7CCSMSDV Simulation and Data Visualisation

Name: Noon Tew

K Number: K1631147

I have chosen the following

Track 1: COVID-19 Data Analysis

Part 1. Analytics

Part a

Q1 : Analyze the spread trend of this virus all over the world. What is the spread over time ?

Q2: Investigating the interventions/ regulations the countries have adopted into slowing the spread of the virus.

Q3: Countries that have introduced vaccine, does it have any effect over the spread of the virus?

Part b

Q1: Analyze the spread trend of this virus all over the world. What is the spread over time ?

To answer this question, we first need to understand the question then identify the elements/ data needed to answer the question.

i. Spread : in many literatures, susceptible-infected-removed (SIR) model[1][2] has been widely adopt in modelling the spread of infectious disease (Equation 1)

$$\frac{dS}{dt} = -\frac{\beta IS}{N}$$
$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

Equation 1: Differential equations for SIR model

Where S = susceptible, I = infected, R= recovered, β = transmission rate (constant), γ = recovery rate(constant). The data needed are the variables of the equation. However, it would be tricky to acquire the data:

1. Recovered: most database sources lack the information of it ,

2. Transmission and recovery rate: different sources have pointed towards different rate: some suggesting the transmission rates vary with the different variants [3]; and various studies pointing to different transmission rate [4], [5].

With the problem associated implementing the SIR model, this article [6] has suggested using the four metrics to track the spread of the COVID-19 :

1. Confirmed cases: it shows the first indicator of the spread of the virus and can warn about the need of extra healthcare support.
2. Deaths: the most straightforward indicator that indicates the severity of the spread of the virus
3. Positive Tests: this indicator is helpful in showing the real number of infections. For example, if the overall number of infections is rising, and the amount of testing stays the same, the confirmed cases will then reveal a much less number of infection that it actually is.
4. Hospitalization: This indicator gives the figure of how many infected individuals are ill enough to be admitted to a hospital for treatment. It helps controlling possible outbreaks of severe cases of the virus.

ii. Trend: Pattern over a period of time. In order to visualize this, we need data that includes susceptible , infected, recovered cases over certain period of time.

iii. Over the world: This includes all the countries in the world. We need aforementioned data in part i of all countries .

Dataset: ECDC dataset

1. This dataset gives a comprehensive data (19 December 2019 till 14 December 2010) on the geographic distribution of Covid-19 cases worldwide, however, it is not sufficient to answer the research question.
2. To only use this dataset to answer the research question, 3 variables are missing in the dataset to adopt Equation 1. However, we can simplify the SIR model to susceptible-infected (SI) model[7] to get rid of the three missing variables.
   - While including a recovered individuals (R) dataset on top of this dataset could be more accurate in modelling the data, but the decision for not doing so is because data from another source could potentially be different from the dataset at hand, thus causing multiple errors overall.
3. It provides the list of all countries and their recorded cases and deaths for last year.

Dataset: COVID-19 (coronavirus) by Our World in Data

1. This dataset, too, has a comprehensive up-to-date data on the geographic distribution of Covid-19 cases worldwide
2. It can be seen as an extension on the provided ECDC dataset: it encompasses majority of the same data that ECDC dataset has (cases and deaths of deach country

daily), but also have recorded other data such as COVID testing, hospitalisation, vaccination.

Link : https://github.com/owid/covid-19-data/tree/master/public/data

---

Q2. Investigating the effectiveness of the interventions/ regulations the countries have adopted into slowing the spread of the virus.

1. Effectiveness: It can be straight forward to measure this metrics, however, the data could be tricky to investigate:
   1.) We first measure the rate of spread of the disease for the time when no intervention/ restriction have been practiced.
   2.) Note the time the intervention is introduced and measure the number of cases daily for a period and get the spread of disease.
   3.) Compare the results to see if is effective. Big difference = high effectiveness, small difference= not so effective

   To measure this, we require the dataset of daily cases of COVID-19. It would not be easy to investigate as there are a large amount of regulations introduced at the same time, and hard to track the effectiveness of each individual restriction without having multiple factors (other restrictions) influencing the data.

