



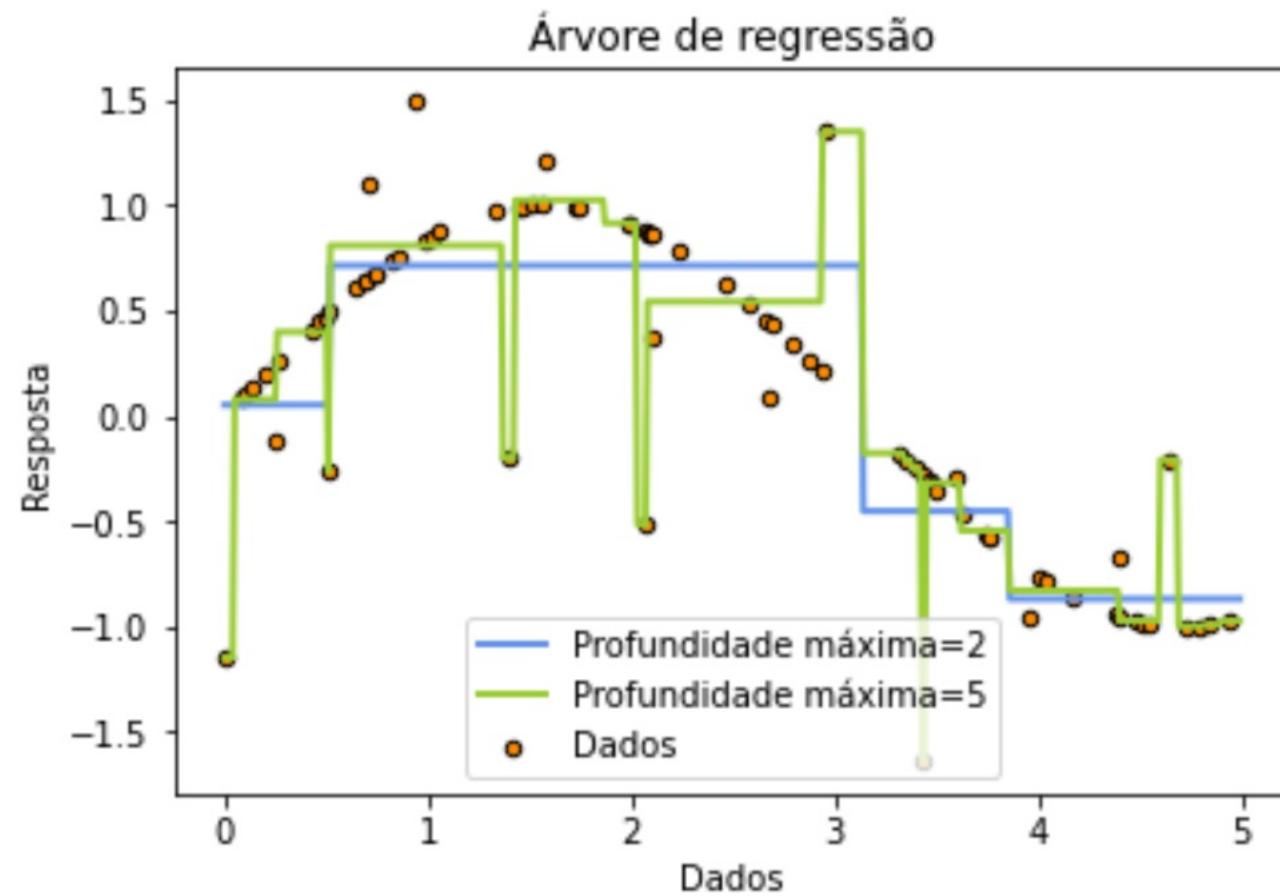
escola  
britânica de  
artes criativas  
& tecnologia

## **Profissão: Cientista de Dados**

Módulo 11: Árvores de regressão

# Introdução: Árvores de regressão

# Árvores de regressão

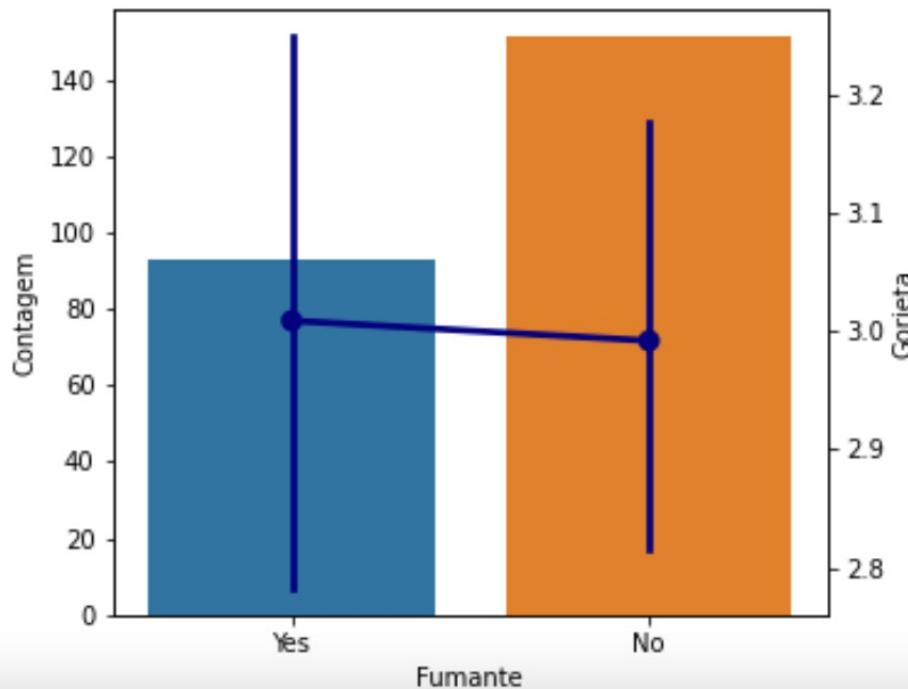
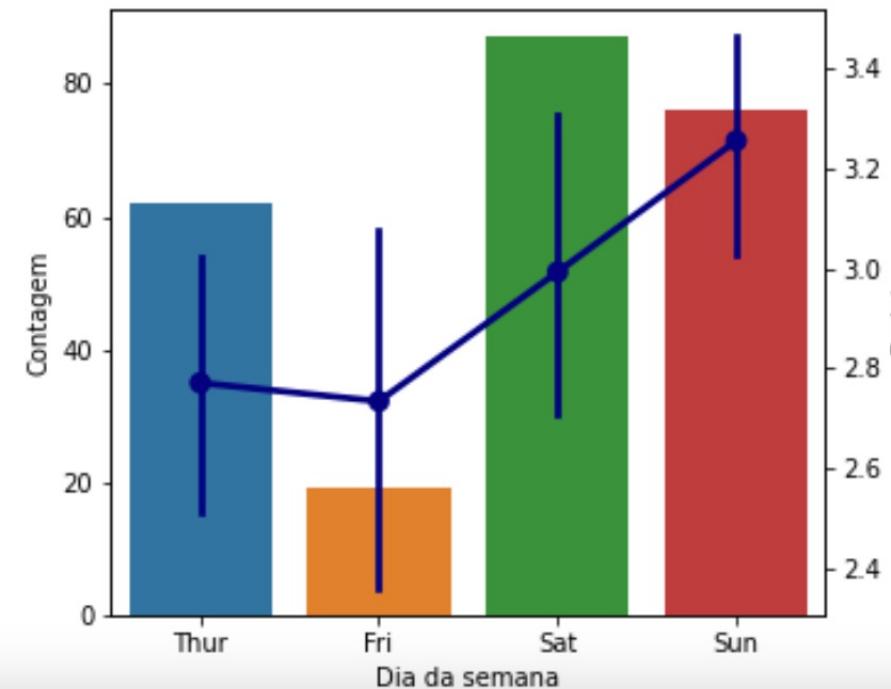
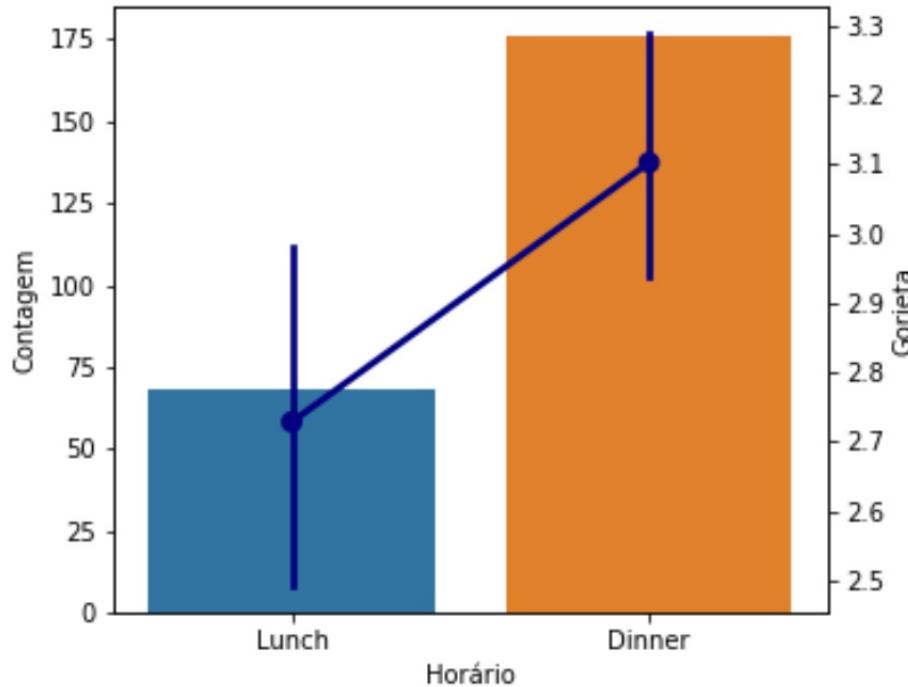
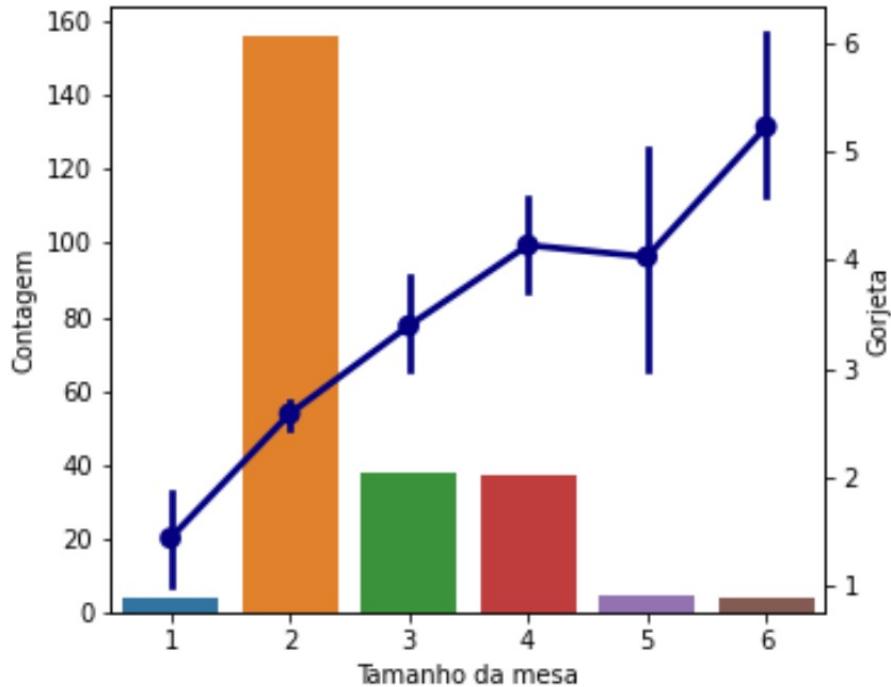


# O problema da gorjeta

---



# Como trabalhar menos e ganhar mais gorjeta?



# Introdução: Medidas de impureza

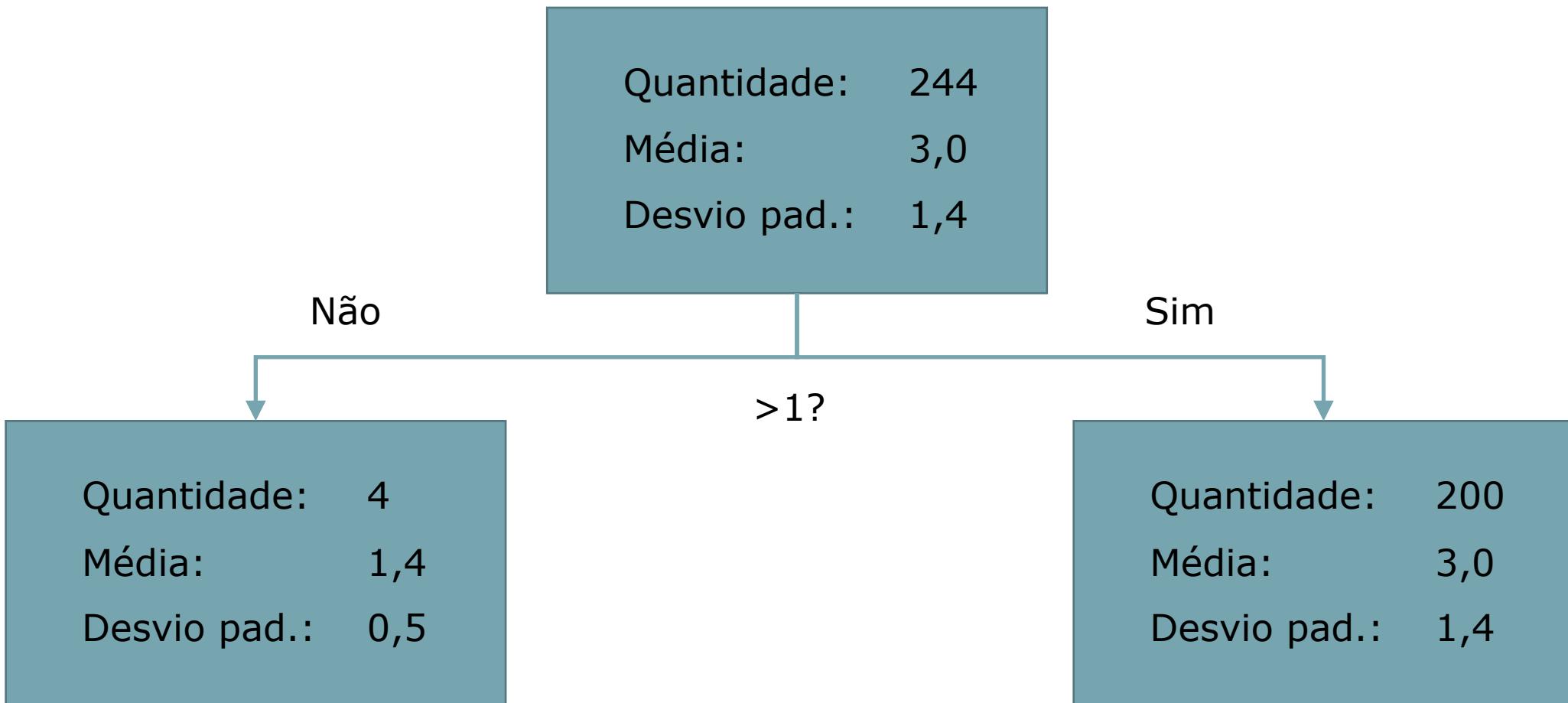
# Como maximizar a gorjeta?

