Lista II: Análise de Dados da CAPES

Dan Nogueira da Silva

2024-03-05

Nesta lista, o objetivo será analisar a produção de teses e dissertações de programas de pós-graduação notas 4, 5, 6 e 7 na Capes, das áreas de Sociologia, entre os anos de 1987 e 2022. A análise explora a evolução da produção ao longo do tempo, a distribuição por subtemas e a concentração regional das defesas, seguindo um recorte de palavras-chave específico. Os dados foram processados e analisados utilizando o software R e pacotes para análise de dados e visualização.

Tendências na produção de teses e dissertações em Sociologia

Metodologia

A análise foi realizada utilizando dados coletados de duas fontes:

Banco de dados de defesas de teses e dissertações da CAPES: Contém informações sobre as defesas realizadas entre 1987 e 2022, incluindo dados sobre o programa, a instituição, o autor, o título, as palavras-chave e a área de avaliação.

Banco de dados de programas de pós-graduação da CAPES: Contém informações sobre os programas de pós-graduação, incluindo o código do programa, o estado e o conceito CAPES.

1. Importando arquivos e arrumando a base

Para começar a análise, foi preciso importar a base de dados contendo as dissertações e teses de programas na CAPES. Optei por utilizar a função `map_df`. Nesse caso, a `map_df` é mais prática que o laço de repetição, por economizar linhas e espaço de memória.

```
defesas <- map_df(c("csvs/capes_1987-1992.csv", "csvs/capes_1993-1998.csv",
                    "csvs/capes_1999-2004.csv", "csvs/capes_2005-2010.csv",
                    "csvs/capes_2011-2016.csv","csvs/capes_2017-2022.csv"),
                    read_delim, delim = ",")
```

Usando `map_df` é possível combinar todas as planilhas contendo as teses e dissertações defendidas entre 1987 a 2022, em um único tibble. Como resultado, a planilha possui 13 variáveis: código do programa, ano, sigla, instituição, nome do programa, grande área, área de conhecimento, área de avaliação, autor, titulo, nivel, palavras-chave e resumo. Cada observação diz respeito a uma defesa.

```
programas <- import("programas.csv")
```

Em seguida, usei a função `import()` do pacote `rio` para carregar a planilha com informações sobre os programas de pós-graduação. A planilha resultante possui 3 variáveis: código do programa, estado e conceito CAPES. Cada observação diz respeito a um programa de pós-graduação.

Para concatenar as informações sobre os programas com as informações sobre as defesas, utilizei a função `left_join` para juntar as informações dos programas partindo de uma variável em comum: o código do programa.

```
defesas_e_programas <- defesas |>
  left_join(programas, by = c("codigo_programa" = "CD_PROGRAMA"))
```

