R: A Swiss Army Knife for Market Research

Enhancing work practices with R

Martin Chan – Consultant, Rainmakers CSI

12th September, 2018

This presentation is about using R in **business environments or conditions where its usage is less intuitively suitable…**

… and a story of how R has been transformative for our work practices

EARL CONFERENCE

Who are we?
What do we do?

What

Answer strategic questions to help our clients grow profitably

Inform strategic decisions through creative and analytical thinking

How

Estimate size of market, and size of opportunity, anticipating trends

Utilise resources available within an organisation (e.g. stakeholder knowledge, existing research)

Develop consumer targeting frameworks – identify core consumers and areas of opportunities

Where

Across multiple industries and markets

- FMCG
- Finance & Insurance
- Travel
- Media

EARL CONFERENCE

A team with a mix of backgrounds from strategy, brand planning, marketing, research…

(where analytics is a key, but only one of the components of our work)

EARL CONFERENCE

RAIN MAKERS
CUSTOMER STRATEGY INTELLIGENCE

What's different about the use of R in our work?

Challenges of Using R

1. **Nature of data: disparate, patchy**

2. **Nature of U&A data**

3. **Client requirement and expectations** (of outputs)

U&A – research with the aim to understand a market and identify growth opportunities by answering questions on whom to target, with what, and how.

(Source: https://www.ipsos.com/en/ipsos-encyclopedia-usage-attitude-surveys-ua)

EARL CONFERENCE

1 Nature of data: disparate, patchy

EARL CONFERENCE

Our data come together in different forms, like pieces of

JIGSAW

Historical survey data – often designed for different purposes and collected from different samples

Population / Demographic data (from census, World Bank research etc.)

Pricing data (e.g. Euromonitor, Nielsen)

Primary research (e.g. U&A, customer satisfaction surveys)

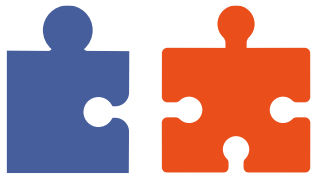
Stakeholder Interviews (Qualitative)

Customer Interviews (Qualitative)

Historical segmentation work (past work produced by client's suppliers)

Challenges in Using R

Disparate and patchy data…



… renders it more difficult to reap the benefits offered by R

1. **Ad-hoc data** – easier to explore in Excel
   Traditionally Excel is an easy tool for these (albeit chaotic) situations when you need to 'play around' to figure out your plan of attack

2. **Data is small or 'non-repetitive'** – not worth automating
   Perception of using R – for automating analysis - is only justifiable if the benefits of reproducibility exceed the costs of setting up code (prominent for small data sets)

3. **Poorly structured / stored data** – more time-consuming to use R to clean
   Condition of the input data (think merged cells in Excel, or data saved in PowerPoint) almost make R even more time-consuming in the short-term

Integrating data types

Integrating analysis from different sources of data under one roof

Package for Qualitative Data Analysis

Qualitative Research
stakeholder interviews, telephone interviews; focus groups, workshops

Output
Understand the “why” and the reasoning behind behaviour, or develop hypotheses / segments – development of a framework of themes and quotes

Qualitative Analysis (Without R)

Interview or group takes place, which is recorded and transcribed into a text or Word file

- **Transcripts are annotated and interpreted**
- **Possible themes / hypotheses are documented in a separate document**
- **Quotes are copied out and pasted into a separate document,** with references to the respondent background and interview number