2. Interventions: intervention includes restrictions/ public health policy aiming to reduce the COVID-19 risk. To visualize this data, we need the all the intervention and the date of each country adopting the intervention.

Dataset: Health Intervention Tracking for COVID-19 (HIT-COVID) Data

1. This dataset gives a full list of intervention adopted by all the countries and the date they adopted the measure.
2. It also provides information on the first death and first case discovered in their respectively country.
3. Using this dataset and conjunction with the ECDC dataset, we can visualise Q2 research question

Link: https://github.com/HopkinsIDD/hit-covid

Q3: Countries that have introduced vaccine, does it have any effect over the spread of the virus?

1. Vaccine: There are different types of vaccines administered for each country. In order to visualise and analyse for the Q3 research question, we need data that have the total number of vaccines administered for each country, the type of vaccines used

2. Spread of the virus: we can use a similar dataset like EDCD to measure the spread of the virus, viewing the before and after vaccination is introduce and investigate if there is any difference of the spread of the virus. However, the vaccine is only introduced in around Dec 2020, where the EDCD dataset we were given only have data from Jan 2020 till Dec 2020, a more recent dataset is required to investigate the difference the effect vaccination has on the spread of the virus.

Dataset: COVID-19 (coronavirus) vaccinations by Our World in Data

1. This dataset provides a comprehensive up-to-date information on confirmed cases, deaths, hospitalizations, testing, and vaccinations of COVID-19 worldwide.
2. This data will be able to provide sufficient information, such as the vaccination information, confirmed cases for us to be calculated and analysed.

Link: https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations

---

Part c

Dataset 1 : EDCD dataset

Dataset 2: COVID-19 (coronavirus) by Our World in Data / COVID-19 (coronavirus) vaccinations by Our World in Data

Dataset 3: Health Intervention Tracking for COVID-19 (HIT-COVID) Data

(Dataset 1, Dataset 2, Dataset 3 will be referred throughout the report as an abbreviation of the datasets mentioned above)

While investigating the data, I found that Dataset 1 provides data that is almost like a subset of Dataset 2. Both datasets provide information on daily cases and deaths of COVID of each countries. A few of the notable differences between both datasets are:

1. As mentioned before, Dataset 2 provides more data on other fields: such as data of country vaccination, hospitalisation, testing etc. Moreover, the data it provides extend from Jan 2020 till today.
2. The number of cases and deaths differ: In January 2020, Dataset 1 seemed to report higher number (both cases and deaths) than Dataset 2. Dataset 2 also reports negative number is cases (which is a result of misclassifying negative cases), where Dataset 1 does not.

Under comparison of Dataset 1 and Dataset 2 to Dataset 3, Dataset 3 provides data that does not reveal quantitative information regarding the spread of the virus, but qualitative information. For example, by combing the two datasets (Dataset 2 and Dataset 3), we might find :

1. Correlation between a certain intervention and number of new cases/ deaths: The intervention is usually be implemented when new cases or deaths reach a threshold.
2. The effectiveness of each methods: which interventions introduced decrease the increase of daily cases the most.

Part 2. Design and Discussion

Part a & b

Combined part a and b for the ease of reading the text and viewing the design.

