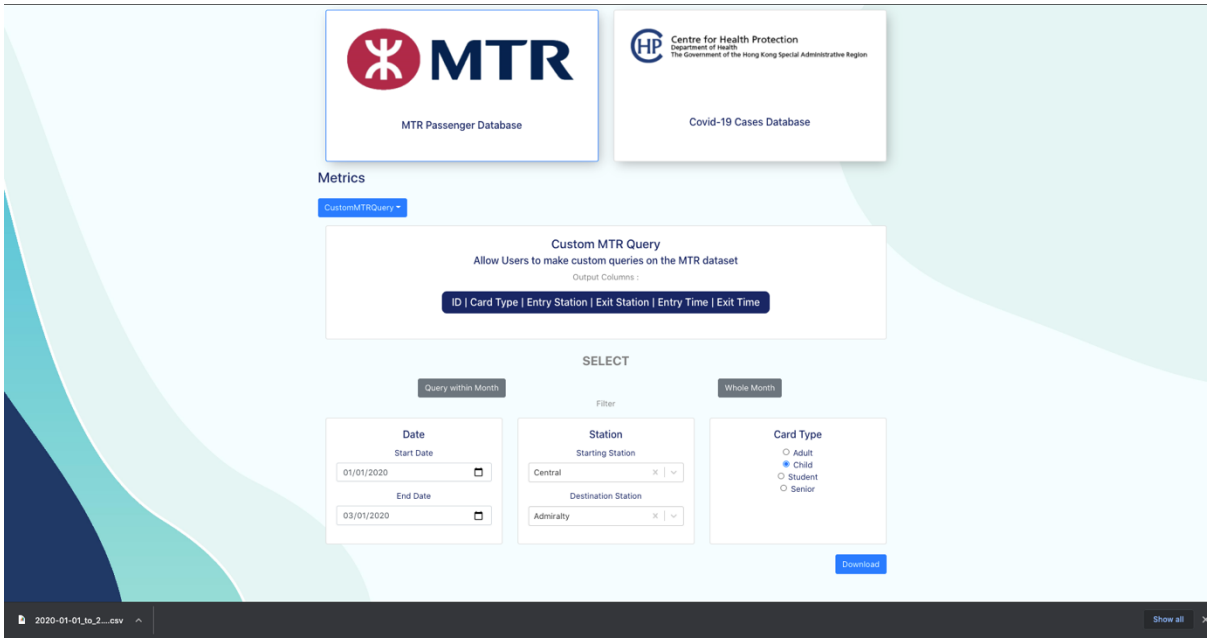

Figure 3.4.4 Platform – raw MTR data query

The second section of the platform represents the dynamically created geospatial visualizations as shown in Figure 3.4.5 and Figure 3.4.6. These visualizations are built on the ArcGIS API provided by ESRI and the data required for the visualizations is retrieved from the backend. Figure 3.4.4 represents the dynamic geospatial visualization of Station Density and COVID-19 during the period 01/01/2020 to 15/04/2020 with the top 10 populated stations during this period. The visualization displayed in Figure 3.4.5 requires the user to input the period, select weather they need the COVID-19 geolocations to be plotted and mention the number of most populated stations.

