Projeto Web-Scraping Dados da Libertadores Wikipedia

A ideia do projeto é realizar web-scraping com dados das tabelas do libertadores presentes na página do wikipedia sobre a competição.

Separei o projeto em 3 etapas:

1º Extração dos dados

2º Manipulação e Limpeza

3º Criação dos gráficos

Raspagem

O código abaixo realizar a raspagem da 1º tabela presente no site utilizando a url e a biblioteca BeatifulSoup. A intenção é pegar dos dados da seguinta tabela presente no site:

Títulos por clubes

Clube	País	Títulos	Vices	Semifinais	Aprov.
Independiente	Argentina	7 (1964, 1965, 1972, 1973, 1974, 1975 e 1984)	0	5 (1966, 1976, 1979, 1985 e 1987)	100%
Boca Juniors	Argentina	6 (1977, 1978, 2000, 2001, 2003 e 2007)	6 (1963, 1979, 2004, 2012, 2018 e 2023)	7 (1965, 1966, 1991, 2008, 2016, 2019 e 2020)	50%
Peñarol	Uruguai	5 (1960, 1961, 1966, 1982 e 1987)	5 (1962, 1965, 1970, 1983 e 2011)	10 (1963, 1967, 1968, 1969, 1972, 1974, 1976, 1979, 1981 e 1985)	50%
River Plate	Argentina	4 (1986, 1996, 2015 e 2018)	3 (1966, 1976 e 2019)	13 (1967, 1970, 1978, 1982, 1987, 1990, 1995, 1998, 1999, 2004, 2005, 2017 e 2020)	57,14%
Estudiantes	Argentina	4 (1968, 1969, 1970 e 2009)	1 (1971)	1 (1983)	80%
Olimpia	Paraguai	3 (1979, 1990 e 2002)	4 (1960, 1989, 1991 e 2013)	5 (1961, 1980, 1982, 1986 e 1994)	42,86%
Nacional	Uruguai	3 (1971, 1980 e 1988)	3 (1964, 1967 e 1969)	7 (1962, 1966, 1972, 1981, 1983, 1984 e 2009)	50%

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

# URL da página da Wikipedia com as tabelas da Libertadores
url = "https://pt.wikipedia.org/wiki/Copa_Libertadores_da_Am%C3%A9rica"

# Fazendo a requisição para a URL
response = requests.get(url)

# Criando o objeto BeautifulSoup
soup = BeautifulSoup(response.text, 'html.parser')

# Encontrando a tabela no HTML com a classe desejada
table = soup.find('table', {'class': 'wikitable sortable'})

# Aqui eu pego as infomações em html da 1º tabela
tabela_wiki = pd.read_html(str(table))[0]
```

Já com as informações em mãos, eu transformo os dados em um dataframe para realizar as manipulações e o gráficos necessários.

Agora eu posso visualizar o dataframe semelhante a tabela do wikipedia

In [ ]:
```
import pandas as pd
import requests
from bs4 import BeautifulSoup
import matplotlib.pyplot as plt

page = requests.get('https://pt.wikipedia.org/wiki/Copa_Libertadores_da_Am%C3%A9rica').t
soup = BeautifulSoup(page, 'html.parser')
table = soup.find('table', class_='wikitable sortable')

