IVAR-Shiny - Interactive Visual Analysis with R Shiny for Exploring COVID-19 Data

Lin Yongyan
Masters of IT in Business (MITB), Singapore Management University (SMU)
yongyan.lin.2020@mitb.smu.edu.sg

Siow Mun Chai
Masters of IT in Business (MITB), Singapore Management University (SMU)
mcsiow.2020@mitb.smu.edu.sg

Tan Wei Li Amanda
Masters of IT in Business (MITB), Singapore Management University (SMU)
amandatan.2020@mitb.smu.edu.sg

ABSTRACT

In today's Data Age, data on COVID-19 are related data available, visualisations and studies.

It consists of two paragraphs.

1. INTRODUCTION

The Coronavirus (COVID-19) has caught the world's attention with the first COVID-19 cases reported in Wuhan, Hubei, China, in December 2019. In the global battle against the virus, countries seek to understand the virus, its spread, impact and more recently, receptivity towards the COVID-19 vaccination. We are currently living in the Data Age, where many COVID-19 related data are made available on the Internet. This has facilitated numerous, but not limited to, epidemiology and statistical studies across the globe.

In the data science realm, many data-driven applications are developed to provide a one-stop information hub for the public. These applications are typically developed using programming languages such as HTML, Java and JavaScript. With the growing popularity of R, and its ability to create web applications using the R Shiny package, the creation of interactive visualisations without having in-depth web programming knowledge has been made possible.

In this paper, we aim to leverage the richness of the COVID-19 data to provide an interactive experience in generating insights and analyses using R Shiny from three key aspects: (1) new cases; (2) deaths; and (3) vaccination receptivity.

2. MOTIVATION OF THE APPLICATION

There are several one-stop applications that allow interactive visualisation of COVID-19 related data across time. These applications typically report number of events i.e. number of new cases/deaths/tests conducted, number of people vaccinated. Deeper exploration and analysis on COVID-19 trends and relationships with other factors or indicators are done in silos and majority of such studies report their findings based on pre-defined variables and specific analysis models.

With this application, we hope to combine and provide an interactive experience for in-depth exploration and analysis of the COVID-19 data. The three key aspects selected for the application are:-

- Predictive analysis of new cases
- Bivariate and multivariate analysis of deaths and death rates with health, economic and population structure indicators
- Exploratory and bivariate analysis of vaccination receptivity with virus perception and demographics

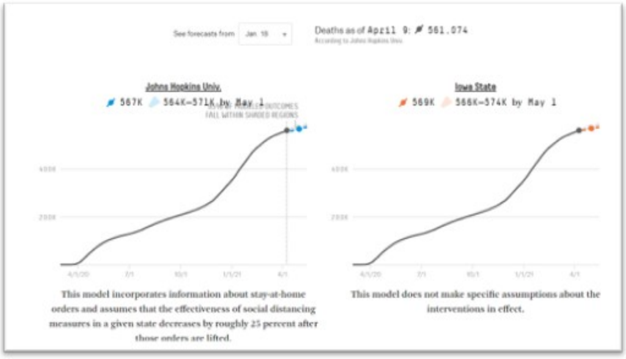
Data is obtained from several sources: Center for Systems Science and Engineering (CSSE) at Johns Hopkins University for COVID-related data; Our World in Data, World Bank, UNdata, United Nations Development Programme (UNDP) for health, economic and population structure indicators; and Imperial College London YouGov COVID-19 Behaviour Tracker Data Hub for survey data on virus perception and vaccination receptivity.

3. REVIEW AND CRITIC ON PAST WORKS

As there are three components to the application, the discussion is done separately for each component.

