Introdução ao R e gráficos

MAT021 - Estatística I B

Rodney Fonseca

14/10/2024

Software R

R é uma linguagem e um ambiente para computação estatística e para preparação de gráficos

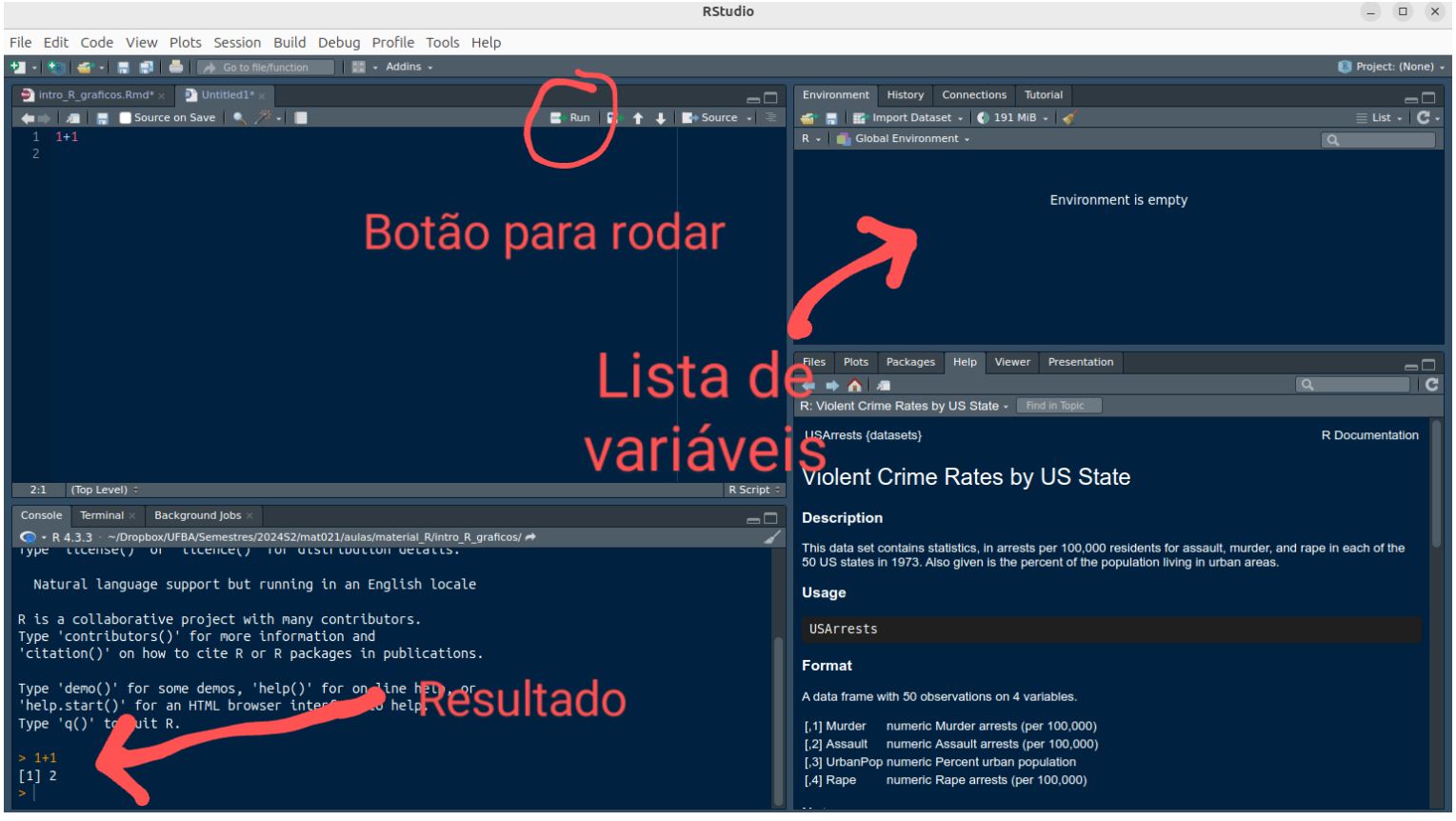
Vantagens

- software gratuito
- tratamento de dados
- ferramentas de diversos níveis para análise de dados
- ferramentas gráficas

