Data Analytics and Business Intelligence
(8696/8697)

Introducing Data Science with Rattle and R

Graham.Williams@togaware.com

Chief 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. An Introduction to Data Mining
2. Why Choose R for Data Science?
3. The Rattle Package for Data Mining
4. Moving Into R
5. Knitting
An Introduction to Data Mining

Why Choose R for Data Science?

The Rattle Package for Data Mining

Moving Into R

Knitting
Data Mining

A data driven analysis to uncover otherwise unknown but useful patterns in large datasets, to discover new knowledge and to develop predictive models, turning data and information into knowledge and (one day perhaps) wisdom, in a timely manner.
Data Mining

- Application of
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - Effective Communications and Intuition

- ...to Datasets that vary by Volume, Velocity, Variety, Value, Veracity

- ...to discover new knowledge
- ...to improve business outcomes
- ...to deliver better tailored services
Data Mining in Research

- **Health Research**
  Adverse reactions using linked Pharmaceutical, General Practitioner, Hospital, Pathology datasets.

- **Astronomy**
  Microlensing events in the Large Magellanic Cloud of several million observed stars (out of 10 billion).

- **Psychology**
  Investigation of age-of-onset for Alzheimer’s disease from 75 variables for 800 people.

- **Social Sciences**
  Survey evaluation. Social network analysis - identifying key influencers.
Data Mining in Government

- **Australian Taxation Office**
  - Lodgment ($110M)
  - Tax Havens ($150M)
  - Tax Fraud ($250M)

- **Immigration and Border Control**
  - Check passengers before boarding

- **Health and Human Services**
  - Doctor shoppers
  - Over servicing
The Business of Data Mining

- SAS has annual revenues of $3B (2013)
- IBM bought SPSS for $1.2B (2009)
- Analytics is >$100B business and >$320B by 2020
- Amazon, eBay/PayPal, Google, Facebook, LinkedIn, . . .
- Shortage of 180,000 data scientists in US in 2018 (McKinsey) . . .
Basic Data Mining Algorithms

- Cluster Analysis (kmeans, wskm)
- Association Analysis (arules)
- Linear Discriminant Analysis (lda)
- Logistic Regression (glm)
- Decision Trees (rpart, wsrpart)
- Random Forests (randomForest, wsr)
- Boosted Stumps (ada)
- Neural Networks (nnet)
- Support Vector Machines (kernlab)
- ...

That’s a lot of tools to learn in R!  
Many with different interfaces and options.
Overview

1. An Introduction to Data Mining
2. Why Choose R for Data Science?
3. The Rattle Package for Data Mining
4. Moving Into R
5. Knitting
Installing R and Rattle

- **First task is to install R**

  As free/libre open source software (FLOSS or FOSS), R and Rattle are available to all, with no limitations on our freedom to use and share the software, except to share and share alike.

- Visit CRAN at [http://cran.rstudio.com](http://cran.rstudio.com)
- Visit Rattle at [http://rattle.togaware.com](http://rattle.togaware.com)

- Linux: Install packages (Ubuntu is recommended)
  
  ```
  $ wajig install r-recommended r-cran-rattle
  ```

- Windows: Download and install from CRAN
- MacOSX: Download and install from CRAN
Why do Data Science with R?

- Most widely used Data Mining and Machine Learning Package
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - But not the nicest of languages for a Computer Scientist!

- Free (Libre) Open Source Statistical Software
  - ... all modern statistical approaches
  - ... many/most machine learning algorithms
  - ... opportunity to readily add new algorithms

That is important for us in the research community
Get our algorithms out there and being used—impact!!!
Why do Data Science with R?

- Most widely used Data Mining and Machine Learning Package
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - But not the nicest of languages for a Computer Scientist!

- Free (Libre) Open Source Statistical Software
  - ... all modern statistical approaches
  - ... many/most machine learning algorithms
  - ... opportunity to readily add new algorithms

That is important for us in the research community
Get our algorithms out there and being used—impact!!!
Why do Data Science with R?

- Most widely used Data Mining and Machine Learning Package
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - But not the nicest of languages for a Computer Scientist!

