UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

VITOR DE SOUZA CRUZEIRO

CIÊNCIA DE DADOS APLICADA AO DESENVOLVIMENTO DE FERRAMENTAS PARA ANÁLISE DE DADOS ELEITORAIS BRASILEIROS EM LINGUAGEM R

BAURU
Novembro/2019

VITOR DE SOUZA CRUZEIRO

CIÊNCIA DE DADOS APLICADA AO DESENVOLVIMENTO DE FERRAMENTAS PARA ANÁLISE DE DADOS ELEITORAIS BRASILEIROS EM LINGUAGEM R

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.
Orientador: Prof. Assoc. Dr. João Pedro Albino

BAURU
Novembro/2019

Vitor de Souza Cruzeiro Ciência de dados aplicada ao desenvolvimento de ferramentas para análise de dados eleitorais brasileiros em linguagem R/ Vitor de Souza Cruzeiro. – Bauru, Novembro/2019- 37 p. : il. (algumas color.) ; 30 cm.
Orientador: Prof. Assoc. Dr. João Pedro Albino
Trabalho de Conclusão de Curso – Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Bacharelado em Ciência da Computação, Novembro/2019.
1. Ciência de dados 2. Visualização de dados eleitorais 3. Dados demográficos 4. Linguagem R

Vitor de Souza Cruzeiro

Ciência de dados aplicada ao desenvolvimento de ferramentas para análise de dados eleitorais brasileiros em linguagem R

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Prof. Assoc. Dr. João Pedro Albino
Orientador
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Profa. Dra. Simone das Graças Domingues Prado
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Prof. Dr. Kelton Augusto Pontara da Costa
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Bauru, 3 de novembro de 2019.

Agradecimentos

Agradeço a Deus, a quem devo a vida. Agradeço, ainda, àqueles que estiveram comigo nos dias maus; aos amigos e familiares em cujos abraços e palavras encontrei conforto; aos que a graduação me permitiu conhecer, que a fizeram parecer mais um circo do que um peso; à Universidade e aos professores, por incentivar alguém como eu a ter esta experiência tão enriquecedora que é a busca pelo conhecimento.

Resumo

Com a popularização da internet, uma quantidade crescente de dados passou a ser produzida e disseminada de forma rápida e acessível. Enquanto este crescimento é benéfico para o acúmulo de dados, pode ser um obstáculo a verificação da acurácia destes. Ainda assim, existem fontes confiáveis que disponibilizam dados verificados com um certo nível de garantia de exatidão. Para diminuir a burocracia e facilitar o espalhamento dos dados, o Governo do Brasil vem utilizando, há alguns anos, a Internet como meio de divulgação de informações que têm sua disseminação obrigatória por lei. Por provirem diretamente do Estado, estes dados são tidos como confiáveis. A maioria dos órgãos das esferas governamentais está sujeita a esta lei, e não é diferente com o Tribunal Superior Eleitoral (TSE), que serve de administrador dos processos eleitorais no país. São objetos deste estudo os dados disponibilizados pelo TSE, que trazem informações sobre pleitos, resultados de eleições e dados demográficos dos eleitores. Assim, este projeto buscou desenvolver um conjunto de funções simples em linguagem R que permitisse a criação de mapas temáticos. O trabalho realizado gerou duas funções que criam mapas distintos oferecendo uma organização e visualização dos dados. Em função das limitações da base de dados do TSE, não foi possível utilizar uma maior gama de dados no projeto. Acredita-se que o trabalho cumpriu o objetivo inicial proposto de oferecer um panorama para demonstrar a evolução entre votações semelhantes ocorridas em períodos distintos.

Palavras-chave: Ciência de dados, visualização de dados eleitorais, dados demográficos, linguagem R.

Abstract

Throughout the years, internet usage has become more common, leading to mass production and distribution of data in a quick and accessible way. Although this has contributed to the increase of the amount of content available worldwide, it has also made the verification process more difficult. Nevertheless, there are reliable sources of information that provide verified data with some level of accuracy. A few years ago, in order to reduce bureaucracy and to help data propagation, the Government of Brazil started leveraging the Internet as a means of publishing public information (which it is required to do by law). Because the data comes directly from the government, it is deemed trustable. Most of the public institutions are subject to this information broadcasting law, and that includes the Tribunal Superior Eleitoral (TSE), which is Brazil's court responsible for the electoral proceedings nationwide. The data provided by this court includes electoral polls, results and demographics on electors, and it's the main source of data for this project. This project aimed at developing a set of simple tools that allow for data visualisation through the creation of maps using the R language. It has generated two functions that plot Brazil maps divided by state which contain data about elections. These maps offer a rich way to visualize and organize this electoral data. Because of some technical difficulties the TSE has been facing, not a lot of data was available to be used. It is believed that this work has achieved its goals by providing the tools necessary to analyse and compare the results of two different pools.

Keywords: Data science, electoral data visualisation, demographic data, R language.

Lista de figuras

Lista de quadros

Lista de abreviaturas e siglas

CRAN	Comprehensive R Archive Network
IBGE	Instituto Brasileiro de Geografia e Estatística
LAI	Lei de Acesso à Informação
PSDB	Partido da Social Democracia Brasileira
PT	Partido dos Trabalhadores
PSL	Partido Social Liberal
PSOL	Partido Socialista
SIC	Serviço de Informação ao Cidadão
TSE	Tribunal Superior Eleitoral
UF	Unidade Federativa

Sumário

1 Introdução

É perceptível o aumento do fluxo de dados na Internet à medida que o serviço passou a ser barateado e disponibilizado à população brasileira. Em 2018, a rede foi utilizada por cerca de 70% dos brasileiros pelo menos uma vez (CETIC.BR, 2019). Principalmente graças às redes sociais, a criação de conteúdo em larga escala também cresceu consideravelmente (AGGARWAL, 2011). Assim, há uma abundância de dados disponíveis através de poucos cliques e uma simples conexão à Internet.

Além disso, a popularização deste serviço permitiu a divulgação em massa de outros dados de natureza acadêmica ou informativa, por exemplo. Estes dados, outrora difíceis de obter e, portanto, objeto de estudo apenas para cientistas, tornaram-se acessíveis a qualquer pessoa com interesse ou curiosidade de procurá-los e analisá-los (BOYD; CRAWFORD, 2012).

Tamanha facilidade e velocidade de geração podem representar um empecilho para obter informações úteis destes dados crus. Neste caso, a quantidade não garante, necessariamente, a qualidade das análises (LABRINIDIS; JAGADISH, 2012). Além disso, nem toda informação presente na Internet é verídica e sua exuberância torna, por vezes, a análise da veracidade mais difícil.

Ainda assim, há fontes verificadas e confiáveis na rede, que se esforçam para garantir a veracidade de seus dados (dadas as limitações existentes). Governos de todo o mundo, por exemplo, têm utilizado a Internet para divulgar seus dados. No Brasil, graças à Lei de Acesso à Informação (LAI), é garantido ao cidadão o acesso às informações públicas de toda a esfera governamental: poderes, entidades, órgãos, Organizações Não Governamentais (ONGs) e administração pública. A requisição é feita *on-line* ou fisicamente no Serviço de Informação ao Consumidor (SIC) do órgão cujas informações deseja-se obter, de tal forma que nem todos os dados são prontamente disponibilizado, por motivos de segurança público ou sigilo (BRASIL, 2011).

Para facilitar o cumprimento da lei, o governo geralmente divulga seus dados antes mesmo de receber uma solicitação na forma desta lei, e a melhor forma para fazê-lo é através da Internet, dada sua facilidade de uso e presença em cerca de três a cada quatro domicílios, segundo o IBGE (2017).

Esta gama de dados governamentais de fácil acesso permite fazer análises relativamente mais robustas (uma vez que a fonte é tida como confiável) do que aquelas feitas a partir de informações oriundas de fontes não verificadas ou que não têm autoria sobre os dados.

Os dados eleitorais, objeto de estudo deste trabalho, não apresentam exceção à lei supracitada e são constantemente atualizados pelo Tribunal Superior Eleitoral (TSE). Sua

base de dados é acessível, contém informações sobre eleições ocorridas desde 1945 (apesar de faltar com informações de alguns pleitos) e pode ser importada para quaisquer *softwares* de computador (TSE, 2019).

A análise destes dados sozinhos permite entender melhor a situação política num dado momento da história. Combinada com outros dados relacionados da mesma época, é possível chegar a conclusões ainda mais acertadas e profundas sobre a população daquele tempo e sua visão política. Outra possibilidade é a de previsão de acontecimentos futuros, através da observação de repetições de padrões históricos e tendências que têm um grande impacto em eleições (MORRIS, 2019).

Enquanto o TSE disponibiliza uma ferramenta *on-line* de visualização dos dados das eleições ocorridas após 2004, não há nenhuma funcionalidade que compare os resultados de duas eleições semelhantes ocorridas em períodos diferentes, dificultando o acompanhamento da evolução dos votos num período de tempo. Além disso, suas funcionalidades deixam a desejar no quesito visualização.

