R-Ladies

# R from scratch

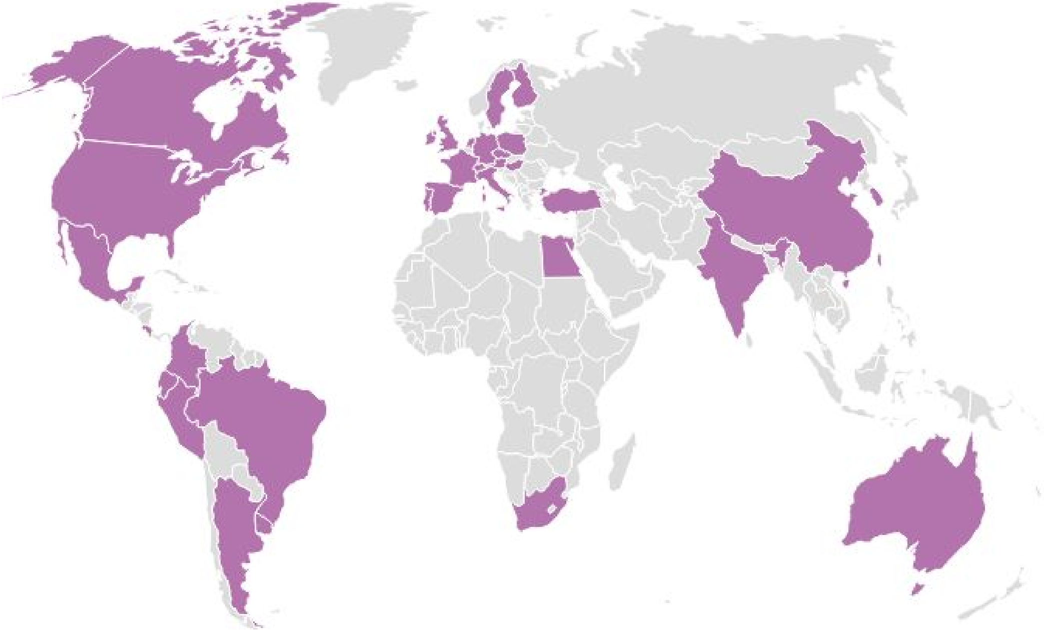
**Kasia Kulma & Claudia Vitolo**

# 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/



- 45+ 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

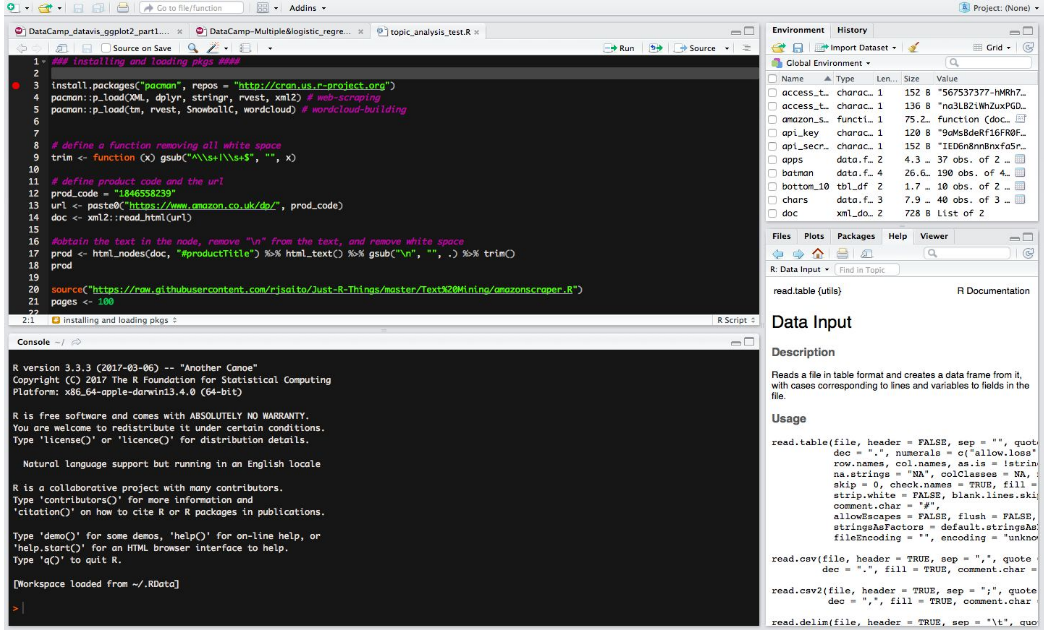
- Freeware, open - source & platform - independent
- **HUGE** choice of analytics and statistics libraries
- Integrates many data - sources
- Efficient data structures (data.frame, tibble, data.table, etc.)
- Produces arguably best Data Visualizations
- Can build interactive web apps (Shiny)
- **Easy**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