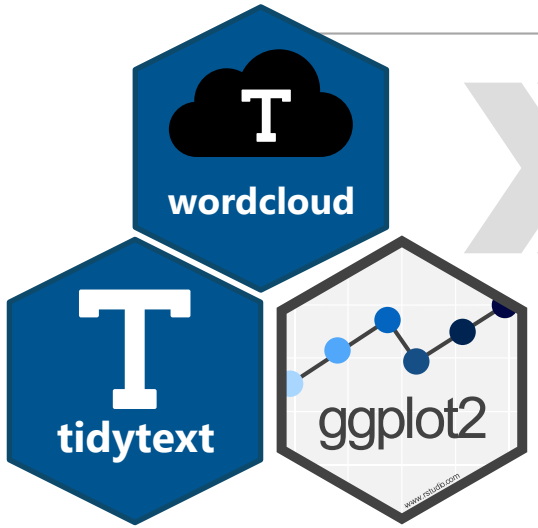
Transcript | Quotes | Themes

Keep refining framework until it is a fair and sensible interpretation of the different themes emerging from the interviews

A framework of themes/hypotheses or a story is developed – and a report is produced using the analysis of the annotations, themes, and quotes

RQDA: What is it used for?

1. A more systematic tool for analysing qualitative data
2. **GUI launched in R for marking up themes**, quotes from qualitative interviews
3. **Creates an analysis output in sqlite database**, allowing further analysis or production of outputs within R
4. **Integrated with R** – use sqlite to extract data, enable exploration of insights through text mining techniques

