Metodologias Informacionais com R

Modulo IV: PNAD Contínua do IBGE

Telmo dos Santos Klipp (telmo.klipp@inpe.br)

Informações Gerais sobre o Curso

- Materiais disponibilizados via Classroom;
- O aprendizado requer a prática, que será constante nas aulas;

Bibliografia Básica:

- Kennedy, R., & Waggoner, P. D. (2021). Introduction to r for social scientists: a tidy programming approach. CRC Press.

Bibliografia Complementar:

- Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). R for data science (2e): import, tidy, transform, visualize, and model data. "O'Reilly Media, Inc.". Disponível em: https://r4ds.hadley.nz/. Acesso em: 14 de junho, 2023. (Online)
- Damiani, A. et. al., (2022). Ciência de Dados em R. Curso-R. Disponível em: https://livro.curso-r.com. Acesso em: 12 de maio, 2023. (Online)
- de Aquino, J. A. (2014). R para cientistas sociais. Editora da UESC (editus). Disponível em: http://www.uesc.br/editora/. Acesso em: 12 de maio, 2023.
- de Oliveira, P. F., Guerra, S., McDonnell, R. (2018). Ciência de Dados com R: Introdução. Editora IBPAD. Disponível em: https://cdr.ibpad.com.br/index.html. Acesso em: 12 de maio, 2023. (Online)

Microdados da PNAD Contínua

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) visa:

> *"Acompanhar as flutuações trimestrais e a evolução, no curto, médio e longo prazos, da força de trabalho, e outras informações necessárias para o estudo do desenvolvimento socioeconômico do País."* (IBGE)

Algumas características da PNAD Contínua são:

- As pesquisas possuem periodicidade mensal (dada por trimestre móvel), trimestral, anual e variável (feita com maior periodicidade ou ocasionalmente).
- As pesquisas apresentam uma amostragem complexa, seguindo um **plano amostral** (ou desenho amostral) cuja seguinte nota do IBGE detalha informações - ex: extratificação, seleção e rotação de amostras, seleção de domicílios, cálculos de pesos de amostras e domicílios, estimadores e precisão.

Microdados da PNAD Contínua

Para trabalhar com os microdados da PNAD Contínua, vamos fazer uso dos pacotes:

- **PNADcIBGE** - possibilita baixar dados da PNAD Contínua.
- **survey** - possibilita produzir diversas operações estatísticas em dados com amostragem complexa.
- **srvyr** - uma extensão do pacote **survey**, que permite o uso de funções do **dplyr** e sua sintaxe.

É necessário a instalação de:

```
install.packages(c("PNADcIBGE", "survey", "srvyr", "tidyverse"))
# Outros pacotes que serão usados
install.packages(c("kableExtra", "patchwork"))
```

E carregar alguns desses pacotes:

```
library(PNADcIBGE)
library(survey)
library(tidyverse)
library(kableExtra)
library(patchwork)
```

Microdados da PNAD Contínua 📊 – pacote **PNADcIBGE**

Permite obter microdados da PNAD Contínua e aplicar o plano amostral para realização de análises. A importação ou carregamento dos microdados pode ser feita de duas formas:

- **Online** – usando a função PNADcIBGE::get_pnadc()
    - Observe **?get_pnadc**
    - Na requisição dos microdados, são baixados de um servidor **FTP** do IBGE o arquivo com os microdados e os arquivos com a documentação necessária ao processamento e leitura desses dados (*input*, dicionários, variáveis de deflação).
- **Offline** – (dados previamente baixados) – usando a função PNADcIBGE::read_pnadc()
    - Observe **?read_pnadc**
    - Essa função permite ler diretamente o arquivo contendo os microdados, dado um arquivo de *input* (um programa de leitura na linguagem SAS) fornecido pelo IBGE.

Microdados da PNAD Contínua 📊 - pacote **PNADcIBGE**

Existem divisões nos microdados da PNAD Contínua, sendo possível baixá-los de quatro formas básicas:

- Dados trimestrais usando o parâmetro **quarter** que permite baixar um trimestre (1 à 4) por vez.

```
dados_2022q4 <- get_pnadc(year = 2022, quarter = 4)
```

- Dados anuais acumulados em determinada visita (entre a 1ª e 5ª por vez) usando o parâmetro **interview**.

```
dados_2022_interview_1 <- get_pnadc(year = 2022, interview = 1)
```

- Dados anuais concentrados em determinado trimestre (entre o 1º e 4º por vez) usando o parâmetro **topic**.

```
dados_2022_anual_trimestre <- get_pnadc(year = 2022, topic = 2)
```

- Dados de temas e tópicos suplementares (questionários específicos) aplicados a morador selecionado.

```
# Dados do módulo de Sensação de Segurança aplicado ao morador selecionado no 4º trimestre.
dados_2022_anual_trimestre <- get_pnadc(year = 2022, topic = 4, selected = TRUE)
```

Microdados da PNAD Contínua 📊 - pacote PNADcIBGE

Outros argumentos da função get_pnadc() são:

- **vars**: Permite carregar apenas variáveis específicas, mantendo-se aquelas relacionadas ao plano amostral.

```
# Obtendo as  variáveis:
#  VD4001 - Condição em relação à força de trabalho; VD4002 - Condição de ocupação
dados_2022q4 <- get_pnadc(year = 2022, quarter = 4, vars = c("VD4001","VD4002"))
```

- **labels** (lógico TRUE ou FALSE): substituir ou não os códigos das variáveis categóricas por texto descritivo.
- **deflator** (lógico TRUE ou FALSE): incluir ou não as variáveis que permitem aplicar deflacionamento.
- **design** (lógico TRUE ou FALSE): aplica ou não o plano amostral (**`survey.design`** ou **`svyrep.design`**).
- **defyear** e **defperiod**: Ano ou trimestre considerados nas variáveis de deflacionamento, respectivamente. Se **defyear** ou **defperiod** forem indicados como NULL e o **deflator** como TRUE, será considerado o ano mais recente, o ano ou o trimestre dos dados, conforme a alternativa de requisição de dados usada.
- **savedir**: Diretório para salvar os microdados e a documentação necessária ao processamento e leitura.

Microdados da PNAD Contínua 📊 - pacote **PNADcIBGE**

Vamos trabalhar com microdados do primeiro trimestre de 2023. Para baixar os dados, considere:

```
pnadc_2023q1 <- get_pnadc(year = 2023, quarter = 1, savedir = "PNADC20231")
```

Como por padrão "design = TRUE", o objeto retornado será do tipo **svyrep.design**, pois o plano amostral gerado considera os pesos replicados por reamostragem do tipo *bootstrap* - para maiores informações, veja o artigo. Também incluirá os *labels* categóricos e as variáveis de deflacionamento.

Como se trata de um conjunto grande de dados, podemos escolher carregar apenas variáveis específicas como o ano (**Ano**), a unidade da federação (**UF**), a condição de ocupação (**VD4002**), entre outras. Além disso, caso se deseje gerar um plano amostral com objeto do tipo **survey.design**, é necessário manter nos dados determinadas variáveis (ex: pesos do domicílio e das pessoas, domínios e índices de projeção), indicar "design = FALSE" e aplicar o plano amostral depois. Veja o exemplo:

```
pnadc_2023q1 <- get_pnadc(year = 2023, quarter = 1, design = FALSE,
                          vars = c("Ano","Trimestre","UF","Capital","Estrato","UPA","posest",
                                   "V1027","V1028","V1029", "2003","2007","V2009","VD3004",
                                   "VD3005","VD4002", "VD4020", "VD4035","ID_DOMICILIO",
                                   "Habitual","Efetivo"), savedir = "PNADC20231")
```

Microdados da PNAD Contínua - pacote PNADcIBGE

Conforme mencionado, caso os dados dados da PNAD Contínua já tenham sido baixados, pode-se carregá-los (*offline*) com a função read_pnadc(). Para isso, basta informar o arquivo dos microdados (ex: PNADC_012023.txt) e o arquivo de *input* (um programa de leitura na linguagem SAS).

```
pnadc_2023q1 <- read_pnadc(microdata = "PNADC20231/PNADC_012023.txt",
                           input_txt = "PNADC20231/input_PNADC_trimestral.txt",
                           vars = c("Ano", "Trimestre", "UF", "Capital", "Estrato", "UPA",
                                    "posest", "V1027", "V1028", "V1029",
                                    "V2003","V2009", "V2007", "VD3004", "VD3005",
                                    "VD4002", "VD4020", "VD4035"))
class(pnadc_2023q1)
```

Os dados serão retornados em um objeto do tipo **tibble**, porém, serão mantidas as variáveis relacionadas ao plano amostral, incluso os pesos replicados (de **V1028001** à **V1028200**) que foram gerados na reamostragem por *bootstrap*. A princípio, para considerar o plano amostral nas estatísticas, temos duas possibilidades:

- Construir um objeto do tipo **survey.design**, contendo as informações de estratificação e os pesos finais.
- Construir um objeto do tipo **svyrep.design**, contendo as informações de estratificação e os pesos replicados.

Microdados da PNAD Contínua 📊 - pacote **PNADcIBGE**

Para gerar o plano amostral com objeto do tipo **survey.design**, não precisamos dos pesos replicados (de **V1028001** à **V1028200**), que podem ser excluídos. Manteremos algumas variáveis relacionadas à estratificação, os pesos finais e as variáveis que desejamos trabalhar, conforme uma das opções:

```
pnadc_2023q1 <- select(pnadc_2023q1, -(V1028001:V1028200)) # Excluí pesos replicados, ou
```

```
pnadc_2023q1 <- pnadc_2023q1 |>
  select(Ano, Trimestre, UF, Capital, Estrato, UPA, posest, V1027, V1028, V1029,
         V2003, V2007, V2009, VD3004, VD3005, VD4002, VD4020, VD4035, ID_DOMICILIO)
```

Dentre as variáveis escolhidas, estão:

- Peso do domicílio e das pessoas, sem calibração (**V1027**) e com calibração (**V1028**).
- Projeção da população por níveis geográficos (**V1029**) e domínios de projeção geográficos (**posest**).
- Número de ordem (**V2003**).
- Identificação do domicílio (**ID_DOMICILIO**).

Para gerar o plano amostral com objeto do tipo **svyrep.design**, podemos ler os dados filtrando as variáveis de interesse, mas sem excluir os pesos replicados, conforme foi realizado nos *slides* anteriores.

Microdados da PNAD Contínua 📊 – pacote **PNADcIBGE**

Agora, podemos aplicar as demais operações aos dados para:

- Substituir os códigos das variáveis categóricas por textos descritivos (*labels*).
- Obter as variáveis de deflacionamento.
- Transformar o objeto do tipo **tibble** para **survey.design** ou **svyrep.design**, caso seja de interesse obter estatísticas considerando o plano amostral com o pacote **survey**.

Essas operações são possíveis através das funções pnadc_labeller(), pnadc_deflator() e pnadc_design(), sendo que ...

Microdados da PNAD Contínua 📊 - pacote PNADcIBGE

... procedemos na forma:

```
# Atribuindo labels as variáveis categóricas
pnadc_2023q1 <- pnadc_labeller(
  data_pnadc = pnadc_2023q1,
  dictionary.file = "PNADC20231/dicionario_PNADC_microdados_trimestral.xls"
)
# Carregando índices de deflação caso se deseje obter variáveis de rendimento
# em valor real.
pnadc_2023q1 <- pnadc_deflator(
  data_pnadc = pnadc_2023q1,
  deflator.file = "PNADC20231/deflator_PNADC_2023_trimestral_040506.xls"
)
# Criação do plano amostral para obter estatísticas coerentes com pacote 'survey'
pnadc_2023q1 <- pnadc_design(data_pnadc = pnadc_2023q1)
```

A função PNADcIBGE::pnadc_design() aplica o plano amostral ao usar internamente as funções survey::svydesign() ou survey::svrepdesign() - gerando objeto do tipo **survey.design** ou **svyrep.design** (caso os pesos replicados sejam mantidos). Agora, vejamos ...

