# R-CsI: A R-ConSumerInsights Business Application

LEE Kern Choong, LEE Yeng Ling, Debbie SIAH Mei Ping; KAM Tin Seong (Supervisor)

## Abstract

With the advent of technologies and increased availability of various sources of customer data, businesses are more able than before to gain invaluable insights about their customers as well as their purchase behaviours through data mining techniques. Data analysis allow the organizations to be customer-centric and develop customer acquisition, development and retention strategies to remain well-position and competitive in the business environment.

While many different analytics applications exist in R for these purposes, it should be noted that a number of them only address certain areas of the overall data lifecycle – for example, data preparation and visualization, or data visualization and analytics. Through the integration of R packages, our team will be creating an application, R-Consumer Insights (R-CsI), that will help users to address the end-to-end data lifecycle – from prepared data, to exploratory data analysis, and to use of data analytics methodologies, namely, Polytomous Latent Class Analysis (poLCA), Parallel Sets Visualisation (PSV), and Market Basket Analysis (MBA) techniques. Through effective data visualization, they enhanced these organizations ability to boost sales and develop strategies such as creating customer profiling based on their buying pattern, develop bundling/cross-sell strategies, finding best product association.

### Introduction

Technology in today's world is advancing faster than ever before. With the concepts of digital transformation, the Internet of Things (IoT) and cloud computing becoming increasingly prevalent, it has also become far easier to obtain and access large amounts of data on a variety of consumer activities in an ever-widening list of industries. By using various visual, statistical and data mining techniques on these data sets, businesses will be able to harness the power of hindsight with regards to customer behavior, allowing them to learn more about the activities, purchases or other transactions made by their customer base. Businesses will then be able to use the insights gleaned from data exploration and discovery to address fundamental issues for customer acquisition, development and retention.

### Motivation

As different customers have different needs and wants, it is only logical to conclude that they will have varying driving reasons for buying products. Therefore, customer segmentation is a very useful data mining technique to allow business to understand the purchasing habits of different groups of customers. In addition, products are rarely purchased in isolation; it is undeniable that customers are more likely to purchase sets of individual items, such as bread

and butter, and hence businesses also stand to gain much from being able to understand the relationship between the purchasing patterns of items. By understanding these differences, business can make better strategic choices about opportunities, product definition, and positioning, and can engage in more effective promotional efforts.

This project aims to discover insights on the segments that exist in a selected retailer's customer base, as well as identify groups of products that are highly associated during purchase. The data was obtained from the Dunnhumby data science company and tracks the purchases of 2500 households over 2 years.

## Implementation

R-CsI was developed using the free and open source statistical software R (https://www.r-project.org) and the R package shiny (https://cran.r-project.org/web/packages/shiny). R is a free and open source software environment for statistical computing and graphics that operates in a Windows, Mac OS X, and Linux environment. Shiny is a free, open source, extensible web applications framework for R that allows the creation of a rich, interactive web interface for users to interact with data in real-time and summarising data as tables, free text, or graphs. A shiny application can either operate on a local machine using a standard web browser to manage the interaction with a local instance of R, or it can operate on an internet connected server (http://www.shinyapps.io). Summary features of R-CsI are provided in Table 1.

**Table 1**

**Features of R-CsI app**

| Feature | Description |
|---|---|
| ***General*** | |
| Software & R packages | R-CsI was developed using the R statistical software and the R package shiny.<br><br>Additional R packages used to support R-CsI are included in Annex A. |
| Availability | R-CsI is available as an online application and as a standalone version for use offline. |
| Compatibility | Online version: any web browser on all desktop or laptop computers<br>Standalone version: all computers with a Windows or Macintosh operating system. |

| | |
|---|---|
| Installation | <u>R or Rstudio</u><br>To run any of the applications locally in R or Rstudio, an installation of the Shiny package (and any dependencies) is required, and ui.R and server.R scripts for the application saved in the same directory (i.e. a folder titled with the application's name, R-CsI). To launch the application from Rstudio, open each ui.R and server.R script in Rstudio and click "RunApp", which will appear in the top right hand corner of the source pane. Launching an application from R requires setting the working directory to where the application folder is located and using the runApp function. When initiating, Shiny will open a Web-browser window for the application.<br><br><u>R-CsI app</u><br>The online version requires no installation.<br>The standalone version is available in a zip folder and needs to be extracted and saved to the computer's hard drive. The extracted R-CsI folder contains portable versions of the R statistical software and the web browser Mozilla Firefox, which are required to run R-CsI, but do not require any installation. The standalone version can simply be launched by double-clicking the start file. |
| | |
| ***Built-in database*** | |
| Dunnhumby dataset | The Dunnhumby dataset contains over 2.5 million transactions of data at the household level, grouped into 7 tables. The details include individual purchases of specific items, categorisation of purchased items, customer demographics, and details of coupons used/redemptions made.<br><br>For this app, 3 Data Tables were used:<br>• HH_DEMOGRAPHIC Table – carries demographic information of 801 households (income, purchase frequency, etc);<br>• TRANSACTION_DATA Table – carries details of over 2.5 million transactions by the involved households;<br>• PRODUCT Table – product information of 92353 unique product ids with their commodity descriptions and categories. |
| | |
| ***Supporting Instruction*** | |
| User Instructions | The user guide provides information on how to set up and work with R-CsI. |