Referências

- Livro Frery, Cribari-Neto (2011) *Elementos de Estatística Computacional Usando Plataformas de Software Livre/Gratuito*
- Página do curso de *Ciência de dados* da profa. Carolina Mota e prof. Gilberto Sassi do DEst-UFBA: https://ufba.netlify.app/paginas/catalogo
- Página do curso *Estatística Computacional com R* do prof. Paulo Justiniano (UFPR) e equipe: http://cursos.leg.ufpr.br/ecr/index.html

Ambiente do RStudio



Operações básicas

Adição

```
1+1
```

```
## [1] 2
```

Subtração

```
5 - 3
```

```
## [1] 2
```

Operações básicas

Multiplicação

```
2 * 3
```

```
## [1] 6
```

Divisão

```
10/5
```

```
## [1] 2
```

Operadores lógicos

Maior/menor que

```
5 > 3
```

```
## [1] TRUE
```

```
3>=3
```

```
## [1] TRUE
```

```
3>3
```

```
## [1] FALSE
```

Operadores lógicos

Igual

```
5 == 3
```

```
## [1] FALSE
```

```
3 == 3
```

```
## [1] TRUE
```

Diferente

```
4 != 2
```

```
## [1] TRUE
```

```
4 != 4
```

```
## [1] FALSE
```

Tipos de dados em R

Número real

```
class(0.5)
```

```
## [1] "numeric"
```

Valor lógico

```
class(TRUE)
```

```
## [1] "logical"
```

Caractere

```
class('UFBA')
```

```
## [1] "character"
```

Variável

- É como uma caixa nomeada. Você pode trocar o conteúdo da caixa, mas o nome permanece o mesmo.

Usamos os símbolos <- ou = para atribuir valores a uma variável

```
x <- 2
```

Para ver o valor da variável, basta rodar o seu nome

```
x
```

```
## [1] 2
```

Variável

Podemos fazer operações com a variável

```
5 * x
```

```
## [1] 10
```

podemos trocar o seu valor

```
x <- 20
x
```

```
## [1] 20
```

Podemos atribuir o valor de operações à outras variáveis

```
idade <- 2 * x
idade
```

```
## [1] 40
```

Funções

- Em matemática, funções recebem um ou mais argumentos e retornam um ou mais valores. Funções em R são similares.

Função para criar uma sequência de 9 números

```
seq(from = 1, to = 9)
```

```
## [1] 1 2 3 4 5 6 7 8 9
```

Vetores

- Vetores são listas indexadas de *variáveis do mesmo tipo*. É como um armário de gavetas numeradas como 1, 2, 3, ... Você pode mudar o conteúdo da gaveta 5 sem alterar o conteúdo da gaveta 1.

Vetores podem ser criados com a fórmula `c(x1, x2, ...)`

```
meu_vetor <- c(1, 2, 5, 8, 1, 3)
```

Valores de elementos de um vetor podem ser vistos assim

```
meu_vetor[3]
```

```
## [1] 5
```

Podemos alterar elementos do vetor

```
meu_vetor[3] <- 99
meu_vetor
```

```
## [1]  1  2 99  8  1  3
```

Conjuntos de dados em R

- Análise de dados geralmente envolvem tabelas com diferentes tipos de variáveis, tanto quantitativas como qualitativas. Em R, diferentes tipos de dados podem ser armazenados em estruturas chamadas *listas* e *dataframes*.

Conjuntos de dados em R

- Análise de dados geralmente envolvem tabelas com diferentes tipos de variáveis, tanto quantitativas como qualitativas. Em R, diferentes tipos de dados podem ser armazenados em estruturas chamadas *listas* e *dataframes*.
- Diversas funções soltam resultados em formas de listas.
- Dataframes geralmente são obtidos quando lemos planilhas com dados.

