One Page R Data Science Template

Graham.Williams@togaware.com

3rd June 2018

Visit https://essentials.togaware.com/onepagers for more Essentials.

Introduce the chapter through two or three paragraphs.

This document serves as a template from which all other chapters derive their basic structure. The intent is that when starting a new chapter we begin by taking a copy of this document as the first version of that chapter. Remove the packages in the next section that are not relevant. Some basic examples are also included to copy and paste or remove as appropriate. *20180603*

On taking this document as the first cut version of any new chapter, once copied edit the new document to replace *template:* with something appropriate for the new chapter, as a single word prefix for each code block.

The chapter consists of a series of sections, each section limited to a single page.

Through this guide new R commands will be introduced. The reader is encouraged to review the command's documentation and understand what the command does. Help is obtained using the `?` command as in:

```
?read.csv
```

Documentation on a particular package can be obtained using the `help=` option of `library()`:

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

This chapter is intended to be hands on. To learn effectively the reader is encouraged to run R locally (e.g., RStudio or Emacs with ESS mode) and to replicate all commands as they appear here. Check that output is the same and it is clear how it is generated. Try some variations. Explore.

1 Packages from the R Library

20180603

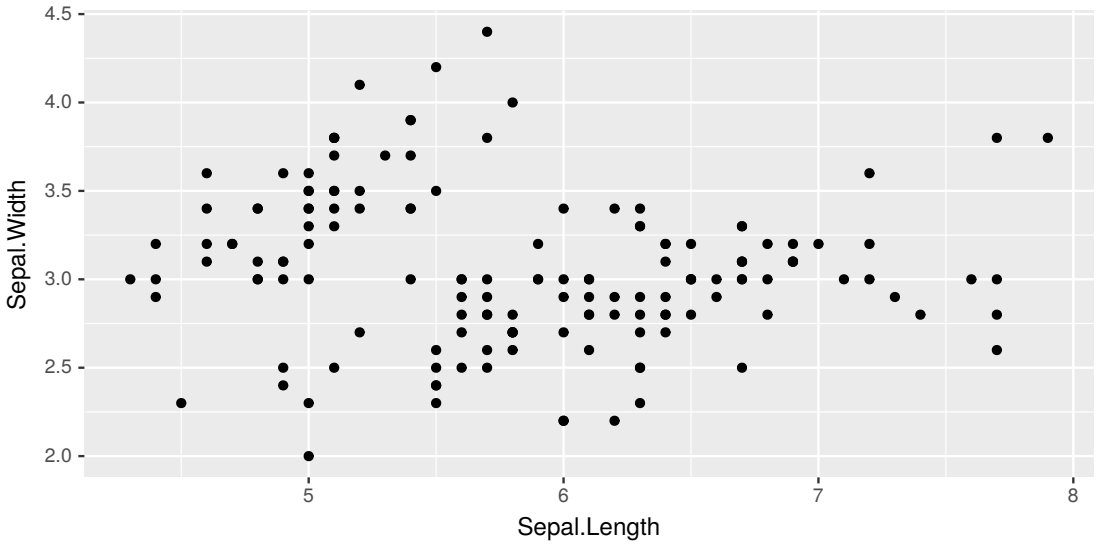
Packages used in this chapter include C50 (Kuhn and Quinlan, 2018), Hmisc (Harrell, 2018), ROCR (Sing *et al.*, 2015), ada (Culp *et al.*, 2016), caret (Kuhn *et al.*, 2018), ckanr (Chamberlain, 2015), diagram (Soetaert, 2017), directlabels (Hocking, 2018), dplyr (Wickham *et al.*, 2018), ggplot2 (Wickham and Chang, 2016), gridExtra (**?**), lubridate (Spinu *et al.*, 2018), magrittr (Bache and Wickham, 2014), nnet (Ripley, 2016), party (Hothorn *et al.*, 2018), pryr (Wickham, 2018a), randomForest (Breiman *et al.*, 2018), readr (Wickham *et al.*, 2017), rpart (Therneau and Atkinson, 2018), rvest (Wickham, 2016), scales (Wickham, 2017a), stringi (Gagolewski *et al.*, 2018), stringr (Wickham, 2018b), tibble (Müller and Wickham, 2018), tidyr (Wickham and Henry, 2018), tidyverse (Wickham, 2017b), xckd (**?**), xtable (Dahl, 2016), scales, and rattle (Williams, 2017b).