Desta forma, entende-se que é possível obter *insights* significativos através da análise correta destes dados. Portanto, torna-se interessante a criação de ferramentas e métodos que facilitem o estudo e consequente transformação destas informações em conhecimento útil.

Assim, este projeto valeu-se da linguagem de programação R e dos dados eleitorais e demográficos disponíveis no site do TSE para o desenvolvimento de uma ferramenta que permite organizar, visualizar e modelar estes dados. Para tal, foram utilizadas técnicas de Ciência de Dados e diversos pacotes que estendem a funcionalidade da linguagem.

O resultado obtido foi uma série de funções em R de simples utilização por pessoas que não detêm avançado conhecimento em computação (requerendo, portanto, apenas conhecimentos básicos sobre a linguagem) que permite a visualização de mapas simples, separados por unidade federativa brasileira, que trazem diferentes informações sobre as eleições para presidente deste país numa única eleição ou ao longo de vários pleitos (mostrando a evolução de um mesmo fenômeno ao longo do tempo).

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta as bases de estudo utilizadas no desenvolvimento; o Capítulo 3 detalha as bibliotecas e extensões utilizadas na criação da ferramenta, além de detalhes sobre este processo; o Capítulo 4 mostra exemplos dos mapas gerados por meio das funções criadas em R e uma breve explanação sobre os resultados obtidos; por fim, o Capítulo 5 exprime anseios para o futuro da ferramenta e discorre sobre algumas dificuldades encontradas durante o seu desenvolvimento.

2 Fundamentação Teórica

Os subtópicos a seguir discorrem brevemente sobre os conceitos utilizados para a realização deste trabalho.

2.1 Ciência de Dados

É inegável que Ciência de Dados, num contexto geral, tornou-se objeto de interesse de uma parte maior da população mundial desde a última década. Isto colaborou para sua popularização, mas fez com que seu real significado se distorcesse. De fato, hoje não é possível afirmar, livre de contestações, que existe uma simples e absoluta definição do que Ciência de Dados realmente é (O'NEIL; SCHUTT, 2013).

De um modo geral, Ciência de Dados é uma coleção de técnicas utilizadas para extrair algo de valor de dados geralmente insignificantes quando vistos fora de contexto. Compreende técnicas de coleta, armazenamento, processamento, análise e transformação de dados com o intuito de obter informações concretas ou até mesmo predições (KOTU; DESHPANDE, 2018).

Apesar de só recentemente ter se tornado popular, os métodos e técnicas que permeiam esta ciência já são comuns e consolidados à área. Hoje, o conceito está inserido no cotidiano de diversas corporações e instituições que frequentemente utilizam análises de dados para tomar suas decisões.

2.2 Estatística Computacional

O paradigma de trabalho de um estatístico não mudou muito ao longo dos séculos. Basicamente, sempre consistiu em observar dados com o correto conhecimento do seu domínio para desenvolver um modelo para estudo e entendimento de um processo de geração de dados. A análise dos dados é ainda utilizada para refinar o modelo (ou até mesmo selecionar um modelo diferente para a natureza particular do problema em questão), escolher valores apropriados para as suas variáveis e tirar conclusões a partir dele (MORI, 2004).

O advento dos computadores ampliou a capacidade de utilização de novos métodos estatísticos, além de permitir aos cientistas da área lidar com quantidades jamais imaginadas de dados, até mesmo de diferentes naturezas, graças às informações avançadas que são hoje possíveis de se obter através destas máquinas. O tempo necessário para testar um modelo e coletar dados também pôde ser drasticamente reduzido.

A visualização dos dados e resultados de modelos teve avanço considerável. Com

aparelhos cada vez mais potentes, fica fácil ver, em tempo real, uma representação gráfica dos dados coletados e resultados produzidos.

Ainda que o objetivo deste trabalho seja apenas produzir uma ferramenta que facilite esta análise (e não criar as análises propriamente ditas), é importante notar que toda esta evolução permitiu a inferência computacional através da facilitação dos testes até mesmo de modelos que não parecem promissores para o campo. A comparação de vários modelos simultaneamente, buscando descobrir qual é mais promissor para uma aplicação, também representa uma grande evolução do fenômeno da computação em estatística.

2.3 A Linguagem R

R é um ambiente e linguagem de programação de código aberto surgido em 1993. É extensível (permitindo o acoplamento de vários outros pacotes que aumentam suas funcionalidades) e foi criado para lidar com estatística computacional e criação de elementos gráficos que representem estes dados (CRAN, 2018).

Além de já contar com diversas funções estatísticas que permitem a análise e modelagem de dados, qualquer um pode criar pacotes que facilitem a execução de tarefas mais complexas (como, por exemplo, a criação de mapas) ou que não estão inclusas na linguagem.

Neste ponto, é interessante notar que a comunidade que utiliza o ambiente é maciça, sendo considerada, segundo Cass (2019), a 5ª linguagem mais popular no mundo, atrás apenas de gigantes do mundo *mobile* e/ou web.

Sua principal conquista é trazer a computação mais próxima dos estatísticos que, por vezes, não são tão familiarizados com os paradigmas de programação como é um profissional da área.

É interessante notar que, apesar de sua grande comunidade, a documentação de suas funções e bibliotecas tende a ser mais pobre em detalhes do que o que o usuário comum está acostumado a ver ao utilizar linguagens como Python ou C. Por ser de código aberto e ter extensibilidade em mente, este efeito é esperado. Os fóruns abertos costumam concentrar grande parte das informações sobre como resolver determinados problemas que podem ser gerados através da utilização destes pacotes, por vezes sendo mais valiosos que a própria documentação fornecida pelo desenvolvedor.

3 Metodologia

Este capítulo discorre sobre os métodos e técnicas de desenvolvimento utilizadas ao longo do projeto.

3.1 Base de dados do TSE

O TSE é o órgão máximo da Justiça Eleitoral e serve como um dos pilares da construção da democracia brasileira. Apesar de ser uma corte e julgar, normalmente, os casos que dizem respeito a sua alçada, também é responsável por organizar as eleições no país.

Desde 1996, o país começou a ver o uso em massa no processo eleitoral das urnas eletrônicas. O voto eletrônico permite uma precisão e velocidade na contagem de votos jamais antes vistas. Graças a isso, hoje é possível saber o resultado de uma eleição que envolve mais de uma centena de milhões de indivíduos em poucas horas.

Este advento favoreceu em muito o registro dos dados para posterior consulta e análise. Por já estar no meio digital, o número de erros a que o processo de transposição de dados está sujeito é muito menor do que quando o procedimento utilizava papel.

O TSE disponibiliza dados relativos às candidaturas, eleitorado, prestação de contas, pesquisas eleitorais e resultados. Alguns dados são coletados desde 1945, mas não contam com a mesma riqueza de detalhes ou precisão características dos processos mais recentes.

Embora os dados abertos do TSE possam ser facilmente acessados por meio do Repositório de dados eleitorais, o seu uso requer um nível básico de conhecimento técnico. Há arquivos tanto no formato tabulado (arquivos .csv) como simples (arquivos .txt básicos). Os repositórios contêm ainda arquivos .pdf que descrevem brevemente como cada arquivo é organizado e o que cada campo representa.

Para a escolha dos anos que este projeto contemplaria, pensou-se em utilizar apenas os dados disponibilizados a partir de 1988, uma vez que corresponde ao ano da atual Constituição Federal (e também das regras do processo eleitoral), que dividiu o Brasil como é hoje (a criação do estado do Tocantins vem deste documento).

Porém, há poucos dados sobre a primeira eleição após esta Constituição, por isso foi desconsiderado o ano de 1989 nas análises do projeto. Além disso, os dados das eleições de 1994 a 2002 que constam na base de dados do TSE estão incompletos, como informa o próprio Tribunal em sua página.

Assim, decidiu-se disponibilizar nesta ferramenta apenas os dados das eleições gerais ocorridas em 2006, 2010, 2014 e 2018. Os dados coletados dos resultados e perfil eleitoral

(únicos repositórios utilizados neste trabalho) nestes anos são, em sua maioria, os mesmos, o que garante uma análise justa.

Foram usados, para a construção do projeto, o arquivo "Votação nominal por município e zona" disponível sob "Resultados" e o arquivo "Eleitorado xxxx" (onde "xxxx"corresponde ao ano da eleição) de cada ano acima citado. Todos os arquivos estão disponíveis no site do TSE[1].

3.2 Os pacotes R

É sabido que há várias funções disponíveis por padrão numa instalação comum do ambiente de programação R. Porém, é possível alavancar ainda mais as suas capacidades através da adição de outros pacotes (ou bibliotecas) que facilitam as etapas de desenvolvimento da aplicação. Com exceção do último, todos fazem parte do *tidyverse*, uma renomada coleção de pacotes R de alta qualidade criados para Ciência de Dados. Para este projeto, foram utilizados os pacotes relacionados nas seções 3.2.1 a 3.2.3 deste Capítulo.

3.2.1 Pacote readr

O objetivo principal deste pacote é fornecer métodos simples e rápidos para a leitura de dados "retangulares", ou seja, bidimensionais, geralmente expressos em tabelas.