Quantidade: 244

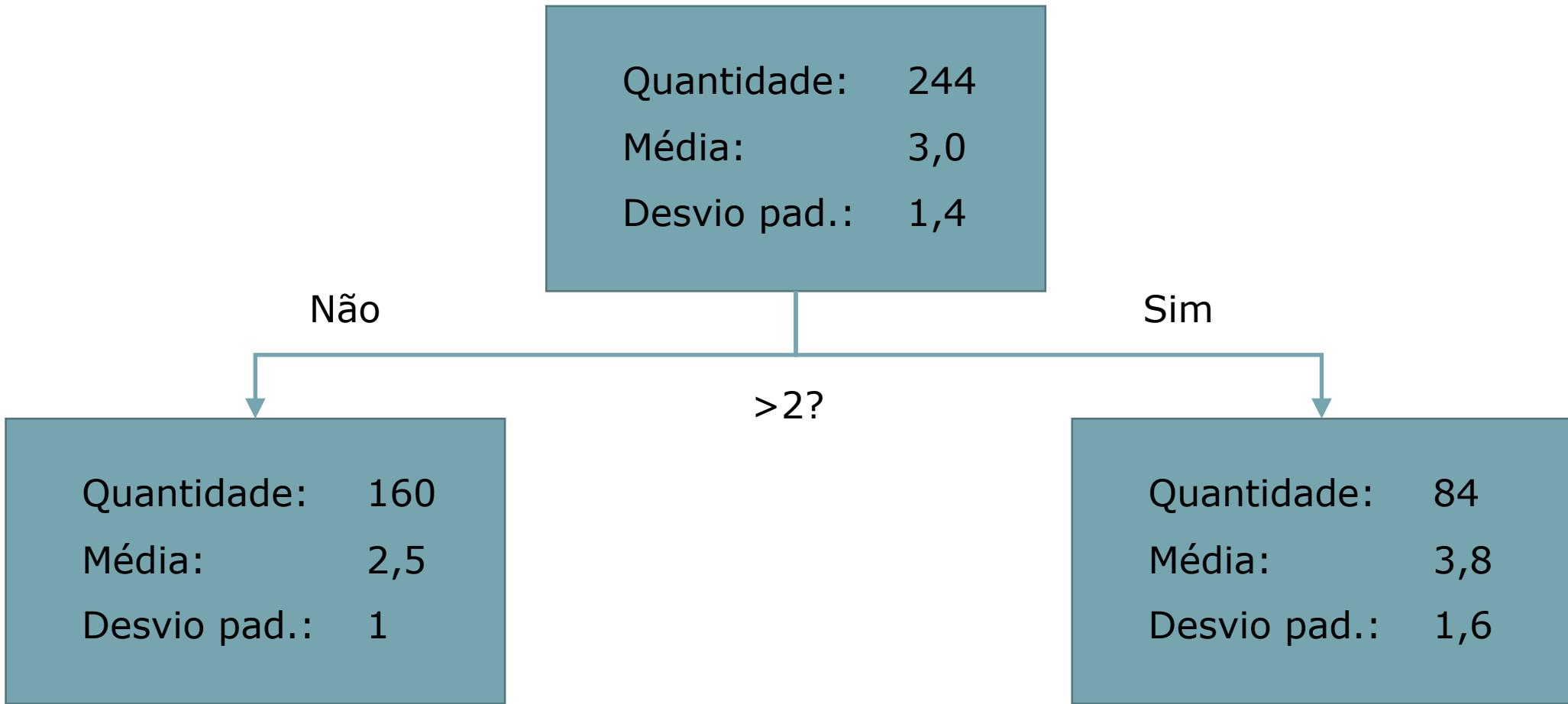
Média: 3,00

Desvio pad.: 1,38

# Como maximizar a gorjeta?



# Como maximizar a gorjeta?



# Critérios

Erro quadrático médio

$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2$$

Erro absoluto médio

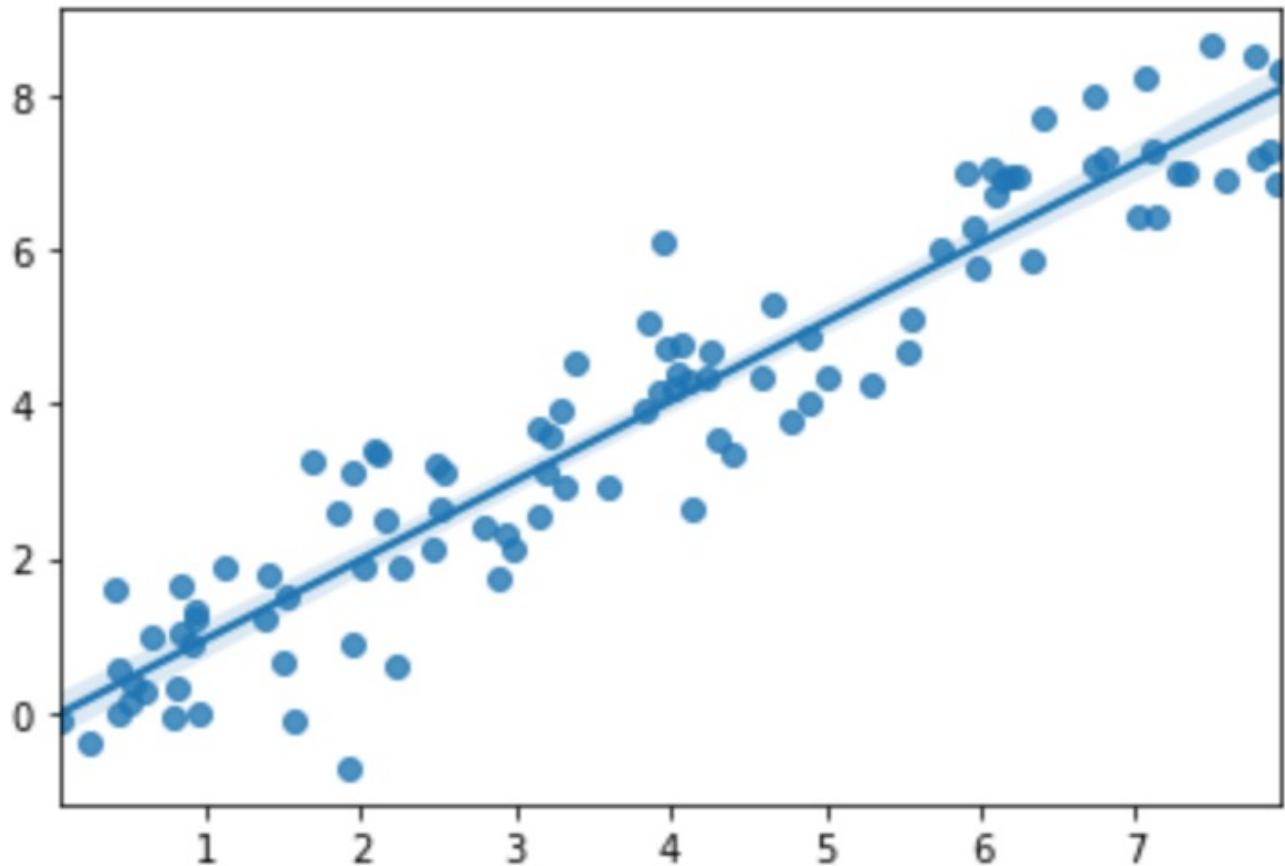
$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Q_m} |y - median(y)_m|$$