Listas

Criando uma lista

```
minha_lista <- list("dois", 2, c(33, 5, 88), TRUE)
class(minha_lista)
```

```
## [1] "list"
```

Acessando elementos da lista

```
minha_lista[[1]]
```

```
## [1] "dois"
```

```
class(minha_lista[[1]])
```

```
## [1] "character"
```

Listas

...se o elemento for um vetor, podemos ver seus elementos

```
minha_lista[[3]][2]
```

```
## [1] 5
```

Os elementos de listas também podem ter nomes

```
lista_nomeada <- list(nome = c("Chico", "Nara"),
                      idade = c(30, 25))
lista_nomeada$nome
```

```
## [1] "Chico" "Nara"
```

```
lista_nomeada$idade[2]
```

```
## [1] 25
```

Dataframes

- Para ler alguma planilha devemos checar o formato do arquivo (*.xls* para excel, *.odt* para libre office, *.txt* para textos, etc.).

Dataframes

- Para ler alguma planilha devemos checar o formato do arquivo (*.xls* para excel, *.odt* para libre office, *.txt* para textos, etc.).
- Nesse material, discutiremos como carregar arquivos .csv (sigla em inglês para *valores separados por vírgula*), que é um dos formatos mais comuns para arquivos de dados.

Dataframes

- Para ler alguma planilha devemos checar o formato do arquivo (*.xls* para excel, *.odt* para libre office, *.txt* para textos, etc.).
- Nesse material, discutiremos como carregar arquivos .csv (sigla em inglês para *valores separados por vírgula*), que é um dos formatos mais comuns para arquivos de dados.
- Para ler um arquivo, é aconselhável que ele esteja na mesma pasta que o seu script/código R. Usaremos a função `read.csv()`, a qual já lê o arquivo como um dataframe.

Lendo arquivo

- Para usar a função `read.csv()`, escrevemos o nome do arquivo em `file` e indicamos se o arquivo tem cabeçalho com a opção `header`. Os dados podem ser então salvos em uma variável com algum nome sugestivo.
- Nesta parte usaremos um exemplo com dados de currículos lattes no arquivo `multidisciplinar_ufba_lattes.csv` (Fonte: http://bi.cnpq.br/painel/formacao-atuacao-lattes/)

```
df_lattes<-read.csv(file='multidiscip_ufba_lattes.csv',
                    header = TRUE)
class(df_lattes)
```

```
## [1] "data.frame"
```

- Para visualizar todos os dados, use a função `View(df_lattes)`

Checando os nomes das variáveis

```
names(df_lattes)
```

```
##  [1] ".Ano"                          ".País.de.Nascimento
##  [3] ".País..Formação."              ".Região..Formação.
##  [5] ".UF..Formação."                ".Cidade..Formação.
##  [7] ".Grande.Área..Formação."       ".Área..Formação."
##  [9] ".Instituição..Formação."       ".Nível..Formação."
## [11] ".País..Atuação."               ".Região..Atuação."
## [13] ".UF..Atuação."                 ".Cidade..Atuação."
## [15] ".Grande.Área..Atuação."        ".Área..Atuação."
## [17] ".Instituição..Atuação."        ".Setor.de.Atividad
## [19] ".Enquadramento..Atuação."      ".Sexo"
## [21] ".Cor.ou.Raça"
```

Checando os valores de uma das variáveis

```
df_lattes$.Grande.Área..Atuação.
```

```
##  [1] "Ciências Exatas e da Terra" "Ciências Sociais Apli
##  [3] "Não informado"              "Ciências Exatas e da
##  [5] "Ciências Humanas"           "Ciências Humanas"
##  [7] "Ciências Humanas"           "Ciências Humanas"
##  [9] "Ciências da Saúde"          "Ciências Humanas"
## [11] "Ciências Humanas"           "Ciências Humanas"
## [13] "Ciências Humanas"           "Ciências Humanas"
## [15] "Ciências Humanas"           "Ciências Humanas"
## [17] "Não informado"              "Não informado"
## [19] "Ciências Sociais Aplicadas" "Ciências Humanas"
## [21] "Ciências Humanas"           "Ciências Humanas"
## [23] "Ciências Humanas"
```

