A Visual Discovery of Covid-19 Behaviours in Singapore

Alex YEO Chia Guan
Singapore Management University, School of Computing and Information Systems

ONG Chee Hong
Singapore Management University, School of Computing and Information Systems

Stella LOH Yun Jia
Singapore Management University, School of Computing and Information Systems
stella.loh@mitb.smu.edu.sg

ABSTRACT

Understanding how are Singaporeans' behaviours change in light of the Covid-19 pandemic is important in furthering the government's efforts to keep the virus under control. Representative questions is one of the ways of gathering information to this question. However, there is a remains a gap between the information collected by various sources and the information actually being used to understand these changing behaviours. This project aims to develop a R Shiny application to visualize the data collected, so as to perform insightful analysis on the data collected from the YouGov survey. The use of R Shiny allows non-technical users to manipulate the visualizations easily, hence allowing business users to be more proactive in generating such insights. Given that the pandemic situation is constantly evolving, we demonstrate the potential of the application through the use of interactive data visualizations to explore and analyze the fast changing data. The analytics and design choices made in the development of the application, initial findings and potential future work are discussed.

1. INTRODUCTION

With the onset of the Covid-19 pandemic since 2019, nations have been scrambling to gather as much data as possible on this novel virus. In a bid to contain the spread, nations have implemented various measures, with most of these approaches centered on the role of an individual to practice preventive behaviours[1]. These measures range from being mandatory (requirement to wear masks once leaving the house) to those which are highly encouraged (proper handwashing). With the implementation of these approaches, daily behaviours as we know it have slowly evolved.

Understanding how perceptions and behaviours have changed in critical in determining the success of Covid-19 measures. Afterall, it is not enforcement that will determine if measures introduced by the government succeed or fail, but the behaviour of the people. A common approach is to collect data from representative surveys, Most of these survey fieldwork and analysis are performed independently for various studies and tend to be on a smaller scale.

Despite the effort and expense on such research initiatives, the real world practice of using these survey data tends to be confined to the particular study and bulk of the data collected is left untapped for insights. The charts presented in the analysis are typically confined to static charts which provide summarized information to the user. They tend to be highly aggregated with the granular details either not presented or presented in supplementary charts.

We have developed an interactive tool to help users better understand the change in Covid-19 behaviours on a more granular level. With this tool we aim to provide users such as a government Covid-19 taskforce practitioner, with an improved way to gain deeper understanding of the changing behaviours using the most effective visualizations and segmentation techniques.

This paper documents the development effort to design and implement the interactive tool for supporting the analysis and visualization of a government Covid-19 taskforce practitioner. The paper consists of seven sections and starts off with an introduction of the paper. This is followed by an overview of the motivation and objectives of the paper and a review of past works. Section 4 explains the data used and methodology employed, including the analytical methods used to visualize and analyze data. Section 5 provides an overview of the interface design, functionality and findings after applying the application to a practical use case. Section 6 then summarizes some observations obtained from end users. Lastly, Section 7 concludes by highlighting the future direction of the research.

2. MOTIVATION AND OBJECTIVES

This research and development is motivated by the difficulty faced by non-technical users in using interactive tools to visualize the data collected and perform analysis on them. Hence this tool aims to fill this gap so as to allow users such as a government Covid-19 taskforce practitioner to i) maximize the usage of survey data collected, ii) be able to independently generate insights from this data. We have identified that one of the main difficulties these non-technical users face when attempting to generate such insights and visual-

izations independently is that they may not know how to update the codes to generate alternative views or to achieve visualizations of various levels of aggregation. The tool was thus aimed to support the following requirements:

1) Visualize data of different levels of aggregation, with minimal technical knowledge required
2) Visualize trends and patterns among responses
3) Ability to display detailed records on-demand

3. LITERATURE REVIEW OF PREVIOUS WORKS

There has been work performed on the visualization of the same dataset collected. These visualizations are static in nature. Some limitations of static charts are that users are unable to manipulate the view of the charts. Furthermore, unless there are adequate labels included, users may not be able to derive the exact value of the variable simply by viewing the chart.

An example of this is seen in Figure 1 where we are able to see an overall trend of higher proportion of respondents wearing a mask outside of their homes. However, the user is unable to obtain the exact percentage for each period. Additionally, users only get one view from this static chart- an overall aggregated view. They are unable to change the view to view the response by gender, for example. Moreover, there is poor use of color as colors representing "Not at all" and "Rarely" responses are similar and hence when overlapped, makes it hard to read the chart. Lastly, the x-axis labels are not presented horizontally for easy reading.