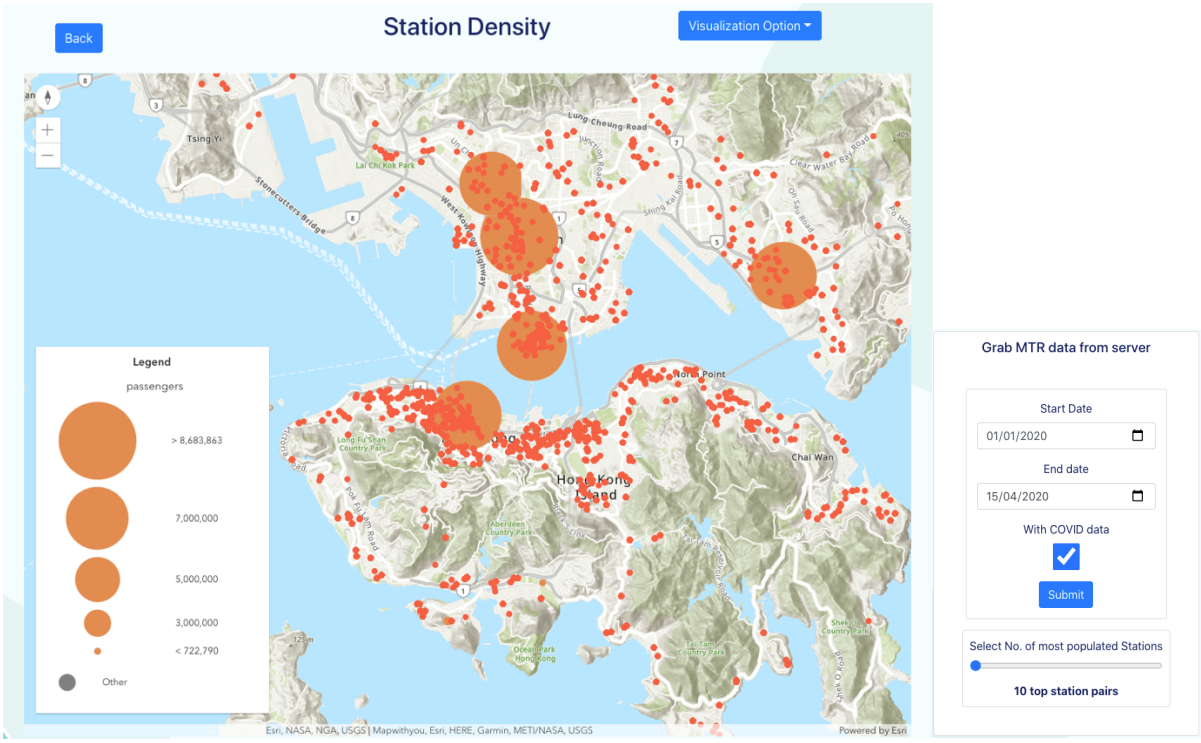

Figure 3.4.5 Station Density and COVID-19 (01/01/2020 – 15/04/2020)

Figure 3.4.6 represents the second dynamic geospatial visualizations of Travel Pattern and COVID-19 for the period 01/01/2020 to 01/04/2020 with the top 160 most populated station pairs. These stations pairs in the visualization are actually connected using their geolocations and not the actual path followed by the MTR.

