Exploração de Dados

Prof. Dr. Leandro Balby Marinho

Ciência de Dados Preditiva

Roteiro

1. Análise Exploratória de Dados

2. Relacionamento entre os dados

3. Amostragem de Dados

Introdução

Objetivo: entender melhor os dados para obter insights e ajudar na tomada de decisão.

As tarefas para atingir esse objetivo envolvem:

- Coleta de dados.
- Pré-processamento dos dados.
- **Sumarização**
- **Análise**
- **Interpretação**

Estatística Descritiva (ou Exploração de Dados): sumarização de dados com números ou figuras.

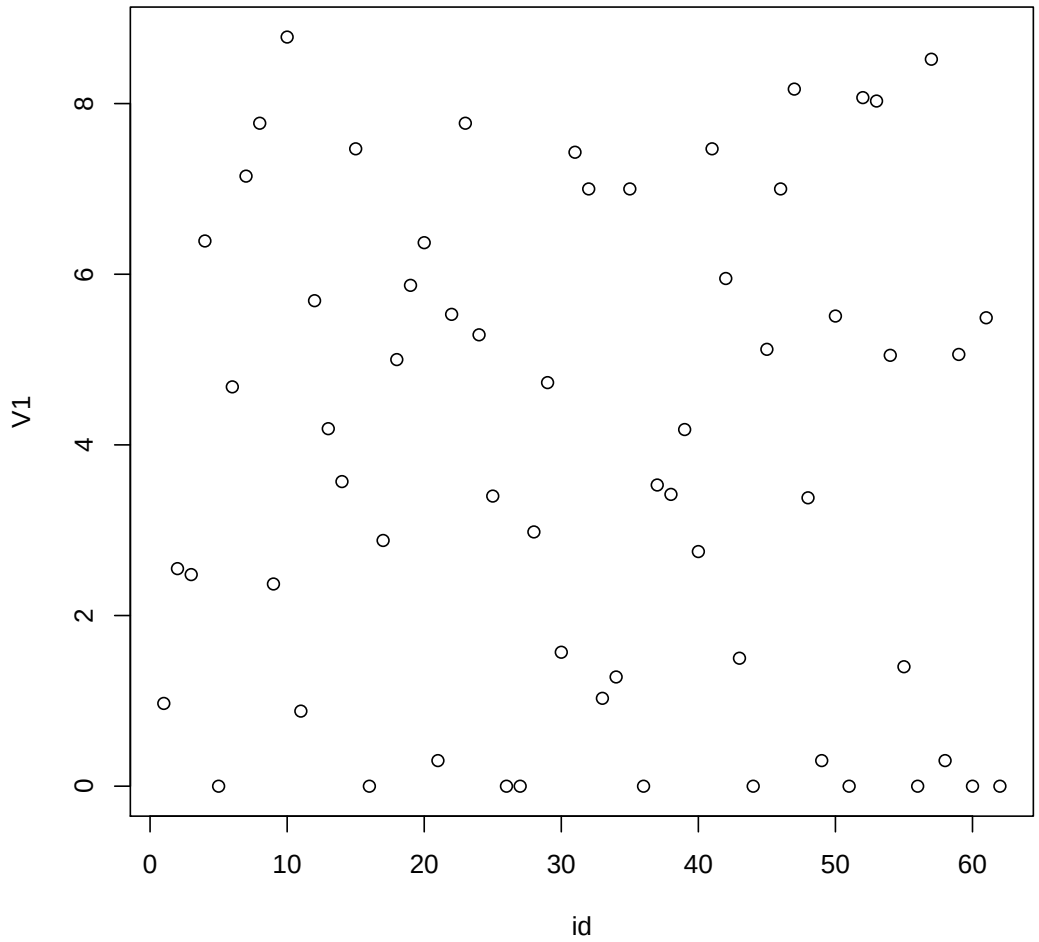
Sumário com 5 números

Considere as notas de uma disciplina abaixo:

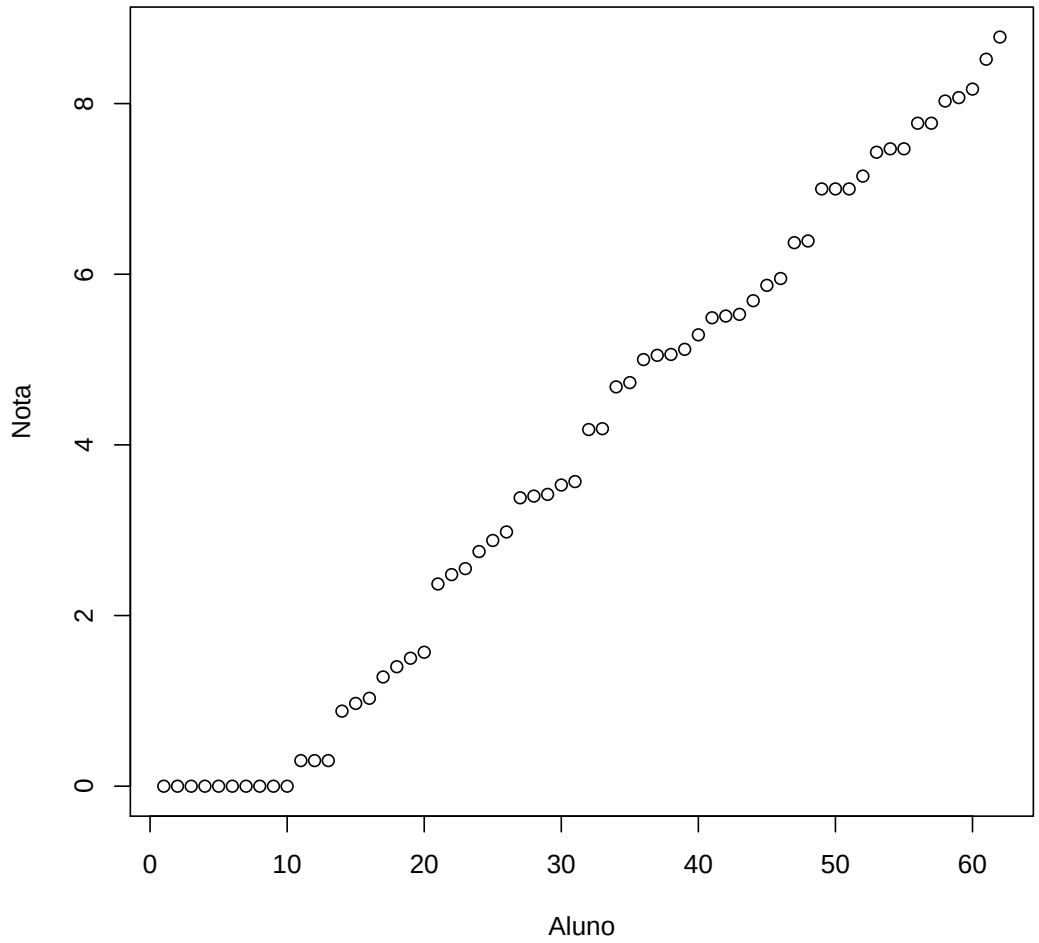
Aluno	Nota
1	1,0
2	2,5
3	2,5
4	6,4
5	0
6	4,7
7	7,2
8	7,8
9	2,4
10	8,9
⋮	⋮
62	0

Como sumarizar esses dados com poucos números?

