

包

周世祥

0 0年 月11日

简介

开发者 是 - 的用户，因为商业化，开发了 . 包。轻松处理 级数据。 . 既是 包 的名字，也是一种数据格式，作为 . 的升级版。

```
library(data.table)
library(magrittr)
```

```
# 通过fwrite函数将部分mtcars数据集的内容输出到mtcars_DT1.csv文件中，并用fread函数读取到R中。
fwrite(mtcars[1:5, 1:3], file="./RawData/mtcars_DT1.csv", row.names =TRUE)
fread("./RawData/mtcars_DT1.csv")
```

```
##           V1  mpg  cyl disp
## 1:      Mazda RX4 21.0   6  160
## 2:    Mazda RX4 Wag 21.0   6  160
## 3:    Datsun 710  22.8   4  108
## 4:  Hornet 4 Drive 21.4   6  258
## 5: Hornet Sportabout 18.7   8  360
```

```
fread("./RawData/mtcars_DT.csv", msip =
```

```
# 从指定的字符串位置开始读取
# fread("./RawData/mtcars_DT1.csv", skip = "710")
# fread("./RawData/mtcars_DT1.csv", skip = "Drive")
```

```
fread("./RawData/mtcars_DT.csv", select = c("V1", "cyl"))
```

```
##           V1 cyl
## 1:      Mazda RX4    6
## 2:    Mazda RX4 Wag    6
## 3:      Datsun 710    4
## 4:    Hornet 4 Drive    6
## 5: Hornet Sportabout    8
```

```
# 选择或丢弃某些列
# fread("./RawData/mtcars_DT.csv", select = c(1, 3))
# fread("./RawData/mtcars_DT.csv", drop = "cyl")
# fread("./RawData/mtcars_DT.csv", drop = 2)
```

，， 数据处理的句式

```
DT %>%
  filter(i) %>%
  select(j) %>%
  group_by()
```

```
## Error in eval(lhs, parent, parent): 找不到对象'DT'
```

数据是从1969年美国一本关于汽车的杂志中提取的数据，对1969款车型10个方面的数据进行整理，如气缸数量，马力，等等。

```
DT <- data.table(mtcars, keep.rownames = TRUE)
unique(DT$rn) #看所有车型
```

```
## [1] "Mazda RX4"           "Mazda RX4 Wag"       "Datsun 710"
## [4] "Hornet 4 Drive"      "Hornet Sportabout"   "Valiant"
## [7] "Duster 360"         "Merc 240D"           "Merc 230"
## [10] "Merc 280"           "Merc 280C"           "Merc 450SE"
## [13] "Merc 450SL"         "Merc 450SLC"         "Cadillac Fleetwood"
## [16] "Lincoln Continental" "Chrysler Imperial"   "Fiat 128"
## [19] "Honda Civic"         "Toyota Corolla"      "Toyota Corona"
## [22] "Dodge Challenger"   "AMC Javelin"         "Camaro Z28"
## [25] "Pontiac Firebird"   "Fiat X1-9"           "Porsche 914-2"
## [28] "Lotus Europa"       "Ford Pantera L"      "Ferrari Dino"
## [31] "Maserati Bora"      "Volvo 142E"
```

```
DT[rn == "Datsun 710"] #查询这款
```

```
##           rn mpg cyl disp hp drat   wt  qsec vs am gear carb
## 1: Datsun 710 22.8   4  108 93 3.85 2.32 18.61  1  1    4    1
```

```
DT[mpg < 18 & cyl == 6 ] #查油耗大的，即每加仑可以运行的里程mpg小，气缸为6的
```

```
##           rn  mpg cyl  disp  hp drat   wt  qsec vs am gear carb
## 1: Merc 280C 17.8   6 167.6 123 3.92 3.44 18.9  1  0    4    4
```

```
#拥有5个档位的，油耗比为21的
DT[gear == 5 | mpg == 21 ] # 逻辑值进行随机组合，data.frame做不到
```

```
##           rn  mpg cyl  disp  hp drat   wt  qsec vs am gear carb
## 1:   Mazda RX4 21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## 2:   Mazda RX4 Wag 21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## 3:   Porsche 914-2 26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## 4:   Lotus Europa 30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## 5: Ford Pantera L 15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## 6:   Ferrari Dino 19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## 7: Maserati Bora 15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
```

```
# 选取DT中的3列
DT[,.(rn,mpg,cyl)] %>%
  head()
```

```
##           rn  mpg cyl
## 1:   Mazda RX4 21.0   6
## 2:   Mazda RX4 Wag 21.0   6
## 3:   Datsun 710 22.8   4
## 4:   Hornet 4 Drive 21.4   6
## 5: Hornet Sportabout 18.7   8
## 6:   Valiant 18.1   6
```

```
# 行列配合在一起设置
models <- c("Merc 240D", "Merc 230", "Merc 280")
DT[rn %in% models, .(rn,mpg,cyl,hp,gear)]
```

```
##           rn  mpg cyl  hp gear
## 1: Merc 240D 24.4   4  62    4
## 2: Merc 230 22.8   4  95    4
## 3: Merc 280 19.2   6 123    4
```

```
# 分组计算，分别对三个不同的列，进行油耗比平均值，
# 最大车重，最大马力的计算，对计算后的数据集按气缸数量从小到大进行排序
DT[,.(mpg_mean = mean(mpg)), by = cyl][order(cyl)]
```

```
##      cyl mpg_mean
## 1:    4 26.66364
## 2:    6 19.74286
## 3:    8 15.10000
```

```
DT[,.(wt_max = max(wt)), by = cyl][order(cyl)]
```

```
##      cyl wt_max
## 1:     4  3.190
## 2:     6  3.460
## 3:     8  5.424
```

```
DT[, .(hp_max = max(hp)), by = cyl][order(cyl)]
```

```
##      cyl hp_max
## 1:     4    113
## 2:     6    175
## 3:     8    335
```

```
DT[vs == 1 , .(hp_max = max(hp)), by = cyl][order(cyl)]
```

```
##      cyl hp_max
## 1:     4    113
## 2:     6    123
```

是 `data.table` 包中的函数，两者完全兼容。代码采用两个连续的中括弧来对数据进行连续处理。与 `base` 包中的管道函数 `%>%` 功能一致。假如是小数据集，建议用 `base` 系列中的计算函数。如果超过百万行以上的数据集，强烈推荐 `data.table` 的计算功能。

参考文献

刘健 邬书豪，《数据科学实战工具详解与案例分析》，机械工业出版社，01 年 月。