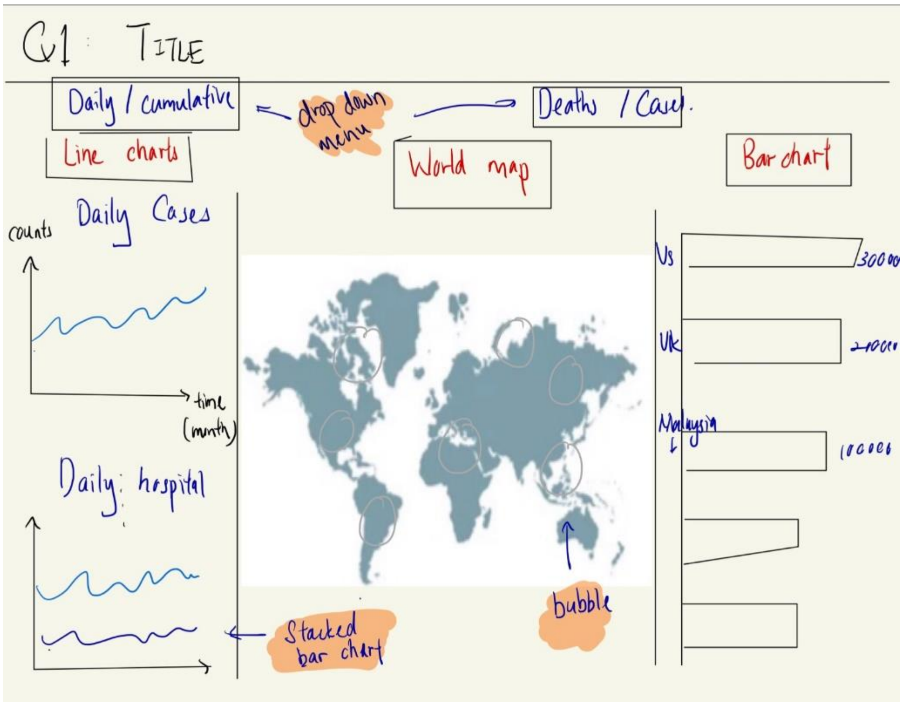

Figure 1: Design 1 . World map is from[8]

Design 1: <u>Q1: Analyze the spread trend of this virus all over the world. What is the spread over time ?</u>

For this design, I want to communicate with my audience using easy understanding and clear visualisation technique. With that in mind, I choose to use multiple simple visualisation tools to present the finding, instead of one complicated visualisation tool.

<u>Data Abstraction:</u>

First, the datasets proposed to show the visualisation of Q1 are Dataset 1 and Dataset 2. The main data to be extracted are :

1. Date : Temporal data
2. Country/ location : Spatial/ Categorical data
3. Cases/ Deaths/ Hospitalisations/ Tests : Quantitative data

Data processing:

First, Dataset 1 and Dataset 2 needs to be combined together so the data are homogenised. The data itself provides daily cases and deaths of each country. But in order to have the visualisation in Design 1, data are needed to be expanded to

1. Daily data worldwide – data such as cases, deaths , tests and hospitalisation are to be summed separately.
2. Cumulative data worldwide - the quantitative data mentioned above are to be summed and brought forward to get the cumulative numbers of each data.
3. Country's daily data – similar to daily data worldwide, but group by country
4. Country's Cumulative data – similar to cumulative data worldwide, but group by country instead.

Visual encoding/ Layouts :

1. Line charts on the left:

I chose to use small multiple line graphs to represent the daily and cumulative graph for data :cases, deaths, hospitalisations, and test. It is perhaps the most straight-forward/ common visualisation, but is effective to easily communicate the numbers of each quantitative data in linear time temporal domain. Y-axis is the value of the data, while x-axis is time in months. Using line, positions on a common scale, can be an effective way to convey the information. Moreover, in the hospitalisation, to demonstrate the two separate data but in the same chart, I use stacked bar chart to with different colour to categories the data.

2. Map in the middle :

To show the daily cases and deaths of the world, I chose to use a Conformal Projection of the world map to cover the spatial data , and using the bubble/circles to overlay onto it to show specific information (data on cases/deaths) of the country. The decision to use Conformal Projection is because the longitude and latitude can be utilised to integrate my data onto the d3 projection. Moreover, as it does not differ too much from the Winkel Tripel Projection, it should not confuse the audience with the shape of the map, but instead allows the audience to focus on the data (bubbles) being shown. The bubbles are half transparent and scale with the value of the data.

3. Bar Chart on the right

The bar chart on the right will show the cumulative quantitative data (cases/deaths/tests/hospitalisations) but grouped by countries. Bar chart is good because it groups the data can encodes 2 attributes : line with horizontal position channel used to map the quantitative attribute; the vertical spatial position channel used to map the categorical attribute. Each bar is accompanied by its values shown on the right, so it can help to eliminates the wrong perception arises from scaling issues.

