Como usar a função filter()?

Gustavo Paterno

12/4/2018

Contents

Introdução

O pacote `dplyr` implementa a **gramática** da manipulação de dados. O pacote oferece diversas funções (verbos) para manipular e organizar tabelas de dados diretamente do R. Para aprender mais sobre o pacote e suas funções visite este site

`filter()`

Neste breve tutorial iremos aprender um pouco sobre a função `filter`. Basicamente ele serve para filtrar um banco de dados baseado em condições definidas pelo usuário. Essa função funciona de forma parecida com a opção **filtro** do excel. A ideia geral é a mesma: filtrar as linhas que atendam certas condições.

Mão na massa

Carregando o pacote

```
library(dplyr) # caso não tenha instalado ainda: install.packages("dplyr")
```

Carregando dados ("iris")

O banco de dados iris mostra o tamanho e comprimento das pétas e sépalas de três espécies de planta. para saber mais sobre o banco de dados utilize o comando `?iris`. Veja as três primeiras linhas do banco de dados abaixo:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa

Carregando bancos de dados:

```
data(iris)
### estrutura dos dados
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 ...
```

```
### Primeiras linhas
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
### Últimas linhas
tail(iris)
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## 145          6.7         3.3          5.7         2.5 virginica
## 146          6.7         3.0          5.2         2.3 virginica
## 147          6.3         2.5          5.0         1.9 virginica
## 148          6.5         3.0          5.2         2.0 virginica
## 149          6.2         3.4          5.4         2.3 virginica
## 150          5.9         3.0          5.1         1.8 virginica
```

```
### Quais espécies existem no banco de dados?
levels(iris$Species)
```

```
## [1] "setosa"     "versicolor" "virginica"
```

Filtrar por categoria

Se eu quiser selecionar apenas os dados das espécie “setosa”?

```
filter(iris, Species == "setosa")
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4         0.2  setosa
## 2           4.9         3.0          1.4         0.2  setosa
## 3           4.7         3.2          1.3         0.2  setosa
## 4           4.6         3.1          1.5         0.2  setosa
## 5           5.0         3.6          1.4         0.2  setosa
## 6           5.4         3.9          1.7         0.4  setosa
## 7           4.6         3.4          1.4         0.3  setosa
## 8           5.0         3.4          1.5         0.2  setosa
## 9           4.4         2.9          1.4         0.2  setosa
## 10          4.9         3.1          1.5         0.1  setosa
```

```
## 11          5.4         3.7          1.5         0.2  setosa
## 12          4.8         3.4          1.6         0.2  setosa
## 13          4.8         3.0          1.4         0.1  setosa
## 14          4.3         3.0          1.1         0.1  setosa
## 15          5.8         4.0          1.2         0.2  setosa
## 16          5.7         4.4          1.5         0.4  setosa
## 17          5.4         3.9          1.3         0.4  setosa
## 18          5.1         3.5          1.4         0.3  setosa
## 19          5.7         3.8          1.7         0.3  setosa
## 20          5.1         3.8          1.5         0.3  setosa
## 21          5.4         3.4          1.7         0.2  setosa
## 22          5.1         3.7          1.5         0.4  setosa
## 23          4.6         3.6          1.0         0.2  setosa
## 24          5.1         3.3          1.7         0.5  setosa
## 25          4.8         3.4          1.9         0.2  setosa
## 26          5.0         3.0          1.6         0.2  setosa
## 27          5.0         3.4          1.6         0.4  setosa
## 28          5.2         3.5          1.5         0.2  setosa
## 29          5.2         3.4          1.4         0.2  setosa
## 30          4.7         3.2          1.6         0.2  setosa
## 31          4.8         3.1          1.6         0.2  setosa
## 32          5.4         3.4          1.5         0.4  setosa
## 33          5.2         4.1          1.5         0.1  setosa
## 34          5.5         4.2          1.4         0.2  setosa
## 35          4.9         3.1          1.5         0.2  setosa
## 36          5.0         3.2          1.2         0.2  setosa
## 37          5.5         3.5          1.3         0.2  setosa
## 38          4.9         3.6          1.4         0.1  setosa
## 39          4.4         3.0          1.3         0.2  setosa
## 40          5.1         3.4          1.5         0.2  setosa
## 41          5.0         3.5          1.3         0.3  setosa
## 42          4.5         2.3          1.3         0.3  setosa
## 43          4.4         3.2          1.3         0.2  setosa
## 44          5.0         3.5          1.6         0.6  setosa
## 45          5.1         3.8          1.9         0.4  setosa
## 46          4.8         3.0          1.4         0.3  setosa
## 47          5.1         3.8          1.6         0.2  setosa
## 48          4.6         3.2          1.4         0.2  setosa
## 49          5.3         3.7          1.5         0.2  setosa
## 50          5.0         3.3          1.4         0.2  setosa
```