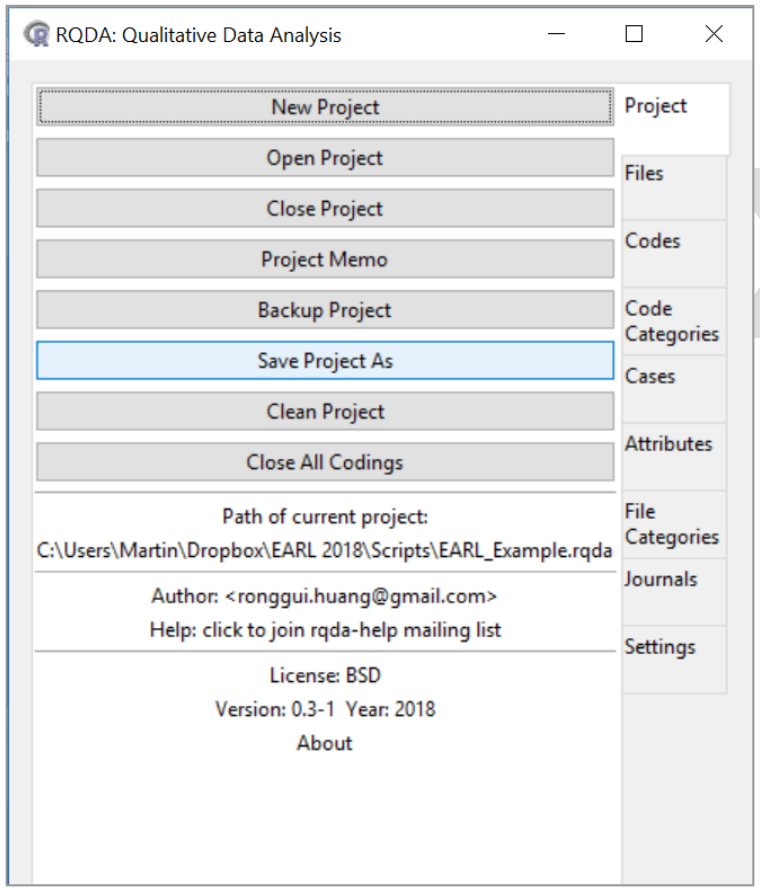


Outputs

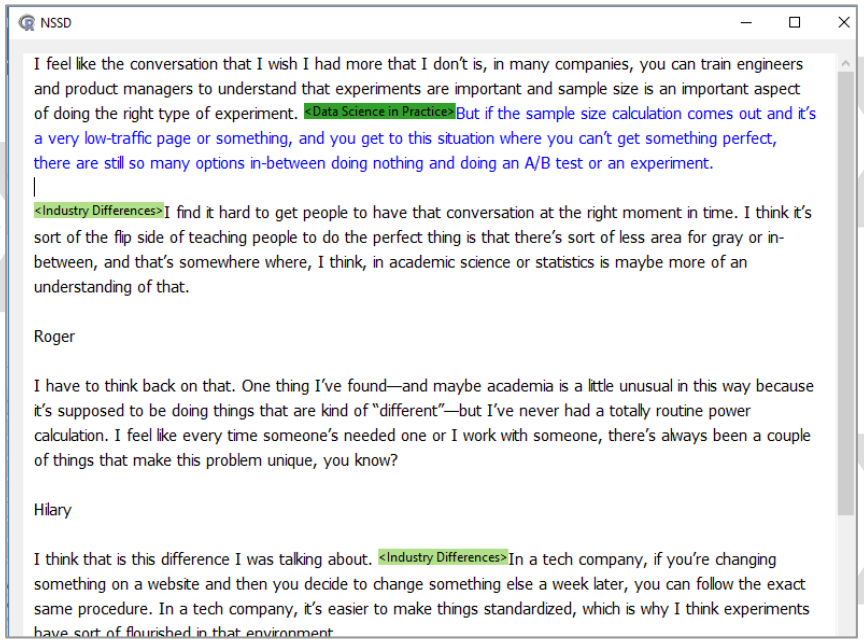
1. Word / N-gram frequency analysis → **Help generate hypotheses**
2. Word cloud → **Visual output for communicating strategic output**
3. Table output → **Ease of retrieving quotes to be used in a presentation**

Integrating data types

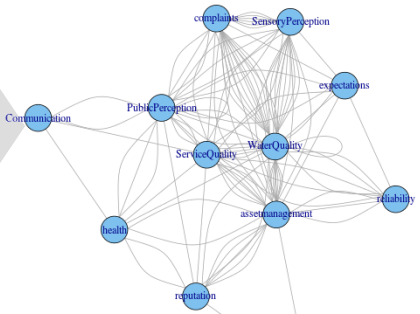
RQDA GUI Interface



Marked up transcript



Outputs



Structure of database

Code ID	Codes / Themes	File	…
1	Data Science in Practice	1, 2, 3	**frequency of occurrence, code themes, etc.**
2	Industry Differences	1, 2	
3	R versus Python	3,5	
…	…	…	

Source:
https://leanpub.com/conversationsondatascience/read_sample

2 Nature of U&A data

U&A – research with the aim to understand a market and identify growth opportunities by answering questions on whom to target, with what, and how.

(Source: https://www.ipsos.com/en/ipsos-encyclopedia-usage-attitude-surveys-ua)

EARL CONFERENCE

Features of a Usage & Attitude Survey (U&A)

