Data Science with R

Introducing and Interacting with R

Graham.Williams@togaware.com

Senior Director and Data Scientist, Analytics
Australian Taxation Office

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

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

Overview

1. R Tool Suite
2. RStudio
3. Introduction to R
4. Knitting

Overview

1. R Tool Suite
2. RStudio
3. Introduction to R
4. Knitting

Tools

- Ubuntu GNU/Linux operating system
  - Feature rich toolkit, up-to-date, easy to install, FLOSS
- RStudio
  - Easy to use integrated development environment, FLOSS
- R Statistical Software Language
  - Extensive, powerful, thousands of contributors, FLOSS
- KnitR
  - Produce beautiful documents, easily reproducible, FLOSS

Using Ubuntu

- Desktop Ubuntu
- Connecting to Analytics Servers
  - Using XWin
  - Using VNC
- Start up RStudio from the Dash

Overview

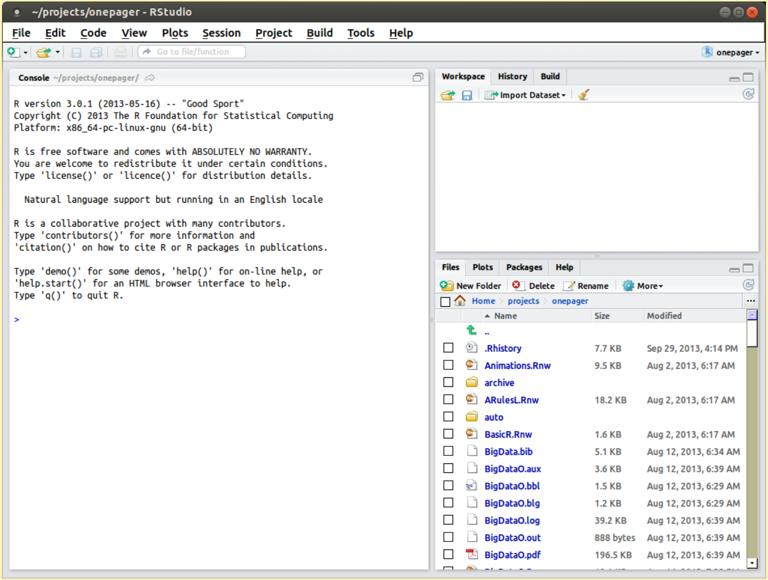
1. R Tool Suite

2. RStudio

3. Introduction to R

4. Knitting

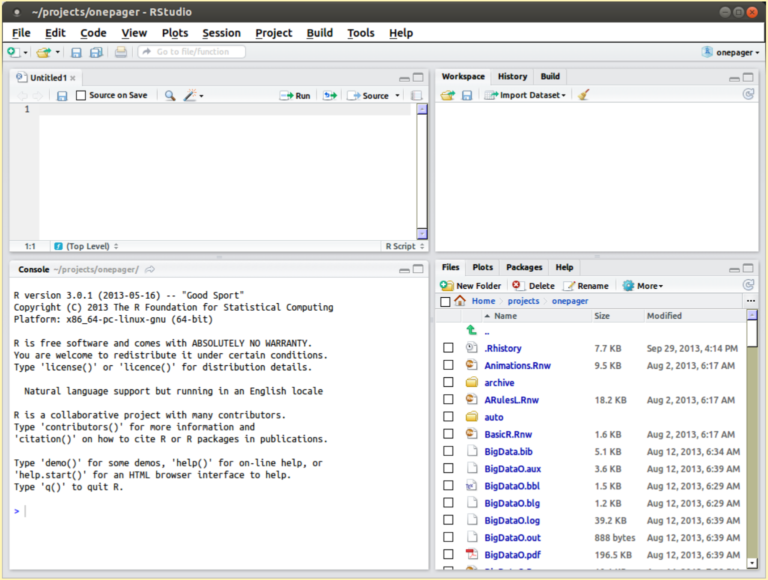
RStudio—The Default Three Panels



```
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.

>
```

Name	Size	Modified
..		
.Rhistory	7.7 KB	Sep 29, 2013, 4:14 PM
Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
archive		
ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
auto		
BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM

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.

