A Visual Discovery of Singaporeans' Media Habits

Cherie Wong Pei Qi, Denny Dunsford, Vinit Nair

Abstract**—** What are the media usage habits of Singaporeans is a question that is important in furthering public communications goals of being increasingly data-led and targeted in its efforts. Representative surveys are a common means of generating answers to this question. However, a substantial amount of information from such surveys remain untapped. Using a range of data visualisation techniques, this project aimed to develop an R Shiny application to maximise the insights obtained from media consumption survey data, and encourage communications practitioners to be more self-reliant in generating insights to their business questions. The analytics and design choices made in the development of the application and initial findings and future work are discussed.

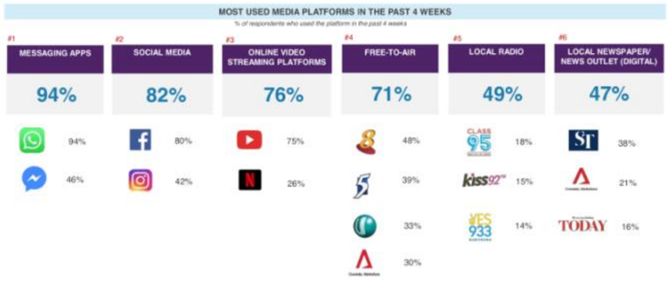
Keywords**—** Media Consumption, Government Communications, Analysis of Survey Data, Interactive Data Visualisation, Audience Segmentation, R Shiny

1 Introduction

Government communications departments in many countries are transforming to be more data-led.[1] This shift has taken place amid higher scrutiny on public spending, and challenges in connecting to audiences across a diffuse range of platforms. This trend is particularly relevant in the Singapore context, with high internet and smartphone penetration rates that enable access to an endless range of online information sources.

Understanding the media consumption habits of audiences is crucial in implementing targeted comms. In addition to data released by traditional media owners (e.g. TV stations, telecommunications operators), a common approach has been to collect data from representative surveys. The survey fieldwork and analysis has often been outsourced to market research firms such as Nielsen, given the heavy resources required especially for survey fieldwork.

Despite the high amount of spending on such research initiatives, the real-world experience of using survey data to understand audiences falls short of expectation, and the bulk of the data is usually left untapped for insights. The market research firms often focus on delivering voluminous reports featuring static charts, and summary tables to the Government client. There are also concerns that the outputs do not adhere to analytics best practices such as helping the reader appreciate the uncertainty in the sample data beyond over-simplified approaches such as suppressing data with a small sample size.



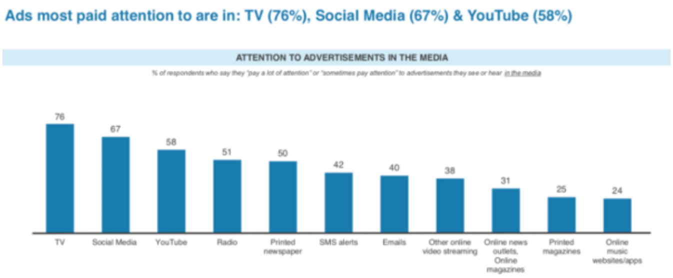


Fig 1. Example of static data visualisations from market research reports. These charts are usually presented in highly aggregated formats, with granular details embedded in supplementary tables. They follow the most conventional analysis techniques (e.g. bar charts and line charts) but may not adhere to best practices such as depicting the uncertainty in the data.

To help government communicators bridge this gap, we present an interactive tool to visually discover media consumption survey data. With this web application, we aim to encourage deeper understanding of Singaporeans' media habits and highlight the different analysis techniques that can help practitioners answer their business questions.

This paper documents our approach to design and develop the interactive application targeted at the government communications practitioner. This introduction is followed in Section 2 by an explanation of our motivation and objectives. Section 3 details the data used and methodology selected. Section 4 provides a visual overview of the final product. Section 5 summarises the findings from a limited useability test. Section 6 concludes the report and offers ideas for further development.