3.1 New cases

Most studies that forecast the number of new cases use time series charts with confidence interval of the predicted values. The use of the confidence interval shows the range in which the predicted values will fall within and provides a sense of the prediction variation. Most predictive models used and the model input parameters are usually pre-defined, with only a handful of studies comparing the results from different models. The visualisation in Figure 1 compares the time-series chart of different assumptions made in the predictive analysis, while not providing information on the predictive


See forecasts from Jan. 18

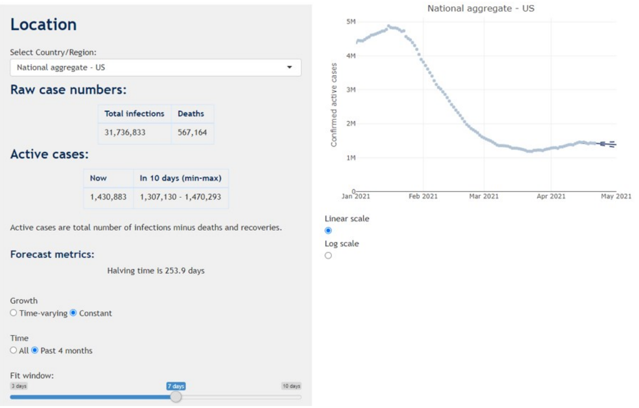
Deaths as of April 9: 561,074

According to Johns Hopkins Univ.

This model incorporates information about stay-at-home orders and assumes that the effectiveness of social distancing measures in a given state decreases by roughly 25 percent after those orders are lifted.

This model does not make specific assumptions about the interventions in effect.

Figure 1: Comparison of predictive models with option for user to select starting month that forecasts will be based on. Source: `https://projects.fivethirtyeight.com/covid-forecasts/`


Location

Select Country/Region: National aggregate - US

Raw case numbers:

Total infections	Deaths
31,736,833	567,164

Active cases:

Now	In 10 days (min-max)
1,430,883	1,307,130 - 1,470,293

Active cases are total number of infections minus deaths and recoveries.

Forecast metrics:

Halving time is 253.9 days

Growth
○ Time-varying ● Constant

Time
○ All ● Past 4 months

Fit window: 3 days 7 days 10 days

Linear scale ●

Log scale ○

Figure 2: Time-series forecasting with model options, model input parameters and graph scale options. Source: `http://us.covid19forecast.science.unimelb.edu.au/`

model used. The model input parameters is also limited to the starting month to based the forecast on.

Another predictive analysis (Figure 2) allows more flexibility to the users, whereby the user can select between two simple models (constant or time-varying growth) with model input parameters such as the time period to calculate the predicted values. Other parameters such as country and graph scales are also available for selection. In this visualisation, the available models are limited and simplistic and there is a lack of model assessment metrics e.g. Root Mean Square Error (RMSE).

Majority of the visualisations reviewed do not allow users to explore and understand the data before proceeding to the forecast.

To allow user to have a more holistic predictive analysis of the new cases, the application will attempt to combine data exploration of the trend, seasonality and anomaly (if any) of the time-series data and predictive modeling. For the predictive modeling, users will be given the option to select and compare the predictive models, and define model parameters such as the date range to be used for the forecasting. Model assessment measures will also be included.

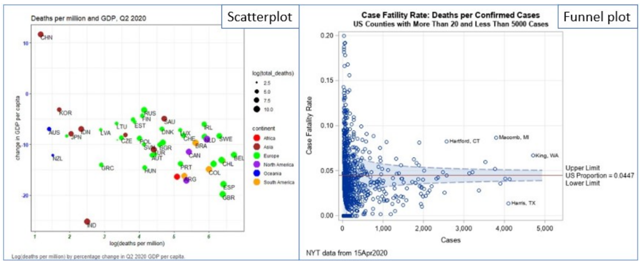


Figure 3: *(a)* Scatterplot on change in GDP per capita against deaths per million [9]. *(b)* Funnel plot on case fatality rate against number of confirmed cases [11].

3.2 Deaths and death rates with health, economic and population structure indicators

All the one-stop COVID-19 applications report the death toll by location using geo-spatial, time-series and/or in tabular form (see Figure 1). There are lesser analyses that study the relationship between deaths or death rates with other indicators, with the majority of them seeking to explain a causal relationship between the COVID-19 numbers and the indicators. The review will focus on the analysis and visualisations used in these analyses.