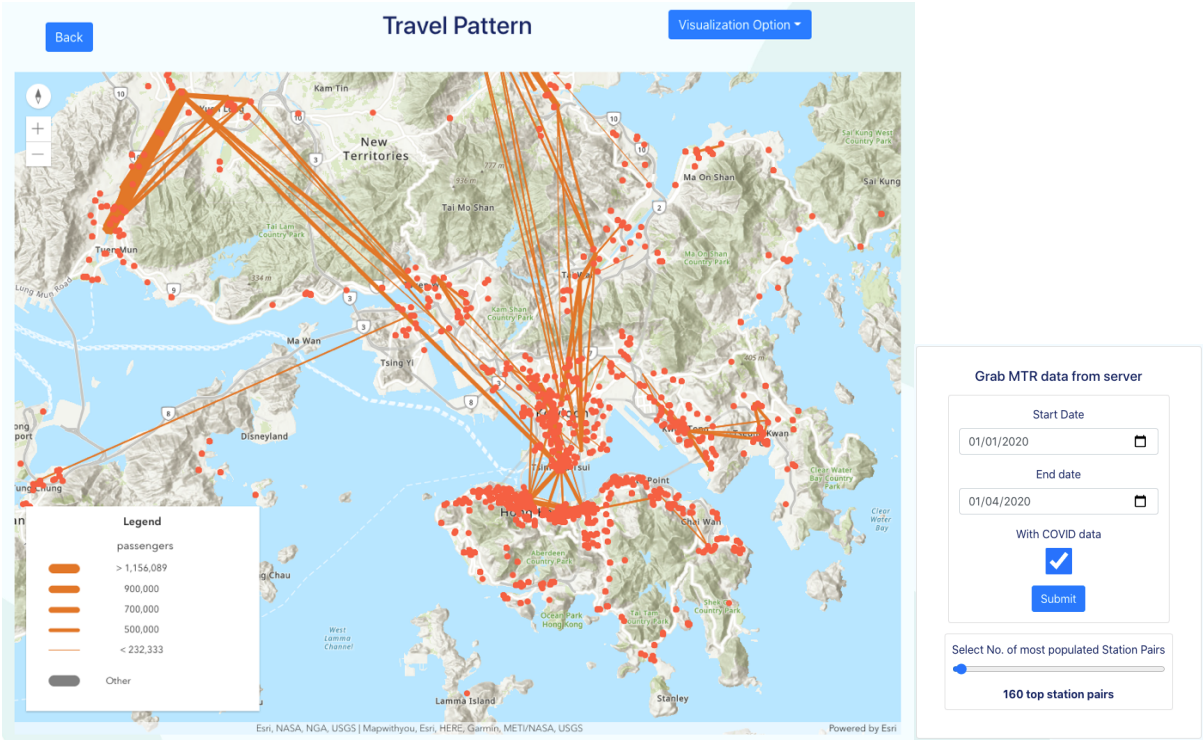

Figure 3.4.6 Travel Pattern and COVID-19 (01/01/2020 – 01/04/2020)

Figure 3.4.7 below represents the final visualization on our platform. The visualization plots our passenger mobility metric i.e. passenger volume against date. We also have the option to split the passenger volume into different card types i.e. Senior, Adult, Child and Student.

Figure 3.4.6 Passenger Mobility (Volume) by card type (01/01/2020 – 28/01/2020)

The final section of the platform provides the users the access to our *'someone like you'* and *'sensor individual'* algorithm as show in Figure 3.4.7 and Figure 3.4.8 respectively. For *'someone like you'* the user is asked to input the interval they want to split the day in and the month of data they want to perform the analysis on. *'Sensor Individuals'* requires the user to only mention the day they want to perform the analysis on. The average time taken by the *'someone like you'* analysis on the platform is about 5 minutes 13 seconds on one month of data, whereas, the average time taken by *'Sensor Individuals'* analysis on the platform is about 13 minutes for only one day of data. The difference in the runtime between the two displays the complexity of *'Sensor Individuals'* analysis.

