Apostila do Minicurso: introdução ao pacote dplyr

Msc.Elisângela C. BiazattiDouglas Vinícius Jossivana Macedo

22 de outubro de 2019

Contents

Chapter 1

Prefácio

Este material foi elaborado com o proprósito de um minicurso, que tem como objetivo apresentar algumas ideias das funções básicas do pacote `dplyr`, podemos usar um computador e um pouco de criatividade para explorar essas idéias em uma variedade de situações. Usamos R com o RStudio para fazer todo o nosso trabalho.

O livro R for data science é o mais indicado para aprender sobre o universo `tidyverse`. Nesse minicurso abordamos mais sobre a gramática das funções básicas do `dplyr` alguns exemplos e exercícios abordados.

1.1 Público-alvo

- Estudantes de estatística que desejam ganhar tempo nos trabalhos da faculdade;
- Acadêmicos com interesse em tornar suas análises e códigos mais legíveis em R.

1.2 Conteúdo:

```
- Primeiro dia (22/10): R básico, swirl, organização de dados, exercícios;
- Segundo dia (23/10): select(), filter(), arrange(), mutate(), summarise(), exercícios;
- Terceiro dia (24/10): agrupar dados, combinar conjuntos de dados.
```

1.3 Pré-requisitos

Chapter 2

Introdução

> "Existem apenas dois tipos de idiomas: os que as pessoas reclamam e os que ninguém usa". - Bjarne Stroustrup

O modelo típico de análise de dados é similar:

Primeiramente, você deve **importar** seus dados para o R. Significa que você pega os dados armazenados em um arquivo, banco de dados ou API da Web e carrega-os em um data frames no R.

Logo após, a ideia é **organizá**-los. Significa armazená-los de forma consistente.

Depois de arrumar os dados, o próximo passo é **transformá**-los. Significa restringir observações de interesse, criar novas variáveis

Depois de organizar os dados com as variáveis necessárias, existem dois mecanismos principais de geração de conhecimento: visualização e modelagem. Eles têm pontos fortes e fracos complementares, portanto qualquer análise real se repetirá entre eles várias vezes.

A **visualização** é uma atividade fundamentalmente humana. Uma boa visualização mostrará coisas que você não esperava, ou fará novas perguntas sobre os dados.

Modelos são ferramentas complementares para visualização. Depois de fazer suas perguntas suficientemente precisas, você pode usar um modelo para respondê-las.

O último passo da ciência de dados é a **comunicação**, uma parte absolutamente crítica de qualquer projeto de análise de dados. Não importa o quão bem seus modelos e visualização levaram você a entender os dados, a menos que você também possa comunicar seus resultados a outras pessoas.

Ao redor de todas essas ferramentas está a programação. A programação é uma ferramenta transversal que você usa em todas as partes do projeto. Você não precisa ser um programador especialista para ser um cientista de dados, mas aprender mais sobre programação compensa, porque se tornar um programador melhor permite automatizar tarefas comuns e resolver novos problemas com maior facilidade.

2.1 R e RStudio

A primeira coisa que você precisa fazer para iniciar o R é instalá-lo no seu computador. O R funciona em praticamente todas as plataformas disponíveis, incluindo os sistemas Windows, Mac OS X e Linux amplamente disponíveis.

Uma nova versão principal do R sai uma vez por ano, e há 2 ou 3 versões menores a cada ano. É uma boa ideia atualizar regularmente. A atualização pode ser um pouco complicada, especialmente para as versões principais, que exigem a reinstalação de todos os seus pacotes.

Há também um ambiente de desenvolvimento integrado (IDE) disponível para o R, construído pelo RStudio. IDE, do inglês **Integrated Development Environment** ou Ambiente de Desenvolvimento Integrado, é um programa de computador que reúne características e ferramentas de apoio ao desenvolvimento de software com o objetivo de agilizar este processo. O RStudio é atualizado duas vezes por ano. Quando uma nova versão estiver disponível, o RStudio informará você.

Você pode ver como instalar o R e o RStudio aqui:

- Instalando o RStudio

Após instalado, o R tem uma interface assim, com apenas o console para digitar comandos:

Experimente um comando: 2+2, cujo output é 4:

```
2 + 2
```

```
## [1] 4
```

E a interface do RStudio é dividida, inicialmente, em 3 partes:

Do lado esquerdo fica o console, onde os comandos podem ser digitados e onde ficam os *outputs*.

