



 from scratch

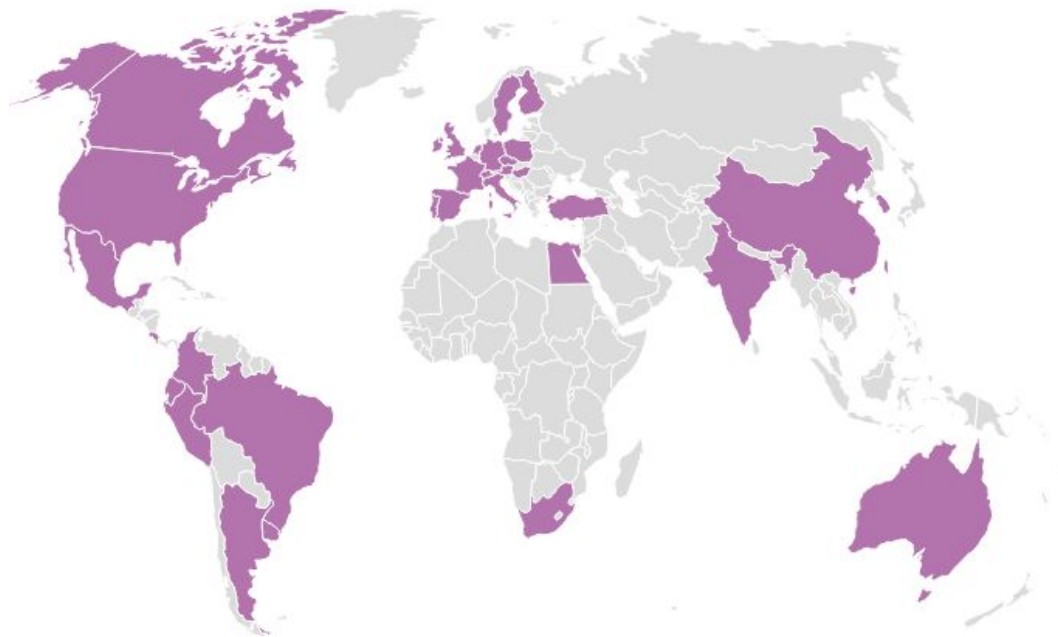
Kasia Kulma & Erle Holgersen

OVERVIEW

1. R-ladies - who are we? (5 mins)
2. What's R and why you should learn it (5 mins)
3. Let's learn some R! (1h 20 mins)
4. Post-session survey (5 mins)

What's R-Ladies?

R-ladies is a world-wide organization to promote gender diversity in the R community: <https://rladies.org/>



- 60+ Chapters worldwide
- 20+ Countries
- 6000+ Members

What's R...

- Open source programming language for statistical computing & graphics
- Originated from language S
- Developed & matured in early 1990's in New Zealand (Auckland University)
- Names to remember: Ross Ihaka & Robert Gentleman

...and why you should learn it

- Freeware, open - source & platform - independent
- **HUGE** choice of analytics and statistics libraries
- Integrates many data - sources
- Efficient data structures (data.frame, tibble, data.table, etc.)
- Produces arguably best Data Visualizations
- Can build interactive web apps (Shiny)
- **Easy** reporting and blogging (RMarkdown, blogdown)
- **AWESOME** community

...and why you should learn it

- Freeware, open - source & platform - independent
- **HUGE** choice of analytics and statistics libraries
- Integrates many data - sources
- Efficient data structures (data.frame, tibble, data.table, etc.)
- Produces arguably best Data Visualizations
- Can build interactive web apps (Shiny)
- **Easy** reporting and blogging (RMarkdown, blogdown)
- **AWESOME** community

...and why you should learn it

- Freeware, open - source & platform - independent
- **HUGE** choice of analytics and statistics libraries
- Integrates many data - sources
- Efficient data structures (data.frame, tibble, data.table, etc.)
- Produces arguably best Data Visualizations
- Can build interactive web apps (Shiny)
- **Easy** reporting and blogging (RMarkdown, blogdown)
- **AWESOME** community

...and why you should learn it

- Freeware, open - source & platform - independent
- **HUGE** choice of analytics and statistics libraries
- Integrates many data - sources
- Efficient data structures (data.frame, tibble, data.table, etc.)
- Produces arguably best Data Visualizations
- Can build interactive web apps (Shiny)
- **Easy** reporting and blogging (RMarkdown, blogdown)
- **AWESOME** community

...and why you should learn it

- Freeware, open - source & platform - independent
- **HUGE** choice of analytics and statistics libraries
- Integrates many data - sources
- Efficient data structures (data.frame, tibble, data.table, etc.)
- Produces arguably best Data Visualizations
- Can build interactive web apps (Shiny)
- **Easy** reporting and blogging (RMarkdown, blogdown)
- **AWESOME** community

