Jennifer Fajardo

Post-Pandemic Play: Evaluating Home Advantage in European Soccer After COVID-19

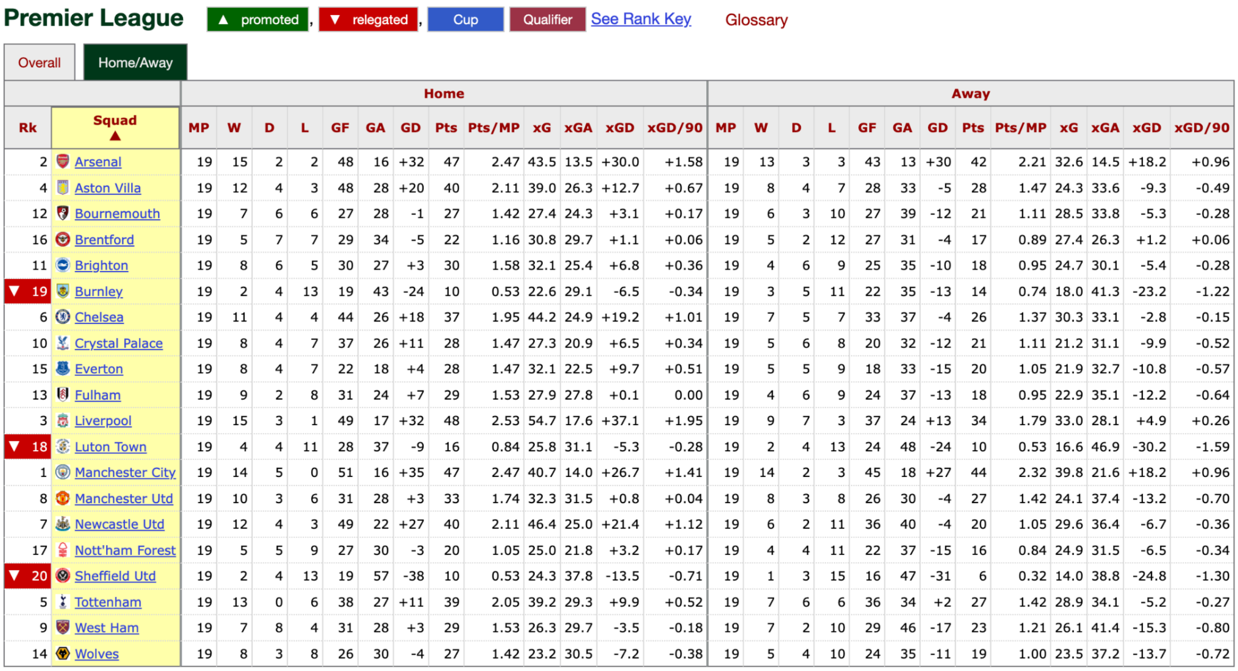
Checkpoint 4

Pipeline Overview & Target Grain

My goal for this checkpoint was to turn raw tables into two clean sheets of data– one for the Premier League and one for La Liga. I was unable to download files directly from FBref.com, so I copied and pasted each season's results manually (Ctrl+C / Ctrl+V) into an Excel workbook. Before copying each table from the site, I sorted the squad name column alphabetically (A-Z) to keep the order as consistent as possible across seasons. In both the Premier League and La Liga, not all teams appear every year as the bottom three teams are relegated at the end of each season and replaced, which explains the little bit of inconsistency across years.

To gather each season's team results on FBref.com I went to Menu → Competitions → La Liga / Premier League → selected the season → then scrolled down to the league table

