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

Figure 1: Comparison of predictive models with option for user to select starting month that forecasts will be based on [2].

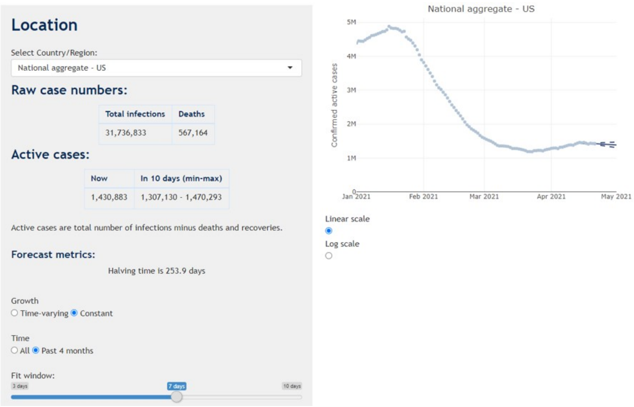

Figure 2: Time-series forecasting with model options, model input parameters and graph scale options [10].

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

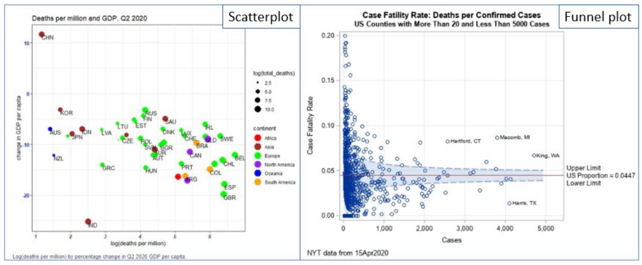

Figure 3: ***(a)*** **Scatterplot on change in GDP per capita against deaths per million [9].** ***(b)*** **Funnel plot on case fatality rate against number of confirmed cases [11].**

such as the date range to be used for the forecasting. Model assessment measures will also be included.

3.2 Deaths and death rates

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 (Figure 3b) 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. It is similar to the scatterplot, and additionally seeks to highlight any anomalies from the expected range of the numerical values based on statistical concepts.

There are very few multivariate analysis done, and of those conducted, most of them are presented in tabular form or described in text. Only one study on regression models[1] presented its findings visually in a scatterplot with fit lines (Figure 4).

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, 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

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%.

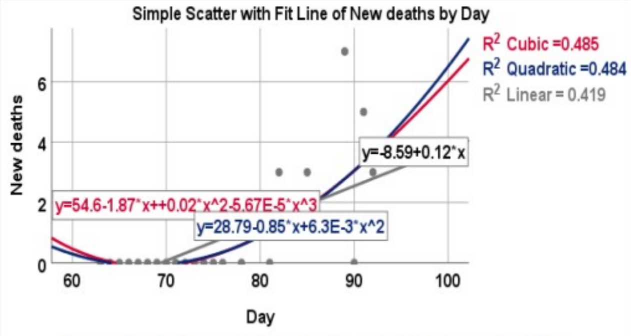

Figure 4: Simple line graph of COVID-19 new deaths by the number of days

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.

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

Likert scales are most commonly presented in pie charts and variations of the bar chart, where proportion of response levels are used to describe the data. Salient points regarding each chart type (Figure 5) are highlighted:-

- Pie charts: depict the proportions of responses to a question, but is difficult to compare and visualize the differences in proportion between the questions
- Grouped column/bar charts: allow easy comparison on frequency of response levels within a question, but is difficult to compare proportions of the response levels across questions
- Stacked bar charts: show proportion of various response levels for each question while allowing sufficient comparison of response levels across questions. Use of colour is important to facilitate user interpretation of the graph.
- Diverging stacked bar charts: similar to stacked bar chart, comparison of response level proportion across questions is made easier by aligning the neutral response along a vertical baseline.

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

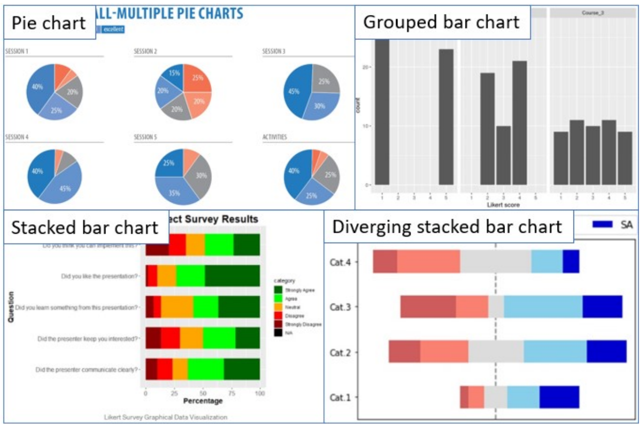

Figure 5: *(From top left)* Pie chart, grouped bar chart, stacked bar chart and diverging stacked bar chart representation of Likert scale survey response levels.

