MXB362 Advanced Visualisation and Data Science

Visualisation Case Study

Visualising the Impact of COVID-19 in Singapore

Name	Dennis Chen
Student ID	n12150801
Date	14/08/2024

Table of Contents

Executive Summary

The motivation for this case study is to produce a visualisation tool that is able to provide up-to-date daily insights on how the pandemic has spread, and thus offer the public ongoing insights into the evolving pandemic landscape. While total case numbers and regional breakdowns provide a quantitative overview of a pandemic, they often fail to effectively communicate the nuances and implications of the current situation to the general public.

This case study aims to be a tool that can be utilised for future pandemic, only using COVID-19 pandemic and specifically in Singapore, as a tool to evaluate the current pandemic visualisations available and improve on it, developing one that can be easily understood by the general public and allow them to make informed decisions regarding the current pandemic situation.

According to the Singapore's government website, one should continue to self-monitor for up to 5 days if they are a close contact to any potential COVID-19 cases (Gov.SG | Updates to Health Protocols, n.d.). Following the guidelines of the Singapore's government, the visualisation produced is an animation that shows the locations of COVID-19 cases within a period of 5 days. The visualisation produced should inform and educate the general public of how the pandemic landscape has changed for the past 5 days and this will help them make informed decisions with how they move on with their day.

This case study takes great inspiration from Visualising SG COVID-19 Spread (Chua, 2021; Chua, n.d.)

Visualising SG COVID-19 Spread

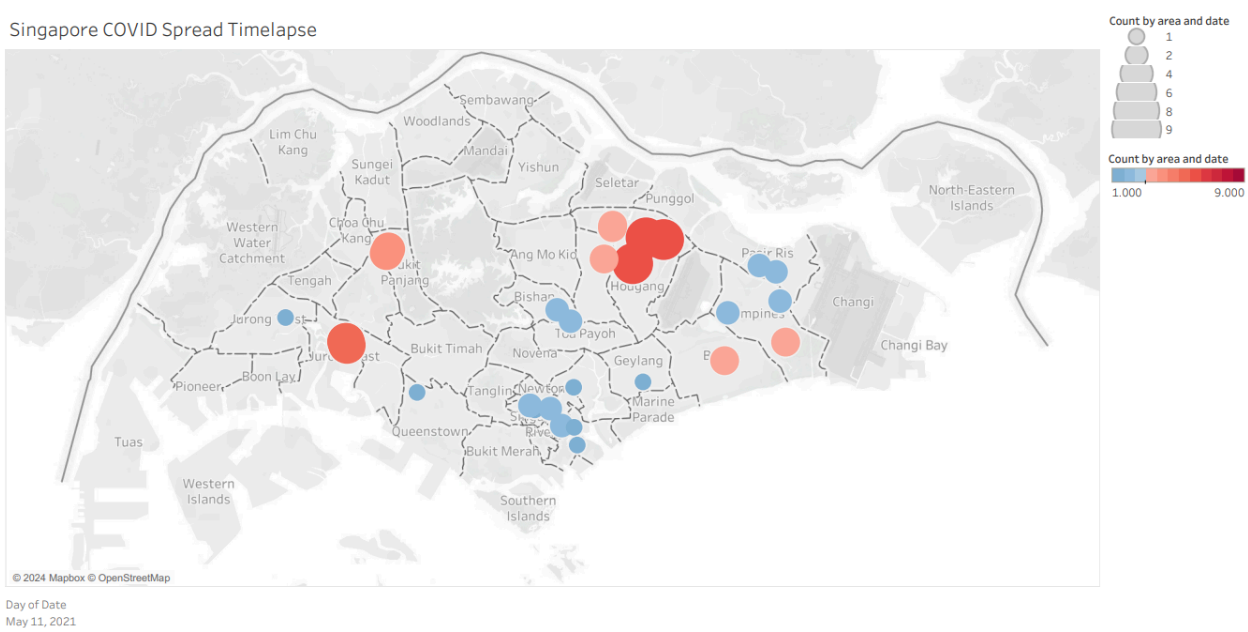

Fig 1. Singapore COVID Spread Book - Done by Hui Shun (Chua, 2021)

An existing visualisation is an animation that shows the daily distribution of COVID-19 cases in Singapore (Fig 1). The size of the circle represents the daily number of cases in that specific area whereas the colour gradient indicates the density of the number of cases till that date.