Challenges of Using R

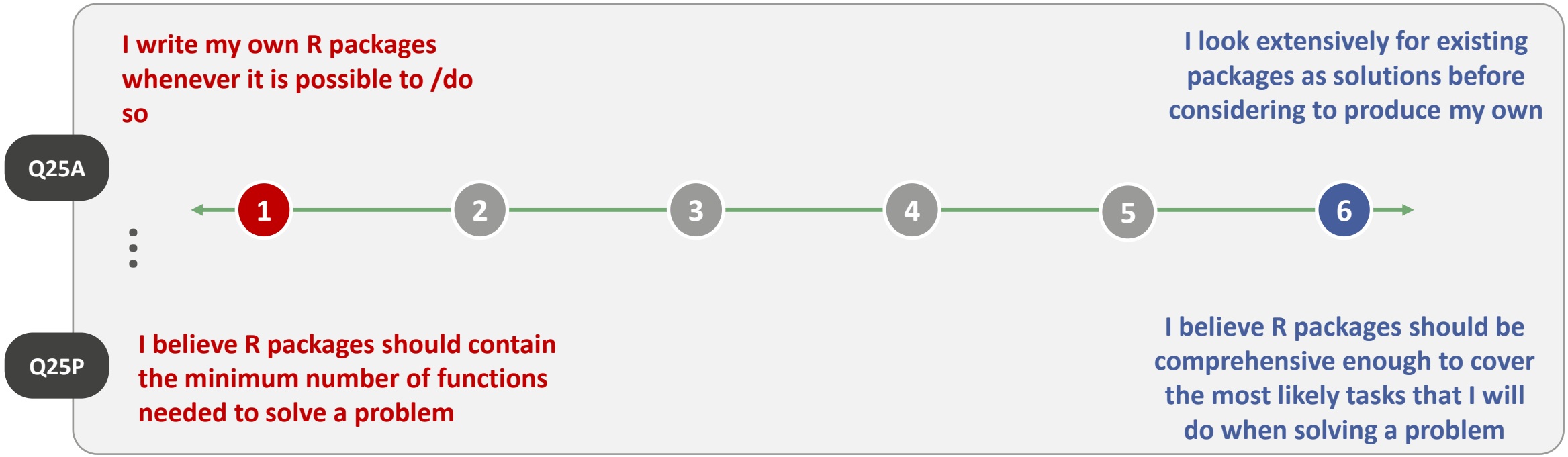
1. **Wide data**
   Large number of variables with relatively few cases / observations

2. **Design-dependent**
   Variables must be very closely interpreted relative to the original research design, e.g. questionnaire wording, rating vs ranking – making it difficult to interpret results within R

Example Question 1

Attitudinal

Q25. For each of the following statements - on a scale of 1 to 6, please select the score which best describes your attitude towards using R packages:



Wording will also slightly differ if you enter the survey as a different respondent – e.g. R package developer, beginner or heavy R user!

Example Question 2

Q27. For each of the following tasks, please select the packages that you are likely to use in carrying out the task… **SELECT A MAXIMUM OF THE TEN MOST LIKELY PACKAGES**

Tasks

1. Building a linear model with an interactive dashboard output
2. Writing a package to store customised analysis functions for a project
3. Producing publishing quality visualisations

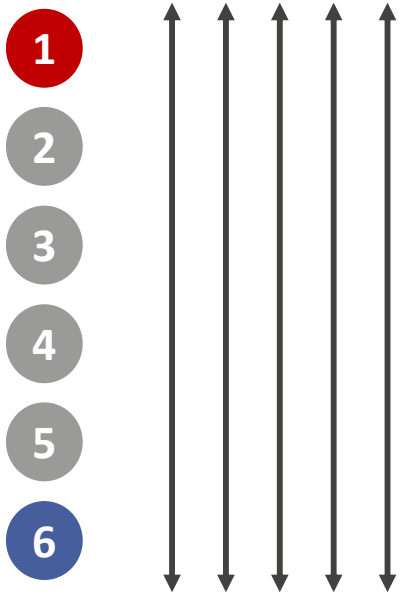
   …
12. Analyse survey data

Packages

1. dplyr
2. tidyr
3. data.table
4. tibble
5. shiny
6. plotly
7. ggplot2

   …
110. Other (Please enter)

Entered through search box




```
241  q27     1    21   q27_1_21
242  q27     2    21   q27_2_21
243  q27     3    21   q27_3_21
244  q27     4    21   q27_4_21
245  q27     5    21   q27_5_21
246  q27     6    21   q27_6_21
247  q27     7    21   q27_7_21
248  q27     8    21   q27_8_21
249  q27     9    21   q27_9_21
250  q27    10    21  q27_10_21
 [ reached getOption("max.print") -- omitted 1070 rows ]
```

Number of variables from one question

Our challenge

1. The meaning represented by each variable is highly dependent on the **exact wording on both ends of the scale**

2. **Sparse data** (particularly in example 2)

3. Handling both the **variable and value labels** effectively through R when there are often a large number of these variables, which are not supported by data frames / tibbles

