SME0822 - Análise Multivariada e Aprendizado Não-Supervisionado

Atividade 1

Francisco Rosa Dias de Miranda - 4402962

Setembro 2022

```
# Bibliotecas do R utilizadas
library(bookdown)
library(pander)
library(tidyverse)
library(ggExtra)
library(ggpubr)
library(knitr)
```

Exercício 1.4

```
dat14 <- read_delim("data/companies.csv", ";", show_col_types = F)
```

Carregamos o conjunto de dados com o lucro e total de vendas das 10 maiores empresas do mundo. A partir do gráfico na Figura @ref(fig:graf1) nota-se uma relação aparentemente linear entre as variáveis.

```
p <- dat14 |> ggplot(aes(x = sales, y = profits)) +
  geom_point() +
  labs(x= "Vendas (bilhões)", y = "Lucro (bilhões)",
       title = "Lucros versus quantidade de vendas" ) +
  theme_pubr()
  ggMarginal(p,type = "histogram")
```

Vamos agora obter algumas medidas descritivas das variáveis lucro e vendas. Na Tabela @ref{tab1} temos as médias, covariâncias e coeficiente de correlação linear de Pearson.

O valor de r obtido indica-nos uma forte associação linear entre as variáveis lucro e vendas, conforme havíamos sugerido a partir dos indícios gráficos da Figura @ref{graf1}.

```
dat14 |> select(sales, profits) |>
  summarise("$\\bar{x_1}$" = mean(sales),"$\\bar{x_2}$" = mean(profits),
            "$s_{11}$" = var(sales), "$s_{22}$" = var(profits),
            "$s_{12}$" = cov(sales,profits), "$r_{12}$" = cor(sales,profits)) |>
  kable(caption = "Médias e covariâncias das 10 maiores empresas do mundo")
```

Table 1: Médias e covariâncias das 10 maiores empresas do mundo

$\bar{x_1}$	$\bar{x_2}$	s_{11}	s_{22}	s_{12}	r_{12}
155.603	14.704	7476.453	26.19032	303.6186	0.686136

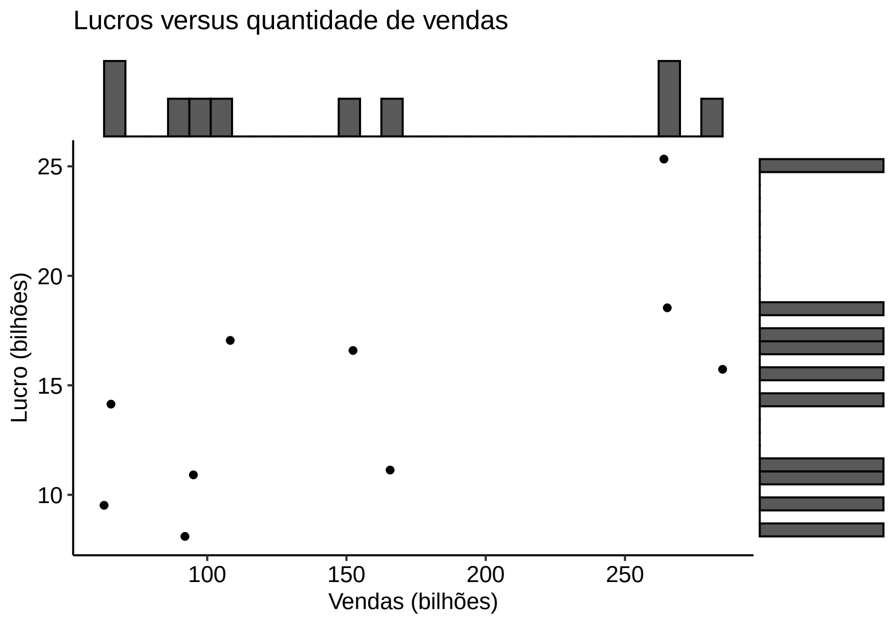


Figure 1: Scatter plot com distribuição marginal do lucro e vendas das 10 maiores empresas do mundo

Exercicio 1.5

Vamos agora estudar a associação entre o total de recursos de cada empresa (x_3) e as variáveis do exercício anterior. A Figura @ref{graf2} sugere-nos um padrão de dispersão linear com correlação negativa, que investigaremos adiante.

```
dat14 |> select(!Company) |>
  pivot_longer(cols = !"assets") |>
  ggplot(aes(y = assets)) +
  geom_point(aes(x = value)) +
  facet_wrap(~name, scales = "free_x") +
  theme_pubclean()
```

```
vars <- dat14 |> select(!Company)

# Vetor de Medias
xbar <- vars |> map(~mean(.))
# Matriz de covariâncias S
S <- cov(vars)
# Matriz de correlações de Pearson
R <- cor(vars)
```

Dessa forma, obtemos:

$$\bar{\boldsymbol{x}} = \begin{bmatrix} 155.603 \\ 14.704 \\ 710.911 \end{bmatrix}$$

$$\boldsymbol{S_n} = \begin{bmatrix} 7476.45 & 303.62 & -35575.96 \\ 303.62 & 26.19 & -1053.83 \\ -35575.96 & -1053.83 & 237054.27 \end{bmatrix}$$

