



# **Tópicos Especiais em Computação I**

---

**Universidade Regional Integrada do Alto Uruguai e das Missões -  
Campus Erechim**

**Prof. Jackson Felipe Magnabosco**

# Seleção dos dados

---



# Introdução

---

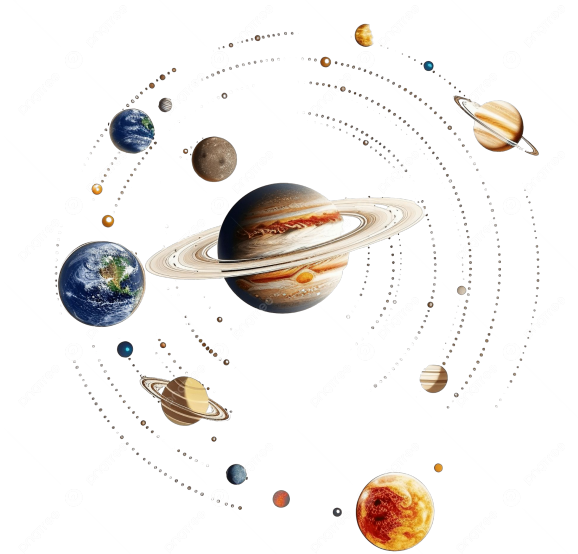
- Definir seleção dos dados.
- Descrever nomenclatura e tipos de dados.
- Aplicar a seleção de dados em uma base de dados.

# Introdução à Seleção de Atributos

---

## Etapas do Processo de Seleção de Atributos

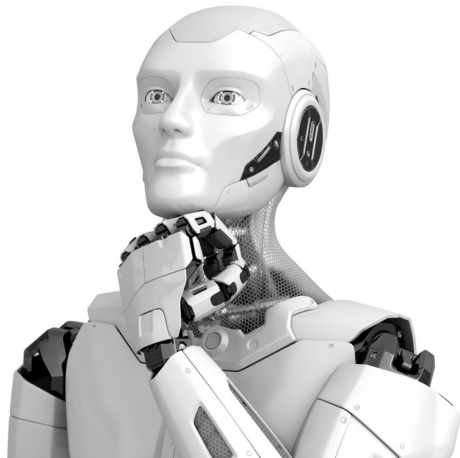
- A seleção de atributos (feature selection) é uma técnica de seleção de dados.
- Permite remover características irrelevantes dos dados brutos, evitando atrapalhar a mineração de dados..



# O Processo de Seleção de Atributos - Input

---

- A entrada no processo é uma base de dados com todos os atributos.
- O primeiro passo é a criação de subconjuntos com atributos candidatos.
- Geralmente, remove-se alguns atributos aleatoriamente para descobrir os mais importantes.

