

RASPAGEM DE DADOS - BRASILEIRÃO 2023

Por: Leandro Dias Vieira



- Instalando bibliotecas

```
In [63]: !pip install BeautifulSoup4
```

```
Requirement already satisfied: BeautifulSoup4 in /usr/local/lib/python3.10/dist-packages (4.11.2)  
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-packages (from BeautifulSoup4) (2.5)
```

```
In [64]: !pip install requests
```

```
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (2.31.0)  
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests) (3.3.2)  
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests) (3.6)  
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests) (2.0.7)  
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests) (2023.11.17)
```

- Importando Bibliotecas

```
In [65]: import requests  
from bs4 import BeautifulSoup  
import pandas as pd  
import numpy as np
```

- Pegando o Link

```
In [66]: url = "https://www.espn.com.br/futebol/classificacao/_/liga/bra.1"
```

- Passando parametros para acessar site

```
In [67]: headers = {"User-Agent" : "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/53
```

- Testando requisicao

```
In [68]: requisicao = requests.get(url, headers = headers)
```

```
In [69]: print(requisicao)
```

```
<Response [200]>
```

```
In [70]: ## print(requisicao.text)
```

- Parseando o site

```
In [71]: site = BeautifulSoup(requisicao.text, "html.parser")
```

- Função Prettify(), para arrumar HTML e Concatenar

```
In [72]: site2 = site.prettify()
```

```
In [73]: ## print(site2)
```

Importando a primeira tabela

- Encontrando a tag da primeira tabela

```
In [74]: tabela = site.find("table")
```

```
In [75]: print(tabela)
```

<table class="Table Table--align-right Table--fixed Table--fixed-left" style="border-collapse: collapse; border-spacing: 0"><colgroup class="Table__Colgroup"><col class="Table__Column"/></colgroup><thead class="Table__header-group Table__THEAD"><tr class="Table__sub-header Table__TR Table__even"><th class="subHeader_item--content Table__TH" title="">2023</th></tr></thead><tbody class="Table__BODY"><tr class="Table__TR Table__TR--sm Table__even" data-idx="0"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">1<abbr data-clubhouse-uid="s:600~t:2029" style="text-decoration: none" title="Palmeiras">PAL</abbr>Palmeiras</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="1"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">2<abbr data-clubhouse-uid="s:600~t:6273" style="text-decoration: none" title="Grêmio">GRE</abbr>Grêmio</div></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="2"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">3<abbr data-clubhouse-uid="s:600~t:7632" style="text-decoration: none" title="Atlético-MG">CAM</abbr>Atlético-MG</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="3"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">4<abbr data-clubhouse-uid="s:600~t:819" style="text-decoration: none" title="Flamengo">FLA</abbr>Flamengo</div></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="4"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">5<abbr data-clubhouse-uid="s:600~t:6086" style="text-decoration: none" title="Botafogo">BOT</abbr>Botafogo</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="5"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">6</div></td></tr></tbody></table>

[pr4 TeamLink__Logo](/futebol/time/_id/6079/red-bull-bragantino)><abbr data-clubhouse-uid="s:600~t:6079" style="text-decoration:none" title="Red Bull Bragantino">BRA</abbr>Red Bull Bragantino</div></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="6"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">7<abbr data-clubhouse-uid="s:600~t:3445" style="text-decoration:none" title="Fluminense">FLU</abbr>Fluminense</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="7"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">8<abbr data-clubhouse-uid="s:600~t:3458" style="text-decoration:none" title="Athletico-PR">CAP</abbr>Athletico-PR</div></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="8"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">9<abbr data-clubhouse-uid="s:600~t:1936" style="text-decoration:none" title="Internacional">INT</abbr>Internacional</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="9"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">10<abbr data-clubhouse-uid="s:600~t:6272" style="text-decoration:none" title="Fortaleza">FOR</abbr>Fortaleza</div></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="10"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">11<abbr data-clubhouse-uid="s:600~t:2026" style="text-decoration:none" title="São Paulo">SAO</abbr>São Paulo</div></td></tr>

