Lista Prática 2

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

a) 9: Functions

b) 10: lapply and sapply

c) 12: Looking at Data

Adicionalmente, instale o curso "Getting and Cleaning Data" usando a seguinte função `swirl::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 editar:

`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, abra o arquivo "lesson.yaml" com o Bloco de Notas, busque na ***linha 205*** *o código* `"Hint: !is.na(c(3, 5, NA, 10)) will negate the previous command, thus telling us what is NOT NA"`*, e* ***adicione um # no início da linha****. Salve o arquivo e acesse a lição 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:

```
outcome = read.csv2("https://fhnishida.github.io/fearp/eco1/outcome-care_modified.csv")
```

[1] Caso tenha problemas, instalar manualmente: `http://swirlstats.com/scn/getclean.html`
[2] Versão 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:*

```
best = function(estado, problema) {
    # Filtre a base de dados pela sigla do estado

    # Reordene a coluna taxa de mortalidade da doença selecionada de forma
    crescente e, em caso de empate, coloque os nomes dos hospitais de forma alfabé
    tica.

    # Retorne o nome do hospital com a menor taxa de mortalidade
    }
```

Como referência, seguem alguns exemplos de output da função:

```
> best("TX", "heart failure")
[1] "FORT DUNCAN MEDICAL CENTER"

> best("MD", "heart attack")
[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 $\alpha = 36$ *e* $\beta = -4$ *e calcule a soma dos desvios quadráticos* ε^2*, tal que o desvio é dado por:* $\varepsilon = mpg - \widehat{mpg}$.

b) Crie uma função `desv_quad(alpha, beta)` *que recebe como inputs possíveis valores de* α *e* β*, e retorna a soma dos desvios 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 valor de* α *e cada coluna corresponde a um valor de* β*. Preencha essa matriz usando a função* `desv_quad()` *e verifique qual par* (α, β) *minimiza o desvio quadrático do modelo acima. (Dica: use a função* `which(..., arr.ind = TRUE)` *na matriz preenchida)*

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

```
gdp = read.csv2("https://fhnishida.github.io/fearp/eco1/data_GDP.csv")
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"?*

c) *Usando* `cut(..., breaks=quantile(...))`*, crie uma nova coluna chamada "groupGDP" que, usando quintis, classifica países em 5 grupos de acordo com seus GDPs. Depois, faça uma tabela entre "groupGDP" e "Income.Group". Quantos países estão entre as nações com maior renda (OECD e nonOECD) e estão entre os quantis em 20% e 40% do GDP?*