Relatório Final

Code

Análise de Bolsas e Auxílios Pagos pelo CNPq

Ennoile Raquel, Felipe William e Reyner Alegria

21 de julho de 2025

1 1. Introdução

Este relatório apresenta uma análise exploratória, detecção de outliers, estatística descritiva e mapas dos dados de Bolsas e Auxílios Pagos pelo CNPq. O foco da análise está nas principais linhas de fomento, áreas de conhecimento e na distribuição geográfica por Unidade da Federação (UF) de origem. Exploraremos também a evolução temporal dos investimentos e uma análise mais granular para a região do Amazonas.

2 2. Metodologia

2.1 2.1. Obtenção dos Dados

Os dados foram extraídos da plataforma Google BigQuery utilizando o pacote `basedosdados`. O código a seguir realiza a autenticação e o download. Para otimizar a execução, o download só é realizado se os arquivos de dados ainda não existirem localmente no diretório `dados/`.

Hide

```
# Este chunk deve ser executado manualmente uma vez para baixar os dados.
# A opção eval=FALSE impede que ele rode toda vez que o relatório for gerado.

# 1. Autenticação (pode pedir para logar no navegador)
bq_auth()

# 2. Defina seu ID de projeto do Google Cloud para cobrança
basedosdados::set_billing_id("projetofinalgr03")

# 3. Cria o diretório para os dados
if (!dir.exists("dados")) dir.create("dados")

# 4. Query para buscar os dados
query <- "
WITH
dicionario_linha_fomento AS (
  SELECT
    chave AS chave_linha_fomento,
    valor AS descricao_linha_fomento
  FROM `basedosdados.br_cnpq_bolsas.dicionario`
  WHERE nome_coluna = 'linha_fomento'
    AND id_tabela = 'microdados'
)
SELECT
  dados.ano AS ano,
  dados.processo AS processo,
  dados.data_inicio_processo,
  dados.data_fim_processo,
  dados.titulo_projeto,
  d.descricao_linha_fomento AS linha_fomento,
  dados.area_conhecimento,
  dados.sigla_uf_origem,
  dados.instituicao_origem,
  dados.sigla_uf_destino,
  dados.sigla_instituicao_destino,
  dados.valor
FROM `basedosdados.br_cnpq_bolsas.microdados` AS dados
LEFT JOIN dicionario_linha_fomento d ON dados.linha_fomento = d.chave_linha_fomento
"

# 5. Executa a query e salva os resultados
dados_cnpq <- basedosdados::read_sql(query,
                                     billing_project_id = basedosdados::get_billing_id()) %>%
  write_csv("dados/cnpq.csv")

# 6. Salva os dados geoespaciais dos estados
estados <- geobr::read_state(year = 2020)
saveRDS(estados, "dados/estados.rds")
```

2.2 2.2. Leitura e Tratamento dos Dados

Após o download, os dados são carregados e pré-processados. O tratamento consiste em substituir valores ausentes (NA) nas colunas categóricas por “Não informado” e remover registros que não possuem valor financeiro. Além disso, adicionamos uma coluna para o código do município de origem, essencial para análises geográficas mais detalhadas. É crucial que o mapeamento das instituições para os códigos de município seja preciso para o funcionamento do mapa.

Hide

```
# CARREGAMENTO DOS DADOS - ESSAS LINHAS FORAM INSERIDAS/CORRIGIDAS AQUI
dados <- read_csv("dados/cnpq.csv")
estados <- readRDS("dados/estados.rds")

dados <- dados %>%
  mutate(
    # Exemplo: Mapeia instituições a códigos de município. MUITO SIMPLIFICADO!
    # Na vida real, você precisaria de um dicionário ou geocodificação.
    code_municipio_origem = case_when(
      grepl("UNIVERSIDADE FEDERAL DO AMAZONAS", instituicao_origem, ignore.case = TRUE) ~ 1302603, # Manaus
      grepl("INSTITUTO FEDERAL DE EDUCACAO CIENCIA E TECNOLOGIA DO AMAZONAS - CAMPUS ITACOATIARA", instituicao_or
         igem, ignore.case = TRUE) ~ 1301902, # Itacoatiara
      grepl("UNIVERSIDADE DO ESTADO DO AMAZONAS - CAMPUS PARINTINS", instituicao_origem, ignore.case = TRUE) ~ 13
         03403, # Parintins
      TRUE ~ NA_real_ # Deixa NA para outros
    )
  ) %>%
  mutate(
    linha_fomento     = replace_na(linha_fomento,     "Não informado"),
    area_conhecimento = replace_na(area_conhecimento, "Não informado")
  ) %>%
  filter(!is.na(valor))

# Verificação do percentual de NAs por coluna após o tratamento
# round(colSums(is.na(dados)) / nrow(dados) * 100, 2)
```

3 3. Análise Exploratória e Descritiva

3.1 3.1. Visão Geral

O conjunto de dados, após o tratamento, possui as seguintes dimensões e características:

```
## Total de registros: 2.838.785
```

```
## Total de colunas: 13
```

Hide

```
skim(dados)
```

Data summary

Name	dados
Number of rows	2838785
Number of columns	13
_____	
Column type frequency:	
character	8
Date	2
numeric	3
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
processo	0	1.00	13	13	0	1342021	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
titulo_projeto	2451295	0.14	1	241	0	126945	0
linha_fomento	0	1.00	7	50	0	21	0
area_conhecimento	0	1.00	5	48	0	135	0
sigla_uf_origem	258749	0.91	2	2	0	27	0
instituicao_origem	228537	0.92	3	75	0	9391	0
sigla_uf_destino	99092	0.97	2	8	0	28	0
sigla_instituicao_destino	12357	1.00	1	16	0	7008	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
data_inicio_processo	2611528	0.08	2012-06-01	2024-03-01	2022-03-16	463
data_fim_processo	2611528	0.08	2016-08-31	2027-06-30	2023-08-31	199

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ano	0	1	2014.57	5.03	2002.00	2011	2015	2019	2022	▂▃▆▆▇
valor	0	1	10074.62	39496.29	-241.51	1600	2800	12100	13588321	▇▁▁▁▁
code_municipio_origem	2825315	0	1302603.00	0.00	1302603.00	1302603	1302603	1302603	1302603	▁▁▇▁▁

3.2 3.2. Estatísticas do Valor das Bolsas

Abaixo, um resumo estatístico da variável `valor`.

