

一份不太简短的 R 语言介绍

黄湘云 新浪 BIP 2017 年 10 月

- 1. 前世
- 2. 今生
- 3. 社区箴言
- 4. 社区领袖
- 5. 数据可视化
- 6. 数据报告
- 7. 学习R

前世

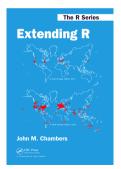
来自何方

R was inspired by the S environment which has been principally developed by John Chambers, with substantial input from Douglas Bates, Rick Becker, Bill Cleveland, Trevor Hastie[1, 2], Daryl Pregibon and Allan Wilks.









今生

创世者

R was initially written by **Robert Gentleman** and **Ross Ihaka** also known as "R & R" of the Statistics Department of the University of Auckland.

Robert Gentleman joined 23andMe in 2015 as VP of computational biology

创世者

R was initially written by **Robert Gentleman** and **Ross Ihaka** also known as "R & R" of the Statistics Department of the University of Auckland.

- Robert Gentleman joined 23andMe in 2015 as VP of computational biology
- Ross Ihaka

• R Development Core Team (Since mid-1997):

- R Development Core Team (Since mid-1997):
- 1. Uwe Ligges (CRAN 官仓管理员)

- R Development Core Team (Since mid-1997):
- 1. Uwe Ligges (CRAN 官仓管理员)
- 2. Paul Murrell (grid 图形系统创始人 [3])

- R Development Core Team (Since mid-1997):
- 1. Uwe Ligges (CRAN 官仓管理员)
- 2. Paul Murrell (grid 图形系统创始人 [3])
- 3. Brian Ripley (MASS 包作者 [4])

- R Development Core Team (Since mid-1997):
- 1. Uwe Ligges (CRAN 官仓管理员)
- 2. Paul Murrell (grid 图形系统创始人 [3])
- 3. Brian Ripley (MASS 包作者 [4])
- 4. Deepayan Sarkar (lattice 包作者 [5])

1. 要到哪里去

R is **GNU S**, a freely available language and environment for statistical computing(统计计算)and graphics(统计图形)

1. 要到哪里去

R is **GNU S**, a freely available language and environment for statistical computing(统计计算)and graphics(统计图形)

1997年4月23日发布版本 0.49至2017年9月28日发布版本 3.4.2(已历20载)

1. 要到哪里去

R is **GNU S**, a freely available language and environment for statistical computing(统计计算)and graphics(统计图形)

1997年4月23日发布版本 0.49至2017年9月28日发布版本 3.4.2 (已历20载)

2. 现在做了什么

1. 要到哪里去

R is **GNU S**, a freely available language and environment for statistical computing(统计计算)and graphics(统计图形)

• 1997年4月23日发布版本 0.49至2017年9月28日发布版本 3.4.2 (已历20载)

2. 现在做了什么

R provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc.

• 可以看看 CRAN 的 TaskViews 栏

1. 要到哪里去

R is **GNU S**, a freely available language and environment for statistical computing(统计计算)and graphics(统计图形)

1997年4月23日发布版本 0.49至2017年9月28日发布版本 3.4.2 (已历20载)

2. 现在做了什么

- 可以看看 CRAN 的 TaskViews 栏
- 还可以看看 Bioconductor 官网

社区箴言

Paul Murrell

If you imagine that this pen is Trellis, then Lattice is not this pen.

-- Paul Murrell (on the difference of Lattice (which eventually was called grid) and Trellis)DSC 2001, Wien (March 2001)

Have you ever wanted to write a book, but not known where to start? Now is a very good time to jump in, because there is currently a very simple recipe for success: just put R in the title and you will have to beat the publishers off with a stick!

-- Paul Murrell
ASA Statistical Computing & Graphics Newsletter 17(2) (November 2006)

Barry Rowlingson

I'd like to prefix all these solutions with 'Here's how to do it, but don't actually do it you crazy fool'. It's on a par with redefining pi, or redefining '+'. And then redefining '<-'. These techniques have their proper place, and that would be in the currently non-existent obfuscated R contest.