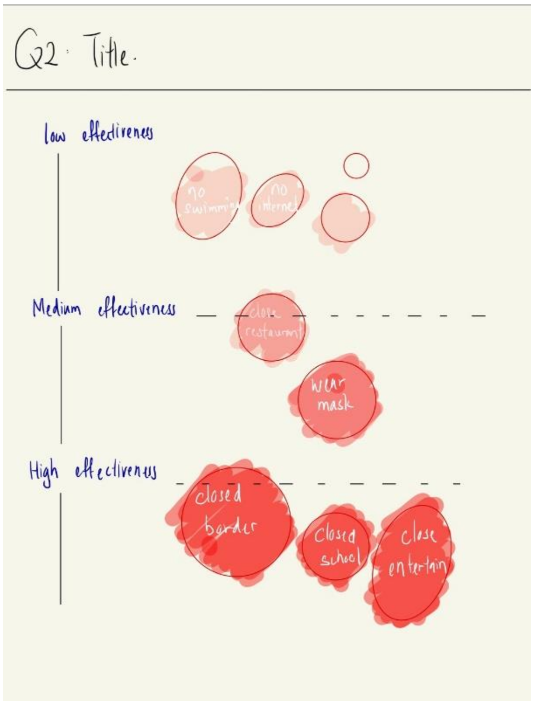

Figure 2: Design 2.

Design 2: Q2: Investigating the effectiveness of the intervention/ regulations the countries have adopted into slowing the spread of the virus.

For this design, I want to compress the information from Dataset 2 and Dataset 3 into an easier understanding visualisation, while keeping as much information as possible.

Data Abstraction:

First, the datasets proposed to show the visualisation of Q2 are Dataset 2 and Dataset 3. The main data to be extracted are :

1. Date : Temporal data
2. Country/ Location : Spatial/ Categorical data
3. Cases/ Deaths/ Hospitalisations/ Tests: Quantitative data
4. Intervention group: Categorical data

Data processing:

Before starting to discuss the layout of design 2, I would need to process both datasets to extract the effectiveness of each intervention (process briefly described in Question 1 part b). Then we need to sort each intervention in ascending order in terms of their effectiveness in slowing the spread of the virus. And finally , we need to process the data to identify the number of countries that have adopted a specific intervention.

Visual encoding/ Layouts :

The proposed design is a graph with y-axis showing the effectiveness of the intervention , while x-axis does not have any value associated with it . Each intervention is represented using a circle and words labelling the inside. Each of the circle are coloured, using red colour, from less saturated at the top (low effectiveness category) to more saturated at the bottom (high effectiveness category).

The visual encoding used for the intervention are circles. Using a circle is a good choice to show :

1. Data under the same category (intervention)
2. Scaling the size of the circle to the number of countries that have adopted the intervention. Although the use of area is usually not encouraged to communicate the quantitative information, but in this case, I want to communicate a relative information rather than a detailed numerical comparison.
    a. In aiding for the quantitative visualisation , I utilise a display box , where when it hovers on, will show the number of countries has adopted this specific intervention.

Y-axis shows effectiveness of the intervention: the visual channel position is utilised.

Colour : red. As there is only one category of data (intervention), only one colour is used. However red colour uses saturation to show the different effectiveness of each intervention.

