Przetwarzanie i wizualizacja danych

prof. UAM dr hab. Tomasz Górecki

tomasz.gorecki@amu.edu.pl

Zakład Statystyki Matematycznej i Analizy Danych
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Warunki zaliczenia

W ciągu ćwiczeń zostaną przeprowadzone 2 kolokwia. Na każdym z nich będzie do zdobycia 50 punktów. W ramach laboratorium będą do przygotowania 3 projekty (w sumie również 100 punktów). Należy zaliczyć zarówno ćwiczenia (od 50 punktów) jak i laboratorium (od 60 punktów). Ocena końcowa z egzaminu będzie wystawiana na bazie sumy uzyskanych punktów zarówno z ćwiczeń (60%) jak i laboratorium (40%).

Literatura


Biecek, P. (2016). *Odkrywać! Ujawniać! Objaśniać! Zbiór esejów o sztuce przedstawiania danych.* Fundacja Naukowa SmarterPoland.pl.


Biecek, P. (2017). *Przewodnik po pakiecie R*. GiS.


Crawley, M.J. (2012). *The R Book*. Wiley.


Gągolewski, M. (2014). *Programowanie w języku R*. PWN.


Górecki, T. (2011). *Podstawy statystyki z przykładami w R*. BTC.


Lander, J.P. (2018). *Język R dla każdego. Zaawansowane analizy i grafika statystyczna.* APN Promise.

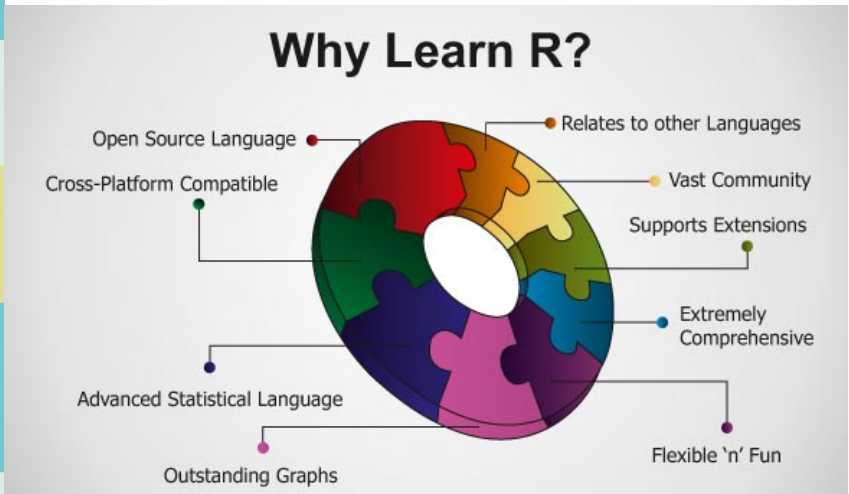

Winke, C.O. (2020). *Podstawy wizualizacji danych. Zasady tworzenia atrakcyjnych wykresów.* Helion.

WHAT IS R?

R is a scripting language for statistical data manipulation and analysis. It was inspired by, and is mostly compatible with, the statistical language S developed by AT&T. The name S, obviously standing for statistics, was an allusion to another programming language developed at AT&T with a one-letter name, C. S later was sold to a small firm, which added a GUI interface and named the result S-Plus.

WHY USE R FOR YOUR STATISTICAL WORK?

1. A public-domain implementation of the widely-regarded statistical language; R/S is the de facto standard among professional statisticians
2. comparable, and often superior, in power to commercial products in most senses
3. available for Windows, Macs, Linux
4. in addition to enabling statistical operations, it's a general programming language, so that you can automate your analyses and create new functions
5. object-oriented and functional programming structure
6. your data sets are saved between sessions, so you don't have to reload each time
7. open-software nature means it's easy to get help from the user community, and lots of new functions get contributed by users, many of which are prominent statisticians



R jest zaawansowanym pakietem statystycznym jak również językiem programowania istniejącym na platformy Windows, Linux oraz MacOS. Objęty jest licencją GNU GPL i oparty na języku S. Język R jest językiem interpretowanym, a nie kompilowanym. Z tego względu program w nim napisany nie będzie tak szybki jak np. program napisany w C++. Można jednak wykorzystać funkcje z bibliotek napisanych np. w C, C++, Python czy Fortran. Można również używać funkcji R w Javie, Pythonie czy C++, jak również w VB for Statistica.

Pakiet umożliwia również tworzenie zaawansowanych wykresów, które mogą zostać zapisane w formatach takich jak PDF i JPG. O sile R stanowi ponad 16 000 bibliotek, przeznaczonych do najróżniejszych zastosowań. Całość wzbogacona jest kompleksową dokumentacją. Dodatkowo bardzo użyteczną cechą R jest dostępność zbiorów danych praktycznie do każdego zagadnienia.

RStudio

RStudio jest zintegrowanym środowiskiem programowania (IDE) dla R. Dostępne jest w dwóch edycjach: RStudio Desktop oraz RStudio Server. Dostępne jest dla Windows, macOS oraz Linuxa. Występuje wersja darmowa (open source) oraz komercyjna. Zostało napisane głównie w C++. Pierwsza wersja beta powstała lutym 2011 roku. Wersja 1.0 została wydana 1 listopada 2016 roku.

RStudio

RStudio IDE Cheat Sheet

learn more at www.rstudio.com

RStudio

The RStudio IDE is an Integrated Development Environment in R that comes in three versions

Desktop IDE
A local version of the IDE for your desktop

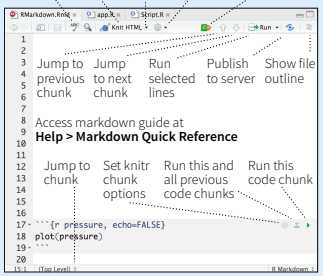
Open Source Server
for larger compute resources and remote access

Professional Server
for teams that share large compute resources, large data, and uniform environments for collaboration

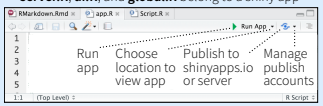
Download all at **www.rstudio.com**. Each provides the same useful interface:

Documents and Apps

Open Shiny, R Markdown, knitr, Sweave, LaTeX, .Rd files and more in Source Pane

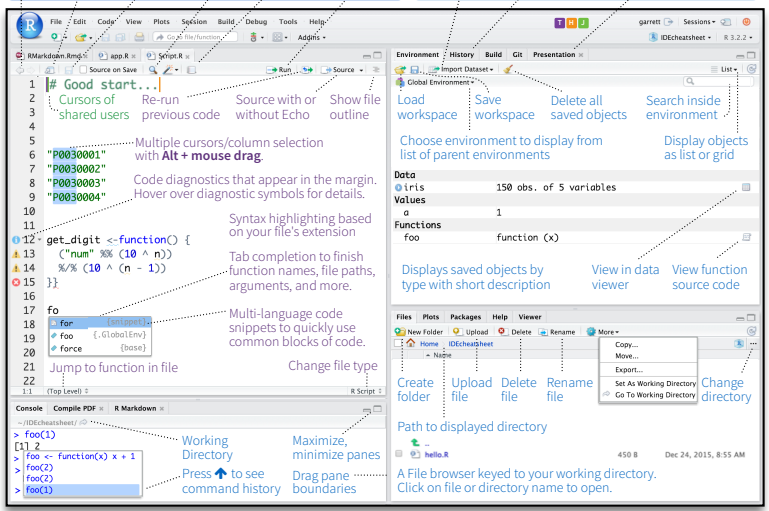


RStudio recognizes that files named **app.R**, **server.R**, **ui.R**, and **global.R** belong to a shiny app



