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

Como iremos trabalhar

- Pouca exposição.
- Tutoriais e exercícios.
- Discussão individual e coletiva das dúvidas.
- Leituras complementares.

Conteúdo da aula de hoje

- Vamos conhecer a linguagem de programação R e o software RStudio.
- Conheceremos os *data frames* e algumas funcionalidades básicas para explorá-los.
- Também aprenderemos a importar, organizar e manipular bases de dados no R através de um de seus pacotes mais famosos: o *dplyr*.

R e RStudio

- É uma linguagem de código aberto voltada à manipulação, análise e visualização de dados. O R pode ser utilizado em todo o processo analítico dos dados, como coleta, tratamento, testes estatísticos e apresentação desses a partir de gráficos e mapas. Uma de suas vantagens é a comunidade bastante ativa que constantemente atualiza os pacotes e traz novas funcionalidades para os usuários.

- O RStudio é a interface de desenvolvimento (IDE) na qual realizaremos nossas análises. É um software livre e gratuito.

Data frames

- Os *data frames* são de extrema importância no R, pois são os objetos que armazenam os nossos dados. Eles são equivalentes a uma tabela do SQL ou uma planilha do Excel.

- Um *data frame* é organizado em linhas e colunas, como no exemplo abaixo:

Tabela 01: Cinco maiores cidades do Brasil em população

Município	Estado	População
São Paulo	SP	11.451.245
Rio de Janeiro	RJ	6.211.423
Brasília	DF	2.817.068
Fortaleza	CE	2.428.678
Salvador	BA	2.418.005

Fonte: IBGE.

Funções básicas para *data frames*

- Visualiza o *data frame*: `View()`
- Retorna a primeira ou última parte do *data frame*: `head()` ou `tail()`
- Exibe o número de linhas no *data frame*: `nrow()`
- Exibe o número de colunas no *data frame*: `ncol()`
- Exibe os nomes das colunas do *data frame*: `names()`
- Exibe as colunas e os atributos do *data frame*: `glimpse()`

O pacote *dplyr*

- Conjunto de funções projetadas para permitir a transformação de *data frames* de maneira intuitiva e fácil de usar.
- Os scripts em R que fazem uso inteligente dos verbos *dplyr* tendem a ficar mais legíveis e organizados, sem perder velocidade de execução.
- A utilização é facilitada com o emprego do operador **pipe** (%>%).

Principais funções do *dplyr*

- Renomeia as colunas: `rename()`
- Seleciona as colunas: `select()`
- Filtra linhas: `filter()`
- Cria/modifica colunas: `mutate()`
- Agrupa a base de dados: `group_by()`
- Sumariza a base de dados: `summarise()`
- Ordena a base: `arrange()`

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**.

Dúvidas?