Se eu quiser selecionar apenas os dados das especies “setosa” e “virginica”? Para isso utilize o comando **%in%** no lugar de == e coloque o nome das espécies concatenados pelo c(“setosa”, “virginica”).

```
filter(iris, Species %in% c("setosa", "virginica"))
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## 1            5.1         3.5          1.4         0.2    setosa
## 2            4.9         3.0          1.4         0.2    setosa
## 3            4.7         3.2          1.3         0.2    setosa
## 4            4.6         3.1          1.5         0.2    setosa
## 5            5.0         3.6          1.4         0.2    setosa
## 6            5.4         3.9          1.7         0.4    setosa
## 7            4.6         3.4          1.4         0.3    setosa
## 8            5.0         3.4          1.5         0.2    setosa
```

```
## 9           4.4         2.9          1.4         0.2    setosa
## 10          4.9         3.1          1.5         0.1    setosa
## 11          5.4         3.7          1.5         0.2    setosa
## 12          4.8         3.4          1.6         0.2    setosa
## 13          4.8         3.0          1.4         0.1    setosa
## 14          4.3         3.0          1.1         0.1    setosa
## 15          5.8         4.0          1.2         0.2    setosa
## 16          5.7         4.4          1.5         0.4    setosa
## 17          5.4         3.9          1.3         0.4    setosa
## 18          5.1         3.5          1.4         0.3    setosa
## 19          5.7         3.8          1.7         0.3    setosa
## 20          5.1         3.8          1.5         0.3    setosa
## 21          5.4         3.4          1.7         0.2    setosa
## 22          5.1         3.7          1.5         0.4    setosa
## 23          4.6         3.6          1.0         0.2    setosa
## 24          5.1         3.3          1.7         0.5    setosa
## 25          4.8         3.4          1.9         0.2    setosa
## 26          5.0         3.0          1.6         0.2    setosa
## 27          5.0         3.4          1.6         0.4    setosa
## 28          5.2         3.5          1.5         0.2    setosa
## 29          5.2         3.4          1.4         0.2    setosa
## 30          4.7         3.2          1.6         0.2    setosa
## 31          4.8         3.1          1.6         0.2    setosa
## 32          5.4         3.4          1.5         0.4    setosa
## 33          5.2         4.1          1.5         0.1    setosa
## 34          5.5         4.2          1.4         0.2    setosa
## 35          4.9         3.1          1.5         0.2    setosa
## 36          5.0         3.2          1.2         0.2    setosa
## 37          5.5         3.5          1.3         0.2    setosa
## 38          4.9         3.6          1.4         0.1    setosa
## 39          4.4         3.0          1.3         0.2    setosa
## 40          5.1         3.4          1.5         0.2    setosa
## 41          5.0         3.5          1.3         0.3    setosa
## 42          4.5         2.3          1.3         0.3    setosa
## 43          4.4         3.2          1.3         0.2    setosa
## 44          5.0         3.5          1.6         0.6    setosa
## 45          5.1         3.8          1.9         0.4    setosa
## 46          4.8         3.0          1.4         0.3    setosa
## 47          5.1         3.8          1.6         0.2    setosa
## 48          4.6         3.2          1.4         0.2    