Hide

```
describe(dados$valor)
```

	vars	n	mean	sd	median	trimmed	mad	min	max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
X1	1	2838785	10074.62	39496.29	2800	6074.309	3024.504	-241.51	13588321

1 row | 1-10 of 14 columns

As estatísticas também foram calculadas para cada linha de fomento:

Hide

```
dados %>%
  group_by(linha_fomento) %>%
  summarise(
    n        = n(),
    Média    = mean(valor, na.rm = TRUE),
    Mediana  = median(valor, na.rm = TRUE),
    'Desvio Padrão' = sd(valor, na.rm = TRUE)
  ) %>%
  arrange(desc(n)) %>%
  kable(
    caption = "Estatísticas do Valor das Bolsas por Linha de Fomento.",
    digits  = 2
  )
```

Estatísticas do Valor das Bolsas por Linha de Fomento.

linha_fomento	n	Média	Mediana	Desvio Padrão

linha_fomento	n	Média	Mediana	Desvio Padrão
Não informado	2377350	10818.13	2800.00	36200.17
Bolsas de Iniciação Científica	178340	1518.82	1200.00	1100.81
Bolsas de Iniciação Científica Júnior	81883	361.52	200.00	310.12
Bolsas de Produtividade em Pesquisa e Tecnologia	60040	10313.26	11000.00	9582.59
Bolsas de Doutorado	51908	9224.29	4334.00	11791.24
Bolsas de Mestrado	29471	6211.07	3000.00	5875.67
Bolsas de Iniciação Tecnológica e Industrial	18125	2060.14	1600.00	998.95
Bolsas de Desenvolvimento Tecnológico e Industrial	15166	11700.24	8320.00	10349.63
Apoio a Projetos de Pesquisas	7865	99988.49	32020.00	391051.32
Bolsas de Extensão em Pesquisa	5105	20040.94	12000.00	18141.75
Bolsas de Apoio Técnico	4923	2194.77	1650.00	1591.46
Bolsas de Pós-doutorado	4188	24588.56	18000.00	24022.55
Bolsas de Fixação de Doutores	2574	14121.50	10500.00	11613.60
Bolsas de Pesquisador/Especialista Visitante	651	16923.59	12500.00	13555.19
Bolsas de Desenvolvimento Científico e Regional	538	27940.43	29400.00	21147.23
Apoio a Participação/Realização de Eventos	343	42723.63	40000.00	23639.66
Apoio a Periódicos Científicos	152	18113.82	15125.00	9181.45
Indefinido	93	3797.20	4000.00	3453.54
Estágio	37	40705.25	38274.99	26459.03
Bolsas de Graduação	24	26347.42	16104.74	32786.12
Bolsas no Exterior	9	10371.82	8307.93	4095.41

3.3 3.3. Detecção de Outliers

Valores atípicos (outliers) na variável `valor` foram identificados usando o critério do Intervalo Interquartil (IQR). A tabela interativa a seguir mostra os registros classificados como outliers.

Hide

```
q1  <- quantile(dados$valor, 0.25, na.rm = TRUE)
q3  <- quantile(dados$valor, 0.75, na.rm = TRUE)
iqr <- q3 - q1

lim_inf <- q1 - 1.5 * iqr
lim_sup <- q3 + 1.5 * iqr

dados <- dados %>%
  mutate(outlier = case_when(
    valor < lim_inf ~ "Inferior",
    valor > lim_sup ~ "Superior",
    TRUE            ~ "Normal"
  ))

datatable(
  dados %>% filter(outlier != "Normal") %>%
    select(ano, sigla_uf_origem, titulo_projeto, valor, outlier),
  options = list(pageLength = 10),
  caption = "Tabela: Outliers Inferiores e Superiores nos Valores das Bolsas."
)
```

Show 10 **entries** **Search:**

Tabela: Outliers Inferiores e Superiores nos Valores das Bolsas.

	ano	sigla_uf_origem	titulo_projeto	valor	outlier
1	2022	SP	IV ENCONTRO INTERNACIONAL DE PESQUISA EM ENFERMAGEM - 80 ANOS DA ESCOLA DE ENFERMAGEM DA UNIVERSIDADE DE SÃO PAULO: PROTAGONISMO NA PESQUISA	70000	Superior
2	2022	SP	APOIO FINANCEIRO PARA REALIZAÇÃO DO 67º CONGRESSO BRASILEIRO DE CERÂMICA, FLORIANÓPOLIS, SC	30000	Superior
3	2022	RJ	ORGANIZAÇÃO E LOGÍSTICA DO XXIX ENCONTRO NACIONAL DE TRATAMENTO DE MINÉRIOS E METALURGIA EXTRATIVA – ENTMME	33000	Superior
4	2022	PB	V COLÓQUIO DE MATEMÁTICA DA REGIÃO NORDESTE	30000	Superior
5	2022	PB	12ª REUNIÃO ANUAL DO INSTITUTO BRASILEIRO DE NEUROPSICOLOGIA E COMPORTAMENTO (IBNEC)	31000	Superior
6	2021	RJ	XXXII CONGRESSO BRASILEIRO DE VIROLOGIA & XVI ENCONTRO DE VIROLOGIA DO MERCOSUL	30000	Superior
7	2022	PR	IX CONGRESSO BRASILEIRO DE METABOLISMO, NUTRIÇÃO E EXERCÍCIO	80000	Superior
8	2022	PR	IV ENCONTRO NACIONAL DE CENTROS E MUSEUS DE CIÊNCIAS - MARINGÁ - PR	150000	Superior
9	2022	MA	PROJETO DE APOIO À REALIZAÇÃO DO 31O. ENCONTRO ANUAL DA COMPÓS - 2022 (ONLINE)	70000	Superior
10	2022	MG	ENCONTRO DE OUTONO DA SOCIEDADE BRASILEIRA DE FÍSICA - EOSBF 2023	90000	Superior

