Introduction to R

Jan-Philipp Kolb

20 Mai, 2019

Introduction round

- Where are you from? What are you studying/working?
- What are your expectations of this course?
- Where do you think you can use Machine Learning in the future?

Preliminaries

- This topic is huge - we concentrate on presenting the application in R
- Usually we have big differences in knowledge and abilities of the participants - please tell, if it is too fast or slow.
- We have many **exercises** because at the end you can only learn on your own
- We have many **examples** - try them out
- If there are questions - always ask
- R is more fun together - ask your neighbor

Content of the course - day 1

- The first section is about laying the foundations in R. We will need all things covered later on.
- The second section is an introduction to the field of machine learning.
- The third part is on regression and classification.

Why R is a good choice ...

- ... because it is an **open source language**
- ... outstanding graphs - **graphics**, **graphics**, **graphics**
- ... relates to other languages - **R can be used in combination with other programs** - e.g. **data linking**
- ...R can be used **for automation**
- ... Vast Community - **you can use the intelligence of other people ;-)**
- ...
- Because of the large comunity
- New statistical methodologies are implemented quite fast
- Because R can be combined with other programs like Postgresql or Python

Constraints

Newer modules in Python

- Machine learning is a field that changes rapidly.
- Some new tools are first developed in Python.
- The package reticulate offers the possibility to use these modules from an R environment.
- Good news - Python is also Open Source

Big Data

- Especially if you work with web data, you quickly have to deal with large amounts of data.
- Therefore one must fall back on databases, which can be used in combination with R.

Content of this part

- Introduction to programming in R

what is relevant for this course.

- How to import data?
- What to do with missing values?
- Parallelization

Import data

```
?read.csv
?read.csv2
```

Using a path to import data

```
path <- 'https://raw.githubusercontent.com/thomaspernet/data_csv
titanic <-read.csv(path)
```

Built in datasets

- A sample dataset is often provided to demonstrate the functionality of a package.
- These records can be loaded using the `data` command.

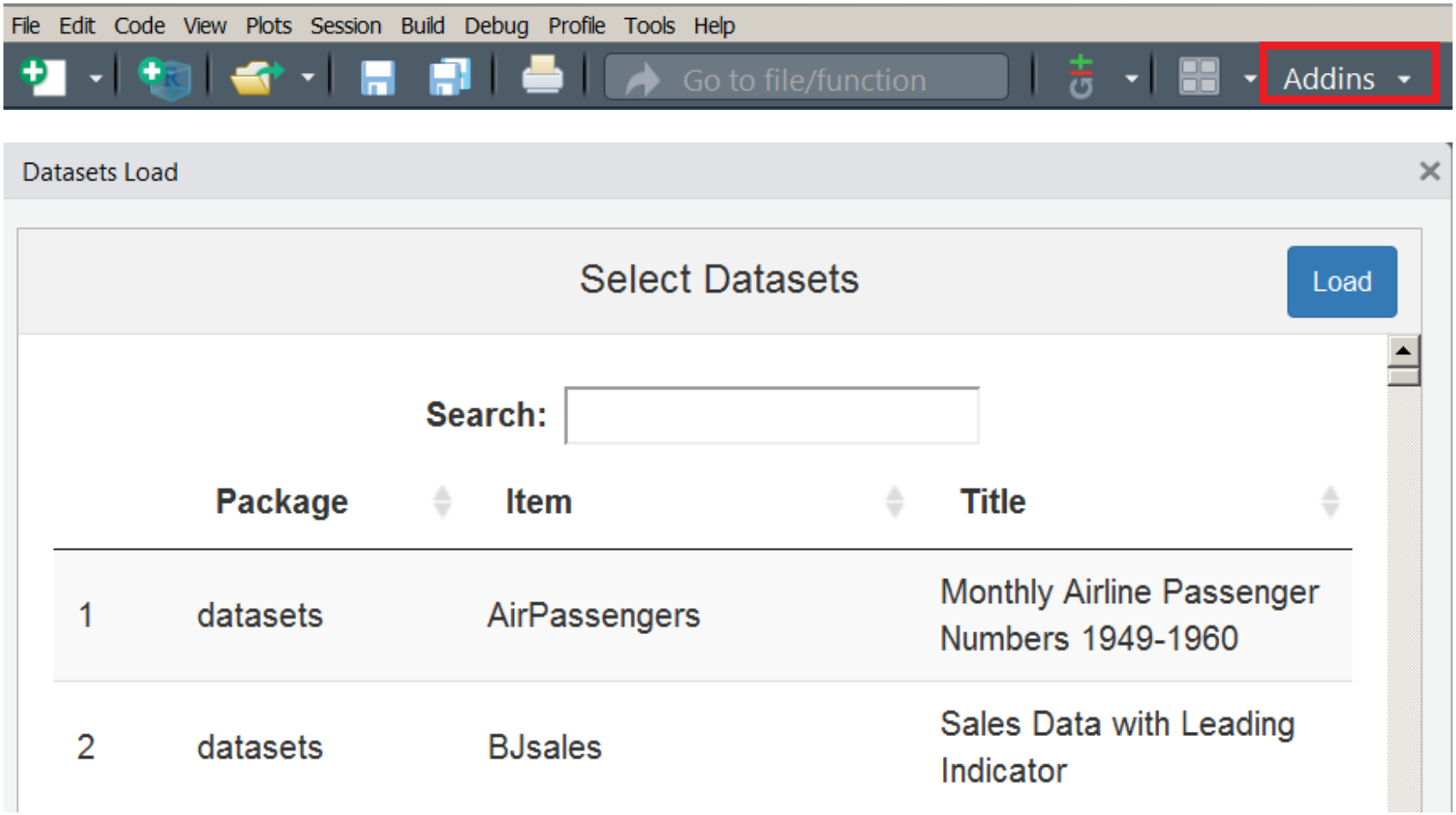
```
data(iris)
```

- There is also a **RStudio Add-In** that helps to find a built-in dataset.

```
install.packages("datasets.load")
```

