Exercício Programa (EP) 03

Ana C V de Melo - Alexandre Locci

MAC113 - FEA diurno – 2020

1 EP: Informações sobre a Covid-19 no mundo

Este EP utiliza um conjunto de informações disponibilizadas pelo "European Union Open Data Portal" (https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data).

2 Preparação

1. Abra o RStudio.
2. Na janela de *Files* (canto direito inferior) crie um diretório com seu nome e/ou identificação do EP.
3. Copie do *moodle* todos os arquivos disponíveis para o EP atual.

4. Atualize o diretório de execução do RStudio para a sessão atual: clique em *Session/Set Working Directory/Choose Directory...* e escolha o diretório onde colocou os seus arquivos.
5. Abra o script disponível e pode iniciar a solução do EP.

3 Sobre este EP

Apesar de existirem atualmente vários dados sobre a Covid-19, muitas vezes queremos informações que não são exibidas diretamente nos sites que armazenam os dados.

3.1 Os dados utilizados:

1. Planilha "mundoCovid19.cvs", na qual temos as seguintes colunas de informações (ela está escrita em inglês):
    - **dateRep**: data completa (dia/mês/ano)
    - **day**: dia
    - **month**: mês
    - **year**: ano
    - **cases**: número de casos confirmados
    - **deaths**: número de óbitos confirmados
    - **countriesAndTerritories**: país
    - **geoId**: identificção geográfica
    - **countryterritoryCode**: código do país
    - **popData2018**: população em 2018
    - **continentExp**: continente do país

Antes de iniciar a solução, veja a planilha para entender as informações disponíveis.

3.2 Objetivo do EP

Extrair novos dados juntando informações dessa planilha. Os dados disponibilizados nas planilhas são difíceis de serem comparados por causa do volume de informação.

Neste EP, vamos construir novas funções para conseguir comparar a quantidade de óbitos nos países considerando semanas de notificação. Temos como objetivo final construir um data-frame onde teremos apenas os países com o maior número de óbitos em cada continente. Isso dá a ideia da evolução do número de óbitos desses países a cada semana, desde janeiro até a última semana na planilha. Para isso, precisamos definir algumas funções que auxiliarão a solução do problema.

4 Sua Tarefa será:

No script parcial deste EP, *EP3.script.R*, temos já definidos os nomes das funções e os seus respectivos parâmetros, assim como os valores resultado. Não mude nada disso. Você pode criar novas funções se achar necessário, criar novas variáveis... não há problema nenhum desde que as funções solicitadas realizem o que foi pedido (vamos testar cada uma das funções na avaliação dos EPs).

O seu script deverá implementar as funções para cada um dos exercícios definidos abaixo. **Não mude os nomes das funções fornecidas no script** porque precisaremos desses nomes para conseguir corrigir os EPs. V. precisa apenas complementar cada uma das funções pedidas.

Inicialmente o script lê a planilha (já implementado no script fornecido):

- "mundoCovid19.csv" (descrita acima)

  obs: Usaremos em especial a configuração (*stringAsFactors = FALSE*) para que os nomes dos países e continentes sejam tratados como caracteres puros (não retire isso do script, é para facilitar as operações que vocês irão realizar).

4.1 Exercício 1

Ler a planilha dada e criar o respectivo data-frame: código disponível no script fornecido. Você só precisa executar essa parte do script.

```
df_covidMundo    <- read.csv(file = "mundoCovid19.csv",
                             header = TRUE, sep = ",",
                             fill = TRUE,
                             stringsAsFactors = FALSE)  # lê a planilha)
```