No, the R-ish (iRish?) way is to index vectors from 1. That's what the R gods intended!

-- Barry Rowlingson (in a discussion how vectors in R could be indexed starting from 0)

R-help (March 2004)

Brian D. Ripley

To paraphrase provocatively, 'machine learning is statistics minus any checking of models and assumptions'.

-- Brian D. Ripley (about the difference between machine learning and statistics) useR! 2004. Vienna (May 2004)

Let's not kid ourselves: the most widely used piece of software for statistics is Excel.

-- Brian D. Ripley ('Statistical Methods Need Software: A View of Statistical Computing')Opening lecture RSS 2002, Plymouth (September 2002)

Peter Dalgaard

The documentation level of R is already much higher than average for open source software and even than some commercial packages (esp. SPSS is notorious for its attitude of "You want to do one of these things. If you don't understand what the output means, click help and we'll pop up five lines of mumbo-jumbo that you're not going to understand either.")

-- Peter Dalgaard R-help (April 2002)

Douglas Bates

```
Zhu Wang: I am trying to create a library which uses some Fortran source files [...]

Douglas Bates: Someone named Martin Maechler will shortly be sending you email regarding the distinction between 'library' and 'package' :-)

-- Zhu Wang and Douglas Bates
R-help (May 2004)
```

最后的话

For the S system, which has forever altered the way people analyze, visualize, and manipulate data S is an elegant, widely accepted, and enduring software system, with conceptual integrity, thanks to the insight, taste, and effort of John Chambers.

-- Association for Computing Machinery
ACM/Software System Award citation (1998)

社区领袖

Roger Bivand

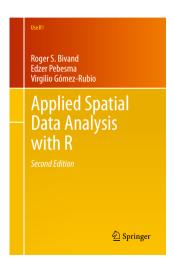
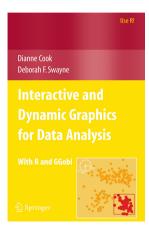


图 1: Applied spatial data analysis with R

Di Cook



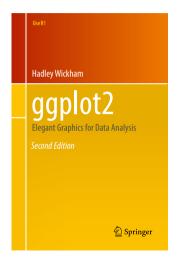


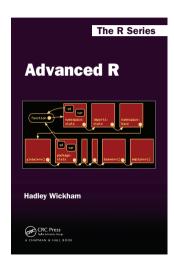
Hadley Wickham



图 2: Hadley Wickham

Hadley Wickham





• 2017, Carson Sievert plotly

- 2017, Carson Sievert plotly
- 2016, Tong He (何通), Tianqi Chen (陈天奇) and Yixuan Qiu (邱怡轩) XGBoost

- 2017, Carson Sievert plotly
- 2016, Tong He (何通), Tianqi Chen (陈天奇) and Yixuan Qiu (邱怡轩) XGBoost
- 2009, Yihui Xie (谢益辉) animation

- 2017, Carson Sievert plotly
- 2016, Tong He(何通), Tianqi Chen(陈天奇)and Yixuan Qiu(邱怡轩)XGBoost
- 2009, Yihui Xie (谢益辉) animation
- 2006, Hadley Wickham reshape and ggplot

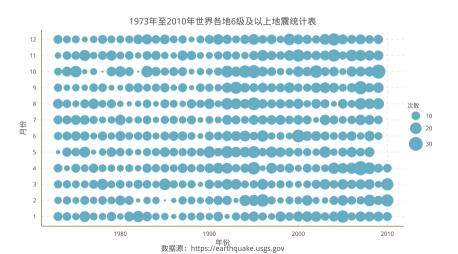
- 2017, Carson Sievert plotly
- 2016, Tong He(何通), Tianqi Chen(陈天奇)and Yixuan Qiu(邱怡轩)XGBoost
- 2009, Yihui Xie (谢益辉) animation
- 2006, Hadley Wickham reshape and ggplot
- 2004, Deepayan Sarkar lattice

