PythonLab

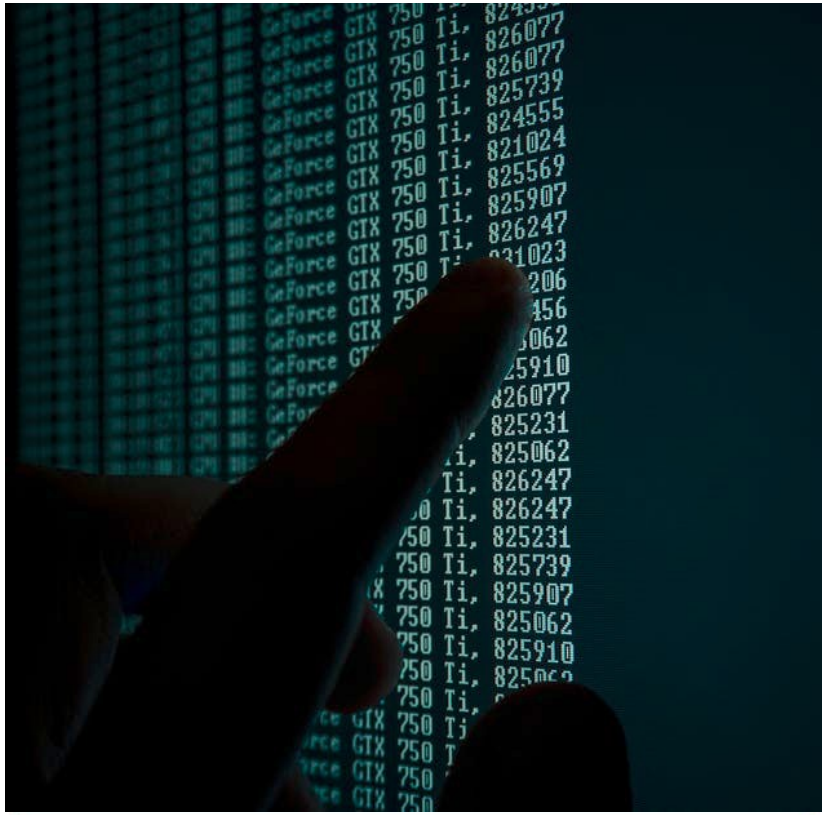
Prof. Dr. Álvaro Campos Ferreira
alvaro.ferreira@idp.edu.br

idp

Dados

O que são dados?

Fatos e números coletados, analisados e sintetizados para a apresentação e interpretação.



Dados

Categorizados

- Escala Nominal
- Escala Ordinal
- Exemplos:
  - Lista de nomes
  - Lista de menções

Quantitativos

- Escala Intervalar
- Exemplos:
  - Lista de salários
  - Preços diários
  - Número de usuários

Dados

Seção transversal

- Todos as observações são realizadas ao mesmo tempo

Série temporal

- Observações realizadas ao longo do tempo

Pandas e DataFrames

Pandas e Dataframes

Vamos usar o Pandas para acessar os nossos dados sobre Pokemons.

```
import pandas as pd
file = “pokemon_data.csv”
df = pd.read_csv(file)
print(df)
```

idp

Indexação de DataFrames

O DataFrames é organizado em colunas que podem ser acessadas a partir de seu nome.

Para determinar os nomes das colunas, pode-se escrever na tela os primeiros valores com a função columns

```
df.columns
```

Indexação de DataFrames

Para selecionar uma coluna de um DataFrame, a sintaxe é a mesma de um dicionário, com o nome da coluna no lugar da chave:

```
df[“Name”]
```

Para selecionar elementos, use a indexação de listas:

```
df[“Name”][0]
```

idp

Indexação de DataFrames

Slices ou fatias são seleções de mais de um elemento de um objeto ao mesmo tempo.

A faixa de índices é indicada entre o sinal “:”.

Ou seja, para selecionar os dez primeiros elementos, usa-se:

```
df[“Name”][0:10]
```

Indexação de DataFrames

Para modificar o valor de um DataFrame, utilize o indicador .loc:

df.loc[0,“Name”] = “Álvaro”

Para selecionar elementos sem fazer referência ao nome da coluna, use a indexação de listas com .iloc:

df.iloc[0,0] = “Xerxes”

idp

Indexação de DataFrames

Para acessar um elemento ou mais a partir de uma operação, pode-se utilizar a seguinte sintaxe:

```
df[df[“Nome”] == “Xerxes”]
```

Isso vai retornar todas as linhas em que o valor da coluna “Nome” seja “Xerxes”.

idp

Funções de DataFrames

DataFrames possuem diversas funções para facilitar a análise exploratória e de estatística descritiva. Algumas das funções que vamos utilizar são:

- describe()
- plot()

Funções de DataFrames

Funções de estatística descritiva:

- describe()
- count()
- sum()
- mean()
- median()
- mode()
- std()
- min()
- max()
- abs()

Funções de DataFrames

Algumas funções importantes para DataFrames são:

- groupby()
- sort_values()
- filter()
- value_counts()
- columns
- head()
- tail()
- values

Tratamento e limpeza de dados

Qualidade de dados

Dimensões da qualidade de dados:

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

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.