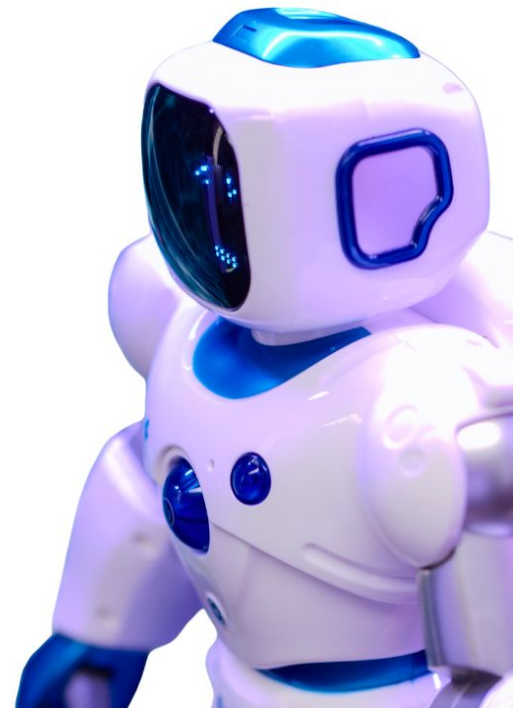


# Aplicação de Mineração

---

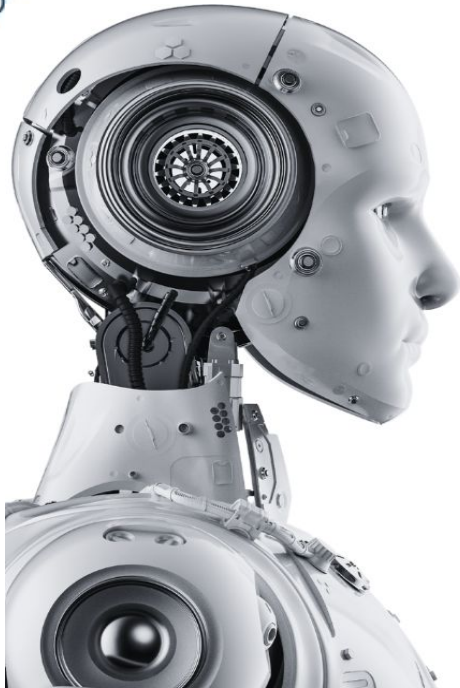
## Aplicação de Mineração no Subconjunto

- Após a remoção de alguns atributos, aplica-se técnicas de mineração nos subconjuntos criados.
- A mineração ajuda a avaliar a relevância dos atributos selecionados.



# Avaliação dos Resultados

---



- Após aplicar a mineração, os resultados são avaliados.
- Se o resultado for satisfatório, a seleção está completa. Caso contrário, o processo é repetido até obter resultados satisfatórios.

# Iteração do Processo

---

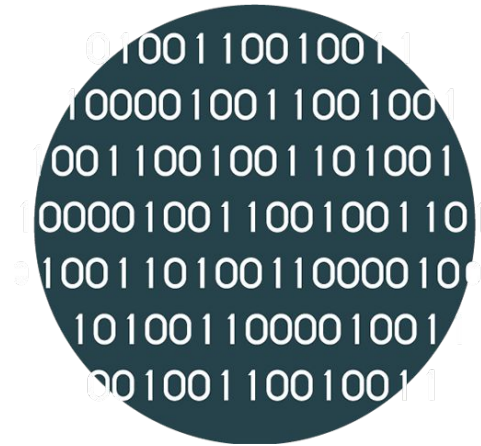
- Caso os resultados não sejam satisfatórios, o processo é repetido várias vezes.
- Isso é feito até que os resultados sejam satisfatórios ou até atingir um número máximo de repetições.





# Resultado Final - Output

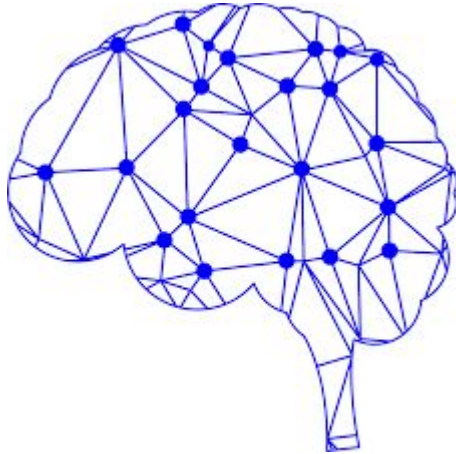
---



- O melhor resultado determina quais são os atributos mais significativos.
- A saída do processo é uma base de dados com os atributos mais relevantes.

# Importância dos Métodos de Seleção

---



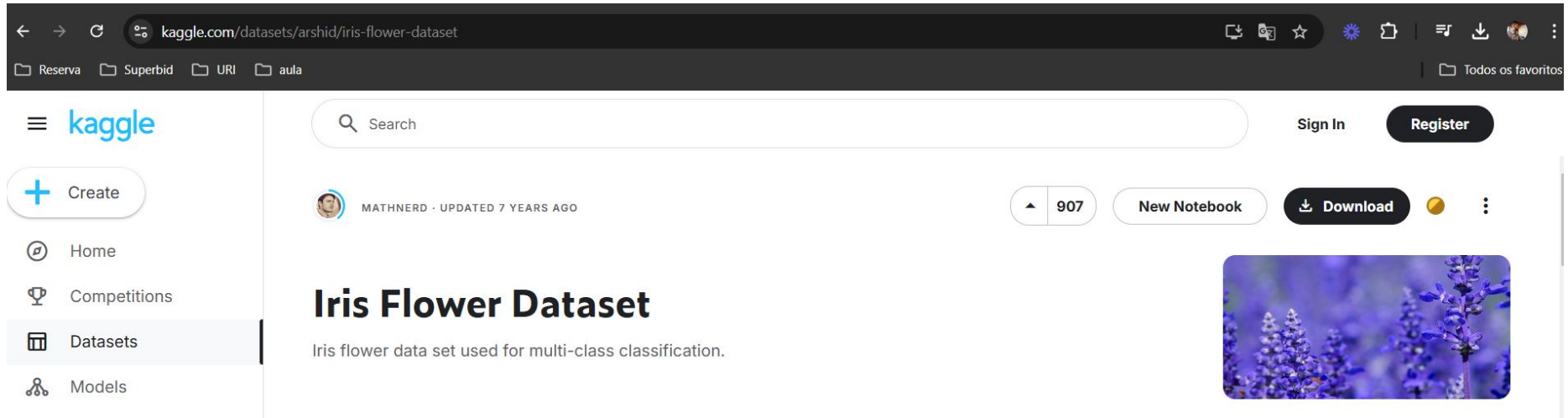
- Os métodos de seleção de atributos atuam como filtros de qualidade para os dados utilizados na mineração.
- Eles podem ser realizados de maneira sistemática ou aleatória, garantindo que apenas os atributos mais relevantes sejam mantidos.
- No final, o objetivo é identificar os atributos que trazem os melhores resultados para o modelo.