2 Motivation and Objectives

This project was motivated by a desire to i) help government communications practitioners maximise the value of the survey data they collect via market research firms, and ii) encourage them to be self-reliant in generating meaningful information and insights. To achieve these aims, the interactive tool was developed to addresses the following requirements:

- What are the results overall and when I drill down into different profiles? – Visualise the data at different levels of aggregation.
- *What is the significance of the survey findings?* – Visualise the statistical differences between responses.

[1] Macnamara, J. (2017). Evaluating public communication: Exploring new models, standards, and best practice. Routledge.

- What are the hidden patterns in the data that can help guide action? – Visualise the similarities among responses.

3 Methodology

The dataset used in this project was a face-to-face street intercept survey of 1,047 Singapore citizens and permanent residents conducted in late-2018. The survey collected data in three areas: i) which media platforms were most commonly used; ii) how frequently were these media platforms used; iii) how much attention was paid to advertising on these media platforms. In addition, extensive demographic information of each respondent was collected.

To enable lay practitioners to effectively explore the dataset involving varied data types – categorical and continuous, geospatial – we developed a flexible tool that adhered to best practices in data visualisation. Notably, there was alignment with the taxonomy of interactive visual analysis[2] (e.g. allowing the user to choose the data and view specification to focus on relevant items; facilitate sharing of the views for discussion and collaboration) and the importance of carefully characterising the uncertainty and variability in the data.[3]

An explanation of the different methods used follows.

3.1 Data

The raw data (prepared by the market research vendor) was provided in wide excel format. First stage preparation of the data was conducted in R using base R and the **dplyr** package to narrow the scope of the data by removing responses to lower-priority survey questions. The demographic variables were grouped logically based on common definitions used by the practitioners (e.g. occupational profiles are viewed in binary terms – PMETs vs Rank-and-File workers). Data on the residential location of respondents was collected as a free text field and required extensive cleaning to align with official URA Master Plan records.[4] This common datafile (in .csv format) was further prepared according to the needs of the different analysis techniques used.

3.2 Shiny Architecture

Development of the interactive tool was done on Shiny, an R package used to build interactive web apps. Shiny is adopted in the data analytics industry because its framework makes it easy to collect input values from a web page, with R code written as output values back to the web page.[5] Input values can be changed by the user at any time, through interaction with customisable widgets. The output values react to changes in inputs, with the resulting outputs being reflected immediately.

The design of the Shiny web app flowed from the key business questions of the data:

i) First tab – Exploratory Data Analysis (EDA) to form a quick understanding of the media consumption results at various levels of detail (e.g. location, socio-economic status).

ii) Second tab – Inferential analysis to help the user understand the statistical significance of the findings.

iii) Third tab – Audience Segmentation to help the user explore if there are hidden patterns in the data that can guide further action, such as customising communications to similar profiles of Singaporeans.

To facilitate exploration by users without any baseline proficiency in data analytics, we included a user guide to accompany each analysis technique featured in the app.

3.3 Analysis Techniques

3.3.1 Bar Chart

A bar in a bar graph encodes a single value by varying its length. We selected this technique for use in instances where we wish to help the audience focus on comparisons of individual values (i.e. comparison of the proportion of responses of a certain category), of which bars are considered to be a best practice visualization technique.[6] To illustrate the ranges of uncertainty in the data, an option to view the error bars which represent the 95% confidence interval was included as a feature.

The **ggplot2** package in R was used to create the bar chart used in the app. ggplot2 is a flexible data visualization package that is consistent with *grammar of graphics* principles whereby plot building blocks (e.g. data, aesthetic mapping, geometric object) are combined to create a graphical display.

3.3.2 Choropleth Map

Choropleth maps are commonly used when data is attached to enumeration units (e.g. regions), standardized to show rates or ratios, and with a continuous statistical surface.[7] The choropleth map is deployed in this app to make it easier for the user to make sense of the usage of media along geographic lines. The choropleth map is built on the number of data classes and choice of classification method. There is no single best choice for this, as it largely depends on the user's specific goals of the analysis (e.g. if the user is looking for a highly generalized view, they may opt for a fewer data classes or an unclassed choropleth). Thus, the app is designed to allow the user to select the options that suits their purpose.