setosa
## 49          5.3         3.7          1.5         0.2    setosa
## 50          5.0         3.3          1.4         0.2    setosa
## 51          6.3         3.3          6.0         2.5 virginica
## 52          5.8         2.7          5.1         1.9 virginica
## 53          7.1         3.0          5.9         2.1 virginica
## 54          6.3         2.9          5.6         1.8 virginica
## 55          6.5         3.0          5.8         2.2 virginica
## 56          7.6         3.0          6.6         2.1 virginica
## 57          4.9         2.5          4.5         1.7 virginica
## 58          7.3         2.9          6.3         1.8 virginica
## 59          6.7         2.5          5.8         1.8 virginica
## 60          7.2         3.6          6.1         2.5 virginica
## 61          6.5         3.2          5.1         2.0 virginica
## 62          6.4         2.7          5.3         1.9 virginica
```

```
## 63           6.8         3.0          5.5         2.1 virginica
## 64           5.7         2.5          5.0         2.0 virginica
## 65           5.8         2.8          5.1         2.4 virginica
## 66           6.4         3.2          5.3         2.3 virginica
## 67           6.5         3.0          5.5         1.8 virginica
## 68           7.7         3.8          6.7         2.2 virginica
## 69           7.7         2.6          6.9         2.3 virginica
## 70           6.0         2.2          5.0         1.5 virginica
## 71           6.9         3.2          5.7         2.3 virginica
## 72           5.6         2.8          4.9         2.0 virginica
## 73           7.7         2.8          6.7         2.0 virginica
## 74           6.3         2.7          4.9         1.8 virginica
## 75           6.7         3.3          5.7         2.1 virginica
## 76           7.2         3.2          6.0         1.8 virginica
## 77           6.2         2.8          4.8         1.8 virginica
## 78           6.1         3.0          4.9         1.8 virginica
## 79           6.4         2.8          5.6         2.1 virginica
## 80           7.2         3.0          5.8         1.6 virginica
## 81           7.4         2.8          6.1         1.9 virginica
## 82           7.9         3.8          6.4         2.0 virginica
## 83           6.4         2.8          5.6         2.2 virginica
## 84           6.3         2.8          5.1         1.5 virginica
## 85           6.1         2.6          5.6         1.4 virginica
## 86           7.7         3.0          6.1         2.3 virginica
## 87           6.3         3.4          5.6         2.4 virginica
## 88           6.4         3.1          5.5         1.8 virginica
## 89           6.0         3.0          4.8         1.8 virginica
## 90           6.9         3.1          5.4         2.1 virginica
## 91           6.7         3.1          5.6         2.4 virginica
## 92           6.9         3.1          5.1         2.3 virginica
## 93           5.8         2.7          5.1         1.9 virginica
## 94           6.8         3.2          5.9         2.3 virginica
## 95           6.7         3.3          5.7         2.5 virginica
## 96           6.7         3.0          5.2         2.3 virginica
## 97           6.3         2.5          5.0         1.9 virginica
## 98           6.5         3.0          5.2         2.0 virginica
## 99           6.2         3.4          5.4         2.3 virginica
## 100          5.9         3.0          5.1         1.8 virginica
```