- 2017, Carson Sievert plotly
- 2016, Tong He(何通), Tianqi Chen(陈天奇)and Yixuan Qiu(邱怡轩)XGBoost
- 2009, Yihui Xie (谢益辉) animation
- 2006, Hadley Wickham reshape and ggplot
- 2004, Deepayan Sarkar lattice
- 2003, Daniel Adler rgl

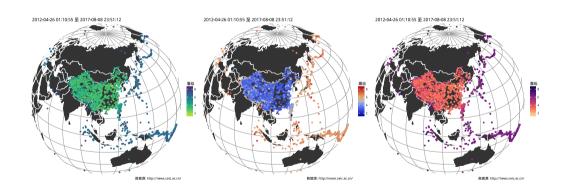
数据可视化

1973-2010 全球 6 级以上地震

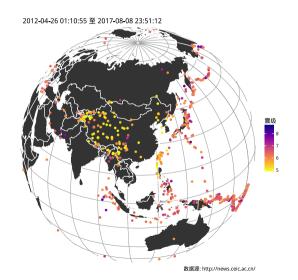
1973-2010 全球 6 级以上地震



2012.4-2017.8 中国 5 级及以上地震



2012.4-2017.8 中国 5 级及以上地震



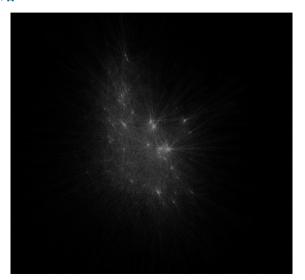
2015年11月28-30日雾霾时空过程@王江浩

Using R packages and education to scale Data Science at Airbnb

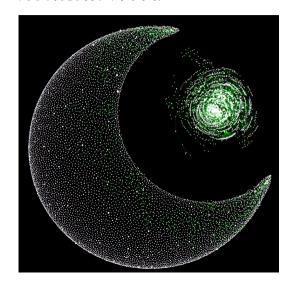


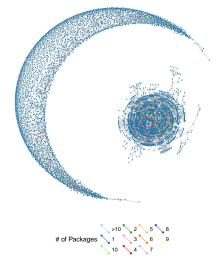
图 4: Ricardo Bion https://github.com/ricardo-bion/medium_visualization

R包依赖关系网络



开发者协作关系网络





Flat map great circle animation example

敬请欣赏

Matthew Leonawicz

Flat map great circle animation example

敬请欣赏

- Matthew Leonawicz
- https://leonawicz.github.io/

数据报告

基于 R Markdown 的分析报告



• R + Markdown + Knitr + RStudio(IDE)

- R + Markdown + Knitr + RStudio(IDE)
- Python + Markdown + IPython + Jupyter + Rodeo(IDE)

- R + Markdown + Knitr + RStudio(IDE)
- Python + Markdown + IPython + Jupyter + Rodeo(IDE)
- 1. Jupyter 不支持缓存

- R + Markdown + Knitr + RStudio(IDE)
- Python + Markdown + IPython + Jupyter + Rodeo(IDE)
- 1. Jupyter 不支持缓存
- 2. Kernels 没有统一接口

- R + Markdown + Knitr + RStudio(IDE)
- Python + Markdown + IPython + Jupyter + Rodeo(IDE)
- 1. Jupyter 不支持缓存
- 2. Kernels 没有统一接口
- 3. 版面太死板, 导出文档格式极其有限

- R + Markdown + Knitr + RStudio(IDE)
- Python + Markdown + IPython + Jupyter + Rodeo(IDE)
- 1. Jupyter 不支持缓存
- 2. Kernels 没有统一接口
- 3. 版面太死板, 导出文档格式极其有限
- 4. 代码和内容不能分离

