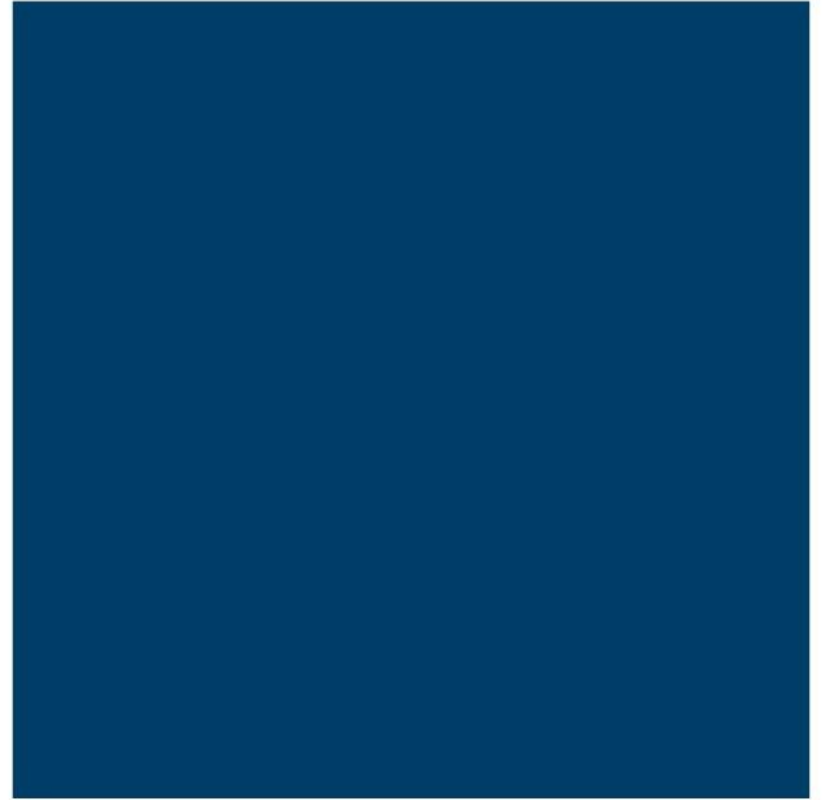


Data science aplicada aos negócios

Prof. Dr. Álvaro Campos
Ferreira

alvaro.ferreira@idp.edu.br

Vamos construir um conjunto de dados



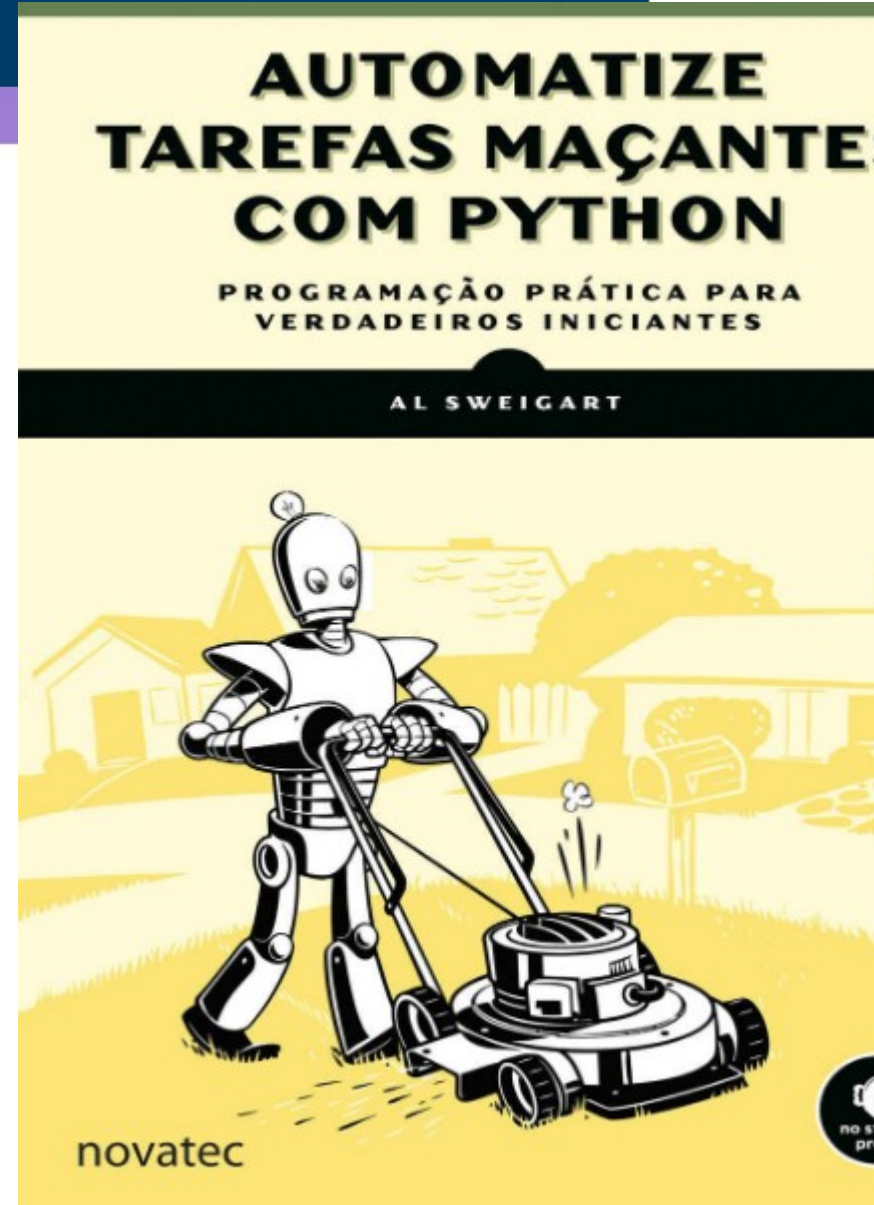
<https://forms.gle/rn1eZGAatVk6LNpk6>

Sugestões de leitura

Sugestões de leitura