- Free (Libre) Open Source Statistical Software
  - . . . all modern statistical approaches
  - . . . many/most machine learning algorithms
  - . . . opportunity to readily add new algorithms

- That is important for us in the research community
  Get our algorithms out there and being used—impact!!!
Why do Data Science with R?

- Most widely used Data Mining and Machine Learning Package
  - Machine Learning
  - Statistics
  - Software Engineering and Programming with Data
  - But not the nicest of languages for a Computer Scientist!

- Free (Libre) Open Source Statistical Software
  - ...all modern statistical approaches
  - ...many/most machine learning algorithms
  - ...opportunity to readily add new algorithms

- That is important for us in the research community
  Get our algorithms out there and being used—impact!!!
“I think it addresses a niche market for high-end data analysts that want free, readily available code. We have customers who build engines for aircraft. I am happy they are not using freeware when I get on a jet.” Anne H. Milley, director of technology product marketing at SAS (New York Times, 7 January 2009).

It’s interesting that SAS Institute feels that non-peer-reviewed software with hidden implementations of analytic methods that cannot be reproduced by others should be trusted when building aircraft engines. (Frank Harrell)
Monthly email traffic on software’s main discussion list.
How Popular is R? Discussion Topics

Number of discussions on popular QandA forums 2013.

Source: http://r4stats.com/articles/popularity/
How Popular is R? R versus SAS

Number of R/SAS related posts to Stack Overflow by week.

Source: http://r4stats.com/articles/popularity/
HOW POPULAR IS R? PROFESSIONAL FORUMS

Registered for the main discussion group for each software.

Source: http://r4stats.com/articles/popularity/
Why Choose R for Data Science?  Popularity of R?

How Popular is R? Used in Analytics Competitions

Software used in data analysis competitions in 2011.

Source: http://r4stats.com/articles/popularity/
How Popular is R? User Survey

Rexer Analytics Survey 2010 results for data mining/analytic tools.

Source: http://r4stats.com/articles/popularity/
Why Choose R for Data Science?

R Skills Attract Good Salaries

- 2014 survey of average US tech salaries by Dice Tech puts R at the top of the list at $115,531
  
  ComputerWorld

- 2013 O’Reilly Strata Conference: Data Scientists use R over other data programming languages and Data Scientists using Open Source earn $130,000 on average.
  
  Revolution Analytics
**What is R?**

R — The Video

A 90 Second Promo from Revolution Analytics

http://www.revolutionanalytics.com/what-is-open-source-r/
Choosing R for Data Science

- Data Science is about Analysing Data;
- R is freely available to all to analyse data;
- R has the most extensive suite of functionality available;
- Nothing else is any longer even close.
OVERVIEW

1. An Introduction to Data Mining
2. Why Choose R for Data Science?
3. The Rattle Package for Data Mining
4. Moving Into R
5. Knitting
Why a GUI for Data Science in R?

- Statistics can be complex and traps await
- **So many** tools in R to deliver insights
- Effective analyses should be scripted
- Scripting also required for repeatability
- R is a language for **programming** with data

How to remember how to do all of this in R?
How to skill up 150 data analysts with Data Mining?
Users of Rattle

Today, Rattle is used world wide in many industries

- Health analytics
- Customer segmentation and marketing
- Fraud detection
- Government

It is used by

- Universities to teach Data Mining
- Within research projects for basic analyses
- Consultants and Analytics Teams across business

It is and will remain freely available.

CRAN and http://rattle.togaware.com
**Installation**

- Rattle is built using R
- Need to download and install R from cran.r-project.org
- Recommend also install RStudio from www.rstudio.org

- Then start up RStudio and install Rattle:
  ```r
  install.packages("rattle")
  ```

- Then we can start up Rattle:
  ```r
  rattle()
  ```

- Required packages are loaded as needed.
A Tour Thru Rattle: Startup

Welcome to Rattle (rattle.togaware.com).

Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for statistical computing and graphics. Together they provide a sophisticated environments for data mining, statistical analyses, and data visualisation.

