



北京邮电大学

BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS



Data Analytics with R
语言数据分析



人人都爱tidyverse

艾新波 / 2018 • 北京



课程体系



R语言数据分析



上部：论道



- 第1章 气象万千、数以等观
- 第2章 所谓学习、归类而已
- 第3章 格言联璧话学习
- 第4章 源于数学、归于工程



中部：执具



- 第5章 工欲善其事必先利其器
- 第6章 基础编程
- 第7章 数据对象



第8章 人人都爱tidyverse

第9章 最美不过数据框



下部 博术



- 第10章 观数以形
- 第11章 相随相伴、谓之关联
- 第12章 既是世间法、自当有分别
- 第13章 方以类聚、物以群分
- 第14章 庐山烟雨浙江潮

如果说优雅也有缺点的话，那就是你需要
艰巨的工作才能得到它，需要良好的教育
才能欣赏它

——Edsger Wybe Dijkstra

Tidy datasets are all alike, but every messy dataset is messy in its own way.

— Hadley Wickham


tidy 


英 ['taɪdɪ] 

美 ['taɪdi] 

adj. 整齐的；相当大的
vt. 整理；收拾；弄整齐

verse 

英 [vɜːs] 

美 [vɜːs] 

n. 诗，诗篇；韵文；诗节
vi. 作诗

tidyverse



tidy = 清爽

tidyxxxx代码界的一股清流

tidyverse

[tidybayes](#)

[tidyboot](#)

[tidycensus](#)

[tidygenomics](#)

[tidygraph](#)

[tidyhydat](#)

[tidyimpute](#)

[tidylog](#)

[tidyLPA](#)

[tidymodels](#)

[tidyposterior](#)

[tidypredict](#)

[tidyquant](#)

[tidyqwi](#)

[tidyr](#)

[tidyRSS](#)

[tidyselect](#)

[tidystats](#)

[tidystopwords](#)

[tidytext](#)

[tidytidbits](#)

[tidytransit](#)

[tidytree](#)

[tidyverse](#)

[tidyxl](#)

Tidy Data and 'Geoms' for Bayesian Models

Tidyverse-Compatible Bootstrapping

Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames

Tidy Verbs for Dealing with Genomic Data Frames

A Tidy API for Graph Manipulation

Extract and Tidy Canadian 'Hydrometric' Data

Imputation the Tidyverse Way

Logging for 'dplyr' Functions

Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software

Easily Install and Load the 'Tidymodels' Packages

Bayesian Analysis to Compare Models using Resampling Statistics

Run Predictions Inside the Database

Tidy Quantitative Financial Analysis

A Convenient API for Accessing United States Census Bureau's Quarterly Workforce Indicator

Easily Tidy Data with 'spread()' and 'gather()' Functions

Tidy RSS for R

Select from a Set of Strings

Create a Tidy Statistics Output File

Customizable Lists of Stopwords in 53 Languages

Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools

A Collection of Tools and Helpers Extending the Tidyverse

Read, Validate, Analyze, and Map Files in the General Transit Feed Specification

A Tidy Tool for Phylogenetic Tree Data Manipulation

Easily Install and Load the 'Tidyverse'

Read Untidy Excel Files

tidytext

tidygraph

tidystats

tidyimpute

tidypredict

tidyxl

tidybayes

tidyboot

tidyverse扩展包套装

```
> library(tidyverse)
```

```
-- Attaching packages ----- tidyverse 1.2.1 --
```

```
✓ ggplot2 2.2.1          ✓ purrr 0.2.4
```

```
✓ tibble 1.4.2           ✓ dplyr 0.7.4
```

```
✓ tidyr 0.8.0            ✓ stringr 1.3.1
```

```
✓ readr 1.1.1           ✓ forcats 0.3.0
```

```
-- Conflicts ----- tidyverse_conflicts() --
```

```
✗ dplyr::filter() masks stats::filter()
```

```
✗ dplyr::lag() masks stats::lag()
```

tidyverse扩展包套装

序号	扩展包	功能
1	ggplot2	data visualisation
2	dplyr	data manipulation
3	tidyr	data tidying
4	readr	data import
5	purrr	functional programming
6	tibble	tibbles, a modern re-imagining of data frames
7	stringr	strings
8	forcats	for Categorical Variables (Factors)

更多内容请参阅：
<https://www.tidyverse.org/>

%>%
magrittr

Ceci n'est pas un pipe.

