Atividade 1

Mikael

11 de janeiro de 2023

Índice

'execute:  
 warning: false  
cache: TRUE  
theme: cerulean  
reference-location: margin  
code-fold: true  
format:  
 html:  
 output-file: Atividade1.html  
 docx:   
 output-file: Atividade1.docx  
 toc: true  
 number-sections: true  
 highlight-style: github  
params:  
 run\_all\_code: "T"  
 overwrite\_tables: "T"  
 use\_db: "T"  
 debug: "F"'

[1] "execute:\n warning: false\ncache: TRUE\ntheme: cerulean\nreference-location: margin\ncode-fold: true\nformat:\n html:\n output-file: Atividade1.html\n docx: \n output-file: Atividade1.docx\n toc: true\n number-sections: true\n highlight-style: github\nparams:\n run\_all\_code: \"T\"\n overwrite\_tables: \"T\"\n use\_db: \"T\"\n debug: \"F\""

# 1. Aula de introdução à ciência de dados

## 1.1 Atividade 1

### 1.1.1 1, 2 e 3) Lendo base de dados e selecionando variáveis

library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

## Lendo dados ----  
df\_saeb <- data.table::fread(here::here("1-Semestre/CienciaDeDados-STC878/Dados/saeb\_2021\_ts\_aluno\_9ef.csv"))  
  
df\_saeb <- df\_saeb |>   
 dplyr::select(ID\_REGIAO, ID\_UF, IN\_PUBLICA, PROFICIENCIA\_LP, PROFICIENCIA\_MT, NU\_TIPO\_NIVEL\_INSE,  
 SEXO = TX\_RESP\_Q01, RACACOR = TX\_RESP\_Q04) |>   
 dplyr::mutate(dplyr::across(everything(), as.character))  
  
df\_saeb\_raw <- df\_saeb

### 1.1.2 4) Alterando categorias das variáveis região e indicador de escola pública

df\_saeb <- df\_saeb |>   
 dplyr::mutate(  
 ID\_REGIAO = case\_when(  
 ID\_REGIAO == "1" ~ "Norte",  
 ID\_REGIAO == "2" ~ "Nordeste",  
 ID\_REGIAO == "3" ~ "Sudeste",  
 ID\_REGIAO == "4" ~ "Sul",  
 ID\_REGIAO == "5" ~ "Centro-Oeste",  
 .default = "Sem Região"  
 ),  
 ID\_UF = case\_when(  
 ID\_UF == "11" ~ "RO",  
 ID\_UF == "12" ~ "AC",  
 ID\_UF == "13" ~ "AM",  
 ID\_UF == "14" ~ "RR",  
 ID\_UF == "15" ~ "PA",  
 ID\_UF == "16" ~ "AP",  
 ID\_UF == "17" ~ "TO",  
 ID\_UF == "21" ~ "MA",  
 ID\_UF == "22" ~ "PI",  
 ID\_UF == "23" ~ "CE",  
 ID\_UF == "24" ~ "RN",  
 ID\_UF == "25" ~ "PB",  
 ID\_UF == "26" ~ "PE",  
 ID\_UF == "27" ~ "AL",  
 ID\_UF == "28" ~ "SE",  
 ID\_UF == "29" ~ "BA",  
 ID\_UF == "31" ~ "G",  
 ID\_UF == "32" ~ "ES",  
 ID\_UF == "33" ~ "RJ",  
 ID\_UF == "35" ~ "SP",  
 ID\_UF == "41" ~ "PR",  
 ID\_UF == "42" ~ "SC",  
 ID\_UF == "43" ~ "RS",  
 ID\_UF == "50" ~ "MS",  
 ID\_UF == "51" ~ "MT",  
 ID\_UF == "52" ~ "GO",  
 ID\_UF == "53" ~ "DF",  
 .default = "Sem Estado"  
 )  
 )

### 1.1.3 5) Média e desvio padrão das proeficiências em Língua Portuguesa e Matemática, por UF:

library(gt)  
  
resultado\_regiao <- df\_saeb |>   
 dplyr::group\_by(ID\_UF) |>   
 dplyr::mutate(  
 media\_lp\_por\_UF = mean(as.numeric(PROFICIENCIA\_LP), na.rm = T),  
 media\_mt\_por\_UF = mean(as.numeric(PROFICIENCIA\_MT), na.rm = T),  
 dp\_lp\_por\_UF = sd(as.numeric(PROFICIENCIA\_LP), na.rm = T),  
 dp\_mt\_por\_UF = sd(as.numeric(PROFICIENCIA\_MT), na.rm = T)  
 ) |>   
 dplyr::distinct(ID\_UF, .keep\_all = T) |>   
 dplyr::select(ID\_UF, media\_lp\_por\_UF, dp\_lp\_por\_UF, media\_mt\_por\_UF, dp\_mt\_por\_UF) |>   
 dplyr::ungroup()  
  
