Atividade 1 - Tópicos Especiais em Estatística Computacional

Tatiana Alejandra Moreno Avila

Web Scraping e Análise Exploratória de Dados

O objetivo desta atividade é realizar uma análise exploratória de dados da página web da OCDE, extraindo as informações por meio de web scraping. Para isso, utilizamos a linguagem de programação R, na qual, através do pacote chromote, que automatiza a navegação web e simula um navegador, realizamos a extração dos dados do site, os quais fazem referência a informações do setor econômico, para posteriormente poder visualizá-los por meio de diferentes gráficos.

Web Scraping

1. Configuração do ambiente

Inicialmente instalamos os pacotes e as bibliotecas necessárias:

```
library(chromote)
library(rvest)
library(dplyr)
library(stringr)
library(purrr)
library(janitor)
library(corrplot)
library(DataExplorer)
library(ggplot2)
```

2. Extração de tabelas com chromote

Realizamos o web scraping através da função chromote para ler a URL correspondente e obter o HTML por meio do rvest

```
$frameId
[1] "DA29A05BEAC3E32B74495DC52DFE0A62"

$loaderId
[1] "155DE62228841A0CAD9E95D12D9E7A9B"

$isDownload
[1] FALSE
```

3.Extraccion de las tablas de datos

```
Número de tabelas encontradas: 2

Processando tabela 1 ...
Processando tabela 2 ...

Tabelas extraídas com sucesso: 2

[1] TRUE
```

4. Visualização da tabela

```
# Carregar apenas a tabela 2
dados <- data.frame(all_tables[2])
# Exibir primeiras linhas
head(dados)
```

```
  Combined.measure Combined.measure_1 Life.expectancy..Female..At.birth
2  Unit of measure    Unit of measure                               Years
3   Reference area               <NA>                                <NA>
4        Australia               <NA>                                85.5
5          Austria               <NA>                                84.6
6          Belgium               <NA>                                84.4
7           Canada               <NA>                                84.8
```

```
  Life.expectancy..Female..65.years Life.expectancy..Male..At.birth
2                             Years                           Years
3                              <NA>                            <NA>
4                              87.9                            81.7
5                              86.9                            80.1
6                              87.1                            80.2
7                              87.7                            80.9
  Life.expectancy..Male..65.years     Employment.rate..From.55.to.59.years
2                           Years Percentage of population in the same age
3                            <NA>                                     <NA>
4                            85.5                                     75.6
5                            83.9                                     77.4
6                            84.1                                     73.1
7                            85.1                                     73.5
      Employment.rate..From.60.to.64.years
2 Percentage of population in the same age
3                                     <NA>
4                                     58.3
5                                     32.3
6                                     39.1
7                                     53.7
      Employment.rate..From.65.to.69.years
2 Percentage of population in the same age
3                                     <NA>
4                                     31.0
5                                     10.4
6                                      7.5
7                                     27.0
  Old.age.to.working.age.ratio..65.years.or.over
2     Percentage of population aged 20-64 years
3                                          <NA>
4                                          28.6
5                                          32.5
6                                          34.0
7                                          31.7
  Effective.labour.market.exit.age..Female
2                                    Years
3                                     <NA>
4                                     64.4
5                                     60.9
6                                     61.3
7                                     63.5
  Effective.labour.market.exit.age..Male
```

```
2                                    Years
3                                     <NA>
4                                     65.1
5                                     61.6
6                                     61.1
7                                     64.9
  Expected.years.after.labour.market.exit..Female
2                                            Years
3                                             <NA>
4                                             23.5
5                                             25.5
6                                             25.2
7                                             24.0
  Expected.years.after.labour.market.exit..Male
2                                          Years
3                                           <NA>
4                                           20.5
5                                           21.6
6                                           22.2
7                                           20.2
```