a new “pipe”-like operator, %>%, with which you may pipe a **value** forward into an expression or **function** call; something along the lines of `x %>% f`, rather than `f(x)`

管道操作符

lhs %>% rhs



value



function

更多用法参见

`?magrittr::`%>``

表达式	用法
<code>x %>% f</code>	<code>f(x)</code>
<code>x %>% f(y)</code>	<code>f(x, y)</code>
<code>y %>% f(x, .)</code>	<code>f(x, y)</code>
<code>z %>% f(x, y, arg = .)</code>	<code>f(x, y, arg = z)</code>
<code>x %>% f %>% g %>% h</code>	<code>h(g(f(x)))</code>

查看数据记录

cjb %>%

head

```
#> # A tibble: 6 x 13
```

```
#> xm          bj xb          yw          sx          wy          zz          ls          dl
```

```
#> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
#> 1 周黎      1101 女          94          82          96          97          97          98
```

```
#> 2 汤海明~   1101 男          87          94          89          95          94          94
```

```
#> 3 舒江辉~   1101 男          92          79          86          98          95          96
```

```
#> 4 翁柯      1101 女          91          84          96          93          97          94
```

```
#> 5 祁强      1101 男          85          92          82          93          87          88
```

```
#> 6 湛容      1101 女          92          82          85          91          90          92
```

```
#> # ... with 4 more variables: wl <dbl>, hx <dbl>,
```

```
#> #      sw <dbl>, wlfk <chr>
```

查看数据记录

```
cjb %>%
```

```
  head(n = 4) #cjb默认为第一个参数
```

```
head(cjb, n = 4)
```

```
#> # A tibble: 4 x 13
```

```
#> xm          bj xb          yw      sx      wy      zz      ls      dl
```

```
#> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
#> 1 周黎      1101 女          94      82      96      97      97      98
```

```
#> 2 汤海明~   1101 男          87      94      89      95      94      94
```

```
#> 3 舒江辉~   1101 男          92      79      86      98      95      96
```

```
#> 4 翁柯      1101 女          91      84      96      93      97      94
```

```
#> # ... with 4 more variables: wl <dbl>, hx <dbl>,
```

```
#> #      sw <dbl>, wlfk <chr>
```

dplyr: data manipulation

序号	扩展包	功能
1	<code>select()</code>	picks variables based on their names
2	<code>mutate()</code>	adds new variables that are functions of existing variables
3	<code>filter()</code>	picks cases based on their values
4	<code>arrange()</code>	changes the ordering of the rows
5	<code>summarise()</code>	reduces multiple values down to a single summary

选择列

```
cjb %>%
```

```
  select(xm, yw, sx) %>%
```

```
  head(n = 3)
```

```
#> # A tibble: 3 x 3
```

```
#>   xm      yw      sx
```

```
#>   <chr>   <dbl> <dbl>
```

```
#> 1 周黎      94      82
```

```
#> 2 汤海明     87      94
```

```
#> 3 舒江辉     92      79
```

选择列

```
cjb %>%
```

```
  select(xm, yw, sx) %>%
```

```
  set_names(c("姓名", "语文", "数学")) %>%
```

```
  head(n = 3)
```

```
#> # A tibble: 3 x 3
```

```
#>   姓名      语文    数学
```

```
#>   <chr>   <dbl> <dbl>
```

```
#> 1 周黎      94     82
```

```
#> 2 汤海明    87     94
```

```
#> 3 舒江辉    92     79
```

选择列

```
cjb %>%
```

```
  select(1, 4:12) %>%
```

```
  head(n = 3)
```

```
#> # A tibble: 3 x 10
```

```
#>   xm      yw      sx      wy      zz      ls      dl      wl      hx
```

```
#>   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
#> 1 周黎      94      82      96      97      97      98      95      94
```

```
#> 2 汤海明~    87      94      89      95      94      94      90      90
```

```
#> 3 舒江辉~    92      79      86      98      95      96      89      94
```

```
#> # ... with 1 more variable: sw <dbl>
```


选择列

```
cjb %>%
```

```
  select(xm, yw:sw) %>%
```

```
  head(n = 3)
```

```
#> # A tibble: 3 x 10
```

```
#>   xm      yw      sx      wy      zz      ls      dl      wl      hx
```

```
#>   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
#> 1 周黎      94      82      96      97      97      98      95      94
```

```
#> 2 汤海明~    87      94      89      95      94      94      90      90
```

```
#> 3 舒江辉~    92      79      86      98      95      96      89      94
```

```
#> # ... with 1 more variable: sw <dbl>
```