Microdados da PNAD Contínua 📊 – pacote **PNADcIBGE**

... o tipo do nosso objeto:

```
class(pnadc_2023q1)
```

```
## [1] "survey.design2" "survey.design"
```

Esse objeto é uma lista cujos elementos são:

```
names(pnadc_2023q1)
```

```
##  [1] "cluster"    "strata"     "has.strata" "prob"       "allprob"
##  [6] "call"       "variables"  "fpc"        "pps"        "postStrata"
```

No item "variables", está o objeto **tibble** com os microdados. Os nomes das variáveis disponíveis podem ser obtidos com:

```
names(pnadc_2023q1$variables)
```

Microdados da PNAD Contínua 📊 - pacote **PNADcIBGE**

Os arquivos de microdados e auxiliares (input, dicionários, indicadores de deflação), usados para aplicar as operações anteriores são fornecidos pelo IBGE.

Esses arquivos podem ser obtidos no site ou em servidor **ftp** do IBGE (https://ftp.ibge.gov.br).

No entanto, é muito mais cômodo obtê-los fazendo uma requisição com a função get_pnadc(). Nesse caso, não é necessário procurar os dados desejados na hierarquia de pastas do referido site ou do servidor **ftp**.

Microdados da PNAD Contínua 📊 – pacote **PNADcIBGE**

Relembrando:

- A função get_pnadc() permite:
    - Baixar os dados de servidor **FTP**.
    - Inclui por padrão *labels* categóricos, variáveis de deflacionamento e plano amostral. A não inclusão requer o uso de parâmetros específicos da função.
- A função read_pnadc() permite:
    - Ler os microdados, requerendo arquivos específicos.
    - A inclusão de *labels* categóricos, variáveis de deflacionamento e plano amostral requer o uso de funções auxiliares.
- Os objetos **survey.design** (usa método do conglomerado primário) e **svyrep.design** (usa método *bootstrap*) não devem apresentar diferenças nos resultados para uma mesma estimativa (ex: média ou proporção de uma variável). No entanto, como possuem métodos diferentes para estimar as variâncias, os resultados para estimativas de erros como desvio padrão ou intervalo de confiança podem diferir.

Microdados da PNAD Contínua 📊 - pacote **PNADcIBGE**

O plano amostral é necessário para obter estatísticas coerentes e métricas sobre essas estatísticas como intervalos de confiança e margem de erro. Para maiores informações, veja as seguintes vídeo-aulas sobre microdados:

- Vídeo 1 - teórico
- Vídeo 2 - sobre PNADC
- Vídeo 3 - sobre Pesquisa de Orçamentos Familiares (POF)

O plano amostral da PNAD Contínua considera dois estágios de seleção com **estratificação** das unidades primárias de amostragem (UPAs). Não obstante, para produzir estatísticas coerentes sobre os dados, temos algumas opções:

- Usar o pacote **survey**, porém, seu objeto do tipo lista não permitirá operações usando o **dplyr**, devendo-se usar outras opções para manipulação de dados.
- Considerar algumas métricas do plano amostral como, por exemplo, indicar os pesos dos domicílios e das pessoas para aplicação de ponderação em funções pertinentes ao gerar as estatíticas desejadas.
- Gerar o objeto combinado dos pacotes **survey** e **srvyr** (possui o plano amostral), o que permitirá o uso de funções do **dplyr** e, consequente, flexibilidade nas manipulações de dados.

Microdados da PNAD Contínua 📊 - estatísticas com **survey**

O objeto do tipo **survey.design** permite produzir estatísticas através de funções do pacote **survey**. Por exemplo, quando se trata de uma variável categórica como **VD4002** (ocupados e desocupados), a função svymean() cálcula/estima a proporção da população em cada nível encontrado, com projeção geográfica conforme a consulta. Vejamos:

Qual a taxa de Ocupados/Desocupados em nível nacional?

```
svymean(~VD4002, pnadc_2023q1, na.rm = T)
```

```
##                              mean SE
## VD4002Pessoas ocupadas     0,9121  0
## VD4002Pessoas desocupadas 0,0879  0
```

Como gerar intervalos de confiança da proporção de Ocupados/Desocupados?

```
confint(svymean(~VD4002, pnadc_2023q1, na.rm = T), level = .95)
```

```
##                               2,5 %  97,5 %
## VD4002Pessoas ocupadas     0,9100 0,91410
## VD4002Pessoas desocupadas 0,0859 0,08997
```

Microdados da PNAD Contínua 📊 – estatísticas com survey

Qual a taxa de Ocupados/Desocupados no Rio Grande do Sul?

```
  svymean(~VD4002, subset(pnadc_2023q1, UF == 'Rio Grande do Sul'), na.rm = T)
```

```
##                                mean SE
## VD4002Pessoas ocupadas     0,9462  0
## VD4002Pessoas desocupadas  0,0538  0
```

Como gerar intervalos de confiança da proporção de Ocupados/Desocupados?

```
  confint(svymean(~VD4002,
                  subset(pnadc_2023q1, UF == 'Rio Grande do Sul'), na.rm = T), level = .95)
```

```
##                                2,5 %  97,5 %
## VD4002Pessoas ocupadas     0,94099 0,95147
## VD4002Pessoas desocupadas  0,04853 0,05901
```

Microdados da PNAD Contínua 📊 - estatísticas com survey

A função svytotal() cálcula/estima a parcela da população, em números totais, em cada nível encontrado, projetando a estimativa geograficamente conforme a consulta. Vejamos:

Quantas pessoas estão ocupadas ou desocupadas no Rio Grande do Sul?

```
svytotal(~VD4002, subset(pnadc_2023q1, UF == 'Rio Grande do Sul'), na.rm=T)
```

```
##                              total    SE
## VD4002Pessoas ocupadas     5925147 53805
## VD4002Pessoas desocupadas   336727 16375
```

Como gerar intervalos de confiança do total de Ocupados/Desocupados?

```
confint(svytotal(~VD4002,
                 subset(pnadc_2023q1, UF == 'Rio Grande do Sul'), na.rm=T), level = .95)
```

```
##                              2,5 %  97,5 %
## VD4002Pessoas ocupadas     5819690 6030603
## VD4002Pessoas desocupadas   304633  368821
```

Microdados da PNAD Contínua - estatísticas com **survey**

A variável **V2009** (numérica) contém a idade dos entrevistados, enquanto **VD3005** (categórica) indica a quantidade de anos de estudo. É possível extrair com gsub() os caracteres numéricos do texto de cada nível categórico, conforme a tabela abaixo. Perceba que, da categoria "**Sem instrução e menos de 1 ano de estudo**", extrai-se o texto '1', assim, este nível pode ser ignorado em estimativas ou atribuído o valor 0.

```
data.frame(VD3005 = pnadc_2023q1$variables$VD3005,
           qtd = gsub("\\D", "", pnadc_2023q1$variables$VD3005)) |> unique() |>
kable() |>
kable_styling(bootstrap_options = c('striped', 'condensed')) |>
scroll_box(width = "1050px", height = "240px")
```

	VD3005	qtd
1	12 anos de estudo	12
3	16 anos ou mais de estudo	16
5	NA	NA
10	9 anos de estudo	9
15	15 anos de estudo	15

Microdados da PNAD Contínua 📊 – estatísticas com **survey**

Quando trabalhamos com variáveis numéricas, a função svymean() cálcula/estima a média, ao invés da proporção. Portanto, podemos extrair os anos de estudo de **VD3005**, converter para "numeric", e estimar a média de anos de estudo da população com 25 anos ou mais de idade no Rio Grande do Sul, conforme:

```
  svymean(~as.numeric(gsub("\\D", "", VD3005)),
          subset(pnadc_2023q1, V2009 >= 25 & UF == 'Rio Grande do Sul'
                 & VD3005 != "Sem instrução e menos de 1 ano de estudo"), na.rm = T)
```

```
##                                        mean   SE
## as.numeric(gsub("\\\\D", "", VD3005)) 10,5 0,07
```

```
  confint(svymean(~as.numeric(gsub("\\D", "", VD3005)),
                  subset(pnadc_2023q1, V2009 >= 25 & UF == 'Rio Grande do Sul'
                         & VD3005 != "Sem instrução e menos de 1 ano de estudo"), na.rm = T),
          level = .95)
```

```
##                                        2,5 % 97,5 %
## as.numeric(gsub("\\\\D", "", VD3005)) 10,36  10,63
```

Microdados da PNAD Contínua 📊 – estatísticas com survey

Qual é a média de anos de estudos da população com 25 anos ou mais de idade na capital Porto Alegre?

```
  svymean(~as.numeric(gsub("\\D", "", VD3005)),
          subset(pnadc_2023q1, V2009 >= 25 & Capital == "Município de Porto Alegre (RS)"
                 & VD3005 != "Sem instrução e menos de 1 ano de estudo"), na.rm = T)
```

```
##                                        mean   SE
## as.numeric(gsub("\\\\D", "", VD3005)) 12,2 0,18
```

```
  confint(svymean(~as.numeric(gsub("\\D", "", VD3005)),
                  subset(pnadc_2023q1, V2009 >= 25 & Capital == "Município de Porto Alegre (RS)"
                         & VD3005 != "Sem instrução e menos de 1 ano de estudo"), na.rm = T),
          level = .95)
```

```
##                                        2,5 % 97,5 %
## as.numeric(gsub("\\\\D", "", VD3005)) 11,87  12,57
```

Microdados da PNAD Contínua 📊 – estatísticas com **survey**

A função svyby() permite estimar quantidades, considerando agrupamentos. No cálculo da estimativa, dispõe-se de outras funções do pacote **survey** (ex: parâmetro **FUN** = svytotal). Vejamos a proporção da população com 25 anos ou mais de idade em cada nível de instrução por sexo, conforme:

```
svyby(formula = ~VD3004, by = ~V2007, design = subset(pnadc_2023q1, V2009 >= 25),
      FUN = svymean, na.rm = T) |> kable() |>
kable_styling(bootstrap_options = c('striped', 'condensed')) |>
scroll_box(width = "1050px", height = "250px")
```

	V2007	VD3004Sem instrução e menos de 1 ano de estudo	VD3004Fundamental incompleto ou equivalente	VD3004Fundamental completo ou equivalente	VD3004Médio incompleto ou equivalente	VD3004Médio completo ou equivalente	V
Homem	Homem	0,0628	0,2846	0,0778	0,0532	0,3050	
Mulher	Mulher	0,0596	0,2612	0,0723	0,0477	0,3028	

Microdados da PNAD Contínua 📊 – estatísticas com **survey**

Vejamos o rendimento médio mensal das pessoas com 14 anos ou mais de idade por estado:

```
svyby(formula = ~VD4020, by = ~UF, design = pnadc_2023q1, FUN = svymean, na.rm = T) |>
 arrange(desc(VD4020)) |>
 kable() |>
 kable_styling(bootstrap_options = c('striped', 'condensed')) |>
 scroll_box(width = "1050px", height = "305px")
```

	UF	VD4020	se
Distrito Federal	Distrito Federal	5347	322,12
São Paulo	São Paulo	3788	106,83
Rio de Janeiro	Rio de Janeiro	3773	101,18
Mato Grosso do Sul	Mato Grosso do Sul	3493	145,19
Santa Catarina	Santa Catarina	3481	57,76
Rio Grande do Sul	Rio Grande do Sul	3454	79,64
Paraná	Paraná	3339	71,55

Microdados da PNAD Contínua 📊 – estatísticas com **survey**

A função svytable() permite produzir tabelas com valores totais de uma variável, considerando tabulamento cruzado (múltiplas variáveis). Vejamos as estimativas da população total por sexo em cada estado:

```
svytable(~UF + V2007, design = pnadc_2023q1) |>
 kable() |>
 kable_styling(bootstrap_options = c('striped', 'condensed')) |>
 scroll_box(width = "1050px", height = "304px")
```

