Lista II: Análise de dados

Estes exercícios devem ser realizados usando funções do `tidyverse`. Será necessário importar dados, combiná-los, modificar e resumir de variáveis e criar algumas visualizações e tabelas. Complementarmente, usar *loops* e criar mapas pode facilitar o trabalho, embora não seja algo necessário. Leia as instruções com atenção antes de começar a escrever o seu código.

Descrição

Nesta lista, seu objetivo será analisar a produção de teses e dissertações de programas dos pós-graduação notas 4, 5, 6 e 7 na Capes das áreas de *Sociologia* ou *Ciência Política e Relações Internacionais*, entre os anos de 1987 e 2022. De quebra, este exercício ajudará você a se familiarizar com os trabalhos finais já produzidos sobre seu problema de pesquisa – facilitando o levantamento de materiais para uma futura revisão de literatura.

Usaremos várias bases de dados nessa tarefa:

- `programas.csv`, contendo informações sobre os programas de pós-graduação em Sociologia e Ciência Política e Relações Internacionais, incluindo seus estados e notas na Capes.
- `ANO.csv`, diversos arquivos CSV com informações sobre todas as teses e dissertações defendidas em dado ano.

Tarefas

1. Importação

Carregue todas as planilhas `ANO.csv` e combine-as em um único *tibble* que contenha todos as teses e dissertações do período analisado. Carregue também a planilha `programas.csv` e faça um *join* com o *tibble* anterior para adicionar informações sobre os programas de pós-graduação. Em seguida, filtre apenas as observações de sua área de interesse e de programas com notas 4, 5, 6 e 7 da Capes. *Faça todas as análises utilizando dados de uma única área.* Caso necessário, crie, modifique e ajuste variáveis antes de prosseguir.

Dicas:

- É possível carregar cada arquivo separadamente e depois combiná-los em um único `tibble` usando a função `bind_rows()` do `tidyverse`.
- É possível usar a função `list.files` para listar todos os arquivos CSV em um diretório; também é possível usar *loops* ou a função `map_df`, do `tidyverse`, para tentar automatizar a importação dos vários arquivos `ANO.csv`.
- Alguns trabalhos na base foram defendidos em programas de pós-graduação posteriormente descontinuados ou transformados, o que significa que suas notas de avaliação têm *missing*. Decida a melhor forma de lidar com estes casos.

2. Seleção de palavras-chave

Escolha 3 palavras-chave relevantes para o seu problema de pesquisa. Use essas palavras para filtrar os títulos e/ou resumos dos trabalhos e selecione apenas as observações que contenham pelo menos uma dessas palavras. Obtenha pelo menos 30 trabalhos antes de prosseguir com a análise.

Dicas:

- Considere todas as variações possíveis de uma palavra-chave (ex.: "democracia" e "Democracia") ou padronize todos os textos para minúsculas. Queremos textos que mencionem qualquer uma das três palavras-chave, isto é, `palavra1 OU palavra2 OU palavra3`.
- É possível usar a função `str_detect` do `tidyverse` para verificar se uma palavra-chave está presente em um texto.

3. Evolução ao longo do tempo

Crie uma visualização que reporte de forma sucinta e informativa a produção de teses e dissertações no seu tema ao longo do tempo. Selecione o tipo de visualização que julgar mais adequado para mostrar a frequência absoluta de trabalhos defendidos por ano.

4. Diferenças regionais

Calcule o total de trabalhos defendidos ao longo de todo o período por estado e crie duas visualizações: em uma, reporte a frequência de trabalhos por região; em outra, reporte a frequência de trabalhos por unidade da federação.

Dicas:

- Para seguir o princípio de *data-ink ratio* do Tufte, organize e ordene geometrias para facilitar a visualização de padrões.
- Considere reportar a frequência de trabalhos por região e por unidade da federação usando mapas para facilitar a visualização de padrões.

5. Produção por programa

Calcule o total de teses e de dissertações defendidas *por programa* de pós-graduação. Feito isso, reporte em uma tabela o número de trabalhos defendidos pelos 10 programas com maior produção. A tabela final deve reportar **4** colunas: nome do programa; nota na Capes; total de dissertações; e total de teses defendidas no programa. Apresente o resultado da tabela ordenando os programas pelo total de teses defendidas, do maior para o menor. *O resultado precisa ser uma tabela, e não output de console*.

Dicas:

- Considere o usar o pacote `gt` para criar tabelas bem formatadas.

6. Exportação

Crie uma base menor que contenha apenas as seguintes variáveis: ano, estado, programa, título, resumo e autor(a). Exporte essa base para uma planilha de Excel – você poderá consultá-la futuramente.

Entrega

Envie um único arquivo PDF, gerado usando o nosso *template* em quarto, no *Google Classroom*, isto é, *não envie scripts ou outros arquivos auxiliares*. Certifique-se também de que o código no seu PDF esteja visível trocando `echo = FALSE` por `echo = TRUE` na seguinte linha do *template*, que está perto do início do documento:

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
```

Cada seção do seu documento (sub-títulos antecedidos por `##`) deve conter o código que você escreveu para responder o item correspondente da tarefa. Fique à vontade para escrever texto adicional para explicar o que você fez em cada seção.