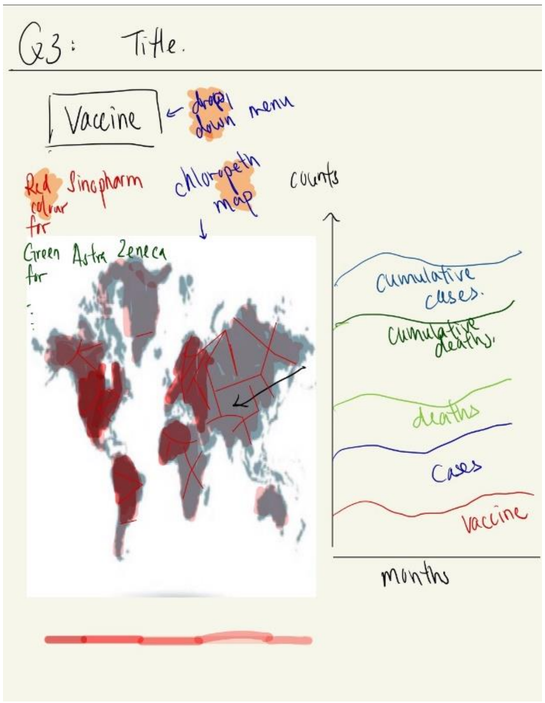

Figure 3. Design 3. World map is from[8]

Design 3: Q3: Countries that have introduced vaccine, does it have any effect over the spread of the virus?

Data Abstraction:

First, the datasets proposed to show the visualisation of Q3 are Dataset 1 and Dataset 2. The main data to be extracted are :

1. Date : Temporal data
2. Country/ Location : Spatial/ Categorical data
3. Cases/ Deaths: Quantitative data
4. Types of vaccines: Categorical data
5. Total Vaccinations : Quantitative data

Data processing:

The data processing will be similar as Design 1, but instead of getting the hospitalisation and test data, we want to obtain the vaccination data. The following data will be needed:

1. Daily data worldwide – data for total vaccinations, cases and deaths are to be summed separately.
2. Cumulative data worldwide - the quantitative data mentioned above are to be summed and brought forward to get the cumulative numbers of each data.
3. Country's daily data – similar to daily data worldwide, but group by country
4. Country's Cumulative data – similar to cumulative data worldwide, but group by country instead.

Visual encoding/ Layouts :

The proposed design for answering Q3 is centered around a choropleth world map, where when a country is clicked, it will display the information of the total vaccination, cumulative deaths, cumulative cases, daily death and daily cases of the country in a stacked area chart. On top of the map, there will also be an option to select the type of vaccination.

1. Choropleth world map

As Design 3 purpose is to answer Q3, where it requires to display the countries (spatial data) and the vaccinations (quantitative data), choropleth map is a good medium to present both data. Choropleth map can manipulate the shades of the colour to scale with the value of total vaccinations of each country. Upon selecting a different type of vaccination, the colour of the choropleth map should also change (red to green etc) to show a different category of quantitative data is used.

2. Stacked area chart

While clicking on a country, a stacked area chart will be prompt, showing the data that includes total vaccination, cumulative deaths, cumulative cases, daily death and daily cases of the country. Stacked area chart gives good visualisation that fit in this case because :

a. Helps to visualise a linear time temporal domain. The data shown will be from the start of the pandemic (Jan 2020) till today.
b. Helps in comparing the data visually. Will be easy to visualise if the introduction of vaccines has any effect over the new cases/ deaths daily.
c. Easy to spot any trend , if there is any.

3. Colour

For this visualisation, it will be utilising a lot of different colour as the visual channel to separate different categories of data : colour for the choropleth map that corresponds to different type of vaccines , colour of stacked area chart.

Part 3. Implementation

To run the implementation, run the “ index.html”

Data processing:

The main datasets used are Dataset 1 and Dataset 2, under “asset/data/Big datasets/”, and the derived datasets from all files under the “asset/data/”. An external dataset is countriesgeo.csv, which contains the longitude and latitude information of each country, is also file under “asset/data/Big datasets/”.

Link: https://developers.google.com/public-data/docs/canonical/countries_csv

All the derived datasets are processed using Python. Below are the files names and description of their processing :

1. **combine.csv**: a dataset that result from integrating Dataset 1 and Dataset 2, contain the main quantitative information: cases, deaths, tests, hospitalization. I take the first month of the Dataset 1 (31 Dec 2019 – 31 Jan 2020) including the deaths and cases of everyday of every country, and append it with Dataset 2 , from 1 Feb 2020 to 5 April 2021. Inner join both data on “date”.

2. **cases.csv** : daily and cumulative data of new COVID-19 cases reported, everyday from 31st December 2019 till 5th April 2021, worldwide. The data results from using combine.csv, grouping by the date, and summing over the case data.