> 
```

	Name	Size	Modified
	..		
□	.Rhistory	7.7 KB	Sep 29, 2013, 4:14 PM
□	Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
□	archive		
□	ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
□	auto		
□	BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
□	BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
□	BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
□	BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
□	BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
□	BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM

Overview

1 R Tool Suite

2 RStudio

3 Introduction to R

4 Knitting

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*

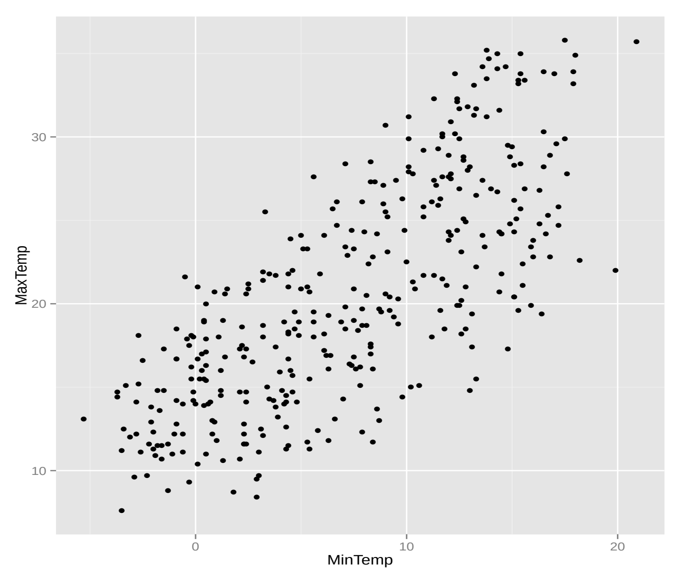
```
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function
```

- Then produce a quick plot using `qplot()`

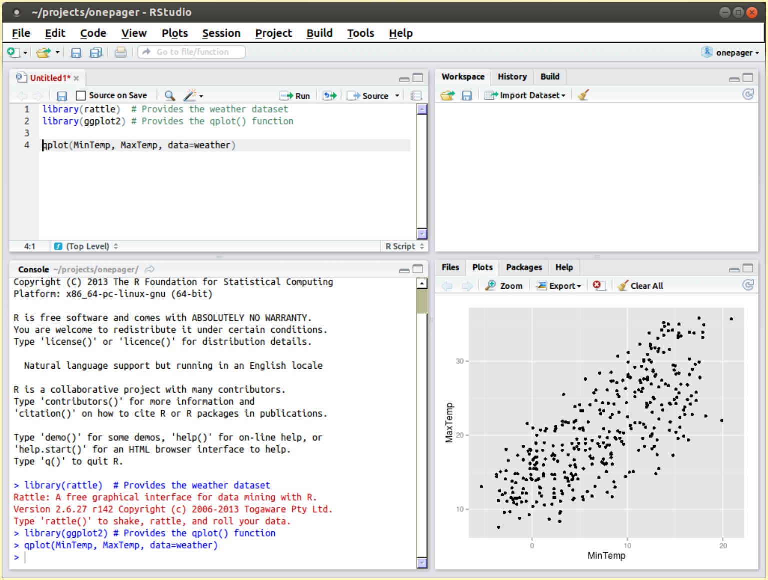
```
ds <- weather
qplot(MinTemp, MaxTemp, data=ds)
```

- Your turn: give it a go.

Scatterplot—Plot



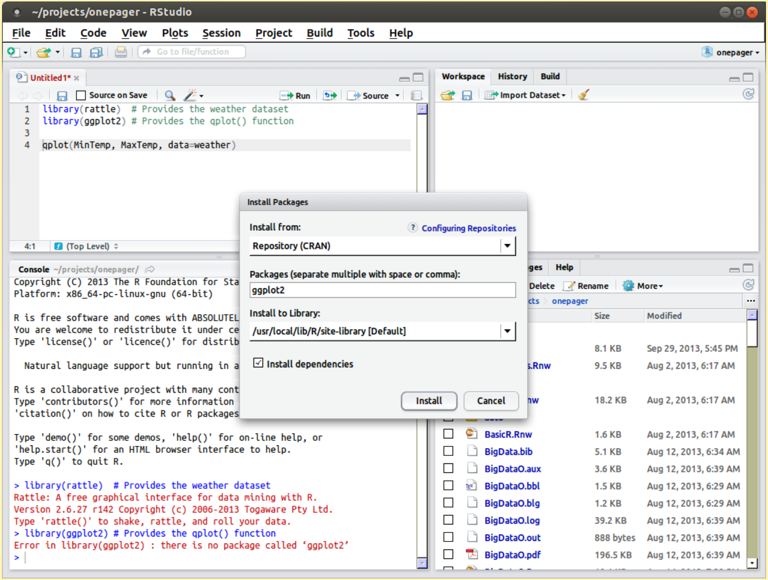
Scatterplot—RStudio



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

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

```
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.

