




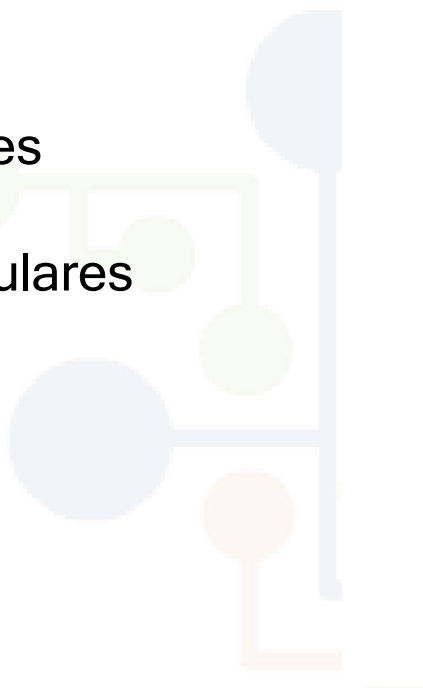
Data Science Academy

# Big Data Analytics com R e Microsoft Azure Machine Learning Módulo 3



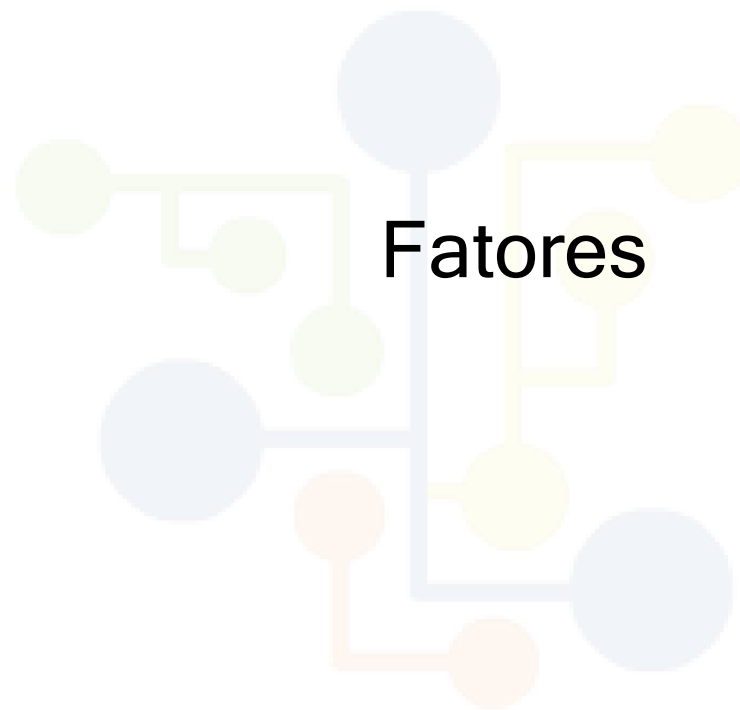
Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)

- 
- 
- Fatores e Funções
  - Pacotes
  - Expressões Regulares
  - Datas
  - Gráficos



Data Science Academy



Data Science Academy

(Categóricas)

