Lista 4: Transformando dados em informação

Ana Luisa Sousa de Oliveira

04 de Setembro de 2022

Transformando dados em informação

Questão 1

a) Importe os bancos de dados "bnames.csv.bz2" e "births.csv" para o R e apresente as 6 primeiras observações (linhas) de cada um dos bancos.

```
bnames <- read.csv("C:\\Users\\Ana Luisa\\Downloads\\RStudio\\Documentos R\\bnames.csv",
                   header=TRUE)

births <- read.csv("C:\\Users\\Ana Luisa\\Downloads\\RStudio\\Documentos R\\births.csv",
                   header=TRUE)
```

```
head(bnames)
```

```
##   year    name     prop sex soundex
## 1 1880    John 0.081541 boy    J500
## 2 1880 William 0.080511 boy    W450
## 3 1880   James 0.050057 boy    J520
## 4 1880 Charles 0.045167 boy    C642
## 5 1880  George 0.043292 boy    G620
## 6 1880   Frank 0.027380 boy    F652
```

```
head(births)
```

```
##   year sex births
## 1 1880 boy 118405
## 2 1881 boy 108290
## 3 1882 boy 122034
## 4 1883 boy 112487
## 5 1884 boy 122745
## 6 1885 boy 115948
```

b) No ano de 2003, quais foram os 10 nomes mais populares de meninas?

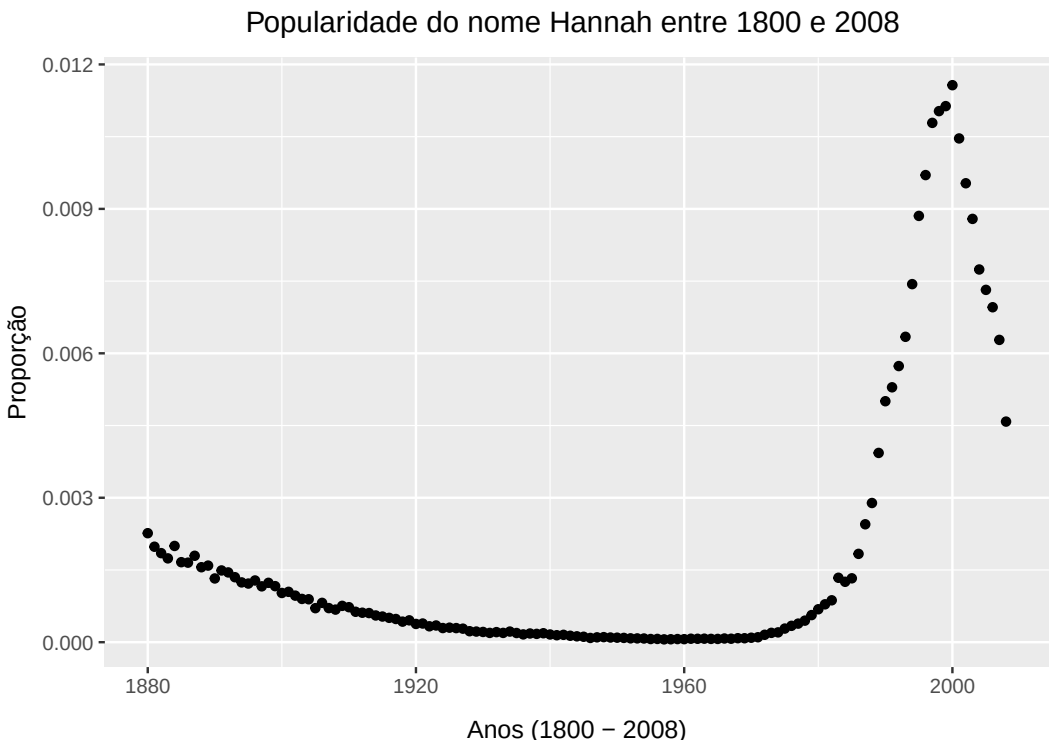
```
meninas <- subset(bnames, sex == "girl")
meninas_2003 <- subset(meninas, year == "2003")
meninas_nomes_2003 <- meninas_2003[order(meninas_2003[["prop"]], decreasing = TRUE),]
library(dplyr)
meninas_2003_top_10_nomes <- select(meninas_2003, name)
head(meninas_2003_top_10_nomes, 10)
```