# Coeficiente de determinação

R2 – coeficiente de determinação

$$R^2 = 1 - \frac{SQM}{SQT}$$

$$= 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$$

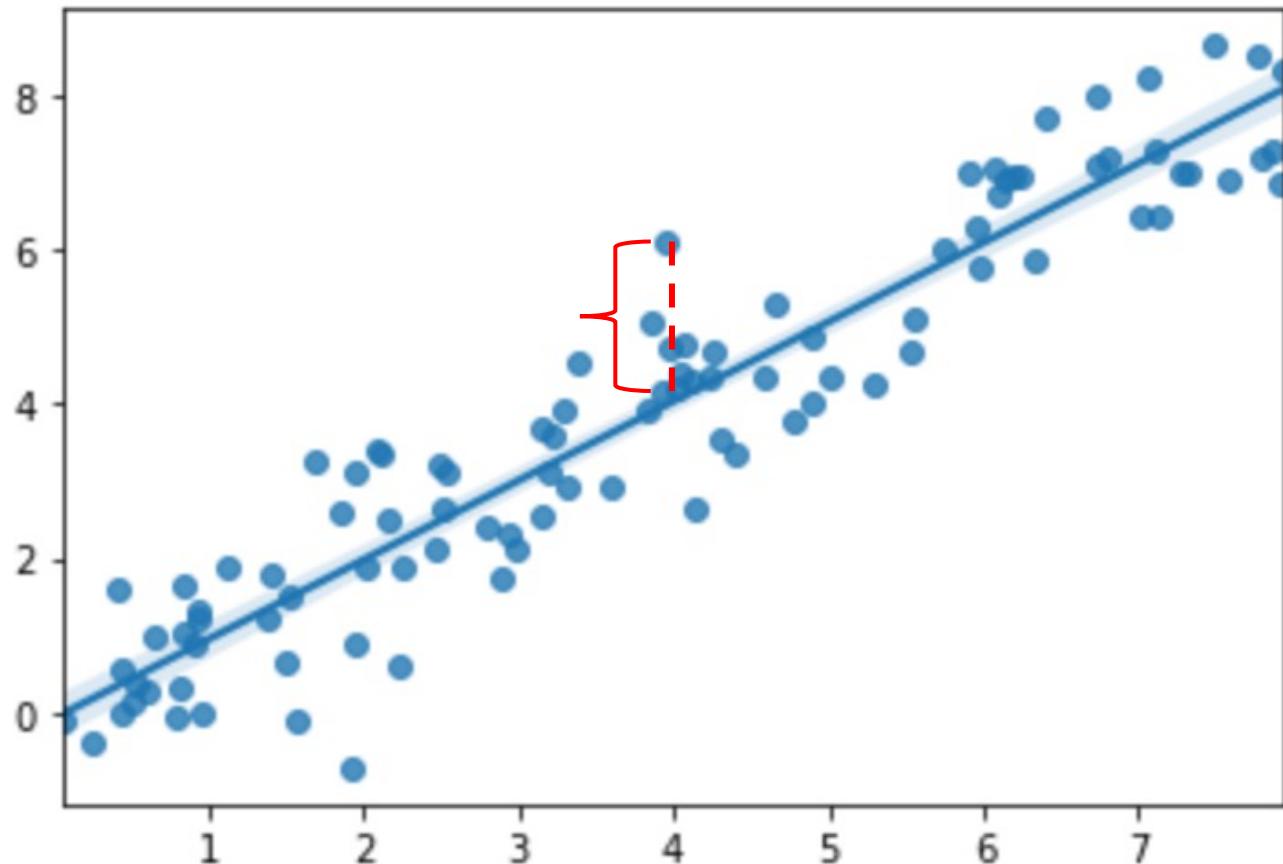


# Coeficiente de determinação

R2 – coeficiente de determinação

$$R^2 = 1 - \frac{SQM}{SQT}$$

$$= 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$$

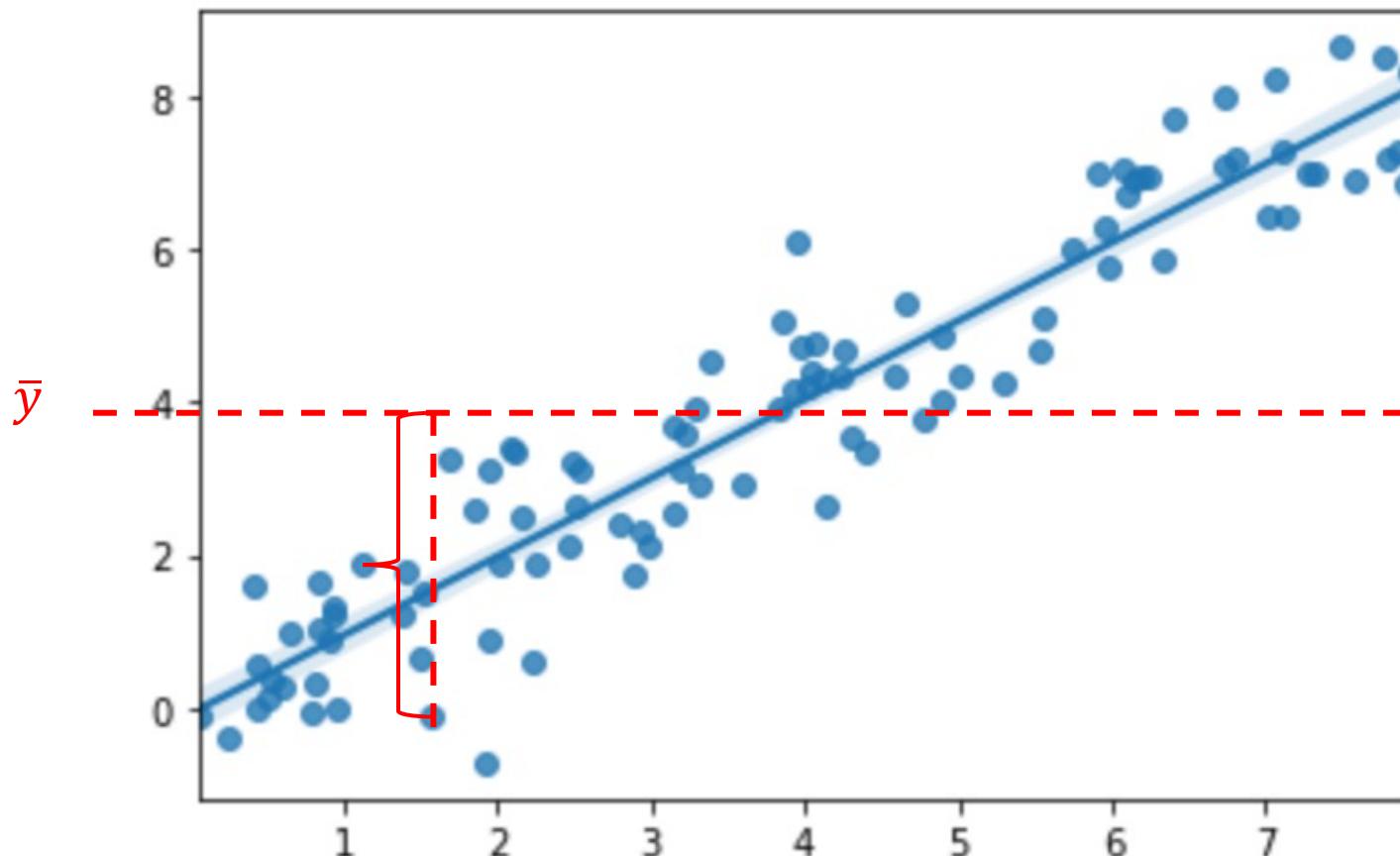


# Coeficiente de determinação

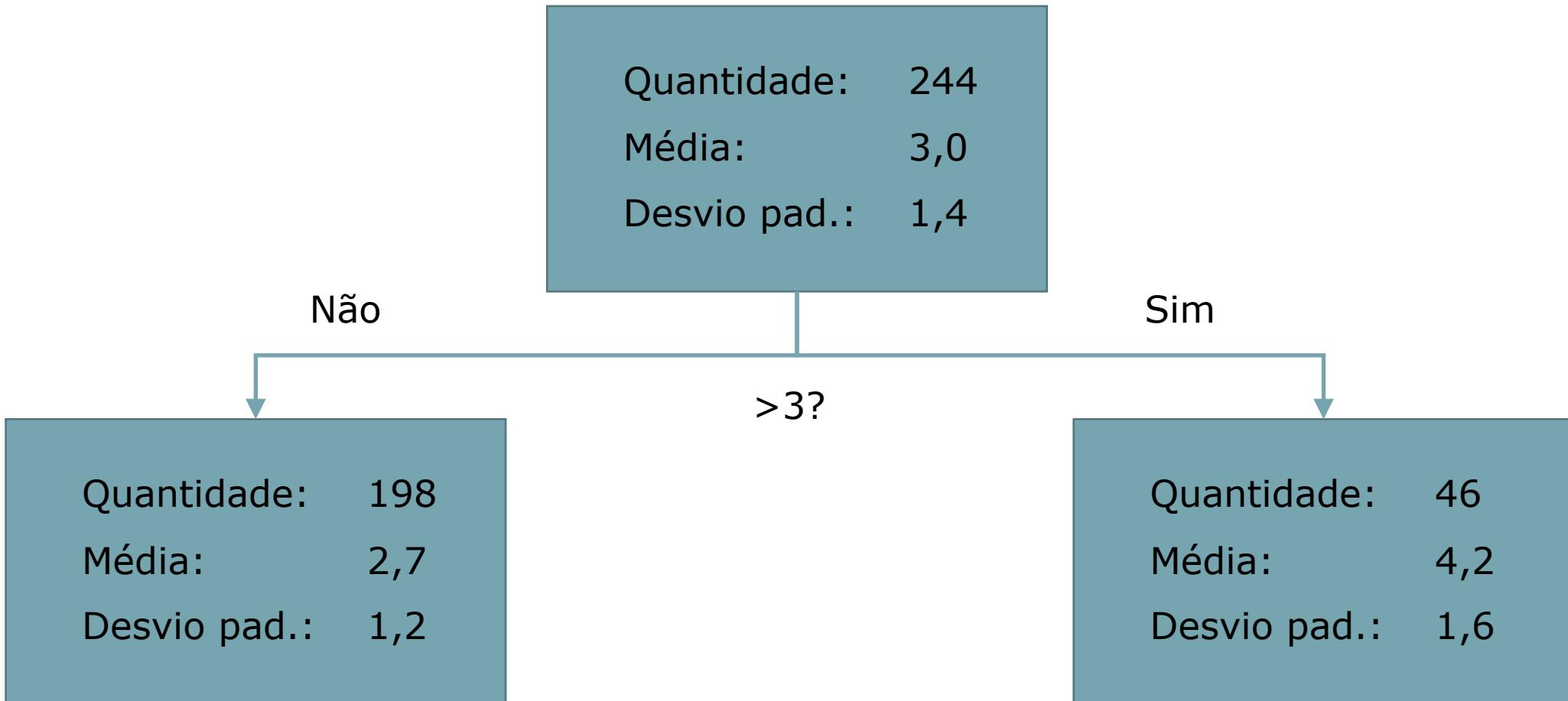
R2 – coeficiente de determinação

$$R^2 = 1 - \frac{SQM}{SQT}$$

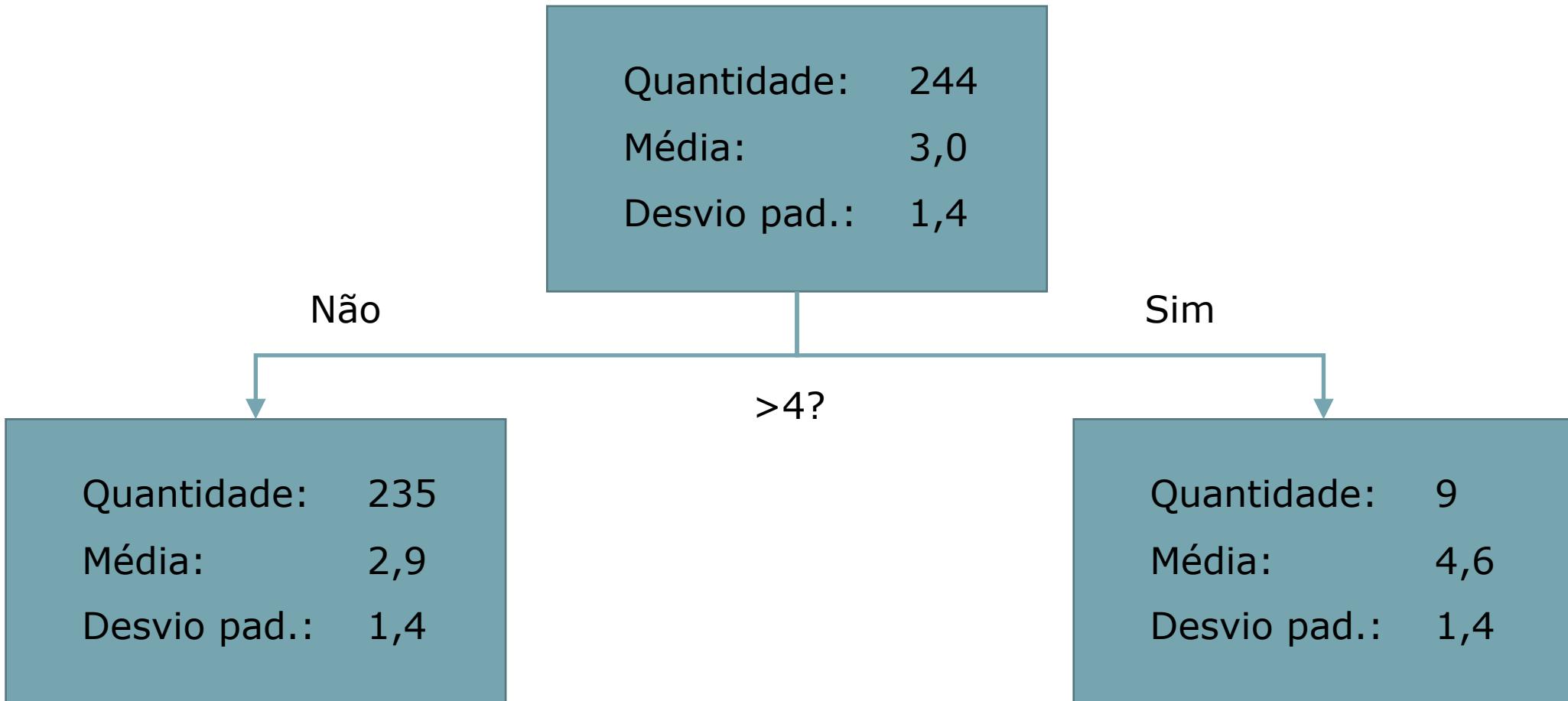
$$= 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$$



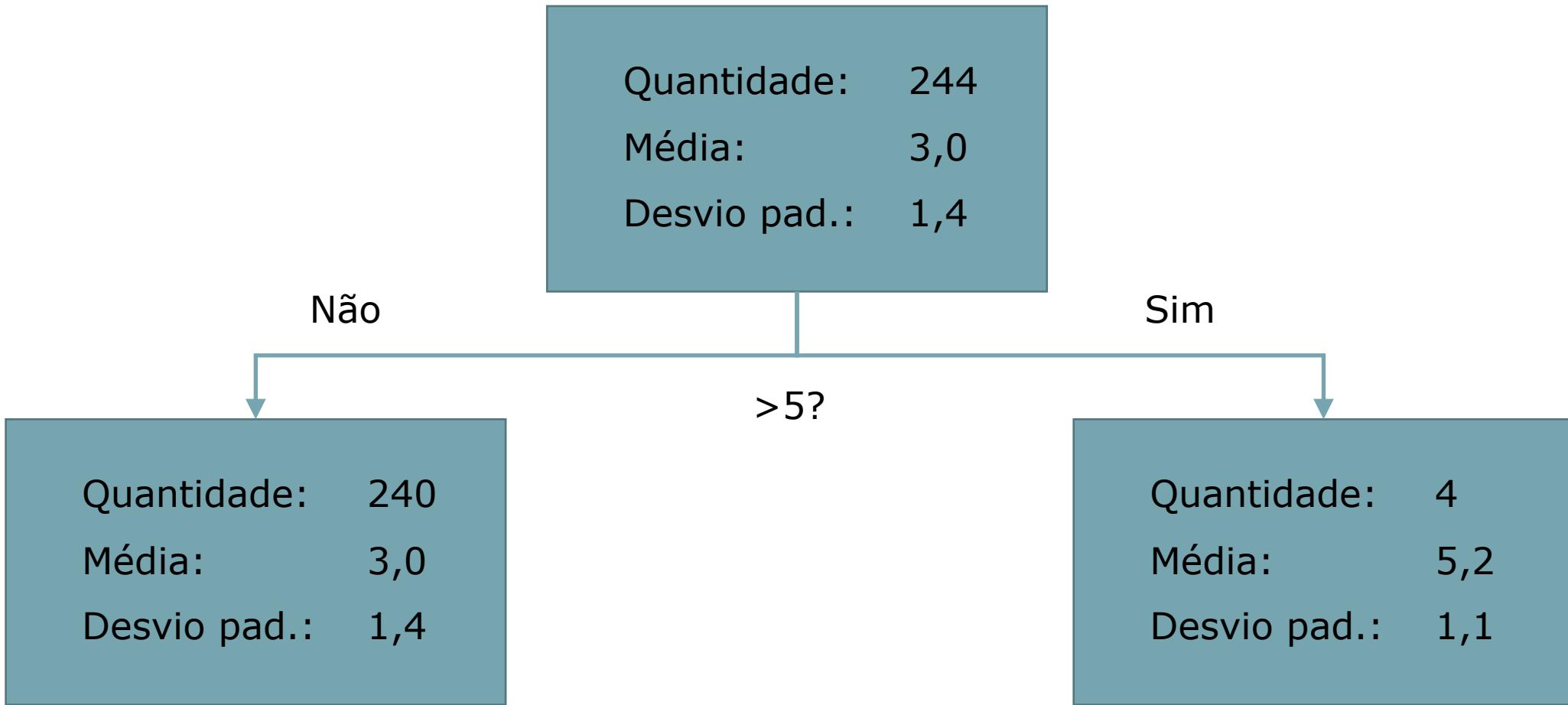
# Como maximizar a gorjeta?



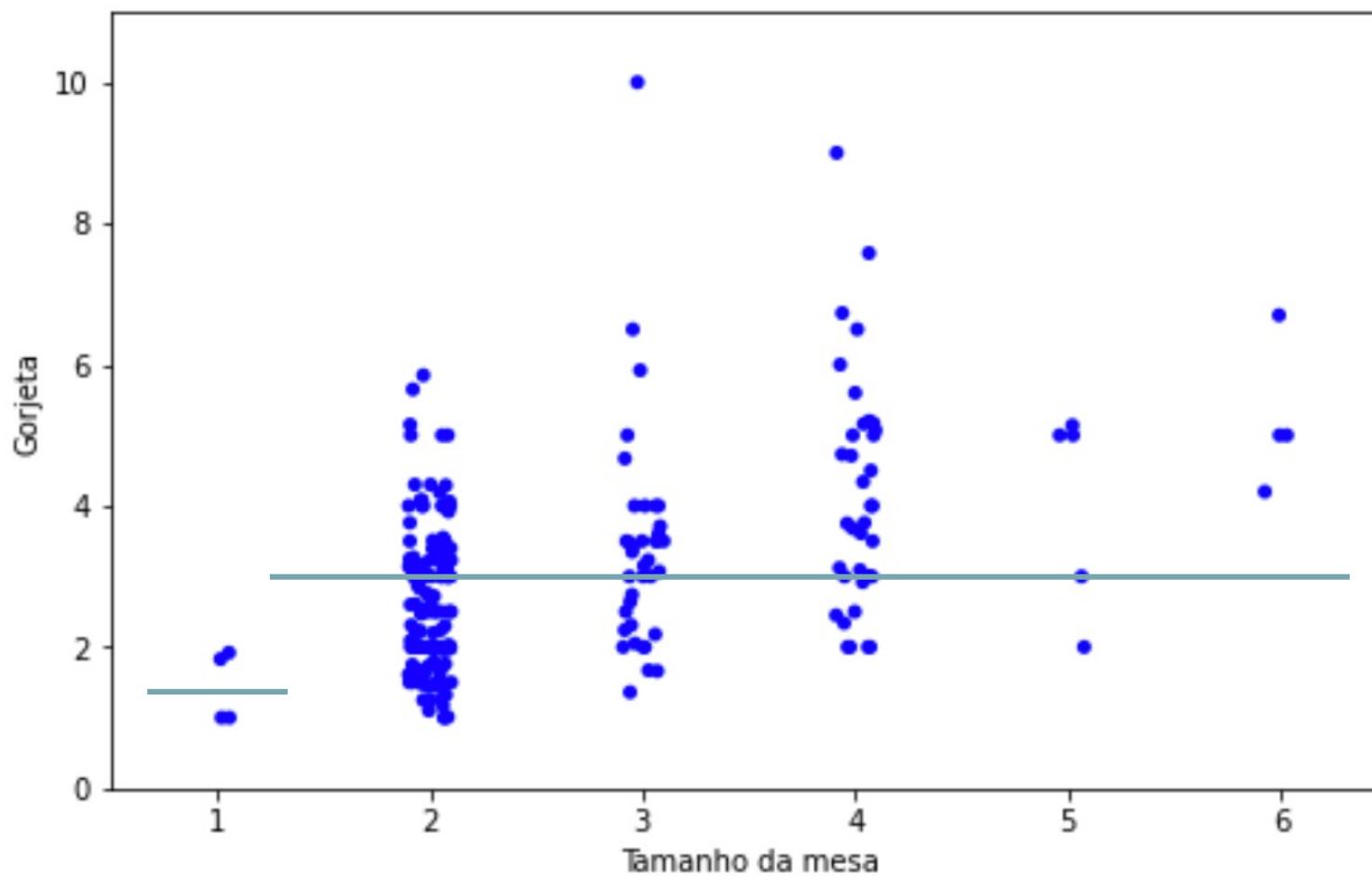
# Como maximizar a gorjeta?