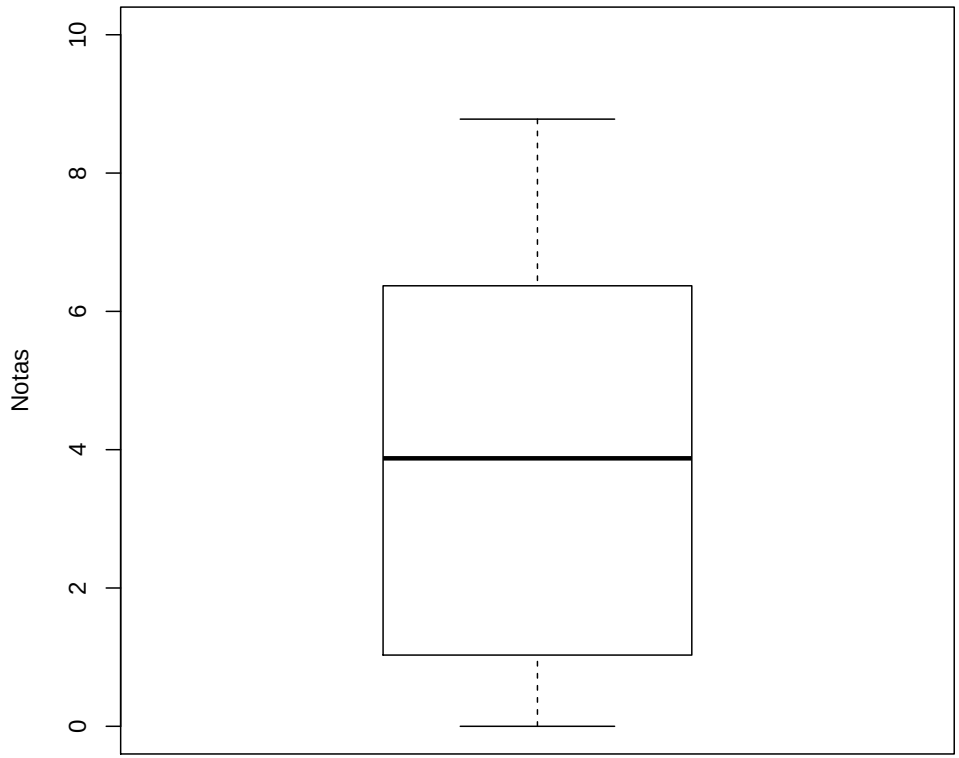
Ideia 1: Scatter Plot



Ideia 2: Scatter Plot com Valores Ordenadas



Ideia 3: Box Plot

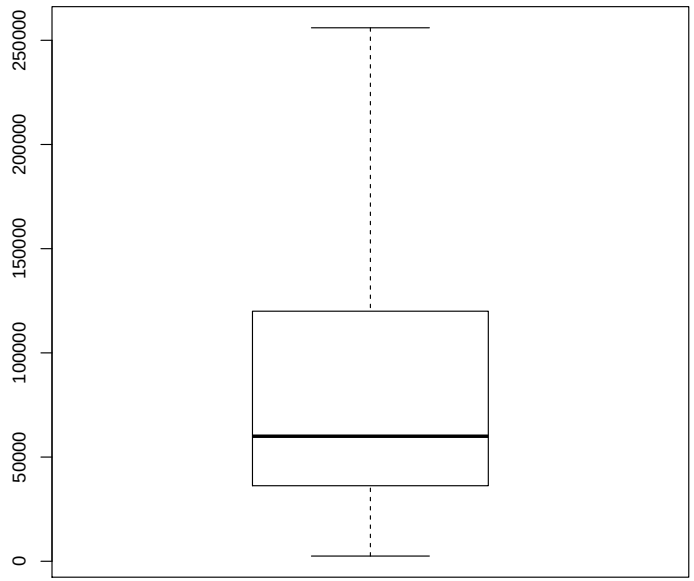


Min.	1st Qu.	Median	3rd Qu.	Max.
0	1.093	3.875	6.265	8.780

Valores Extremos

Considere os salários dos jogadores do Barcelona em 2015.

Jogador	Salário (libras esterlinas)
1	55.000
2	33.500
3	15.500
4	130.000
5	130.000
6	90.000
7	40.000
8	55.500
9	75.000
10	60.000
11	60.000
12	50.000
13	120.000
14	35.000
15	120.000
16	150.000
17	75.000
18	85.000
19	37.500
20	20.000
21	5.000
22	150.000
23	256.000
24	60.000
25	200.000
26	5.000
27	2.500



Min.	1st Qu.	Median	3rd Qu.	Max.
2.500	36.250	60.000	120.000	256.000

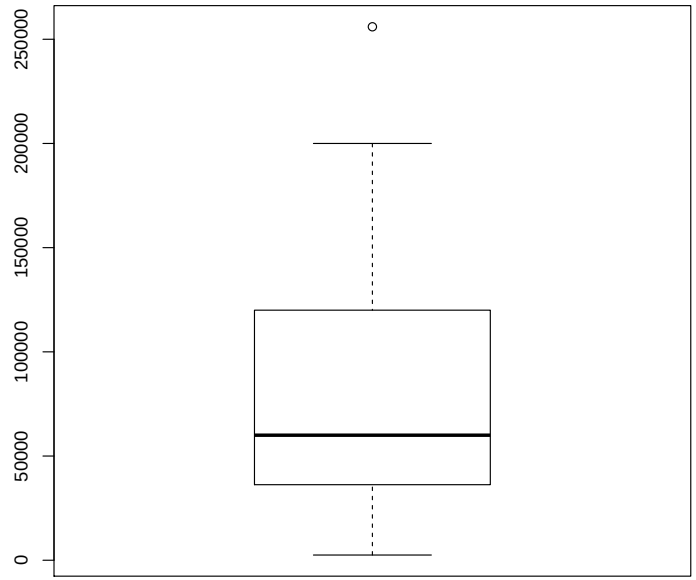
Box Plots com Valores Extremos (Outliers)

- Outliers são valores muito maiores ou menores que o resto dos dados.
- Boxplots podem ser usados para identificar outliers.
- Primeiro calcula-se o IQR (Inter Quartile Range), i.e., $Q3 - Q1$.
- Um outlier é considerado qualquer valor maior ou menor que 1,5 vezes o IQR, i.e,
  - Maior que $Q3 + 1,5 \times IQR$ ou
  - Menor que $Q1 - 1,5 \times IQR$.

Valores Extremos

Considere novamente os salários dos jogadores do Barcelona em 2015.