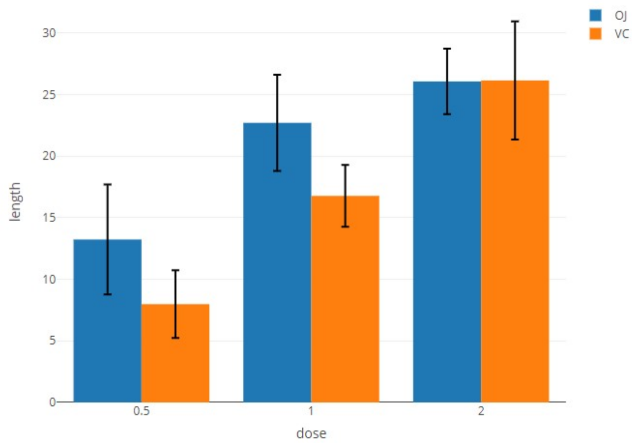

Figure 6: Bar plots with error bars

deviate from the actual viewpoint of the population. Confidence intervals give an indication of that uncertainty and can be represented with error bars (Figure 6).

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 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 7). For correlations between **categorical** variables, the UpSet plot (Figure 8) allows users to see how frequently each combination or intersection of different fac-

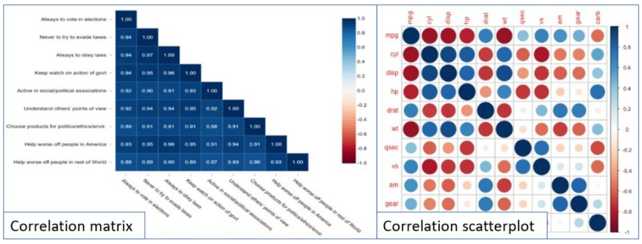

Figure 7: Correlation matrix and correlation scatterplot

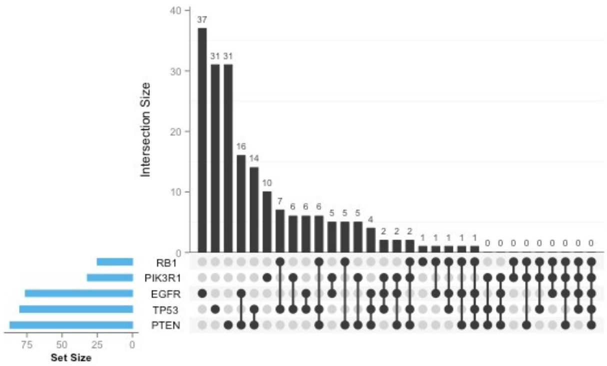

Figure 8: UpSet plot

tors 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

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

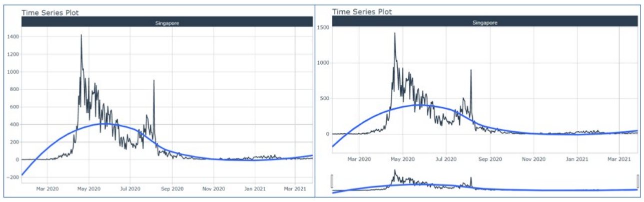

Figure 9: Development of base exploratory time series plot

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 design principle of the application is that the application should 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.

4.1 New cases

In the prediction of new cases, considerations of how the user would proceed with the analysis were included in the system architecture. In time-series forecasting, it is important to understand the characteristics of the time-series data, i.e. trend, cycle, seasonality etc., so as to better interpret and even improve the forecast results. The design also take into account the constantly evolving COVID-19 situation that users are in. New data is available on a daily basis and users would want to analyse the latest data. As such, two components on exploring the dataset (using the `timetk` package) and a function for user to upload the latest dataset are added to the predictive analysis of new confirmed COVID-19 cases.