ss="hide-mobile">São Paulo</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="11"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">12CuiabáCuiabá</div></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="12"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">13CorinthiansCorinthians</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="13"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">14CruzeiroCruzeiro</div></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="14"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">15Vasco da GamaVasco da Gama</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="15"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">16BahiaBahia</div></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="16"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">17<abbr data-clubhouse-uid="s:600~t:2674" style="text-decoration:none" title="Santos">SAN</abbr>Santos</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="17"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">18<abbr data-clubhouse-uid="s:600~t:3395" style="text-decoration:none" title="Goiás">GOI</abbr>Goiás</div></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="18"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">19<abbr data-clubhouse-uid="s:600~t:3456" style="text-decoration:none" title="Coritiba">CF C</abbr>Coritiba</div></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="19"><td class="Table__TD"><div class="team-link flex items-center clr-gray-03">20<abbr data-clubhouse-uid="s:600~t:6154" style="text-decoration:none" title="América-MG">AMG</abbr>América-MG</div></td></tr></tbody></table>

- Criando DataFrame

In [76]: `df = pd.DataFrame(columns=["Time"])`

In [77]: `df`

Out[77]: Time

- Fazendo a Inserção das Linhas

In [78]:

```
for linha in tabela.tbody.find_all("tr"):
    for coluna in linha.find_all("td"):
        for div in coluna.find_all("div"):
            span = div.find("span", class_="hide-mobile")
            nometime = span.find_all("a")
            if (nometime != []):
                time = nometime[0].text.strip(" ")
                df = pd.concat([df, pd.DataFrame.from_records([{"Time": time}])])
```

```
In [79]: df
```

```
Out[79]:
```

	Time
0	Palmeiras
0	Grêmio
0	Atlético-MG
0	Flamengo
0	Botafogo
0	Red Bull Bragantino
0	Fluminense
0	Athletico-PR
0	Internacional
0	Fortaleza
0	São Paulo
0	Cuiabá
0	Corinthians
0	Cruzeiro
0	Vasco da Gama
0	Bahia
0	Santos
0	Goiás
0	Coritiba
0	América-MG

Importando a segunda tabela

- Encontrando a tag da segunda tabela

```
In [80]: tabela2 = site.find("table", class_="Table Table--align-right")
```

```
In [81]: print(tabela2)
```

Jogos	Vitórias	Empate	Derrotas	Gols pró
GP	Gols contra	Saldo de gols	Pontos	PTS
38	20	10	8	64
31	70	33	70	38
5	12	63	56	7
52	32	20	66	10
38	19	9	56	42
32	18	10	58	38