4.2 Exercício 2

Quando trabalhamos com dados, principalmente em grande quantidade, precisamos fazer alguns ajuste nestes dados para que não atrapalhem o processamento. Sendo assim, faça uma função (subs_NA_por_zero) que dada uma coluna de um data-frame, substitui todos os objetos "NA" pelo valor 0. Iremos substituir todos os objetos "NA" da coluna "deaths" do data-frame "df_covidMundo" pelo valor 0 (alguns valores iniciais estavam vazios). **Essa função já deve ter sido implementada no EP2. Pode copiá-la para este novo EP. Execute-a para que o seu data-frame não tenha mais a informação NA na coluna de "deaths"**.

```
head(df_covidMundo)
```

```
##      dateRep day month year cases deaths countriesAndTerritories geoId
## 1 18/05/2020  18     5 2020   262      1             Afghanistan    AF
```

```
## 2 17/05/2020  17     5 2020     0     NA             Afghanistan    AF
## 3 16/05/2020  16     5 2020  1063     32             Afghanistan    AF
## 4 15/05/2020  15     5 2020   113      6             Afghanistan    AF
## 5 14/05/2020  14     5 2020   259      3             Afghanistan    AF
## 6 13/05/2020  13     5 2020   280      5             Afghanistan    AF
##   countryterritoryCode popData2018 continentExp
## 1                  AFG    37172386         Asia
## 2                  AFG    37172386         Asia
## 3                  AFG    37172386         Asia
## 4                  AFG    37172386         Asia
## 5                  AFG    37172386         Asia
## 6                  AFG    37172386         Asia
```

```
df_covidMundo["deaths"] <- subs_NA_por_zero(df_covidMundo$deaths)
head(df_covidMundo)
```

```
##      dateRep day month year cases deaths countriesAndTerritories geoId
## 1 18/05/2020  18     5 2020   262      1             Afghanistan    AF
## 2 17/05/2020  17     5 2020     0      0             Afghanistan    AF
## 3 16/05/2020  16     5 2020  1063     32             Afghanistan    AF
## 4 15/05/2020  15     5 2020   113      6             Afghanistan    AF
## 5 14/05/2020  14     5 2020   259      3             Afghanistan    AF
## 6 13/05/2020  13     5 2020   280      5             Afghanistan    AF
##   countryterritoryCode popData2018 continentExp
## 1                  AFG    37172386         Asia
## 2                  AFG    37172386         Asia
## 3                  AFG    37172386         Asia
## 4                  AFG    37172386         Asia
## 5                  AFG    37172386         Asia
## 6                  AFG    37172386         Asia
```

A função *head()* imprime na tela as primeiras linhas de um data-frame...

Observe que na linha 2 o NA foi substituído por zero!

4.3 Exercício 3

Dado um continente, qual o país que registrou o maior número de óbitos naquele continente durante todo o período da planilha? Por exemplo, podemos perguntar qual país da "America" que registrou o maior número de óbitos.

Faça uma função (pais_mais_obitos_continente), que dados como parâmetros de entrada um data-frame com as informações existentes no df_covidMundo e um continente, retorne o país com o maior número de óbitos naquele continente e o total de óbitos no país. Essa função está parcialmente definida no script fornecido.

Aqui segue o exemplo de uso dessa função para o continente "America".

```
result <- pais_mais_obitos_continente(df_covidMundo, "America")
cat("\n País com maior número de óbitos no continente: America - número de óbitos \n")
```

```
##
##  País com maior número de óbitos no continente: America - número de óbitos
```

```
cat(result[1], " -- ", result[2])
```

```
## United_States_of_America  --  89562
```

Veja que podemos fazer perguntas semelhantes para outros continentes.

4.4 Exercício 4

Agora que já sabemos como calcular o país com o maior número de óbitos em um continente (vide exercício anerior), queremos gerar um data-frame com a informação de todos os continentes e os respectivos países que registraram o maior número de óbitos.

Para gerar esse data-frame, você deverá fazer uma função (continentes_paises_obitos) que tem como parâmetros de entrada um data-frame com as informações existentes no df_covidMundo, e dá como resultado um data-frame com a informação de todos os continentes, os respectivos países que registraram o maior número de óbitos e esse número de óbitos.

Usando a função definida, criamos um novo data-frame e imprimimos o resultado a seguir.

```
df_cont_paises <- continentes_paises_obitos(df_covidMundo)
cat("\n Países com maior número de óbitos em cada continente")
```

```
##
##  Países com maior número de óbitos em cada continente
```

```
print(df_cont_paises)
```

```
##   Continente                     Pais NumObitos
## 1       Asia                     Iran      6988
## 2     Europe           United_Kingdom     34636
## 3     Africa                    Egypt       630
## 4    America United_States_of_America     89562
## 5    Oceania                Australia        98
```

4.5 Exercício 5

Neste exercício, você já tem duas funções implementadas (dadas no script fornecido) para calcular a qual semana do ano pertence uma data (semana_do_ano) e uma outra que gera um vetor de números de semanas que começa em uma data inicial e termina em uma data final (gera_vetor_semanas). Neste exercício, você precisa apenas executar as definições das funções e alguns exemplos de como elas são usadas. As datas são dadas no formato (dd/mm/aaaa).