See the Help menu for extensive support in using Rattle. The Togaware Desktop Data Mining Survival Guide includes Rattle documentation and is available from datamining.togaware.com.

Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See Help -> About for details.

Rattle Version 2.6.27 r108. Copyright 2006-2013 Togaware Pty Ltd
Rattle is a registered trademark of Togaware Pty Ltd

To Begin: Choose the data source, specify the details, then click the Execute button.
A Tour Thru Rattle: Loading Data
### Pearson's product-moment correlation

<table>
<thead>
<tr>
<th></th>
<th>MaxTemp</th>
<th>Rainfall</th>
<th>WindSpeed9am</th>
<th>Pressure9am</th>
<th>Temp9am</th>
<th>Temp3pm</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaxTemp</td>
<td>-.09</td>
<td></td>
<td>-.15</td>
<td>.27</td>
<td>.87</td>
<td>.99</td>
</tr>
<tr>
<td>Rainfall</td>
<td>.17</td>
<td>.31</td>
<td>.08</td>
<td>-.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WindSpeed9am</td>
<td>-.40</td>
<td>.14</td>
<td>-.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pressure9am</td>
<td>-.46</td>
<td>.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temp9am</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temp3pm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Save**  **Print**  **Close**
A Tour Thru Rattle: Explore Correlations
A Tour Thru Rattle: Hierarchical Cluster
A Tour Thru Rattle: Decision Tree

Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 256

node), split, n, loss, yval, (yprob)
   * denotes terminal node

1) root 256 41 No (0.83984375 0.16015625)
   2) Pressure3pm< 1011.9 204 16 No (0.92156863 0.07843137)
      4) Cloud3pm< 7.5 195 10 No (0.94871795 0.05128205) *
      5) Cloud3pm>=7.5 9 3 Yes (0.33333333 0.66666667) *
   3) Pressure3pm< 1011.9 52 25 No (0.51923077 0.48076923)
      6) Sunshine>=8.85 25 5 No (0.80000000 0.20000000) *
      7) Sunshine< 8.85 27 7 Yes (0.25925926 0.74074074) *

Classification tree:
rpart(formula = RainTomorrow ~ ., data = crs$dataset[crs$strain,
    c(crs$input, crs$target)], method = "class", parms = list(split = "information"),
    control = rpart.control(use surrogate = 0, max surrogate = 0))

Variables actually used in tree construction:
   [1] Cloud3pm  Pressure3pm Sunshine

Root node error: 41/256 = 0.161616

The Decision Tree model has been built. Time taken: 0.09 secs
A Tour Thru Rattle: Decision Tree Plot
A Tour Thru Rattle: Random Forest

Call:
randomForest(formula = RainTomorrow ~ ..,
data = crs$dataset[crs$sample, c(crs$input, crs$target)],
nTree = 500, mtry = 4, importance = TRUE, replace = FALSE, na.action = na.roughfix)

Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 4

OOB estimate of error rate: 13.28%

Confusion matrix:

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>207</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>26</td>
</tr>
</tbody>
</table>

Analysis of the Area Under the Curve (AUC)

The Random Forest model has been built. Time taken: 0.87 secs
A Tour Thru Rattle: Risk Chart

Risk Chart Random Forest weather.csv [test] RainTomorrow

- Risk Scores
- Risk Chart
- Performance (%)
- Caseload (%)
- Lift
- RainTomorrow (92%)
- Rain in MM (97%)
- Precision

Risk Chart

<table>
<thead>
<tr>
<th>Risk Scores</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caseload (%)</td>
<td>0</td>
<td>20</td>
<td>40</td>
<td>60</td>
<td>80</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

0.1 0.2 0.3 0.4 0.5 0.6 0.7
Risk Scores

Lift
4
3
2
1
22%

Copyright © 2014, Graham.Williams@togaware.com
Rattle Interface Notes

- Work through the tabs from left to right
- After setting up a tab we need to Execute it
- Projects save the current Rattle state
- Projects can be restored at a later time
OVERVIEW