Ao analisar a nova base, percebi alguns missings na variável `UF`. Tentei arrumar a lista me guiando a partir da coluna `siglas_ies`. Além disso, incluí a variável `regiao`, que vai ser útil mais à frente.

```
banco_tidy <- defesas_e_programas |>
  mutate(UF = case_when(
    str_detect(sigla_ies, "RJ|RIO|UENF|UFF|UCAM") ~ "RJ",
    str_detect(sigla_ies, "SP|UNICAMP") ~ "SP",
    str_detect(sigla_ies, "ES|UVV") ~ "ES",
    str_detect(sigla_ies, "AC") ~ "AC",
    str_detect(sigla_ies, "AL") ~ "AL",
    str_detect(sigla_ies, "AP") ~ "AP",
    str_detect(sigla_ies, "UFAM") ~ "AM",
    str_detect(sigla_ies, "BA") ~ "BA",
    str_detect(sigla_ies, "CE|FJN|UFC") ~ "CE",
    str_detect(sigla_ies, "DF|UNB") ~ "DF",
    str_detect(sigla_ies, "GO|UFG") ~ "GO",
    str_detect(sigla_ies, "MA") ~ "MA",
    str_detect(sigla_ies, "MT") ~ "MT",
    str_detect(sigla_ies, "MS|UFGD") ~ "MS",
    str_detect(sigla_ies, "MG") ~ "MG",
    str_detect(sigla_ies, "PA") ~ "PA",
    str_detect(sigla_ies, "PB|UFCG") ~ "PB",
    str_detect(sigla_ies, "PR|UEL") ~ "PR",
    str_detect(sigla_ies, "PE|UNIVASF") ~ "PE",
    str_detect(sigla_ies, "PI") ~ "PI",
    str_detect(sigla_ies, "RN") ~ "RN",
    str_detect(sigla_ies, "RS|UFRGS") ~ "RS",
    str_detect(sigla_ies, "RO") ~ "RO",
    str_detect(sigla_ies, "RR") ~ "RR",
```

```
    str_detect(sigla_ies, "SC") ~ "SC",
    str_detect(sigla_ies, "SE") ~ "SE",
    str_detect(sigla_ies, "TO") ~ "TO",
    TRUE ~ "Outros"
  )) |>
  mutate(regiao = case_when(
    UF %in% c("AM", "RR", "AP", "PA", "TO", "RO", "AC") ~ "Norte",
    UF %in% c("MA", "PI", "CE", "RN", "PE", "PB", "SE", "AL", "BA") ~ "Nordeste",
    UF %in% c("MT", "MS", "GO", "DF") ~ "Centro Oeste",
    UF %in% c("PR", "SC", "RS") ~ "Sul",
    UF %in% c("SP", "RJ", "ES", "MG") ~ "Sudeste",
    TRUE ~ "Não reportado"
  ))
```

Após "dar uma arrumada" no banco, filtrei dele apenas as observações que continham programas com notas CAPES maiores que 4 em minha área de interesse (sociologia).

```
sociologia <- banco_tidy |>
  filter(CONCEITO != "NA|3|A") |>
  filter(str_detect(nome_programa, "SOCIOLOGIA"))
```

Por algum motivo os *missings* continuavam a ser considerados quando eu utilizava a lógica `CONCEITO == 4|5|6|7`, então achei melhor filtrar usando a negação das observações indesejadas.

2. Seleção de palavras-chave

Para o meu desenho de pesquisa, as 3 palavras-chave mais interessantes são **`ensino superior`**, **`desigualdade`** e **`educação`**. Filtrei a base para mostrar apenas defesas que se enquadravam em pelo menos uma das 3 palavras-chave.

```
trabalhos_relevantes <- sociologia |>
 filter(str_detect(palavras_chave,"ensino superior|desigualdade|educação"))
```