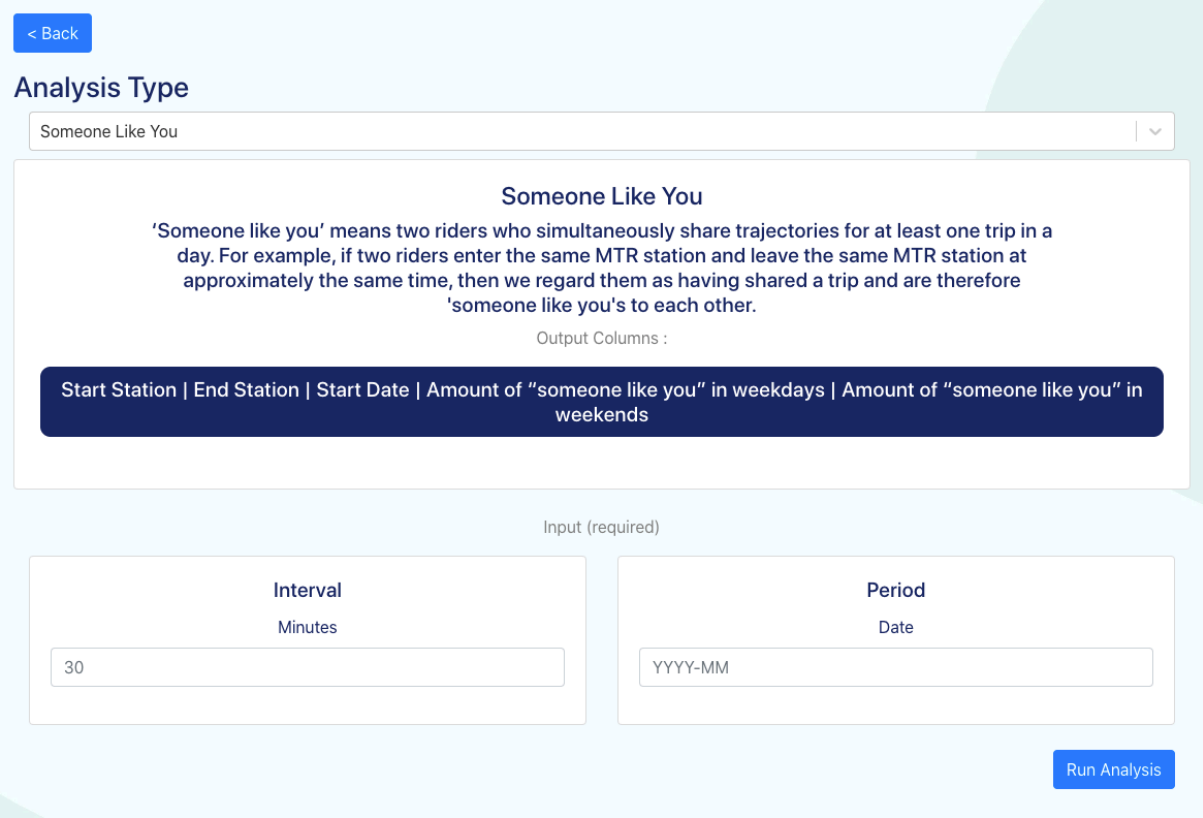

Figure 3.4.7 Platform - *'Someone like you'*

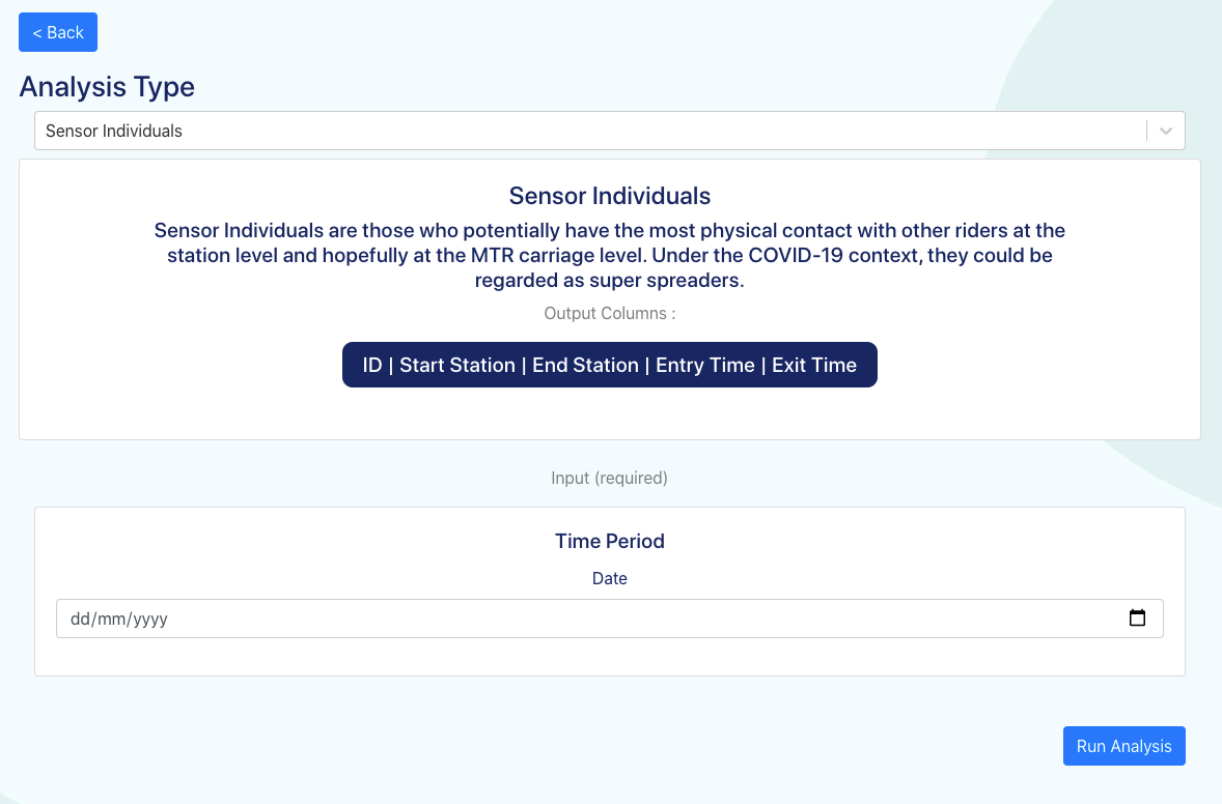

Figure 3.4.8 Platform – *'Sensor Individuals'*

4 Challenges and Limitations

4.1 COVID-19 and MTR relation

Despite of performing various analysis techniques we were not able to find any concrete results ensuring the role of the MTR in the spread of the virus. Most of the relations between the two datasets were due to high population densities of different areas, and even these relations were not consistent throughout Hong Kong. Additionally, Hong Kong government has been able to contain the spread of the virus, therefore, there is no consistent data in regards to the spread of the virus.

Another issue was to relate the two datasets, i.e. we do not have any method to map these COVID-19 cases to the MTR customer id. We believe if we had access to this information, we could further analyse the datasets to see if any of the cases with no previous links might have had a link with another COVID-19 patient who travelled in the MTR on a similar path and around the same time.

