
# Análise Exploratória de Dados (EDA)

Ciência de dados II

**Professor: Gabriel Machado Lunardi**
gabriel.lunardi@ufsm.br

# Relembrando...

- Dados: é uma coleção de **objetos** e seus **atributos**;
- Um **atributo** é uma propriedade ou **característica** de um objeto
  - Exemplos: a cor dos olhos de uma pessoa, a sua temperatura, etc.
  - Atributos também são conhecidos como variáveis, campos, características.
- Uma coleção de atributos descreve um objeto
  - Objetos também são conhecidos como registro, caso, exemplo, entidade, instância.

**Atributos/variáveis**

**Objetos**

| Nome | Idade | Renda | Pagador |
|---|---|---|---|
| João | <30 | Média | Bom |
| Ana | 41..50 | Alta | Bom |
| Pedro | 41..50 | Alta | Bom |
| Maria | 41..50 | Baixa | Ruim |
| Paulo | <30 | Baixa | Ruim |
| Aldo | >60 | Alta | Ruim |

# Tipos de variáveis

- **Atributo qualitativo**
  - Nominal: nome, cor dos olhos, estado civil, tipo sanguíneo.
  - Ordinal: classificação em uma competição (1º, 2º, etc), nível de escolaridade, satisfação com um serviço prestado.
- **Atributo quantitativo**
  - Discreto: idade, número de filhos, número de carros na garagem.
  - Contínuo: peso, altura, volume, temperature.

# EDA

É comumente empregada em qualquer processo de análise de dados, **antes** de aplicar técnicas estatísticas mais complexas ou modelos preditivos.

- ✓ **Entender os Dados**: Familiarizar-se com o conjunto de dados, incluindo sua estrutura, tipos de variáveis e distribuição.
- ✓ **Identificar Padrões e Relações**: Detectar correlações, tendências e padrões que possam existir entre as variáveis.
- ✓ **Detectar Anomalias**: Identificar outliers, valores faltantes e erros nos dados.
- ✓ **Preparar os Dados**: Identificar a necessidade de transformações, normalizações ou limpeza dos dados para análises subsequentes.

# EDA

- ✓A **média aritmética e o desvio padrão** são medidas muito utilizadas.
- ✓Porém, essas medidas descrevem de forma ótima **distribuições de frequências simétricas**.
- ✓Numa distribuição assimétrica seus valores são bastante afetados pelos valores discrepantes – outliers (não são medidas resistentes).

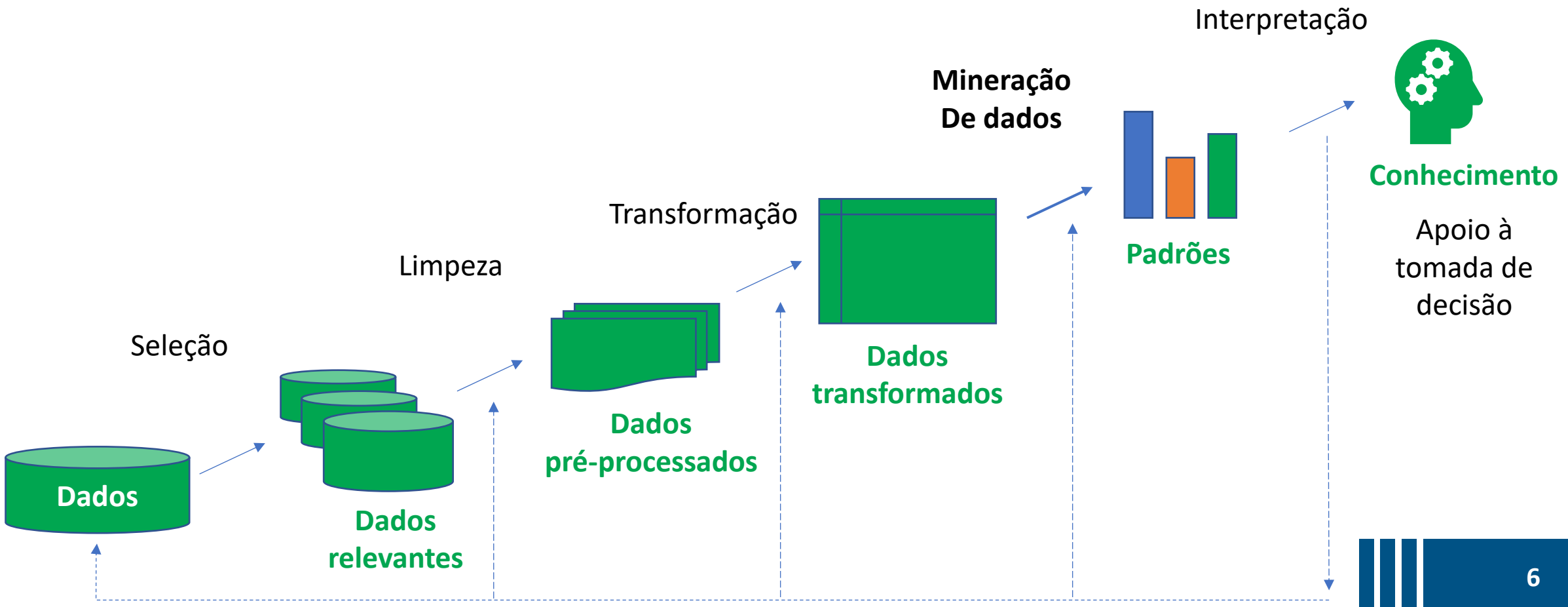
Em 1970, **John Turkey** propôs técnicas que contornavam esses problemas. O conjunto dessas técnicas recebeu a denominação de **Análise Exploratória de Dados.**