1. An Introduction to Data Mining
2. Why Choose R for Data Science?
3. The Rattle Package for Data Mining
4. Moving Into R
5. Knitting
Data Scientists are Programmers of Data

But...  
- Data scientists are programmers of data 
- A GUI can only do so much 
- R is a powerful statistical language

Data Scientists Desire...  
- Scripting 
- Transparency 
- Repeatability 
- Sharing
FROM GUI TO CLI — RATTLE’S LOG TAB

# Rattle is Copyright (c) 2006-2013 Togaware Pty Ltd.
#
# Rattle timestamp: 2013-05-13 16:49:53 x86_64-pc-linux-gnu
# Rattle version 2.6.27 user 'gjw'

# Export this log textview to a file using the Export button or the Tools
# menu to save a log of all activity. This facilitates repeatability. Exporting
# to file 'myrf01.R', for example, allows us to the type in the R Console
# the command source('myrf01.R') to repeat the process automatically.
# Generally, we may want to edit the file to suit our needs. We can also directly
# edit this current log textview to record additional information before exporting.

# Saving and loading projects also retains this log.

library(rattle)

# This log generally records the process of building a model. However, with very
# little effort the log can be used to score a new dataset. The logical variable
# 'building' is used to toggle between generating transformations, as when building
# a model, and simply using the transformations, as when scoring a dataset.

building <- TRUE
score <- !building

# The colorspace package is used to generate the colours used in plots, if available.

library(colorspace)
FROM GUI TO CLI — RATTLE’S LOG TAB

```r
# Rattle timestamp: 2013-05-13 17:35:07 x86_64-pc-linux-gnu

# Random Forest

# The 'randomForest' package provides the 'randomForest' function.
require(randomForest, quietly=TRUE)

# Build the Random Forest model.
set.seed(crs$seed)
crs$rf <- randomForest(RainTomorrow ~ .,
data=crs$dataset[crs$sample, c(crs$input, crs$target)],
nTree=500,
mtry=4,
importance=TRUE,
na.action=na.roughfix,
replace=FALSE)

# Generate textual output of 'Random Forest' model.
crs$rf

# The 'pROC' package implements various AUC functions.
require(pROC, quietly=TRUE)

# Calculate the Area Under the Curve (AUC).
```

http://togaware.com
Step 1: Load the Dataset

dname <- "weather"
d <- get(dname)
dim(d)

## [1] 366 24

names(d)

## [1] "Date" "Location" "MinTemp" "...
## [5] "Rainfall" "Evaporation" "Sunshine" "...
## [9] "WindGustSpeed" "WindDir9am" "WindDir3pm" "...
## [13] "WindSpeed3pm" "Humidity9am" "Humidity3pm" "...
....
**Step 2: Observe the Data — Observations**

```r
head(ds)
```

```
## Date Location MinTemp MaxTemp Rainfall Evapot...  
## 1 2007-11-01 Canberra 8.0 24.3 0.0 ...  
## 2 2007-11-02 Canberra 14.0 26.9 3.6 ...  
## 3 2007-11-03 Canberra 13.7 23.4 3.6 ...  
``` ...

```r
tail(ds)
```

```
## Date Location MinTemp MaxTemp Rainfall Evapot...  
## 361 2008-10-26 Canberra 7.9 26.1 0 ...  
## 362 2008-10-27 Canberra 9.0 30.7 0 ...  
## 363 2008-10-28 Canberra 7.1 28.4 0 ...  
``` ...

---

http://togaware.com

Copyright © 2014, Graham.Williams@togaware.com
**Step 2: Observe the Data — Structure**

```r
str(ds)
```

```
# 'data.frame': 366 obs. of 24 variables:
# $ Date     : Date, format: "2007-11-01" "2007-11-...
# $ Location : Factor w/ 49 levels "Adelaide","Alba...
# $ MinTemp  : num 8 14 13.7 13.3 7.6 6.2 6.1 8.3  ...
# $ MaxTemp  : num 24.3 26.9 23.4 15.5 16.1 16.9 1...
# $ Rainfall : num 0 3.6 3.6 39.8 2.8 0 0.2 0 0 16...
# $ Evaporation : num 3.4 4.4 5.8 7.2 5.6 5.8 4.2 5.6...
# $ Sunshine : num 6.3 9.7 3.3 9.1 10.6 8.2 8.4 4....
# $ WindGustDir : Ord.factor w/ 16 levels "N"<"NNE"<"N...
# $ WindGustSpeed: num 30 39 85 54 50 44 43 41 48 31  
# $ WindDir9am : Ord.factor w/ 16 levels "N"<"NNE"<"N...
# $ WindDir3pm : Ord.factor w/ 16 levels "N"<"NNE"<"N...
```
# Step 2: Observe the Data — Summary