Jogador	Salário (libras esterlinas)
1	55.000
2	33.500
3	15.500
4	130.000
5	130.000
6	90.000
7	40.000
8	55.500
9	75.000
10	60.000
11	60.000
12	50.000
13	120.000
14	35.000
15	120.000
16	150.000
17	75.000
18	85.000
19	37.500
20	20.000
21	5.000
22	150.000
23	256.000
24	60.000
25	200.000
26	5.000
27	2.500



Min.	1st Qu.	Median	3rd Qu.	Max.
2.500	36.250	60.000	120.000	256.000

Média

Seja x_i o valor da i-ésima observação, a média amostral é calculada por:

$$\bar{x} = \sum_{1}^{n} \frac{x_i}{n}$$

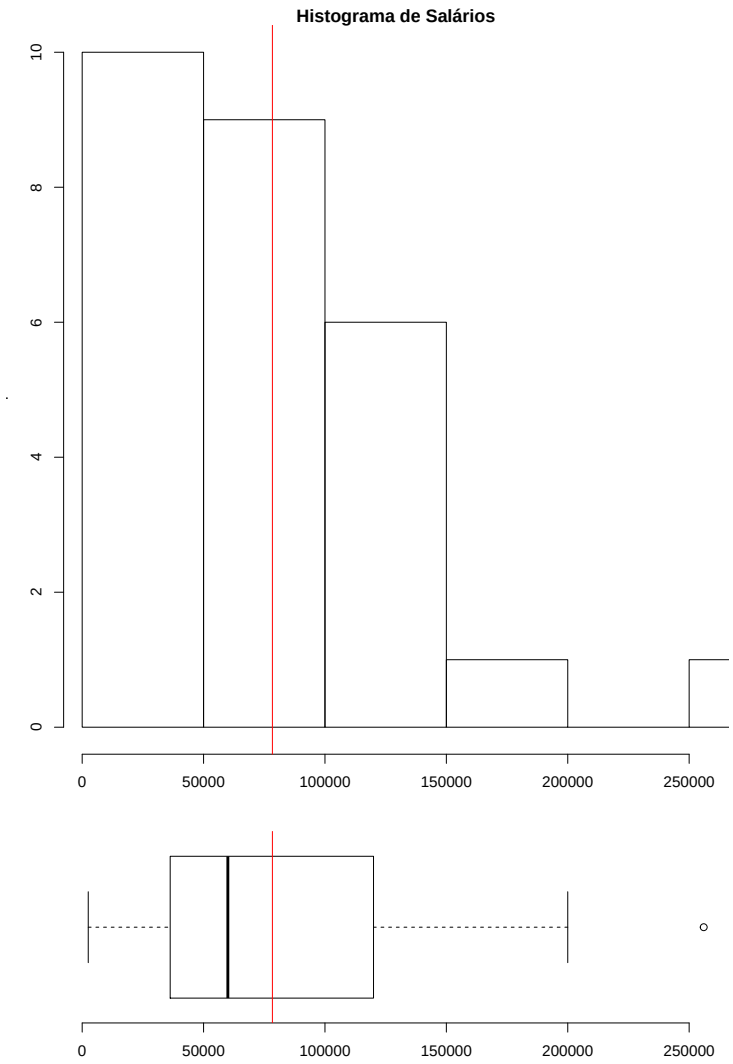
Considerando o exemplo do slide passado a média é:

a) Maior que a mediana.

b) Menor que a mediana.

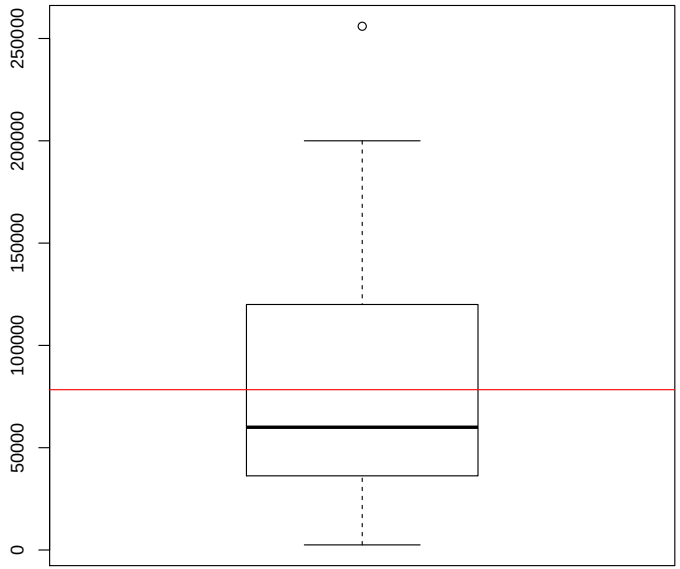
c) Igual a mediana.

Média


Histograma de Salários
0
2
4
6
8
10
0
50000
100000
150000
200000
250000
0
50000
100000
150000
200000
250000

Média

A média (linha vermelha) é bem maior que a mediana pois o outlier a puxa para cima.



A média não é uma estatística robusta pois é afetada por valores extremos.

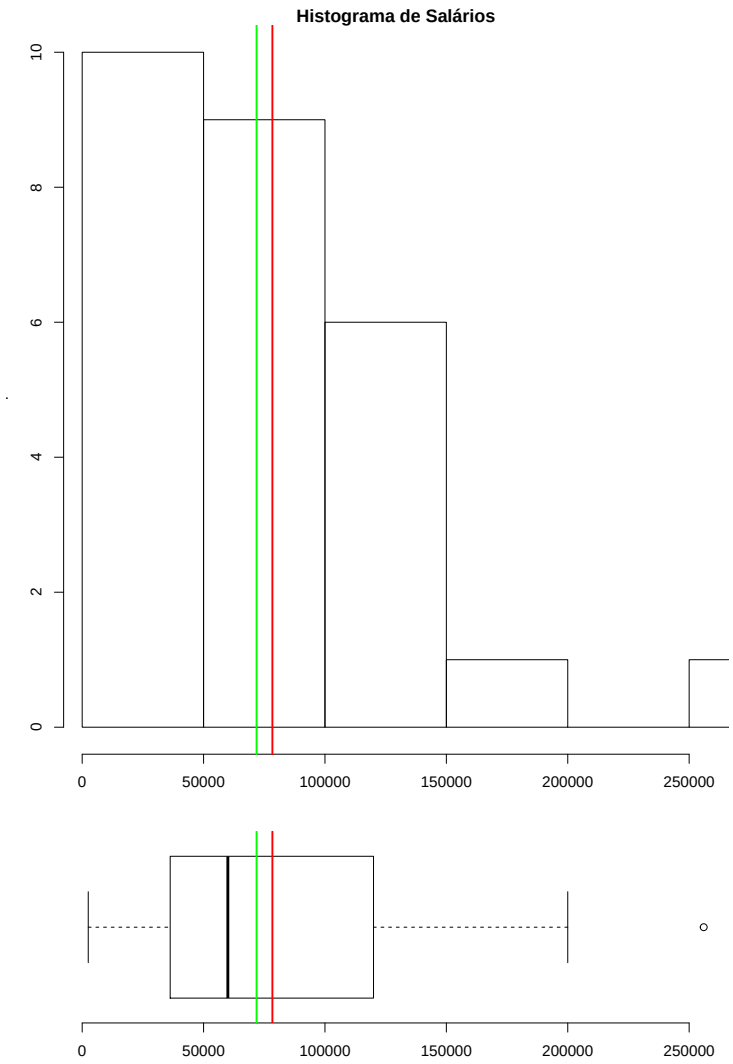
Média Aparada

A média aparada é a média retirando-se os $X\%$ maiores e menores valores. Ela ameniza o efeito de outliers e portanto é considerada uma estatística robusta.