This visualisation can be rather misleading with the size of the circle. For each size of the circle, one is unable to tell at first glance how the increase in cases affects the increase in size of the circle. For the areas with more cases, the size of the circle is so big that it seems to cover up the entire district. Looking at this in the perspective of the general audience, they might think that the entire estate is a hotspot of the pandemic when in fact it could just be a particular neighbourhood in the district.

Other than that, having 2 different colour gradients to indicate the density of cases in the area might not be necessary. The excessive use of colour can create visual clutter which reduces the effectiveness of the visualisation, and might make it more challenging to interpret the visualisation for the general audience who might not understand how to read the legend. In this case, a single colour gradient is sufficient as the intention is to only represent a single set of data.

Dataset

The proposed raw dataset is a list that includes public places COVID-19 cases had visited for more than 30 minutes (Chua, 2020). The raw dataset has been cleaned and prepared and can be found in Google Sheets (*SG-COVID-Locations*, n.d.).

The main data that will be visualised will be the following:

- Time
- Date
- Latitude and Longitude Coordinates
- Area (Estate names)

The Date and Time data provides a temporal dimension to the analysis, whereas the geographic coordinates for each COVID-19 case is crucial for mapping and spatial analysis. Having the Area data provides a more general classification, which is useful for grouping and comparing different estates.

Visualisation Techniques

The goal of this case study visualisation is to provide up to date insights regarding the current pandemic landscape and this will be achieved through the following visualisation techniques.

Technique 1: Choropleth Maps

Through the usage of choropleth maps, it provides an easy way to visualise the density of COVID-19 cases across Singapore and show the level of variability across the different estates in Singapore. This utilises the Area data within our dataset.

Technique 2: Spatio-Temporal Animation - Scatter Plot

Doing a spatio-temporal animation of the different data points on the map will allow the general public to see the areas which the COVID-19 cases have visited. This helps provide a visualisation of how COVID-19 spread and how the public behaviour evolved during the pandemic. This utilities the geographic coordinates data of our dataset.

Visualisation Environment and Tools

1. Data Manipulation
   1.1. R
      1.1.1. dplyr – A grammar of data manipulation, providing a consistent set of verbs to solve data manipulation challenges.
2. Visualisation
   2.1. Python
      2.1.1. matplotlib – Useful for creating static and animated visualizations.
      2.1.2. pandas – Data manipulation and analysis tool
      2.1.3. geopandas – Adds support for geographic data, allowing spatial operations on geometric types.
      2.1.4. numpy – Provides support for numerical computations.
3. Development on
   3.1. Jupyter Notebooks – Offers a user-friendly environment for coding and visualisation
4. Video Editing
   4.1. Microsoft Clipchamp

Results and Outputs

1. Data Cleaning

The current dataset (*SG-COVID-Locations*, n.d.) have some minor formatting inconsistencies.

Specifically, the “Time” column had instances of missing spacing around dashes (e.g., "0000h-0000h"), which were corrected by ensuring spaces are placed before and after the dash (e.g., "0000h - 0000h"). Additionally, occurrences of en-dashes ("–") were replaced with proper hyphens ("-") for consistency.

In the “Date” column, some entries that had included unnecessary time components (e.g., "00/00/0000 0:00:00") were standardised by removing the time, ensuring all dates followed a uniform format without additional time information.

Source File can be found here.

2. Data Preparation using R

Since the project utilises geographical data, it will require a GIS map of Singapore (Urban Redevelopment Authority [URA], 2024). The preparation involved extracting relevant values from the GeoJSON "Description" field to create a new dataframe. This dataframe was merged with location data based on geographic boundaries identified by area names. The "Time" column was standardised and combined with the "Date" column to form a new Datetime column. Unnecessary columns were removed, and the cleaned dataframe was exported to *"covid_cases_cleaned.csv"*.

