Data Manipulation

Marcella Pedro

04/10/2021

Manipulação de Dados

É muito comum manipular datasets ao analisar dados. Abaixo seguem alguns comandos mais básicos e importantes para essa atividade.

Instalar Pacotes

Antes mesmo de importar bibliotecas, pode ser necessário instalar o pacote e assim poder importa-lo ao projeto. As bibliotecas ficam alocadas no repositório Comprehensive R Archive Network (CRAN) junto de muitas informações relevantes. No exemplo abaixo o código irá instalar a biblioteca **readr**.

```
install.packages("readr")
```

Importar Bibliotecas

O comando abaixo demostra como importar uma biblioteca uma vez que já instalada.

A biblioteca **readr** é destinada para realizar a leitura de arquivos com dados tabulados e separados por um delimitador.

CSV é um exemplo de dados em formato de tabela separados por vírgula ou ponto e vírgula.

```
library(readr)
```

Importando Dataset

No trecho abaixo estamos importando o dataset do Titanic disponível no Kaggle:

Obs.: No RStudio existe uma forma de importar datasets via interface, através do botão "Import Dataset", escolhendo o tipo de arquivo a ser importado e internamente o próprio RStudio já executa a importação da biblioteca necessária e a importação do arquivo.

```
titanic <- read_csv("titanic.csv")
```

```
## Rows: 891 Columns: 12

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Name, Sex, Ticket, Cabin, Embarked
## dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare
```

```
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Informações sobre o Dataset

Para checar os dados que acabamos de importar, aqui vão algumas funções importantes:

Quantidade de Colunas

```
ncol(titanic)
```

```
## [1] 12
```

Quantidade de linhas

```
nrow(titanic)
```

```
## [1] 891
```

Nome das Colunas

```
colnames(titanic)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

Ou se preferir...

```
names(titanic)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

Visualizar primeiras linhas

```
head(titanic)
```

```
## # A tibble: 6 x 12
##   PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin
##         <dbl>    <dbl>  <dbl> <chr>   <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
## 1           1        0      3 Braund~ male     22     1     0 A/5 2~  7.25 <NA>
## 2           2        1      1 Cuming~ fema~    38     1     0 PC 17~ 71.3  C85
## 3           3        1      3 Heikki~ fema~    26     0     0 STON/~  7.92 <NA>
## 4           4        1      1 Futrel~ fema~    35     1     0 113803 53.1  C123
## 5           5        0      3 Allen,~ male     35     0     0 373450  8.05 <NA>
## 6           6        0      3 Moran,~ male     NA     0     0 330877  8.46 <NA>
## # ... with 1 more variable: Embarked <chr>
```

Visualizar últimas linhas

```
tail(titanic)
```

```
## # A tibble: 6 x 12
##   PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin
##         <dbl>    <dbl>  <dbl> <chr>   <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
## 1         886        0      3 "Rice,~ fema~    39     0     5 382652 29.1  <NA>
## 2         887        0      2 "Montv~ male     27     0     0 211536 13    <NA>
## 3         888        1      1 "Graha~ fema~    19     0     0 112053 30    B42
## 4         889        0      3 "Johns~ fema~    NA     1     2 W./C.~ 23.4  <NA>
## 5         890        1      1 "Behr,~ male     26     0     0 111369 30    C148
## 6         891        0      3 "Doole~ male     32     0     0 370376  7.75 <NA>
## # ... with 1 more variable: Embarked <chr>
```

Gráficos

Um exemplo muito simples de como visualizar os dados de forma gráfica:

```
library(ggplot2)

qplot(x = PassengerId,
      y = Age,
      color = Sex,
      data = titanic)
```

```
## Warning: Removed 177 rows containing missing values (geom_point).
```