Showing 1 to 10 of 232,279 entries Previous 1 2 3 4 5 ... 23,228 Next

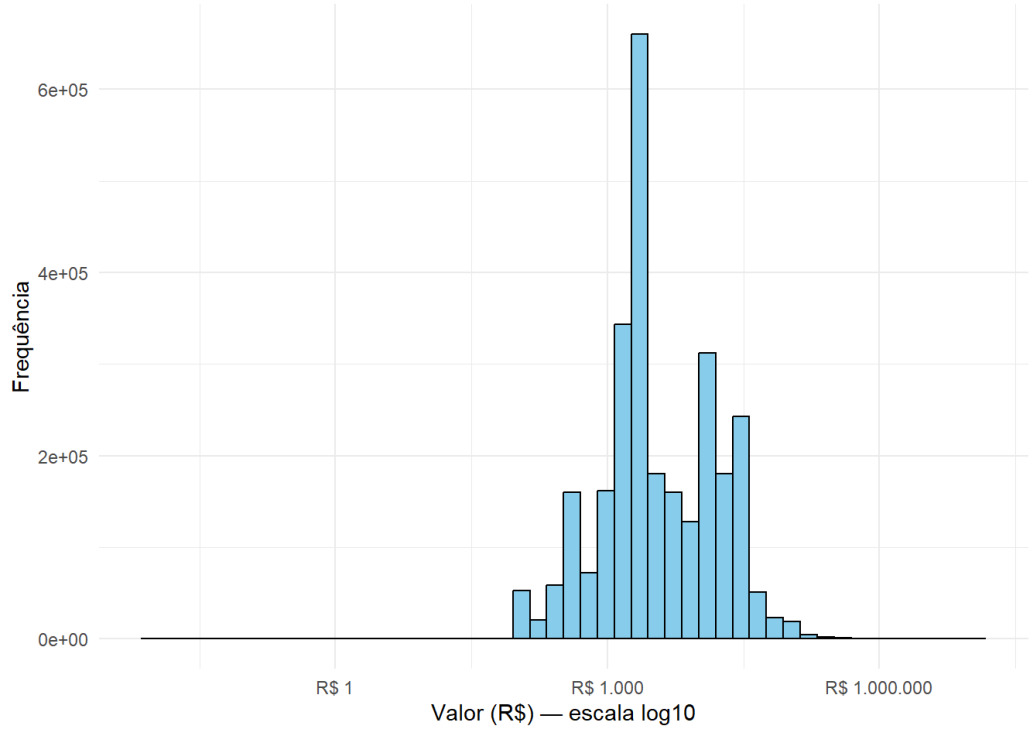
4 4. Visualizações

4.1 4.1. Distribuição dos Valores

O histograma a seguir exibe a distribuição dos valores das bolsas. Devido à grande assimetria e presença de valores muito altos, foi aplicada uma escala logarítmica no eixo X para melhor visualização da concentração dos dados.

Hide

```
library(scales)
ggplot(dados, aes(x = valor)) +
  geom_histogram(bins = 50, fill = "skyblue", color = "black") +
  scale_x_log10(
    labels = dollar_format(prefix = "R$ ", big.mark = ".", decimal.mark = ","),
    breaks = scales::log_breaks(n = 5)
  ) +
  labs(
    x = "Valor (R$) – escala log10",
    y = "Frequência"
  ) +
  theme_minimal()
```



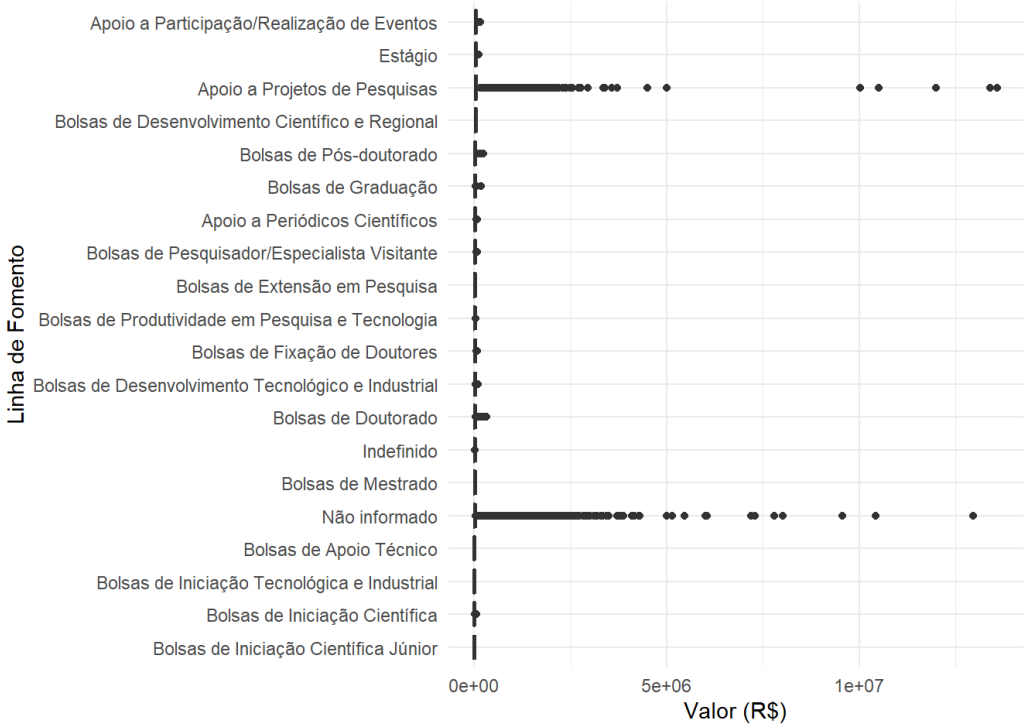
Distribuição dos Valores das Bolsas (R$).

O que fizemos: 1. Carregamos o pacote **scales** para formatar o eixo X em reais.
2. Usamos `scale_x_log10()` para comprimir a cauda longa e espalhar os valores menores.
3. Ajustamos os rótulos com `dollar_format()` e quebras logarítmicas legíveis com `log_breaks()`.

Agora você verá a forma real da distribuição, sem o “super-bar” dominando o gráfico.

Hide

```
# Filtra as 20 principais linhas para uma melhor visualização
top_linhas <- dados %>% count(linha_fomento, sort = TRUE) %>% slice_max(n, n = 20) %>% pull(linha_fomento)

dados %>%
  filter(linha_fomento %in% top_linhas) %>%
  ggplot(aes(x = reorder(linha_fomento, valor, FUN=median), y = valor)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = "Linha de Fomento", y = "Valor (R$)") +
  theme_minimal()
```

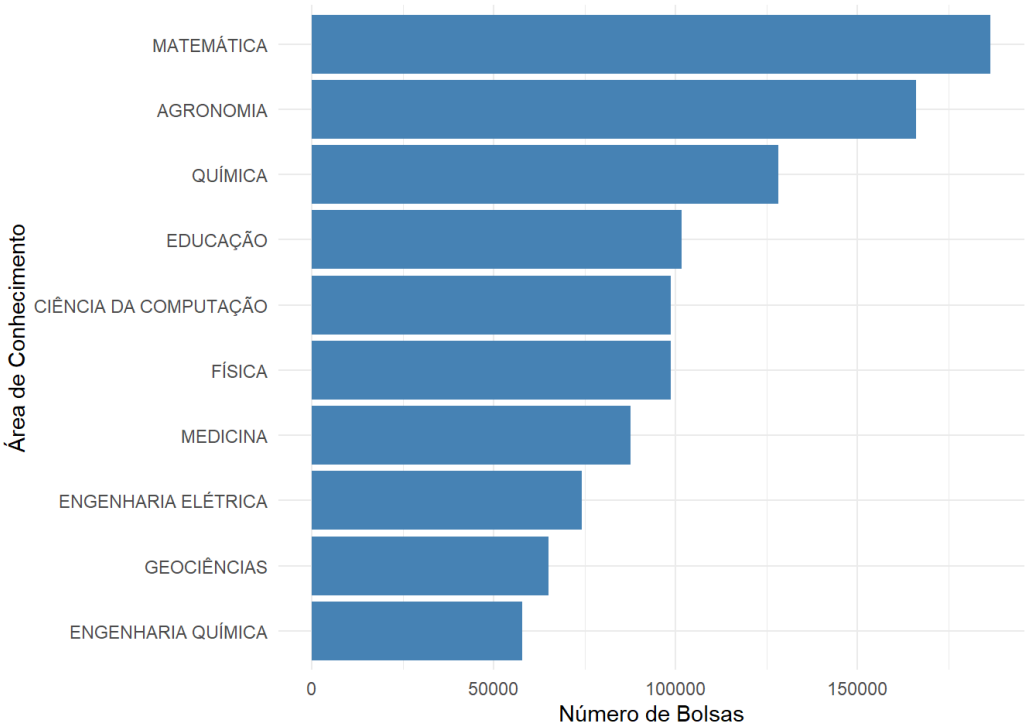


Dispersão do Valor das Bolsas por Linha de Fomento.

