In [1]:
```
import pandas as pd
import numpy as np

df_2014 = pd.read_csv('data/brasileirao_2014_2018.txt', sep = ',')
df_2019 = pd.read_csv('data/brasileirao_2019_2022.txt', sep = ',')
```

In [2]:
```
df_2019.columns # verifica colunas
```

Out[2]:
```
Index(['Rk', 'Squad', 'MP', 'W', 'D', 'L', 'GF', 'GA', 'GD', 'Pts', 'Pts/MP',
       'xG', 'xGA', 'xGD', 'xGD/90', 'MP.1', 'W.1', 'D.1', 'L.1', 'GF.1',
       'GA.1', 'GD.1', 'Pts.1', 'Pts/MP.1', 'xG.1', 'xGA.1', 'xGD.1',
       'xGD/90.1'],
      dtype='object')
```

In [3]:
```
# dropa colunas desnecessárias

cols_drop_2019 = ['xG', 'xGA', 'xGD', 'xGD/90', 'xG.1', 'xGA.1', 'xGD.1', 'xGD/90.1']
df_2019 = df_2019.drop(cols_drop_2019, axis = 1)
```

In [4]:
```
df_concat = pd.concat([df_2014, df_2019]).reset_index() # junta dataframes
df_concat
```

Out[4]:

	index	Rk	Squad	MP	W	D	L	GF	GA	GD	...	Pts/MP	MP.1	W.1	D.1	L.1	GF.1	GA.1	GD.1	Pts.1	Pts/MP.1
0	0	1	Cruzeiro	19	15	2	2	43	17	26	...	2.47	19	9	6	4	24	21	3	33	1.74
1	1	2	São Paulo	19	11	6	2	32	16	16	...	2.05	19	9	4	6	27	24	3	31	1.63
2	2	3	Internacional	19	15	0	4	37	16	21	...	2.37	19	6	6	7	16	25	-9	24	1.26
3	3	4	Corinthians	19	12	6	1	32	15	17	...	2.21	19	7	6	6	17	16	1	27	1.42
4	4	5	Atlético Mineiro	19	12	5	2	28	12	16	...	2.16	19	5	6	8	23	26	-3	21	1.11
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
175	75	16	Cuiabá	19	6	8	5	17	15	2	...	1.37	19	4	3	12	14	27	-13	15	0.79
176	76	17	Ceará	19	4	8	7	20	22	-2	...	1.05	19	3	8	8	14	19	-5	17	0.89
177	77	18	Atl Goianiense	19	6	5	8	22	23	-1	...	1.21	19	2	7	10	17	34	-17	13	0.68
178	78	19	Avaí	19	7	4	8	24	26	-2	...	1.32	19	2	4	13	10	34	-24	10	0.53
179	79	20	Juventude	19	2	9	8	18	27	-9	...	0.79	19	1	4	14	11	42	-31	7	0.37

180 rows × 21 columns

In [5]:
```
# cria coluna dos anos de cada campeonato

df_concat['Year'] = 0

# //: trunca o número como inteiro. Se usasse só / iria aumentar de 0.05 em 0.05, pois inicia em 1 e 1/20 = 0.05.
years = [2014 + (i // 20) for i in range(len(df_concat))]

df_concat.loc[:, 'Year'] = years
```

In [6]:
```
df_concat.isna().sum() # check NA
```

Out[6]:
```
index       0
Rk          0
Squad       0
MP          0
W           0
D           0
L           0
GF          0
GA          0
GD          0
Pts         0
Pts/MP      0
MP.1        0
W.1         0
D.1         0
L.1         0
GF.1        0
GA.1        0
GD.1        0
Pts.1       0
Pts/MP.1    0
Year        0
dtype: int64
```