Aprenda e relembre Python, sua sintaxe e seus usos.

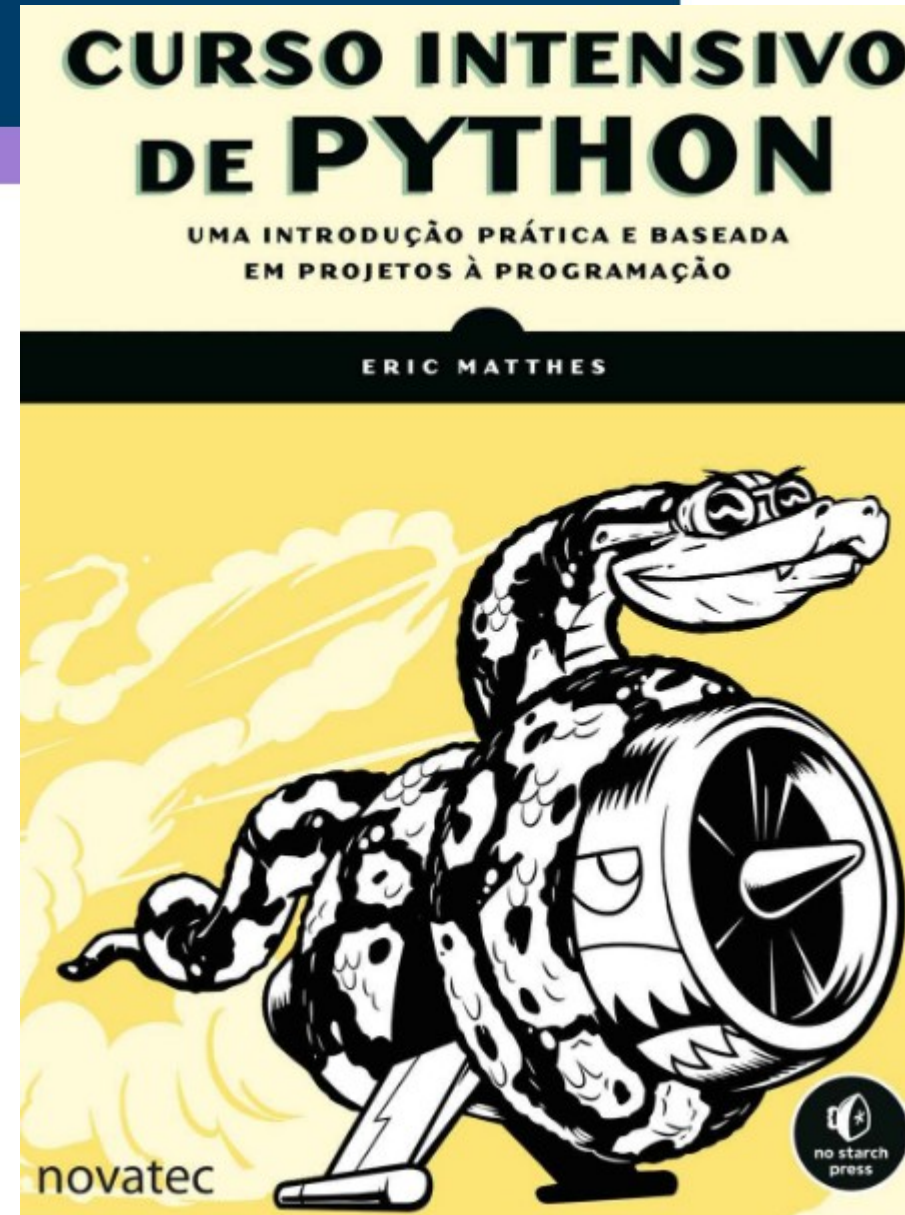
- Automatize as tarefas maçantes com Python (Al Sweigart, 2015)
- Curso Intensivo de Python (Eric Matthes, 2016)



Sugestões de leitura

Aprenda e relembre Python, sua sintaxe e seus usos.