summary(ds)

<table>
<thead>
<tr>
<th>Date</th>
<th>Location</th>
<th>MinTemp</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min: 2007-11-01</td>
<td>Canberra</td>
<td>-5.3...</td>
<td></td>
</tr>
<tr>
<td>1st Qu.: 2008-01-31</td>
<td>Adelaide</td>
<td>2.3...</td>
<td></td>
</tr>
<tr>
<td>Median: 2008-05-01</td>
<td>Albany</td>
<td>7.4...</td>
<td></td>
</tr>
<tr>
<td>Mean: 2008-05-01</td>
<td>Albury</td>
<td>7.2...</td>
<td></td>
</tr>
<tr>
<td>3rd Qu.: 2008-07-31</td>
<td>AliceSprings</td>
<td>12.5...</td>
<td></td>
</tr>
<tr>
<td>Max: 2008-10-31</td>
<td>BadgerysCreek</td>
<td>20.9...</td>
<td></td>
</tr>
<tr>
<td>(Other)</td>
<td></td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rainfall</th>
<th>Evaporation</th>
<th>Sunshine</th>
<th>Wind...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min: 0.00</td>
<td>Min: 0.20</td>
<td>Min: 0.00</td>
<td>NW ...</td>
</tr>
<tr>
<td>1st Qu.: 0.00</td>
<td>1st Qu.: 2.20</td>
<td>1st Qu.: 5.95</td>
<td>NNW ...</td>
</tr>
<tr>
<td>Median: 0.00</td>
<td>Median: 4.20</td>
<td>Median: 8.60</td>
<td>E ...</td>
</tr>
</tbody>
</table>

...
Step 2: Observe the Data — Variables

```
id <- c("Date", "Location")
target <- "RainTomorrow"
risk <- "RISK_MM"
(ignore <- union(id, risk))

## [1] "Date" "Location" "RISK_MM"

(vars <- setdiff(names(ds), ignore))

## [1] "MinTemp"  "MaxTemp"  "Rainfall"  "..."
## [5] "Sunshine" "WindGustDir" "WindGustSpeed" "..."
## [9] "WindDir3pm" "WindSpeed9am" "WindSpeed3pm" "..."
## [13] "Humidity3pm" "Pressure9am" "Pressure3pm" "...

....
Step 3: Clean the Data — Remove Missing

```r
dim(ds)
## [1] 366 24

sum(is.na(ds[vars]))
## [1] 47

ds <- ds[-attr(na.omit(ds[vars]), "na.action")]
```
Step 3: Clean the Data — Remove Missing

```r
dim(ds)
## [1] 328 24

sum(is.na(ds[vars]))
## [1] 0
```
**Step 3: Clean the Data—Target as Categoric**

```r
summary(ds[[target]])

## RainTomorrow
##    Min.   1st Qu.    Median      Mean   3rd Qu.    Max.
##   0.0000    0.0000    0.0000    0.1830    0.0000    1.0000

ds[[target]] <- as.factor(ds[[target]])
levels(ds[[target]]) <- c("No", "Yes")
```
Step 3: Clean the Data—Target as Categoric

```r
summary(ds[target])
```

```r
table(ds$target)
```

## RainTomorrow
## 0: 268
## 1:  60

```

![Bar chart showing the distribution of RainTomorrow](http://togaware.com)

Copyright © 2014, Graham.Williams@togaware.com

http://togaware.com
Step 4: Prepare for Modelling

(form <- formula(paste(target, "~ .")))