...and why you should learn it

- Freeware, open - source & platform - independent
- **HUGE** choice of analytics and statistics libraries
- Integrates many data - sources
- Efficient data structures (data.frame, tibble, data.table, etc.)
- Produces arguably best Data Visualizations
- Can build interactive web apps (Shiny)
- **Easy** reporting and blogging (RMarkdown, blogdown)
- **AWESOME** community

...and why you should learn it

- Freeware, open - source & platform - independent
- **HUGE** choice of analytics and statistics libraries
- Integrates many data - sources
- Efficient data structures (data.frame, tibble, data.table, etc.)
- Produces arguably best Data Visualizations
- Can build interactive web apps (Shiny)
- **Easy** reporting and blogging (RMarkdown, blogdown)
- **AWESOME** community

...and why you should learn it

- Freeware, open - source & platform - independent
- **HUGE** choice of analytics and statistics libraries
- Integrates many data - sources
- Efficient data structures (data.frame, tibble, data.table, etc.)
- Produces arguably best Data Visualizations
- Can build interactive web apps (Shiny)
- **Easy** reporting and blogging (RMarkdown, blogdown)
- **AWESOME** community

Let's learn some R!

- Installing R and RStudio
- RStudio Projects
- Data import and exploration
- Data manipulation & basic Data Visualisation
- Taking R to the next level

Installing R and RStudio

1. Install R from CRAN (<https://cran.r-project.org/>)
2. Install R's IDE: RStudio (<https://www.rstudio.com/>)
3. Open RStudio Console

The screenshot shows the RStudio interface with a script editor, a console, and an environment pane. The script editor contains R code for installing the 'pacman' package and scraping data from Amazon. The console shows the R version and platform information. The environment pane shows the loaded packages and data frames.

```
1 ## installing and loading pkgs ##
2
3 install.packages("pacman", repos = "http://cran.us.r-project.org")
4 pacman::p_load(OML, dplyr, stringr, rvest, xml2) # web-scraping
5 pacman::p_load(tm, rvest, SnowballC, wordcloud) # wordcloud-building
6
7
8 # define a function removing all white space
9 trim <- function(x) gsub("\\s+", "", x)
10
11 # define product code and the url
12 prod_code <- "1846558239"
13 url <- paste0("https://www.amazon.co.uk/dp/", prod_code)
14 doc <- xml2::read_html(url)
15
16 # obtain the text in the node, remove "\n" from the text, and remove white space
17 prod <- html_nodes(doc, "#productTitle") %>% html_text() %>% gsub("\n", "", .) %>% trim()
18 prod
19
20 source("https://raw.githubusercontent.com/rjsohta/Just-R-Things/master/Texts2QMining/amazonscraper.R")
21 pages <- 100
22
```

Console:

```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.

[Workspace loaded from ~/.RData]
>
```

Environment:

Name	Type	Len.	Size	Value
access_t...	charac...	1	152 B	"567537377-HMRh7...
access_t...	charac...	1	136 B	"na3l82lWhZuxPGD...
amazon_s...	function	1	75.2	function (doc...
api_key	charac...	1	120 B	"9wMs8deRf16FR8F...
api_sec...	charac...	1	152 B	"IED6n8nn8nxf5r...
apps	data.f...	2	4.3	37 obs. of 2
batman	data.f...	4	26.6	190 obs. of 4
bottom_10	tbl_df	2	1.7	10 obs. of 2
chars	data.f...	3	7.9	40 obs. of 3
doc	xml2_o...	2	728 B	List of 2

Data Input

read.table (utils) R Documentation

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"",
  as.is = TRUE, na.strings = "NA", colClasses = NA,
  skip = 0, check.names = TRUE, fill = NA,
  as.is.white = FALSE, blank.lines.skip = NA,
  comment.char = "#",
  allowEscapes = FALSE, flush = FALSE,
  stringsAsFactors = default.stringsAsFactors(),
  fileEncoding = "", encoding = "unknown")

read.csv(file, header = TRUE, sep = ",", quote = "\"",
  as.is = TRUE, na.strings = "NA", colClasses = NA,
  skip = 0, check.names = TRUE, fill = NA,
  as.is.white = FALSE, blank.lines.skip = NA,
  comment.char = "#",
  allowEscapes = FALSE, flush = FALSE,
  stringsAsFactors = default.stringsAsFactors(),
  fileEncoding = "", encoding = "unknown")

