

Minicurso - Tidyverse

Gustavo Almeida Silva

O objetivo deste minicurso é introduzir estudantes com um nível mínimo de conhecimento em R aos pacotes da biblioteca Tidyverse. A partir de problemas reais de análise de dados, apresentaremos técnicas de coleta, limpeza e organização de dados. Para isso, será introduzido o operador pipe: “|>” ou “%>%” e utilizaremos principalmente os pacotes: dplyr e tidyr

Conteúdo Programático

Aula 1: Introdução e motivação ao uso Tidyverse, introdução ao operador Pipe

Nesse primeira aula teremos uma motivação para utilização ao tidyverse, mostrando os pacotes que serão utilizados no minicurso como: dplyr, tidyr, readxl, readr e purrr, e também pacotes mais avançados que não serão abordados no curso, mas que são do tidyverse e utilizam do jeito ‘tidy’ de se programar, como: tidymodels(modelagem), tsibble e fable(dados e modelos de série temporal).

Além disso, o operador pipe será introduzido, operador esse fundamental para se utilizar o tidyverse

Sequência

- Introdução ao minicurso e ao tidyverse
- Motivação: tidymodels(modelagem), tsibble e fable(dados e modelos de série temporal)
- Operador Pipe, para que serve e como utilizar

Aula 2: Carregando Conjunto e Limpeza Inicial de dados e um tibble

Nessa segunda aula os pacote de carregamentos de dados readr e readxl serão introduzidos, o pacote janitor será utilizado para uma limpeza inicial dos dados, e o data.frame do tidyverse, chamado de tibble será introduzido, onde abordaremos as principais diferenças entre um data.frame, um tibble e data.table

Sequência

- Introdução ao conjunto de dados utilizado
- Introdução ao readr e readxl
- Limpeza Inicial com o pacote Janitor
- Entendendo o que é um tibble

Aula 3: Limpeza e Organização dos Dados

A terceira aula marca a introdução dos pacotes dplyr e tidyr ao minucurso. As funções principais de cada pacote serão exemplificadas, mostrando o fluxo de limpeza e organização padrão tidy.

Sequência

- Introdução as funções: **mutate**, **select**, **filter** do pacote Dplyr
- Introdução as funções: **pivot_longer** e **pivot_wider** do pacote Tidyr

Aula 4: Agrupamento e Joins

A quarta aula tem como objetivo demonstrar as formas de agrupamento de dados realizados via Dplyr, além disso a funções de junção de 2 conjunto de dados, os chamados **joins** serão introduzidas

Sequência

- Introdução as funções: **group_by**, **summarise**, **reframe** do pacote Dplyr, utilizados na parte de agrupamento de dados
- Introdução as funções: **joins** do pacote Dplyr, utilizadas em operações entre pares de data.frame

Aula 5: Teste Final

A quinta e última aula tem como objetivo ser um teste final para os alunos onde dado um conjunto de dados 'cru', uma limpeza e organização será realizada. Funções não apresentadas o curso podem ser utilizadas, buscando estimular a curiosidade do aluno em buscar o funcionamento e objetivo de cada uma

Refêrencias:

- Boehmke, Bradley C. (2016) Data Wrangling with R (<http://link.springer.com/book/10.1007%2F978-3-319-45599-0>)
- Grolemond, G & Wickham, H (2023): R for Data Science (<https://r4ds.hadley.nz>)
- Wickham, H. (2014): Tidy Data <https://www.jstatsoft.org/article/view/v059i10>