Data analysis and visualization using R

The R toolbox

Michiel Noback

november 2015

Why do statistical programming?

Complete reproducibility and flexibility

- Keeping an exact log of all your mouse clicks in Excel may be problematic. . .
- Redoing an analysis with adjustments is easy
- Store the analysis, not its output

The R toolbox

Overview

This presentation will introduce you to a toolbox that will serve you well during your data quests.

- The R programming language
- The R studio IDE (Integrated Development Environment)
- R Markdown as documenting and reporting tool
- swirl as training environment

Tool 1: R itself

The R programming language

Wikipedia says:
“R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis”

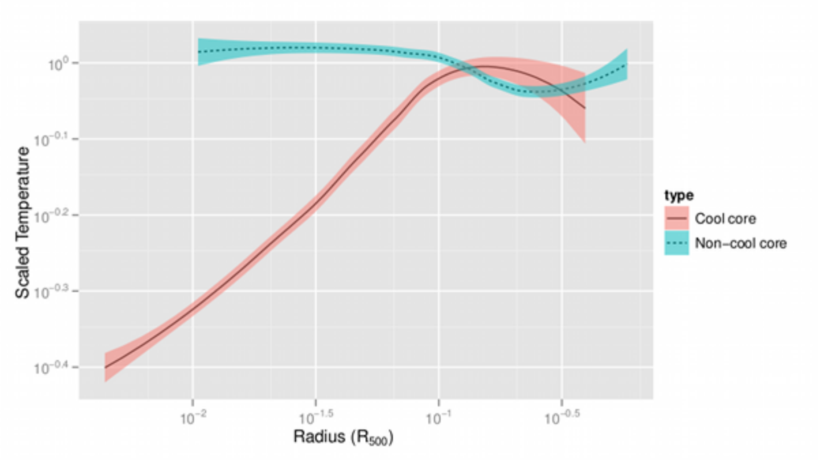
R is just a programming language

- R is a programming language like any other.
- It knows about character data, numbers, lists, functions (. . . but **has no dicts**)
- Like Python, it has an interactive mode

R is **not** just a programming language

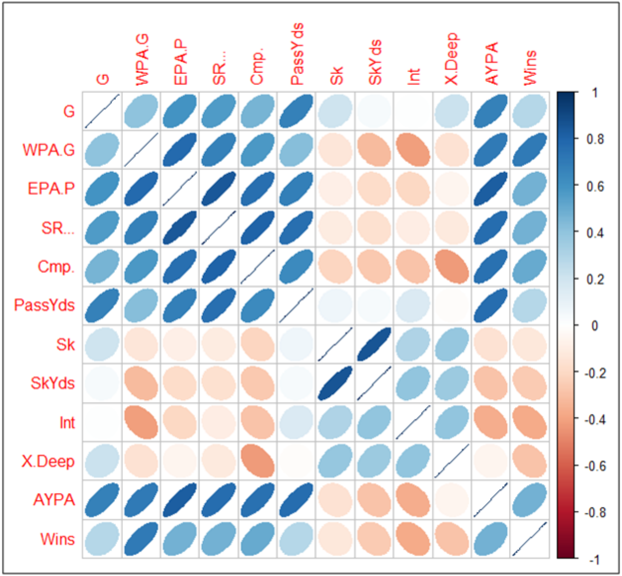
- It has embedded advanced graphical support
- It has extensive support for statistical work

R showcase (1)

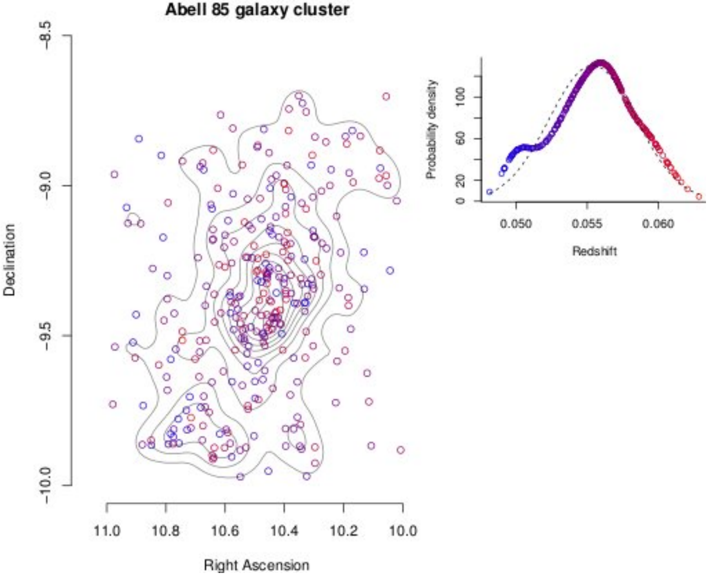


Scaled Temperature
10^{0}
$10^{-0.1}$
$10^{-0.2}$
$10^{-0.3}$
$10^{-0.4}$
10^{-2}
$10^{-1.5}$
10^{-1}
$10^{-0.5}$
Radius (R_{500})
type
Cool core
Non−cool core