> library(rattle)  # Provides the weather dataset
Rattle: A free graphical interface for data mining with R.
Version 2.6.27 r142 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)
> 
```

Missing Packages—Tools→Install Packages. . .



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

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

```
Copyright (C) 2013 The R Foundation for Sta
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTEL
You are welcome to redistribute it under ce
Type 'license()' or 'licence()' for distrib

  Natural language support but running in a

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

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 r142 Copyright (c) 2006-2013 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> library(ggplot2) # Provides the qplot() function
Error in library(ggplot2) : there is no package called ‘ggplot2’
> 
```

Install Packages

Install from: Repository (CRAN) — ? Configuring Repositories

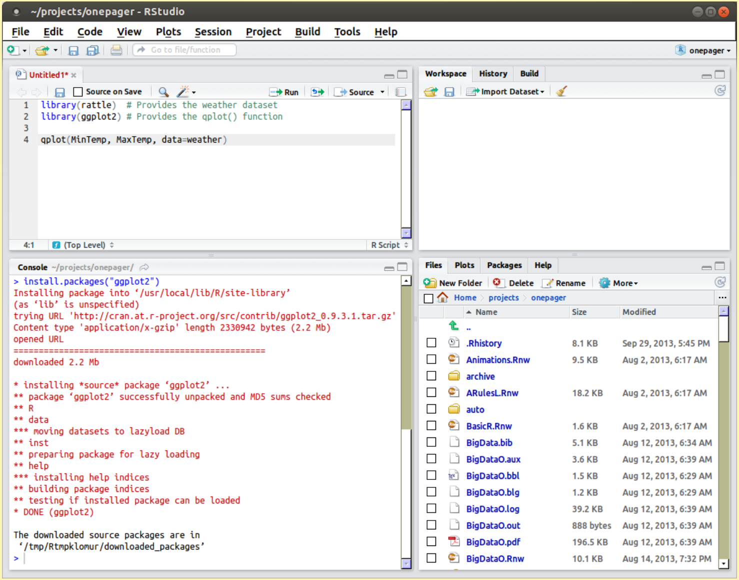
Packages (separate multiple with space or comma): ggplot2

Install to Library: /usr/local/lib/R/site-library [Default]

☑ Install dependencies

	Size	Modified
	8.1 KB	Sep 29, 2013, 5:45 PM
s.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
nw	18.2 KB	Aug 2, 2013, 6:17 AM
BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM

RStudio—Installing ggplot2



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

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

```
> install.packages("ggplot2")
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
trying URL 'http://cran.at.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
==================================================
downloaded 2.2 Mb

* installing *source* package ‘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
 ‘/tmp/Rtmpklomur/downloaded_packages’
> 
```

Home > projects > onepager

	Name	Size	Modified
	..		
□	.Rhistory	8.1 KB	Sep 29, 2013, 5:45 PM
□	Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
□	archive		
□	ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
□	auto		
□	BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
□	BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
□	BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
□	BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
□	BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
□	BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.Rnw	10.1 KB	Aug 14, 2013, 7:32 PM

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.

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.

Basic R

