

Lista 1

2025-11-03

Pacotes necessários

Para essa lista de exercícios serão necessários os pacotes *data.table* e *dplyr*, para leitura e manipulação das bases de dados, respectivamente.

```
# instale os pacotes com o comando abaixo
# install.packages("tidyverse") # obs: o pacote dplyr faz parte do pacote tidyverse
# install.packages("data.table")

# após a instalação, importe os pacotes
library(tidyverse)
library(data.table)
```

Exercício 1 : Leitura dos dados

A função **fread** da biblioteca *data.table* é uma versão otimizada das funções **read.csv** e **read.csv2** para grandes bases de dados.

a) Leia novamente as bases de dados da CGIL de 2024 e 2025 utilizando a função **fread**. (Guarde cada uma das bases em uma variável)

Dica : Consulte o **help(fread)** para mais informações sobre a leitura.

b) Junte as duas bases de dados em uma só variável.

Exercício 2 : Limpeza e manipulação dos dados

O pacote **dplyr** é focado em manipulação de bases de dados. A pesar de ter um grande número de funções, a maioria delas tem nomes intuitivos, por exemplo:

- **filter** : filtra as linhas de acordo com as condições dadas;
- **select** : seleciona as colunas indicadas;
- **group_by** : agrupa os dados de acordo com as colunas escolhidas;
- **count** : conta a quantidade de elementos na coluna;
- **summarise** : agrupa os dados das colunas indicadas de acordo com as medidas escolhidas.

Esses comandos podem ser encadeados para recrear as tabelas vistas ao longo do curso, como por exemplo os dados filtrados

```
dados24 <- read.csv2("CGIL_CNIg_2024.csv", fileEncoding = "UTF-8")
dados25 <- read.csv2("CGIL_CNIg_jan-set2025.csv", fileEncoding = "UTF-8")
dados <- rbind(dados24, dados25)

dados_final <- dados %>%
  filter(modalidade == "CGIL", andamento == "DEFERIDO") %>%
  filter((ano == 2024 & mes == 9) | (ano == 2025 & mes %in% c(8,9)))
```

Ou, para replicar a tabela de frequênciapor gênero

```
dados_final %>%
  select(ano, mes, genero) %>%
  group_by(ano, mes, genero) %>%
  summarise(n = n(), .groups = "drop") %>%
  pivot_wider(names_from = c(mes, ano),
  names_prefix = "d_",
  values_from = n)

## # A tibble: 2 x 4
##   genero d_9_2024 d_8_2025 d_9_2025
##   <chr>     <int>     <int>     <int>
## 1 F          291      312      588
## 2 M         2298     3703     4084
```

- a) Utilize as funções do pacote dyplr para recriar a tabela de frquênciade paises por mes e ano.
- b) A partir da tabela encontrada na letra a) crie uma nova coluna com o total de frequencias ao longo dos anos.
- c) Agora selecione os 10 paises de maior frequênciatotal. (Dica: help(arange))