Estes dados são geralmente inseridos num objeto do tipo *data.frame*, estrutura utilizada em R para o armazenamento de dados de duas dimensões (linhas e colunas). A estrutura ainda disponibiliza funções para consultar seu cabeçalho, número de colunas, número de linhas, etc.

Sua principal função é a *read_delim()*, que recebe como parâmetro uma cadeia de caracteres representando o caminho para um arquivo tabulado (ou separado por um ou mais caracteres) e uma série de campos adicionais, como *col_names*, um campo binário que indica a presença ou ausência de cabeçalho no arquivo lido; *skip*, que recebe um número inteiro correspondente ao número de linhas que deve ser ignorado no início da leitura; entre tantos outros.

Há ainda funções para tipos específicos de arquivos, como *read_csv()* e *read_tsv()*, que otimizam a leitura de arquivos .csv e .tsv, respectivamente. Neste trabalho, apenas utilizou-se o *read_delim()*.

Foi de grande ajuda a existência do parâmetro *col_types* em conjunto com a função *cols_only()*, que permite escolher quais colunas do arquivo fonte serão de fato importadas para o *data.frame*. Uma vez que os arquivos fornecidos pelo TSE são ricos em detalhes que, por vezes, não são utilizados para as consultas definidas no escopo deste projeto, importar

[1] http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais

apenas as colunas necessárias promoveu um ganho significativo em velocidade de execução e armazenamento utilizado.

3.2.1.1 Leitura dos arquivos de resultados das eleições

O Código 3.1 traz um exemplo de leitura do arquivo dos resultados das eleições de 2018 por município e zona.

Código 3.1 – Leitura dos dados da eleição para presidente de 2018

```
1  poll <- read_delim(
2      "dados/votacao_candidato_munzona_2018/votacao_candidato_munzona_2018_
           BRASIL.csv",
3      delim=";",
4      quote='"',
5      col_types = cols_only(
6          ANO_ELEICAO = "c",
7          SG_UF = "c",
8          CD_MUNICIPIO = "i",
9          NM_MUNICIPIO = "c",
10         NR_TURNO = "i",
11         NR_ZONA = "i",
12         CD_CARGO = "i",
13         NR_CANDIDATO = "i",
14         NM_URNA_CANDIDATO = "c",
15         NR_PARTIDO = "i",
16         SG_PARTIDO = "c",
17         QT_VOTOS_NOMINAIS = "i"))
```

Esta função, presente no arquivo filter_tables.R, é responsável por ler e inserir dados do arquivos *votacao_candidato_munzona_2018_BRASIL.csv* em um *data.frame* nomeado "poll". O arquivo original possui 38 colunas, mas apenas as descritas entre as linhas 6 e 17 são utilizadas.

Seguido do nome de cada coluna nas linhas mencionadas anteriormente há uma atribuição de caractere, que denota qual tipo de dado a coluna contém: se "c", uma cadeia de caracteres; se "i", um número inteiro. Semelhantemente, é possível utilizar o caractere "?" para que a função "adivinhe" o tipo. É o que ocorre na leitura dos arquivos de 2006 a 2014, mostrado no código 3.2.

Código 3.2 – Leitura dos dados das eleições para presidente de 2006 a 2014

```
1  header <- c("ANO_ELEICAO", "NR_TURNO", "SG_UF", "CD_MUNICIPIO", "NM_
       MUNICIPIO", "NR_ZONA", "CD_CARGO", "NM_URNA_CANDIDATO", "SG_PARTIDO", "
       QT_VOTOS_NOMINAIS")
2  poll <- read_delim(
```

```
3     "dados/votacao_candidato_munzona_xxxx/votacao_candidato_munzona_xxxx_BR
          .txt",
4     delim=";",
5     quote='"',
6     locale = readr::locale(encoding = "latin1"),
7     col_names=FALSE,
8     col_types = cols_only(
9         X3 = "?",
10        X4 = "?",
11        X6 = "?",
12        X7 = "?",
13        X9 = "?",
14        X10 = "?",
15        X11 = "?",
16        X15 = "?",
17        X24 = "?",
18        X29 = "?"))
19    names(poll) <- header
```

No Código 3.2, "xxxx"corresponde ao ano da eleição (2006, 2010 ou 2014).

Uma vez que há apenas registros inteiramente numéricos ou inteiramente alfabéticos, é seguro utilizar este caractere coringa certo de que a transposição dos dados para o R não será prejudicada.

Por ser um arquivo de texto (.txt), é preciso especificar sua codificação para que não haja problemas com a leitura de caracteres especiais, como os acentos na língua portuguesa. Esta etapa ocorre na linha 6 e segue o formato recomendado pelo TSE nos documentos explicativos dos arquivos disponibilizados no site.

Além dessas etapas adicionais, é preciso manualmente inserir a primeira linha do *data.frame* (*header*, visto que apenas os arquivos de 2018 contêm esta estrutura completa. Isto ocorre na linha 19 utilizando a cadeia de caracteres criada na linha 1.

3.2.1.2 Leitura dos arquivos do perfil do eleitorado

Valendo-se da mesma biblioteca e dos mesmos métodos, são lidos os dados do perfil do eleitorado, como pode ser visualizado no Código 3.3.

Código 3.3 – Leitura dos dados demográficos das eleições gerais de 2006 a 2014

```
1   poll <- read_delim(
2     "dados/perfil_eleitorado_xxxx.txt",
3     delim=";",
4     quote='"',
5     locale = readr::locale(encoding="latin1"),
6     col_names = FALSE,
```

```
7       col_types = cols_only(
8           X2 ="?",
9           X5 = "?",
10          X7 = "?",
11          X8 = "?",
12          X9 = "?"))
13    names(poll) <- c(
14      "SG_UF",
15      "NR_ZONA",
16      "DS_FAIXA_ETARIA",
17      "DS_GRAU_ESCOLARIDADE",
18      "QT_ELEITORES_PERFIL")
```

Neste exemplo, "xxxx" corresponde ao ano da eleição (2006, 2010 ou 2014).

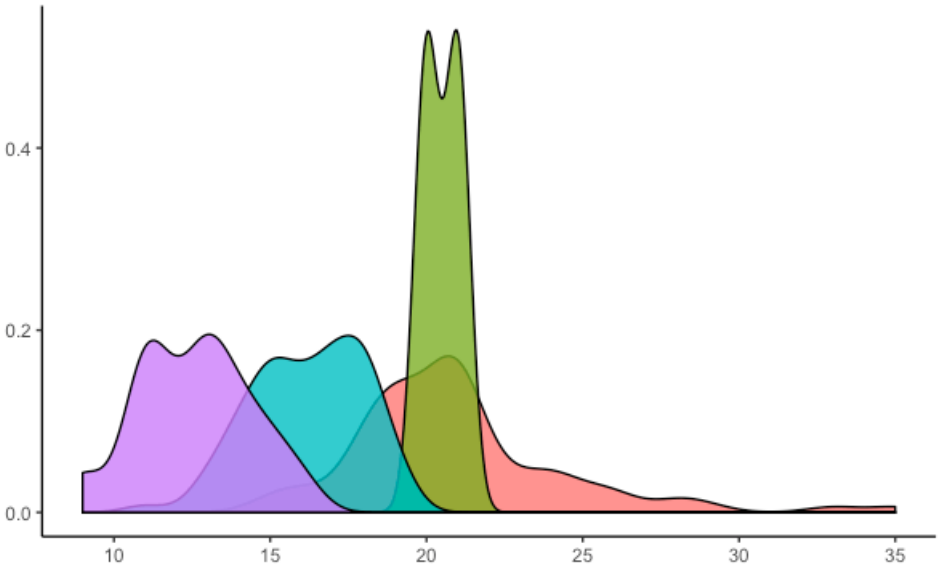
A leitura dos dados demográficos da última eleição é feita de forma mais simples por já conter o cabeçalho. Ocorre de forma análoga à dos resultados do mesmo ano, observadas as mudanças do nome das colunas e do arquivo.

3.2.2 Pacote ggplot2

Esta é uma biblioteca para a criação de gráficos em geral através de comandos simples que instruem como deve ser a passagem dos dados para as imagens. É considerada uma das bibliotecas mais completas em R para esta finalidade.

A *ggplot2* permite criar uma gama muito diversificada de gráficos, desde os mais simples (como gráficos de barra ou pizza) até os mais complexos (correlogramas e histogramas, por exemplo). Também é possível exportar processos mais complexos em funções que facilitam seu uso. A Figura 1 apresenta um gráfico produzido através da biblioteca.

Figura 1 – Exemplo de gráfico de densidade construído com as funções da biblioteca *ggplot2*



Fonte: Prabhakaran (2017)

Para os mapas, geralmente utilizam-se polígonos encapsulados na classe *SpatialPolygonsDataFrame* do ambiente R, que checa a existência de uma equivalência aos IDs (elementos únicos que identificam uma informação) de cada polígono com as informações a serem desenhadas.