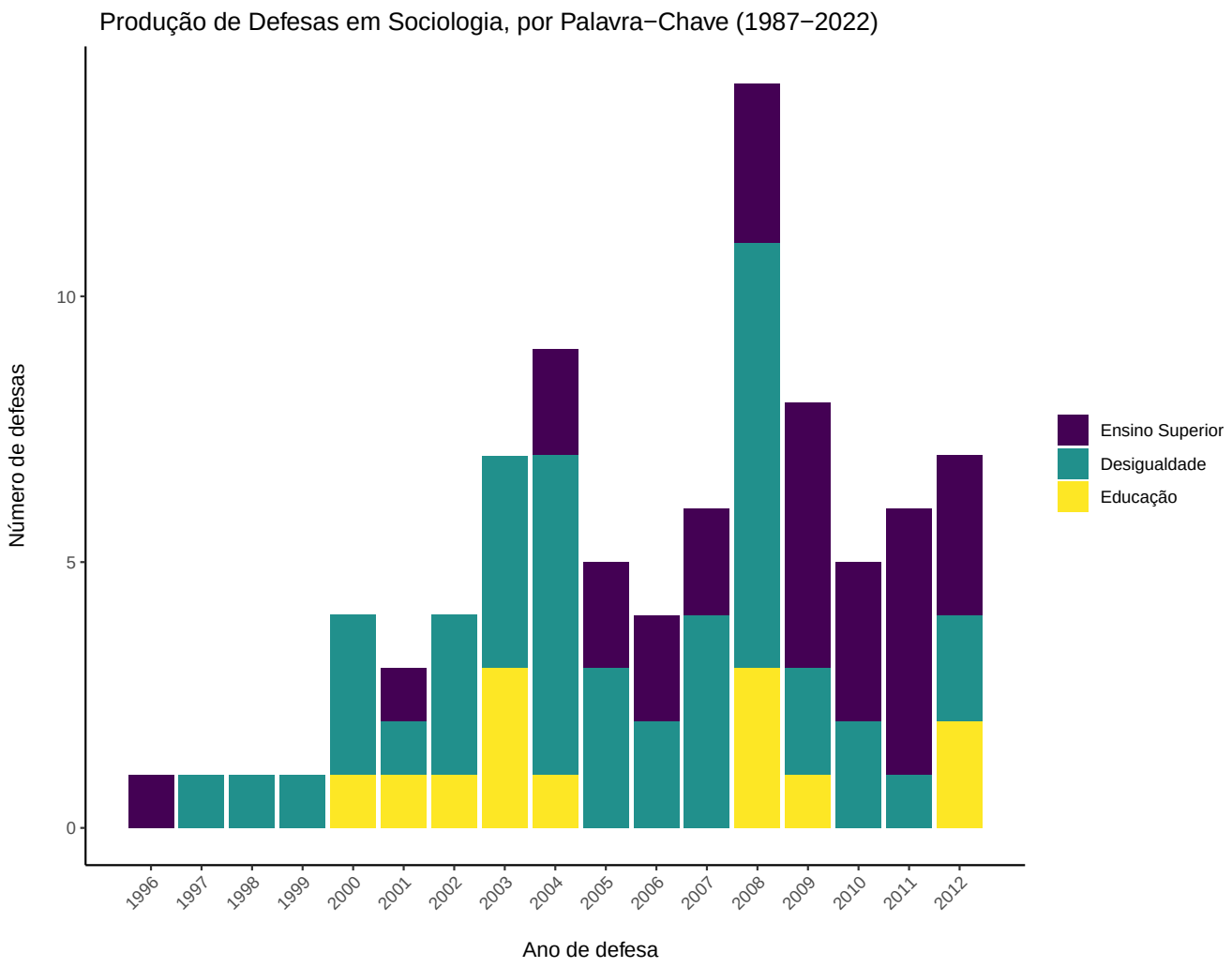
3. Evolução ao longo do tempo

O objetivo aqui é criar uma visualização que reporte de forma sucinta e informativa a produção de teses e dissertações do meu interesse, por ano. Sendo assim, era preciso criar uma coluna para contagem das ocorrências de defesas por ano e palavra-chave. Adicionei a variável **`subtema`**, que classificava as defesas por palavra-chave correspondente. Depois contei quantas ocorrências cada variável **`ano`** haviam em relação a cada observação da variável **`subtema`**.Dessa forma consigo um output com uma tabela 'long' contendo 3 variáveis: **`ano`**, **`subtema`** e **`frequência`**. Assim consigo os dados ideais para visualizar a primeira tabela da lista!

```r
ocorrencias_por_ano <- trabalhos_relevantes |>
 mutate(subtema = case_when(
  str_detect(palavras_chave, "ensino superior") ~ "Ensino Superior",
  str_detect(palavras_chave, "desigualdade") ~ "Desigualdade",
  str_detect(palavras_chave, "educação") ~ "Educação",
  TRUE ~ "Outros"  # catchError
 )) |>
 count(ano, subtema) |>
 rename(frequencia = n)
```

