

Lista Prática 2

Exercício 1 (*swirl*). Usando o pacote *swirl*, acesse o curso “R Programming” e faça as seguintes lições:

a) 10: *lapply and sapply*

b) 12: *Looking at Data*

Adicionalmente, carregue o pacote *swirl* e instale o curso “Getting and Cleaning Data” usando a seguinte função `install_course("Getting and Cleaning Data")`¹ e faça as seguintes lições:

e) 1: *Manipulating Data with dplyr*

e) 2: *Grouping and Chaining with dplyr*

Ao instalar o curso e entrar na lição “1: Manipulating Data with dplyr” ocorrerá um erro e aparecerá uma mensagem informando um endereço de arquivo que precisamos substituir:

`C:/.../Getting_and_Cleaning_Data/Manipulating_Data_with_dplyr/lesson.yaml`

Copie o endereço até a pasta do arquivo (“C:/.../Manipulating_Data_with_dplyr/”) e cole em uma barra de endereço para acessar a pasta. Então, substitua o arquivo “lesson.yaml” pelo disponibilizado em: <https://fhnishida.github.io/fearp/rec2301/lesson.yaml>

Depois, acesse a lição “1: Manipulating Data with dplyr” no *swirl* novamente.

Exercício 2. A base de dados deste exercício vem do site Hospital Compare (<http://hospitalcompare.hhs.gov>) administrado pelo Departamento de Saúde e Serviços Humanos dos EUA. O propósito deste site é prover dados e informação sobre a qualidade dos tratamentos de mais de 4 mil hospitais certificados pelo Medicare. O arquivo que utilizaremos é `outcome-care_modified.csv`², que contém informação sobre taxa de mortalidade de 30 dias para alguns problemas de saúde.

Comece carregando a base de dados com o seguinte código, olhe suas primeiras 10 linhas e observe a sua estrutura:

```
1 outcome = read.csv2("https://fhnishida.github.io/fearp/ecol/outcome-care_modified.csv")
```

¹Caso tenha problemas, instalar manualmente: <http://swirlstats.com/scn/getclean.html>

²Versão resumida da base de dados fornecida no curso da John Hopkins no Coursera

Como exercício, escreva a função `best()` que terá dois argumentos: (I) a sigla de um Estado americano, e (II) o nome do problema de saúde. Essa função, a partir base de dados `outcome`, retorna um vetor de texto com o nome do hospital que possui a melhor (menor) taxa de mortalidade (Death Rate), dadas sigla de Estado e problema de saúde. O nome do hospital está na coluna “Hospital.Name” e os problemas de saúde podem ser “heart attack”, “heart failure” e “pneumonia”. Se houver empate de melhor hospital, o nome do hospital a ser retornado pela função é dado pela ordem alfabética – se “Hospital A” e “Hospital B” estiverem empatados, deve-se retornar “Hospital A”. A função deve ter a seguinte estrutura:

```
1 best = function(estado, problema) {
2   # Filtre a base de dados 'outcome' pela sigla do estado
3
4   # Reordene a coluna taxa de mortalidade da doença selecionada de
5   # forma crescente e, em caso de empate, coloque os nomes dos
6   # hospitais de forma alfabética.
7
8   # Retorne o nome do hospital com a menor taxa de mortalidade
9 }
```

Como referência, seguem alguns exemplos de output da função:

```
1 > best("TX", "heart failure")
2 [1] "FORT DUNCAN MEDICAL CENTER"
3
4 > best("MD", "heart attack")
5 [1] "JOHNS HOPKINS HOSPITAL, THE"
```

Usando a função `best()` escrita, quais são os resultados retornados para os casos:

- a) `estado = "SC"` e `problema = "heart attack"`
- b) `estado = "NY"` e `problema = "pneumonia"`
- c) `estado = "AK"` e `problema = "pneumonia"`

Exercício 3. Considere a base de dados `mtcars` (nativa no R) que contém o consumo de combustível e mais 10 características automobilísticas para 32 carros (modelos 1973-74). Em particular, queremos analisar a relação entre o consumo de combustível (`mpg`, em milhas por galão) e o peso do automóvel (`wt`, em mil libras) pelo seguinte modelo:

$$mpg = \alpha + \beta wt + \varepsilon.$$

- a) Assuma $\hat{\alpha} = 36$ e $\hat{\beta} = -4$ e calcule a soma dos erros quadráticos $\hat{\varepsilon}^2$, tal que o desvio é dado por: $\hat{\varepsilon} = mpg - \widehat{mpg}$. Note que `mpg` e $\widehat{mpg} = \hat{\alpha} + \hat{\beta}wt$ são, respectivamente, vetores com os valores observados e preditos/ajustados da variável de consumo de combustível.
- b) Crie uma função `erro_quad(alpha, beta)` que recebe como inputs possíveis valores de α e β , e retorna a soma dos erros quadráticos do modelo acima a partir da base de dados `mtcars`.

- c) Assuma os seguintes vetores de possíveis valores de α e de β :
`alpha_grid = seq(34, 38, length=11)` e `beta_grid = seq(-6, -2, length=11)`
Crie uma matriz de dimensão 11×11 em que cada linha corresponde a um possível valor de α e cada coluna corresponde a um possível valor de β . Preencha essa matriz usando repetições (`for` ou `while`) e a função `erro_quad()`. Depois, verifique qual par (α, β) minimiza o erro quadrático do modelo acima. (Dica: use a função `which(..., arr.ind = TRUE)` na matriz preenchida para obter os índices de linha e coluna.)

Exercício 4. Carregue bases de dados de GDP para 190 países ranqueados e de informações educacionais a partir do seguinte código:

```
1 gdp = read.csv2("https://fhnishida.github.io/fearp/eco1/data_GDP.csv")
2 educ = read.csv2("https://fhnishida.github.io/fearp/eco1/EDSTATS_Country.csv")
```

- a) Mescle as bases de dados a partir do código de país de 3 dígitos. Quantos códigos tiveram correspondência em ambas bases? Organize a base de dados mesclada de forma decrescente no GDP (de modo que USA fica na última linha). Qual é o país na 13^a linha?
- b) Quais são os rankings médios para os grupos de renda “High income: OECD” e “High income: nonOECD”?