PythonLab

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

idp

Dados

Onde encontrar dados?

Onde encontrar dados?

1) Google Data Set Search
2) OpenDataSUS
3) Kaggle
4) Webscrapping

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’)
```

Pandas e DataFrames

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

Detecção de outliers

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

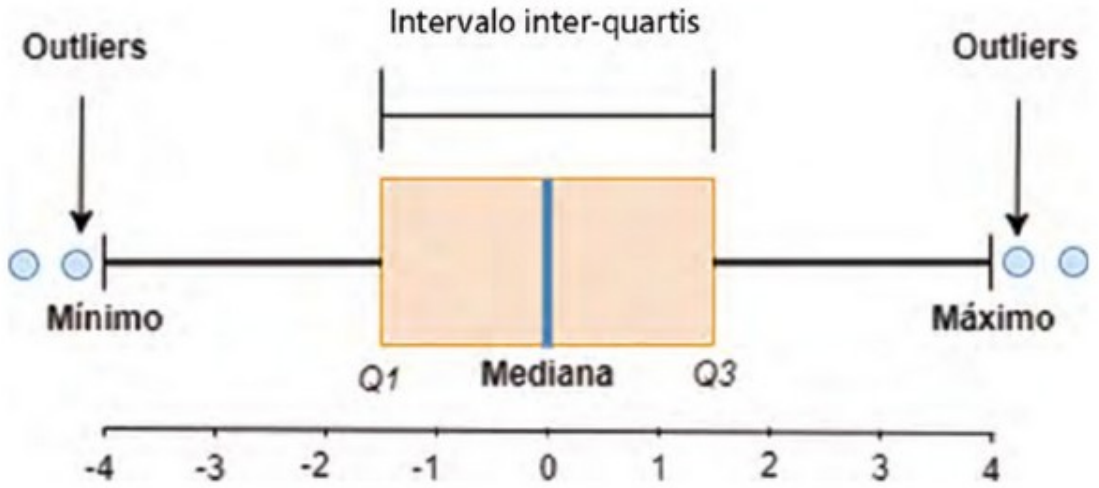

Figura 1. Diferentes partes do *box-plot*.

Detecção de outliers

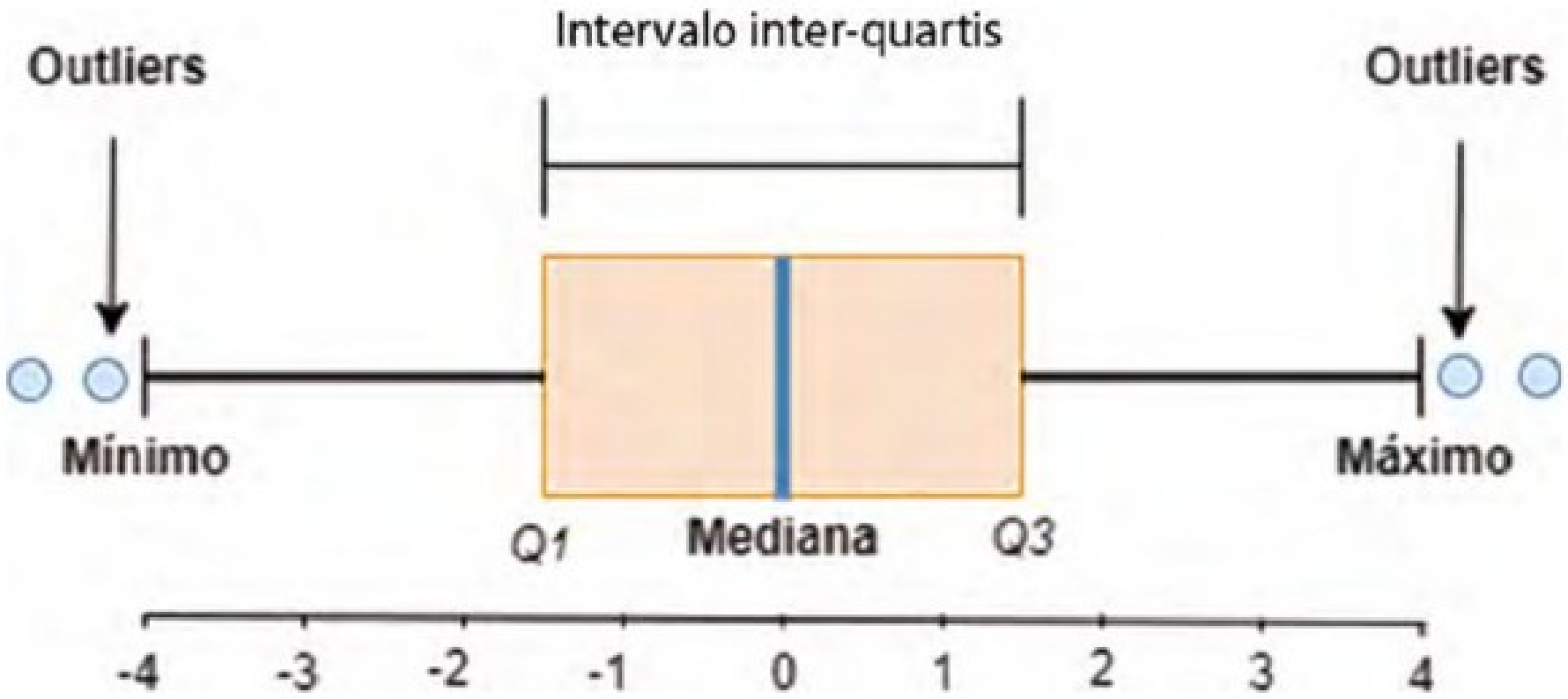

Figura 1. Diferentes partes do *box-plot*.

Visualização de dados

Visualização de dados

Existem várias formas de se visualizar dados. Em geral, esses tipos se dividem em:

- Relacionamentos
- Comparação
- Distribuição
- Composição

Relacionamentos

Relacionamentos são os gráficos que mostram as relações entre dois conjuntos de dados.

- Scatter Plot
- Bubble Plot

Comparação

Para comparar dois conjuntos de dados graficamente, utiliza-se:

- Gráfico de linha (Line Plot)
- Gráfico de barras (Bar Plot)

Distribuição

Para visualizar a distribuição de valores,

- Histograma
- Box Plot

INSTITUTO BRASILEIRO DE ENSINO,
DESENVOLVIMENTO E PESQUISA