Data Science with R

Documenting Projects with KnitR

Graham.Williams@togaware.com

Data Scientist
Australian Taxation Office

Adjunct Professor, Australian National University
Adjunct Professor, University of Canberra
Fellow, Institute of Analytics Professionals of Australia

Graham.Williams@togaware.com
http://datamining.togaware.com
Overview

1. Motivation
2. Using RStudio
3. Basic LaTeX Markup
4. Incorporating R Code
5. Formatting Tables and Plots
6. Getting Sophisticated
7. Summary
**Overview**

1. Motivation
2. Using RStudio
3. Basic LaTeX Markup
4. Incorporating R Code
5. Formatting Tables and Plots
6. Getting Sophisticated
7. Summary
**Why is Reproducibility Important?**

Your Research Leader or Executive drops by and asks:

- “Remember that research you did last year? I’ve heard there is an update on the data that you used. Can you add the new data in and repeat the same analysis?”

- “Jo Bloggs did a great analysis of the company returns data just before she left. Can you get someone else to analyse the new data set using the same methods, and so produce an updated report that we can present to the Exec next week?”

- “The fraud case you provided an analysis of last year has finally reached the courts. We need to ensure we have a clear trail of the data sources, the analyses performed, and the results obtained, to stand up in court. Could you document these please.”
Literate Data Mining Overview

- One document to intermix the analysis, code, and results
- Authors productive with narrative and code in one document
- Sweave (Leisch 2002) and now KnitR (Yihui 2011)
- Embed R code into $\LaTeX$ documents for typesetting
- KnitR also supports publishing to the web
**Why Reproducible Data Mining?**

- **Automatically** regenerate documents when code, data, or assumptions change.

- Eliminate errors that occur when transcribing results into documents.

- Record the context for the analysis and decisions made about the type of analysis to perform in the one place.

- Document the processes to provide integrity for the conclusions of the analysis.

- Share approach with others for peer review and for learning from each other—engender a continuous learning environment.
Prime Objective: Trustworthy Software

Those who receive the results of modern data analysis have limited opportunity to verify the results by direct observation. Users of the analysis have no option but to trust the analysis, and by extension the software that produced it. This places an obligation on all creators of software to program in such a way that the computations can be understood and trusted. This obligation I label the Prime Directive.

John Chambers (2008)
Software for Data Analysis: Programming with R
Beautiful Output by KnitR

\textit{KnitR combined with} \LaTeX \textit{will}

- Intermix analysis and results of analysis
- Automatically generate graphics and tables
- Support reproducible and transparent analysis
- Produce the best looking reports.
Beautiful Output by Default

The reader wants to read the document and easily do so!

- Code highlighting is done automatically
- Default theme is carefully designed
- Many other themes are available
- R Code is “properly” reformatted
- Analyses (Graphs and Tables) automatically included.
OVERVIEW

1 Motivation

2 Using RStudio

3 Basic LaTeX Markup

4 Incorporating R Code

5 Formatting Tables and Plots

6 Getting Sophisticated

7 Summary
A suite of Free and Open Source Software — FLOSS

- RStudio — Creating, managing, compiling documents
- \LaTeX — Markup language for typesetting a document
- R — Statistical analysis language
- KnitR — Integrator of typesetting and analysis
Using RStudio

- Simplified interaction with R, \LaTeX, and KnitR
- Executes R code one line at a time
- Formats \LaTeX documents and provides and spell checking
- A single click compile to PDF and synchronised views

Demonstrate: Startup and explore RStudio.
Overview

1 Motivation
2 Using RStudio
3 Basic LaTeX Markup
4 Incorporating R Code
5 Formatting Tables and Plots
6 Getting Sophisticated
7 Summary
**Introducing LaTeX**

- A text markup language rather than a WYSIWYG.
- Based on \TeX{} from 1977 — very stable and powerful.
- \LaTeX{} is easier to use macro package built on \TeX{}.
- Ensures consistent style (layout, fonts, tables, maths, etc.)
- Automatic indexes, footnotes and references.
- Documents are well structured and are clear text.
- Has a learning curve.
Basic \LaTeX\ Usage

\documentclass{article}
\begin{document}
\end{document}

Demonstrate Create a new Sweave document in RStudio
Basic LaTeX Markup

Structures

\documentclass{article}
\begin{document}
\section{Introduction}
...
\subsection{Concepts}
...
\end{document}
\documentclass{article}
\begin{document}
\begin{itemize}
\item ABC
\item DEF
\end{itemize}

This if \textbf{bold} text or \textit{italic} text, ...