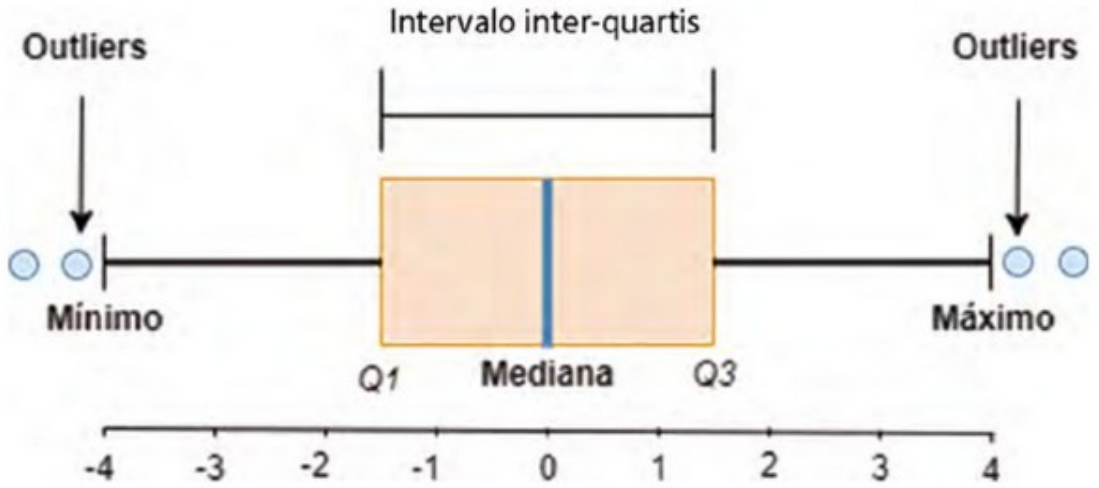


Figura 1. Diferentes partes do *box-plot*.

Detecção de outliers

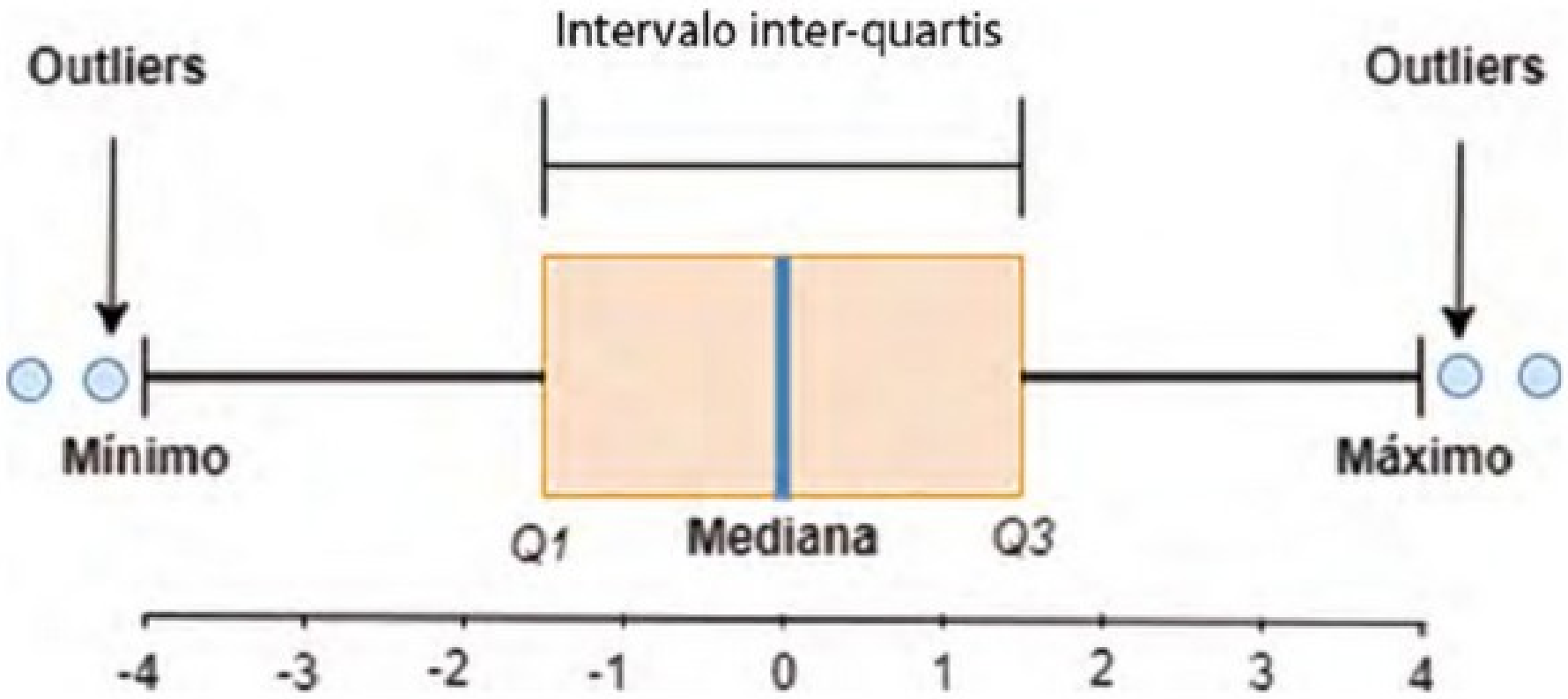


Figura 1. Diferentes partes do *box-plot*.

Webscrapping

Webscrapping

É legal ou ilegal fazer webscrapping?

- É possível acidentalmente realizar muitos requests e isso é visto como um ataque
- Seu IP pode ser banido para algum site ou serviço

idp

Webscrapping

É perfeitamente legal se feito de forma responsável!

- Robots.txt

  https://twitter.com/robots.txt

- Requests
- Delays

idp

Webscrapping

Existem dados em tabelas que podem ser utilizados diretamente pelo Pandas.

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

Webscrapping

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’)
```

INSTITUTO BRASILEIRO DE ENSINO,
DESENVOLVIMENTO E PESQUISA