Explorando dados no R - Parte 2

Disciplina de Ciência de Dados 2025.2 - UESC

Gabriel Rodrigues

2025-09-09

PARTE 1 - Prospecção básica

Importar datasets básicos do R

```
# Carregar pacote
library(datasets)
# Checar manual do pacote
library(help = "datasets")
```

A partir da importação do pacote `datasets`, vamos importar o dataset "iris".

```
## Selecionar dataset específico
iris <- datasets::iris
### Checar composição
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

```
dim(iris)
```

```
[1] 150   5
```

Alguns valores das tabelas podem ser vazios (`NA`), é bom compreender a completude dos dados.

```
## Selecionar dataset específico
iris <- datasets::iris

## Mostrar numero de valores ausentes (NAs)
sum(is.na(iris))
```

```
[1] 0
```

```
### Mostrar NAs por coluna
colSums(is.na(iris))
```

```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
           0            0            0            0            0
```

A função `summary` também mostra a quantidade de `NA`:

```
## Selecionar dataset específico
iris <- datasets::iris

## Sumário básico
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

PARTE 2 - Conceitos Importantes de EDA

Relembrando - Funções de Medidas de Tendência Central

```
## Selecionar dataset específico
iris <- datasets::iris

### Média
mean(iris$Sepal.Length)
```

```
[1] 5.843333
```

```
mean(iris$Petal.Length)
```

```
[1] 3.758
```

```
### Mediana
median(iris$Sepal.Length)
```

```
[1] 5.8
```

```
median(iris$Petal.Length)
```

```
[1] 4.35
```

Funções de Medidas de Dispersão

```
## Selecionar dataset específico
iris <- datasets::iris

### Mínimo e Máximo
min(iris$Sepal.Length)
```

```
[1] 4.3
```

```r
max(iris$Sepal.Length)
```

```
[1] 7.9
```

```r
### Variância e Desvio Padrão (sd)
var(iris$Sepal.Length)
```

```
[1] 0.6856935
```

```r
sd(iris$Sepal.Length)
```

```
[1] 0.8280661
```

```r
## O desvio padrão é representado pela mesma unidade de medida dos dados
```

Funções de Medidas de Relacionamento

COVARIÂNCIA - Calcula a covariância entre duas variáveis. Ela indica a direção da relação linear entre elas. Uma covariância positiva indica que as variáveis se movem na mesma direção (quando uma aumenta, a outra tende a aumentar) e vice-versa. **Ponto de atenção**: O valor da covariância não tem uma interpretação padronizada. Seu valor depende das unidades de medida das variáveis.

```r
## Selecionar dataset específico
iris <- datasets::iris

cov(iris[, c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width")])
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

```r
### Ou
cov(iris[, c(1:4)])
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

CORRELAÇÃO - Calcula a correlação, que também mede a força e a direção da relação linear entre duas variáveis. A grande vantagem é que o valor de correlação é padronizado e varia de -1 a 1.

- Correlação próxima de 1: Relação linear positiva forte.
- Correlação próxima de -1: Relação linear negativa forte.
- Correlação próxima de 0: Nenhuma relação linear.

```r
## Selecionar dataset específico
iris <- datasets::iris

# Função de correlação
cor(iris[, c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width")])
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

```r
### Ou
cor(iris[, c(1:4)])
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

PARTE 3 - Pacotes específicos para EDA

Instalar pacote `skimr` para criar um sumário geral de nossos dados.

```r
## Selecionar dataset específico
iris <- datasets::iris

# Instalar pacote "skimr"
# install.packages("skimr")

# Importar pacotes
library(skimr)

# Testar com nosso dataset
skim(iris)
```

Table 1: Data summary

Name	iris
Number of rows	150
Number of columns	5
_____	
Column type frequency:	
factor	1
numeric	4
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Species	0	1	FALSE	3	set: 50, ver: 50, vir: 50

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Sepal.Length	0	1	5.84	0.83	4.3	5.1	5.80	6.4	7.9	
Sepal.Width	0	1	3.06	0.44	2.0	2.8	3.00	3.3	4.4	
Petal.Length	0	1	3.76	1.77	1.0	1.6	4.35	5.1	6.9	
Petal.Width	0	1	1.20	0.76	0.1	0.3	1.30	1.8	2.5	