The **tmap** and **leaflet** packages in R were used to create the interactive choropleth map. The tmap package offers a flexible, layer-based approach to create static thematic maps that is based on the *grammar of graphics*. The output object of tmap can also be saved as an interactive webmap using the leaflet package.

3.3.3 Parallel Sets

Parallel Sets is a visualization method for exploration of categorical data, focusing on the data frequencies instead of individual data points. The technique is built on the axis layout of parallel coordinates, with boxes representing the categories of data (e.g. Media Type, Age group) and parallelograms between the axes showing the relations between categories.[8] Our app enables the user to interactively remap the data to up to five levels of categorization, which creates a wider format of data exploration than usually possible.

[2] Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. Communications of the ACM, 55(4), 45–54.

[3] Koomey, J. G. (2006). Best practices for understanding quantitative data.

[4] https://data.gov.sg/dataset/master-plan-2014-planning-area-boundary-web?resource_id=2ab23cb2-b1a4-4b1a-a9e1-b9cad0ac159b

[5] https://shiny.rstudio.com/articles/basics.html

[6] Few, S. (2006). Data visualization: Rules for encoding values in graph.

[7] https://www.axismaps.com/guide/univariate/choropleth/

[8] R. Kosara, F. Bendix, and H. Hauser, "Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data," IEEE Transactions on Visualization and Computer Graphics, vol. 12, no. 4, pp. 558-568, 2006.

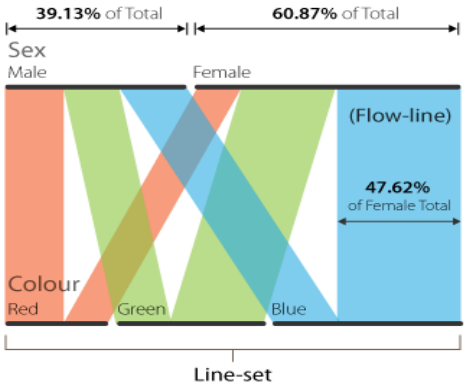


Fig 2. Example of a Parallel Sets plot[9].

The **parsetR** package in R was used to create the interactive Parallel Sets chart in the app. parsetR was developed to allow to integrate parallel sets into existing R workflows.

3.3.4 Upset Plot

UpSet plots offer a convenient way to visualize set data by frequency. UpSet visualizes set intersections in a matrix layout to enable interpretation of quantitative data such as the number of elements in the aggregate level and intersections between sets.[10]

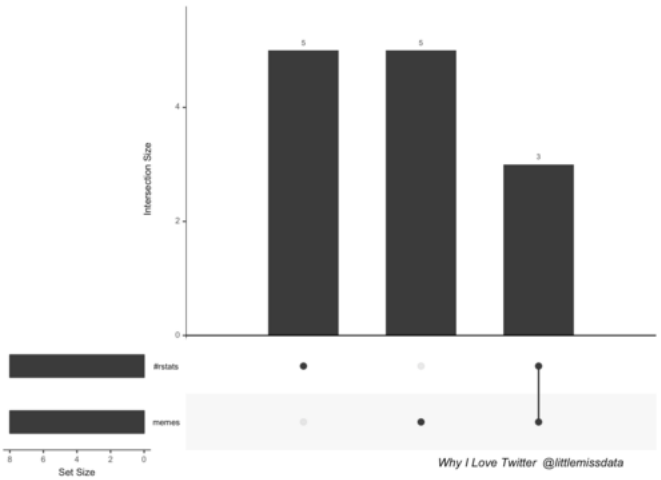


Fig 3. Reading an UpSet plot: The bottom left hand side horizontal bar chart shows the entire size of each set. The vertical bar chart on the upper right hand side shows the sizes of isolated set participation.[11]

The **UpSetR** package was used to build the Upset plots in the app. This package is implemented using ggplot2 and provides support to multiple input data formats, and visualization of attributes to enable users to explore the intersections between the sets. To facilitate exploration, the user is given the option to select the number of sets, select the method of ordering of intersections and filter the plot to a specific age group of interest.