## Qualitativas

### Nominais

- Profissão
- Sexo
- Religião

### Ordinais

- Escolaridade
- Classe Social
- Fila

## Quantitativas

### Discretas


- Número de Filhos
- Número de carros
- Número de acessos

### Contínuas

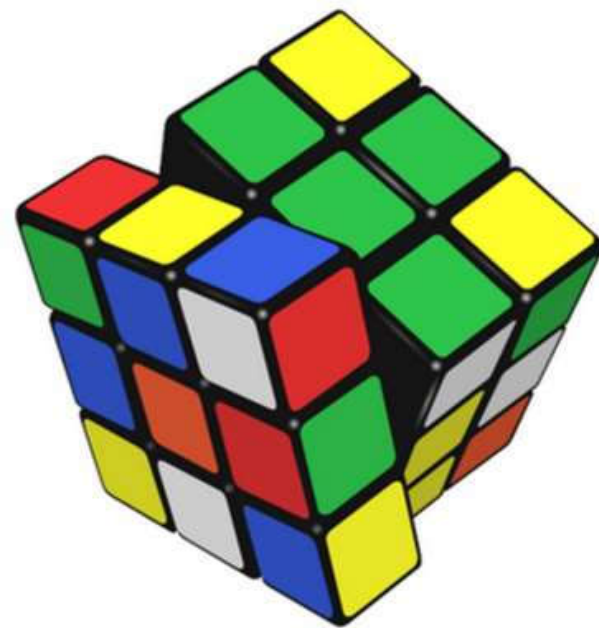
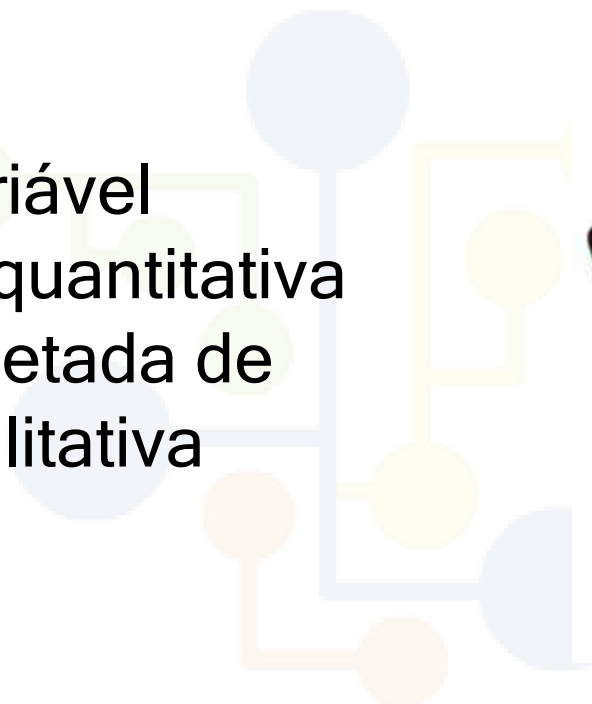
- Altura
- Peso
- Salário



Data Science Academy

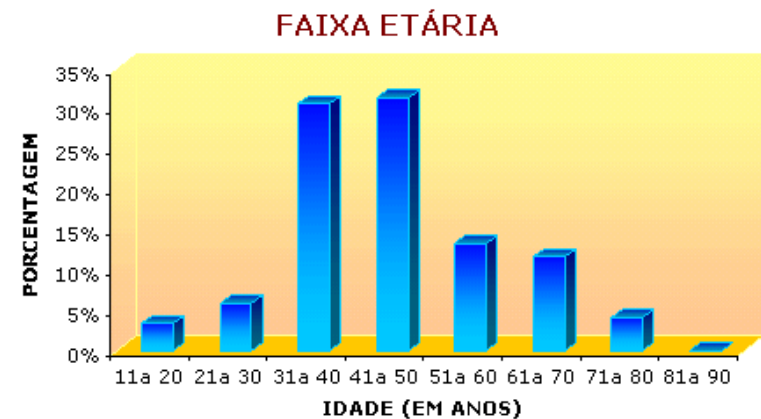


Uma variável  
originalmente quantitativa  
pode ser coletada de  
forma qualitativa



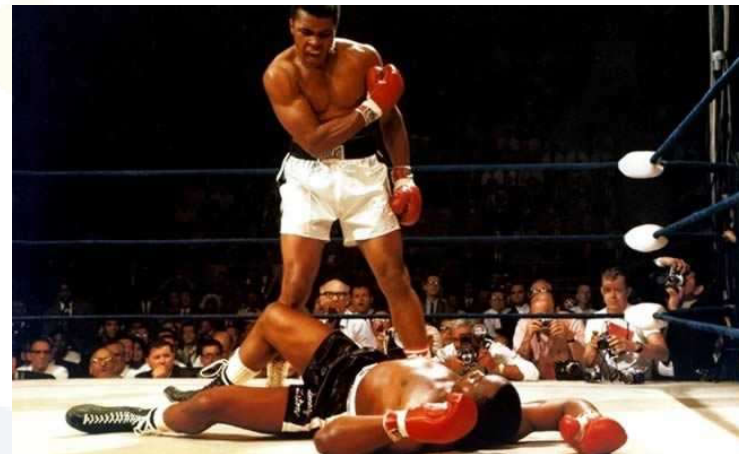
Data Science Academy

Por exemplo, a variável idade, medida em anos completos, é quantitativa (contínua); mas, se for informada apenas a faixa etária (0 a 5 anos, 6 a 10 anos, etc...), é qualitativa (ordinal)