# Como maximizar a gorjeta?

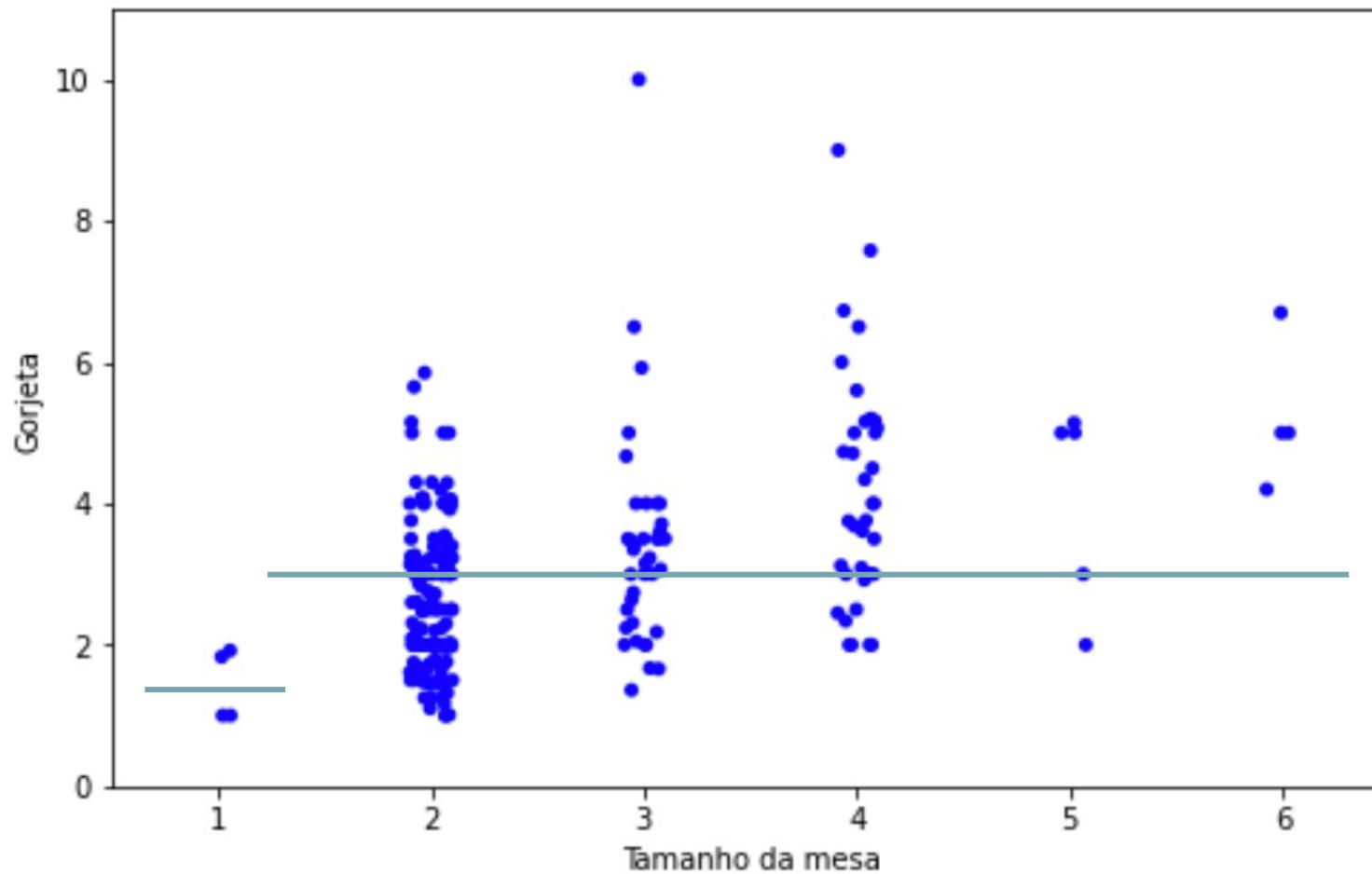


# Gorjeta por tamanho da mesa



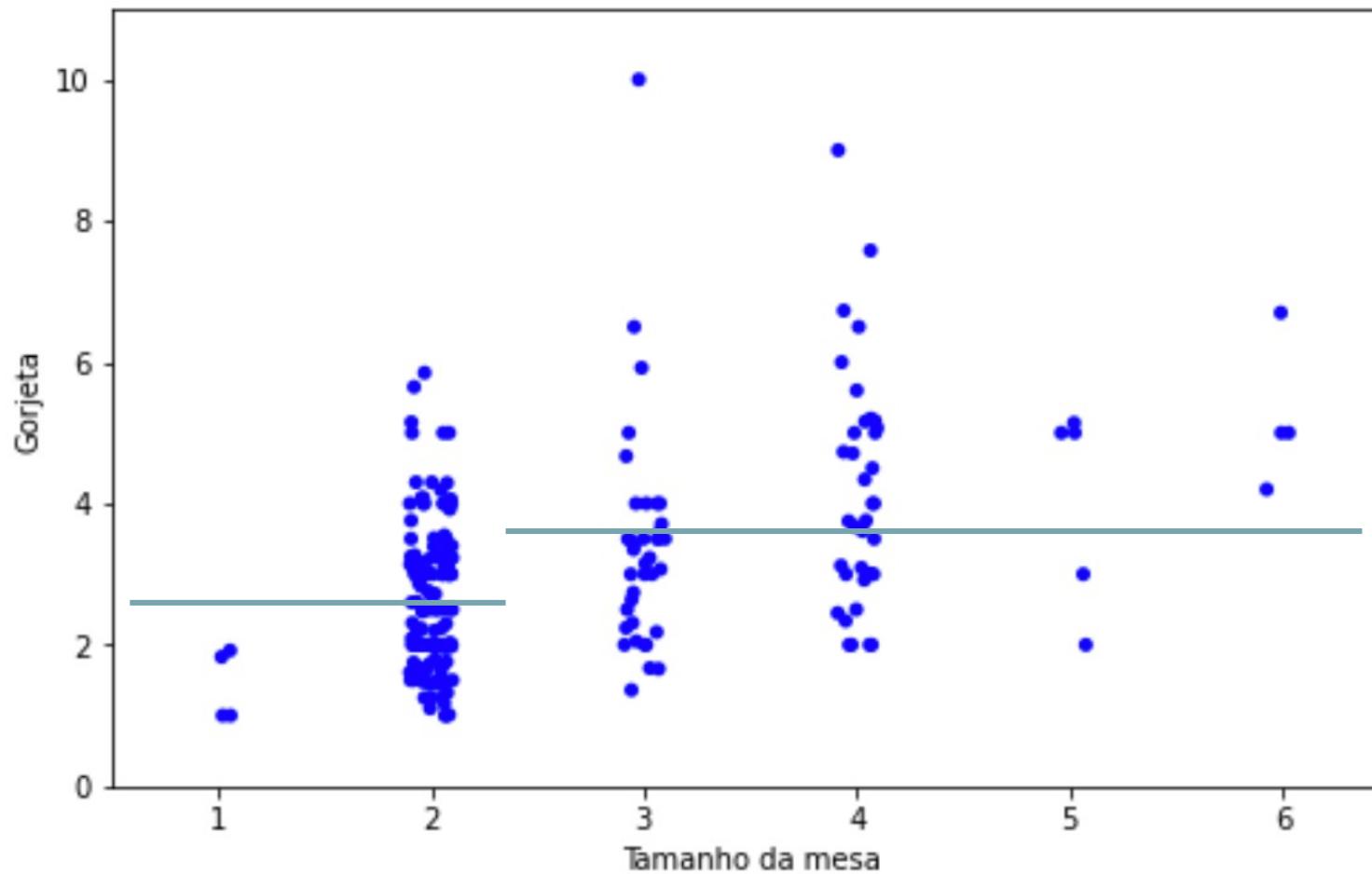
# Melhor quebra

# Gorjeta por tamanho da mesa



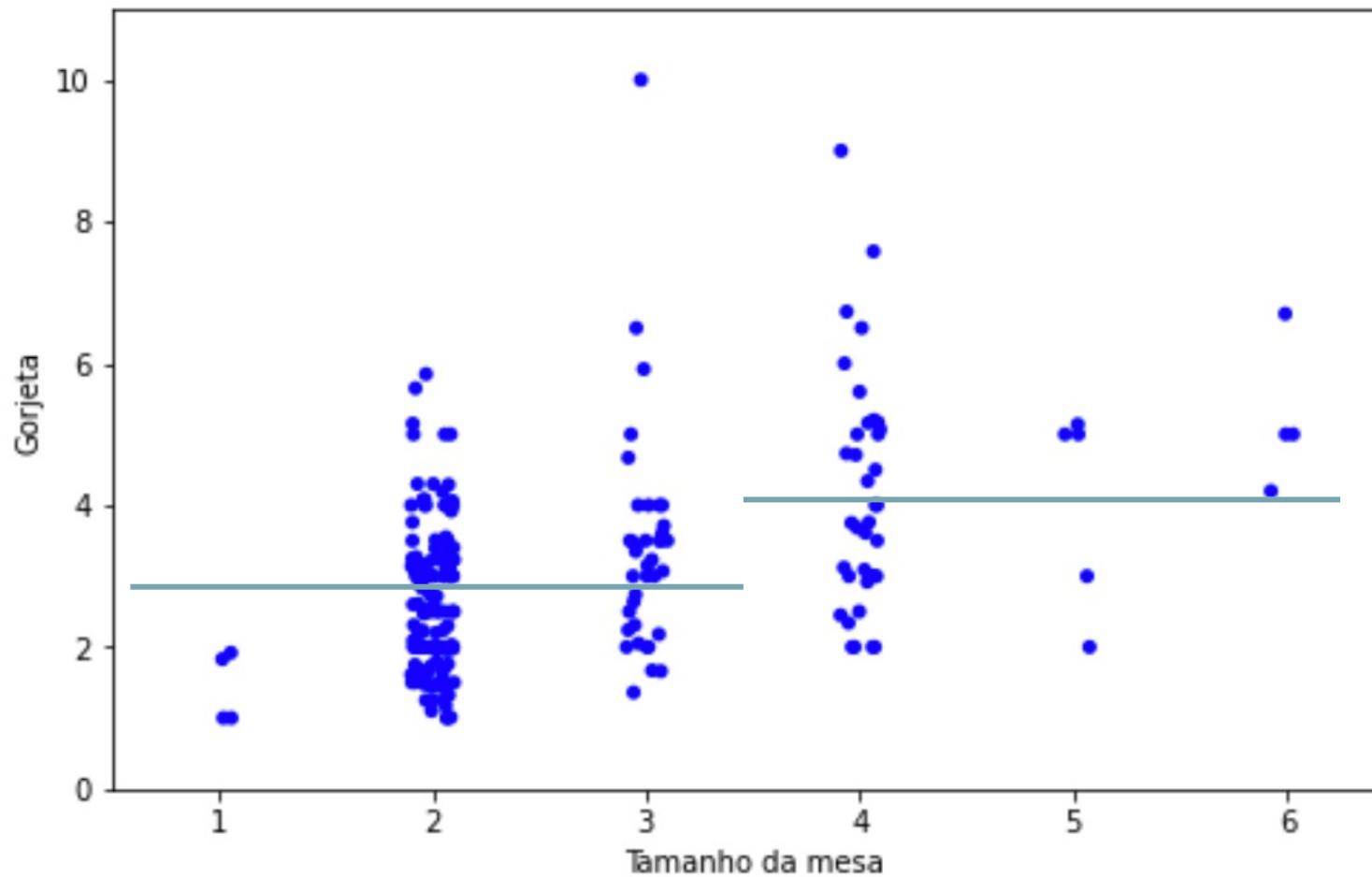
EQM = 1,87

# Gorjeta por tamanho da mesa



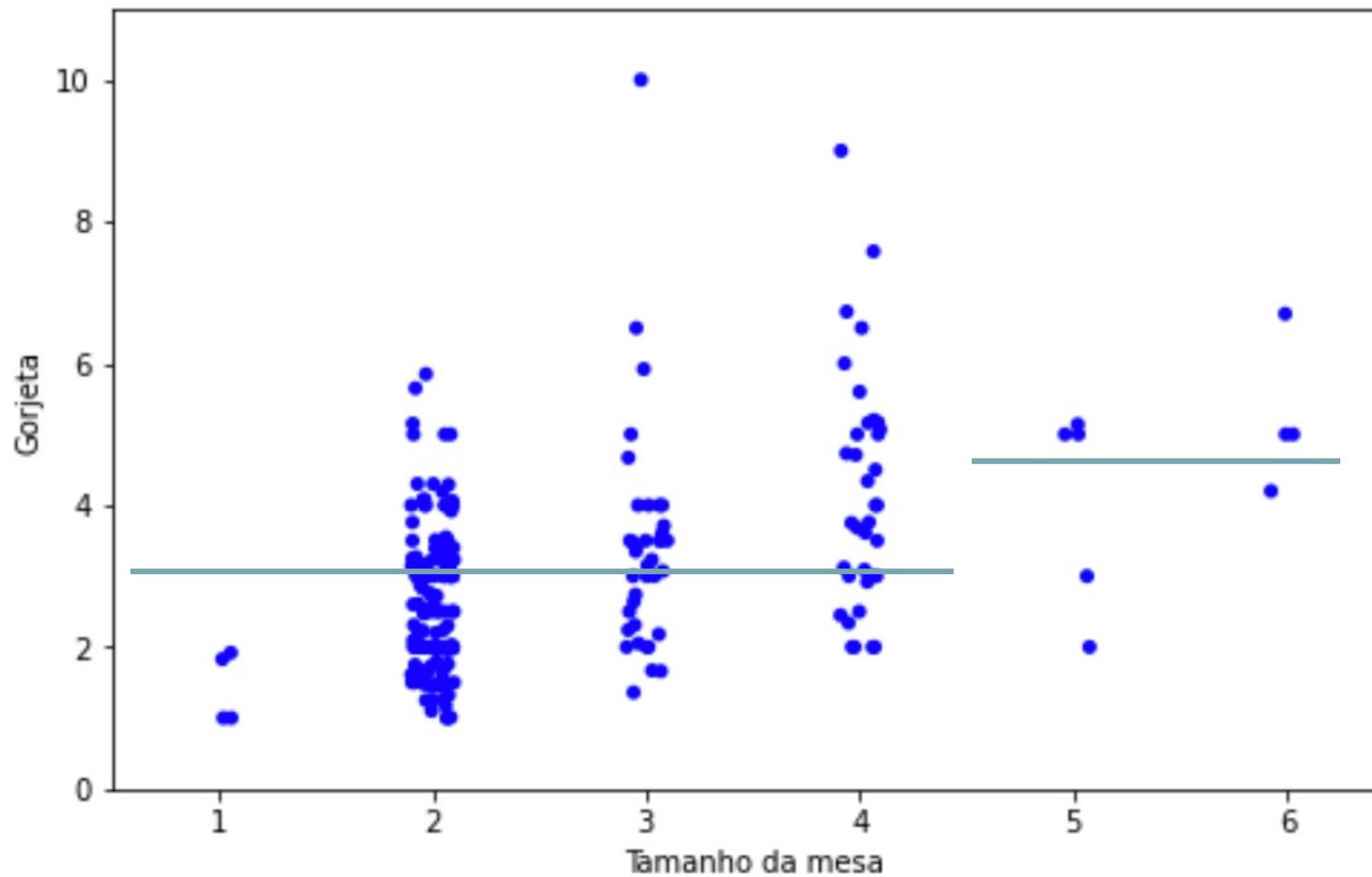
EQM = 1,53

# Gorjeta por tamanho da mesa



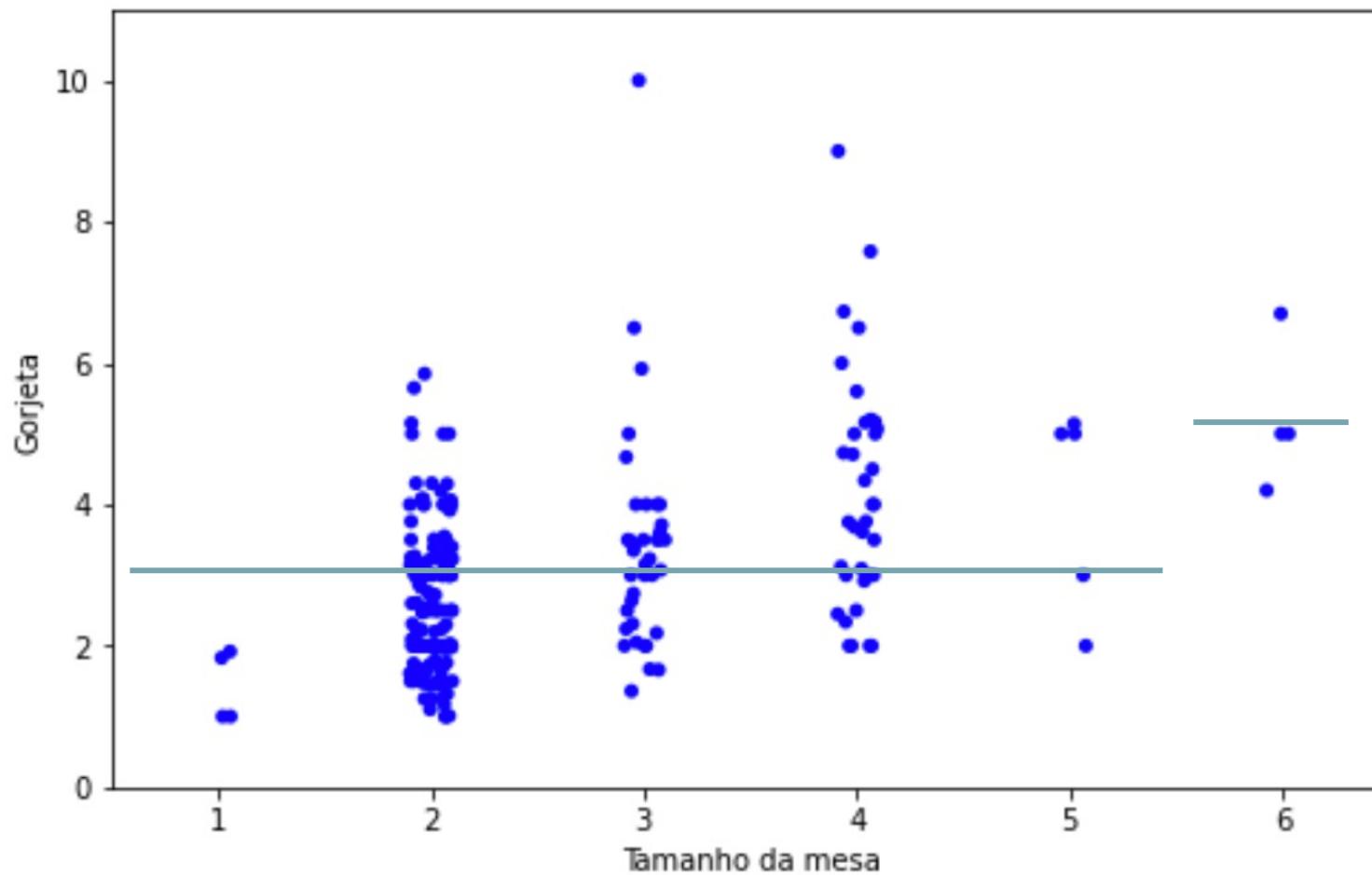
EQM = 1,56

# Gorjeta por tamanho da mesa



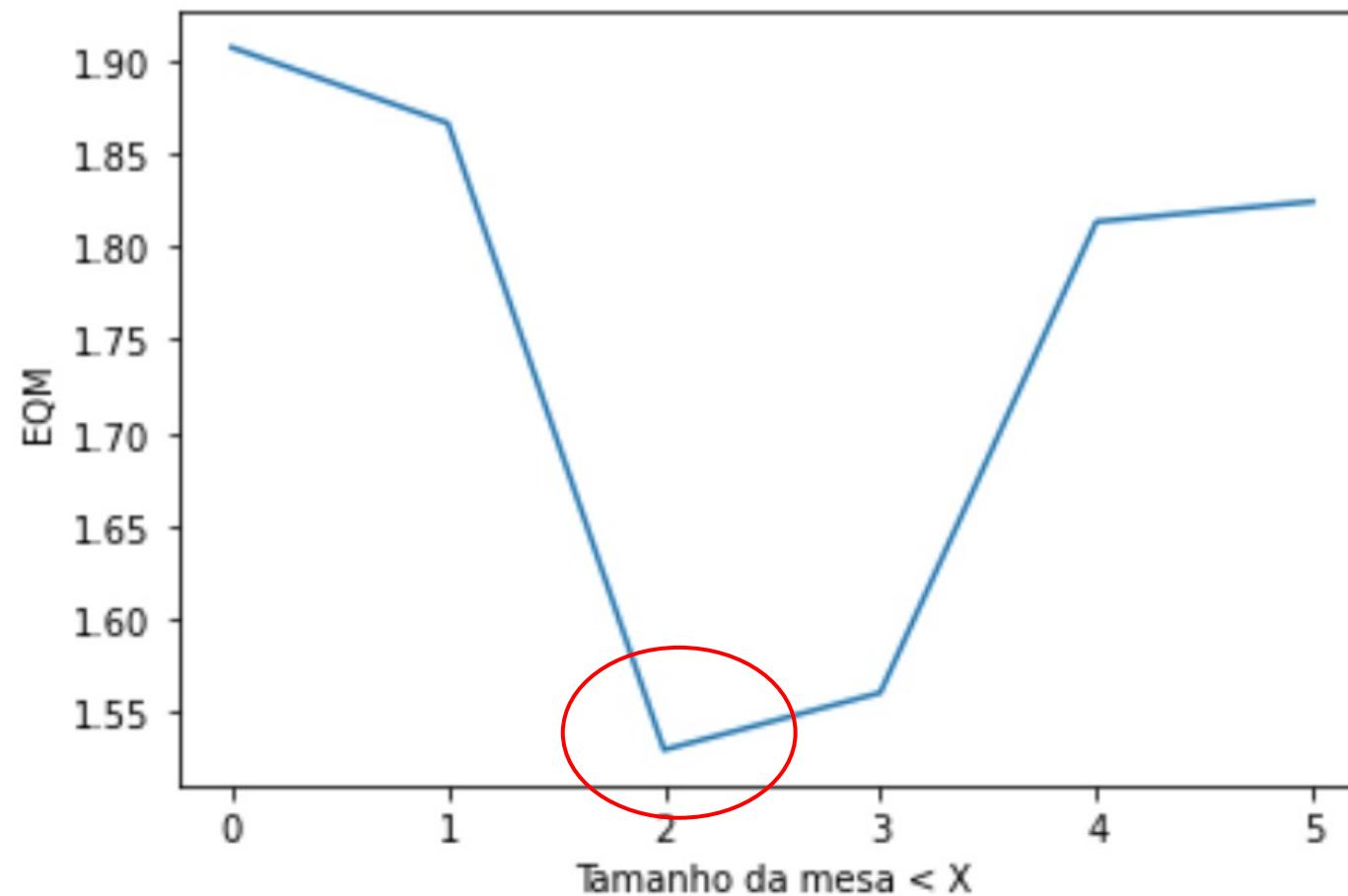
EQM = 1,81

# Gorjeta por tamanho da mesa



EQM = 1,82

# EQM por quebra de size



# Correlaç̄o e covariânc̄ia

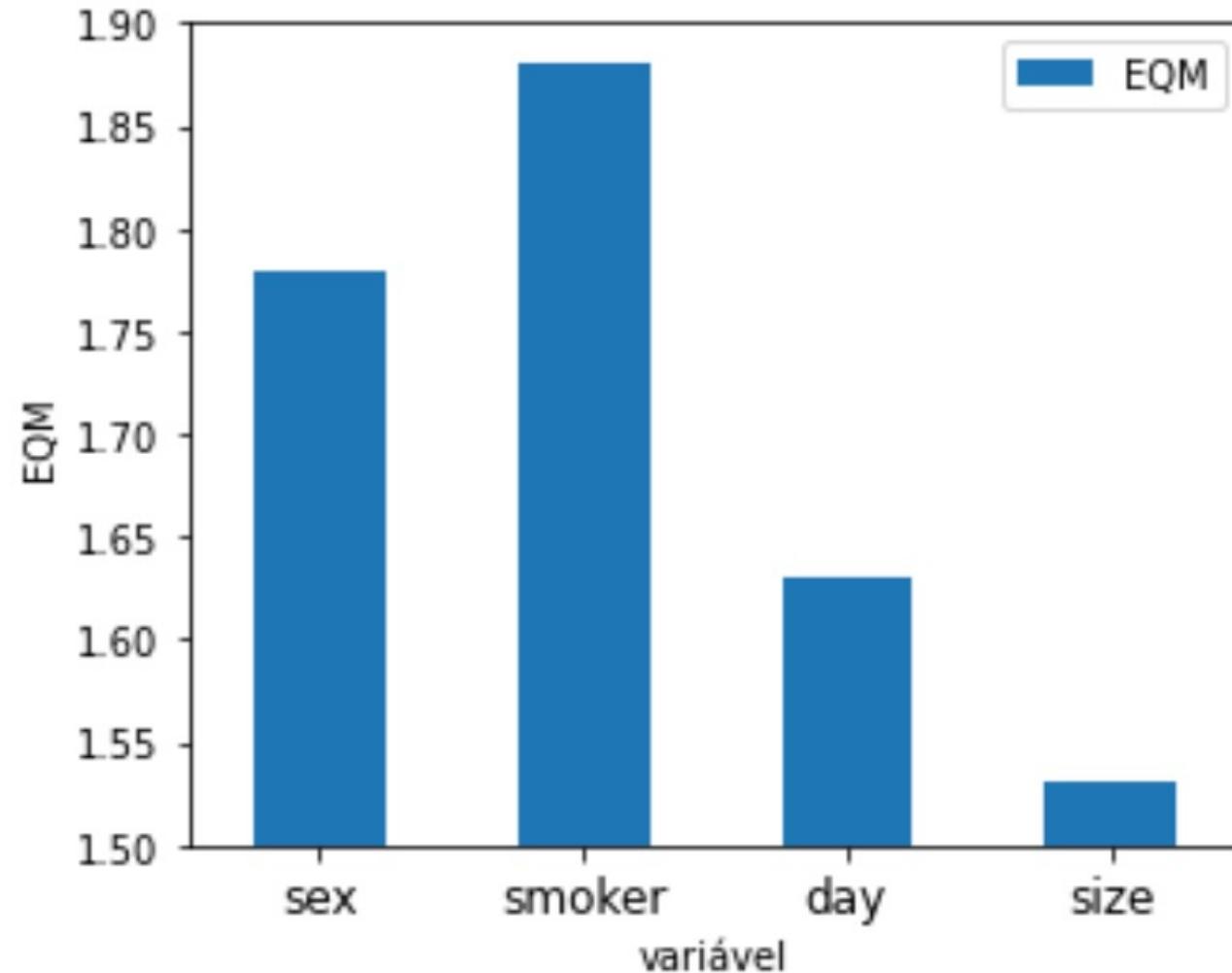
# EQM por variável

O algoritmo busca menor EQM em cada quebra.

Faz isso para cada variável.

A variável com menor EQM é selecionada.

O algoritmo continua até um critério de parada ser atingido.



# Variância

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

Variância: medida de dispersão dos dados em torno da média.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{N - 1}$$