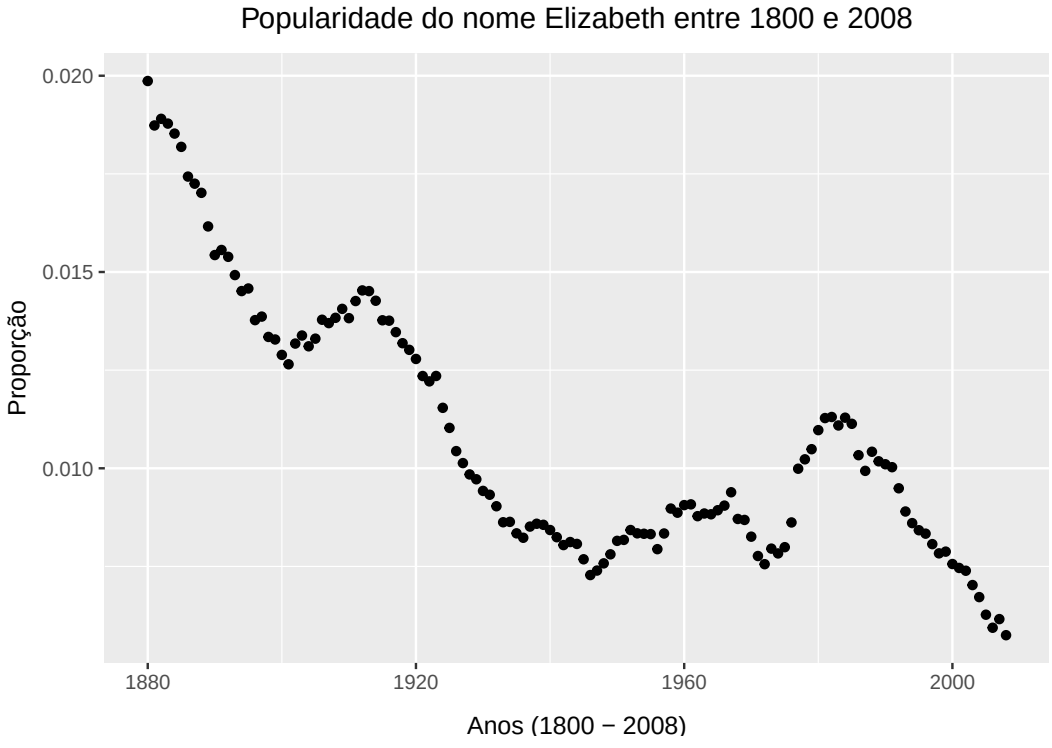
```
##             name
## 252001     Emily
## 252002      Emma
## 252003   Madison
## 252004    Hannah
## 252005    Olivia
## 252006   Abigail
## 252007    Alexis
## 252008    Ashley
## 252009 Elizabeth
## 252010  Samantha
```

c) Escolha seus 2 nomes preferidos da letra b) e faça um gráfico da evolução da popularidade deles ao longo dos anos (1880 - 2008). Considere apenas o sexo feminino.

```
hannah <- subset(meninas, name == "Hannah")
library(dplyr)
hannah_prop <- select(hannah, year, prop)
library(ggplot2)
p <- ggplot(data = hannah_prop) +
  geom_point(mapping = aes(x = year, y = prop)) +
  ggtitle("Popularidade do nome Hannah entre 1800 e 2008 ") +
  xlab('Anos (1800 - 2008)') +
  ylab('Proporção') +
  theme(plot.title = element_text(hjust = 0.5, vjust = 2),
        axis.title.x = element_text(vjust = -2, hjust = 0.5),
        axis.title.y = element_text(hjust = 0.5, vjust = 2))
```

```
elizabeth <- subset(meninas, name == "Elizabeth")
elizabeth_prop <- select(elizabeth, year, prop)
q <- ggplot(data = elizabeth_prop) +
  geom_point(mapping = aes(x = year, y = prop)) +
  ggtitle("Popularidade do nome Elizabeth entre 1800 e 2008") +
  xlab('Anos (1800 - 2008)') +
  ylab('Proporção') +
  theme(plot.title = element_text(hjust = 0.5, vjust = 2),
        axis.title.x = element_text(vjust = -2, hjust = 0.5),
        axis.title.y = element_text(hjust = 0.5, vjust = 2))
```


Popularidade do nome Hannah entre 1800 e 2008
0.012
0.009
0.006
0.003
0.000
Proporção
1880
1920
1960
2000
Anos (1800 − 2008)


Popularidade do nome Elizabeth entre 1800 e 2008
0.020
0.015
0.010
Proporção
1880
1920
1960
2000
Anos (1800 − 2008)

d) Combine os bancos de dados "bnames.csv.bz2" e "births", e adicione uma nova coluna chamada 'n' que mostre o número total de bebês nascidos em cada ano para cada nome.

```
juntos <- left_join(bnames, births)
nascidos_por_nome <- mutate(juntos, n = round(prop * births))
```

e) Descreva sucintamente como usar a função 'str_sub' do pacote 'stringr'.