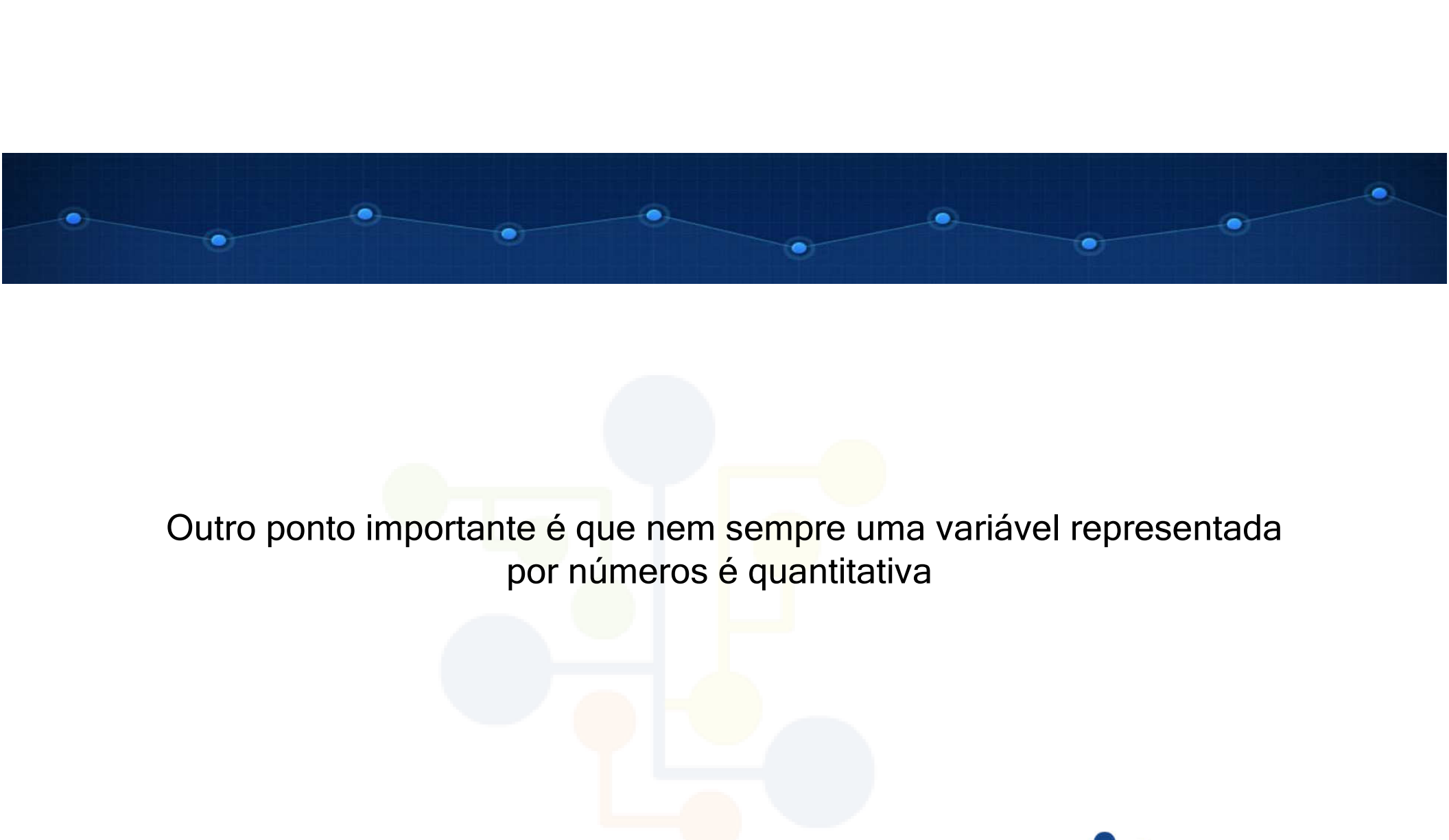


Data Science Academy

Outro exemplo é o peso dos lutadores de boxe, uma variável quantitativa (contínua) se trabalhamos com o valor obtido na balança, mas qualitativa (ordinal) se o classificarmos nas categorias do boxe (peso-pena, peso-leve, peso-pesado, etc.)