Variância amostral: estimativa não viésada da variância com uma amostra.

# Desvio padrão

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

Desvio padrão: Tem a mesma unidade de medida da variável.

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

Desvio padrão amostral: Não viesado para a variância amostral.

# Covariância e correlação

$$\text{cov}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Covariância: Medida de associação entre duas variáveis.

Domínio:  $[-\infty, +\infty]$

Unidade de mensuração:  $U(x).U(y)$

$$\text{cor}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}}$$

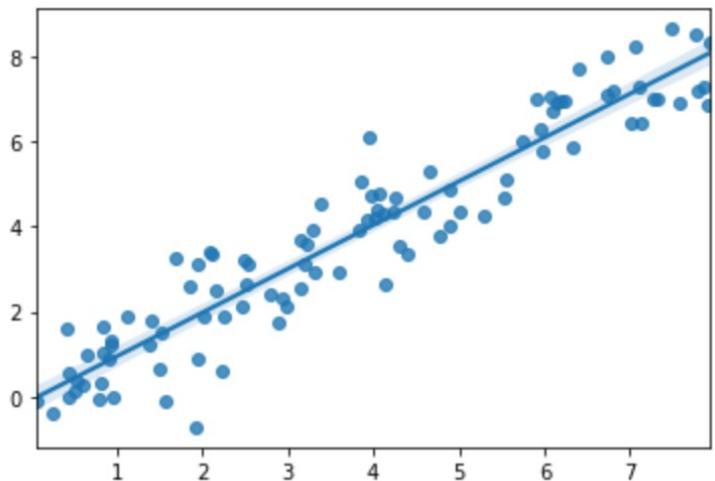
Desvio padrão amostral: Não viesado para a variância amostral.

Domínio:  $[-1, +1]$

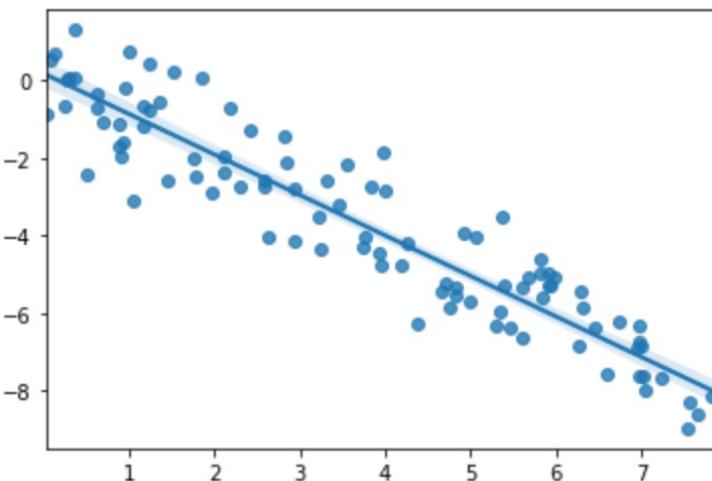
Unidade de mensuração: sem unidade

$$\text{cor}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

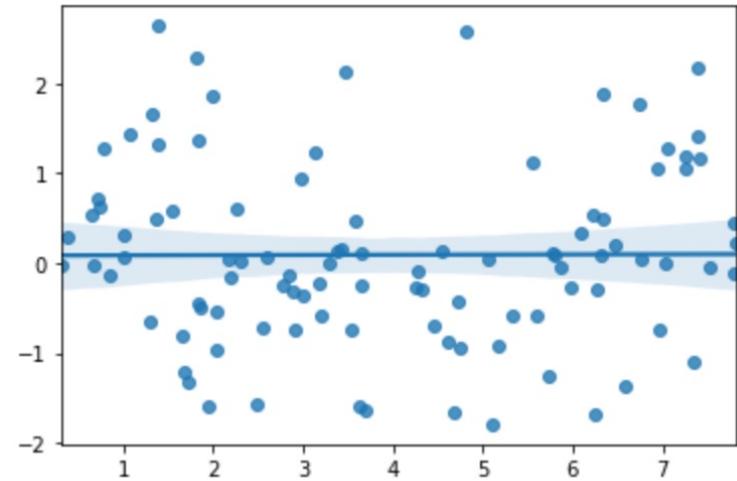
# Correlação



Correlação positiva

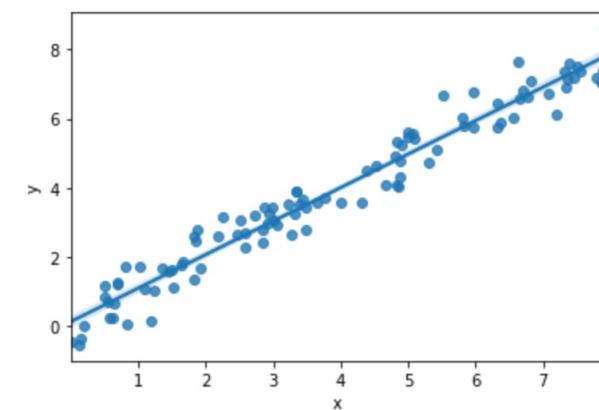
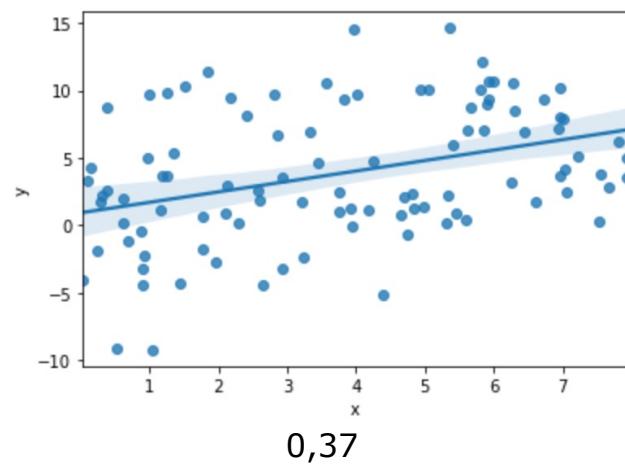