Write Code

Navigate tabs | Open in new window | Save | Find and replace | Compile as notebook | Run selected code

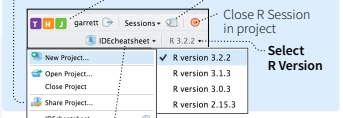


R Support

Import data file with wizard | History of past commands to run/add to source | Display .RPres slideshows **File > New File > R Presentation**

RStudio Pro Features

Share Project with Collaborators | Active shared collaborators | Start **new R Session** in current project



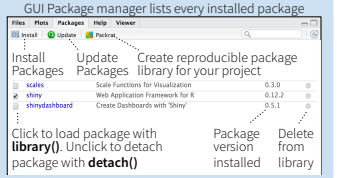
Project System

File > New Project

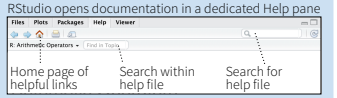
RStudio saves the call history, workspace, and working directory associated with a project. It reloads each when you re-open a project.

RStudio opens plots in a dedicated Plots pane

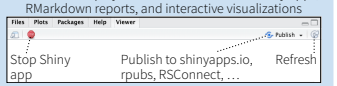
GUI Package manager lists every installed package



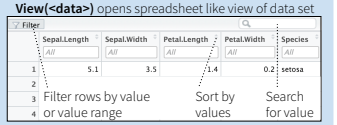
RStudio opens documentation in a dedicated Help pane



Viewer Pane displays HTML content, such as Shiny apps, RMarkdown reports, and interactive visualizations



View(<data>) opens spreadsheet like view of data set

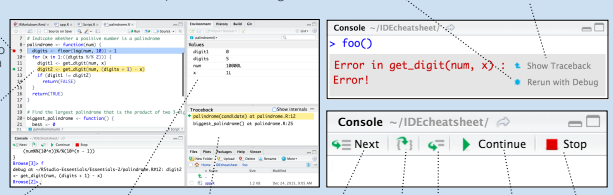


Debug Mode

Open with **debug()**, **browse()**, or a breakpoint. RStudio will open the debugger mode when it encounters a breakpoint while executing code.

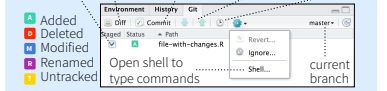
Launch debugger mode from origin of error

Open traceback to examine the functions that R called before the error occurred



Version Control with Git or SVN

Turn on at **Tools > Project Options > Git/SVN**

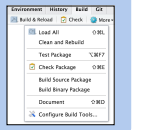


Package Writing

File > New Project > New Directory > R Package

Turn project into package,
Enable roxygen documentation with **Tools > Project Options > Build Tools**

Roxygen guide at **Help > Roxygen Quick Reference**



More cheat sheets at **http://www.rstudio.com/resources/cheatsheets/**

Learn more at **support.rstudio.com** • RStudio IDE 0.99.832 • Updated: 01/16

RStudio

1 LAYOUT	Windows/Linux	Mac
Move focus to Source Editor	Ctrl+1	Ctrl+1
Move focus to Console	Ctrl+2	Ctrl+2
Move focus to Help	Ctrl+3	Ctrl+3
Show History	Ctrl+4	Ctrl+4
Show Files	Ctrl+5	Ctrl+5
Show Plots	Ctrl+6	Ctrl+6
Show Packages	Ctrl+7	Ctrl+7
Show Environment	Ctrl+8	Ctrl+8
Show Git/SVN	Ctrl+9	Ctrl+9
Show Build	Ctrl+0	Ctrl+0

2 RUN CODE	Windows/Linux	Mac
Search command history	Ctrl+↑	Cmd+↑
Navigate command history	↑/↓	↑/↓
Move cursor to start of line	Home	Cmd+←
Move cursor to end of line	End	Cmd+→
Change working directory	Ctrl+Shift+H	Ctrl+Shift+H
Interrupt current command	Esc	Esc
Clear console	Ctrl+L	Ctrl+L
Quit Session (desktop only)	Ctrl+Q	Cmd+Q
Restart R Session	Ctrl+Shift+F10	Cmd+Shift+F10
Run current line/selection	Ctrl+Enter	Cmd+Enter
Run current (retain cursor)	Alt+Enter	Option+Enter
Run from current to end	Ctrl+Alt+E	Cmd+Option+E
Run the current function	Ctrl+Alt+F	Cmd+Option+F
Source a file	Ctrl+Shift+O	Cmd+Shift+O
Source the current file	Ctrl+Shift+S	Cmd+Shift+S
Source with echo	Ctrl+Shift+Enter	Cmd+Shift+Enter

3 NAVIGATE CODE	Windows /Linux	Mac
Goto File/Function	Ctrl+.	Ctrl+.
Fold Selected	Alt+L	Cmd+Option+L
Unfold Selected	Shift+Alt+L	Cmd+Shift+Option+L
Fold All	Alt+O	Cmd+Option+O
Unfold All	Shift+Alt+O	Cmd+Shift+Option+O
Go to line	Shift+Alt+G	Cmd+Shift+Option+G
Jump to	Shift+Alt+J	Cmd+Shift+Option+J
Switch to tab	Ctrl+Shift+.	Ctrl+Shift+.
Previous tab	Ctrl+F11	Ctrl+F11
Next tab	Ctrl+F12	Ctrl+F12
First tab	Ctrl+Shift+F11	Ctrl+Shift+F11
Last tab	Ctrl+Shift+F12	Ctrl+Shift+F12
Navigate back	Ctrl+F9	Cmd+F9
Navigate forward	Ctrl+F10	Cmd+F10
Jump to Brace	Ctrl+P	Ctrl+P
Select within Braces	Ctrl+Shift+Alt+E	Ctrl+Shift+Alt+E
Use Selection for Find	Ctrl+F3	Cmd+E
Find in Files	Ctrl+Shift+F	Cmd+Shift+F
Find Next	Win: F3, Linux: Ctrl+G	Cmd+G
Find Previous	W: Shift+F3, L: Ctrl+Shift	Cmd+Shift+G
Jump to Word	Ctrl+ ←/→	Option+ ←/→
Jump to Start/End	Ctrl+↑/↓	Cmd+↑/↓