Sumarizando valores em tabelas

```
tb_area_atuacao <- table(df_lattes$.Grande.Área..Atuação.)
tb_area_atuacao
```

```
##
##           Ciências da Saúde Ciências Exatas e da Terra
##                           1                           2
##            Ciências Humanas Ciências Sociais Aplicadas
##                          15                           2
##               Não informado
##                           3
```

Exercícios

- Como você pode ver os valores da variável *área de atuação* no dataframe **df_lattes**?
- Qual é a *área de atuação* da sexta observação nesse dataframe?
- O que significa verificar `df_lattes$.Área..Atuação.[10] != df_lattes$.Área..Atuação.[11]`?
- Qual é a *cidade de atuação* que mais aparece nos dados no dataframe **df_lattes**?

Gráficos de variáveis qualitativas

Gráfico de barras

- O gráfico em barras consiste em construir colunas ou barras cujo comprimento/altura é proporcional à magnitude da contagem para cada valor da variável.

Gráfico de barras

- O gráfico em barras consiste em construir colunas ou barras cujo comprimento/altura é proporcional à magnitude da contagem para cada valor da variável.
- Tais barras são dispostas paralelamente umas às outras, horizontal ou verticalmente.
- Geralmente o comprimento da barra corresponde a uma contagem ou frequência de cada valor da variável.

- Para criar um gráfico de barras, podemos contar o número de observações para cada valor/atributo da variável qualitativa
- A função `table()` pode ser usada para fazer tal contagem

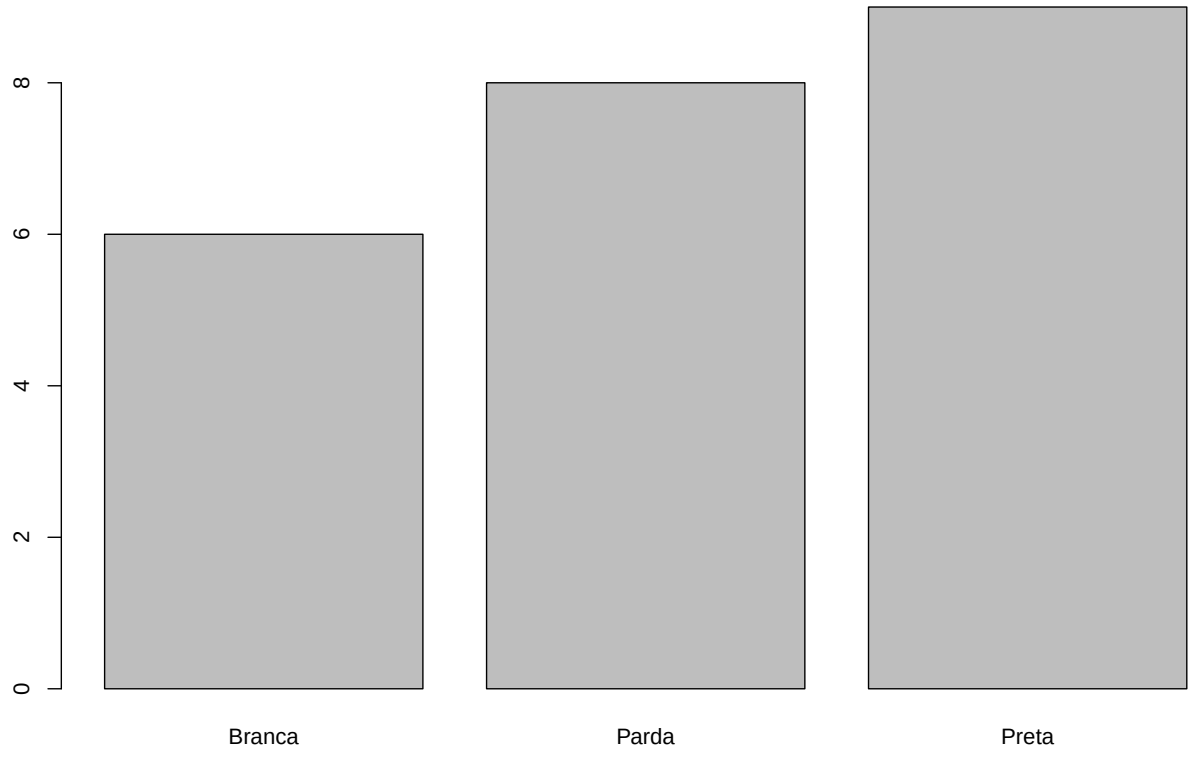
Criando uma tabela da variável Raça

```
tb_raca <- table(df_lattes$.Cor.ou.Raça)
tb_raca
```

```
##
## Branca  Parda  Preta
##      6      8      9
```