df = pd.read_html(str(table))[0]
tabela_wiki = pd.DataFrame(df)
tabela_wiki.head(17)
```

Out[ ]:

	Clube	País	Títulos	Vices	Semifinais	Aprov.
0	Independiente	Argentina	7 (1964, 1965, 1972, 1973, 1974, 1975 e 1984)	0	5 (1966, 1976, 1979, 1985 e 1987)	100%
1	Boca Juniors	Argentina	6 (1977, 1978, 2000, 2001, 2003 e 2007)	6 (1963, 1979, 2004, 2012, 2018 e 2023)	7 (1965, 1966, 1991, 2008, 2016, 2019 e 2020)	50%
2	Peñarol	Uruguai	5 (1960, 1961, 1966, 1982 e 1987)	5 (1962, 1965, 1970, 1983 e 2011)	10 (1963, 1967, 1968, 1969, 1972, 1974, 1976,...	50%
3	River Plate	Argentina	4 (1986, 1996, 2015 e 2018)	3 (1966, 1976 e 2019)	13 (1967, 1970, 1978, 1982, 1987, 1990, 1995, ...	57,14%
4	Estudiantes	Argentina	4 (1968, 1969, 1970 e 2009)	1 (1971)	1 (1983)	80%
5	Olimpia	Paraguai	3 (1979, 1990 e 2002)	4 (1960, 1989, 1991 e 2013)	5 (1961, 1980, 1982, 1986 e 1994)	42,86%
6	Nacional	Uruguai	3 (1971, 1980 e 1988)	3 (1964, 1967 e 1969)	7 (1962, 1966, 1972, 1981, 1983, 1984 e 2009)	50%
7	Palmeiras	Brasil	3 (1999, 2020 e 2021)	3 (1961, 1968 e 2000)	5 (1971, 2001, 2018, 2022 e 2023)	50%
8	São Paulo	Brasil	3 (1992, 1993 e 2005)	3 (1974, 1994 e 2006)	4 (1972, 2004, 2010 e 2016)	50%
9	Grêmio	Brasil	3 (1983, 1995 e 2017)	2 (1984 e 2007)	5 (1996, 2002, 2009, 2018 e 2019)	60%
10	Santos	Brasil	3 (1962, 1963 e 2011)	2 (2003 e 2020)	4 (1964, 1965, 2007 e 2012)	60%
11	Flamengo	Brasil	3 (1981, 2019 e 2022)	1 (2021)	2 (1982 e 1984)	75%
12	Cruzeiro	Brasil	2 (1976 e 1997)	2 (1977 e 2009)	2 (1967 e 1975)	50%
13	Internacional	Brasil	2 (2006 e 2010)	1 (1980)	4 (1977, 1989, 2015 e 2023)	66,67%
14	Atlético Nacional	Colômbia	2 (1989 e 2016)	1 (1995)	2 (1990 e 1991)	66,67%
15	Colo-Colo	Chile	1 (1991)	1 (1973)	3 (1964, 1967 e 1997)	50%
16	Fluminense	Brasil	1 (2023)	1 (2008)	0	50%

Agora vem uma parte muito importante. Podemos observar que na coluna títulos temos os valores que cada time conquistou junto dos anos em parenteses. O problema que não é possível realizar gráficos assim, já que está em formato de string a coluna de títulos.

Para isso é necessário limpar os parenteses da coluna e transformar em float para ser um dados numerico. A função remove_parentheses tem esse objetivo.

In [ ]:
```
import re
# Função para remover o que está entre parênteses e converter para float
remove_parentheses = lambda s: float(re.sub(r'\([^)]*\)', '', s))
# Lendo novamente podemos perceber que os parenteses sumiram e podemos criar o gráfico a
tabela_wiki["Títulos"] = tabela_wiki["Títulos"].apply(remove_parentheses)
tabela_wiki
```

Out[ ]:

	Clube	País	Títulos	Vices	Semifinais	Aprov.
0	Independiente	Argentina	7.0	0	5 (1966, 1976, 1979, 1985 e 1987)	100%

1	Boca Juniors	Argentina	6.0	6 (1963, 1979, 2004, 2012, 2018 e 2023)	7 (1965, 1966, 1991, 2008, 2016, 2019 e 2020)	50%
