数据分析工具实践
DPLYR包

刘春晓 1600022704

大纲

- 一、dplyr的简介和安装
- 二、dplyr包安装及载入
- 三、dplyr包的下述六个函数用法
- 四、dplyr包的总结评价

一、DPLYR的简介和安装

- dplyr包是Hadley Wickham的新作，主要用于**数据清洗**和**整理**，该包专注dataframe数据格式，从而大幅提高了数据处理速度，并且提供了与其它数据库的接口。

二、DPLYR包安装及载入

- >install.packages("dplyr")
- >library(dplyr)

三、DPLYR包的下述六个函数用法：

- 筛选: filter()
- 排列: arrange()
- 选择: select()
- 变形: mutate()
- 汇总: summarise()
- 分组: group_by()

3.1 筛选：FILTER()

- 按给定的逻辑判断筛选出**符合要求**的子数据集

```
filter(mtcars_df,mpg==21,hp==110)

#  A tibble: 2 x 11
    mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1    21     6   160   110   3.9 2.620 16.46     0     1     4     4
2    21     6   160   110   3.9 2.875 17.02     0     1     4     4
```

3.1 筛选: FILTER()~续

- 按给定的逻辑判断筛选出符合要求的子数据集, 类似于 base::subset() 函数
- 例如:
- filter(mtcars_df, mpg == 21, hp== 110)

3.2 排列: ARRANGE()

- 按给定的列名依次对行进行排序

```
arrange(mtcars_df, disp) #可对列名加 desc(disp) 进行倒序

# A tibble: 32 x 11
     mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1   33.9     4  71.1    65  4.22 1.835 19.90     1     1     4     1
2   30.4     4  75.7    52  4.93 1.615 18.52     1     1     4     2
3   32.4     4  78.7    66  4.08 2.200 19.47     1     1     4     1
4   27.3     4  79.0    66  4.08 1.935 18.90     1     1     4     1
5   30.4     4  95.1   113  3.77 1.513 16.90     1     1     5     2
6   22.8     4 108.0    93  3.85 2.320 18.61     1     1     4     1
7   21.5     4 120.1    97  3.70 2.465 20.01     1     0     3     1
8   26.0     4 120.3    91  4.43 2.140 16.70     0     1     5     2
9   21.4     4 121.0   109  4.11 2.780 18.60     1     1     4     2
10  22.8     4 140.8    95  3.92 3.150 22.90     1     0     4     2
# ... with 22 more rows
```

3.2 排列: ARRANGE()~续

- 按给定的列名依次对行进行排序.
- 例如:
- arrange(mtcars_df, disp)
- 对列名加 desc() 进行倒序;

- arrange(mtcars_df, desc(disp))
- 这个函数和 plyr::arrange() 是一样的, 类似于 order()

3.3 选择: SELECT()

```
select(mtcars_df, disp:wt)

# A tibble: 32 x 4
    disp    hp  drat    wt
*  <dbl> <dbl> <dbl> <dbl>
1  160.0   110  3.90 2.620
2  160.0   110  3.90 2.875
3  108.0    93  3.85 2.320
4  258.0   110  3.08 3.215
5  360.0   175  3.15 3.440
6  225.0   105  2.76 3.460
7  360.0   245  3.21 3.570
8  146.7    62  3.69 3.190
9  140.8    95  3.92 3.150
10 167.6   123  3.92 3.440
# ... with 22 more rows
```

3.3 选择: SELECT()~续

- 用列名作参数来选择子数据集:
- select(mtcars_df, disp—wt)
- 还可以用 : 来连接列名, 没错, 就是把列名当作数字一样使用:
- select(mtcars_df, disp—wt)
- 用 - 来排除列名:
- select(mtcars_df, disp—wt, -(Drat:3.21))

3.4 变形：MUTATE()

```
mutate(mtcars_df,
  NO = 1:dim(mtcars_df)[1])

# A tibble: 32 x 12
     mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb    NO
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>
1   21.0     6 160.0   110  3.90 2.620 16.46     0     1     4     4     1
2   21.0     6 160.0   110  3.90 2.875 17.02     0     1     4     4     2
3   22.8     4 108.0    93  3.85 2.320 18.61     1     1     4     1     3
4   21.4     6 258.0   110  3.08 3.215 19.44     1     0     3     1     4
5   18.7     8 360.0   175  3.15 3.440 17.02     0     0     3     2     5
6   18.1     6 225.0   105  2.76 3.460 20.22     1     0     3     1     6
7   14.3     8 360.0   245  3.21 3.570 15.84     0     0     3     4     7
8   24.4     4 146.7    62  3.69 3.190 20.00     1     0     4     2     8
9   22.8     4 140.8    95  3.92 3.150 22.90     1     0     4     2     9
10  19.2     6 167.6   123  3.92 3.440 18.30     1     0     4     4    10
# ... with 22 more rows
```

3.4 变形: MUTATE()~续

- 对已有列进行数据运算并添加为新列:

- mutate(mtcars_df,
- No=1：dim（mtcars_df）[1])

3.5 汇总：SUMMARISE()

- 对数据框调用其它函数进行**汇总**操作，返回一维的结果：

```
summarise(mtcars_df,
    mdisp = mean(disp, na.rm = TRUE))
# A tibble: 1 x 1
     mdisp
     <dbl>
1 230.7219
```

3.5 汇总：SUMMARISE()~续

- 对数据框调用其它函数进行汇总操作, 返回一维的结果:
- summarise(mtcars_df,
- mdisp=mean(disp,na.rm=TRUE)))

- 等同于 plyr::summarise(), 原文说该函数功能尚不是非常有用, 大概以后的更新会加强吧.

3.6 分组: GROUP_BY()

- 当对数据集通过group_by()添加了分组信息后，mutate()，arrange()和 summarise() 函数会自动对这些 tbl 类数据执行分组操作。

```
cars <- group_by(mtcars_df, cyl)
countcars <- summarise(cars, count = n()) # count = n()用来计算次数

# A tibble: 3 x 2
    cyl count
  <dbl> <int>
1     4    11
2     6     7
3     8    14
```

四、DPLYR包的总评

- R包dplyr可用于处理R内部或者外部的结构化数据，相较于plyr包，dplyr专注接受dataframe对象，大幅提高了速度,并且提供了更稳健的数据库接口。同时，dplyr包可用于操作Spark的dataframe。

- dplyr包还有一部分高级应用，但是因为时间等原因无法继续向大家分享。希望有时间再继续用dplyr包进行后续的高级应用工作。

谢谢