BDSE Onboarding weeks workshop Python and Pandas dataframes

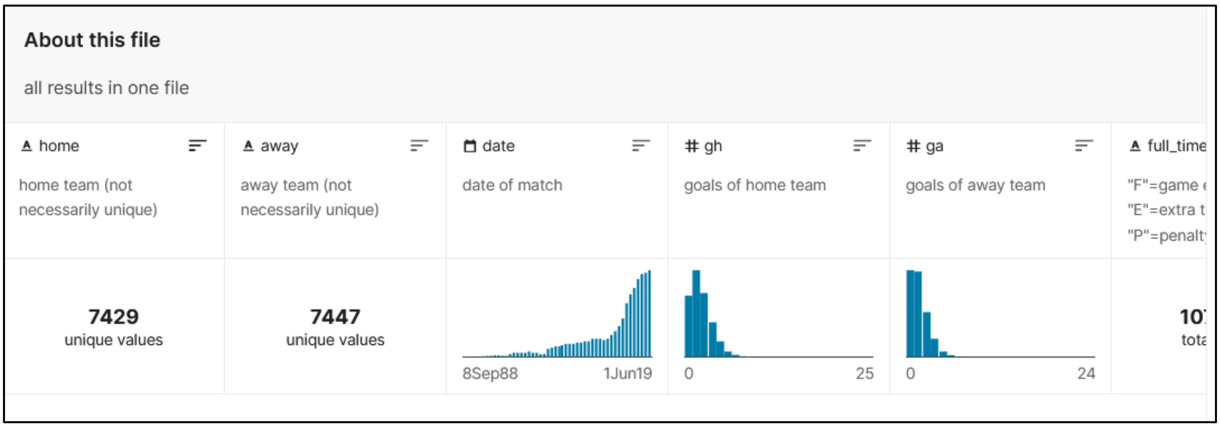
Prerequisites

You

- Have started the online course on DataCamp: https://app.datacamp.com/learn/courses/data-manipulation-with-pandas
- Have a local environment set up to program Python. Could be Visual Studio Code, Pycharm or whatever
- Know how to “pip” …
- Know where “pip” will store your packages …
- …

Exercise Pandas

1. Go to https://www.kaggle.com/datasets/schochastics/domestic-football-results-from-1888-to-2019
2. Download the data: football_results.csv
3. Read the data into a dataframe
4. Print first rows of the dataframe

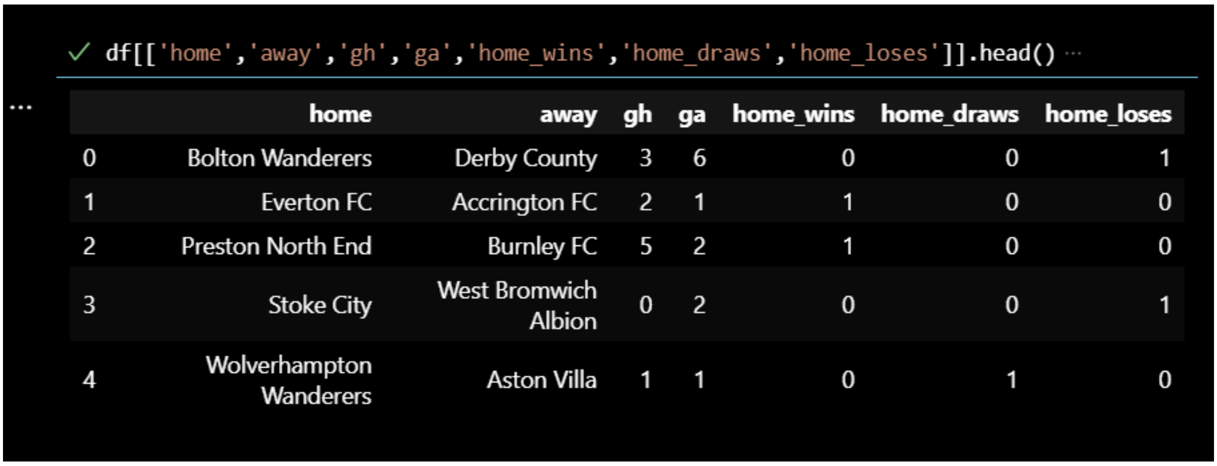


About this file

all results in one file

home	away	date	gh	ga	full_time
home team (not necessarily unique)	away team (not necessarily unique)	date of match	goals of home team	goals of away team	"F"=game ("E"=extra t "P"=penalt
7429 unique values	7447 unique values	8Sep88 1Jun19	0 25	0 24	10 tota