No lado superior direito há duas abas:

-i) *Environment*, que é onde ficam armazendos os objetos criados, bases de dados importadas, etc; e

-ii) *History*, onde ficam o histórico dos comandos executados.

A forma mais eficiente e prática de usar o R ou o RStudio é através de um *script*. No RStudio, vá em *File* → *New File* → *R Script*. A interface agora fica dividida em 4 partes:

No *script* você pode digitar comandos a serem executados e também comentários.

2.2 *swirl*

O *swirl* é um pacote do R construído para transformar o console em uma ferramenta interativa para aprender R. *swirl* ensina programação de R e ciência de dados interativamente, no seu próprio ritmo e diretamente no console do R. Para entender melhor do projeto, veja http://swirlstats.com/. Em http:

//swirlstats.com/students. Nestes endereços de são dados os detalhes sobre como usar o *swirl*. Uma vez intalado e carregado o pacote, você é levado a efetuar tarefas:

O *swirl* dá acesso às tarefas de cursos de R que estão disponíveis também no Coursera, como o *R Programming*: *The basics of programming in R*, em https://pt.coursera.org/learn/r-programming. Além deste, estão disponíveis no *swirl*: *Regression Models: The basics of regression modeling in R, Statistical Inference: The basics of statistical inference in R, e Exploratory Data Analysis: The basics of exploring data in R.*

2.3 Universo `tidyverse`

O tidyverse é uma coleção opinativa de pacotes R projetados para ciência de dados. Todos os pacotes compartilham uma filosofia de design, gramática e estruturas de dados subjacentes.

Os princípios fundamentais do `tidyverse` são:

- 1.Reutilizar estruturas de dados existentes;
- 2.Organizar funções simples usando o pipe;

- 3.Aderir à programação funcional;
- 4.Projetado para ser usado por seres humanos.

Assim como o processo típico do passo a passo apresentando anteriormente para análise de dados, o `tidyverse` é a ferramenta que o ajuda eficientemente a executar este processo.

```
library(tidyverse) #Carregar o pacote.
tidyverse_logo() #Logo
```

```
## * __  _    __   .    o          *  .
##  / /_(_)__/ /_ ___  _____ _______ ___
## / __/ / _  / // / |/ / -_) __(_-</ -_)
## \__/_/\_,_/\_, /|___/\__/_/ /___/\__/
##      *  . /___/      o      .       *
```

Chapter 3

O que é Dados organizados?

> "Famílias felizes são todas iguais; toda família infeliz é infeliz à sua maneira." – Leo Tolstoi

> "Os conjuntos de dados organizados são todos iguais, mas todos os conjuntos de dados confusos são confusos à sua maneira." – Hadley Wickham

Você vai precisar instalar os pacotes `tidyr`, `devtools` e `DSR`. Para instalar tidyr e devtools, abra o RStudio e execute o comando:

```
install.packages(c("tidyr", "devtools"))
```

`DSR` é uma coleção de conjuntos de dados. Para instalar DSR, execute o comando:

```
devtools::install_github("garrettgman/DSR")
```

Os dados tabulares podem ser organizados de várias maneiras. Os conjuntos de dados abaixo mostram os mesmos dados organizados de quatro maneiras diferentes, sendo que possuem as mesmas variáveis: país, ano, população e casos. Mas cada conjunto organiza os valores em forma de layout diferente.

```
library(DSR)
# Primeiro conjunto de dados.
table1
```

```
## # A tibble: 6 x 4
```

```
##   country      year  cases population
##   <fct>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

```
# Segundo conjunto de dados.
table2
```

```
## # A tibble: 12 x 4
##    country      year key             value
##    <fct>       <int> <fct>           <int>
##  1 Afghanistan  1999 cases             745
##  2 Afghanistan  1999 population   19987071
##  3 Afghanistan  2000 cases            2666
##  4 Afghanistan  2000 population   20595360
##  5 Brazil       1999 cases           37737
##  6 Brazil       1999 population  172006362
##  7 Brazil       2000 cases           80488
##  8 Brazil       2000 population  174504898
##  9 China        1999 cases          212258
## 10 China        1999 population 1272915272
## 11 China        2000 cases          213766
## 12 China        2000 population 1280428583
```

```
# Terceiro conjunto de dados.
table3
```

```
## # A tibble: 6 x 3
##   country      year rate
##   <fct>       <int> <chr>
## 1 Afghanistan  1999 745/19987071
## 2 Afghanistan  2000 2666/20595360
## 3 Brazil       1999 37737/172006362
## 4 Brazil       2000 80488/174504898
## 5 China        1999 212258/1272915272
## 6 China        2000 213766/1280428583
```

O último conjunto de dados é uma coleção de duas tabelas.