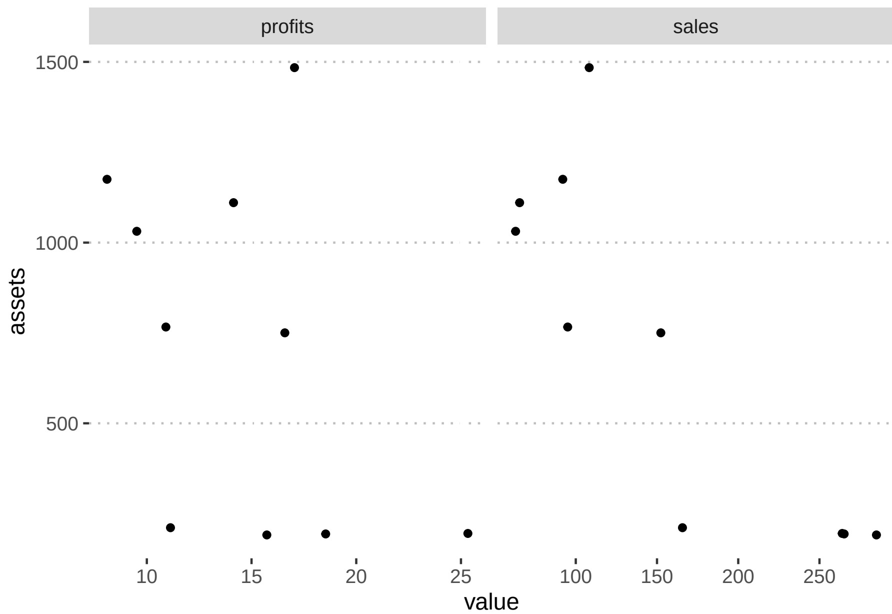


Figure 2: Scatter plot com distribuição marginal do lucro e vendas das 10 maiores empresas do mundo

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.6861 & -0.8451 \\ 0.6861 & 1 & -0.4229 \\ -0.8451 & -0.4229 & 1 \end{bmatrix}$$

Exercicio 1.22

```
data22 <- read_delim("data/oxygen.csv")
```

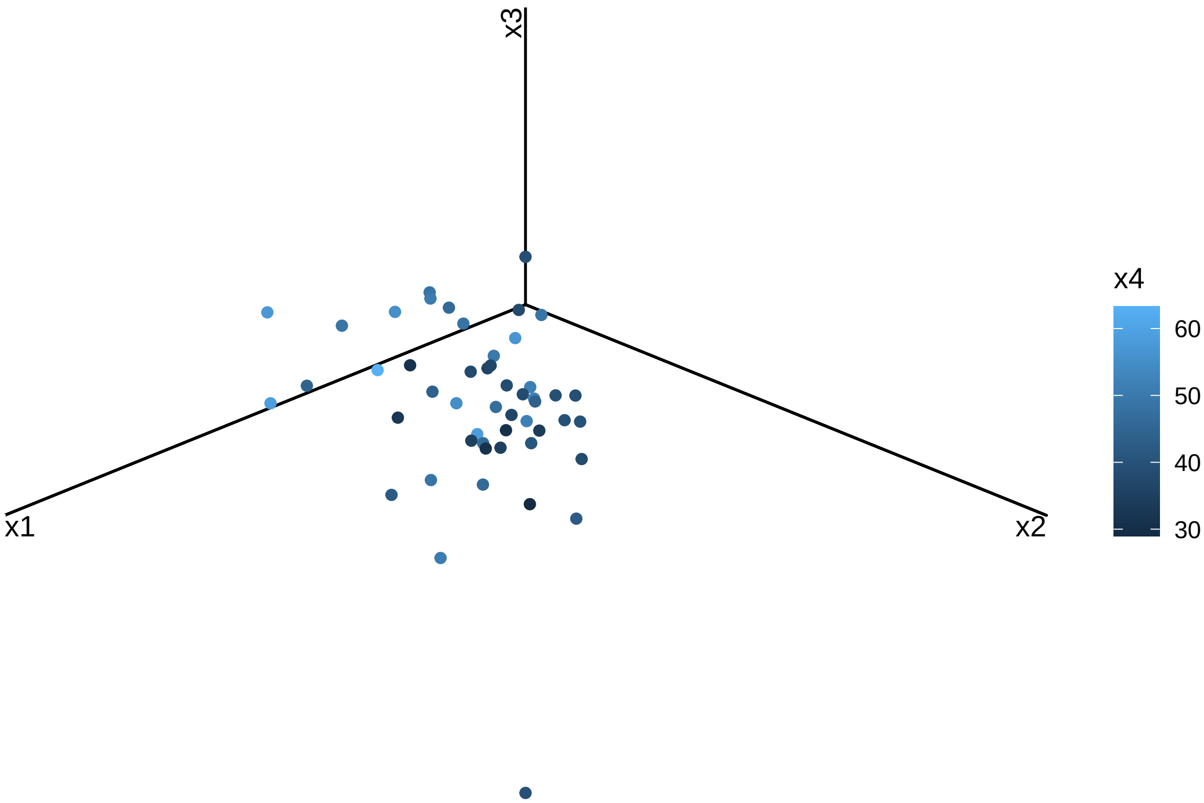
```
## Rows: 50 Columns: 5
## -- Column specification --------------------------------------------------------------------------
## Delimiter: ";"
## chr (1): sex
## dbl (4): x1, x2, x3, x4
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
a<- data22[c(2,3,4)]
```

