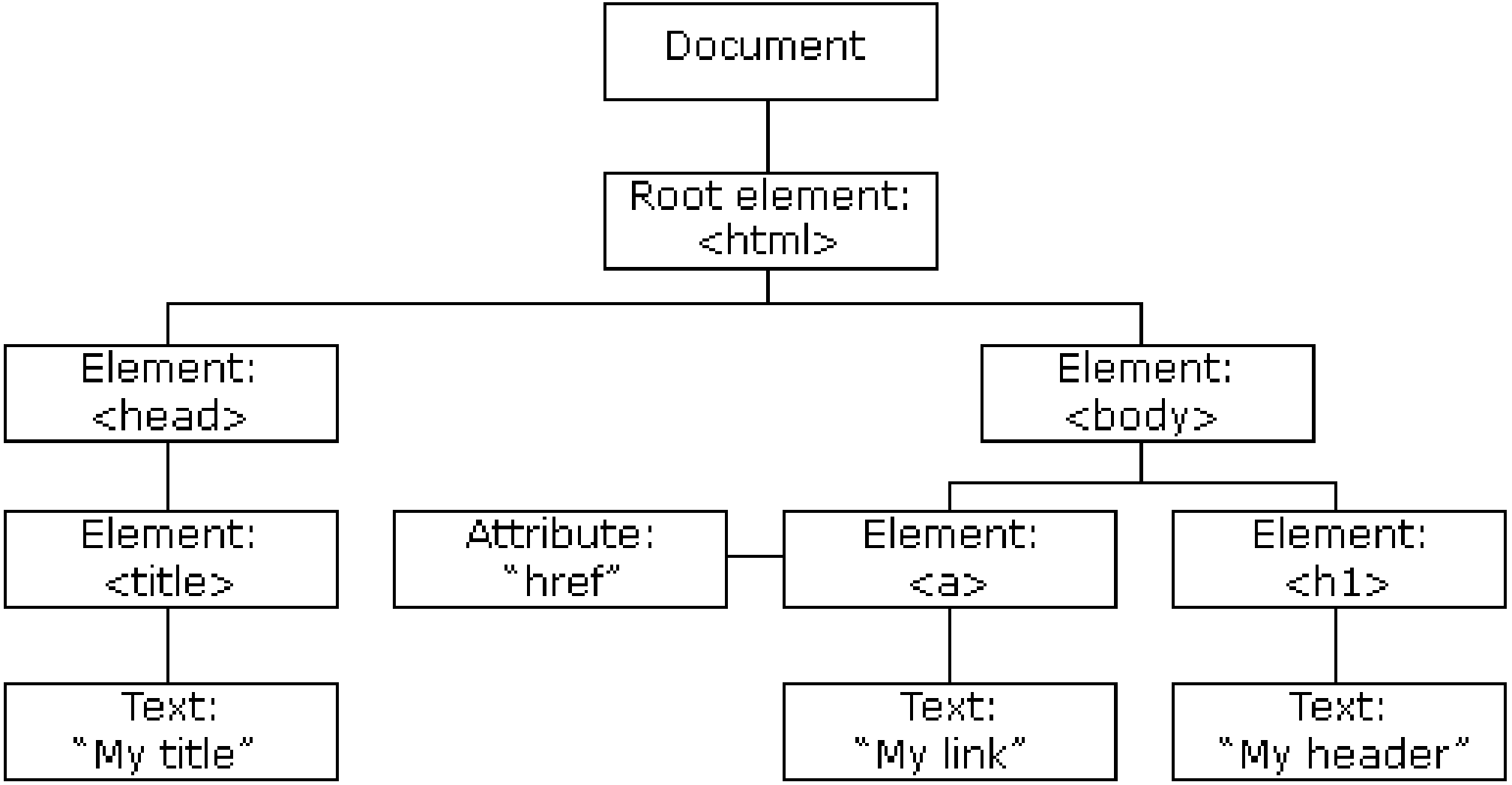
Pensamento Computacional e Lógica de Programação

Prof. Dr. Álvaro Campos Ferreira

HTML e suas tags

O DOM – Document Object Model



Web Scraping

Web Scraping

Os 10 sites mais raspados da Web:

- Mercado Livre
- Indeed
- Tripadvisor
- Google
- Páginas amarelas
- Yelp
- Walmart
- eBay
- Amazon

Web Scraping

Para selecionar apenas a tabela que queremos, usamos o argumento match na função read_html().

```
import pandas as pd
url = ‘https://pt.wikipedia.org/wiki/COVID-19’
html = pd.read_html(url,match=’Frequência’)
```

Web Scraping

Dados obtidos da página estão na linguagem HTML, que divide os dados no que chama de tags.