resultado\_regiao |>  
 gt::gt() |>  
 tab\_header(title = "Média e desvio padrão das proeficiências em Língua Portuguesa e Matemática") |>  
 cols\_label(  
 ID\_UF = "UF",  
 media\_lp\_por\_UF = "Média de LP",  
 dp\_lp\_por\_UF = "DP de LP",  
 media\_mt\_por\_UF = "Média de MT",  
 dp\_mt\_por\_UF = "DP de MT") |>  
 tab\_source\_note(source\_note = "Fonte: SAEB") |>  
 tab\_style(  
 style = list(cell\_fill(color = "#F4F4F4")),  
 locations = cells\_body(columns = ID\_UF)  
 ) |>  
 fmt\_number(  
 columns = c(media\_lp\_por\_UF, media\_mt\_por\_UF, dp\_lp\_por\_UF, dp\_mt\_por\_UF),  
 decimals = 2,  
 use\_seps = FALSE  
 )

Table 1: Média e desvio padrão das proeficiências em Língua Portuguesa e Matemática

| UF | Média de LP | DP de LP | Média de MT | DP de MT |
| --- | --- | --- | --- | --- |
| RO | 0.04 | 0.86 | 0.01 | 0.81 |
| AC | 0.05 | 0.87 | -0.02 | 0.82 |
| AM | -0.08 | 0.89 | -0.15 | 0.86 |
| RR | -0.09 | 0.91 | -0.15 | 0.88 |
| PA | -0.20 | 0.83 | -0.29 | 0.78 |
| AP | -0.12 | 0.84 | -0.28 | 0.77 |
| TO | -0.03 | 0.89 | -0.02 | 0.86 |
| MA | -0.32 | 0.86 | -0.41 | 0.83 |
| PI | -0.03 | 0.89 | -0.06 | 0.87 |
| CE | 0.21 | 0.91 | 0.16 | 0.94 |
| RN | -0.08 | 0.87 | -0.15 | 0.82 |
| PB | -0.11 | 0.87 | -0.17 | 0.83 |
| PE | -0.03 | 0.88 | -0.05 | 0.86 |
| AL | -0.11 | 0.91 | -0.09 | 0.93 |
| SE | -0.20 | 0.87 | -0.21 | 0.81 |
| BA | -0.16 | 0.86 | -0.22 | 0.80 |
| G | 0.16 | 0.91 | 0.12 | 0.87 |
| ES | 0.08 | 0.94 | 0.13 | 0.88 |
| RJ | 0.03 | 0.88 | -0.05 | 0.81 |
| SP | 0.24 | 0.89 | 0.19 | 0.85 |
| PR | 0.16 | 0.89 | 0.20 | 0.84 |
| SC | 0.28 | 0.86 | 0.30 | 0.81 |
| RS | 0.22 | 0.87 | 0.18 | 0.82 |
| MS | 0.08 | 0.88 | 0.07 | 0.83 |
| MT | 0.00 | 0.90 | -0.03 | 0.86 |
| GO | 0.19 | 0.88 | 0.14 | 0.86 |
| DF | 0.15 | 0.90 | 0.12 | 0.84 |
| Fonte: SAEB | | | | |

# data.table::fwrite(resultado\_regiao, here::here("1-Semestre/CienciaDeDados-STC878/R/Atividade1/resultado\_regiao.csv"))  
  
# resultado\_regiao\_gt |> gtsave(filename = here::here("1-Semestre/CienciaDeDados-STC878/R/Atividade1/resultado\_regiao.html"))  
  
library(DT)  
  
## Preparando tabela para impressão  
  
resultado\_regiao\_dt <- resultado\_regiao |>   
 dplyr::mutate(media\_lp\_por\_UF = round(media\_lp\_por\_UF, 2), dp\_lp\_por\_UF = round(dp\_lp\_por\_UF, 2),  
 media\_mt\_por\_UF = round(media\_mt\_por\_UF, 2), dp\_mt\_por\_UF = round(dp\_mt\_por\_UF, 2)) |>   
 dplyr::rename(UF = ID\_UF, "Média de LP" = media\_lp\_por\_UF, "DP de LP" = dp\_lp\_por\_UF,   
 "Média de MT" = media\_mt\_por\_UF, "DP de MT" = dp\_mt\_por\_UF) |>   
 datatable(options = list(  
 order = list(list(2, 'asc'), list(4, 'asc'))  
))  
  
resultado\_regiao\_dt |> htmlwidgets::saveWidget(here::here("1-Semestre/CienciaDeDados-STC878/R/Atividade1/resultado\_regiao.html"))  
  