In [7]:
```
df_concat['Squad'].unique() # check repetidos
```

Out[7]:
```
array(['Cruzeiro', 'São Paulo', 'Internacional', 'Corinthians',
       'Atlético Mineiro', 'Fluminense', 'Grêmio', 'Atl Paranaense',
       'Santos', 'Flamengo', 'Sport Recife', 'Goiás', 'Figueirense',
       'Coritiba', 'Chapecoense', 'Palmeiras', 'Vitória', 'Bahia',
       'Botafogo (RJ)', 'Criciúma', 'Ponte Preta', 'Avaí',
       'Vasco da Gama', 'Joinville', 'Santa Cruz', 'América (MG)',
       'Atl Goianiense', 'Ceará', 'Paraná', 'Fortaleza', 'CSA',
       'Bragantino', 'Cuiabá', 'Juventude'], dtype=object)
```

In [8]:
```
# cria dataframes separados para resultados em casa e fora

cols_drop_home = ['index', 'MP', 'W', 'D', 'L', 'GF', 'GA', 'GD', 'Pts', 'Pts/MP']
cols_drop_away = ['index', 'MP.1', 'W.1', 'D.1', 'L.1', 'GF.1', 'GA.1', 'GD.1', 'Pts.1', 'Pts/MP.1']
df_home = df_concat.drop(cols_drop_away, axis = 1)
df_away = df_concat.drop(cols_drop_home, axis = 1)
```

In [9]:
```
# seleciona colunas

df_concat = df_concat[['Year','Rk', 'Squad', 'MP', 'W', 'D', 'L', 'GF', 'GA', 'GD', 'Pts', 'Pts/MP',
            'MP.1', 'W.1', 'D.1', 'L.1', 'GF.1', 'GA.1', 'GD.1', 'Pts.1', 'Pts/MP.1']]

df_home = df_home[['Year','Rk', 'Squad', 'MP', 'W', 'D', 'L', 'GF', 'GA', 'GD', 'Pts', 'Pts/MP']]
df_away = df_away[['Year','Rk', 'Squad', 'MP.1', 'W.1', 'D.1', 'L.1', 'GF.1', 'GA.1', 'GD.1', 'Pts.1', 'Pts/MP.1']]
```

In [10]:
```
# renomeia colunas

cols_new_df_concat = ['Ano', 'Pos', 'Clube',
                      'Jc', 'Vc', 'Ec', 'Dc', 'GPc', 'GCc', 'SDc', 'Ptsc', 'Ptsc/Jc',
                      'Jf', 'Vf', 'Ef', 'Df', 'GPf', 'GCf', 'SDf', 'Ptsf', 'Ptsf/Jf']
df_concat.columns = cols_new_df_concat

cols_new_df_home = ['Ano', 'Pos', 'Clube', 'Jc', 'Vc', 'Ec', 'Dc', 'GPc', 'GCc', 'SDc', 'Ptsc', 'Ptsc/Jc']
df_home.columns = cols_new_df_home

cols_new_df_away = ['Ano', 'Pos', 'Clube', 'Jf', 'Vf', 'Ef', 'Df', 'GPf', 'GCf', 'SDf', 'Ptsf', 'Ptsf/Jf']
df_away.columns = cols_new_df_away
```

In [11]:
```
# agrupa resultados fora e em casa

df_total = df_concat.copy()

df_total['J'] = df_total['Jc'] + df_total['Jf']
df_total['V'] = df_total['Vc'] + df_total['Vf']
df_total['E'] = df_total['Ec'] + df_total['Ef']
df_total['D'] = df_total['Dc'] + df_total['Df']
df_total['GP'] = df_total['GPc'] + df_total['GPf']
df_total['GC'] = df_total['GCc'] + df_total['GCf']
df_total['SD'] = df_total['SDc'] + df_total['SDf']
df_total['Pts'] = df_total['Ptsc'] + df_total['Ptsf']
df_total['Pts/J'] = (df_total['Ptsc/Jc'] + df_total['Ptsf/Jf'])/2

df_full = df_total.copy() # tabelão
df_total = df_total[['Ano', 'Pos', 'Clube', 'J', 'V', 'E', 'D', 'GP', 'GC', 'SD', 'Pts', 'Pts/J']]
```

