

# Análise Exploratória de Dados: Estatística Descritiva

---



# Tópicos da Aula

01

Fundamentos da EDA

Conceitos básicos e objetivos da análise exploratória

02

Medidas de Tendência Central

Média, mediana e moda

03

Medidas de Dispersão

Amplitude, variância, desvio padrão e coeficiente de variação

04

Medidas de Posição

Quartis, percentis e identificação de outliers

05

Medidas de Forma

Assimetria e curtose

06

Estatísticas Multivariadas

Correlação e covariância

Durante este curso, aprenderemos como transformar dados brutos em insights valiosos através da estatística descritiva.

# O que é Análise Exploratória de Dados?

A Análise Exploratória de Dados (EDA) é o processo sistemático de:

- Investigar conjuntos de dados para descobrir padrões ocultos
- Detectar anomalias e valores atípicos
- Testar hipóteses preliminares
- Verificar suposições através de resumos estatísticos
- Utilizar representações gráficas para melhor compreensão dos dados

A EDA é frequentemente o primeiro passo crucial em qualquer análise de dados, servindo como base para decisões analíticas subsequentes.



# Objetivos da Análise Exploratória de Dados

1

## Compreender a Estrutura dos Dados

Identificar tipos de variáveis, distribuições e qualidade dos dados antes da análise formal.

2

## Identificar Padrões e Relacionamentos

Descobrir tendências, associações e conexões entre variáveis que podem não ser evidentes à primeira vista.

3

## Detectar Outliers e Anomalias

Encontrar valores atípicos que podem representar erros de coleta ou casos especiais dignos de investigação adicional.

4

## Formular Hipóteses

Desenvolver questões e suposições baseadas em evidências para análises estatísticas posteriores mais aprofundadas.

5

## Orientar Escolhas Metodológicas

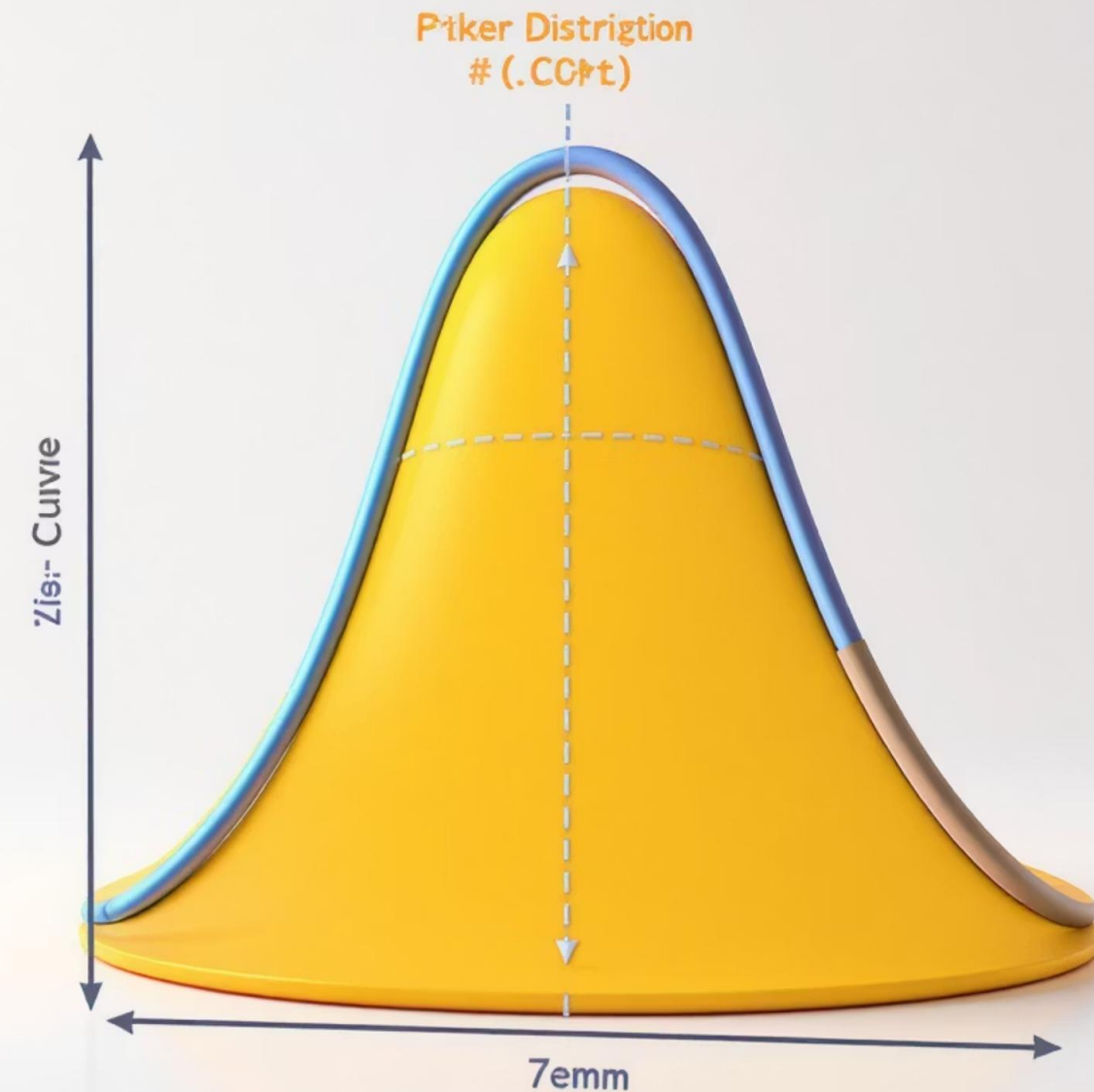
Definir quais modelos estatísticos ou técnicas de visualização serão mais adequados para os dados em questão.



## Medidas de Tendência Central

As medidas de tendência central são estatísticas que identificam um valor central ou típico em um conjunto de dados. As três principais medidas são a média aritmética, a mediana e a moda.

Estas medidas nos ajudam a compreender onde os dados tendem a se concentrar, fornecendo um ponto de referência para análises posteriores.



## Média Aritmética

**Definição:** A média aritmética é a soma de todos os valores do conjunto dividida pelo número total de observações.

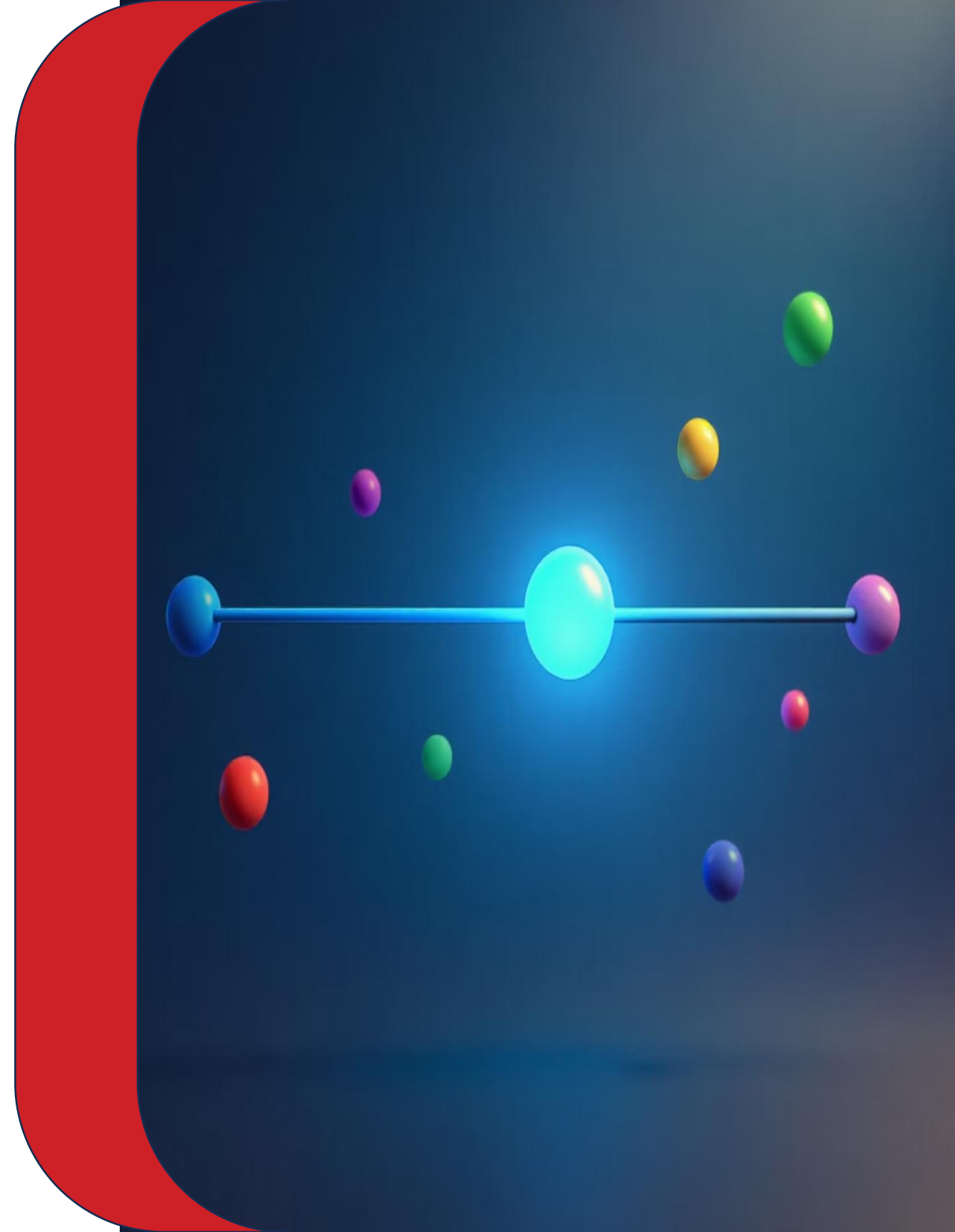
**Fórmula:**

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

**Características:**

- Altamente sensível a valores extremos (outliers)
- Pode não existir como valor real no conjunto de dados
- Ideal para distribuições aproximadamente simétricas
- Utilizada como base para muitas outras estatísticas

A média é afetada por todos os valores do conjunto, o que a torna menos representativa quando existem outliers significativos.



# Mediana

**Definição:** A mediana é o valor central que divide o conjunto ordenado em duas partes iguais.

## Como calcular:

1. Ordenar os dados em ordem crescente
2. Se  $n$  for ímpar: mediana = valor na posição  $(n+1)/2$
3. Se  $n$  for par: mediana = média dos valores nas posições  $n/2$  e  $(n/2)+1$

## Características:

- Robusta a outliers (não é afetada por valores extremos)
- Sempre existe no conjunto de dados (ou como interpolação)
- Melhor representante para distribuições assimétricas



# Moda

**Definição:** A moda é o valor (ou valores) que ocorre(m) com maior frequência em um conjunto de dados.

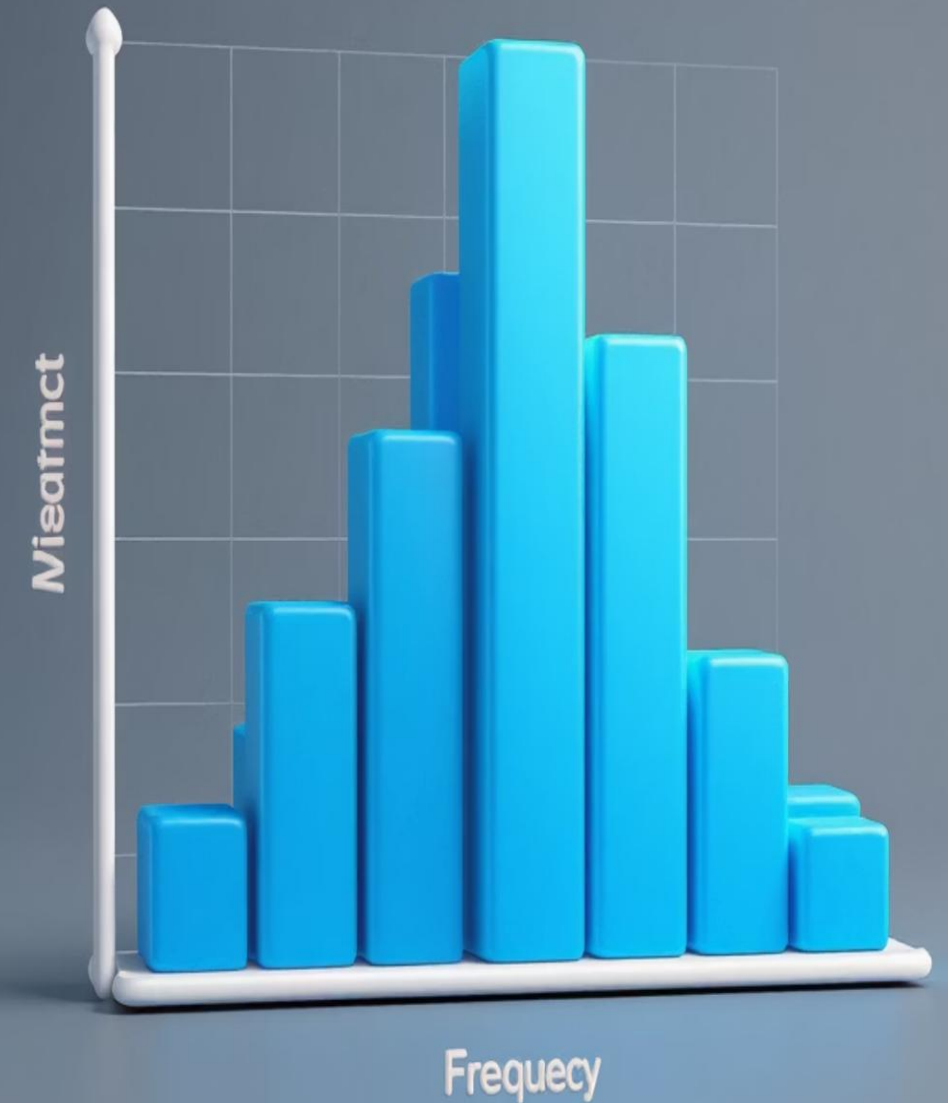
## Tipos de distribuições:

- **Unimodal:** possui apenas uma moda
- **Bimodal:** possui duas modas distintas
- **Multimodal:** possui mais de duas modas
- **Amodal:** não possui moda definida (todos os valores têm a mesma frequência)

## Características:

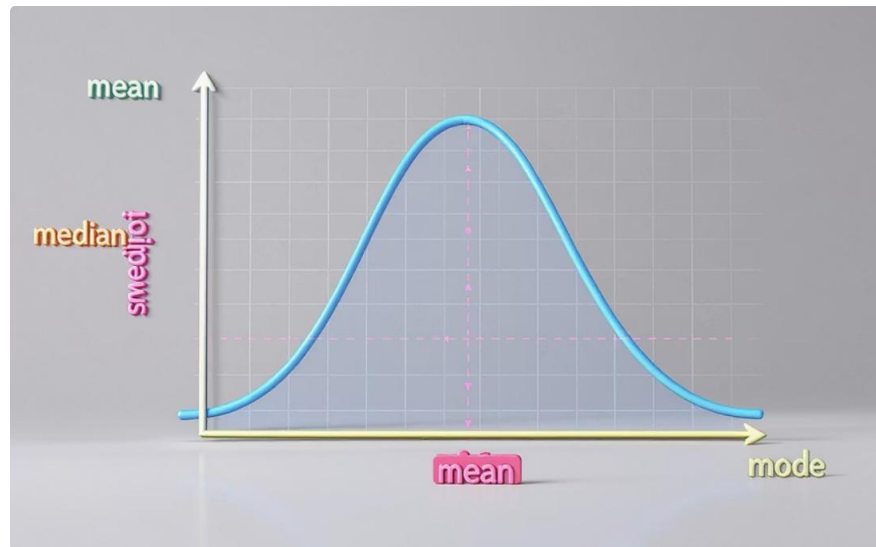
- Única medida adequada para dados categóricos
- Não é afetada por valores extremos
- Pode não ser única ou não existir

A moda é particularmente útil para dados categóricos e distribuições com picos claros de frequência.





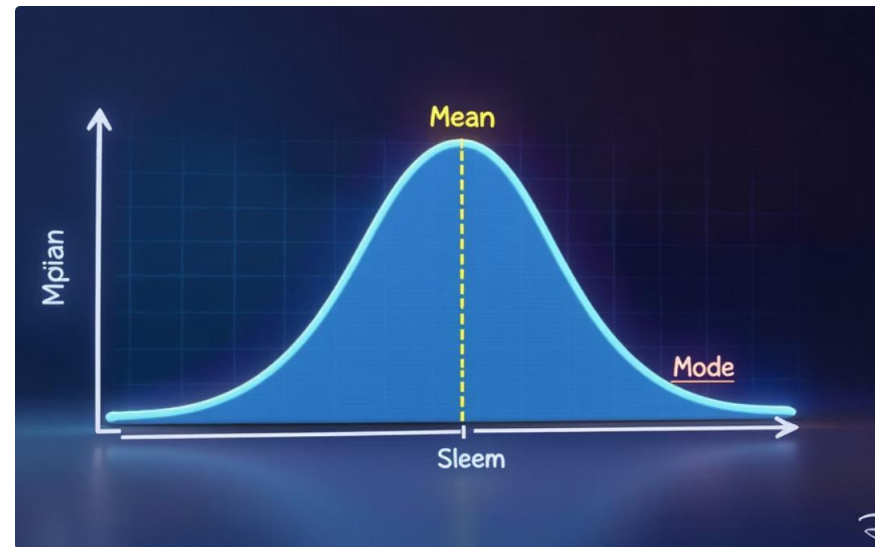
# Comparação das Medidas de Tendência Central



## Distribuição Simétrica

Média  $\approx$  Mediana  $\approx$  Moda

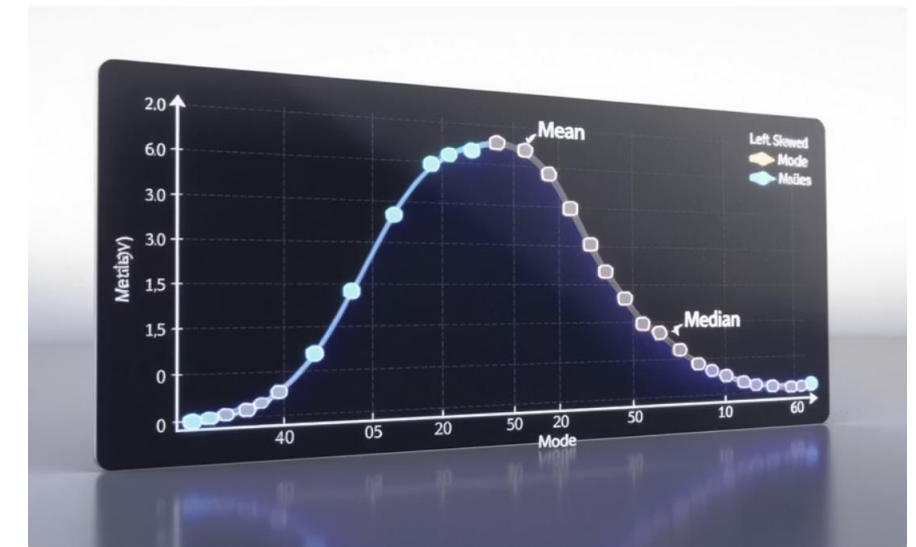
Nas distribuições perfeitamente simétricas, as três medidas coincidem no mesmo ponto.