\end{document}
RStudio Support for \LaTeX

RStudio provides excellent support for working with \LaTeX documents. Helps to avoid having to know too much about \LaTeX. Best illustrated through a demonstration:

- Format menu
  - Section commands
  - Font commands
  - List commands
  - Verbatim/Block commands
- Spell Checker
- Compile PDF

*Demonstrate:* Start a new document, add contents, format to PDF.
Incorporating R Code

Overview

1. Motivation
2. Using RStudio
3. Basic LaTeX Markup
4. Incorporating R Code
5. Formatting Tables and Plots
6. Getting Sophisticated
7. Summary
Incorporating R Code

- We insert R code in a *Chunk* starting with `<< >>=`
- We terminate the Chunk with `@`
- Save LaTeX with extension Rnw

```
<<simple_example>>=
x <- sum(1:10)
x
@  
```

*Produces*

```
x <- sum(1:10)
x
## [1] 55
```

- *Demonstrate:* Do this in RStudio
Incorporating R Code

Making You Look Good

<<format_example>>=
for(i in 1:5){j<-cos(sin(i)*i^2)+3;print(j-5)}
@

for(i in 1:5)
{
  j <- cos(sin(i)*i^2)+3
  print(j-5)
}

## [1] -1.334
## [1] -2.88
## [1] -1.704
## [1] -1.103
....
R Within the Text

- Include information about data within the narrative.
- We can do that with `\Sexpr{...}`.

Our dataset has \Sexpr{nrow(ds)} observations of \Sexpr{ncol(ds)} variables.

Becomes

Our dataset has 88768 observations of 24 variables.

Better Still: \Sexpr{format(nrow(ds), big.mark="","")}

Our dataset has 88,768 observations of 24 variables.
OVERVIEW

1 Motivation
2 Using RStudio
3 Basic \LaTeX\ Markup
4 Incorporating R Code
5 Formatting Tables and Plots
6 Getting Sophisticated
7 Summary
### A Simple Table

```r
library(xtable)
obs <- sample(1:nrow(weatherAUS), 8)
vars <- 2:6
xtable(weatherAUS[obs, vars])
```

<table>
<thead>
<tr>
<th>Location</th>
<th>MinTemp</th>
<th>MaxTemp</th>
<th>Rainfall</th>
<th>Evaporation</th>
</tr>
</thead>
<tbody>
<tr>
<td>27986 Wollongong</td>
<td>8.80</td>
<td>16.60</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>75317 Perth</td>
<td>17.90</td>
<td>32.90</td>
<td>0.00</td>
<td>11.60</td>
</tr>
<tr>
<td>59168 Townsville</td>
<td>23.50</td>
<td>32.70</td>
<td>0.00</td>
<td>12.60</td>
</tr>
<tr>
<td>22167 SydneyAirport</td>
<td>13.30</td>
<td>18.40</td>
<td>6.20</td>
<td>5.60</td>
</tr>
<tr>
<td>17652 Richmond</td>
<td>5.70</td>
<td>15.60</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>9385 Newcastle</td>
<td>21.80</td>
<td>29.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>33537 MountGinini</td>
<td>9.50</td>
<td>18.10</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>68294 Albany</td>
<td>8.90</td>
<td>16.30</td>
<td>2.80</td>
<td></td>
</tr>
</tbody>
</table>

http://togaware.com

Copyright © 2014, Graham.Williams@togaware.com
### Table: Exclude Row Names

```r
print(xtable(weatherAUS[obs, vars],
             include.rownames=FALSE))
```