Data Science Academy



Outro ponto importante é que nem sempre uma variável representada por números é quantitativa



Data Science Academy



O número do telefone de uma pessoa, o número da casa, o número de sua identidade. Às vezes o sexo do indivíduo é registrado na planilha de dados como 1 se macho e 2 se fêmea, por exemplo. Isto não significa que a variável sexo passou a ser quantitativa!



Data Science Academy



# Fatores

Fatores representam uma maneira muito eficiente para armazenar valores de caracteres, porque cada caracter único é armazenado apenas uma vez e os dados são armazenados como um vetor de inteiros



Data Science Academy



## Fatores Ordenados

`factor(..., ord=T)` ou `ordered()`



Data Science Academy



# Funções

Tudo que você atribui com:

`<-`

vira um objeto no R



Data Science Academy



# Funções

## Anônimas

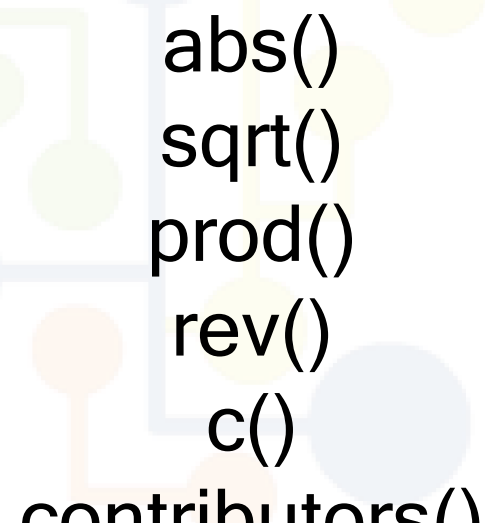
```
> teste_func <- sapply(c(1:10), function(x) {x %% 2 == 0})  
> teste_func  
[1] FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE
```



Data Science Academy



## Funções Built-in



abs()  
sqrt()  
prod()  
rev()  
c()  
contributors()



Data Science Academy




# Família de Funções Apply



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



```
lista1 <- list(a = (1:10), b = (45:77))
```

```
valor_a = 0
valor_b = 0
for (i in lista1$a){
  valor_a = valor_a + i
}
for (j in lista1$b){
  valor_b = valor_b + j
}
print(valor_a)
print(valor_b)
```

`sapply(lista1, sum)`

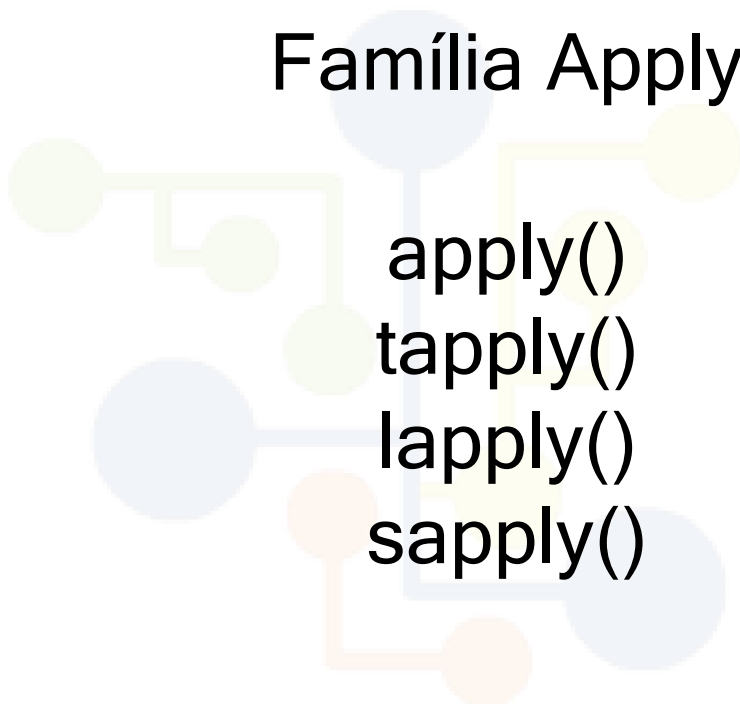


Data Science Academy





## Família Apply



`apply()`  
`tapply()`  
`lapply()`  
`sapply()`



Data Science Academy



apply()

apply(X, MARGIN, FUN, ...)

x = matriz ou dataframe

Margin = linha ou coluna

FUN = função a ser aplicada



Data Science Academy



# lapply()

Recebe um vetor ou lista e aplica uma  
função a cada elemento

`lapply(X, FUN, ...)`



Data Science Academy



apply()