	Homem	Mulher
Rondônia	913158	920632
Acre	454192	455050
Amazonas	2082833	2116128
Roraima	302445	297635
Pará	4464052	4407105
Amapá	453198	440894
Tocantins	797670	823003

Microdados da PNAD Contínua 📊 – estatísticas com **survey**

Vejamos as estimativas das pessoas ocupadas/desocupadas em cada estado:

```
svytable(~UF + VD4002, design = pnadc_2023q1) |>
 kable() |>
 kable_styling(bootstrap_options = c('striped', 'condensed')) |>
 scroll_box(width = "1050px", height = "330px")
```

	Pessoas ocupadas	Pessoas desocupadas
Rondônia	797385	26010
Acre	298492	32255
Amazonas	1703240	199990
Roraima	255397	18504
Pará	3696377	402641
Amapá	373963	51764
Tocantins	750756	55727

Microdados da PNAD Contínua 📊 – estatísticas com **survey**

Também é possível obter a proporção com auxilio da função prop.table() do **rbase**:

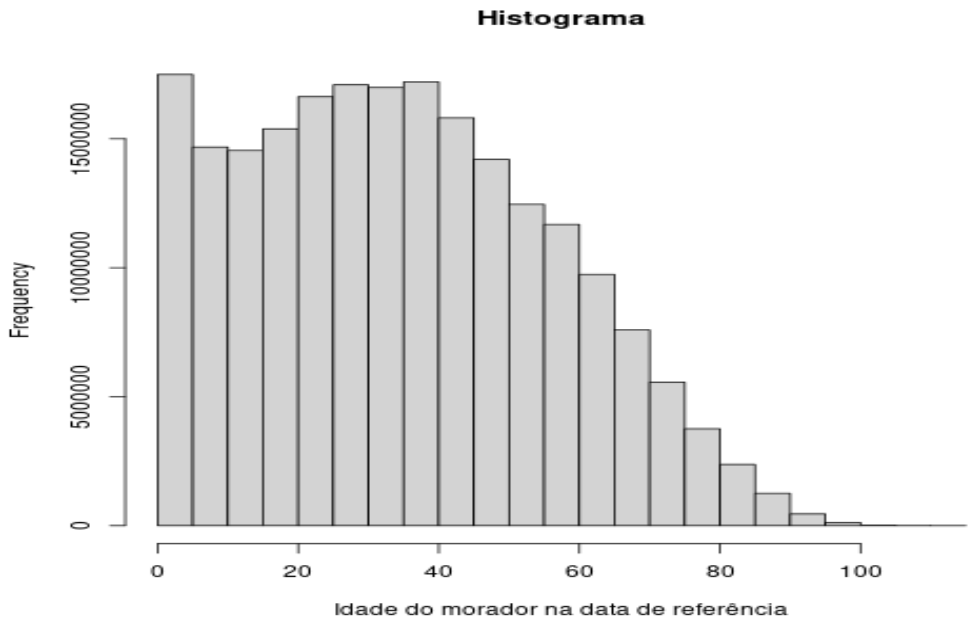
```
prop.table(svytable(~UF + VD4002, design = pnadc_2023q1), margin = 1) |>
 kable() |>
 kable_styling(bootstrap_options = c('striped', 'condensed')) |>
 scroll_box(width = "1050px", height = "330px")
```

	Pessoas ocupadas	Pessoas desocupadas
Rondônia	0,9684	0,0316
Acre	0,9025	0,0975
Amazonas	0,8949	0,1051
Roraima	0,9324	0,0676
Pará	0,9018	0,0982
Amapá	0,8784	0,1216
Tocantins	0,9309	0,0691

Microdados da PNAD Contínua 📊 – estatísticas com **survey**

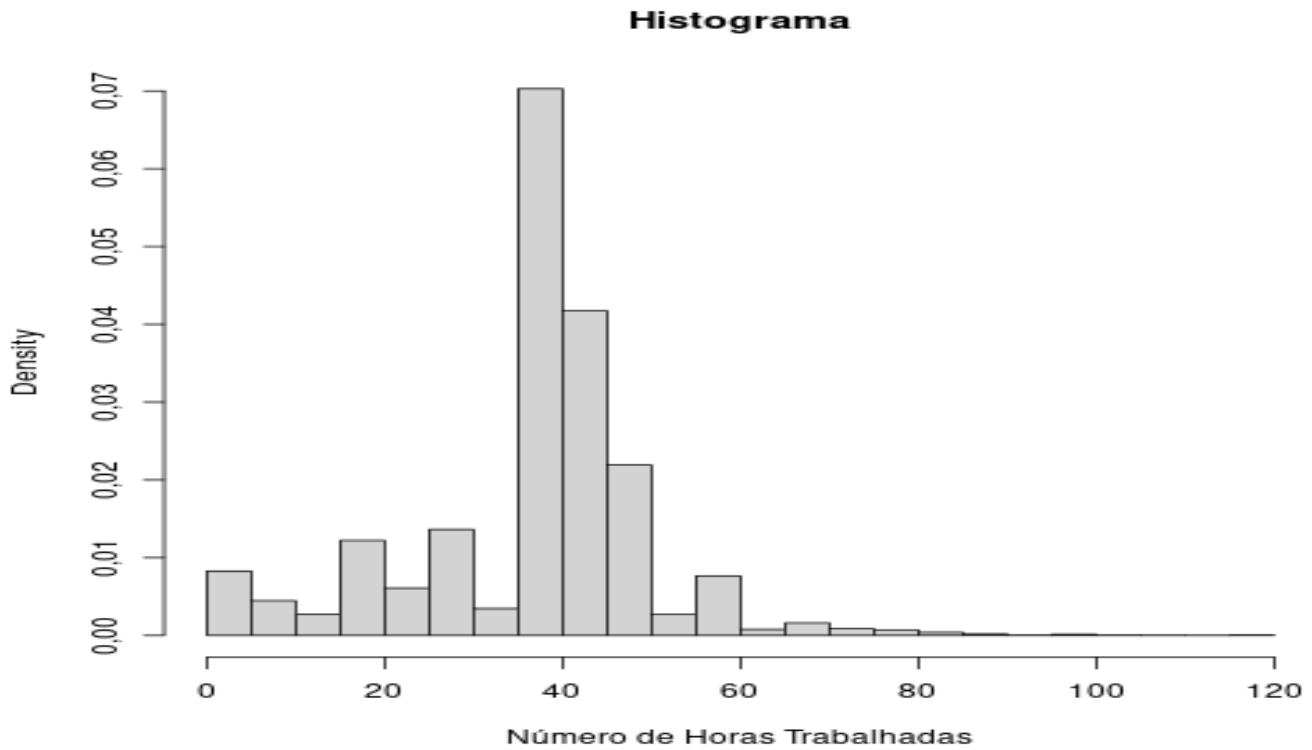
O pacote **survey** possui outras funções como svyvar(), svyratio(), svyquantile(), entre outras. Também possui funções para gerar gráficos como svyplot(), svyhist() e svysmooth(). Veja os exemplos:

```
svyhist(formula = ~V2009, design = pnadc_2023q1, freq = TRUE,
        main = "Histograma", xlab = "Idade do morador na data de referência")
```



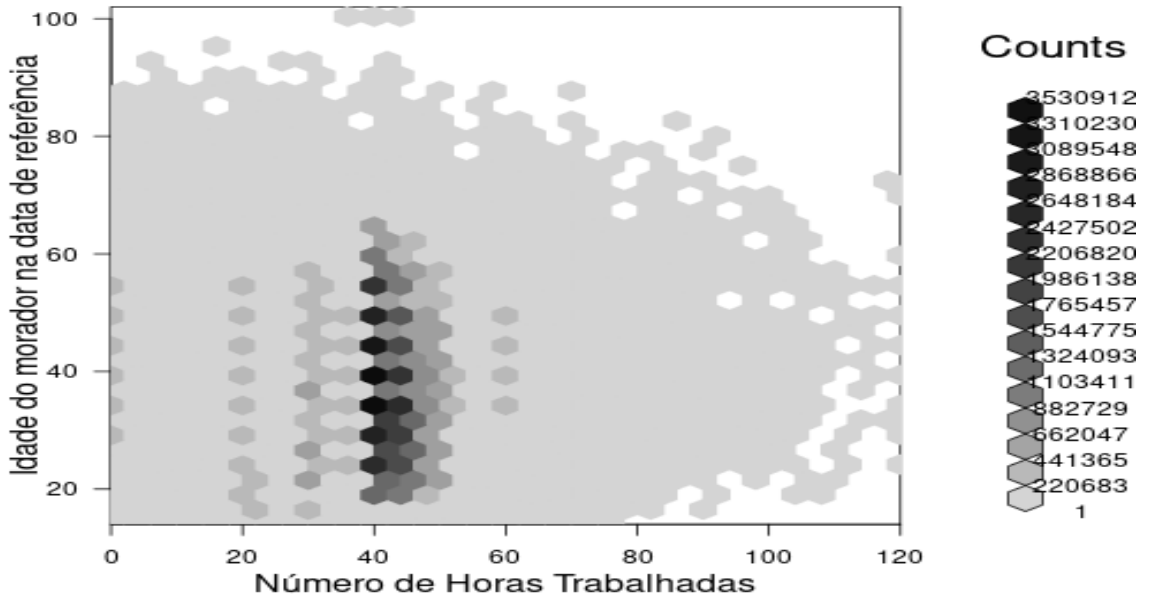
Microdados da PNAD Contínua 📊 – estatísticas com **survey**

```
svyhist(formula = ~VD4035, design = pnadc_2023q1,
        main = "Histograma", xlab = "Número de Horas Trabalhadas")
```



Microdados da PNAD Contínua 📊 – estatísticas com **survey**

```
svyplot(formula = V2009~VD4035, design = pnadc_2023q1,
                style = "grayhex", xlab = "Número de Horas Trabalhadas",
                ylab = "Idade do morador na data de referência")
```



Microdados da PNAD Contínua 📊 – estatísticas ponderadas

Nos próximos exemplos trabalharemos com os seguintes dados:

```
pnadc2023_1trim <- read_pnadc(microdata = "PNADC20231/PNADC_012023.txt",
                              input_txt = "PNADC20231/input_PNADC_trimestral.txt",
                              vars = c("Ano","Trimestre","UF", "Capital","UPA",
                                       "Estrato", "posest", "V1027","V1028", "V1029",
                                       "V2003","V2007","V2009","V2010",
                                       "V3007","VD3004","VD3005",
                                       "VD4001","VD4002","VD4005","VD4020"))
pnadc2023_1trim <-  pnadc2023_1trim |>
  # Manter os códigos dos estados e anos de estudo
  mutate(code_state = as.numeric(UF),
         anos_de_estudo = as.numeric(VD3005))
```

Não iremos aplicar o plano amostral, no momento. No entanto, manteremos os pesos replicados (de **V1028001** à **V1028200**) no conjunto de dados para gerar o plano amostral com objeto do tipo **svyrep.design**, posteriormente. Além disso, criaremos uma variável com o código de cada unidade da federação (**code_state**). Para atribuir *labels* categóricos e obter variáveis de deflação procedemos ...

