北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

数据对象

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

数据对象

当前位置

设备或人工采集到的数据
（对现实系统进行观测、记录）

其它数据集合

单一属性特征

多属性特征

多属性其他组合

5 列表

属性类型相同、长度相同

单一维度数据集合

3 矩阵

属性类型不同、长度相同

本质上是列表

定距定比

1 向量

6 数据框
（关系表）

形式上相同

二维到高维

定类定序

2 因子

4 数组

观数以形——探索性数据分析

相随相伴、谓之关联——关联

既是世间法、自当有分别——分类

方以类聚、物以群分——聚类

矩阵的创建

#单变量观测值可以用向量或因子存储

#假如对观测对象的多个属性同时进行记录（多变量）

#若这些数据是同质的，宜采用矩阵作为一个整体进行存储

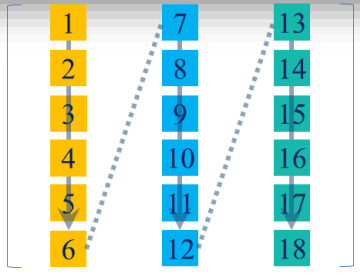
#依然以学生成绩这份数据为例

```
xm <- c("周黎", "汤海明", "舒江辉", "翁柯", "祁强", "湛容")
yw <- c(94, 87, 92, 91, 85, 92)
sx <- c(82, 94, 79, 84, 92, 82)
wy <- c(96, 89, 86, 96, 82, 85)
```

矩阵的创建

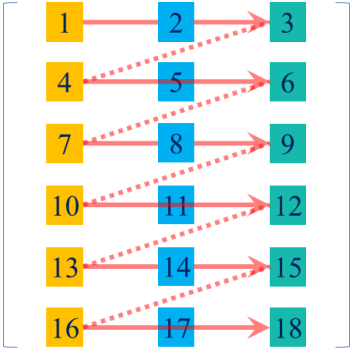
```
#语文、数学、外语三科成绩作为一个整体
ysw <- matrix(c(94, 87, 92, 91, 85, 92,
                82, 94, 79, 84, 92, 82,
                96, 89, 86, 96, 82, 85),
              ncol = 3)
colnames(ysw) <- c("yw", "sx", "wy")
row.names(ysw) <- xm
View(ysw)
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85



矩阵的创建


```
#假如数据本身就是“站”着的
ysw <- matrix(
    c(94, 82, 96,
      87, 94, 89,
      92, 79, 86,
      91, 84, 96,
      85, 92, 82,
      92, 82, 85),
   byrow = TRUE, #注意byrow = 参数的设置
   ncol = 3)
colnames(ysw) <- c("yw", "sx", "wy")
row.names(ysw) <- xm
```



矩阵的基本性质

```
#矩阵的基本性质
colnames(ysw)
#> [1] "yw" "sx" "wy"
row.names(ysw)
#> [1] "周黎"   "汤海明" "舒江辉" "翁柯"   "祁强"   "湛容"
nrow(ysw) #行数
#> [1] 6
ncol(ysw) #列数
#> [1] 3
dim(ysw) #行数和列数
#> [1] 6 3
dimnames(ysw) #行列名称
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

访问矩阵的子集

```
#子集的访问依然是通过[]
#由于矩阵是二维的，需要','来分别指定行和列
ysw[1, ] #第一个同学语文、数学、外语得分
ysw["周黎", ] #同上
#> yw sx wy
#> 94 82 96
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

访问矩阵的子集

```
#子集的访问依然是通过[]
#由于矩阵是二维的，需要','来分别指定行和列
ysw[, 1] #语文成绩
ysw[, "yw"] #同上
#>  周黎     汤海明舒江辉
#>  9487     92
#>  翁柯     祁强   湛容
#>  9185     92
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

访问矩阵的子集

```
#子集的访问依然是通过[]
#由于矩阵是二维的，需要','来分别指定行和列
ysw[1, 1] #第一个同学的第一门课得分
ysw["周黎", "yw"] #第一个同学的第一门课得分
#> [1] 94
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

访问矩阵的子集

```
#子集的访问依然是通过[]
#由于矩阵是二维的，需要','来分别指定行和列
ysw["周黎", 2:3]
ysw[1, c("sx", "wy")]
#> sx wy
#> 82 96
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

访问矩阵的子集

```
#子集的访问依然是通过[]
#由于矩阵是二维的，需要','来分别指定行和列
ysw[1, -1]
#> sx wy
#> 82 96
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

行列重排

```
#列重新排序
ysw[, c("sx", "yw", "wy")]
ysw[, c(2, 1, 3)]

#>      sx   yw   wy
#> 周黎     82   94   96
#> 汤海明   94   87   89
#> 舒江辉 79     92   86
#> 翁柯     84   91   96
#> 祁强     92   85   82
#> 湛容     82   92   85
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

	sx	yw	wy
周黎	82	94	96
汤海明	94	87	89
舒江辉	79	92	86
翁柯	84	91	96
祁强	92	85	82
湛容	82	92	85

行列重排

```
#行进行排序：按照数学成绩进行排序
(order_sx <- order(ysw[, "sx"],
      decreasing = TRUE))
#> [1] 2 5 4 1 6 3
ysw[order_sx, ]
#>          yw sx wy
#> 汤海明     87 94 89
#> 祁强       85 92 82
#> 翁柯       91 84 96
#> 周黎       94 82 96
#> 湛容       92 82 85
#> 舒江辉     92 79 86
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

	yw	sx	wy
汤海明	87	94	89
祁强	85	92	82
翁柯	91	84	96
周黎	94	82	96
湛容	92	82	85
舒江辉	92	79	86

矩阵合并