- R + Markdown + Knitr + RStudio(IDE)
- Python + Markdown + IPython + Jupyter + Rodeo(IDE)
- 1. Jupyter 不支持缓存
- 2. Kernels 没有统一接口
- 3. 版面太死板, 导出文档格式极其有限
- 4. 代码和内容不能分离
- 5. ipynb 基于 json 数据格式存储

- R + Markdown + Knitr + RStudio(IDE)
- Python + Markdown + IPython + Jupyter + Rodeo(IDE)
- 1. Jupyter 不支持缓存
- 2. Kernels 没有统一接口
- 3. 版面太死板, 导出文档格式极其有限
- 4. 代码和内容不能分离
- 5. ipynb 基于 json 数据格式存储
- 6. Jupyter 不支持文本行内计算

- R + Markdown + Knitr + RStudio(IDE)
- Python + Markdown + IPython + Jupyter + Rodeo(IDE)
- 1. Jupyter 不支持缓存
- 2. Kernels 没有统一接口
- 3. 版面太死板, 导出文档格式极其有限
- 4. 代码和内容不能分离
- 5. ipynb 基于 json 数据格式存储
- 6. Jupyter 不支持文本行内计算
- 7. • •

- R + Markdown + Knitr + RStudio(IDE)
- Python + Markdown + IPython + Jupyter + Rodeo(IDE)
- 1. Jupyter 不支持缓存
- 2. Kernels 没有统一接口
- 3. 版面太死板, 导出文档格式极其有限
- 4. 代码和内容不能分离
- 5. ipynb 基于 json 数据格式存储
- 6. Jupyter 不支持文本行内计算
- 7. ...
- 处处学朕(RStudio), 处处学得不像, 朕是以宽仁治国, 他(Rodeo)是以宽仁收买人心

ggplot2 画廊 I

```
[1] "qqalt"
                        "gganimate"
                                            "ggbeeswarm"
[4] "ggbio"
                        "ggBrackets"
                                            "ggChernoff"
    "ggCompNet"
                        "ggcorrplot"
                                            "ggcyto"
    "ggdendro"
                        "ggdmc"
                                            "gge"
[13] "ggedit"
                        "ggeffects"
                                            "ggene"
[16] "ggenealogy"
                        "ggExtra"
                                            "ggforce"
[19] "ggformula"
                        "ggfortify"
                                            "ggghost"
[22] "ggguitar"
                        "gghalfnorm"
                                            "ggimage"
[25] "ggiraph"
                        "ggiraphExtra"
                                            "ggjoy"
[28] "gglasso"
                        "gglogo"
                                            "ggloop"
[31] "ggm"
                        "ggmap"
                                            "ggmcmc"
[34] "ggmosaic"
                        "ggnetwork"
                                            "ggparallel"
[37] "ggplot2"
                        "ggplot2.SparkR"
                                            "ggplot2movies"
```

ggplot2 画廊 II

[40]	"ggplotFL"	"ggplotgui"	"ggpmisc"
[43]	"ggpolypath"	"ggpubr"	"ggpval"
[46]	"ggQC"	"ggradar"	"ggRandomForests"
[49]	"ggraph"	"ggraptR"	"ggrepel"
[52]	"ggridges"	"ggR0C"	"ggsci"
[55]	"ggseas"	"ggseqlogo"	"ggsignif"
[58]	"ggsn"	"ggspatial"	"ggspectra"
[61]	"ggstance"	"ggswissmaps"	"ggtech"
[64]	"ggtern"	"ggThemeAssist"	"ggthemes"
[67]	"ggthemr"	"ggtree"	"ggvis"

• 理论基础扎实 (ggvis) [6]

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)
 - % http://ggplot2.tidyverse.org/reference/

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)
 - % http://ggplot2.tidyverse.org/reference/
- 生态喜人(69+包)

