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

Overview

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

KnitR combined with LaTeX 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

Supporting Technology

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

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

Structures

```
\documentclass{article}

\begin{document}

\section{Introduction}

...

\subsection{Concepts}

...

\end{document}
```

Formats

```
\documentclass{article}

\begin{document}

\begin{itemize}
  \item ABC
  \item DEF
\end{itemize}

This if \textbf{bold} text or \textbf{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 abuot 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.

Overview

Incorporating R Code

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

  This Chunk

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

  Produces

  ```
  x <- sum(1:10)
  x
  ```

  ```
  ## [1] 55
  ```

- *Demonstrate:* Do this in RStudio

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

A Simple Table

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

	Location	MinTemp	MaxTemp	Rainfall	Evaporation
27986	Wollongong	8.80	16.60	0.00	
75317	Perth	17.90	32.90	0.00	11.60
59168	Townsville	23.50	32.70	0.00	12.60
22167	SydneyAirport	13.30	18.40	6.20	5.60
17652	Richmond	5.70	15.60	1.00	
9385	Newcastle	21.80	29.00	0.00	
33537	MountGinini	9.50	18.10	2.00	
68294	Albany	8.90	16.30	2.80	3.60

Table: Exclude Row Names

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

Location	MinTemp	MaxTemp	Rainfall	Evaporation
Wollongong	8.80	16.60	0.00	
Perth	17.90	32.90	0.00	11.60
Townsville	23.50	32.70	0.00	12.60
SydneyAirport	13.30	18.40	6.20	5.60
Richmond	5.70	15.60	1.00	
Newcastle	21.80	29.00	0.00	
MountGinini	9.50	18.10	2.00	
Albany	8.90	16.30	2.80	3.60

Table: Limit Number of Digits

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

Location	MinTemp	MaxTemp	Rainfall	Evaporation
Wollongong	8.8	16.6	0.0	
Perth	17.9	32.9	0.0	11.6
Townsville	23.5	32.7	0.0	12.6
SydneyAirport	13.3	18.4	6.2	5.6
Richmond	5.7	15.6	1.0	
Newcastle	21.8	29.0	0.0	
MountGinini	9.5	18.1	2.0	
Albany	8.9	16.3	2.8	3.6

Table: Tiny Font