**As mais elementares, são:**

- Resumo dos cinco números
- Box-plot

# EDA no processo de KDD

"É um processo de várias etapas, não trivial, **interativo** e **iterativo**, para a identificação de **padrões** válidos, novos e potencialmente úteis a partir de um grande conjunto de dados" (FAYYAD, 1996).



# EDA

A técnica pode ser classificada de acordo com a quantidade de variáveis que analisa ao mesmo tempo e pode utilizar recursos visuais ou não.

- ✓ **EDA Uni variada**: usa uma única variável para entender sua distribuição e características.
- ✓ **EDA Bivariada**: usa a relação entre duas variáveis para identificar correlações e padrões.
- ✓ **EDA Multivariada**: explora interações entre três ou mais variáveis para entender padrões complexos.

# EDA univariada – principais recursos estatísticos

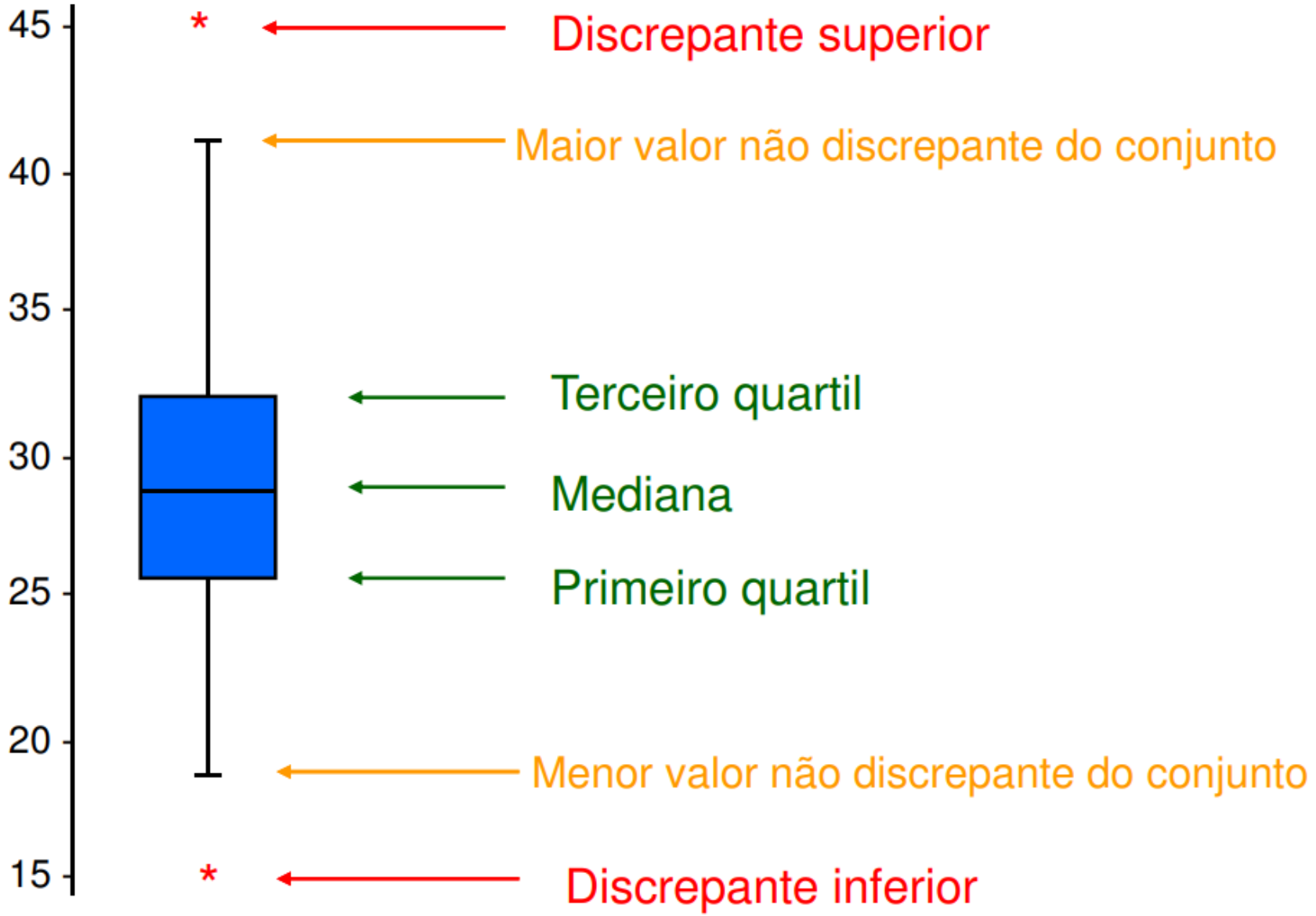
1. **Distribuição de Frequência**:
    1. **Histogramas**: Para variáveis contínuas.
    2. **Gráficos de Barras**: Para variáveis categóricas.
2. **Medidas de Tendência Central**:
    1. **Média**: Valor médio dos dados.
    2. **Mediana**: Valor que divide os dados ao meio.
    3. **Moda**: Valor mais frequente nos dados.
3. **Medidas de Dispersão**:
    1. **Variância e Desvio Padrão**: Medem a dispersão dos dados em torno da média.
    2. **Amplitude**: Diferença entre o maior e o menor valor.
4. **Identificação de Outliers**:
    1. **Boxplots**: Para visualizar a dispersão e identificar outliers.
    2. **Z-Scores**: Identifica pontos que estão a uma certa distância da média em termos de desvios padrão.

# EDA univariada – Box-plot

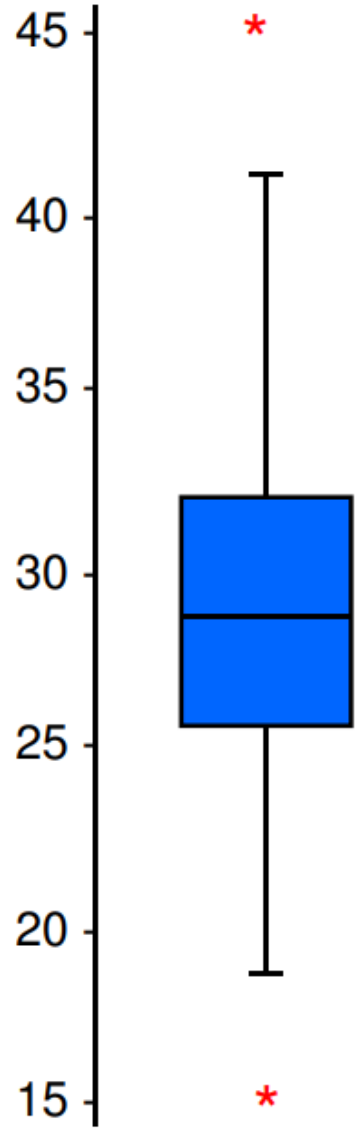
A informação dada pelo resumo de cinco números pode ser apresentada na forma de um box-plot que agrega uma série de informações sobre a distribuição:

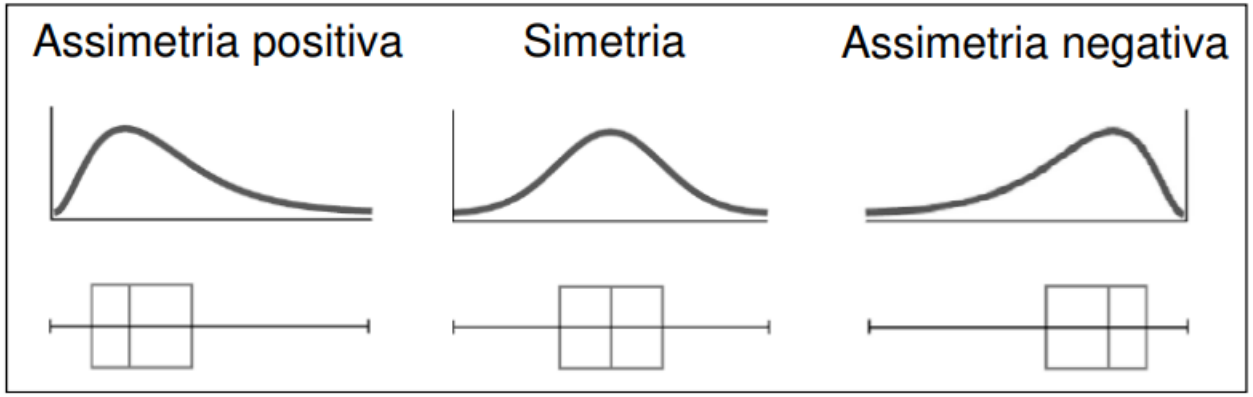
- ✓ posição
- ✓ dispersão
- ✓ assimetria
- ✓ caudas
- ✓ dados discrepantes (outliers)

# Box-plot


45
40
35
30
25
20
15
*
*
Discrepante superior
Maior valor não discrepante do conjunto
Terceiro quartil
Mediana
Primeiro quartil
Menor valor não discrepante do conjunto
Discrepante inferior