## Assimetria Positiva

Média  $>$  Mediana  $>$  Moda

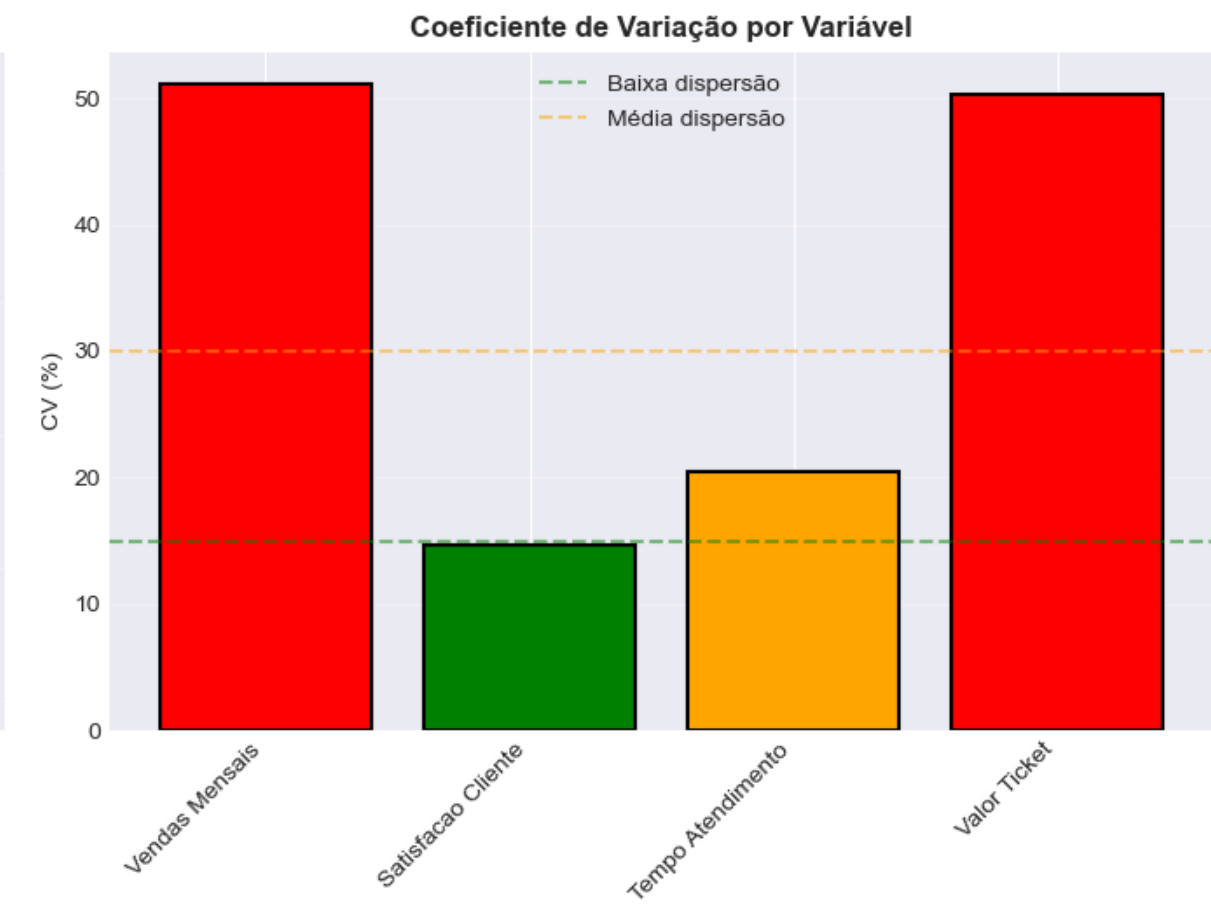
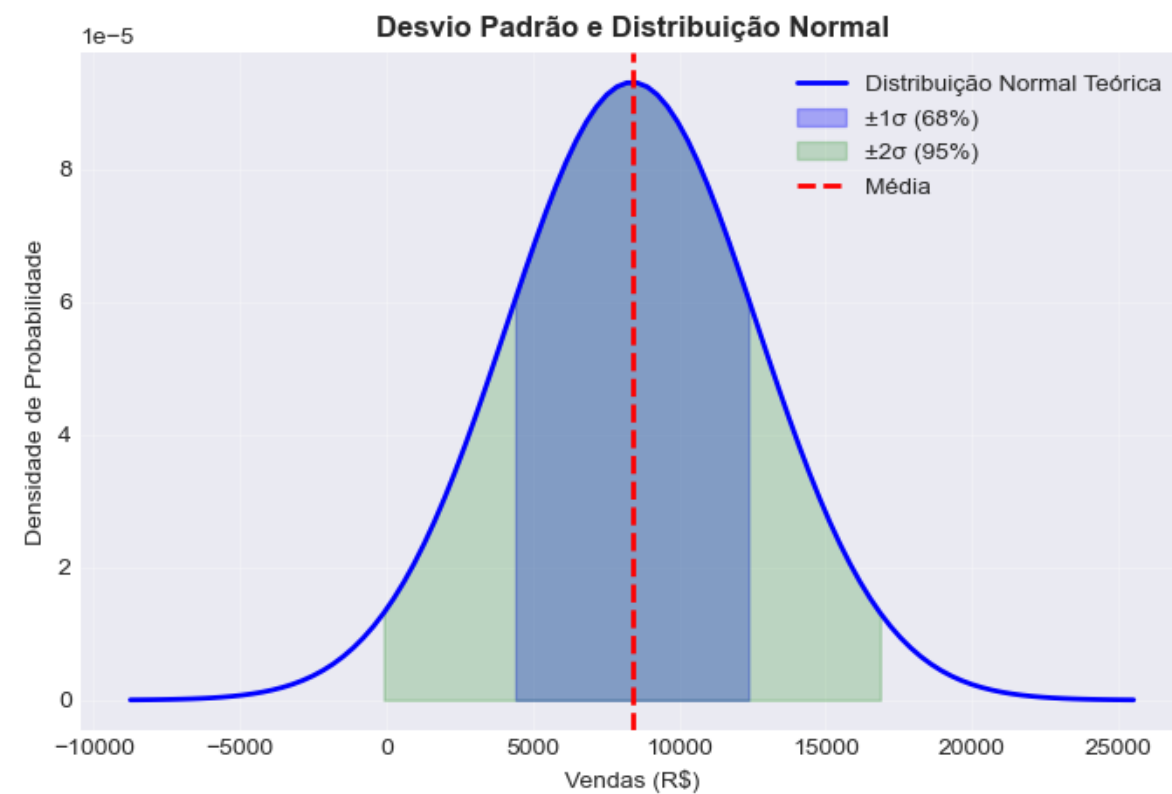
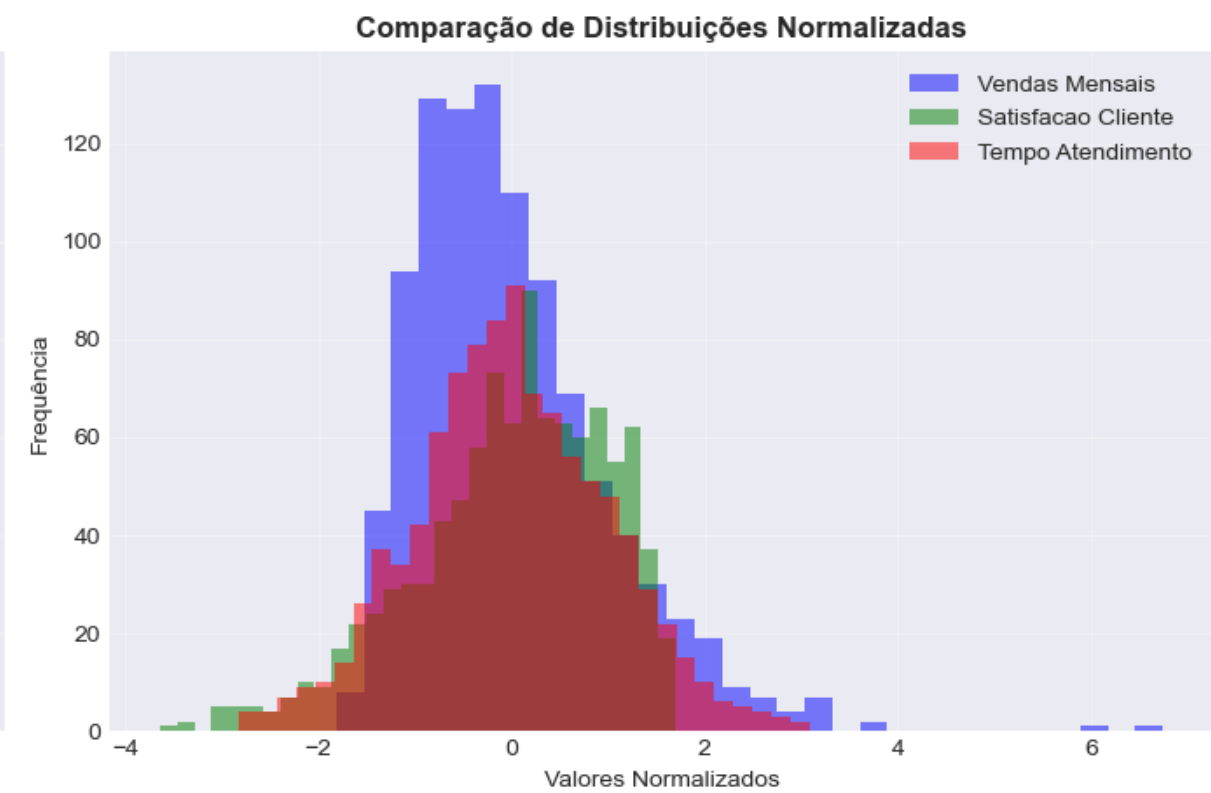
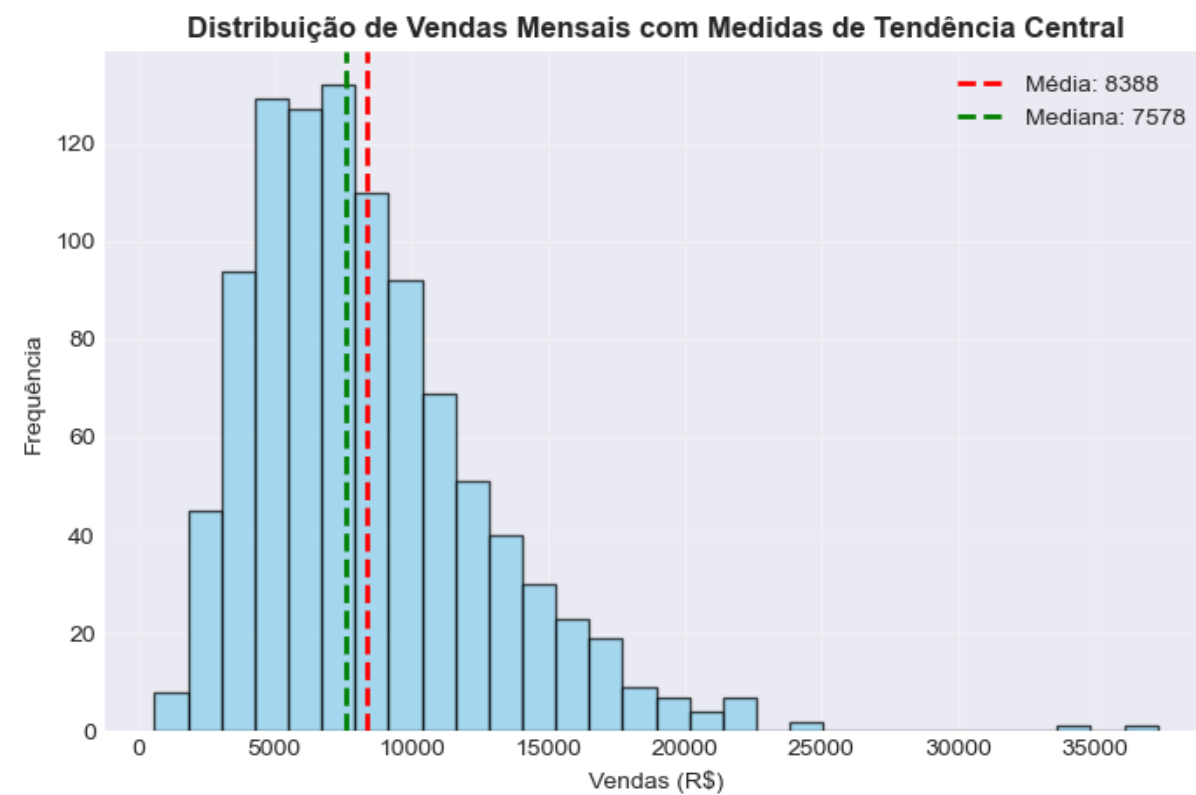
A cauda à direita "puxa" a média para valores mais altos, enquanto a moda permanece no pico da distribuição.



## Assimetria Negativa

Média  $<$  Mediana  $<$  Moda

A cauda à esquerda "puxa" a média para valores mais baixos, afastando-a do pico da distribuição.



# Medidas de Dispersão

Enquanto as medidas de tendência central nos dizem onde os dados se concentram, as **medidas de dispersão** nos informam o quanto os dados estão espalhados ou variando em torno desses valores centrais.

As principais medidas de dispersão incluem a amplitude, a variância, o desvio padrão e o coeficiente de variação. Essas estatísticas são essenciais para compreender a variabilidade e consistência dos dados.

## Amplitude (Range)

**Definição:** A amplitude é a diferença entre o maior e o menor valor de um conjunto de dados.

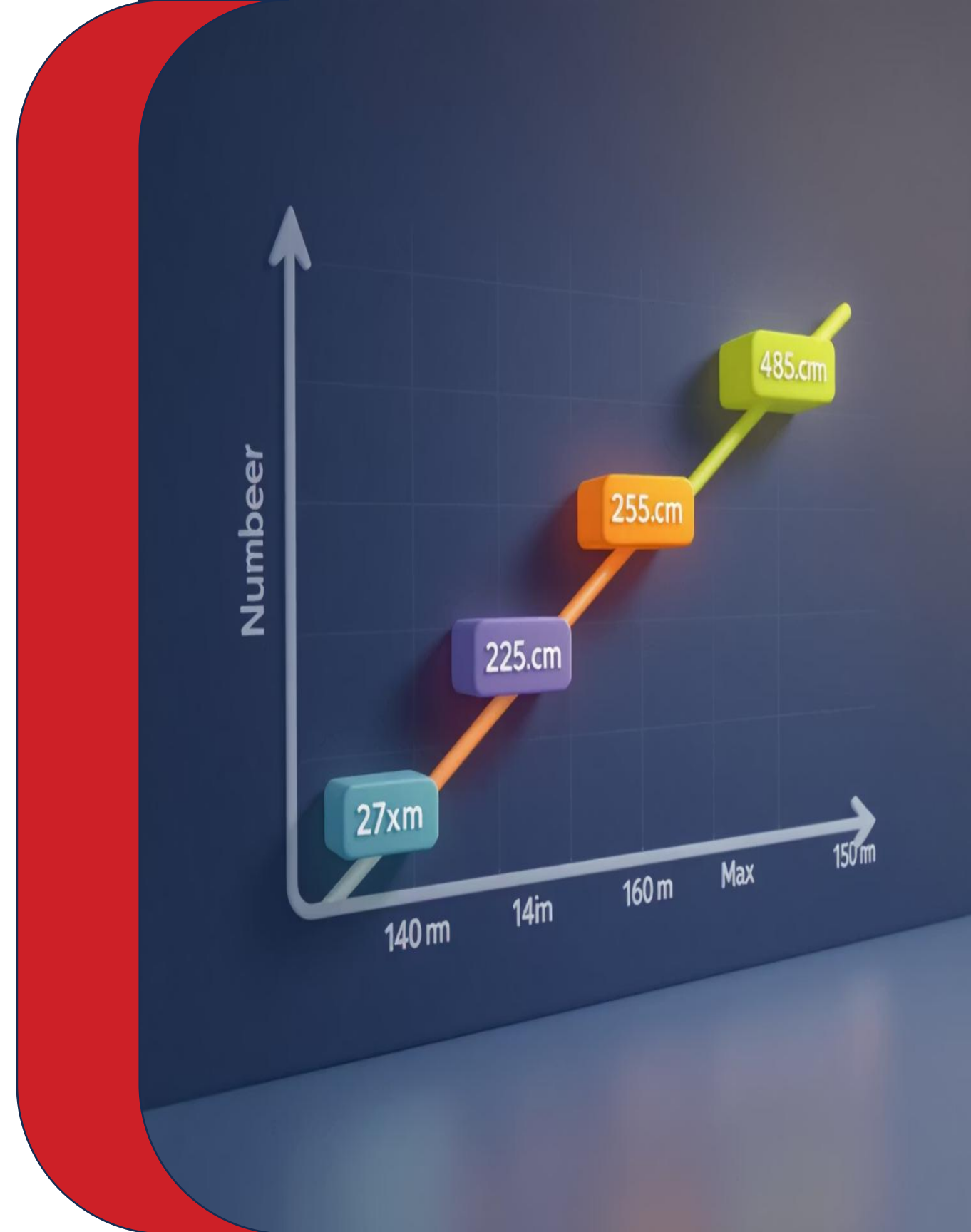
**Fórmula:**

$$Amplitude = X_{max} - X_{min}$$

**Características:**

- Medida mais simples de dispersão
- Extremamente sensível a outliers
- Considera apenas os valores extremos, ignorando a distribuição interna
- Útil para uma análise rápida e preliminar

A amplitude fornece uma visão rápida do espalhamento total dos dados, mas não informa sobre a distribuição dos valores intermediários.



# Variância

**Definição:** A variância é a média dos quadrados dos desvios em relação à média aritmética.

**Fórmula:**

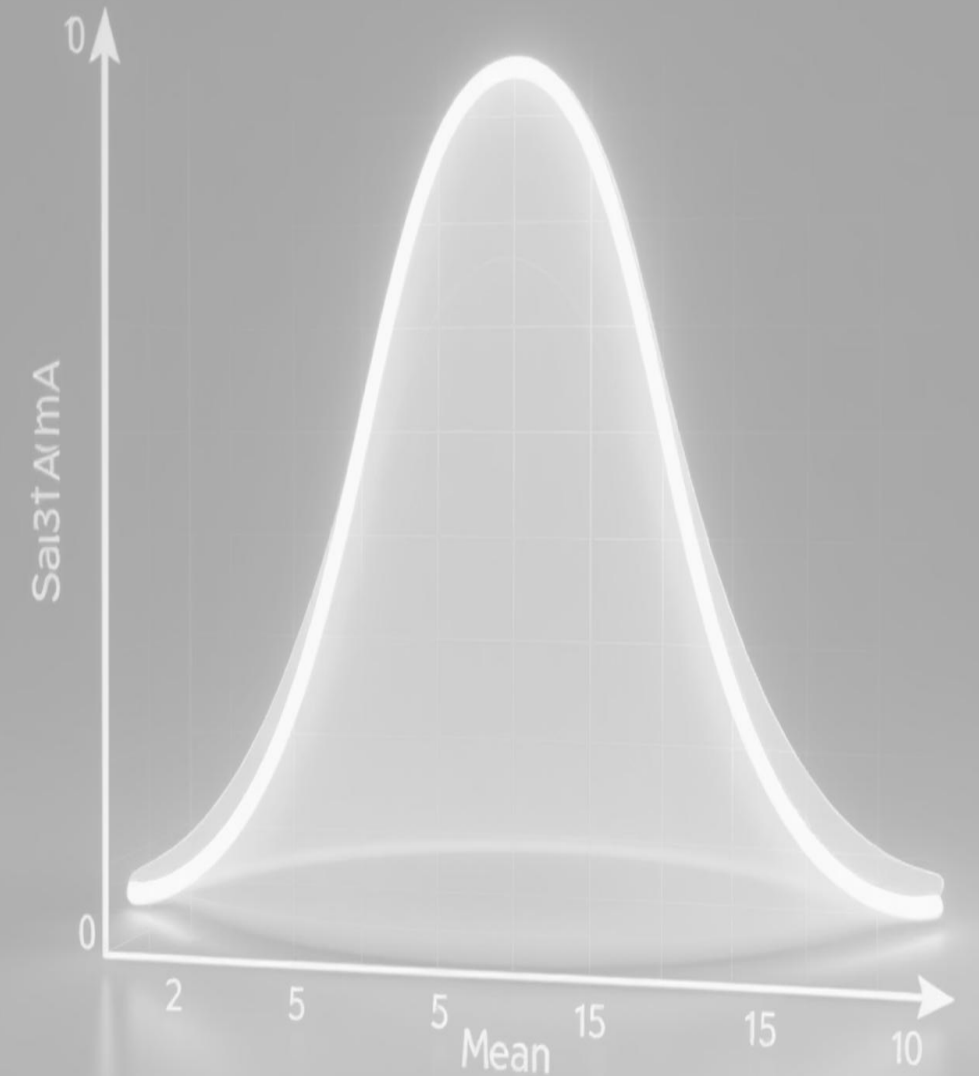
$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

Onde:

- $\sigma^2$  = variância
- $x_i$  = valor individual
- $\mu$  = média aritmética
- $n$  = número total de observações

**Características:**

- Considera todos os valores do conjunto de dados
- Unidade de medida é o quadrado da unidade original dos dados
- Base para muitas técnicas estatísticas avançadas
- Sensível a outliers devido ao quadrado dos desvios





# Desvio Padrão

**Definição:** O desvio padrão é a raiz quadrada da variância, representando a dispersão dos dados na mesma unidade da variável original.