## Review and Critique of Past Works

Many R applications have been written on individual part (either clustering or MBA) but none available that deal with data analytics from prepared data to provision of customer insights (from data mining techniques) in totality. Therefore, to address this, the R-CsI app is intended to be fuller coverage in application on customers, which allows users to perform data preparation, visual exploratory data discovery, and finally, gaining insightful data analysis of the datasets seen.

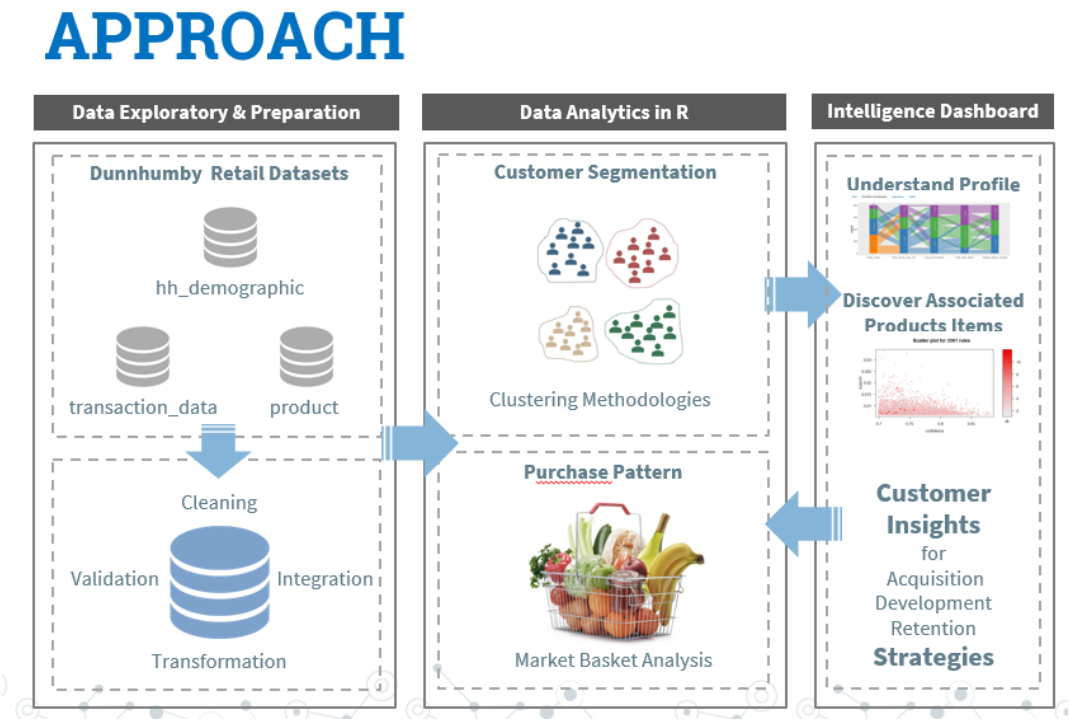
## Visual Design Framework & Discussion

### Idea Conceived

The features of this Shiny application, R-CsI, were conceptualized and developed by project Group 4 during the SMU course term 1 2018-2019 under the guidance of Prof Kam Tin Seong. By using various visual, statistical and data mining techniques on the Dunnhumby – The Complete Journey dataset, businesses will be able to harness the power of hindsight with regards to customer behaviour, allowing them to learn more about the activities, purchases or other transactions made by their customer base.

### Analytics Approach

A summary of our creation process:



## User Dashboard Design

The design of R-CsI dashboard visualizations. We used the Shiny dashboard package as we can easily create flexible and interactive dashboards with R shiny. The side panel navigation provides for the analytic tasks and the detailed tabs in each analytic task provides specific visualization and data tables. By having separate detailed tabs, the user does not need to scroll up and down and can instead toggle freely between tabs to access the different visualisations. The dashboard structure is shown in Figure 1.