The scatterplot (Figure 3a) is useful in showing the relationship between two independent variables. Scales can be employed to encode useful variables not represented on the plot. However, it may be difficult to clearly differentiate points on the plot when the number of data points increases.

The funnel plot is another graph that shows the relationship between two variables that are dependent on each other e.g. case fatality rate against the number of confirmed cases, where case fatality rate is calculated as a ratio of the number of deaths to the number of confirmed cases. There are two similar visualisations created specifically for COVID-19 case fatality rate for counties in the US by a SAS researcher Rick Wicklin (see Figure 4). The funnel plot seeks to highlight any anomalies from the expected range of the numerical values based on statistical concepts. The drawback of the visualisation is that the funnel plots are static with no interactivity: users are unable to identify the other data points that are not labelled and do further exploration with other variables.

3.2.1 Multiple regression model

The visualisations discussed thus far are bivariate in nature: analysis of each factor with the number of deaths. There are very few multivariate analysis done, and of those conducted, most of them are presented in tabular form or described in text. There is one study on regression models to predict the number of COVID-19 new deaths, which presents its findings visually in the research paper (see Figure 5).

There are gaps in the current visualisations in supporting the intended analysis. The majority of interactive visualisations are univariate analysis presented on maps or in time series,

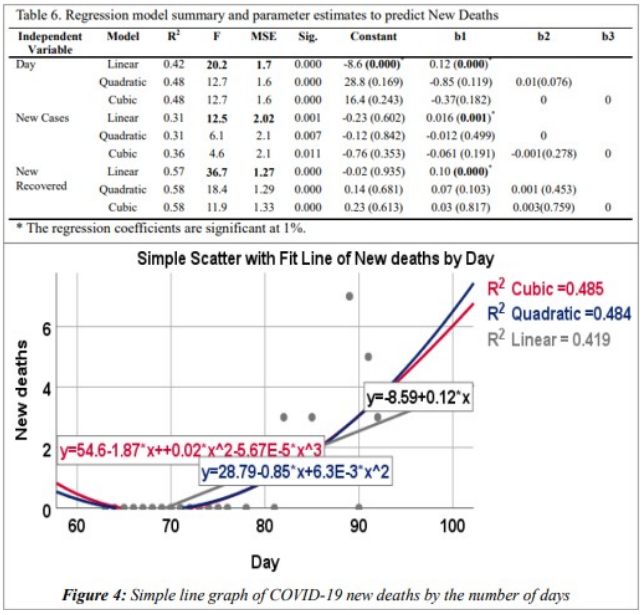


Table 6. Regression model summary and parameter estimates to predict New Deaths

Independent Variable	Model	R^2	F	MSE	Sig.	Constant	b1	b2	b3
Day	Linear	0.42	**20.2**	**1.7**	0.000	-8.6 **(0.000)***	0.12 **(0.000)***		
	Quadratic	0.48	12.7	1.6	0.000	28.8 (0.169)	-0.85 (0.119)	0.01(0.076)	
	Cubic	0.48	12.7	1.6	0.000	16.4 (0.243)	-0.37(0.182)	0	0
New Cases	Linear	0.31	**12.5**	**2.02**	0.001	-0.23 (0.602)	0.016 **(0.001)***		
	Quadratic	0.31	6.1	2.1	0.007	-0.12 (0.842)	-0.012 (0.499)	0	
	Cubic	0.36	4.6	2.1	0.011	-0.76 (0.353)	-0.061 (0.191)	-0.001(0.278)	0
New Recovered	Linear	0.57	**36.7**	**1.27**	0.000	-0.02 (0.935)	0.10 **(0.000)***		
	Quadratic	0.58	18.4	1.29	0.000	0.14 (0.681)	0.07 (0.103)	0.001 (0.453)	
	Cubic	0.58	11.9	1.33	0.000	0.23 (0.613)	0.03 (0.817)	0.003(0.759)	0

* The regression coefficients are significant at 1%.