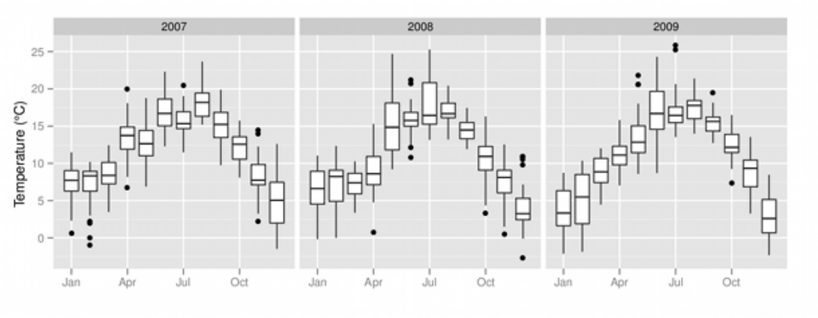
R showcase (2)


G
WPA.G
EPA.P
SR...
Cmp.
PassYds
Sk
SkYds
Int
X.Deep
AYPA
Wins
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

R showcase (3)


Abell 85 galaxy cluster
Declination
-8.5
-9.0
-9.5
-10.0
11.0
10.8
10.6
10.4
10.2
10.0
Right Ascension
Probability density
0
20
60
100
0.050
0.055
0.060
Redshift

R showcase (4)



What is funny about R

In R, **EVERYTHING** lives inside a vector. There are no singular variables of atomic data types, only vectors of length zero or greater:

```
x <- 42
x
```

```
## [1] 42
```

```
x[1]
```

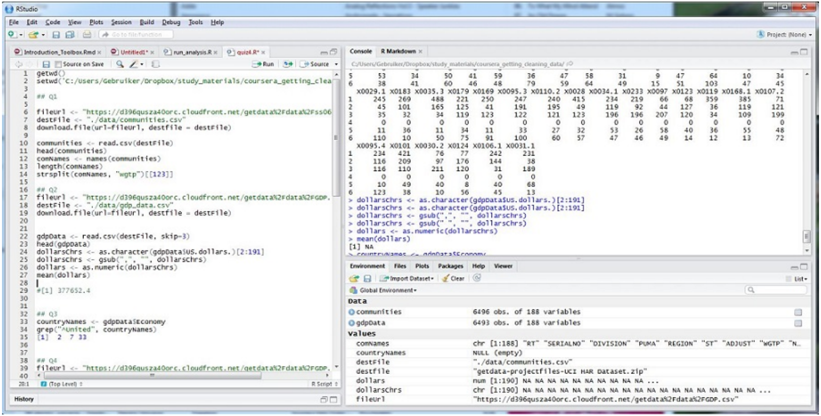
```
## [1] 42
```

```
x[1][1][1][1]
```

```
## [1] 42
```