# Seleção de atributos utilizando a linguagem R e o RStudio

---

# Iris flower dataset

<https://www.kaggle.com/datasets/arshid/iris-flower-dataset?resource=download>



The screenshot shows the Kaggle website interface. At the top, the browser address bar displays the URL: `kaggle.com/datasets/arshid/iris-flower-dataset`. The page header includes the Kaggle logo, a search bar, and links for "Sign In" and "Register". The left sidebar contains navigation links: "Home", "Competitions", "Datasets" (which is highlighted), and "Models". The main content area features the dataset title "Iris Flower Dataset" by user "MATHNERD", updated 7 years ago. It shows 907 votes and options to "New Notebook" or "Download". A description states: "Iris flower data set used for multi-class classification." A thumbnail image of purple iris flowers is visible on the right.

**Iris Flower Dataset**

Iris flower data set used for multi-class classification.

## Dados: Iris flower dataset

---

150 amostras das três espécies de flores iris:



Setosa



Virginica



Versicolor



## Dados

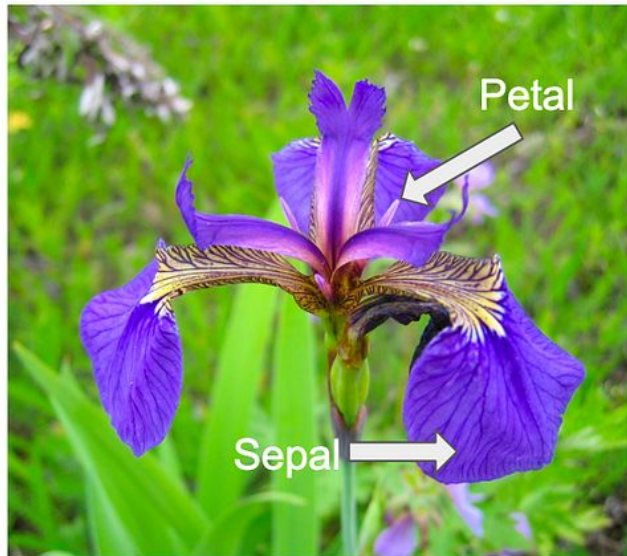
---

- Comprimento das sépalas
- Comprimento das pétalas
- Largura das sépalas
- Largura das pétalas
- Espécie que corresponde determinada amostra

