Project based exam 2021: Covid-19 Data Analysis

7CCSMSDV Simulation and Data Visualisation
Humphrey Curtis, *k20103881@kcl.ac.uk*
Kings College London
Department of Informatics
For Professor Rita Borgo

15th April 2021

Contents

1 Part 1: Analytics

1.1 Section A

Propose two or more exploratory research questions beyond Q1, label them as Q2 and Q3.

1.1.1 Proposal of research questions

For part 1 and section A, I have proposed four exploratory research questions – two of which will require access to datasets beyond the those initially provided by Kings College London (KCL) faculty. Key details of the reasons for the research question choices are noted in 1.1.2. Initially I have chosen track 1 and the research question is as follows:

Question 1 (Q1): *"Analyse the spread trend of this virus all over the world. What is the spread over time?"*

My secondary research question supplements the initial research question but focuses on a deeper analysis of the coronavirus spread over time. In particular, with a focus on the spread differences between sovereign nation states with differing human development or human development index (HDI)[1] and whether that has notably impacted the underlying coronavirus spread data:

Question 2 (Q2): *"What is the spread of the virus in the developing versus developed world. What initial insights can be reached?"*

In contrast, my third research question changes course from the initial two research questions and requires supplementary datasets. Here, I focus not on the virus but on the *resolution* to the virus notably the emergence of vaccinations[2] to combat the virus, Foley (2020). However, my third research question does deliberately align with the initial research question by focusing on the global developments (i.e., the differing spread across all nations worldwide):

Question 3 (Q3): *"Analyse the global rollout of vaccines over the world. What nations have rolled out vaccinations most effectively?"*

My fourth question supplements the third research question by also studying the roll-out of vaccinations. However, deliberately the question is angled to emulate the second research question by focusing on vaccination rollout in the developing versus developed world. Here, I am once again trying to focus on whether the differences between sovereign nation states HDI have notably impacted the vaccine rollout and intranational accessibility of vaccines, Kay (2021):

Question 4 (Q4): *"What are the key differences in the vaccine roll out in the developing versus developed world. Has the vaccine rollout been unequal?"*

[1] Human Development Index is a statistic composite index of life expectancy, education and per capita indicators, which are used to rank countries into four tiers of human development, Sagar and Najam (1998).

[2] A Covid-19 vaccination is intended to provide acquired immunity against severe acute respiratory syndrome coronavirus 2, the virus causing coronavirus disease 2019 (Covid-19), Foley (2020).

1.1.2 Reasons for choosing research questions

There are a range of reasons for choosing these four research questions. For research question 1, I favoured track one provided by Kings College London. Whilst, for research question 2, I instead wanted to use the dataset provided to effectively visualise if human development indicators (HDI) have had a substantial impact on the spread of the coronavirus and the responsive global vaccination rollout. Indeed, viewing the spread of coronavirus and consequent vaccine drive through the spectre of human development is very interesting and has proven to be a thriving locus of emerging research for many data scientists, Shahbazi and Khazaei (2020); Ross et al. (2020); Roghani and Panahi (2021).

In contrast, research questions 3 and 4 focus on emerging datasets for the solution to the global pandemic i.e., vaccinations. Indeed, research questions 3 and 4 are an expansion on the initial first two research questions yet equally aligned by focusing on global pandemic developments. In particular, research question 3 will look to understand the spread of vaccinations globally and visualise which nations have most effectively sourced and vaccinated their populations, Shahbazi and Khazaei (2020). In contrast, question 4 seeks to clarify if there are any inequalities in the rollout of vaccines between the developing and developed world (i.e., the nation's respective HDI). For instance, if nations with higher human development and wealth are able to afford more vaccines and vaccinate their populations faster, Roghani and Panahi (2021).

From this pool of research and four research questions, it is then feasible to gain an understanding of global developments of the virus spread and then consequent global response via the rollout of vaccinations. Coupled with visualisations on the differing pandemic outlook between the developed and developing world – thereby effectively framing the pandemic using human development.

1.2 Section B

Explain what type of data, beyond the one provided, could be used to answer Q1 and each one of the questions you proposed in part a. Assess the fitness of each dataset between the ones we provided and other resources you have found, you would potentially be using to answer the questions.

1.2.1 Further datasets to answer research questions

In this section, I will firstly, provide the datasets that can be used to effectively answer the research questions proposed in 1.1.1. Following this, I will provide a critical assessment of the datasets including their fitness and the types of data that can be used to answer the underlying research questions.

Beyond the ECDC global historical dataset up to the 14th of December 2020 sourced and provided by KCL. There are a number of other publicly available datasets noted in the table over the next 2 pages to provide answers for the 4 research questions proposed:

Table 1: Further datasets and links

Research questions	**Links**
1	• https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide • https://ourworldindata.org/coronavirus-data • https://github.com/CSSEGISandData/COVID-19 • https://disease.sh/ • https://corona.lmao.ninja/docs/
2	• https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide • https://ourworldindata.org/coronavirus-data • https://github.com/CSSEGISandData/COVID-19 • https://disease.sh/ • https://corona.lmao.ninja/docs/ • https://ourworldindata.org/human-development-index • https://data.un.org/

Research questions	Links
3	• https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations • https://ourworldindata.org/coronavirus-data • https://corona.lmao.ninja/docs/
4	• https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations • https://corona.lmao.ninja/docs/ • https://ourworldindata.org/human-development-index • https://data.un.org/

1.2.2 Assessment of datasets

Other types of data to answer the research questions

Having noted a wide range of datasets in the above table, now I will provide an assessment of the datasets and the types of data that can be used to answer the four research questions. Following a discussion of types of data, I can then turn to assessing the datasets and whether they have collected records for data variables needed to answer the research questions.

Beyond the ECDC data for research questions 1 and 2 provided by KCL it would be useful to have access to more recent datasets because the ECDC dataset terminates on the 14th of December 2020. A more up to date dataset could provide discrete quantitative case counts and collected records up to the present day (e.g., further temporally significant data) – enabling the spread all over the world to be visualised throughout the entirety of the global pandemic.

Equally, for research questions 1 and 2 it would be useful to have access to a dataset which effectively collects records for variables relating to testing with the nation states e.g., discrete quantitative (e.g., number of tests) or statistical quantitative data (e.g., number of tests as a ratio of population size). Indeed, the ECDC dataset does not collect records for testing variables – therefore it is difficult to verify the accuracy of some of its underlying discrete quantitative data (i.e., case count) for certain nations.

For example, certain nation states may lack adequate health infrastructure or the wealth to purchase numerous Coronavirus tests for their populations – therefore the number of reported Coronavirus cases and deaths collected by the ECDC dataset may be unreliable and too low for these nations with less testing. Therefore, a dataset which collects records relating to testing will ultimately result in a more accurate visualisation and assessment of data for research questions 1 and 2.

On the other hand, specifically, for research question 2, the ECDC dataset does not provide a comparison between developing and developed nations. Therefore, another dataset would be required to provide data collected for variables focused on nations development status. Consequently, with a dataset classifying nations as developed or developing and the ECDC dataset – research question 2 could then subsequently be appropriately visualised. Notably, nations states development could be classified in many different data types such as Boolean developed or undeveloped (e.g., yes or no), Ordinal (e.g., GDP/HDI) or Nominal (e.g., rankings). However, for this research paper, the internationally recognised ordinal data ranking of HDI will be favoured to visualise nation states contrasting levels of development.

Instead of the ECDC dataset, for research questions 3 and 4, discrete quantitative data (e.g., number of vaccines) and statistical quantitative data (e.g., number of vaccines as a ratio of population size) would be required to provide accurate visualisations on the global rollout of vaccines. Furthermore, this data would have to contain a temporal (i.e., chronological timestamps) to determine the rollout of vaccinations globally.