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)
 - % http://ggplot2.tidyverse.org/reference/
- 生态喜人(69+包)
 - % http://www.ggplot2-exts.org/

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)
 - % http://ggplot2.tidyverse.org/reference/
- 生态喜人(69+包)
 - % http://www.ggplot2-exts.org/
- 学习资源、书籍丰富

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)
 - % http://ggplot2.tidyverse.org/reference/
- 生态喜人(69+包)
 - % http://www.ggplot2-exts.org/
- 学习资源、书籍丰富
 - % http://www.cookbook-r.com/

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)
 - % http://ggplot2.tidyverse.org/reference/
- 生态喜人(69+包)
 - % http://www.ggplot2-exts.org/
- 学习资源、书籍丰富
 - % http://www.cookbook-r.com/
 - % http://www.sthda.com/english/

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)
 - % http://ggplot2.tidyverse.org/reference/
- 生态喜人(69+包)
 - % http://www.ggplot2-exts.org/
- 学习资源、书籍丰富
 - % http://www.cookbook-r.com/
 - % http://www.sthda.com/english/
 - % http://ggplot2.org/

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)
 - % http://ggplot2.tidyverse.org/reference/
- 生态喜人(69+包)
 - % http://www.ggplot2-exts.org/
- 学习资源、书籍丰富
 - % http://www.cookbook-r.com/
 - % http://www.sthda.com/english/
 - % http://ggplot2.org/
- 爆栈网 18,000+ 问题

- 理论基础扎实 (ggvis) [6]
- 功能全面(435个函数)
 - % http://ggplot2.tidyverse.org/reference/
- 生态喜人(69+包)
 - % http://www.ggplot2-exts.org/
- 学习资源、书籍丰富
 - % http://www.cookbook-r.com/
 - % http://www.sthda.com/english/
 - % http://ggplot2.org/
- 爆栈网 18,000+ 问题
 - % https://stackoverflow.com/questions/tagged/ggplot2

光说不练假把式

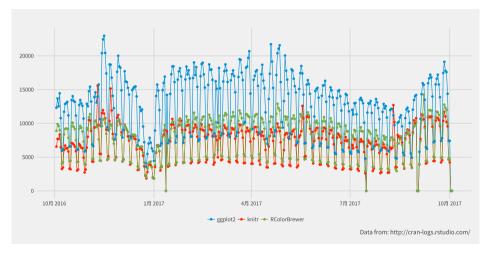


图 6: 每日下载量

饼图

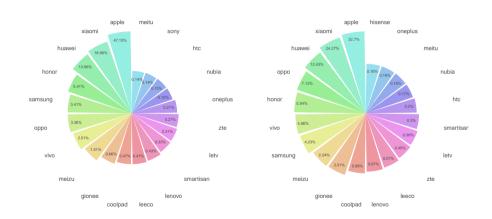
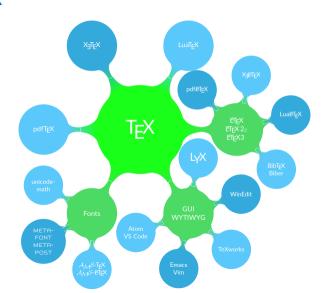


图 7: 一线和二三线各品牌手机份额对比

了解一点 LATEX



学习R

自带 R 包共 29 个

```
Pkgs <- sapply(list.files(R.home('library')), function(x)</pre>
  packageDescription(pkg = x,fields = "Priority"))
names(Pkgs[Pkgs == 'base' & !is.na(Pkgs)])
 [1] "base"
                 "compiler"
                            "datasets"
                                         "graphics"
                                                     "grDevices"
                 "methods"
                             "parallel" "splines"
 [6] "grid"
                                                     "stats"
[11] "stats4"
                "tcltk"
                             "tools"
                                         "utils"
names(Pkgs[Pkgs == 'recommended' & !is.na(Pkgs)])
 [1] "boot"
                  "class"
                               "cluster"
                                            "codetools"
                                                         "foreign"
 [6] "KernSmooth" "lattice"
                               "MASS"
                                            "Matrix"
                                                         "macv"
[11] "nlme"
                  "nnet"
                               "rpart"
                                            "spatial"
                                                         "survival"
```