3.3.5 Analysis of Variance (ANOVA)

One-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent groups. The research interest from this dataset is whether mean media consumption levels differs across age groups for any given media platform (note – each respondent provided an estimated percentage of time they spend on each media platform). The app enables the user to select their desired media platform and test statistic (i.e. non-parametric or Bayes Factor) to show in the visual display. In addition, the univariate distributions for each age group are displayed in the panel, with user input on the number of bins.

The **ggstatsplot** package, an extension of ggplot2, was used to create the visuals for the ANOVA tests. ggstatsplot supports the most common types of statistical tests used by analysts, with the advantage of providing details of the statistical tests directly within the plots themselves.

3.3.6 Heat Map

Heatmaps visualize data through variations in coloring. They are useful for cross-examining multivariate data, through placing variables in the columns and observation in rows and gradient coloring the cells within the table, with the darker shade depicting a higher proportion and vice versa.[12] Heatmaps are good for showing variance across multiple variables and the "heatmaply" package allows us to apply hierarchical clustering across both, the rows and columns.

heatmaply is an R package for building interactive cluster heatmap, based on ggplot2 and plotly.js engine. It produces similar heatmaps as d3heatmap, with the advantage of speed (plotly.js is able to handle larger size matrix), and the ability to zoom from the dendrogram. heatmaply uses the seriation package to find an optimal ordering of rows and columns. Optimal means to optimize the Hamiltonian path length that is restricted by the dendrogram structure. This, in other words, means to rotate the branches so that the sum of distances between each adjacent leaf (label) will be minimized. heatmaply supports a variety of hierarchical clustering algorithms (e.g. Ward, Mcquitty).

3.3.7 Latent Class Analysis (LCA)

Identification of groups based on their media habits is a natural business question because of its relevance to the targeting of public communications. LCA is a statistical method used to group individuals into classes of an unobserved (i.e. latent) variable on the basis of their responses to a set of nominal, ordinal or continuous observed variables.[13] LCA was selected as the appropriate analytics technique because it can be used with data with non-normal distributions that show heteroskedasticity or have heterogeneity of variance.[14]

In choosing the optimal number of latent classes to retain, there is no universal agreement among researchers on what is the best criteria. Often researchers rely on Bayesian information criterion (BIC) or Akaike information criterion (AIC) together with a qualitative assessment of the interpretability of the model in deciding on the optimal model. Thus, we have enabled the user to select the

[9] https://datavizcatalogue.com/methods/parallel_sets.html

[10] Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics*, *20*(12), 1983-1992.

[11] https://www.littlemissdata.com/blog/set-analysis

[12] https://en.wikipedia.org/wiki/Heat_map

[13] Porcu, M., & Giambona, F. (2017). Introduction to latent class analysis with applications. *The Journal of Early Adolescence*, *37*(1), 129-158.

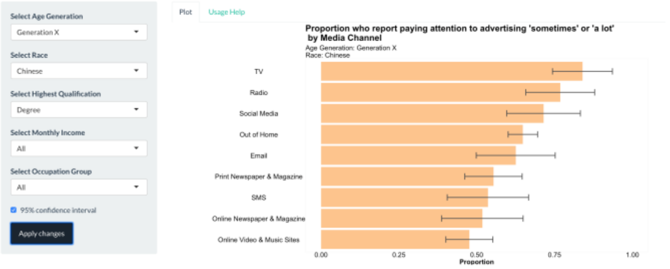
[14] Logan, J. A., & Pentimonti, J. M. (2016). Introduction to latent class analysis for reading fluency research. In *The Fluency Construct* (pp. 309-332). Springer, New York, NY.

desired number of latent classes in the app. To aid the user in understanding the conceptual clarity of the selected classes, we have also visualized the probability estimates for the different media groups across each of the latent classes.

The **poLCA** R package was used for the latent class analysis modeling in the app. The package generates the class probabilities, conditional response probabilities and relevant fit statistics necessary for analysis. ggplot2 was used to improve the visual clarity of the default plots produced in poLCA.

4 DESCRIPTION OF PRODUCT AND FINDINGS

The final interactive plots in the published web application, and an illustration of the potentially wide range of insights that can be gleaned from the application are briefly described below.