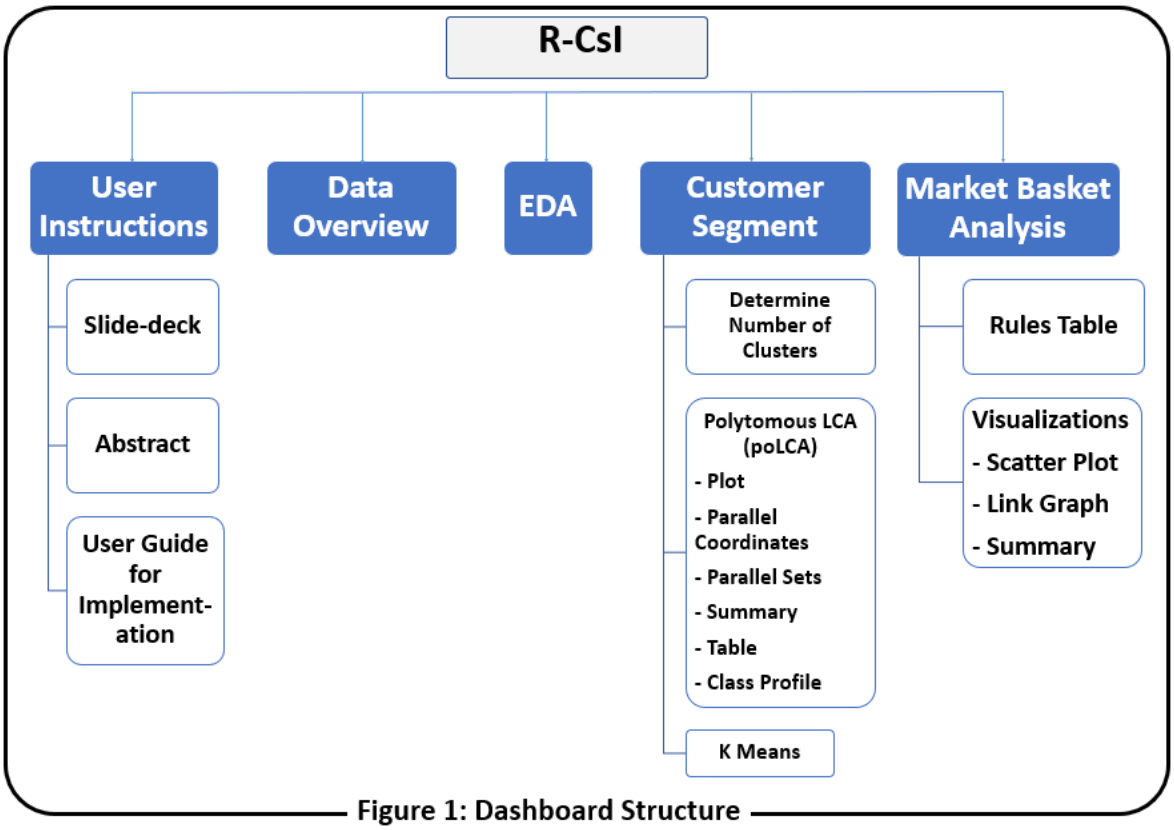



**Figure 1: Dashboard Structure**

## Data Tables

The R package DT and JavaScript library DataTables are used for displaying data in tables. DataTables is a JavaScript library to render HTML tables that can be paginated, filtered and sorted. The R package DT is an interface to DataTables based on htmlwidgets, and users do not need to know JavaScript to render HTML tables in Shiny or R Markdown.

## Datasets Preparation

**Customer Segmentation**

In order to perform polytomous Latent Class Analysis (poLCA), 4 of the numerical variables, were first binned and renamed. Recency, Frequency, and Monetary variables, along with the Number of different stores visited, were used to understand general customer behaviour over the previous 107 weeks (approximately) 2 years. The variables are:

| Original Name in Data | Renamed as |
|---|---|
| **Frequency** | Freq_of_Purchase |
| **Days.Since.Last.Transaction** | Time_Since_Last_Txn |
| **Total.Transaction.Spending** | Total_Amt_Spent |
| **Num.Diff.Stores.Visited** | Unique_Stores_Visited |

Each factor was binned as follows:

| Level | Freq_of_ Purchase | Time_Since_ Last_Txn | Total_Amt_Spent | Unique_Stores _Visited |
|---|---|---|---|---|
| **1 – Low** | 0-107 (Once Weekly) | 0-2 (< ½ a week) | 0-2140 (<$20/Week) | <=4 |
| **2 – Medium** | 108-214 (Twice Weekly) | 3-6 ( ½ to 1 week) | 2140-5350 ($20-$50/Week) | 5-7 |
| **3 - High** | >214 (>Twice Weekly) | >6 (> 1 week) | >5350 (>$50/Week) | >7 |

The four binned variables were then used for poLCA.