Packages are loaded here into the R session from the local library folders. Any packages that are missing can be installed using `utils::install.packages()`.

```
# Load required packages from local library into the R session.

library(C50)          # Original C5.0 decision tree builder.
library(Hmisc)        # Escape special LaTeX charaters.
library(ROCR)         # Use prediction() for evaluation.
library(ada)          # Model: Adaptive boosting ada().
library(caret)        # Unified model builder.
library(ckanr)        # Access data from CKAN.
library(diagram)      # Produce a flowchart.
library(directlabels) # Dodging labels for ggplot2.
library(dplyr)        # Wrangling: tbl_df(), group_by(), print().
library(ggplot2)      # Visualise data.
library(lubridate)    # Dates and time.
library(magrittr)     # Data pipelines: %>% %<>% %T>% equals().
library(nnet)         # Neural network model builder.
library(party)        # Conditional trees ctree() cforest().
library(pryr)         # Meta tools: object_size(), mem_used().
library(randomForest) # Model: randomForest() na.roughfix() for missing data.
library(rattle)       # Support: normVarNames(), riskchart(), weather.
library(rattle.data)  # Datasets: weatherAUS.
library(readr)        # Modern and efficient data reader.
library(readxl)       # Read Excel spreadsheets: read_excel().
library(rpart)        # Model: decision tree rpart().
library(rvest)        # Scrape data from the Internet.
library(scales)       # commas(), percent().
library(stringi)      # String concat operator: %s+%.
library(stringr)      # str_replace_all().
library(tibble)       # Table data frame: rownames_to_column(), glimpse().
library(tidyr)        # Tidy the dataset: gather().
library(tidyverse)    # Load the full tidyverse suite.
library(xkcd)         # Fancy handwritten font.
library(xtable)       # Generate LaTeX tables.
```


2 Sample Graphic



Here we illustrate the inclusion of a graphic. It may be appealing to include the actual graphic at the top of the page, followed by a narrative, and then followed by the actual code to generate the graphic. The narrative itself might incorporate a description of the code, or restrict itself to interpreting the story the graphic tells. Alternatively the description of the code may come after the code listing itself. *20180603*

```
iris %>%
  ggplot(aes(x=Sepal.Length, y=Sepal.Width)) +
  geom_point()
```


3 Command Summary

At the end of the chapter include a summary of the commands introduced in this chapter.

This chapter has introduced, demonstrated and described the following R packages, functions, commands, operators, and datasets: *20180603*

`ggplot()` *Function from ggplot2. Construct the canvas for a new plot without yet adding anything to the canvas.*

4 Exercises

Create a set of exercises to review the chapter material and develop beyond what has been covered.

20180603

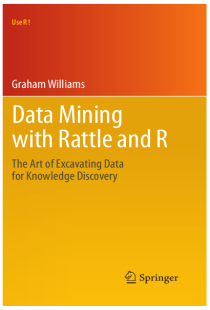
Exercise 1 A title for the table of contents.

A relatively straightforward exercise to build up some confidence.

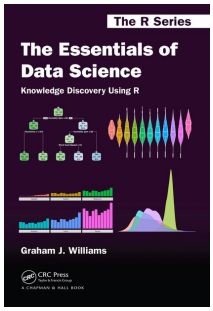
Exercise 2 Another title for the table of contents.

Each is increasingly complex and challenging, taking it slowly to get to the most difficult.

5 Further Reading and Acknowledgements



The Rattle Book (Williams, 2011), published by Springer, provides a comprehensive introduction to data mining and analytics using Rattle and R. It is available from Amazon. Rattle provides a graphical user interface through which the user is able to load, explore, visualise, and transform data, and to build, evaluate, and export models. Through its Log tab it specifically aims to provide an R template which can be exported and serve as the starting point for further programming with data in R.



The Essentials of Data Science book (Williams, 2017a), published by CRC Press, provides a comprehensive introduction to data science through programming with data using R. It is available from Amazon. The book provides a template based approach to doing data science and knowledge discovery. Templates are provided for data wrangling and model building. These serve as generic starting points for programming with data, and are designed to require minimal effort to get started. Visit `https://essentials.togaware.com` for further guides and templates.

List any other resources of use that extend beyond what we have presented in this chapter.

Other resources include:

- Under development.