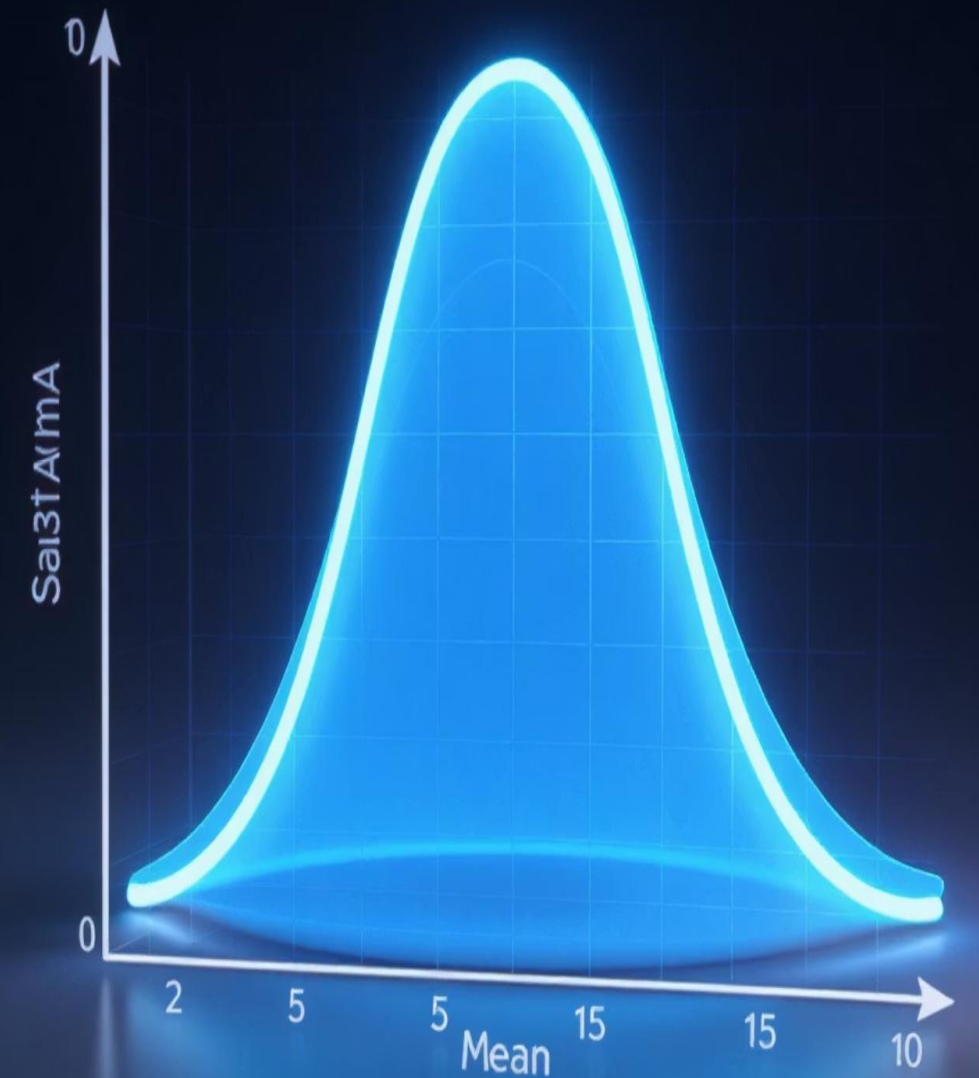
**Fórmula:**

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

**Interpretação:**

- Indica, em média, o quanto os valores se afastam da média aritmética
- Quanto maior o desvio padrão, maior a dispersão dos dados
- Na distribuição normal, aproximadamente 68% dos dados estão a  $\pm 1\sigma$  da média

O desvio padrão é uma das medidas de dispersão mais utilizadas por estar na mesma unidade que os dados originais.



# Coeficiente de Variação (CV)

**Definição:** O coeficiente de variação é uma medida relativa de dispersão que expressa o desvio padrão como porcentagem da média.

**Fórmula:**

$$CV = \frac{\sigma}{\mu} \times 100 \%$$

**Interpretação:**



**CV < 15%**

Baixa dispersão

Dados bastante homogêneos



**15% ≤ CV ≤ 30%**

Média dispersão

Dados com dispersão moderada



**CV > 30%**

Alta dispersão

Dados muito heterogêneos

O CV permite comparar a variabilidade de diferentes conjuntos de dados, mesmo quando as unidades de medida são diferentes.

# Desvio Padrão

**Definição:** O desvio padrão é a raiz quadrada da variância, representando a dispersão dos dados na mesma unidade da variável original.

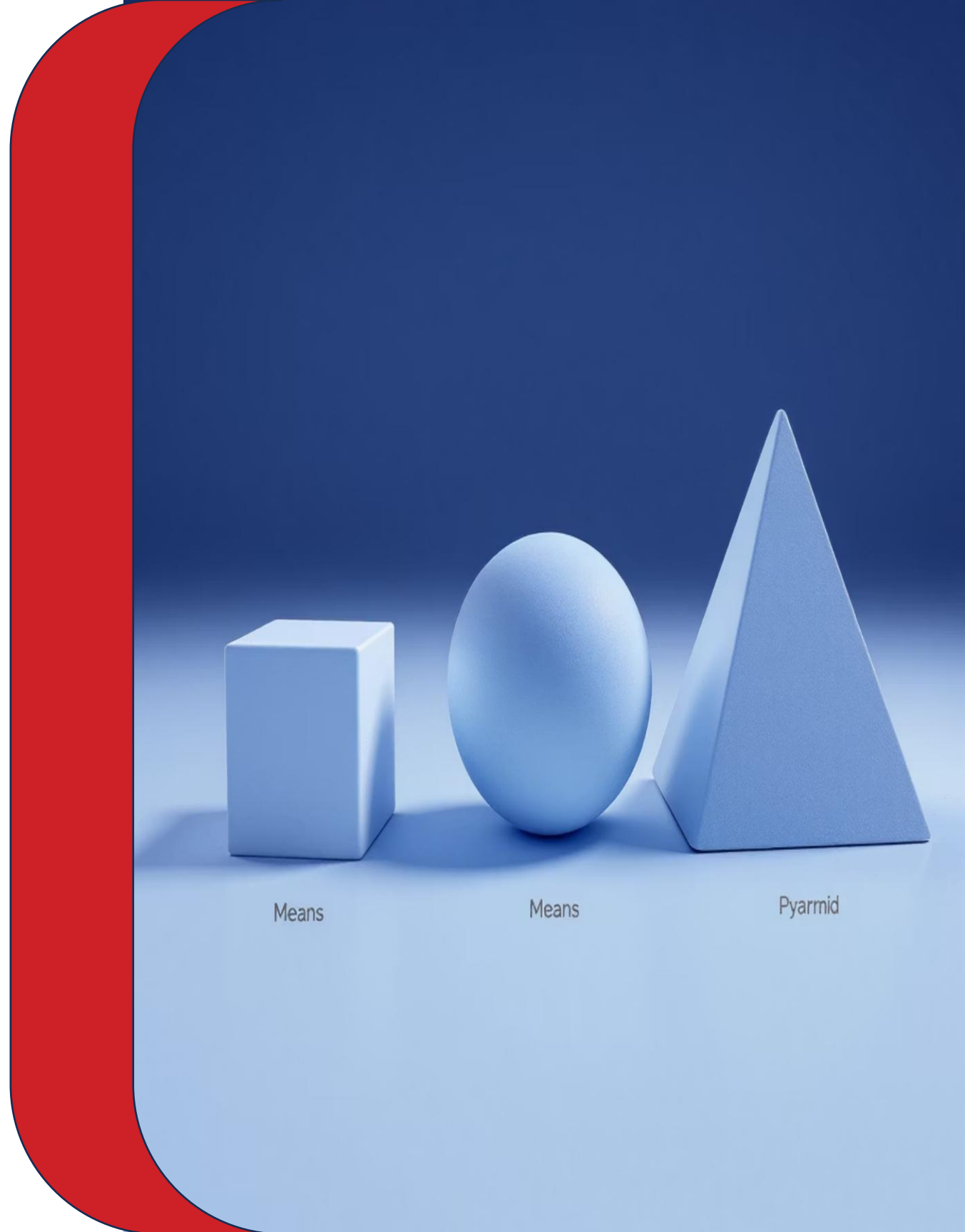
**Fórmula:**

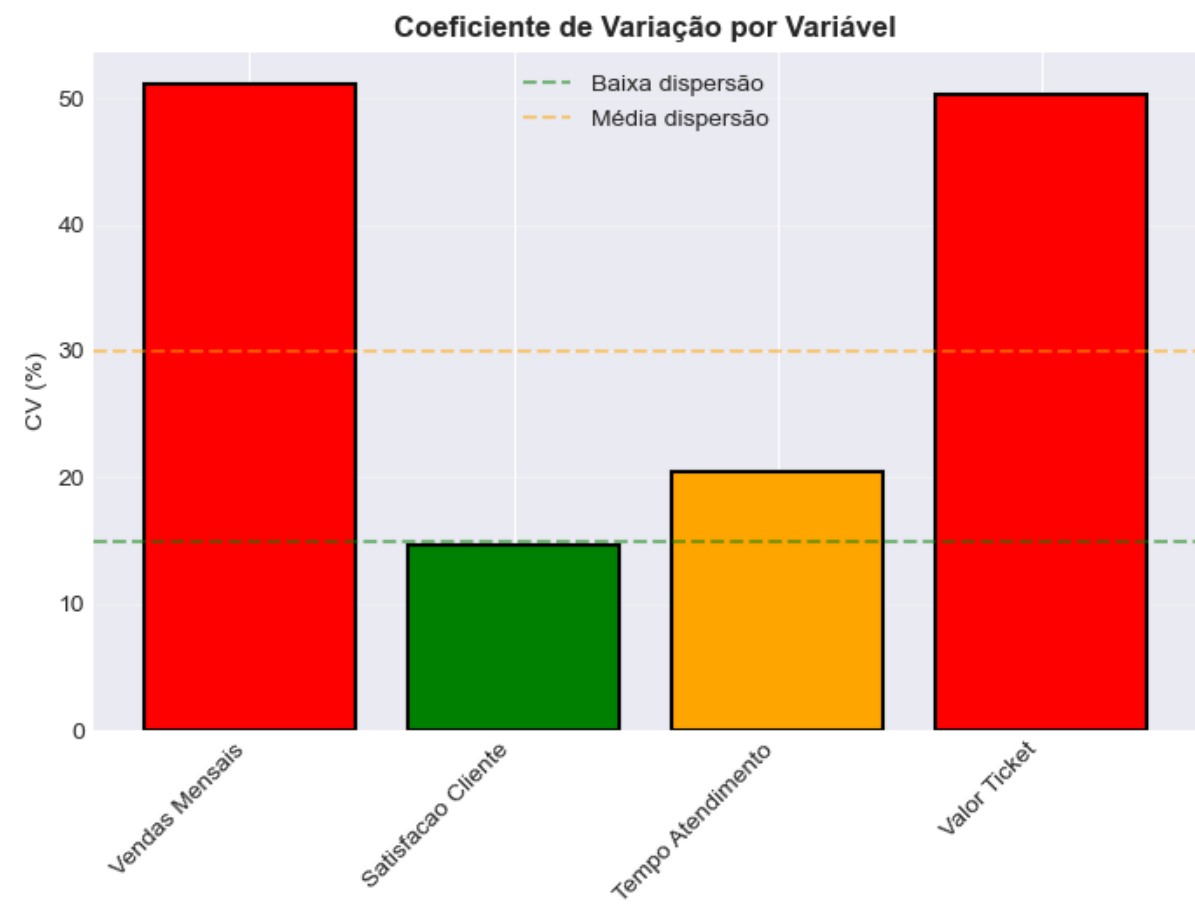
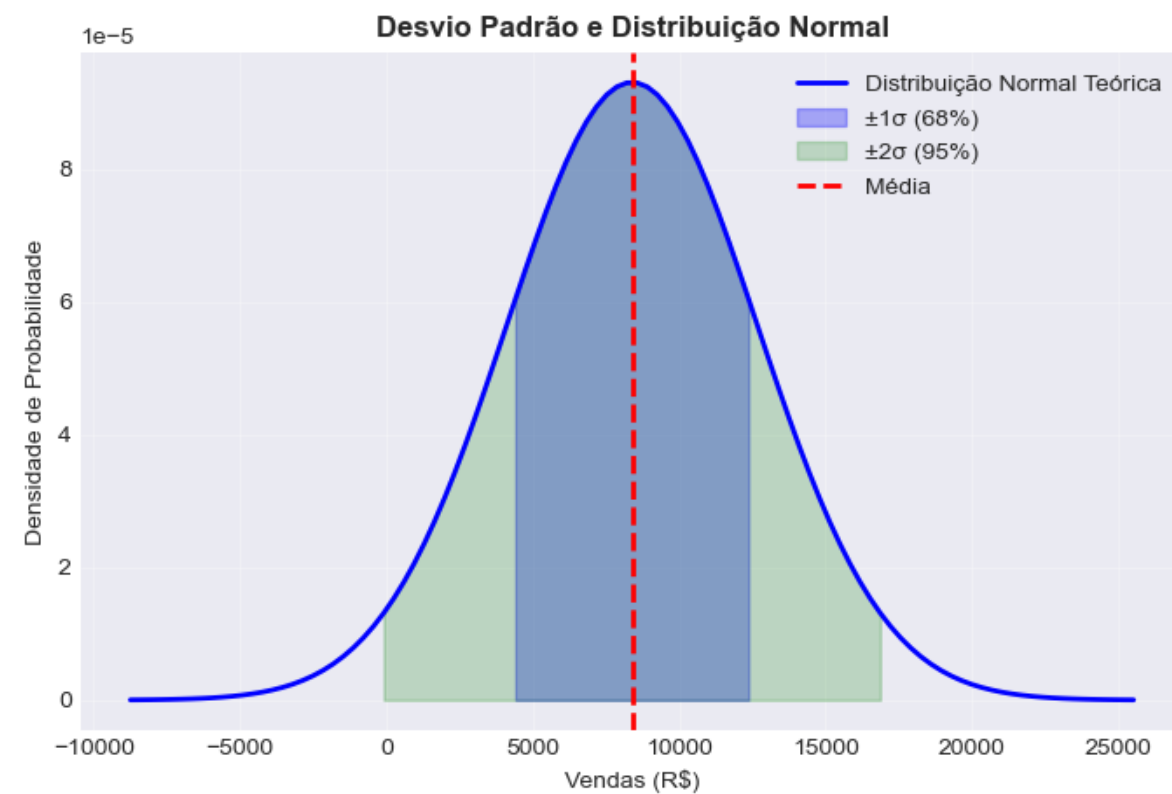
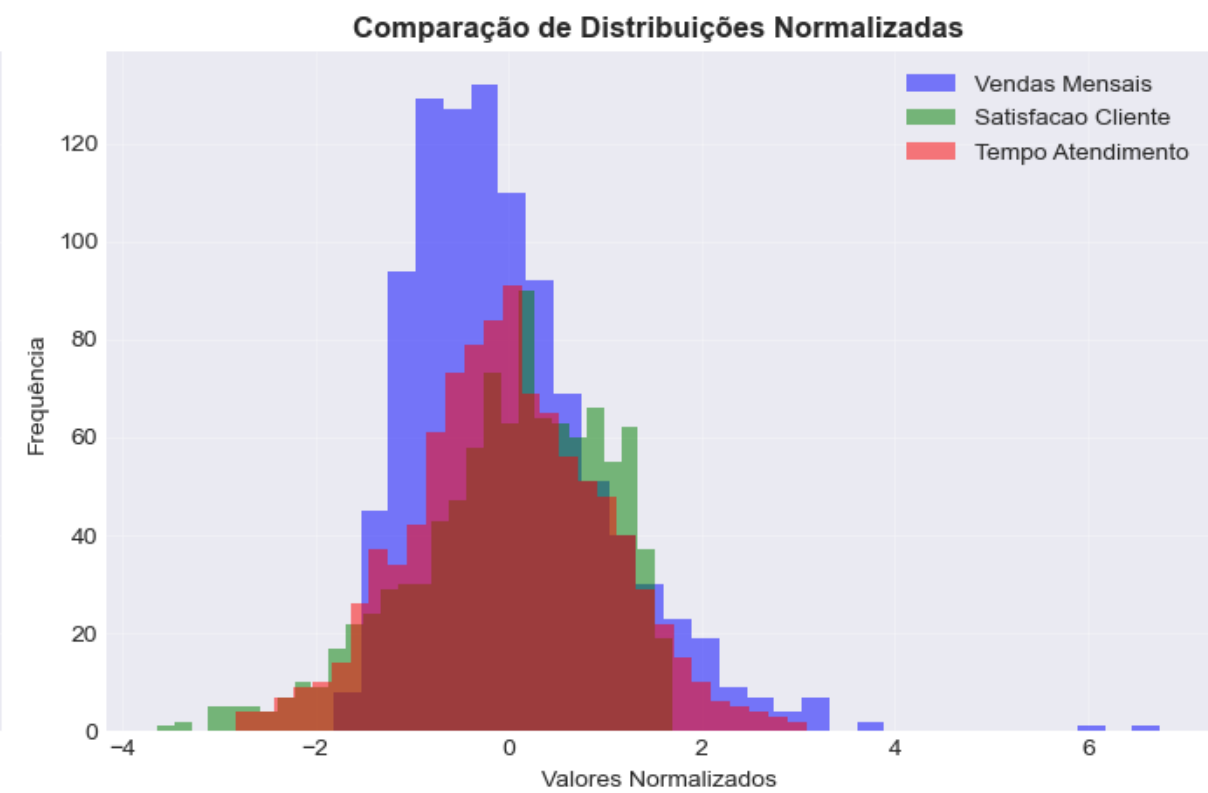
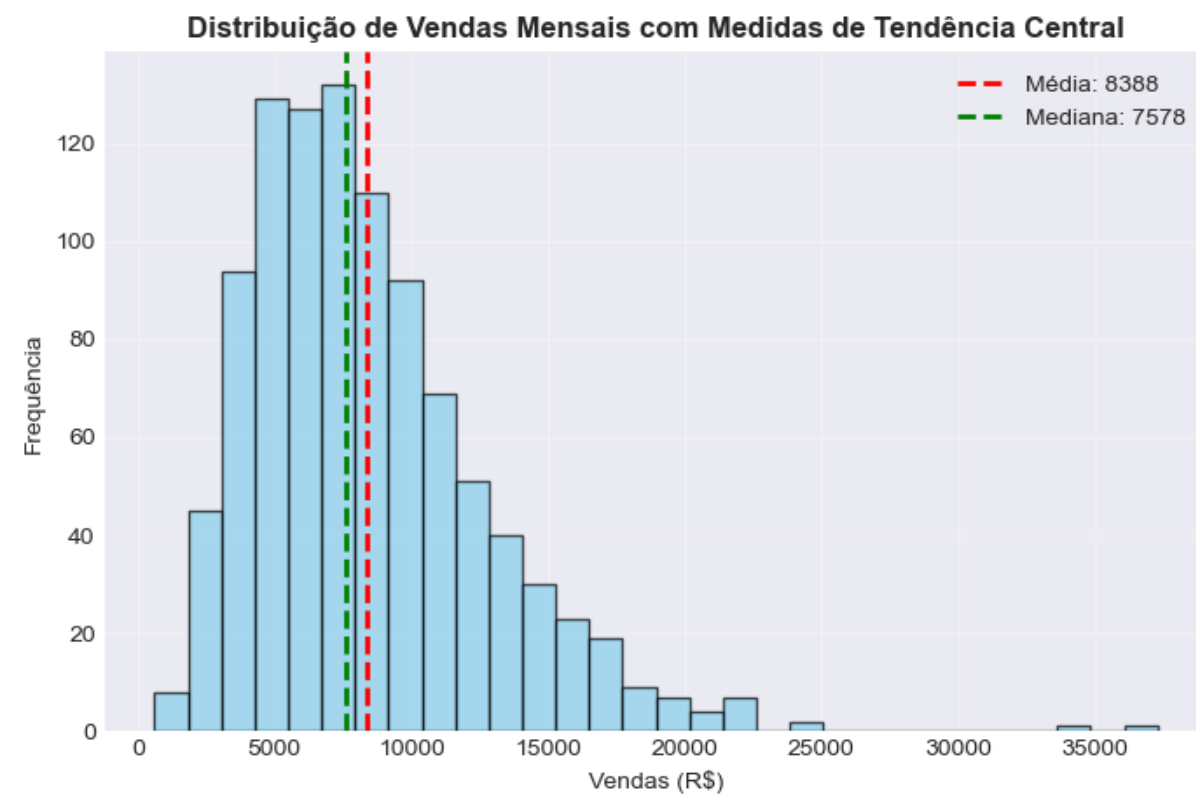
$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$$

**Interpretação:**

- Indica, em média, o quanto os valores se afastam da média aritmética
- Quanto maior o desvio padrão, maior a dispersão dos dados
- Na distribuição normal, aproximadamente 68% dos dados estão a  $\pm 1\sigma$  da média

O desvio padrão é uma das medidas de dispersão mais utilizadas por estar na mesma unidade que os dados originais.







