Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

04 agosto 2022

Lista 1

Prof. Guilherme Rodrigues

Computação em Estatística para dados e cálculos massivos

Tópicos especiais em Estatística 1

Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

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* de todos os arquivos disponíveis no endereço eletrônico a seguir. *https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao/resource/5093679f-12c3-4d6b-b7bd-07694de54173?inner_span=True*

Extra: Faça isso automatizando os downloads e direcionando-os para uma mesma pasta. Esta pasta deve ser criada com código R. Sugestão: faça com que a máquina confira se a pasta existe e crie a pasta apenas se não existir.

```
pasta <- "dados"
if (file.exists(pasta)) {
  print("A pasta já existe")
  rm(pasta)
} else {
  dir.create(pasta)
  rm(pasta)
}

link <- c("https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAC/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAC/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAC/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAL/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAL/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAL/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAM/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAM/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAM/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAP/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAP/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DAP/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DBA/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DBA/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DBA/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DCE/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DCE/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DCE/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DDF/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DDF/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DDF/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DES/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DES/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DES/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DGO/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DGO/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DGO/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMA/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMA/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMA/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMG/part-00000-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMG/part-00001-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMG/part-00002-8c
          ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMS/part-00000-8c
```

```
         ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMS/part-00001-8c
         ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMS/part-00002-8c
         ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMT/part-00000-8c
         ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMT/part-00001-8c
         ,"https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIPNI/COVID/uf/uf%3DMT/part-00002-8c
         )

nome_destino <- c("./dados/AC1.csv","./dados/AC2.csv","./dados/AC3.csv",
                  "./dados/AL1.csv","./dados/AL2.csv","./dados/AL3.csv",
                  "./dados/AM1.csv","./dados/AM2.csv","./dados/AM3.csv",
                  "./dados/AP1.csv","./dados/AP2.csv","./dados/AP3.csv",
                  "./dados/BA1.csv","./dados/BA2.csv","./dados/BA3.csv",
                  "./dados/CE1.csv","./dados/CE2.csv","./dados/CE3.csv",
                  "./dados/DF1.csv","./dados/DF2.csv","./dados/DF3.csv",
                  "./dados/ES1.csv","./dados/ES2.csv","./dados/ES3.csv",
                  "./dados/GO1.csv","./dados/GO2.csv","./dados/GO3.csv",
                  "./dados/MA1.csv","./dados/MA2.csv","./dados/MA3.csv",
                  "./dados/MG1.csv","./dados/MG2.csv","./dados/MG3.csv",
                  "./dados/MS1.csv","./dados/MS2.csv","./dados/MS3.csv",
                  "./dados/MT1.csv","./dados/MT2.csv","./dados/MT3.csv")

# Rodar o comando abaixo apenas uma vez para baixar!!!!
# download.file(link, nome_destino) {como já baixei, deixo esta parte como comentário}
```

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. Descreva brevemente o banco de dados.

Extra: explore essa amostra sem o comando explícito de download.

```
p_load(vroom,fs,tidyfst)

AC1 <- vroom("./dados/AC1.csv",
             locale = locale("br", encoding = "UTF-8"),
             num_threads = threads)
```

Descrição: Este banco de dados é a primeira de três partes que compõe o banco de dados sobre registros de vacinação - COVID 19 do estado do Acre. Além desse estado, trabalharemos também com os registros dos estados do Alagoas, Amazonas, Amapá, Bahia, Ceará, Espírito Santo, Goiás, Maranhão, Minas Gerais, Mato Grosso do Sul, Mato Grosso, além do Distrito Federal.

A disponibilidade dos arquivos é feita pelo DataSUS de forma pública e gratuita. Para mais informações, acesse DataSUS.

O dado de cada um dos estados é disponibilizado na forma de arquivos *.csv*, divididos em 3 partes. O tamanho dos arquivos difere levemente, ocupando entre *55Mb* e *70Mb* de memória rígida cada.

As tabelas constam informações individuais de cada vacinado. Cada linha da tabela é um indivíduo. Cada coluna da tabela é uma observação, característica, variável, do indivíduo.

c) Quantos arquivos totalizam nossos dados? Qual é o tamanho total (em Megabytes) de todos os arquivos?