## Market Basket Analysis

According to Gulalkari N. (2016) "Implementing Apriori Algorithm in R", the key points to note on dataset preparation prior to implementing Apriori Algorithm:

(a) The data must have the following three columns (reference Figure 2):

**BasketID or Member_number**: An ID that can help distinguish different purchases by different basket or customers. [*In this case, it is the **BASKETID** being used.]*

**Date:** The date of transaction. [*In this case, Day 1 in the dataset is converted to date and assumed to be 1/1/ 2013.]*

**ItemDescription:** The description of the actual item that was bought. *[In this case, it is the **COMMODITYDESC** being used.]*

**Figure 2: Dataset for MBA**

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | householdkey | BASKETID | COMMODITYDESC | DEPARTMENT | DAY | DATE | STOREID | TRANSTIME | BuyCode | LikelyClass |
| 2 | 1004 | 27972546776 | COLD CEREAL | GROCERY | 83 | 25/3/2013 | 316 | 1629 | Buy L21 mths | Class1 |
| 3 | 1004 | 27972546776 | FRZN NOVELTIES/WTR ICE | GROCERY | 83 | 25/3/2013 | 316 | 1629 | Buy L21 mths | Class1 |
| 4 | 1004 | 27972546776 | FRZN NOVELTIES/WTR ICE | GROCERY | 83 | 25/3/2013 | 316 | 1629 | Buy L21 mths | Class1 |
| 5 | 1004 | 27972546776 | FLUID MILK PRODUCTS | GROCERY | 83 | 25/3/2013 | 316 | 1629 | Buy L21 mths | Class1 |

(b) Data cleaning and manipulations using R:

- The basket format must have a column as a unique identifier of each transaction. The BASKETID should be of numeric data type and then sort the dataframe based on BASKETID.
- Using ddply, offered by package plyr – convert the dataframe into transactions format such that we have all the items bought at the same time in one row.

## The Application

**Visualization**

**POLYTOMOUS LCA (poLCA)**

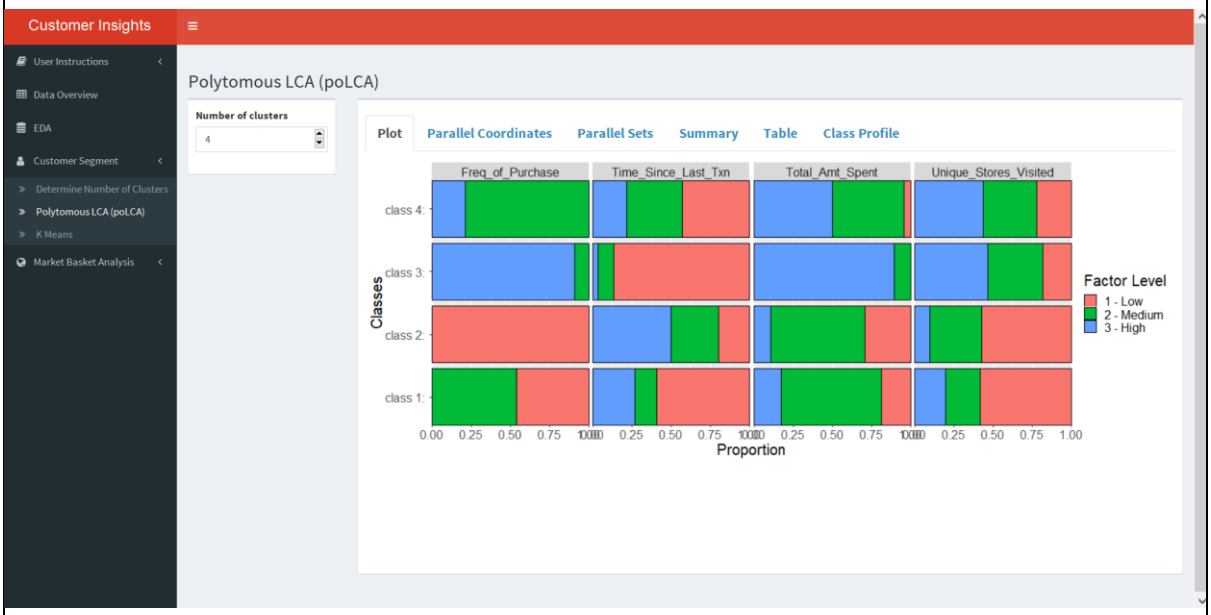



**Chart type:** Stacked Bar Chart

**R Package:** plys, poLCA, reshape2

**Interpretation:** Using a stacked bar chart, the app shows how customers have been segmented into different classes to identify differing cluster profiles, as well as allows for easy comparison between the clusters.