# Medidas de Posição: Quartis e Percentis

As medidas de posição dividem um conjunto ordenado de dados em partes iguais, permitindo compreender como os valores estão distribuídos ao longo do intervalo.

Os percentis indicam a posição relativa de um valor em relação ao conjunto de dados, enquanto os quartis são casos especiais de percentis que dividem os dados em quatro partes iguais.



# Quartis

## Primeiro Quartil (Q1)

Também conhecido como 25º percentil, representa o valor abaixo do qual estão 25% dos dados ordenados.

## Segundo Quartil (Q2)

Corresponde à mediana (50º percentil), valor que divide os dados ordenados em duas partes iguais.

## Terceiro Quartil (Q3)

Também conhecido como 75º percentil, representa o valor abaixo do qual estão 75% dos dados ordenados.

Os quartis são pontos de referência importantes que nos permitem dividir o conjunto de dados em quatro partes com igual número de observações, facilitando a compreensão da distribuição dos valores.

# Amplitude Interquartílica (IQR)

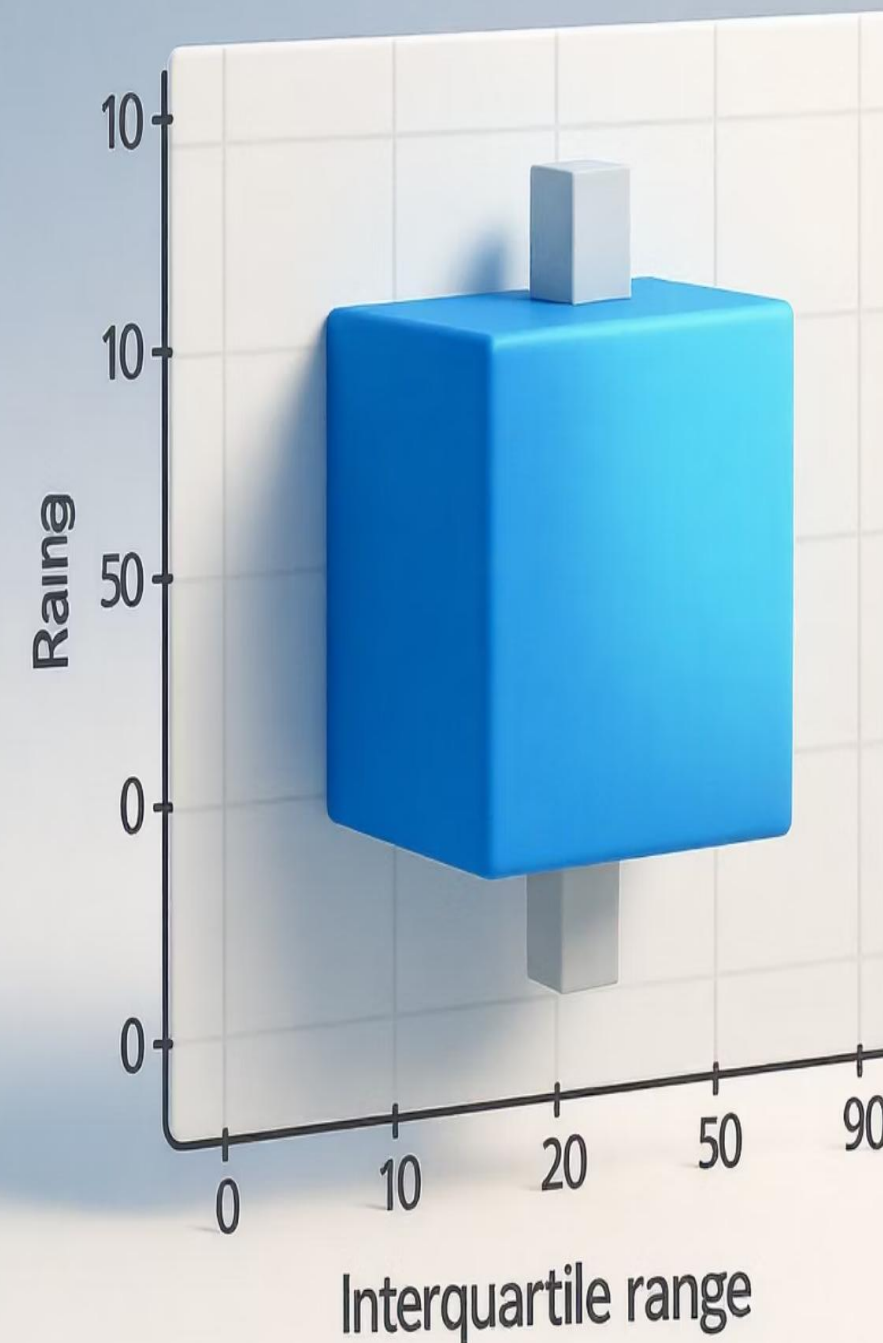
**Definição:** A amplitude interquartílica é a diferença entre o terceiro e o primeiro quartil.

**Fórmula:**

$$IQR = Q3 - Q1$$

**Características:**

- Medida robusta de dispersão, não afetada por outliers
- Representa a faixa onde estão concentrados 50% dos dados centrais
- Base para a construção de boxplots
- Utilizada para identificação de valores atípicos



# Identificação de Outliers com IQR

O método do IQR é uma técnica robusta para identificação de outliers, baseada nos

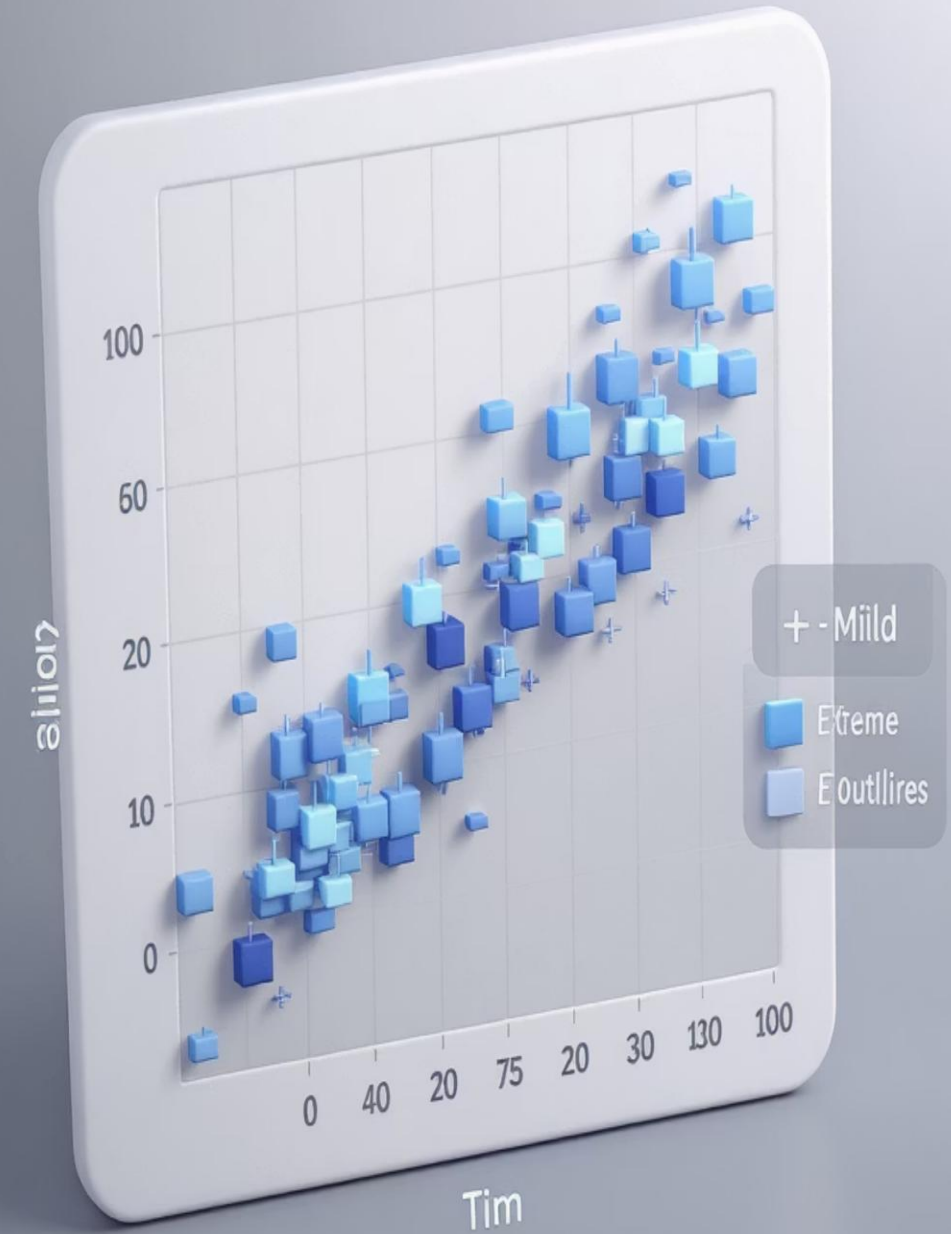
## **Critérios para Outliers:**

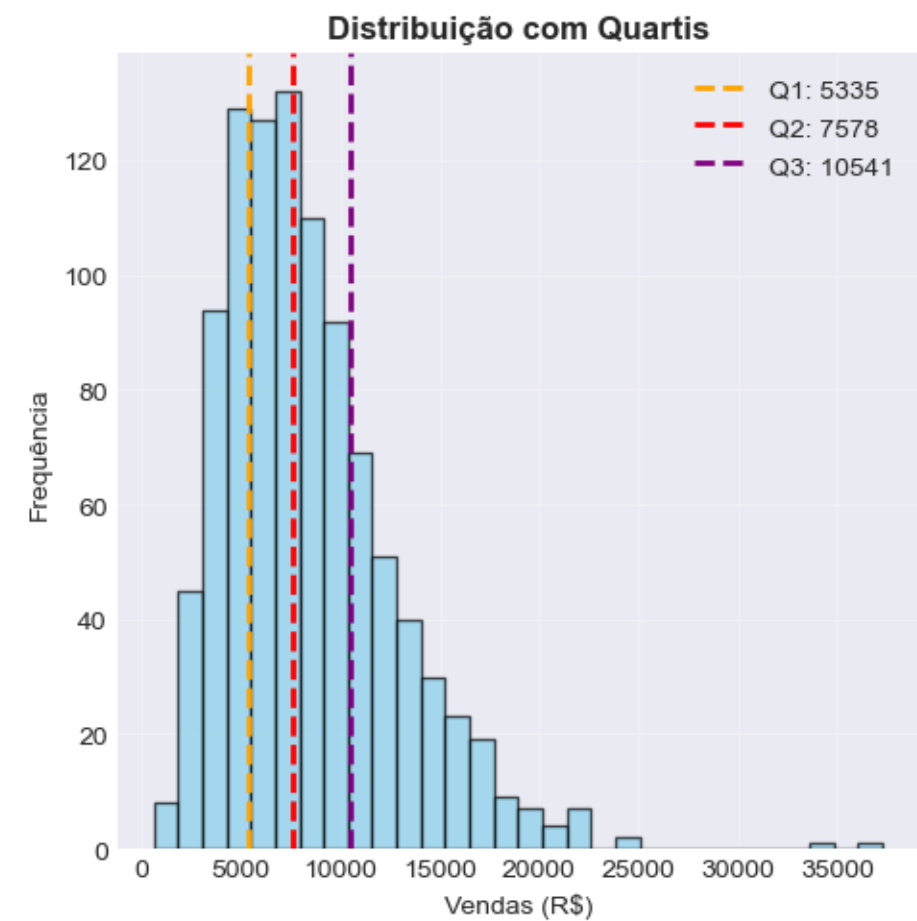
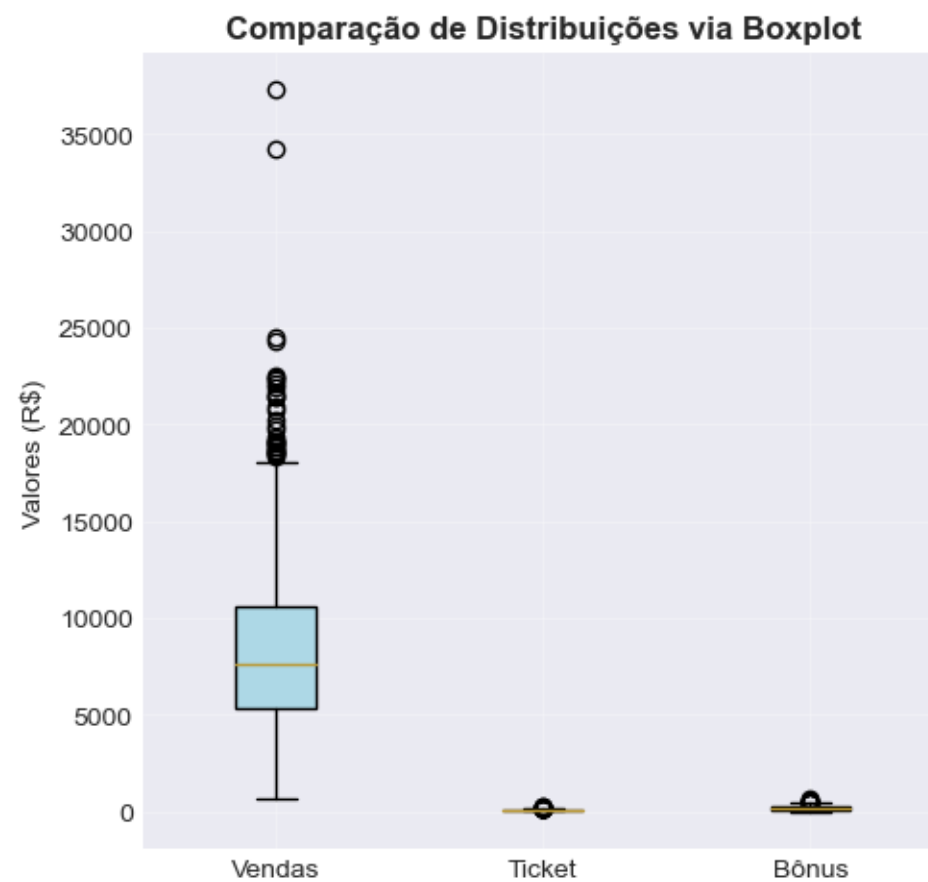
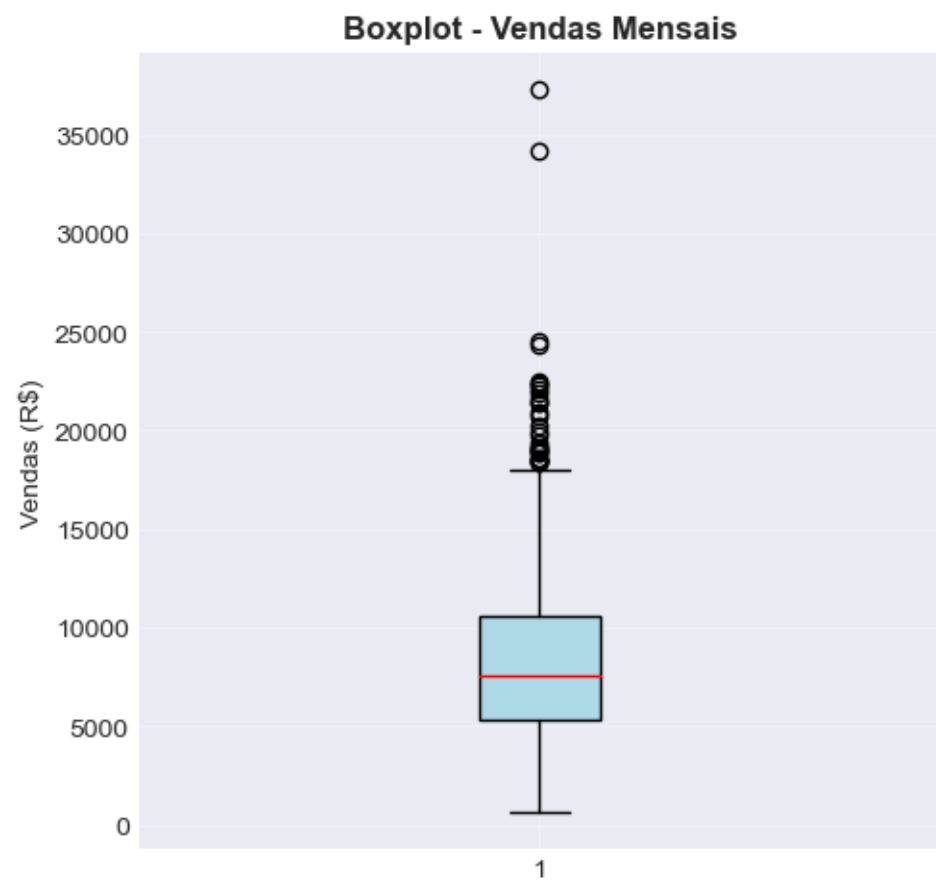
### **Outliers suaves:**

