Jennifer Fajardo

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

Checkpoint 3

Project Summary & Data Overview

This project analyzes whether the traditional home-field advantage in professional European soccer declined during and after the COVID-19 pandemic. I am focusing on the following three periods across the Premier League and La Liga…

(1) Pre-COVID: 2018-2019
(2) COVID: 2020-2021
(3) Post-COVID: 2023-2024

To evaluate this question, I am using publicly available match-level data on results, goals, scored, and stadium attendance. This data is essential for calculating home win percentages, goal differentials, and the relationship between crowd presence and home performance.

Data Source Summary

My dataset draws from three sources:

- Official Premier League Website – Provides official match schedules, final scores, and crowd attendance. This data is recorded and published directly by the league during and after each game.
- Official la Liga Website – Also contains official match information recorded and published by the league directly.
- FBref.com – An open-source soccer statistics platform that aggregates match results, advanced team metrics, and attendance through StatsBomb and Hudl databases. Data is updated regularly and downloadable in CSV format. This source will serve as a supplement if the official league sites have missing attendance figures or gaps across seasons.

Data Structure & Content

The dataset will include regular-season stats from 2018-2024 for both the Premier League and La Liga. I will organize everything into one clean table. Each row will represent a team (20 teams per league, 40 teams total), and the columns will be grouped by season. Under each season, I will track two key metrics: home goal differential and average home attendance.

To build this, I will download 12 files from FBref– one file per league for each of the six seasons. After that, I'll combine them into one easy-to-read Excel sheet for analysis.

Data Completeness & Consistency

It looks like the Premier League and La Liga websites don't let me download their data directly, so I'll pull most of the information from FBref and double-check it against the official league sites. This actually works well because FBref is easier to use and provides more complete attendance data than La Liga's site.

Because COVID led to cancelled or rescheduled matches, some seasons may have irregular numbers of games. I'll make sure the match counts are equal for each team and season before analyzing the data.

Quality Issues & Potential Biases

The Premier League and La Liga may have brought fans back at different times, which could affect the comparison. There are also team-level differences, for example, some clubs have much stronger fan bases, louder stadiums, or different home environments that could create additional bias. To avoid this bias, I may create separate tables for each league and analyze them individually.

Initial Cleaning / Preparation Plan / Next Steps

(1) *Combine and organize all files:* I will download the 12 FBref files (one per league per season) and merge them into a single clean Excel table.

(2) *Verify match counts:* Because COVID seasons had cancelled or rescheduled games, I will confirm that each team has a full and accurate set of matches for every year. Any irregular season lengths will be documented and handled before calculating averages.

(3) *Cross-check key metrics against official league sites:* Although FBref will be my main source, I will verify goal totals and attendance patterns using the Premier League and La Liga websites. This helps ensure accuracy, especially in seasons where attendance reporting varies.

(4) *Clean and format metrics:* I will calculate each team's home goal differential and average home attendance for every season if not given. All numbers will be checked for outliers or errors, and I will standardize formats so the final table is easy to read.

(5) *Prepare separate league tables if needed:* If the data shows meaningful differences, I will split the dataset into two separate league tables for cleaner comparisons.

(6) *Run basic exploratory checks:* Once the table is complete, I will run summary statistics and visual checks to confirm that the data trends make sense before moving into deeper analysis.