3. **country_cases_geo.csv** : cumulative data of new COVID-19 cases, hospitalisation, tests of each country reported. The data also have the longitude and latitude data for each country. The data results from using combine.csv integrating with countriesgeo.csv (inner join on “location”), grouping by the countries, and summing over the all the data for each category.

4. **country_deaths_geo.csv** : cumulative data of new COVID-19 deaths, hospitalisation, tests of each country reported. The data also have the longitude and latitude data for each country. The data results from using combine.csv integrating with countriesgeo.csv csv (inner join on “location”), grouping by the countries, and summing over the all the data for each category.

5. **deaths.csv** : daily and cumulative data of new COVID-19 deaths reported, everyday from 31st December 2019 till 5th April 2021, worldwide. The data results from using combine.csv, grouping by the date, and summing over the case data.

6. **hospital.csv** : daily and cumulative data of new COVID-19 hospitalisation and COVID-19 ICU hospitalisation reported, everyday from 31st December 2019 till 5th April 2021, worldwide. The data results from using combine.csv, grouping by the date, and summing over the case data.

7. **test.csv** : daily and cumulative data of new COVID-19 positive test reported, everyday from 31st December 2019 till 5th April 2021, worldwide. The data results from using combine.csv, grouping by the date, and summing over the case data.

8. **owid-covid-geo.csv** : a dataset that result from integrating Dataset 2 and countriesgeo.csv. It provides all the quantitative information: cases, deaths, tests, hospitalization, and also the geographical location (longitude and latitude) of each location. Performed Inner join on “location”.

Each dataset is utilise for :

Line/area chart: cases.csv , deaths.csv, hospital.csv, test.csv

Bubble map : owid-covid-geo.csv

Bar chart: country_cases_geo.csv, country_deaths_geo.csv

User interaction:

1. Scroll bar

It is included in the first column and the third column of the visualisation (counting from left to right). It is the simplest level of user interaction included, that aids the visualisation as a whole by reducing the compromisation of the figures size. Without using the scroll bar, the graphs are then needed to reduce in size, resulting harder/ erroneous reading of the details.

Reference link(s):

i. https://github.com/Grsmto/simplebar
ii. https://www.youtube.com/watch?v=74eaw_nM5tY&t=183s

2. Dropdown menu selection

It is included in the first column and between the second and third column:

1. In the first column: it allows selection between daily or cumulative data on cases, deaths , tests, hospitalisation on the on the line chart in the first column
2. Between the second and third column: it allows selection between cases or deaths data to be displayed for each country.

Including the menu selection data helps to simplify the visualisation, without having overlay multiple data on top of each other. Therefore, it can easily highly the data to be investigated.

Reference link(s):

i. https://www.d3-graph-gallery.com/graph/line_select.html

3. Tooltips

There are 3 types of tooltips utilised in this visualisation: line charts, bubblemap, and Bar charts.

Line chart: There are a total of 8 line charts utilised in this visualisation, showing the daily and cumulative data of cases, deaths, hospitalisations and tests. All the line charts are created by appending a line, x-axis and y-axis onto an svg. An area with the same colour as the line are then appended under the line to create the chart. Then, I appended tooltip onto the svg, listening to events touchmove and mousemove, to display information of the data on the selected part: the date of the selected part and the data associated with it.

Bubblemap: The Bubblemap is created by creating a map and appending bubbles/ circles over it. The size of the bubbles scale with the data (cases/ deaths): the higher the number count of cases/ deaths, the bigger the bubble. Each bubble are located on top of the map using, the longitude and latitude information of each country to aid with the positioning. The tooltips I appended onto the bubbles, where when the mouse hover over, the colour of the bubble will change, indicating the selected bubble. After the mouse leave, the colour will then change to a different colour, indicating that bubble has been selected before. As there are many bubbles on top of the map, the tooltip not only aids in visualising the numerical data (cases/deaths), it also helps to reduce the complexity while reading the data by “ticking off” which bubbles has been browsed (“pop” effect).