Análise exploratória de dados

Vamos ver a estrutura dos dados

```
# Verificar estrutura dos dados
str(dados)
```

```
'data.frame':   57 obs. of  14 variables:
 $ Combined.measure                                : chr  "Unit of measure" "Reference area" "
 $ Combined.measure_1                              : chr  "Unit of measure" NA NA NA ...
 $ Life.expectancy..Female..At.birth               : chr  "Years" NA "85.5" "84.6" ...
 $ Life.expectancy..Female..65.years               : chr  "Years" NA "87.9" "86.9" ...
 $ Life.expectancy..Male..At.birth                 : chr  "Years" NA "81.7" "80.1" ...
 $ Life.expectancy..Male..65.years                 : chr  "Years" NA "85.5" "83.9" ...
 $ Employment.rate..From.55.to.59.years            : chr  "Percentage of population in the sa
 $ Employment.rate..From.60.to.64.years            : chr  "Percentage of population in the sa
 $ Employment.rate..From.65.to.69.years            : chr  "Percentage of population in the sa
 $ Old.age.to.working.age.ratio..65.years.or.over : chr  "Percentage of population aged 20-6
 $ Effective.labour.market.exit.age..Female        : chr  "Years" NA "64.4" "60.9" ...
 $ Effective.labour.market.exit.age..Male          : chr  "Years" NA "65.1" "61.6" ...
```

```
 $ Expected.years.after.labour.market.exit..Female: chr  "Years" NA "23.5" "25.5" ...
 $ Expected.years.after.labour.market.exit..Male  : chr  "Years" NA "20.5" "21.6" ...
```

Notemos que inicialmente os dados são de classe categórica, vamos eliminar as colunas e linhas vazias

```
#Eliminar as duas primeiras linhas
dados<-data.frame(dados[-c(1,2), -2])

head(dados)
```

```
  Combined.measure Life.expectancy..Female..At.birth
4        Australia                               85.5
5          Austria                               84.6
6          Belgium                               84.4
7           Canada                               84.8
8            Chile                               81.9
9         Colombia                               77.1
  Life.expectancy..Female..65.years Life.expectancy..Male..At.birth
4                              87.9                            81.7
5                              86.9                            80.1
6                              87.1                            80.2
7                              87.7                            80.9
8                              85.6                            77.2
9                              82.8                            70.3
  Life.expectancy..Male..65.years Employment.rate..From.55.to.59.years
4                            85.5                                 75.6
5                            83.9                                 77.4
6                            84.1                                 73.1
7                            85.1                                 73.5
8                            82.9                                 64.8
9                            79.8                                 61.2
  Employment.rate..From.60.to.64.years Employment.rate..From.65.to.69.years
4                                 58.3                                 31.0
5                                 32.3                                 10.4
6                                 39.1                                  7.5
7                                 53.7                                 27.0
8                                 54.1                                 33.4
9                                 48.3                                 35.8
  Old.age.to.working.age.ratio..65.years.or.over
4                                           28.6
5                                           32.5
```

```
6                                          34.0
7                                          31.7
8                                          20.9
9                                          14.5
  Effective.labour.market.exit.age..Female
4                                     64.4
5                                     60.9
6                                     61.3
7                                     63.5
8                                     63.7
9                                     60.7
  Effective.labour.market.exit.age..Male
4                                    65.1
5                                    61.6
6                                    61.1
7                                    64.9
8                                    67.3
9                                    67.8
  Expected.years.after.labour.market.exit..Female
4                                            23.5
5                                            25.5
6                                            25.2
7                                            24.0
8                                            21.7
9                                            21.2
  Expected.years.after.labour.market.exit..Male
4                                           20.5
5                                           21.6
6                                           22.2
7                                           20.2
8                                           16.2
9                                           13.0
```

Em seguida, identificamos as variáveis numéricas para transformá-las de categóricas para numéricas.

```r
# Identificar colunas que parecem numéricas mas são caracteres
columnas_a_convertir <- sapply(dados, function(x) {
  all(grepl("^[0-9.]+$", x) | is.na(x))
})

# Converter apenas essas colunas
dados[columnas_a_convertir] <- lapply(dados[columnas_a_convertir], as.numeric)
```