4 WRITE CODE	Windows /Linux	Mac
Attempt completion	Tab or Ctrl+Space	Tab or Cmd+Space
Navigate candidates	↑/↓	↑/↓
Accept candidate	Enter, Tab, or →	Enter, Tab, or →
Dismiss candidates	Esc	Esc
Undo	Ctrl+Z	Cmd+Z
Redo	Ctrl+Shift+Z	Cmd+Shift+Z
Cut	Ctrl+X	Cmd+X
Copy	Ctrl+C	Cmd+C
Paste	Ctrl+V	Cmd+V
Select All	Ctrl+A	Cmd+A
Delete Line	Ctrl+D	Cmd+D
Select	Shift+[Arrow]	Shift+[Arrow]
Select Word	Ctrl+Shift+ ←/→	Option+Shift+ ←/→
Select to Line Start	Alt+Shift+←	Cmd+Shift+←
Select to Line End	Alt+Shift+→	Cmd+Shift+→
Select Page Up/Down	Shift+PageUp/Down	Shift+PageUp/Down
Select to Start/End	Shift+Alt+↑/↓	Cmd+Shift+↑/↓
Delete Word Left	Ctrl+Backspace	Ctrl+Opt+Backspace
Delete Word Right		Option+Delete
Delete to Line End		Ctrl+K
Delete to Line Start		Option+Backspace
Indent	Tab (at start of line)	Tab (at start of line)
Outdent	Shift+Tab	Shift+Tab
Yank line up to cursor	Ctrl+U	Ctrl+U
Yank line after cursor	Ctrl+K	Ctrl+K
Insert yanked text	Ctrl+Y	Ctrl+Y
Insert <-	Alt+-	Option+-
Insert %>%	Ctrl+Shift+M	Cmd+Shift+M
Show help for function	F1	F1
Show source code	F2	F2
New document	Ctrl+Shift+N	Cmd+Shift+N
New document (Chrome)	Ctrl+Alt+Shift+N	Cmd+Shift+Alt+N
Open document	Ctrl+O	Cmd+O
Save document	Ctrl+S	Cmd+S
Close document	Ctrl+W	Cmd+W
Close document (Chrome)	Ctrl+Alt+W	Cmd+Option+W
Close all documents	Ctrl+Shift+W	Cmd+Shift+W
Extract function	Ctrl+Alt+X	Cmd+Option+X
Extract variable	Ctrl+Alt+V	Cmd+Option+V
Reindent lines	Ctrl+I	Cmd+I
(Un)Comment lines	Ctrl+Shift+C	Cmd+Shift+C
Reflow Comment	Ctrl+Shift+/	Cmd+Shift+/
Reformat Selection	Ctrl+Shift+A	Cmd+Shift+A
Select within braces	Ctrl+Shift+E	Ctrl+Shift+E
Show Diagnostics	Ctrl+Shift+Alt+P	Cmd+Shift+Alt+P
Transpose Letters		Ctrl+T
Move Lines Up/Down	Alt+↑/↓	Option+↑/↓
Copy Lines Up/Down	Shift+Alt+↑/↓	Cmd+Option+↑/↓
Add New Cursor Above	Ctrl+Alt+Up	Ctrl+Alt+Up
Add New Cursor Below	Ctrl+Alt+Down	Ctrl+Alt+Down
Move Active Cursor Up	Ctrl+Alt+Shift+Up	Ctrl+Alt+Shift+Up
Move Active Cursor Down	Ctrl+Alt+Shift+Down	Ctrl+Alt+Shift+Down
Find and Replace	Ctrl+F	Cmd+F
Use Selection for Find	Ctrl+F3	Cmd+E
Replace and Find	Ctrl+Shift+J	Cmd+Shift+J

5 DEBUG CODE	Windows/Linux	Mac
Toggle Breakpoint	Shift+F9	Shift+F9
Execute Next Line	F10	F10
Step Into Function	Shift+F4	Shift+F4
Finish Function/Loop	Shift+F6	Shift+F6
Continue	Shift+F5	Shift+F5
Stop Debugging	Shift+F8	Shift+F8

6 VERSION CONTROL	Windows/Linux	Mac
Show diff	Ctrl+Alt+D	Ctrl+Option+D
Commit changes	Ctrl+Alt+M	Ctrl+Option+M
Scroll diff view	Ctrl+↑/↓	Ctrl+↑/↓
Stage/Unstage (Git)	Spacebar	Spacebar
Stage/Unstage and move to next	Enter	Enter

7 MAKE PACKAGES	Windows/Linux	Mac
Build and Reload	Ctrl+Shift+B	Cmd+Shift+B
Load All (devtools)	Ctrl+Shift+L	Cmd+Shift+L
Test Package (Desktop)	Ctrl+Shift+T	Cmd+Shift+T
Test Package (Web)	Ctrl+Alt+F7	Cmd+Alt+F7
Check Package	Ctrl+Shift+E	Cmd+Shift+E
Document Package	Ctrl+Shift+D	Cmd+Shift+D

8 DOCUMENTS AND APPS	Windows/Linux	Mac
Preview HTML (Markdown, etc.)	Ctrl+Shift+K	Cmd+Shift+K
Knit Document (knitr)	Ctrl+Shift+K	Cmd+Shift+K
Compile Notebook	Ctrl+Shift+K	Cmd+Shift+K
Compile PDF (TeX and Sweave)	Ctrl+Shift+K	Cmd+Shift+K
Insert chunk (Sweave and Knitr)	Ctrl+Alt+I	Cmd+Option+I
Insert code section	Ctrl+Shift+R	Cmd+Shift+R
Re-run previous region	Ctrl+Shift+P	Cmd+Shift+P
Run current document	Ctrl+Alt+R	Cmd+Option+R
Run from start to current line	Ctrl+Alt+B	Cmd+Option+B
Run the current code section	Ctrl+Alt+T	Cmd+Option+T
Run previous Sweave/Rmd code	Ctrl+Alt+P	Cmd+Option+P
Run the current chunk	Ctrl+Alt+C	Cmd+Option+C
Run the next chunk	Ctrl+Alt+N	Cmd+Option+N
Sync Editor & PDF Preview	Ctrl+F8	Cmd+F8

Previous plot	Ctrl+Alt+F11	Cmd+Option+F11
Next plot	Ctrl+Alt+F12	Cmd+Option+F12

Show Keyboard Shortcuts	Alt+Shift+K	Option+Shift+K

Why RStudio Server Pro?

Do everything you would do with the open source server with a commercial license, support, and more.

- edit the same project at the same time as others
- switch easily from one version of R to a different version
- open and run multiple R sessions simultaneously
- see what you and others are doing on your server
- tune your resources to improve performance
- integrate with your authentication, authorization, and audit practices

Download a free 45 day evaluation at
www.rstudio.com/products/rstudio-server-pro/

 More cheat sheets at **http://www.rstudio.com/resources/cheatsheets/** Learn more at **support.rstudio.com** • RStudio IDE 0.99.832 • Updated: 01/16

Pierwsze kroki

Pomoc

help() lub ?

Jeśli w nazwie znajdują się pewne znaki szczególne lub słowa kluczowe języka np. takie jak if, for, function, to nazwę funkcji wpisujemy w cudzysłowie. W przypadku gdy nie pamiętamy nazwy funkcji ale znamy temat wpisujemy:

help.search()

Pomoc na temat pakietu uzyskamy wpisując:

library(help = nazwa_pakietu)

Pierwsze kroki

Komentarz

#

Operator przypisania

=, <-, ->

Pakiety

Instalacja

```
install.packages(nazwa_pakietu, dependencies = T)
```

Ładowanie

```
library(nazwa_pakietu)
```

Usunięcie

```
detach(package:nazwa_pakietu)
```

Struktury danych – typy proste

- numeryczny,
- zespolony,
- logiczny – jedna z dwóch wartości prawda (TRUE/T) lub fałsz (FALSE/F),
- znakowy – napisy, łańcuchy znaków. Powinny być zawarte pomiędzy znakami ' lub ". W łańcuchu można używać znaków sterujących, które poprzedzone są \ (np. \n – nowa linia, \t – tabulator itd.).