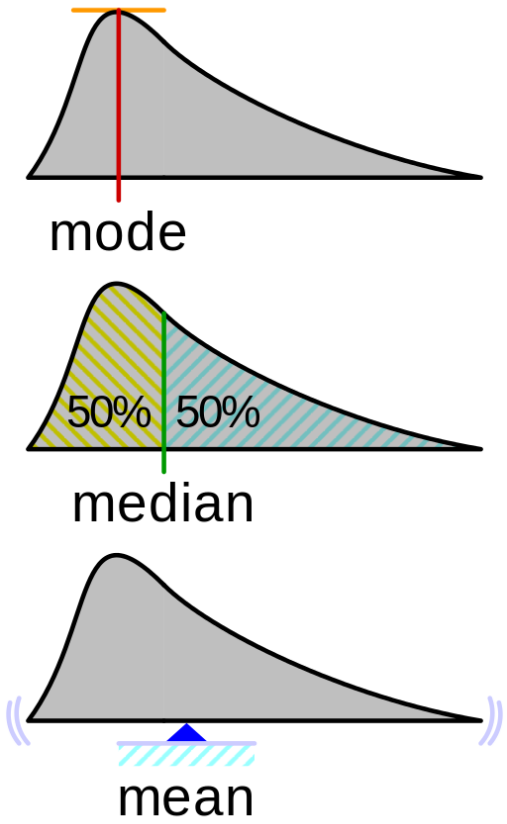
Na tabela abaixo temos a mediana, média e média aparada retirando-se os 10% maiores e menores salários.

Mediana	Média	Média Aparada (10%)
60.000	78.351,85	71.826,09

Média Aparada



Moda, Média e Mediana



Dispersão dos Dados

Como medimos a dispersão dos dados em relação aos valores centrais?

- Ideia 1: Valor máximo - Valor mínimo.
- Ideia 2: IQR
- Ideia 3: Desvio Padrão

Variância (populacional):

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Desvio padrão (populacional):

$$\sigma = \sqrt{\sigma^2}$$

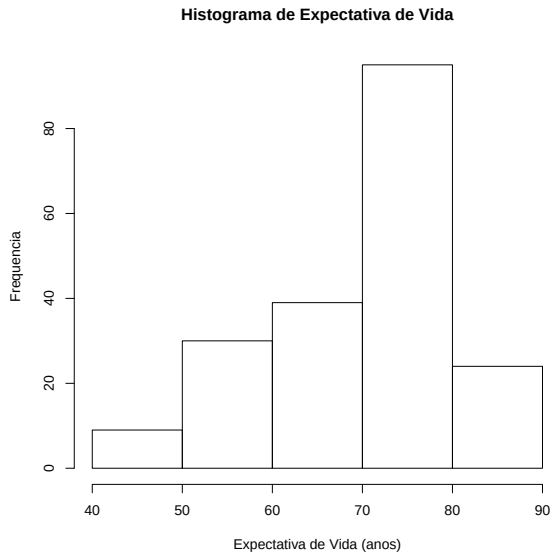
Dispersão dos Dados

Considerando novamente o exemplo dos salários do Barcelona:

	dados originais	aparados	robusta?
Mediana	60.000	60.000	Sim
Média	78.351,85	72.826,09	Não
range	253.500	145.000	Não
IQR	83.750	66.250	Sim
d.p.	61.835	43.139	Não

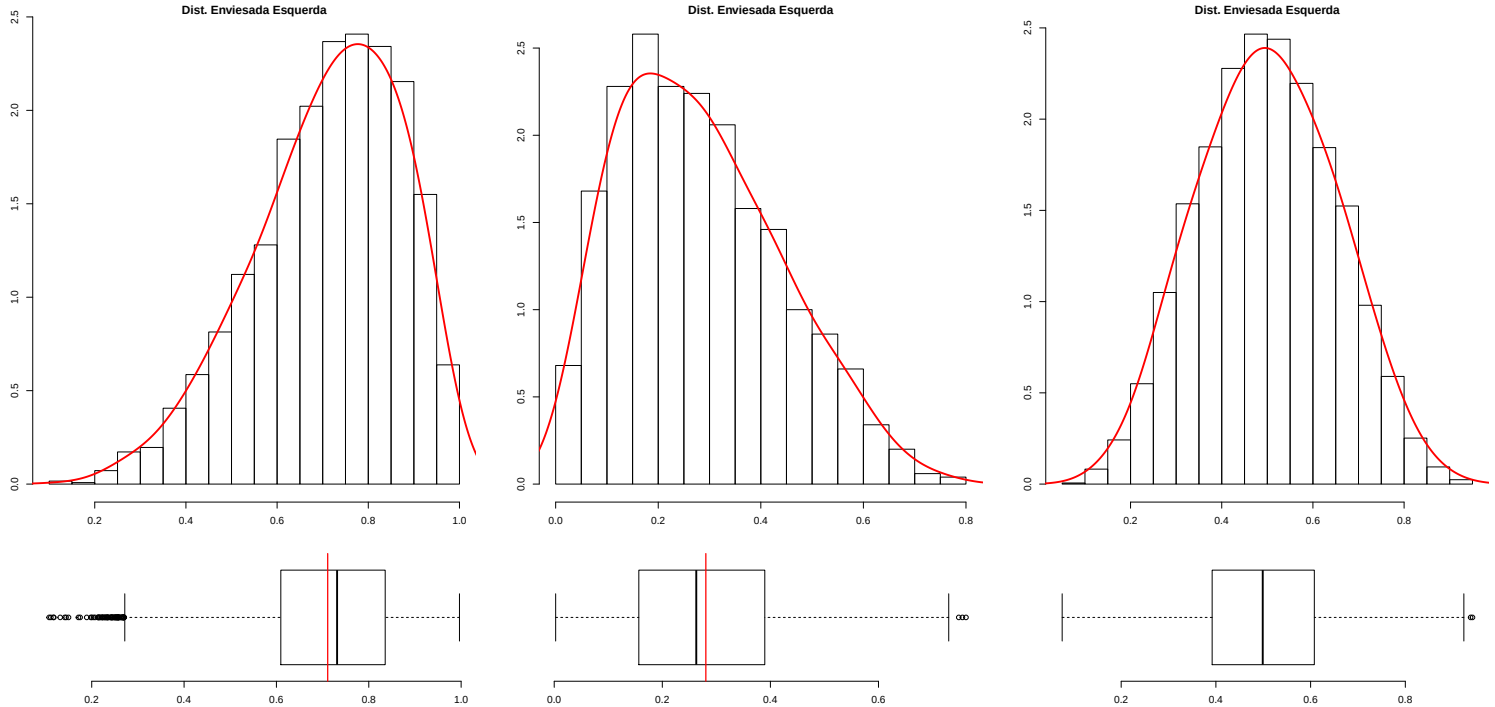
Distribuição dos Dados

- Podemos visualizar a distribuição de variáveis quantitativas por meio de histogramas.
- O eixo x é composto por intervalos de valores (também chamados de bins).
- O eixo y contém a quantidade de observações que caem dentro de cada bin.
- O histograma abaixo mostra a distribuição das expectativas de vida (em ano) de 197 países.



O que podemos aprender com os histogramas?

- Quantidade de picos: unimodal, bimodal, etc.
- Simetria: simétricos, enviesados para a direita ou esquerda.
- Dispersão dos dados e outliers.