Para explorar as tags de um site, usa-se a ferramenta de desenvolvimento do navegador, F12 no Firefox.

Web Scraping

Uma vez identificadas as tags de interesse, filtramos o resultado:

```
soup.findAll('div')
for tag in soup.findAll('div'):
    print(tag.text)
```

idp

Web Scraping

Seletores e Xpath são duas formas de selecionar o conteúdo de interesse na página.

Tratamento e limpeza de dados

Qualidade de dados

Dimensões da qualidade de dados:

- Completude ou integridade
- Conformidade
- Validade
- Acurácia e precisão

Qualidade de dados

Problemas	Soluções
Erros de dados	Utilizar entrada de dados automatizada, formulários *web* para entrada de dados individuais, com checagem de integridade de dados, menus de localização e botões de opção
Dados duplicados	Redesenhar o modelo de dados e normalizar o banco de dados relacional
Dados comprometidos	Implementar uma abordagem de defesa profunda à segurança de dados
Dados faltando	Tornar campos obrigatórios nos formulários de entrada de dados

Fonte: Adaptado de Turban *et al.* (2013).

Qualidade de dados

Soluções para quando já não se pode corrigir a coleta:

- Descartar o registro ruim
- Atribuir um valor de sentinela
- Atribuir o valor médio ou mais frequente
- Calcular um valor substituto
- Atribuir um valor baseado em valores vizinhos

idp

Valores faltantes

Valores faltantes são representados por **nan** em DataFrames e não são a mesma coisa que valores nulos.

- Faltas (completamente) aleatórias
- Faltas não aleatórias

Outliers

Outliers são valores encontrados fora de um intervalo razoável com relação às outras observações. É um ponto diferente dos demais, não necessariamente um erro. Sua identificação é importante e pode possuir informações interessantes sobre o sistema.

Detecção de outliers

Uma maneira de detectar outliers é verificando se estão diferindo muito dos demais. Uma forma simples é através do intervalo interquartil.

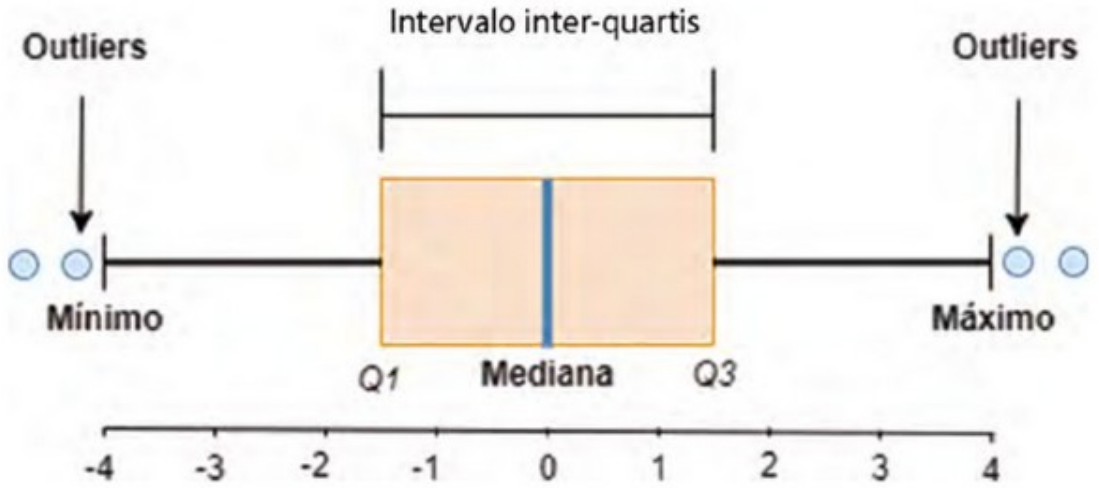


Figura 1. Diferentes partes do *box-plot*.