In the exploratory time-series analysis, the time series plot is straight-forward and is interactive. Figure 9 shows the development of the base plot. The `timetk` package also provided diagnostic plots to study the trend, seasonality and remainder, anomalies and lag if present in the time-series. The UI is simple, whereby users need only to select the country of interest to commence their exploration (Figure #).

The focus of the predictive analysis is to allow users to compare and assess the performance of the predictive models provided. The design caters to two types of users: (1)

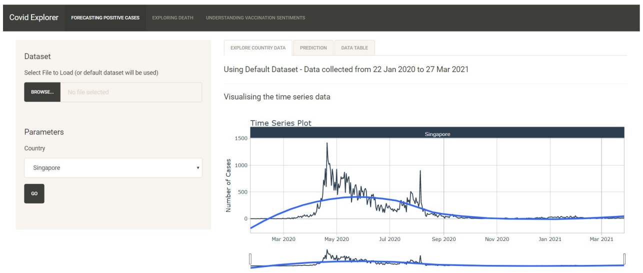

Figure 10: Final UI for exploratory time-series analysis

Tool	Input Parameters
Exploratory time-series	Country
Predictive time-series	Country, forecast date range, forecast horizon
- model comparison	Models
- advance analysis	Selected models
>ETS model	Error, trend, season
>Prophet model	Growth, change-point range, season

Table 1: Parameters for the exploratory time-series and predictive time-series analyses

trained user with basic knowledge in predictive time-series analysis; and (2) trained user with good knowledge and experience in predictive time-series analysis. The first type of user would be more interested to know what are the available models for predictive time-series analysis and to compare these models, whereas the second type of user is better equipped to understand and change more advanced model parameters.

Basic model parameters, such as country, date range in which the forecast is based on, and the forecast horizon, would be provided as a basic feature to explore the predictive time-series analysis. For the first type of users, the emphasis is on the comparison between models, and this comparison would be based on the same input parameters. The functionality to customise the input parameters for selected models would be provided for the more knowledgeable user (Table 1). For both groups of user, it is important to have evidence-based comparison based on established metrics such as the mean absolute error (MAE), root mean square error (RMSE), R-squared (RSQ) etc. The figures ## show the UI of the predictive time-series analysis.

4.2 Deaths and death rates

The bivariate and multivariate analyses employed to explore the relationship between total number of deaths, death rates, health, economic and population structure indicators are the scatterplot (`ggstatsplot` package), funnel plot (`ggplot2` package) and multiple linear regression (`olsrr` package).

The key design principle behind this sub-module is to allow users to interactively explore the relationships of the number of deaths or death rates with the other variables, and to provide users with the flexibility to choose plot or model

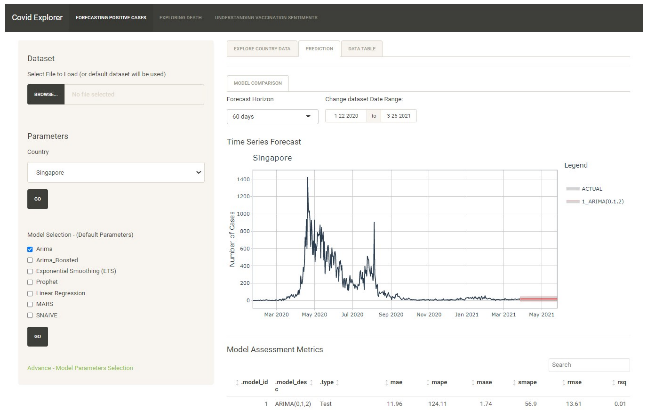

Figure 11: Final UI for predictive time-series analysis

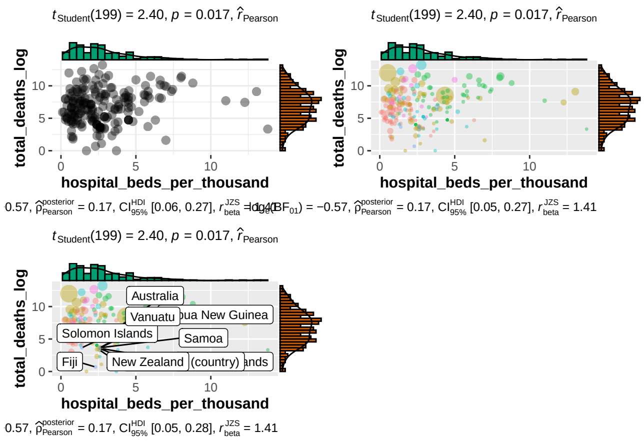

Figure 12: Development of the base scatterplot to include colour and size scales, and data labels

parameters. The scatterplot from the `ggstatsplot` package was chosen for the statistical tests conducted on the variables, to cater for the statistically trained user. For the scatterplot and funnel plot, additional considerations were made as to what information should be displayed in the base plot. Figure 9 shows the development process of the base scatterplot. The parameters for user selection were identified for each analysis (Table 2).