2	Peñarol	Uruguai	5.0	5 (1962, 1965, 1970, 1983 e 2011)	10 (1963, 1967, 1968, 1969, 1972, 1974, 1976,...	50%
3	River Plate	Argentina	4.0	3 (1966, 1976 e 2019)	13 (1967, 1970, 1978, 1982, 1987, 1990, 1995, ...	57,14%
4	Estudiantes	Argentina	4.0	1 (1971)	1 (1983)	80%
...	...	...	...	...	...	...
67	Junior Barranquilla	Colômbia	0.0	0	1 (1994)	0%
68	Emelec	Equador	0.0	0	1 (1995)	0%
69	Independiente Medellín	Colômbia	0.0	0	1 (2003)	0%
70	Cúcuta Deportivo	Colômbia	0.0	0	1 (2007)	0%
71	Defensor Sporting	Uruguai	0.0	0	1 (2014)	0%

72 rows × 6 columns

Agora estamos próximos de criar os gráficos, para isso criei um novo dataframe apenas com as informações do Clube e Títulos para plotar. Escolhi o head(17) para pegar os 17 primeiros clubes com mais títulos, para incluir o Fluminense último campeão.

In [ ]:
```
# Criando um dataframe novo com os dados dos clubes e titulos
top_17_winners = tabela_wiki[["Clube", "Títulos"]].head(17)
top_17_winners
```

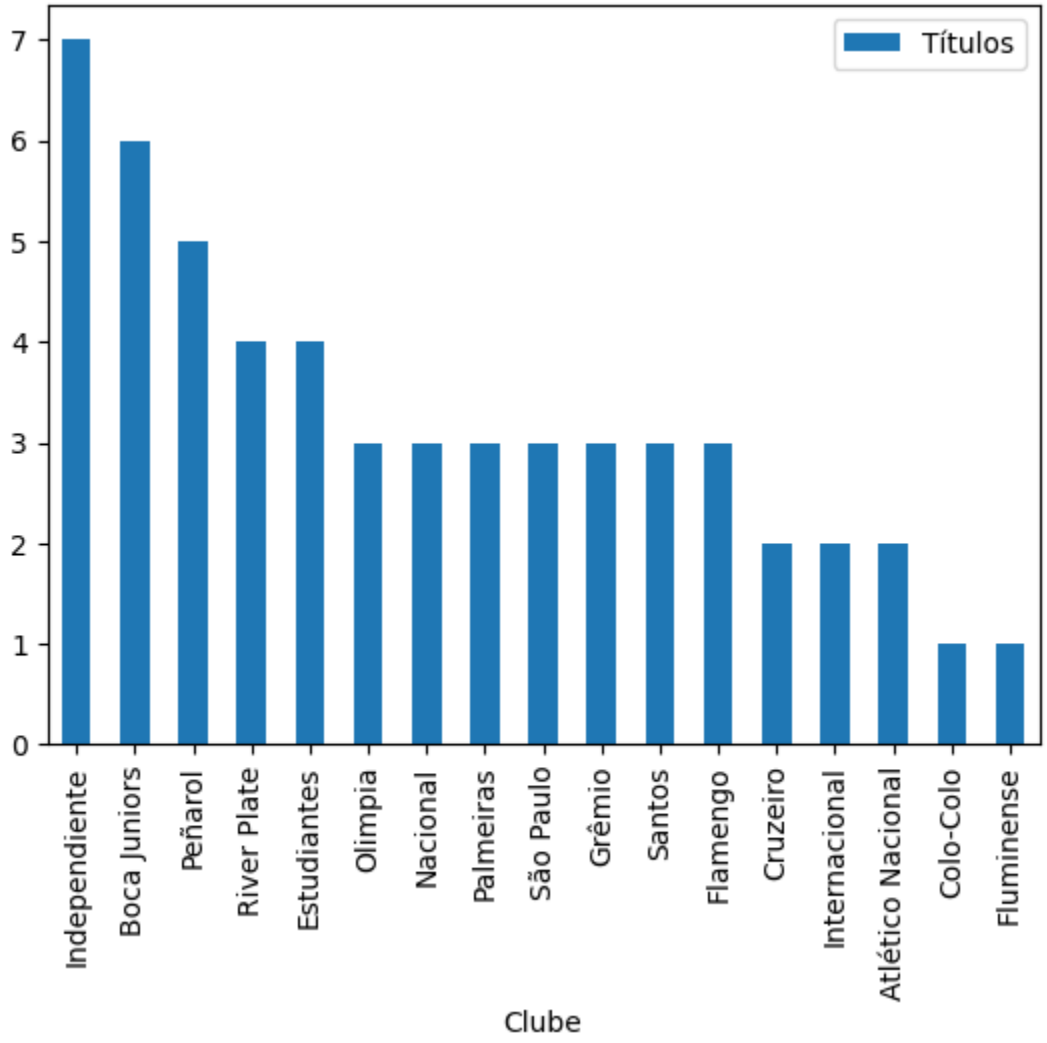
Out[ ]:

	Clube	Títulos
0	Independiente	7.0
1	Boca Juniors	6.0
2	Peñarol	5.0
3	River Plate	4.0
4	Estudiantes	4.0
5	Olimpia	3.0
6	Nacional	3.0
7	Palmeiras	3.0
8	São Paulo	3.0
9	Grêmio	3.0
10	Santos	3.0
11	Flamengo	3.0
12	Cruzeiro	2.0
13	Internacional	2.0
14	Atlético Nacional	2.0
15	Colo-Colo	1.0
16	Fluminense	1.0

Criando o gráfico de barras para visualizarmos os times que mais vencerão a competição. Podemos ver que indepiende da Argentina venceu incrível 7x

In [ ]:
```
# Cria o gráfico de barras
top_17_winners.plot.bar(x='Clube', y='Títulos')
```

Out[ ]:
```
<Axes: xlabel='Clube'>
```



Agora realizei um filtro para pegar apenas os clubes brasileiros e comparar a quantidade de títulos de cada clube. Vamos ver quem tem mais!

In [ ]:
```
top_Brasil = tabela_wiki.loc[(df['País'] == 'Brasil') & (df["Títulos"]>= 1)]
top_Brasil
```

Out[ ]:

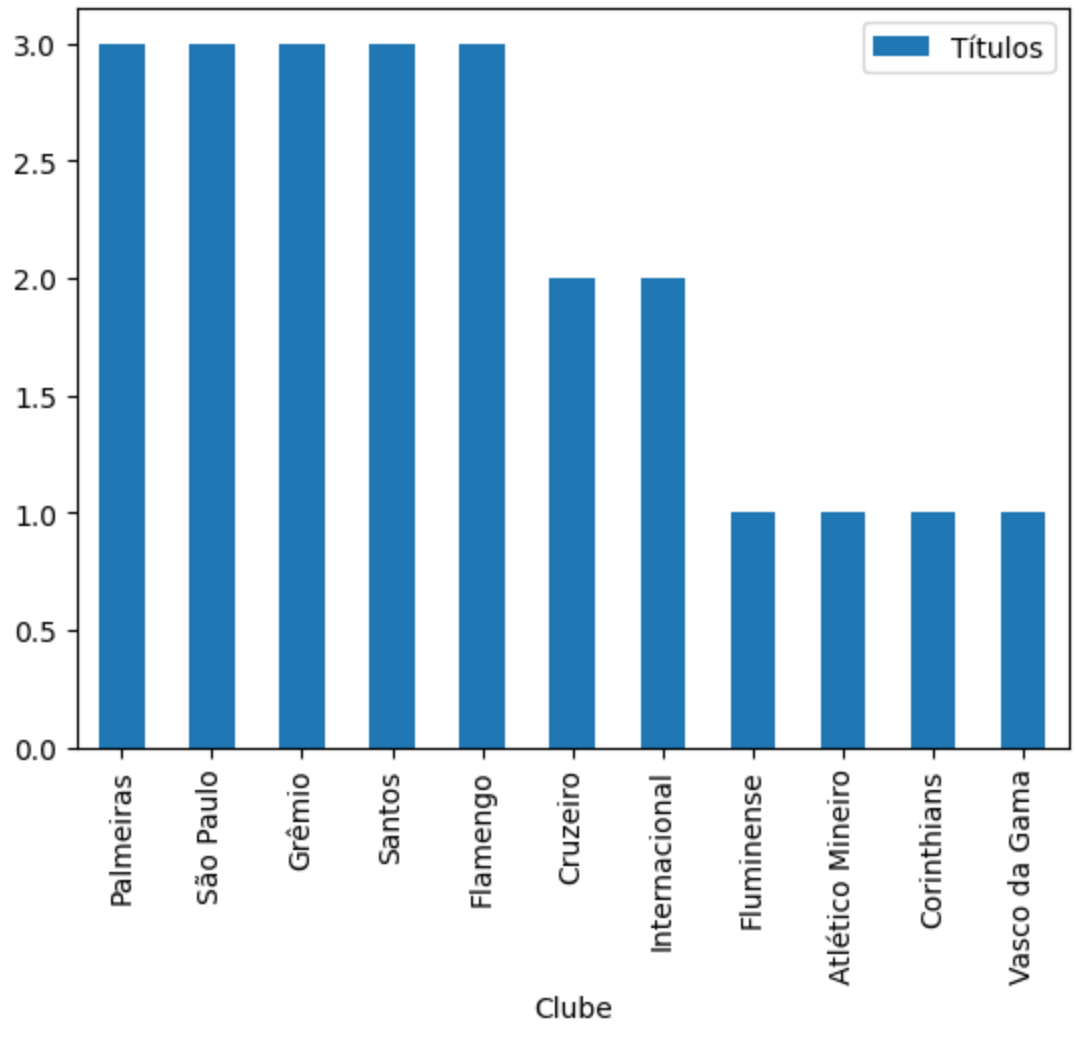
	Clube	País	Títulos	Vices	Semifinais	Aprov.
7	Palmeiras	Brasil	3.0	3 (1961, 1968 e 2000)	5 (1971, 2001, 2018, 2022 e 2023)	50%
8	São Paulo	Brasil	3.0	3 (1974, 1994 e 2006)	4 (1972, 2004, 2010 e 2016)	50%
9	Grêmio	Brasil	3.0	2 (1984 e 2007)	5 (1996, 2002, 2009, 2018 e 2019)	60%
10	Santos	Brasil	3.0	2 (2003 e 2020)	4 (1964, 1965, 2007 e 2012)	60%
11	Flamengo	Brasil	3.0	1 (2021)	2 (1982 e 1984)	75%
12	Cruzeiro	Brasil	2.0	2 (1977 e 2009)	2 (1967 e 1975)	50%
13	Internacional	Brasil	2.0	1 (1980)	4 (1977, 1989, 2015 e 2023)	66,67%
16	Fluminense	Brasil	1.0	1 (2008)	0	50%
21	Atlético Mineiro	Brasil	1.0	0	2 (1978 e 2021)	100%
23	Corinthians	Brasil	1.0	0	1 (2000)	100%
24	Vasco da Gama	Brasil	1.0	0	0	100%

Criando o gráfico de barras para saber quem tem mais títulos dos clubes brasileiros

Entre os brasileiros temos 5 times com 3 conquitas: Palmeiras, Santos, Flamengo, Grêmio e São Paulo.

In [ ]:
```