# Let’s learn some R!

- Installing R and RStudio
- Projects and Data Pipeline
- 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



```r
### installing and loading pkgs ####

install.packages("pacman", repos = "http://cran.us.r-project.org")
pacman::p_load(XML, dplyr, stringr, rvest, xml2) # web-scraping
pacman::p_load(tm, rvest, SnowballC, wordcloud) # wordcloud-building


# define a function removing all white space
trim <- function (x) gsub("^\\s+|\\s+$", "", x)

# define product code and the url
prod_code = "1846558239"
url <- paste0("https://www.amazon.co.uk/dp/", prod_code)
doc <- xml2::read_html(url)

#obtain the text in the node, remove "\n" from the text, and remove white space
prod <- html_nodes(doc, "#productTitle") %>% html_text() %>% gsub("\n", "", .) %>% trim()
prod

source("https://raw.githubusercontent.com/rjsaito/Just-R-Things/master/Text%20Mining/amazonscraper.R")
pages <- 100
```

```
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]

>
```

| □ | Name | Type | Len... | Size | Value |
|---|---|---|---|---|---|
| □ | access_t... | charac... | 1 | 152 B | "567537377-hMRh7... |
| □ | access_t... | charac... | 1 | 136 B | "na3LB2iWhZuxPGD... |
| □ | amazon_s... | functi... | 1 | 75.2... | function (doc... |
| □ | api_key | charac... | 1 | 120 B | "9aMsBdeRf16FR0F... |
| □ | api_secr... | charac... | 1 | 152 B | "IED6n8nnBnxfa5r... |
| □ | 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 | xml_do... | 2 | 728 B | List of 2 |

read.table {utils} R Documentation

Data Input

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.