- Valores  $< Q1 - 1.5 \times IQR$
- Valores  $> Q3 + 1.5 \times IQR$

### **Outliers extremos:**

- Valores  $< Q1 - 3 \times IQR$
- Valores  $> Q3 + 3 \times IQR$



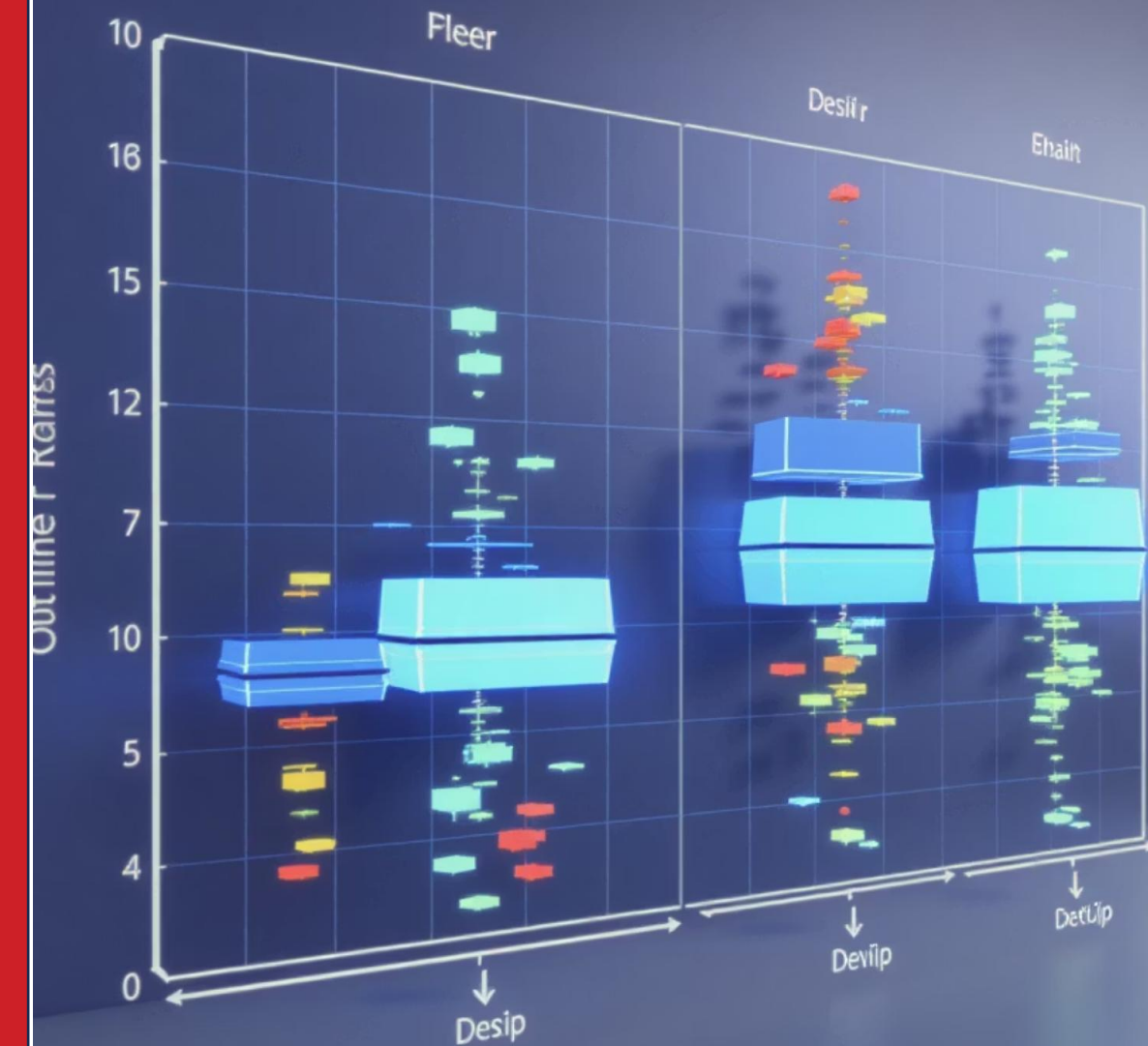


# Aplicação: Análise de Outliers em Boxplots

Os boxplots são representações visuais poderosas que mostram simultaneamente:

- A mediana (linha no centro da caixa)
- O primeiro e terceiro quartis (limites da caixa)
- A variabilidade dos dados (altura da caixa)
- A presença de valores atípicos (pontos individuais)
- A simetria ou assimetria da distribuição

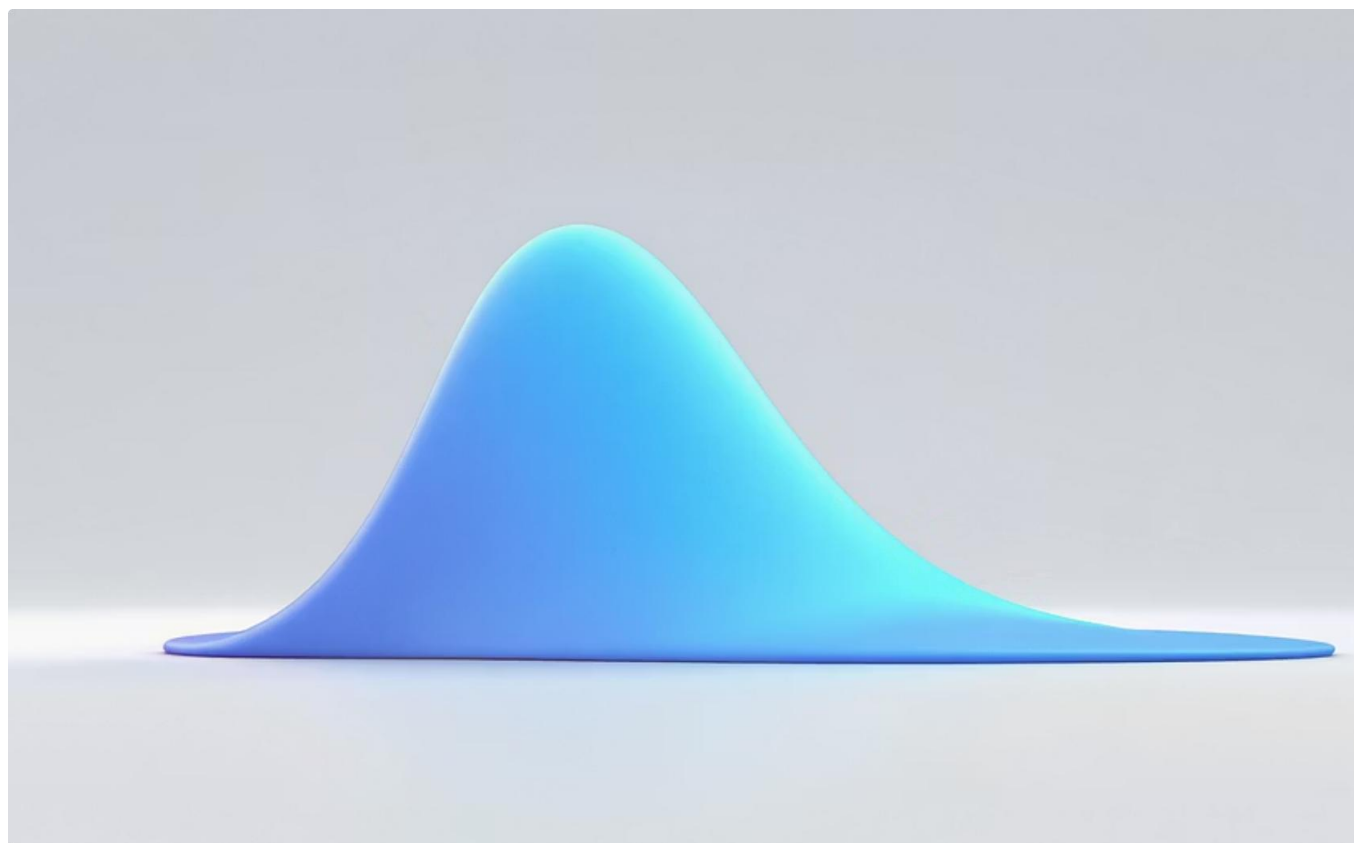
A análise de outliers via boxplots permite identificar rapidamente anomalias nos dados e comparar múltiplas distribuições lado a lado.





# Medidas de Forma da Distribuição

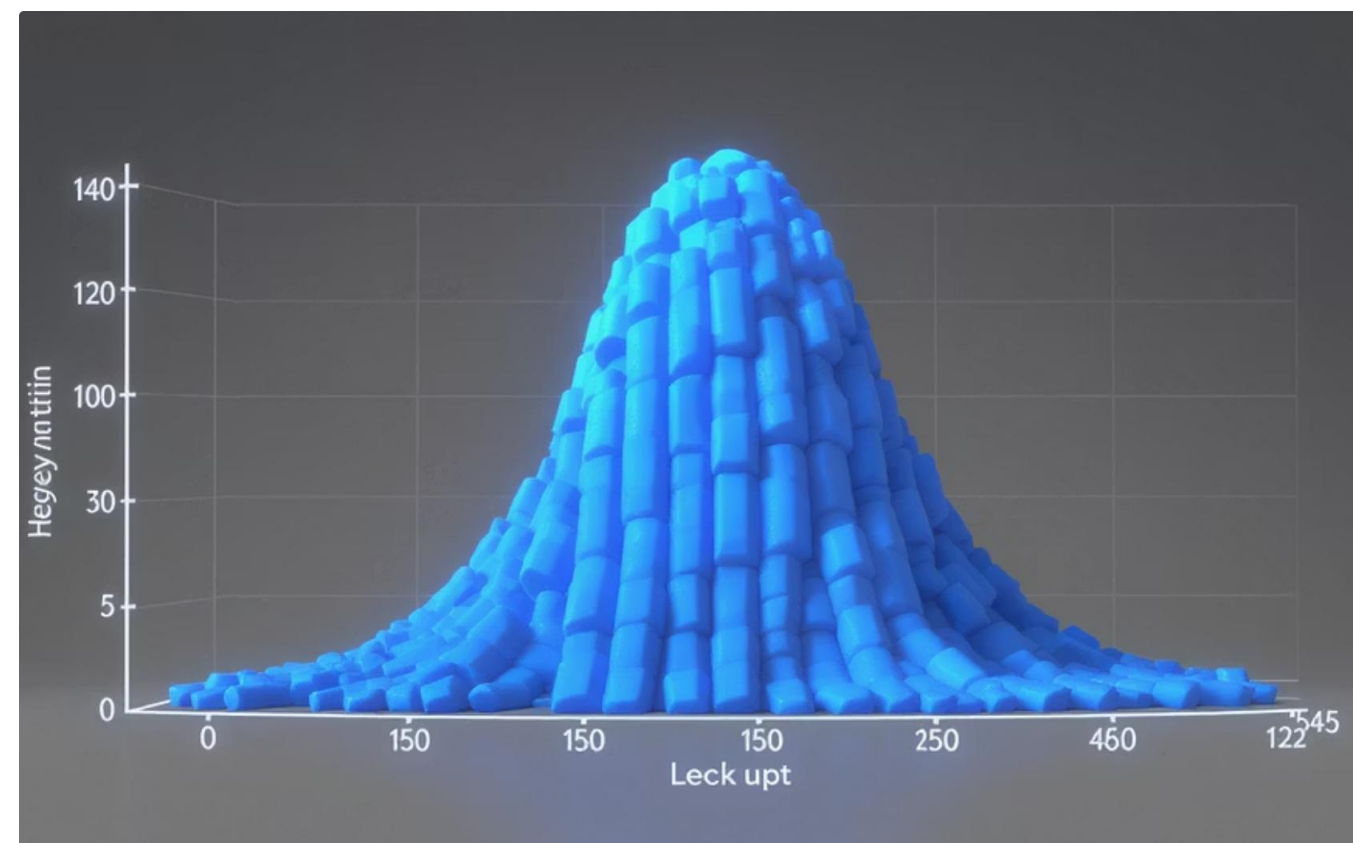
As medidas de forma descrevem características da distribuição dos dados que vão além das medidas de tendência central e dispersão.



## Assimetria (Skewness)

Mede o grau e direção da assimetria de uma distribuição, indicando se os dados estão concentrados em valores menores ou maiores.

Essas estatísticas fornecem informações importantes sobre o formato da distribuição, ajudando a entender seu comportamento e a escolher métodos analíticos apropriados.



## Curtose (Kurtosis)

Quantifica o "achatamento" da distribuição, indicando a concentração de valores próximos à média e o comportamento das caudas.

# Assimetria (Skewness)

**Definição:** A assimetria mede o grau em que uma distribuição se desvia da simetria.

**Interpretação:**

- $Skew \approx 0$ : distribuição aproximadamente simétrica
- $Skew > 0$ : assimetria positiva (cauda à direita)
- $Skew < 0$ : assimetria negativa (cauda à esquerda)

A assimetria influencia a relação entre as medidas de tendência central e pode afetar a escolha de técnicas estatísticas apropriadas.



Negative skew



Symmetrical



Positive Skew

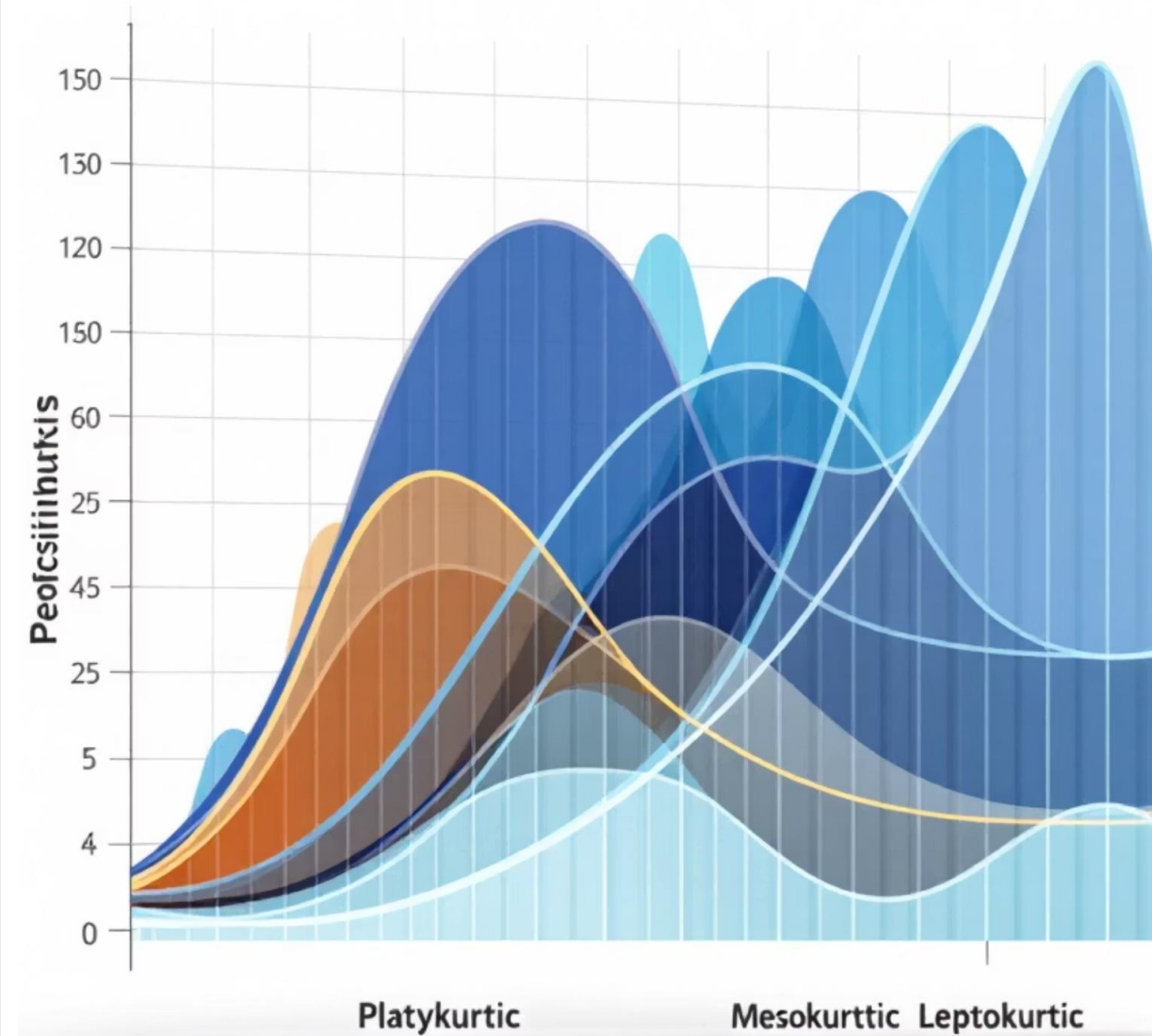
## Curtose (Kurtosis)