Lastly, with respect to research question 4, much like question 2 another dataset would be required to provide data collected for variables focused on nations development status. As noted previously, nation states development could be represented in many different data types such as Boolean developed or undeveloped (e.g., yes or no), Ordinal (e.g., GDP/HDI) or Nominal (e.g., rankings).

The fitness of each of the datasets

ECDC dataset strengths and weaknesses

Focusing on the strengths of the ECDC dataset. The dataset usefully comes prepared in three formats – csv, json and xlsx. Additionally, the dataset also contains many useful variables and records. The data is chronological and temporally ordered with timestamps for each day for the respective nation identified nominally by country name and country ID. Indeed, this structure enables extraction of useful discrete quantitative data (i.e., cases, deaths, population data) and continuous data (i.e., cumulative number of cases per 100,000 for 14 days).

The ECDC dataset can also be accurately queried and summarised into useful geological landmasses – via the nominal categorical data of continent. Other reasons the ECDC dataset is strong is that it has limited empty or null values and is therefore consequently very complete and consistent. Furthermore, the dataset comes from a source with an excellent reputation for data collection and accuracy namely the European Centre for Disease Prevention and Control, ECDC (2021). Equally, another advantage is that the ECDC dataset provides data updates on a monthly basis which is key for research question 1 and determining the spread of the virus over time.

Weaknesses of the ECDC dataset were briefly mentioned in the previous section – there are a variety of factors. For instance, the dataset terminates on the 14th of December and does not provide records for the variables up to the present day. Another weakness of the ECDC dataset is that it critically does not provide a collected variable for the amount of testing occurring within the relevant nations. Consequently, this undermines the validity of the data because it is difficult to determine if countries are equivalently testing for Coronavirus cases. For example, based on insights noted by other sources – it is very difficult to effectively compare the UK which conducted 27 Coronavirus tests per 1,000 with South Africa that managed 9 tests per 1,000 – almost 1/3 fewer tests, Soy (2020); Butcher (2020). Therefore, the UK could simply have a relatively higher number of respective cases than South Africa due to better reporting, health infrastructure and Coronavirus testing.

The ECDC dataset also suffers from other weaknesses. These include that the data gives quite a broad global snapshot of nations Coronavirus trend for 2020. Yet, it would have been useful to have more granular data – for example, provincial/county data or geographical (i.e., spatial coordinate longitude/latitude) related to spread of the virus within the very nation. To deduce if Coronavirus been localised to cities or certain provinces of the nation. Instead, the ECDC dataset provides a much broader collection of records for the underlying variables. Lastly, the ECDC dataset does not provide a summary of the underlying data – data is instead presented for the nation on a month-by-month basis. Consequently, for summary or totals of Coronavirus cases and death within a nation across 2020 external scripts to parse and collate data appropriately will need to be written.

Other datasets strengths and weaknesses

Having assessed the ECDC dataset provided by KCL for Track 1, now I will turn to the supplementary datasets sourced to answer the four research questions. Initially, the

John Hopkins University global dataset strengths include its accuracy, consistency and completeness as well as a usable API which provides new data every 10 minutes, CSSEGIS (2021). The dataset also offers variable data concerning recoveries, deaths, critical and cases and provincial data or states if in the US. Furthermore, the dataset also comes in csv, json and xlsx whilst data is provided up to the present day. Weaknesses of the dataset are that it provides a huge quantity of data for each day – external scripts would have to be written to effectively parse for the data from the start of each month over a prolonged period of time. Once again, the dataset does not critically include data concerning testing for every nation – compromising the underlying reliability.

The Our World in Data (OWID) dataset strengths include the fact that it is updated daily, its accuracy consistency and completeness as well as a usable API for web applications, OWID (2021). The dataset also offers the most extensive range of variables - providing confirmed cases, deaths, hospitalisations, testing and vaccinations throughout the duration for the Covid-19 pandemic. The complete dataset is provided in csv, xlsx and json format. Other interesting and useful variables with collected records within the dataset of interest included a range of useful health and population metrics from density, to age ranges, cardiovascular health, hand washing facilities, diabetes, smoking rates and HDI. However, weaknesses of the dataset are that it provides so much data that once again external scripts will have to be written to effectively parse and collect useful insights from the extensive dataset.

Disease.sh has provided a public API and records to extract government reported Coronavirus data for specific nations every 24 hours, Disease.sh (2021). Strengths include that the data comes from reputable government sources with high accuracy and the dataset also includes all available data presented by governments including tests, deaths and recoveries. However, weaknesses of the dataset are that it is not consistent and complete in structure – data is often null or missing. Furthermore, there are a large number of countries that are not supported with no records collected.

For determining the developmental status of nations for research questions 2and 4 the UN's HDI data offers an effective ranking of nations development on a scale from 0 to 1 for the years 1990-2019. Strengths include its classification of developing nations versus developed. However, weaknesses include its failure to include HDI data for 2020-1. Whilst, for vaccine data, the OWID dataset offers publicly available data for the vaccine rollout across countries globally. The dataset also provides key data in xslx, csv and json format.

Concluding remarks and comparisons between datasets

In terms of overall comparison between the datasets provided. For research questions, 1 and 2 the John Hopkins University, OWID and ECDC datasets offers a very rich and accurate observation of key variables such as number of cases, deaths and countries throughout the duration of the Coronavirus global pandemic. In contrast, the Disease.sh dataset is not quite as strong – containing limited and missing or null records for certain critical variables. Whilst for research questions 3 and 4, the OWID dataset offers the highest high accuracy and completeness for vaccination data. Lastly, for determining the developmental status of nations for research questions 2 and 4 the UN HDI data offers a complete, effective and internationally recognised dataset for nations developmental status.

1.3 Section C

Explain if and how the datasets described in b, and including the one provided, could be correlated?

For this section, I initially decided to look at specific positive correlations between the datasets before turning to potential negative correlations. From this analysis, I will effectively conclude on overall correlations between the datasets.

1.3.1 Positive correlations between datasets

The ECDC dataset provided by KCL would certainly have a largely positive correlation with the JHU, OWID and Disease.sh datasets, ECDC (2021); CSSEGIS (2021); OWID (2021); Disease.sh (2021). Indeed, all four datasets observe and collect records for similar variables and from similar sources. In terms of data sourcing, the ECDC dataset screens up to 500 sources to collect Covid-19 figures from 196 countries. This includes official websites of Ministries of Health, public health institutes and other national authorities. Similarly, the OWID dataset uses the WHO, the ECDC directly, the JHU directly, ONS and government announcements. Additionally, the JHU dataset is similarly based upon the WHO, the ECDC dataset, and a range of US data sources. Finally, the Disease.sh dataset uses Worldometers, the JHU directly and a range of government sources. Certainly, all four datasets use very similar legitimate and internationally recognised data sources - which will ultimately result in positively correlated datasets and Covid-19 data visualisations built on top of all the datasets records.

Following on from this initial analysis, upon deeper inspection and directly browsing the global visualisations and figures generated from the datasets (pleases see Figure 1) shows that they are indeed very similar with minor outliers. All four datasets have been used to effectively create dynamic global visualisations. Indeed, the ECDC and JHU datasets in particular have effective and interactive visualisation hubs available for public use. Notably, all four datasets similarly have Covid-19 data missing and null for certain specific nations such as Turkmenistan and North Korea.

Instead, with regards to the vaccine and HDI datasets. There would be an expected positive correlation between the OWID vaccine dataset and the UN dataset for HDI. Indeed, it is expected that nations with a higher HDI will be more effective at purchasing vaccines from the international pharmaceutical market due to more wealth and GDP. Equally, due to better health infrastructures it is expected that nations with a higher HDI will be able to rollout vaccinations more quickly and effectively – due to better health infrastructure, more trained medical staff and a variety of other key factors.