Versão mais amigável do lapply

sapply(X, FUN, ..., simplify = TRUE, USE.NAMES = TRUE)



Data Science Academy



## tapply()

tapply(X, INDEX, FUN = NULL, ..., simplify = TRUE)

Os vetores podem ser divididos em diferentes subsets e as funções aplicadas a estes subsets



Data Science Academy



# mapply()

## Versão multivariada da sapply()



Data Science Academy



vapply()

Similar a sapply() mas possui um tipo específico que deve ser retornado



Data Science Academy



by()

Versão orientada a objetos da `tapply()`  
aplicada em dataframes




Data Science Academy





# A família apply - uma forma elegante de fazer loops

- # apply() - arrays e matrizes
- # tapply() - os vetores podem ser divididos em diferentes subsets
- # lapply() - vetores e listas
- # sapply() - versão amigável da lapply
- # vapply() - similar a sapply, com valor de retorno modificado
- # rapply() - similar a lapply()
- # eapply() - gera uma lista
- # mapply() - similar a sapply, multivariada
- # by



Data Science Academy



# Se você estiver trabalhando com os objetos:

# list, numeric, character (list/vecor) => sapply ou lapply  
# matrix, data.frame (agregação por coluna) => by / tapply  
# Operações por linha ou operações específicas => apply



Data Science Academy



# Expressões Regulares



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



## Expressões Regulares

Recurso usado para verificar se existe um padrão em uma string ou vetor de caracteres



Data Science Academy



`grepl()`

Retorna TRUE quando um padrão é encontrado



Data Science Academy



grep()

Retorna um vetor de índices dos caracteres  
que contém o padrão especificado



Data Science Academy



sub()

Substitui o primeiro caracter encontrado de  
acordo com o padrão especificado



Data Science Academy



gsub()

Substitui todos os caracteres encontrados de acordo com o padrão especificado



Data Science Academy





# Trabalhando com Datas



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Data - representado por Date

Armazenados como número de dias desde  
1 de Janeiro de 1970



Data Science Academy



Time - representado por POSIXct

Armazenados como número de segundos  
desde 1 de Janeiro de 1970



Data Science Academy



## Formatando Data

```
# %d: dia do mês em 2 dígitos (13)
# %m: mês em 2 dígitos (01)
# %y: ano em 2 dígitos (82)
# %Y: ano em 4 dígitos (1982)
# %A: dia da semana (Friday)
# %a: dia da semana abreviado (Fri)
# %B: mês (July)
# %b: mês abreviado (Jul)
```



Data Science Academy



## Formatando Hora



- # %H: hora (00-23)
- # %M: minuto
- # %S: segundo
- # %T: formado reduzido para %H:%M:%S