# resultado\_regiao\_dt

#### 1.1.3.1 *Resultado*

Existem evidências de disparidade entre os estados, com o resultado mais baixo sendo -0.32 de nota média em português para o estado de Maranhão, e 0.27 de nota média de português no estado de Santa Catarina, para matemática temos um resultado parecido mantendo os mesmos estados com notas baixas e altas, apenas variando um pouco em valores. Porém, vemos que os desvios padrões são bem altos, ou seja, temos uma variação bem grande nestes dados.

### 1.1.4 6) Calculando as médias e desvios padrões das proeficiências em Língua Portuguêsa e Matemática por sexo e tipo de escola (pública ou particular)

# library(gt)  
  
resultado\_sexo\_tipoEscola <- df\_saeb |>   
 dplyr::filter(SEXO != "." & SEXO != "\*") |>   
 dplyr::group\_by(SEXO, IN\_PUBLICA) |>   
 dplyr::mutate(  
 media\_lp\_por\_sex\_esc = mean(as.numeric(PROFICIENCIA\_LP), na.rm = T),  
 media\_mt\_por\_sex\_esc = mean(as.numeric(PROFICIENCIA\_MT), na.rm = T),  
 dp\_lp\_por\_sex\_esc = sd(as.numeric(PROFICIENCIA\_LP), na.rm = T),  
 dp\_mt\_por\_sex\_esc = sd(as.numeric(PROFICIENCIA\_MT), na.rm = T)  
 ) |>   
 dplyr::mutate(SEXO = if\_else(SEXO == "A", "Masculino", if\_else(SEXO == "B", "Feminino", "Branco ou nulo")),  
 IN\_PUBLICA = if\_else(IN\_PUBLICA == "0", "Privada", if\_else(IN\_PUBLICA == "1", "Pública", "Inválido"))) |>   
 dplyr::distinct(SEXO, IN\_PUBLICA, .keep\_all = T) |>   
 dplyr::select(SEXO, IN\_PUBLICA, media\_lp\_por\_sex\_esc, dp\_lp\_por\_sex\_esc, media\_mt\_por\_sex\_esc, dp\_mt\_por\_sex\_esc) |>   
 dplyr::arrange(SEXO, desc(IN\_PUBLICA)) |>   
 dplyr::ungroup()  
  
library(gt)  
  
resultado\_sexo\_tipoEscola |>   
 gt::gt() |>   
 tab\_header(title = "Média e desvio padrão das proeficiências em Língua Portuguesa e Matemática por Sexo e Raça/Cor") |>   
 cols\_label(  
 SEXO = "Sexo",  
 IN\_PUBLICA = "Escola pública ou particular",  
 media\_lp\_por\_sex\_esc = "Média de LP",  
 dp\_lp\_por\_sex\_esc = "DP de LP",  
 media\_mt\_por\_sex\_esc = "Média de MT",  
 dp\_mt\_por\_sex\_esc = "DP de MT") |>   
 tab\_source\_note(source\_note = "Fonte: SAEB") |>   
 tab\_style(  
 style = list(cell\_fill(color = "#F4F4F4")),  
 locations = cells\_body(columns = c(SEXO, IN\_PUBLICA))  
 ) |>  
 fmt\_number(  
 columns = c(media\_lp\_por\_sex\_esc, media\_mt\_por\_sex\_esc, dp\_lp\_por\_sex\_esc, dp\_mt\_por\_sex\_esc),  
 decimals = 2,  
 use\_seps = FALSE  
 )

Table 1: Média e desvio padrão das proeficiências em Língua Portuguesa e Matemática por Sexo e Raça/Cor

| Sexo | Escola pública ou particular | Média de LP | DP de LP | Média de MT | DP de MT |
| --- | --- | --- | --- | --- | --- |
| Feminino | Pública | 0.19 | 0.86 | -0.04 | 0.84 |
| Feminino | Privada | 0.81 | 0.77 | 0.66 | 0.83 |
| Masculino | Pública | -0.05 | 0.92 | 0.10 | 0.88 |
| Masculino | Privada | 0.57 | 0.91 | 0.79 | 0.88 |
| Fonte: SAEB | | | | | |