A função *str_sub* reciclará todos os argumentos para que tenham o mesmo comprimento que o argumento mais longo. Se algum argumento for de comprimento 0, a saída será um vetor de caractere de comprimento zero.

f) Usando a função 'str_sub', adicione uma coluna no banco de dados criado na letra d) contendo a inicial (primeira letra) de cada nome.

```
library(stringr)
letras <- stringr::str_sub(nascidos_por_nome[ , 2], start = 0, end = 1)
primeiras_letras <- mutate(nascidos_por_nome, letter = letras)
```

g) Qual é a inicial mais popular para meninos e para meninas considerando os dados de todos os anos (Considere a soma das frequências absolutas das iniciais - coluna 'n' - durante todo o período)? **Dica: utilize as funções group_by, summarise, dcast e arrange.**

```
meninas_completo <- subset(primeiras_letras, sex == "girl")
meninos_completo <- subset(primeiras_letras, sex == "boy")
freq_letras_meninas <- meninas_completo %>%
        group_by(letter) %>%
        dplyr::summarize(total = sum(n))
freq_letras_meninas_ordem <- arrange(freq_letras_meninas, desc(total))
head(freq_letras_meninas_ordem, 1)
```

```
## # A tibble: 1 x 2
##   letter    total
##   <chr>     <dbl>
## 1 M      16997360
```

Para as meninas a letra mais popular é **M**.

```
meninos_completo <- subset(primeiras_letras, sex == "boy")
freq_letras_meninos <- meninos_completo %>%
        group_by(letter) %>%
        dplyr::summarize(total = sum(n))
freq_letras_meninos_ordem <- arrange(freq_letras_meninos, desc(total))
head(freq_letras_meninos_ordem, 1)
```

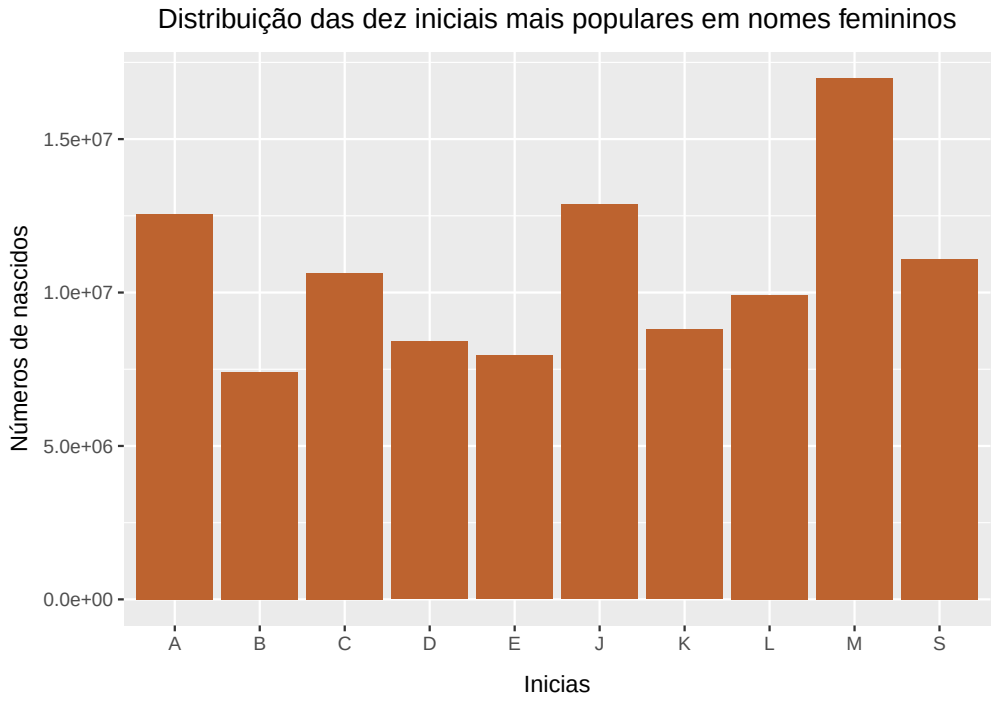
```
## # A tibble: 1 x 2
##   letter    total
##   <chr>     <dbl>
## 1 J      26368455
```

Para os meninos a letra mais popular é **J**.

h) Represente graficamente as informações encontradas na letra g).