Data Science Academy



# Pacote lubridate



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



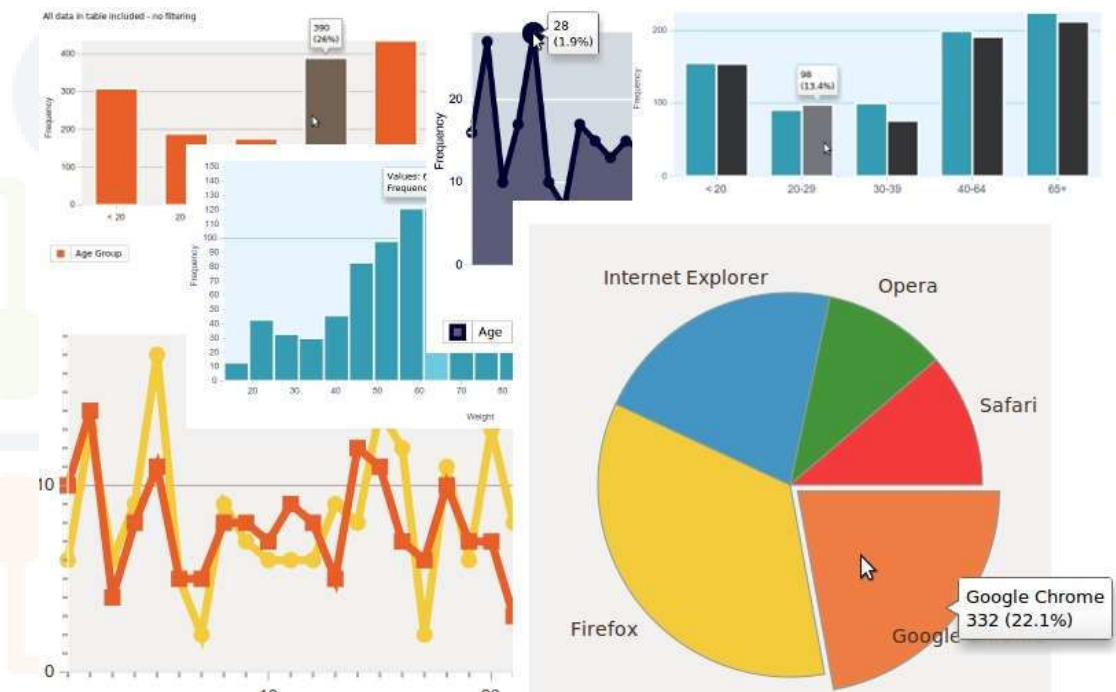
# Visualização de Dados

É a representação de dados em formato gráfico



Data Science Academy

# Visualização de Dados



Data Science Academy





# Pacote lubridate



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)




# Visualização de Dados

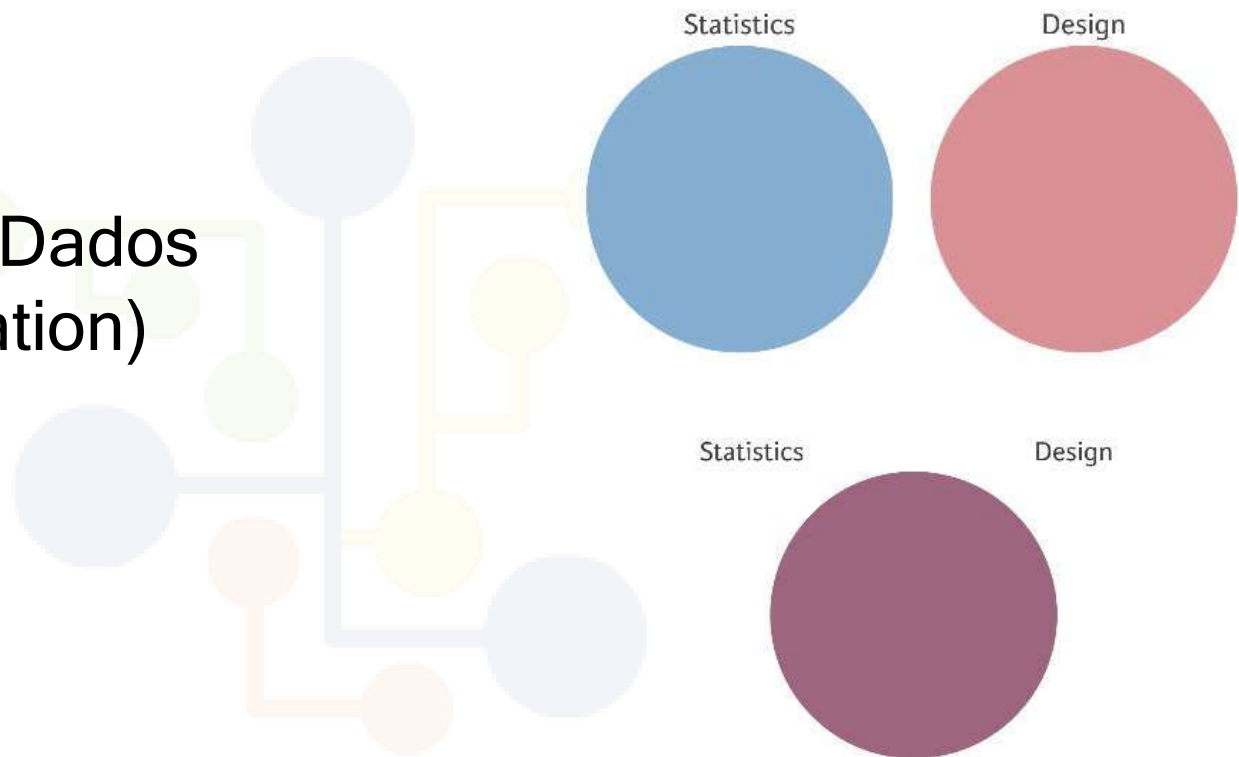
É a representação de dados em formato gráfico



