Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

16 de novembro de 2022

Lista 1: Computação eficiente (dados em memória)

Computação em Estatística para dados e cálculos massivos

Tópicos especiais em Estatística 2

Prof. Guilherme Rodrigues

César Augusto Fernandes Galvão (aluno colaborador)

Gabriel Jose dos Reis Carvalho (aluno colaborador)

1. **As questões deverão ser respondidas em um único relatório *PDF* ou *html*, produzido usando as funcionalidades do *Rmarkdown* ou outra ferramenta equivalente.**
2. **O aluno poderá consultar materiais relevantes disponíveis na internet, tais como livros, *blogs* e artigos.**
3. **O trabalho é individual. Suspeitas de plágio e compartilhamento de soluções serão tratadas com rigor.**
4. **Os códigos *R* utilizados devem ser disponibilizados na integra, seja no corpo do texto ou como anexo.**
5. **O aluno deverá enviar o trabalho até a data especificada na plataforma Microsoft Teams.**
6. **O trabalho será avaliado considerando o nível de qualidade do relatório, o que inclui a precisão das respostas, a pertinência das soluções encontradas, a formatação adotada, dentre outros aspectos correlatos.**
7. **Escreva seu código com esmero, evitando operações redundantes, visando eficiência computacional, otimizando o uso de memória, comentando os resultados e usando as melhores práticas em programação.**

Nessa lista, utilizamos os pacotes `vroom` e `data.table` para analisar, com rapidez computacional e eficiente uso de memória, dados públicos sobre a vacinação contra a Covid-19.

Questão 1: leitura eficiente de dados

a) Utilizando códigos R, crie uma pasta (chamada *dados*) em seu computador e faça o *download* dos arquivos referentes aos estados do Acre, Alagoas, Amazonas e Amapá, disponíveis no endereço eletrônico a seguir. *https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao/resource/5093679f-12c3-4d6b-b7bd-07694de54173?inner_span=True*

Dica: Veja os slides sobre *web scraping* disponibilizados na página da equipe na plataforma MS Teams, em *Materiais de estudo*, na aba *arquivos*; Eles permitem a imediata identificação dos endereços dos arquivos a serem baixados. Use *wi-fi* para fazer os downloads!

b) Usando a função `p_load` (do pacote `pacman`), carregue o pacote `vroom` (que deve ser usado em toda a Questão 1) e use-o para carregar o primeiro dos arquivos baixados para o R (*Dados AC - Parte 1*). Descreva brevemente o banco de dados.

c) Qual é o tamanho total (em Megabytes) de todos os arquivos baixados (use a função `file.size`)? Qual é o espaço ocupado pelo arquivo *Dados AC - Parte 1* na memória do `R` (use a função `object.size`) e no Disco rígido (*HD*)? Comente os resultados.

d) Repita o procedimento do item **b)**, mas, dessa vez, carregue para a memória apenas os casos em que a vacina aplicada foi a Janssen. Para tanto, faça a filtragem usando uma conexão `pipe()`. Observe que a filtragem deve ser feita durante o carregamente, e não após ele.

Quantos megabites deixaram de ser carregados para a memória RAM (ao fazer a filtragem durante a leitura, e não no próprio `R`)?

e) Carregue para o `R` **todos** os arquivos da pasta de uma única vez (usando apenas um comando `R`, sem métodos iterativos), trazendo apenas os casos em que a vacina aplicada foi a Janssen.

Questão 2: manipulação de dados

a) Utilizando o pacote `data.table`, repita o procedimento do item **1e)**, agora mantendo, durante a leitura, todas as vacinas e apenas as colunas `estabelecimento_uf`, `vacina_descricao_dose` e `estabelecimento_municipio_codigo`. Use o pacote `geobr` para obter os dados sobre as regiões de saúde do Brasil (comando `geobr::read_health_region()`). O pacote `geobr` não está mais disponível para download no CRAN; Para instalá-lo, use o link https://cran.r-project.org/src/contrib/Archive/geobr/.

A tabela que relaciona o código do IBGE (`estabelecimento_municipio_codigo`, na tabela de vacinação) e o código de saúde (`code_health_region`, na tabela de regiões de saúde) está disponível pelo link https://sage.saude.gov.br/paineis/regiaoSaude/lista.php?output=html& e nos arquivos da lista.

b) Junte (*join*) os dados da base de vacinações com o das regiões de saúde e descreva brevemente o que são as regiões (use documentação do governo, não se atenha à documentação do pacote). Em seguida, crie as variáveis descritas abaixo:

1. Quantidade de vacinados por região de saúde;
2. Condicionalmente, a *faixa de vacinação* por região de saúde (alta ou baixa, em relação à mediana da distribuição de vacinações).

Crie uma tabela com as 5 regiões de saúde com menos vacinados em cada *faixa de vacinação*.

c) Utilizando o pacote `dtplyr`, repita o procedimento do item **b)** (lembre-se das funções `mutate`, `group_by`, `summarise`, entre outras). Exiba os resultados.

d) Com o pacote `microbenchmark`, comparare o tempo de execução dos itens **b)** e **c)**. Isso é, quando se adota o `data.table` e o `dtplyr`, respectivamente.

Extra: Inclua na comparação a execução usando o próprio `dplyr`. Para isso, primeiro conversta os 3 objetos do item **a)** para a classe `tibble`.