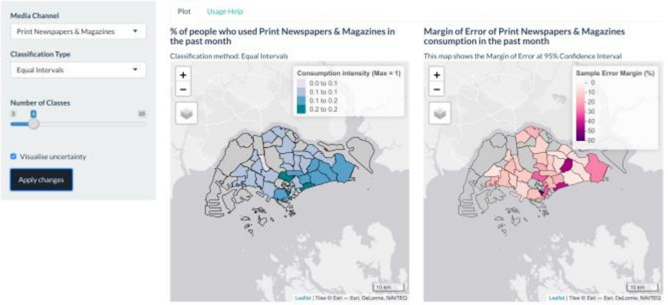


Select Age Generation: Generation X
Select Race: Chinese
Select Highest Qualification: Degree
Select Monthly Income: All
Select Occupation Group: All
☑ 95% confidence interval

Apply changes
Plot
Usage Help
Proportion who report paying attention to advertising 'sometimes' or 'a lot' by Media Channel
Age Generation: Generation X
Race: Chinese
TV
Radio
Social Media
Out of Home
Email
Print Newspaper & Magazine
SMS
Online Newspaper & Magazine
Online Video & Music Sites
0.00
0.25
0.50
0.75
1.00
Proportion

Fig 4. Interactive Bar Chart

At the highest level of aggregation, it is clear that TV and Social Media are where the largest percentage of the population pay attention to advertising, with relatively low attention paid to newspaper advertising. This high level view suggests a higher allocation of advertising spend on TV and Social Media. However, the app easily reveals the visible variations within population groups that will need to be considered depending on the comms goals (e.g. suprisingly high attention to advertising on out-of-home platforms such as bus stops among Millennials).

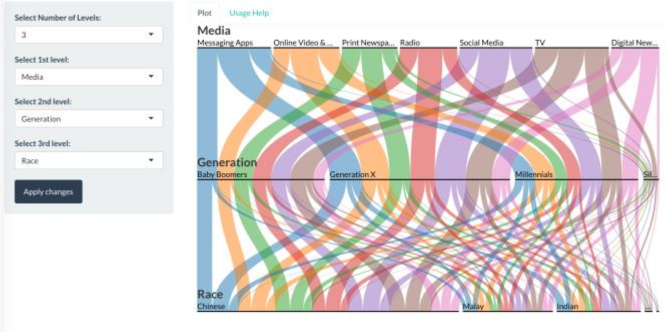


Media Channel: Print Newspapers & Magazines
Classification Type: Equal Intervals
Number of Classes: 3 (range 3–10)
☑ Visualise uncertainty

Apply changes
Plot
Usage Help
% of people who used Print Newspapers & Magazines in the past month
Classification method: Equal Intervals
Consumption intensity (Max = 1)
0.0 to 0.1
0.1 to 0.1
0.1 to 0.2
0.2 to 0.2
Margin of Error of Print Newspapers & Magazines consumption in the past month
This map shows the Margin of Error at 95% Confidence Interval
Sample Error Margin (%)
0
10
20
30
40
50
60
10 km
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

Fig 5. Interactive Choropleth map

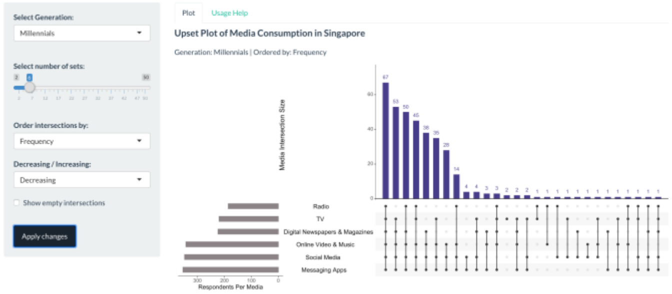
At a glance the user is able to get a sense of whether there is a geographic bias towards usage of certain media platforms. For example, the usage of print newspapers & magazines in Novena appears relatively high. However, because the sample margin of error is also depicted on the chart (in the case of Novena it is +/- 30%), the user can quickly note whether they are looking at signal or noise.