# Cria o gráfico de barras
top_Brasil.plot.bar(x='Clube', y='Títulos')
```

Out[ ]:
```
<Axes: xlabel='Clube'>
```



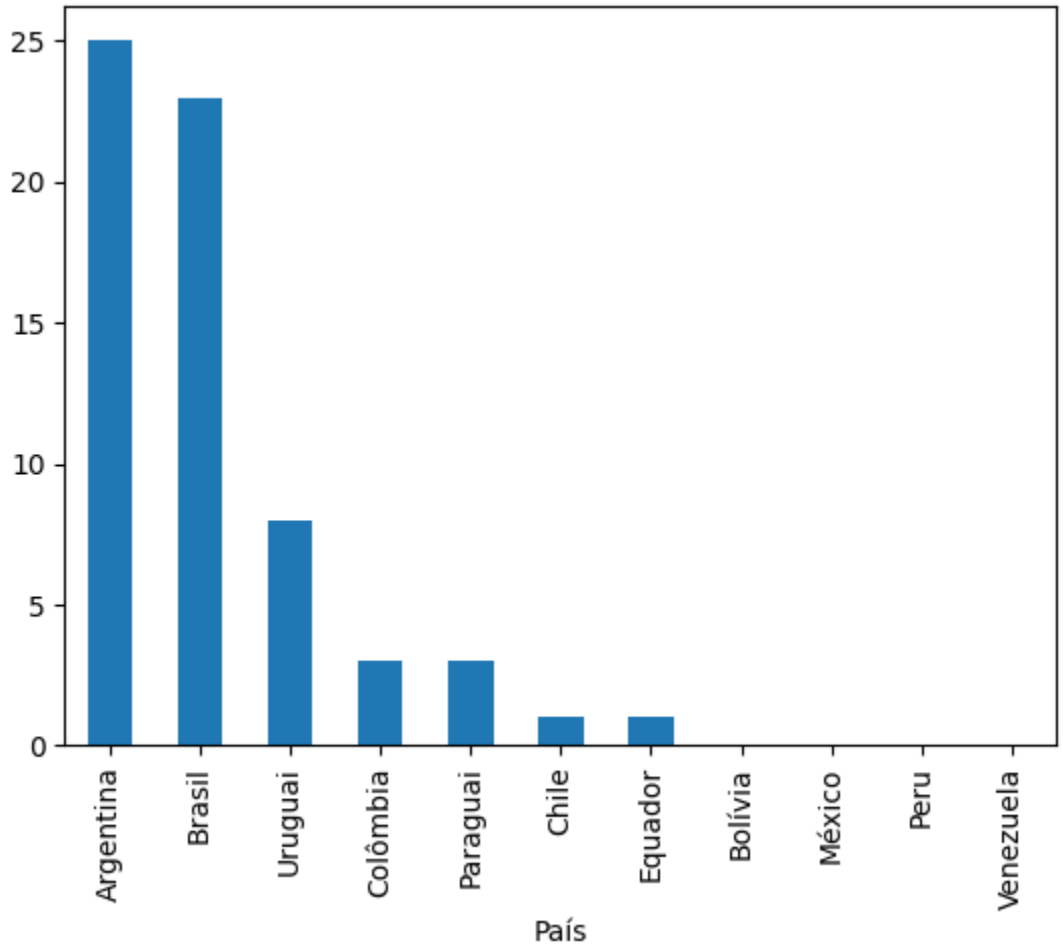
Agora irei agrupar entre títulos e país para descobrir qual país ganhou mais a competição. Infelizmente é a Argentina.

Como irei agrupar, precisarei utilizar o groupby e uma função que nesse caso será o sum(), já que quero saber o total.

In [ ]:
```
agrupando_pais = tabela_wiki.groupby('País')['Títulos'].sum()
ordenado_agrupando_pais = agrupando_pais.sort_values(ascending=False)
```

Criando o gráfico com os países com mais títulos:

In [ ]:
```
ordenado_agrupando_pais.plot(kind='bar')
plt.show()
```



O resultado ficou parecido com a tabela abaixo. Eu poderia ter pego essa tabela e ter realizado o gráfico com as colunas País e Títulos. Mas optei fazer por groupby:

Títulos por países

País	Títulos	Vices	Aprov.	Clubes campeões
Argentina	25	13	**65,79 %**	8
Brasil	23	18	**56,1 %**	11
Uruguai	8	8	**50 %**	2
Colômbia	3	7	**30 %**	2
Paraguai	3	5	**37,5 %**	1
Chile	1	5	**16,67 %**	1
Equador	1	3	**25 %**	1
México	0	3	0%	0
Peru	0	2	0%	0
Bolívia	0	0	–	0
Venezuela	0	0	–	0

Agora irei pegar os dados da 3º tabela presente no site do wikipedia para realizar 2 análises:

- 1º Times brasileiros com mais participações;

- 2º Times brasileiros com mais vitórias.

Ranking de pontos

De 1960 até 2022, foram realizadas 63 edições da Copa Libertadores da América.[carece de fontes] Nesse período, os 30 maiores clubes pontuadores na competição, foram os seguintes:

Pos.	Clube	País	Part	Tít	J	V	E	D	GP	GC	Pts	Dif.
1	River Plate	Argentina	38	4	381	186	101	94	634	396	**659**	=
2	Nacional	Uruguai	49	3	407	172	109	126	566	445	**625**	=
3	Peñarol	Uruguai	48	5	375	166	80	129	560	455	**578**	=
4	Boca Juniors	Argentina	31	6	316	165	80	71	476	275	**575**	=
5	Olimpia	Paraguai	44	3	332	129	96	107	469	418	**482**	=
6	Cerro Porteño	Paraguai	43	0	325	118	92	115	410	426	**446**	=
7	Palmeiras	Brasil	22	3	222	125	42	55	429	226	**417**	=
8	Grêmio	Brasil	21	3	207	108	43	56	318	189	**367**	=
9	Colo-Colo	Chile	35	1	247	97	55	95	345	347	**346**	=
10	Bolívar	Bolívia	36	0	245	96	54	95	360	375	**342**	▲ (1)
11	São Paulo	Brasil	21	3	199	96	48	55	307	192	**336**	▼ (1)
12	América de Cali	Colômbia	21	0	208	91	59	58	299	229	**332**	=

Raspando os dados da tabela sobre ranking de pontos para retirar as informações que preciso para analisar os times com mais vitórias e participações

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


page = requests.get('https://pt.wikipedia.org/wiki/Copa_Libertadores_da_Am%C3%A9rica').t
soup = BeautifulSoup(page, 'html.parser')
table = soup.find_all('table', class_='wikitable sortable')

df = pd.read_html(str(table))[3]
tabela_participações = pd.DataFrame(df)
tabela_participações
# df.to_csv("elections.csv", index=False)
```