Agora podemos ver que algumas variáveis mudaram de numéricas para categóricas

```
str(dados)
```

```
'data.frame':   55 obs. of  13 variables:
 $ Combined.measure                                : chr  "Australia" "Austria" "Belgium" "Ca
 $ Life.expectancy..Female..At.birth               : num  85.5 84.6 84.4 84.8 81.9 77.1 80 81
 $ Life.expectancy..Female..65.years               : num  87.9 86.9 87.1 87.7 85.6 82.8 83.8 8
 $ Life.expectancy..Male..At.birth                 : num  81.7 80.1 80.2 80.9 77.2 70.3 74.8 7
 $ Life.expectancy..Male..65.years                 : num  85.5 83.9 84.1 85.1 82.9 79.8 81.1 8
 $ Employment.rate..From.55.to.59.years            : num  75.6 77.4 73.1 73.5 64.8 61.2 61.9 8
 $ Employment.rate..From.60.to.64.years            : num  58.3 32.3 39.1 53.7 54.1 48.3 46.4 5
 $ Employment.rate..From.65.to.69.years            : num  31 10.4 7.5 27 33.4 35.8 21.8 14.9 1
 $ Old.age.to.working.age.ratio..65.years.or.over : num  28.6 32.5 34 31.7 20.9 14.5 17.5 35
 $ Effective.labour.market.exit.age..Female        : num  64.4 60.9 61.3 63.5 63.7 60.7 62.2 6
 $ Effective.labour.market.exit.age..Male          : num  65.1 61.6 61.1 64.9 67.3 67.8 66.7 6
 $ Expected.years.after.labour.market.exit..Female: num  23.5 25.5 25.2 24 21.7 21.2 21.1 21
 $ Expected.years.after.labour.market.exit..Male  : num  20.5 21.6 22.2 20.2 16.2 13 14.8 15
```

A seguir, temos algumas estatísticas básicas de cada uma das variáveis

```
# Dimensões do dados
cat("Dimensões do dados: ", dim(dados), "\n\n")
```

```
Dimensões do dados:  55 13
```

```
# Resumo estatístico
summary(dados)
```

```
 Combined.measure   Life.expectancy..Female..At.birth
 Length:55          Min.   :64.20
 Class :character   1st Qu.:79.88
 Mode  :character   Median :83.40
                    Mean   :81.85
                    3rd Qu.:84.80
                    Max.   :87.80
                    NA's   :3
 Life.expectancy..Female..65.years Life.expectancy..Male..At.birth
```

```
Min.   :78.00                     Min.   :58.60
1st Qu.:83.95                     1st Qu.:73.12
Median :86.05                     Median :78.30
Mean   :85.32                     Mean   :76.70
3rd Qu.:87.12                     3rd Qu.:80.75
Max.   :89.90                     Max.   :82.50
NA's   :3                         NA's   :3
Life.expectancy..Male..65.years Employment.rate..From.55.to.59.years
Min.   :75.30                   Min.   :40.30
1st Qu.:80.03                   1st Qu.:66.38
Median :83.05                   Median :75.50
Mean   :82.16                   Mean   :72.85
3rd Qu.:84.35                   3rd Qu.:80.17
Max.   :85.50                   Max.   :88.60
NA's   :3                       NA's   :7
Employment.rate..From.60.to.64.years Employment.rate..From.65.to.69.years
Min.   :20.40                        Min.   : 4.80
1st Qu.:41.08                        1st Qu.:13.55
Median :53.95                        Median :21.70
Mean   :51.51                        Mean   :23.06
3rd Qu.:62.60                        3rd Qu.:31.27
Max.   :79.70                        Max.   :50.90
NA's   :7                            NA's   :7
Old.age.to.working.age.ratio..65.years.or.over
Min.   : 4.40
1st Qu.:23.20
Median :31.30
Mean   :29.06
3rd Qu.:35.30
Max.   :55.40
NA's   :1
Effective.labour.market.exit.age..Female
Min.   :58.40
1st Qu.:61.40
Median :62.95
Mean   :63.07
3rd Qu.:64.47
Max.   :69.20
NA's   :5
Effective.labour.market.exit.age..Male
Min.   :60.3
1st Qu.:63.0
Median :64.2
```

```
 Mean   :64.3
 3rd Qu.:65.5
 Max.   :69.8
 NA's   :6
 Expected.years.after.labour.market.exit..Female
 Min.   :18.10
 1st Qu.:21.15
 Median :22.60
 Mean   :22.79
 3rd Qu.:23.90
 Max.   :27.80
 NA's   :16
 Expected.years.after.labour.market.exit..Male
 Min.   :13.00
 1st Qu.:16.60
 Median :18.50
 Mean   :18.44
 3rd Qu.:20.30
 Max.   :23.30
 NA's   :16
```