Para visualizar, escolhi o gráfico de barras empilhadas para ver quantas observações foram feitas para cada ano do eixo x, qualificando a frequência por palavras-chave. Assim consigo ver quantas teses foram defendidas em cada ano, ao mesmo tempo em que consigo ver a frequência de cada subtema que considerei relevante para minha pesquisa. O gráfico em barras empilhadas é perfeito para visualizar a relação entre uma variável numérica e uma variável categórica.

```r
ggplot(ocorrencias_por_ano, aes(x = ano, y = frequencia, fill = subtema)) +
    geom_bar(stat = "identity", position = "stack") +
    labs(title = "  Produção de Defesas em Sociologia, por Palavra-Chave (1987-2022)",
         x = "
            Ano de defesa", y = "Número de defesas
        ") +
    scale_x_continuous(breaks = unique(ocorrencias_por_ano$ano)) +
    theme_classic() +
    theme(panel.grid.minor = element_blank())+
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    scale_fill_viridis_d(labels = c("Ensino Superior", "Desigualdade",
                                    "Educação", "Outros"),
                         name = element_blank())
```



4. Diferenças regionais

Para esta seção, foram calculados o total de trabalhos defendidos ao longo de todo o período por estado a fim de criar duas visualizações: em uma, é reportada a frequência de trabalhos por região; em outra, a frequência de trabalhos por unidade da federação.

Mapeando da frequência de trabalhos por Unidade da Federação usando geobr

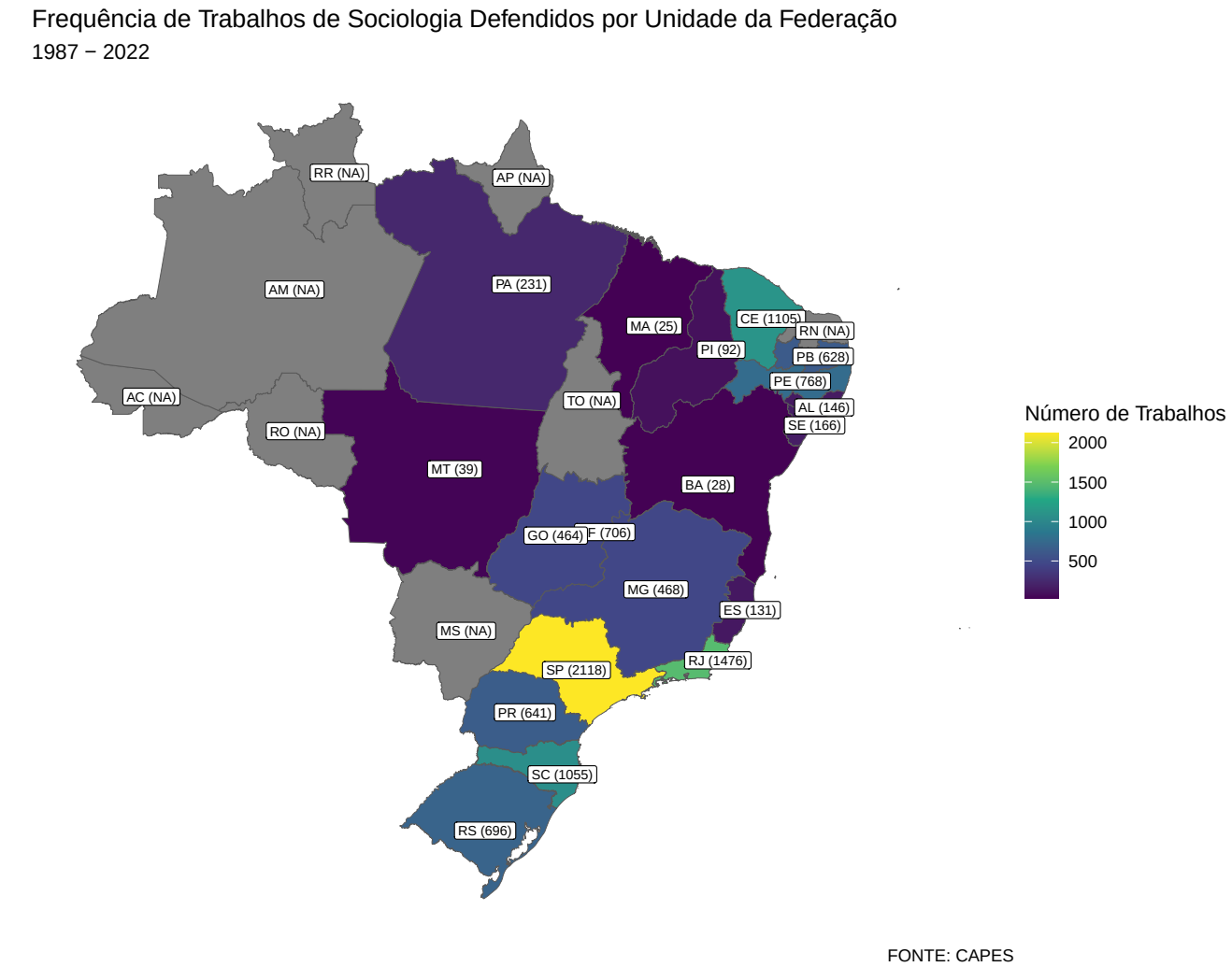
Usando as funções `read_state()` e `read_region()` do pacote `geobr` é possível obter as coordenadas de todos os estados e regiões do Brasil. Para evitar outputs indesejados no documento final, utilizei o atributo `showProgress = FALSE`.

```
coordenadas_estados <- read_state(showProgress = FALSE)
coordenadas_regioes <- read_region(showProgress = FALSE)
```