```
ysw1 <- matrix(
    c(94, 87, 92, 91, 85, 92,
      82, 94, 79, 84, 92, 82,
      96, 89, 86, 96, 82, 85),
  ncol = 3,
  dimnames = list(
    c("周黎", "汤海明", "舒江辉", "翁柯", "祁强", "湛容"),
    c("yw", "sx", "wy")
  )
)
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

矩阵合并

```
ysw2 <- matrix(
    c(88, 81,
      72, 89,
      86, 87),
   ncol = 3,
   dimnames = list(
     c("穆伶俐", "韦永杰"),
     c("yw", "sx", "wy")
   )
)
```

	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85

	yw	sx	wy
穆伶俐	88	72	86
韦永杰	81	89	87

矩阵合并

```
ysw <- rbind(ysw1, ysw2)
```

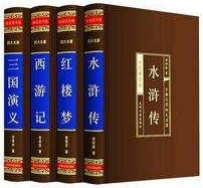
	yw	sx	wy
周黎	94	82	96
汤海明	87	94	89
舒江辉	92	79	86
翁柯	91	84	96
祁强	85	92	82
湛容	92	82	85
穆伶俐	88	72	86
韦永杰	81	89	87

矩阵合并

```
zzls <- matrix(
  c(97, 97,
    95, 94,
    98, 95,
    93, 97,
    93, 87,
    91, 90,
    94, 87,
    97, 94),
  ncol = 2, byrow = TRUE,
  dimnames = list(
    c("周黎", "汤海明", "舒江辉", "翁柯",
      "祁强", "湛容", "穆伶俐", "韦永杰"),
    c("zz", "ls")
  )
)
```

	zz	ls
周黎	97	97
汤海明	95	94
舒江辉	98	95
翁柯	93	97
祁强	93	87
湛容	91	90
穆伶俐	94	87
韦永杰	97	94

矩阵合并



#得到成绩表cjb如下

```
cjb <- cbind(ysw, zzls)
```

	yw	sx	wy	zz	ls
周黎	94	82	96	97	97
汤海明	87	94	89	95	94
舒江辉	92	79	86	98	95
翁柯	91	84	96	93	97
祁强	85	92	82	93	87
湛容	92	82	85	91	90
穆伶俐	88	72	86	94	87
韦永杰	81	89	87	97	94

矩阵的基本操作

```
rowSums(cjb) #每个同学的总成绩
#>  周黎   汤海明  舒江辉   翁柯   祁强   湛容  穆伶俐   韦永杰
#>  466    459     450     461    439    440    427     448
```

	yw	sx	wy	zz	ls	
周黎	94	82	96	97	97	→ 466
汤海明	87	94	89	95	94	→ 459
舒江辉	92	79	86	98	95	→ 450
翁柯	91	84	96	93	97	→ 461
祁强	85	92	82	93	87	→ 439
湛容	92	82	85	91	90	→ 440
穆伶俐	88	72	86	94	87	→ 427
韦永杰	81	89	87	97	94	→ 448

矩阵的基本操作

```
colMeans(cjb) #各门课的平均分
#>  yw      sx      wy      zz      ls
#>  88.75   84.25   88.38   94.75   92.62
```

	yw	sx	wy	zz	ls
周黎	94	82	96	97	97
汤海明	87	94	89	95	94
舒江辉	92	79	86	98	95
翁柯	91	84	96	93	97
祁强	85	92	82	93	87
湛容	92	82	85	91	90
穆伶俐	88	72	86	94	87
韦永杰	81	89	87	97	94
	↓	↓	↓	↓	↓
	88.75	84.25	88.38	94.75	92.62

矩阵的基本操作

```
#更一般的方法
apply(cjb, 1, sum)
#> 周黎   汤海明   舒江辉   翁柯   祁强   湛容   穆伶俐   韦永杰
#> 466    459     450     461    439    440    427     448
apply(cjb, 2, mean)
#> yw    sx    wy    zz    ls
#> 88.75 84.25 88.38 94.75 92.62
round(apply(cjb, 2, sd), digits = 2)
#> yw   sx   wy   zz   ls
#> 4.33 7.23 5.10 2.43 4.10
```

	yw	sx	wy	zz	ls	
周黎	94	82	96	97	97	466
汤海明	87	94	89	95	94	459
舒江辉	92	79	86	98	95	450
翁柯	91	84	96	93	97	461
祁强	85	92	82	93	87	439
湛容	92	82	85	91	90	440
穆伶俐	88	72	86	94	87	427
韦永杰	81	89	87	97	94	448
	88.75	84.25	88.38	94.75	92.62	

矩阵的基本操作

```
#可以自定义函数
coefficient_of_variation <- function(x) {
    sd(x) / mean(x)
}
apply(cjb, 2, coefficient_of_variation)
#当然，也可以采用匿名函数
apply(cjb, 2, function(x) {
    sd(x) / mean(x)
})
#> yw    sx    wy    zz    ls
#> 0.049 0.086 0.058 0.026 0.044
```

apply函数族

- apply代表了一种数据处理模式
- split-apply-combine模式
- apply函数族，以及tidyverse包等

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