1.3.2 Negative correlations between datasets

Turning to negative correlations between the datasets. The underlying records for key variables within the JHU, OWID and ECDC dataset have been collected differently. The JHU and OWID dataset collected records for each nation every day up to the present day, whilst conversely the ECDC dataset provides daily records throughout just 2020. Additionally, as noted in the previous section, the specific sources used to collect records for the datasets are overwhelmingly similar yet ever so slightly different which will inevitably

result in discrepancies and potentially outliers when visualising the datasets at a a more granular and specific level. For instance, it is expected that the ECDC dataset is clearly EU-centric and will contain very accurate data for the EU. In contrast, the OWID and Disease.sh datasets are very global using a range of global sources. Lastly, the JHU dataset is US-based and uses a number of US news sources for data - resulting in accurate and potentially more specific data for the US.

This minor negative correlation between datasets is notably exposed when very closely granularly studying the visualisations and overall figures generated using the datasets. For example, the JHU global visualisation is tethered to specific regions and coordinates within each nation. Whilst, the ECDC and OWID visualisations produce chloropleth maps - generating overall colours for the entire nation. Therefore the datasets initially provide very different levels of granularity. Furthermore, the datasets calculate marginally different overall summed figures. At the time of writing:

- ECDC reports extended to the present day reports 123,636,852 global cases
- ECDC reports 2,721,891 global deaths
- JHU reports 127,319,002 global cases
- JHU reports 2,785,838 global deaths
- OWID reports 126,360,661 global cases
- OWID reports 2,783,768 deaths
- Disease.sh reports 127,185,164 global cases
- Disease.sh reports 2,783,800 deaths

Indeed, as we can see from these figures there are minor underlying discrepancies in the number of cases and deaths being reported by the four data sets. This proves that all the datasets are ever so slightly negative correlated with respect to their Covid-19 reporting and consequent different records for key variables.

Initial insights and findings from data scientists suggest there will be in the long term negative correlations between the ECDC, JHU and OWID coronavirus datasets and the OWID dataset for vaccinations. Indeed, over the same time period it would be expected that several of the ECDC, JHU and OWID coronavirus dataset variables would decrease in size for several nations whilst variables within the OWID vaccination dataset would increase over the same period. However, this negative correlation assumes that the vaccine rollout has been successful within the nations and that enough vaccines have been administered to reduce the spread of Coronavirus globally.

1.3.3 Concluding remarks on correlations and uncertainties

From the data and contrasting perspectives of a variety of data scientists: Foley (2020); Kay (2021); Shahbazi and Khazaei (2020); Ross et al. (2020); Roghani and Panahi (2021); Soy (2020) and global public health administrators, at present there is an uncertain and debated correlation between the ECDC, JHU and OWID coronavirus datasets and the UN dataset for HDI. Indeed, the spread of the Coronavirus was originally expected to be lower in

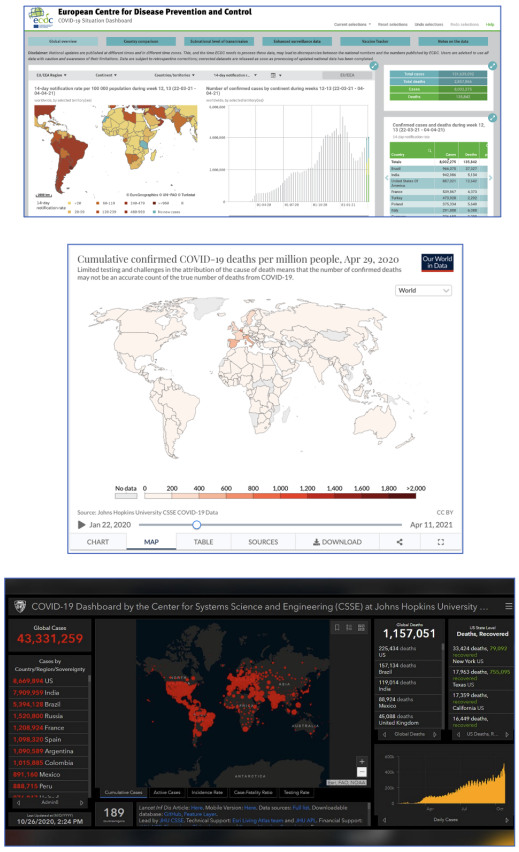

Figure 1: Positively correlated and similar Coronavirus global visualisations generated using from top - the ECDC, OWID and JHU datasets.

developed nations with a higher HDI i.e., more developed nations. In contrast, the spread of the Coronavirus was originally expected to be higher in less developed nations with a lower HDI i.e., developing nations. However, this readily assumes that the Coronavirus will spread more effectively across developing nations due to their poorer health infrastructure, worse living conditions, difficulties maintaining social distancing and a range of other factors.

The underlying data shows that this original assumption is not true in all cases and that certain countries with a high HDI have faced a significant Coronavirus spread versus equivalent developing nations. Yet once again, it is very difficult to discern if this correlation is accurate from the data given the highly unequal rates of testing between developed and developing nations. Therefore, ultimately over time a clearer correlation on the underlying spread of coronavirus versus HDI will be reached.

To conclude, this section on assessing datasets - it is clear collectively between that all the datasets have both positive, slightly negative and even uncertain correlations which is interesting and useful to consider when developing effective data visualisations.

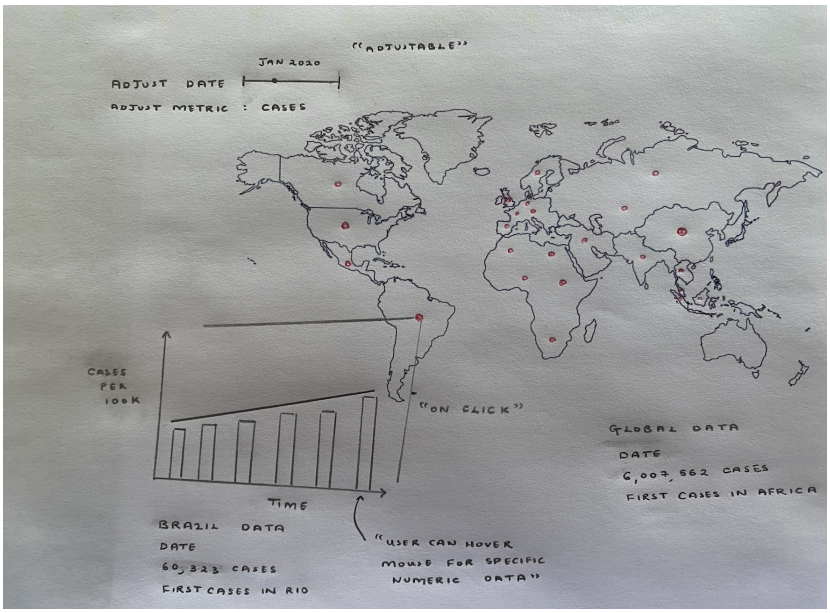

Figure 2: Sketch for Design 1 - users are provided with a global bubble chart map and snapshot of global statistics - users can adjust the date, the bubble chart metric for cases, cases per 100K and other metrics denoting global spread. Users can also click on the proportionally sized bubbles and will be provided with a hover-able histogram for current spread trends of that specific nation.

2 Part 2: Design and discussion

2.1 Overview of design process

For the design and prototyping stage, I decided to employ a variety of ideation methods, which can be divided into three key stages. For each research question – initially, I employed sketching, followed by application in PowerPoint/Excel before lastly building a low fidelity prototype in D3. For the sketching stage, I used two main techniques: firstly, I tried collect a variety of visualisations for inspiration before secondly, trying to sketch a highly effective visualisation. In contrast, the D3 low fidelity development focused on exploration, fast development and creatively challenging initial assumptions.