Neste projeto, foi utilizada a biblioteca *abjData* (explicada em mais detalhes no subtópico 3.2.3), que lida mais diretamente com as funções do *ggplot2* para construir o mapa. Assim, utiliza-se a função *(plot_map(dataset))* do Código 3.4 para construir os mapas.

Código 3.4 – Função para desenho de mapas

```
plot_map <- function(dataset){
  dataset %>% inner_join(abjData::br_uf_map) %>% {
      ggplot(.) +
        geom_map(aes(x = long, y = lat,
                     map_id = id, fill = variavel),
                     color = 'black', map = ., data = .) +
        theme_void() + coord_equal()}
}
```

O parâmetro *dataset*, presente no código, corresponde a uma tabela que contém duas colunas: uma com as siglas das Unidades Federativas (UFs) brasileiras e outra com um dado quantitativo referente à intensidade de uma cor no gráfico. O Quadro 1 apresenta um exemplo da estrutura desta tabela e a Figura 2 representa o mapa gerado através da inserção destes dados na função.

Quadro 1 – Exemplo de tabela usada para desenho do mapa

id	**variavel**
AC	10
AL	22
AP	23
AM	40
BA	10
CE	30
DF	55
ES	40
GO	13
MA	8
MT	7
MS	-44
MG	10
PA	6
PB	52
PR	5
PE	10
PI	1
RJ	5
RN	21
RS	10
RO	22
RR	23
SC	40
SP	32
SE	20
TO	0

Fonte: Elaborado pelo autor.

Figura 2 – Mapa gerado através dos dados do Quadro 1 e a função *plot_map()*

Fonte: Elaborado pelo autor.

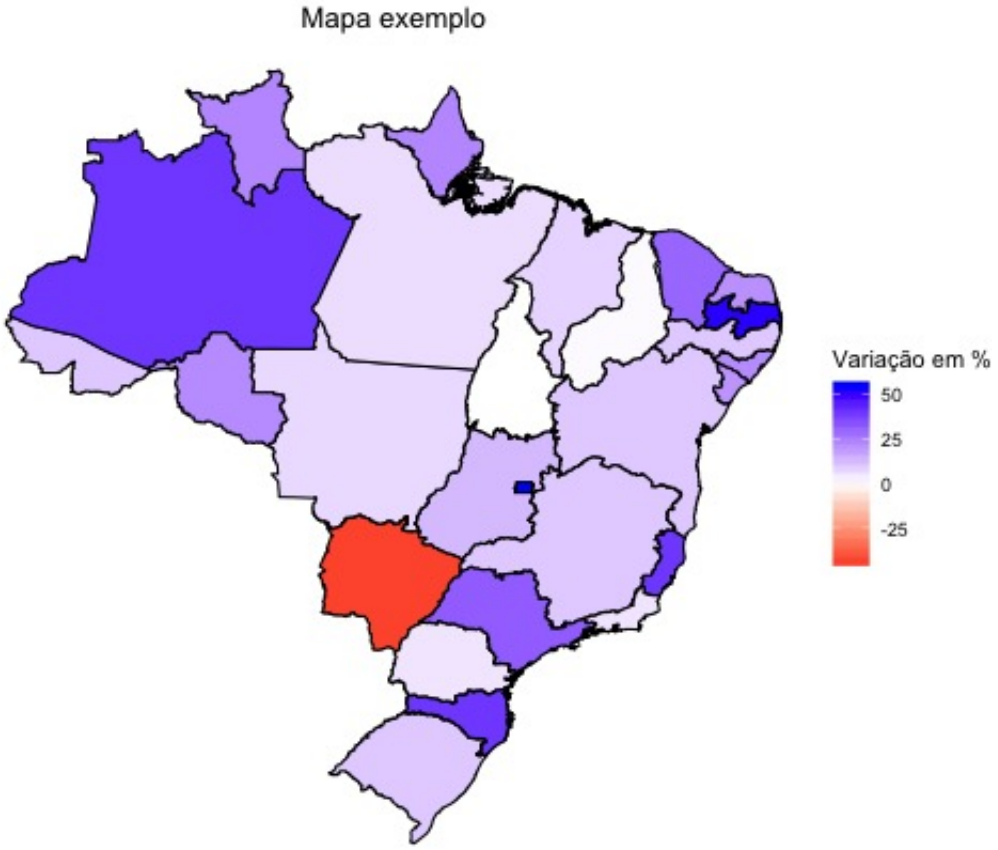
Nota-se que as UFs às quais foram atribuídos números maiores receberam uma coloração mais forte de azul (ou, no caso do valor negativo, de vermelho), contrastando com as que tiveram números menores. Esta regra é utilizada em todos os mapas gerados através deste trabalho para mostrar o grau de intensidade de uma grandeza em uma UF.

São necessárias poucas linhas de código para adicionar outros elementos ao mapa, como um título e legenda, mostradas no Código 3.5. A Figura 3 denota as alterações produzidas.

Código 3.5 – Função para desenho de mapas

```
df_exemplo %>%
  plot_map()+
  scale_fill_gradient2(name = "Variacao em %", low="red",
                       midpoint = 0, high="blue",
                       mid = "white",
                       limits = c(min(dd$variavel),
                                  max(dd$variavel))) +
  ggtitle("Mapa exemplo") +
  theme(plot.title = element_text(hjust = 0.5))
```

Figura 3 – Mapa da Figura 2 modificado com elementos visuais de gráficos



Fonte: Elaborado pelo autor.

df_exemplo corresponde ao *data.frame* que contém os dados explicitados no Quadro 1. O código das linhas 3 a 7 determinam as cores do degradê e a legenda do mapa. Na linha 8, define-se o título do mapa e, na 9, seu tamanho e posicionamento.

3.2.3 Bibliotecas minimamente utilizadas

O projeto também requereu outras bibliotecas, além das padrões do R, que não demandam a criação de um subtópico inteiro por terem sido apenas superficialmente utilizadas. É o caso do *dplyr*, uma biblioteca especializada em manipulação de dados através de funções que, como seus nomes sugerem (*filter, select, arrange,* etc.), realizam operações simples para resolver os problemas mais comuns desta etapa.

Foi utilizado, ainda, o pacote *abjData*, fornecido pela Associação Brasileira de Jurimetria. Coube a ele criar a base dos mapas produzidos neste trabalho através da função *geom_map* do

ggplot2 e de um *data.frame* que contém o mapeamento dos polígonos e as siglas dos estados brasileiros.

Assim, é necessário apenas uma estrutura que contenha duas colunas para traçar o mapa: a sigla oficial da UF e um valor que determinará a cor (ou tonalidade de cor) da qual este elemento será colorizado na renderização do mapa final. Em todos os casos, este é um valor numérico que expressa a grandeza de uma dimensão (quão maior ou menor foi a vantagem de votos entre uma eleição e outra, a diferença entre os votos recebidos em um estado e outro, etc.), mas também poderia assumir o valor de uma cadeia de caracteres correspondente à cor sólida que o elemento deve assumir (como "blue" ou "red").

4 Funções e Resultados Obtidos

Este trabalhou produziu duas funções em R as quais geram mapas distintos, utilizando-se dos dados crus (brutos) do TSE: *plot_demographic_data* e *single_party_evolution*. Os tópicos a 4.1 e 4.2 descrevem em mais detalhes o funcionamento das funções e apresentam os gráficos que foram gerados através delas.

4.1 Função single_party_evolution(year1, year2, party, round)

Recebe como parâmetros dois números naturais distintos (correspondentes ao ano de realização de duas eleições), uma cadeia de caracteres equivalente à sigla de um partido político e um número natural entre 1 e 2, que representa o turno da votação. O mapa gerado apresentará o índice de variação (em pontos percentuais) da porcentagem de votos que o partido em questão recebeu em cada UF entre os dois anos informados.

Através deste mapa, é possível perceber como uma UF tem se deslocado no espectro político ao longo dos anos. Se, por exemplo, uma UF que tradicionalmente vota em partidos de esquerda começa a diminuir a porcentagem de votos que confere ao partido representante deste "lado" político, pode-se esperar que numa próxima eleição esta diferença se intensifique. Isto pode indicar ao partido uma necessidade de intensificar suas campanhas naquela UF, mesmo que ela tenha sido historicamente fiel às suas ideias.

Esta função não se limita apenas a partidos grandes. É possível traçar a evolução de partidos pequenos também, desde que ele tenha participado das eleições para presidente nos dois anos analisados. É válido notar que o ano das eleições a serem comparadas não precisa ser sequencial: pode-se comparar, por exemplo, as eleições de 2010 e 2014 ou de 2014 e 2010, gerando resultados diferentes; ou as de 2006 e 2018, que contam com 12 anos de diferença entre si.

A análise é simples: leem-se os dados dos arquivos correspondentes às eleições dos anos *year1* e *year2*, utilizando o método descrito na Seção 3.2.1.1. Depois, filtrando-se as duas tabelas pelos campos correspondentes (turno e partido), obtem-se um vetor que contém a quantidade total de votos nominais que o partido recebeu em cada eleição e em cada UF. Dividindo estes valores pelo total de votos nominais que cada UF coletou, tem-se a porcentagem de votos do partido naquela UF.