**Interactivity**: The user is able to choose the number of clusters that the data will be segmented into.

**PARALLEL SETS PLOT**

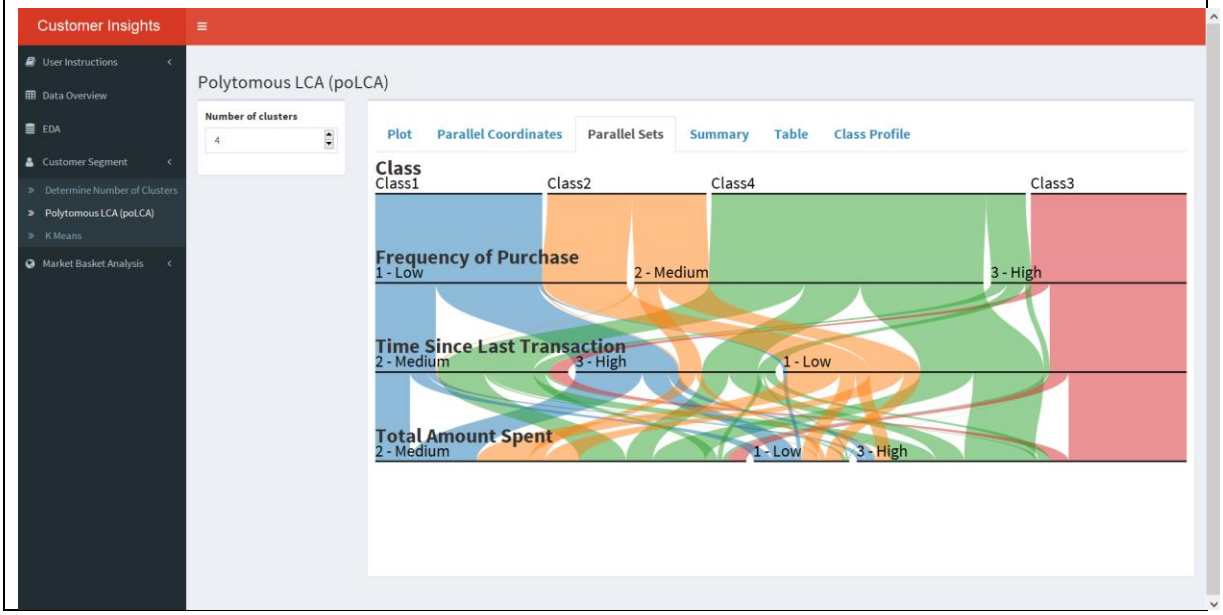


**Chart type:** Parallel Sets Plot
**R Package:** parsetR (Note: parsetR is not available on CRAN. For instructions on how to use the parsetR library, please refer to http://timelyportfolio.github.io/parsetR/.
**Interpretation:** A Parallel Sets Plot allows users to visualize the relationship between the variables used in the analysis. The graph is read from top (parent) to bottom (child), and consists of a list of nodes (factor levels) and paths (the lines between the nodes).
**Interactivity**: A mouseover of any of the paths in the graph will provide a tooltip showing data such as the percentage customers on each path, as well as the exact profile of that line. All parent and child paths for preceding and subsequent variables will also be highlighted.