Figure 1

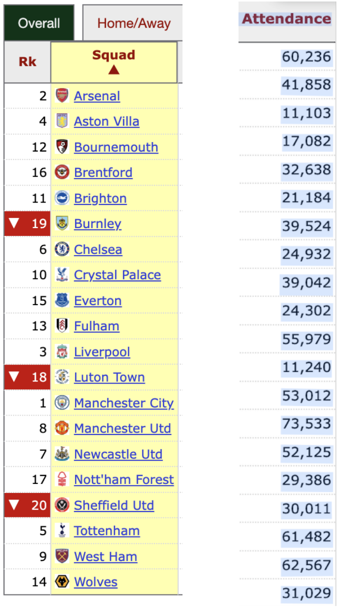


		Home													Away												
Rk	Squad ▲	MP	W	D	L	GF	GA	GD	Pts	Pts/MP	xG	xGA	xGD	xGD/90	MP	W	D	L	GF	GA	GD	Pts	Pts/MP	xG	xGA	xGD	xGD/90
2	Arsenal	19	15	2	2	48	16	+32	47	2.47	43.5	13.5	+30.0	+1.58	19	13	3	3	43	13	+30	42	2.21	32.6	14.5	+18.2	+0.96
4	Aston Villa	19	12	4	3	48	28	+20	40	2.11	39.0	26.3	+12.7	+0.67	19	8	4	7	28	33	-5	28	1.47	24.3	33.6	-9.3	-0.49
12	Bournemouth	19	7	6	6	27	28	-1	27	1.42	27.4	24.3	+3.1	+0.17	19	6	3	10	27	39	-12	21	1.11	28.5	33.8	-5.3	-0.28
16	Brentford	19	5	7	7	29	34	-5	22	1.16	30.8	29.7	+1.1	+0.06	19	5	2	12	27	31	-4	17	0.89	27.4	26.3	+1.2	+0.06
11	Brighton	19	8	6	5	30	27	+3	30	1.58	32.1	25.4	+6.8	+0.36	19	4	6	9	25	35	-10	18	0.95	24.7	30.1	-5.4	-0.28
▼ 19	Burnley	19	2	4	13	19	43	-24	10	0.53	22.6	29.1	-6.5	-0.34	19	3	5	11	22	35	-13	14	0.74	18.0	41.3	-23.2	-1.22
6	Chelsea	19	11	4	4	44	26	+18	37	1.95	44.2	24.9	+19.2	+1.01	19	7	5	7	33	37	-4	26	1.37	30.3	33.1	-2.8	-0.15
10	Crystal Palace	19	8	4	7	37	26	+11	28	1.47	27.3	20.9	+6.5	+0.34	19	5	6	8	20	32	-12	21	1.11	21.2	31.1	-9.9	-0.52
15	Everton	19	8	4	7	22	18	+4	28	1.47	32.1	22.5	+9.7	+0.51	19	5	5	9	18	33	-15	20	1.05	21.9	32.7	-10.8	-0.57
13	Fulham	19	9	2	8	31	24	+7	29	1.53	27.9	27.8	+0.1	0.00	19	4	6	9	24	37	-13	18	0.95	22.9	35.1	-12.2	-0.64
3	Liverpool	19	15	3	1	49	17	+32	48	2.53	54.7	17.6	+37.1	+1.95	19	9	7	3	37	24	+13	34	1.79	33.0	28.1	+4.9	+0.26
▼ 18	Luton Town	19	4	4	11	28	37	-9	16	0.84	25.8	31.1	-5.3	-0.28	19	2	4	13	24	48	-24	10	0.53	16.6	46.9	-30.2	-1.59
1	Manchester City	19	14	5	0	51	16	+35	47	2.47	40.7	14.0	+26.7	+1.41	19	14	2	3	45	18	+27	44	2.32	39.8	21.6	+18.2	+0.96
8	Manchester Utd	19	10	3	6	31	28	+3	33	1.74	32.3	31.5	+0.8	+0.04	19	8	3	8	26	30	-4	27	1.42	24.1	37.4	-13.2	-0.70
7	Newcastle Utd	19	12	4	3	49	22	+27	40	2.11	46.4	25.0	+21.4	+1.12	19	6	2	11	36	40	-4	20	1.05	29.6	36.4	-6.7	-0.36
17	Nott'ham Forest	19	5	5	9	27	30	-3	20	1.05	25.0	21.8	+3.2	+0.17	19	4	4	11	22	37	-15	16	0.84	24.9	31.5	-6.5	-0.34
▼ 20	Sheffield Utd	19	2	4	13	19	57	-38	10	0.53	24.3	37.8	-13.5	-0.71	19	1	3	15	16	47	-31	6	0.32	14.0	38.8	-24.8	-1.30
5	Tottenham	19	13	0	6	38	27	+11	39	2.05	39.2	29.3	+9.9	+0.52	19	7	6	6	36	34	+2	27	1.42	28.9	34.1	-5.2	-0.27
9	West Ham	19	7	8	4	31	28	+3	29	1.53	26.3	29.7	-3.5	-0.18	19	7	2	10	29	46	-17	23	1.21	26.1	41.4	-15.3	-0.80
14	Wolves	19	8	3	8	26	30	-4	27	1.42	23.2	30.5	-7.2	-0.38	19	5	4	10	24	35	-11	19	1.00	23.5	37.2	-13.7	-0.72

**this example are stats from the 2023-24 Premier League season*

To collect the average home attendance per game, I navigated to the **Overall** tab shown above and copied the attendance column into a new column I made in the Excel file. I made sure the squad names were sorted alphabetically to keep the data aligned.