Criando um gráfico de barras da variável

```
barplot(tb_raca)
```



Nomeando os elementos do gráfico de barras

```
barplot(tb_raca, ylab = "Contagem", xlab = "Cor",
        main = "Número de pessoas por cor")
```

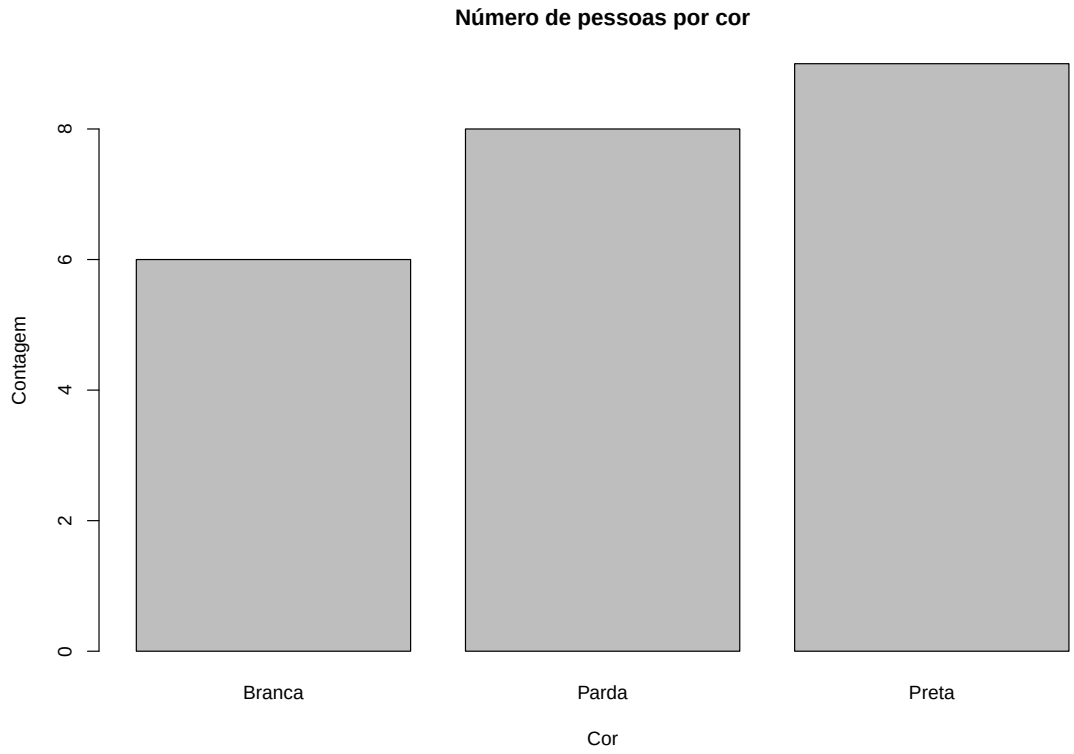


Gráfico de setores

- O gráfico de setores, também conhecido como gráfico de "pizza", representa a composição de partes de um todo, geralmente em porcentagem.
- Consiste num círculo representando o todo, dividido em setores que correspondem às partes de maneira proporcional.