Simple Scatter with Fit Line of New deaths by Day
R^2 Cubic =0.485
R^2 Quadratic =0.484
R^2 Linear = 0.419
y=-8.59+0.12*x
y=54.6-1.87*x++0.02*x^2-5.67E-5*x^3
y=28.79-0.85*x+6.3E-3*x^2
New deaths
0
2
4
6
60
70
80
90
100
Day

Figure 4: Simple line graph of COVID-19 new deaths by the number of days

Figure 4: ***(Top)*** **Table showing the regression models summaries and parameter estimates to predict new COVID-19 deaths.** ***(Bottom)*** **Scatterplot with fit lines comparing the regression models built to predict new COVID-19 deaths by day [1].**

while the bivariate and multivariate analysis of the country indicators and the number of deaths are largely static. The application will attempt to create interactive visualisations for bivariate (scatterplot and funnel plot) and multivariate analysis (multiple linear regression). The focus will be on the cumulative or total number of deaths, so that more meaningful relationships can be observed between the COVID-19 related data and national aggregate indicators.

3.3 Vaccination receptivity

The review of current visual analytic techniques of survey data can be categorised into three areas: (1) representation of Likert scales; (2) visualising uncertainty; and (3) visualising correlation.

3.3.1 Representation of Likert scales

There are several ways to visually represent Likert scales: (1) grouped column/bar charts; (2) pie charts; (3) stacked bar charts; (4) diverging stacked bar charts; and (5) numerical expression of scores.

Grouped column/bar charts

Responses by questions (or other categories) can be grouped and represented as column or bar charts, with the Likert score represented by each column/bar, as shown in Figure 6.

While bar charts allow easy comparison on the count (frequency) of responses across the different Likert scores, they do not allow the easy comparison of proportions of the responses (e.g., proportion of respondents who strongly agree or agree to a question).

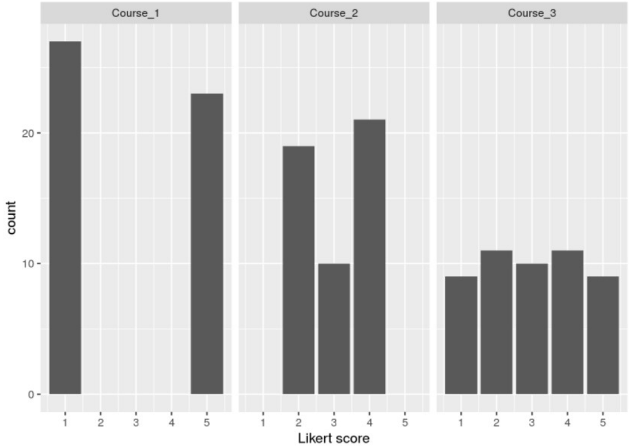


Figure 5: Grouped bar chart representing responses to a Likert scale survey

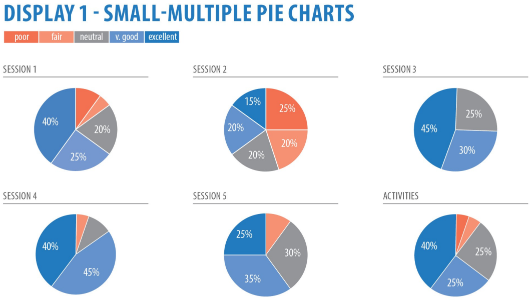


Figure 6: Multiple pie charts representing responses to a Likert scale survey

Pie charts

While pie charts depict the proportions of responses to a question, it is difficult to compare and visualize the differences in proportion between the questions (Figure 7).

Stacked bar charts

Stacked bar charts clearly shows the proportion of the various responses for each question asked in a survey (Figure 8). The use of different tones of colours (i.e. light green to dark green) provides the reader with an idea of increasing levels of responses (i.e. Agree to Strongly Agree).

A benefit of stacked bar charts is that it allows readers to compare different proportions across different bars since each set of stacked bar sums up to 100%. In Figure 8, it is evident that while a large proportion agreed that they liked the presentation, a lower proportion actually felt that they learnt something from the presentation.