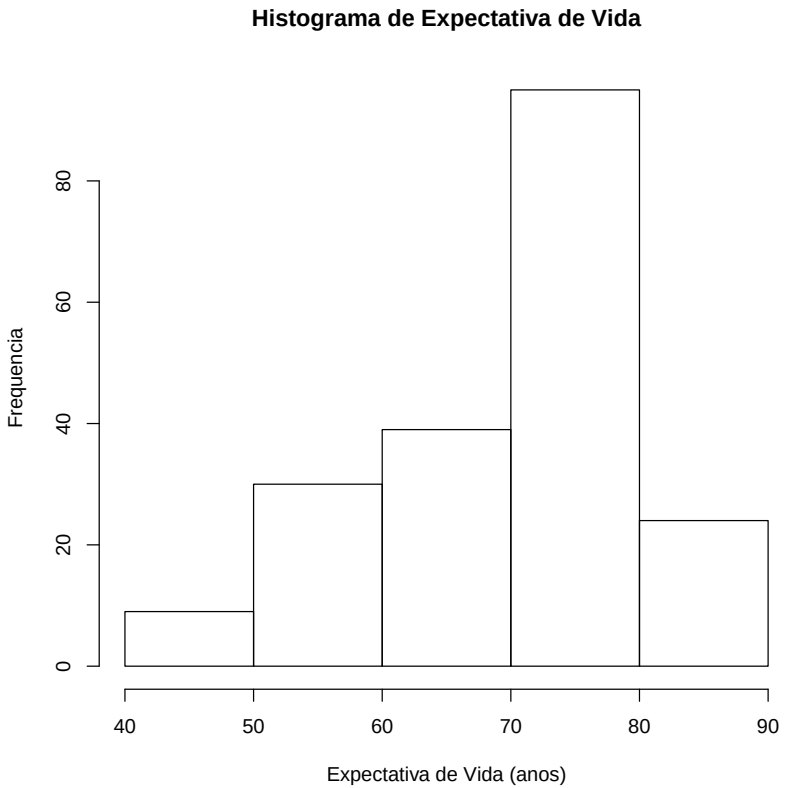


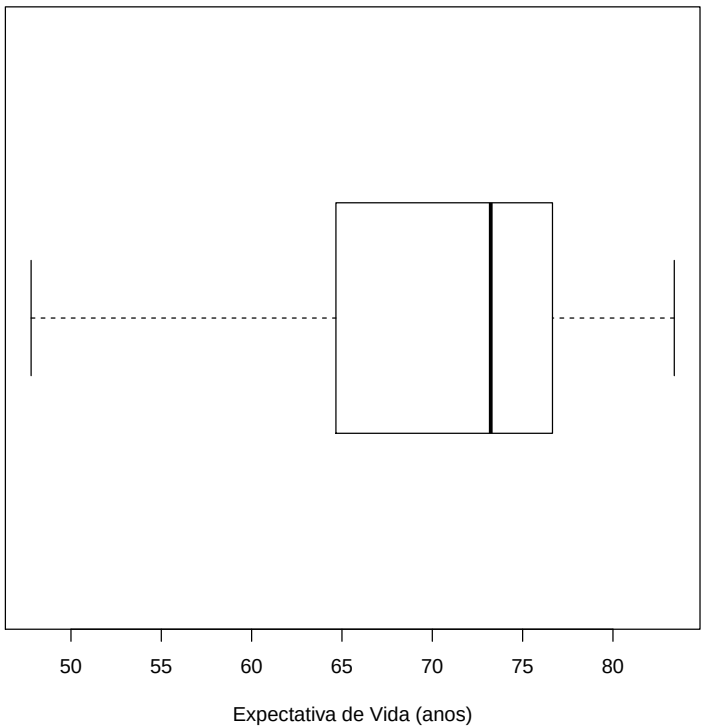
Distribuição dos Dados

Considere novamente o histograma de expectativa de vida. Ele é enviesado à direita, esquerda ou é simétrico?

Distribuição dos Dados

Considere novamente o histograma de expectativa de vida. Ele é enviesado à direita, esquerda ou é simétrico?





Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
47.79	64.67	73.24	69.86	76.65	83.39

Variáveis Categóricas (ou qualitativas)

- Exemplos: Sexo (M,F), Cor (Azul, Vermelho, etc.) e Nacionalidade (Brasileiro, Chileno, etc.)
- Não podemos realizar operações matemáticas com essas variáveis (e.g. máximo, mínimo e média).
- Podemos contar quantas observações ocorrem em cada nível da variável.

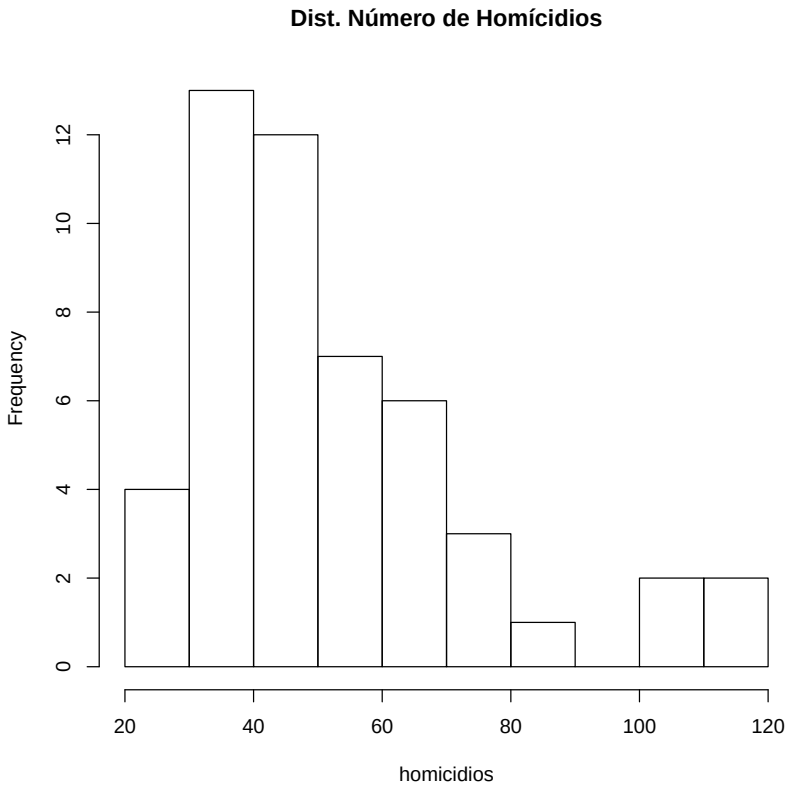
Exemplo Variável Quantitativa vs Qualitativa