This explains the prevalence of **GUI analysis software** such as SPSS and Q

EARL CONFERENCE

The solution to overcoming this is to use a mixture of the following that helps us automate certain workflows:

- 

- for loops / apply functions
- merge functions
- User-defined functions

Generalised Approach

Create empty list

Loop* through scenarios / categories

Operations to wrangle / analyse data, e.g. group_by(), summarise(), mutate(), spread(), gather()

Output: either as a tibble / data.frame or **mschart** object

Assign output to member of list

→ Create a combined output using merge_recurse or merge_all

→ Export Excel output using writexl::write_excel()

→ Export PowerPoint output using mschart package

+ timestamp to outputs to document analysis

*Or in the form of combining sapply and a UDF, if possible – for computation efficiency; but for loops benefit from readability

Survey Analysis (With R)

Import

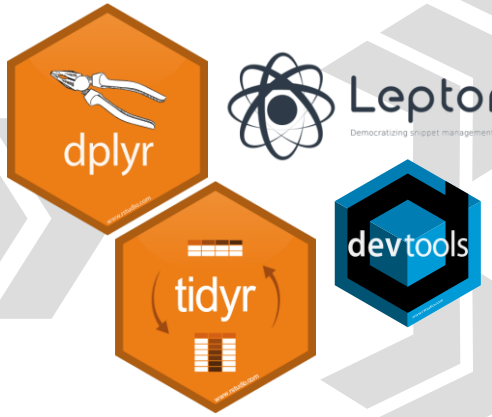
Import files as SAV (SPSS) or CSV files; sometimes serialised as RDS to save loading time and memory

Prepare

Extract, cleaning, and structuring data to prepare for analysis

Analyse

Run different iterations



apply functions, loops, and custom functions for generating the outputs – usually in the form of summary tables

Lepton providing an alternative to R packages for managing ad-hoc functions

Output

Summary tables, e.g. contingency tables

VBA for automating formatting

PowerPoint outputs – occasionally using SVG (vector) for visual quality

EARL CONFERENCE

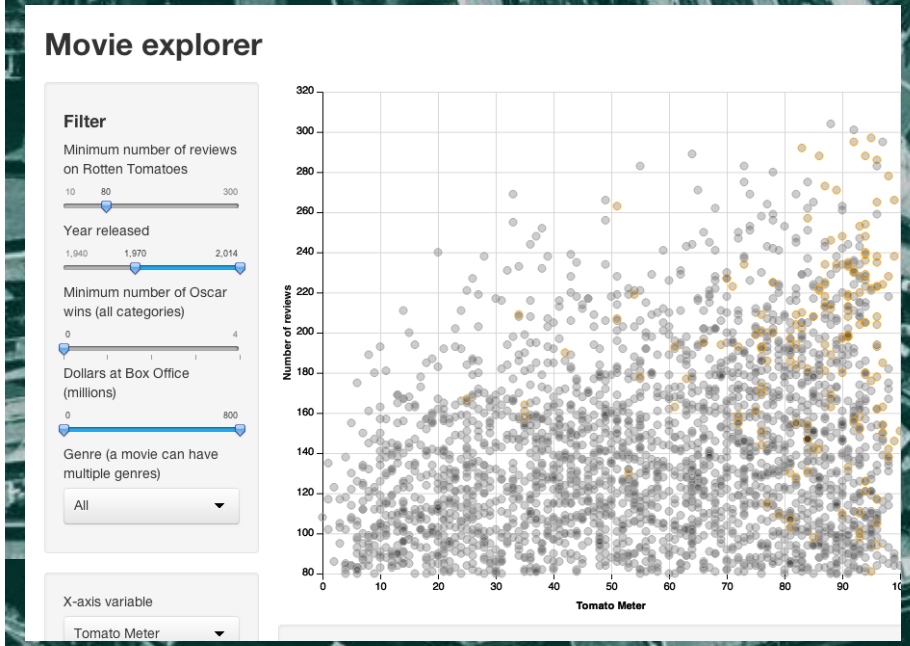
Outcome

- **Reproducibility** – can trace back to exactly what you have done
- **A quick and organised way of exploring data through iterations** (using custom functions)