增加或修改某些列

```
cjb %>%
```

```
  mutate_at(vars(bj, xb, wlfk), factor) %>%
```

```
  mutate(zcj = rowSums(.[4:12])) %>%
```

```
  arrange(desc(zcj)) %>%
```

```
  tail(n = 2)
```

```
#> # A tibble: 2 x 14
```

```
#>   xm      bj      xb      yw      sx      wy      zz      ls      dl
```

```
#>   <chr> <fct> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
#> 1 滑亚  1113  男      33      46      30      65      82      76
```

```
#> 2 张良平~ 1115  男      0      0      0      0      0      0
```

```
#> # ... with 5 more variables: wl <dbl>, hx <dbl>,
```

```
#> #   sw <dbl>, wlfk <fct>, zcj <dbl>
```

让更改生效

#采用%<>%操作符

```
cjb %<>%
```

```
mutate_at(vars(bj, xb, wlfk), factor) %>%
```

```
mutate(zcj = rowSums(.[4:12])) %>%
```

```
arrange(desc(zcj))
```

cjb <- cjb %>% #和上述语句等价

```
mutate_at(vars(bj, xb, wlfk), factor) %>%
```

```
mutate(zcj = rowSums(.[4:12])) %>%
```

```
arrange(desc(zcj))
```

选择行

```
cjb %>%
```

```
  filter(yw < 60)
```

```
#> # A tibble: 2 x 13
```

```
#>   xm      bj xb      yw      sx      wy      zz      ls      dl
```

```
#>   <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
#>   1 滑亚    1113 男      33      46      30      65      82      76
```

```
#>   2 张良平~  1115 男       0       0       0       0       0       0
```

```
#> # ... with 4 more variables: wl <dbl>, hx <dbl>,
```

```
#> #   sw <dbl>, wlfk <chr>
```

选择行

```
cjb %>%
```

```
  filter_at(vars(4:12), any_vars(. < 60))
```

```
#> # A tibble: 52 x 13
```

```
#> xm          bj xb          yw      sx      wy      zz      ls
```

```
#> <chr>  <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
#> 1 凌诗雨  1101 女          84      55      95      90      86
```

```
#> 2 邓洪涛  1101 男          79      74      75      95      97
```

```
#> 3 罗云    1101 女          92      69      85      91      84
```

```
#> # ... with 49 more rows, and 5 more variables:
```

分组统计

xm	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	zgj
汤海明	男	87	94	89	95	94	94	90	90	89	822
舒江辉	男	92	79	86	98	95	96	89	94	87	816
...
韦永杰	男	81	89	87	97	94	96	81	88	83	796
邵友生	男	88	82	91	89	81	98	89	98	75	791
周黎	女	94	82	96	97	97	98	95	94	88	841
翁柯	女	91	84	96	93	97	94	82	90	83	810
...
穆伶俐	女	88	72	86	94	87	88	89	98	94	796
龚兰秀	女	88	77	95	94	84	94	87	94	82	795
舒亚	女	94	81	88	91	85	98	81	88	88	794

分组统计

```
cjb %>%
```

```
  filter(zcj != 0) %>%
```

```
  group_by(xb) %>%
```

```
  summarise(count = n(),  
            max = max(zcj),  
            mean = mean(zcj),  
            min = min(zcj))
```

```
# # A tibble: 2 x 5
```

```
#   xb      count    max  mean  min
```

```
# <fct> <int> <dbl> <dbl> <dbl>
```

```
# 1 男      368    885  793.  523
```

```
# 2 女      406    879  797.  647
```

xm	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	zcj
汤海明	男	87	94	89	95	94	94	90	90	89	822
舒江辉	男	92	79	86	98	95	96	89	94	87	816
...
韦永杰	男	81	89	87	97	94	96	81	88	83	796
邵友生	男	88	82	91	89	81	98	89	98	75	791
周黎	女	94	82	96	97	97	98	95	94	88	841
翁柯	女	91	84	96	93	97	94	82	90	83	810
...
穆伶俐	女	88	72	86	94	87	88	89	98	94	796
龚兰秀	女	88	77	95	94	84	94	87	94	82	795
舒亚	女	94	81	88	91	85	98	81	88	88	794