Bar chart: There are a total of 2 horizontal bar charts utilised in this visualisation, showing the cumulative cases and deaths of COVID-19 up to 5th April of each country. The bar charts are created by appending a rectangle that scales with the cases/ deaths count , x-axis and y-axis onto an svg. The bar charts are arranged in descending order of the cases/deaths counts so the bar chart can also show the ranking in terms of the “riskiness” of the countries. Then, I appended tooltip onto the svg, listening to events mousemove and mouseout, to display information of the data on the selected part. A selected country will show its cumulative deaths/ cases, hospitalisation, and test cases up from 31 Dec 2019 till 5 April 2021. If there is no information regarding a specific field, for example , if no test data for Philippines, it will then display “ No data” in the following field. The tooltip helps to give an up-to-date summary of all the data being investigated in the visualisation.

Reference link(s):

i. https://www.d3-graph-gallery.com/graph/bubblemap_template.html
ii. https://www.d3-graph-gallery.com/graph/bubble_tooltip.html
iii. https://bl.ocks.org/larsenmtl/e3b8b7c2ca4787f77d78f58d41c3da91
iv. http://www.d3noob.org/2013/01/adding-title-to-your-d3js-graph.html#:~:text=What%20we%20want%20to%20do,append(%22text%22)%20
v. https://observablehq.com/@bsaienko/animated-bar-chart-with-tooltip
vi. https://www.d3-graph-gallery.com/backgroundmap
vii. https://www.d3-graph-gallery.com/graph/barplot_horizontal.html

4. Slider

There are two sliders implemented in the visualisation, each controlling a different map (cases / deaths). However, the building of both sliders is similar , but just appending onto different sets of data. The slider implemented allows the user to control the date between 31 Dec 2019 to 5 April 2021. While sliding through the date, for every date, the bubble will change its size according to the number of cases/ deaths. The slider acts as an important element that helps to investigate the trend of spread of virus over time.

Reference link(s):

i. https://observablehq.com/@bradvoracek/d3-simple-slider
ii. http://dev.centrogeo.org.mx/viz_desaparecidos/lib/d3.slider/

References

[1] I. Cooper, A. Mondal, and C. G. Antonopoulos, ‘A SIR model assumption for the spread of COVID-19 in different communities’, *Chaos, Solitons & Fractals*, vol. 139, p. 110057, Oct. 2020, doi: 10.1016/j.chaos.2020.110057.

[2] L. Laguzet and G. Turinici, ‘Individual Vaccination as Nash Equilibrium in a SIR Model with Application to the 2009–2010 Influenza A (H1N1) Epidemic in France’, *Bull Math Biol*, vol. 77, no. 10, pp. 1955–1984, Oct. 2015, doi: 10.1007/s11538-015-0111-7.

[3] ‘What do we know about the new COVID-19 variants? - Public health matters’. https://publichealthmatters.blog.gov.uk/2021/02/05/what-do-we-know-about-the-new-covid-19-variants/ (accessed Apr. 22, 2021).

[4] P. Sun, X. Lu, C. Xu, W. Sun, and B. Pan, ‘Understanding of COVID-19 based on current evidence’, *Journal of Medical Virology*, vol. 92, no. 6, pp. 548–551, 2020, doi: https://doi.org/10.1002/jmv.25722.

[5] D. Fanelli and F. Piazza, ‘Analysis and forecast of COVID-19 spreading in China, Italy and France’, *Chaos, Solitons & Fractals*, vol. 134, p. 109761, May 2020, doi: 10.1016/j.chaos.2020.109761.

[6] R. Yeip, ‘Four Ways to Track the Spread of Coronavirus—and Why None of Them Is Perfect’, *Wall Street Journal*, Jul. 11, 2020.

[7] K. M. Sutton, ‘Discretizing the SI Epidemic Model’, p. 20.

[8] ‘Worldmap Images, Stock Photos & Vectors | Shutterstock’. https://www.shutterstock.com/search/worldmap (accessed Apr. 23, 2021).