6 References

Bache SM, Wickham H (2014). *magrittr: A Forward-Pipe Operator for R*. R package version 1.5, URL https://CRAN.R-project.org/package=magrittr.

Breiman L, Cutler A, Liaw A, Wiener M (2018). *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.6-14, URL https://CRAN.R-project.org/package=randomForest.

Chamberlain S (2015). *ckanr: Client for the Comprehensive Knowledge Archive Network ('CKAN') 'API'*. R package version 0.1.0, URL https://CRAN.R-project.org/package=ckanr.

Culp M, Johnson K, Michailidis G (2016). *ada: The R Package Ada for Stochastic Boosting*. R package version 2.0-5, URL https://CRAN.R-project.org/package=ada.

Dahl DB (2016). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2, URL https://CRAN.R-project.org/package=xtable.

Gagolewski M, Tartanus B, , other contributors; IBM, other contributors; Unicode, Inc (2018). *stringi: Character String Processing Facilities*. R package version 1.2.2, URL https://CRAN.R-project.org/package=stringi.

Harrell Jr FE (2018). *Hmisc: Harrell Miscellaneous*. R package version 4.1-1, URL https://CRAN.R-project.org/package=Hmisc.

Hocking TD (2018). *directlabels: Direct Labels for Multicolor Plots*. R package version 2018.05.22, URL https://CRAN.R-project.org/package=directlabels.

Hothorn T, Hornik K, Strobl C, Zeileis A (2018). *party: A Laboratory for Recursive Partytioning*. R package version 1.3-0, URL https://CRAN.R-project.org/package=party.

Kuhn M, Quinlan R (2018). *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.2, URL https://CRAN.R-project.org/package=C50.

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, the R Core Team, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C, Hunt T (2018). *caret: Classification and Regression Training*. R package version 6.0-80, URL https://CRAN.R-project.org/package=caret.

Müller K, Wickham H (2018). *tibble: Simple Data Frames*. R package version 1.4.2, URL https://CRAN.R-project.org/package=tibble.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ripley B (2016). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. R package version 7.3-12, URL https://CRAN.R-project.org/package=nnet.

Sing T, Sander O, Beerenwinkel N, Lengauer T (2015). *ROCR: Visualizing the Performance of Scoring Classifiers*. R package version 1.0-7, URL https://CRAN.R-project.org/package=ROCR.

Soetaert K (2017). *diagram: Functions for Visualising Simple Graphs (Networks), Plotting Flow Diagrams*. R package version 1.6.4, URL https://CRAN.R-project.org/package=diagram.

Spinu V, Grolemund G, Wickham H (2018). *lubridate: Make Dealing with Dates a Little Easier*. R package version 1.7.4, URL https://CRAN.R-project.org/package=lubridate.

Therneau T, Atkinson B (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13, URL https://CRAN.R-project.org/package=rpart.

Wickham H (2016). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.2, URL https://CRAN.R-project.org/package=rvest.

Wickham H (2017a). *scales: Scale Functions for Visualization*. R package version 0.5.0, URL https://CRAN.R-project.org/package=scales.

Wickham H (2017b). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1, URL https://CRAN.R-project.org/package=tidyverse.

Wickham H (2018a). *pryr: Tools for Computing on the Language*. R package version 0.1.4, URL https://CRAN.R-project.org/package=pryr.

Wickham H (2018b). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.3.1, URL https://CRAN.R-project.org/package=stringr.

Wickham H, Chang W (2016). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 2.2.1, URL https://CRAN.R-project.org/package=ggplot2.

Wickham H, François R, Henry L, Müller K (2018). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.5, URL https://CRAN.R-project.org/package=dplyr.

Wickham H, Henry L (2018). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.8.1, URL https://CRAN.R-project.org/package=tidyr.

Wickham H, Hester J, Francois R (2017). *readr: Read Rectangular Text Data*. R package version 1.1.1, URL https://CRAN.R-project.org/package=readr.

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.

Williams GJ (2017a). *The Essentials of Data Science: Knowledge discovery using R*. The R Series. CRC Press.

Williams GJ (2017b). *rattle: Graphical User Interface for Data Science in R*. R package version 5.1.0, URL https://CRAN.R-project.org/package=rattle.

This document, sourced from TemplateO.Rnw bitbucket revision 240, was processed by KnitR version 1.20 of 2018-02-20 10:11:46 UTC and took 3.6 seconds to process. It was generated by gjw on Ubuntu 18.04 LTS.