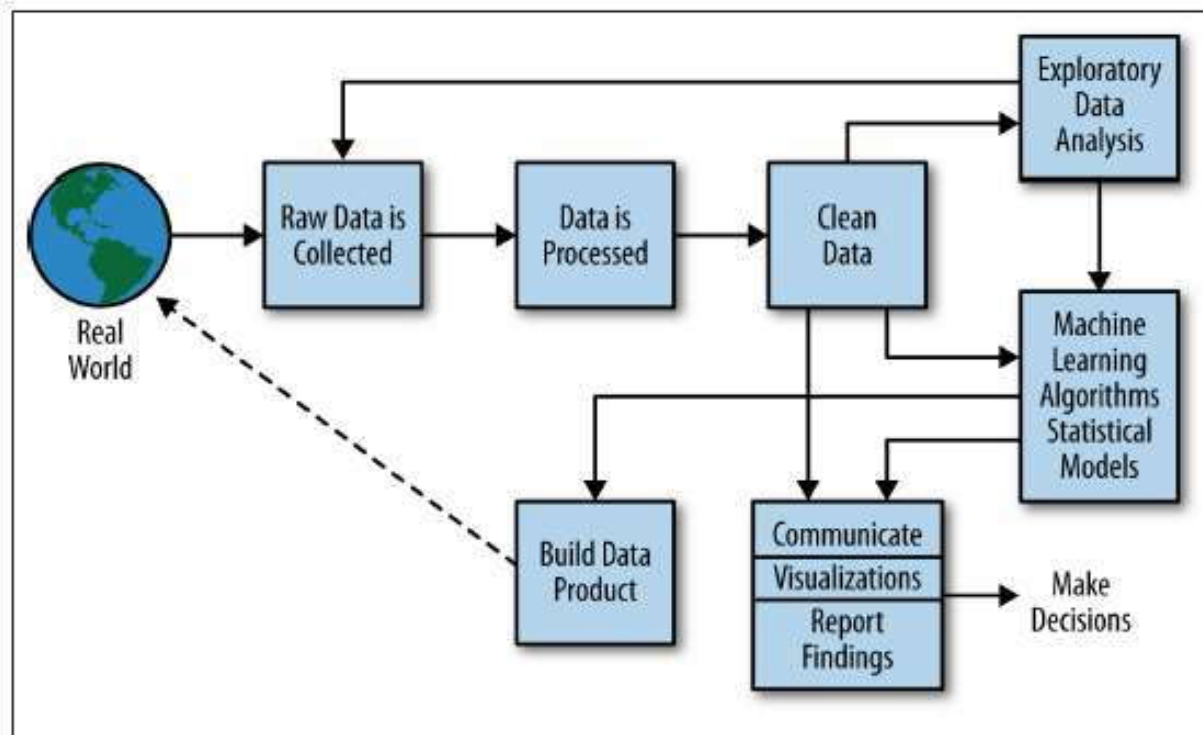
Data Science Academy



# Visualização de Dados (Data Visualization) (DataViz)



Data Science Academy



Data Science Academy



[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



# O que são Gráficos?

O gráfico é uma representação com forma geométrica construída de maneira exata e precisa a partir de informações numéricas obtidas através de pesquisas e organizadas em uma tabela



Data Science Academy




## Pacote Básico de Plotagem (Base Plotting System)

- **graphics** - contém as funções gráficas básicas, incluindo plot, hist e boxplot
- **grDevices** - contém as implementações de dispositivos gráficos como X11, pdf, PostScript, png, etc.