4.2 4.2. Principais Categorias

Hide

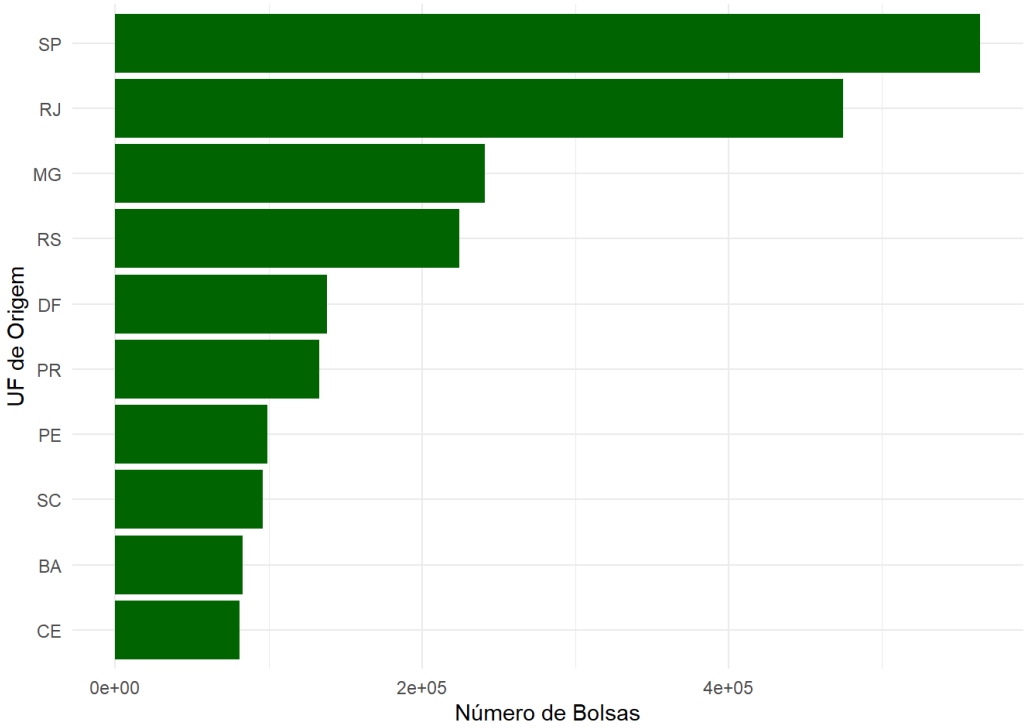
```
dados %>%
  filter(area_conhecimento != "Não informado") %>%
  count(area_conhecimento, sort = TRUE) %>%
  slice_max(n, n = 10) %>%
  ggplot(aes(x = reorder(area_conhecimento, n), y = n)) +
    geom_col(fill = "steelblue") +
    coord_flip() +
    labs(x = "Área de Conhecimento", y = "Número de Bolsas") +
    theme_minimal()
```



Top 10 Áreas de Conhecimento com maior número de bolsas.

Hide

```
dados %>%
  filter(!is.na(sigla_uf_origem)) %>%
  count(sigla_uf_origem, sort = TRUE) %>%
  slice_max(n, n = 10) %>%
  ggplot(aes(x = reorder(sigla_uf_origem, n), y = n)) +
    geom_col(fill = "darkgreen") +
    coord_flip() +
    labs(x = "UF de Origem", y = "Número de Bolsas") +
    theme_minimal()
```



Top 10 UFs de Origem com maior número de bolsistas.

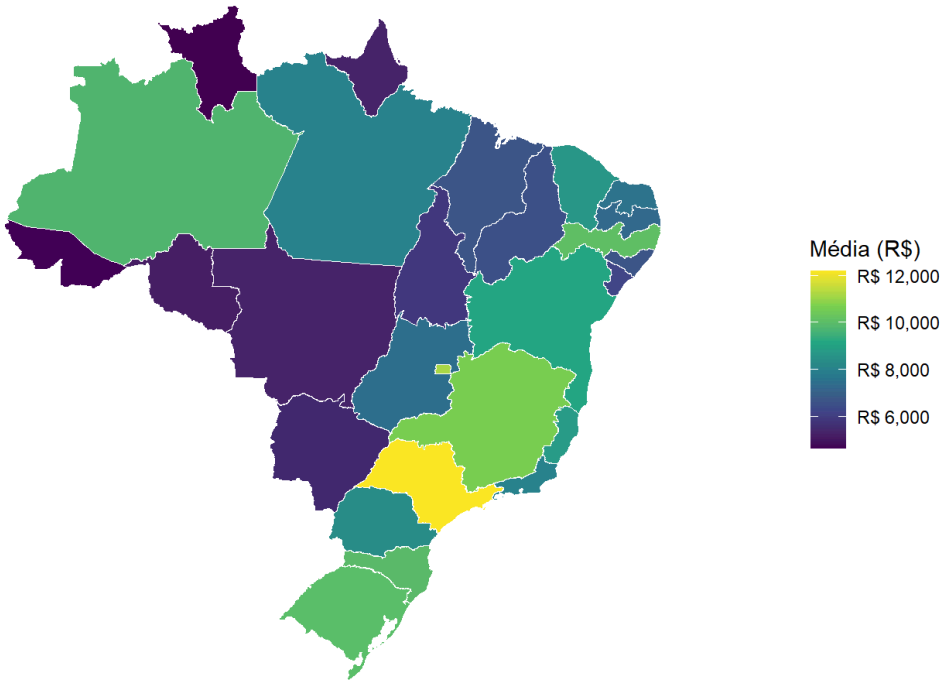
5 5. Análise Geográfica e Inferencial

5.1 5.1. Mapa Coroplético

O mapa abaixo ilustra o valor médio das bolsas distribuído pelas Unidades da Federação de origem dos pesquisadores.

Hide

```
uf_summary <- dados %>%
  filter(!is.na(sigla_uf_origem)) %>%
  group_by(sigla_uf_origem) %>%
  summarise(valor_medio = mean(valor, na.rm = TRUE))

map_data <- estados %>%
  left_join(uf_summary, by = c("abbrev_state" = "sigla_uf_origem"))

ggplot(map_data) +
  geom_sf(aes(fill = valor_medio), color = "white", size=0.1) +
  scale_fill_viridis_c(na.value = "grey80", labels = scales::dollar_format(prefix="R$ ")) +
  labs(fill = "Média (R$)") +
  theme_void()
```



Valor Médio das Bolsas por UF de Origem.