Também podemos identificar os valores Na's de cada uma das variáveis

```
# VALORES Na's
# Verificar valores nulos por coluna
print(colSums(is.na(dados)))
```

```
                              Combined.measure
                                             0
                 Life.expectancy..Female..At.birth
                                             3
                 Life.expectancy..Female..65.years
                                             3
                   Life.expectancy..Male..At.birth
                                             3
                   Life.expectancy..Male..65.years
                                             3
             Employment.rate..From.55.to.59.years
                                             7
             Employment.rate..From.60.to.64.years
                                             7
             Employment.rate..From.65.to.69.years
```

```
                                                 7
 Old.age.to.working.age.ratio..65.years.or.over
                                                 1
       Effective.labour.market.exit.age..Female
                                                 5
         Effective.labour.market.exit.age..Male
                                                 6
Expected.years.after.labour.market.exit..Female
                                                16
  Expected.years.after.labour.market.exit..Male
                                                16
```

```
cat("Total de valores nulos:", sum(is.na(dados)), "\n")
```

```
Total de valores nulos: 77
```

Para poder visualizar os dados de melhor forma nos gráficos, procedemos a tentar limpar os dados faltantes.

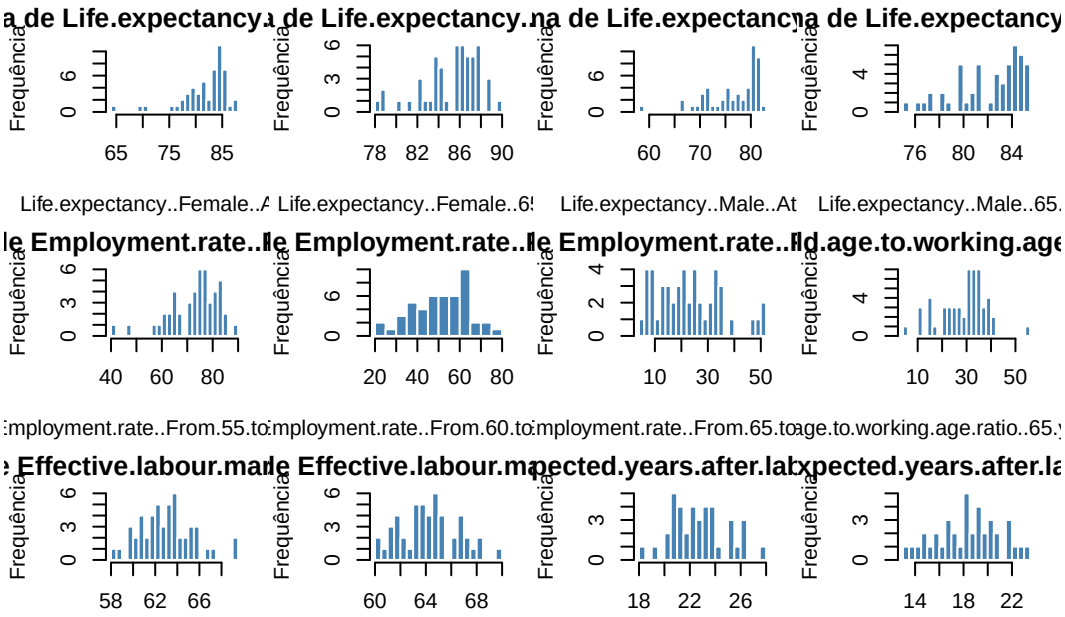
```
#| echo: true
#| warning: false

#Análise numérica

numeric_cols <- dados %>% select(where(is.numeric))
categorical_cols <- dados %>% select(where(is.factor) | where(is.character))

dados_clean <- dados[complete.cases(numeric_cols), ]
cat("\nLinhas após eliminar NA das colunas numéricas:", nrow(dados_clean), "/", nrow(dados),
```

```
Linhas após eliminar NA das colunas numéricas: 39 / 55
```