For the Shiny app UI, careful considerations were made not to overwhelm the user. The output is dependent on the type of variables that the user selects i.e. funnel plot is more appropriate for death rate analysis, MLR is more appropriate for exploring relationship with multiple independent variables. The UI is designed based on how the user would conduct the analysis, instead of the type of output. The output is dependent on the variables that the user select, and the parameters are in turn dependent on the output that is to be shown. Due to the multiple layers of conditions, the `conditionalPanel` function is employed in the R Shiny code. The UIs of the three outputs are shown in Figures 10-12.

Analysis	Input Parameters	Output Options
Scatterplot	Statistical test, confidence level used in test, marginal distribution type, regression smooth line	Range of data labels to display
Funnel plot	Confidence interval for the expected range of values	None as able to achieve interactivity
MLR	Regression model, selection methods	Plot model results, plot model diagnostics

Table 2: Parameters for the scatterplot, funnel plot and multiple linear regression (MLR) analyses

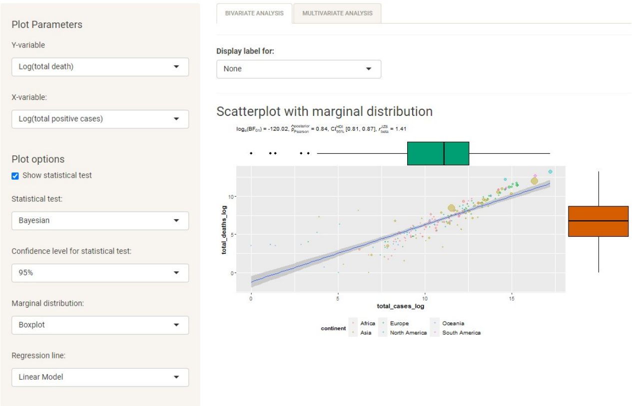

Figure 13: Final UI for scatterplot

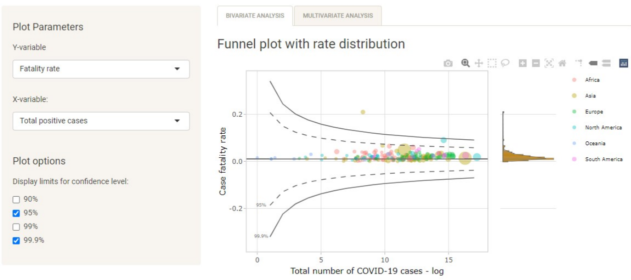

Figure 14: Final UI for funnel plot

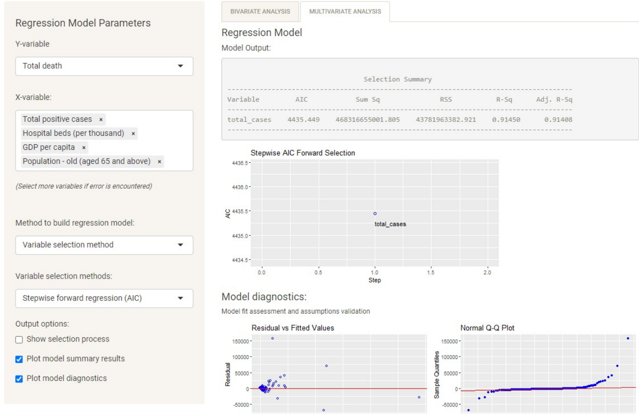

Figure 15: Final UI for multivariate analysis

Exploratory Tool	Input Parameters
Diverging stacked bar chart	Survey question
Bar plot with error bar	Response level Confidence interval for error bar
UpSet plot	Country, response level, survey question
Density plot, mosaic plot	Country, x- and y-variables

Table 3: Parameters for the diverging stacked bar chart, bar plot with error bar, UpSet plot, density plot and mosaic plot analyses

It is possible to collapse the bivariate and multivariate analysis tabs and generate the output solely based on the variables that the user has chosen. It is also possible to build a regression model with only one independent variable. As such, the decision was made to separate the UI by the type of analysis.