5.2 5.2. Teste de Hipótese (Exemplo)

Para ilustrar uma análise inferencial, realizamos um Teste t para comparar o valor médio das bolsas da Região Norte (representada por `AM`, Amazonas, por estar na localização atual) com o da Região Sudeste (`SP`, `RJ`, `MG`, `ES`). A hipótese nula (H0) é que não há diferença significativa entre as médias das duas regiões.

O Teste t resultou em uma estatística de **-2.396** com um p-valor de **0.0166**.

Conclusão do Teste: Como o p-valor (é menor que 0.05), nós rejeitamos a hipótese nula. Isso sugere que há uma diferença estatisticamente significativa entre o valor médio das bolsas concedidas para proponentes do Amazonas em comparação com os da Região Sudeste neste conjunto de dados.

6 6. Análises Adicionais e Aprofundamento

Para enriquecer a análise, exploramos a distribuição temporal das bolsas, o valor médio por área de conhecimento, a relação entre valor e duração do processo, e a proporção de linhas de fomento ao longo do tempo. Além disso, uma análise geográfica mais granular foi realizada para o estado do Amazonas.

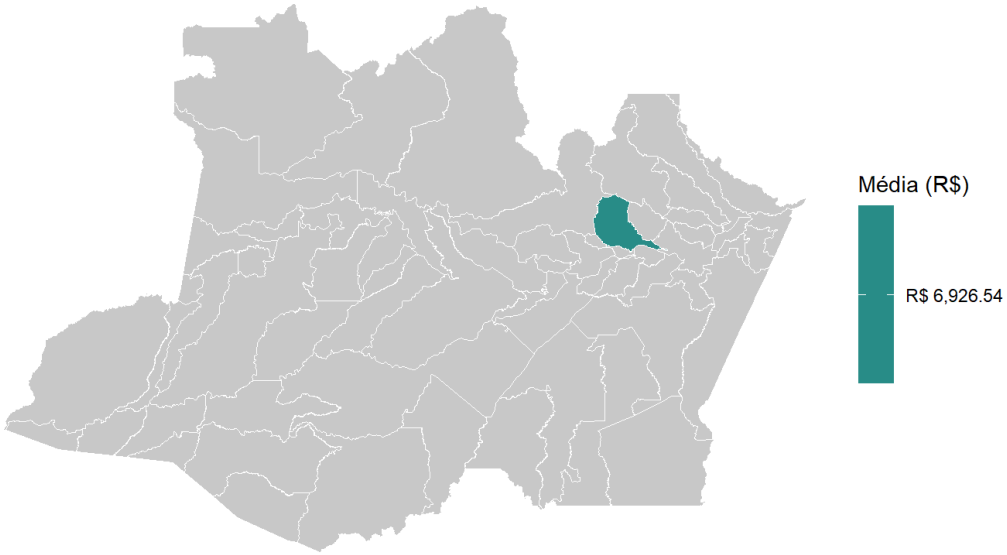
6.1 6.1. Análise Geográfica no Amazonas (por Município)

Focando no estado do Amazonas, os mapas abaixo ilustram a distribuição do valor médio das bolsas e do número de bolsas por município de origem, permitindo uma visão mais detalhada da distribuição do investimento do CNPq na região. É crucial que a coluna `code_municipio_origem` em seus dados esteja preenchida corretamente para que esses mapas funcionem. Se ela estiver vazia ou incorreta, os mapas não exibirão dados.

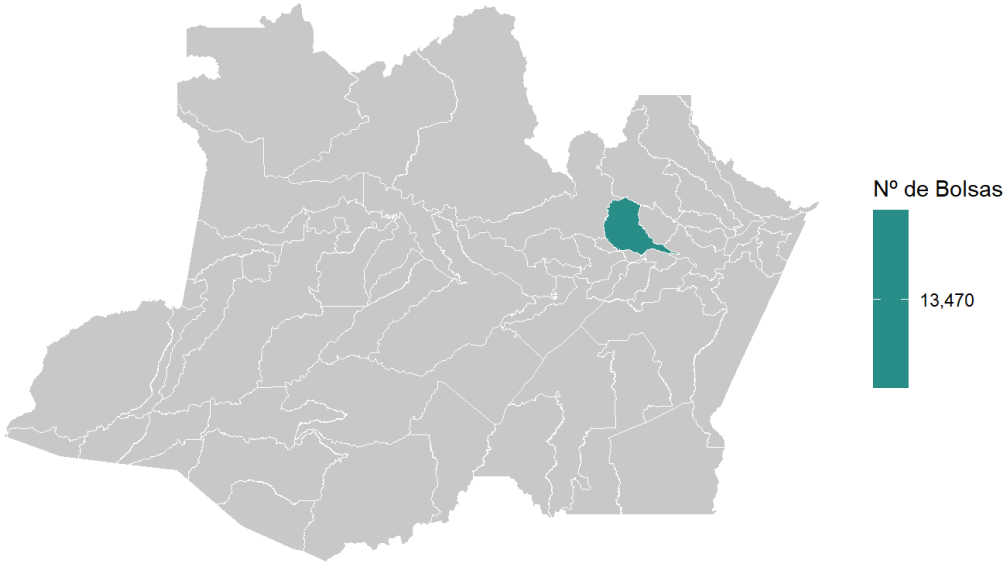
Hide

```
cnpq_am <- dados %>%
  filter(sigla_uf_origem == "AM")

municipios_am <- tryCatch({
  geobr::read_municipality(code_muni = 13, year = 2020)
}, error = function(e) {
  message("Erro ao baixar dados de municípios do Amazonas. Verifique conexão ou versão do geobr.")
  NULL
})

if (!is.null(municipios_am)) {
  uf_am_summary_municipio <- cnpq_am %>%
    filter(!is.na(code_municipio_origem)) %>%
    group_by(code_municipio_origem) %>%
    summarise(valor_medio = mean(valor, na.rm = TRUE),
              num_bolsas = n())

  if (nrow(uf_am_summary_municipio) == 0) {
    cat("AVISO: Nenhuma bolsa com 'code_municipio_origem' válido encontrada para o Amazonas. Os mapas não serão g
        erados. Por favor, verifique a coluna `code_municipio_origem` na sua base de dados.\n")
  } else {
    map_data_am_municipio <- municipios_am %>%
      left_join(uf_am_summary_municipio, by = c("code_muni" = "code_municipio_origem"))

    # Mapa do Valor Médio por Município no AM
    print(ggplot(map_data_am_municipio) +
      geom_sf(aes(fill = valor_medio), color = "white", size=0.1) +
      scale_fill_viridis_c(na.value = "grey80", labels = scales::dollar_format(prefix="R$ ")) +
      labs(
        title = "Valor Médio das Bolsas por Município de Origem no Amazonas",
        fill  = "Média (R$)"
      ) +
      theme_void())

    cat("\n\n") # Espaço entre os gráficos

    # Mapa do Número de Bolsas por Município no AM
    print(ggplot(map_data_am_municipio) +
      geom_sf(aes(fill = num_bolsas), color = "white", size=0.1) +
      scale_fill_viridis_c(na.value = "grey80", labels = scales::comma) +
      labs(
        title = "Número de Bolsas por Município de Origem no Amazonas",
        fill  = "Nº de Bolsas"
      ) +
      theme_void())
  }
} else {
  cat("Não foi possível gerar mapas de município no AM devido a problemas com o download dos dados geoespaciais.
       Verifique sua conexão com a internet ou a instalação do pacote 'geobr'.\n")
}
```

