Lista 2

Marcelo Saito

30 agosto, 2023

carregando pacotes

```
library(PNADcIBGE)
library(tidyverse)
library(tidylog)
library(ggplot2)
```

1

```
data <- get_pnadc(year=2017,
quarter=4,
selected=FALSE,
vars=c("Ano", "Trimestre", "UF", "V2007", "VD4020", "VD4035"),
design=FALSE,
savedir=tempdir())

data <- data %>%
select(Ano, Trimestre, UF, V2007, VD4020, VD4035)

data <- data %>%
rename(Sexo = V2007,
Renda = VD4020,
Horas_trabalhadas = VD4035)
```

2 Calcule:

i) Renda Média;

```
renda_media <- data %>%
  summarise(renda_media = mean(Renda, na.rm = TRUE))

print(renda_media)
```

```
## # A tibble: 1 x 1
##   renda_media
##         <dbl>
## 1       1931.
```

ii) Variância da renda;

```
var_renda <- data %>%
  summarise(var_renda = var(Renda, na.rm = TRUE))

print(var_renda)
```

```
## # A tibble: 1 x 1
##   var_renda
##       <dbl>
## 1  9543677.
```

iii) Renda média dos homens e das mulheres;

```
renda_media_h <- data %>%
  filter(Sexo == "Homem") %>%
  summarise(renda_media_h = mean(Renda, na.rm = TRUE))

print(renda_media_h)
```

```
## # A tibble: 1 x 1
##   renda_media_h
##           <dbl>
## 1         2078.
```

```
renda_media_m <- data %>%
  filter(Sexo == "Mulher") %>%
  summarise(renda_media_m = mean(Renda, na.rm = TRUE))

print(renda_media_m)
```

```
## # A tibble: 1 x 1
##   renda_media_m
##           <dbl>
## 1         1721.
```

iv) a renda média em cada estado brasileiro;

```
renda_media_estado <- data %>%
  group_by(UF) %>%
  summarise(renda_media_estado = mean(Renda, na.rm = TRUE))
```

```
## group_by: one grouping variable (UF)
```

```
## summarise: now 27 rows and 2 columns, ungrouped
```

```
print(renda_media_estado)
```

```
## # A tibble: 27 x 2
##    UF        renda_media_estado
##    <fct>                  <dbl>
##  1 Rondônia               1811.
##  2 Acre                   1574.
##  3 Amazonas               1593.
##  4 Roraima                2035.
##  5 Pará                   1394.
```

```
##  6 Amapá                  2109.
##  7 Tocantins              1780.
##  8 Maranhão               1075.
##  9 Piauí                  1292.
## 10 Ceará                  1264.
## # i 17 more rows
```

v) Covariância entre a renda e o número de horas trabalhadas;

```
cov_renda_h <- data %>%
  summarise(covariancia = cov(Horas_trabalhadas, Renda, use = "pairwise.complete.obs"))

print(cov_renda_h)
```

```
## # A tibble: 1 x 1
##   covariancia
##         <dbl>
## 1       5777.
```

3 Exemplifique a veracidade da equação, considerando X = Renda, Y = Horas trabalhadas, a = 2 e b = 3.

$$E[aX + bY] = a * E[X] + b * E[Y]$$

resolução:

```
a <- 2

b <- 3

e_x <- data %>%
  summarise(media = mean(Renda, na.rm = TRUE))

e_y <- data %>%
  summarise(media = mean(Horas_trabalhadas, na.rm = TRUE))

l_e <- data %>%
  summarise(esquerdo = mean(a * Renda + b * Horas_trabalhadas, na.rm = TRUE))

l_d <- a * e_x + b * e_y

l_e == l_d
```

```
##      esquerdo
## [1,]    FALSE
```

```
print(l_e)
```

```
## # A tibble: 1 x 1
##   esquerdo
##      <dbl>
## 1    3975.
```

```r
print(l_d)
```