Source File can be found here.
New CSV File can be found here.

Further processing included re-importing the cleaned data, converting "Datetime" to POSIXct format, and extracting the date part while converting columns to character to avoid factor issues. The data was aggregated by date and location name, counting occurrences. To ensure complete coverage, a new dataframe with all date-location combinations was merged with the aggregated data. NA values were replaced with zeros, and cumulative case counts were calculated for each location. The final cumulative cases dataframe was exported to *"cumulative_cases.csv"* for further analysis.

Source File can be found here.
New CSV File can be found here.

3. Visualisation Technique 1: Animated Choropleth Map

I have experimented to make Choropleth Map with 3 different libraries, namely - Plotly, Folium and Matplotlib.

Plotly offers an interactive and visually appealing way to create Choropleth maps. It provides built-in animations, which makes it easy to display changes over time. Plotly's strength lies in its interactivity, allowing users to zoom, hover, and pan over the map to explore the data in more detail. However, I did not manage to get this to work and render animation after trying for some time and decided to try other libraries.

Folium is focused on simplicity and the ability to render maps directly in Jupyter notebooks. While it excels in embedding interactive maps into web pages, Folium does not inherently support animations.

Matplotlib, though primarily known for static plots, can also be used for creating Choropleth maps. For animation, Matplotlib's FuncAnimation can be used, but it requires more manual effort to set up compared to Plotly. Matplotlib offers greater customization but lacks the ease of interaction and built-in animation capabilities of the other two libraries.

Choosing the library to use for producing animation
The target audience for this case study primarily consists of the general public, who are not necessarily data scientists or GIS experts. Their primary need is to understand the spread and density of COVID-19 cases over time in an easy-to-follow visualisation.

I chose to produce the animation through Matplotlib instead of opting for more interactivity. Interactive elements like zooming, panning, and hovering over data points are useful for exploring complex datasets, but they can also add a layer of complexity that might confuse non-expert users. A static or pre-animated visualisation allows viewers to passively consume the information without needing to interact, ensuring the message is conveyed without additional cognitive load.

Source File can be found here. (Visualisation Technique 2: Choropleth Maps of COVID-19 Cases in Singapore)
Output GIF can be found here.

4. Visualisation Technique 2: Animated Scatter Plot

Using Matplotlib, I experimented with different ways to animate the locations of COVID-19 cases across Singapore. The animations aimed to visualise the spread of the virus over time using various temporal perspectives:

1. **Sequential Animation of Individual Cases (Entire Dataset):**
   Individual COVID-19 cases were plotted one by one across Singapore, covering the entire dataset. Each point represented a confirmed case, and the animation gradually revealed the spread of the virus as new cases were added to the map over time.

2. **Rolling 5-Day Window Animation of Individual Cases:**
   This animation focused on visualising individual COVID-19 cases within a rolling 5-day window. The map continuously updated to show new cases while older ones faded out, offering a dynamic view of how the virus spread over short periods.

3. **Rolling 5-Day Window Animation of Daily Cases:**
   Here, daily COVID-19 cases were aggregated and visualised within a rolling 5-day window. The animation showed clusters of cases appearing and dissipating, helping to highlight patterns of daily infections within specific areas.

Source File can be found here. (Visualisation Technique 1: Spatio-Temporal Animation of Covid-19 Cases)

Output GIF can be found:
1. Sequential Animation of Individual Cases (Entire Dataset). Link
2. Rolling 5-Day Window Animation of Individual Cases. Link
3. Rolling 5-Day Window Animation of Daily Cases. Link

5. Combination of both techniques

Both visualisation techniques serve complementary purposes in conveying the spread of COVID-19 across Singapore. The choropleth maps provide a clear, high-level overview of case density over time, while the animated scatter plots offer a more granular, location-based view of individual cases. Combining both approaches provides a well-rounded narrative of how the pandemic evolved spatially and temporally, allowing users to explore the data from multiple perspectives.

