MBA
USP
ESALQ

DATA WRANGLING

Prof. Dr. Wilson Tarantin Junior

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados, é do professor.

Preparação de Dados no R

Data wrangling

- O que é data wrangling?
  - É o processo de transformar a base de dados de sua estrutura original para a nova estrutura que permitirá a extração de informações
  - Portanto, dado que há uma ou mais bases de dados em mãos, é uma etapa de preparação, organização, manipulação do banco de dados
    - É o processo que ocorre antes da extração de estatísticas descritivas, criação de gráficos e estimação de modelos, por exemplo
  - Dificilmente, as bases de dados estão disponíveis na estrutura mais adequada

Data wrangling

- Utilizaremos, principalmente, o **dplyr**
  - O dplyr é um pacote contido no tidyverse
  - Contém funções úteis para a manipulação/preparação de bancos de dados
  - Materiais para referência:
    - https://dplyr.tidyverse.org/
    - https://github.com/rstudio/cheatsheets/blob/master/data-transformation.pdf
    - Wickham, H. & Grolemund, G. **R for Data Science**: https://r4ds.had.co.nz/index.html

Data wrangling

- Funções frequentemente utilizadas no software R
  - **Pipe**: encadeamento de diversas funções em sequência
  - **Rename**: alteração de nomes de variáveis
  - **Mutate**: alteração de conteúdo das variáveis e criação de novas variáveis
  - **Filter**: seleção de observações com base em critérios lógicos
  - **Select**: seleção de variáveis
  - **Summarise**: criação de tabelas com medidas resumo (estatísticas descritivas)
  - **Group by**: agrupamento das observações com base em critérios
  - **Join**: junção (*merge*) de bancos de dados

Iterações com Pacote “purrr”

purrr (tidyverse)

- Pacote purrr: **funções map**
  - Funções que facilitam a aplicação de processos iterativos por meio de códigos que são mais simples de escrever e ler, além de serem sucintos
  - Para simplificar operações que precisam ser repetidas muitas vezes, como a aplicação de certa função em diversas variáveis do banco de dados
  - Materiais para consulta:
    - Wickham, H. & Grolemund, G. **R for Data Science**: https://r4ds.had.co.nz/index.html
    - https://github.com/rstudio/cheatsheets/blob/master/purrr.pdf

Criação de Projects e Scripts R Markdown

R Markdown

- **Introdução ao R Markdown**
- **Formatação básica do texto**
- **Inserção de fórmulas**
- **Chunks**
- **Gerando outputs (HTML; PDF, DOC)**

- Material para referência:
  - https://rmarkdown.rstudio.com/index.html

Referências

Wickham, H. & Grolemund, G. (2017) **R for Data Science.** O'Reilly

OBRIGADO!

linkedin.com/in/wilson-tarantin-junior-359476190/