Table_TR		stat-cell">37		span		Table_TR		stat-cell">21	
Table_TR		stat-cell">64		span		Table_TR		stat-cell">17	
Table_TR		stat-cell">11		span		Table_TR		stat-cell">10	
Table_TR		stat-cell">49		span		Table_TR		stat-cell">35	
Table_TR		stat-cell">14		span		Table_TR		stat-cell">62	
Table_TR		stat-cell">38		span		Table_TR		stat-cell">8	
Table_TR		stat-cell">14		span		Table_TR		stat-cell">51	
Table_TR		stat-cell">47		span		Table_TR		stat-cell">4	
Table_TR		stat-cell">56		span		Table_TR		stat-cell">14	
Table_TR		stat-cell">14		span		Table_TR		stat-cell">10	
Table_TR		stat-cell">51		span		Table_TR		stat-cell">43	
Table_TR		stat-cell">8		span		Table_TR		stat-cell">56	
Table_TR		stat-cell">15		span		Table_TR		stat-cell">10	
Table_TR		stat-cell">13		span		Table_TR		stat-cell">46	
Table_TR		stat-cell">45		span		Table_TR		stat-cell">1	
Table_TR		stat-cell">55		span		Table_TR		stat-cell">38	
Table_TR		stat-cell">15		span		Table_TR		stat-cell">9	
Table_TR		stat-cell">14		span		Table_TR		stat-cell">45	
Table_TR		stat-cell">44		span		Table_TR		stat-cell">1	
Table_TR		stat-cell">54		span		Table_TR		stat-cell">38	
Table_TR		stat-cell">14		span		Table_TR		stat-cell">11	
Table_TR		stat-cell">13		span		Table_TR		stat-cell">40	
Table_TR		stat-cell">38		span		Table_TR		stat-cell">2	
Table_TR		stat-cell">53		span		Table_TR		stat-cell">38	
Table_TR		stat-cell">14		span		Table_TR		stat-cell">9	
Table_TR		stat-cell">15		span		Table_TR		stat-cell">40	
Table_TR		stat-cell">39		span		Table_TR		stat-cell">1	
Table_TR		stat-cell">51		span		Table_TR		stat-cell">38	
Table_TR		stat-cell">12		span		Table_TR		stat-cell">14	
Table_TR		stat-cell">12		span		Table_TR		stat-cell">47	
Table_TR		stat-cell">48		span		Table_TR		stat-cell">-1	
Table_TR		stat-cell">50		span		Table_TR		stat-cell">38	
Table_TR		stat-cell">11		span		Table_TR		stat-cell">14	
Table_TR		stat-cell">13		span		Table_TR		stat-cell">35	
Table_TR		stat-cell">32		span		Table_TR		stat-cell">3	
Table_TR		stat-cell">47		span		Table_TR		stat-cell">47	
Table_TR		stat-cell">47		span		Table_TR		stat-cell">47	

```
s="stat-cell">38</span></td><td class="Table__TD"><span class="stat-cell">12</span>
</td><td class="Table__TD"><span class="stat-cell">9</span></td><td class="Table__TD"><span class="stat-cell">17</span></td><td class="Table__TD"><span class="stat-cell">41</span></td><td class="Table__TD"><span class="stat-cell">51</span></td><td class="Table__TD"><span class="stat-cell">-10</span></td><td class="Table__TD"><span class="stat-cell">45</span></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="15"><td class="Table__TD"><span class="stat-cell">38</span></td><td class="Table__TD"><span class="stat-cell">12</span></td><td class="Table__TD"><span class="stat-cell">8</span></td><td class="Table__TD"><span class="stat-cell">18</span></td><td class="Table__TD"><span class="stat-cell">50</span></td><td class="Table__TD"><span class="stat-cell">53</span></td><td class="Table__TD"><span class="stat-cell">-3</span></td><td class="Table__TD"><span class="stat-cell">44</span></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="16"><td class="Table__TD"><span class="stat-cell">38</span></td><td class="Table__TD"><span class="stat-cell">11</span></td><td class="Table__TD"><span class="stat-cell">10</span></td><td class="Table__TD"><span class="stat-cell">17</span></td><td class="Table__TD"><span class="stat-cell">39</span></td><td class="Table__TD"><span class="stat-cell">64</span></td><td class="Table__TD"><span class="stat-cell">-25</span></td><td class="Table__TD"><span class="stat-cell">43</span></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="17"><td class="Table__TD"><span class="stat-cell">38</span></td><td class="Table__TD"><span class="stat-cell">9</span></td><td class="Table__TD"><span class="stat-cell">11</span></td><td class="Table__TD"><span class="stat-cell">18</span></td><td class="Table__TD"><span class="stat-cell">36</span></td><td class="Table__TD"><span class="stat-cell">53</span></td><td class="Table__TD"><span class="stat-cell">-17</span></td><td class="Table__TD"><span class="stat-cell">38</span></td></tr><tr class="Table__TR Table__TR--sm Table__even" data-idx="18"><td class="Table__TD"><span class="stat-cell">38</span></td><td class="Table__TD"><span class="stat-cell">8</span></td><td class="Table__TD"><span class="stat-cell">6</span></td><td class="Table__TD"><span class="stat-cell">24</span></td><td class="Table__TD"><span class="stat-cell">41</span></td><td class="Table__TD"><span class="stat-cell">73</span></td><td class="Table__TD"><span class="stat-cell">-32</span></td><td class="Table__TD"><span class="stat-cell">30</span></td></tr><tr class="filled Table__TR Table__TR--sm Table__even" data-idx="19"><td class="Table__TD"><span class="stat-cell">38</span></td><td class="Table__TD"><span class="stat-cell">5</span></td><td class="Table__TD"><span class="stat-cell">9</span></td><td class="Table__TD"><span class="stat-cell">24</span></td><td class="Table__TD"><span class="stat-cell">42</span></td><td class="Table__TD"><span class="stat-cell">81</span></td><td class="Table__TD"><span class="stat-cell">-39</span></td><td class="Table__TD"><span class="stat-cell">24</span></td></tr></tbody></table>
```

- Criando outro DF para armazenar resultado

```
In [82]: df2 = pd.DataFrame(columns=["Jogos", "Vitórias", "Empates", "Derrotas", "Gols Pró",
```