The choropleth map helps the audience grasp broader patterns in the data, such as the concentration of cases in specific districts or regions, while the animated scatter plots give insight into the day-to-day dynamics of virus transmission. Together, these visualisations cater to the general public's need to understand the spread of COVID-19 in a clear and intuitive manner.

Source File can be found here. (Final Visualisation)
Output GIF can be found here.

6. Legend Capturing

When producing the animation, I have set the plot to not show the legend. Instead, I have separately captured the image of the legend which will be put onto the final animation later on.

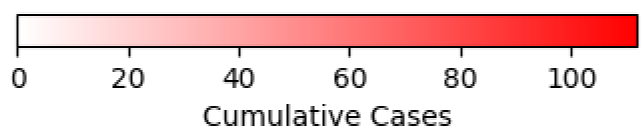

Fig 2. Visualisation Legend Image

Source File can be found here. (Legend)
Output PNG can be found here.

7. Capturing Final Frame of Animation

For a general audience, instant information is often more useful than a long animation. Many viewers prefer to quickly understand the overall picture, such as the total impact of COVID-19 at a glance, without waiting for an animation to unfold. To address this need, a static image capturing the final frame of the animation can be an effective solution.

The final frame of the animation have been captured and the legend have been appended onto the image.

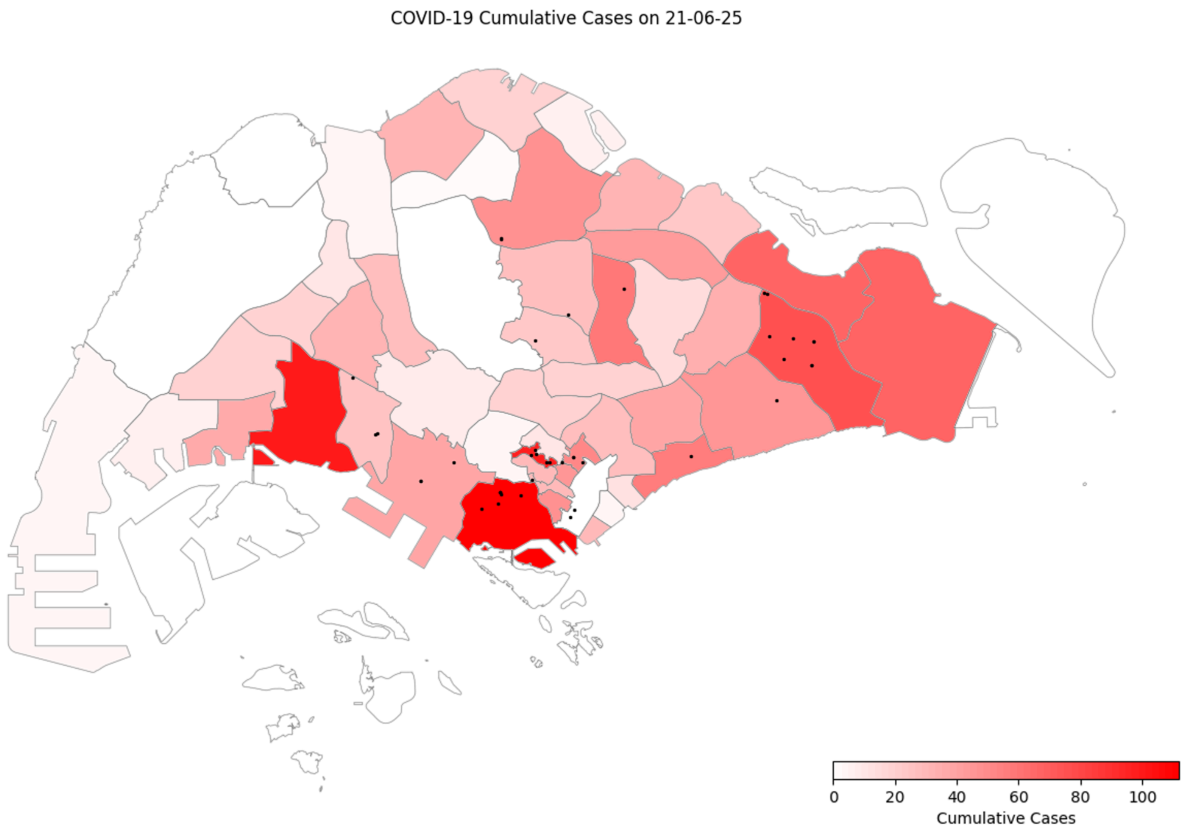

Fig 3. Final Frame Output Image

Source File can be found here. (Final Visualisation)
Output PNG without legend can be found here.
Output PNG with legend can be found here.

8. Merging Legend with Final Visualisation in Microsoft ClipChamp

The final step in producing a cohesive visualisation of COVID-19 data involves merging the animation (without the legend) with the separately captured legend image. Using Microsoft ClipChamp, a user-friendly video editing tool, you can easily combine both the animation and the legend for the final output.

Output MP4 can be found here.
View the visualisation on Youtube.

Exegesis of the Effectiveness of Visualisation Outputs

1. Lines