<table>
<thead>
<tr>
<th>Location</th>
<th>MinTemp</th>
<th>MaxTemp</th>
<th>Rainfall</th>
<th>Evaporation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wollongong</td>
<td>8.80</td>
<td>16.60</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Perth</td>
<td>17.90</td>
<td>32.90</td>
<td>0.00</td>
<td>11.60</td>
</tr>
<tr>
<td>Townsville</td>
<td>23.50</td>
<td>32.70</td>
<td>0.00</td>
<td>12.60</td>
</tr>
<tr>
<td>SydneyAirport</td>
<td>13.30</td>
<td>18.40</td>
<td>6.20</td>
<td>5.60</td>
</tr>
<tr>
<td>Richmond</td>
<td>5.70</td>
<td>15.60</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Newcastle</td>
<td>21.80</td>
<td>29.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>MountGinini</td>
<td>9.50</td>
<td>18.10</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>Albany</td>
<td>8.90</td>
<td>16.30</td>
<td>2.80</td>
<td>3.60</td>
</tr>
</tbody>
</table>
**Table: Limit Number of Digits**

```r
print(xtable(weatherAUS[obs, vars],
             digits=1),
      include.rownames=FALSE)
```

<table>
<thead>
<tr>
<th>Location</th>
<th>MinTemp</th>
<th>MaxTemp</th>
<th>Rainfall</th>
<th>Evaporation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wollongong</td>
<td>8.8</td>
<td>16.6</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Perth</td>
<td>17.9</td>
<td>32.9</td>
<td>0.0</td>
<td>11.6</td>
</tr>
<tr>
<td>Townsville</td>
<td>23.5</td>
<td>32.7</td>
<td>0.0</td>
<td>12.6</td>
</tr>
<tr>
<td>SydneyAirport</td>
<td>13.3</td>
<td>18.4</td>
<td>6.2</td>
<td>5.6</td>
</tr>
<tr>
<td>Richmond</td>
<td>5.7</td>
<td>15.6</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Newcastle</td>
<td>21.8</td>
<td>29.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>MountGinini</td>
<td>9.5</td>
<td>18.1</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Albany</td>
<td>8.9</td>
<td>16.3</td>
<td>2.8</td>
<td>3.6</td>
</tr>
</tbody>
</table>
vars <- 2:8
print(xtable(weatherAUS[obs, vars],

digits=0),

size="tiny",
include.rownames=FALSE)

<table>
<thead>
<tr>
<th>Location</th>
<th>MinTemp</th>
<th>MaxTemp</th>
<th>Rainfall</th>
<th>Evaporation</th>
<th>Sunshine</th>
<th>WindGustDir</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wollongong</td>
<td>9</td>
<td>17</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td>WSW</td>
</tr>
<tr>
<td>Perth</td>
<td>18</td>
<td>33</td>
<td>0</td>
<td>13</td>
<td>13</td>
<td>SE</td>
</tr>
<tr>
<td>Townsville</td>
<td>24</td>
<td>33</td>
<td>0</td>
<td>13</td>
<td>13</td>
<td>ENE</td>
</tr>
<tr>
<td>SydneyAirport</td>
<td>13</td>
<td>18</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>SSE</td>
</tr>
<tr>
<td>Richmond</td>
<td>6</td>
<td>16</td>
<td>1</td>
<td></td>
<td></td>
<td>NE</td>
</tr>
<tr>
<td>Newcastle</td>
<td>22</td>
<td>29</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MountGinini</td>
<td>10</td>
<td>18</td>
<td>2</td>
<td></td>
<td></td>
<td>SSW</td>
</tr>
<tr>
<td>Albany</td>
<td>9</td>
<td>16</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
### Table: Column Alignment

```r
vars <- 2:8
print(xtable(weatherAUS[obs, vars],
    digits=0,
    align="rlrrrrrrr"),
    size="tiny")
```

<table>
<thead>
<tr>
<th>Location</th>
<th>MinTemp</th>
<th>MaxTemp</th>
<th>Rainfall</th>
<th>Evaporation</th>
<th>Sunshine</th>
<th>WindGustDir</th>
</tr>
</thead>
<tbody>
<tr>
<td>27986</td>
<td>Wollongong</td>
<td>9</td>
<td>17</td>
<td>0</td>
<td>12</td>
<td>WSW</td>
</tr>
<tr>
<td>75317</td>
<td>Perth</td>
<td>18</td>
<td>33</td>
<td>0</td>
<td>13</td>
<td>SE</td>
</tr>
<tr>
<td>59168</td>
<td>Townsville</td>
<td>24</td>
<td>33</td>
<td>0</td>
<td>13</td>
<td>ENE</td>
</tr>
<tr>
<td>22167</td>
<td>SydneyAirport</td>
<td>13</td>
<td>18</td>
<td>6</td>
<td>6</td>
<td>SSE</td>
</tr>
<tr>
<td>17652</td>
<td>Richmond</td>
<td>6</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>NE</td>
</tr>
<tr>
<td>9385</td>
<td>Newcastle</td>
<td>22</td>
<td>29</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33537</td>
<td>MountGinini</td>
<td>10</td>
<td>18</td>
<td>2</td>
<td></td>
<td>SSW</td>
</tr>
<tr>
<td>68294</td>
<td>Albany</td>
<td>9</td>
<td>16</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>
### Table: Caption

```r
print(xtable(weatherAUS[obs, vars],
             digits=1,
             caption="This is the table caption."),
       size="tiny")
```

