# Carregando e armazenando dados em vários formatos de arquivos

Trabalhando com HTML

# Carregando e armazenando dados

## Trabalhando com HTML

Existem muitas bibliotecas para ler e gravar *HTML* e *XML*, como o *Beautiful Soup*, *lxml* e *html5lib*.

O *Pandas* possui uma função embutida denominada *read_html* que utiliza bibliotecas como *lxml* e *Beautiful Soup* para fazer parse automaticamente de tabelas existentes em arquivos *HTML* para objetos *DataFrame*.

# Carregando e armazenando dados

## Trabalhando com HTML

Usando o *read_html* informando apenas o caminho do arquivo *HTML*, por padrão o *Pandas* tenta fazer o parse de dados tabulares contidos em *tags <table></table>*.

Vamos criar um arquivo HTML de exemplo, com uma tabela para carregarmos os dados da tabela com o *Pandas*.

# Carregando e armazenando dados

## Trabalhando com HTML

```
<html>
    <head>
        <meta charset="UTF-8">
        <title>Municípios</title>
    </head>
<body>
    Tabela:
    <table>
        <tr>
            <th>Município</th>
            <th>População</th>
        </tr>
        <tr>
            <td>Serra</td>
            <td>507.598</td>
        </tr>
        <tr>
            <td>Vila Velha</td>
            <td>486.208</td>
        </tr>
...
```

```
...
        <tr>
            <td>Cariacica</td>
            <td>387.368</td>
        </tr>
        <tr>
            <td>Vitória</td>
            <td>358.267</td>
        </tr>
        <tr>
            <td>Cachoeiro de Itapemirim</td>
            <td>220.670</td>
        </tr>
        <tr>
            <td>Linhares</td>
            <td>173.555</td>
        </tr>
        <tr>
            <td>São Mateus</td>
            <td>130.611</td>
        </tr>
    </table>
</body>
</html>
```

# Carregando e armazenando dados

## Trabalhando com HTML

Veja o conteúdo do arquivo quando abrimos usando um browser.

Tabela:

| Município | População |
|---|---|
| Serra | 507.598 |
| Vila Velha | 486.208 |
| Cariacica | 387.368 |
| Vitória | 358.267 |
| Cachoeiro de Itapemirim | 220.670 |
| Linhares | 173.555 |
| São Mateus | 130.611 |

# Carregando e armazenando dados

## Trabalhando com HTML

Veja agora como fazemos para carregar o arquivo usando o *read_html* do *Pandas*

E também seu resultado.

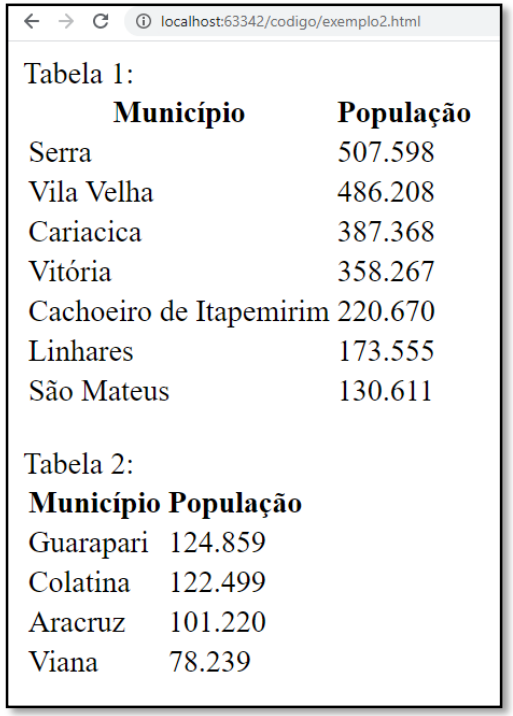
```
import pandas as pd

tabela = pd.read_html('exemplo1.html')
print(tabela)
```

```
[                  Município  População
0                      Serra    507.598
1                 Vila Velha    486.208
2                   Cariacica    387.368
3                     Vitória    358.267
4  Cachoeiro de Itapemirim     220.670
5                    Linhares    173.555
6                  São Mateus    130.611]
```