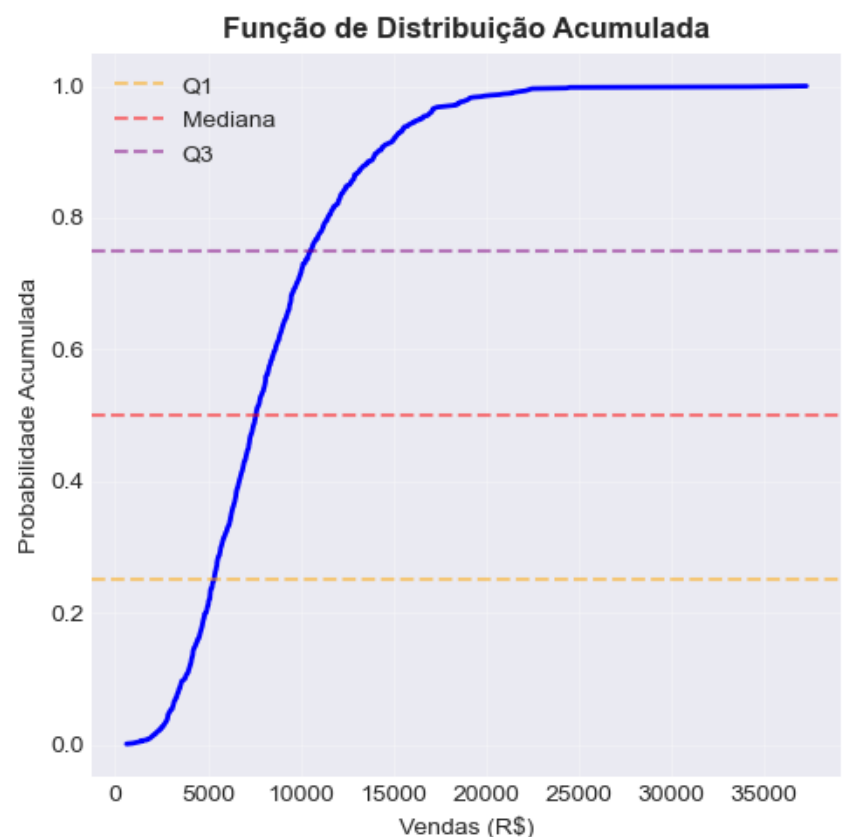
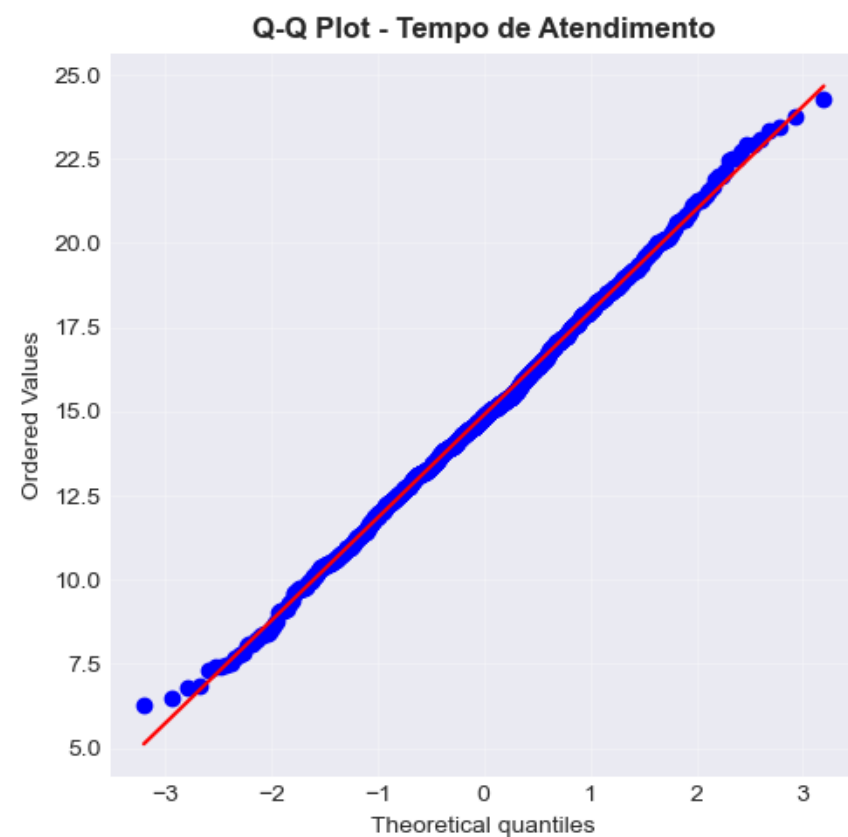
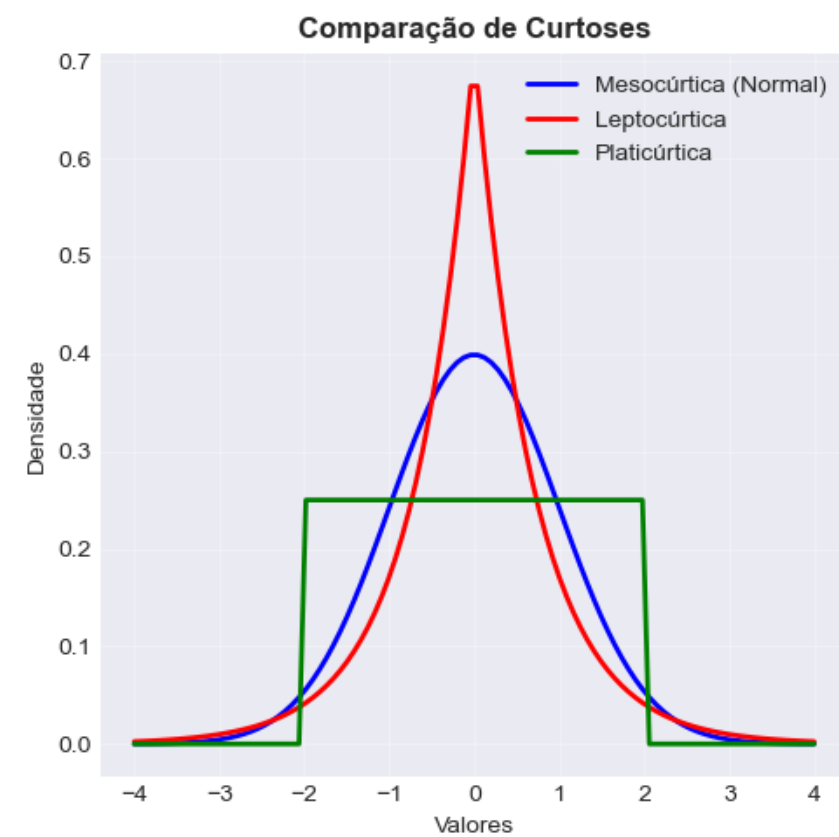
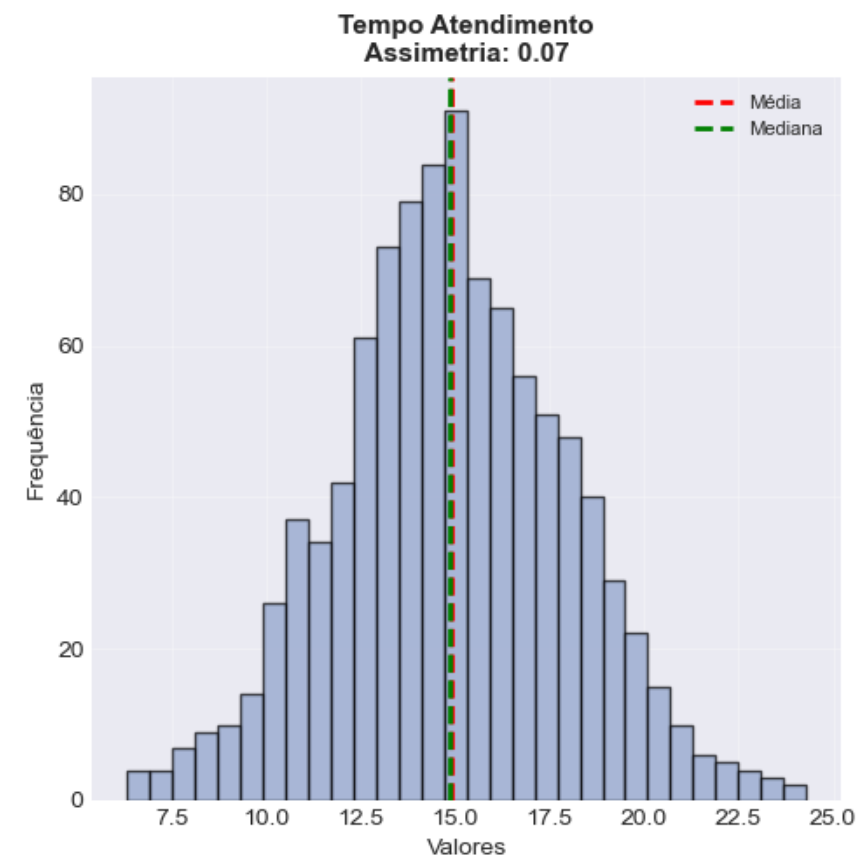
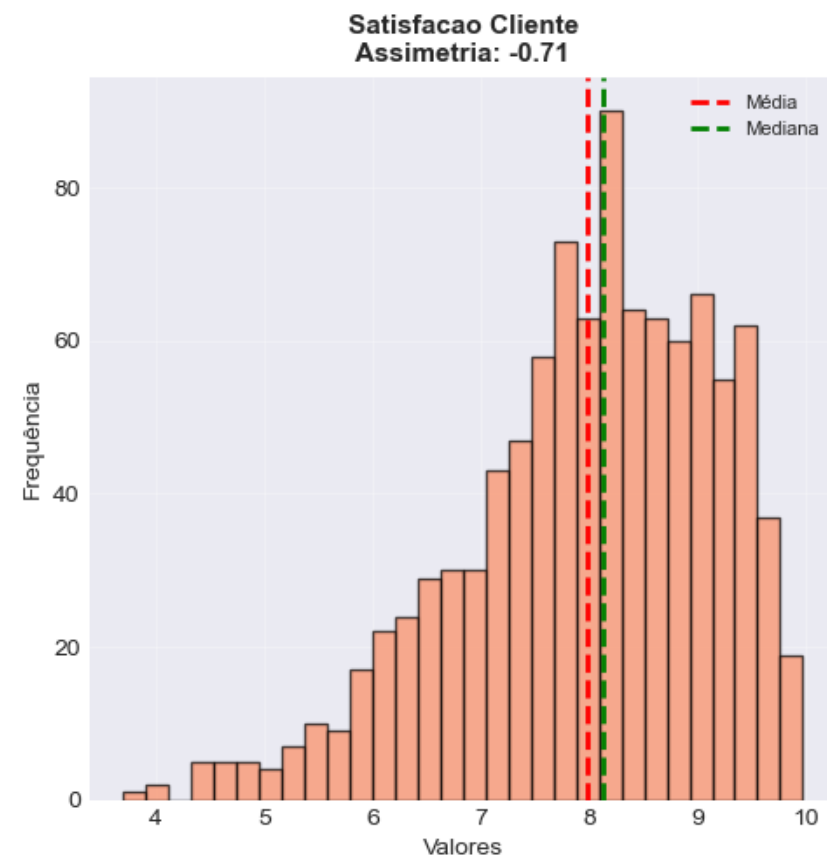
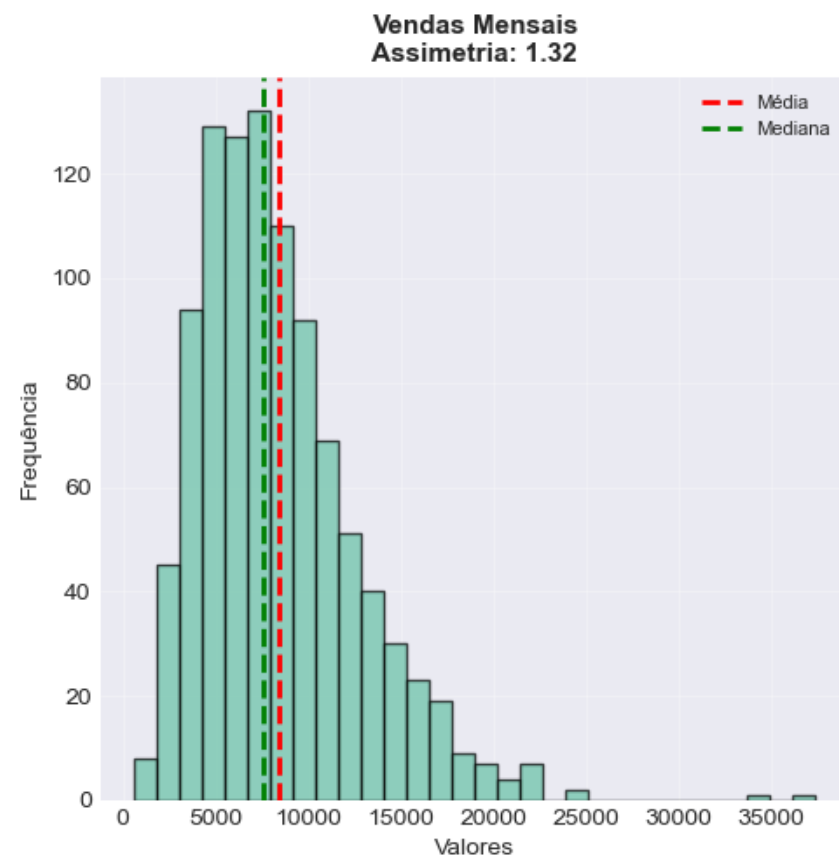
**Definição:** A curtose mede o grau de "achatamento" ou "pico" de uma distribuição em comparação com a distribuição normal.

### Interpretação:

- $Kurt \approx 0$ : mesocúrtica (similar à normal)
- $Kurt > 0$ : leptocúrtica (mais pontiaguda)
- $Kurt < 0$ : platicúrtica (mais achatada)

Uma curtose alta indica maior concentração de valores próximos à média e/ou caudas mais pesadas, enquanto uma curtose baixa indica distribuição mais uniforme.







# Impacto da Assimetria e Curtose na Análise de Dados

## Escolha de Medidas Descritivas

Em distribuições assimétricas, a mediana é geralmente mais representativa que a média. Distribuições com alta curtose podem exigir medidas robustas.

## Seleção de Testes Estatísticos

Muitos testes paramétricos assumem normalidade (simetria e curtose específicas). Assimetria ou curtose extremas podem exigir testes não paramétricos.

## Transformações de Dados

Dados com alta assimetria frequentemente necessitam de transformações (logarítmica, raiz quadrada, etc.) para normalizar sua distribuição.

## Interpretação de Resultados

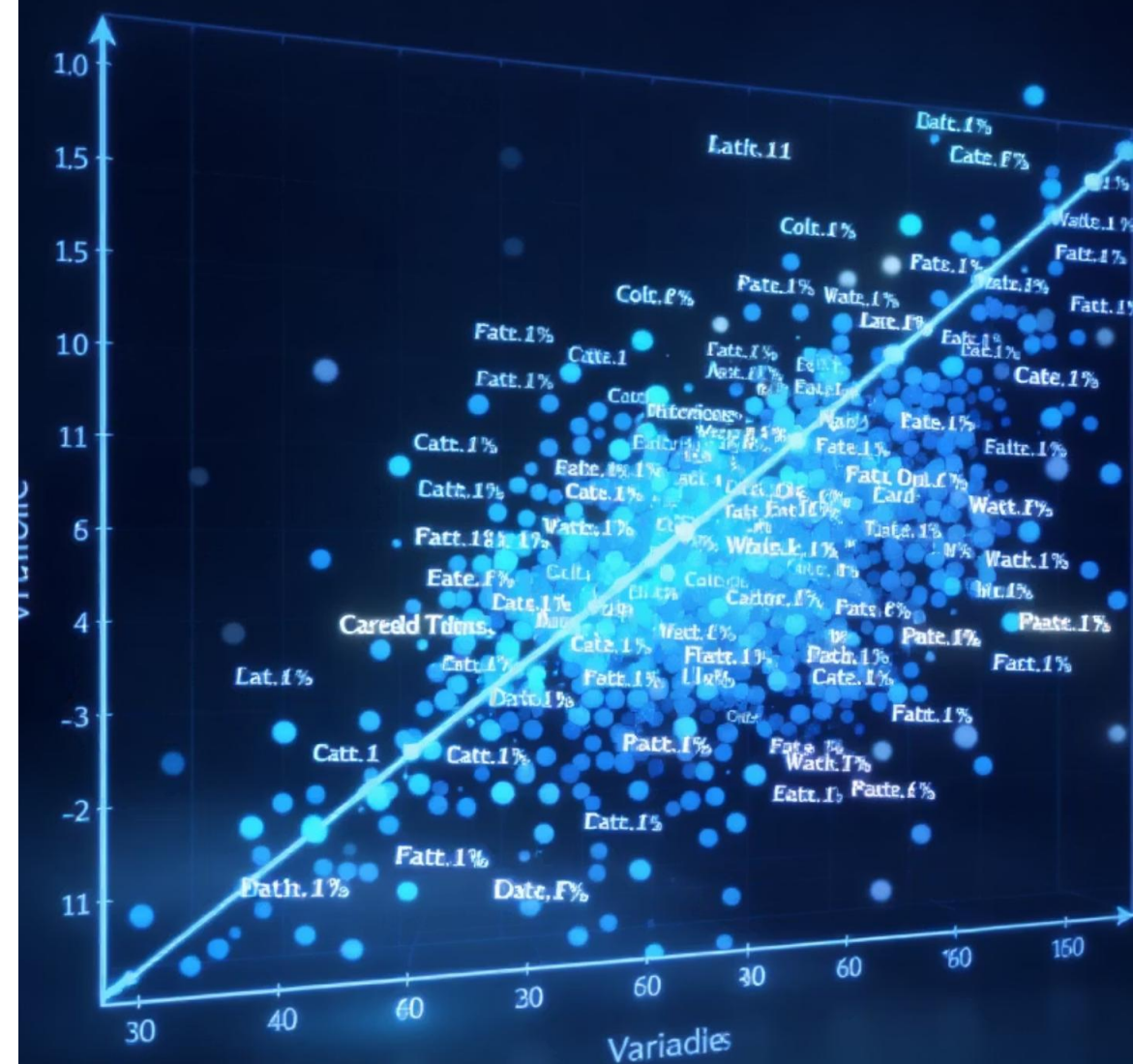
A forma da distribuição afeta a probabilidade de valores extremos e pode impactar significativamente a interpretação de resultados e intervalos de confiança.



# Estatísticas Descritivas Multivariadas

As estatísticas multivariadas analisam relações entre duas ou mais variáveis simultaneamente, permitindo identificar padrões de associação.

As principais medidas multivariadas incluem a correlação e a covariância, que quantificam a força e direção da relação linear entre variáveis.



## Covariância

**Definição:** A covariância mede como duas variáveis variam conjuntamente em relação às suas médias.

**Fórmula:**

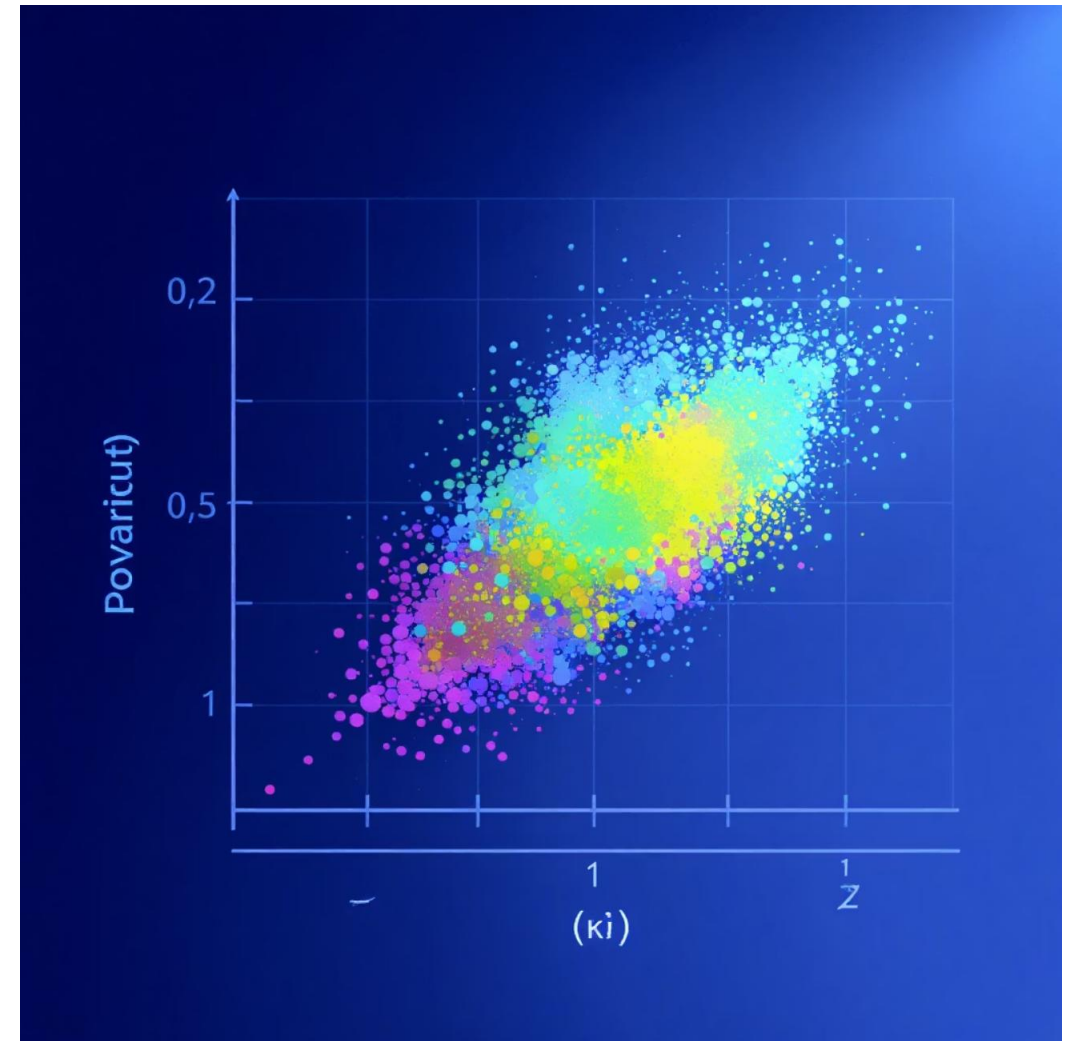
$$\sigma_{xy} = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{n}$$

**Interpretação:**

- **Covariância > 0:** relação positiva (variáveis tendem a aumentar/diminuir juntas)
- **Covariância < 0:** relação negativa (uma variável aumenta quando a outra diminui)
- **Covariância ≈ 0:** ausência de relação linear

**Limitações:**

- Dependente da escala das variáveis
- Difícil de interpretar em termos absolutos



A covariância indica a direção da relação, mas sua magnitude é difícil de interpretar por depender da escala das variáveis.

# Correlação de Pearson

**Definição:** O coeficiente de correlação de Pearson é uma versão padronizada da covariância, que mede a força e direção da relação linear entre duas variáveis.

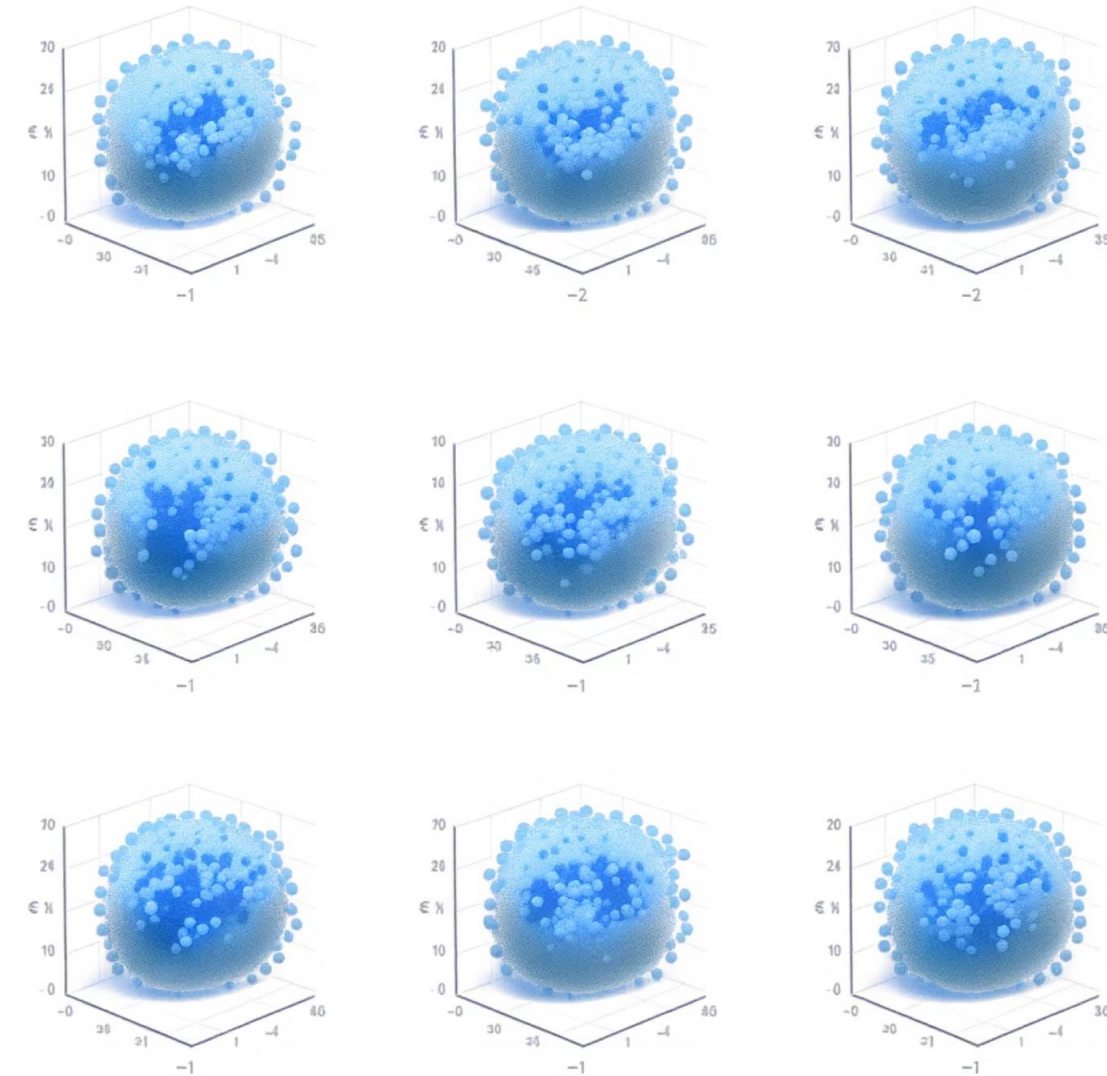
**Fórmula:**

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum (x_i - \mu_x)^2 \sum (y_i - \mu_y)^2}}$$

**Características:**

- Varia entre -1 (correlação negativa perfeita) e +1 (correlação positiva perfeita)
- Independente da escala das variáveis
- Mede apenas relações lineares
- Sensível a outliers

A correlação de Pearson facilita a comparação da força de diferentes relações por ser padronizada entre -1 e +1.