<table>
<thead>
<tr>
<th>Location</th>
<th>MinTemp</th>
<th>MaxTemp</th>
<th>Rainfall</th>
<th>Evaporation</th>
<th>Sunshine</th>
<th>WindGustDir</th>
</tr>
</thead>
<tbody>
<tr>
<td>27986 Wollongong</td>
<td>8.8</td>
<td>16.6</td>
<td>0.0</td>
<td>11.6</td>
<td>11.6</td>
<td>WSW</td>
</tr>
<tr>
<td>75317 Perth</td>
<td>17.9</td>
<td>32.9</td>
<td>0.0</td>
<td>12.6</td>
<td>12.6</td>
<td>SE</td>
</tr>
<tr>
<td>59168 Townsville</td>
<td>23.5</td>
<td>32.7</td>
<td>0.0</td>
<td>12.6</td>
<td>12.6</td>
<td>ENE</td>
</tr>
<tr>
<td>22167 SydneyAirport</td>
<td>13.3</td>
<td>18.4</td>
<td>6.2</td>
<td>5.6</td>
<td>0.0</td>
<td>SSE</td>
</tr>
<tr>
<td>17652 Richmond</td>
<td>5.7</td>
<td>15.6</td>
<td>1.0</td>
<td></td>
<td></td>
<td>NE</td>
</tr>
<tr>
<td>9385 Newcastle</td>
<td>21.8</td>
<td>29.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33537 MountGinini</td>
<td>9.5</td>
<td>18.1</td>
<td>2.0</td>
<td></td>
<td></td>
<td>SSW</td>
</tr>
<tr>
<td>68294 Albany</td>
<td>8.9</td>
<td>16.3</td>
<td>2.8</td>
<td>3.6</td>
<td>5.5</td>
<td></td>
</tr>
</tbody>
</table>

### Table: This is the table caption.
library(ggplot2)
cities <- c("Canberra", "Darwin", "Melbourne", "Sydney")
ds <- subset(weatherAUS, Location %in% cities & ! is.na(Temp3pm))
g <- ggplot(ds, aes(Temp3pm, colour=Location, fill=Location))
g <- g + geom_density(alpha = 0.55)
print(g)
Getting Sophisticated

Overview

1 Motivation
2 Using RStudio
3 Basic LaTeX Markup
4 Incorporating R Code
5 Formatting Tables and Plots
6 Getting Sophisticated
7 Summary
Demonstration
**Actual Examples**

- Linked Risk Visualisations
- Visualising Clusters
- Siebel Case Profile Attachments
- Specifications of Rule-Based Model Logic
### OVERVIEW

1. **Motivation**
2. **Using RStudio**
3. **Basic \LaTeX{} Markup**
4. **Incorporating R Code**
5. **Formatting Tables and Plots**
6. **Getting Sophisticated**
7. **Summary**
Summary

Document as we go to record all modelling activity

Ensure transparency, repeatability, sharing

Mature technology: \LaTeX{} and R

Modern support: KnitR and RStudio
Further Reading

- http://onepager.togaware.com/
- http://yihui.name/knitr/
- http://www.rstudio.org/
- http://bcb.dfci.harvard.edu/~aedin/courses/ReproducibleResearch/ReproducibleResearch.pdf
Further Reading

- http://onepager.togaware.com/
- http://yihui.name/knitr/
- http://www.rstudio.org/
- http://bcb.dfci.harvard.edu/~aedin/courses/ReproducibleResearch/ReproducibleResearch.pdf