In [12]:
```
# ajusta casas decimais

df_total['Pts/J'] = np.round(df_total['Pts/J'], 2)
df_home['Ptsc/Jc'] = np.round(df_home['Ptsc/Jc'], 2)
df_away['Ptsf/Jf'] = np.round(df_away['Ptsf/Jf'], 2)
df_full['Pts/J'] = np.round(df_full['Pts/J'], 2)
```

In [13]:
```
# grava os csvs

df_total.to_csv('data/df_total.csv', index = False)
df_home.to_csv('data/df_home.csv', index = False)
df_away.to_csv('data/df_away.csv', index = False)
df_full.to_csv('data/df_full.csv', index = False)
```

Desafio: realizar uma análise de dados sobre futebol a partir de dados históricos disponíveis no site https://fbref.com/en/comps/24/history/Serie-A-Seasons.

Passo a passo da solução:

- 1. Coleta dos dados em formato .txt através da fonte.
- 1. Concatenação dos arquivos .txt de cada ano de campeonato em um único arquivo .txt.
- 1. Criação de hipóteses com base nos dados disponíveis.
- 1. Limpeza dos dados e derivação de atributos usando o Python.
- 1. Análise exploratória dos dados com o Power BI a partir de filtros interativos e segmentações.
- 1. Validação das hipóteses geradas na etapa 3.

Premissas:

- A análise contempla exclusivamente o campeonato brasileiro série A.
- A análise contempla exclusivamente as tabelas entre os anos de 2014 a 2022.
- A análise destaca os resultados obtidos pelos clubes como mandante e como visitante.
- A análise destaca os resultados dos 15 clubes de maior torcida do país segundo pesquisa do instituto AtlasIntel realizada em abril de 2023 (mais detalhes em: https://ge.globo.com/futebol/noticia/2023/04/25/maiores-torcidas-do-brasil-pesquisa-atlas-mostra-flamengo-corinthians-e-sao-paulo-no-top-3.ghtml).

Hipóteses relacionadas aos pontos ganhos dos clubes:

- 1. Todo ano o clube que mais pontua como visitante é campeão: verdadeira.
- 1. Todo ano o clube que mais pontua como mandante é campeão: falsa. Em 2016 o clube que mais pontuou como mandante foi o Santos (47) e o campeão foi o Palmeiras, em 2020 foi o Atlético Mineiro (46) e o campeão foi o Flamengo e em 2022 foi o Internacional (44) e o campeão foi o Palmeiras.
- 1. O clube que mais somou pontos no geral é o clube que mais vezes foi campeão: verdadeira. Palmeiras é o clube com mais títulos (3) e mais pontos somados no geral (595).
- 1. O clube que mais somou pontos como visitante é o clube que mais vezes foi campeão: verdadeira. Palmeiras é o clube com mais títulos (3) e mais pontos somados como visitante (247).

Hipóteses relacionadas às vitórias dos clubes:

- 1. Todo ano o clube que mais vence é campeão: verdadeira.
- 1. O clube que mais venceu no geral é o clube que mais vezes foi campeão: falsa. O Flamengo é o clube com mais vitórias no geral (173) e o Palmeiras é o clube com mais títulos (3).

Hipóteses relacionadas aos empates dos clubes:

- 1. O clube que mais empatou no geral nunca foi campeão: verdadeira. O São Paulo é o clube que mais empatou (108).
- 1. O clube campeão do ano nunca ficou entre os 5 clubes que mais empataram: verdadeira. O Palmeiras de 2022 foi o campeão que mais chegou próximo de ficar entre 5 clubes que mais empataram, ficando na 8ª posição.
- 1. O clube campeão do ano nunca empatou mais de 10 vezes: falsa. O Palmeiras em 2018 e em 2022 ultrapassou a marca, fazendo 11 e 12 empates respectivamente.

Hipóteses relacionadas aos gols dos clubes:

- 1. O clube com maior saldo de gols é o clube com mais títulos: falsa. O flamengo é o clube com maior saldo de gols (172) e o Palmeiras é clube com mais títulos (3).
- 1. O clube com mais gols marcados é o clube com mais títulos: falsa. O flamengo é o clube com mais gols marcados (534) e o Palmeiras é o clube com mais títulos (3).