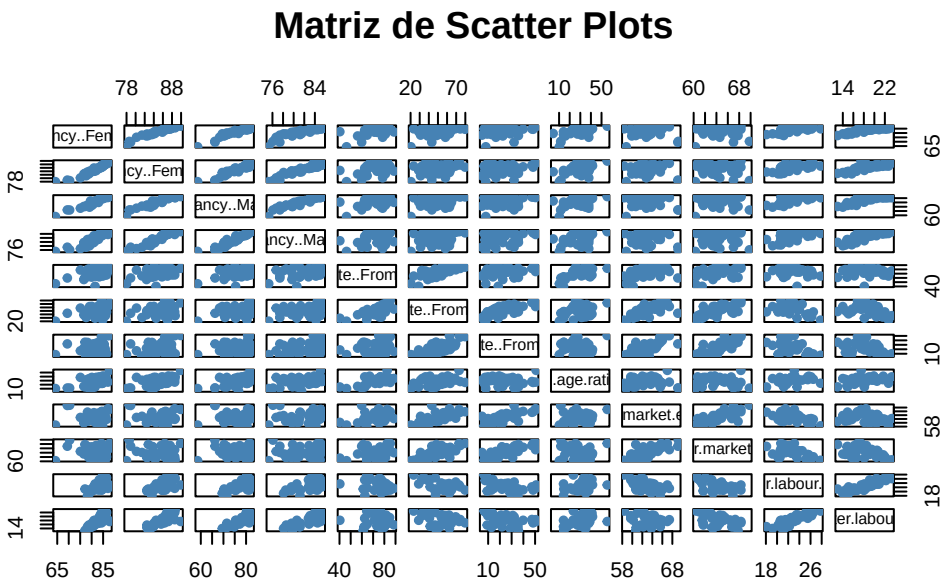
```
numeric_cols_clean <- dados_clean %>% select(where(is.numeric))
categorical_cols_clean <- dados_clean %>% select(where(is.factor) | where(is.character))
```

A seguir, podemos observar alguns gráficos.

```r
# Histogramas para variáveis numéricas
if(ncol(numeric_cols) > 0) {

  # Calcular layout para grid
  n_plots <- ncol(numeric_cols)
  n_cols <- ceiling(sqrt(n_plots))
  n_rows <- ceiling(n_plots / n_cols)

  # Configurar layout
  par(mfrow = c(n_rows, n_cols))
  par(mar = c(4, 4, 2, 1))

  for(col in names(numeric_cols)) {
    hist(numeric_cols[[col]],
         main = paste("Histograma de", col),
         xlab = col,
         ylab = "Frequência",
         col = "steelblue",
         border = "white",
         breaks = 20)
  }

  # Resetar layout
  par(mfrow = c(1, 1))
}
```




```
# Scatter plots entre variáveis numéricas
if(ncol(numeric_cols) > 1) {

  # Criar matriz de scatter plots
  pairs(numeric_cols,
        main = "Matriz de Scatter Plots",
        pch = 19,
        col = "steelblue",
        cex = 0.8,
        gap = 0.5)
}
```