A quantidade de arquivos é 39. O tamanho total desses arquivos é 2521516539b (2.35Gb)

d) Repita o procedimento do item b), mas, dessa vez, carregue para a memória apenas os casos em que a vacina aplicada foi a Astrazeneca. 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`)?

```
filtro <- "findstr ASTRAZENECA/FIOCRUZ C:\\Users\\toled\\Documents\\GitHub\\CE3\\dados\\AC1.csv"

AC1_AZ <- vroom(pipe(filtro),
                locale = locale("br", encoding = "UTF-8"),
                num_threads = threads)

economia <- format(object.size(AC1)-object.size(AC1_AZ),units="auto")
```

43.8 Mb deixaram de ser carregados para a memória RAM.

e) Carregue para o `R` todos os arquivos da pasta de uma única vez (usando apenas um comando `R`, sem métodos iterativos).

```
pasta_arquivos <- "./dados/"
nomes_arquivos <- list.files(pasta_arquivos)
nomes_arquivos <- str_c(pasta_arquivos, nomes_arquivos)

df <- vroom(nomes_arquivos,
            locale = locale("br", encoding = "UTF-8"),
            num_threads = threads)
```

Questão 2: manipulação de dados

a) Utilizando o pacote `data.table`, repita o procedimento do item 1e), agora mantendo, durante a leitura, apenas as 3 primeiras colunas. Use o pacote `geobr` para obter os dados sobre as regiões de saúde do Brasil (procure as funções do `geobr`). Junte (*join*) os dados da base de vacinações com o das regiões de saúde.

Descreva brevemente o que são as regiões (use documentação do governo, não se atenha à documentação do pacote).

```
p_load(data.table)

df <- nomes_arquivos %>%
  map(fread,
      drop=c(4:17,19:28,30:32),
      nThread=threads) %>%
  rbindlist()

p_load(geobr)

rs <- read_health_region(
  year = 2013,
  macro = FALSE,
  simplified = FALSE,
  showProgress = TRUE
  )
```

```
rs <- as.data.table(rs)

index <- fread("./index.csv",select=(4:5),nThread = threads)

colnames(df)[4] <- "regiao ibge"
colnames(index) <- c("regiao ibge","regiao saude")
index$`regiao ibge`<- as.factor(index$`regiao ibge`)
index$`regiao saude` <- as.factor(index$`regiao saude`)
df$`regiao ibge` <- as.factor(df$`regiao ibge`)
colnames(rs)[1] <- "regiao saude"
rs$`regiao saude` <- as.factor(rs$`regiao saude`)

df <- merge(df,index, by="regiao ibge")

junto <- merge(df, rs, by = "regiao saude")
```

Região de Saúde: espaço geográfico contínuo constituído por agrupamentos de Municípios limítrofes, delimitado a partir de identidades culturais, econômicas e sociais e de redes de comunicação e infraestrutura de transportes compartilhados, com a finalidade de integrar a organização e o planejamento de ações e serviços de saúde. Fonte:SciELO

b) No *datatable* obtido no item a), crie as variáveis descritas abaixo considerando apenas os pacientes registrados para a segunda dose:

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*.

Observação: os itens **a)** e **b)** podem ser executados de modo encadeado, usando o operador de pipe.

Tabelas:

regiao saude	n	name_health_region	code_state	abbrev_state	name_state
31041	345	Manhuaçu	31	MG	Minas Gerais
31046	322	Nanuque	31	MG	Minas Gerais
31050	267	Passos Piumhi	31	MG	Minas Gerais
31067	315	Sete Lagoas	31	MG	Minas Gerais
31076	339	Vespasiano	31	MG	Minas Gerais