Figure 2



Rk	Squad ▲	Attendance
2	Arsenal	60,236
4	Aston Villa	41,858
12	Bournemouth	11,103
16	Brentford	17,082
11	Brighton	32,638
▼ 19	Burnley	21,184
6	Chelsea	39,524
10	Crystal Palace	24,932
15	Everton	39,042
13	Fulham	24,302
3	Liverpool	55,979
▼ 18	Luton Town	11,240
1	Manchester City	53,012
8	Manchester Utd	73,533
7	Newcastle Utd	52,125
17	Nott'ham Forest	29,386
▼ 20	Sheffield Utd	30,011
5	Tottenham	61,482
9	West Ham	62,567
14	Wolves	31,029

I repeated this process three times for each league for the 2018-19, 2020-21, and 2023-24 seasons. You will see this in the first spreadsheet, **Raw Data**, in the attached workbook "DATASET LL+PL.xlsx." This raw dataset is what feeds into the next two spreadsheets, **Clean Data LL** and **Clean Data PL.**

- A row in the clean tables = one team x one season x one league
  - All clean tables include the same fields: Attendance, MP, W, WP, Pts/MP, Home GD, Away GD, PG

ID & Mapping Strategy

The unique key in my cleaned dataset is a combination of the columns squad name and season. There were no name collisions as each squad was spelled out completely in the raw data, making it clear to read. The only challenge I came across was team turnover across seasons. To keep the final tables comparable, I included only the teams that remained in the league for all three seasons. Each league has 20 teams per season, but only 12 teams stayed consistent in La Liga and 14 in the Premier League. These two final clean tables are at the top of both **clean data** spreadsheets.

Data Cleaning, Transformation, and Validation

I began by copying the **Raw Data** sheet so I could work to clean it on a separate version while keeping the original untouched. My step-by-step cleaning process is outlined below.

1. Normalized season names to **YYYY-YY**

2. Standardized league labels to **La Liga** and **Premier League**

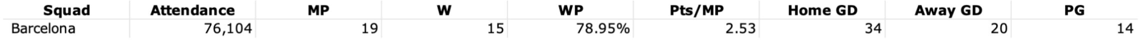
3. Deleted the home and away stats I did not need: Rank, D, L, GF, GA, Pts, xG, xGA, xGD, xGD/90

4. Kept the following columns of stats: Squad, Attendance, MP, W, Pts/MP, Home GD, Away GD
    - No units were converted

5. Computed the following new variables:
    - Win Percentage (WP) = home wins ÷ total home games
        - Why? If home advantage declined, home win % should drop during COVID.
    - Performance Gap (PG) = home goal differential - away goal differential
        - Why? large positive gap would mean strong home-field advantage

6. Once extra data columns were deleted and new variables were computed, I duplicated the spreadsheet naming one **Clean Data LL** and the other **Clean Data PL.**
    - Dataset Size
      Total tables = 6 **one per season per league*
      La Liga: 20 teams x 3 seasons = 60 rows
      Premier League: 20 teams x 3 season = 60 rows

    - Example Row

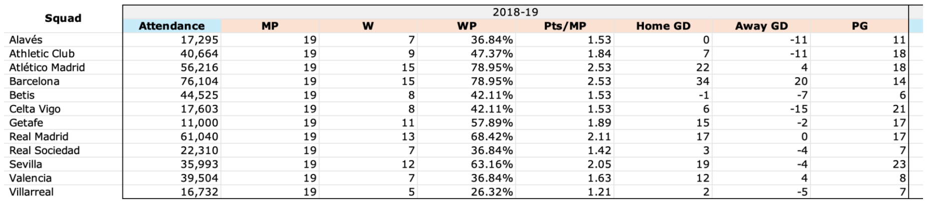


Squad	Attendance	MP	W	WP	Pts/MP	Home GD	Away GD	PG
Barcelona	76,104	19	15	78.95%	2.53	34	20	14

7. To create **one** final clean table for each league, I identified which teams appeared in all three seasons. Any teams that were relegated or promoted in a given year were removed so the comparison would remain consistent across seasons.
    - This final table is one long dataset with columns grouped by season. Each season group consists of all key metrics. Example shown below.