Por exemplo:

```
semana_do_ano("01/01/2020")
```

```
## [1] 0
```

```
semana_do_ano("04/01/2020")
```

```
## [1] 0
```

```
semana_do_ano("05/01/2020")
```

```
## [1] 1
```

```
semana_do_ano("08/01/2020")
```

```
## [1] 1
```

Veja que a semana do ano da data "01/01/2020" é 0 (zero). Nesta função, uma semana começa sempre no domingo. Portanto, como essa data é uma 4a.feira (o ano começou na 4a.feira), essa é ainda a semana zero do ano. Da mesma forma, como o dia "04/01/2020" é um sábado, ele também está na semana zero, ao passo que dia "05/01/2020" é um domingo e conta como o primeiro dia da semana 1 do ano de 2020. Teste outras datas para entender como funciona.

Podemos também gerar um vetor com os números das semanas de uma data inicial a uma data final. Por exemplo,

```
gera_vetor_semanas("01/01/2020", "14/03/2020")
```

```
##  [1]  0  1  2  3  4  5  6  7  8  9 10
```

```
gera_vetor_semanas("01/01/2020", "18/03/2020")
```

```
##  [1]  0  1  2  3  4  5  6  7  8  9 10 11
```

Veja que os vetores gerados começaram com a semana zero (já explicado acima) e terminaram com o número da semana da data final. Teste para outras datas...

```
gera_vetor_semanas("01/05/2020", "14/06/2020")
```

```
## [1] 17 18 19 20 21 22 23 24
```

4.6 Exercício 6

Uma outra coisa interessante para saber, é como o número de óbitos evoluíram em cada país ao longo do tempo. Uma forma de visualizar esses dados de forma mais fácil é considerar essa evolução semana-a-semana. Como já sabemos calcular a semana do ano e gerar os vetores de semanas até a última data na planilha, podemos agora calcular quantos óbitos foram contabilizados em cada semana do ano para cada país (inclusive para o Brasil).

Neste exercício, em particular, queremos ver a evolução dos óbitos a cada semana dos países que foram apontados como de maior número de óbitos em cada continente.

Para auxiliar nisso, você deverá fazer uma função (obitos_por_semana_no_pais) que tem como parâmetros de entrada um data-frame com as informações existentes no df_covidMundo para um país em particular e o vetor de semanas que queremos contabilizar. Ela deverá dar como resultado um vetor contendo o número de óbitos no país em cada uma das semanas.

Usando a função que deverá ser definida, podemos, por exemplo, ver a evolução dos óbitos no Brasil com o código abaixo. Estamos considerando as datas do início de janeiro até a última data contabilizada na planilha.

```
## datas consideradas
data_ini = "01/01/2020"
data_fim = "18/05/2020"

v_semanas <- gera_vetor_semanas(data_ini, data_fim) # vetor de semanas
df_brasil <- subset(df_covidMundo, df_covidMundo$countriesAndTerritories == "Brazil")
v_semana_obitos_pais <- obitos_por_semana_no_pais(df_brasil, v_semanas)
print(data.frame(Semana = v_semanas, Brazil = v_semana_obitos_pais))
```

```
##    Semana Brazil
## 1       0      0
## 2       1      0
## 3       2      0
## 4       3      0
## 5       4      0
## 6       5      0
## 7       6      0
## 8       7      0
## 9       8      0
## 10      9      0
```

```
## 11     10      0
## 12     11     11
## 13     12     81
## 14     13    267
## 15     14    697
## 16     15   1085
## 17     16   1529
## 18     17   2659
## 19     18   3568
## 20     19   4920
## 21     20   1301
```

Cuidado com a leitura dos resultados... na semana 20 parece que os óbitos estão diminuindo, mas na realidade estamos contabilizando apenas 2 dias nesta semana porque na planilha consta apenas esses 2 dias da semana.