Em seguida, para garantir uma boa visualização das frequências, as defesas foram agrupadas por estado. Depois foi somado o total de trabalhos por unidade da federação, e esses dados foram unidos com as coordenadas correspondentes, extraídas do pacote `geobr`.

```
defesas_por_estado <- sociologia |>
  group_by(UF) |>
  summarise(total_trabalhos_uf = n()) |>
  full_join(coordenadas_estados, by = c("UF" = "abbrev_state")) |>
  st_as_sf()
```

Mapa representando a frequência de defesas por estado



É possível observar que São Paulo (SP) é o estado que mais produz trabalhos na área da sociologia atuantes nas pesquisas sobre educação, desigualdades e ensino superior.Seguido de Rio de Janeiro (RJ) e Ceará (CE), com 2118, 1476 e 1105 produções de pós-graduandos, respectivamente.

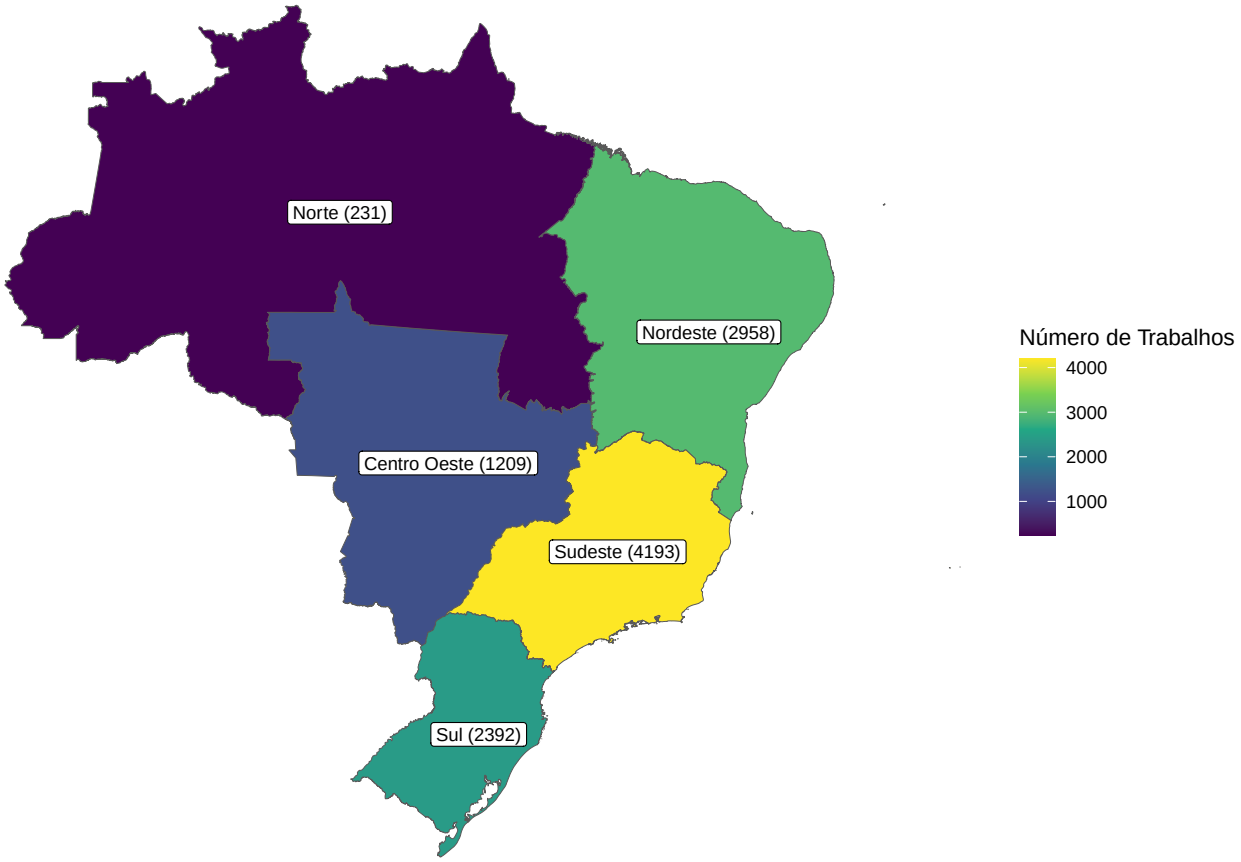
Mapa da frequência de trabalhos por região

Utilizando a mesma lógica construída no mapa anterior, o mapa abaixo resume o número de trabalhos defendidos entre as regiões do Brasil.

```
defesas_por_regiao <- sociologia |>
  group_by(regiao) |>
  summarise(total_trabalhos_regiao = n()) |>
  full_join(coordenadas_regioes, by = c("regiao" = "name_region")) |>
  st_as_sf()


ggplot(defesas_por_regiao, aes(fill = total_trabalhos_regiao, label =  paste(regiao, " (", total_