## RainTomorrow ~ .

(nobs <- nrow(ds))

## [1] 328

train <- sample(nobs, 0.70*nobs)
length(train)

## [1] 229

test <- setdiff(1:nobs, train)
length(test)

## [1] 99
library(randomForest)
model <- randomForest(form, ds[train, vars], na.action=na.omit)
model

##
## Call:
## randomForest(formula=form, data=ds[train, vars], ... 
## Type of random forest: classification 
## Number of trees: 500 
## No. of variables tried at each split: 4 
## ....
Step 6: Evaluate the Model—Risk Chart

```r
pr <- predict(model, ds[test,], type="prob")[,2]
riskchart(pr, ds[test, target], ds[test, risk],
    title="Random Forest – Risk Chart",
    risk=risk, recall=target, thresholds=c(0.35, 0.15))
```

Random Forest – Risk Chart

- Risk Scores
- Lift
- Caseload (%)
- Performance (%)
- RainTomorrow (98%)
- RISK_MM (97%)
- Precision

Copyright © 2014, Graham.Williams@togaware.com
Tools

- Ubuntu GNU/Linux operating system
  - Feature rich toolkit, up-to-date, easy to install, FLOSS

- RStudio
  - Easy to use integrated development environment, FLOSS
  - Powerful alternative is Emacs (Speaks Statistics), FLOSS

- R Statistical Software Language
  - Extensive, powerful, thousands of contributors, FLOSS

- KnitR and \LaTeX
  - Produce beautiful documents, easily reproducible, FLOSS
Moving Into R

Using Ubuntu

- Desktop Operating System (GNU/Linux)
- Replacing Windows and OSX

- The GNU Tool Suite based on Unix — significant heritage
- Multiple specialised single task tools, working well together
- Compared to single application trying to do it all
- Powerful data processing from the command line: grep, awk, head, tail, wc, sed, perl, python, most, diff, make, paste, join, patch, . . .

- For interacting with R — start up RStudio from the Dash
**RStudio—The Default Three Panels**

[Image of RStudio interface]

R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
RStudio—With R Script File—Editor Panel

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

http://togaware.com

Copyright © 2014, Graham.Williams@togaware.com
Scatterplot—R Code

Our first little bit of R code:

- Load a couple of packages into the R library

```
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function
```

- Then produce a quick plot using qplot()

```
ds <- weather
qplot(MinTemp, MaxTemp, data=ds)
```

- Your turn: give it a go.
SCATTERPLOT—R CODE

Our first little bit of R code:

- Load a couple of packages into the R library

```r
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function
```

- Then produce a quick plot using qplot()

```r
d <- weather
qplot(MinTemp, MaxTemp, data=d)
```

- Your turn: give it a go.
SCATTERPLOT—PLOT
**Scatterplot—RStudio**

```
library(rattle)  # Provides the weather dataset
library(ggplot2)  # Provides the qplot() function

qplot(MinTemp, MaxTemp, data=weather)
```

R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type ‘license()’ or ‘licence()’ for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors. Type ‘contributors()’ for more information and ‘citation()’ on how to cite R or R packages in publications.

Type ‘demo()’ for some demos, ‘help()’ for on-line help, or ‘help.start()’ for an HTML browser interface to help. Type ‘q()’ to quit R.

> library(rattle)  # Provides the weather dataset
Rattle: A free graphical interface for data mining with R.
Version 2.6.27-142 Copyright (c) 2006-2013 Togaware Pty Ltd.
Type ‘rattle()’ to shake, rattle, and roll your data.
> library(ggplot2)  # Provides the qplot() function
> qplot(MinTemp, MaxTemp, data=weather)
```
Moving Into R
Installing Packages

Missing Packages→Tools→Install Packages...