Podstawowe funkcje operujące na napisach

chartr(stary, nowy, napis)	Zamienia określone znaki na inne
grep(wyrażenie regularne, dane)	Poszukuje wystąpień wyrażenia w danych
nchar(napis)	Liczba znaków
paste(napis1, napis2)	Łączy napisy
strsplit(napis, wyrażenie regularne)	Dzieli napis
strtrim(napis, ile)	Obcina napis
sub(wyrażenie regularne, nowy tekst, napis)	Zamienia pierwsze wystąpienie łańcucha (funkcja gsub zamienia wszystkie wystąpienia)
tolower(napis)	Zamiana znaków na małe
toupper(napis)	Zamiana znaków na wielkie

Wyrażenia regularne

Basic Regular Expressions in R
Cheat Sheet

Character Classes

[[:digit:]] or \\d	Digits; `[0-9]`
\\D	Non-digits; `[^0-9]`
[[:lower:]]	Lower-case letters; `[a-z]`
[[:upper:]]	Upper-case letters; `[A-Z]`
[[:alpha:]]	Alphabetic characters; `[A-z]`
[[:alnum:]]	Alphanumeric characters `[A-z0-9]`
\\w	Word characters; `[A-z0-9_]`
\\W	Non-word characters
[[:xdigit:]] or \\x	Hexadec. digits; `[0-9A-Fa-f]`
[[:blank:]]	Space and tab
[[:space:]] or \\s	Space, tab, vertical tab, newline, form feed, carriage return
\\S	Not space; `[^[:space:]]`
[[:punct:]]	Punctuation characters; `!"#$%&'()*+,-./:;<=>?@[]^_`{|}~`
[[:graph:]]	Graphical char.; `[[:alnum:][:punct:]]`
[[:print:]]	Printable characters; `[[:alnum:][:punct:]\\s]`
[[:cntrl:]] or \\c	Control characters; `\n`, `\r` etc.

Special Metacharacters

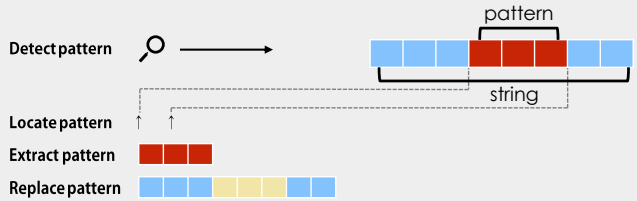
\n	New line
\r	Carriage return
\t	Tab
\v	Vertical tab
\f	Form feed

Lookaraounds and Conditionals*

(?=)	Lookahead (requires PERL = TRUE), e.g. `(?=yx)`: position followed by 'xy'
(?!)	Negative lookahead (PERL = TRUE); position NOT followed by pattern
(?<=)	Lookbehind (PERL = TRUE), e.g. `(?<=yx)`: position following 'xy'
(?<!)	Negative lookbehind (PERL = TRUE); position NOT following pattern
?(if)then	If-then-condition (PERL = TRUE); use lookaheads, optional char. etc in if-clause
?(if)then|else	If-then-else-condition (PERL = TRUE)

*see, e.g. http://www.regular-expressions.info/lookaround.html
http://www.regular-expressions.info/conditional.html

Functions for Pattern Matching




```
> string <- c("Hiphopopotamus", "Rhymenoceros", "time for bottomless lyrics")
> pattern <- "t.m"
```

Detect Patterns

grep(pattern, string)
`[1] 1 3`

grep(pattern, string, value = TRUE)
`[1] "Hiphopopotamus"`
`[2] "time for bottomless lyrics"`

grepl(pattern, string)
`[1] TRUE FALSE TRUE`

stringr::str_detect(string, pattern)
`[1] TRUE FALSE TRUE`

Locate Patterns

regexpr(pattern, string)
find starting position and length of first match

gregexpr(pattern, string)
find starting position and length of all matches

stringr::**str_locate(string, pattern)**
find starting and end position of first match

stringr::**str_locate_all(string, pattern)**
find starting and end position of all matches

Split a String using a Pattern

strsplit(string, pattern) or stringr::**str_split(string, pattern)**

Extract Patterns

regmatches(string, regexpr(pattern, string))
extract first match `[1] "tam" "tim"`

regmatches(string, gregexpr(pattern, string))
extracts all matches, outputs a list
`[[1]] "tam" [[2]] character(0) [[3]] "tim" "tom"`

stringr::**str_extract(string, pattern)**
extract first match `[1] "tam" NA "tim"`

stringr::**str_extract_all(string, pattern)**
extract all matches, outputs a list

stringr::**str_extract_all(string, pattern, simplify = TRUE)**
extract all matches, outputs a matrix

stringr::**str_match(string, pattern)**
extract first match + individual character groups

stringr::**str_match_all(string, pattern)**
extract all matches + individual character groups

Replace Patterns

sub(pattern, replacement, string)
replace first match

gsub(pattern, replacement, string)
replace all matches

stringr::**str_replace(string, pattern, replacement)**
replace first match

stringr::**str_replace_all(string, pattern, replacement)**
replace all matches

Character Classes and Groups

.	Any character except \n
|	Or, e.g. `(a|b)`
[…]	List permitted characters, e.g. `[abc]`
[a-z]	Specify character ranges
[^…]	List excluded characters
(…)	Grouping, enables back referencing using \\N where N is an integer

Anchors

^	Start of the string
$	End of the string
\\b	Empty string at either edge of a word
\\B	NOT the edge of a word
\\<	Beginning of a word
\\>	End of a word

Quantifiers

*	Matches at least 0 times
+	Matches at least 1 time
?	Matches at most 1 time; optional string
{n}	Matches exactly n times
{n,}	Matches at least n times
{,n}	Matches at most n times
{n,m}	Matches between n and m times

General Modes

By default R uses *POSIX extended regular expressions*. You can switch to *PCRE regular expressions* using `PERL = TRUE` for base or by wrapping patterns with `perl()` for stringr.

All functions can be used with literal searches using `fixed = TRUE` for base or by wrapping patterns with `fixed()` for stringr.

All base functions can be made case insensitive by specifying `ignore.cases = TRUE`.

Escaping Characters

Metacharacters (`. * +` etc.) can be used as literal characters by escaping them. Characters can be escaped using `\\` or by enclosing them in `\\Q...\\E`.

Case Conversions

Regular expressions can be made case insensitive using `(?i)`. In backreferences, the strings can be converted to lower or upper case using `\\L` or `\\U` (e.g. `\\L\\1`). This requires `PERL = TRUE`.

Greedy Matching

By default the asterisk * is greedy, i.e. it always matches the longest possible string. It can be used in lazy mode by adding `?`, i.e. `*?`.

Greedy mode can be turned off using `(?U)`. This switches the syntax, so that `(?U)a*` is lazy and `(?U)a*?` is greedy.

Note

Regular expressions can conveniently be created using rex::**rex()**.

Updated: 09/16

Wyrażenia regularne

Work with strings with stringr : : CHEAT SHEET

The **stringr** package provides a set of internally consistent tools for working with character strings, i.e. sequences of characters surrounded by quotation marks.