Exkurs RStudio Addins

- Oben rechts befindet sich ein Button Addins



The titanic dataset

```
kable(head(titanic))
```

X	pclass	survived	name	sex
1	1	1	Allen, Miss. Elisabeth Walton	fema
2	1	1	Allison, Master. Hudson Trevor	male
3	1	0	Allison, Miss. Helen Loraine	fema
4	1	0	Allison, Mr. Hudson Joshua Creighton	male
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	fema
6	1	1	Anderson, Mr. Harry	male

Exercise

Load the the built-in dataset `swiss` and answer the following questions:

- How many observations and variables are available?
- What is the scale level of the variables?

Create an interactive data table

The function scan to import data

```
?scan
```

The R-package data.table

Get an overview

```
data(airquality)
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Overview with data.table

```
library(data.table)
airq <- data.table(airquality)
airq
```

Exercise

```
x <- c(0.109, 0.359, 0.63, 0.996, 0.515, 0.142, 0.017, 0.829, 0.
```

- Compute the logarithm of `x`, return suitably lagged and iterated differences,
- compute the exponential function and round the result

```
## [1]  3.3  1.8  1.6  0.5  0.3  0.1 48.8  1.1
```

The pipe operator

```
library(magrittr)

# Perform the same computations on `x` as above
x %>% log() %>%
    diff() %>%
    exp() %>%
    round(1)
```

```
## [1]  3.3  1.8  1.6  0.5  0.3  0.1 48.8  1.1
```

How to deal with missing values

```
?na.omit

airq

##       Ozone Solar.R Wind Temp Month Day
##   1:     41     190  7.4   67     5   1
##   2:     36     118  8.0   72     5   2
##   3:     12     149 12.6   74     5   3
##   4:     18     313 11.5   62     5   4
##   5:     NA      NA 14.3   56     5   5
##  ---
## 149:     30     193  6.9   70     9  26
## 150:     NA     145 13.2   77     9  27
## 151:     14     191 14.3   75     9  28
## 152:     18     131  8.0   76     9  29
## 153:     20     223 11.5   68     9  30

na.omit(airq)
```

Clean the titanic data set

```
library(dplyr)
library(magrittr)
# Drop variables
clean_titanic <- titanic %>%    mutate(pclass = factor(pclass, l
    survived = factor(survived, levels = c(0, 1), labels = c('No
na.omit()
```

- `pclass = factor(pclass, levels = c(1,2,3), labels= c('Upper', 'Middle', 'Lower'))`: Add label to the variable pclass. 1 becomes Upper, 2 becomes MIddle and 3 becomes lower
- `factor(survived, levels = c(0,1), labels = c('No', 'Yes'))`: Add label to the variable survived. 1 Becomes No and 2 becomes Yes
- `na.omit()`: Remove the NA observations

Get an overview of the data

```
glimpse(clean_titanic)
```

```
## Observations: 1,045
## Variables: 13
## $ X           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,
## $ pclass      <fct> Upper, Upper, Upper, Upper, Upper, Upper, U
## $ survived    <fct> Yes, Yes, No, No, No, Yes, Yes, No, Yes, No
## $ name        <fct> "Allen, Miss. Elisabeth Walton", "Allison,
## $ sex         <fct> female, male, female, male, female, male, f
## $ age         <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000,
## $ sibsp       <int> 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0, 0, 0
## $ parch       <int> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ ticket      <fct> 24160, 113781, 113781, 113781, 113781, 1995
## $ fare        <dbl> 211.3375, 151.5500, 151.5500, 151.5500, 151
## $ cabin       <fct> B5, C22 C26, C22 C26, C22 C26, C22 C26, E12
## $ embarked    <fct> S, S, S, S, S, S, S, S, S, C, C, C, C, S, S
## $ home.dest <fct> "St Louis, MO", "Montreal, PQ / Chestervill
```

Example Data - Housing Values in Suburbs of Boston

```
library(MASS)
data <- Boston

kable(head(data))
```

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptra
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	1
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	1
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	1
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	1
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	1
0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	1

Normalize your data

```
maxs <- apply(data, 2, max)
mins <- apply(data, 2, min)
scaled <- as.data.frame(scale(data, center = mins, scale = maxs
```

The command sample

- We can use this command to draw a sample.
- We need the command later to split our dataset into a test and a training dataset.

```
sample(1:10,3,replace=T)
```

```
## [1] 1 8 7
```

```
sample(1:10,3,replace=T)
```

```
## [1] 2 1 6
```

Set a seed

- `set.seed` is the recommended way to specify seeds.
- If we set a seed, we get the same result for random events.
- This function is mainly required for simulations.

```
set.seed(234)
sample(1:10,3,replace=T)
```

```
## [1] 8 8 1
```

```
set.seed(234)
sample(1:10,3,replace=T)
```

```
## [1] 8 8 1
```

Time measurement

```
start_time <- Sys.time()
ab <- runif(10000000)
end_time <- Sys.time()

end_time - start_time

## Time difference of 0.382 secs
```

How many cores are available

```
library(doParallel)
detectCores()
```

```
## [1] 4
```

Make cluster

```
cl <- makeCluster(detectCores())
registerDoParallel(cl)

start_time <- Sys.time()
ab <- runif(10000000)
end_time <- Sys.time()

end_time - start_time


## Time difference of 0.313 secs

stopCluster(cl)

?parallel::makeCluster
```

The swirl package

```
install.packages("swirl")

library("swirl")
swirl()
```

Resources

- Course materials for the Data Science Specialization