Valor Médio das Bolsas por Município de Origem no Amazonas



Número de Bolsas por Município de Origem no Amazonas

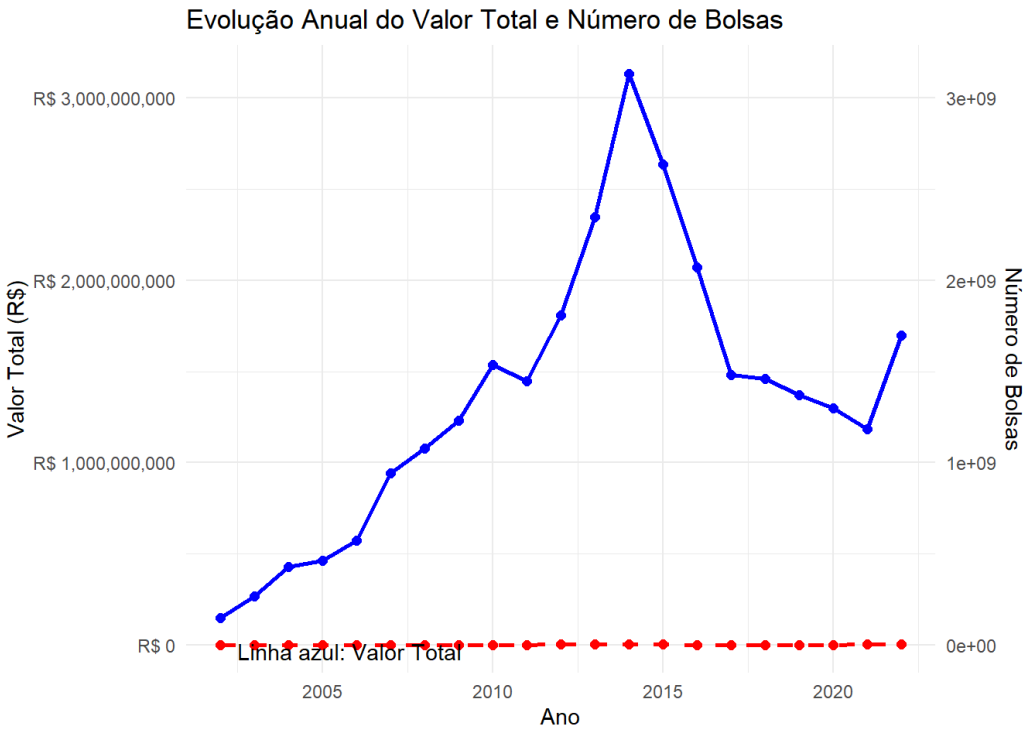


6.2 6.2. Distribuição das Bolsas ao Longo do Tempo.

Este gráfico permite visualizar a evolução do valor total e do número de bolsas concedidas pelo CNPq ano a ano.

Hide

```
dados %>%
  group_by(ano) %>%
  summarise(total_valor = sum(valor, na.rm = TRUE),
            num_bolsas = n()) %>%
  ggplot(aes(x = ano)) +
  geom_line(aes(y = total_valor), color = "blue", size = 1) +
  geom_point(aes(y = total_valor), color = "blue", size = 2) +
  geom_line(aes(y = num_bolsas), color = "red", linetype = "dashed", size = 1) +
  geom_point(aes(y = num_bolsas), color = "red", size = 2) +
  scale_y_continuous(
    name = "Valor Total (R$)",
    labels = scales::dollar_format(prefix = "R$ "),
    sec.axis = sec_axis(~., name = "Número de Bolsas")
  ) +
  labs(
    title = "Evolução Anual do Valor Total e Número de Bolsas",
    x = "Ano"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom") +
  annotate("text", x = min(dados$ano) + 0.5, y = max(dados$valor, na.rm = TRUE) * 0.9,
           label = "Linha azul: Valor Total\nLinha vermelha: Número de Bolsas",
           hjust = 0, vjust = 1, color = "black")
```

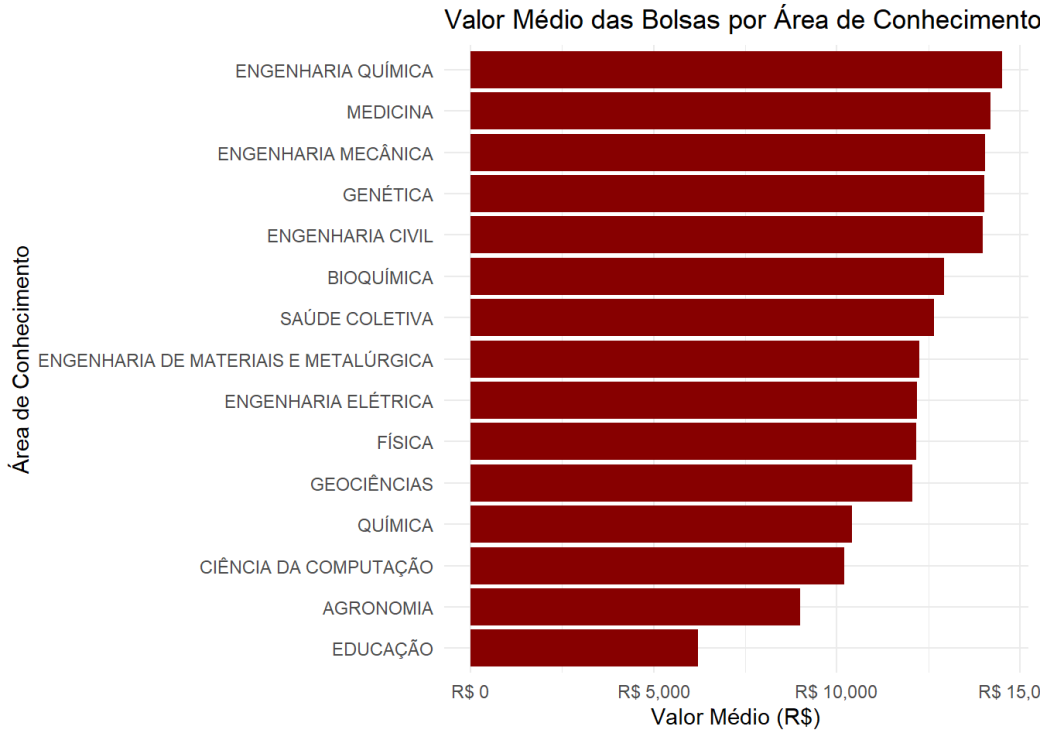


6.3 6.3. Valor Médio por Área de Conhecimento.

