

CIÊNCIA DE DADOS

INTRODUÇÃO

Universidade Federal de Santa Maria - UFSM
Prof. Moisés da Silva Melo

Introdução



-
- O R é uma linguagem de programação projetada para análise estatística.
 - O R é um software livre (gratuito) com código aberto.
 - A comunidade de usuários e desenvolvedores do R é bastante ativa.
 - O R é frequentemente integrado com outras linguagens de programação, como Python e SQL

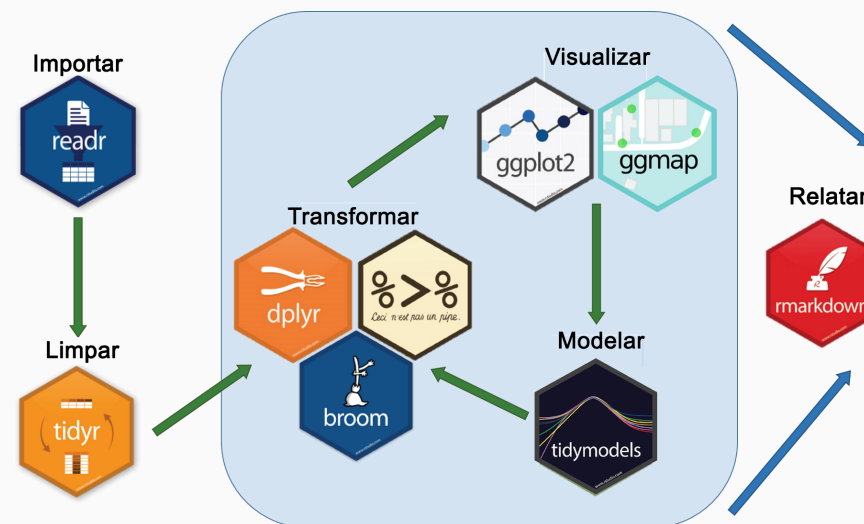


-
- O R Studio é um ambiente de desenvolvimento integrado (IDE) utilizado para executar códigos em R.
 - O R Studio tem uma interface de usuário mais amigável e simplificada em comparação ao R base.
 - O R Studio oferece um ambiente integrado que inclui painéis, janelas e ferramentas que facilitam o desenvolvimento, a análise de dados e a visualização.

Organização e manipulação

Pacote tidyverse

- O `tidyverse` é uma coleção abrangente de pacotes para R, desenvolvida por Hadley Wickham e outros colaboradores, com o propósito de fornecer um conjunto completo de ferramentas essenciais para um fluxo de trabalho eficaz em ciência de dados e análise estatística.
- O `tidyverse` oferece eficiência e coesão em todo o processo de análise de dados, desde a limpeza e transformação dos dados até a criação de gráficos informativos.



Pacote readr - Importar dados

- O pacote `{readr}` é utilizado para importar arquivos de texto, como `.txt`, `.csv` e outros formatos comuns. Ele fornece funções para ler esses tipos de arquivos de forma eficiente e conveniente.
- As principais funções são:
 - `read_csv()`: Para ler arquivos separados por vírgula.
 - `read_csv2()`: Para ler arquivos separados por ponto-e-vírgula.
 - `read_delim()`: Para ler arquivos separados por um delimitador genérico. O argumento `delim=` indica qual caractere separa cada coluna no arquivo de texto.

Tidy Data

- Tidy Data (dados organizados) é um conceito introduzido pelo estatístico e cientista de dados Hadley Wickham em seu artigo de 2014 intitulado **Tidy Data**.
- O formato de dados organizados é projetado para simplificar a manipulação, limpeza, transformação, análise e visualização de dados.
- Isso é especialmente útil ao utilizar ferramentas e pacotes de análise de dados que foram projetados com o conceito de "tidy data" em mente, como os pacotes do tidyverse no R
- As três propriedades mais importantes de uma base tidy são:
 - Cada variável é uma coluna
 - Cada observação é uma fila
 - Cada célula é uma única medição

each column a variable

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation












Pacote dplyr - Transformando dados

- O `dplyr` é um pacote extremamente útil para realizar transformações de dados em R.
- O `dplyr` fornece um conjunto de funções simples e intuitivas para realizar operações comuns de transformação de dados, como selecionar, filtrar, agregar, agrupar, ordenar, sumarizar, criar variáveis, dentre outras.
- Algumas das funções mais usadas do dplyr incluem:
 - `select()`: Usada para selecionar colunas específicas de um conjunto de dados.
 - `arrange()`: Usada para reordenar as linhas de um conjunto de dados com base em uma ou mais colunas.
 - `filter()`: Usada para filtrar linhas com base em condições específicas.
 - `rename()` Usada para para alterar os nomes de variáveis (colunas) em um conjunto de dados.
 - `mutate()`: Permite criar novas colunas ou modificar as existentes com base em cálculos ou operações.
 - `group_by()`: Utilizada para agrupar dados com base em uma ou mais variáveis.
 - `summarize()`: Permite resumir dados dentro de grupos criados com `group_by()`.

Pacote dplyr - Transformando dados

- Principais operadores lógicos (critérios) para usar com o pacote `dplyr`
- `:` Seleciona um intervalo consecutivos.
- `!` Seleciona o complementar.
- `&` Seleciona a interseção.
- `|` Seleciona a interseção.
- `is.na()` Seleciona os `NA`
- `!is.na()` Seleciona os não `NA`
- `>` Seleciona os valores maiores que
- `<` Seleciona os valores menores que
- `==` Seleciona os valores iguais a
- `<=` Seleciona os valores menores ou iguais a
- `>=` Seleciona os valores maiores ou iguais a
- `<=` Seleciona os valores diferentes de

Pacote dplyr - Transformando dados

PAÍS	 OURO	 PRATA	 BRONZE
1°  Estados Unidos	72	45	54
2°  Brasil +	37	47	39
3°  Canadá	35	32	38
4°  México	35	22	32
5°  Colômbia	14	20	14
6°  Cuba	11	6	7
7°  Peru	7	2	13
8°  Chile	6	19	13

Exercício

- Usando os microdados do Sistema de Informação sobre Nascidos Vivos (SINASC) referentes ao ano de 2022, realize as seguintes manipulações no R.
 1. Importe o banco de dados `sinasc_2022.csv`.
 2. Selecione as seguintes variáveis (colunas): `UF`, `IDADEMAE`, `SEXO`, `RACACOR`, `PESO`.
 3. Transforme as variáveis `SEXO` e `RACACOR` para a classe `factor`.
 4. Transforme a variável `PESO` para a classe `numeric`.
 5. Substitua os códigos usados para as variáveis `SEXO` e `RACACOR` pela descrição contida no arquivo `Dicionario_Sinasc.pdf`.
 6. Calcule a média de idade das mães dos nascidos em 2022 por `UF`. Em seguida, ordene as médias por ordem decrescente.
 7. Calcule a média de idade das mães dos nascidos em 2022 por `RACACOR`. Em seguida, ordene as médias por ordem decrescente.
 8. Calcule o percentual de nascidos vivos com baixo peso ao nascer (<2500g) por `SEXO`.