Microdados da PNAD Contínua 📊 - estatísticas ponderadas

```
# obtendo labels categóricos e variáveis de deflação
pnadc2023_1trim <- pnadc_labeller(
  data_pnadc = pnadc2023_1trim,
  dictionary.file = "PNADC20231/dicionario_PNADC_microdados_trimestral.xls"
)
pnadc2023_1trim <- pnadc_deflator(
  data_pnadc = pnadc2023_1trim,
  deflator.file = "PNADC20231/deflator_PNADC_2023_trimestral_040506.xls"
)
```

```
class(pnadc2023_1trim)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

Após as operações realizadas, o objeto **pnadc2023_1trim** será do tipo **tibble**, possibilitando o uso do **dplyr** e **ggplot2**. Na geração de estatísticas, podem ser consideradas funções ponderadas - ex: weighted.mean(), weighted.median() ou weighted.quantile(). No entanto, as estatísticas geradas podem divergir daquelas obtidas com os pacotes **survey/srvyr** pois não consideram algumas variáveis do plano amostral como o **Estrato**, entre outras.

Microdados da PNAD Contínua 📊 – estatísticas ponderadas

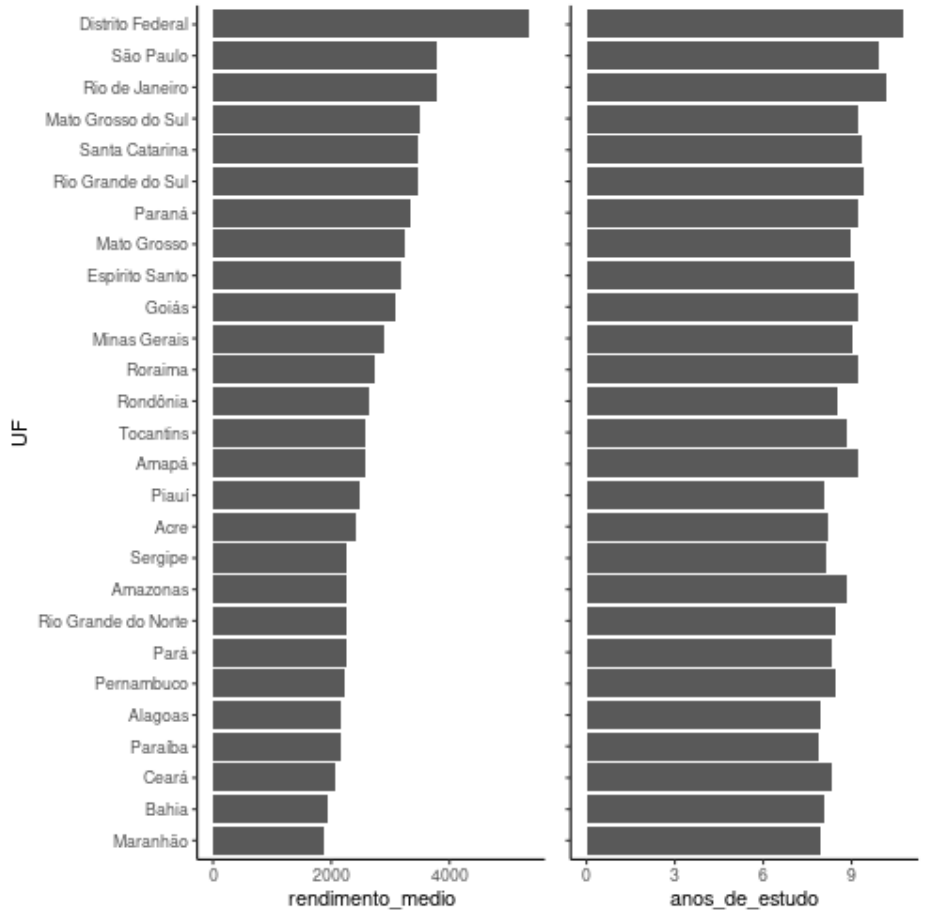
```
rendimento_mensal_estados <- pnadc2023_1trim |>
  summarise(rendimento_medio = weighted.mean(VD4020, w = V1028, na.rm = TRUE),
            anos_de_estudo = weighted.mean(anos_de_estudo, w = V1028, na.rm = TRUE),
            .by = "UF") |>
  drop_na() |>
  arrange(desc(rendimento_medio))
rendimento_mensal_estados |> kable() |>
  kable_styling(bootstrap_options = c('striped', 'condensed')) |>
  scroll_box(width = "1050px", height = "210px")
```

UF	rendimento_medio	anos_de_estudo
Distrito Federal	5347	10,731
São Paulo	3788	9,951
Rio de Janeiro	3773	10,173
Mato Grosso do Sul	3493	9,244

Microdados da PNAD Contínua 📊 - estatísticas ponderadas

O objeto **rendimento_mensal_estados** armazena as estimativas para o rendimento médio mensal e anos de estudo das pessoas com 14 anos ou mais de idade por estado, observável nos gráficos:

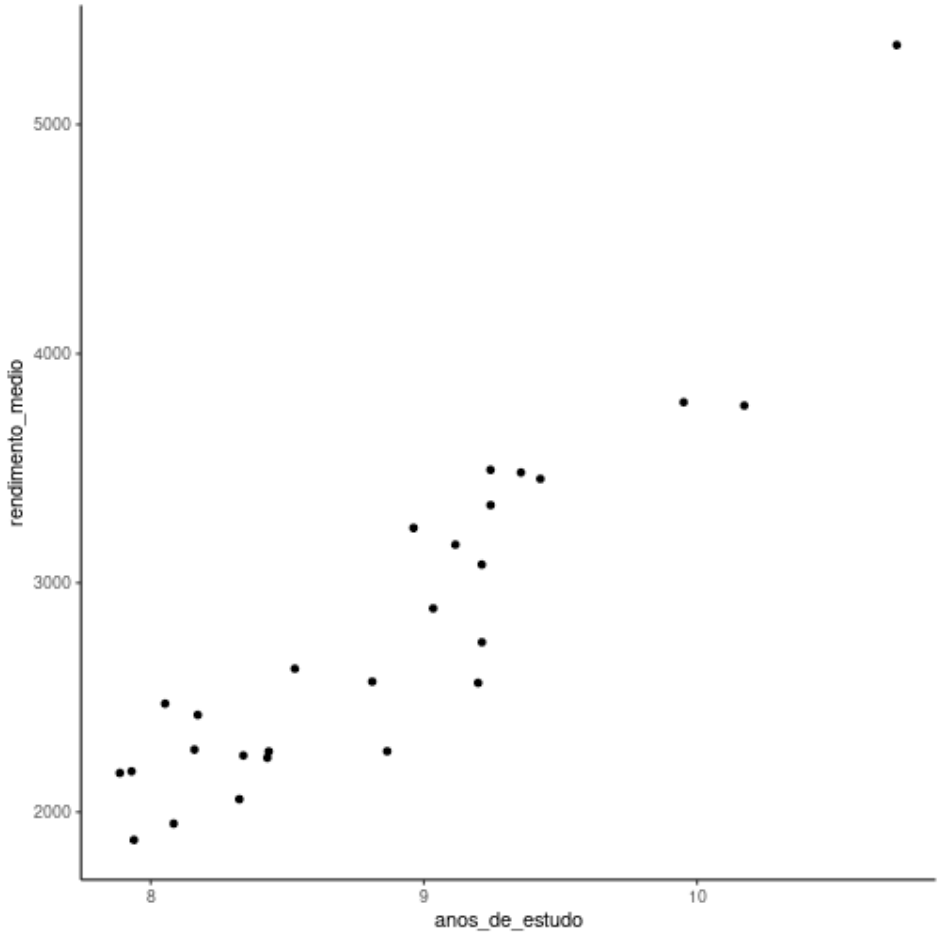
```
dt <- rendimento_mensal_estados |>
  mutate(
    UF = fct_reorder(UF, rendimento_medio)
  )
 p <- ggplot(dt) + theme_classic()
 p1 <- p +
   geom_col(aes(rendimento_medio, UF))
 p2 <- p +
   geom_col(aes(anos_de_estudo, UF)) +
   theme(axis.text.y = element_blank(),
         axis.title.y = element_blank())
 p1 + p2 # unindo gráficos com 'patchwork'
```



Microdados da PNAD Contínua 📊 - estatísticas ponderadas

Podemos observar o rendimento médio mensal por anos de estudo, considerando os estados, através de gráfico de dispersão:

```
ggplot(rendimento_mensal_estados,
       aes(anos_de_estudo,
       rendimento_medio)) +
geom_point() +
theme_classic()
```



Microdados da PNAD Contínua 📊 – estatísticas ponderadas

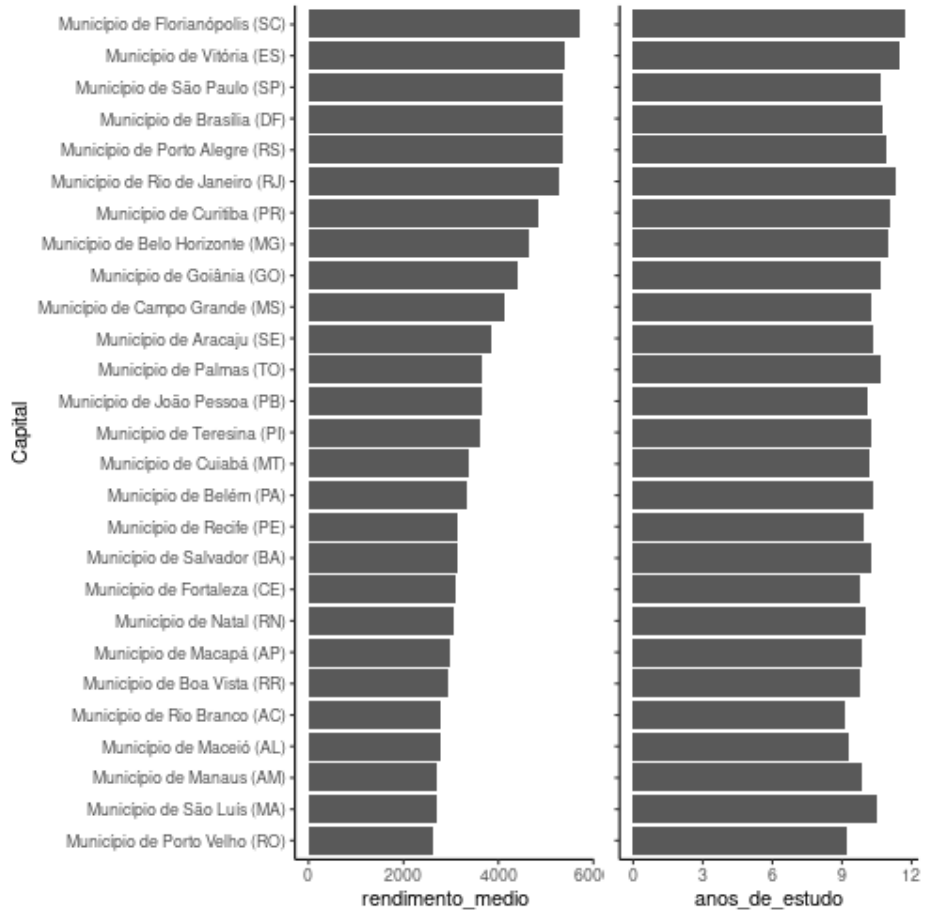
```
rendimento_mensal_capitais <- pnadc2023_1trim |>
  summarise(rendimento_medio = weighted.mean(VD4020, w = V1028, na.rm = TRUE),
            anos_de_estudo = weighted.mean(anos_de_estudo, w = V1028, na.rm = TRUE),
            .by = "Capital") |>
  drop_na() |>
  arrange(desc(rendimento_medio))
rendimento_mensal_capitais |> kable() |>
  kable_styling(bootstrap_options = c('striped', 'condensed')) |>
  scroll_box(width = "1050px", height = "210px")
```

Capital	rendimento_medio	anos_de_estudo
Município de Florianópolis (SC)	5723	11,706
Município de Vitória (ES)	5391	11,441
Município de São Paulo (SP)	5353	10,682
Município de Brasília (DF)	5347	10,731

Microdados da PNAD Contínua 📊 – estatísticas ponderadas

O objeto **rendimento_mensal_capitais** armazena as estimativas do rendimento médio mensal e anos de estudo das pessoas com 14 anos ou mais de idade nas capitais, observável nos gráficos:

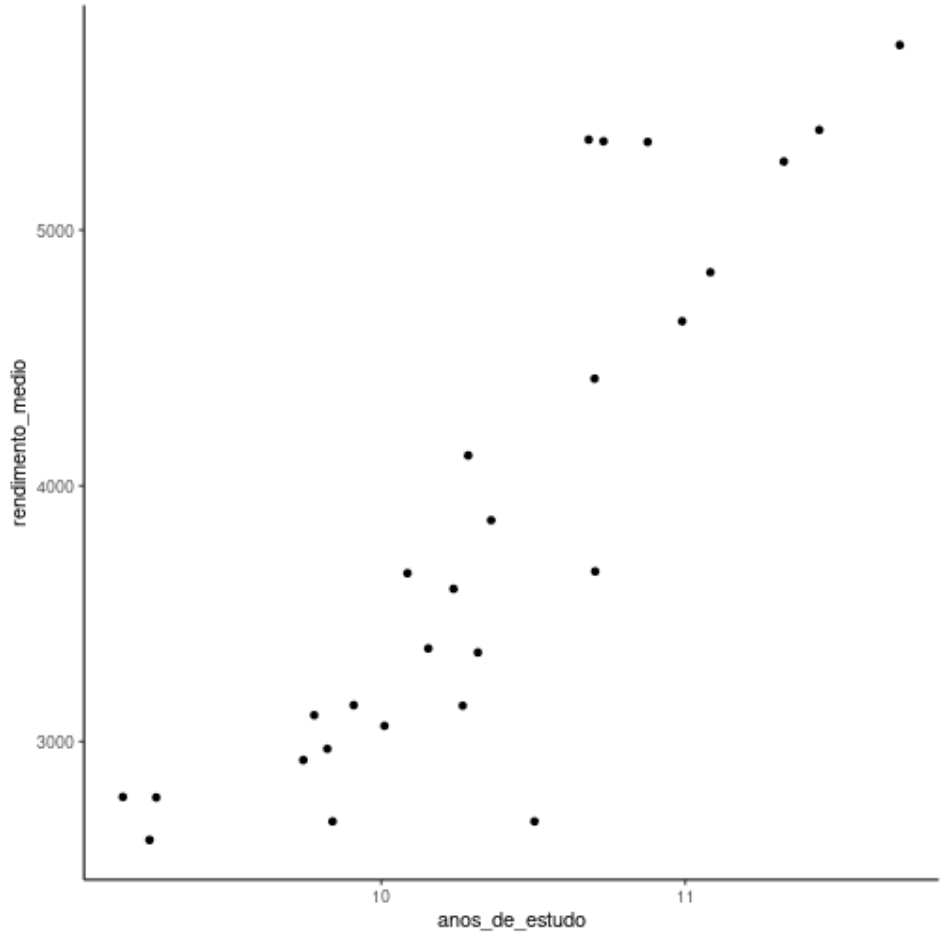
```
dt <- rendimento_mensal_capitais |>
  mutate(
    Capital = fct_reorder(Capital,
                          rendimento_medio)
  )
 p <- ggplot(dt) + theme_classic()
 p1 <- p +
   geom_col(aes(rendimento_medio,Capital))
 p2 <- p +
   geom_col(aes(anos_de_estudo,Capital))+
   theme(axis.text.y = element_blank(),
         axis.title.y = element_blank())
 p1 + p2
```



Microdados da PNAD Contínua 📊 - estatísticas ponderadas

Podemos observar o rendimento médio mensal por anos de estudo, considerando as capitais, através de gráfico de dispersão:

```
rendimento_mensal_capitais |>
  ggplot(aes(anos_de_estudo,
             rendimento_medio)) +
  geom_point() +
  theme_classic()
```



Microdados da PNAD Contínua 📊 – estatísticas com **srvyr**

Agora, vamos usar de forma combinada **survey** e **srvyr**. Antes, necessitamos carregar o pacote **srvyr**:

```
library(srvyr)
```

Aplicaremos o plano amostral nos nossos dados para obter um objeto **svyrep.design**:

```
pnadc2023_1trim <- PNADcIBGE::pnadc_design(data_pnadc = pnadc2023_1trim)
class(pnadc2023_1trim)
```

```
## [1] "svyrep.design"
```

E tranformaremos novamente o objeto para poder trabalhar com **dplyr** e **ggplot2**:

```
pnadc2023_1trim <- srvyr::as_survey(.data = pnadc2023_1trim)
class(pnadc2023_1trim)
```

```
## [1] "tbl_svy"       "svyrep.design"
```

Microdados da PNAD Contínua 📊 - estatísticas com **srvyr**

Esse objeto continua sendo uma lista cujos elementos são:

```
names(pnadc2023_1trim)
```

```
##  [1] "type"             "scale"            "rscales"          "rho"
##  [5] "call"             "combined.weights" "variables"        "pweights"
##  [9] "repweights"       "degf"             "mse"
```

No item "variables" está o objeto **tibble** com os microdados, as variáveis disponíveis são:

```
names(pnadc2023_1trim$variables)
```

Observe que é possível usar funções do pacote **dplyr** no novo objeto:

```
glimpse(pnadc2023_1trim)
```

Da mesma forma, funções de outros pacotes do **tidyverse** podem ser usadas. Além disso, o pacote **srvyr** possui funções próprias como survey_mean(), survey_ratio(), survey_total(), survey_count(), entre outras.

Microdados da PNAD Contínua 📊 – estatísticas com **srvyr**

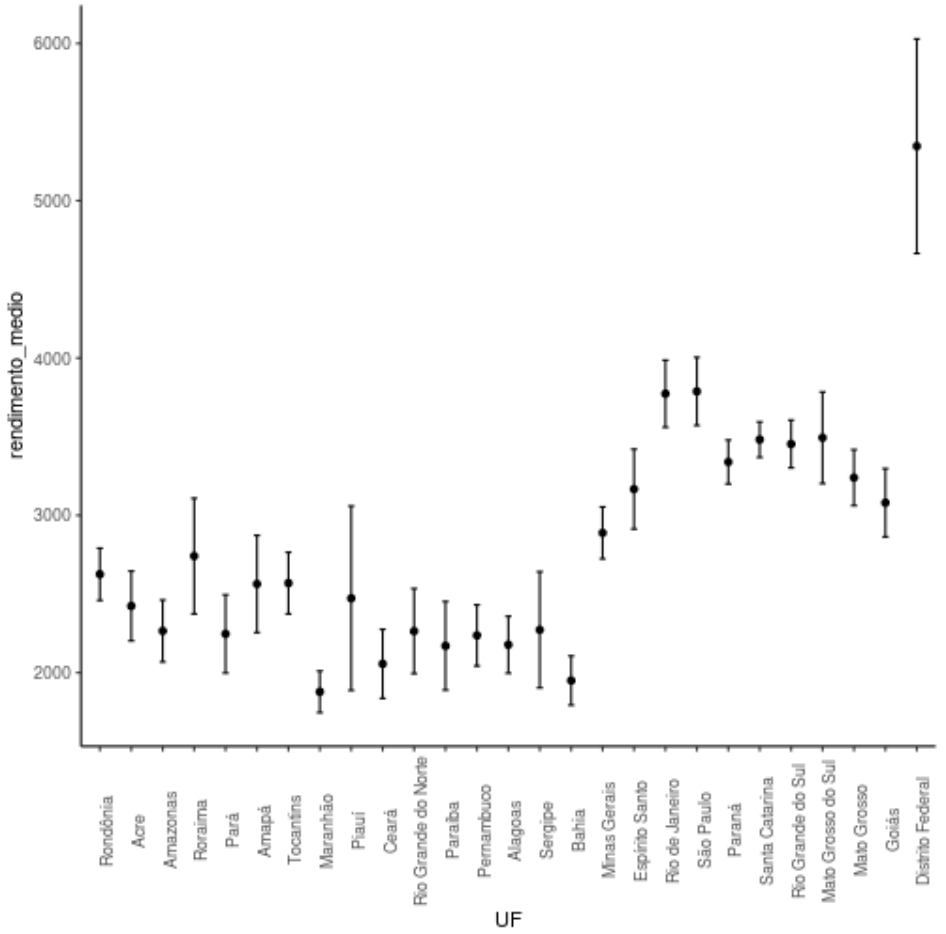
```
rendimento_mensal_estados <- pnadc2023_1trim |>
  group_by(UF) |>
  summarise(rendimento_medio = survey_mean(VD4020, na.rm = T, vartype = "ci")) |>
  arrange(desc(rendimento_medio))
rendimento_mensal_estados |>
  kable() |>
  kable_styling(bootstrap_options = c('striped', 'condensed')) |>
  scroll_box(width = "1050px", height = "250px")
```

UF	rendimento_medio	rendimento_medio_low	rendimento_medio_upp
Distrito Federal	5347	4665	6028
São Paulo	3788	3572	4004
Rio de Janeiro	3773	3560	3987
Mato Grosso do Sul	3493	3202	3784
Santa Catarina	3481	3369	3594

Microdados da PNAD Contínua 📊 – estatísticas com **srvyr**

O objeto **rendimento_mensal_estados** contém a estimativa do rendimento médio mensal das pessoas com 14 anos ou mais de idade por unidade da federação, assim como, o intervalo de confiança dessa estimativa - ambos visualizáveis na forma:

```
rendimento_mensal_estados |>
  ggplot(aes(UF, rendimento_medio)) +
  geom_point() +
  geom_errorbar(
    aes(ymin = rendimento_medio_low,
        ymax = rendimento_medio_upp),
    width = 0.2) +
  theme_classic() +
  theme(
    axis.text.x = element_text(angle = 90)
  )
```



Microdados da PNAD Contínua 📊 – estatísticas com **srvyr**

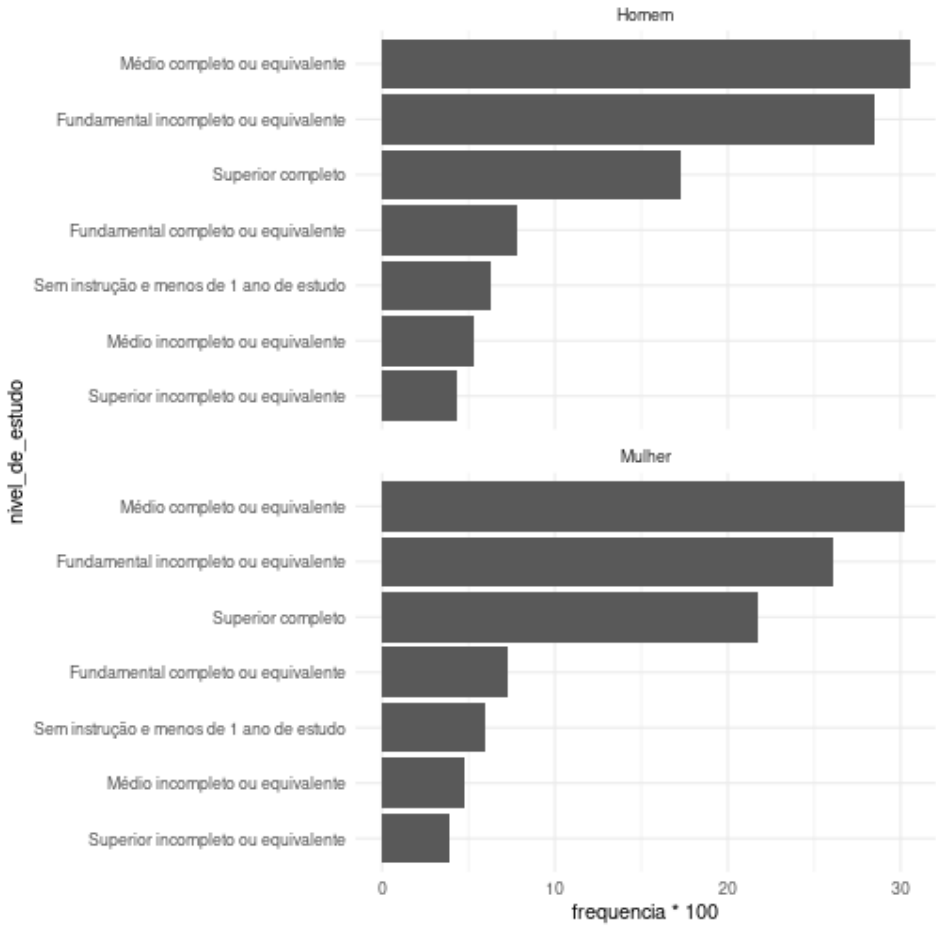
```
proporcao_nivel_estudo <- pnadc2023_1trim |>
  filter(V2009 >= 25) |>
  group_by(V2007, VD3004) |>
  summarise(frequencia = survey_mean())
proporcao_nivel_estudo |>
  kable() |>
  kable_styling(bootstrap_options = c('striped', 'condensed')) |>
  scroll_box(width = "1050px", height = "250px")
```

V2007	VD3004	frequencia	frequencia_se
Homem	Sem instrução e menos de 1 ano de estudo	0,0628	0,0009
Homem	Fundamental incompleto ou equivalente	0,2846	0,0021
Homem	Fundamental completo ou equivalente	0,0778	0,0010
Homem	Médio incompleto ou equivalente	0,0532	0,0009
Homem	Médio completo ou equivalente	0,3050	0,0020

Microdados da PNAD Contínua 📊 – estatísticas com **srvyr**

O objeto **proporcao_nivel_estudo** contém a proporção estimada da população com 25 anos ou mais de idade em cada nível de instrução, considerando **V2007** – visualizável na forma:

```
proporcao_nivel_estudo |>
  mutate(
    nivel_de_estudo = fct_reorder(
      VD3004, frequencia
    )
  ) |>
  drop_na() |>
  ggplot(aes(frequencia * 100,
             nivel_de_estudo)) +
  geom_col() +
  facet_wrap(~V2007, nrow = 2) +
  theme_minimal()
```



Microdados da PNAD Contínua 📊 – Relembrando

- **survey.design/svyrep.design** podem apresentar diferenças para estimativas de erros (ex: desvio padrão), mas as estimativas obtidas para variáveis como totais, médias e proporções, devem ser idênticas.

- Estatísticas com **survey**:

  - Consideram o plano amostral.
  - Não permitem manipulação dos dados com funções do pacote **dplyr** e correlatos.
  - Custo computacional pode ser alto conforme à complexidade da operação e quantidade de dados.

- Estatísticas com **survey/srvyr**:

  - Consideram o plano amostral.
  - Permitem usar funções do pacote **dplyr** e correlatos.
  - Custo computacional pode ser alto conforme à complexidade da operação e quantidade de dados.

- Estatísticas ponderadas:

  - Consideram os pesos dos domicílios e das pessoas, mas não outras variáveis do plano amostral.
  - Podem combinar manipulação de dados usando **dplyr** com funções que aplicam ponderação.
  - Custo computacional bem inferior em relação às funções dos pacotes **survey/srvyr**.

Microdados da PNAD Contínua 📊 - Exercícios

Exercício 1: Gere a estimativa do rendimento médio mensal das pessoas com idade igual ou superior à 14 anos, segundo as unidades da federação e a variável **V2007**. Use o pacote **srvyr** e operações do **dplyr**.

```
rendimento_mensal <- pnadc2023_1trim |>
  group_by(UF, V2007) |>
  summarise(rendimento_medio = survey_mean(VD4020, na.rm = TRUE)) |>
  tidyr::drop_na() |>
  arrange(UF)
```

Microdados da PNAD Contínua 📊 – Exercícios

Exercício 1:

```
rendimento_mensal |>
  kable() |>
  kable_styling(bootstrap_options = c('striped', 'condensed')) |>
  scroll_box(width = "1050px", height = "330px")
```

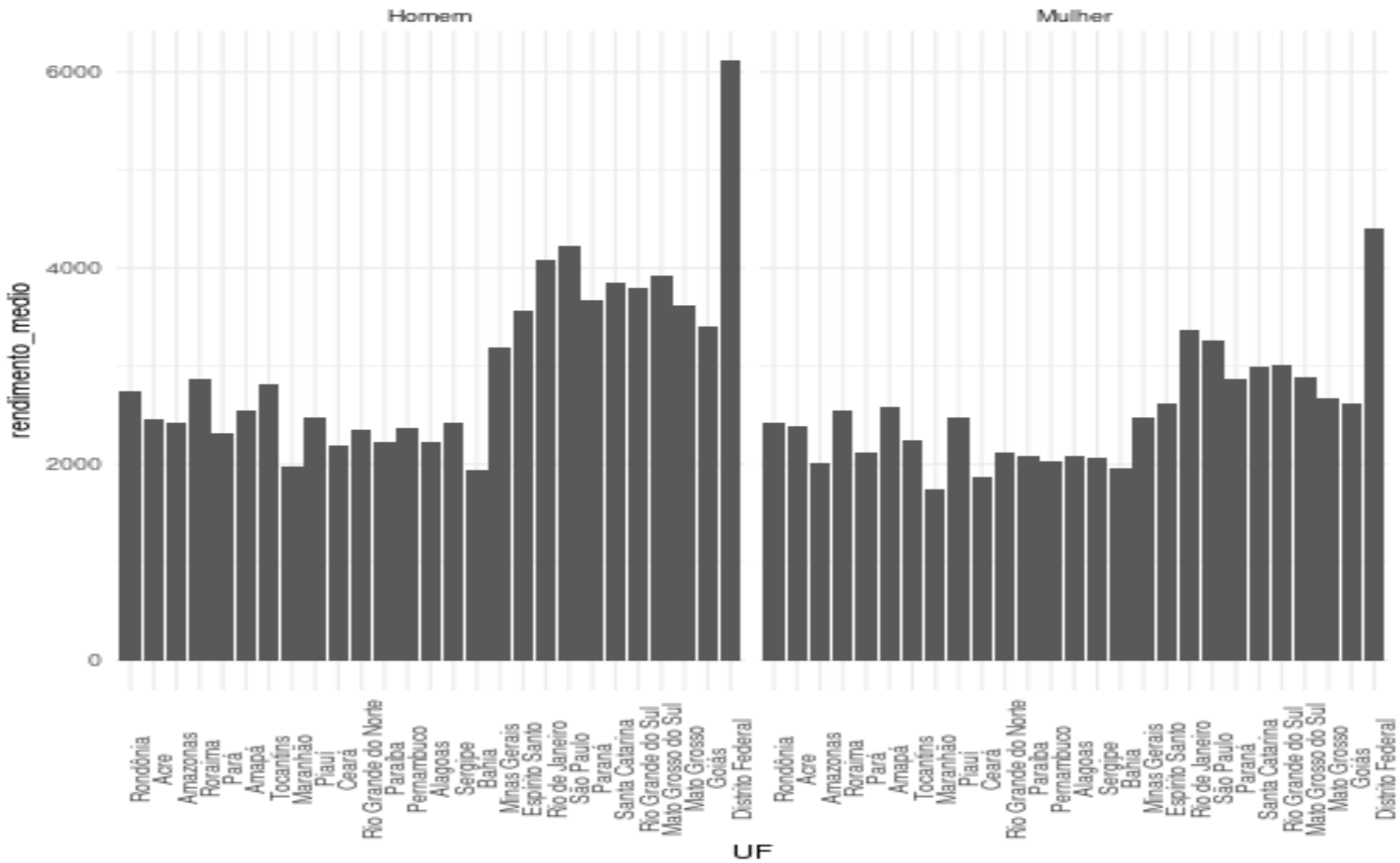
UF	V2007	rendimento_medio	rendimento_medio_se
Rondônia	Homem	2746	94,17
Rondônia	Mulher	2432	110,48
Acre	Homem	2451	126,33
Acre	Mulher	2382	118,31
Amazonas	Homem	2422	126,62
Amazonas	Mulher	2018	83,20
Roraima	Homem	2865	191,24

Microdados da PNAD Contínua 📊 - Exercícios

Exercício 1: Gere gráficos de colunas do rendimento médio mensal, segundo a variável **V2007**, conforme:

```
rendimento_mensal |>
  ggplot(aes(UF, rendimento_medio)) +
  geom_col() +
  facet_wrap(~V2007, nrow = 1) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90))
```

Microdados da PNAD Contínua 📊 – Exercícios



Microdados da PNAD Contínua 📊 - Exercícios

Exercício 2: Estime o rendimento médio mensal das pessoas com idade igual ou superior à 14 anos, bem como, a média dos anos de estudo. Agrupe os dados por estado e pela variável **V2007**. Use os pacotes **srvyr** e **dplyr**.

```
rendimento_mensal <- pnadc2023_1trim |>
  group_by(UF, V2007) |>
  summarise(rendimento_medio = survey_mean(VD4020, na.rm = TRUE, vartype = "ci"),
            anos_de_estudo = survey_mean(anos_de_estudo, na.rm = TRUE),
  ) |>
  tidyr::drop_na() |>
  arrange(UF)
```

Microdados da PNAD Contínua 📊 - Exercícios

Exercício 2:

```
rendimento_mensal |>
  kable() |>
  kable_styling(bootstrap_options = c('striped', 'condensed')) |>
  scroll_box(width = "1050px", height = "250px")
```

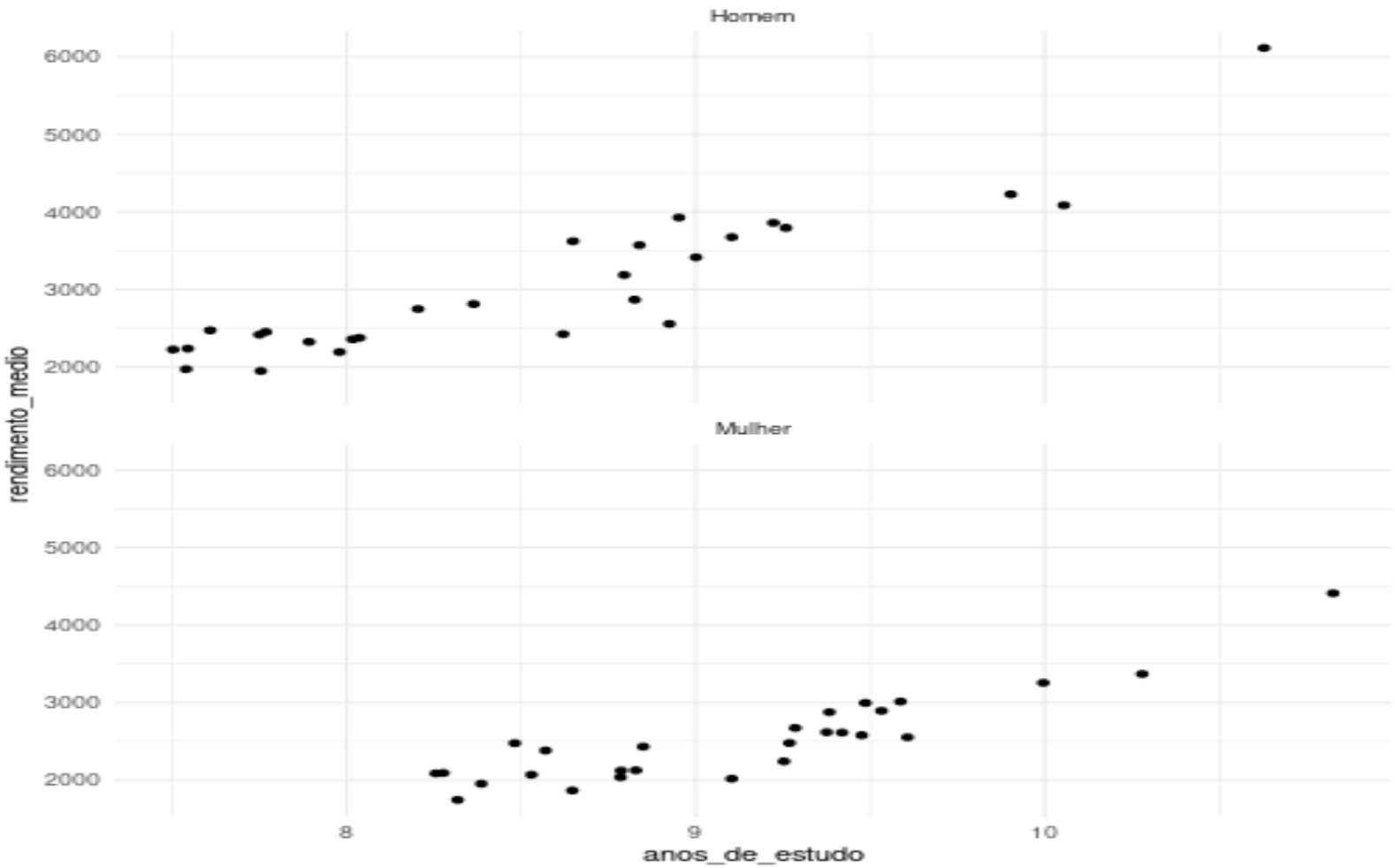
UF	V2007	rendimento_medio	rendimento_medio_low	rendimento_medio_upp	anos_de_estud
Rondônia	Homem	2746	2560	2931	8,20
Rondônia	Mulher	2432	2214	2649	8,85
Acre	Homem	2451	2202	2700	7,76
Acre	Mulher	2382	2149	2616	8,57
Amazonas	Homem	2422	2172	2671	8,62

Microdados da PNAD Contínua 📊 - Exercícios

Exercício 2: Gere gráficos de dispersão do rendimento médio mensal por anos de estudo, segundo a variável **V2007**, conforme:

```
rendimento_mensal |>
  ggplot(aes(anos_de_estudo, rendimento_medio)) +
  geom_point() +
  facet_wrap(~V2007, nrow = 2) +
  theme_minimal()
```

Microdados da PNAD Contínua 📊 – Exercícios

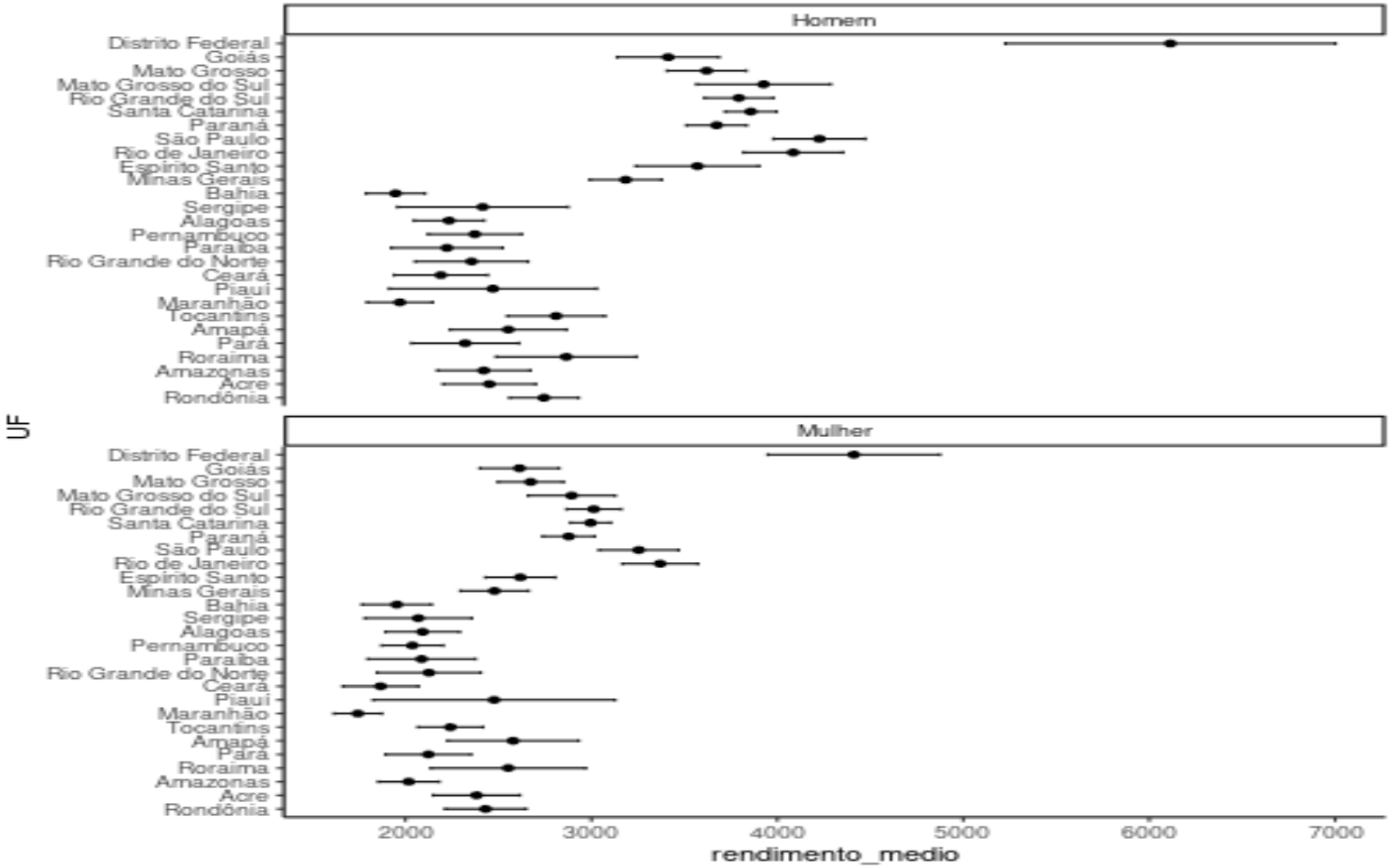


Microdados da PNAD Contínua 📊 - Exercícios

Exercício 2: Gere sub-gráficos (segundo **V2007**) contendo o rendimento médio mensal e intervalo de confiança para cada estado, conforme:

```
rendimento_mensal |>
  ggplot(aes(rendimento_medio, UF)) +
  geom_point() +
  geom_errorbar(
    aes(xmin = rendimento_medio_low,
        xmax = rendimento_medio_upp),
    width = 0.2) +
  facet_wrap(~V2007, nrow = 2) +
  theme_classic()
```

Microdados da PNAD Contínua 📊 – Exercícios



Microdados da PNAD Contínua 📊 - Exercícios

Exercício 3: Estime o rendimento médio mensal das pessoas com idade igual ou superior à 14 anos por nível de instrução (**VD3004**). Agrupe os dados por estado.

```
rendimento_mensal <- pnadc2023_1trim |>
  filter(VD3004 != "NA") |>
  group_by(UF, VD3004) |>
  summarise(rendimento_medio = survey_mean(VD4020, na.rm = TRUE, vartype = "ci")) |>
  tidyr::drop_na() |>
  arrange(UF)
```

Microdados da PNAD Contínua 📊 - Exercícios

Exercício 3:

```
rendimento_mensal |>
  kable() |>
  kable_styling(bootstrap_options = c('striped', 'condensed')) |>
  scroll_box(width = "1050px", height = "250px")
```

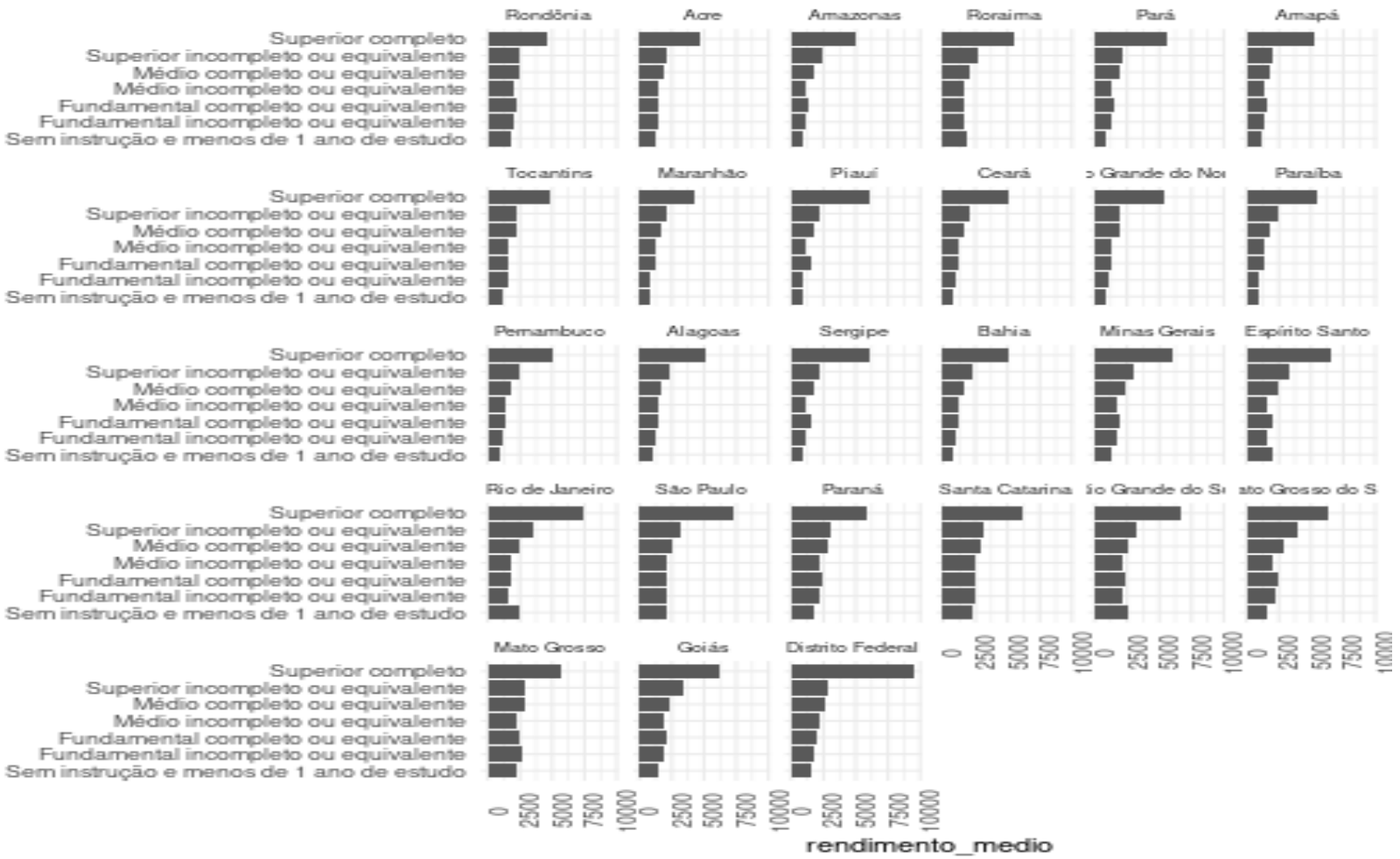
UF	VD3004	rendimento_medio	rendimento_medio_low	rendimento_medio_upp
Rondônia	Sem instrução e menos de 1 ano de estudo	1907,8	1493,6	2322,0
Rondônia	Fundamental incompleto ou equivalente	1959,9	1796,3	2123,4

Microdados da PNAD Contínua 📊 - Exercícios

Exercício 3: Gere gráficos de colunas com as estimativas do rendimento médio mensal por nível de instrução em cada estado, conforme:

```
rendimento_mensal |>
  ggplot(aes(rendimento_medio, VD3004)) +
  geom_col() +
  facet_wrap(~UF) +
  theme_minimal() +
  theme(axis.title.y = element_blank(),
        axis.text.x = element_text(angle = 90),
        strip.text = element_text(size = 7))
```

Microdados da PNAD Contínua – Exercícios



Microdados da PNAD Contínua 🌎 - visualização espacial

Considere os seguintes pacotes:

```
library(geobr)     # fornece mapas e conjuntos de dados do Brasil
library(sf)        # fornece representações e operações sobre dados espaciais
library(ggspatial) # fornece anotações (ex: escala) e outras operações em mapas
```

Vamos carregar um mapa do Brasil com a subdivisão dos estados:

```
states <- read_state( year = 2020, showProgress = FALSE)
```

```
glimpse(states)
```

```
## Rows: 27
## Columns: 6
## $ code_state   <dbl> 11, 12, 13, 14, 15, 16, 17, 21, 22, 23, 24, 25, 26, 27, 2…
## $ abbrev_state <chr> "RO", "AC", "AM", "RR", "PA", "AP", "TO", "MA", "PI", "CE…
## $ name_state   <chr> "Rondônia", "Acre", "Amazônas", "Roraima", "Pará", "Amapá…
## $ code_region  <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, …
## $ name_region  <chr> "Norte", "Norte", "Norte", "Norte", "Norte", "Norte", "No…
## $ geom         <MULTIPOLYGON [°]> MULTIPOLYGON (((-65,38 -10,..., MULTIPOLYGON…
```

Microdados da PNAD Contínua 🌎 – visualização espacial

Vamos estimar novamente o rendimento médio mensal, conforme o Exercício 3. Só que dessa vez, agruparemos os dados considerando a variável **code_state**.

