# Premier League Data Inventory & Schema Reference

Project: Predicting the 2025/26 Premier League Champions.  
This document catalogues each dataset you loaded, including the data grain, coverage, important columns and suggested join keys. Use it as a quick schema reference while you engineer features and build the predictive model.

## 1. E0 (1).csv — EPL Match Results & Betting Odds

Grain: One Premier‑League match.

Coverage: Seasons 1993/94 → 2023/24 (≈ 106 columns).

Key identifiers: Div, Date, Time, HomeTeam, AwayTeam.

Outcome fields: FTHG, FTAG, FTR, HTHG, HTAG.

Bookmaker odds columns: B365H/D/A, Pinnacle, BetVictor, … plus Asian Handicap (CAHH/CAHA) and Over/Under (CO<2.5>, …).

Odds aggregates: MaxH/D/A, AvgH/D/A, MaxCAHH etc.

Candidate PK: (Date, HomeTeam, AwayTeam).

Typical joins: Club name ↔ dimension tables.

## 2. Attendance Data.csv — Stadium Capacity & Average Attendance

Grain: One club‑stadium record.

Columns: #, Club, Stadium, Capacity, Avg Attendance.

Clean‑up: Remove commas, cast to int.

Join on Club.

## 3. Club Value.csv — Squad Market Valuations (as of 1 Jul 2025)

Grain: One club snapshot.

Columns: #, Club, Club Value Jul 1 2025, Current value, % (YoY change).

Parse strings like “€1.33bn” ➜ float(€).

Join on Club.

## 4. Club wages.csv — Team Wage Bill 2024/25

Grain: Club‑season.

Columns: Rk, Squad (club), # Pl, Weekly Wages, Annual Wages, % Estimated.

Currency field includes £ plus €/$ in parentheses — strip symbols.

Join on Squad ↔ Club.

## 5. Injury list 2002‑2016.csv — Historical Injury Log

Grain: One injury incident.

Columns: Date Reported, Injury, Injury (group), Player, Season, Team, Number of Records, Winner.

Use: derive injury‑burden features per club/season.

Join on Team and/or Player.

## 6. player\_salaries.csv — 2024/25 Player‑Level Wages

Grain: One contracted player.

Columns: Player Name, Weekly Wage, Yearly Salary, Age, Position, Nationality, Team.

Join on Player Name (to stats) or Team (to club tables).

## 7. players\_epl\_14‑15.csv — Player Performance Metrics 2014/15 → 2021/22

Grain: Player‑season.

Key columns: id, player\_name, games, minutes (time), goals, xG, assists, xA, shots, key\_passes, cards, position, team\_title, npg, npxG, xGChain, xGBuildup.

Join on player\_name / team\_title.

## 8. shots\_epl\_14‑15.csv — Event‑Level Shot Data

Grain: Every shot event.

Important columns: id (event), minute, result, X, Y, xG, player, h\_a, situation, season, shotType, match\_id, h\_team, a\_team, h\_goals, a\_goals, date, player\_assisted, lastAction.

Aggregate for shooting‑quality features per player or club.

## 9. Posession data 24‑24.csv — Team Stats Season 2024/25

Grain: Club‑season.

Columns: CLUB, POSS%, SEASON, M, W, D, L, G, GC, XG, XGA, SH, SOT, PASS–ATT, PASS, FOUL, YC, RC, Unnamed: 18.

Join on CLUB name.

## 10. Premier League Managers.csv — Historical Manager Stints

Grain: Manager‑club stint.

Columns: Name, Nat., Club, From, Until, Duration (days), Years in League, Ref.

First row blank → drop.

Join on Club; derive current manager & tenure length.