The decision to use thinner lines to separate the different areas of Singapore enhances clarity. Thin lines create subtle boundaries that distinguish regions without overwhelming the viewer, making the map more readable and visually pleasing.

2. Colours

The use of a monochromatic colour scheme ensures consistency and makes it easier to interpret the data. Monochromatic gradients, often ranging from light to dark shades of a single colour, help communicate the density or magnitude of cases in a clear, non-distracting way. This approach emphasises the spread of COVID-19 without introducing unnecessary colour distractions.

3. Clear Labels

The visualisation uses clear and concise labels, limited to the legend and a title indicating the cumulative COVID-19 cases and the corresponding date. The title provides context for the time frame being displayed, while the legend helps viewers interpret the data with ease. This minimalist approach ensures that essential information is communicated effectively without cluttering the visual, allowing the audience to focus on the spread of cases across Singapore.

4. Negative Space

The visualisation effectively makes use of negative space to ensure that the visualisation stands out. This technique prevents the graphic from becoming cluttered and allows the viewer's eyes to focus on the essential elements—such as case density and geographic boundaries. Negative space creates breathing room, improving readability and making the visualisation more aesthetically pleasing.

5. Minimised Complexity

The visualisation minimises complexity by focusing solely on essential elements: the geographic areas, the cumulative case counts, and the date. By removing unnecessary details and interactivity, the visualisation offers a clean and straightforward presentation that avoids overwhelming the audience. By presenting only key details, the visualisation enables the audience to grasp the core message quickly and efficiently, which is crucial for effective communication in public health contexts.