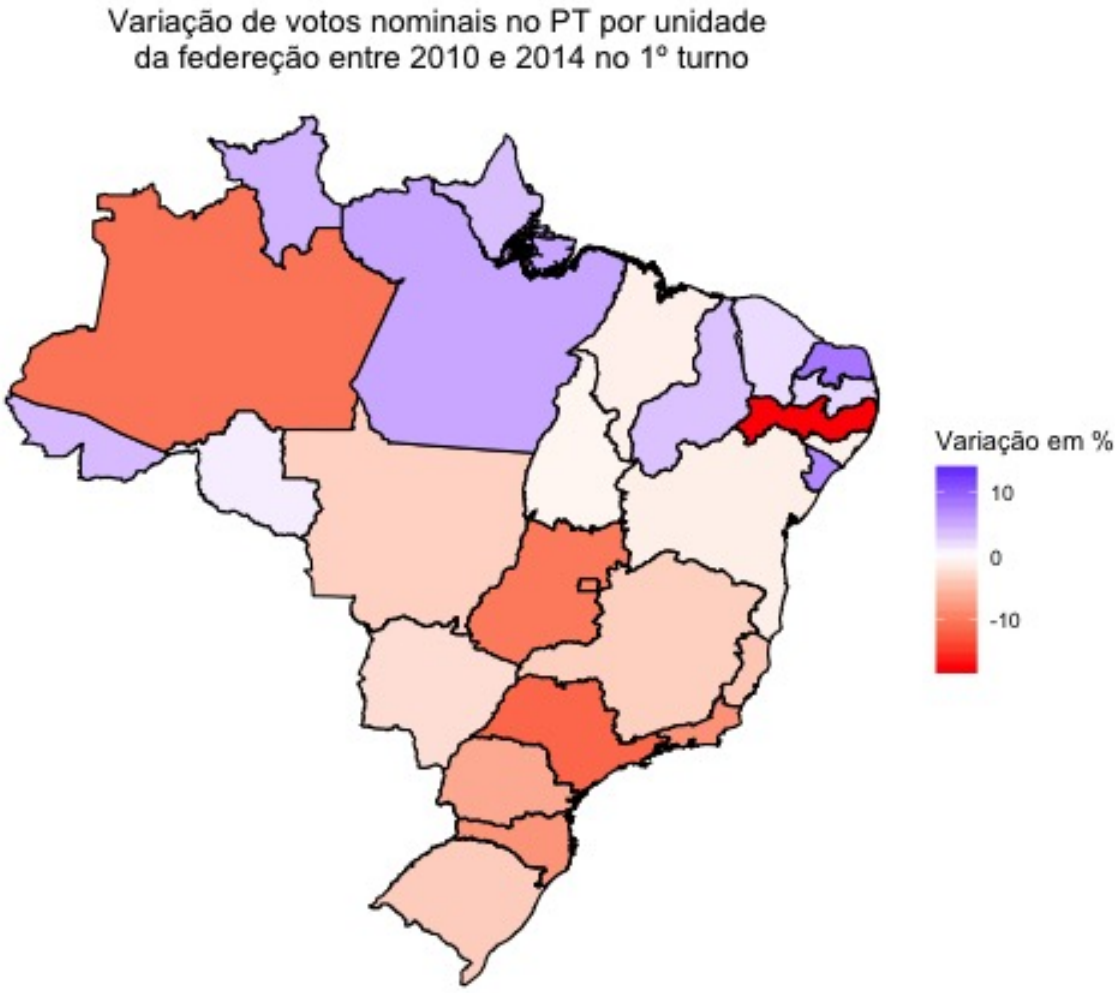
Obtidos estes dados, subtraem-se os dois vetores para chegar na diferença na porcentagem de votos que o partido observou entre as duas eleições. Basta, então, inseri-los num *data.frame* semelhante ao descrito na Seção 3.2.2 e então utilizar as funções do *ggplot2* para decorar o mapa. Alguns resultados gerados por esta função podem ser conferidos na Figura

4 e demais figuras dste subtópico. Estes exemplos citam o Partido dos Trabalhadores (PT), Partido da Social Democracia Brasileira (PSDB) e Partido Socialista (PSOL).

Nota-se que, para intervalos de tempo que compreendem mais de uma eleição, apenas é comparado, pontualmente, a primeira com a última eleição requisitada nos parâmetros da função.

Através da observação da Figura 4, nota-se que, em comparação a 2010, o PT obteve uma porcentagem menor de votos em 2014 em todas as UFs do sul, sudeste e centro-oeste. Enquanto na maioria dos estados do nordeste o partido aumentou sua margem de votos, Pernambuco destoou este padrão ao apresentar uma queda de cerca de 15%.

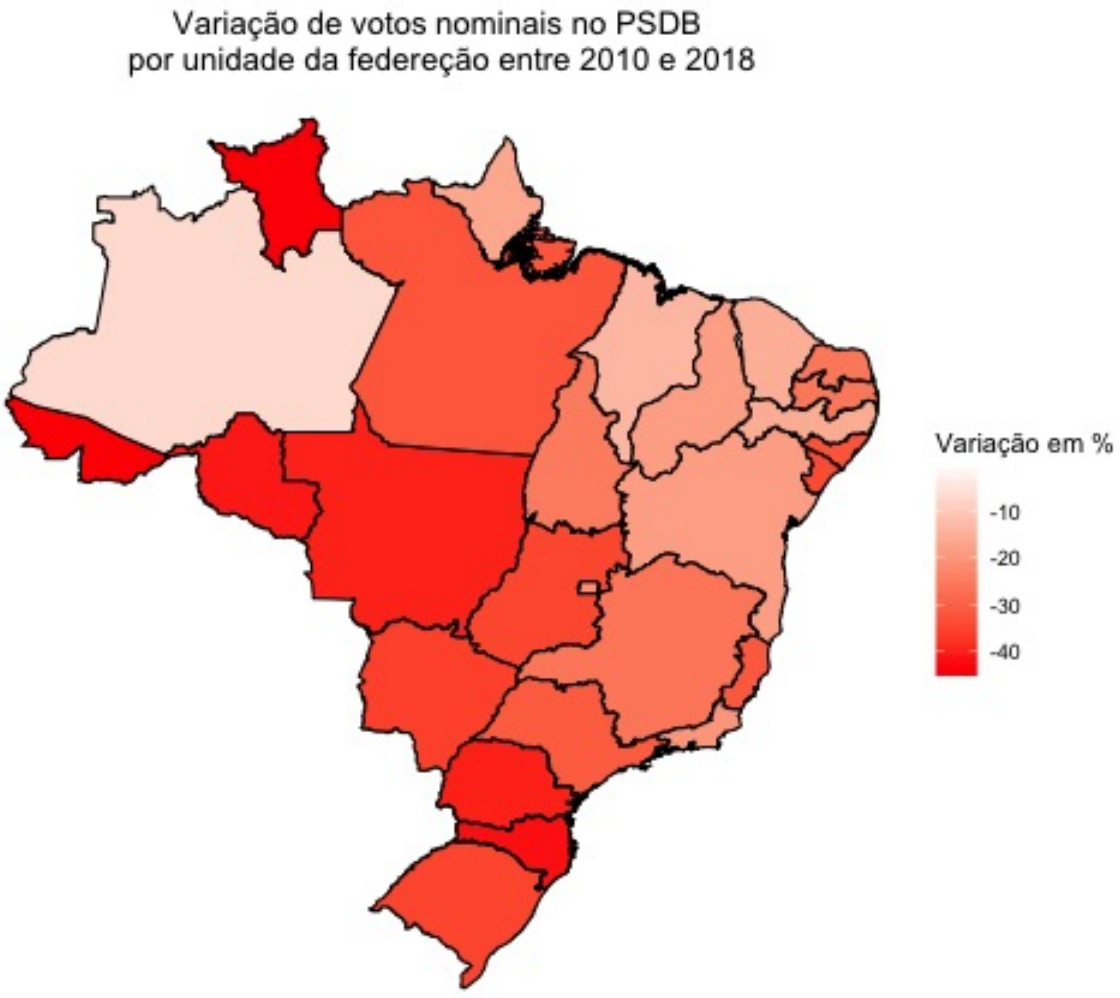
Figura 4 – Mapa da variação de votos nominais no PT por UF nos anos de 2010 e 2014 (1º turno)



Fonte: Elaborado pelo autor.

Já na Figura 5 percebe-se facilmente o que aconteceu com o PSDB, partido que disputava o pódio com o PT desde 1994, na última eleição. O partido apresentou uma queda brusca de desempenho em todo o país e, pela primeira vez em 24 anos, não chegou perto de disputar o 2º turno, dando lugar ao Partido Social Liberal (PSL) . O gráfico mostra claramente este declínio em comparação com a eleição de 2010, denunciando quedas de cerca de 40% em algumas UFs.