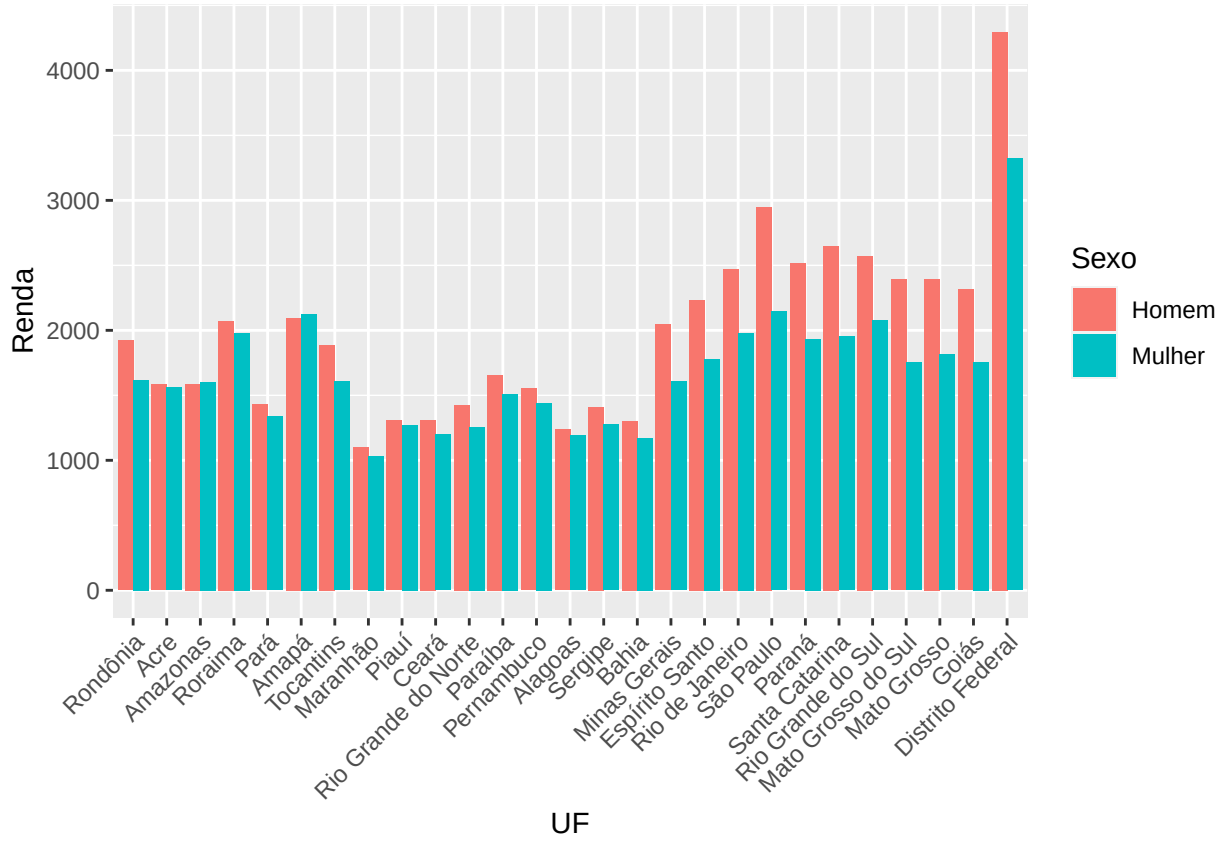
```
##      media
## 1 3974.108
```

4 Apresente um gráfico que permita visualização adequada da média da renda por estado brasileiro e sexo.

resolução:

```r
renda_uf_sexo <- data %>%
  aggregate(Renda ~ UF + Sexo, FUN = mean)

ggplot(renda_uf_sexo, aes(x = UF, y = Renda, fill = Sexo)) +
  geom_bar(stat = "summary", position = "dodge")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



5 Agora trabalharemos explicitamente com a esperança condicional. Note que essa lógica estava implícita nas questões anteriores. Assuma duas variáveis aleatórias, X e Y, tais que X = renda e Y = horas trabalhadas. Calcule:

$$E[X|10 \leq Y \leq 20]$$

```
e_xtalq_10y20 <- data %>%
  filter(Horas_trabalhadas >= 10 & Horas_trabalhadas <= 20) %>%
  summarise(media = mean(Renda, na.rm = TRUE))

print(e_xtalq_10y20)
```

```
## # A tibble: 1 x 1
##   media
##   <dbl>
## 1  940.
```

$$E[X|Y \geq 20]$$