**CLUSTER PROFILE HISTOGRAM**

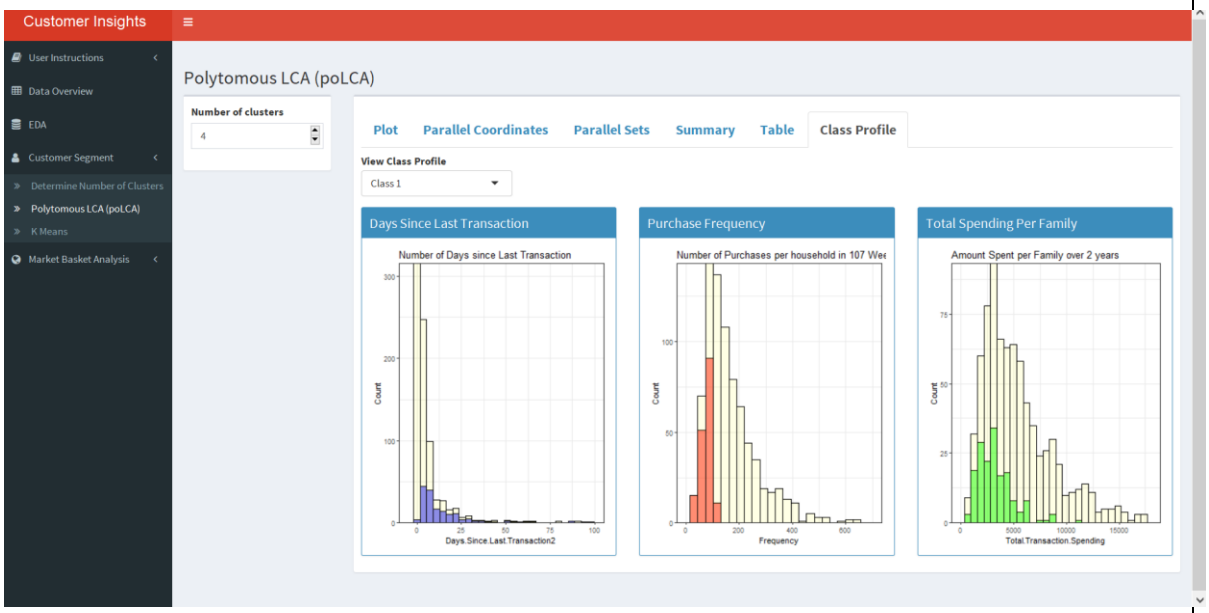
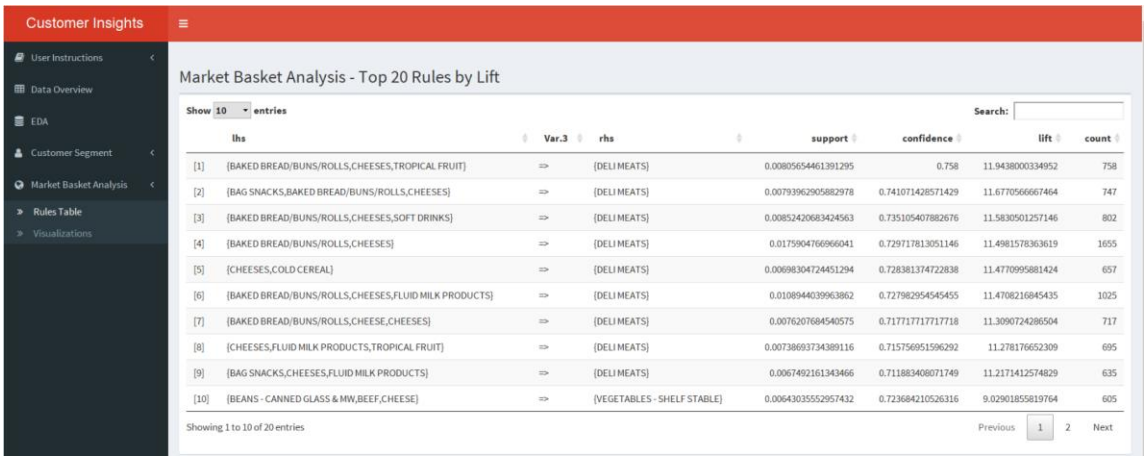