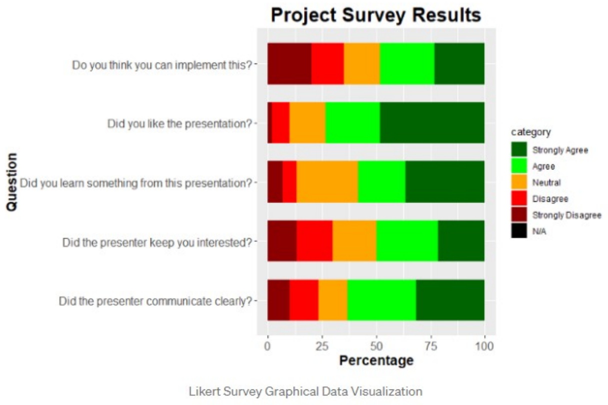


Figure 7: Stacked bar charts representing responses to Likert scale survey

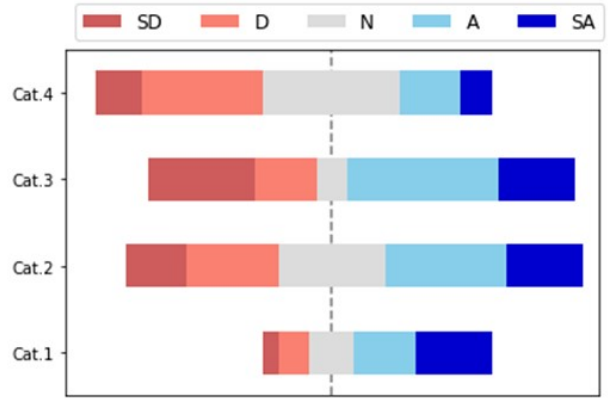


Figure 8: Diverging stacked bar charts representing responses to Likert scale survey

Diverging stacked bar chart

Another way of representing responses to a Likert scale survey is the diverging stacked bar chart (Figure 9). This visualization is similar to the Stacked Bar Chart but differs in that the bars have a common vertical baseline located in the the centre of the diagram. The lengths of the segments of the bar charts are proportional to the number of responses for each value of the Likert scale for each question. Segments which represent favourable responses are usually on the right of the baseline while those that are unfavourable are on the left. Neutral responses are located on the central baseline.

The diverging stacked bar chart is most appropriate to represent responses from Likert scale surveys as it allows easy interpretation and comparison of multiple categories.

Numerical representation of Likert score

While it may seem appropriate to compare the mean Likert score of each category, there is a potential pitfall in doing so. Referring to Figure 10, the mean score for the three courses are similar, yet it is evident that the distribution and spread of the responses are very different.