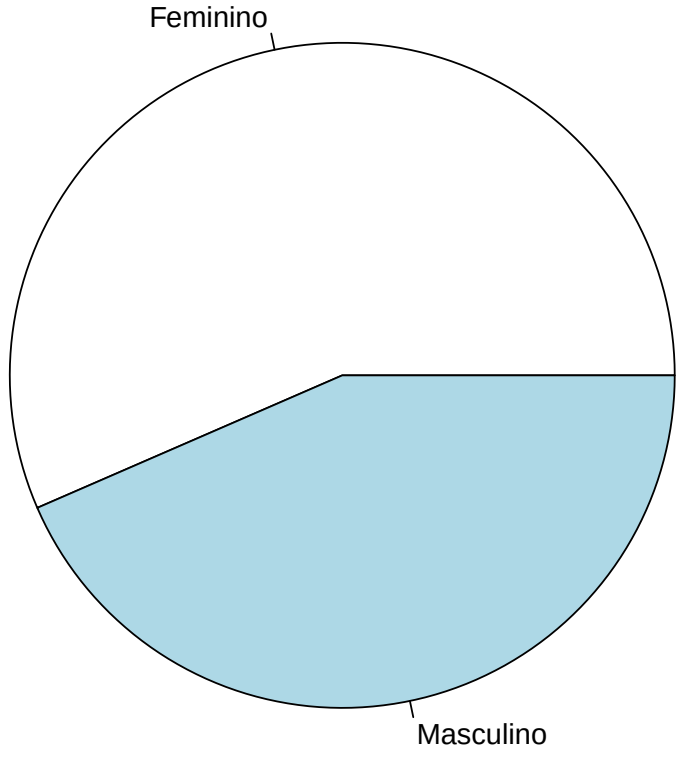
Contagens para a variável sexo

```r
tb_sexo <- table(df_lattes$.Sexo)
tb_sexo
```

```
##
##  Feminino Masculino
##        13        10
```

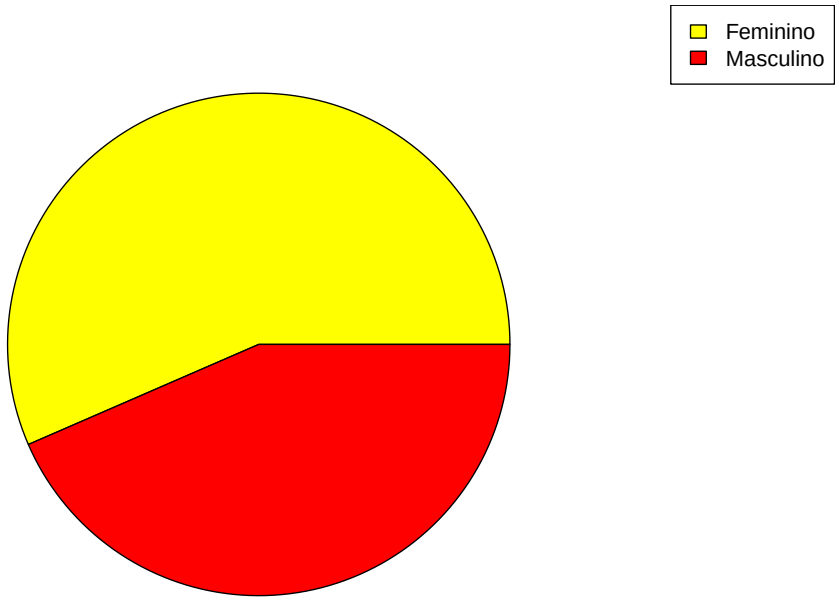
Criando um gráfico de setores da variável sexo

```
pie(tb_sexo)
```