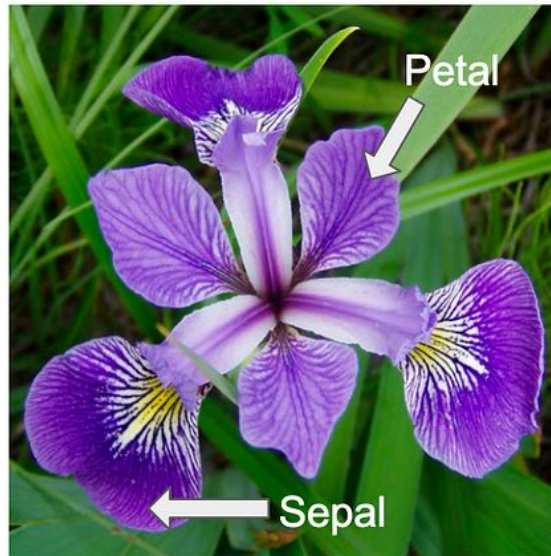


# Dados

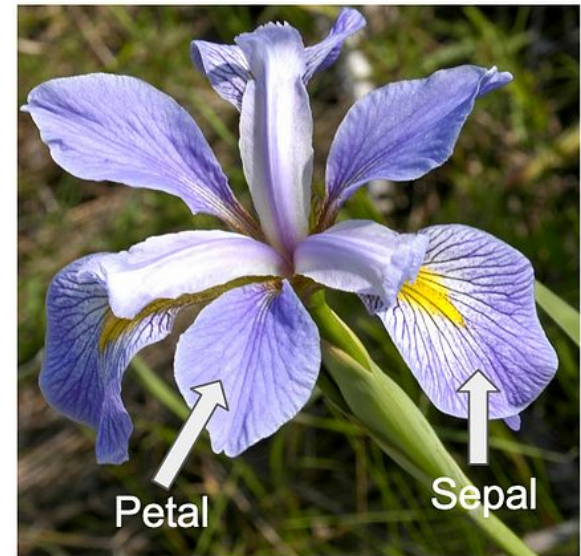
*Iris setosa*



*Iris versicolor*



*Iris virginica*



# Linguagem de programação

---





R

---



# R - baixar

vps.fmvz.usp.br/CRAN/

Reserva Superbid URI aula



CRAN

[Mirrors](#)

[What's new?](#)

[Search](#)

[CRAN Team](#)

About R

## The Comprehensive R Archive Network

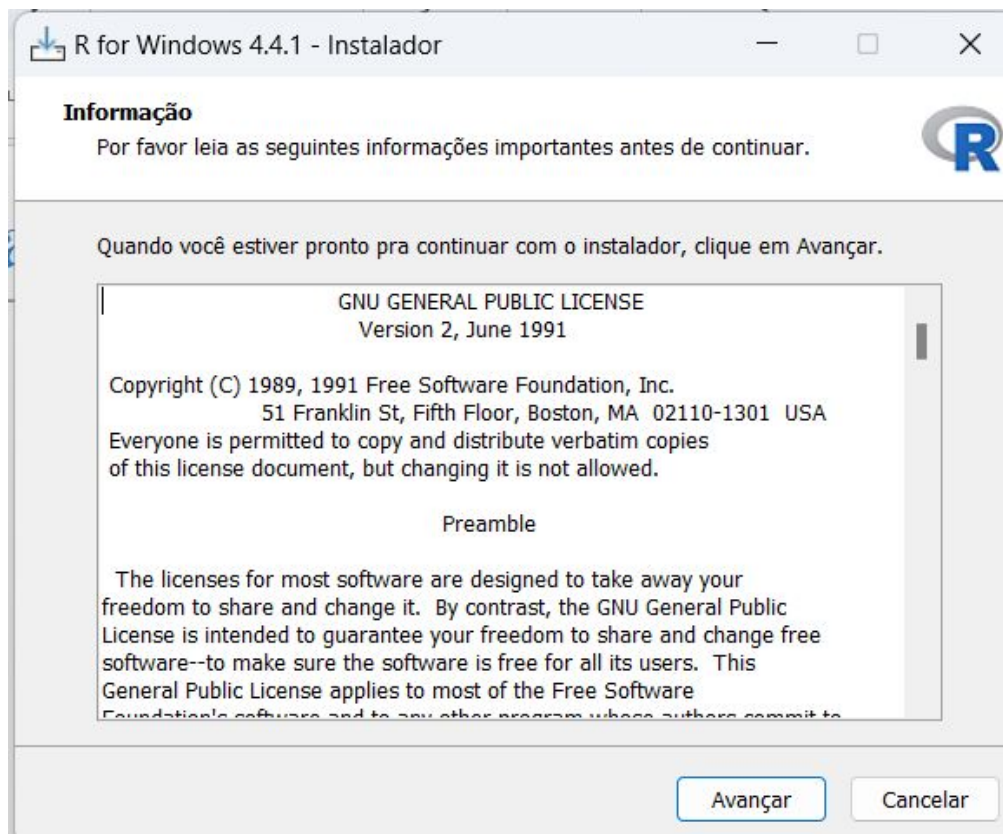
### Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux \(Debian, Fedora/Redhat, Ubuntu\)](#)
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

# R - instalar



# Ferramentas

---



**RStudio**

---



# RStudio - baixar



Want to learn about core or advanced workflows in RStudio?  
Explore the [RStudio User Guide](#) or the [Getting Started](#) section.