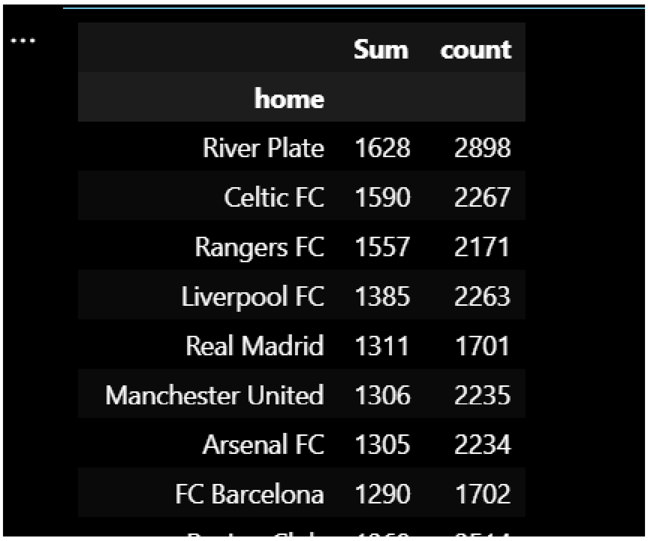
5. Check the information on the website, i.e.
   a. Number of unique ‘home’ teams: 7429
   b. Number of ‘away teams: 7447
6. Find the teams that are mentioned in the ‘away ‘ column but never in the ‘home’ column
   a. Hint turn both columns into a socalled ‘set’
   b. And google to find a way to find the difference between two sets
   c. The number of away teams – home teams should be 35. Note 7447-7429= 16! So, is also the other way around : there are home teams which are never mentioned as away team (17)
7. Add a column ‘home_wins’, that should contain a 1 if the team won, a 0 (zero) otherwise . Use lambda function
8. Similar add column ‘home_draws’ and ‘home_loses’. It should look like



```
df[['home','away','gh','ga','home_wins','home_draws','home_loses']].head()
```

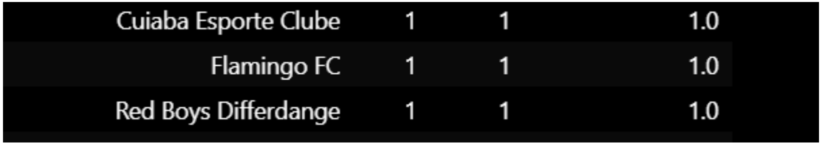
	home	away	gh	ga	home_wins	home_draws	home_loses
0	Bolton Wanderers	Derby County	3	6	0	0	1
1	Everton FC	Accrington FC	2	1	1	0	0
2	Preston North End	Burnley FC	5	2	1	0	0
3	Stoke City	West Bromwich Albion	0	2	0	0	1
4	Wolverhampton Wanderers	Aston Villa	1	1	0	1	0

9. Which team won the most home games in total? Hint , sum ‘home_wins’ for each home team and order desc. It should look like the picture below. Count has been added to be able to calculate % of won games in the next exercise



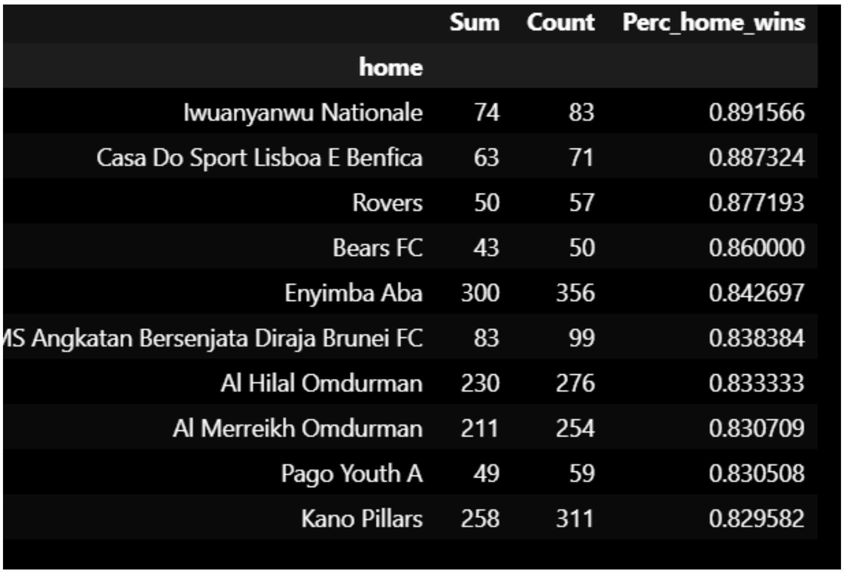
	Sum	count
home		
River Plate	1628	2898
Celtic FC	1590	2267
Rangers FC	1557	2171
Liverpool FC	1385	2263
Real Madrid	1311	1701
Manchester United	1306	2235
Arsenal FC	1305	2234
FC Barcelona	1290	1702

10. In order to pick out unbeatable home teams we need a percentage (numberOfWins)/(NumberofGames). But when to figure out which team is the best you should consider teams with at least , say 50 home games. Since there are teams with only 1 home game, which they won, are these teams considered to be the best? Maybe not…