2.2 Design 1 - Analyse the spread trend of this virus all over the world. What is the spread over time?

2.2.1 Visualisations for inspiration

For this design I was strongly influenced by the Coronavirus global bubble map charts – favoured and developed by many websites tracking the international spread of the Coronavirus. Good examples of pre-existing bubble map charts used for inspiration include:

- The WHO Coronavirus Dashboard - `https://covid19.who.int/`
- The Microsoft Covid-19 tracker - `https://www.bing.com/covid`

- The John Hopkins University Coronavirus dashboard - `https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6`

The popularity of using a bubble map chart is most probably as a consequence of the bubble map charts effectiveness at visualising location and proportion simplistically and effectiveness at denoting global trends for the Coronavirus.

2.2.2 Design rationale

Visual encodings used

An interactive bubble map was chosen as the design to analyse the spread trend of the Coronavirus all over the world. The marks used are points as bubbles, which use the visual channel and vary in size proportional to the number of cases present within the nation they are representing. In terms of colour a simple single colour is used (i.e., red) on top of a simple Mercator map projection in white and black. Additionally, a simple legend is utilised in the corner of the visualisation to denote the current date, global total number of cases and continent in which cases have most recently emerged. The visualisation should have two levels of user interaction: (1) to enable the user to set the date and (2) to enable the user to click on a nation / its bubble and be presented with the relevant case numbers within that nation via figures and a histogram.

Encoding choice and data appropriateness

The encoding choice is highly appropriate for the underlying data provided by Kings College London from the ECDC. The visualisation matches with the ECDC data provided in four ways.

Firstly, the data provides the changes in cases for each nation over a year – in much the same way this visualisation will intuitively encode the global data changes over time. Secondly, the visualisation provides effective contextual understanding with the size of the simple symbol (i.e., the transparent red bubbles) changing for each nation in relation to number of coronavirus cases. Thirdly, the legend updates on the basis of the ECDC data which can sum the number of coronavirus cases for all nations. Fourthly, the data can be depicted on a more granular level via useful user interaction in which the user can click bubbles and see the trends for countries through figures and a histogram.

The only data manipulation necessary would be to tie the ECDC data to each nation – the underlying Mercator map projection and its latitude / longitude – this could be performed via utilising a secondary data source which provided the map projection and the relevant coordinates of the nations.

Strengths and weaknesses of the visualisation and encoding choice

Strengths of the visualisation include its clarity at representing a sophisticated data source. Indeed, the visualisation enables ease in understanding the distribution of the relative spread trend of this virus all over the world tagged to time. Additionally, the use of the Mercator map projection provides the ability to compare across several nations at a glance and contextualise the data within the real world. Furthermore, the visualisation

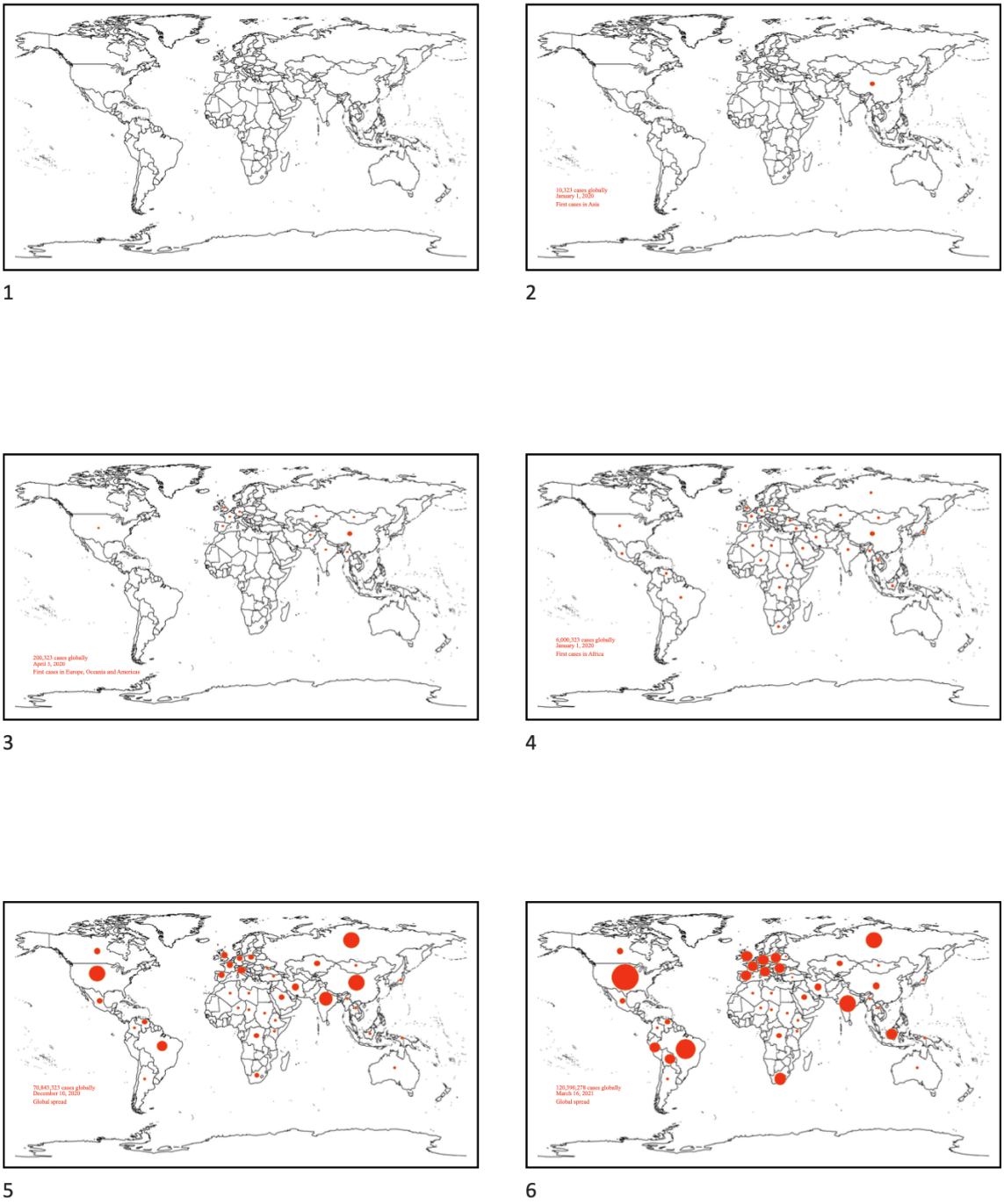

Figure 3: Low fidelity development in Powerpoint of Design 1 - the user should be able to set the date and click on a bubble to get shown relevant case numbers for the respective nation including a histogram of that nations data over the time period.

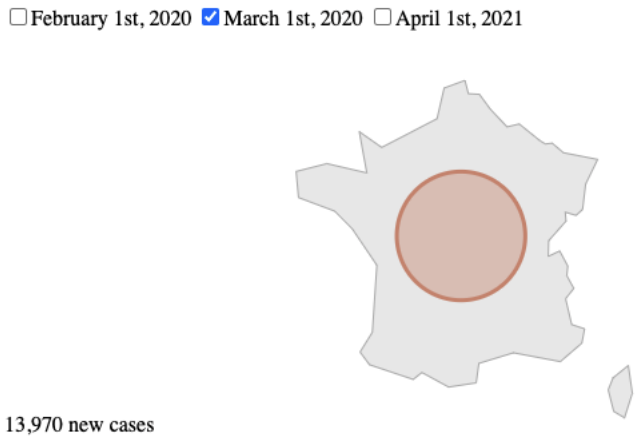

Figure 4: High fidelity and fast development of Design 1 using D3 to test the concept for the nation of France - the data is extracted from an underlying JSON file.

is deliberately refined with limited to colour so as not to overwhelm the users cognitive processing. The final strength of the visualisation is that it does depict granular data relatively effectively via users clicking on bubbles they can effectively see histograms and accurate figures tied to specific regions.