Detect Matches

str_detect(string, **pattern**) Detect the presence of a pattern match in a string.
str_detect(fruit, "a")

str_which(string, **pattern**) Find the indexes of strings that contain a pattern match.
str_which(fruit, "a")

str_count(string, **pattern**) Count the number of matches in a string.
str_count(fruit, "a")

str_locate(string, **pattern**) Locate the positions of pattern matches in a string. Also **str_locate_all**. *str_locate(fruit, "a")*

Subset Strings

str_sub(string, start = 1L, end = -1L) Extract substrings from a character vector.
str_sub(fruit, 1, 3); str_sub(fruit, -2)

str_subset(string, **pattern**) Return only the strings that contain a pattern match.
str_subset(fruit, "b")

str_extract(string, **pattern**) Return the first pattern match found in each string, as a vector. Also **str_extract_all** to return every pattern match. *str_extract(fruit, "[aeiou]")*

str_match(string, **pattern**) Return the first pattern match found in each string, as a matrix with a column for each ( ) group in pattern. Also **str_match_all**.
str_match(sentences, "(a|the) ([^]+)")

Manage Lengths

str_length(string) The width of strings (i.e. number of code points, which generally equals the number of characters). *str_length(fruit)*

str_pad(string, width, side = c("left", "right", "both"), pad = " ") Pad strings to constant width. *str_pad(fruit, 17)*

str_trunc(string, width, side = c("right", "left", "center"), ellipsis = "...") Truncate the width of strings, replacing content with ellipsis.
str_trunc(fruit, 3)

str_trim(string, side = c("both", "left", "right")) Trim whitespace from the start and/or end of a string. *str_trim(fruit)*

Mutate Strings

str_sub() **<-** value. Replace substrings by identifying the substrings with str_sub() and assigning into the results.
str_sub(fruit, 1, 3) <- "str"

str_replace(string, **pattern**, replacement) Replace the first matched pattern in each string. *str_replace(fruit, "a", "-")*

str_replace_all(string, **pattern**, replacement) Replace all matched patterns in each string. *str_replace_all(fruit, "a", "-")*

str_to_lower(string, locale = "en")[1] Convert strings to lower case.
str_to_lower(sentences)

str_to_upper(string, locale = "en")[1] Convert strings to upper case.
str_to_upper(sentences)

str_to_title(string, locale = "en")[1] Convert strings to title case. *str_to_title(sentences)*

Join and Split

str_c(..., sep = "", collapse = NULL) Join multiple strings into a single string.
str_c(letters, LETTERS)

str_c(..., sep = "", **collapse = NULL**) Collapse a vector of strings into a single string.
str_c(letters, collapse = "")

str_dup(string, times) Repeat strings times times. *str_dup(fruit, times = 2)*

str_split_fixed(string, **pattern**, n) Split a vector of strings into a matrix of substrings (splitting at occurrences of a pattern match). Also **str_split** to return a list of substrings.
str_split_fixed(fruit, " ", n=2)

glue::**glue**(..., .sep = "", .envir = parent.frame(), .open = "{", .close = "}") Create a string from strings and {expressions} to evaluate. *glue::glue("Pi is {pi}")*

glue::**glue_data**(.x, ..., .sep = "", .envir = parent.frame(), .open = "{", .close = "}") Use a data frame, list, or environment to create a string from strings and {expressions} to evaluate. *glue::glue_data(mtcars, "{rownames(mtcars)} has {hp} hp")*

Order Strings

str_order(x, decreasing = FALSE, na_last = TRUE, locale = "en", numeric = FALSE, ...)[1] Return the vector of indexes that sorts a character vector. *x[str_order(x)]*

str_sort(x, decreasing = FALSE, na_last = TRUE, locale = "en", numeric = FALSE, ...)[1] Sort a character vector.
str_sort(x)

Helpers

str_conv(string, encoding) Override the encoding of a string. *str_conv(fruit,"ISO-8859-1")*

str_view(string, **pattern**, match = NA) View HTML rendering of first regex match in each string. *str_view(fruit, "[aeiou]")*

str_view_all(string, **pattern**, match = NA) View HTML rendering of all regex matches.
str_view_all(fruit, "[aeiou]")

str_wrap(string, width = 80, indent = 0, exdent = 0) Wrap strings into nicely formatted paragraphs. *str_wrap(sentences, 20)*

RStudio

[1] See bit.ly/ISO639-1 for a complete list of locales.

Wyrażenia regularne

Need to Know

Pattern arguments in stringr are interpreted as regular expressions *after any special characters have been parsed.*

In R, you write regular expressions as *strings*, sequences of characters surrounded by quotes (**""**) or single quotes(**''**).

Some characters cannot be represented directly in an R string . These must be represented as **special characters**, sequences of characters that have a specific meaning., e.g.

Special Character	Represents
\\	\
\"	"
\n	new line

Run **?"'"** to see a complete list

Because of this, whenever a \ appears in a regular expression, you must write it as \\ in the string that represents the regular expression.

Use **writeLines**() to see how R views your string after all special characters have been parsed.

```
writeLines("\\.")
# \.

writeLines("\\ is a backslash")
# \ is a backslash
```

INTERPRETATION

Patterns in stringr are interpreted as regexs To change this default, wrap the pattern in one of:

regex(pattern, ignore_case = FALSE, multiline = FALSE, comments = FALSE, dotall = FALSE, ...) Modifies a regex to ignore cases, match end of lines as well of end of strings, allow R comments within regex's , and/or to have . match everything including \n.
str_detect("I", regex("i", TRUE))

fixed() Matches raw bytes but will miss some characters that can be represented in multiple ways (fast). *str_detect("\u0130", fixed("i"))*

coll() Matches raw bytes and will use locale specific collation rules to recognize characters that can be represented in multiple ways (slow). *str_detect("\u0130", coll("i", TRUE, locale = "tr"))*

boundary() Matches boundaries between characters, line_breaks, sentences, or words. *str_split(sentences, boundary("word"))*

Regular Expressions -

Regular expressions, or *regexps*, are a concise language for describing patterns in strings.

MATCH CHARACTERS

see <- function(rx) str_view_all("abc ABC 123\t.!?\\(){}\n", rx)

string (type this)	regexp (to mean this)	matches (which matches this)	example	
	a (etc.)	a (etc.)	see("a")	abc ABC 123 .!?\(){}
\\.	\.	.	see("\\.")	abc ABC 123 .!?\(){}
\\!	\!	!	see("\\!")	abc ABC 123 .!?\(){}
\\?	\?	?	see("\\?")	abc ABC 123 .!?\(){}
\\\\	\\	\	see("\\\\")	abc ABC 123 .!?\(){}
\\(	\(	(	see("\\(")	abc ABC 123 .!?\(){}
\\)	\)	)	see("\\)")	abc ABC 123 .!?\(){}
\\{	\{	{	see("\\{")	abc ABC 123 .!?\(){}
\\}	\}	}	see("\\}")	abc ABC 123 .!?\(){}
\\n	\n	new line (return)	see("\\n")	abc ABC 123 .!?\(){}
\\t	\t	tab	see("\\t")	abc ABC 123 .!?\(){}
\\s	\s	any whitespace (**\S** *for non-whitespaces*)	see("\\s")	abc ABC 123 .!?\(){}
\\d	\d	any digit (**\D** *for non-digits*)	see("\\d")	abc ABC 123 .!?\(){}
\\w	\w	any word character (**\W** *for non-word chars*)	see("\\w")	abc ABC 123 .!?\(){}
\\b	\b	word boundaries	see("\\b")	abc ABC 123 .!?\(){}
	[:digit:][1]	digits	see("[:digit:]")	abc ABC 123 .!?\(){}
	[:alpha:][1]	letters	see("[:alpha:]")	abc ABC 123 .!?\(){}
	[:lower:][1]	lowercase letters	see("[:lower:]")	abc ABC 123 .!?\(){}
	[:upper:][1]	uppercase letters	see("[:upper:]")	abc ABC 123 .!?\(){}
	[:alnum:][1]	letters and numbers	see("[:alnum:]")	abc ABC 123 .!?\(){}
	[:punct:][1]	punctuation	see("[:punct:]")	abc ABC 123 .!?\(){}
	[:graph:][1]	letters, numbers, and punctuation	see("[:graph:]")	abc ABC 123 .!?\(){}
	[:space:][1]	space characters (i.e. \s)	see("[:space:]")	abc ABC 123 .!?\(){}
	[:blank:][1]	space and tab (but not new line)	see("[:blank:]")	abc ABC 123 .!?\(){}
	.	every character except a new line	see(".")	abc ABC 123 .!?\(){}