4.2 Server Performance

The performance of the server has led to many inefficiencies throughout the course of the project. Using a relational database certainly helped the project in creating relations within tables but at the cost creating inefficiencies later. In addition to the slow performance of the MySQL database, the hincare server has low processing power and computation capabilities. The reason behind placing the database on the server was to keep the data secure. Therefore, using a NoSQL database or other powerful databases could help increase the performance but at the cost of losing the relation between different tables.

As a result, most of the varies over the dataset take a long time to return a result. For example, simple queries like passenger volume took 2 minutes for the whole month of data and complex queries like sensor individuals took 13 minutes on a single day of data indicating the need of a better database system.

4.3 Database structure

The database contains multiple tables (for every month, January to August for both 2019 and 2020) with the same columns for the MTR data to provide quicker access to the data. This led to creating a lot of tedious code while programming the backend as Django works with a model-view-template, i.e. we need to create a model for each table. Creating one table for all the MTR data was not an option, as it would lead to greater inefficiencies.

4.4 ArcGIS

ArcGIS was one of the most sophisticated tool used in our project and helped us create in depth complex visualisations. The major limitation was understanding the software as it was not as intuitive and needed a lot more knowledge to explore its full potential. This knowledge was not as readily available, therefore, our team lacked the required experience in this field to explore the full potential of the software.

The API that has been used in the platform to create the visualizations is not as developed as the software itself, therefore, lacks a few features. For example, the API does support heat maps for COVID-19 spread similar to the visualizations in section 3.3.

5 Project Timeline

Table 5.1 below shows in detail timeline of the implementation of our project and also highlights the key events throughout the year.

Table 5.1.1 Project Plan

Date	Tasks	Status
September 2020	Pre-processing of data: o COVID-19 data o MTRC data Previous work / Literature review: o Familiar Strangers o Someone like you o Sensor Individuals Other tasks: o Develop a system to access MySQL on hincare server	Completed
October 2020	Database: o Onloading pre-processed data Phase 1 Deliverables: o Detailed Project Plan o Project Website (WordPress) Mobility Trend Analysis: o Visualisations o Insights Others: o Compiling results from analysis. o Innovation Wing poster	Completed

	○ Presentation at the Tam Wing Fan Innovation Wing opening.	
January 2021	Geo-Spatial Analysis: ○ Exploring ArcGIS ○ Visualisations FYP First presentation: ○ Compiling results (Visualizations and statistic) Phase 2 Deliverables: ○ Preliminary implementation ○ Detailed interim report Re-evaluation of project scope to build a platform with the following features ○ Query ○ Visualizations ○ Advanced Analysis	Completed
February 2021	InnoSpark competition: ○ Application ○ Presentation (Runners up – top 8 projects) Platform: ○ Setup ○ Develop methods to connect to server ○ Prototype	

March 2020	Platform: o Query Tool (frontend and backend) o Authentication system o Acquiring FYP domain Progress Presentation: o MTR officials	
April 2021	Platform: o Visualizations o Advanced Analysis (Someone like you and Sensor Individuals) Phase 3 deliverables: o Finalized tested implementation o Final report Final presentation	Completed
May 2021	Project exhibition o Presentation and video	In progress

6 Future works

6.1 Contact and Behaviour based research

The algorithm to both our contact and behaviour based research techniques (*'Someone like you'* and *'Sensor Individuals'*) have been implemented and tested. The results from both the algorithms can now be used to create visualisations. For *'someone like you'* analysis we could expect a similar result to the Beijing research paper mentioned in section 1.3. Whereas, for 'Sensor Individuals' we would expect to create spatio-temporal patterns. These visualizations deal with high level of complexities due to large volumes of data, therefore, need experience with software like ArcGIS.

Our *'Sensor Individual'* algorithm currently does not account for placing the passengers into different carriages, therefore, modifying the algorithm to increase the accuracy of the model also needs to be done. Implementing an equal distribution of the passengers throughout the MTR carriages would involve high level complexity as we need to account for every passenger on the MTR during a particular time period (might be on different routes). We expect this part of the algorithm to further increase the complexity and lead to a longer run time, therefore, needing more computational power.

6.2 Alternative Database

MySQL supported the scope of our project well and helped us achieve most of the tasks in our scope without any issues but we realised the increase in runtime with our queries becoming more complex.