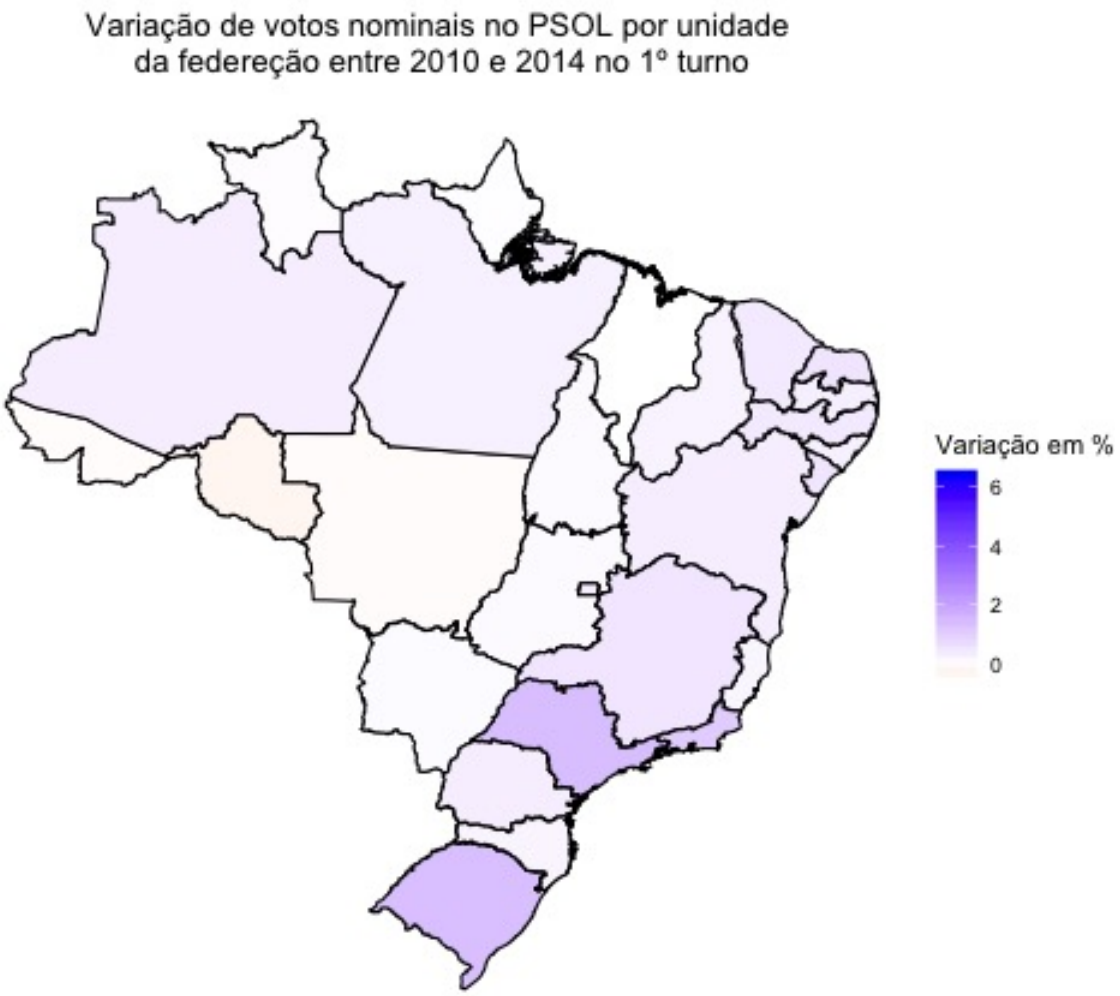
Figura 5 – Mapa da variação de votos nominais no PSDB por UF nos anos de 2010 e 2018 (1º turno)



Fonte: Elaborado pelo autor.

A Figura 6 é um "contra-exemplo" do gráfico anterior que se faz valer de um partido considerado "nanico", isto é, que não costuma impactar o resultado das eleições, com porcentagens de votos às vezes inexpressivas. O PSOL, historicamente, nunca teve uma votação expressiva. Mas nota-se que, entre 2010 e 2014, aumentou ou manteve igual seus percentuais de votos em todas as UFs.

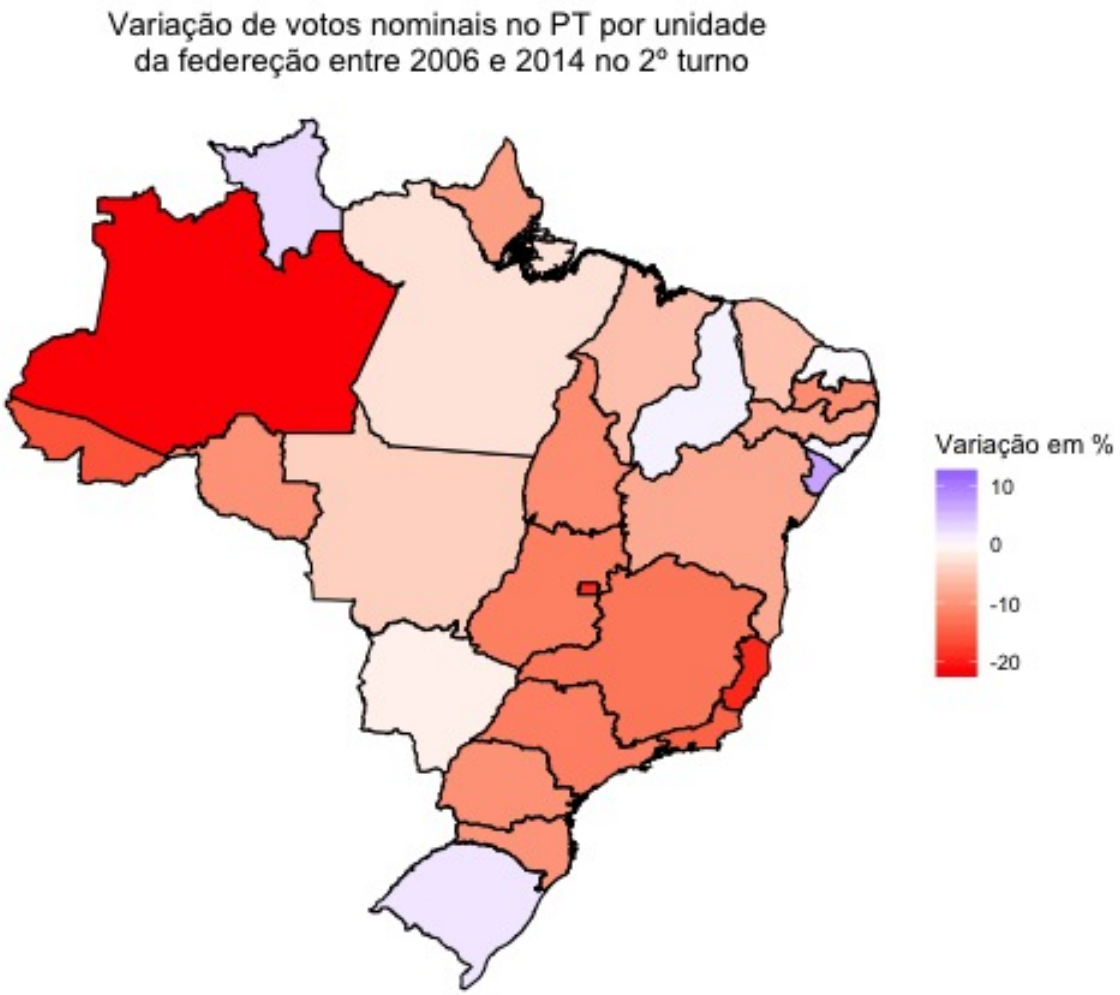
Figura 6 – Mapa da variação de votos nominais no PSOL por UF nos anos de 2010 e 2014 (1º turno)



Fonte: Elaborado pelo autor.

Por último, a Figura 7 traz um exemplo de 2º turno. Por três eleições seguidas (2002, 2006 e 2010), o PT manteve uma margem positiva considerável (porém decadente) sobre o PSDB no turno decisivo. Entretanto, em 2014, vivenciou sua vitória mais acirrada até hoje, com menos de 2 pontos percentuais de diferença. Este fato é observado na Figura 7, que mostra como o partido perdeu votos em quase todas as Unidades Federativas no período citado.

Figura 7 – Mapa da variação de votos nominais no PT por UF entre 2006 e 2014 (2º turno)



Fonte: Elaborado pelo autor.

4.2 Função plot_demographic_data(year, party, round, state)

Cruzando-se os dados relativos à eleitorado e os resultados das eleições, criou-se esta função. Seus parâmetros são um número natural correspondente ao ano de uma eleição, uma

cadeia de caracteres representando um partido, outro número natural que simboliza o número do turno e uma última cadeia de caracteres que determina um grupo de eleitores numa situação (como, por exemplo, "17 ANOS" para se referir à idade ou "SUPERIOR INCOMPLETO" para o grau de escolaridade).

Não é possível saber, especificamente, através dos dados fornecidos pelo TSE, qual foi o contingente de votos que um "grupo social" (conjunto de indivíduos com características semelhantes) proporcionou a um candidato, mas é possível obter dados relativos às zonas eleitorais e gerar outras estatísticas baseando-se nestas informações.