[1] Many base R functions require classes to be wrapped in a second set of [ ], e.g. **[[:digit:]]**

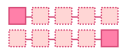
ALTERNATES

alt <- function(rx) str_view_all("abcde", rx)

regexp	matches	example	
ab|d	or	alt("ab|d")	abcde
[abe]	one of	alt("[abe]")	abcde
[^abe]	anything but	alt("[^abe]")	abcde
[a-c]	range	alt("[a-c]")	abcde

ANCHORS

anchor <- function(rx) str_view_all("aaa", rx)



regexp	matches	example	
^a	start of string	anchor("^a")	aaa
a$	end of string	anchor("a$")	aaa

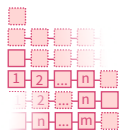
LOOK AROUNDS

look <- function(rx) str_view_all("bacad", rx)

regexp	matches	example	
a(?=c)	followed by	look("a(?=c)")	bacad
a(?!c)	not followed by	look("a(?!c)")	bacad
(?<=b)a	preceded by	look("(?<=b)a")	bacad
(?<!b)a	not preceded by	look("(?<!b)a")	bacad

QUANTIFIERS

quant <- function(rx) str_view_all(".a.aa.aaa", rx)



regexp	matches	example	
a?	zero or one	quant("a?")	.a.aa.aaa
a*	zero or more	quant("a*")	.a.aa.aaa
a+	one or more	quant("a+")	.a.aa.aaa
a{n}	exactly **n**	quant("a{2}")	.a.aa.aaa
a{n,}	**n** or more	quant("a{2,}")	.a.aa.aaa
a{n, m}	between **n** and **m**	quant("a{2,4}")	.a.aa.aaa

GROUPS

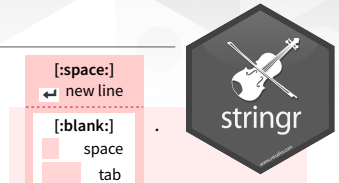
ref <- function(rx) str_view_all("abbaab", rx)

Use parentheses to set precedent (order of evaluation) and create groups

regexp	matches	example	
(ab|d)e	sets precedence	alt("(ab|d)e")	abcde

Use an escaped number to refer to and duplicate parentheses groups that occur earlier in a pattern. Refer to each group by its order of appearance

string (type this)	regexp (to mean this)	matches (which matches this)	example (the result is the same as ref("abba"))	
\\1	\1 (etc.)	first () group, etc.	ref("(a)(b)\\2\\1")	abbaab



Struktury danych – sprawdzanie typu i wielkości

Klasa

class()

Struktura

str()

Długość

length()

Struktury danych – obiekty – wektor

Wprowadzanie danych

Proste dane wprowadzamy do R za pomocą funkcji c(). Tak wprowadzone dane stają się wektorem. Wektor może zawierać dane jedynie jednego typu.

Struktury danych – obiekty – wektor

Generowanie ciągów liczb

: – generuje liczby z podanego przedziału,
seq – generuje liczby z podanego przedziału, przy czym można podać krok (by) i długość (length),
rep – generuje ciąg składający się z powtórzeń innego ciągu.

Struktury danych – obiekty – wektor

Indeksowanie

Wykonujemy je poprzez użycie nawiasu kwadratowego. W wyniku indeksowania uzyskuje się również wektor.

Struktury danych – obiekty – wektor

Operacje arytmetyczne

Na wektorach można wykonywać praktycznie wszystkie operacje arytmetyczne. Przy czym np. kwadrat wektora jest wektorem złożonym z kwadratów jego składowych, iloczyn wektorów jest wektorem, którego każda składowa jest iloczynem odpowiednich składowych mnożonych wektorów.

Struktury danych – obiekty – czynnik

Czynnik jest specjalną strukturą danych w R, przechowującą oprócz danych również liczbę wystąpień każdej wartości (w wielu językach programowania nazywany jest typem wyliczeniowym). Strukturę taką tworzymy za pomocą funkcji factor, natomiast liczbę wystąpień danego składnika otrzymamy funkcją table. Na tablicach można wykonywać także bardziej zaawansowane operacje. Służy do tego funkcja tapply, która działa na całej tablicy. Tablicę można, za pomocą polecenia, addmargins dodatkowo rozszerzyć o wiersze i kolumny podsumowujące. Warta uwagi jest jeszcze funkcja by (będąca w zasadzie wraperem na funkcję tapply), która pozwala na podsumowanie określonej zmiennej według innej zmiennej (najczęściej grupującej).

Struktury danych – obiekty – czynnik

Istnieje również struktura zwana uporządkowanym czynnikiem.
Tworzymy ją za pomocą funkcji ordered.

Struktury danych – obiekty – tablica

Tablica jest wektorem, zawierającym dodatkowe dane określające uporządkowanie elementów. Najczęściej stosowana jest tablica dwuwymiarowa czyli macierz. Indeksowanie tablic odbywa się podobnie do wektorów, w nawiasie kwadratowym podajemy „współrzędne" indeksowanego elementu. W razie pominięcia współrzędnej wynikiem indeksowania jest cały wiersz lub kolumna. Tablice tworzone są kolumnowo. Tablice można tworzyć z istniejących wektorów używając funkcji dim. Innymi, bardziej naturalnymi funkcjami tworzącymi tablice są matrix (dwuwymiarowe) i array (większe niż dwuwymiarowe). Do wykonywania operacji na wszystkich wierszach lub kolumnach macierzy równocześnie służy funkcja apply(macierz, wymiar, funkcja).

Podstawowe operacje na macierzach (1)

Funkcja	Działanie
outer(A, B, '*')	Tworzy z dwóch tablic większą tablicę wielowymiarową. Wymiary tej tablicy są połączeniem wektorów wymiarów dwóch tablic, zaś jej zawartość stanowią wszystkie możliwe kombinacje iloczynów (lub innych operacji) pomiędzy elementami. Ostatnim argumentem jest nazwa funkcji operującej na dwóch zmiennych. Istnieje zatem możliwość przeprowadzenia dowolnych operacji pomiędzy tablicami przez utworzenie własnej funkcji.
cbind(a, b)	Tworzy tablicę z podanych wektorów, poprzez umieszczenie ich kolumnami w nowo tworzonej tabeli.
rbind(a, b)	Tworzy tablicę z podanych wektorów, poprzez umieszczenie ich wierszami w nowo tworzonej tabeli.
t(A)	Transpozycja
det(A)	Wyznacznik
A%*%B	Iloczyn macierzy, samo * daje iloczyn po elementach
diag(A)	W przypadku wektora daje macierz z elementami tego wektora na przekątnej. W przypadku macierzy daje wektor o elementach przekątniowych macierzy.