Weaknesses of the visualisation include, that using bubbles varying by area size as a visual channel is an ineffective choice because users struggle to accurately visualise size. In response to this criticism to use another visual channel such as length (i.e., a 3D map projection) or more colours (i.e., a choropleth map) is perhaps arguably unnecessarily sophisticated and does not clearly show the relative spread of the Coronavirus over time.

2.3 Design 2 - What is the spread of the virus in the developing versus developed world. What initial insights can be reached?

2.3.1 Visualisations for inspiration

With regards to this design, I was strongly influenced by Hans Rosling's renowned Gapminder bubble chart visualisation `https://www.gapminder.org/tools/?from=world`

There are also a number of other HDI centric graphs using similar Cartesian x-y axis with bubble's used for denoting country, the area / size of the bubbles to denote population and the colour of the bubbles for continent. Other influential visualisations include:

- World Data Visualisation prize winner by Dimiter Toshkov - `http://www.dimiter.eu/Visualizationsfiles/WDVP.html`
- OWID HDI graph - `https://ourworldindata.org/human-development-index`

Notably. bubble graphs are effective at depicting 3 numeric variables rather than just two as denoted by a common scatter plot.

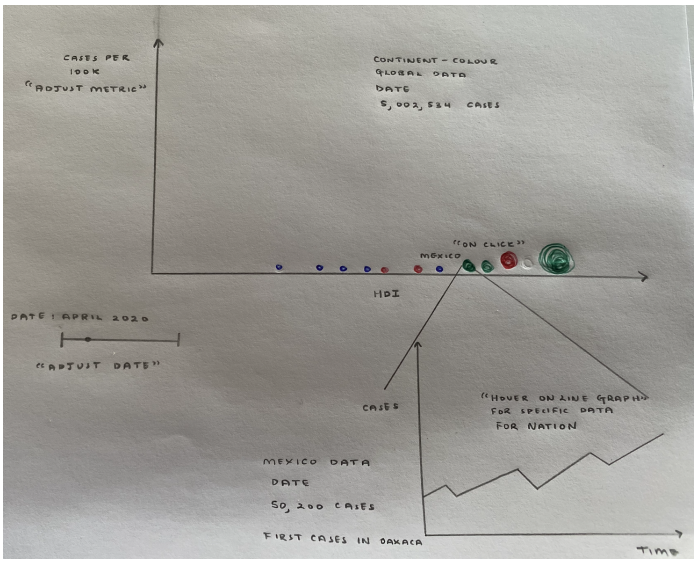

Figure 5: Sketch for Design 2 - users are provided with a bubble chart and snapshot of global statistics - users can adjust the date, the bubble chart metric (colours denote continent) and other metrics denoting global spread versus HDI. Users can also click on the proportionally sized bubbles and will be provided with a hover-able line chart for current spread trends of that specific nation.

2.3.2 Design rationale

Visual encodings used

An interactive bubble chart was used to encode the spread of the virus in the developing versus the developed world. The chart uses two scales – a y-axis denoting the number of Coronavirus cases per 100,000 inhabitants and an x-axis denoting HDI. Following this, countries are denoted by bubbles with the respective area of the bubbles representing the number of coronavirus deaths per 100,000 inhabitants (higher number of deaths i.e., larger bubbles). Lastly, the colour of the bubbles denotes the 6 continents of the countries (i.e., blue Africa, green the Americas, yellow Europe, purple Asia and red Oceania). Much like the first visualisation, the visualisation should have two levels of user interaction: (1) to enable the user to set the date and (2) to enable the user to click on a nation / its bubble and be presented with the relevant case numbers within that nation denoted via a line graph.

Encoding choice and data appropriateness

The encoding choices are highly appropriate for with the ECDC dataset provided by KCL. The visualisation is supported by the underlying data in three ways.

Firstly, the ECDC dataset provides the y-axis (i.e., cases per 100K) and the area size of the country's bubbles (i.e., deaths per 100K) – as the ECDC dataset contains both deaths, cases and census population data. Secondly, the ECDC datasets supports the colour encoding of the bubbles – by providing continent data for each nation. Thirdly,

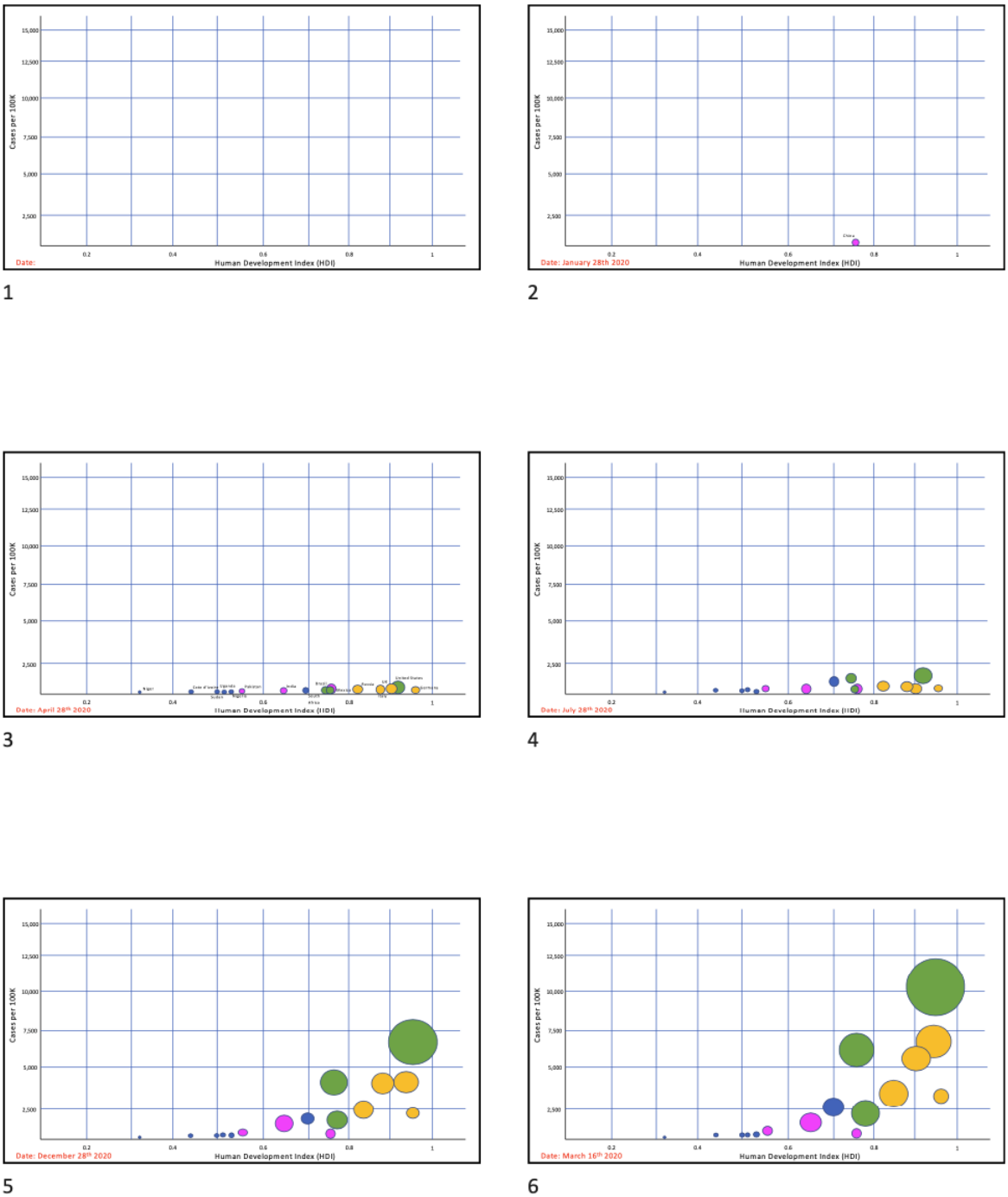

Figure 6: Low fidelity development in Powerpoint of Design 2 - the user should be able to interactively set the date generating multiple versions of the visulisation (as denoted above), adjust the axes in respect of case number per 100K and HDI and click on bubbles for a closer inspection of the underlying data.