Da mesma forma, podemos também ver a evolução dos óbitos nos países com o maior número de óbitos em cada continente semana-a-semana.

```
## Resumo dos países com o maior número de óbitos por continente - cada semana

## datas consideradas
data_ini = "01/01/2020"
data_fim = "18/05/2020"

v_semanas <- gera_vetor_semanas(data_ini, data_fim) # vetor de semanas
df_paises_semanas_obitos <- data.frame(v_semanas)   # primeira coluna do df
                                                    # a ser gerado (as semanas)
df_cont_paises <- continentes_paises_obitos(df_covidMundo) #df - continentes e países
                                                           #com maior número de óbitos
v_paises <- df_cont_paises$Pais                     # o vetor de países

# criar um data-frame com o número de óbitos dos países por semana
for (pais in v_paises){
  # cria o df do país
  df_pais <- subset(df_covidMundo, df_covidMundo$countriesAndTerritories == pais)
  # cria um vetor com o num de óbitos (por semana) para o dado pais
  v_semana_obitos_pais <- obitos_por_semana_no_pais(df_pais, v_semanas)
  # acrescenta a coluna de óbitos por semana para o país ao df a ser gerado
  df_paises_semanas_obitos <- cbind(df_paises_semanas_obitos,
                                    v_semana_obitos_pais)
}

# coloca o header no data-frame (Semana e o nome dos países) no df gerado
names(df_paises_semanas_obitos) <- c("Semana",as.character(v_paises))

cat("\n Resumo dos países com o maior número de óbitos nos continentes - \n
     \t de: ", data_ini, "  até: ", data_fim, "\n " )
```

```
##
##  Resumo dos países com o maior número de óbitos nos continentes -
##
##           de:  01/01/2020   até:  18/05/2020
##
```

```
print(df_paises_semanas_obitos)
```

```
##    Semana Iran United_Kingdom Egypt United_States_of_America Australia
## 1       0    0              0     0                         0         0
## 2       1    0              0     0                         0         0
## 3       2    0              0     0                         0         0
## 4       3    0              0     0                         0         0
## 5       4    0              0     0                         0         0
## 6       5    0              0     0                         0         0
## 7       6    0              0     0                         0         0
## 8       7    4              0     0                         0         0
## 9       8   30              0     0                         0         0
## 10      9   90              1     0                        14         2
## 11     10  390              9     2                        33         1
## 12     11  919            184     5                       213         4
## 13     12  945            967    17                      1447         6
## 14     13  782           3300    28                      5450        17
## 15     14 1072           6299    83                     11620        24
## 16     15  726           6119    70                     18277        13
## 17     16  616           5913    89                     13963        12
## 18     17  517           4718   112                     14051        14
## 19     18  450           3731    97                     12112         4
## 20     19  361           2757    89                     10388         1
## 21     20   86            638    38                      1994         0
```

4.7 Dicas importantes:

- Não modifique o que está escrito no script fornecido. Ele foi feito para ajudá-los a resolver o problema.
- Resolva o problema na ordem em que é sugerida. Alguns precisam de conhecimentos dos exercícios anteriores.

4.8 Exemplo de Execução e como o seu EP será avaliado:

- Exemplos de execução mostrados em cada exercício.
- a correção será feita mediante comparação dos resultados geradas pelo seu EP com os resultados gerados pelos nossos scripts para os dados atualizados. Vamos basicamente fazer novas perguntas usando as funções que foram pedidas em cada exercício e comparar com os nossos resultados.

4.9 O que está sendo fornecido no *moodle*

- a planilha:
  a. *covidMundo19.csv*
- o script da solução parcial do EP
  a. *EP03.script.R*

4.10 O que V. deve entregar:

- V. deve gravar o seu script solução no arquivo *EP02.script.R* (parte da solução já está no script fornecido). **ESTE É O ÚNICO ARQUIVO A SER ENVIADO NO MOODLE - NÃO ENVIE ARQUIVOS .ZIP**.
- Envie o seu arquivo solução para o *moodle* (**precisa ser esse arquivo mesmo e só ele**). Tenha certeza de que você gravou todas as modificações que fez no arquivo antes de entregar.
- Guarde uma cópia para você de todas as listas e EPs que fizer durante a disciplina.
- Observe a **data de entrega**. Só serão recebidos os EPs (pelo próprio sistema) até aquela data.