Jennifer Fajardo
November 10, 2025

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

Topic Summary

This project analyzes whether the traditional "home advantage" in professional European soccer has declined since the COVID-19 pandemic. My focus is measuring how crowd absence and subsequent return affected team performance outcomes. The research question remains the same: *Has the frequency and magnitude of home wins decreased in major European soccer leagues since COVID-19, and has this trend persisted after fans returned?* I will compare win rates and goal differentials before, during, and after the pandemic using consistent metrics across the Premier League and La Liga.

Understanding whether home advantage has declined is important for multiple stakeholders. Coaches and team analysts rely on home-field factors to plan strategies, while sports bettors and data analysts use these trends to model outcomes. If home advantage has diminished, it shifts how teams prepare for games and how leagues interpret competitive balance while not relying on home-field energy. Additionally, this topic contributes to the broader study of how external factors, like fan attendance, affect athlete performance and psychology in professional sports.

Data Scope & Sources

The focus is on individual soccer matches from the Premier League and La Liga. The dataset includes only regular season games, excluding friendlies, qualifiers, and playoffs. This setup will allow reliable league-level comparisons across three distinct pandemic periods: Pre-COVID (2017–2019), COVID (2020–2021), and Post-COVID (2022–2024).

Primary data will be collected directly from the official Premier League and La Liga websites, both of which publish publicly accessible match results, team statistics, and attendance figures. To ensure completeness, supplementary information will also be obtained from Fbref.com, which compiles open-source football statistics through the Hudl StatsBomb database. Data will be accessed via direct CSV downloads. This source will help me if there are missing values or attendance reporting gaps on the official league sites.

Key Metrics & Modeling Direction

1. Home Win Percentage: *Ratio of home wins to total home matches in each season.*
2. Goal Differential: *Average goals scored by home teams minus goals scored by away teams per match.*
3. Attendance Impact Index: *Relationship between reported stadium attendance and home win probability across the sample.*

First, I might develop a multiple linear or logistic regression model to test whether stadium attendance (representing fan atmosphere) predicts home win probability and goal differential. Then, ANOVA and t-tests can be used to compare mean home win percentages and goal differentials across pre-, during-, and post-COVID periods.

Mini Literature Scan

- Leitner MC, Daumann F, Follert F, Richlan F. (2023): found that home advantage dropped sharply during COVID-19 “ghost games,” mainly due to reduced crowd pressure and referee bias.
- H. Almeida, C., & S. Leite, W. (2021): showed that the decline varied by league, suggesting that contextual factors like team strength and competition style influenced whether the home advantage fully disappeared or only weakened.
- Benz, L.S., Lopez, M.J. (2023): used Poisson regression to show pandemic-period shifts in home/away scoring patterns across multiple European leagues.

Risks & Ethical Considerations

I will be using only public match and attendance data, so privacy risks are minimal. The main challenges involve data consistency and potential bias in interpretation. To reduce analytical bias, I’ll control for factors like team strength and season differences when comparing results. All data will be cited properly, and findings will be reported objectively without favoring any team or league.

Next Steps

Before Checkpoint 3, I will collect and organize all match and attendance data from the Premier League, La Liga, and Fbref for the 2017-2024 seasons. I’ll clean and format the data so variables and dates align across sources, then run basic exploratory checks like summary statistics and visualizations to confirm accuracy before starting the analysis phase.