Analisamos o valor médio das bolsas para as 15 áreas de conhecimento com o maior investimento total, revelando onde os recursos financeiros são mais concentrados por bolsa.

Hide

```
top_areas_valor <- dados %>%
  filter(area_conhecimento != "Não informado") %>%
  group_by(area_conhecimento) %>%
  summarise(valor_total = sum(valor, na.rm = TRUE)) %>%
  arrange(desc(valor_total)) %>%
  slice_head(n = 15) %>%
  pull(area_conhecimento)

dados %>%
  filter(area_conhecimento %in% top_areas_valor) %>%
  group_by(area_conhecimento) %>%
  summarise(valor_medio = mean(valor, na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(area_conhecimento, valor_medio), y = valor_medio)) +
  geom_col(fill = "darkred") +
  coord_flip() +
  labs(
    title = "Valor Médio das Bolsas por Área de Conhecimento (Top 15 por Valor Total)",
    x = "Área de Conhecimento",
    y = "Valor Médio (R$)"
  ) +
  scale_y_continuous(labels = scales::dollar_format(prefix = "R$ ")) +
  theme_minimal()
```

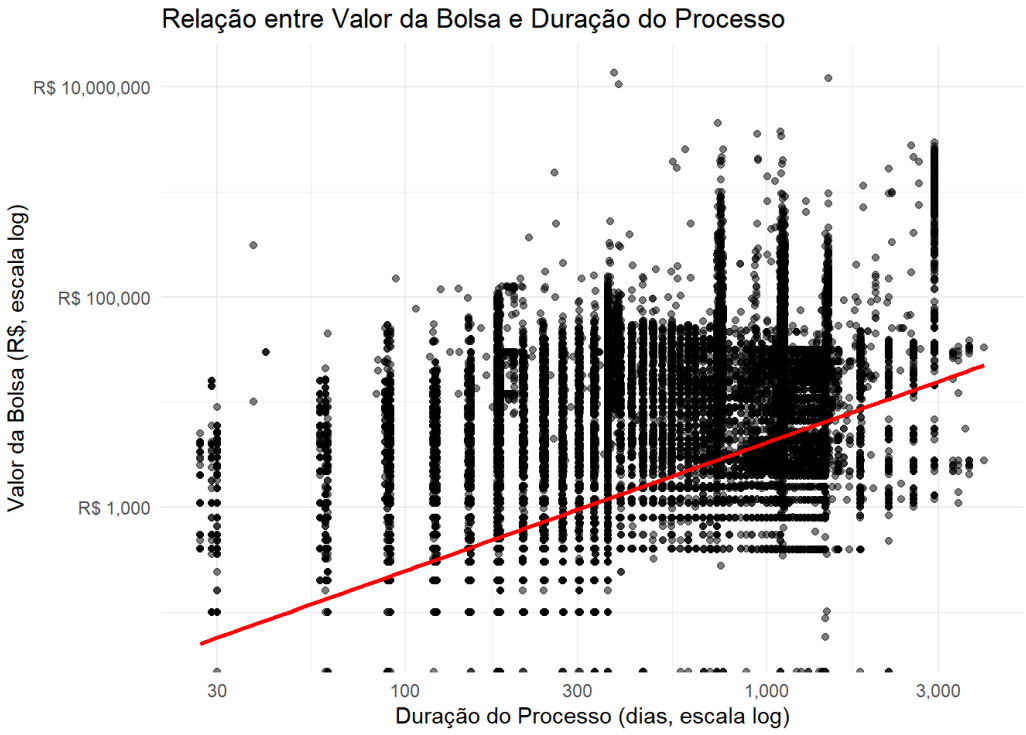


6.4 6.4. Relação entre Valor da Bolsa e Duração do Processo.

Este gráfico de dispersão com uma linha de regressão auxilia a entender se existe uma correlação entre o valor financeiro da bolsa e a sua duração em dias.

Hide

```
dados_com_duracao <- dados %>%
  mutate(
    data_inicio_processo = as.Date(data_inicio_processo),
    data_fim_processo = as.Date(data_fim_processo),
    duracao_dias = as.numeric(data_fim_processo - data_inicio_processo)
  ) %>%
  filter(!is.na(duracao_dias), duracao_dias > 0)

ggplot(dados_com_duracao, aes(x = duracao_dias, y = valor)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  scale_x_log10(labels = scales::comma) +
  scale_y_log10(labels = scales::dollar_format(prefix = "R$ ")) +
  labs(
    title = "Relação entre Valor da Bolsa e Duração do Processo",
    x = "Duração do Processo (dias, escala log)",
    y = "Valor da Bolsa (R$, escala log)"
  ) +
  theme_minimal()
```

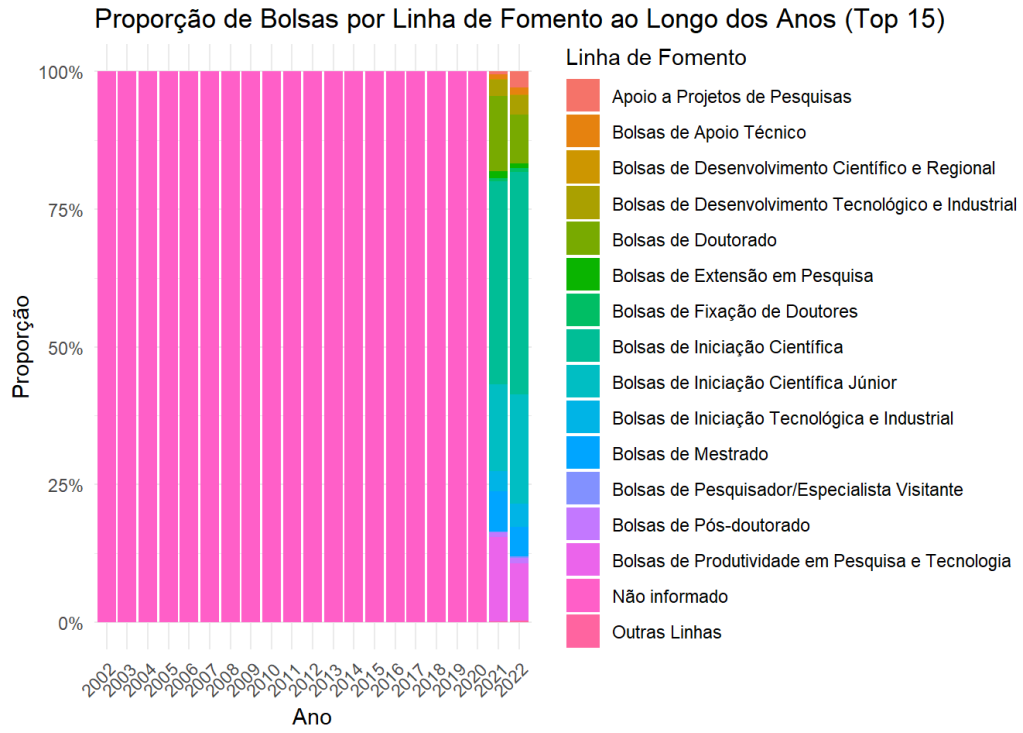


6.5 6.5. Proporção de Linhas de Fomento ao Longo dos Anos.