Figure 3
LA LIGA

**teams that played in all 3 seasons (12/20)*

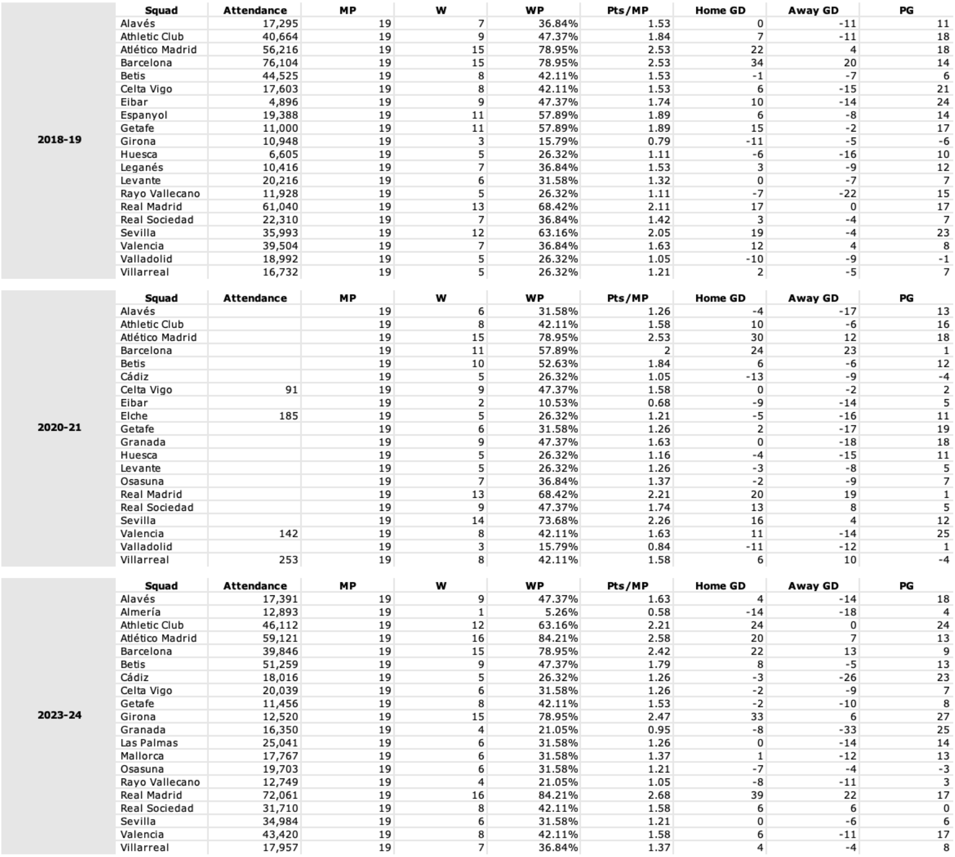


Squad	2018-19							
	Attendance	MP	W	WP	Pts/MP	Home GD	Away GD	PG
Alavés	17,295	19	7	36.84%	1.53	0	-11	11
Athletic Club	40,664	19	9	47.37%	1.84	7	-11	18
Atlético Madrid	56,216	19	15	78.95%	2.53	22	4	18
Barcelona	76,104	19	15	78.95%	2.53	34	20	14
Betis	44,525	19	8	42.11%	1.53	-1	-7	6
Celta Vigo	17,603	19	8	42.11%	1.53	6	-15	21
Getafe	11,000	19	11	57.89%	1.89	15	-2	17
Real Madrid	61,040	19	13	68.42%	2.11	17	0	17
Real Sociedad	22,310	19	7	36.84%	1.42	3	-4	7
Sevilla	35,993	19	12	63.16%	2.05	19	-4	23
Valencia	39,504	19	7	36.84%	1.63	12	4	8
Villarreal	16,732	19	5	26.32%	1.21	2	-5	7

** 2020-21 data is to the right of this, followed by 2023-24 data*

- ○ 6 team rows were deleted from the Premier League sheet, 8 team rows were deleted from La Liga sheet
- ○ Each league spreadsheet now begins with this final clean table containing only the teams that stayed in the league for all three seasons, followed by the three separate season tables with all 20 teams below it.