Podstawowe operacje na macierzach (2)

Funkcja	Działanie
solve(A, b)	Rozwiązuje układy równań liniowych, jako pierwszy parametr podajemy macierz współczynników, a jako drugi wektor wyrazów wolnych. Jeśli nie podamy drugiego parametru funkcja obliczy macierz odwrotną.
colSums(A)	Wektor sum kolumn macierzy. Dla sumy wierszy mamy funkcję rowSums, podobnie można policzyć średnie dla kolumn i wierszy za pomocą funkcji colMeans oraz rowMeans odpowiednio.
eigen(x)	Wektory oraz wartości własne
svd(A)	dekompozycja SVD macierzy
qr(A)	dekompozycja QR macierzy
chol(A)	dekompozycja Choleskiego macierzy
kronecker(A, B)	iloczyn Kroneckera dwóch macierzy

Struktury danych – obiekty – lista

Lista jest uporządkowanym zbiorem elementów różnego typu. Do tworzenia list służy funkcja list. Jeśli chcemy odwołać się do konkretnego elementu listy to indeks elementu należy podać w podwójnych nawiasach kwadratowych. Każdy z elementów listy może mieć nazwę i takie nazwane listy spotyka się najczęściej (można się odwoływać do nazw, co nie jest możliwe w przypadku wektorów). Nazwy elementów listy można skracać do takiej długości, która wystarcza do jednoznacznej ich identyfikacji. Za pomocą funkcji lapply można wykonywać działania na całych listach (można również użyć funkcji sapply, która w wyniku działania daje wektor, funkcja lapply daje listę). W przypadku wielokrotnego wykonywania tej samej operacji (np. generowanie danych) można również wykorzystać funkcję replicate, będącą wraperem na funkcję sapply.

Struktury danych – obiekty – ramka danych

Ramka danych to specyficzna struktura R. Najprościej można określić ją jako macierz, w której poszczególne kolumny mogą zawierać wartości różnego typu. Można sobie wyobrazić, że wiersze takiej ramki to kolejne obserwacje w naszym doświadczeniu, a kolumny to cechy, które obserwujemy (nie wszystkie muszą być ilościowe). W rzeczywistości ramka jest szczególnym typem listy. Do utworzenia takiej struktury służy funkcja data.frame.

Struktury danych – obiekty – ramka danych

Dołączenie

attach()

Odłączenie

detach()

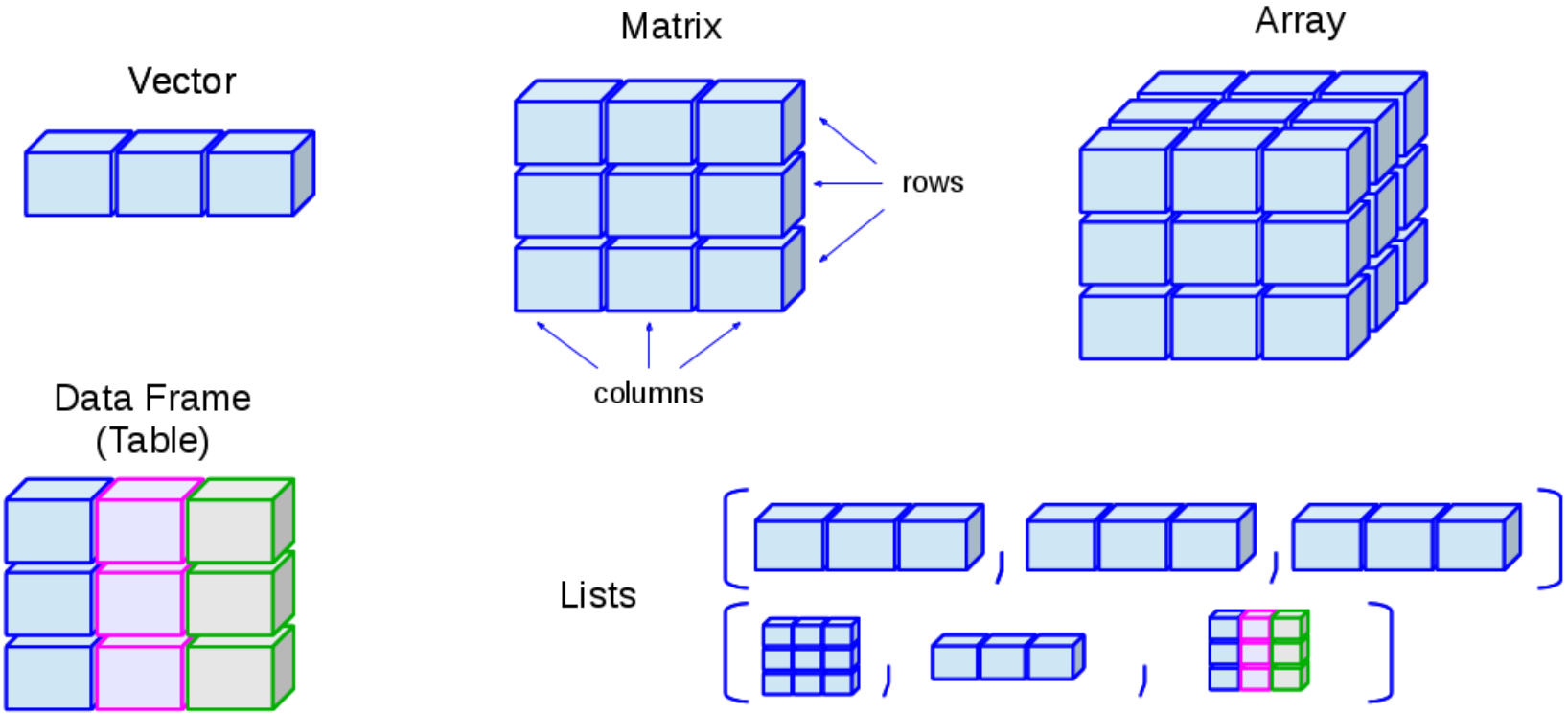
Podzbiór

subset()

Obiekty oraz typy

Obiekt	Typy	Różne typy?
Wektor	numeryczny, znakowy, zespolony, logiczny	Nie
Czynnik	numeryczny, znakowy	Nie
Tabela	numeryczny, znakowy, zespolony, logiczny	Nie
Ramka	numeryczny, znakowy, zespolony, logiczny	Tak
Lista	numeryczny, znakowy, zespolony, logiczny funkcja, wyrażenie	Tak

Obiekty oraz typy



Odczytywanie i zapisywanie danych

Odczyt zbioru danych

read.table(), load()

Nazwy kolumn, wierszy

names(), rownames(), colnames(), dimnames()

Odczytywanie i zapisywanie danych

Zapis zbioru danych

write.table(), save()

Karta pomocy

Base R
Cheat Sheet

Getting Help

Accessing the help files

?mean
Get help of a particular function.
help.search('weighted mean')
Search the help files for a word or phrase.
help(package = 'dplyr')
Find help for a package.