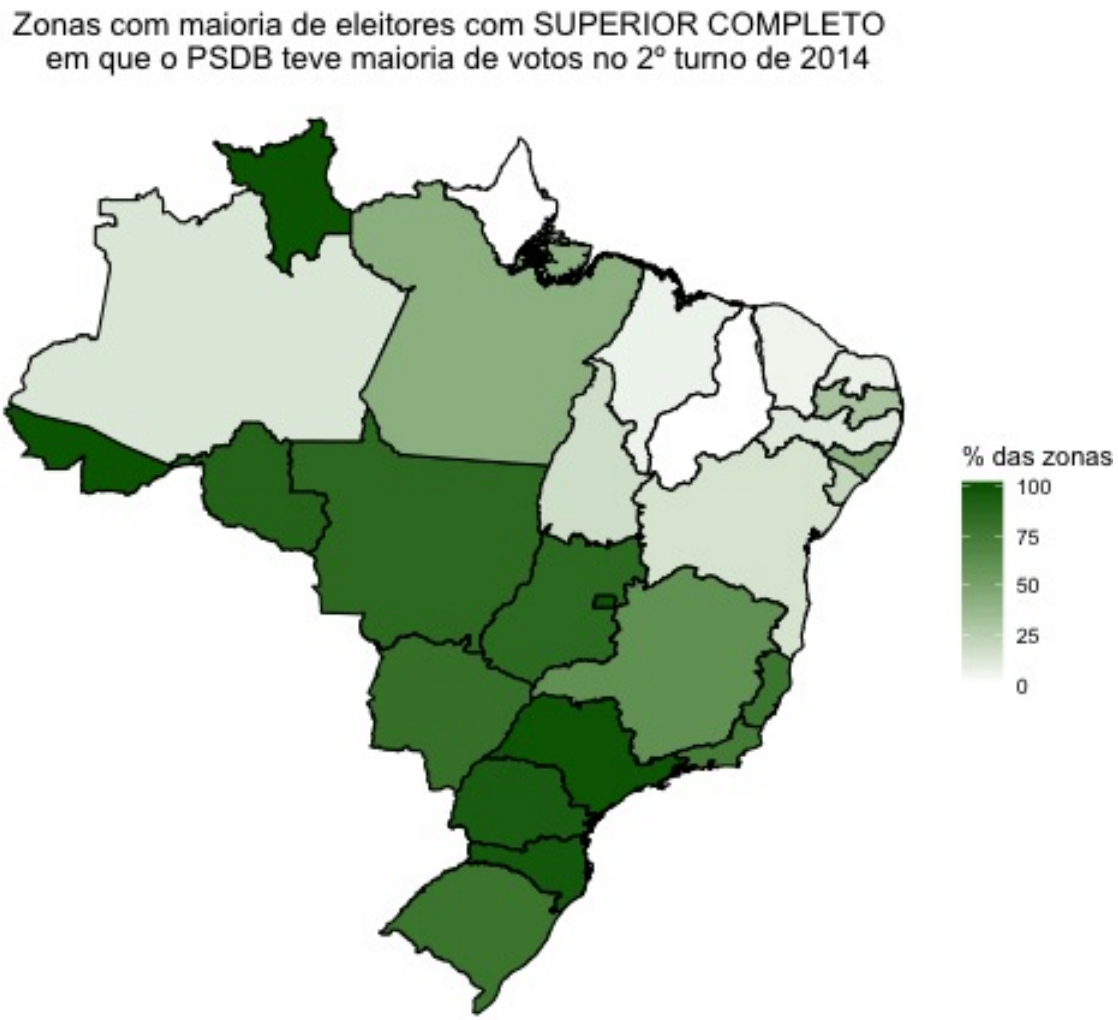
Esta função gera um mapa que responde à pergunta "em quantas zonas cuja maioria dos eleitores pertence a um grupo específico, um partido ganhou a maioria absoluta dos votos, por estado?". Para tanto, calcula-se a porcentagem média de eleitores para cada condição de um grupo de atributos (idade ou grau de escolaridade) por estado. Isto permite saber quais zonas possuem uma maioria de eleitores daquele grupo específico. Estes dados provêm da leitura dos arquivos descrita na Seção 3.2.1.2.

Então, são coletados os dados referentes ao ano da eleição e turno fornecidos anteriormente, de forma semelhante ao que acontece na função *single_party_evolution*. Após isso, faz-se a análise do total dos votos por zona eleitoral para descobrir se o candidato do partido informado nos parâmetros da função obteve a maioria dos votos nominais. O processo é repetido para todas as zonas de cada UF.

Sabendo-se em quais zonas o candidato levou a maioria dos votos e em quais zonas há uma observância maior de um grupo social específico, basta dividir estes valores para se chegar à porcentagem de zonas que cumprem com a condição imposta. Seguem alguns exemplos.

É interessante notar, na Figura 8, como o PSDB levou perto de 100% das zonas com maioria de eleitores com ensino superior completo em estados com grande eleitorado, como São Paulo e Paraná (1º e 5º estados mais populosos do país, respectivamente).

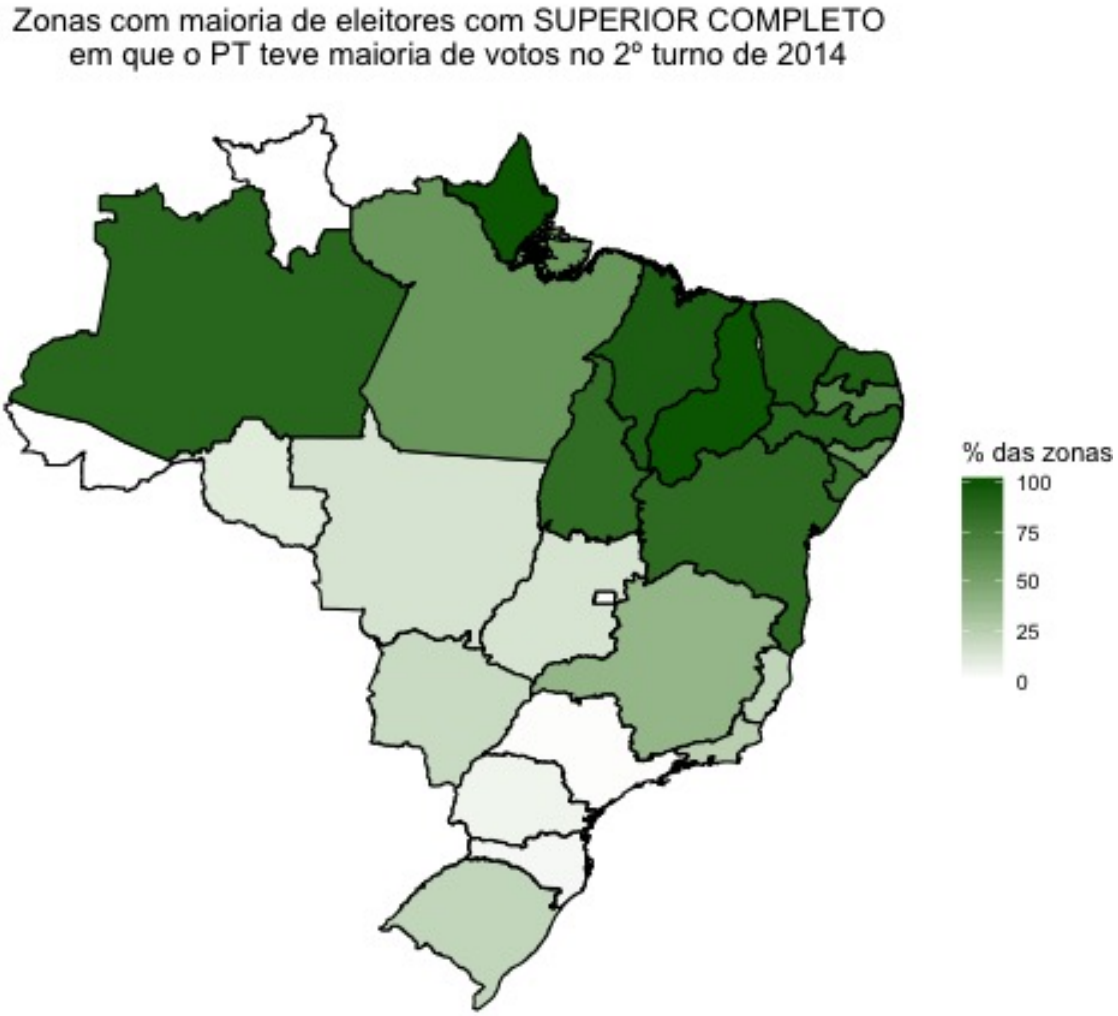
Figura 8 – Mapa das zonas com maioria de eleitores com ensino superior completo em que o PSDB obteve maioria de votos em 2014 (2º turno)



Fonte: Elaborado pelo autor.

Analogamente, ao analisar, nas mesmas condições, a situação do PT, como vista na Figura 9, obteve-se o "complemento"do mapa. O partido garantiu maioria destas zonas em todos os estados do nordeste e em alguns do norte, o que representa um contingente bem menor se comparado àqueles citados no mapa da Figura 8 que o PSDB conseguiu vencer.