Tool 2: RStudio

RStudio is the IDE for R



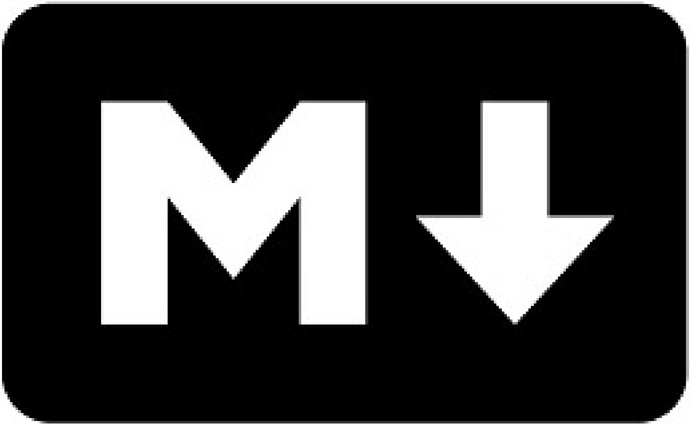
```
getwd()
setwd('C:/users/Gebruiker/Dropbox/study_materials/coursera_getting_clea

## Q1

fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06
destFile <- "./data/communities.csv"
download.file(url=fileUrl, destfile = destFile)

communities <- read.csv(destFile)
head(communities)
comNames <- names(communities)
length(comNames)
strsplit(comNames, "wgtp")[[123]]

## Q2
fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.
destFile <- "./data/gdp_data.csv"
download.file(url=fileUrl, destfile = destFile)


gdpData <- read.csv(destFile, skip=3)
head(gdpData)
dollarsChrs <- as.character(gdpData$US.dollars.)[2:191]
dollarsChrs <- gsub(",", "", dollarsChrs)
dollars <- as.numeric(dollarsChrs)
mean(dollars)

#[1] 377652.4


## Q3
countryNames <- gdpData$Economy
grep("^United", countryNames)
[1]  2  7 33


## Q4
fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.
```

C:/Users/Gebruiker/Dropbox/study_materials/coursera_getting_cleaning_data/

```
5       53     34      50    41    59      36      47    58      31     9    47    64      10      34
6       38     41      60    46    48      79      59    64      49    15    51   103      47      45
  X0029.1 X0183 X0035.3 X0179 X0169 X0095.3 X0110.2 X0028 X0034.1 X0233 X0097 X0123 X0119 X0168.1 X0107.2
1      245   269     488   221   250     247     240   415     234   219    66    68   359     385      71
2       45   101     165   125    41     191     195    49     119    92    44   127    36     119     121
3       35    32      34   119   123     122     121   123     196   196   207   120    34     109     199
4        0     0       0     0     0       0       0     0       0     0     0     0     0       0       0
5       11    36      11    34    11      33      27    32      53    26    58    40    36      55      48
6      110    10      50    75    91     100      60    57      47    46    49    14    12      13      72
  X0095.4 X0101 X0030.2 X0124 X0106.1 X0031.1
1     234   421      76    77     242     231
2     116   209      97   176     144      38
3     116   110     211   120      31     189
4       0     0       0     0       0       0
5      10    49      40     8      40      68
6     123    38      10    56      45      13
> dollarsChrs <- as.character(gdpData$US.dollars.)[2:191]
> dollarsChrs <- as.character(gdpData$US.dollars.)[2:191]
> dollarsChrs <- gsub(",", "", dollarsChrs)
> dollarsChrs <- gsub(" ", "", dollarsChrs)
> dollars <- as.numeric(dollarsChrs)
> mean(dollars)
[1] NA
> countryNames <- gdpData$Economy
```

```
Data
communities       6496 obs. of 188 variables
gdpData           6493 obs. of 188 variables
Values
comNames          chr [1:188] "RT" "SERIALNO" "DIVISION" "PUMA" "REGION" "ST" "ADJUST" "WGTP" "N…
countryNames      NULL (empty)
destFile          "./data/communities.csv"
destfile          "getdata-projectfiles-UCI HAR Dataset.zip"
dollars           num [1:190] NA NA NA NA NA NA NA NA NA NA ...
dollarsChrs       chr [1:190] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA ...
fileUrl           "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"
```

Tool 3: RMarkdown

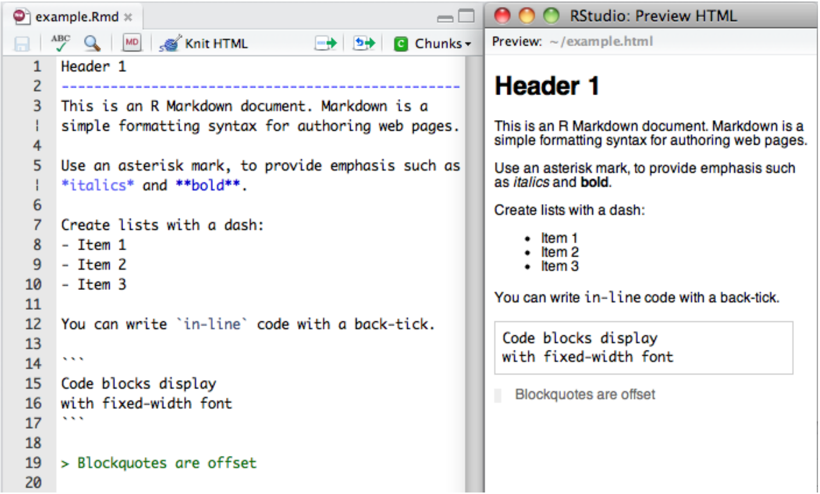


RMarkdown is a great reporting tool

You can use it

- to create reports in word, pdf or html
- you can use it to create presentations (such as this one)
- support for it is integrated in RStudio
- to embed R code that will be executed to generate your report