Select Number of Levels: 3
Select 1st level: Media
Select 2nd level: Generation
Select 3rd level: Race

Fig 6. Interactive Parallel Sets

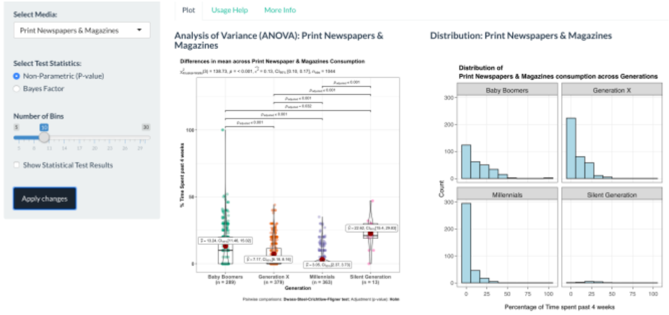
Hovering over the parallel set lines allows the user to gather information about the number of respondents within a flow category. and with one view the user can understand distribution of income and how it varies between generations and the way they consume media.



Select Generation: Millennials
Select number of sets: 6 (range 2–50)
Order intersections by: Frequency
Decreasing / Increasing: Decreasing
☐ Show empty intersections

Fig 7. Interactive Upset Plot

Hovering over the parallel set lines allows the user to gather information about the number of respondents within a flow category and with one view the user can understand distribution of income and how it varies between generations

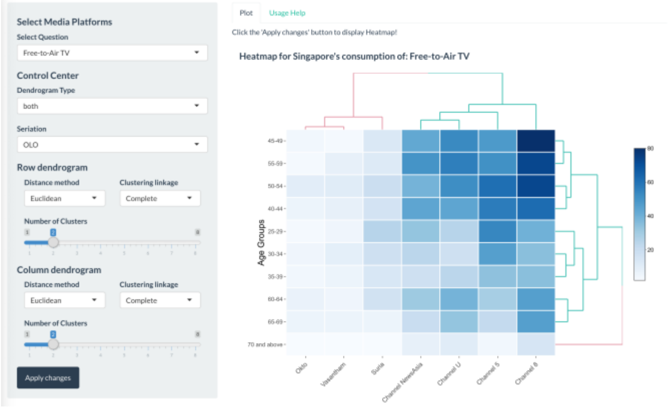


Select Media: Print Newspapers & Magazines
Select Test Statistics: ● Non-Parametric (P-value) ○ Bayes Factor
Number of Bins: 10 (range 5–30)
☐ Show Statistical Test Results

Apply changes
Plot
Usage Help
More Info
Analysis of Variance (ANOVA): Print Newspapers & Magazines
Differences in mean across Print Newspaper & Magazines Consumption
% Time Spent past 4 weeks
0
50
100
Baby Boomers (n = 289)
Generation X (n = 379)
Millennials (n = 363)
Silent Generation (n = 13)
Generation
Distribution: Print Newspapers & Magazines
Distribution of Print Newspapers & Magazines consumption across Generations
Baby Boomers
Generation X
Millennials
Silent Generation
Count
0
100
200
300
0
25
50
75
100
Percentage of Time spent past 4 weeks

Fig 8. Interactive ANOVA

The user is able to confirm that there are statistically significant differences in mean usage numbers across generational groups for most media platforms. As the distribution charts reveal skewedness in the data, the user can get a better idea about using a non-parametric test in this case.



Select Media Platforms

Select Question: Free-to-Air TV

Control Center

Dendrogram Type: both

Seriation: OLO

Row dendrogram

Distance method: Euclidean

Clustering linkage: Complete

Number of Clusters: 2

Column dendrogram

Distance method: Euclidean

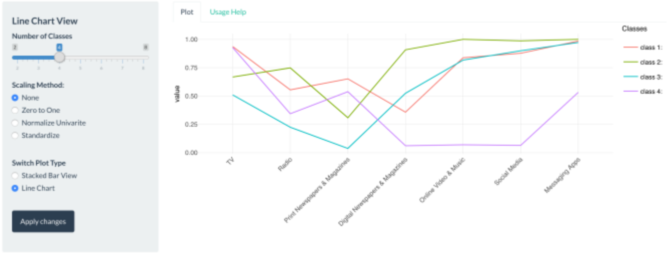
Clustering linkage: Complete

Number of Clusters: 2

Click the 'Apply changes' button to display Heatmap!

