Data Science with R
Introducing R

Graham.Williams@togaware.com

9th June 2014

Visit http://onepager.togaware.com/ for more OnePageR's.

In this laboratory we use the **heart** dataset to explore and introduce R.

The required packages for this module include:

```
library(ggplot2)
```

As we work through this chapter, new R commands will be introduced. Be sure to review the command's documentation and understand what the command does. You can ask for help using the ? command as in:

```
?read.csv
```

We can obtain documentation on a particular package using the *help=* option of `library()`:

```
library(help=rattle)
```

This chapter is intended to be hands on. To learn effectively, you are encouraged to have R running (e.g., RStudio) and to run all the commands as they appear here. Check that you get the same output, and you understand the output. Try some variations. Explore.

1 Working on Heart Dataset

Using knitr (Xie, 2014) within RStudio, and the **heart** dataset, produce a report, including summaries and interesting plots of the dataset. Submit the PDF version of your report for assessment.

1. Create a new KnitR document within RStudio.

2. Add a title and author and date. Ensure you can Compile PDF.

3. Create a section for loading the dataset.

4. Load the **heart** from the file `heart.csv` using R's `read.csv()` to load the data into R.

   ```
   ds <- read.csv("data/heart.csv")
   ```

5. Include some basic textual information about the dataset.

6. Add a new section to report on some basic univariate summaries of the dataset.

7. Add a new section to include some plots of the data. For the plots include some description to help the reader understand what is interesting about the plot. You will use ggplot2 (Wickham and Chang, 2014) for this.

8. Add a section to summarise any interesting observations you made of the dataset.

9. Add a final section to illustrate the use of the following knitr options: echo, eval, results, fig.width fig.height, out.width, out.height.

2   Further Reading

The Rattle Book, published by Springer, provides a comprehensive introduction to data mining and analytics using Rattle and R. It is available from Amazon. Other documentation on a broader selection of R topics of relevance to the data scientist is freely available from http://datamining.togaware.com, including the Datamining Desktop Survival Guide.

This module is one of many OnePageR modules available from http://onepager.togaware.com. In particular follow the links on the website with a * which indicates the generally more developed OnePageR modules.

Other resources include:

- Documentation within RStudio to identify some functions to provide the summaries of the dataset.

- Stack Overflow is a great question and answer resource.
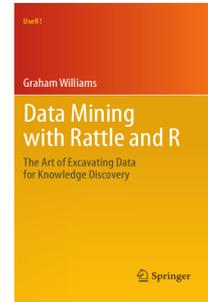
3   References

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Wickham H, Chang W (2014). *ggplot2: An implementation of the Grammar of Graphics.* R package version 1.0.0, URL http://ggplot2.org,https://github.com/hadley/ggplot2.

Williams GJ (2009). "Rattle: A Data Mining GUI for R." *The R Journal*, **1**(2), 45–55. URL http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf.

Williams GJ (2011). *Data Mining with Rattle and R: The art of excavating data for knowledge discovery.* Use R! Springer, New York. URL http://www.amazon.com/gp/product/1441998896/ref=as_li_qf_sp_asin_tl?ie=UTF8&tag=togaware-20&linkCode=as2&camp=217145&creative=399373&creativeASIN=1441998896.

Xie Y (2014). *knitr: A general-purpose package for dynamic report generation in R.* R package version 1.6, URL http://CRAN.R-project.org/package=knitr.

This document, sourced from IntroRR.Rnw revision 419, was processed by KnitR version 1.6 of 2014-05-24 and took 0.9 seconds to process. It was generated by gjw on nyx running Ubuntu 14.04 LTS with Intel(R) Xeon(R) CPU W3520 @ 2.67GHz having 4 cores and 12.3GB of RAM. It completed the processing 2014-06-09 10:37:55.