# Interpretação da Correlação de Pearson



## Correlação Forte Positiva

$r$  entre 0,7 e 1,0

Indica forte tendência de ambas as variáveis aumentarem ou diminuírem juntas, quase em relação linear perfeita.



## Correlação Moderada Positiva

$r$  entre 0,3 e 0,7

Sugere tendência moderada das variáveis se moverem na mesma direção, com alguma dispersão nos dados.



## Correlação Fraca ou Nula

$r$  entre -0,3 e 0,3

Indica pouca ou nenhuma relação linear entre as variáveis. Podem existir relações não lineares não detectadas por este coeficiente.

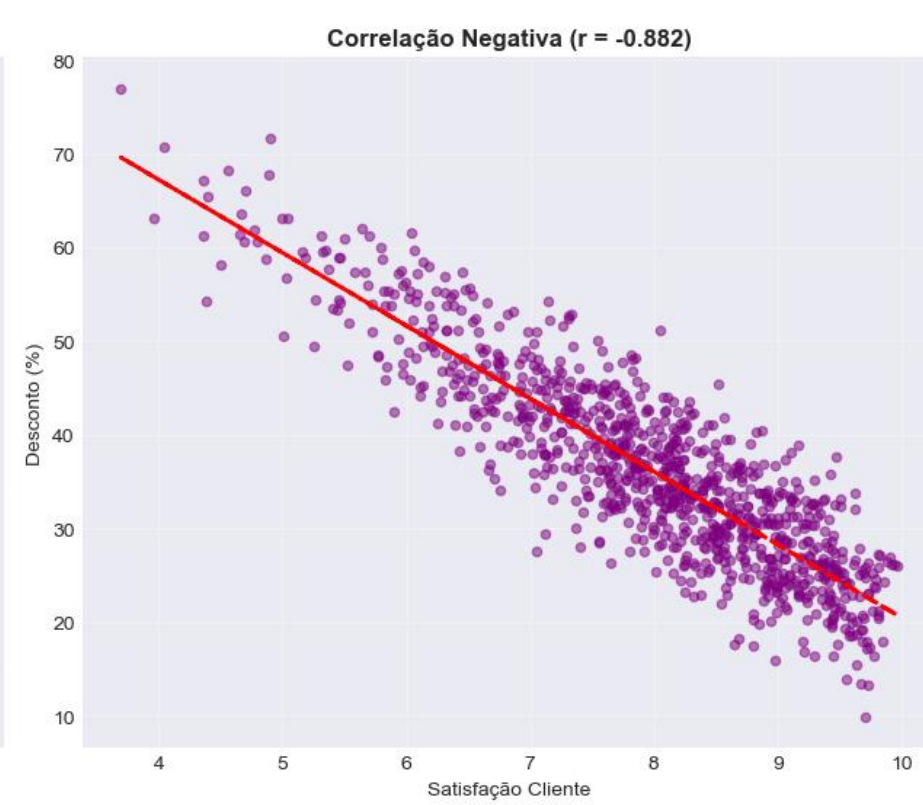
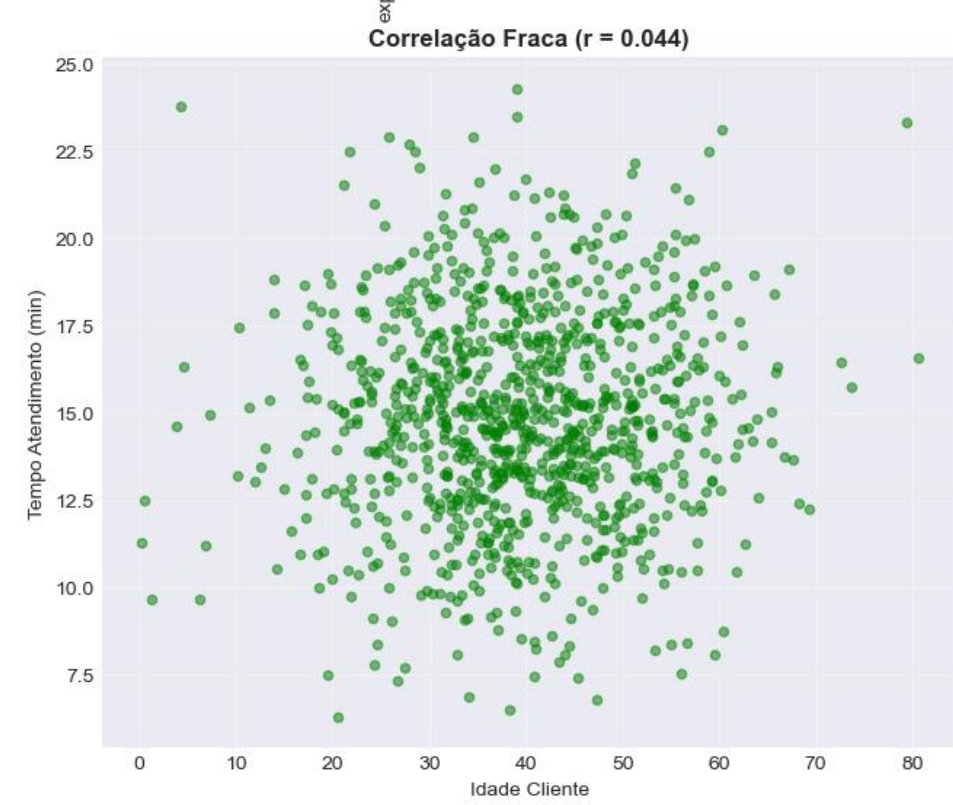
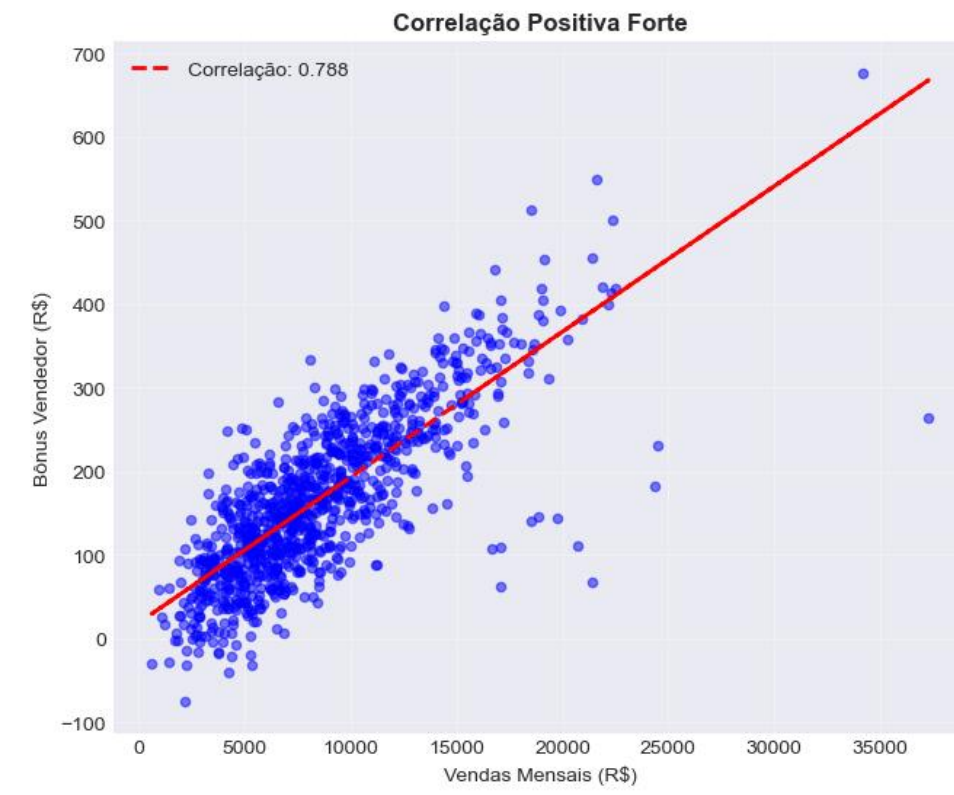
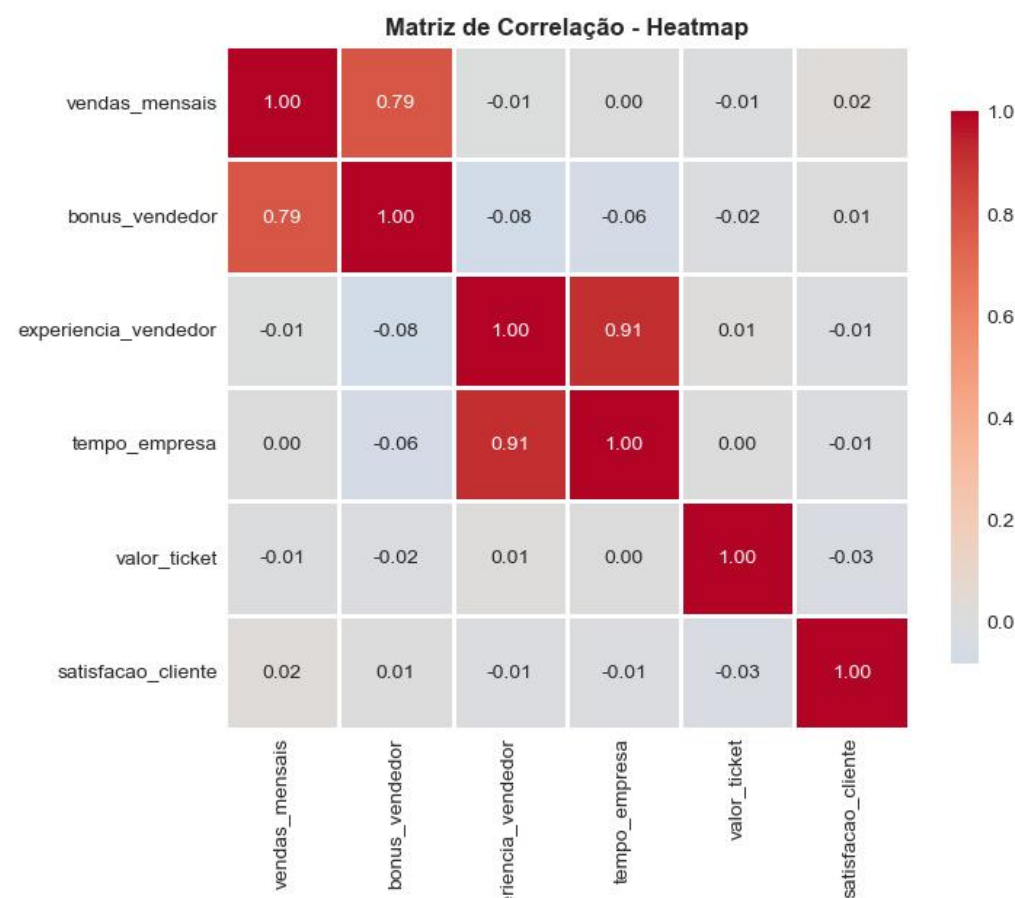


## Correlação Negativa

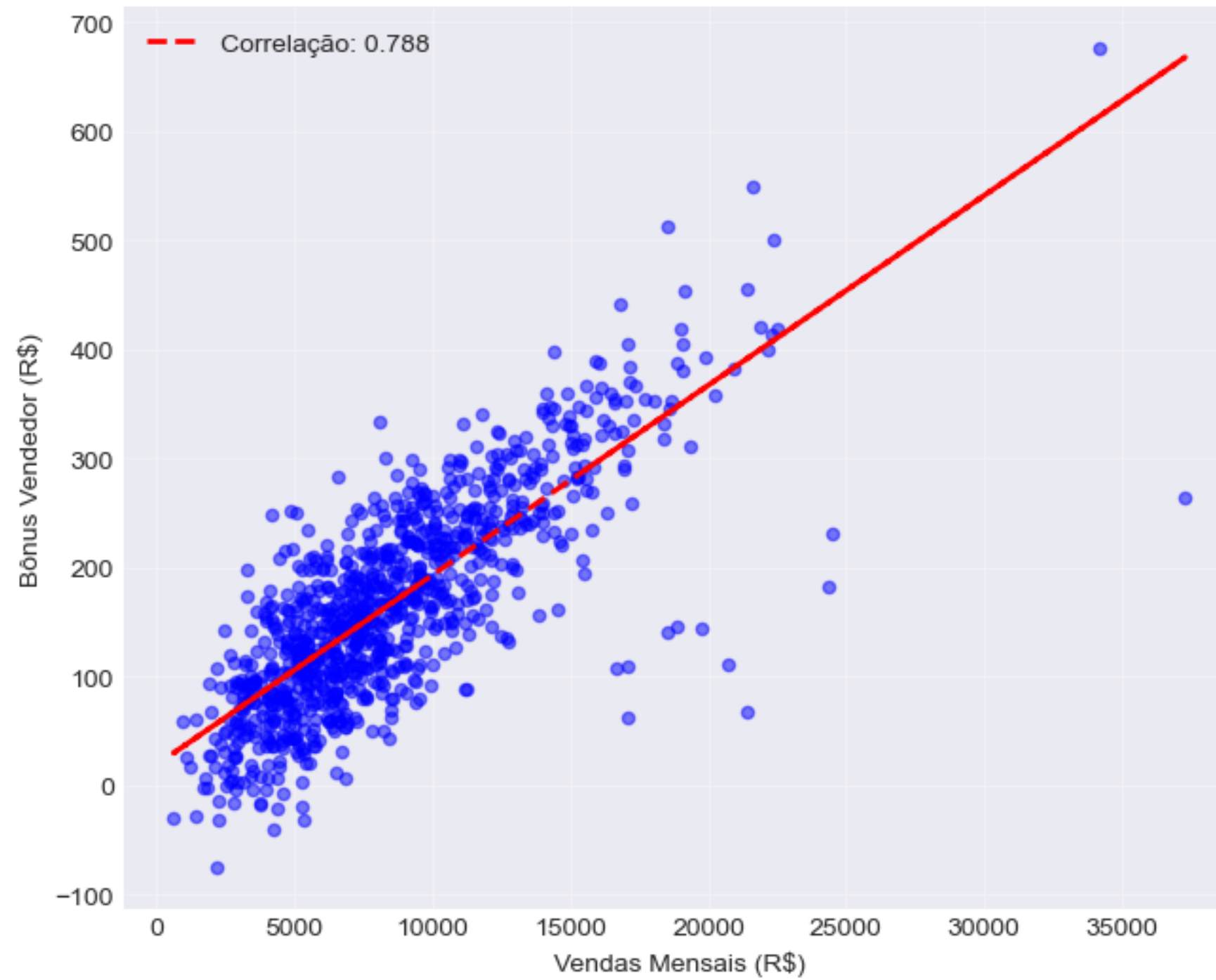
$r$  entre -1,0 e -0,3

Mostra tendência de uma variável aumentar enquanto a outra diminui. Quanto mais próximo de -1, mais forte é esta relação inversa.

