

Introdução à Programação e Ciência de Dados para a Gestão Pública

Rebeca de Jesus Carvalho

FGV CEPESP

rebeca.jesus.carvalho@gmail.com

Tópicos da aula

- 1 Introdução
- 2 Bases relacionais com a gramática do *dplyr*
- 3 R e Power BI
- 4 Laboratório

Conteúdo da aula de hoje

- Realizaremos uma breve recapitulação dos assuntos abordados até aqui.
- Veremos como combinar *data frames* de diferentes origens que se relacionam por meio de uma ou mais variáveis chave.
- Para isso, utilizaremos novos verbos do *dplyr*, de sufixo `_join` para trabalhar com bases de dados relacionais.
- E, como atividade bônus, aprenderemos como integrar R e Power BI.

O que aprendemos até agora

- Conhecemos a gramática do *dplyr* e suas facilidades na manipulação de dados, especialmente com o emprego do operador **pipe** (`%>%`).
- Aprendemos a filtrar, selecionar e renomear colunas: `filter()`, `select()` e `rename()`.
- Também vimos como ordenar, contar, recodificar, criar novas colunas, agrupar e sumarizar nossos dados: `arrange()`, `count()`, `recode()`, `mutate()`, `group_by()` e `summarise()`.

O que aprendemos até agora

- Descobrimos algumas das funções básicas disponíveis para explorar um *data frame*: `head()`, `glimpse()`, `names()`, `table()` etc.
- Exploramos os operadores aritméticos, relacionais e lógicos e seus diferentes usos.
- Discutimos as diferentes cláusulas condicionais `if`, `else` e `else if` e suas aplicações no R.
- E, finalmente, conhecemos os benefícios das estruturas de repetição: `while` e `for`.

O que aprendemos até agora

- No último encontro, aprendemos a importar e exportar bases de dados no R, principalmente com o pacote *readr*: `read_csv` e `write_csv`.
- Vimos as vantagens de se utilizar o pacote *janitor* para alterar a estética de nossas tabelas: `clean_names()`.
- E, por fim, exploramos a diversidade de visualizações possíveis através da utilização da gramática de gráficos do *ggplot2*.

Combinando *data frames*

- Combinar *data frames* é necessário quando as informações que serão utilizadas na análise estão presentes em mais de uma fonte de dados.
- As variáveis usadas para conectar um par de *data frames* são chamadas de chaves. Uma chave é uma variável (ou conjunto de variáveis) que identifica exclusivamente uma observação. Em casos simples, uma única variável é suficiente para identificar uma observação.

Tipos de *join*

- Há 6 tipos de “*joins*” em R:
 - Interseção dos dois conjuntos, x e y: `inner_join()`;
 - Mantém todas as observações em x, ou seja, à esquerda: `left_join()`;
 - Mantém todas as observações em y, ou seja, à direita: `right_join()`;
 - Mantém todas as observações em x e y: `full_join()`;
 - Mantém todas as observações em x que correspondem a y: `semi_join()`; e
 - Elimina todas as observações em x que correspondem a y: `anti_join()`.

Visão geral das funções de *join* do *dplyr*

ID	X1
1	a1
2	a2

ID	X2
2	b1
3	b2

inner_join

ID	X1	X2
2	a2	b1

left_join

ID	X1	X2
1	a1	NA
2	a2	b1

right_join

ID	X1	X2
2	a2	b1
3	NA	b2

full_join

ID	X1	X2
1	a1	NA
2	a2	b1
3	NA	b2

semi_join

ID	X1
2	a2

anti_join

ID	X1
1	a1

Integrando R e Power BI

- É possível importar e modelar dados dentro do Power BI com scripts de R, tais quais os que trabalhamos no curso, além de produzir gráficos ou rodar algoritmos implementados na linguagem.
- Entre as vantagens dessa integração, destacam-se, sobretudo:
 - A possibilidade de criar visualizações personalizadas, que não estão disponíveis nativamente no Power BI;
 - A vasta biblioteca de pacotes estatísticos do R que podem ser incorporados ao Power BI; e
 - Importação e análise de dados de diferentes fontes diretamente no Power BI.

Laboratório

Agora é o momento de partir para a ação! Temos dois tutoriais agendados para hoje, prontos para serem explorados. Eles estão disponibilizados no GitHub, e você pode acessá-los clicando neste **link**.

Se deixou algum tutorial inacabado dos encontros anteriores, comece por eles. Caso contrário, prossiga.

Dúvidas?