```
e_xtalq_ymaior20 <- data %>%
  filter(Horas_trabalhadas >= 20) %>%
  summarise(media = mean(Renda, na.rm = TRUE))

print(e_xtalq_ymaior20)
```

```
## # A tibble: 1 x 1
##   media
##   <dbl>
## 1 2015.
```

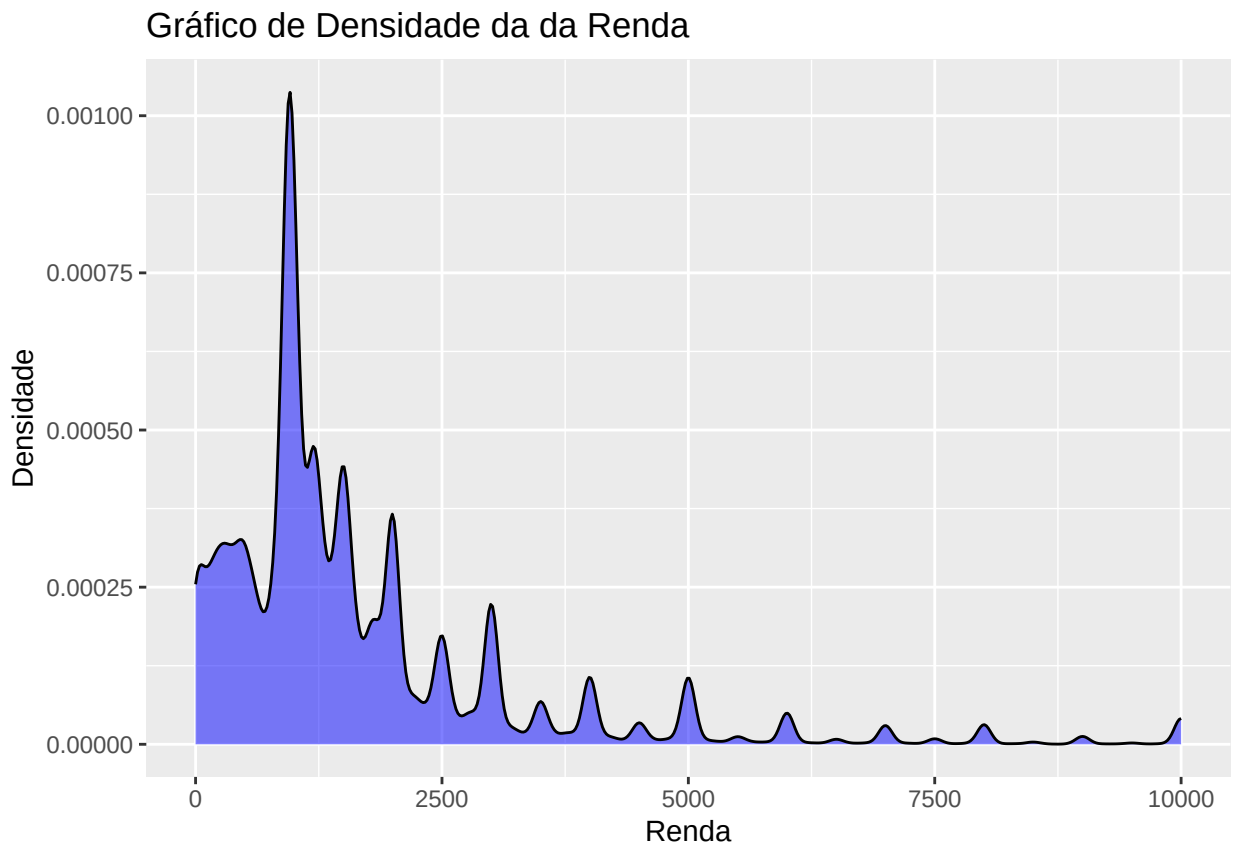
6 Para os itens seguintes (i a iv), remova todas as observações cuja renda seja superior a 10.000 reais.

i) apresente um gráfico de dispersão da variável renda. Interprete;

```
renda_10k <- data %>%
  filter(Renda <= 10000)

# ggplot(renda_10k, aes(x = Renda , y = Horas_trabalhadas)) +
#   geom_point()

ggplot(renda_10k, aes(x = Renda)) +
geom_density(fill = "blue", alpha = 0.5) +
labs(title = "Gráfico de Densidade da da Renda",
x = "Renda",
y = "Densidade")
```



ii) qual é a probabilidade de que, ao retirarmos aleatoriamente uma observação (um indivíduo) dessa base de dados, sua renda esteja entre 1000 e 2000 reais? Apenas para propósitos didáticos, ignore o erro amostral e trate a sua base de dados como uma população (não faça isso em sua pesquisa);