A partir do gráfico 3d, podemos visualizar dois outliers nos cantos do plot.

```
library(gg3D)

data22 |>
  ggplot(aes(x=x1, y=x2, z=x3, color = x4)) +
  theme_void() +
  axes_3D() +
    stat_3D() +
      labs_3D(
    labs=c("x1", "x2", "x3"),
    hjust=c(0,1,1), vjust=c(1, 1, -0.2), angle=c(0, 0, 90))
```



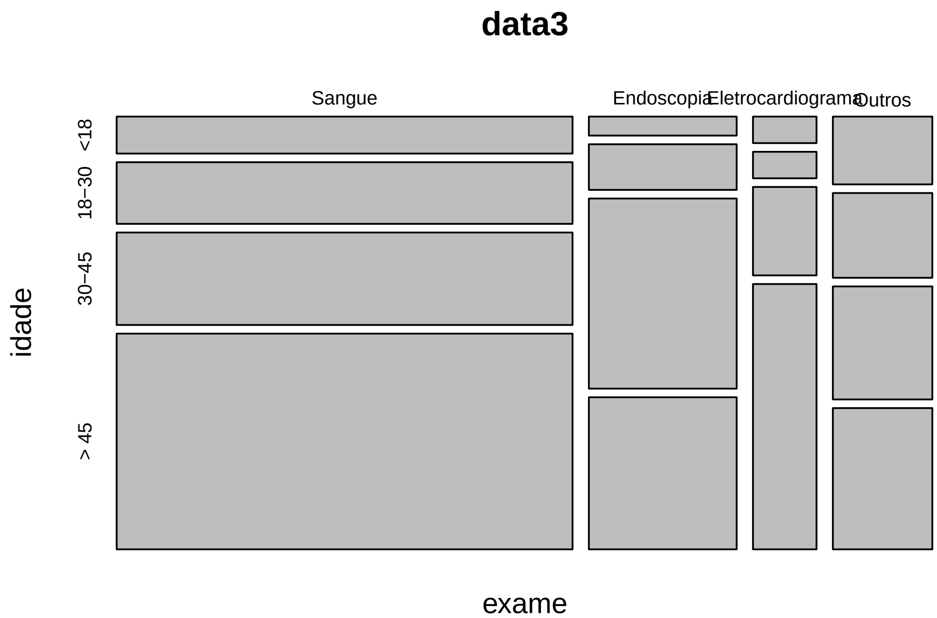
Exercicio 3

```
data3 <- matrix(c( 60, 10, 6, 24,
                   100, 24, 6, 30,
                   150, 100, 20, 40,
                   350, 80,60,50), 4, 4)

dimnames(data3) = list(
  exame = c("Sangue", "Endoscopia", "Eletrocardiograma", "Outros"),
  idade = c("<18", "18-30", "30-45", "> 45")
  )
```

Podemos visualizar graficamente a associação entre Exame e Idade ao plotarmos um gráfico de mosaico com as duas variáveis, na Figura @ref{graph3}.

```
mosaicplot(data3)
```



O teste qui-quadrado é um teste estatístico que verifica o quão distante os valores observados estão dos esperados sob independência. Na tabela @ref{tab3} Obtivemos um p-valor menor do que 0.05, dessa forma rejeitamos a hipótese nula, de independência entre as variáveis.

Assim, a um nível de significância de 5% podemos concluir que existe associação entre as variáveis Exame e Idade.

```
pander(chisq.test(as.table(data3)),
      caption = "Teste de independência entre as variáveis Exame e Idade")
```

Table 2: Teste de independência entre as variáveis Exame e Idade

Test statistic	df	P value
78.2	9	3.685e-13 * * *

O Coeficiente T de Tschuprow é uma medida de associação entre duas variáveis nominais entre 0 e 1. Na Tabela 2 Obtivemos $T = 0,15$, que indica uma associação fraca entre as variáveis Exame e Idade.

```
library(DescTools)

TschuprowT(data3)
```

```
## [1] 0.1532396
```

Exercicio 6

A partir dos coeficientes de $Q(x_1, x_2, x_3)$ obtemos

```
M <- matrix(c( 2, -1, 2,
              -1, 1, 0,
              2, 0, -3), 3,3, byrow = T)
```

$$A = \begin{bmatrix} 2 & -1 & 2 \\ -1 & 1 & 0 \\ 2 & 0 & -3 \end{bmatrix}$$