Correlação negativa

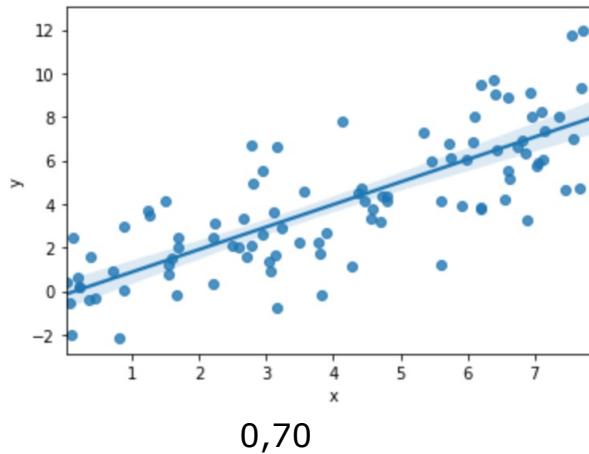


Correlação muito baixa

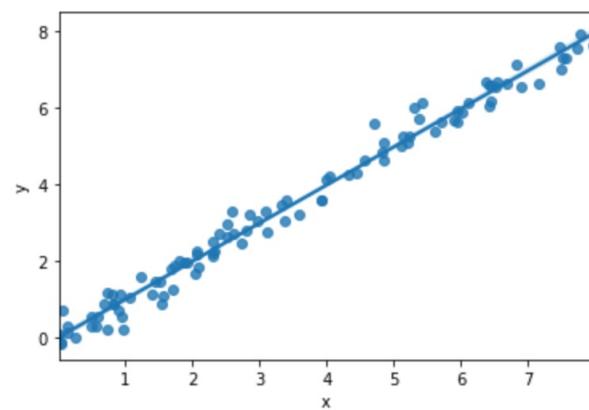
# Correlação



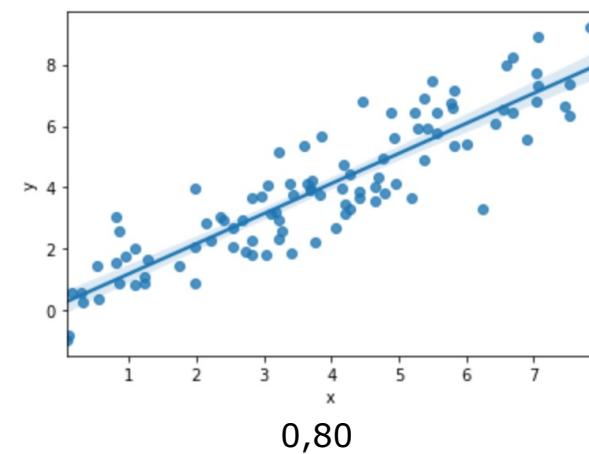
0,97



0,70



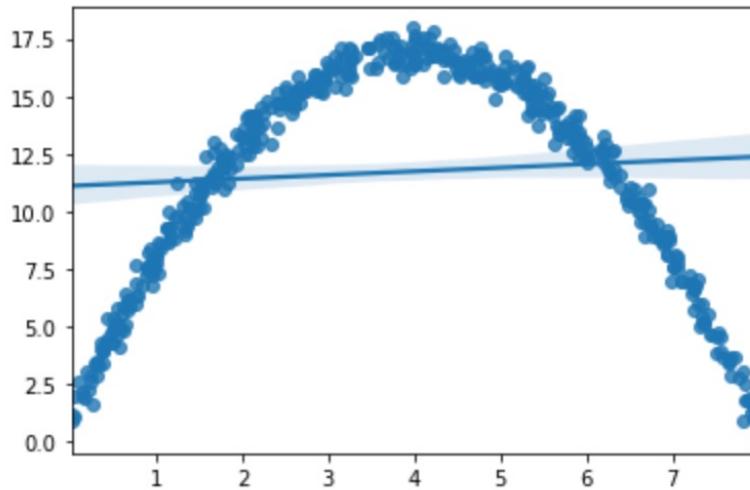
0,99



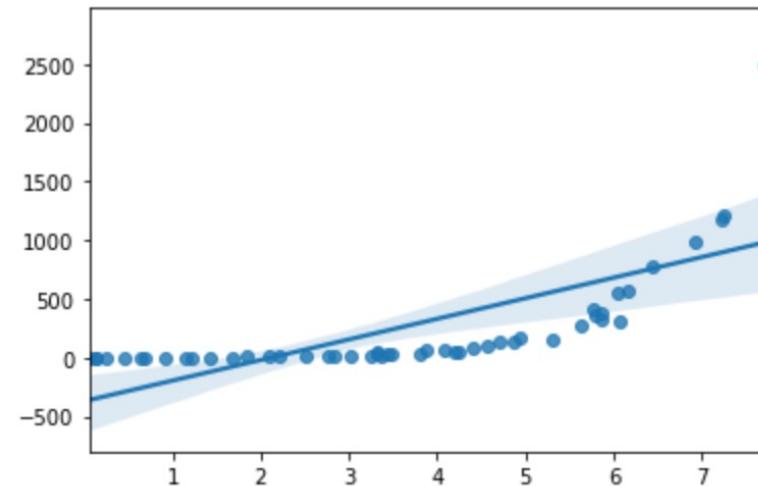
1,00

# Correlação

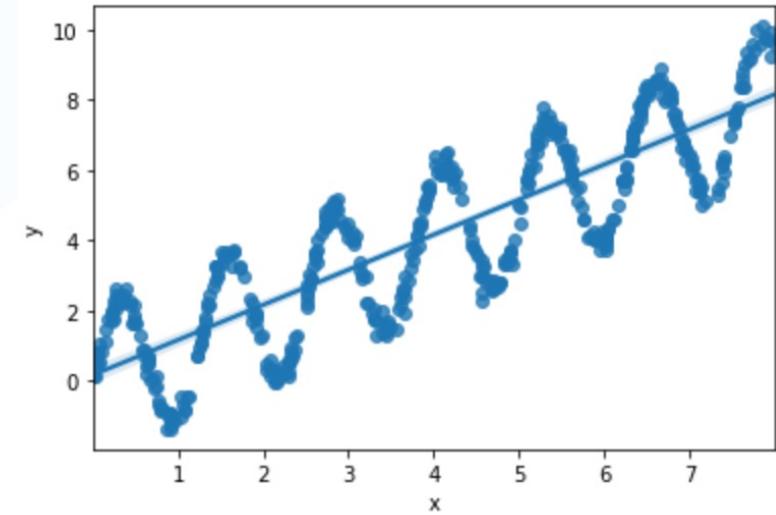
## Medida de associação linear



Correlação muito baixa



Correlação positiva  
com não linearidade



Correlação linear não  
captura toda a associação

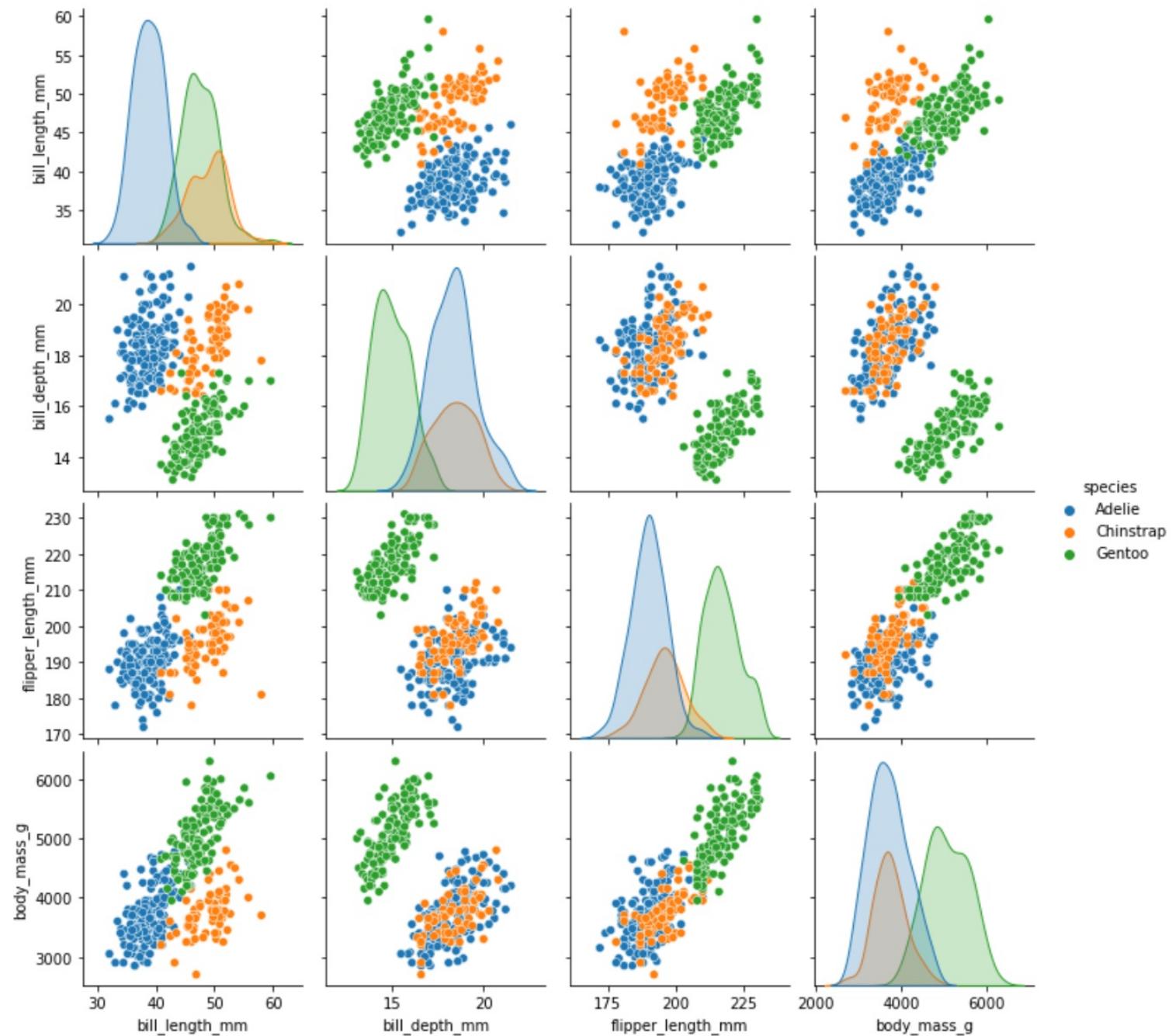
# Matriz de correlação

# Matriz de variância e covariância

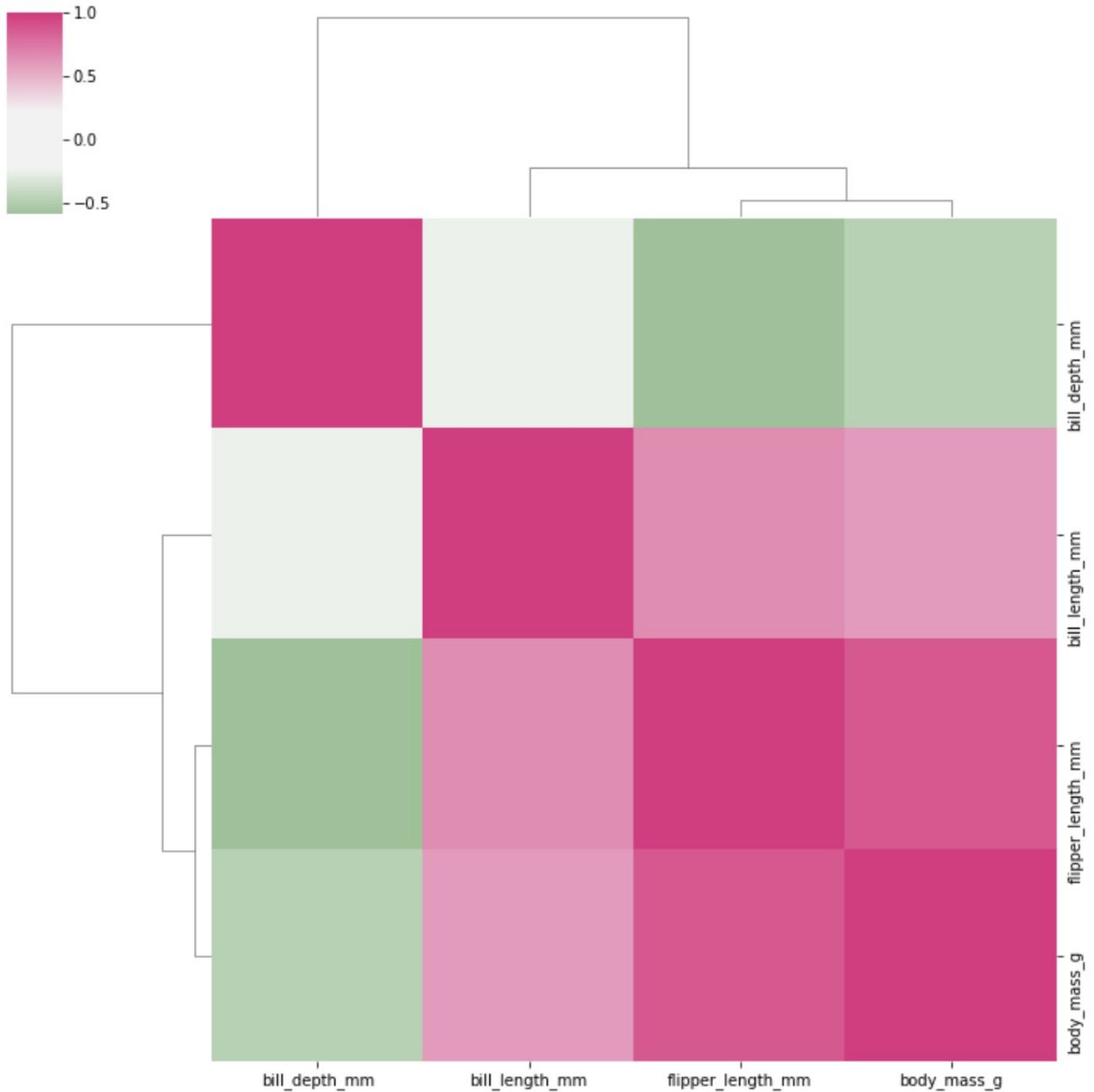
$$\mathbf{S}^2 = \begin{bmatrix} S_1^2 & S_{1,2} & \dots & S_{1,p} \\ S_{2,1} & S_2^2 & & S_{2,p} \\ \vdots & \ddots & \ddots & \vdots \\ S_{p,1} & S_{p,2} & \dots & S_p^2 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} 1 & r_{1,2} & \dots & r_{1,p} \\ r_{2,1} & 1 & & r_{2,p} \\ \vdots & \ddots & \ddots & \vdots \\ r_{p,1} & r_{p,2} & \dots & 1 \end{bmatrix}$$