```
# Quarto conjunto de dados.
table4 # cases
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
##   <fct>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

```
table5 # population
```

```
## # A tibble: 3 x 3
##   country         `1999`     `2000`
##   <fct>            <int>      <int>
## 1 Afghanistan   19987071   20595360
## 2 Brazil       172006362  174504898
## 3 China       1272915272 1280428583
```

R segue um conjunto de convenções que tornam um layout de dados tabulares muito mais fácil de trabalhar do que outros. Seus dados serão mais fáceis de trabalhar no R se seguirem três regras:

- 1.Cada variável no conjunto de dados é colocada em sua própria coluna;
- 2.Cada observação é colocada em sua própria linha;
- 3.Cada valor é colocado em sua própria célula.

Os dados que satisfazem essas regras são conhecidos como dados organizados. Observe que `table1` são dados organizados.

Em `table1`, cada variável é colocada em sua própria coluna, cada observação em sua própria linha e cada valor em sua própria célula.

3.1 Introdução ao `tidyr`

O pacote tidyr tem como principal objetivo transformar um data frame para o formato tidy, ou limpo.

De acordo com as regras ditas anteriormente, um dado limpo é aquele com formato *long*, ou seja, com mais linhas. O outro formato é chamado de *wide*, com mais colunas. No caso deste exemplo, ano é uma variável, logo é necessário existir uma coluna com os valores de ano. O valor relacionado a UF naqueles anos também é outra variável, então precisa de uma coluna pra representá-lo. Além disso, a própria UF precisa de uma coluna.

O `tidyr` possui duas funções principais:

`gather`: Transforma um `tibble` *wide* em *long*, ou seja, transforma os dados no formato *tidy*.

`spread`: Transforma um `tibble` *long* em *wide*, ou seja, transforma dados que estão no formato *tidy* em formato não *tidy*.

Além disso, existem duas funções que podem ser importantes na nossa análise: `separate` e `unite`, que separa uma coluna em duas e vice versa.

3.1.1 `gather`

Vamos criar um `tibble` no formato *wide* e transformá-lo em um dado *tidy*:

```
library(tibble)
tb <- tibble(uf = c("RJ", "SP"), `2017` = c(10, 11), `2018` = c(11, 10))
tb
```

```
## # A tibble: 2 x 3
##   uf    `2017` `2018`
##   <chr>  <dbl>  <dbl>
## 1 RJ        10     11
## 2 SP        11     10
```

Chapter 4

Transformação de Dados com `dplyr`

4.1 Introdução

A visualização é uma ferramenta importante para a geração de *insights*, mas é raro você obter os dados exatamente da forma correta de que precisa. Freqüentemente, você precisará criar algumas novas variáveis ou resumos, ou talvez apenas queira renomear as variáveis ou reordenar as observações para tornar os dados um pouco mais fáceis de trabalhar. Você aprenderá como fazer tudo isso (e muito mais!) Neste capítulo, que ensinará como transformar seus dados usando o pacote `dplyr` e um novo conjunto de dados em voos partindo de Nova York em 2013.