More about an object

str(iris)
Get a summary of an object's structure.
class(iris)
Find the class an object belongs to.

Using Packages

install.packages('dplyr')
Download and install a package from CRAN.

library(dplyr)
Load the package into the session, making all its functions available to use.

dplyr::select
Use a particular function from a package.

data(iris)
Load a built-in dataset into the environment.

Working Directory

getwd()
Find the current working directory (where inputs are found and outputs are sent).

setwd('C://file/path')
Change the current working directory.

Use projects in RStudio to set the working directory to the folder you are working in.

Vectors

Creating Vectors

c(2, 4, 6)	2 4 6	Join elements into a vector
2:6	2 3 4 5 6	An integer sequence
seq(2, 3, by=0.5)	2.0 2.5 3.0	A complex sequence
rep(1:2, times=3)	1 2 1 2 1 2	Repeat a vector
rep(1:2, each=3)	1 1 1 2 2 2	Repeat elements of a vector

Vector Functions

sort(x) Return x sorted.
rev(x) Return x reversed.
table(x) See counts of values.
unique(x) See unique values.

Selecting Vector Elements

By Position

x[4]	The fourth element.
x[-4]	All but the fourth.
x[2:4]	Elements two to four.
x[-(2:4)]	All elements except two to four.
x[c(1, 5)]	Elements one and five.

By Value

x[x == 10]	Elements which are equal to 10.
x[x < 0]	All elements less than zero.
x[x %in% c(1, 2, 5)]	Elements in the set 1, 2, 5.

Named Vectors

x['apple']	Element with name 'apple'.

Programming

For Loop

```
for (variable in sequence){
   Do something
}
```

Example

```
for (i in 1:4){
   j <- i + 10
   print(j)
}
```

While Loop

```
while (condition){
   Do something
}
```

Example

```
while (i < 5){
   print(i)
   i <- i + 1
}
```

If Statements

```
if (condition){
   Do something
} else {
   Do something different
}
```

Example

```
if (i > 3){
   print('Yes')
} else {
   print('No')
}
```

Functions

```
function_name <- function(var){
   Do something
   return(new_variable)
}
```

Example

```
square <- function(x){
   squared <- x*x
   return(squared)
}
```

Reading and Writing Data

Also see the **readr** package.

Input	Ouput	Description
df <- read.table('file.txt')	write.table(df, 'file.txt')	Read and write a delimited text file.
df <- read.csv('file.csv')	write.csv(df, 'file.csv')	Read and write a comma separated value file. This is a special case of read.table/ write.table.
load('file.RData')	save(df, file = 'file.Rdata')	Read and write an R data file, a file type special for R.

Conditions

a == b	Are equal	a > b	Greater than	a >= b	Greater than or equal to	is.na(a)	Is missing
a != b	Not equal	a < b	Less than	a <= b	Less than or equal to	is.null(a)	Is null

Learn more at **web page or vignette** • package version • Updated: 3/15

Karta pomocy

Types

Converting between common data types in R. Can always go from a higher value in the table to a lower value.

`as.logical`	`TRUE, FALSE, TRUE`	Boolean values (TRUE or FALSE).
`as.numeric`	`1, 0, 1`	Integers or floating point numbers.
`as.character`	`'1', '0', '1'`	Character strings. Generally preferred to factors.
`as.factor`	`'1', '0', '1', levels: '1', '0'`	Character strings with preset levels. Needed for some statistical models.

Maths Functions

`log(x)`	Natural log.	`sum(x)`	Sum.
`exp(x)`	Exponential.	`mean(x)`	Mean.
`max(x)`	Largest element.	`median(x)`	Median.
`min(x)`	Smallest element.	`quantile(x)`	Percentage quantiles.
`round(x, n)`	Round to n decimal places.	`rank(x)`	Rank of elements.
`signif(x, n)`	Round to n significant figures.	`var(x)`	The variance.
`cor(x, y)`	Correlation.	`sd(x)`	The standard deviation.

Variable Assignment

```
> a <- 'apple'
> a
[1] 'apple'
```

The Environment

`ls()`	List all variables in the environment.
`rm(x)`	Remove x from the environment.
`rm(list = ls())`	Remove all variables from the environment.

You can use the environment panel in RStudio to browse variables in your environment.

Matrices

`m <- matrix(x, nrow = 3, ncol = 3)`
Create a matrix from x.

`m[2, ]` - Select a row
`m[ , 1]` - Select a column
`m[2, 3]` - Select an element

`t(m)` Transpose
`m %*% n` Matrix Multiplication
`solve(m, n)` Find x in: m * x = n

Lists

`l <- list(x = 1:5, y = c('a', 'b'))`
A list is a collection of elements which can be of different types.

`l[[2]]`	`l[1]`	`l$x`	`l['y']`
Second element of l.	New list with only the first element.	Element named x.	New list with only element named y.

Also see the **dplyr** package.

Data Frames

`df <- data.frame(x = 1:3, y = c('a', 'b', 'c'))`
A special case of a list where all elements are the same length.

x	y
1	a
2	b
3	c

List subsetting

`df$x` `df[[2]]`

Understanding a data frame

`View(df)` See the full data frame.
`head(df)` See the first 6 rows.

Matrix subsetting

`df[ , 2]`
`df[2, ]`
`df[2, 2]`

`nrow(df)` Number of rows.
`ncol(df)` Number of columns.
`dim(df)` Number of columns and rows.

`cbind` - Bind columns.
`rbind` - Bind rows.

Strings

Also see the **stringr** package.

`paste(x, y, sep = ' ')`	Join multiple vectors together.
`paste(x, collapse = ' ')`	Join elements of a vector together.
`grep(pattern, x)`	Find regular expression matches in x.
`gsub(pattern, replace, x)`	Replace matches in x with a string.
`toupper(x)`	Convert to uppercase.
`tolower(x)`	Convert to lowercase.
`nchar(x)`	Number of characters in a string.

Factors

`factor(x)` Turn a vector into a factor. Can set the levels of the factor and the order.

`cut(x, breaks = 4)` Turn a numeric vector into a factor by 'cutting' into sections.

Statistics

`lm(y ~ x, data=df)` Linear model.
`glm(y ~ x, data=df)` Generalised linear model.
`summary` Get more detailed information out a model.

`t.test(x, y)` Perform a t-test for difference between means.
`pairwise.t.test` Perform a t-test for paired data.

`prop.test` Test for a difference between proportions.
`aov` Analysis of variance.

Distributions

	Random Variates	Density Function	Cumulative Distribution	Quantile
Normal	`rnorm`	`dnorm`	`pnorm`	`qnorm`
Poisson	`rpois`	`dpois`	`ppois`	`qpois`
Binomial	`rbinom`	`dbinom`	`pbinom`	`qbinom`
Uniform	`runif`	`dunif`	`punif`	`qunif`

Plotting

Also see the **ggplot2** package.

`plot(x)` Values of x in order.
`plot(x, y)` Values of x against y.
`hist(x)` Histogram of x.

Dates

See the **lubridate** package.

Learn more at **web page or vignette** • package version • Updated: 3/15