# Matriz de dispersão

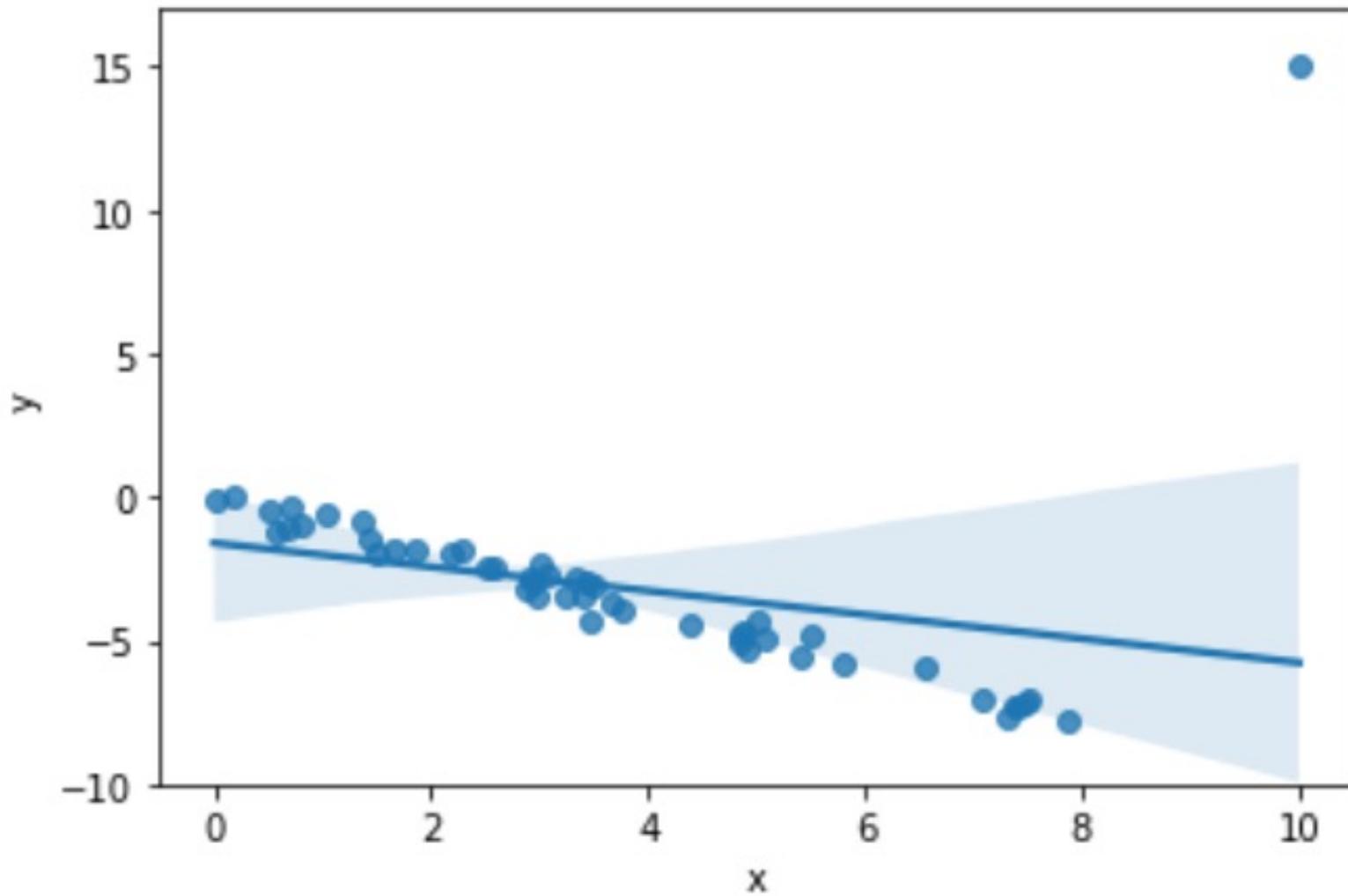


# Clustermapper

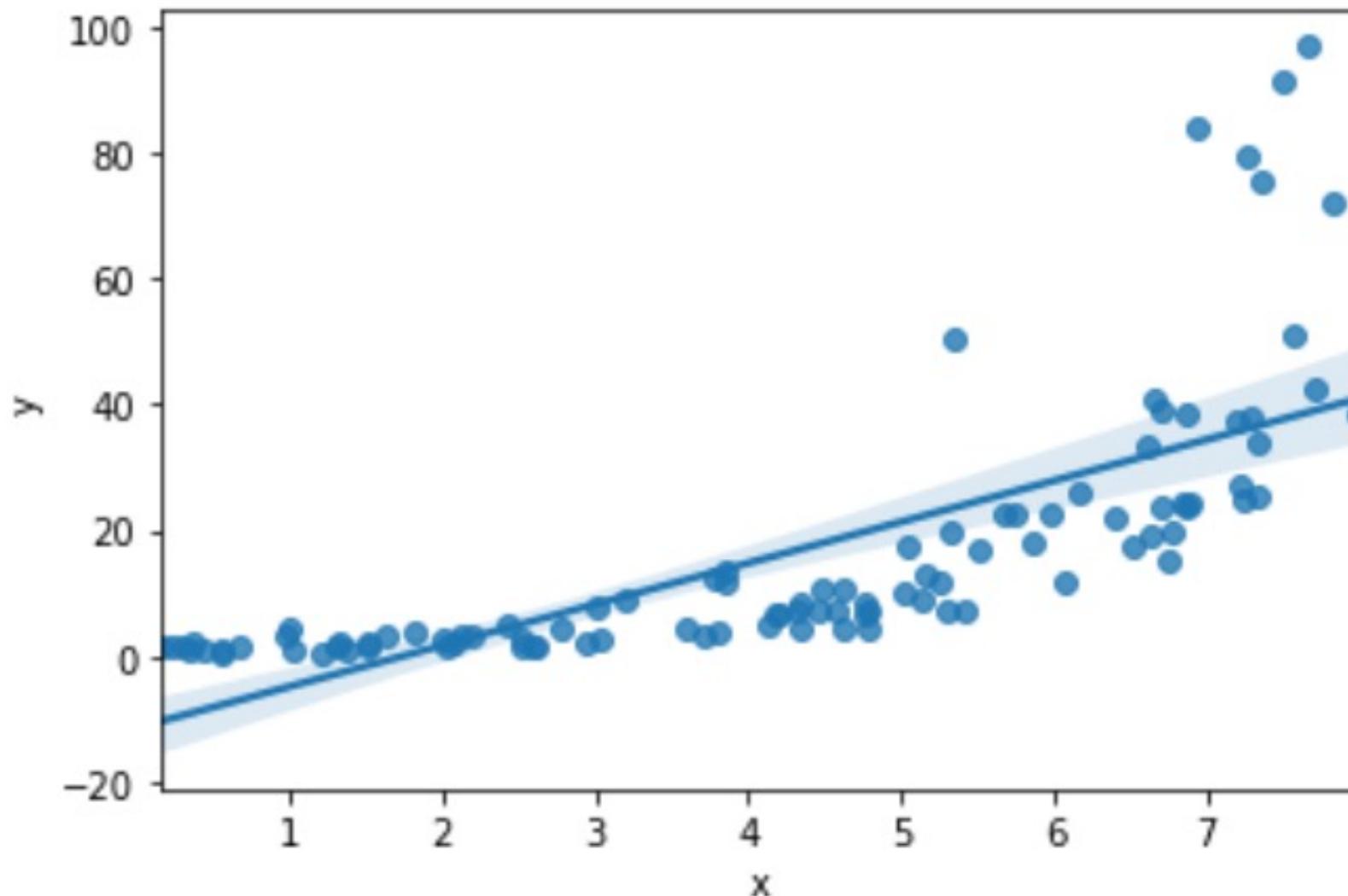


# Outliers

# Outliers



# Outliers



# Ranking (posto)

Índice	Valor	Rank
0	12.1	5
1	4	2
2	7	4
3	3	1
4	5	3

In [139]:

```
1 df4_rank = pd.concat([df4, df4.rank()], axis = 1)
2
3 df4_rank.columns = ['x', 'y', 'x_rank', 'y_rank']
4 df4_rank
```

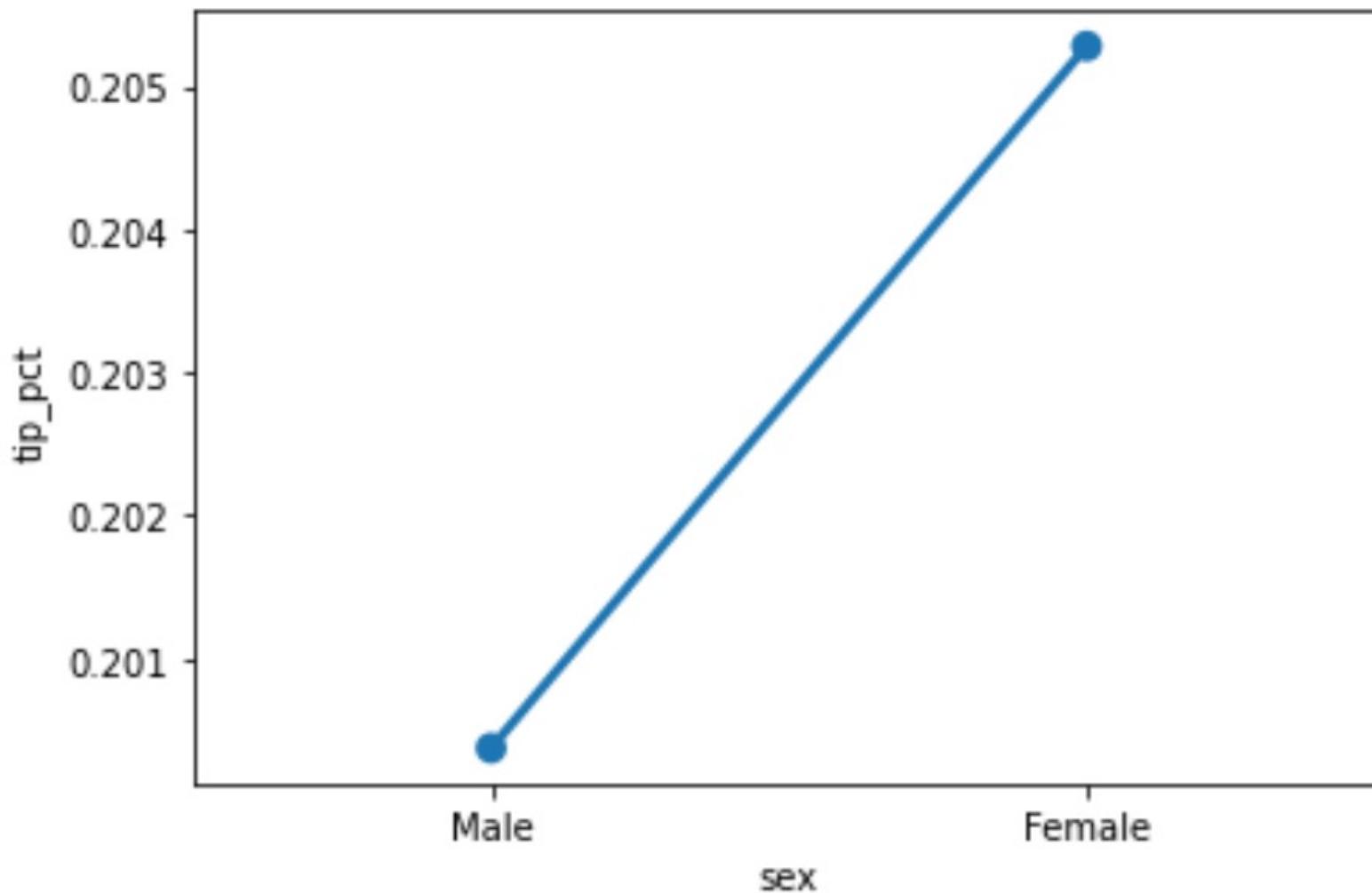
Out [139]:

	x	y	x_rank	y_rank
0	0.409429	1.079676	7.0	3.0
1	3.975723	40.489423	49.0	42.0
2	1.528080	5.608276	12.0	14.0
3	5.702499	461.775559	69.0	73.0
4	2.975968	17.708827	31.0	31.0
...	...	...	...	...

