Atividade nº 4 – Importação de Dados e Gráficos

Instruções para entrega da lista:

a) O relatório de respostas da lista (desenvolvimento, comandos, resultados,
b) saídas gráficas e **comentários**) deve estar apresentada em documento com extensão `.pdf` ou `.html`, **gerado em R Markdown**. Apresente também o *script* correspondente, em extensão `.R`, com todos os comandos utilizados na solução da presente lista.
c) Os arquivos com o relatório de respostas e o script deverão ser denominados, respectivamente, `064-241_At04-SEUNOME-SEUSOBRENOME.pdf` (ou `.html`).
d) Não esqueça de **se identificar no preâmbulo do arquivo**, além de rotular corretamente as questões cujos comandos e resultados você estará apresentando.
e) Apresente todos os comandos (todos os comandos que funcionaram!) que utilizou para obter os resultados solicitados.
f) **Preserve a ordem** das questões e responda brevemente suas justificativas e comentários.
g) O upload do relatório (extensão `.pdf` ou `.html`) e do script (extensão `.R`) deverão ser efetuados **exclusivamente** no Moodle, até a data marcada.
h) Não hesite em procurar o **Fórum de Dúvidas** do Moodle, caso tenha alguma dúvida com relação à solução da presente lista de exercícios. Caso não resolva, acione o professor. Acostume-se a interagir para obter sugestões de solução das dúvidas.

1. *Uso dos comandos* `read.table()` *e* `read.csv()`. Crie os *data frames* indicados abaixo:

    a. Use a função `read.csv()` e importe o arquivo `Census at School-500.csv` diretamente da URL: https://www.stat.auckland.ac.nz/~wild/d2i/FutureLearn/Census.at.School.500_ages9-15.csv. Qual o contexto deste conjunto de dados? Qual a dimensão deste conjunto de dados? Quais os nomes das variáveis? Use agora a função `read.table()` para importar esse conjunto de dados diretamente da URL citada acima.

    b. Instale o arquivo .csv em um local de sua conveniência e use a função `file.choose()` como argumento da função `read.csv()` para importar o arquivo para o R. Descreva a operação. Liste quais os prós e os contras de se utilizar esse procedimento. Qual o contexto deste conjunto de dados? Qual a dimensão deste conjunto de dados? Quais os nomes das variáveis? Use agora a função `read.table()` para importar esse conjunto de dados diretamente da URL citada acima.

2. O conjunto de dados `skulls{ade4}` apresenta medidas feitas em crânios masculinos da área de Tebas, no Egito. Há cinco amostras de 30 crânios cada uma do período pré-dinástico primitivo (cerca de 4.000 a.C.), do período pré-dinástico antigo (cerca de 3.300 a.C.), das 12ª e 13ª dinastias (cerca de 1.850 a.C.), do período Ptolemaico (cerca de 200 a.C.) e do período Romano (cerca de 150 d.C.). São apresentadas quatro medidas para cada crânio: V1 (amplitude máxima do

crânio); V2 (altura basilobregmática do crânio); V3 (comprimento basiloalveolar do crânio) e V4 (altura nasal do Crânio).

a. Manipule e organize o banco de dados como se segue:
   i. Renomeie as variáveis: `V1` por `ACr`, `V2` por `BBr`, `V3` por `BAl` e `V4` por `ANs`.
   ii. Crie a variável categórica `periodo`, com os níveis `1: período pré-dinástico primitivo`, `2: período pré-dinástico antigo`, `3: 12ª e 13ª dinastias`, `4: período Ptolemaico` e `5: período Romano`. Os 150 sujeitos do banco de dados estão ordenados em ordem crescente de idade, 30 para cada um dos cinco períodos (linha `1`: nível `1` até linha `150`: nível `5`)
   iii. Crie a variável quantitativa `idade`, adotando os seguintes valores: `−4000` para os sujeitos do nível `1` da variável `periodo`; `−3300`, para os sujeitos do nível `2`; `−1850`, para os sujeitos do nível `3`; `−200`, para os sujeitos do nível `4` e `150`, para os sujeitos do nível `5`.

b. Amplie a análise exploratória desses dados usando o R:
   i. Calcule a média de cada uma das medidas, por período. Apresente os resultados em uma matriz (tabela), nomeando o nome de cada linha com o nome do período correspondente (`Primitivo`, `Antigo`, `Dinastias`, `Ptolemaico` e `Romano`)
   ii. Construa um gráfico de linhas das médias de cada uma das medidas (eixo vertical) por `idade` (eixo horizontal).

c. Visualizando o gráfico construído, você diria que as medidas médias mudaram ao longo do tempo? Que conjecturas você levanta sobre o tema?