Project Timeline

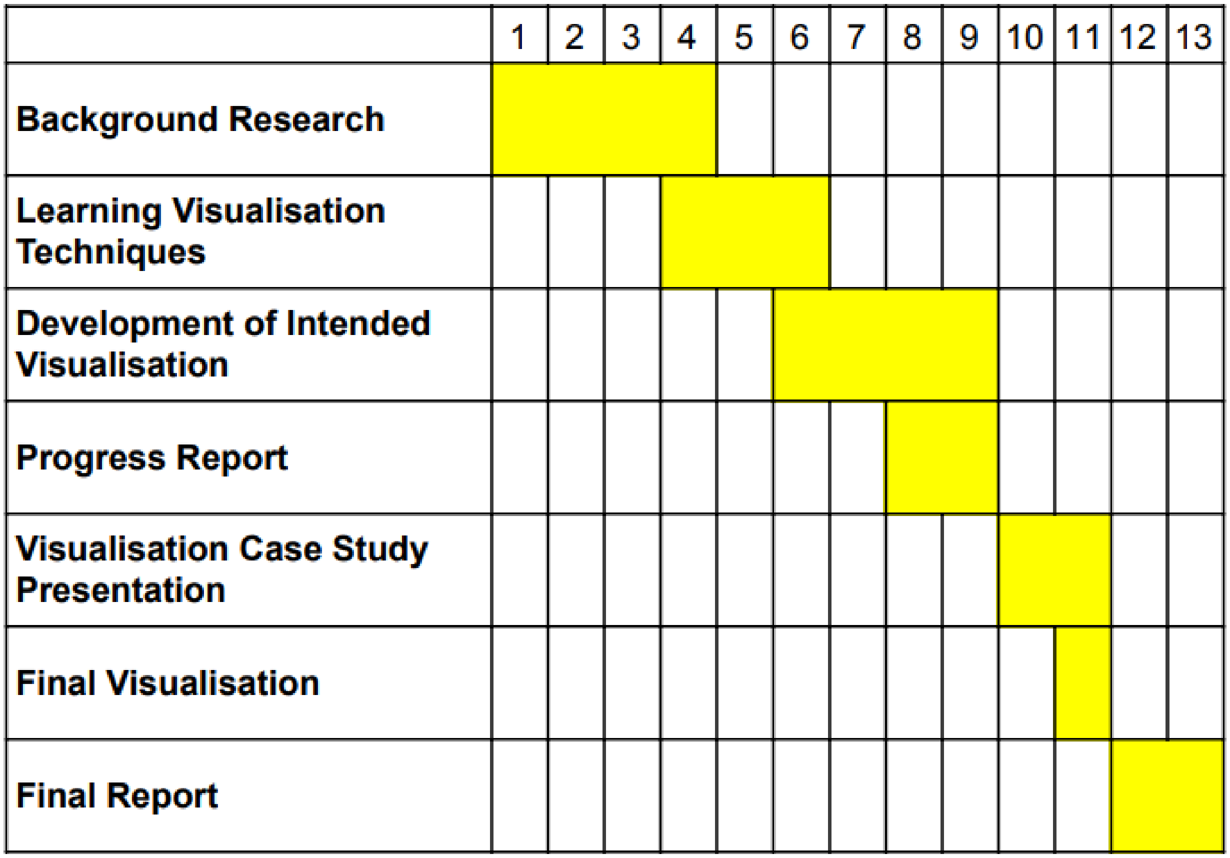

Fig 2. Gantt Chart of Case Study (MXB362, n.d.)

Problems and Risks

1. The initial frames of the animation reveal a discrepancy between the scatter plot and the choropleth map, which do not align as expected. Despite confirming that the datasets used for both visualisations are consistent, the misalignment persists throughout the animation. This issue may originate from a mismatch between the initial dataset and the GIS map of Singapore, indicating a potential problem with the integration of the dataset into the geographical map. Resolving this misalignment will require a thorough examination of the dataset's geographic references and how they correspond to the features represented on the map.
2. The decision to reduce complexity to the bare essentials—focusing only on Singapore's geographic areas, case counts, and date—makes the visualisation well-suited for a Singaporean audience familiar with the locations. However, this simplicity **limits its effectiveness for an international audience**, as non-locals are unlikely to recognise specific areas or districts. Without interactive features or additional geographic context, viewers outside Singapore may find it difficult to fully understand the spread of COVID-19 across the country.

Project Journal

Date	Progress	Time Allocated
29/07/2024	Ideation of potential case study topics	4 Hours
02/08/2024	Checking on current implementation of chosen case study topic	2 Hours
05/08/2024	Checking with tutor the feasibility of case study topic	10 Mins
13/08/2024	Finalising case study proposal	2 Hours
31/08/2024	Dataset Cleaning	1 Hour
01/09/2024	Data Preparation - Extraction of Relevant Details	1 Hour
01/09/2024	Data Preparation - Further Processing	30 Mins
06/09/2024	Leaning about making choropleth maps on Python	2 Hours
06/09/2024	Leaning about making scatter plots on Python	20 Mins
06/09/2024	Learning about spatial and temporal animation on Python	2 Hours
07/09/2024	Experimentation with Animated Choropleth Maps (Plotly)	5 Hours
08/09/2024	Experimentation with Animated Choropleth Maps (Folium)	3 Hours
08/09/2024	Experimentation with Animated Choropleth Maps (Matplotlib)	3 Hours
13/09/2024	Producing Animation (Choropleth Maps)	1 Hour
14/09/2024	Experimentation with Animated Scatter Plots (Matplotlib)	1.5 Hours
14/09/2024	Producing Animation (Scatter Plots)	30 Mins
15/09/2024	Combining both Visualisation Technique	2 Hours
15/09/2024	Producing Animation	30 Mins
15/09/2024	Capturing final frame	10 Mins
16/09/2024	Output legend image	10 Mins
18/09/2024	Finalising progress report	2 Hours
08/10/2024	Finalising case study report	3 Hours