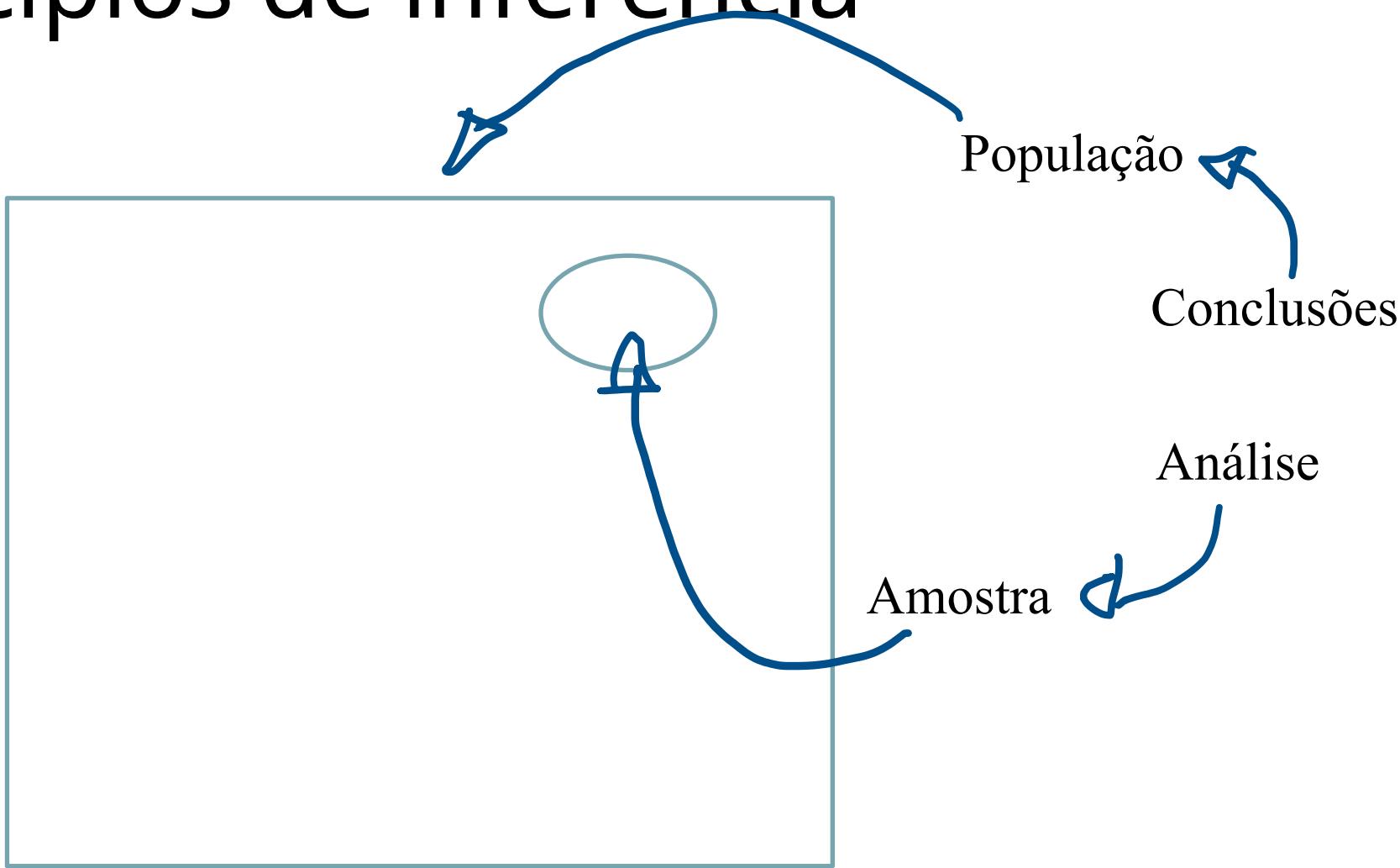
# Explicativas qualitativas

Comparando médias

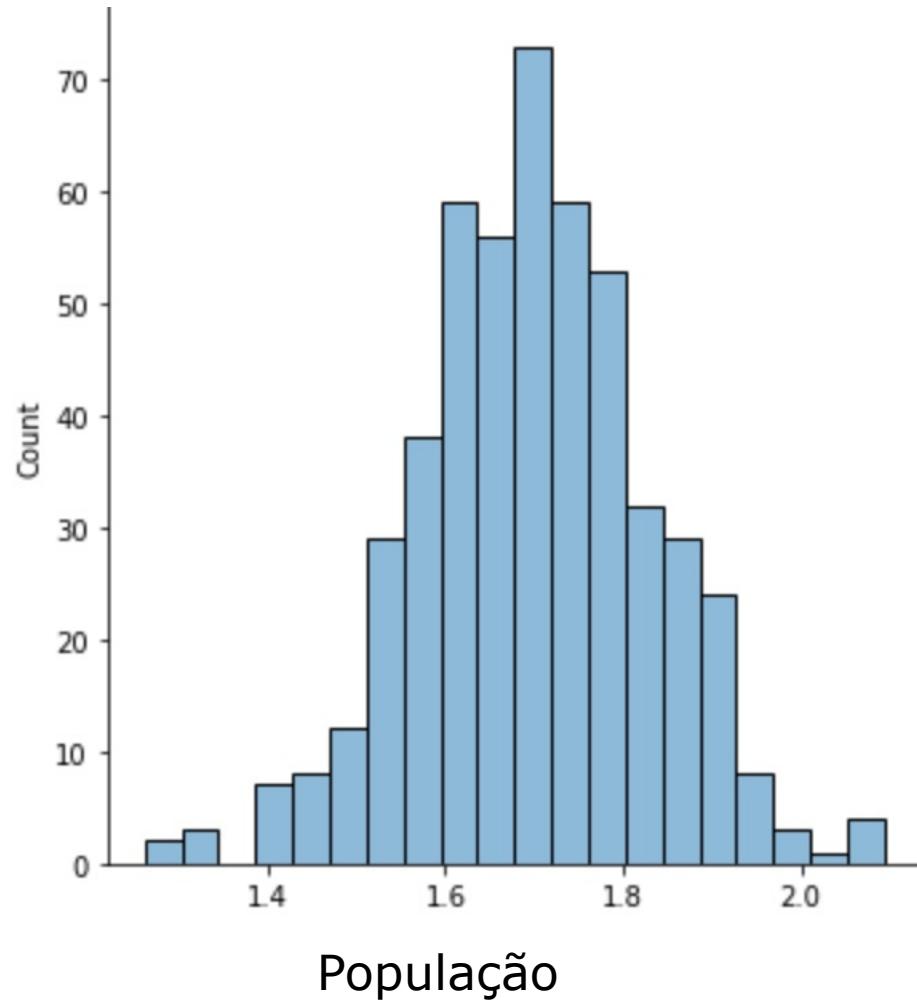
# Comparação de médias



# Princípios de inferência



# Princípios de inferência



Coletamos 10 pessoas aleatoriamente:

Média amostral = 1,62

Média amostral 2 = 1,78

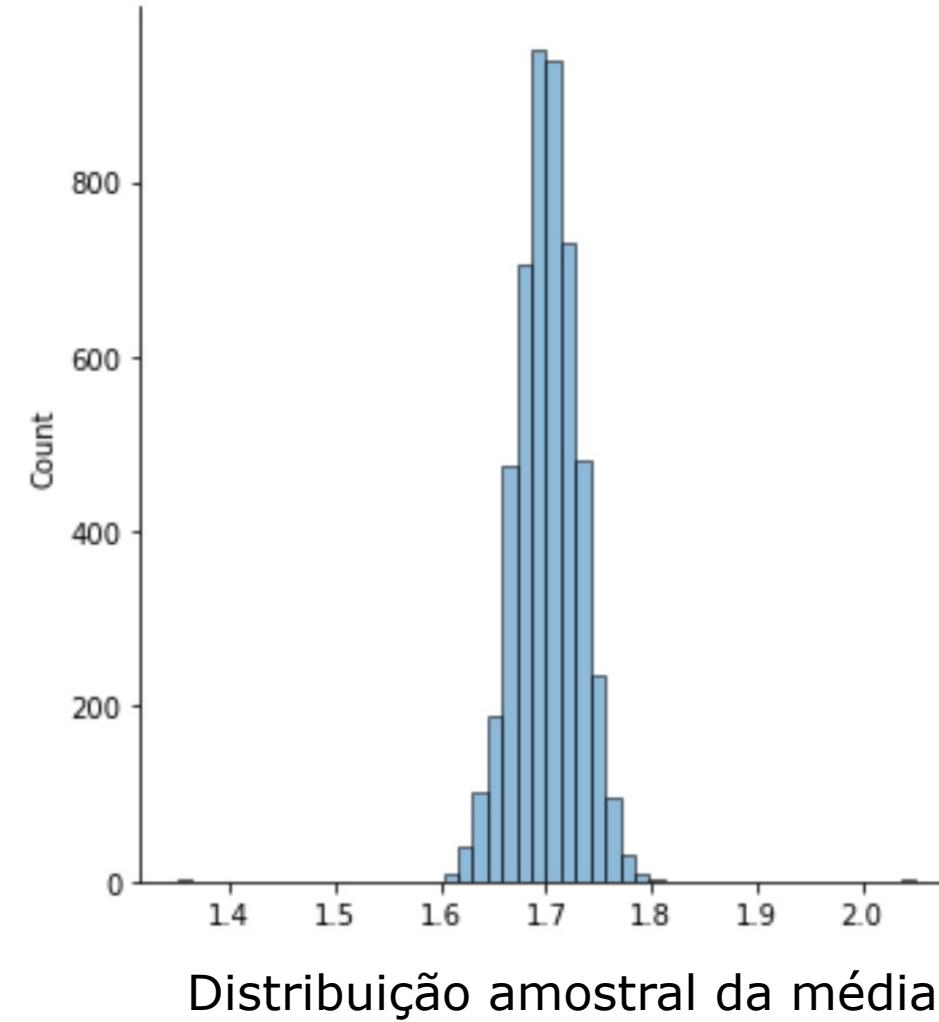
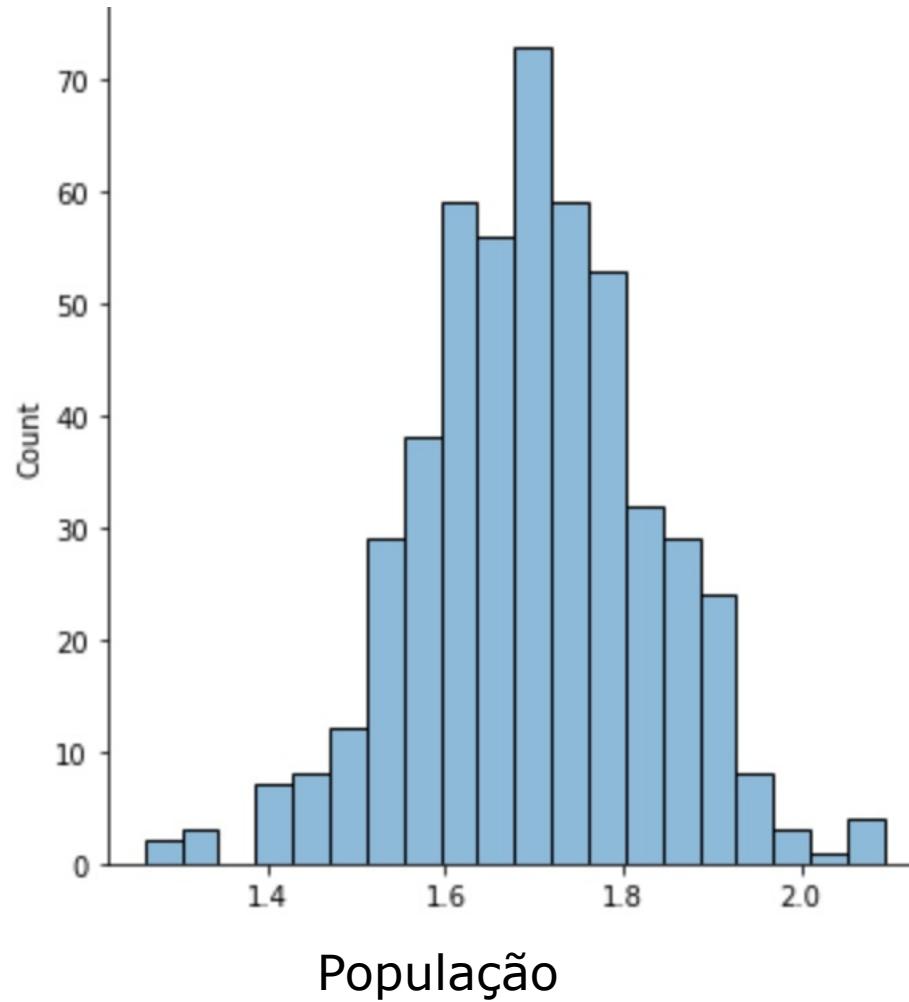
.

.

.

Média amostral 20 = 1,75

# Princípios de inferência



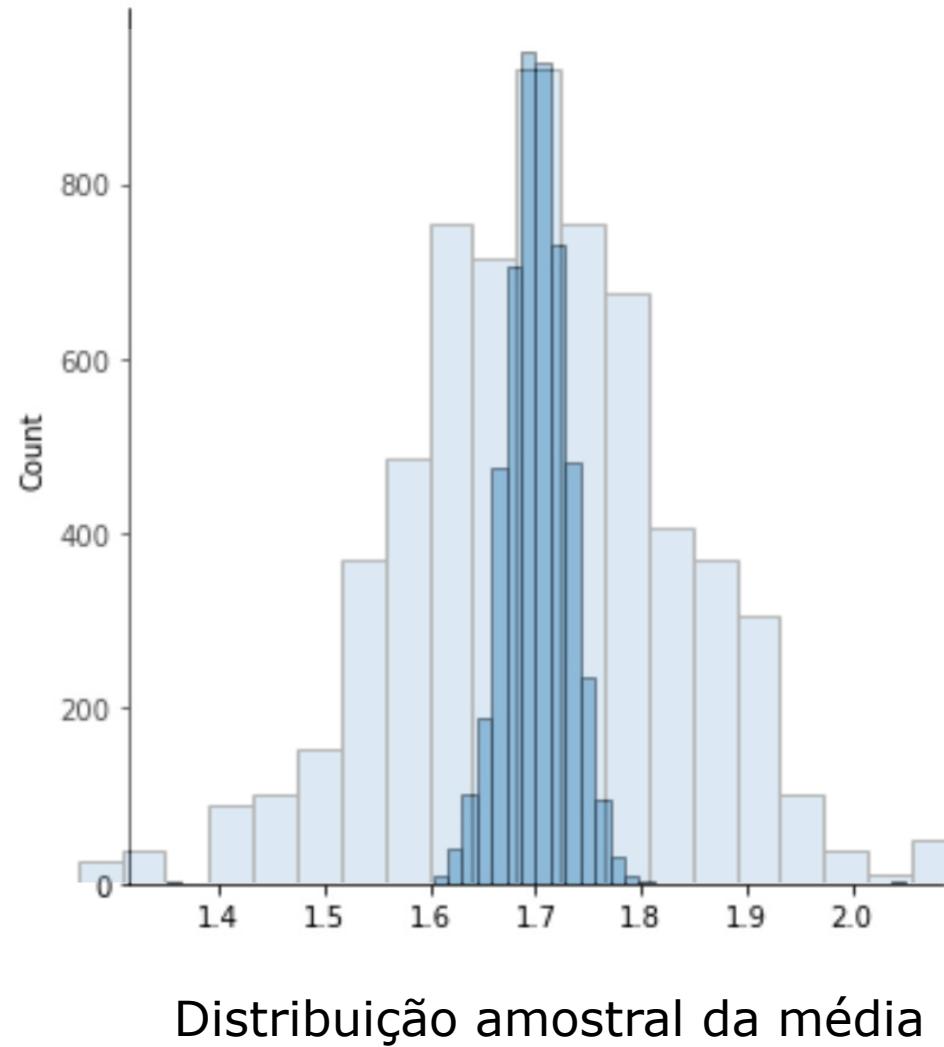
# Erro padrão

O erro padrão é uma estimativa do desvio padrão do parâmetro de interesse.

Estimador consistente: O erro padrão diminui conforme aumentamos a amostra

No caso, o desvio padrão da média amostral.

$$\text{erro padrão} = \sigma/\sqrt{n}$$



# Erro padrão

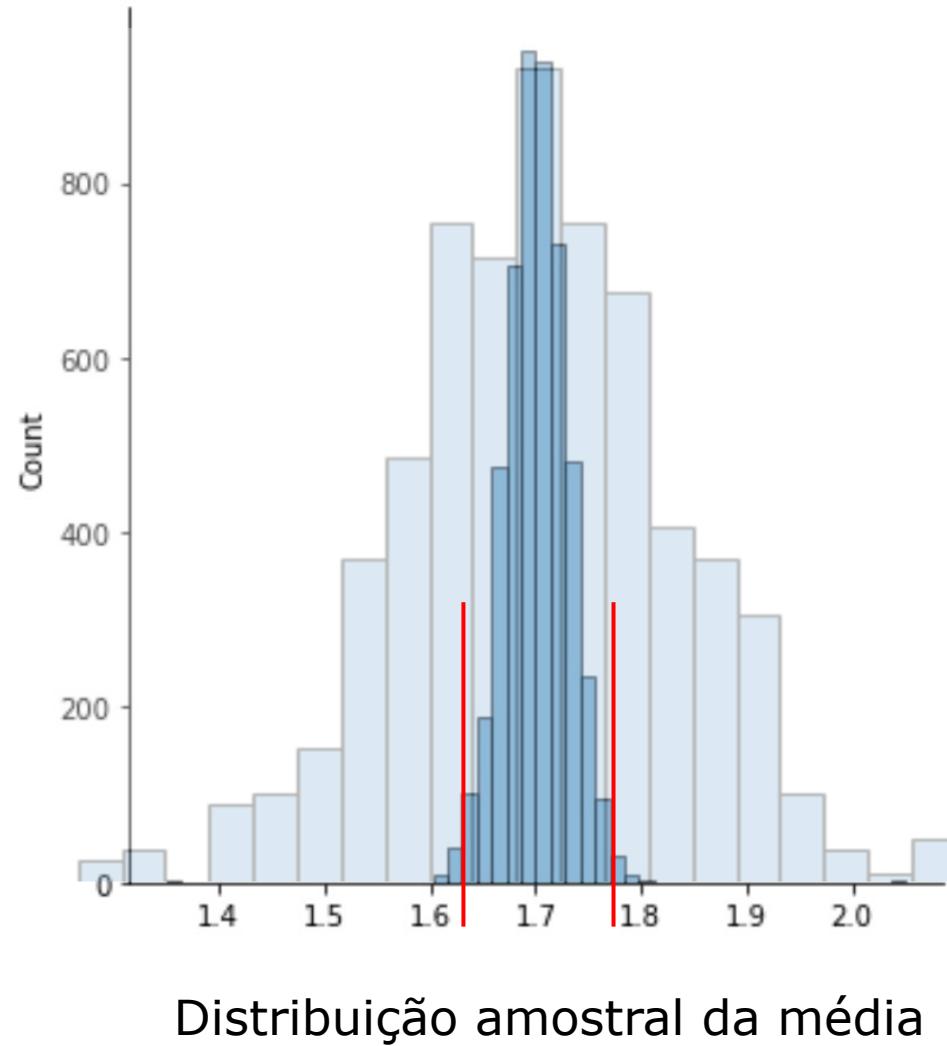
O erro padrão é uma estimativa do desvio padrão do parâmetro de interesse.

Estimador consistente: O erro padrão diminui conforme aumentamos a amostra

No caso, o desvio padrão da média amostral.

$$\text{erro padrão} = \sigma/\sqrt{n}$$

$$IC95\% \sim [\bar{x} \pm 2 * ep]$$



# Gráfico de perfis de médias

# Comparação de médias