3. *Uso dos comandos* `scan()` *e* `lower.tri()`. Nesse exercício, trabalha-se os fundamentos da função `scan()`. Antes de prosseguir, leia a seção 7.2 de VENABLES et al. (2020) e pesquise sobre a função `lower.tri()` que irá auxiliá-lo a montar a matriz de correlações (`R`) do exercício. O conjunto de dados `E9-14.DAT` contém os elementos da diagonal de R e aqueles que estão acima (ou abaixo) dessa diagonal. Eles fazem parte de uma matriz de correlações (portanto simétrica) relativa às medidas ossos de crânio, pernas e asas de frangos *white leghorn*. Os dados são aqueles usados em estudo de Dunn (1928). A matriz é simétrica de ordem 6x6. Suas colunas (ou linhas) referem-se às correlações das seguintes variáveis:

```
X1: comprimento de crânio
X2: largura do crânio
X3: comprimento do fêmur
X4: comprimento da tíbia
X5: comprimento do úmero
X6: comprimento da ulna
```

a. Calcule o traço e o determinante da matriz de correlações R.

b. Use o comando `eigen()` para calcular os autovetores da matriz de correlações e calcule a proporção de cada um deles em relação ao traço da matriz R.
c. Compare o traço da matriz R com a soma de seus autovalores. Compare o determinante da matriz R com o produto de seus autovalores

4. As variáveis `speed` e `dist` do conjunto de dados `cars{datasets}` apresentam, respectivamente, a velocidade (mph) e a distância de parada (ft) de 50 carros.

    a. Use a função `plot` para construir um gráfico de dispersão de `speed` (horizontal) vs. `dist` (vertical).
    b. Revise o `plot` básico rotulando o eixo horizontal com "Velocidade, em mpg" e o eixo vertical com "Distância de parada, em ft", adicione um título ao gráfico.
    c. Revise novamente o gráfico alterando o símbolo de plotagem do caracter default (círculos abertos) para triângulos preenchidos em vermelho (`col = "red", pch = 17`).

    Suponha agora que você deseja comparar os ajustes de modelo linear e quadrático aos dados (`speed`, `dist`). Construa esses dois modelos usando os códigos abaixo:

```
modelo.linear <- lm(dist ~ speed, data = cars)
modelo.quadratico <- lm(dist ~ speed + I(speed^2),
data = cars)
```

    d. Construa um diagrama de dispersão de `speed` (horizontal) e `dist` (vertical).
    e. Use a função `abline` com argumento `modelo.linear` e sobreponha o ajuste linear obtido, o tipo de linha deve ser "`dotted`", com o dobro da largura de linha *default*.
    f. Use a função `lines` com argumento `modelo.quadratico` e sobreponha o ajuste quadrático obtido, o tipo de linha deve ser "`longdash`", com o dobro da largura de linha *default*.
    g. Acrescente uma legenda para mostrar os tipos de linha dos ajustes linear e quadrático.
    h. Refazer os itens (d a (g) usando duas cores contrastantes (digamos, vermelho e azul) para os dois ajustes.
    i. Construa um gráfico de resíduos para o ajuste linear com o comando:

        ```
        plot(modelo.linear$residual ~ speed, data = cars)
        ```

    j. Use a função `abline` e adicione no gráfico de resíduos uma linha horizontal azul e grossa (`lwd = 3`) passando por zero.
    k. Existem dois grandes resíduos positivos neste gráfico. Aplique duas vezes a função de `text`, denominando cada um desses resíduos extremos com o rótulo "`POS`" em azul.
    l. Rotule o menor resíduo negativo no gráfico com o rótulo "`NEG`" em vermelho.

5. Suponha que você esteja interessado em apresentar a curva da função de densidade de probabilidade de três membros da família beta. A função de densidade de probabilidade de uma beta com parâmetros de forma `a` e `b` (denotada por `Beta(a, b)`) é dada por:

$$f(y) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1},\ 0 < y < 1$$

Pode-se o gráfico da curva de uma densidade beta, digamos, com os parâmetros de forma `a = 5` e `b = 2`, usando a função `curve()`:

```
curve(dbeta(x, 5, 2), from = 0, to = 1)
```

a. Aplique três vezes a função `curve()` para apresentar as densidades `Beta(2, 6)`, `Beta(4, 4)` e `Beta(6, 2)` no mesmo gráfico. Note que o comando `curve` com o argumento `add = TRUE` adicionará a curva à janela gráfica em uso.
b. Use o comando descrito abaixo para apresentar no título do gráfico a expressão da função de densidade de probabilidade beta:

```
title(expression(f(y)==frac(1,B(a,b))*
      y^{a-1}*(1-y)^{b-1}))
```

c. Use A função `text()` e rotule cada uma das curvas beta com os valores correspondentes dos parâmetros de forma `a` e `b`.
d. Refaça o gráfico construído em (a) usando cores e tipos de linha diferentes para cada uma das três curvas da função de densidade de probabilidade beta.
e. Em vez de usar a função de `text()`, adicione uma legenda ao gráfico que mostre a cor e o tipo de linha de cada uma das três curvas de densidade beta.

Referência:

- ALBERT, J.; RIZZO, M. R by example. New York: Springer, 2012.
- DUNN, L. C. The effect of inbreeding on the bones of the fowl. *Storrs Agricultural Experimental Station Bulletin*, v. 52, p. 1-112, 1928.
- MANLY, B. F. J. *Métodos estatísticos multivariados: uma introdução*. Porto Alegre: Bookman, 2008.
- VENABLES, W. N.; SMITH, D. M., R Core Team. An introduction to R: Notes on R: a programming environment for data analysis and graphics. Vienna, Austria: R Foundation for Statistical Computing, 2020.