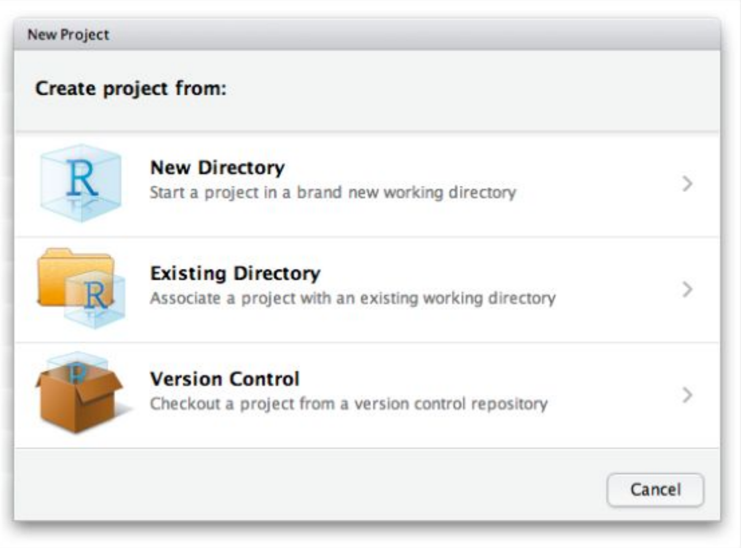
Usage

```
read.table(file, header = FALSE, sep = "", quot
           dec = ".", numerals = c("allow.loss"
           row.names, col.names, as.is = !strin
           na.strings = "NA", colClasses = NA, 
           skip = 0, check.names = TRUE, fill =
           strip.white = FALSE, blank.lines.ski
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAs
           fileEncoding = "", encoding = "unkno

read.csv(file, header = TRUE, sep = ",", quote
         dec = ".", fill = TRUE, comment.char =

read.csv2(file, header = TRUE, sep = ";", quote
          dec = ",", fill = TRUE, comment.char

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

YOUR
TURN!

# RStudio Projects

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


New Project

Create project from:

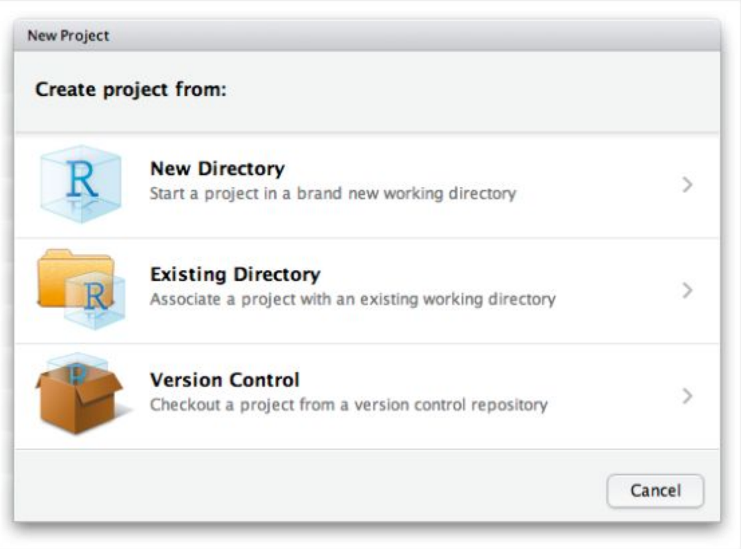
New Directory
Start a project in a brand new working directory

Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository

# RStudio Projects

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



New Project

Create project from:

New Directory
Start a project in a brand new working directory

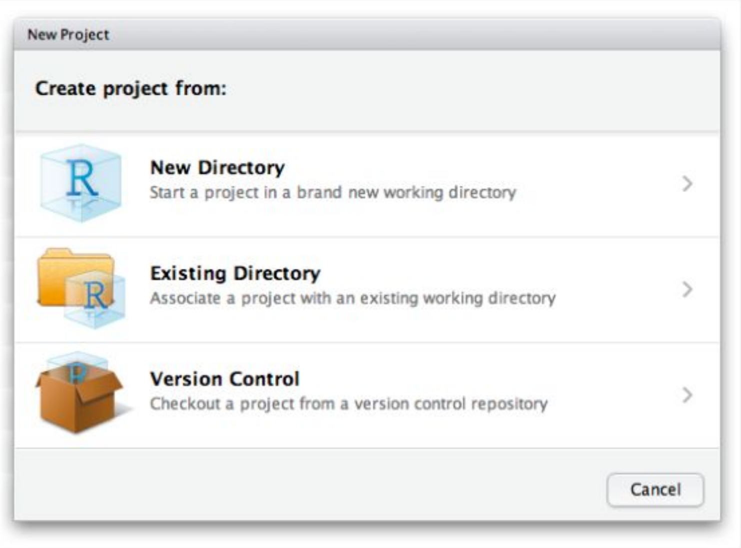
Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository

```
# 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



New Project

Create project from:

**New Directory**
Start a project in a brand new working directory

**Existing Directory**
Associate a project with an existing working directory

**Version Control**
Checkout a project from a version control repository

Cancel

```
# shows current working directory
getwd()
```

~~# changes current working directory~~
~~setwd()~~

YOUR
TURN!

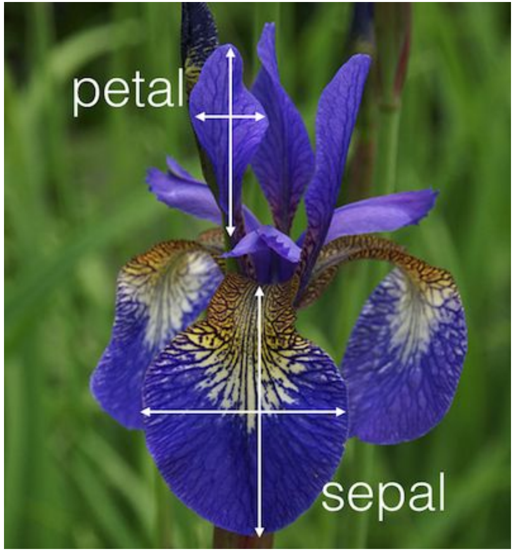
# Data Pipeline in R

- A separate R script for each major step in Data Analysis (e.g. data import, data cleaning, etc.)
- Save the results at every stage as .RData file and load it in the next one

```
save.image() #saves your current workspace as .RData file

save() # saves chosen data objects

load() # loads chosen .RData file
```

# 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
```



YOUR
TURN!

# 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)
```

YOUR
TURN!

# 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

```r
# 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
```

YOUR
TURN!

# Data manipulation & basic Data Visualisation

```r
## 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")
```

YOUR
TURN!

# Data manipulation & basic Data Visualisation

```r
## 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")
```

YOUR
TURN!

# 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)