```
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)      # Struncture of the variables in the dataset.
```

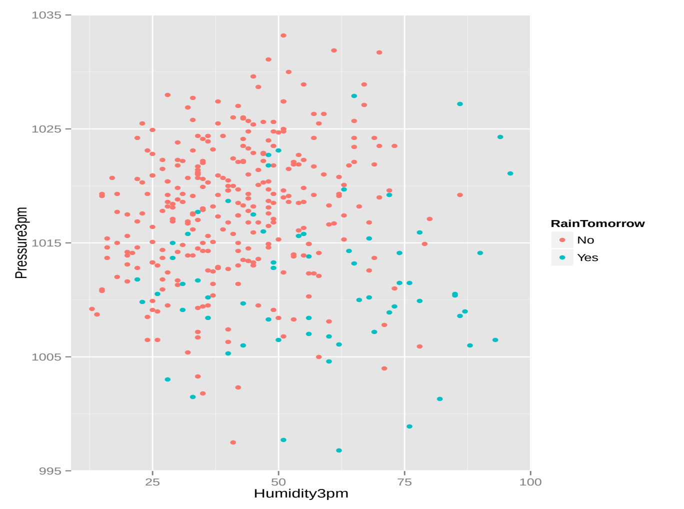
```
## 'data.frame': 366 obs. of  24 variables:
##  $ Date         : Date, format: "2007-11-01" "2007-11-...
##  $ Location     : Factor w/ 46 levels "Adelaide","Alba...
##  $ MinTemp      : num  8 14 13.7 13.3 7.6 6.2 6.1 8.3 ...
....
```

Basic R

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

##       Date                     Location      MinTemp      ...
##  Min.   :2007-11-01   Canberra     :366   Min.   :-5.30   ...
##  1st Qu.:2008-01-31   Adelaide     :  0   1st Qu.: 2.30   ...
##  Median :2008-05-01   Albany       :  0   Median : 7.45   ...
##  Mean   :2008-05-01   Albury       :  0   Mean   : 7.27   ...
##  3rd Qu.:2008-07-31   AliceSprings :  0   3rd Qu.:12.50   ...
##  Max.   :2008-10-31   BadgerysCreek:  0   Max.   :20.90   ...
##                       (Other)      :  0                   ...
##     Rainfall      Evaporation       Sunshine      WindGust...
##  Min.   : 0.00   Min.   : 0.20   Min.   : 0.00   NW     : ...
##  1st Qu.: 0.00   1st Qu.: 2.20   1st Qu.: 5.95   NNW    : ...
##  Median : 0.00   Median : 4.20   Median : 8.60   E      : ...
##  Mean   : 1.43   Mean   : 4.52   Mean   : 7.91   WNW    : ...
##  3rd Qu.: 0.20   3rd Qu.: 6.40   3rd Qu.:10.50   ENE    : ...
....
```

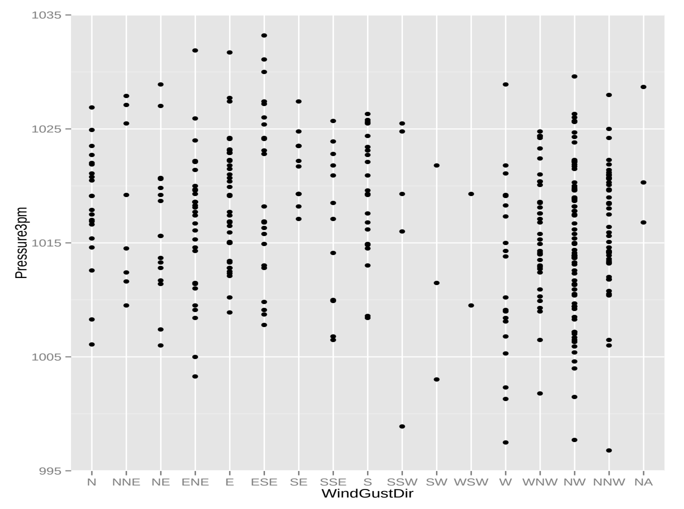
Visual Summaries—Add A Little Colour

```
qplot(Humidity3pm, Pressure3pm, colour=RainTomorrow, data=ds)
```



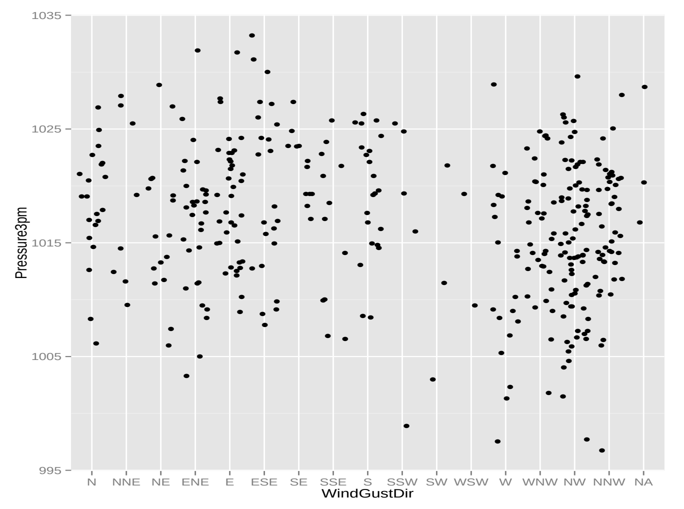
Visual Summaries—Careful with Categorics

```
qplot(WindGustDir, Pressure3pm, data=ds)
```