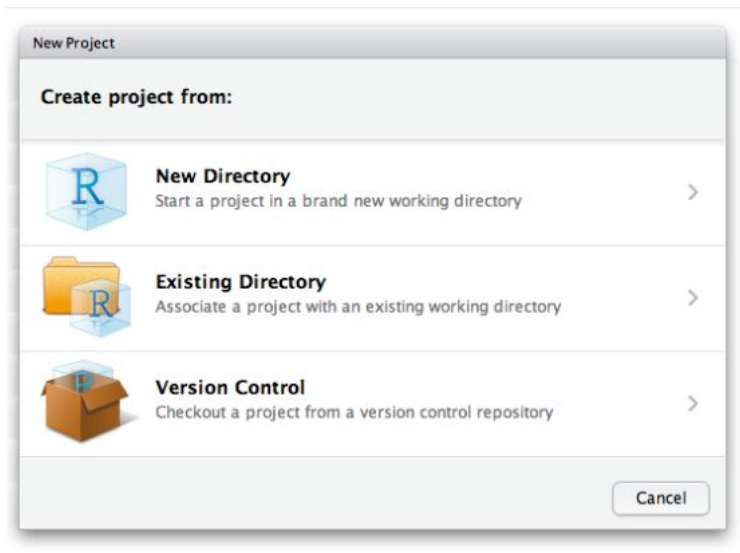
read.csv2(file, header = TRUE, sep = ";", quote = "\"",
  as.is = TRUE, na.strings = "NA", colClasses = NA,
  skip = 0, check.names = TRUE, fill = NA,
  as.is.white = FALSE, blank.lines.skip = NA,
  comment.char = "#",
  allowEscapes = FALSE, flush = FALSE,
  stringsAsFactors = default.stringsAsFactors(),
  fileEncoding = "", encoding = "unknown")

read.delim(file, header = TRUE, sep = "\t", quote = "\"",
```



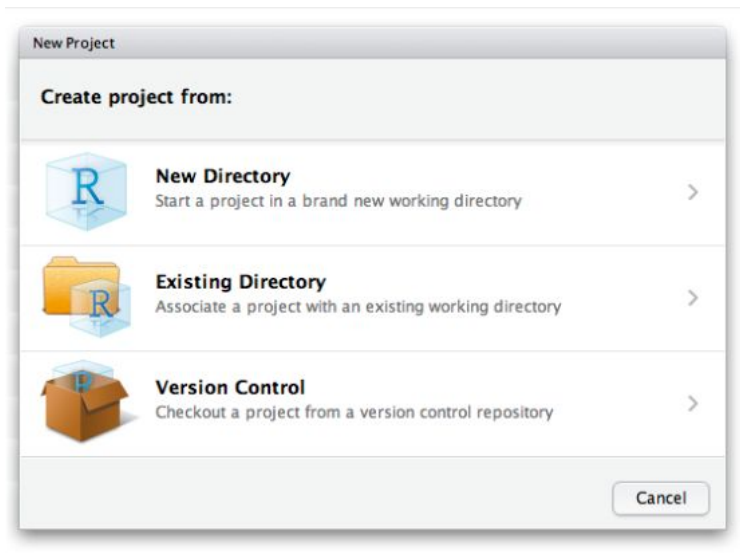
RStudio Projects

R Projects - divide your work into multiple contexts, each with their own working directory, workspace, history, and source documents



RStudio Projects

R Projects - divide your work into multiple contexts, each with their own working directory, workspace, history, and source documents

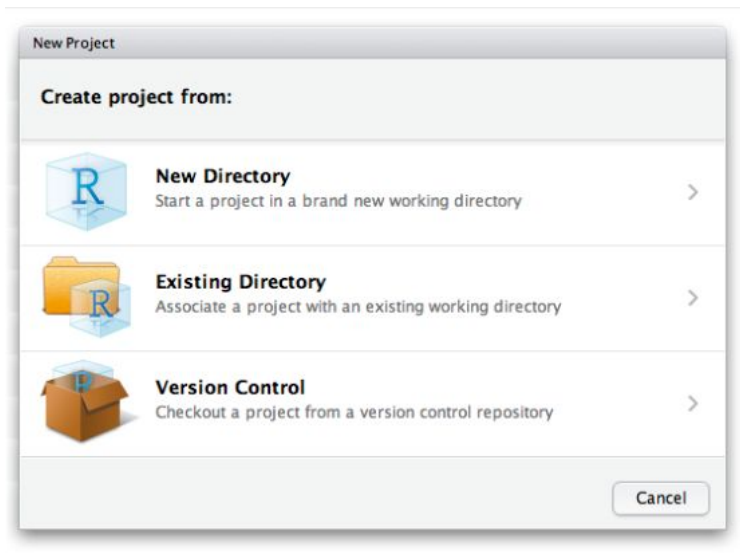


shows current working directory
`getwd()`

changes current working directory
`setwd()`

RStudio Projects

R Projects - divide your work into multiple contexts, each with their own working directory, workspace, history, and source documents



shows current working directory
getwd()

changes current working directory
setwd()



Data import and exploration

In - built datasets

`data(iris) # loads data into the workspace`

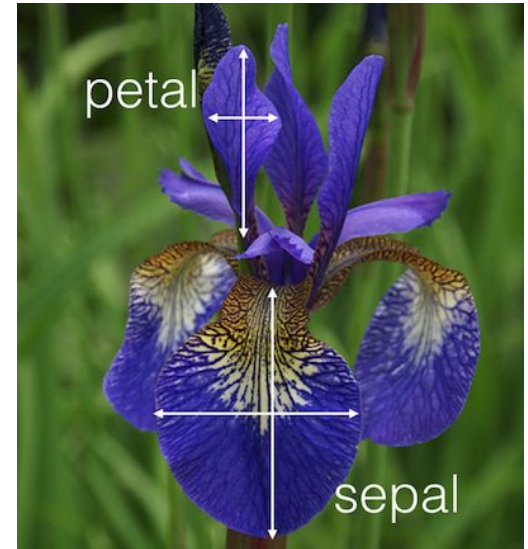
`head(iris) # views the top 6 rows`

`head(iris, 10) # views the top 10 rows`

`tail(iris, 10) # views bottom 10 rows`

`str(iris) # explores data structure`

`summary(iris) # summarises data`





Data import and exploration

saving and importing .csv files

saves a data.frame in a .csv file