RMarkdown is really basic



```
Header 1
--------------------------------------------
This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages.

Use an asterisk mark, to provide emphasis such as *italics* and **bold**.

Create lists with a dash:
- Item 1
- Item 2
- Item 3

You can write `in-line` code with a back-tick.

```
Code blocks display
with fixed-width font
```

> Blockquotes are offset

```

# Header 1

This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages.

Use an asterisk mark, to provide emphasis such as *italics* and **bold**.

Create lists with a dash:

- Item 1
- Item 2
- Item 3

You can write `in-line` code with a back-tick.

```
Code blocks display
with fixed-width font
```

> Blockquotes are offset

source http://rmarkdown.rstudio.com/

Tool 4: swirl

Learn R, in R.

swirl

- Swirl is a very nice tool to train your R brain
- It is a framework for interactive lessons in R
- I recommend doing those!
- See swirlstats.com

Let's get dirty

- OK, before we're going into the gritty details of R, let me try to convince you why it could be nice to learn it.
- R has a very nice package of demo datasets
- ChickWeight is one of these - a dataframe with 578 weight measurements of chicks on different diets

```
head(ChickWeight)
```

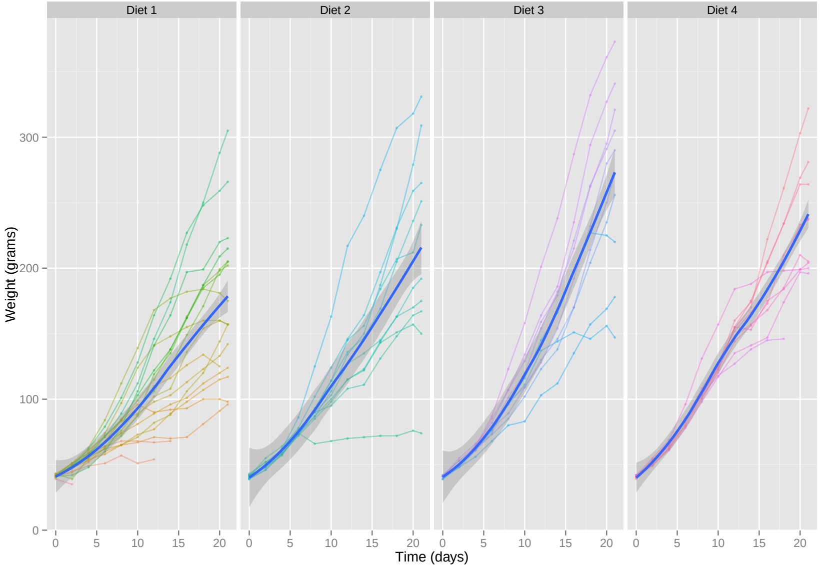
```
##   weight Time Chick Diet
## 1     42    0     1    1
## 2     51    2     1    1
## 3     59    4     1    1
## 4     64    6     1    1
## 5     76    8     1    1
## 6     93   10     1    1
```

Plotting ChickWeight

- I want to explore weight trends over time for the different diets
- Done using the ggplot2 package in only a few lines of code
- *I challenge you to try this in Excel!!*

```r
library(ggplot2)
#make nicer labels
ChickWeight$Diet <- factor(ChickWeight$Diet,
        labels = c("Diet 1", "Diet 2", "Diet 3", "Diet 4"))
ggplot(ChickWeight, aes(x = Time, y = weight)) + facet_gri
    geom_point(aes(color = Chick), alpha = 0.3, size = 1) +
    geom_line(aes(color = Chick), alpha = 0.3) +
    geom_smooth(method = "loess", size = 1) +
    theme(legend.position="none") +
    xlab("Time (days)") + ylab("Weight (grams)")
```

- picture on next slide


Diet 1
Diet 2
Diet 3
Diet 4
Weight (grams)
0
100
200
300
0
5
10
15
20
Time (days)

The end

- This concludes the overview of the toolbox and demo use case
- In the next presentation we'll cover some basic R and then come back to visit R markdown