```
vars <- 2:8
print(xtable(weatherAUS[obs, vars],
             digits=0),
      size="tiny",
      include.rownames=FALSE)
```

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir
Wollongong	9	17	0			WSW
Perth	18	33	0	12	12	SE
Townsville	24	33	0	13	13	ENE
SydneyAirport	13	18	6	6	0	SSE
Richmond	6	16	1			NE
Newcastle	22	29	0			
MountGinini	10	18	2			SSW
Albany	9	16	3	4	6	

Table: Column Alignment

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

	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir
27986	Wollongong	9	17	0			WSW
75317	Perth	18	33	0	12	12	SE
59168	Townsville	24	33	0	13	13	ENE
22167	SydneyAirport	13	18	6	6	0	SSE
17652	Richmond	6	16	1			NE
9385	Newcastle	22	29	0			
33537	MountGinini	10	18	2			SSW
68294	Albany	9	16	3	4	6	

Table: Caption

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

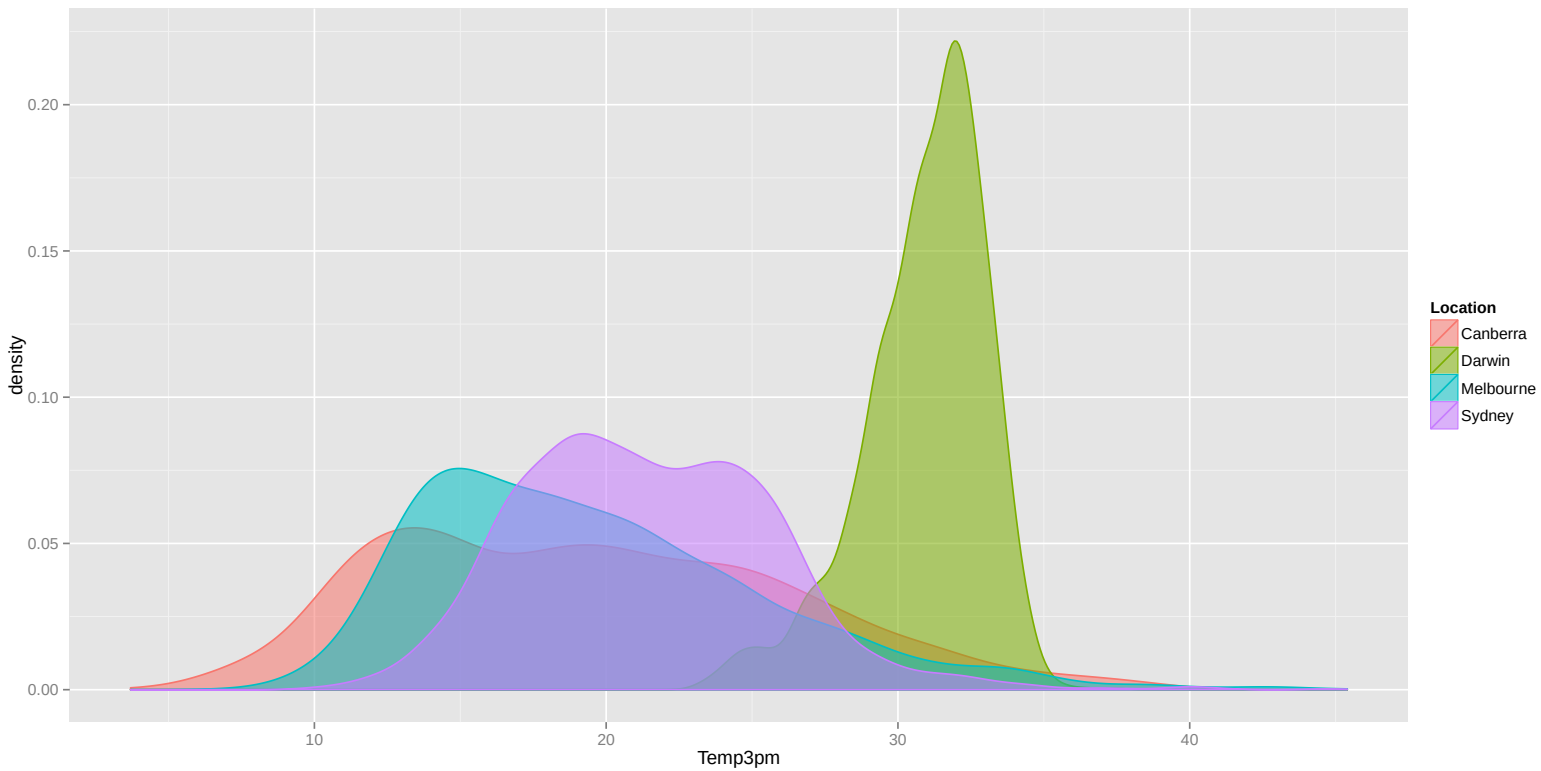
	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir
27986	Wollongong	8.8	16.6	0.0			WSW
75317	Perth	17.9	32.9	0.0	11.6	11.6	SE
59168	Townsville	23.5	32.7	0.0	12.6	12.6	ENE
22167	SydneyAirport	13.3	18.4	6.2	5.6	0.0	SSE
17652	Richmond	5.7	15.6	1.0			NE
9385	Newcastle	21.8	29.0	0.0			
33537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

Table : This is the table caption.

Overview

Plots

```
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)
```



Overview

Advanced Topic—KnitR and ESS and Emacs

Demonstration

Actual Examples

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

Overview

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://yihui.name/slides/2012-knitr-RStudio.html
- http://bcb.dfci.harvard.edu/~aedin/courses/ReproducibleResearch/ReproducibleResearch.pdf
- https://dl.dropbox.com/u/233041/Bios301/lecture2_knitr.html
- http://biostat.mc.vanderbilt.edu/wiki/pub/Main/ReproducibleResearchTutorial/HarrellScottTutorial-useR2012.pdf

Further Reading

- http://onepager.togaware.com/
- http://yihui.name/knitr/
- http://www.rstudio.org/
- http://yihui.name/slides/2012-knitr-RStudio.html
- http://bcb.dfci.harvard.edu/~aedin/courses/ReproducibleResearch/ReproducibleResearch.pdf
- https://dl.dropbox.com/u/233041/Bios301/lecture2_knitr.html
- http://biostat.mc.vanderbilt.edu/wiki/pub/Main/ReproducibleResearchTutorial/HarrellScottTutorial-useR2012.pdf