Since, the performance of the database was not scalable when it came to complex queries such as *'Sensor Individuals'*, trying other NoSQL database technologies is essential to ensure the scalability of the project in the future to perform more complex analysis techniques. NoSQL databases such as Cassandra (by Facebook), OrientDB and MongoDB are a popular choice for Big Data projects (TOP 10 Open Source Big Data Databases, 2016). For instance, OrientDB can store up to 150,000 documents per second (TOP 10 Open Source Big Data Databases, 2016). Therefore, to further explore the potential of our data it is important to explore these technologies.

6.3 ArcGIS updates

It is expected that in the future releases of ArcGIS JavaScript API (version 4.9) that features like heat map would be supported while creating multiple layers in the visualization. Therefore, regularly updating the visualizations with the latest updates from ESRI would be essential to create more accurate and readable results. The final result on the web application would be expected to be similar to the visualizations in section 3.3.

6.4 COVID-19 live data stream

Building a data stream to regularly update the COVID-19 dataset would be helpful as this data stream would provide us with regular COVID-19 updates which can be used to analyse MTR data side-by-side with COVID-19 data. The reason behind not automating the pre-processing steps and creating a live stream was because we do not regularly receive the data from the MTRC and COVID-19 data on its own is not as useful for our project as there are a lot more public websites already analysing this data.

6.5 Digital Confidentially Agreement

To eliminate the tedious process of sharing the confidentiality agreement with other HKU staff members who request access to the system, a system could be developed to automate this process by verifying the user's HKU domain email address and then signing a digital confidentiality agreement to submit a request to access the system. The request would be forwarded to our supervisor, who can decide if the person should be provided with access to the system. Once approved, the user can access the platform.

7 Conclusion

Even though it has been about 1 year and 4 months since Hong Kong reported its first COVID-19 case, the city has still not figured out a way to eliminate the spread of the virus completely and with the new mutations of the virus being spread, COVID-19 pandemic is nowhere close to over. The need to understand and control the spread of this virus is becoming increasingly important, therefore, indicating the importance of projects like ours.

Over the last eight months, our team has made substantial progress to fully utilize the resources provided to us by analysing the two datasets side-by-side using various powerful programming languages and software. Despite not finding any concrete relations between the datasets due to the limitations stated in section 4, we still believe that these datasets have potential and if analysed further with more complex techniques could provide us with promising results. Therefore to ensure the further research and development of the MTR and COVID-19 dataset, we believe we have created a very useful and powerful platform to ensure the distribution of the data to other researchers in a safe and efficient method. This platform will not only help distribute the data to other researchers but also encourage and motivate them to find other hidden patterns in the data by having access to the high-level analysis performed by us. We hope to see contributions to our data repository from other researchers as our platform would provide a reliable and convenient method for these researchers to share and combine their results.

8 References

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

UCI Health. (2020, 04 29). Retrieved from www.ucihealth.org: https://www.ucihealth.org/blog/2020/04/why-is-covid19-so-dangerous#:~:text=It's%20not%20the%20most%20contagious,infecting%20another%20two%20or%20three

Worldometer. (2021, 04 08). Retrieved from Worldometer: https://www.worldometers.info/coronavirus/country/china-hong-kong-sar/

Centre for Health Protection. (2020, January 31). Retrieved from Centre for Health Protection: https://www.chp.gov.hk/en/features/102764.html

(2017). *Public Transport Strategy Study.* Hong Kong: Transport and Housing Bureau. Retrieved from Transport Department.

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

Boyd, D. &. (2011, September 21). Retrieved from https://ssrn.com/abstract=1926431

Connall, M. (2021). *Sigma*. Retrieved from https://www.sigmacomputing.com/blog/top-20-big-data-statistics/

Zhang, F. (2016). Who are My Familiar Strangers? Revealing Hidden Friend Relations and Common Interests from Smart Card Data.

Jiangping , Z., & Yang, Y. (2018). Someone like you: Visualising co-presences of metro riders in Beijing. *Environment and Planning A: Economy and Space*, 752-755.

What is a relational database? (2020, October 27). Retrieved from https://www.oracle.com/hk/database/what-is-a-relational-database/

What is MySQL? (2020, October 27). Retrieved from https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html

TOP 10 Open Source Big Data Databases. (2016, 06 13). Retrieved from Bitnine: https://bitnine.net/blog-useful-information/top-10-open-source-big-data-databases/