Out[ ]:

	Pos.	Clube	País	Part	Tít	J	V	E	D	GP	GC	Pts	Dif.
0	1	River Plate	Argentina	38	4	381	186	101	94	634	396	659	NaN
1	2	Nacional	Uruguai	49	3	407	172	109	126	566	445	625	NaN
2	3	Peñarol	Uruguai	48	5	375	166	80	129	560	455	578	NaN
3	4	Boca Juniors	Argentina	31	6	316	165	80	71	476	275	575	NaN
4	5	Olimpia	Paraguai	44	3	332	129	96	107	469	418	482	NaN
5	6	Cerro Porteño	Paraguai	43	0	325	118	92	115	410	426	446	NaN
6	7	Palmeiras	Brasil	22	3	222	125	42	55	429	226	417	NaN
7	8	Grêmio	Brasil	21	3	207	108	43	56	318	189	367	NaN
8	9	Colo-Colo	Chile	35	1	247	97	55	95	345	347	346	NaN
9	10	Bolívar	Bolívia	36	0	245	96	54	95	360	375	342	(1)
10	11	São Paulo	Brasil	21	3	199	96	48	55	307	192	336	(1)
11	12	América de Cali	Colômbia	21	0	208	91	59	58	299	229	332	NaN
12	13	Universidad Católica	Chile	29	0	236	88	59	89	348	340	323	NaN