分组统计

xm	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	zcj
汤海明	男	87	94	89	95	94	94	90	90	89	822
舒江辉	男	92	79	86	98	95	96	89	94	87	816
祁强	男	85	92	82	93	87	88	95	94	93	809
韦永杰	男	81	89	87	97	94	96	81	88	83	796
邰友生	男	88	82	91	89	81	98	89	98	75	791
周黎	女	94	82	96	97	97	98	95	94	88	841
翁柯	女	91	84	96	93	97	94	82	90	83	810
湛容	女	92	82	85	91	90	92	82	98	90	802
穆伶俐	女	88	72	86	94	87	88	89	98	94	796
龚兰秀	女	88	77	95	94	84	94	87	94	82	795
舒亚	女	94	81	88	91	85	98	81	88	88	794

tidyr:长宽变换

`spread(., key, value)`

xm	xb	yw	sx	wy
汤海明	男	87	94	89
舒江辉	男	92	79	86
祁强	男	85	92	82

宽的数据

xm	xb	key	value
汤海明	男	yw	87
汤海明	男	sx	94
汤海明	男	wy	89
舒江辉	男	yw	92
舒江辉	男	sx	79
舒江辉	男	wy	86
祁强	男	yw	85
祁强	男	sx	92
祁强	男	wy	82

长的数据

`gather(., key, value, yw:wy)`

分组统计

xm	xb	key	value
汤海明	男	yw	87
汤海明	男	sx	94
汤海明	男	wy	89
舒江辉	男	yw	92
舒江辉	男	sx	79
舒江辉	男	wy	86
祁强	男	yw	85
祁强	男	sx	92
祁强	男	wy	82

xm	xb	key	value
汤海明	男	yw	87
舒江辉	男	yw	92
祁强	男	yw	85
汤海明	男	sx	94
舒江辉	男	sx	79
祁强	男	sx	92
汤海明	男	wy	89
舒江辉	男	wy	86
祁强	男	wy	82

分组统计

```
cjb %>%
```

```
  gather(key = ke_mu, value = cheng_ji, yw:sw) %>%
```

```
  arrange(xm)
```

```
#> # A tibble: 6,975 x 7
```

```
#>   xm      bj    xb    wlfk    zcj ke_mu cheng_ji
#>   <chr>  <fct> <fct> <fct> <dbl> <chr>    <dbl>
#> 1 艾春莲 1103 女    文科    713 yw      86
#> 2 艾春莲 1103 女    文科    713 sx      59
#> 3 艾春莲 1103 女    文科    713 wy      87
#> 4 艾春莲 1103 女    文科    713 zz      89
#> 5 艾春莲 1103 女    文科    713 ls      85
#> 6 艾春莲 1103 女    文科    713 dl      92
#> # ... with more rows
```

$$6975 = 775 \times 9$$

分组统计

#按科目进行汇总统计

```
cjb %>%
```

```
  filter(zcj != 0) %>%
```

```
  gather(key = ke_mu, value = cheng_ji, yw:sw) %>%
```

```
  group_by(ke_mu) %>%
```

```
  summarise(max = max(cheng_ji),
```

```
             mean = mean(cheng_ji),
```

```
             median = median(cheng_ji),
```

```
             min = min(cheng_ji)) %>%
```

```
  arrange(desc(mean))
```

分组统计

```
#> # A tibble: 9 x 6
```

```
#> ke_mu      max  mean median   min    sd
#> <chr>    <dbl> <dbl>  <dbl> <dbl> <dbl>
#> 1 dl      100   93.0    94    70   4.87
#> 2 zz      100   92.3    93    65   4.56
#> 3 hx      100   91.7    94    52   7.60
#> 4 ls      100   89.1    90     0   7.66
#> 5 wy       99   87.5    88    30   7.00
#> 6 yw       96   87.4    88    33   4.94
#> 7 sw      100   86.4    88    55   8.26
#> 8 sx      100   86.2    89    26  10.5
#> 9 wl      100   81.2    83    21  12.1
```

后话：优雅的R

hipsteR

hipsteR re-educating people who learned R before it was cool

1. Switch to knitr
2. Learn [Hadley Wickham](#)'s packages
3. Adopt the pipe operator
4. Consider RStudio
5. CRAN is huge, and there's also GitHub
6.

<http://kbroman.org/hipsteR>

A decorative blue border with rounded corners frames the entire slide. Two thin blue crosshair symbols are positioned on the right and left sides of the text.

谢谢聆听
Thank you

教师个人联系方式

艾新波

手机: 13641159546

QQ: 23127789

微信: 13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址: 北京邮电大学科研楼917室

课程网址: <https://github.com/byaxb/RDataAnalytics>