## Preparando tabela para impressão  
  
resultado\_sexo\_tipoEscola\_dt <- resultado\_sexo\_tipoEscola |>   
 dplyr::mutate(media\_lp\_por\_sex\_esc = round(media\_lp\_por\_sex\_esc, 2), dp\_lp\_por\_sex\_esc = round(dp\_lp\_por\_sex\_esc, 2),  
 media\_mt\_por\_sex\_esc = round(media\_mt\_por\_sex\_esc, 2), dp\_mt\_por\_sex\_esc = round(dp\_mt\_por\_sex\_esc, 2)) |>   
 dplyr::rename(Sexo = SEXO, Escola = IN\_PUBLICA,   
 "Média de LP" = media\_lp\_por\_sex\_esc, "DP de LP" = dp\_lp\_por\_sex\_esc,   
 "Média de MT" = media\_mt\_por\_sex\_esc, "DP de MT" = dp\_mt\_por\_sex\_esc) |>   
 datatable(options = list(  
 order = list(list(1, 'asc'), list(2, 'desc'))  
))   
  
resultado\_sexo\_tipoEscola\_dt |> htmlwidgets::saveWidget(here::here("1-Semestre/CienciaDeDados-STC878/R/Atividade1/resultado\_sexo\_tipoEscola.html"))  
  
# resultado\_sexo\_tipoEscola\_dt

#### 1.1.4.1 *Resultado*

Podemos observar uma pequena variáção entre os sexos Masculino e Feminino, porém a diferença é bem mais expressiva se compararmos os tipos de escola, entre públicas e privadas, em ambos os sexos.

### 1.1.5 7) Calculando as mediana, média e desvios padrões na proeficiência dos alunos em Língua Portuguêsa e Matemática por categoria de nível social

# library(gt)  
  
resultado\_nivel <- df\_saeb |>   
 dplyr::filter(!is.na(NU\_TIPO\_NIVEL\_INSE)) |>   
 dplyr::group\_by(NU\_TIPO\_NIVEL\_INSE) |>   
 dplyr::mutate(  
 media\_lp\_por\_nivel = mean(as.numeric(PROFICIENCIA\_LP), na.rm = T),  
 media\_mt\_por\_nivel = mean(as.numeric(PROFICIENCIA\_MT), na.rm = T),  
 dp\_lp\_por\_nivel = sd(as.numeric(PROFICIENCIA\_LP), na.rm = T),  
 dp\_mt\_por\_nivel = sd(as.numeric(PROFICIENCIA\_MT), na.rm = T),  
 mediana\_lp\_por\_nivel = mean(as.numeric(PROFICIENCIA\_LP), na.rm = T),  
 mediana\_mt\_por\_nivel = mean(as.numeric(PROFICIENCIA\_MT), na.rm = T)  
 ) |>   
 dplyr::distinct(NU\_TIPO\_NIVEL\_INSE, .keep\_all = T) |>   
 dplyr::select(NU\_TIPO\_NIVEL\_INSE, media\_lp\_por\_nivel, dp\_lp\_por\_nivel, mediana\_lp\_por\_nivel,   
 media\_mt\_por\_nivel, dp\_mt\_por\_nivel, mediana\_mt\_por\_nivel) |>   
 dplyr::arrange(NU\_TIPO\_NIVEL\_INSE) |>   
 dplyr::ungroup()  
  
library(gt)  
  
resultado\_nivel |>   
 gt::gt() |>   
 tab\_header(title = "Média e desvio padrão das proeficiências em Língua Portuguesa e Matemática por Sexo e Raça/Cor") |>   
 cols\_label(  
 NU\_TIPO\_NIVEL\_INSE = "Nível Socioeconômico",  
 media\_lp\_por\_nivel = "Média de LP",  
 dp\_lp\_por\_nivel = "DP de LP",  
 mediana\_lp\_por\_nivel = "Mediana de LP",  
 media\_mt\_por\_nivel = "Média de MT",  
 dp\_mt\_por\_nivel = "DP de MT",  
 mediana\_mt\_por\_nivel = "Mediana de MT") |>   
 tab\_source\_note(source\_note = "Fonte: SAEB") |>   
 tab\_style(  
 style = list(cell\_fill(color = "#F4F4F4")),  
 locations = cells\_body(columns = c(NU\_TIPO\_NIVEL\_INSE))  
 ) |>  
 fmt\_number(  
 columns = c(media\_lp\_por\_nivel, media\_mt\_por\_nivel, dp\_lp\_por\_nivel, dp\_mt\_por\_nivel,  
 mediana\_lp\_por\_nivel, mediana\_mt\_por\_nivel),  
 decimals = 2,  
 use\_seps = FALSE  
 )