```r
library(rattle) # Provides the weather dataset
library(ggplot2) # Provides the qplot() function

qplot(MinTemp, MaxTemp, data=weather)
```
RSudio—Installing ggplot2

```
> install.packages("ggplot2")

Installing package into 'C:/Users/username/R/site-library' (as 'lib' is unspecified)
trying URL 'http://cran.r-project.org/src/contrib/ggplot2_0.9.3.1.tar.gz'
Content type 'application/x-gzip' length 2330942 bytes (2.2 Mb)
opened URL

Installing ‘ggplot2’ ...
** package ‘ggplot2’ successfully unpacked and MD5 sums checked
** R
** data
*** moving datasets to lazyload DB
** inst
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (ggplot2)
```

The downloaded source packages are in
"C:/Users/username/R/Rtmpклонор/downloaded_packages"
RStudio— Keyboard Shortcuts

These will become very useful!

- **Editor:**
  - Ctrl-Enter will send the line of code to the R console
  - Ctrl-2 will move the cursor to the Console

- **Console:**
  - UpArrow will cycle through previous commands
  - Ctrl-UpArrow will search previous commands
  - Tab will complete function names and list the arguments
  - Ctrl-1 will move the cursor to the Editor

Your turn: try them out.
RSudio—Keyboard Shortcuts

These will become very useful!

- Editor:
  - Ctrl-Enter will send the line of code to the R console
  - Ctrl-2 will move the cursor to the Console

- Console:
  - UpArrow will cycle through previous commands
  - Ctrl-UpArrow will search previous commands
  - Tab will complete function names and list the arguments
  - Ctrl-1 will move the cursor to the Editor

Your turn: try them out.
library(rattle)  # Load the weather dataset.
head(weather)  # First 6 observations of the dataset.

## Date Location MinTemp MaxTemp Rainfall Evapora...
## 1 2007-11-01 Canberra  8.0  24.3  0.0   ... 
## 2 2007-11-02 Canberra 14.0  26.9  3.6   ... 
## 3 2007-11-03 Canberra 13.7  23.4  3.6   ... 
...

str(weather)  # Structure of the variables in the dataset.

## 'data.frame': 366 obs. of  24 variables:
## $ Date : Date, format: "2007-11-01" "2007-11-... 
## $ Location : Factor w/ 49 levels "Adelaide","Alba...
## $ MinTemp  : num 8 14 13.7 13.3 7.6 6.2 6.1 8.3 ...
## ....
**Basic R**

```r
summary(weather)  # Univariate summary of the variables.
```

<table>
<thead>
<tr>
<th></th>
<th>Date</th>
<th>Location</th>
<th>MinTemp</th>
<th>Evaporation</th>
<th>Sunshine</th>
<th>WindGust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>2007-11-01</td>
<td>Canberra</td>
<td>Min.</td>
<td>2007-11-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Qu.</td>
<td>2008-01-31</td>
<td>Adelaide</td>
<td>1st Qu.</td>
<td>2008-01-31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>2008-05-01</td>
<td>Albany</td>
<td>Median</td>
<td>2008-05-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2008-05-01</td>
<td>Albury</td>
<td>Mean</td>
<td>2008-05-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>2008-07-31</td>
<td>AliceSprings</td>
<td>3rd Qu.</td>
<td>2008-07-31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max.</td>
<td>2008-10-31</td>
<td>BadgerysCreek</td>
<td>Max.</td>
<td>2008-10-31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Other)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rainfall</td>
<td>0.00</td>
<td></td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.00</td>
<td></td>
<td>2.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>0.00</td>
<td></td>
<td>4.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.43</td>
<td></td>
<td>4.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>0.20</td>
<td></td>
<td>6.40</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...

http://togaware.com

Copyright © 2014, Graham.Williams@togaware.com
Visual Summaries—Add A Little Colour

qplot(Humidity3pm, Pressure3pm, colour=RainTomorrow, data=ds)
**Visual Summaries—Careful with Categoricals**