Alterando elementos de um gráfico de setor

```
pie(tb_sexo, col = c('yellow', 'red'), labels = "")
legend("topright", legend = names(tb_sexo),
       fill = c('yellow', 'red'))
```



Gráficos de variáveis quantitativas

Histograma

- O histograma é um gráfico de barras contíguas, com as bases proporcionais aos intervalos das classes e a área de cada retângulo proporcional à respectiva frequência/contagem.
- Útil para saber quais valores da variável qualitativa ocorrem com maior frequência.

- Usaremos um exemplo com um banco de dados que já está salvo no R: *USArrest*
- Esse banco de dados tem estatísticas de taxas de assassinatos, assaltos e estupros de *cada estado* dos EUA durante o ano de 1973. As taxas são número de casos para cada 100.000 pessoas.

Carregando os dados USArrest

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

Histograma da variável assassinatos

```
hist(USArrests$Murder, main = 'Histrograma',
     ylab = 'Frequência', xlab = 'Taxa (por 100.000)')
```

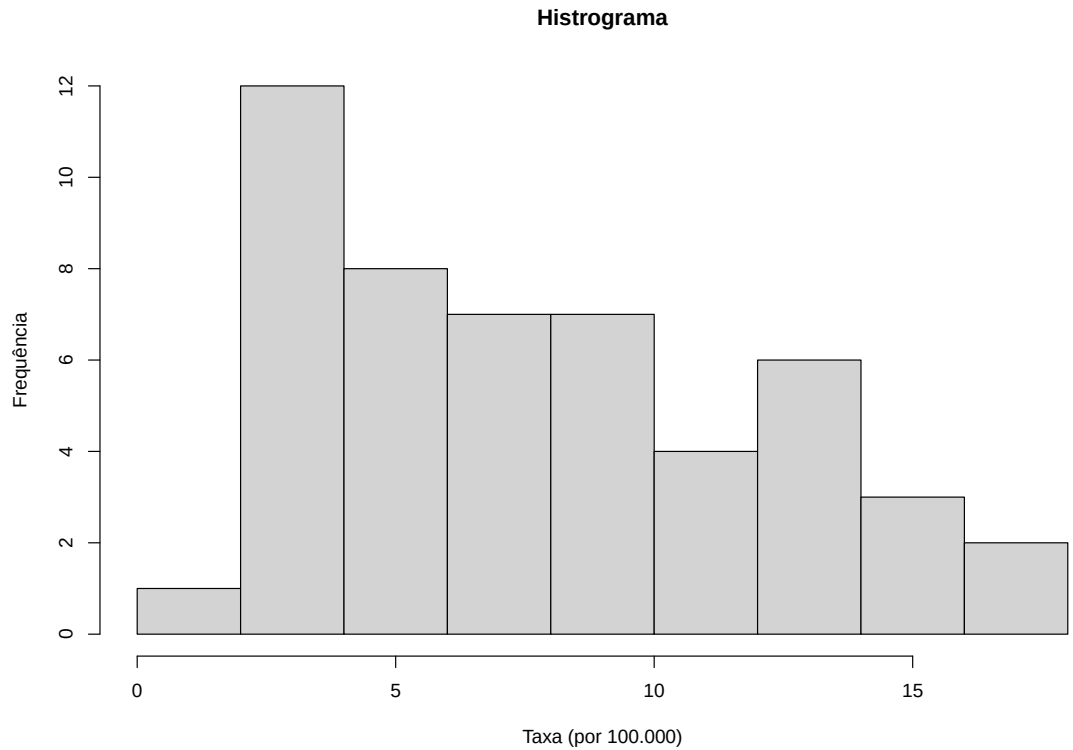


Gráfico de dispersão

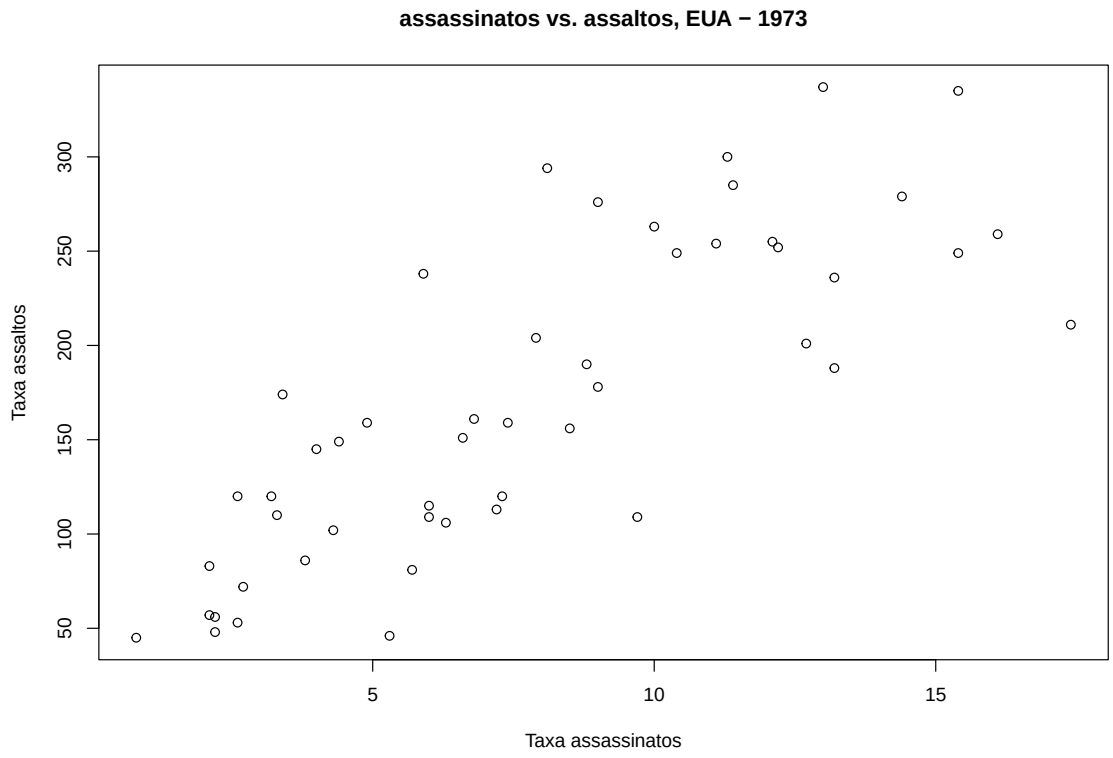
- Mostra a relação entre duas variáveis quantitativas
- Cada ponto no gráfico corresponde aos valores de uma variável para uma das observações
- Útil para identificar *correlação* entre variáveis

Exemplo:

- Para os dados dos EUA, temos as taxas de assassinato e assalto para cada estado
- Num gráfico de dispersão dessas taxas, cada ponto corresponderá a um estado

Gráfico de dispersão das variáveis assassinatos e assalto

```
plot(USArrests$Murder, USArrests$Assault,
     main = 'assassinatos vs. assaltos, EUA - 1973',
     ylab = 'Taxa assaltos', xlab = 'Taxa assassinatos')
```



Exercícios

- Carregue no R os dados *ipeadata_consumo_energia.csv* sobre consumo de energia usando a função `read.csv()` (Fonte: http://www.ipeadata.gov.br/). Um dos argumentos da função deverá ser `dec=','` para indicar que vírgulas marcam os decimais nesses dados.
- Faça um histograma da variável `consumo_energia_2023`. O que você consegue concluir?
- Aplique a função `log()` na variável `consumo_energia_2023` salvando o resultado numa variável chamada `log_consumo`. Faça um histograma da variável `log_consumo`. Qual mudança ocorreu em relação ao gráfico do item anterior?