```
intervalo_1k_2k <- data %>%
  filter(Renda >= 1000 & Renda <= 2000) %>%
  nrow()


total10k <- renda_10k %>%
  nrow()

p <- round(intervalo_1k_2k/total10k * 100, 2)

print(p)
```

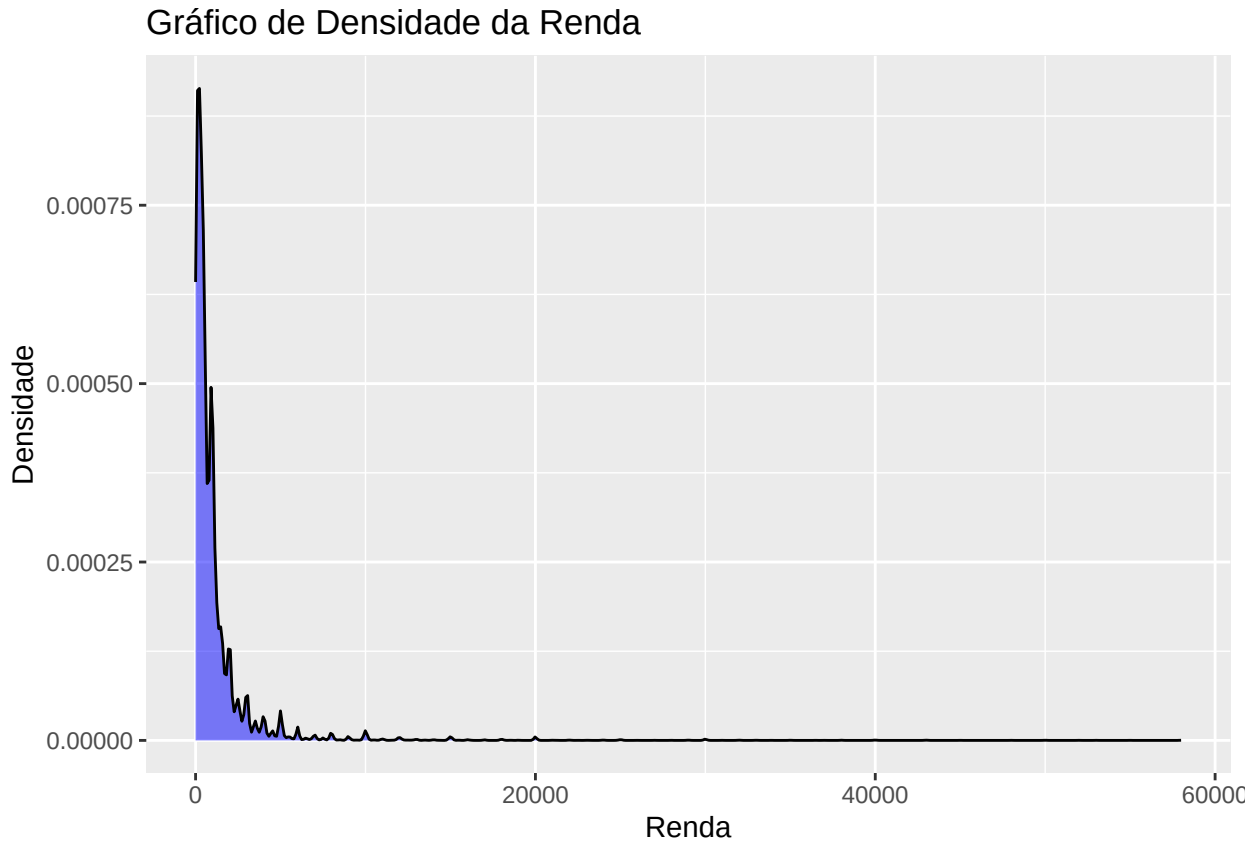
```
## [1] 38.9
```

iii) apresente um gráfico de dispersão da renda dado que as horas trabalhadas (Y) sejam menores ou iguais a 20;

```
h_menorigual_20 <- data %>%
  filter(Horas_trabalhadas <= 20)
```

```r
# ggplot(h_menorigual_20, aes(x = Renda, y = Horas_trabalhadas)) +
#   geom_point()

ggplot(h_menorigual_20, aes(x = Renda)) +
geom_density(fill = "blue", alpha = 0.5) +
labs(title = "Gráfico de Densidade da Renda",
x = "Renda",
y = "Densidade")
```



iv) calcule:

$$P(1000 < X < 2000 | Y \leq 20)$$

```r
x_1 <- data %>%
  filter(Renda > 1000 & Renda <2000)

y_1 <- data %>%
  filter(Horas_trabalhadas <= 20)
```