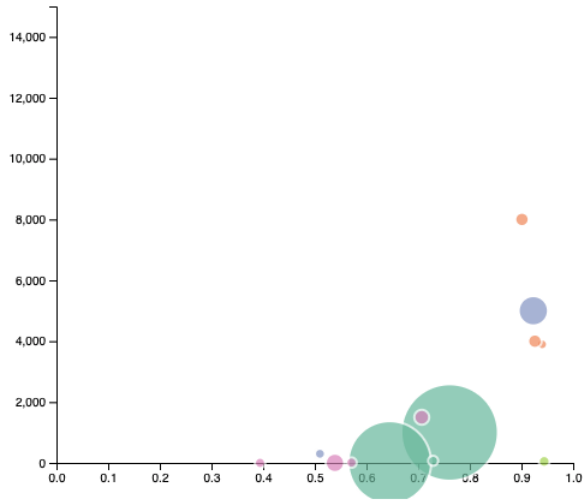

Figure 7: High fidelity and fast development of Design 2 using D3 to test the concept of a bubble chart - dummy data dynamically adapts the bubbles colour based on continent and proportion relative to COVID data within a separate JSON file.

reproducing the graph over multiple time periods would be possible as the underlying ECDC data provides day by day data for each nation.

Further data required to complete the graph would be a simple dataset providing the countries and their associated HDI in order to create the x-axis and countries related index – this data can be provided from the UN dataset.

Strengths and weaknesses of the visualisation and encoding choice

There are many strengths of the visualisation, most notably the thorough use of the ECDC dataset and the encoding of four different variables via colour, area size and two axes. Indeed, the visualisation very effectively encodes trends in the data regarding differences in the spread of the coronavirus in the developing virus developed world across the x-axis. Furthermore, the visualisation even depicts the voracity of the virus within each nation by denoting recorded deaths per 100K with the fluctuating size of the bubbles and shift upwards or downwards in relation to the y-axis.

Weaknesses of the visualisation are that potentially, the bubbles could get cluttered with more nations on the graph. In response to this, the bubbles could be made transparent and if the bubbles are dynamic if clicked upon providing data concerning the nation the bubble represents – the graph should be clear. An additional potential weakness is that the graph uses too many visual channels and that the underlying data for the developing world is potentially inaccurate. Rebuking this argument, the use of multiple visual channels including the bubbles makes the visualisation very accessible - colour choices were even deliberate for a colour blind accessible palette. Furthermore, the accuracy of the underlying data for the developing world is a wider contextual point unrelated to the visualisation itself.

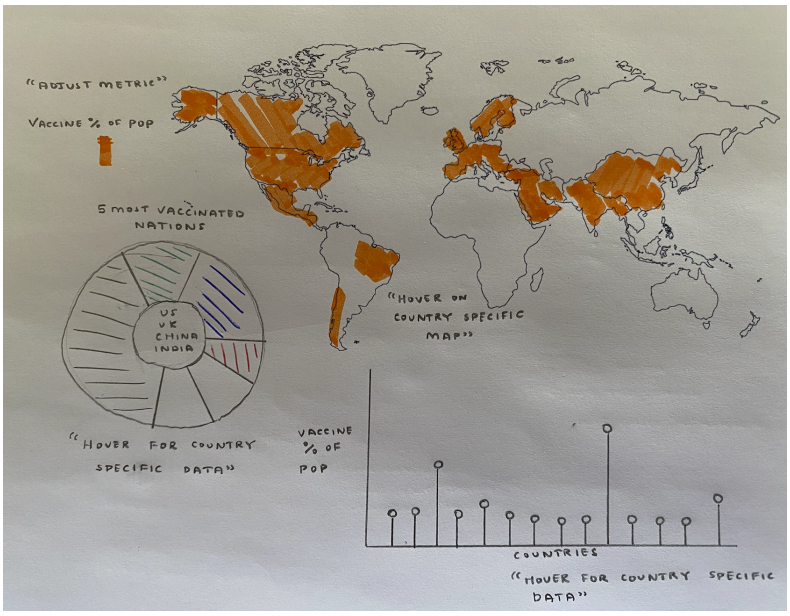

Figure 8: Sketch for Design 3 - users are provided with a hover-able Mercator chloropleth map projection and snapshot of global statistics - users can adjust the chloropleth metric provided in accessible orange. Users are also provided with a hover-able donut chart of the most vaccinated nations and a hover-able lollipop chart for the global trends of the vaccine rollout and number of people vaccinated.

2.4 Design 3 - Analyse the global rollout of vaccines over the world. What can be noted about the global rollout of vaccinations?

2.4.1 Visualisations for inspiration

For the visualisation there is a wide range of inspiration as choropleth maps are widely used and have been favoured to represent data for Coronavirus cases and for the global rollout of vaccines:

- WHO coronavirus and vaccine data represented by choropleth maps - `https://covid19.who.int/`
- ECDC coronavirus data represented by a variety of chlorophleth maps - `https://vaccinetracker.ecdc.europa.eu/public/extensions/COVID-19/covid-19.html`

2.4.2 Design rationale

Visual encodings used

An interactive choropleth map utilising Mercator map projection was chosen to analyse the global roll out of vaccines all over the world. A systematic list and description of visual map variables used by a choropleth map was originally developed by Bertin (1967) with

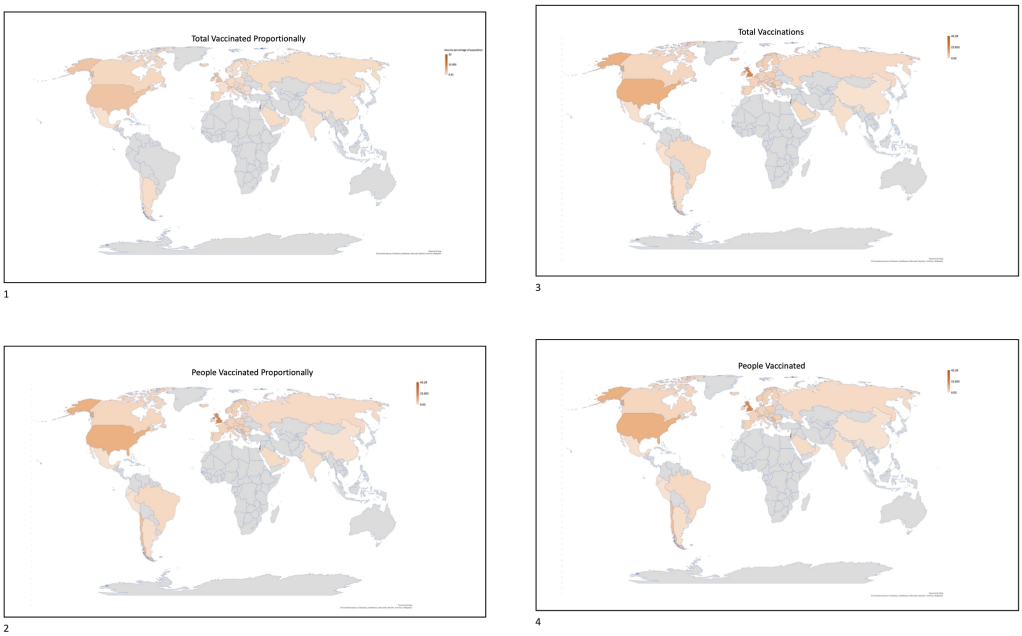

Figure 9: Low fidelity development in Excel of Design 3 - the user should be able to interactively adapt the vaccination metrics generating multiple versions of the visualisation (as denoted above) and closely inspect the data using lollipop and donut charts. High fidelity development in D3 did not occur as I implemented this visualisation for Part 3.

seven visual variables (position, size, shape, orientation, colour, value, texture) that was extended by Morrison (1974) to nine visual variables and even more extended to an even more pronounced list by MacEachren (1995) with twelve visual variables [(1) location, (2) size, (3) shape, (4) orientation, (5) colour hue, (6) colour value, (7) texture, (8) colour saturation, (9) arrangement, (10) crispness, (11) resolution, and (12) transparency].