4.3 Vaccination receptivity

This sub-module considers how the user would explore a survey data. Users would want to compare proportion of responses across countries for a particular survey question (diverging stacked bar chart and bar plot with error bar) and/or find out if certain profile of respondents (based on reported socio-demographic or perception questions in the survey) have an impact on vaccination receptivity (UpSet plot, density plot and mosaic plot). The parameters to support interactivity and exploration differ for the two analyses (Table 3).

Within the association analysis plots, the UpSet plot is able to support multiple variables, whereas the density plot and mosaic plot are more appropriate for bivariate association. As such, the UI is designed to have three separate tabs for each group of identified analysis.

5. DEMONSTRATION

Figure 16: Final UI for exploratory analysis

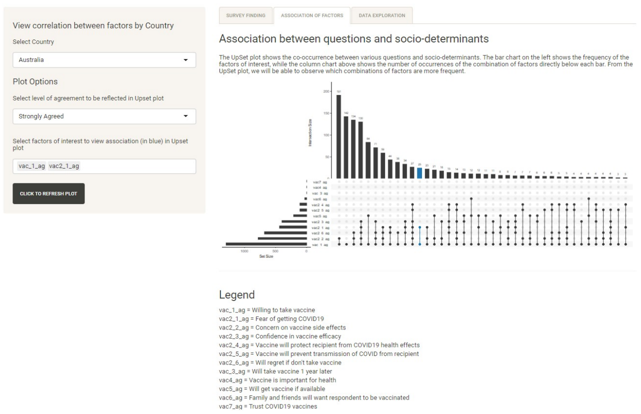

Figure 17: Final UI for association analysis

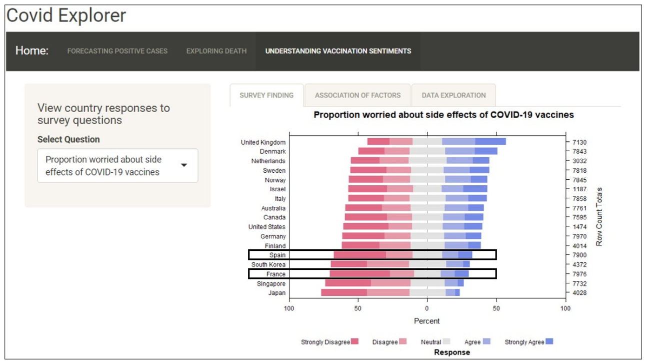

Figure 18: Diverging stacked bar chart showing the proportion of respondents who are worried about side effects of COVID-19 vaccines

Vaccination is believed to play a key role in curbing the spread of COVID-19 and bringing an end to the global pandemic. However, vaccination rates are slow with some countries and it is important to understand the reasons of not wanting to be vaccinated, in order to develop targeted strategies to increase vaccination rates. We will demonstrate the use case of understanding vaccination receptivity with our application.

Using the *diverging stacked bar chart*, we are able to visualize the responses to the different questions and hence get a deeper understanding of the possible association and determinants behind the attitudes and inclination towards vaccination. The question on whether one is worried about the side effects of COVID-19 vaccines is selected.

From Figure ##, it is observed that Spain and France have a relatively high proportion of people who strongly agreed that they were worried about the side effects of the COVID-19 vaccine. Comparing it to their vaccination rates taken from [8], this could explain why Spain and France have was one of the countries that were more inclined to vaccination.

Singapore on the other hand, had most respondents disagreeing that there were worried about getting COVID-19. Closer examination would be needed to understand the factors that could be affecting different countries' inclination towards vaccination.

6. DISCUSSION

7. FUTURE WORK

Due to resource constraints, only three aspects of the COVID-19 data were selected for further exploration in our application. The application could be developed to include other aspects such as COVID-19 testing, hospitalisation and vaccination rates in the future. In the current application, some functionalities were not included and may be considered for future works.

New cases

- Exploratory time-series analysis: explore and compare

trends across different countries
- Predictive time-series analysis: compare two or more models based on different input parameters, and include more models for advance analysis

Deaths and death rates

- Inclusion of more global indicators
- Scatterplot: further customisation of statistical test to be used, customisation of scale encodings
- Funnel plot: customisation of scale encodings
- MLR: include more regression models

Vaccination receptivity

-

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.