Figure 4



**all 20 teams per season*

	Squad	Attendance	MP	W	WP	Pts/MP	Home GD	Away GD	PG
2018-19	Alavés	17,295	19	7	36.84%	1.53	0	-11	11
	Athletic Club	40,664	19	9	47.37%	1.84	7	-11	18
	Atlético Madrid	56,216	19	15	78.95%	2.53	22	4	18
	Barcelona	76,104	19	15	78.95%	2.53	34	20	14
	Betis	44,525	19	8	42.11%	1.53	-1	-7	6
	Celta Vigo	17,603	19	8	42.11%	1.53	6	-15	21
	Eibar	4,896	19	9	47.37%	1.74	10	-14	24
	Espanyol	19,388	19	11	57.89%	1.89	6	-8	14
	Getafe	11,000	19	11	57.89%	1.89	15	-2	17
	Girona	10,948	19	3	15.79%	0.79	-11	-5	-6
	Huesca	6,605	19	5	26.32%	1.11	-6	-16	10
	Leganés	10,416	19	7	36.84%	1.53	3	-9	12
	Levante	20,216	19	6	31.58%	1.32	0	-7	7
	Rayo Vallecano	11,928	19	5	26.32%	1.11	-7	-22	15
	Real Madrid	61,040	19	13	68.42%	2.11	17	0	17
	Real Sociedad	22,310	19	7	36.84%	1.42	3	-4	7
	Sevilla	35,993	19	12	63.16%	2.05	19	-4	23
	Valencia	39,504	19	7	36.84%	1.63	12	4	8
	Valladolid	18,992	19	5	26.32%	1.05	-10	-9	-1
	Villarreal	16,732	19	5	26.32%	1.21	2	-5	7

	Squad	Attendance	MP	W	WP	Pts/MP	Home GD	Away GD	PG
2020-21	Alavés		19	6	31.58%	1.26	-4	-17	13
	Athletic Club		19	8	42.11%	1.58	10	-6	16
	Atlético Madrid		19	15	78.95%	2.53	30	12	18
	Barcelona		19	11	57.89%	2	24	23	1
	Betis		19	10	52.63%	1.84	6	-6	12
	Cádiz		19	5	26.32%	1.05	-13	-9	-4
	Celta Vigo	91	19	9	47.37%	1.58	0	-2	2
	Eibar		19	2	10.53%	0.68	-9	-14	5
	Elche	185	19	5	26.32%	1.21	-5	-16	11
	Getafe		19	6	31.58%	1.26	2	-17	19
	Granada		19	9	47.37%	1.63	0	-18	18
	Huesca		19	5	26.32%	1.16	-4	-15	11
	Levante		19	5	26.32%	1.26	-3	-8	5
	Osasuna		19	7	36.84%	1.37	-2	-9	7
	Real Madrid		19	13	68.42%	2.21	20	19	1
	Real Sociedad		19	9	47.37%	1.74	13	8	5
	Sevilla		19	14	73.68%	2.26	16	4	12
	Valencia	142	19	8	42.11%	1.63	11	-14	25
	Valladolid		19	3	15.79%	0.84	-11	-12	1
	Villarreal	253	19	8	42.11%	1.58	6	10	-4

	Squad	Attendance	MP	W	WP	Pts/MP	Home GD	Away GD	PG
2023-24	Alavés	17,391	19	9	47.37%	1.63	4	-14	18
	Almería	12,893	19	1	5.26%	0.58	-14	-18	4
	Athletic Club	46,112	19	12	63.16%	2.21	24	0	24
	Atlético Madrid	59,121	19	16	84.21%	2.58	20	7	13
	Barcelona	39,846	19	15	78.95%	2.42	22	13	9
	Betis	51,259	19	9	47.37%	1.79	8	-5	13
	Cádiz	18,016	19	5	26.32%	1.26	-3	-26	23
	Celta Vigo	20,039	19	6	31.58%	1.26	-2	-9	7
	Getafe	11,456	19	8	42.11%	1.53	-2	-10	8
	Girona	12,520	19	15	78.95%	2.47	33	6	27
	Granada	16,350	19	4	21.05%	0.95	-8	-33	25
	Las Palmas	25,041	19	6	31.58%	1.26	0	-14	14
	Mallorca	17,767	19	6	31.58%	1.37	1	-12	13
	Osasuna	19,703	19	6	31.58%	1.21	-7	-4	-3
	Rayo Vallecano	12,749	19	4	21.05%	1.05	-8	-11	3
	Real Madrid	72,061	19	16	84.21%	2.68	39	22	17
	Real Sociedad	31,710	19	8	42.11%	1.58	6	6	0
	Sevilla	34,984	19	6	31.58%	1.21	0	-6	6
	Valencia	43,420	19	8	42.11%	1.58	6	-11	17
	Villarreal	17,957	19	7	36.84%	1.37	4	-4	8

Data Dictionary is included in the Excel Workbook attached.