# Carregando e armazenando dados

## Trabalhando com HTML

O arquivo exemplo2.html possui duas tabelas, veja a página, o programa e o resultado.




```python
import pandas as pd


tabela = pd.read_html('exemplo2.html')
print("============================")
print(tabela[0])
print("============================")
print(tabela[1])
print("============================")
```

```
============================
                Município  População
0                  Serra    507.598
1              Vila Velha    486.208
2               Cariacica    387.368
3                 Vitória    358.267
4  Cachoeiro de Itapemirim    220.670
5                Linhares    173.555
6              São Mateus    130.611
============================
   Município  População
0  Guarapari    124.859
1   Colatina    122.499
2    Aracruz    101.220
3      Viana     78.239
============================
```

# Carregando e armazenando dados

## Trabalhando com HTML

Podemos também informar como parâmetro o endereço de uma página *html*. Veja um exemplo tirado da documentação do *Pandas* onde é carregada uma página da agência governamental *FDIC* (*Federal Deposit Insurance Corporation*) dos Estados Unidos, que mostra falências de bancos.

# Carregando e armazenando dados

## Trabalhando com HTML

```
import pandas as pd

url = 'https://www.fdic.gov/bank/individual/failed/banklist.html'
dfs = pd.read_html(url)
print(dfs)
```

```
[                              Bank Name           City  ST   CERT                 Acquiring Institution       Closing Date
0        City National Bank of New Jersey         Newark  NJ  21111                     Industrial Bank   November 1, 2019
1                          Resolute Bank         Maumee  OH  58317                   Buckeye State Bank   October 25, 2019
2                  Louisa Community Bank         Louisa  KY  58112  Kentucky Farmers Bank Corporation   October 25, 2019
3                   The Enloe State Bank         Cooper  TX  10716                  Legend Bank, N. A.       May 31, 2019
4     Washington Federal Bank for Savings        Chicago  IL  30570                  Royal Savings Bank  December 15, 2017
..                                    ...            ...  ..    ...                                 ...               ...
554                  Superior Bank, FSB        Hinsdale  IL  32646              Superior Federal, FSB      July 27, 2001
555                  Malta National Bank           Malta  OH   6629                  North Valley Bank       May 3, 2001
556        First Alliance Bank & Trust Co.      Manchester  NH  34264  Southern New Hampshire Bank & Trust   February 2, 2001
557     National State Bank of Metropolis      Metropolis  IL   3815            Banterra Bank of Marion  December 14, 2000
558                      Bank of Honolulu        Honolulu  HI  21029                 Bank of the Orient   October 13, 2000

[559 rows x 6 columns]]
```

# Carregando e armazenando dados

## Trabalhando com HTML

Podemos utilizar o parâmetro *match* (corresponder) do *read_html* para retornar uma tabela que contenha um conteúdo específico.

Em nosso exemplo2.html temos duas tabelas, vamos trazer somente a tabela que contenha o município “Viana”.

# Carregando e armazenando dados

## Trabalhando com HTML

```
import pandas as pd

cidade = 'Viana'
lista = pd.read_html('exemplo2.html', match=cidade)
print(lista)
```

```
[   Município  População
0  Guarapari    124.859
1   Colatina    122.499
2    Aracruz    101.220
3      Viana     78.239]
```

# Carregando e armazenando dados

## Trabalhando com HTML

Por padrão, o read_html já entende como cabeçalho as tags <th> ou <td> localizados em um <thead>.

Se quisermos especificar uma linha como cabeçalho podemos usar o parâmetro header com o índice da linha. Por exemplo, se no exemplo anterior usarmos header=1, a linha do município Guarapari será considerada o header da tabela.