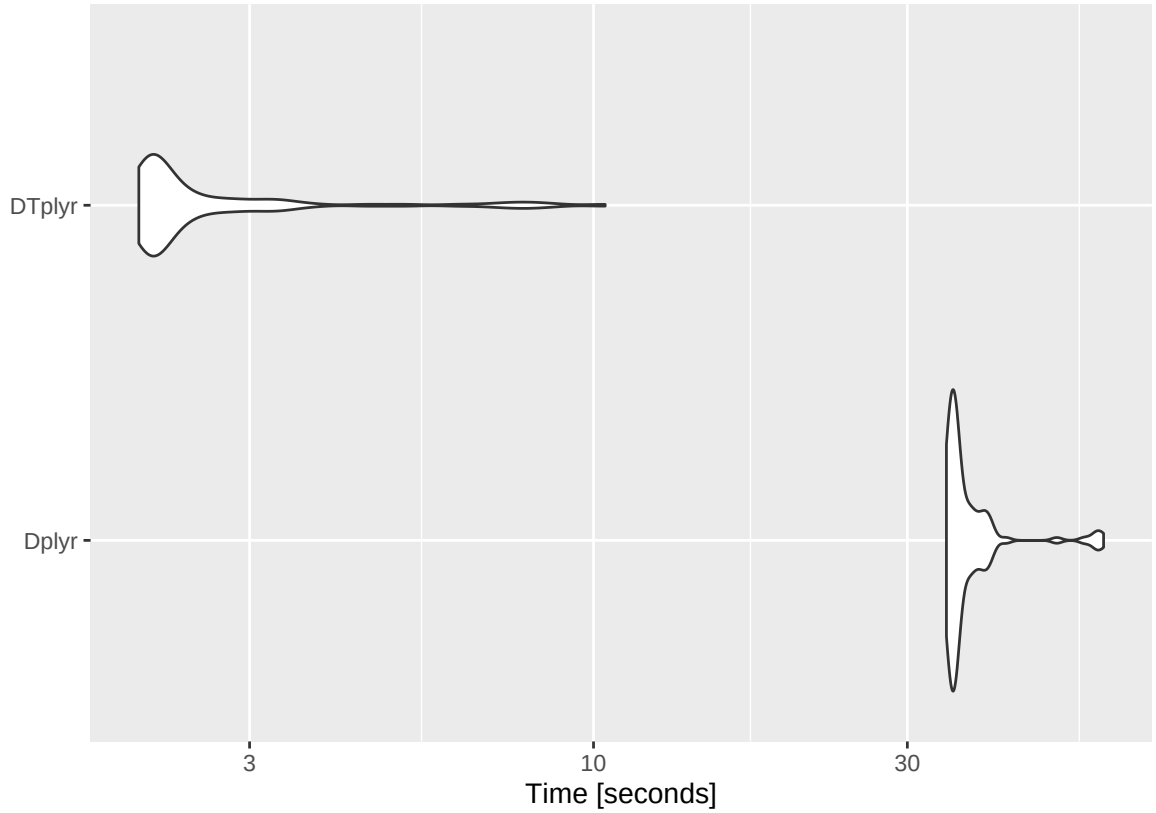
regiao saude	n	name_health_region	code_state	abbrev_state	name_state
29019	2963	Ribeira do Pombal	29	BA	Bahia
29026	3261	Teixeira de Freitas	29	BA	Bahia
13006	3288	Regional Purus	13	AM	Amazônas
29018	3292	Porto Seguro	29	BA	Bahia
51015	3351	Vale do Peixoto	51	MT	Mato Grosso

c) Utilizando o pacote `dtplyr`, repita o procedimento dos itens a) e b) (lembre-se das funções `mutate`, `group_by`, `summarise`, entre outras). Garanta que você conseguiu criar um objeto com *lazy evaluation* e outro resgatado todos os dados para a memória. Exiba os resultados.

De fato, se olharmos o espaço ocupado em memória do objeto original 1.5238714×10^9, do objeto *lazy* 1.5238727×10^9e dos objetos em conjunto 1.5238727×10^9, observamos o objeto *lazy* não está ocupando espaço extra em memória em relação ao objeto original.

d) Com o pacote `microbenchmark`, comparare o tempo de execução do item c) quando se adota as funções do `dtplyr` e do `dplyr`.

Tempo de execução dos códigos usando `dplyr` e `dtplyr`:



Plotando o `microbenchmark` realizado 100 vezes, podemos observar que o tempo de execução do código em `dtplyr` é consideravelmente mais rápido se comparado ao tempo de execução com o `dplyr`.

Considerações finais:

Este PDF é um relatório formatado a fim de entrega do trabalho de forma direta e elegante. Para acesso ao código completo com comentários sobre os processos, favor observar o arquivo *esboco.r* que segue em anexo na entrega.