  geom_sf()+
  geom_sf_label(fill = "white", size = 3, nudge_x = 0.5)+
  scale_fill_viridis_c(name = "Número de Trabalhos")+
  ggtitle("Frequência de Trabalhos de Sociologia Defendidos por Região")+
  theme_void()+
  labs(subtitle = "1987 - 2022",
       caption = "FONTE: CAPES")
```

Frequência de Trabalhos de Sociologia Defendidos por Região
1987 – 2022



FONTE: CAPES

5. Produção por programa

O objetivo é obter uma tabela contendo o número de trabalhos defendidos pelos 10 programas com maior produção entre 1987 e 2022. Nesta seção, foi calculado o total de teses e dissertações defendidas por cada programa de pós-graduação em sociologia.

```
trabalhos_por_programa <- sociologia |>
  mutate(tese_ou_defesa = case_when(
    str_detect(nivel, "Mestrado|MESTRADO|MESTRADO PROFISSIONAL") ~ "dissertacao",
    str_detect(nivel, "Doutorado|DOUTORADO") ~ "tese",
  )) |>
  count(sigla_ies, nome_programa, tese_ou_defesa, CONCEITO) |>
  rename(trabalhos = n)

trabalhos_por_programa <- pivot_wider(trabalhos_por_programa, names_from = tese_ou_defesa, values
  mutate(tese = case_when(
    tese > 0  ~ tese,
```

```
    TRUE ~ 0
  )) |>
  mutate(total_defesas = dissertacao + tese) |>
  arrange(-total_defesas)|>
  slice(1:10)|>
  select(-total_defesas)
```