入门

• R-admin.pdf 安装和管理 79

入门

- R-admin.pdf 安装和管理 79
- R-intro.pdf 导论 105

入门

- R-admin.pdf 安装和管理 79
- R-intro.pdf 导论 105
- R-data.pdf 数据导入与导出 37

入门

- R-admin.pdf 安装和管理 79
- R-intro.pdf 导论 105
- R-data.pdf 数据导入与导出 37
- R-FAQ.html & rw-FAQ.html 常见问题解答 100+

入门

- R-admin.pdf 安装和管理 79
- R-intro.pdf 导论 105
- R-data.pdf 数据导入与导出 37
- R-FAQ.html & rw-FAQ.html 常见问题解答 100+
- Sweave.pdf 文学化编程 (literate programming [7]) 15

入门

- R-admin.pdf 安装和管理 79
- R-intro.pdf 导论 105
- R-data.pdf 数据导入与导出 37
- R-FAQ.html & rw-FAQ.html 常见问题解答 100+
- Sweave.pdf 文学化编程 (literate programming [7]) 15

进阶

• R-lang.pdf 精髓 60

入门

- R-admin.pdf 安装和管理 79
- R-intro.pdf 导论 105
- R-data.pdf 数据导入与导出 37
- R-FAQ.html & rw-FAQ.html 常见问题解答 100+
- Sweave.pdf 文学化编程 (literate programming [7]) 15

- R-lang.pdf 精髓 60
- R-ints.pdf 底层 64

入门

- R-admin.pdf 安装和管理 79
- R-intro.pdf 导论 105
- R-data.pdf 数据导入与导出 37
- R-FAQ.html & rw-FAQ.html 常见问题解答 100+
- Sweave.pdf 文学化编程 (literate programming [7]) 15

- R-lang.pdf 精髓 60
- R-ints.pdf 底层 64
- R-exts.pdf 扩展包开发 185

入门

- R-admin.pdf 安装和管理 79
- R-intro.pdf 导论 105
- R-data.pdf 数据导入与导出 37
- R-FAQ.html & rw-FAQ.html 常见问题解答 100+
- Sweave.pdf 文学化编程 (literate programming [7]) 15

- R-lang.pdf 精髓 60
- R-ints.pdf 底层 64
- R-exts.pdf 扩展包开发 185
- fullrefman.pdf 完整参考手册(补充) 3587

入门

- R-admin.pdf 安装和管理 79
- R-intro.pdf 导论 105
- R-data.pdf 数据导入与导出 37
- R-FAQ.html & rw-FAQ.html 常见问题解答 100+
- Sweave.pdf 文学化编程 (literate programming [7]) 15

- R-lang.pdf 精髓 60
- R-ints.pdf 底层 64
- R-exts.pdf 扩展包开发 185
- fullrefman.pdf 完整参考手册(补充) 3587
- NEWS.O & NEWS.1 & NEWS.2 & NEWS.pdf 更新历史(补充) 300+

适合自己的才是最好的

英文

• https://cran.r-project.org/other-docs.html

中文

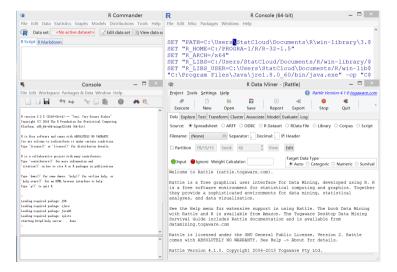
适合自己的才是最好的

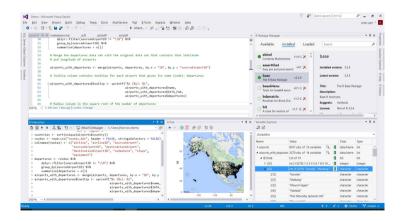
英文

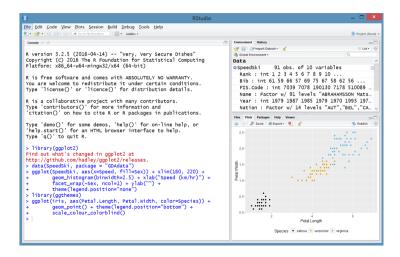
• https://cran.r-project.org/other-docs.html