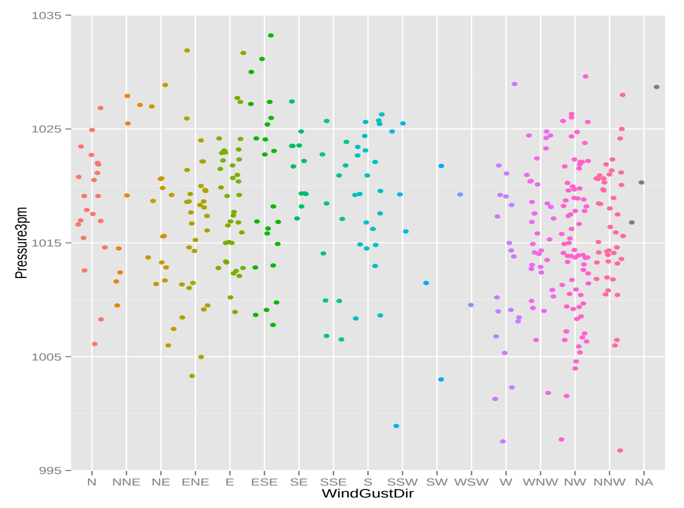
Visual Summaries—Add A Little Jitter

```
qplot(WindGustDir, Pressure3pm, data=ds, geom="jitter")
```

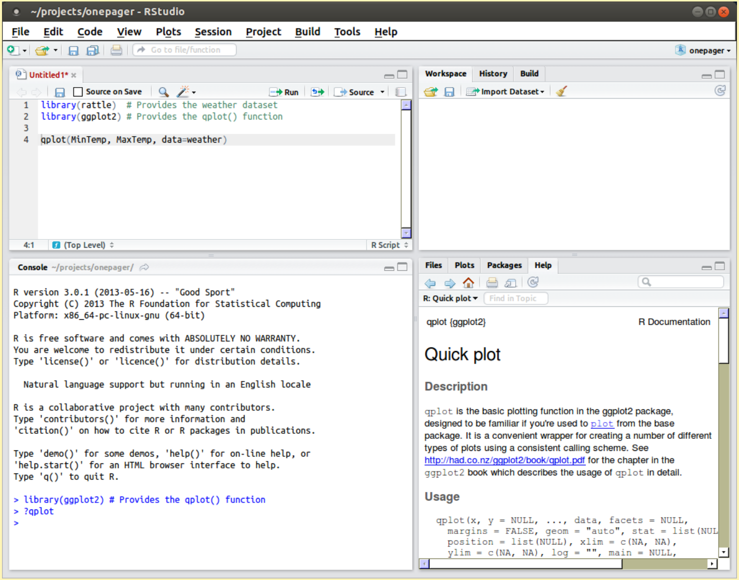


Visual Summaries—And Some Colour

```
qplot(WindGustDir, Pressure3pm, data=ds, colour=WindGustDir, geom="jitter")
```



Getting Help—Precede Command with ?



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

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

```
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.

> library(ggplot2) # Provides the qplot() function
> ?qplot
>
```

qplot {ggplot2} R Documentation

Quick plot

Description

qplot is the basic plotting function in the ggplot2 package, designed to be familiar if you're used to plot from the base package. It is a convenient wrapper for creating a number of different types of plots using a consistent calling scheme. See http://had.co.nz/ggplot2/book/qplot.pdf for the chapter in the ggplot2 book which describes the usage of qplot in detail.