```
write.csv(iris, file = "20171019_iris_local.csv")
```

imports a .csv file and saves it in a new data object

```
local_iris <- read.csv("20171019_iris_local.csv")
```

```
str(local_iris)
```



Data manipulation & basic Data Visualisation

numeric to integer

```
local_iris$Petal.Length  
<-as.integer(local_iris$Petal.Length)
```

numeric to character

```
local_iris$Petal.Width  
<-as.character(local_iris$Petal.Width)
```


Data manipulation & basic Data Visualisation

new logical var

```
local_iris$is_setosa <- local_iris$Species == "setosa"
```

new numeric variables

```
local_iris$sepal_sum <- local_iris$Sepal.Width +  
local_iris$Sepal.Length
```

```
local_iris$petal_sum <- local_iris$Petal.Width +  
local_iris$Petal.Length
```



Data manipulation & basic Data Visualisation

simple scatter plots in base R

```
plot(iris$Sepal.Length ~ iris$Sepal.Width) # basic plot
```

base plot + main title

```
plot(iris$Sepal.Length ~ iris$Sepal.Width, main = "Sepal length by  
sepal width")
```

base plot + main title + axis labels

```
plot(iris$Sepal.Length ~ iris$Sepal.Width, main = "Sepal length by  
sepal width", ylab = "Sepal length", xlab = "Sepal Width")
```



Data manipulation & basic Data Visualisation

simple bar plots in base R

```
boxplot(iris$Sepal.Length ~ iris$Species) # base plot
```

base plot + main title

```
boxplot(iris$Sepal.Length ~ iris$Species, main = "Sepal length by  
species")
```

base plot + main title + axis labels

```
boxplot(iris$Sepal.Length ~ iris$Species, main = "Sepal length by  
species", ylab = "Sepal length", xlab = "Species")
```



Need help?

Each function comes with a documentation page that can be visualised typing:

- `help(name_of_function)` or
- `?name_of_function`

Other suggestions:

- Join R-Ladies - already done!
- Google your problem
- Browse <http://rseek.org/> to find out which packages are available for a given topic
- Join an R users forum (e.g. [R-help-archive - Google Groups](#) or the RStudio community forum <https://community.rstudio.com/>)
- Post a question on <https://stackoverflow.com>

Taking R to the next level

Cheat Sheets & Reference Guides

- R Reference Card (<http://cran.r-project.org/doc/contrib/Short-refcard.pdf>)
- Writing R extensions (<http://bit.ly/1H0U02a>)
- Google's R Style Guide (<https://google.github.io/styleguide/Rguide.xml>)
- RStudio cheatsheets:
 - Data Visualization (<http://bit.ly/1Foy1Lb>)
 - Package Development (<http://bit.ly/1CfbTD2>)
 - Data Wrangling (<http://bit.ly/1y2nh3f>)
 - R Markdown (<http://bit.ly/1BluuT5>)
 - R Markdown Reference Guide (<http://bit.ly/1L2tC7U>)
 - Shiny (<http://bit.ly/1GiGArG>)

Taking R to the next level

Great

tutorials:

- edx MiT course (<https://www.edx.org/course/analytics-edge-mitx-15-071x-2>)
- DataCamp (<https://www.datacamp.com/>)
- Coursera (<https://www.coursera.org/learn/r-programming>)
- Kaggle Tutorial (<https://www.kaggle.com/mrisdal/titanic/exploring-survival-on-the-titanic>)