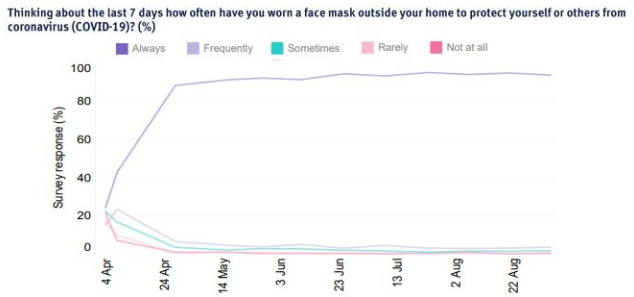

Figure 1: Limitation of static line charts

Similarly, in Figure 2 below, the static chart is presented in a highly aggregated format. Users will require supplementary charts should they wish to view more granular details. Furthermore, the chart has no x-axis labels and hence users may not be able to understand it if they did not read the commentary. Similar to the point made above, given the static nature of the chart, the users are unable to obtain the exact percentage for each period. Lastly, the use of color is unnecessary to comprehend the information represented on the chart and may even distract the user and hence is essentially Chartjunk[2].

There are certain geospatial variables in the survey data collected. This allows information to be mapped onto a

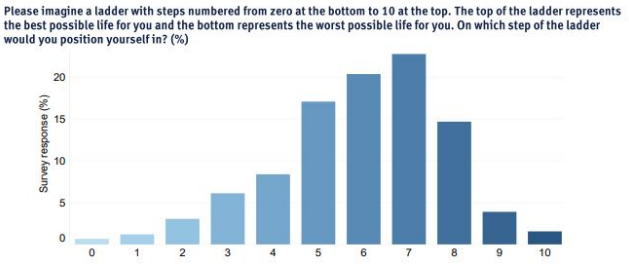

Figure 2: Limitation of static bar charts

choropleth map as seen in Figure 3 below. There are several critics for the chart in Figure 3:

- There is no legend label to specify what the shades of blue are representing. The shades represent a range from 37 to 82 and based on the commentary, the user may be able to derive that the shades represent government stringency score. However, the user will not register what is the range they should be seeing- does the score range from 0 to 100?
- Poor use of colors. The dark blue color over the China area of the map obscures the map label, making it hard for the user to read which country the area is representing. The label should have been amended to a lighter color for easy reference.
- Not all countries are labelled. Hence for some countries for which responses have been collected (hence colored in shades of blue), users are unable to make out which countries those shaded areas are representing.

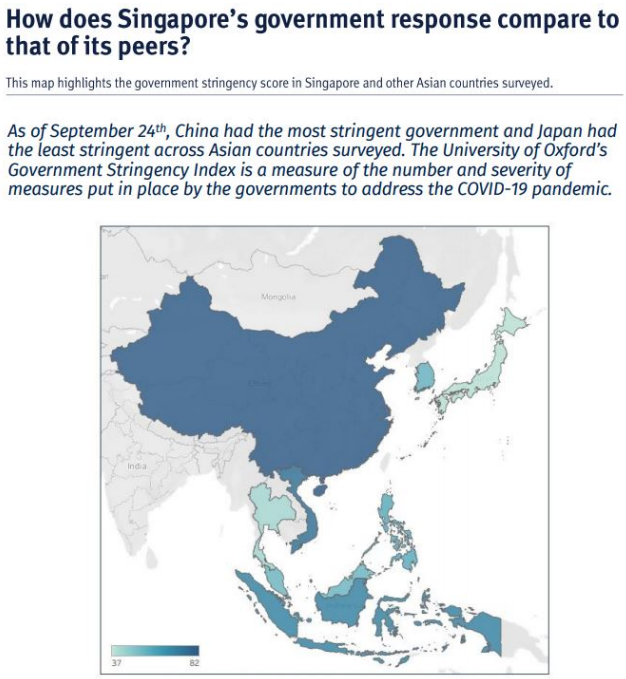

Figure 3: Limitation of static chloropeth map

4. METHODOLOGY

The dataset used in this project is gathered through a representative survey conducted by a partnership between YouGov and Institute of Global Health Innovation (IGHI) at Imperial College London. The questions in the survey are designed to cover on testing, symptoms and ability and willingness to self-isolate if needed. It also looks at behaviours and extent of compliance with the 20 common preventative measures. Contextual information such as gender, age, region, household size, household children, health conditions, employment status and date of survey response have also been collected.