Filtrar por número

Se eu quiser filtrar pelo valor de alguma variável, por exemplo, selecionar apenas as linhas nas quais a Sepal.Length é maior que 7?

```
filter(iris, Sepal.Length  > 7)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## 1           7.1         3.0          5.9         2.1 virginica
## 2           7.6         3.0          6.6         2.1 virginica
## 3           7.3         2.9          6.3         1.8 virginica
## 4           7.2         3.6          6.1         2.5 virginica
## 5           7.7         3.8          6.7         2.2 virginica
## 6           7.7         2.6          6.9         2.3 virginica
## 7           7.7         2.8          6.7         2.0 virginica
```

```
## 8           7.2         3.2          6.0         1.8 virginica
## 9           7.2         3.0          5.8         1.6 virginica
## 10          7.4         2.8          6.1         1.9 virginica
## 11          7.9         3.8          6.4         2.0 virginica
## 12          7.7         3.0          6.1         2.3 virginica
```

Se eu quiser filtrar pelo valor de alguma variável, por exemplo, selecionar apenas as linhas nas quais a Sepal.Length é menor que 5?

```
filter(iris, Sepal.Length  < 5)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1           4.9         3.0          1.4         0.2     setosa
## 2           4.7         3.2          1.3         0.2     setosa
## 3           4.6         3.1          1.5         0.2     setosa
## 4           4.6         3.4          1.4         0.3     setosa
## 5           4.4         2.9          1.4         0.2     setosa
## 6           4.9         3.1          1.5         0.1     setosa
## 7           4.8         3.4          1.6         0.2     setosa
## 8           4.8         3.0          1.4         0.1     setosa
## 9           4.3         3.0          1.1         0.1     setosa
## 10          4.6         3.6          1.0         0.2     setosa
## 11          4.8         3.4          1.9         0.2     setosa
## 12          4.7         3.2          1.6         0.2     setosa
## 13          4.8         3.1          1.6         0.2     setosa
## 14          4.9         3.1          1.5         0.2     setosa
## 15          4.9         3.6          1.4         0.1     setosa
## 16          4.4         3.0          1.3         0.2     setosa
## 17          4.5         2.3          1.3         0.3     setosa
## 18          4.4         3.2          1.3         0.2     setosa
## 19          4.8         3.0          1.4         0.3     setosa
## 20          4.6         3.2          1.4         0.2     setosa
## 21          4.9         2.4          3.3         1.0 versicolor
## 22          4.9         2.5          4.5         1.7  virginica
```

Filtrar por intervalo

Se eu quiser filtrar pelo **intervalo** de alguma variável, por exemplo, selecionar apenas as linhas nas quais a Sepal.Length está entre 6.5 e 7 (maior que 6.5 e menor7)? Utilize o comando `&` para adicionar condições.

```
filter(iris, Sepal.Length > 6.5 & Sepal.Length < 7)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1           6.9         3.1          4.9         1.5 versicolor
## 2           6.6         2.9          4.6         1.3 versicolor
## 3           6.7         3.1          4.4         1.4 versicolor
## 4           6.6         3.0          4.4         1.4 versicolor
## 5           6.8         2.8          4.8         1.4 versicolor
## 6           6.7         3.0          5.0         1.7 versicolor
## 7           6.7         3.1          4.7         1.5 versicolor
## 8           6.7         2.5          5.8         1.8  virginica
## 9           6.8         3.0          5.5         2.1  virginica
## 10          6.9         3.2          5.7         2.3  virginica
## 11          6.7         3.3          5.7         2.1  virginica
## 12          6.9         3.1          5.4         2.1  virginica
```

```
## 13          6.7         3.1          5.6         2.4  virginica
## 14          6.9         3.1          5.1         2.3  virginica
## 15          6.8         3.2          5.9         2.3  virginica
## 16          6.7         3.3          5.7         2.5  virginica
## 17          6.7         3.0          5.2         2.3  virginica
```

Filtrar por número e categoria ao mesmo tempo

Se eu quiser selecionar apenas as linhas da espécies setosa com sépalas maiores que 5.5? Novamente, utilize o comando `&` para adicionar condições. Como no `ggplot` a função filter vai adicionando camadas com novas condições.

```
filter(iris, Species == "setosa", Sepal.Length > 5.5)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.8         4.0          1.2         0.2  setosa
## 2          5.7         4.4          1.5         0.4  setosa
## 3          5.7         3.8          1.7         0.3  setosa
```

Ou pelo intervalo (>5.5 e < 6)

```
filter(iris, Species == "setosa", Sepal.Length > 5.5 & Sepal.Length < 6)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.8         4.0          1.2         0.2  setosa
## 2          5.7         4.4          1.5         0.4  setosa
## 3          5.7         3.8          1.7         0.3  setosa
```

Se por um acaso as condições que você definir não existirem, o resultado será um data.frame vazio. Veja por exemplo, se eu solicitar um filtro com o nome da espécie escrito errado ("setoza") ou um tamaho de pétala maior do que existe nos dados.

```
# Nome do nível da variável incorreto
filter(iris, Species == "setoza")
```

```
## [1] Sepal.Length Sepal.Width  Petal.Length Petal.Width  Species
## <0 rows> (or 0-length row.names)
```

```
# valores que não existem no banco de dados
filter(iris, Sepal.Length > 22)
```

```
## [1] Sepal.Length Sepal.Width  Petal.Length Petal.Width  Species
## <0 rows> (or 0-length row.names)
```

Dicas úteis

Para descobrir quais variáveis de um banco de dados

```
colnames(iris)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
## [5] "Species"
```

Para descobrir quais níveis de uma variável (ex. quais espécies estão dentro da variável Species?). A função `distinct` retorna os nomes dos elementos diferentes dentro de uma coluna (neste caso Species) de um banco de dados.

```
distinct(iris, Species)
```

```
##      Species
## 1     setosa
## 2 versicolor
## 3  virginica
```