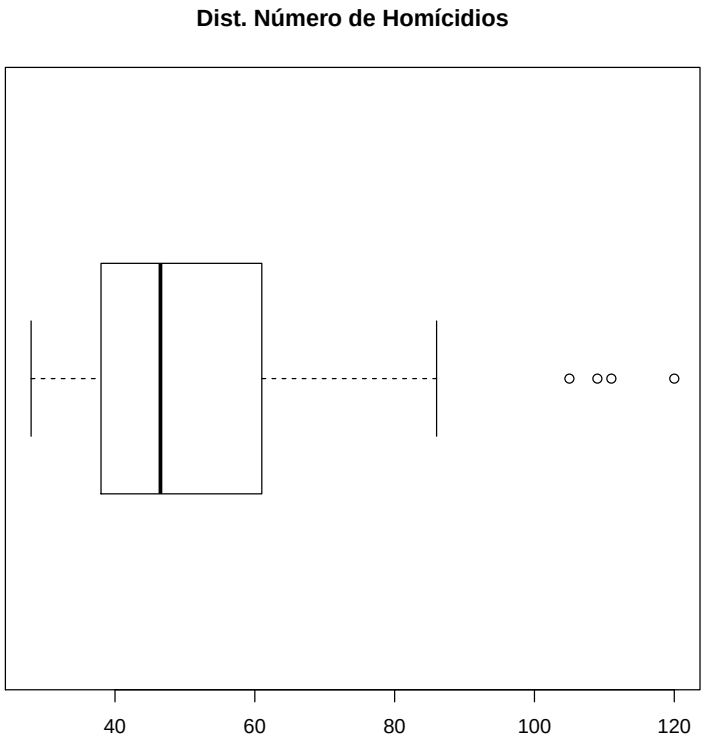
Considere os dados sobre as 50 cidades mais violentas do mundo por número de homícidios:

Cidade	País	Nr. Homícidios
Caracas	Venezuela	120
San Pedro Sula	Honduras	111
San Salvador	El Salvador	109
Acapulco	México	105
Maturín	Venezuela	86
Distrito Central	Honduras	74
Valencia	Venezuela	72
Palmira	Colômbia	71
Cidade do Cabo	África do Sul	66
Cali	Colômbia	64
Ciudade Guayana	Venezuela	62
Fortaleza	Brasil	61
⋮	⋮	⋮
Obregón	México	28

Visualizando a Variável Quantitativa

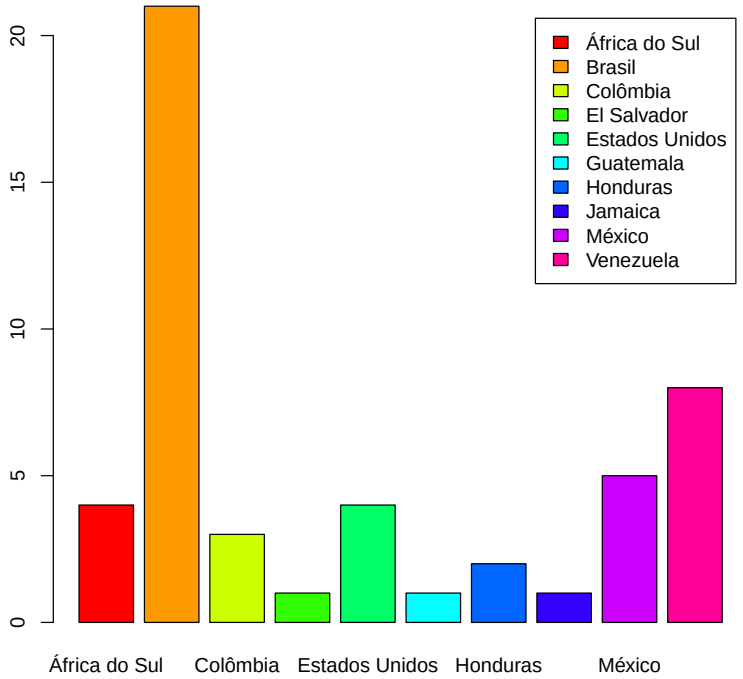
Para summarizar a parte quantitativa dos dados podemos usar histogramas e boxplots.





Visualizando Variáveis Categóricas

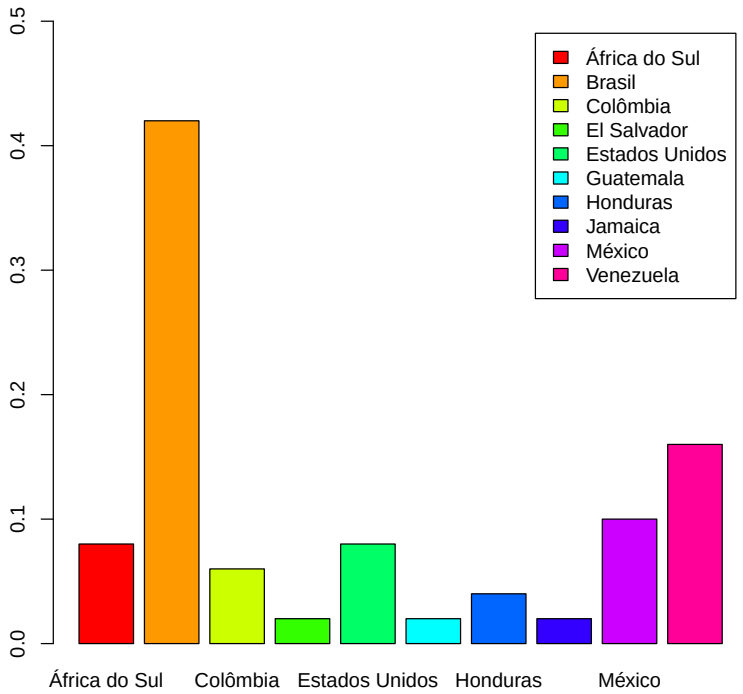
Qual o número de cidades no ranking por país?



Ideia 1: Contar a quantidade de cidades no ranking por país.

Visualizando Variáveis Categóricas

Qual a porcentagem de cidades no ranking por país?



Ideia 2: Dividir a contagem pelo total de cidades.

Exercícios

a) Calcule a média, mediana, moda e desenhe o boxplot do seguinte conjunto de dados:
$(1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 15, 16, 16)$

b) A distribuição acima é enviesada? Para que lado?

c) Considere os salários no Brasil como um todo. Como você acha que é a distribuição? Desenhe um histograma e boxplot que captura sua intuição.

d) Considere as idades dos alunos dessa turma. Como você acha que é a distribuição? Desenhe um histograma e boxplot que captura sua intuição.