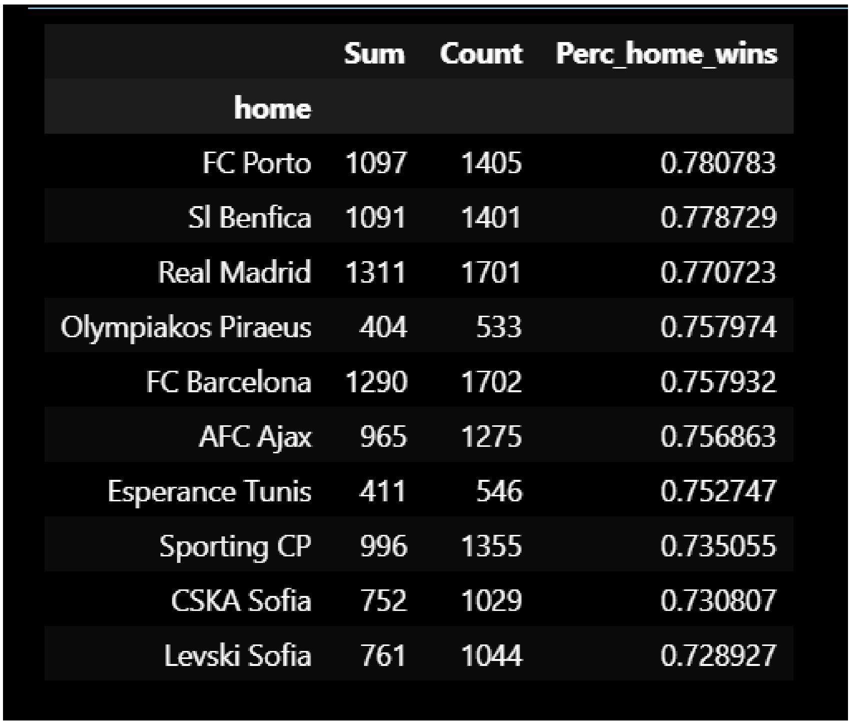
Cuiaba Esporte Clube	1	1	1.0
Flamingo FC	1	1	1.0
Red Boys Differdange	1	1	1.0

So a restriction to the minimum amount of home games will help. This would be the list for the minimum set at 50



	Sum	Count	Perc_home_wins
home			
Iwuanyanwu Nationale	74	83	0.891566
Casa Do Sport Lisboa E Benfica	63	71	0.887324
Rovers	50	57	0.877193
Bears FC	43	50	0.860000
Enyimba Aba	300	356	0.842697
IS Angkatan Bersenjata Diraja Brunei FC	83	99	0.838384
Al Hilal Omdurman	230	276	0.833333
Al Merreikh Omdurman	211	254	0.830709
Pago Youth A	49	59	0.830508
Kano Pillars	258	311	0.829582

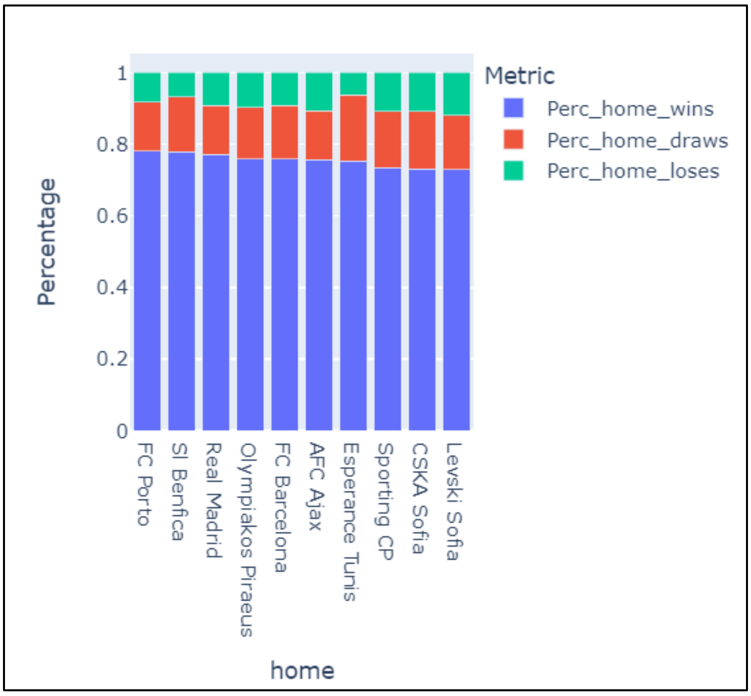
Or , minimum set at 500



	Sum	Count	Perc_home_wins
home			
FC Porto	1097	1405	0.780783
Sl Benfica	1091	1401	0.778729
Real Madrid	1311	1701	0.770723
Olympiakos Piraeus	404	533	0.757974
FC Barcelona	1290	1702	0.757932
AFC Ajax	965	1275	0.756863
Esperance Tunis	411	546	0.752747
Sporting CP	996	1355	0.735055
CSKA Sofia	752	1029	0.730807
Levski Sofia	761	1044	0.728927

Advanced Python

11. To start with some visualisations you will produce a plot like below


Percentage
1
0.8
0.6
0.4
0.2
0
FC Porto
SI Benfica
Real Madrid
Olympiakos Piraeus
FC Barcelona
AFC Ajax
Esperance Tunis
Sporting CP
CSKA Sofia
Levski Sofia
home
Metric
Perc_home_wins
Perc_home_draws
Perc_home_loses