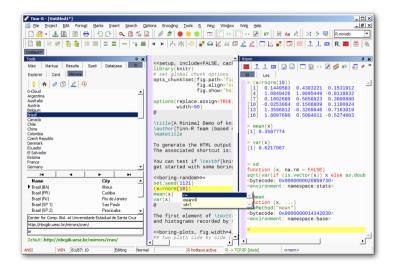
中文

• 常见问题解答(刘思喆等搜集于中文论坛)









尽情探索吧

help.start()





Manuals

An Introduction to R Writing R Extensions R Data Import/Export The R Language Definition
R Installation and Administration
R Internals

Reference

<u>Packages</u>

Search Engine & Keywords

Miscellaneous Material

About R License NEWS Authors Frequently Asked Questions User Manuals

Resources
Thanks
Technical papers

Material specific to the Windows port

最后回答一个大家都关心的问题

R 语言的优劣势是什么

• 知乎问答 https://www.zhihu.com/question/19611094

R的若干基因及争论

超棒的 R 资源列表

最后回答一个大家都关心的问题

R 语言的优劣势是什么

• 知乎问答 https://www.zhihu.com/question/19611094

R的若干基因及争论

• 谢益辉 https://yihui.name/cn/2012/09/equal-and-arrow/

超棒的 R 资源列表

最后回答一个大家都关心的问题

R 语言的优劣势是什么

• 知乎问答 https://www.zhihu.com/question/19611094

R的若干基因及争论

• 谢益辉 https://yihui.name/cn/2012/09/equal-and-arrow/

超棒的 R 资源列表

• 覃文峰 https://github.com/qinwf/awesome-R

• **x**iangyunfaith @ outlook.com

- **x**iangyunfaith @ outlook.com
- **(7)** Cloud2016

- **x**iangyunfaith @ outlook.com
- **(2)** Cloud2016
- **Capital of Statistics https://cosx.org/members/**

- **x**iangyunfaith @ outlook.com
- Cloud2016
- **Capital of Statistics https://cosx.org/members/**
- **%** Sophorae

致谢

• 谢益辉 https://yihui.name/ for knitr and bookdown

致谢

- 谢益辉 https://yihui.name/ for knitr and bookdown
- 统计之都 https://cosx.org/

致谢

- 谢益辉 https://yihui.name/ for knitr and bookdown
- 统计之都 https://cosx.org/
- 论坛 https://d.cosx.org/

- [1] Material from the book's webpage, R port, and packaging by Kjetil B Halvorsen. ElemStatLearn: Data Sets, Functions and Examples from the Book: "The Elements of Statistical Learning, Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani and Jerome Friedman, 2015. R package version 2015.6.26.
- [2] Gareth James, Daniela Witten, Trevor Hastie, and Rob Tibshirani. *ISLR: Data for an Introduction to Statistical Learning with Applications in R*, 2017. R package version 1.1.
- [3] Paul Murrell. RGraphics: Data and Functions from the Book R Graphics, Second Edition, 2016. R package version 2.0-14.
- [4] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.
- [5] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R. Springer, New York, 2008.* ISBN 978-0-387-75968-5.
- [6] Leland Wilkinson. The Grammar of Graphics. Springer-Verlag New York, 2005.
- [7] Donald Knuth. Literate programming. The Computer Journal, 27(2):97-111, 1984.