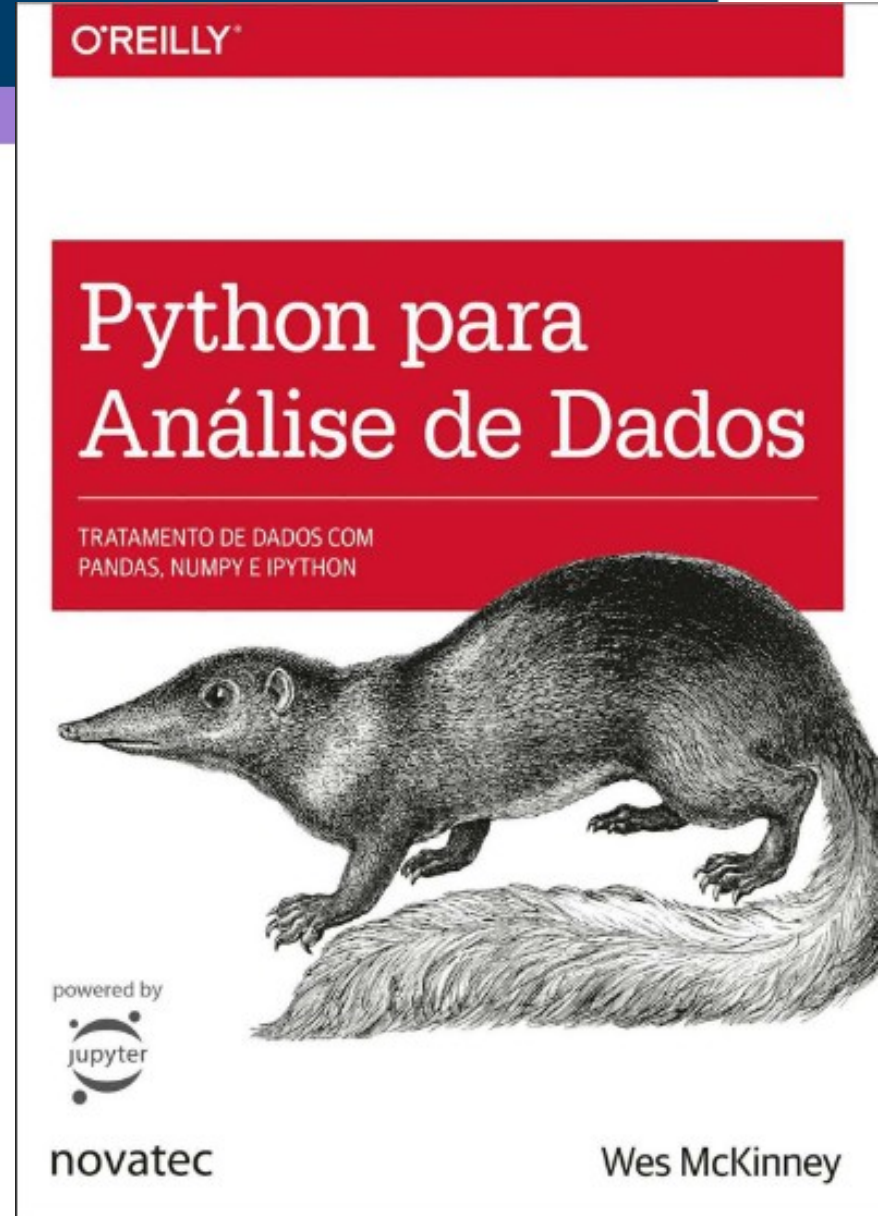
- Automatize as tarefas maçantes com Python (Al Sweigart, 2015)
- Curso Intensivo de Python (Eric Matthes, 2016)



Sugestões de leitura

Relembre estatística e dados com Python.

- Python para análise de dados (Wes McKinney, 2018)



Sugestões de leitura

Relembre estatística e dados com Python.

- Preparação e análise exploratória de dados (Rafael Ferreira, 2021)
- Visualização de dados (Alessandra Milani, 2020)
- Data science do zero: noções fundamentais com Python (Joel Grus, 2016)

Análise Exploratória

Análise exploratória

Etapas da estatística descritiva.