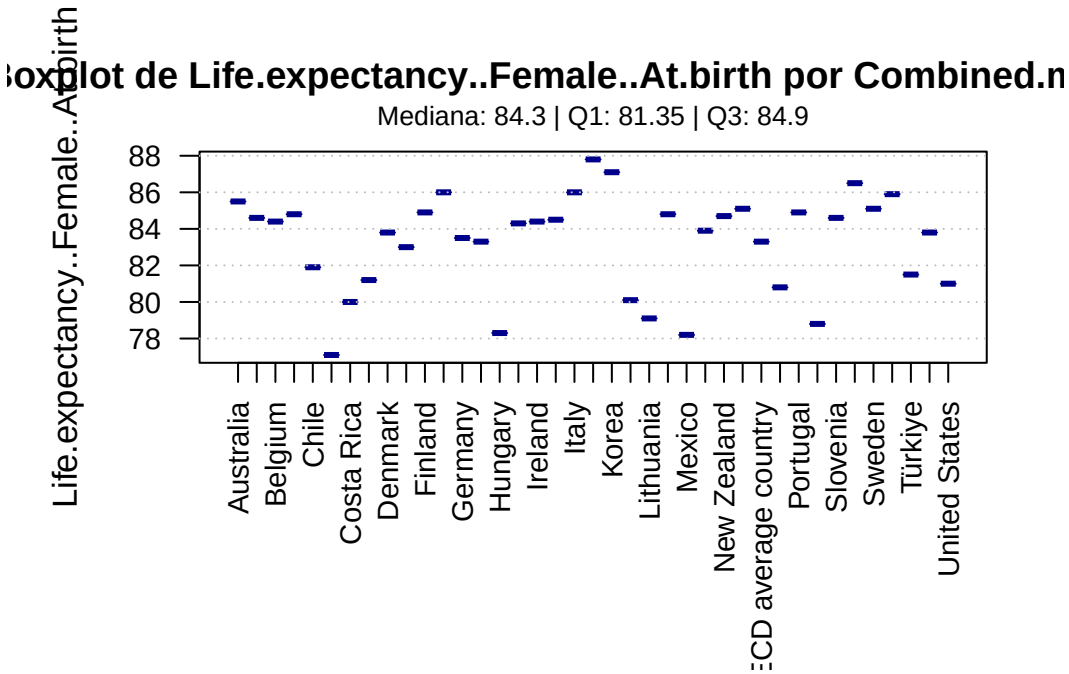
Notemos que os histogramas de cada uma das variáveis, nos quais se percebe que a maioria das variáveis apresenta distribuições aproximadamente normais, com leves assimetrias positivas.

Opción 2: Dividir en múltiples chunks (recomendado para muchos gráficos)

```
# Verificar que hay variables disponibles
if(ncol(numeric_cols) > 0 && ncol(categorical_cols) > 0) {

  # Seleccionar la primera variable numérica y la primera categórica
  num_var <- names(numeric_cols)[1]
  cat_var <- names(categorical_cols)[1]

  cat("Generando boxplot de:", num_var, "por", cat_var, "\n")

  # Configurar ventana gráfica para un solo plot
  par(mar = c(8, 5, 4, 2))  # Márgenes ajustados

  # Crear el boxplot
  boxplot(as.formula(paste(num_var, "~", cat_var)),
          data = dados_clean,
          main = paste("Boxplot de", num_var, "por", cat_var),
          xlab = "",
```

```
          ylab = num_var,
          col = "lightblue",
          border = "darkblue",
          las = 2,  # Rotar etiquetas del eje X
          cex.axis = 0.9,  # Tamaño del texto de los ejes
          cex.lab = 1.1,   # Tamaño de las etiquetas
          cex.main = 1.2)  # Tamaño del título

  # Añadir grid para mejor lectura
  grid(nx = NA, ny = NULL, col = "gray", lty = "dotted")

  # Añadir estadísticas descriptivas en el título secundario
  stats <- boxplot.stats(dados_clean[[num_var]])
  mtext(paste("Mediana:", round(median(dados_clean[[num_var]], na.rm = TRUE), 2),
              "| Q1:", round(stats$stats[2], 2),
              "| Q3:", round(stats$stats[4], 2)),
        side = 3, line = 0.5, cex = 0.8)

} else {
  if(ncol(numeric_cols) == 0) {
    cat("No hay variables numéricas disponibles\n")
  }
  if(ncol(categorical_cols) == 0) {
    cat("No hay variables categóricas disponibles\n")
  }
}
```

```
Generando boxplot de: Life.expectancy..Female..At.birth por Combined.measure
```



Além do estudo dos diferentes boxplots, podemos observar diferenças significativas.

A base de dados possui muitos valores faltantes, o que limita o estudo da análise exploratória, tornando difícil realizar correlações entre as variáveis por não haver dados suficientes. É necessária uma limpeza mais robusta utilizando diferentes metodologias para tratar os dados faltantes.