EARL CONFERENCE

3 **Client requirement and expectations** (of outputs)

In many business environments there is still a preference of (perceived) **simple and reliable outputs** over interactive dashboard outputs like Shiny or Tableau



EARL CONFERENCE

There are many reasons for this – sometimes it may be the **nature of the data** (small, ad-hoc datasets that require a lot of cleaning), but also driven by **certain specific needs**:

1. **Clients can easily adapt content for their own use** when sharing with internal stakeholders
2. **No requirement to know any code** (access to data not dependable on internal skill or supplier)
3. **Perceived to have fewer 'moving parts',** e.g. what do I when there is an error
4. **Output is not merely data visualisation** – often a requirement to heavily annotate or to construct a conceptual framework
5. **Familiarity…**

Correlations: Age (Indices) and **Reach (as % of all 15+ adults)** Results not entirely intuitive - to investigate

Sample Weight	Reach (%)
London_15_24	0.1575727
London_25_34	0.3078648
London	-0.0950570
	-0.2475772
	-0.3399903
	-0.1429540
	0.3111817

Correlations: Age **f all 15+ adults)**

Results not entirely

	Reach (%)
	0.3716040
	062723
London	18
London_4	
London_55	

Sample Weight	Reach (%)
London_65	-0.2855705
London_All_men 25_34 yrs	0.6457612

Familiarity (Style)

"Content looks great, could you put this in a format that looks less academic?"

EARL CONFERENCE

We work with partners and clients from very varied backgrounds of technical knowledge and use different kinds of software.

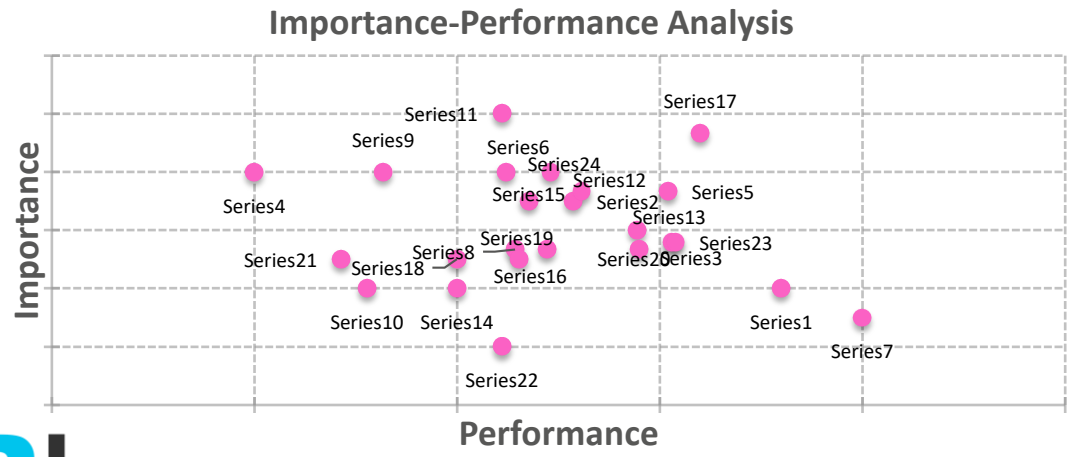
For better or worse, Excel and PowerPoint are still the dominant vessels for communicating analysis and findings.

Thankfully, R has the versatility to accommodate these needs!