The predominant encoding from the choropleth map is colour hue i.e., lower values equates to less vaccinated countries as a percentage of population being lighter transitioning to more vaccinated countries as a darker hue. In terms of colour, a simple monochromatic orange was chosen because it is accessible to those suffering with colour blindness and visual difficulties. Additionally, a slightly off-tone white background was chosen to provide suitable contrast with the orange, grey countries and blue borders. In terms of interaction, the user should be able to view data from each nation – so they can effectively visualise the global rollout of vaccinations as well as inspect the overall data more closely via a donut chart and lollipop chart.

Encoding choice and data appropriateness

The encoding choice is appropriate for the underlying data provided by the OWID public dataset for vaccinations. The OWID dataset provides the key metrics that underpin the visualisation – namely people vaccinated per hundred and total vaccinations as well as population size by nation. The utilisation of percentage of population vaccinated is deliberate because it prevents the use of misleading absolute numbers and figures.

Additionally, the data has been collected throughout the duration pandemic from a range of credible sources notably government reporting, the ECDC and John Hopkins University (JHU) – thus it is possible to create a dynamic visualisation of the rollout of vaccinations across the world with a series of choropleth maps. Furthermore, the dataset also provides necessary geographical data to produce a Mercator map projection.

Strengths and weaknesses of the visualisation and encoding choice

Initially, the visualisation has a wide range of strengths. As noted by Schiewe, "choropleth maps effectively and intuitively display spatial patterns". Indeed, the monochromatic orange hues enable a wide range of specific values to be represented regarding the percentage of nations inhabitants vaccinated. Furthermore, the Mercator map projection is visually compelling and effective – displaying large amounts of information and effectively showing visual patterns. For instance, this provides many different levels of analysis from general overall patterns to the detection of more specific details. Lastly, the choropleth map enables hotspots to be effectively detected in relation to location and area (i.e., several neighbouring countries which have vaccinated large proportions of the population).

However, the visualisation does also suffer from weaknesses. These include that the map uses an average to represent defined areas, not providing granular or detailed information of internal conditions (e.g., have as many inhabitants been vaccinated in London as in Liverpool). In response to this limitation, a more detailed version of the map could be made providing accessible data concerning specific regions of countries to mitigate against this weakness. Furthermore, the data for certain nations concerning the number of inhabitants vaccinated by geolocation is not always available. Additionally, the Mercator projection is misleading in terms of its representation of nations area and size. In response to this weakness an Equal area map projection could instead be used – however the Mercator projection is more widely used and effectively fits with users pre-existing mental models.

2.5 Design 4 - What are the key differences in the vaccine roll out in the developing versus developed world. Has the vaccine rollout been unequal?

2.5.1 Visualisations for inspiration

With respect to this visualisation and the use of a Sankey diagram to denote vaccines and the rollout between developed and developing countries – was slightly more uncommon. Nonetheless, more broadly there are range of influential Sankey diagrams which represent data very effectively – this includes:

- One notable example of an effective visualisation of Covid cases was found under - `https://covid.lonnygomes.com/`.
- Minard's classic Sankey diagram of Napoleon's invasion of Russia.

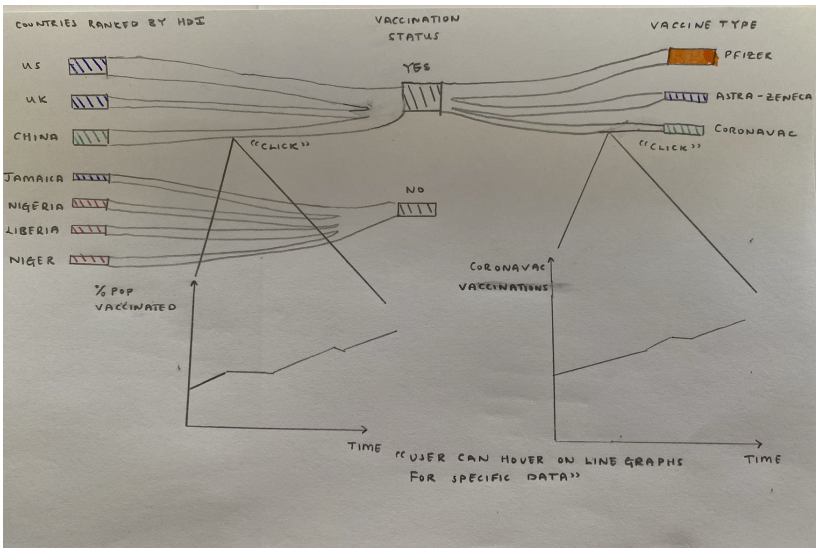

Figure 10: Sketch for Design 4 - users are provided with a hover-able Sankey diagram and snapshot of global statistics - countries are ranked by HDI and the thickness of the nodes denotes the proportion of that metric. Users can click on the nodes to get a more specific line graph of the metric data over time. Users can then hover on the line graph to get more specific numerical data of nations vaccination status and the type of vaccine used by the nations e.g. Pfizer, Oxford AstraZeneca.

2.5.2 Design rationale

Visual encodings used

An interactive Sankey diagram to visualise the vaccine roll out in the developing versus the developed world. Consequently, the diagram utilises multiple different encodings – notably: size/shape (i.e., flow thickness) and ordering on the y-axis of the Sankey diagram according to HDI.

The size/shape (i.e., flow thickness/width of the bands) denotes the percentage of population vaccinated for each respective nation. Therefore, this captures that nations with a thicker flow will have a higher proportion of their population successfully vaccinated with one dose. In contrast, the ordering on the y-axis presents the countries according to HDI with the more developed countries at the top of the y-axis and less developed countries towards the bottom of the axis.

Encoding choice and data appropriateness

The encoding choice is appropriate for the underlying data provided by the OWID public dataset for vaccinations. The OWID dataset provides the key metrics that underpin the visualisation – namely people vaccinated per hundred, current HDI and total vaccinations as well as population size by nation. The utilisation of percentage of population vaccinated is deliberate because it prevents the use of misleading absolute numbers and figures.