Análise descritiva dos dados

Uma descrição óbvia de um conjunto de dados está nos próprios dados.

- Número de observações
- Valores mínimos e máximos
- Ordenação dos valores

Análise descritiva dos dados

Quando temos muitos dados para analisar cada observação, precisamos de novas ferramentas.

- Tendências centrais
- Dispersão
- Correlação

Análise descritiva dos dados

Tendências centrais

- Média
- Mediana

Análise descritiva dos dados

Dispersão

- Amplitude
- Variância
- Desvio-padrão

Análise descritiva dos dados

Correlação

- Covariância
- Correlação

A correlação varia entre -1 (anticorrelação perfeita) e 1 (correlação perfeita)

Tratamento e limpeza de dados

Qualidade de dados

Dimensões da qualidade de dados:

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

Detecção de outliers

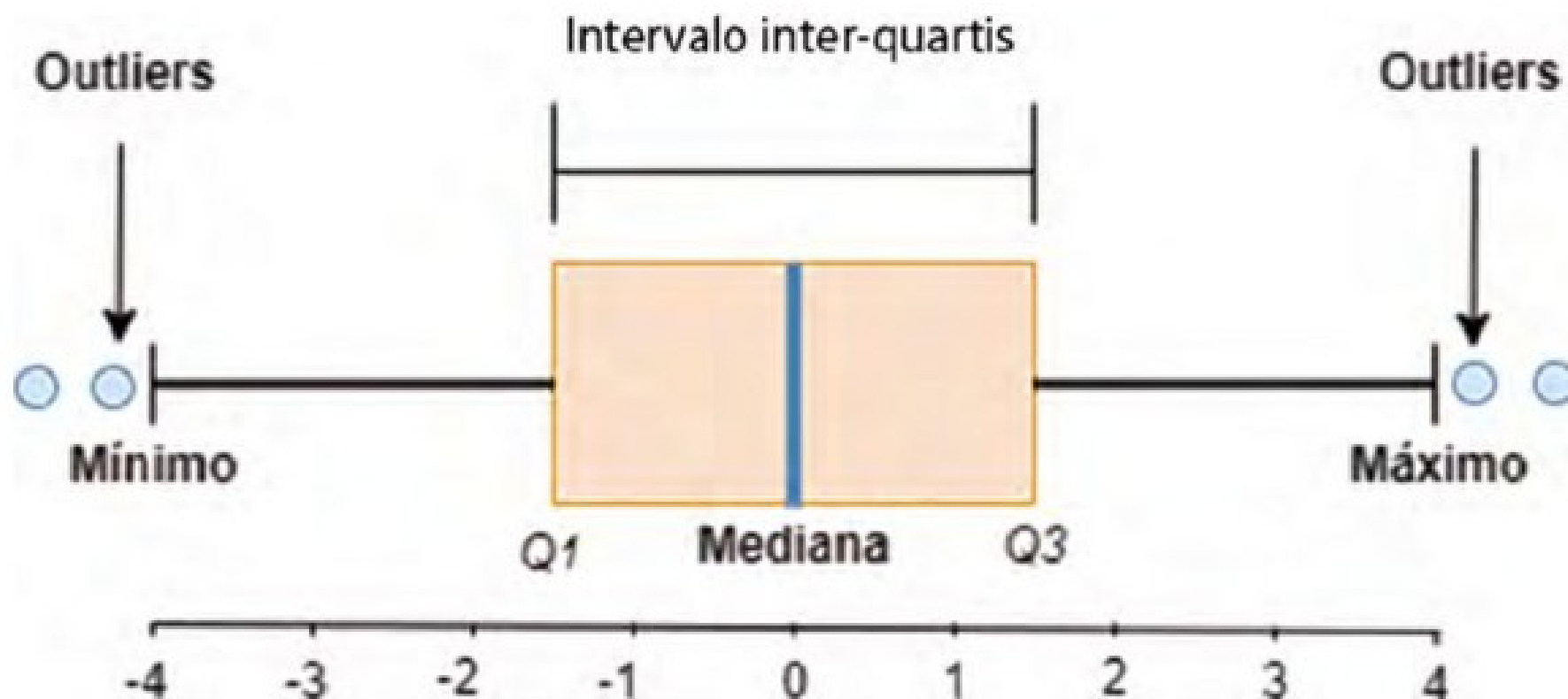


Figura 1. Diferentes partes do *box-plot*.

Detecção de outliers

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

```
import seaborn as sns  
sns.boxplot(x=dados_poke['Attack'])
```

Visualização de dados



INSTITUTO BRASILEIRO DE ENSINO,
DESENVOLVIMENTO E PESQUISA