13	14	Cruzeiro	Brasil	17	2	166	95	32	39	307	158	317	NaN
14	15	Flamengo	Brasil	18	3	162	93	33	36	316	179	312	(4)
15	16	Barcelona de Guayaquil	Equador	28	0	235	83	60	92	277	289	309	(1)
16	17	Santos	Brasil	16	3	153	83	32	38	287	173	281	(1)
17	18	Atlético Nacional	Colômbia	23	2	187	78	46	63	252	207	280	NaN
18	19	Universitario	Peru	33	0	227	70	70	87	265	313	280	(2)
19	20	Vélez Sarsfield	Argentina	17	1	151	76	36	39	217	146	264	(2)
20	21	Estudiantes	Argentina	16	4	143	78	26	39	190	122	260	(4)
21	22	Sporting Cristal	Peru	37	0	234	66	60	108	287	371	258	(2)
22	23	Independiente	Argentina	20	7	154	72	39	43	211	143	255	(2)
23	24	Emelec	Equador	29	0	221	68	44	109	238	314	248	NaN
24	25	Libertad	Paraguai	21	0	170	66	43	61	215	210	241	(1)
25	26	Internacional	Brasil	14	2	140	67	39	34	202	124	240	(3)
26	27	Corinthians	Brasil	16	1	132	66	32	34	216	127	230	(1)
27	28	Deportivo Cali	Colômbia	21	0	160	63	34	63	224	210	223	(1)
28	29	LDU Quito	Equador	20	1	161	62	36	63	238	184	222	(2)
29	30	The Strongest	Bolívia	28	0	177	60	34	83	216	300	214	(1)