```r
qplot(WindGustDir, Pressure3pm, data=ds)
```
Visual Summaries—Add A Little Jitter

qplot(WindGustDir, Pressure3pm, data=ds, geom="jitter")
**Visual Summaries—And Some Colour**

\[
\texttt{qplot(WindGustDir, Pressure3pm, data=ds, colour=WindGustDir, geom="jitter")}
\]
Getting Help—Precede Command with ?
OVERVIEW

1. An Introduction to Data Mining
2. Why Choose R for Data Science?
3. The Rattle Package for Data Mining
4. Moving Into R
5. Knitting
CREATE A KnitR DOCUMENT: NEW → R Sweave
We wish to use KnitR rather than the older Sweave processor

In RStudio we can configure the options to use knitr:

- Select Tools→Options
- Choose the Sweave group
- Choose knitr for *Weave Rnw files using*:
- The remaining defaults should be okay
- Click **Apply** and then **OK**
Simple KnitR Document

Insert the following into your new KnitR document:

\title{Sample KnitR Document}
\author{Graham Williams}
\maketitle

\section*{My First Section}

This is some text that is automatically typeset by the LaTeX processor to produce well formatted quality output as PDF.

Your turn—Click Compile PDF to view the result.
Insert the following into your new KnitR document:

\title{Sample KnitR Document}
\author{Graham Williams}
\maketitle

\section*{My First Section}

This is some text that is automatically typeset by the LaTeX processor to produce well formatted quality output as PDF.

Your turn—Click **Compile PDF** to view the result.
Our First KnitR Document

---

Knitting

---

Simple KnitR Document

---

http://togaware.com

Copyright © 2014, Graham.Williams@togaware.com
Simple KnitR Document—Resulting PDF

Result of Compile PDF

Sample KnitR Document
Graham Williams
September 29, 2013

My First Section
This is some text that is automatically typeset by the LaTeX processor to produce well-formatted quality output as PDF.
KnitR: Add R Commands

R code can be used to generate results into the document:

```r
<<echo=FALSE, message=FALSE>>=
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function
ds <- weather
qplot(MinTemp, MaxTemp, data=ds)
@
```

Your turn—Click Compile PDF to view the result.
**KnitR: Add R Commands**

R code can be used to generate results into the document:

```r
<<echo=FALSE, message=FALSE>>=
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function
ds <- weather
qplot(MinTemp, MaxTemp, data=ds)
@
```

Your turn—Click **Compile PDF** to view the result.
KnitR Document With R Code

This is some text that is automatically typeset by the LaTeX processor to produce well formatted quality output as PDF.

library(rattle) # Provides the weather dataset
library(ggplot2) # Provides the qplot() function
dw <- weather
qplot(MinTemp, MaxTemp, data=dw)

R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-unknown-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
Sample KnitR Document
Graham Williams
September 30, 2013

My First Section
This is some text that is automatically typeset by the LaTeX processor to produce well-formatted output as PDF.

![Scatter plot](image.png)
LaTeX Basics

\subsection*{...} % Introduce a Sub Section

\subsubsection*{...} % Introduce a Sub Sub Section

\textbf{...} % Bold font
\textit{...} % Italic font

\begin{itemize}
\item ...
\item ...
\item ...
\end{itemize}

% A bullet list

Plus an extensive collection of other markup and capabilities.
KnitR Basics

```r
# Do not display the R code
echo=FALSE

# Evaluate the R code
eval=TRUE

# Hide the results of the R commands
results="hide"

# Extend figure width from 7 to 10 inches
fig.width=10

# Extend figure height from 7 to 8 inches
fig.height=8

# Fit figure 80% page width
out.width="0.8\textwidth"

# Fit figure 50% page height
out.height="0.5\textetheight"
```

Plus an extensive collection of other options.
RESOURCES AND REFERENCES

- **OnePageR**: [http://onepager.togaware.com](http://onepager.togaware.com) – Tutorial Notes
- **Rattle**: [http://rattle.togaware.com](http://rattle.togaware.com)
- **Guides**: [http://datamining.togaware.com](http://datamining.togaware.com)
- **Practise**: [http://analystfirst.com](http://analystfirst.com)

- **Book**: Data Mining using Rattle/R
- **Chapter**: Rattle and Other Tales
- **Paper**: A Data Mining GUI for R — R Journal, Volume 1(2)
Lecture Summary

- Data Science—Analytics—Data Mining;
- Rattle as a GUI for Quickly Analysing Data;
- The Power is with R.