Fig 9. Interactive Heatmap

The heatmap provides a more granular view of media usage at the channel level (e.g. Channel 5 in the TV category). For example, the user can identify that among TV viewers, Channel 8 is relatively highly consumed among those in their 40s and 50s. The dendrogram provides insight into the relative distance between other data points.



Line Chart View

Number of Classes: 4

Scaling Method:
● None
○ Zero to One
○ Normalize Univarite
○ Standardize

Switch Plot Type
○ Stacked Bar View
● Line Chart

Fig 10. Interactive LCA

Exploration of the LCA results for different number of classes reveals that a three or four class model provides sufficient clarity and distinction between the groups. Models with five or more classes, are less interpretable.

5 LIMITED USEABILITY TEST

An optimal approach to the development of a user-centered visualisation application would incorporate awareness-building sessions with users up front to stimulate ideas and widen their eyes on the potential application to be developed. This would improve the quality of their feedback. The entire process would be supported through a process of patchwork prototyping.[15]

The development of the first iteration of this application relied on the domain knowledge of group members. However, two end-users from MOM were invited to test the prototype and provide feedback via email, with one response received by our submission deadline. The key points from the written feedback is reflected below:

i) Found the Shiny application to be more versatile than Tableau dashboards. This is because the ability to drill down in the views was comparable to Tableau, but with the advantage of pairing with statistical tests and advanced analytical techniques (LCA, hierarchical clustering) within the same dashboard.

ii) Organisation of different visualisations under the tabs at the top of the application was easier to navigate than individual slides or sheets on Tableau.

iii) The 'usage help' that accompanied each visualisation was useful guidance and helped to keep the main visualisation clean.

iv) Unable to host and access web-based applications on Singapore Government-issued laptops, which can limit adoption of this application in the public service context.

v) Concerns about access to the Shiny application, if sensitive datasets are used.

6 CONCLUSION AND FURTHER IDEAS

This paper set out the development of a web application targeted at government communications professionals to encourage deeper understanding of a survey dataset on Singaporeans' media habits. This was motivated by gaps in the current level of analysis of such datasets in the public service, and a desire to provide a tool for lay users to generate meaningful answers to their business questions.

The application was developed using the Shiny architecture on R, supported with a range of statistical packages to provide users with a suite of techniques to derive insights from the data. Given the emphasis on visual analytics in this project, the various components of the application were developed with due consideration of the taxonomy of interactive visualisations (i.e. Data and View Specification, View Manipulation, Process and Provenance).

Initial user feedback gave the team confidence that the Shiny application can provide the right balance between usability and depth of analysis compared to erstwhile approaches such as market research firms' reports featuring static visualisations, and interactive dashboards on Tableau.

We suggest that further development of the app address the following ideas:

i) Alternate app hosting solutions, such as a personal cloud server. While Shiny apps are easily hosted on shinyapps.io there are perennial concerns among public sector about access to data that may not be publicly available – this includes surveys commissioned to market research companies. Hosting the app on a private server can provide tighter control over who can access the Shiny application, which can go some way to addressing these security concerns.

ii) Additional user interaction features within each view in the app. For example, users could be given the option to highlight certain sets of interest in the UpSet plot which can speed up the interpretation of findings. As another example, the bar chart could incorporate an option to compare the current view with the finding for a different media channel (i.e. represented as a reference line) to make comparisons easier as users may not be able to retain earlier views in their memory.

ACKNOWLEDGMENTS

The authors thank Associate Professor, KAM Tin Seong, *Singapore Management University* for his support and guidance, and the *Ministry of Manpower* for granting access to the survey dataset for research purposes.

[15] Koh, L. C., Slingsby, A., Dykes, J., & Kam, T. S. (2011, July). Developing and applying a user-centered model for the design and implementation of information visualization tools. In *2011 15th International Conference on Information Visualisation* (pp. 90-95). IEEE.