Filtrando o país para apenas selecionar os times brasileiros

In [ ]:
```
tabela_participações_Br = tabela_participações[tabela_participações["País"] == "Brasil"]
tabela_participações_Br
```

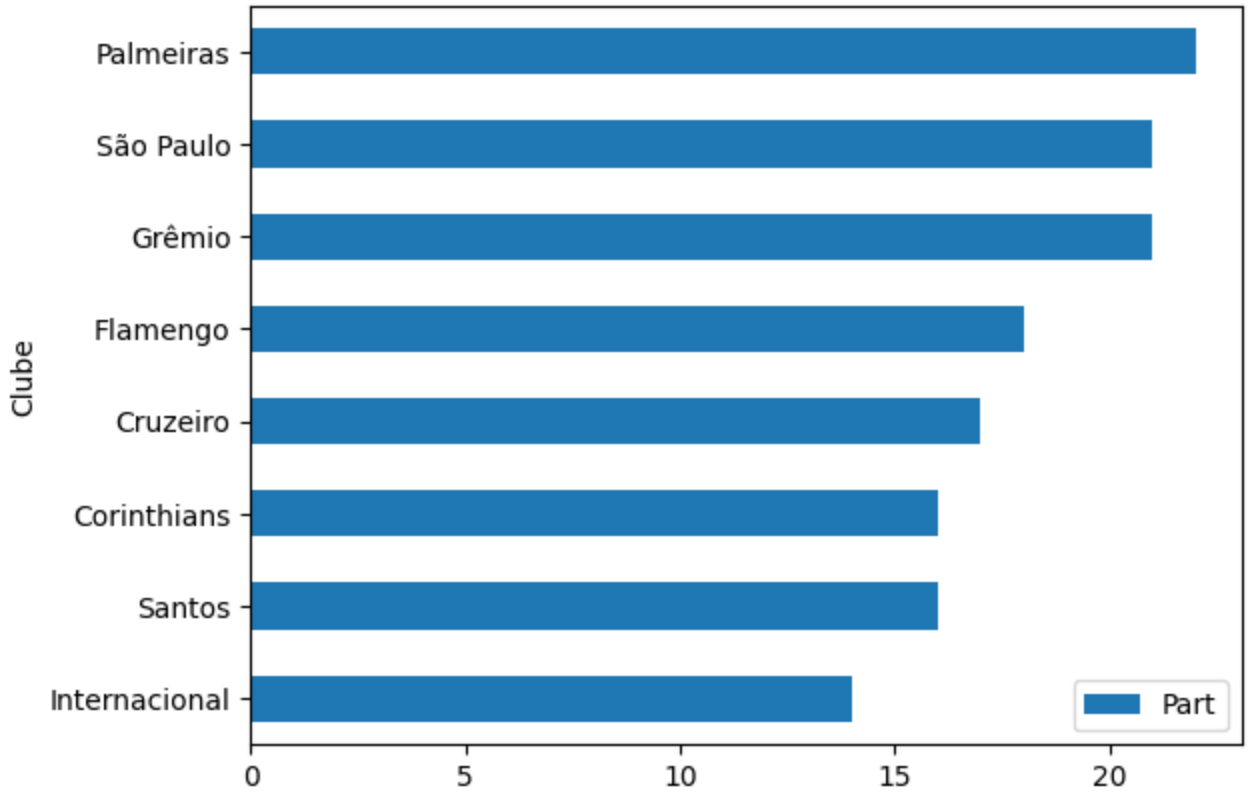
Out[ ]:

	Pos.	Clube	País	Part	Tít	J	V	E	D	GP	GC	Pts	Dif.
6	7	Palmeiras	Brasil	22	3	222	125	42	55	429	226	417	NaN
7	8	Grêmio	Brasil	21	3	207	108	43	56	318	189	367	NaN
10	11	São Paulo	Brasil	21	3	199	96	48	55	307	192	336	(1)
13	14	Cruzeiro	Brasil	17	2	166	95	32	39	307	158	317	NaN
14	15	Flamengo	Brasil	18	3	162	93	33	36	316	179	312	(4)
16	17	Santos	Brasil	16	3	153	83	32	38	287	173	281	(1)
25	26	Internacional	Brasil	14	2	140	67	39	34	202	124	240	(3)
26	27	Corinthians	Brasil	16	1	132	66	32	34	216	127	230	(1)