```
rendimento_mensal <- pnadc2023_1trim |>
  filter(VD3004 != "NA") |>
  group_by(code_state, VD3004) |>
  summarise(rendimento_medio = survey_mean(VD4020, na.rm = TRUE, vartype = "ci")) |>
  tidyr::drop_na()
```

```
glimpse(rendimento_mensal)
```

```
## Rows: 189
## Columns: 5
## Groups: code_state [27]
## $ code_state           <dbl> 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 1…
## $ VD3004               <fct> Sem instrução e menos de 1 ano de estudo, Fundame…
## $ rendimento_medio     <dbl> 1907,8, 1959,9, 2141,0, 1937,5, 2362,3, 2537,8, 4…
## $ rendimento_medio_low <dbl> 1493,6, 1796,3, 1863,5, 1713,7, 2182,7, 2072,7, 4…
## $ rendimento_medio_upp <dbl> 2322,0, 2123,4, 2418,4, 2161,3, 2541,9, 3002,9, 5…
```

Microdados da PNAD Contínua 🌎 - visualização espacial

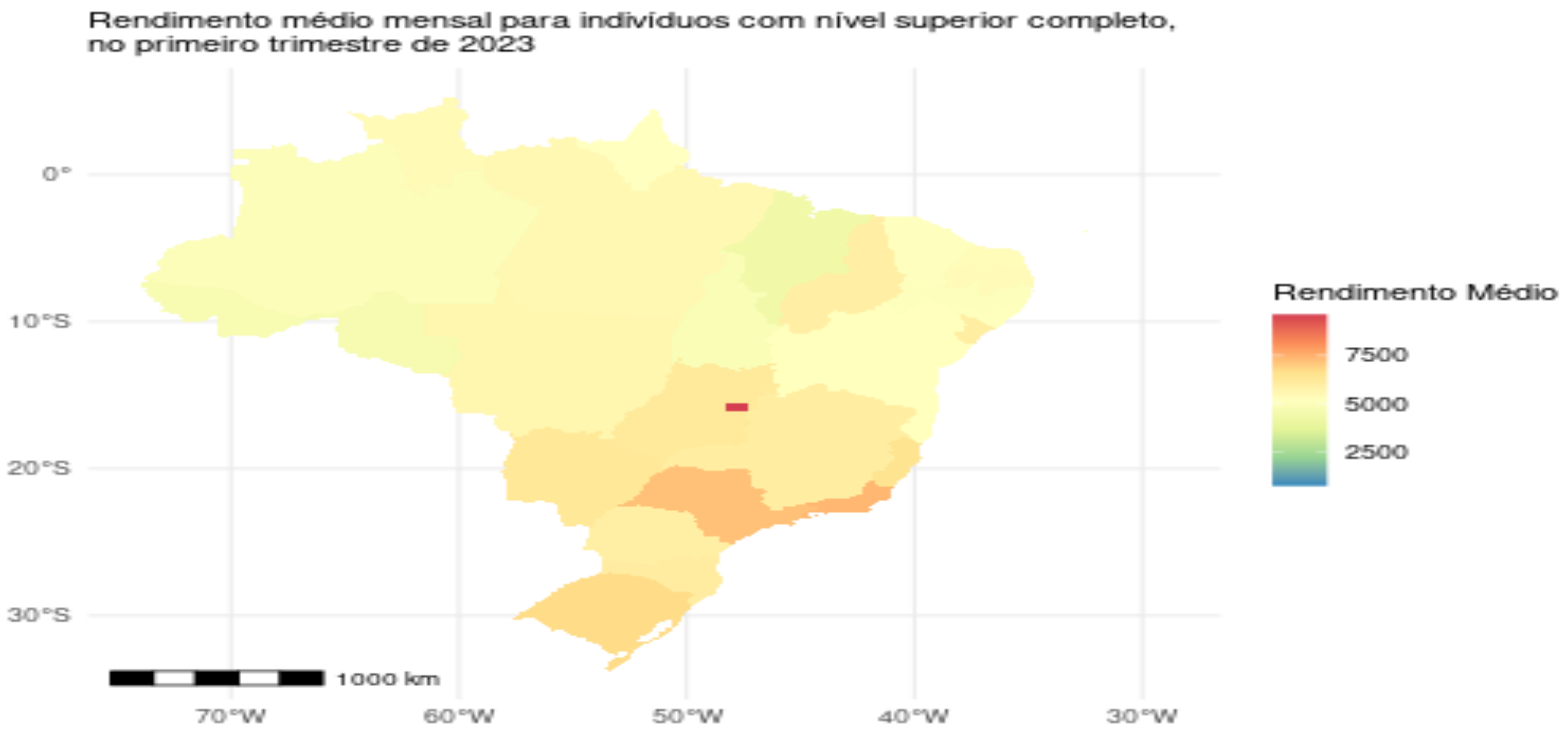
Agora, adicionaremos os dados de rendimento médio mensal ao conjunto de dados do mapa.

```
states <- dplyr::left_join(states, rendimento_mensal, by = "code_state")
```

Finalmente, podemos produzir um mapa usando o ggplot2, conforme:

```
titulo <- paste0("Rendimento médio mensal para indivíduos com nível superior completo,\n",
                 "no primeiro trimestre de 2023")
states |>
  filter(VD3004 == "Superior completo") |>
  ggplot() +
  geom_sf(data = states, aes(fill = rendimento_medio), color = NA) +
  labs(subtitle = titulo) +
  scale_fill_distiller(palette = "Spectral", name = "Rendimento Médio") +
  theme_minimal() +
  ggspatial::annotation_scale()
```

Microdados da PNAD Contínua 🌍 – visualização espacial

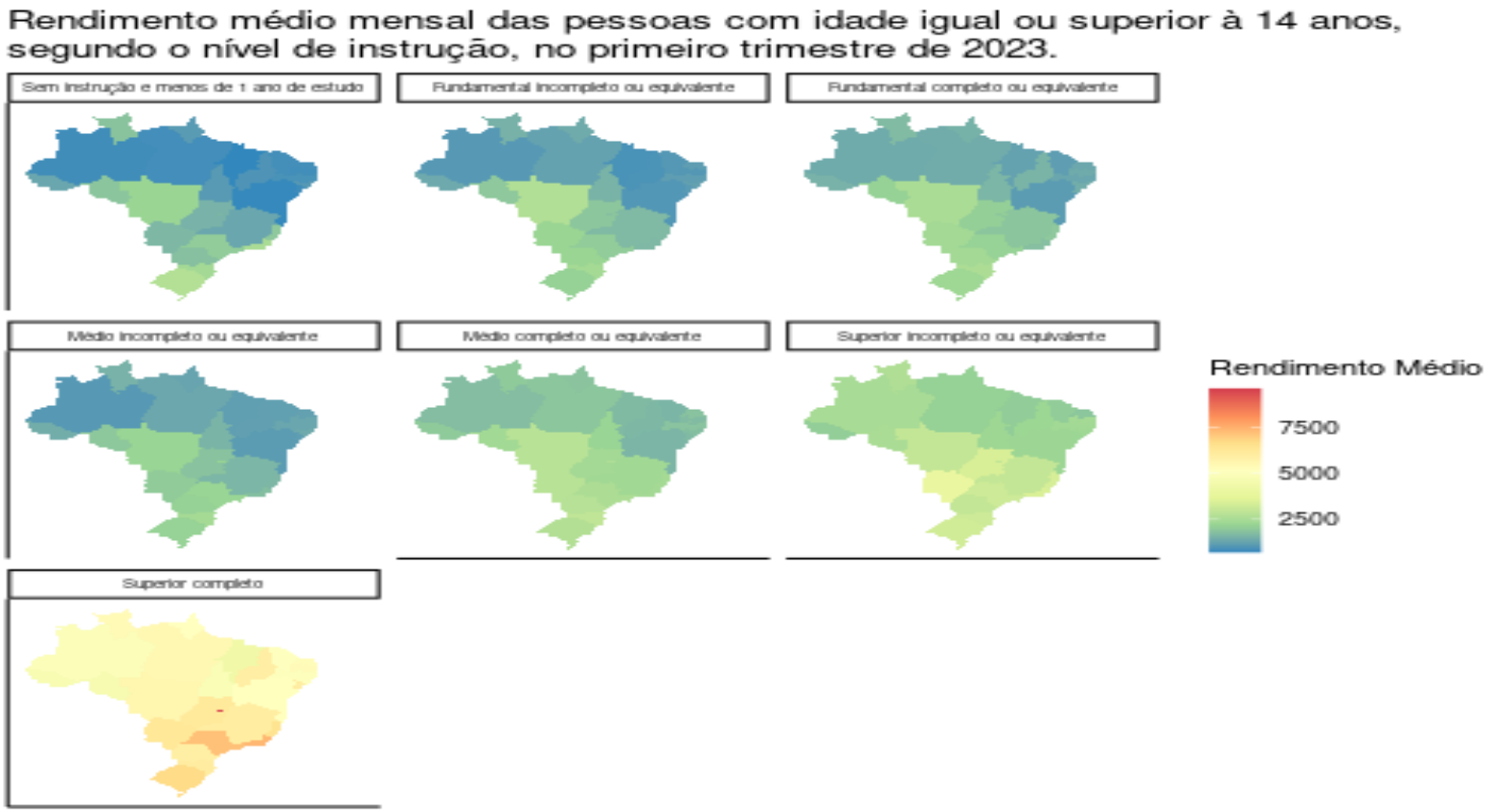


Microdados da PNAD Contínua 🌎 - visualização espacial

Também podemos produzir um mapa, conforme:

```
titulo <- paste0("Rendimento médio mensal das pessoas com idade igual ou superior à 14 anos,",
                 "\nsegundo o nível de instrução, no primeiro trimestre de 2023.")
states |> ggplot() +
  geom_sf(data = states, aes(fill = rendimento_medio), color = NA) +
  facet_wrap(~VD3004) +
  labs(title = titulo) +
  scale_fill_distiller(palette = "Spectral", name = "Rendimento Médio") +
  theme_classic() +
  theme(axis.text = element_blank(),
        axis.title = element_blank(),
        axis.ticks = element_blank(),
        strip.text = element_text(size = 6))
```

Microdados da PNAD Contínua 🌍 – visualização espacial



Microdados da PNAD Contínua

Há uma diversidade de bons materiais que orientam o uso dos microdados da PNAD Contínua através do R:

- Documentação online desenvolvido por um dos criadores do pacote **PNADcIBGE**.
- Livro "POR DENTRO DA PNAD CONTÍNUA: Uma introdução ao tratamento de dados usando o R".
- Curso de R disponível no repositório do Git Hub: Introdução à Programação em R.
- Materiais online sobre o pacote **srvyr**:
  - Link 1.
  - Link 2.

Metotologias Informacionais com 

Muito Obrigado pela Atenção!