Este gráfico de barras empilhadas mostra como a participação de cada linha de fomento mudou percentualmente ao longo dos anos, indicando possíveis shifts nas prioridades de investimento do CNPq.

Hide

```
# Definir um limite para o número de linhas a serem exibidas (ex: Top 15)
top_n_linhas_fomento <- 15

# Calcular as principais linhas de fomento por número total de bolsas
principais_linhas <- dados %>%
  count(linha_fomento, sort = TRUE) %>%
  slice_head(n = top_n_linhas_fomento) %>%
  pull(linha_fomento)

dados %>%
  # Agrupar linhas de fomento que não estão nas principais como "Outras Linhas"
  mutate(linha_fomento_agrupada = ifelse(linha_fomento %in% principais_linhas,
                                         linha_fomento, "Outras Linhas")) %>%
  group_by(ano, linha_fomento_agrupada) %>%
  summarise(count = n(), .groups = 'drop') %>%
  group_by(ano) %>%
  mutate(proportion = count / sum(count)) %>%
  ggplot(aes(x = factor(ano), y = proportion, fill = linha_fomento_agrupada)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(
    title = "Proporção de Bolsas por Linha de Fomento ao Longo dos Anos (Top 15)",
    x = "Ano",
    y = "Proporção",
    fill = "Linha de Fomento"
  ) +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



6.6 6.6. Mapa com Instituições do Amazonas

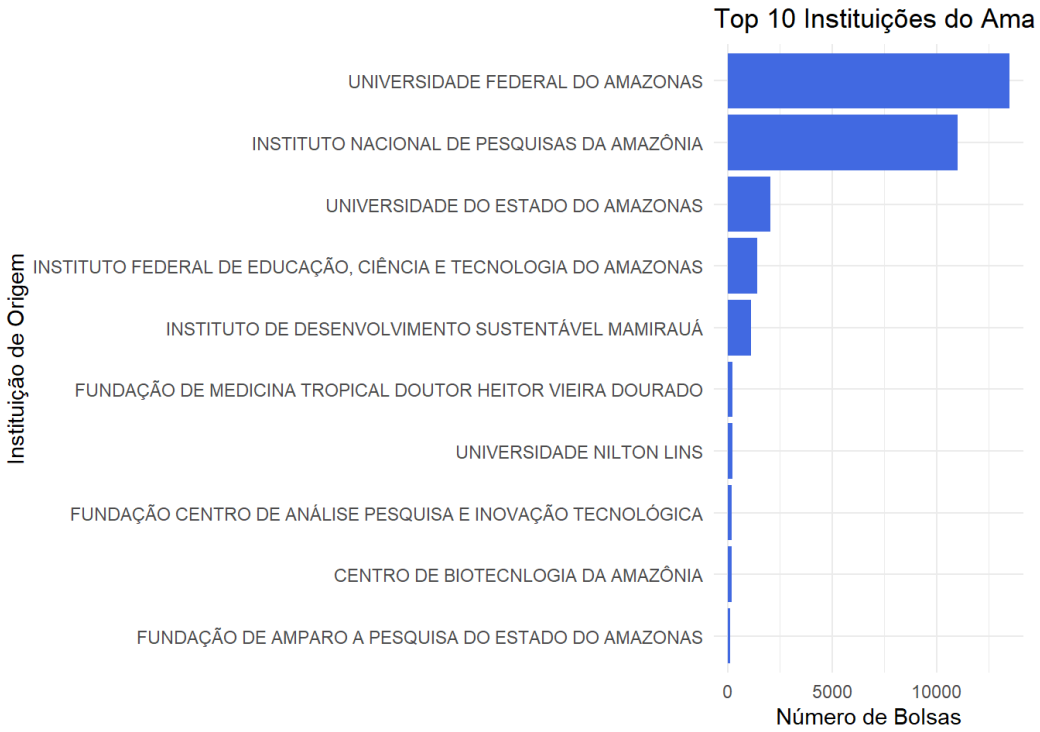
Hide

```r
# Instalar pacotes caso ainda não estejam disponíveis
# install.packages("geobr")
# install.packages("ggrepel")

library(geobr)
library(sf)
library(ggplot2)
library(ggrepel)

# Carrega os municípios do Amazonas (UF 13)
municipios_am <- geobr::read_municipality(code_muni = 13, year = 2020)

# Coordenadas geográficas das principais instituições (exemplos)
instituicoes_am <- data.frame(
  instituicao = c(
    "UNIVERSIDADE FEDERAL DO AMAZONAS",
    "INSTITUTO FEDERAL DO AMAZONAS - CAMPUS ITACOATIARA",
    "UNIVERSIDADE DO ESTADO DO AMAZONAS - CAMPUS PARINTINS"
  ),
  lon = c(-60.025, -58.449, -56.735),
  lat = c(-3.101, -2.763, -2.637)
)

# Plotando o mapa
ggplot() +
  geom_sf(data = municipios_am, fill = "gray95", color = "gray70") +
  geom_point(data = instituicoes_am, aes(x = lon, y = lat), color = "blue", size = 3) +
  geom_text_repel(data = instituicoes_am, aes(x = lon, y = lat, label = instituicao), size = 3) +
  labs(
    title = "Localização das Principais Instituições no Amazonas com Bolsas CNPq",
    x = "Longitude", y = "Latitude"
  ) +
  theme_minimal()
```

6.7 6.7. Comparativo: Instituição que Mais Recebeu Bolsas no Amazonas

Hide

```
dados %>%
  filter(sigla_uf_origem == "AM") %>%
  count(instituicao_origem, name = "num_bolsas", sort = TRUE) %>%
  slice_max(num_bolsas, n = 10) %>%
  ggplot(aes(x = reorder(instituicao_origem, num_bolsas), y = num_bolsas)) +
  geom_col(fill = "royalblue") +
  coord_flip() +
  labs(
    title = "Top 10 Instituições do Amazonas que Mais Receberam Bolsas do CNPq",
    x = "Instituição de Origem",
    y = "Número de Bolsas"
  ) +
  theme_minimal()
```



6.8 7. Conclusão

Este relatório detalhou o processo de extração, tratamento e análise dos dados de Bolsas e Auxílios do CNPq. A análise exploratória e as visualizações destacaram a distribuição assimétrica dos valores, com concentração em montantes menores, e revelaram as principais linhas de fomento, áreas de conhecimento e a variação geográfica por UF. As novas análises aprofundaram a compreensão temporal, as relações entre variáveis (valor e duração), e a distribuição regional detalhada, especialmente para o Amazonas por município. O teste de hipótese exemplificou como análises inferenciais podem ser aplicadas. Como próximos passos, recomenda-se a aplicação de testes estatísticos mais robustos e modelos de regressão para investigar os determinantes do valor das bolsas, e um aprofundamento na geocodificação de instituições para refinar ainda mais os mapas por município.