Criando o gráfico para visualizarmos quais times mais participaram

In [ ]:
```
Br_ordenado = tabela_participações_Br.sort_values('Part')
Br_ordenado.reset_index().plot.barh(x='Clube', y='Part')
```

Out[ ]:
```
<Axes: ylabel='Clube'>
```



Agora irei ordenar os times brasileiros com mais vitórias

In [ ]:
```
Br_ordenado_vitorias = tabela_participações_Br.sort_values('V')
Br_ordenado_vitorias[["Clube","V"]].sort_values(by="V",ascending=False)
```

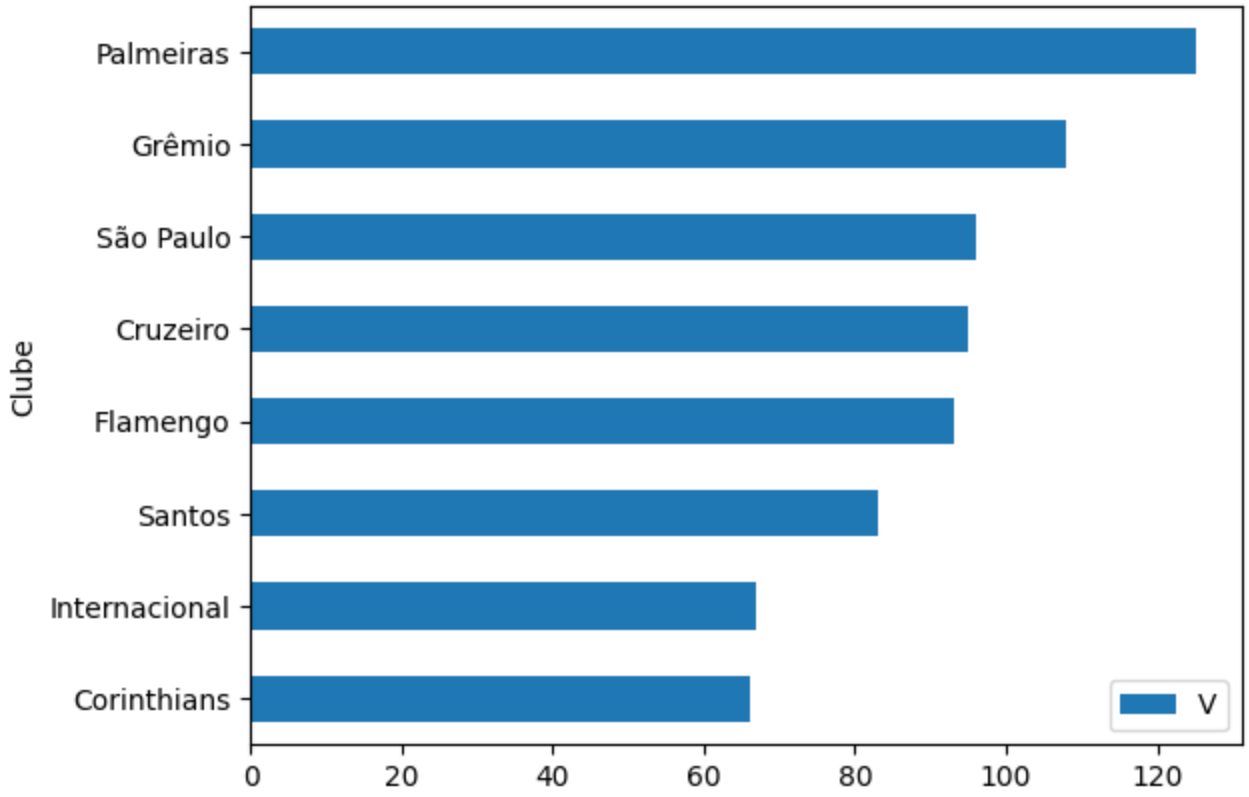
Out[ ]:

	Clube	V
6	Palmeiras	125
7	Grêmio	108
10	São Paulo	96
13	Cruzeiro	95
14	Flamengo	93
16	Santos	83
25	Internacional	67
26	Corinthians	66

Criando o gráfico para visualizar os times brasileiros mais vitoriosos

In [ ]:
```
Br_ordenado_vitorias.reset_index().plot.barh(x='Clube', y='V')
```

Out[ ]:
```
<Axes: ylabel='Clube'>
```


Palmeiras
Grêmio
São Paulo
Cruzeiro
Flamengo
Santos
Internacional
Corinthians
Clube
V
0
20
40
60
80
100
120