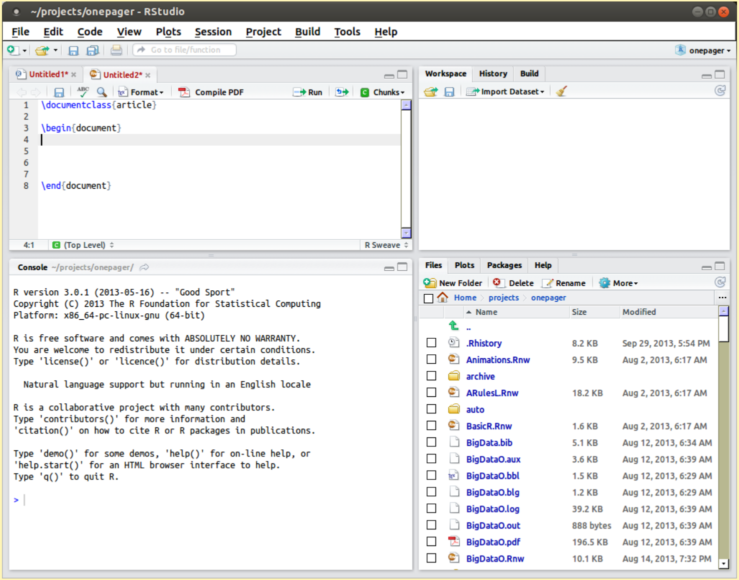
Usage

```
  qplot(x, y = NULL, ..., data, facets = NULL,
    margins = FALSE, geom = "auto", stat = list(NUL
    position = list(NULL), xlim = c(NA, NA),
    ylim = c(NA, NA), log = "", main = NULL,
```

Overview

1. R Tool Suite
2. RStudio
3. Introduction to R
4. Knitting

Create a KnitR Document: New→R Sweave



```
\documentclass{article}

\begin{document}




\end{document}
```

```
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.

> 
```

	Name	Size	Modified
	..		
□	.Rhistory	8.2 KB	Sep 29, 2013, 5:54 PM
□	Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
□	archive		
□	ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
□	auto		
□	BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
□	BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
□	BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
□	BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
□	BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
□	BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.Rnw	10.1 KB	Aug 14, 2013, 7:32 PM

Setup KnitR

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.

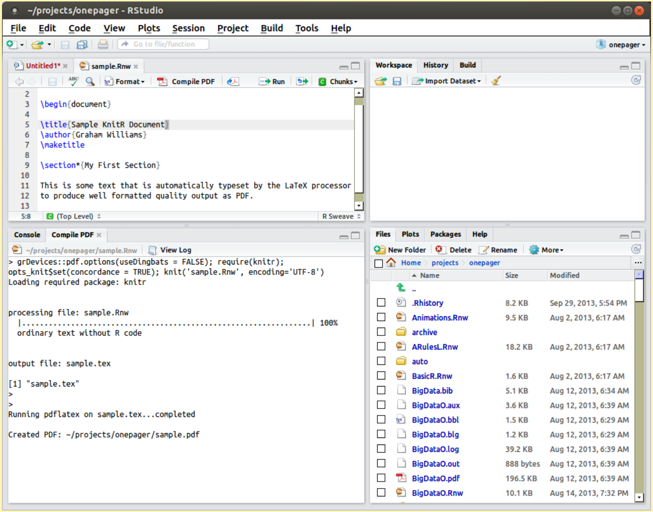
Simple KnitR Document

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.

Simple KnitR Document



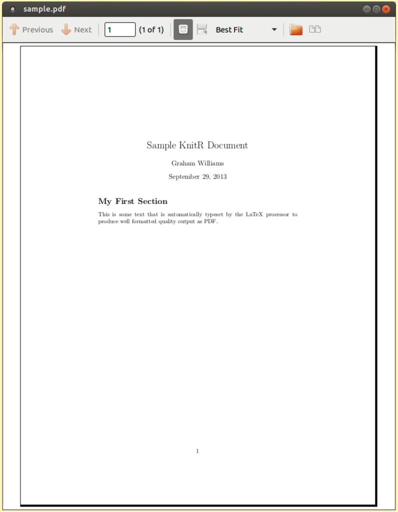
```

\begin{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.