4.2 Pré-requisitos

Neste capítulo, vamos nos concentrar em como usar o pacote `dplyr`, outro membro central do `tidyverse`. Ilustraremos as ideias principais usando dados do pacote `nycflights13` e usaremos o `ggplot2` para nos ajudar a entender os dados.

```
library(nycflights13)
library(tidyverse) # ou isoladamente: library(dplyr).
```

4.3 nycflights13

Esse data frames contém todos os 336.776 vôos que partiram de Nova York em 2013. Os dados são do Bureau of Transportation Statistics dos EUA e estão documentados em `?flights`.

```
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515         2      830
##  2  2013     1     1      533            529         4      850
##  3  2013     1     1      542            540         2      923
##  4  2013     1     1      544            545        -1     1004
##  5  2013     1     1      554            600        -6      812
##  6  2013     1     1      554            558        -4      740
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

Para ver todo o conjunto de dados, você pode executar o `View(flights)` que abrirá o conjunto de dados no visualizador do RStudio. Imprime de forma distinta do data frame, porque é um `tibble`. O que é um *tibble*? *Tibbles* são similares aos *data frames*, porém diferentes em dois aspectos: **impressão** e **indexação**

Na impressão no console, os *tibbles* apresentam apenas as dez primeiras linhas e todas as colunas que cabem na tela, tornando mais fácil o trabalho com grandes volumes de dados. Além disso, cada coluna apresenta o seu tipo, algo semelhante ao apresentado quando utilizamos a função `str()`. A segunda diferença, não menos importante, é a forma de indexação. Para indexar um `tibble` devemos utilizar o nome completo da variável que desejamos. Caso contrário, ocorrerá um erro.

Ainda sobre a indexação, sempre que indexarmos um `tibble` usando `[`, o resultado será outro `tibble`. Usando `[[` o resultados será um vetor.

Abreviações de letras sob os nomes das colunas. Eles descrevem o tipo de cada variável:

-`int` significa números inteiros;

`-dbl` significa números duplos ou reais;

`-chr` significa vetores de caracteres ou seqüências de caracteres;

`-dttm` significa data e hora (uma data + uma hora).

`-lgl` significa vetores lógicos que contêm apenas `TRUE` ou `FALSE`;

`-fctr` significa fatores, que R usa para representar variáveis categóricas com valores possíveis fixos.

`-data` significa data.

Existem outros tipos comuns de variáveis que não são usadas neste conjunto de dados.

Em síntese, *data frames* são tabelas de dados. Em seu formato, são bem parecidos com as matrizes, no entanto, possuem algumas diferenças significativas. Podemos idealizar os *data frames* como sendo matrizes em que cada coluna pode armazenar um tipo de dado diferente. Logo, estamos lidando com um objeto bem mais versátil do que as matrizes e os vetores.

Uma das funções básicas mais importantes para começarmos a trabalhar com *data frames* é a `str()`. Essa função dá uma visão clara da estrutura do nosso objeto, bem como informa os tipos de dados existentes.

4.3.1 Exercícios

- 1.Qual a diferença entre uma matriz e um data frame no R?
- 2.Os data frames podem ser indexados com a mesma sintaxe utilizada para matrizes?
- 3.Qual função básica que utilizamos para verificar a estrutura dos dados de um data frame?

4.4 Operador *pipe* %>%

Os tubos são uma ferramenta poderosa para expressar claramente uma sequência de várias operações. O pipe, `%>%` vem do pacote `magrittr` de Stefan Milton

Bache. Pacotes no `tidyverse` carregam `%>%` automaticamente, para que normalmente não carregue o magrittr explicitamente.

Para começar a utilizar o *pipe*, instale e carregue o pacote `magrittr`.

```
install.packages("magrittr")
library(magrittr)
```

Para mais informações sobre o *pipe*, outros operadores relacionados e exemplos de utilização, visite a página Ceci n'est pas un pipe

4.4.1 Exercícios

- 1. Reescreva a expressão abaixo utilizando o `%>%`.

  round(mean(sum(1:10)/3), digits = 1)

Dica: utilize a função `magrittr::divide_by()`. Veja o help da função para mais informações.

- 2. Reescreva o código abaixo utilizando o `%>%`.

  x <- rnorm(100) x.pos <- x[x>0] media <- mean(x.pos) saida <- round(media, 1)

- 3. Sem rodar, diga qual a saída do código abaixo. Consulte o `help` das funções caso precise.

  2 %>% add(2) %>% c(6, NA) %>% mean(na.rm = T) %>% equals(5)

4.5 `filter()`

`filter()` permite agrupar observações com base em seus valores. O primeiro argumento da função é o nome do data frames. Por exemplo, podemos selecionar todos os valores

Chapter 5

Manipulando Data Frames com `dplyr`

5.1 Data Frames

Chapter 6

Final Words

We have finished a nice book.