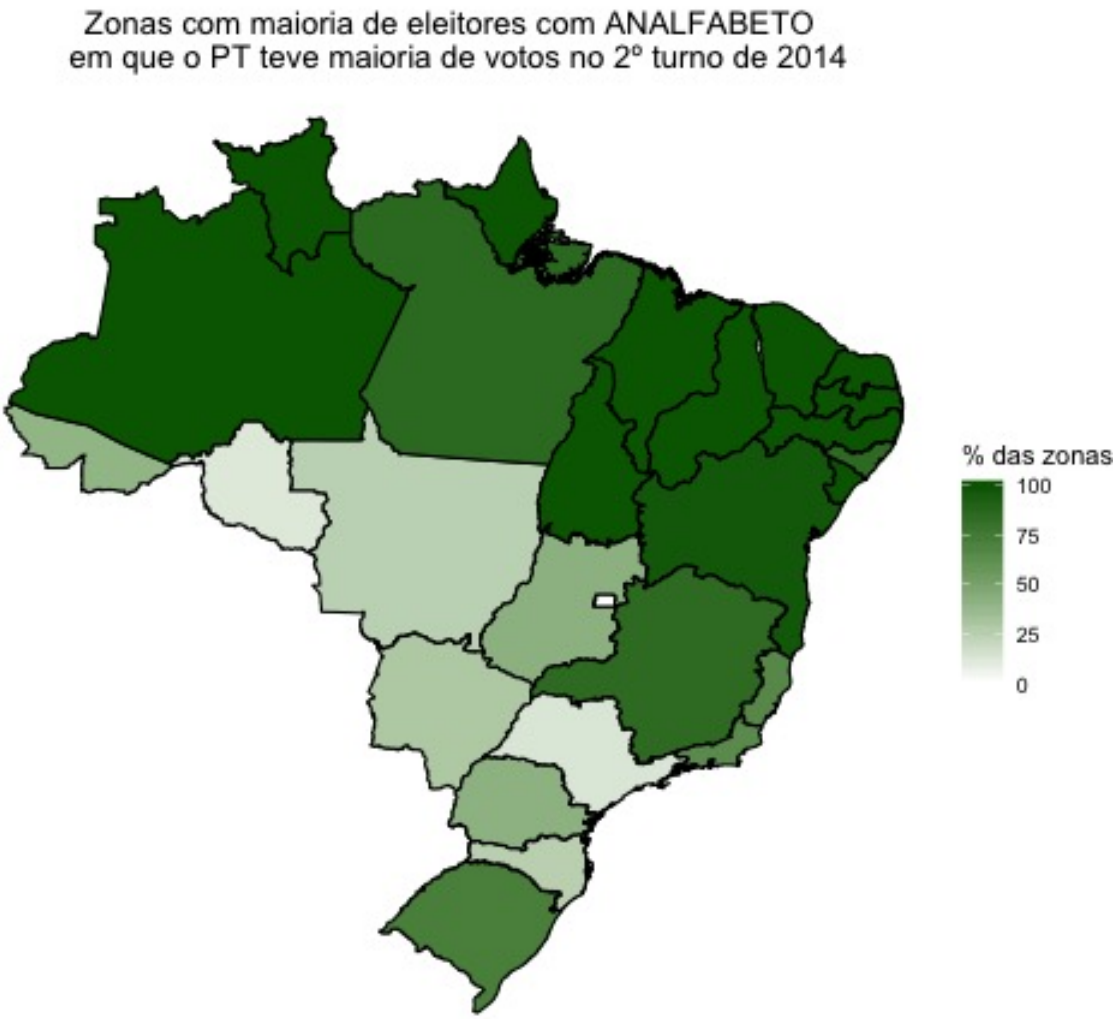
Figura 9 – Mapa das zonas com maioria de eleitores com ensino superior completo em que o PT obteve maioria de votos em 2014 (2º turno)



Fonte: Elaborado pelo autor.

Por fim, vê-se na Figura 10 que o PT ganhou a maioria das zonas com média acima do normal de analfabetos votantes em todo o norte e nordeste e, mesmo tendo menos votos na maioria dos estados do sul, sudeste e centro-oeste, conseguiu garantir uma porcentagem expressiva dessas zonas.

Figura 10 – Zonas com maioria de eleitores analfabetos em que o PT obteve maioria de votos em 2014 (2º turno)



Fonte: Elaborado pelo autor.

5 Conclusão

Apesar de não ter sido possível utilizar dados de várias eleições como previsto no início do desenvolvimento do projeto, o trabalho ainda assim cumpriu seu objetivo proposto de prover um panorama que mostre a evolução entre duas votações semelhantes ocorridas em períodos distintos. Assumindo que o TSE continue com o excelente trabalho que vem mostrando recentemente quanto à coleta e disponibilização de dados cada vez mais ricos, a ferramenta pode ser expandida para incluir os resultados destes próximos processos eleitorais.

Também acredita-se que o objetivo de tornar a ferramenta acessível a indivíduos que não tenham tanto conhecimento sobre a linguagem e/ou programação em geral tenha sido alcançado, uma vez que as funções geradas são de simples utilização e não requerem que o usuário conheça previamente as bibliotecas utilizadas para a confecção dos mapas e leitura dos dados (como as que compõem o *tidyverse*).

5.1 Para o futuro

Posteriormente, espera-se que o projeto se torne uma verdadeira biblioteca do R, distribuída pelo *Comprehensive R Archive Network* (CRAN). Para isso, são necessários alguns passos adicionais para compatibilizá-la (como a criação de um arquivo explícito de documentação e descrição das funções da biblioteca, o *upload* dos arquivos para um repositório público, etc.) de tal forma que esteja pronta para distribuição. Objetiva-se, ainda, estender as consultas para além do cargo de presidente, uma vez que também estão disponíveis dados relativos às eleições de governadores, prefeitos, deputados federais e estaduais, vereadores e senadores. Para isto, alguns cuidados adicionais devem ser levados em conta, mas sua implementação consiste, basicamente, em simples mudanças na função *filter* presente na maioria dos arquivos de código-fonte do projeto.

Referências

AGGARWAL, C. C. An introduction to social network data analytics. In: *Social network data analytics*. [S.l.]: Springer, 2011. p. 1–15.

BOYD, D.; CRAWFORD, K. Critical questions for big data. *Information, Communication & Society*, Routledge, v. 15, n. 5, p. 662–679, 2012. Disponível em: <https://doi.org/10.1080/1369118X.2012.678878>. Acesso em: 28 Out. 2019.

BRASIL. *Lei nº 12.527, de 18 de novembro de 2011*. Brasília, DF, 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 21 Mar. 2019.

CASS, S. *The Top Programming Languages 2019*. IEEE, 2019. Disponível em: <https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019>.

CENTRO REGIONAL DE ESTUDOS PARA O DESENVOLVIMENTO DA SOCIEDADE DA INFORMAçãO. *Pesquisa sobre o Uso das Tecnologias de Informação e Comunicação nos domicílios brasileiros - TIC Domicílios 2018*. 2019. Disponível em: <https://www.cetic.br/tics/domicilios/2018/individuos/C1/expandido>.

CRAN. *What is R?* 2018. Disponível em: <https://www.r-project.org/about.html>. Acesso em: 29 out. 2019.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua*. 2017. Disponível em: <https://www.ibge.gov.br/estatisticas-novoportal/sociais/trabalho/17270-pnad-continua.html?edicao=23205>. Acesso em: 21 Mar. 2019.

KOTU, V.; DESHPANDE, B. *Data Science: Concepts and Practice*. Elsevier Science, 2018. ISBN 9780128147627. Disponível em: <https://books.google.com.br/books?id=-nt8DwAAQBAJ>. Acesso em: 25 Out. 2019.

LABRINIDIS, A.; JAGADISH, H. V. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 5, n. 12, p. 2032–2033, 2012.

MORI, Y. *Handbook of Computational Statistics: Concepts and Methods*. Springer Berlin Heidelberg, 2004. (Springer handbooks of computational statistics). ISBN 9783540404644. Disponível em: <https://books.google.com.br/books?id=MqEBj59xEQoC>. Acesso em: 26 Out. 2019.

MORRIS, G. E. *A Guide to Analyzing (American) Political Data in R*. 2019. Disponível em: <https://www.thecrosstab.com/project/r-politics-guide/guide.html>. Acesso em: 17 Mar. 2019.

O'NEIL, C.; SCHUTT, R. *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, 2013. (Listen Alaska). ISBN 9781449363901. Disponível em: <https://books.google.com.br/books?id=ycNKAQAAQBAJ>. Acesso em: 28 Out. 2019.

PRABHAKARAN, S. *Top 50 ggplot2 Visualizations - The Master List (With Full R Code)*. 2017. Disponível em: <http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>. Acesso em: 28 out. 2019.

TRIBUNAL SUPERIOR ELEITORAL. *Repositório de dados eleitorais*. 2019. Disponível em: <http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais>. Acesso em: 18 Mar. 2019.