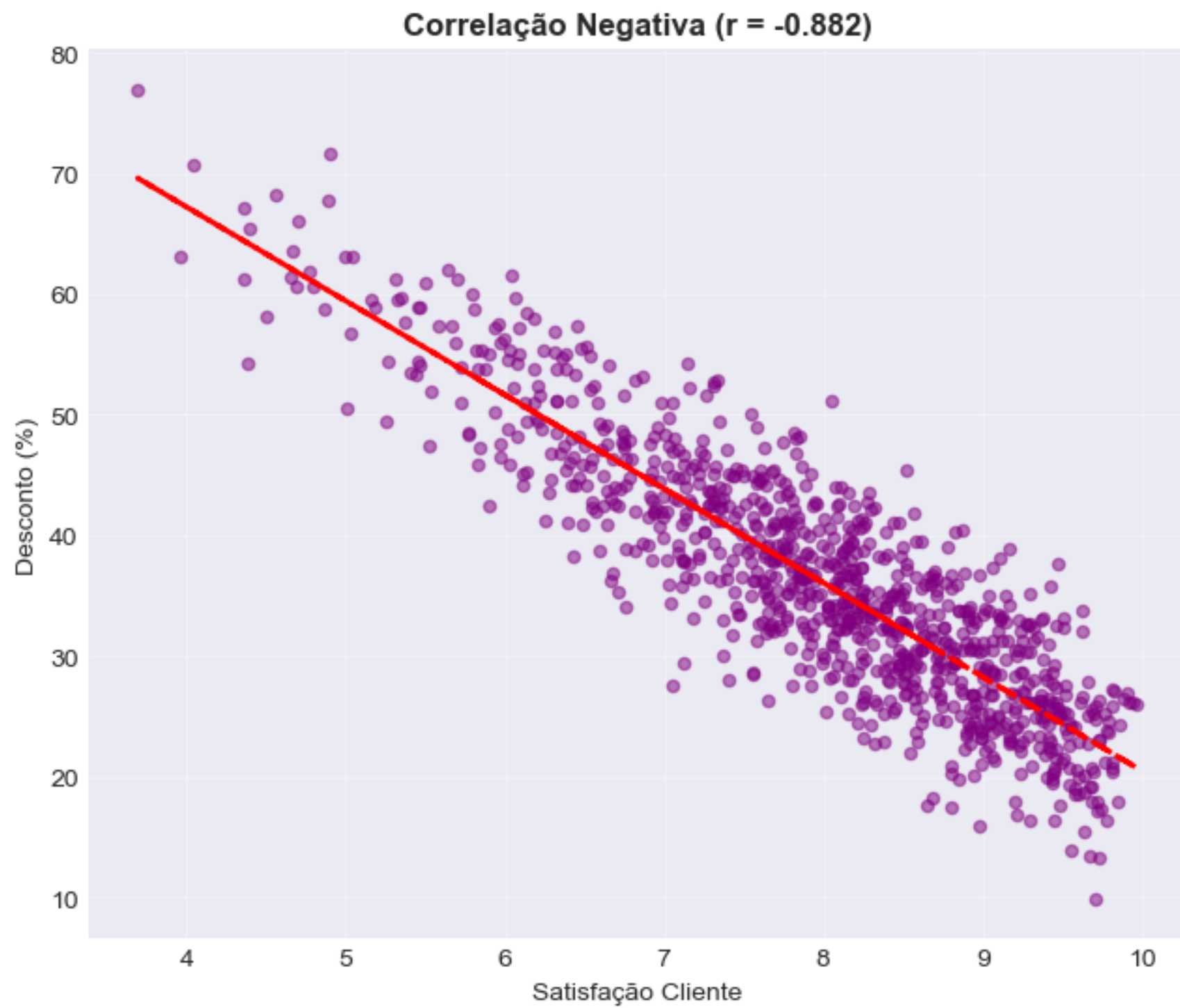




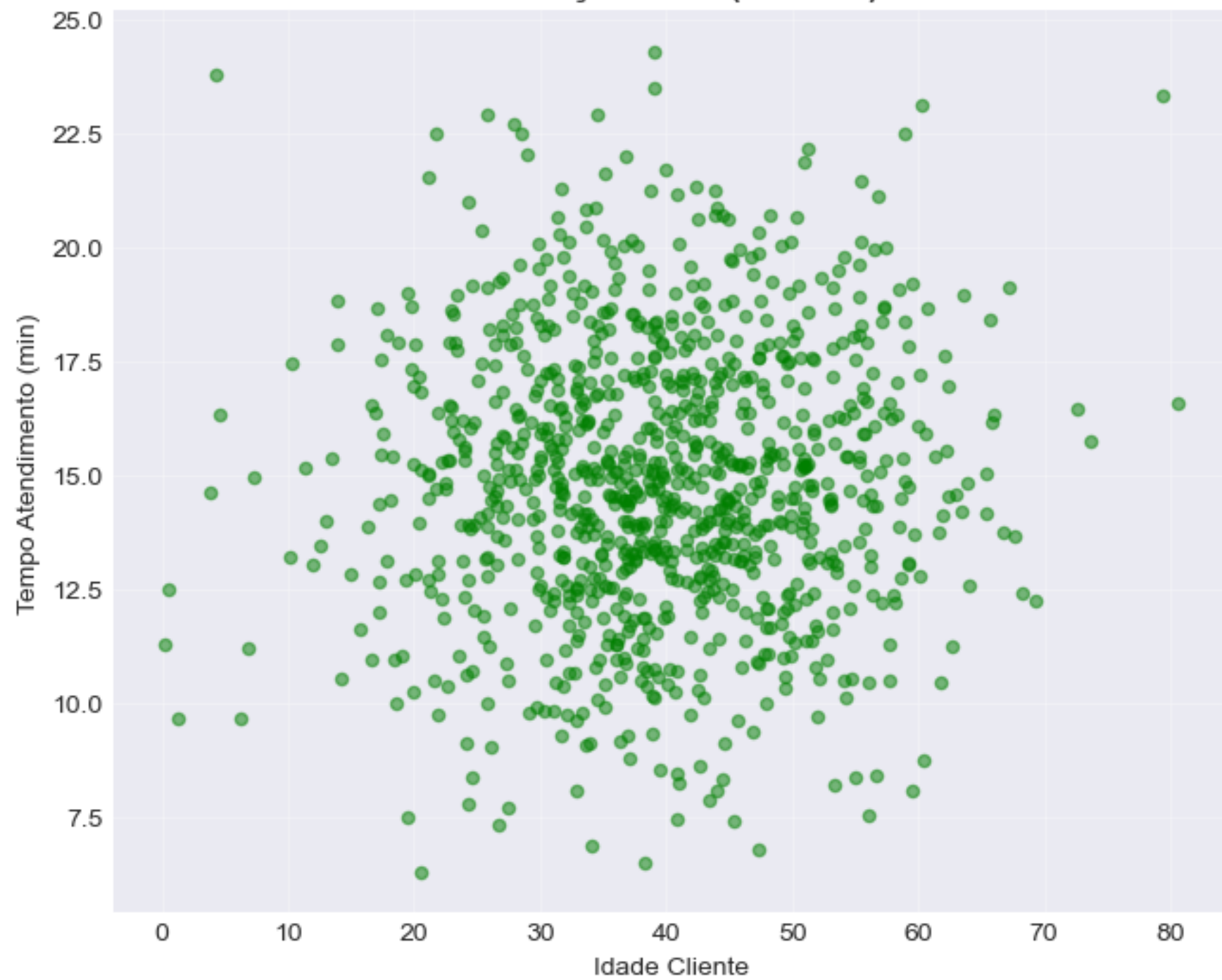
**Correlação Positiva Forte**







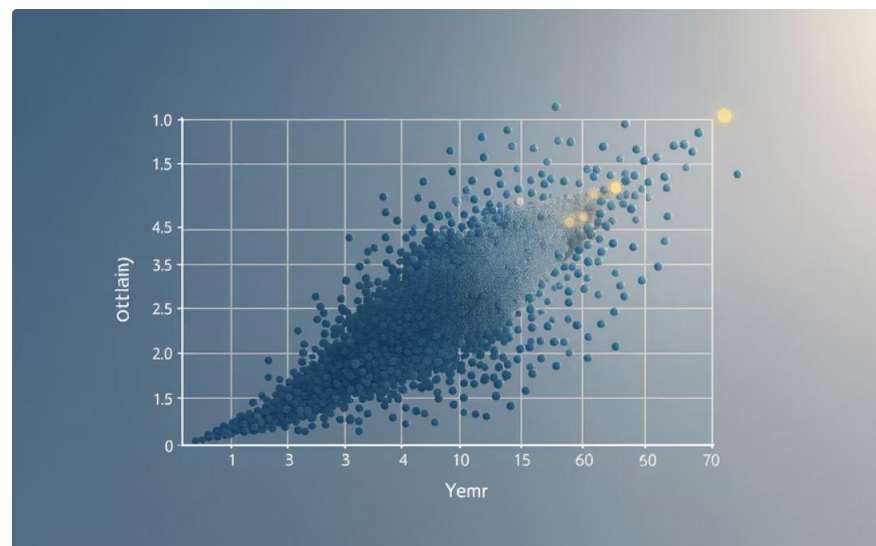
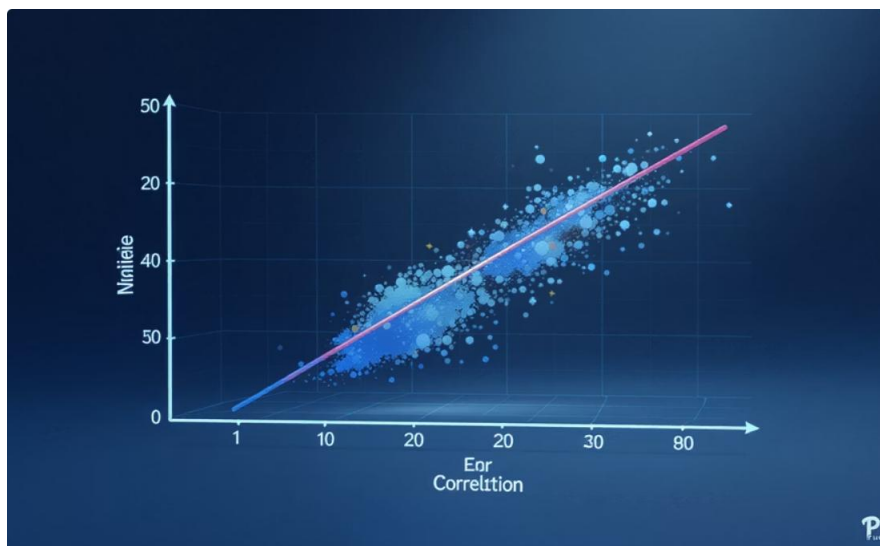
Correlação Fraca ( $r = 0.044$ )



Matriz de Correlação - Heatmap



# Limitações da Correlação



## Não Detecta Relações Não Lineares Sensibilidade a Outliers

A correlação de Pearson mede apenas a força da relação linear. Relações não lineares fortes podem apresentar correlação próxima a zero.

Valores atípicos podem distorcer significativamente o valor da correlação, levando a conclusões errôneas sobre a relação entre variáveis.

## Correlação $\neq$ Causalidade

A existência de correlação não implica em relação causal. Variáveis podem estar correlacionadas devido a um terceiro fator comum.

É fundamental complementar a análise de correlação com visualizações de dados e considerar o contexto para uma interpretação adequada.

# Resumo: Estatística Descritiva na Análise Exploratória



A estatística descritiva fornece as ferramentas fundamentais para transformar dados brutos em informações significativas, permitindo identificar padrões, anomalias e relações que orientarão análises mais avançadas.

# RESUMO ESTATÍSTICO COMPLETO DO DATASET

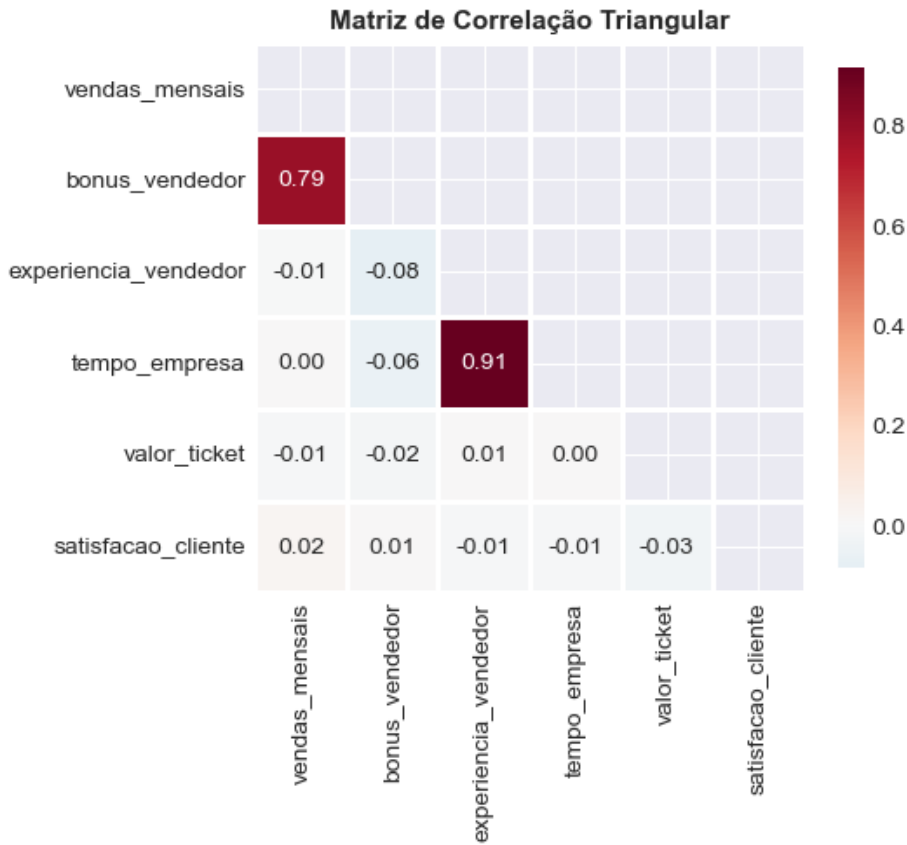
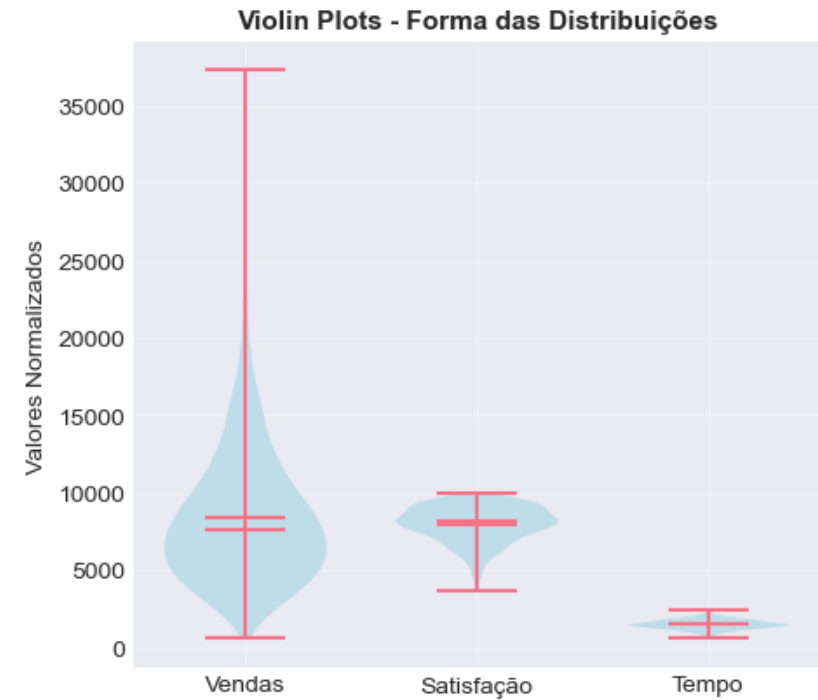
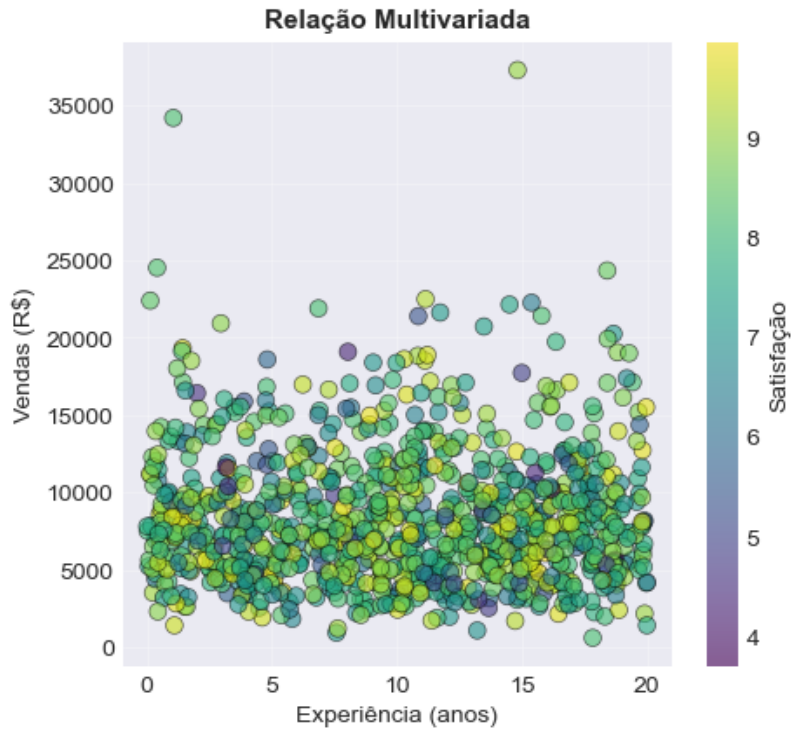
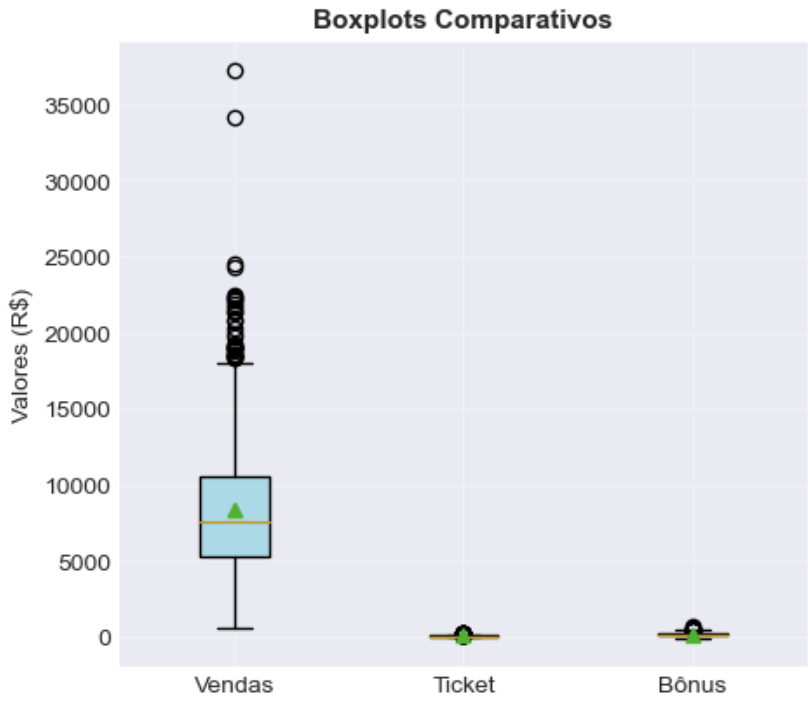
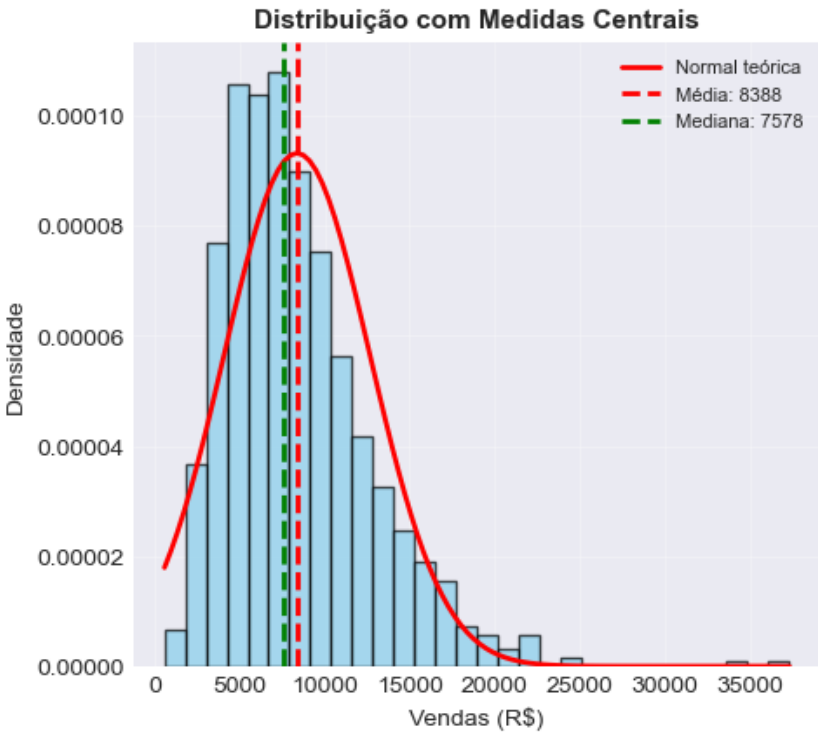
## Estatísticas Descritivas Completas:

	count	mean	std	min	25%	50% \
vendas_mensais	1000.0	8387.65	4284.65	606.83	5335.02	7577.56
satisfacao_cliente	1000.0	7.97	1.18	3.69	7.28	8.13
tempo_atendimento	1000.0	14.89	3.05	6.30	12.93	14.83
numero_produtos	1000.0	3.96	1.72	1.00	3.00	4.00
valor_ticket	1000.0	61.13	30.72	7.68	39.30	54.85
idade_cliente	1000.0	39.33	11.78	0.15	31.72	39.07
experiencia_vendedor	1000.0	10.10	5.69	0.01	5.15	10.27
bonus_vendedor	1000.0	165.01	94.60	-75.13	95.39	156.02
tempo_empresa	1000.0	8.28	4.87	0.01	4.01	8.55
desconto	1000.0	36.27	10.39	9.93	28.64	35.17

	75%	max	CV%	Assimetria	Curtose
vendas_mensais	10541.00	37289.21	51.08	1.32	3.60
satisfacao_cliente	8.90	9.97	14.74	-0.71	0.24
tempo_atendimento	16.99	24.29	20.48	0.07	-0.03
numero_produtos	5.00	11.00	43.37	0.52	0.27
valor_ticket	74.70	234.41	50.26	1.43	2.99
idade_cliente	47.32	80.53	29.95	-0.05	0.24
experiencia_vendedor	15.05	20.00	56.36	-0.06	-1.16
bonus_vendedor	224.98	676.23	57.33	0.64	0.93
tempo_empresa	12.01	20.79	58.84	0.10	-1.01
desconto	42.65	76.97	28.63	0.55	0.20



ANÁLISE EXPLORATÓRIA DE DADOS - PAINEL COMPLETO



Resumo Estatístico - Vendas Mensais

Medida	Valor
Média	R\$ 8387.65
Mediana	R\$ 7577.56
Desvio Padrão	R\$ 4284.65
CV%	51.1%
Assimetria	1.32
Curtose	3.60
Q1	R\$ 5335.02
Q3	R\$ 10541.00
IQR	R\$ 5205.98
Outliers	30 detectados