Additionally, the data has been collected throughout the duration pandemic from a

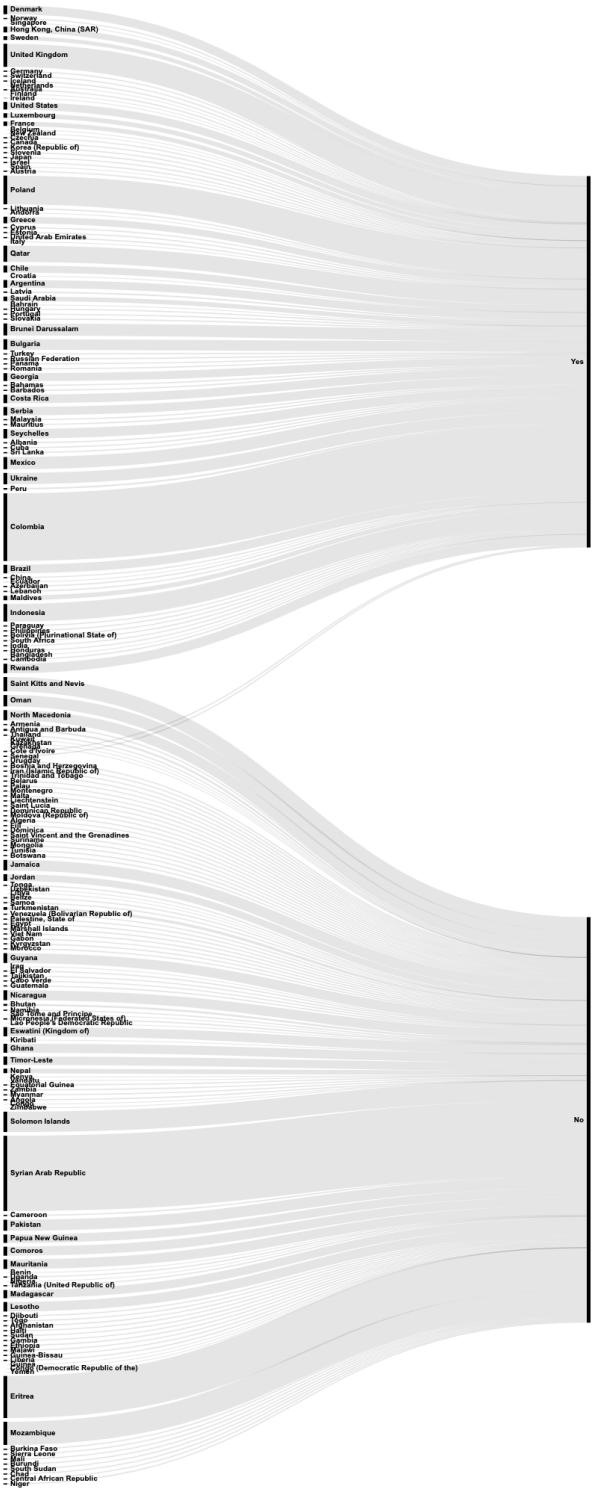

Figure 11: Low fidelity development of Design 4 in RAWGraphs 2.0. The y-axis denotes the HDI standing of countries - users should be able to, set the date (to see the trend over time), click on a nation's node enabling the node to change colour respective to other nations and be highlighted as well as be provided with a small line graph denoting the recent vaccination over time.

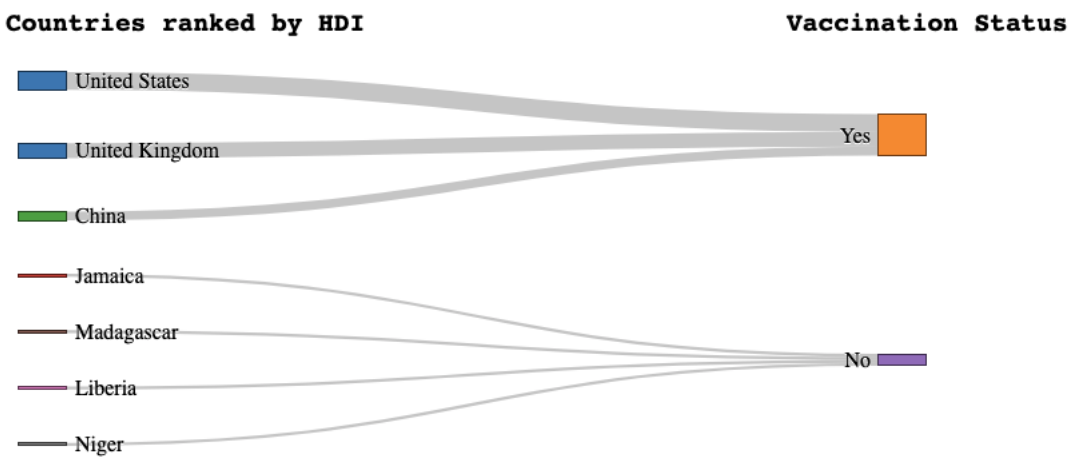

Figure 12: High fidelity and fast development of Design 4 using D3 to test the concept of a Sankey diagram - dummy Coronavirus data dynamically adapts the nodes and width of the link, colour based on continent and ranking based on HDI sourced from a single JSON file.

range of credible sources notably government reporting, the ECDC and John Hopkins University (JHU) – thus it is possible to create a dynamic Sankey visualisation of the rollout of vaccinations.

Strengths and weaknesses of the visualisation and encoding choice

The Sankey diagram visualisation has a range of strengths. Firstly, the Sankey diagram helps effectively depict the trends between the developed and the developing world in terms of vaccine rollout and vaccine accessibility. Secondly, coupled with the axis being ordered in terms of HDI – it is very easy to see the significant difference in flow between the developed and developing world – the major statistical significance of the inequality. Thirdly, the Sankey diagram makes it very easy to detect inconsistencies in the data (i.e., very undeveloped countries having access to vaccines).

There are three key weaknesses of the Sankey diagram. Initially, it would be quite hard for users to digest because it is a more rare visualisation and does not readily easily fit with users pre-existing mental models. For instance, it is very difficult to discern if two nations flow have similar widths. Additionally and secondly it is difficult to gain a more specific understanding of the data when the flows have similar widths. Indeed, this is certainly a heightened problem if the nodes are not spaced out well. Thirdly, particular flows from the Sankey diagram are overlapping and the diagram is slightly cluttered which can reduce the effectiveness of users clearly discerning insights.

Nonetheless, the visualisation I have developed does have a range of positive and graphical advantages. The Sankey graphic illustration of vaccinations is very visually striking and capably illustrates the major transfers and flows of global vaccine supply. Indeed, it is very easy to see the dominant advantages of developed countries and upon closer inspection the few underdeveloped countries that exist as outliers successfully accessing vaccines.

Furthermore, it would be even better to build on this initial Sankey and denote the vaccine source (i.e. the types of vaccines) - effectively determing the types of vaccines nations are getting. For instance, poorer countries may only have access to vaccines which are less recommended by global health authorities (i.e., the WHO) - which still demonstrates a clear vaccine inequality between developed and developing nations.

Acknowledgements

The author acknowledges the generous support and help during difficult circumstances from the lecturers and teaching assistants at Kings College London, Department of Informatics.

References

Butcher, T. K. . B. (2020, Oct). Covid in europe: How much testing do other countries do?

CSSEGIS (2021). Cssegisanddata/covid-19.

Disease.sh (2021).

ECDC (2021). Homepage: European centre for disease prevention and control.

Foley, K. E. (2020). Vaccines won't eradicate covid-19-and that's ok.

Kay, C. (2021). Available at `https://www.bloomberg.com/news/articles/2021-03-25/curbs-by-world-s-biggest-vaccine-exporter-to-hit-poorest-nations`.

OWID (2021). owid/covid-19-data.

Roghani, A. and S. Panahi (2021). The global distribution of covid-19 vaccine: The role of macro-socioeconomics measures. *medRxiv*.

Ross, J., C. M. Diaz, and J. L. Starrels (2020). The disproportionate burden of covid-19 for immigrants in the bronx, new york. *JAMA internal medicine 180*(8), 1043–1044.

Sagar, A. D. and A. Najam (1998). The human development index: a critical review. *Ecological economics 25*(3), 249–264.

Shahbazi, F. and S. Khazaei (2020). Socio-economic inequality in global incidence and mortality rates from coronavirus disease 2019: an ecological study. *New microbes and new infections 38*, 100762.

Soy, A. (2020, May). Lack of covid-19 testing undermines africa's 'success'.