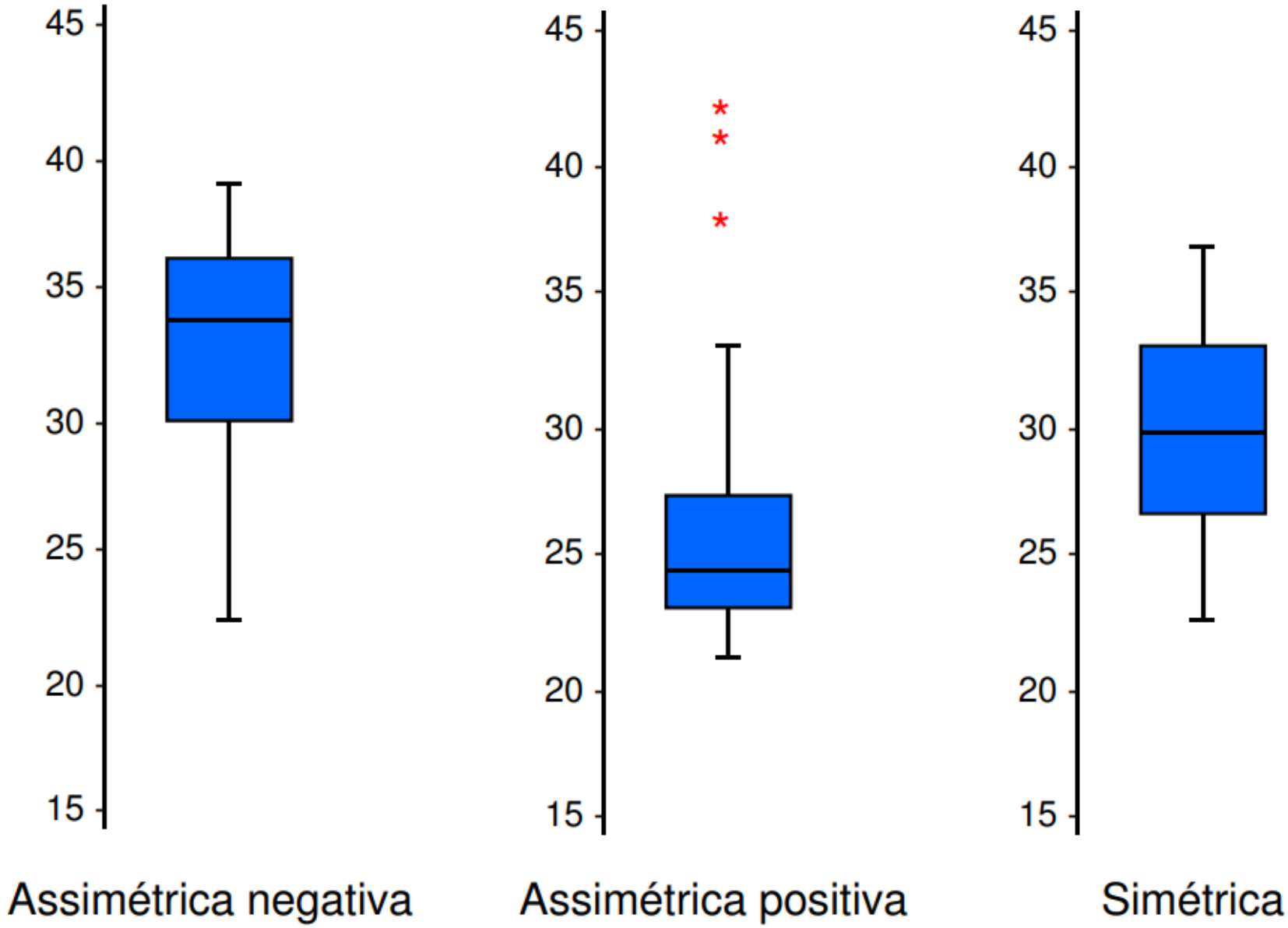
# Box-plot




- A posição central dos valores é dada pela mediana e a dispersão pela amplitude interquartílica.
- As posições relativas da **mediana** e dos **quartis** e o formato dos bigodes dão uma noção da simetria e do tamanho das caudas da distribuição.

# Box-plot exemplos



# EDA bivariada – principais técnicas

1. **Scatterplots**: Para visualizar a relação entre duas variáveis contínuas.

2. **Gráficos de Barras Agrupados**: Para comparar categorias de duas variáveis qualitativas.

3. **Correlação**:
    1. **Coeficiente de Pearson**: Mede a correlação linear entre duas variáveis contínuas.
    2. **Matriz de Correlação**: Para visualizar correlações entre múltiplas variáveis.

# EDA multivariada – principais técnicas

1. **Gráficos de Dispersão Múltiplos (Pair Plots)**: Para visualizar relações entre várias variáveis ao mesmo tempo.

2. **Heatmaps**: Para visualizar matrizes de correlação entre múltiplas variáveis.

1. **Gráficos 3D**: Para visualizar relações entre três variáveis.

# Limpeza de dados: introdução

- Tem por objetivo garantir a **qualidade dos dados**. Quanto maior for a qualidade dos dados, melhor será o resultado final.

- Dados reais são naturalmente confusos, ruidosos e de baixa qualidade, afetando os esforços do processamento de dados.

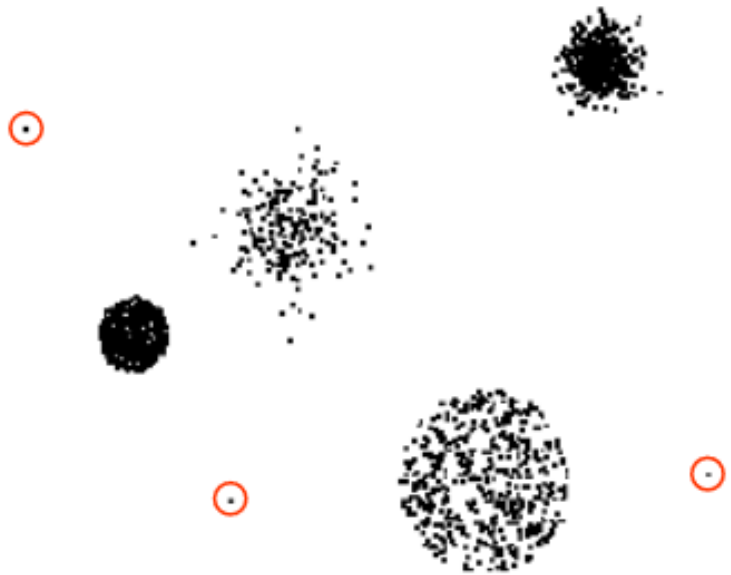
- Exemplo: Usar dados de baixa qualidade para detectar o risco de crédito aos clientes de um banco.
  - Alguns “bons” clientes podem ter o empréstimo negado.
  - Mais empréstimos podem ser concedidos a clientes “ruins”.