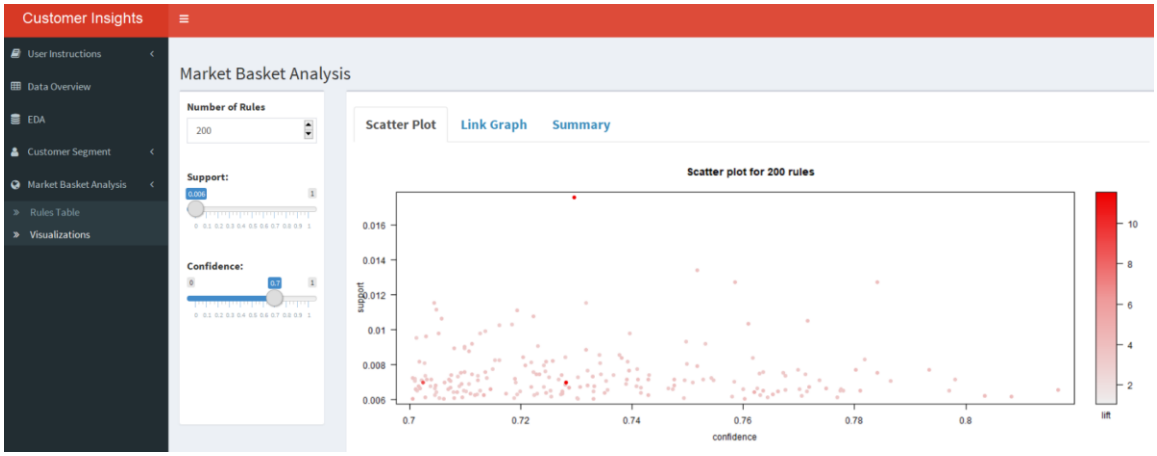
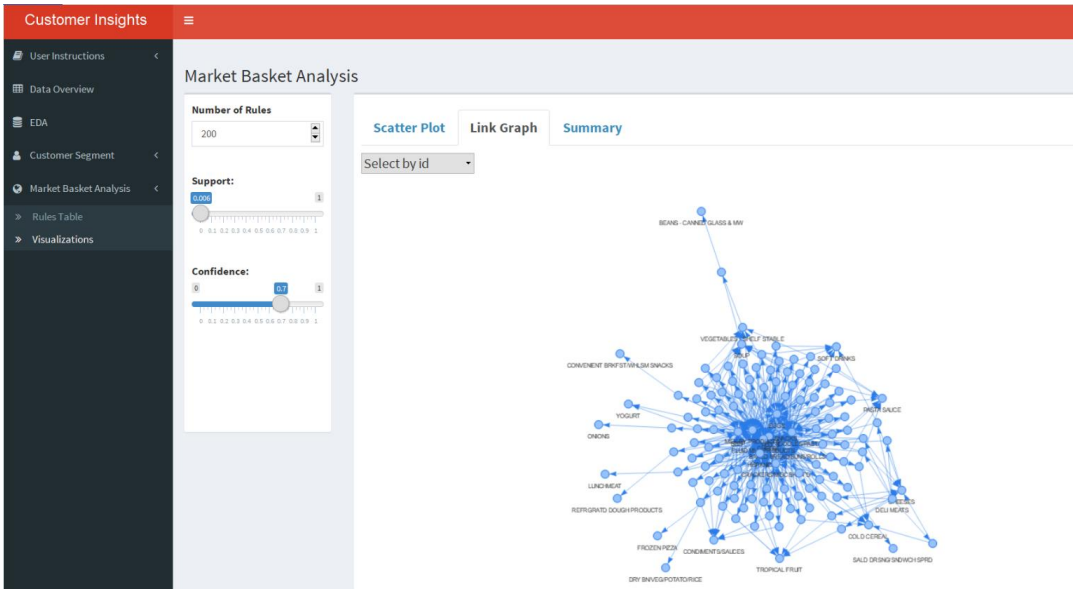