Conclusions

The case study aims to develop a visualisation tool that provides daily insights into the COVID-19 pandemic's spread, focusing specifically on Singapore. It seeks to enhance public understanding by improving existing pandemic visualisations, which often fail to convey essential nuances. By leveraging data on COVID-19 cases in public places, the study employs choropleth maps and spatio-temporal animations to represent case density and location over time, respectively. The visualisations aim to educate the public, enabling informed decisions based on a clear understanding of the evolving pandemic landscape. Data cleaning and preparation were conducted using R and Python, while visualisation outputs were created with tools like Matplotlib and Microsoft Clipchamp. The final visualisations use a monochromatic colour scheme, clear labels, and effective use of negative space to enhance readability and ensure the audience can quickly grasp the pandemic's impact without cognitive overload.

For the future, this case study can aim to integrate real-time data into the visualisation tool which will significantly enhance its relevance and usability during pandemics. By incorporating live feeds from reliable sources such as health organisations and government databases, users will have immediate access to the latest statistics, trends, and developments regarding the spread of the disease. This dynamic approach not only ensures that the information presented is up-to-date but also enables users to make informed decisions based on current conditions. Ultimately, real-time data integration can facilitate better public awareness and engagement, empowering individuals and communities to respond more effectively to ongoing health crises.

References

Chua, H. X. (2021, June 30). *Data VIS Singapore Jun 2021 MeetUp: Visualising SG COVID-19 Spread* [Video]. YouTube. https://www.youtube.com/watch?v=OFRegA91bAM

Chua, H. S. (n.d.). *GitHub - huishun98/SG-COVID-data-automated: This project prepares data for a visualisation showing the spread of COVID-19 in Singapore across time*. GitHub. https://github.com/huishun98/SG-COVID-data-automated?tab=readme-ov-file

Chua, H. S. (2021, June 5). *Singapore COVID Spread Book*. Tableau Public. https://public.tableau.com/app/profile/hui.shun/viz/SingaporeCOVIDSpreadBookAutoUpdate/Dashboard

Chua, H. X. (2020, May 30). *Public Places Visited by Singapore Covid-19 Cases*. data.world. https://data.world/hxchua/public-places-visited-in-singapore-by-covid-19-cases

SG-COVID-Locations. (n.d.). Google Docs. https://docs.google.com/spreadsheets/d/19EPRvGyAMnYZn9LwfFa4UqYrt-MQ94-qEUg76vbl2Gs/edit?gid=0#gid=0

Gov.SG | Updates to health protocols. (n.d.). Default. https://www.gov.sg/article/updates-to-health-protocols

MXB362. (n.d.). Google Docs. https://docs.google.com/spreadsheets/d/1ZSmeVb8H-TFp9U5e9KdF5GGieY2Bom53Z6UJeNp_jmw/edit?usp=sharing

Urban Redevelopment Authority [URA]. (2024, June). *Master Plan 2019 Planning Area Boundary (No Sea)*. data.gov.sg. https://data.gov.sg/datasets/d_4765db0e87b9c86336792efe8a1f7a66/view

Appendices

Github Repository can be found here.

Visualising SG COVID-19 Spread (Referenced Visualisation) can be found here.

SG-COVID-Locations Dataset can be found here.

Master Plan 2019 Planning Area Boundary (No Sea) Dataset can be found here.