Roteiro

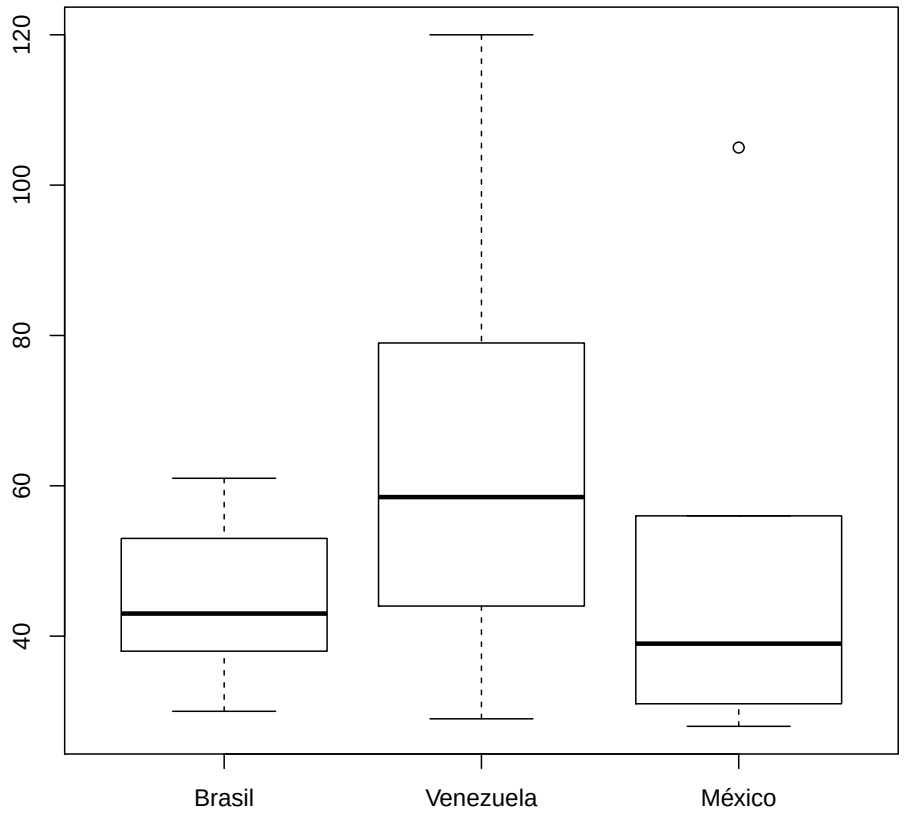
1. Análise Exploratória de Dados

2. Relacionamento entre os dados

3. Amostragem de Dados

Relação entre variáveis qualitativas e quantitativas

Ideia: Visualizar a distribuição da variável quantitativa para cada valor da variável qualitativa.



Relação entre variáveis categóricas

Ideia 1: Construir uma tabela de contingência contendo as frequências de cada nível das variáveis.

Gênero	frequência	frequência relativa
Masculino	5.457	0,545
Feminino	4.543	0,454
Total	10.000	1,000

Saiu	frequência	frequência relativa
Sim	2.037	0,203
Não	7.963	0,796
Total	10.000	1,000

Os dados acima se referem aos dados de clientes de um banco onde "Saiu" indica se um cliente saiu ou continua no banco.

Relação entre variáveis categóricas

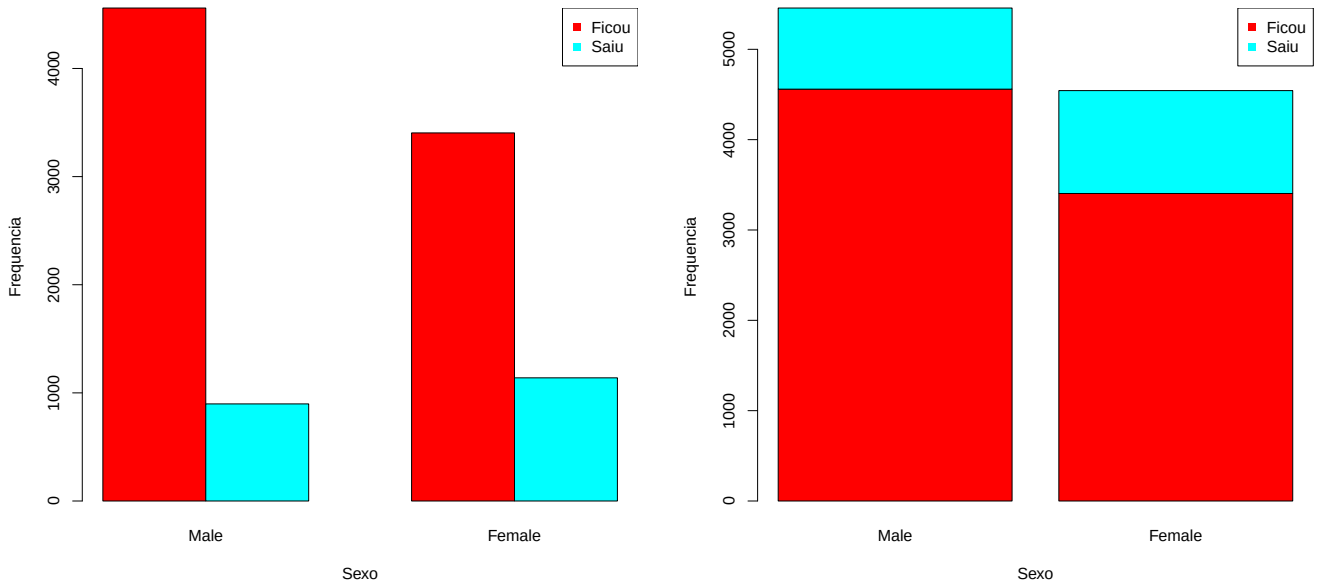
Ideia 1: Construir uma tabela de contingência contendo as frequências de cada nível das variáveis.

	Gênero	
Saiu	Masculino	Feminino
Sim	898	1.139
Não	4.559	3.404

	Gênero	
Saiu	Masculino	Feminino
Sim	0,089	0,113
Não	0,455	0,340

Relação entre variáveis categóricas

Ideia 2: Usar barplots com vários níveis.



Quem mais deixou o banco homens ou mulheres?

Usando proporções

Distribuição marginal é a distribuição de apenas uma variável em uma tabela de contingência.

	Sexo		
Saiu	Masculino	Feminino	Total
Sim	898	1.139	2.037
Não	4.559	3.404	7.963
	5.457	4.543	10.000

A distribuição fica nas margens da tabela.

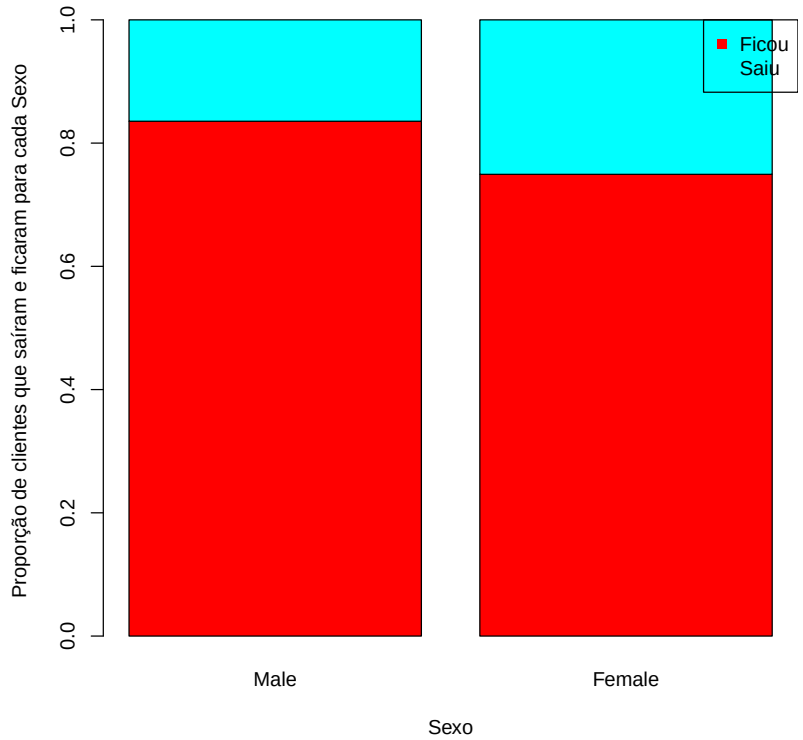
Distribuição condicional de uma variável categórica é sua distribuição com um valor fixo da segunda variável.

	Sexo	
Saiu	Masculino	Feminino
Sim	0,164	0,250
Não	0,835	0,749

Distribuição condicional de *Saiu* dado *Gênero*.