Transformei a coluna `tese_ou_defesa` em duas variáveis: `dissertacao` e `tese`, alongando a base de `trabalhos_por_programa`. Para o cálculo, utilizei a função `case_when()` para filtrar os níveis (mestrado-doutorado) e contabilizar o número de trabalhos totais.

A tabela final possui 4 informações importantes: nome do programa, nota CAPES, total de dissertações e total de teses defendidas no programa.

```
trabalhos_por_programa |>
  gt() |>
  tab_header(title = "Programas com o maior número de trabalhos em sociologia 1987-2022") |>
  tab_source_note(source_note = "Fonte: CAPES.") |>
  cols_label(sigla_ies = md("**Sigla**"), nome_programa = md("**Nome do Programa**"), CONCEITO =
```

Programas com o maior número de trabalhos em sociologia 1987-2022

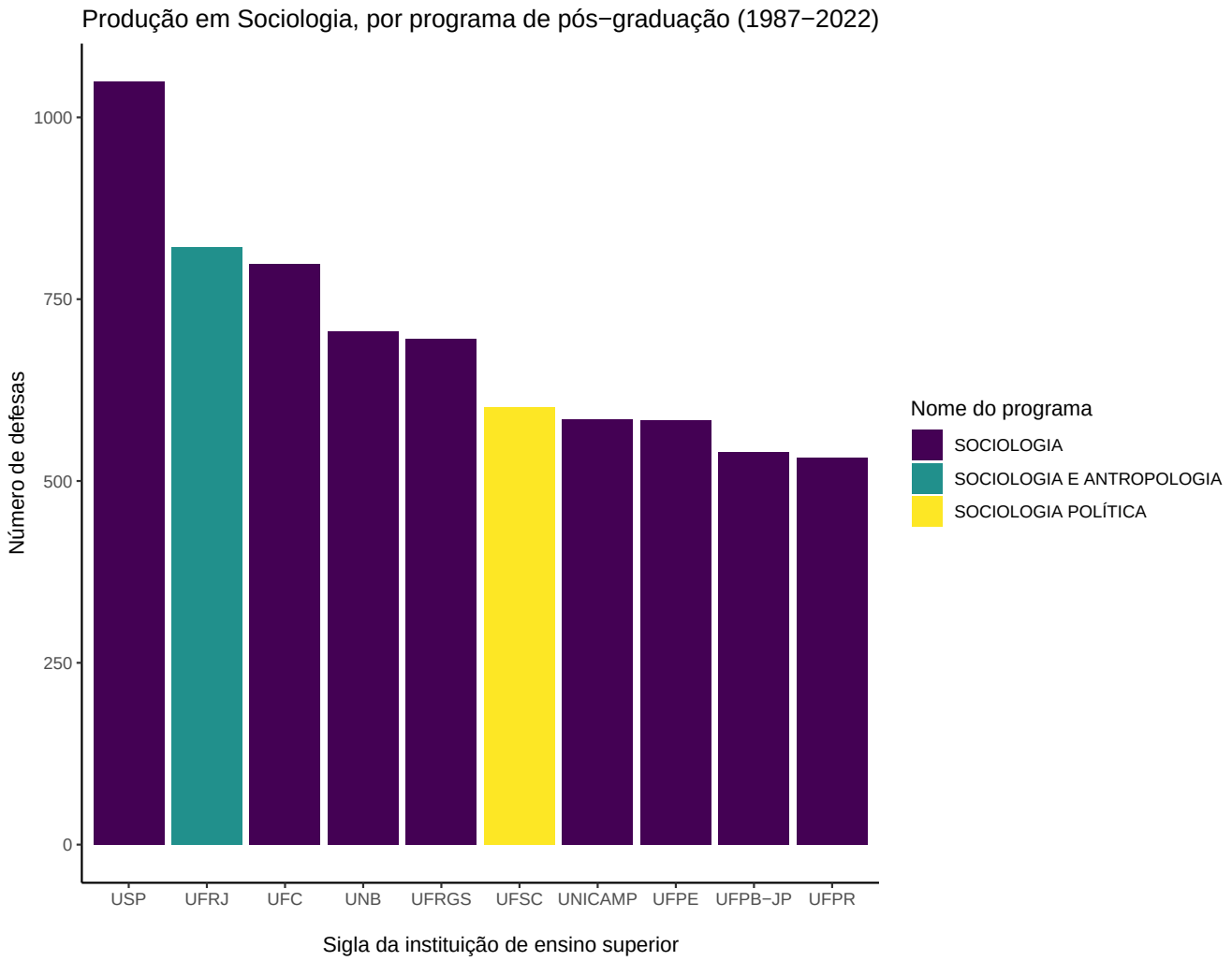
Sigla	Nome do Programa	Nota CAPES	Dissertações	Teses
USP	SOCIOLOGIA	6	489	560
UFRJ	SOCIOLOGIA E ANTROPOLOGIA	7	530	292
UFC	SOCIOLOGIA	5	533	265
UNB	SOCIOLOGIA	7	378	328
UFRGS	SOCIOLOGIA	7	451	245
UFSC	SOCIOLOGIA POLÍTICA	5	448	153
UNICAMP	SOCIOLOGIA	6	434	151
UFPE	SOCIOLOGIA	5	356	228
UFPB-JP	SOCIOLOGIA	5	379	161
UFPR	SOCIOLOGIA	5	363	169