Detecção de outliers

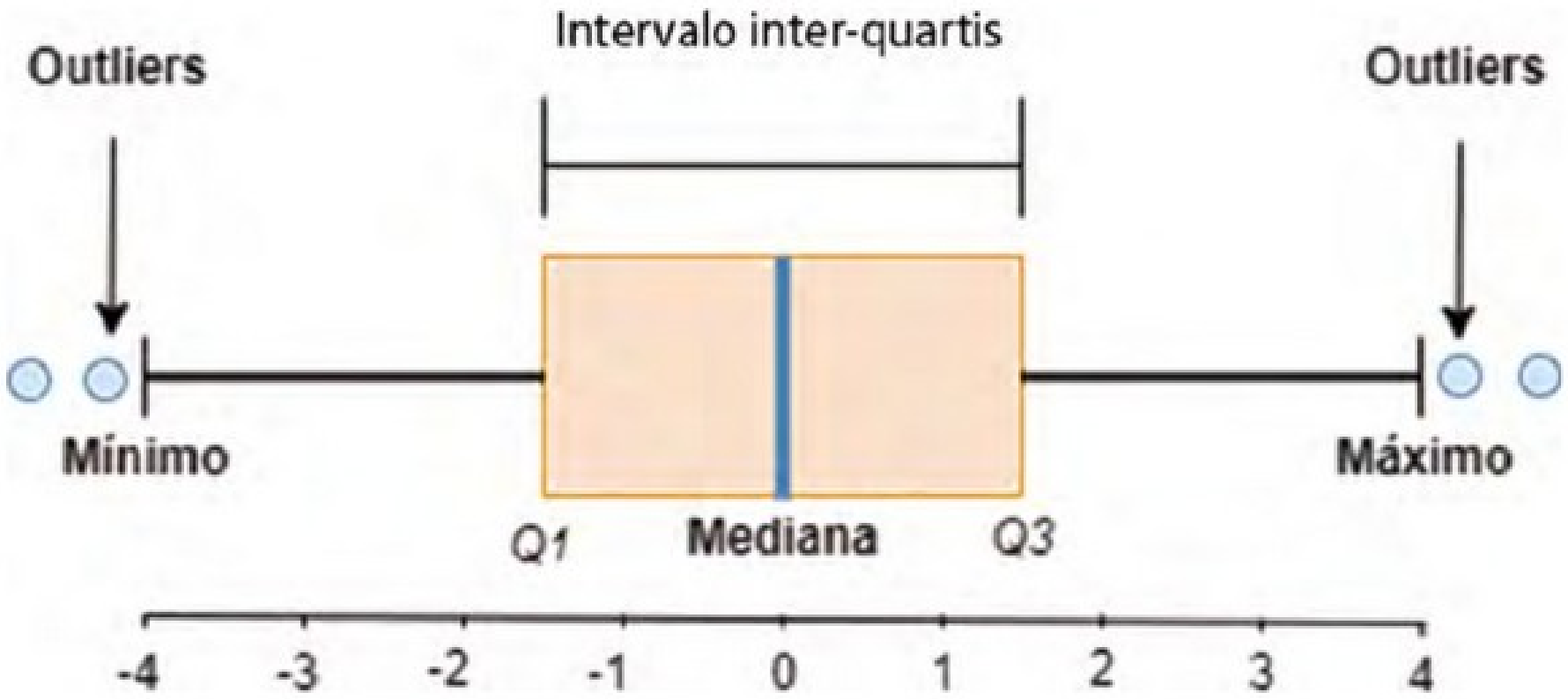


Figura 1. Diferentes partes do *box-plot*.

Limpando dados numéricos e datas

Limpando dados numéricos e datas

Vamos em geral salvar os resultados em novos DataFrames para garantir a integridade dos dados.

- Função split()
- Função strip()
- Compreensões de lista

INSTITUTO BRASILEIRO DE ENSINO,
DESENVOLVIMENTO E PESQUISA