# Carregando e armazenando dados

## Trabalhando com HTML

```
import pandas as pd

cidade = 'Viana'
lista = pd.read_html('exemplo2.html', match=cidade, header=1)
print(lista)
```

```
[  Guarapari  124.859
0  Colatina  122.499
1   Aracruz  101.220
2     Viana   78.239]
```

# Carregando e armazenando dados

## Trabalhando com HTML

Podemos definir também o índice da coluna com index_col=número_da_coluna.

Para especificar a coluna que contém os nomes dos municípios como índice podemos fazer conforme próximo exemplo.

# Carregando e armazenando dados

## Trabalhando com HTML

```
import pandas as pd

cidade = 'Viana'
lista = pd.read_html('exemplo2.html', match=cidade, index_col=0)
print(lista)
```

```
[            População
Município
Guarapari     124.859
Colatina      122.499
Aracruz       101.220
Viana          78.239]
```

# Carregando e armazenando dados

## Trabalhando com HTML

Podemos ignorar algumas linhas utilizando skiprows=número_de_linhas.

```
import pandas as pd

cidade = 'Viana'
lista = pd.read_html('exemplo2.html', match=cidade, skiprows=2)
print(lista)
```

```
[  Colatina  122.499
0  Aracruz  101.220
1    Viana   78.239]
```

# Carregando e armazenando dados

## Trabalhando com HTML

Vamos criar agora o exemplo3.html à partir do exemplo2.html com as seguintes alterações:

- Adicionar um id na tabela 1.
- Adicionar um estilo na tabela 2.
- Remover os dados da população do município de Guarapari.

```
    Tabela 1:
    <table id="tabela1">
        <tr>
            <th>Município</th>
            <th>População</th>
        </tr>
        <tr>
            <td>Serra</td>
            <td>507.598</td>
        </tr>
        ...
        <tr>
            <td>São Mateus</td>
            <td>130.611</td>
        </tr>
    </table><br>
...
```

```
    Tabela 2:
    <table style="width:100%">
        <tr>
            <th>Município</th>
            <th>População</th>
        </tr>
        <tr>
            <td>Guarapari</td>
            <td></td>
        </tr>
        ...
        <tr>
            <td>Viana</td>
            <td>78.239</td>
        </tr>
    </table>
```

# Carregando e armazenando dados

## Trabalhando com HTML

Como no arquivo exemplo3.html os dados da população de Guarapari não existem, será exibido "Dados ausentes" (NaN) para este município.

Para que não seja exibido "NaN" podemos usar o parâmetro keep_default_na igual a False.

# Carregando e armazenando dados

## Trabalhando com HTML

```
import pandas as pd

lista1 = pd.read_html('exemplo3.html', match='Viana')
lista2 = pd.read_html('exemplo3.html', match='Viana', keep_default_na=False)
print(lista1)
print(lista2)
```

```
[   Município  População
0  Guarapari        NaN
1   Colatina    122.499
2    Aracruz    101.220
3      Viana     78.239]
[   Município População
0  Guarapari
1   Colatina   122.499
2    Aracruz   101.220
3      Viana    78.239]
```

# Carregando e armazenando dados

## Trabalhando com HTML

Podemos especificar a tabela que queremos utilizando atributos, como por exemplo, o id, classe ou outro atributo qualquer que identifique a tabela.

```
import pandas as pd

lista1 = pd.read_html('exemplo3.html', attrs={'id': 'tabela1'})
lista2 = pd.read_html('exemplo3.html', attrs={'style': 'width:100%'})
print(lista1)
print("================================")
print(lista2)
```

```
[                 Município  População
0                     Serra    507.598
1                Vila Velha    486.208
2                  Cariacica    387.368
3                   Vitória    358.267
4  Cachoeiro de Itapemirim    220.670
5                  Linhares    173.555
6                São Mateus    130.611]
================================
[   Município  População
0  Guarapari        NaN
1   Colatina    122.499
2    Aracruz    101.220
3      Viana     78.239]
```

# FIM