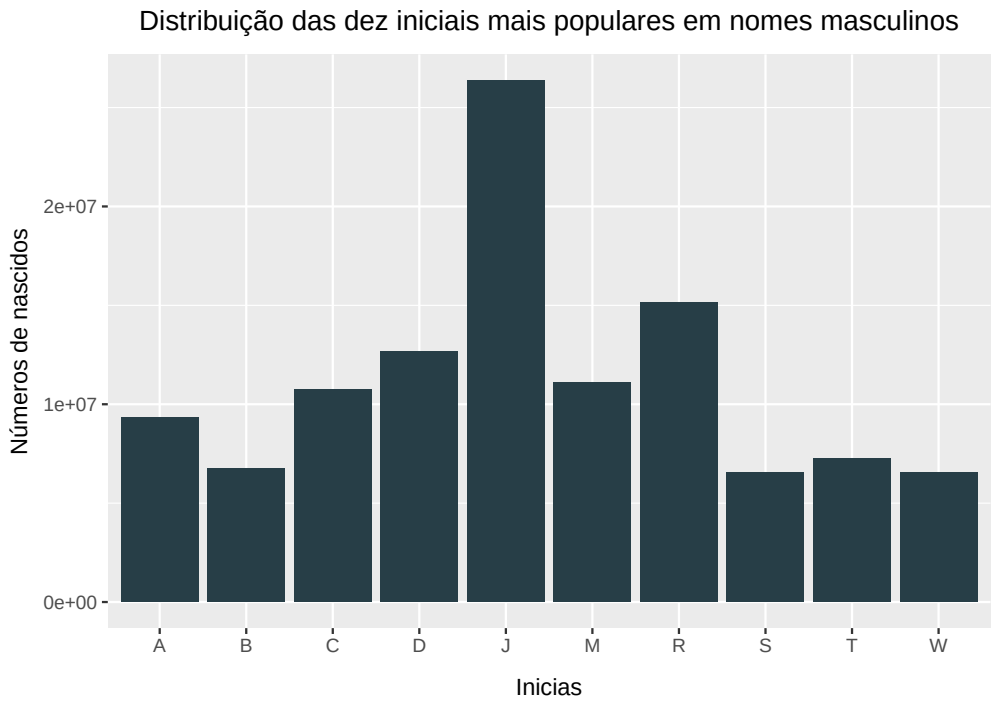
```
top10_letras_meninas <- arrange(freq_letras_meninas_ordem[1:10, ], desc(total))

ggplot() +
  geom_bar(data=top10_letras_meninas, aes(x=letter, y=total),
           stat="identity", fill = "#BD632F") +
  ggtitle("Distribuição das dez iniciais mais populares em nomes femininos") +
  xlab('Inicias') +
  ylab('Números de nascidos') +
  theme(plot.title = element_text(hjust = 0.5, vjust = 2),
        axis.title.x = element_text(vjust = -2, hjust = 0.5),
        axis.title.y = element_text(hjust = 0.5, vjust = 2))
```




```
top10_letras_meninos <- arrange(freq_letras_meninos_ordem[1:10, ], desc(total))

ggplot() +
  geom_bar(data=top10_letras_meninos, aes(x=letter, y=total),
           stat="identity", fill = "#273E47") +
  ggtitle("Distribuição das dez iniciais mais populares em nomes masculinos") +
  xlab('Inicias') +
  ylab('Números de nascidos') +
  theme(plot.title = element_text(hjust = 0.5, vjust = 2),
        axis.title.x = element_text(vjust = -2, hjust = 0.5),
        axis.title.y = element_text(hjust = 0.5, vjust = 2))
```


Distribuição das dez iniciais mais populares em nomes masculinos
Números de nascidos
2e+07
1e+07
0e+00
A
B
C
D
J
M
R
S
T
W
Inicias