Usando proporções

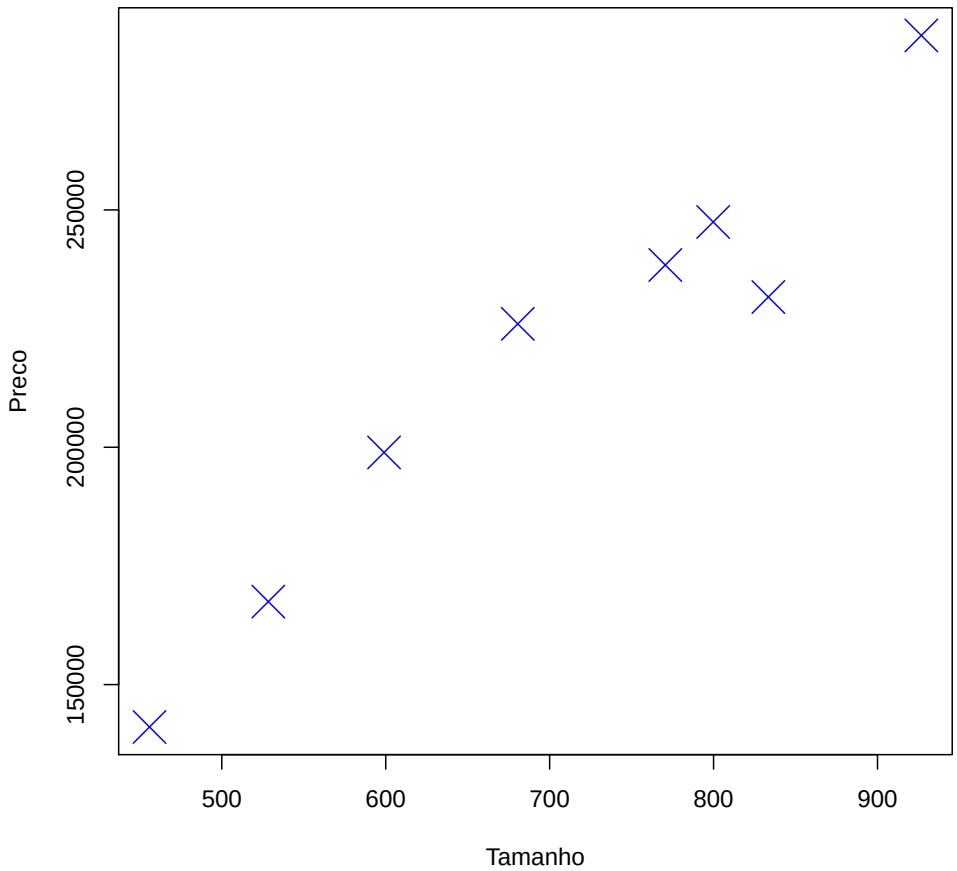
Note que usando a distribuição condicional fica mais fácil comparar as variáveis.



Quem mais deixou o banco homens ou mulheres?

Relação entre variáveis quantitativas

Ideia: Verificar como as variáveis variam em conjunto.



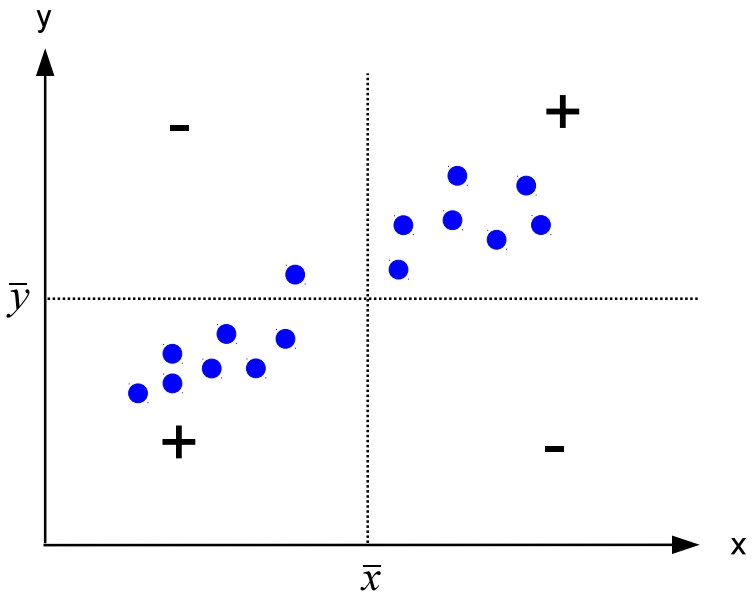
Correlação

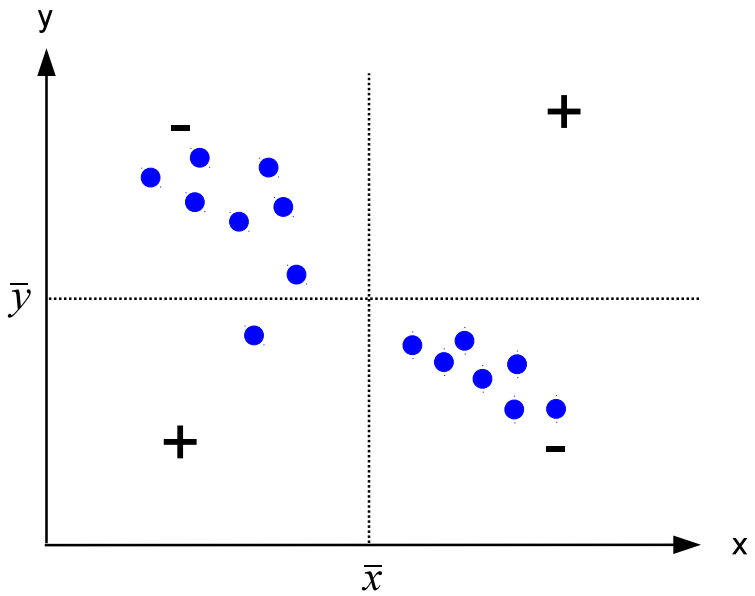
- Variável 1: $x_1, x_2, \ldots, x_n$
- Variável 2: $y_1, y_2, \ldots, y_n$

$$\text{Correlação} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Se $x_i = y_i$ ou $y_i = ax_i + b\,(a > 0)$ então correlação = +1.
- Se $y_i = ax_i + b\,(a < 0)$ então correlação = -1.

Correlação





Roteiro

1. Análise Exploratória de Dados
2. Relacionamento entre os dados
3. Amostragem de Dados

Amostragem de Dados

Inferência Estatística: tirar conclusões com base nos dados.

- População: grupo que estamos interessados em tirar conclusões.
- Censo: coleção de dados de toda população.
- Amostra: Um subconjunto da população.
- Estatística: valor calculado dos dados observados. Usado para estimar uma caracaterística (parâmetro) da população.

De forma a garantir uma boa estimativa do parâmetro (boa generalização) precisamos de uma amostra **representativa**.

Amostragem Randômica

Como selecionar uma amostra representativa? **Randomização**

Alguns métodos importantes de amostragem randômica são:

- Amostra Randômica Simples (ARS): Cada amostra possível de tamanho n da população tem a mesma probabilidade de ser escolhida.
- Amostra Estratificada: Divide a população em subgrupos não sobrepostos e escolhe uma ARS dentro de cada subgrupo.