Table 1: Média e desvio padrão das proeficiências em Língua Portuguesa e Matemática por Sexo e Raça/Cor

| Nível Socioeconômico | Média de LP | DP de LP | Mediana de LP | Média de MT | DP de MT | Mediana de MT |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | -0.46 | 0.83 | -0.46 | -0.48 | 0.84 | -0.48 |
| 2 | -0.22 | 0.85 | -0.22 | -0.28 | 0.83 | -0.28 |
| 3 | -0.05 | 0.86 | -0.05 | -0.13 | 0.82 | -0.13 |
| 4 | 0.06 | 0.87 | 0.06 | 0.00 | 0.83 | 0.00 |
| 5 | 0.17 | 0.87 | 0.17 | 0.13 | 0.83 | 0.13 |
| 6 | 0.28 | 0.88 | 0.28 | 0.26 | 0.83 | 0.26 |
| 7 | 0.35 | 0.90 | 0.35 | 0.38 | 0.85 | 0.38 |
| 8 | 0.36 | 0.96 | 0.36 | 0.45 | 0.94 | 0.45 |
| Fonte: SAEB | | | | | | |

## Preparando tabela para impressão  
  
resultado\_nivel\_dt <- resultado\_nivel |>   
 dplyr::mutate(  
 NU\_TIPO\_NIVEL\_INSE = case\_when(  
 NU\_TIPO\_NIVEL\_INSE == "1" ~ "Nível 1",  
 NU\_TIPO\_NIVEL\_INSE == "2" ~ "Nível 2",  
 NU\_TIPO\_NIVEL\_INSE == "3" ~ "Nível 3",  
 NU\_TIPO\_NIVEL\_INSE == "4" ~ "Nível 4",  
 NU\_TIPO\_NIVEL\_INSE == "5" ~ "Nível 5",  
 NU\_TIPO\_NIVEL\_INSE == "6" ~ "Nível 6",  
 NU\_TIPO\_NIVEL\_INSE == "7" ~ "Nível 7",  
 NU\_TIPO\_NIVEL\_INSE == "8" ~ "Nível 8",  
 .default = "Sem nível"  
 )) |>   
 dplyr::mutate(media\_lp\_por\_nivel = round(media\_lp\_por\_nivel, 2), dp\_lp\_por\_nivel = round(dp\_lp\_por\_nivel, 2),  
 media\_mt\_por\_nivel = round(media\_mt\_por\_nivel, 2), dp\_mt\_por\_nivel = round(dp\_mt\_por\_nivel, 2),  
 mediana\_lp\_por\_nivel = round(mediana\_lp\_por\_nivel, 2), mediana\_mt\_por\_nivel = round(mediana\_mt\_por\_nivel, 2)) |>   
 dplyr::rename("Nível Socioeconômico" = NU\_TIPO\_NIVEL\_INSE,  
 "Média de LP" = media\_lp\_por\_nivel, "DP de LP" = dp\_lp\_por\_nivel,   
 "Média de MT" = media\_mt\_por\_nivel, "DP de MT" = dp\_mt\_por\_nivel,   
 "Mediana de LP" = mediana\_lp\_por\_nivel, "Mediana de MT" = mediana\_mt\_por\_nivel) |>   
 datatable(options = list(  
 order = list(list(1, 'asc'))  
))  
  
resultado\_nivel\_dt |> htmlwidgets::saveWidget(here::here("1-Semestre/CienciaDeDados-STC878/R/Atividade1/resultado\_nivel.html"))  
  
# resultado\_nivel\_dt

#### 1.1.5.1 *Resultado*

A evolução conforme o nível econômico também é bem clara, mostrando como os níveis mais baixos possuem os menores valores e como a média só cresce conforme o nível socioeconômico aumenta.

### 1.1.6 8) Salvando resultado

data.table::fwrite(resultado\_regiao, here::here("1-Semestre/CienciaDeDados-STC878/R/Atividade1/resultado\_regiao.csv"))  
data.table::fwrite(resultado\_sexo\_tipoEscola, here::here("1-Semestre/CienciaDeDados-STC878/R/Atividade1/resultado\_sexo\_tipoEscola.csv"))  
data.table::fwrite(resultado\_nivel, here::here("1-Semestre/CienciaDeDados-STC878/R/Atividade1/resultado\_nivel.csv"))