As the dataset consists of different data types (categorical, continuous and geospatial), we have developed the application to allow non-technical users to effectively explore the data set by enabling them to interact with the application to select the view and data to focus on. With the interactive features, users can select the variables to view and the aggregation level to view them at.

4.1 Data Preparation

The raw data collected was provided in CSV format. The following steps were taken to prepare the data:

i) We noted there were certain variables in the dataset for which most responses were blank. We first read the CSV file into R, taking care to include only the first 78 columns as the columns after had minimal responses and thus were dropped.

ii) The region variable had 2 components in the response, with region and town separated by a "-." dplyr was used to separate the variable into two separate columns, named region and town.

iii) dplyr was also used to rename variable names as the variables were originally named based on the question code and a separate codebook was provided for users to reference which codes represented each question. We thus renamed the variable name to something more intuitive so as to save the user from constantly having to reference to the codebook.

4.2 User Interface Design

The interactive application was developed on Shiny, a R package used to build interactive web apps. The Shiny web framework simplifies collecting input values from a web page, and having the results of R code written as output values back out to the web page[3]. Non-technical users can interact with the customisable widgets and update the input values, which in turn changes the output values which are reflected immediately.

Keeping the needs of business users in mind, we have designed the application as such:

i) EDA tab - Exploratory Data Analysis (EDA) allows business users to explore the data using different variables and aggregation levels, keeping in mind visualization best practices.

ii) Cluster Analysis tab - Segmentation allows business users to identify patterns in the data, which can guide future actions such as implementing targeted policies to influence behaviours based on the cluster profile.

4.3 Analytical Techniques

4.3.1 Bar Chart

Bar charts present categorical data using rectangular bars with lengths proportional to the values that they represent. They are effective in helping users compare individual values to one another[4]. To help users visualize the variability of data, we have also added an option to include 95% confidence error bars into the visualization.

The ggplot2 package was used to design the bar charts. ggplot2 is a system for creating graphics, based on The Grammar of Graphics.

4.3.2 Line Chart

Line charts display information as a series of data points connected by straight line segments. They are effective in their ability to show trends and patterns of change[4].

The ggplot2 package was used to design the line charts.

4.3.3 Parallel Sets

Parallel Sets is a visualization method which allows interactive exploration of categorical data that shows data frequencies instead of the individual data points. The method is based on the axis layout of parallel coordinates, with boxes representing the categories and parallelograms between the axes showing the relations between categories[5]. In the interactive parallel set within the application, the user can visualize the data up to five levels of categorization, which allows them to explore the data using different variables and levels. The tooltip function also gives users information on the number of counts and proportion fulfilling each flow.

The parset package was used to design the parallel sets. Using the infrastructure provided by htmlwidgets, parset allows easy integration of parallel sets into R workflow.

4.3.4 Choropleth Map

Choropleth maps are used to represent statistical data through various shading patterns on predetermined geographic areas. They are effective at utilizing data to represent variability of the desired measurement, across a region[6]. With the use of choropleth maps, the user can explore patterns along geographic lines. In the interactive choropleth map within the application, the user can select the type of data classes and classification method and observe how their choices changes the map results. This allows users to select the option that best suits their needs.

The tmap package was used to design the choropleth maps. It offers a flexible, layer-based approach to create thematic maps and is based on the grammar of graphics[7].

4.3.5 Latent Class Analysis (LCA)

Latent Class Analysis (LCA) is a statistical method for identifying unobserved class membership among subjects using

categorical and/or continuous observed variables. LCA was selected as the dataset contained both categorical and continuous variables. Furthermore, LCA can be used with data with non-normal distributions that show heteroskedasticity or have heterogeneity of variance[8]. Using LCA, business users will be able to identify clusters based on their demographic and health information. This allows for the design and implementation of targeted initiatives.

Traditionally, model fit indices such as Akaike information Criteria (AIC) or Bayesian information criteria (BIC) have been used to identify the ideal number of latent classes. However, there is no consensus as to which model fit index identifies the best model. In the interactive LCA model within the application, the user can select their desired number of latent classes.

The mclust package was used to design the LCA models. It is a R package for model-based clustering, classification, and density estimation based on finite normal mixture modelling.

5. DESCRIPTION OF PRODUCT AND FINDINGS

The data used for analysis contained more than 30,000 responses collected over a period from April 2020 to July 2021. Date and timestamp of the responses are collected, which allows us to add the time dimension to visualizations to understand how behaviours may be changing over time.