Data Science Academy





# Pacote Básico de Plotagem (Base Plotting System)

Os plots são objetos construídos através de funções e com atributos



Data Science Academy

# Tipos de Pontos



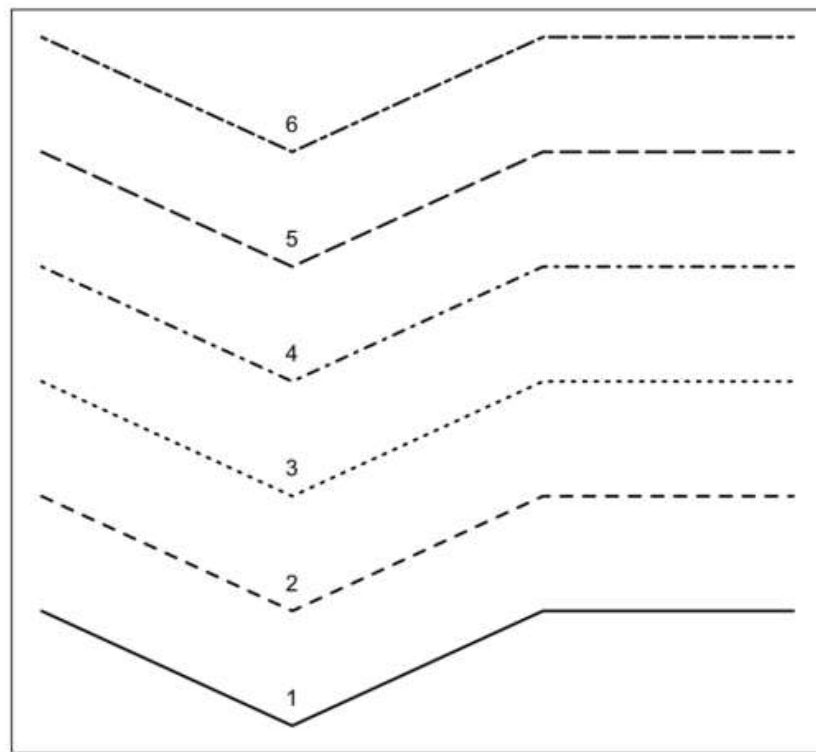
○ 1	▽ 6	⬡ 11	● 16	● 21
△ 2	⊠ 7	⊞ 12	▲ 17	■ 22
⊕ 3	⋆ 8	⊗ 13	◆ 18	◇ 23
⊗ 4	⬠ 9	◻ 14	● 19	▲ 24
◇ 5	⊕ 10	■ 15	● 20	▽ 25



Data Science Academy








# Tipos de Linhas



Data Science Academy

Peso e  
Tamanho

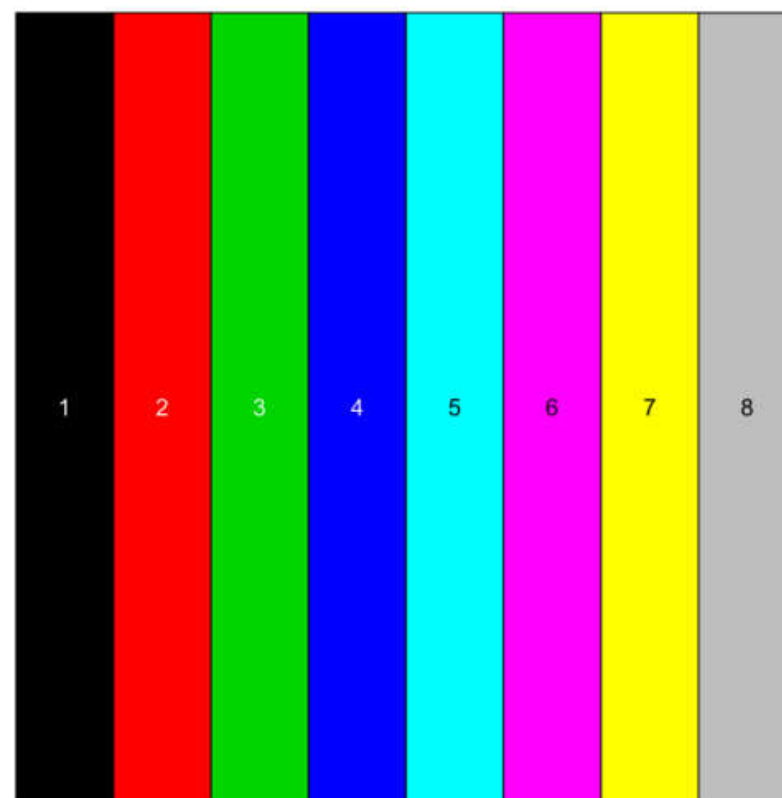
				
0.5	1	1.5	2	2.5



Data Science Academy



Cores



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



# Gramática dos Gráficos



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Elemento	Descrição
Dados	O conjunto de dados a ser analisado
Estética	A escala em que nós mapeamos os dados
Geometria	Os elementos visuais usados para representar os dados
Facets	Visualizar o gráfico em porções menores
Estatística	Representação e análise dos dados
Coordenadas	A área na qual o gráfico será construído
Temas	Visual geral do gráfico



Data Science Academy