Fonte: CAPES.

Visualização em barras dos programas com maior número de trabalhos em sociologia

```
ggplot(trabalhos_por_programa, aes(x =reorder(sigla_ies, -(tese+dissertacao)),
                                   y=  tese + dissertacao, fill = nome_programa)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Produção em Sociologia, por programa de pós-graduação (1987-2022)",
       x = "
```

```
      Sigla da instituição de ensino superior", y = "Número de defesas") +
  theme_classic() +
  theme(panel.grid.minor = element_blank())+
  scale_fill_viridis_d(name = "Nome do programa")
```



6. Exportação

Nesta última etapa, foi criada uma base menor contendo apenas as seguintes variáveis: ano, estado, programa, título, resumo e autor(a). Essa base está exportada em uma planilha `.csv`, para consultas futuras.

```
resumo <- trabalhos_relevantes |>
  select(ano, UF, nome_programa, titulo, resumo, autor)
```

```
# write_csv(resumo, "resumo.csv")
```

Conclusão

A experiência de fazer essa lista foi incrível. Pela primeira vez me senti pesquisador durante o mestrado. Essa lista me fez entender o fluxo de limpeza de bases de dados, me fez aprender a aprender, me trouxe também algumas noites sem dormir tentando achar formas de aprimorar o código e a documentação do mesmo.Em resumo, a jornada de elaboração da lista foi desafiadora e estimulante. Ao longo do processo, aprimorei minhas habilidades em análise de dados, além de reafirmar minha paixão por Sociologia e Programação.

Enfrentei desafios técnicos e pessoais, dediquei noites ao trabalho `árduo`, mas essa experiência desafiadora com certeza reforçou meu compromisso com a pesquisa e me deixou ansioso para continuar seguindo por esse caminho. Quero escrever todos os meus trabalhos em `.qmd` a partir de agora.