Automating PowerPoint charts with **officeR** and **mschart** packages (David Gohel)

Applications in automating Importance Performance Analysis in PowerPoint

- Scatterplots with variable labels



EARL CONFERENCE

Generalised Approach

Read in a pre-set PPTX template and assign to an object

Create function to manipulate data, and return a **mschart object** using functions from the **mschart** package

Create function that adds the chart object to the PPTX object

Write PPTX object

1. Create a list of data frames for 'converting' into chart objects
2. Read in Template
3. Function for appending slides to PPT object (within R)
4. Loop sequence that creates formatted chart objects from 'output list'
5. Assigning them to 'output_combined_doc'
6. Print PPT with timestamp

```
#### Commence PowerPoint production ####
output_list <- list(NULL) # Initalise a list

segment_char <- c("Segment A","Segment B","Segment C")

for(i in 1:length(segment_char)){
  str <- segment_char[[i]]

  produce_tables(segment_char[[i]]) %>% # custom function for analysis / dplyr pipeline
    # takes segment name (character) as argument
    append_to_list(output_list,str) # assign table to list
}

source_doc <- read_pptx("Source/Source_PowerPoint.pptx") # Read in PowerPoint Template File
output_combined_doc <- source_doc # Initialise an "Output" document

# Function for adding newly created slides to PowerPoint file
gen_chart <- function(file,chart){
  add_slide(file, layout="Content_1",master = "Empty Slide with Chart") %>%
    ph_with_chart(chart=chart)
}

for(i in 1:length(output_list)){
  segname <- names(output_list)[[i]]
  title <- paste("Satisfaction Importance Matrix -",segname)

  output_list[[i]] %>%
    separate('.',sep="/",into="label",extra="drop") %>% # Text to columns abbreviating effect
    drop_na() %>%
    ms_scatterchart(x="Satisfaction",y="Importance",group="label") %>%
    chart_labels(title=title) %>%
    chart_data_labels(position="b", show_legend_key = FALSE,show_serie_name = TRUE) %>%
    chart_labels_text(values=fp_text(color="black",font.size=10,font.family="Calibri")) %>%
    chart_data_fill(values="#4F81BD")%>%
    chart_data_stroke(values="#4F81BD")%>%
    chart_data_size(values=6) -> p_chart

  output_combined_doc <- output_combined_doc %>%
    gen_chart(p_chart)
}

print(output_combined_doc,
      target=timed_fn("Output/Importance Satisfaction Slides",".pptx"))
```