## 1: Install R

RStudio requires R 3.6.0+. Choose a version of R that matches your computer's operating system.

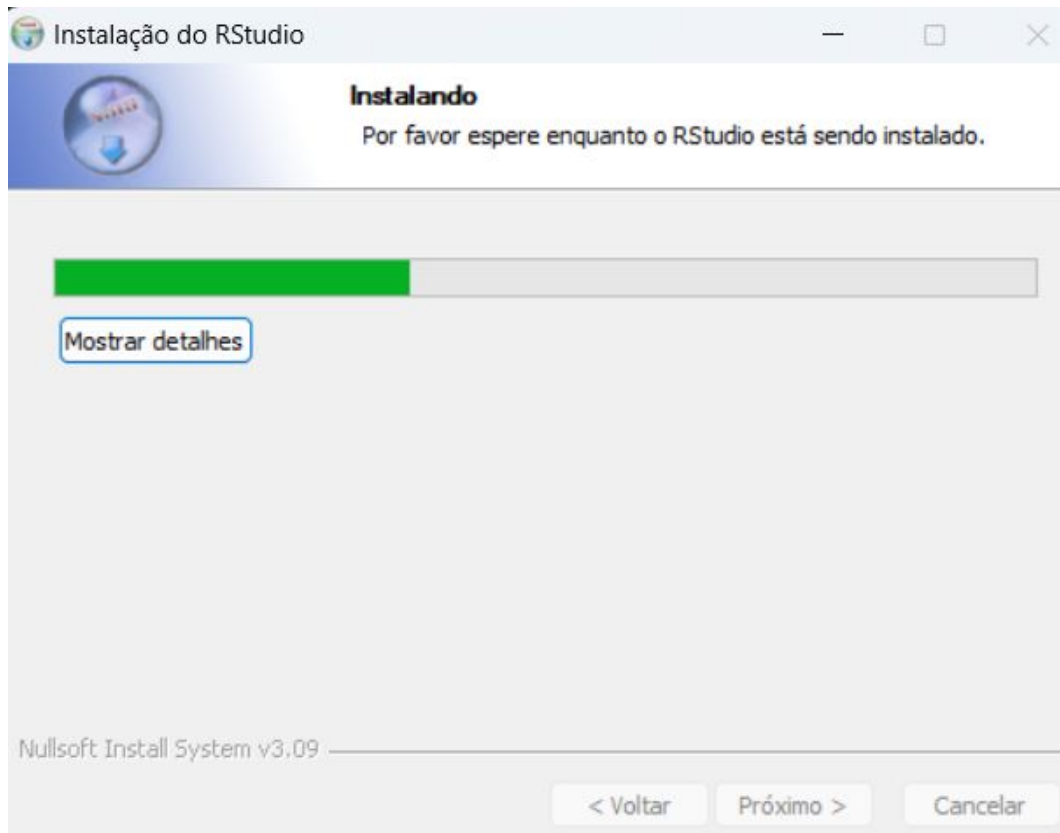
*R is not a Posit product. By clicking on the link below to download and install R, you are leaving the Posit website. Posit disclaims any obligations and all liability with respect to R and the R website.*

## 2: Install RStudio

[DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS](#)

Size: 265.55 MB | [SHA-256: 513216FE](#) | Version: 2024.09.0+375 |  
Released: 2024-09-23

# RStudio - instalação



# Hora de codificar!

Vamos mergulhar no R e  
construir nossas análises.

## Hands on





# Resultados

---

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Environment History Connections

Project: (None)