**Chart type:** Histogram
**R Package:** ggplot
**Interpretation:** The cluster profile charts overlay the data for each cluster on that for the population as a whole. This allows users to compare between the cluster and the data and see how the cluster is distributed among the rest of the population.
**Interactivity**: Users are able to select the cluster which they would like to compare to the original dataset.

## Visualization: Market Basket Analysis

| Visualization |
|---|
| **DATA TABLE**<br>  <br><br>**Chart type:** Data Table<br>**R Package:** DT<br>**Interpretation:** Using a Data Table to list the top 20 generated association rules sorted by lift. |
| **SCATTER PLOT**<br>  <br><br>**Chart type:** Scatter Plot<br>**R Package:** arules, arulesViz, ggplot2, plotly<br>**Interpretation:** Using a scatter plot to visualize the generated association rules in arulesViz with two interest measures on the axes (eg. confidence vs support) and a third measure (lift) with shading. Bayardo, Jr. and Agrawal (1999) argue that the most interesting rules (sc-optimal rules) reside on the support/confidence border, which can be clearly seen (highlighted in dark red) in the above plot.<br>**Interactivity:** Set parameters (support/confidence) to generate valid association rules for MBA and render plot. |

## NETWORK GRAPH



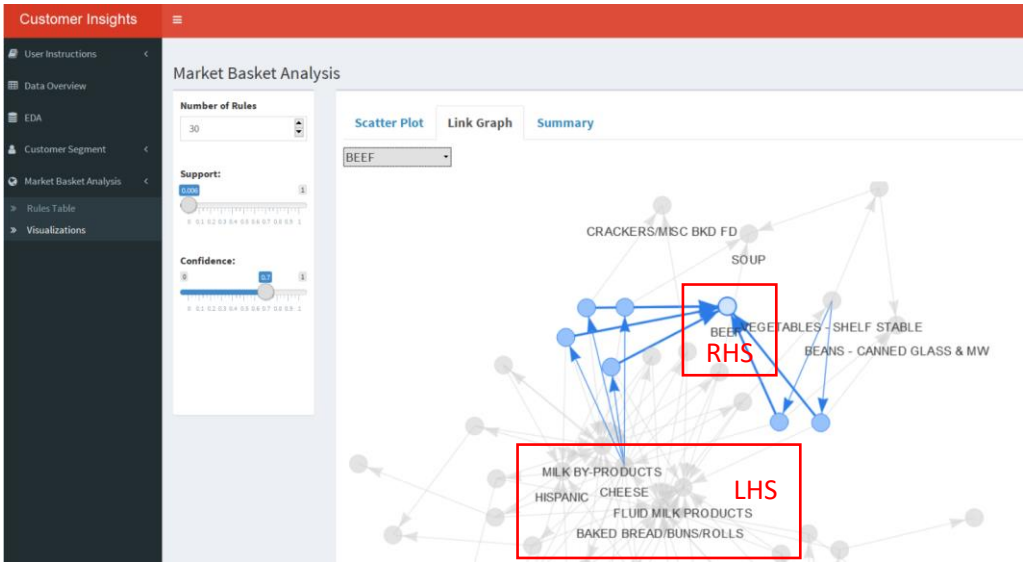




**Chart type:** Network Graph

**R Package:** visNetwork

**Interpretation:** Using Network Graph, the business user will be able to identify rules regarding customer purchase of products that generally occur within the same transaction. The arrows (edges) are the rules visualizing the causal relationship between the LHS (antecedent) and RHS (consequent). LHS (antecedent) items will have arrows (base of arrows) pointing to the rule nodes whereas RHS (consequent) items will have the arrows head pointing to nodes.

**Interactivity:** Set parameters (support/confidence) to visualize the causal relationship/association between the LHS and RHS items. User is able to select by item to visualize the association between LHS and RHS for specific item.

## Future Developments

At present, the binning process is done prior to visualization using the app. In the future, we would like to create a way for users to bin both numerical and non-numerical variables (by allowing the creation of user-specified value-ordered levels). It would also be good if users would be able to select the number of bins used to get a deeper look at whether there may be different, possibly better, results given a differing number of bins.

## Acknowledgments

We would like to thank Prof Kam Tin Seong for introducing us to the world of visual analytics and R, as well as his invaluable guidance and advice for our project.

## References

[1] Dunnhumby dataset – The Complete Journey. Retrieved from https://www.dunnhumby.com/careers/engineering/sourcefiles

[2] Gulalkari N. (2016). *Implementing Apriori Algorithm in R*. Retrieved from https://datascienceplus.com/implementing-apriori-algorithm-in-r/

[3] Bayardo, Jr RJ, Agrawal R (1999). *Mining the most interesting rules*. In KDD'99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 145-154. ACM

[4] Kosara, Robert, Fabian Bendix, and Helwig Hauser. "Parallel sets: Interactive exploration and visual analysis of categorical data." Visualization and Computer Graphics, IEEE Transactions on 12.4 (2006): 558-568.

[5] Russell, K. (2015, September 17). Introduction to parsetR. Retrieved December 5, 2018, from http://timelyportfolio.github.io/parsetR/

[6] Linzer, D., & Lewis, J. (2016, August 29). Package 'poLCA'. Retrieved December 6, 2018, from https://cran.r-project.org/web/packages/poLCA/poLCA.pdf

## Annex A

**List of R Packages Used**

| | |
|---|---|
| **1** | ggplot2 |
| **2** | shinydashboard |
| **3** | plyr |
| **4** | dplyr |
| **5** | DT |
| **6** | tidyverse |
| **7** | cluster |
| **8** | factoextra |
| **9** | purrr |
| **10** | scatterplot3d |
| **11** | poLCA |
| **12** | reshape2 |
| **13** | knitr |
| **14** | readr |
| **15** | stringr |
| **16** | arules |
| **17** | arulesViz |
| **18** | readxl |
| **19** | lubridate |
| **20** | networkD3 |
| **21** | ggiraph |
| **22** | RcolorBrewer |
| **23** | visNetwork |
| **24** | igraph |
| **25** | kableExtra |
| **26** | R.devices |
| **27** | devtools |
| **28** | parsetR |