```

```
> grDevices::pdf.options(useDingbats = FALSE); require(knitr);
opts_knit$set(concordance = TRUE); knit('sample.Rnw', encoding='UTF-8')
Loading required package: knitr


processing file: sample.Rnw
  |.................................................................| 100%
  ordinary text without R code


output file: sample.tex

[1] "sample.tex"
>
>
Running pdflatex on sample.tex...completed

Created PDF: ~/projects/onepager/sample.pdf
```

	Name	Size	Modified
	..		
□	.Rhistory	8.2 KB	Sep 29, 2013, 5:54 PM
□	Animations.Rnw	9.5 KB	Aug 2, 2013, 6:17 AM
□	archive		
□	ARulesL.Rnw	18.2 KB	Aug 2, 2013, 6:17 AM
□	auto		
□	BasicR.Rnw	1.6 KB	Aug 2, 2013, 6:17 AM
□	BigData.bib	5.1 KB	Aug 12, 2013, 6:34 AM
□	BigDataO.aux	3.6 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.bbl	1.5 KB	Aug 12, 2013, 6:29 AM
□	BigDataO.blg	1.2 KB	Aug 12, 2013, 6:29 AM
□	BigDataO.log	39.2 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.out	888 bytes	Aug 12, 2013, 6:39 AM
□	BigDataO.pdf	196.5 KB	Aug 12, 2013, 6:39 AM
□	BigDataO.Rnw	10.1 KB	Aug 14, 2013, 7:32 PM

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.

1

KnitR: Add R Commands

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

```
<<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.

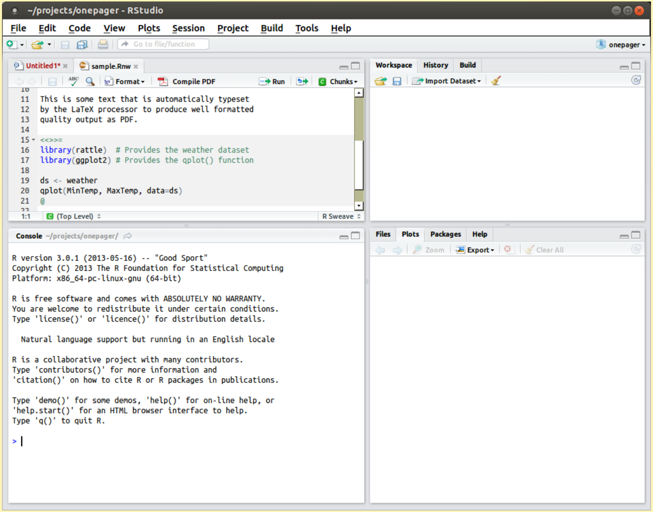
KnitR: Add R Commands

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

```
<<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

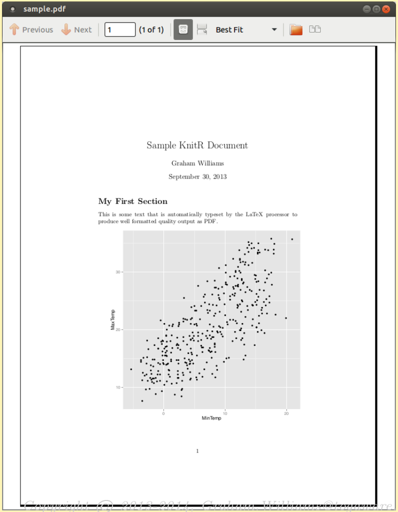


```
This is some text that is automatically typeset
by the LaTeX processor to produce well formatted
quality output as PDF.

<<>>=
library(rattle)  # Provides the weather dataset
library(ggplot2) # Provides the qplot() function

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

```
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.

> 
```

Simple KnitR Document—Resulting PDF with Plot

Result of **Compile PDF**



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 quality output as PDF.

1

LaTeX Basics

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

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

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

\begin{itemize}            % A bullet list
  \item ...
  \item ...
\end{itemize}
```

Plus an extensive collection of other markup and capabilities.

KnitR Basics

```
echo=FALSE         # Do not display the R code
eval=TRUE          # Evaluate the R code

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

fig.width=10       # Extend figure width from 7 to 10 inches
fig.height=8       # Extend figure height from 7 to 8 inches

out.width="0.8\\textwidth"     # Fit figure 80% page width
out.height="0.5\\textheight"   # Fit figure 50% page height
```

Plus an extensive collection of other options.

Thank You

Question Time

This document, sourced from IntroRL.Rnw revision 282, was processed by KnitR version 1.5 of 2013-09-28 and took 2.4 seconds to process. It was generated by gjw on nyx running Ubuntu 13.10 with Intel(R) Xeon(R) CPU W3520 @ 2.67GHz having 4 cores and 12.3GB of RAM. It completed the processing 2014-02-14 06:19:56.