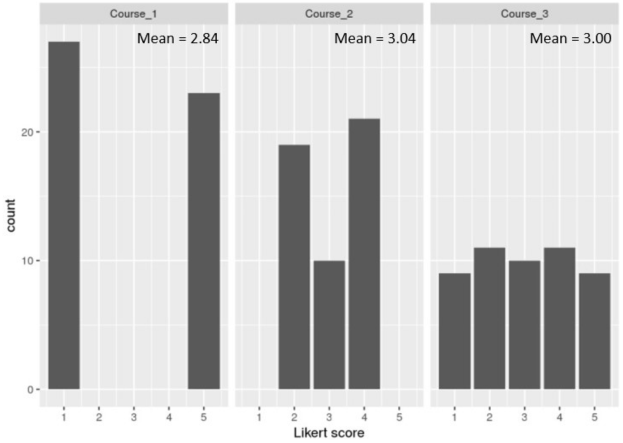


Figure 9: Grouped bar chart with mean score of Likert scale survey

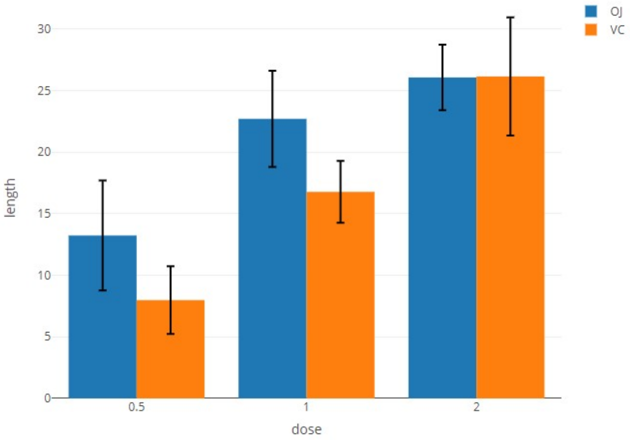


Figure 10: Bar plots with error bars

It would be more appropriate to make comparisons based on the proportion of responses that meet a certain value (e.g. proportion of respondents who gave a 4 or 5 score).

3.3.2 Visualising uncertainty

As surveys are usually conducted on a small sample, there is some degree of uncertainty that the survey results may deviate from the actual viewpoint of the population. Confidence intervals give an indication of that uncertainty and can be represented with error bars (Figure 11).

3.3.3 Visualising correlation

Insights on the relationship between survey responses can be gained from studying the correlation to understand if there are certain determinants or factors that affect the response of certain questions. It would be useful to investigate if vaccination inclination is dependent on certain sociodemographical factors (e.g., age, gender, household size or

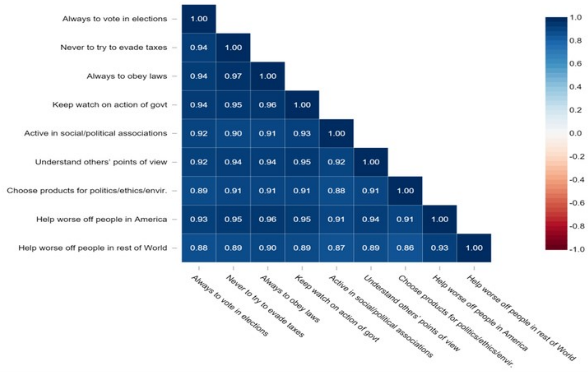


Figure 11: Correlation matrix

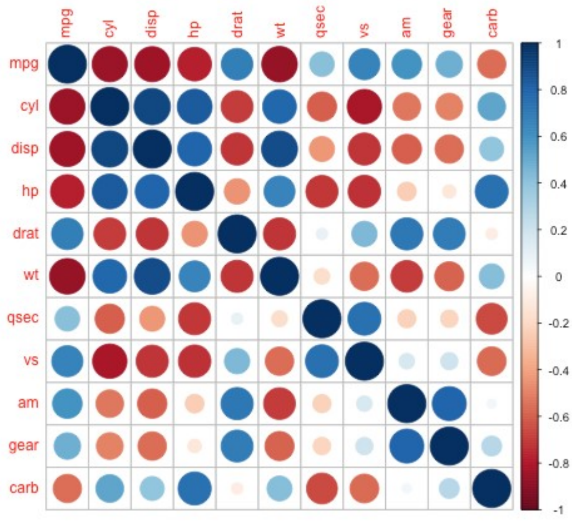


Figure 12: Correlation scatterplot

number of children in the household) or certain attitudes or beliefs (e.g., confidence of vaccine efficacy, or concerns on the side effects of the vaccine).

Correlation matrix or correlation scatterplot are common methods used to depict the correlation between **continuous** variables (Figure 12 and Figure 13). For correlations between **categorical** variables, the UpSet plot can be used via the UpSetR function (Figure 14).

The UpSet plot allows the user to see how frequently each combination or intersection of different factors takes place. Combinations that occur more frequently indicate a stronger correlation between the factors in the combination.

The application will employ the use of the diverging stacked bar chart to visualise the responses from a Likert scale survey, bar plot with error bars to show the uncertainty in survey data and the UpSet plot to understand associations between the categorical variables in the survey.