**Limpeza de dados**
problemas

- ✓Ruídos (*outliers*)
- ✓Dados errados
- ✓Dados falsos
- ✓Valores faltantes
- ✓Dados duplicados

# Limpeza de dados: problemas

- ***Outliers***
  - são objetos de dados com características que são consideravelmente diferentes da maioria dos outros objetos de dados no conjunto de dados.
  - Exemplo 1: A média de alturas de 10 mil pessoas em que algumas observações apresentam valor maior que 2,5m (possível outlier?)
  - Exemplo 2: Análise de agrupamento
  - O que fazer?
    - Próxima aula!



# Limpeza de dados: problemas

- **Dados faltantes *(missing values)***
  - Valores de atributos para um objeto que não foram preenchidos por alguma razão.
  - Exemplo 1 : a pessoa se negou a informar a sua idade ou peso.
  - Exemplo 2: atributos não são aplicáveis a todos os casos. Salário anual não é aplicável para pessoas que são crianças.
  - Exemplo 3: Alguns algoritmos de aprendizado de máquina (**árvores de decisão**), utilizados na fase de mineração, **não lidam bem** com valores ausentes por exemplo.

# Limpeza de dados: problemas

- **Dados faltantes *(missing values)***
  - O que fazer? (Possibilidades a depender do objetivo e dos dados)
    - Eliminar esses dados ou variáveis (atributos) ou ignorá-los durante a análise
    - Estimar os valores faltantes com base em outros
      - Complemento manual
      - Complemento com valor constante global: ex: “desconhecido”
      - Complementar com o valor mais provável
      - Complementar com o valor médio do atributo

# Limpeza de dados: problemas

- **Dados duplicados**
  - Conjuntos de dados podem incluir dados que são duplicados. Isso acontece, muitas vezes, ao juntar dados de **diferentes fontes**.
    - Combinar dados de duas ou mais tabelas em um BD relacional (joins) com SQL
    - Lembre-se disso quando utilizar a função merge() do Pandas
  - O que fazer? (Possibilidades a depender do objetivo e dos dados)
    - Eliminar as duplicatas

# Limpeza de dados

- ✓ Exemplo: conjunto de dados Titanic

| Atributo | Descrição | Tipo |
|---|---|---|
| survival | Se sobreviveu ao acidente | Binário {0 – Não, 1 – Sim} |
| pclass | Classe do passageiro | Categórico {1, 2, 3} primeira, segunda, terceira |
| name | Nome do passageiro | Categórico |
| Sex | Gênero do passageiro | Categórico |
| age | Idade do passageiro | Numérico discreto |
| sibsp | Soma do número de irmãos, cunhados e cônjuge | Numérico discreto |
| Parch | Soma do número de pais e filhos | Numérico discreto |
| ticket | Número da passagem | Numérico discreto |
| Fare | Valor da passagem | Numérico contínuo |
| Cabin | Número da cabine | Numérico discreto |
| Embarked | Porto de embarque | Categórico {C, Q, S} Cherbourg; Queenstown; Southampton |

# Atividade em aula

- ✓ Vamos juntos remover outliers e valores faltantes do conjunto de dados Titanic.

- ✓ Jupyter Notebook no Google Colab.

- ✓ Utilizando algum dataset que tenha chamado a sua atenção nas aulas anteriores, procure valores faltantes e outliers utilizando o conhecimento adquirido nesta aula.