To achieve reliable results, we have prepared the data before performing analysis. We noted several variables in the i14_health section had missing responses, which made up about 10% of the population. We have also created a new variable named "Pre-existing conditions" which specifies if the respondent has any existing health conditions and is derived from the i14_health section responses. For those missing responses, they were categorized as "Not Specified." Redundant data were removed to avoid bulky data and to improve the performance of the analysis and exploration. We did not perform any aggregation of data so thay visualizations can be created at a granular level.

Our assessment shows that the application provides users with an easy to use interface which can potentially provide a wide range of insights, which are illustrated below.

The parallel sets allows users to hover over each flow to observe the percentage of respondents within the flow category. Hence users can obtain insights on the distribution of demographic factors and how it varies with the behaviours observed.

From Figure XX below, we observe that there are almost an equal proportion of each gender in the respondents. A large proportion (90%) of respondents always wear masks when outside. Conversely, fewer respondents (72%) always avoid large gatherings and always avoid crowds (45%). We also observe that a large proportion of respondents with pre-existing conditions tend to practice these preventative measures frequently, if not always.

Figure 4: Interactive Parallel Sets

6. DISCUSSION

Based on a small scale user-study conducted, we have compiled the following feedback from potential end users:

i) Users found the ability to interact with the visualizations useful. Users can easily use the drop-down and radio buttons to make their selections rather than drag and drop like in Tableau, making the application more intuitive to use for beginners.

ii) User guide provided made it easier for non-technical users to use the application.

iii) Organization of the application into EDA and Clustering Analysis tabs, and grouping the relevant visualizations into each tab allowed users to navigate the application easily, as opposed to other visualization software like Tableau where these visualizations tend to be organized in sheets.

iv) An area of improvement is to include an option for the user to compare one chart view against another so as to make comparisons easier.

7. CONCLUSION

The demonstration of the application's potential using the above case study highlights its ability to enable users to explore and analyze the survey data collected, so as to gain valuable insights on the changing behaviours and perceptions, which can be used to drive future policy decisions and initiatives.

We have identified several areas for further development of the application:

i) Given the rapidly evolving pandemic landscape, data quickly becomes irrelevant. Hence, data needs to be constantly refreshed and updated for newer responses. The stored survey data can be replaced with real-time updated survey responses collected for the most up to date analysis and exploration to stay ahead of the curve.

ii) The scope of the application can also be extended to cover a larger geographic area to incorporate comparison across multiple countries.

iii) Additional interactive features can be added to the application, including the option for the user to compare

one chart view against another so as to make comparisons easier, as the human mind may not be able to retain earlier views in memory.

8. ACKNOWLEDGMENTS

The authors thank Associate Professor, KAM Tin Seong, Singapore Management University for his support and guidance.

9. REFERENCES

- [1] Tan THY, Toh MPHS, Vasoo S, et al. Coronavirus Disease 2019 (COVID-19): The Singapore Experience. A Review of the First Eight Months. Ann Acad Med Singap 2020;49:764-78.
- [2] Edward Tufte, The Visual Display of Quantitative Information (1983, 2001), 106-121
- [3] Shiny - The basic parts of a Shiny app. (2021). Retrieved 11 August 2021, from https://shiny.rstudio.com/articles/basics.html
- [4] Few, S. (2006). Retrieved 11 August 2021, from https://www.perceptualedge.com/articles/b-eye/encoding_values_in_graph.pdf
- [5] Kosara, R., Bendix, F., & Hauser, H. (2006). Parallel Sets: interactive exploration and visual analysis of categorical data. IEEE Transactions On Visualization And Computer Graphics, 12(4), 558-568. doi: 10.1109/tvcg.2006.76
- [6] DeLorenzo, N., & Dugger, A. (2021). Choropleth Map. Retrieved 11 August 2021, from https://www.arcgis.com/apps/MapJournal/index.html?appid=75eff041036d40cf8e70df99641004ca#:~:text=Choropleth%20maps%20are%20po
- [7] Hahn, N. (2021). 2 tmap | Making Maps with R. Retrieved 11 August 2021, from https://bookdown.org/nicohahn/making_maps_with_r5/docs/tmap.html
- [8] Logan, Jessica & Pentimonti, Jill. (2016). Introduction to Latent Class Analysis for Reading Fluency Research. 10.1007/978-1-4939-2803-3_11