R 138 MiB

Global Environment

Data

- # iris 150 obs. of 6 vari...
- # iris\_so... 150 obs. of 6 vari...
- # petal 150 obs. of 3 vari...
- # sepal 150 obs. of 3 vari...

```
1 # Carregando a base de dados 'iris'
2 data(iris)
3
4 # Exibindo as primeiras linhas do dataset
5 head(iris)
6
7 # Simples remoção de atributos
8 # Apagando colunas usando NULL
9 iris$sepal.Length = NULL
10 iris$petal.width = NULL
11
12 # Exibindo a estrutura do dataset após a remoção
13 str(iris)
14
15 # Carregando novamente a base original para outras operações
16 data(iris)
17
18 # Seleção de subconjuntos - Separar sépalas e pétalas
19 sepal <- subset(iris, select = c(Sepal.Length, Sepal.width, Species))
20 petal <- subset(iris, select = c(Petal.Length, Petal.width, Species))
21
22 # Exibindo as primeiras linhas de cada subconjunto
23 head(sepal)
24 head(petal)
25
26 # Estatísticas básicas
27 summary(iris) # Resumo estatístico de todas as colunas
28 mean(iris$sepal.Length) # Média do comprimento da sépala
29 sd(iris$petal.width) # Desvio-padrão da largura da pétala
30
31 # Criando uma nova coluna com a proporção entre comprimento e largura da pétala
32 iris$Petal.Ratio <- iris$Petal.Length / iris$Petal.width
33
34 # Verificando as primeiras linhas para conferir a nova coluna
35 head(iris)
36
37 # Ordenando o dataset por comprimento da pétala em ordem decrescente
38 iris.sorted <- iris[order(iris$Petal.Length, decreasing = TRUE), ]
39 head(iris.sorted)
40
41 # Contagem do número de observações por espécie
42 table(iris$Species)
43
44 # Gráfico simples - Dispersão entre comprimento e largura da sépala
45 plot(iris$sepal.Length, iris$sepal.width,
46      col = iris$Species,
47      pch = 16,
48      xlab = "Comprimento da Sépala",
49      ylab = "Largura da Sépala",
50      main = "Gráfico de Dispersão - Sépalas")
51
52 # Salvando um subconjunto como CSV
53 write.csv(sepal, "sepal_subset.csv", row.names = FALSE)
54
```

54:1 (Top Level) 5

R Script 5

```
R - R44.1
> Contagem do número de observações por espécie
> table(iris$Species)
      setosa versicolor virginica 
       50         50         50 

> # Gráfico simples - Dispersão entre comprimento e largura da sépala
> plot(iris$sepal.Length, iris$sepal.width,
+      col = iris$Species,
+      pch = 16,
+      xlab = "Comprimento da Sépala",
+      ylab = "Largura da Sépala",
+      main = "Gráfico de Dispersão - Sépalas")
> # salvando um subconjunto como CSV
> write.csv(sepal, "sepal_subset.csv", row.names = FALSE)
> data(iris)
> # seleção de subconjuntos - Separar sépalas e pétalas
> sepal <- subset(iris, select = c(Sepal.Length, Sepal.width, Species))
> petal <- subset(iris, select = c(Petal.Length, Petal.width, Species))
> # Criando uma nova coluna com a proporção entre comprimento e largura da pétala
> iris$Petal.Ratio <- iris$Petal.Length / iris$Petal.width
> iris.sorted <- iris[order(iris$Petal.Length, decreasing = TRUE), ]
> head(iris.sorted)
   Sepal.Length Sepal.width Petal.Length Petal.width Species Petal.Ratio
119           7.7         2.6          6.9          2.3 virginica  3.000000
118           7.7         2.8          6.7          2.2 virginica  3.045455
123           7.7         2.8          6.7          2.0 virginica  3.350000
106           7.6         3.0          6.6          2.1 virginica  3.142857
112           7.8         3.8          6.4          1.9 virginica  3.200000
108           7.3         2.9          6.3          1.8 virginica  3.500000

> table(iris$Species)
      setosa versicolor virginica 
       50         50         50 

> # Gráfico simples - Dispersão entre comprimento e largura da sépala
> plot(iris$sepal.Length, iris$sepal.width,
+      col = iris$Species,
+      pch = 16,
+      xlab = "Comprimento da Sépala",
+      ylab = "Largura da Sépala",
+      main = "Gráfico de Dispersão - Sépalas")
> # salvando um subconjunto como CSV
> write.csv(sepal, "sepal_subset.csv", row.names = FALSE)
```

**Gráfico de Dispersão - Sépala**

Largura da Sépala

Comprimento da Sépala



## Disponível em:

---

- **Github:**  
**<https://github.com/jacksonn455/Analise-Exploratoria-com-o-Dataset-Iris>**

# Dúvidas ou sugestões ?





# Tópicos Especiais em Computação I

---

**Universidade Regional Integrada do Alto Uruguai e das Missões -  
Campus Erechim**

**Prof. Jackson Felipe Magnabosco**