EARL CONFERENCE

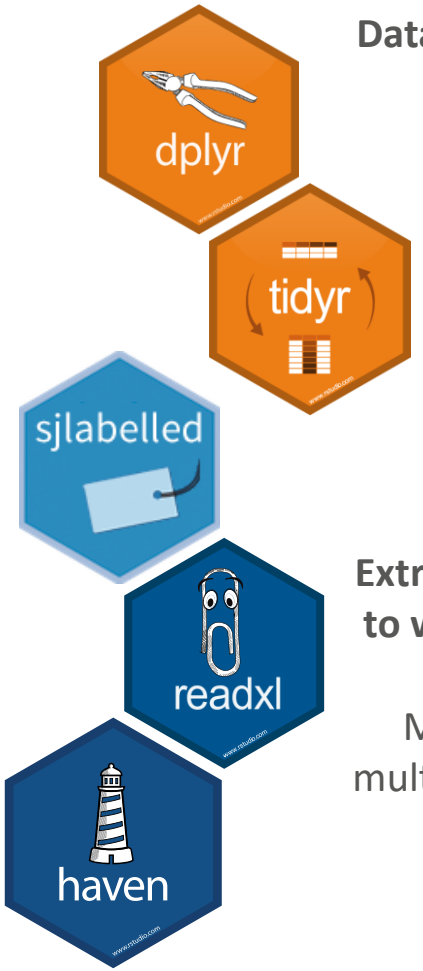
```vba
Sub ScatterLabelsTweak()

Dim sld As Slide
Dim shp As Shape
Dim sr As Series
Dim chrt As Chart
Dim i, j, k, m As Long

r = 28.3464567 'r converts cm to points

    For Each sld In ActivePresentation.Slides
        For Each shp In sld.Shapes

            If shp.HasChart Then
                shp.Height = 12.83 * r
                shp.Width = 22.73 * r
                shp.Left = 1.34 * r
                shp.Top = 3.88 * r

                shp.Chart.ChartTitle.Font.Size = 14
                shp.Chart.ChartTitle.Font.Name = "Calibri"

                With shp.Chart.Axes(xlValue) 'Y-axis
                    .TickLabelPosition = xlNone
                    .AxisTitle.Font.Size = 14
                    .AxisTitle.Font.Name = "Calibri"
                End With
                With shp.Chart.Axes(xlCategory) 'X-axis
                    .TickLabelPosition = xlNone
                    .AxisTitle.Font.Size = 14
                    .AxisTitle.Font.Name = "Calibri"
                End With
                shp.Chart.HasLegend = False

                j = shp.Chart.FullSeriesCollection.Count
                Debug.Print j
                For i = 1 To j
                    shp.Chart.FullSeriesCollection(i).Format.Fill.Visible = msoTrue
                    shp.Chart.FullSeriesCollection(i).Format.Line.Visible = msoTrue

                    shp.Chart.FullSeriesCollection(i).Format.Fill.ForeColor.RGB = RGB(94, 98, 68)
                    shp.Chart.FullSeriesCollection(i).Format.Line.ForeColor.RGB = RGB(94, 98, 68)
                    shp.Chart.FullSeriesCollection(i).HasLeaderLines = True
                    k = shp.Chart.SeriesCollection(i).Points.Count
                    Debug.Print k
                    For m = 1 To k
                        shp.Chart.SeriesCollection(i).Points(m).DataLabel.Font.Size = 8
                        shp.Chart.SeriesCollection(i).Points(m).DataLabel.Font.Name = "Calibri"
                    Next
                Next

            End If

    Next shp
    Next sld

End Sub
```

PowerPoint VBA

- Scatter plot style formatting
- Axes formatting
- Repositions and resizes charts
- Loops through entire PowerPoint document

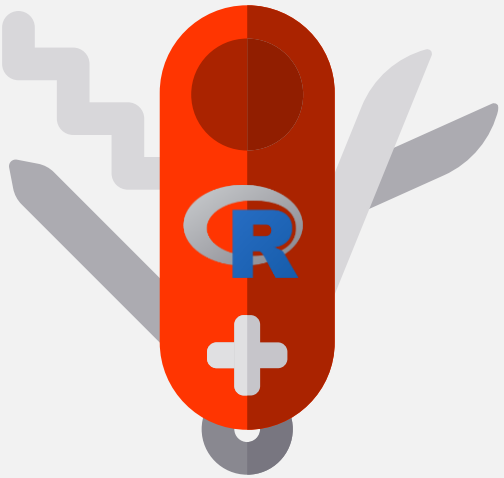
EARL CONFERENCE

Summary



Data cleaning and manipulation that is readable and reproducible
Also easy to cache in functions – significant improvement to Excel

Sheer versatility and efficiency offered by user-defined functions, apply functions and loops



PowerPoint chart automation with **OfficeR** and **mschart**

(if necessary – publishing quality vector image outputs with SVG using ggplot2)

Extremely versatile tools to work with the kind of data we receive
Multi-level, loop-level, multiple choice type data

Making Qualitative Data Analysis systematic with **RQDA** – significant improvement on traditional analysis

Ability to automate or conduct reproducible survey analysis
Survey analysis with questionr or survey

EARL CONFERENCE

Thank you!

martin.chan@rainmakerscsi.com

Rainmakers CSI Ltd
71 Gloucester Place
London W1U 8JW
T +44 (0) 20 3691 8100
E info@rainmakerscsi.com
rainmakerscsi.com