- Fazendo a inserção das linhas

```
In [83]: for linha in tabela2.tbody.find_all("tr"):
    coluna = linha.find_all("td")
    if (coluna != []):
        jogos = coluna[0].text.strip(' ')
        vitorias = coluna[1].text.strip(' ')
        empates = coluna[2].text.strip(" ")
        derrotas = coluna[3].text.strip(" ")
        gols_pro = coluna[4].text.strip(" ")
        gols_contra = coluna[5].text.strip(" ")
        saldo_de_gols = coluna[6].text.strip(" ")
        pontos = coluna[7].text.strip(" ")
        df2 = pd.concat([df2, pd.DataFrame.from_records([{"Jogos": jogos, "Vitórias": \
```

```
In [84]: df2
```

```
Out[84]:
```

	Jogos	Vitórias	Empates	Derrotas	Gols Pró	Gols Contra	Saldo de Gols	Pontos
0	38	20	10	8	64	33	+31	70
0	38	21	5	12	63	56	+7	68
0	38	19	9	10	52	32	+20	66
0	38	19	9	10	56	42	+14	66
0	38	18	10	10	58	37	+21	64
0	38	17	11	10	49	35	+14	62
0	38	16	8	14	51	47	+4	56
0	38	14	14	10	51	43	+8	56
0	38	15	10	13	46	45	+1	55
0	38	15	9	14	45	44	+1	54
0	38	14	11	13	40	38	+2	53
0	38	14	9	15	40	39	+1	51
0	38	12	14	12	47	48	-1	50
0	38	11	14	13	35	32	+3	47
0	38	12	9	17	41	51	-10	45
0	38	12	8	18	50	53	-3	44
0	38	11	10	17	39	64	-25	43
0	38	9	11	18	36	53	-17	38
0	38	8	6	24	41	73	-32	30
0	38	5	9	24	42	81	-39	24

Concatenando as tabelas

```
In [85]: tabela_final = pd.concat([df,df2], axis=1)
tabela_final
```

Out[85]:

	Time	Jogos	Vitórias	Empates	Derrotas	Gols Pró	Gols Contra	Saldo de Gols	Pontos
0	Palmeiras	38	20	10	8	64	33	+31	70
0	Grêmio	38	21	5	12	63	56	+7	68
0	Atlético-MG	38	19	9	10	52	32	+20	66
0	Flamengo	38	19	9	10	56	42	+14	66
0	Botafogo	38	18	10	10	58	37	+21	64
0	Red Bull Bragantino	38	17	11	10	49	35	+14	62
0	Fluminense	38	16	8	14	51	47	+4	56
0	Athletico-PR	38	14	14	10	51	43	+8	56
0	Internacional	38	15	10	13	46	45	+1	55
0	Fortaleza	38	15	9	14	45	44	+1	54
0	São Paulo	38	14	11	13	40	38	+2	53
0	Cuiabá	38	14	9	15	40	39	+1	51
0	Corinthians	38	12	14	12	47	48	-1	50
0	Cruzeiro	38	11	14	13	35	32	+3	47
0	Vasco da Gama	38	12	9	17	41	51	-10	45
0	Bahia	38	12	8	18	50	53	-3	44
0	Santos	38	11	10	17	39	64	-25	43
0	Goiás	38	9	11	18	36	53	-17	38
0	Coritiba	38	8	6	24	41	73	-32	30
0	América-MG	38	5	9	24	42	81	-39	24

Arrumando índices

- Começando por 1, e não por 0

```
In [86]: tabela_final.index = np.arange(1, len(tabela_final)+1)
```

```
In [87]: tabela_final
```

Out[87]:

	Time	Jogos	Vitórias	Empates	Derrotas	Gols Pró	Gols Contra	Saldo de Gols	Pontos
1	Palmeiras	38	20	10	8	64	33	+31	70
2	Grêmio	38	21	5	12	63	56	+7	68
3	Atlético-MG	38	19	9	10	52	32	+20	66
4	Flamengo	38	19	9	10	56	42	+14	66
5	Botafogo	38	18	10	10	58	37	+21	64
6	Red Bull Bragantino	38	17	11	10	49	35	+14	62
7	Fluminense	38	16	8	14	51	47	+4	56
8	Athletico-PR	38	14	14	10	51	43	+8	56
9	Internacional	38	15	10	13	46	45	+1	55
10	Fortaleza	38	15	9	14	45	44	+1	54
11	São Paulo	38	14	11	13	40	38	+2	53
12	Cuiabá	38	14	9	15	40	39	+1	51
13	Corinthians	38	12	14	12	47	48	-1	50
14	Cruzeiro	38	11	14	13	35	32	+3	47
15	Vasco da Gama	38	12	9	17	41	51	-10	45
16	Bahia	38	12	8	18	50	53	-3	44
17	Santos	38	11	10	17	39	64	-25	43
18	Goiás	38	9	11	18	36	53	-17	38
19	Coritiba	38	8	6	24	41	73	-32	30
20	América-MG	38	5	9	24	42	81	-39	24

Ajustando Tipagem dos Dados

```
In [88]: tabela_final["Jogos"] = pd.to_numeric(tabela_final["Jogos"])
tabela_final["Vitórias"] = pd.to_numeric(tabela_final["Vitórias"])
tabela_final["Empates"] = pd.to_numeric(tabela_final["Empates"])
tabela_final["Derrotas"] = pd.to_numeric(tabela_final["Derrotas"])
tabela_final["Gols Pró"] = pd.to_numeric(tabela_final["Gols Pró"])
tabela_final["Gols Contra"] = pd.to_numeric(tabela_final["Gols Contra"])
tabela_final["Saldo de Gols"] = pd.to_numeric(tabela_final["Saldo de Gols"])
tabela_final["Pontos"] = pd.to_numeric(tabela_final["Pontos"])
tabela_final.dtypes
```

```
Out[88]: Time          object
         Jogos        int64
         Vitórias     int64
         Empates       int64
         Derrotas      int64
         Gols Pró      int64
         Gols Contra   int64
         Saldo de Gols int64
         Pontos        int64
         dtype: object
```

Top 5

- Top 5 times com menos vitórias

```
In [89]: tabela_final.sort_values(by="Vitórias", ascending= True).head(5)
```

```
Out[89]:
```

	Time	Jogos	Vitórias	Empates	Derrotas	Gols Pró	Gols Contra	Saldo de Gols	Pontos
20	América-MG	38	5	9	24	42	81	-39	24
19	Coritiba	38	8	6	24	41	73	-32	30
18	Goiás	38	9	11	18	36	53	-17	38
17	Santos	38	11	10	17	39	64	-25	43
14	Cruzeiro	38	11	14	13	35	32	3	47

- Top 5 times com mais empates

```
In [90]: tabela_final.sort_values(by="Empates", ascending= False).head(5)
```

```
Out[90]:
```

	Time	Jogos	Vitórias	Empates	Derrotas	Gols Pró	Gols Contra	Saldo de Gols	Pontos
14	Cruzeiro	38	11	14	13	35	32	3	47
13	Corinthians	38	12	14	12	47	48	-1	50
8	Athletico-PR	38	14	14	10	51	43	8	56
11	São Paulo	38	14	11	13	40	38	2	53
18	Goiás	38	9	11	18	36	53	-17	38

- Top 5 times que mais fizeram gols

```
In [91]: tabela_final.sort_values(by="Gols Pró", ascending= False).head(5)
```

Out[91]:

	Time	Jogos	Vitórias	Empates	Derrotas	Gols Pró	Gols Contra	Saldo de Gols	Pontos
1	Palmeiras	38	20	10	8	64	33	31	70
2	Grêmio	38	21	5	12	63	56	7	68
5	Botafogo	38	18	10	10	58	37	21	64
4	Flamengo	38	19	9	10	56	42	14	66
3	Atlético-MG	38	19	9	10	52	32	20	66

- Top 5 times com mais saldo de gols

```
In [92]: tabela_final.sort_values(by="Saldo de Gols", ascending=False).head(5)
```

Out[92]:

	Time	Jogos	Vitórias	Empates	Derrotas	Gols Pró	Gols Contra	Saldo de Gols	Pontos
1	Palmeiras	38	20	10	8	64	33	31	70
5	Botafogo	38	18	10	10	58	37	21	64
3	Atlético-MG	38	19	9	10	52	32	20	66
4	Flamengo	38	19	9	10	56	42	14	66
6	Red Bull Bragantino	38	17	11	10	49	35	14	62