4. DESIGN FRAMEWORK

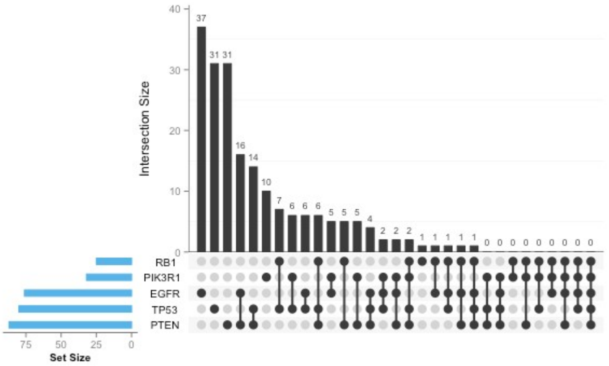


Figure 13: UpSet plot

The purpose of the application is to provide users with an interactive visual experience for in-depth exploration and analysis of the COVID-19 data. With this in mind, the design focus would be centred on *user interaction* and *user experience* with the R Shiny application.

User interaction is not a new concept and has been applied in web-based learning as *learner-content interaction*[6] and product design as *interaction design*[5] for many years. It is concerned with how the user interact with a product or application that meets the users' needs. Interaction design is closely linked to user experience (UX) design, where the design of the product is centred around the experience of the user. In the design of this application, the following areas are considered:-

- Who: the target audience and user of the application
- Why: the motivation behind using the application
- What: functionalities and features of the application
- How: interaction with the functionalities and features

The application is designed for users who wish to find out more about COVID-19 beyond the "standard" reported figures. These users are likely to be curious individuals with an inquisitive mind, and are likely to have some experience and knowledge in statistical analysis. As such, the application needs to be flexible to support exploration of the data while providing clear and easy-to-understand information with statistical metrics.

Data visualisation is the quickest way to communicate information in a clear and easy-to-understand format. Statistical metrics would be presented visually, where appropriate and relevant. To meet the need of flexible data exploration, interactivity would be a key feature of the application, where by the user is able to (1) select and try different combination of variables and parameters to explore the data; and (2) interact with the visualisation to discover trends and insights. As with all applications, the user interface should be kept as simple as possible, and where not possible, to provide information or markers to direct how the user should interact with the application. Finer details of how the design framework is applied to each aspect (new cases, deaths and vaccination receptivity) is discussed in the next few sections.

- take screenshot of main new case tab - briefly explain the interactivity features and visualisation and statistical stuff

4.1 New cases

4.2 Deaths

4.3 Vaccination receptivity

5. DEMONSTRATION

*USE CASE ***

5.1 New cases

5.2 Deaths

5.3 Vaccination receptivity

6. DISCUSSION

7. FUTURE WORK

Due to resource constraints,

8. CONCLUSION

The use of interactive techniques

References

[1] Argawu, A.S. 2020. Regression Models for Predictions of COVID-19 New Cases and New Deaths Based on May/June Data in Ethiopia. Cold Spring Harbor Laboratory Press.

[2] Best, R. and Boice, J. 2021. Where the Latest COVID-19 Models Think We're Headed - And Why They Disagree.

[3] Center for Systems Science and Engineering (CSSE) 2020. COVID-19 Data Repository.

[4] Google 2021. U.S. COVID-19 Public Forecasts.

[5] Interaction Design Foundation 2002. User Experience (UX) Design.

[6] Northrup, P. 2001. A Framework for Designing Interactivity into Web-Based Instruction. Educational Technology Publications, Inc.

[7] Pagano, B. 2021. US - Covid-19 Modeling.

[8] Roser, M. et al. 2020. Coronavirus Pandemic (COVID-19).

[9] Smithson, M. 2020. Data from 45 countries show containing COVID vs saving the economy is a false dichotomy.

[10] University of Melbourne 2021. United States of America Coronavirus 10-day forecast.

[11] Wicklin, R. 2020. Visualize the case fatality rate for COVID-19 in US counties.