WHY ARSENAL IS ALWAYS 2ND PLACE

By Matheus Grover

PROBLEM AND MOTIVATION

- Five seasons under Coach Mikel Arteta = 1 Trophy 3 consecutive 2nd place finishes
- 2022-23 collapse: 248 / 296 days in 1st, finished 2nd
- 8th -> 8th -> 5th -> 2nd -> 2nd -> 2nd
- No CL -> No CL -> No CL -> No CL -> Quarter Finals -> Semi Finals
- Every Arsenal fan knows our problem; we're consistently getting better but still haven't pushed over the line yet.

HYPOTHESIS

- After discussing with my Applied Consultant, he theorized that a tactical shift after spring 2023 changed how we play.
- The Pre-shift Arsenal has:
- Lower possession
- Higher xG
- Less cards
- Conceded and scored more goals
- In 2022-23 teams played more attacking against us

DATA

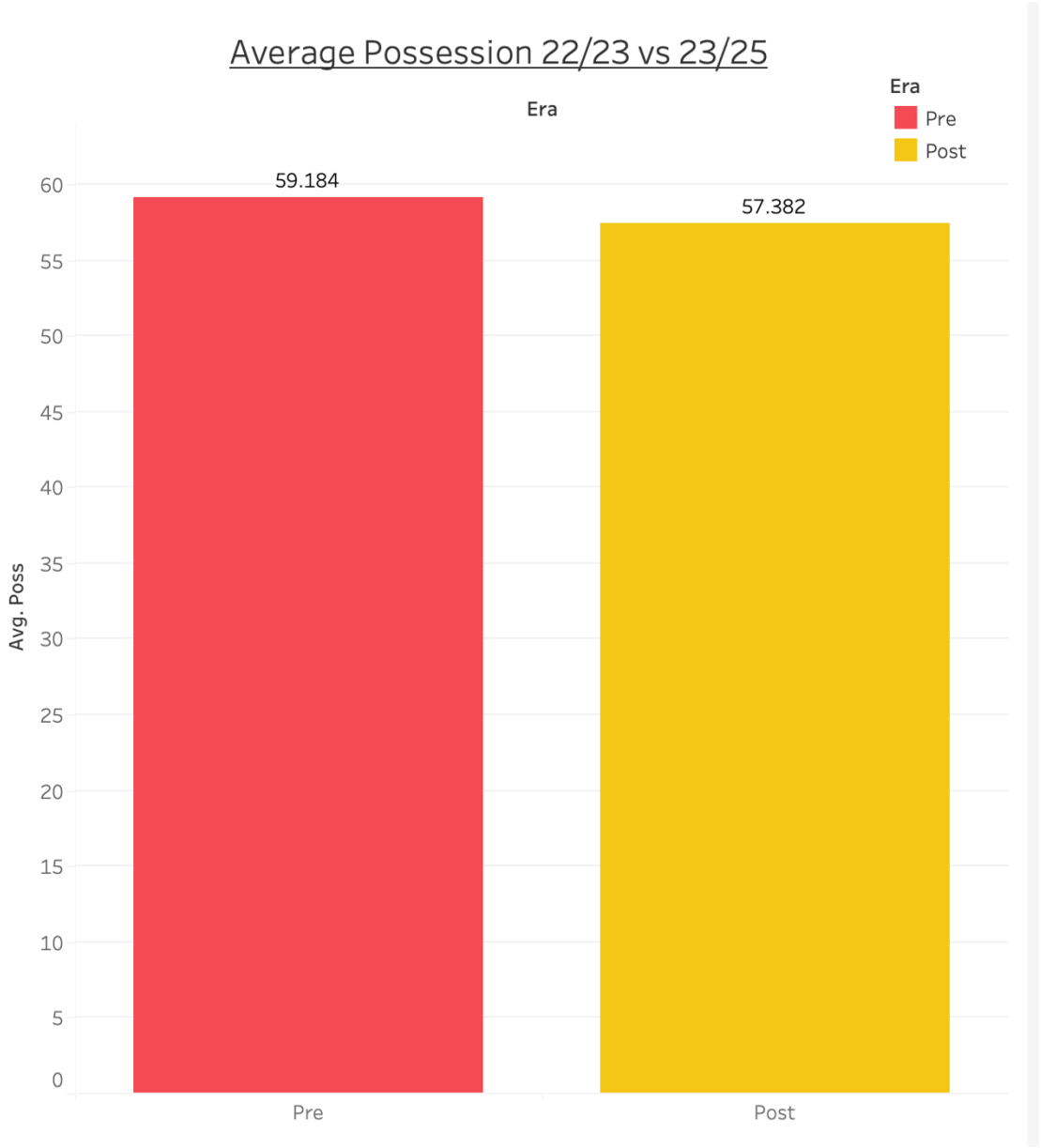
- Fbref match logs: Arsenal, Man City, Liverpool (2022-2025)
- Added discipline cards through a left join of a different google sheet
- Created Variables:

1. Season, Era (Pre = 2022/23, Post = 2023/24, 2024/25)
2. Opponent Difficulty(Top -> Non-League)
3. Formation Style: (Defensive/Balanced/Attacking)
4. Points, Levels for Formation and Difficulty

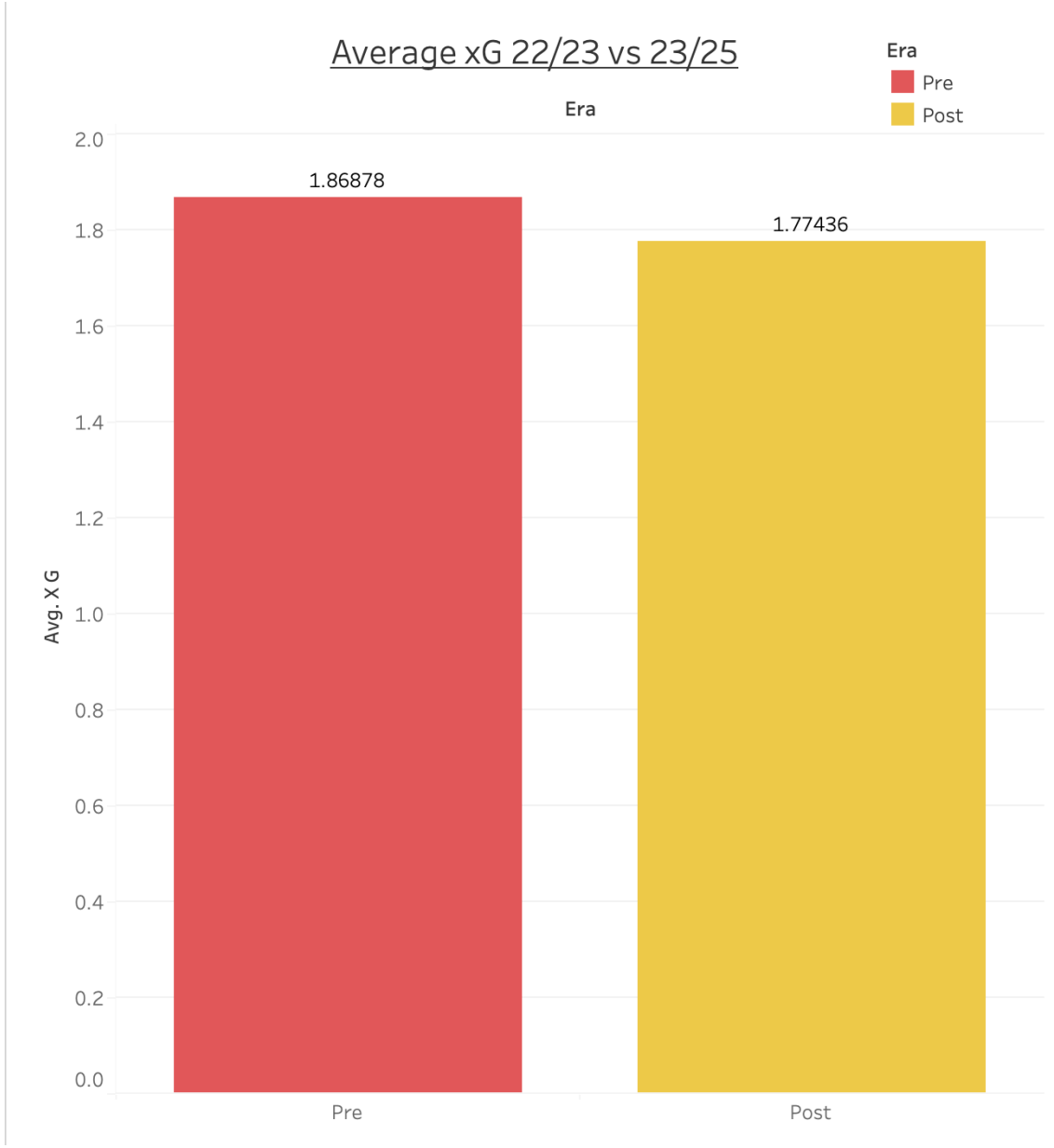
EDA

- Two forms of EDA, before I opened the data, I wrote down 9 predictions about what it would look like. Then created 15+ tableau figures answering these questions.
- This helped me get a sense of what I was working with, what I was missing, what I was right and wrong about.
- After I created all the earlier variables, I still wasn't ready for analysis, my consultant and I drafted up the Hypothesis questions and I created another round of figures answering them.

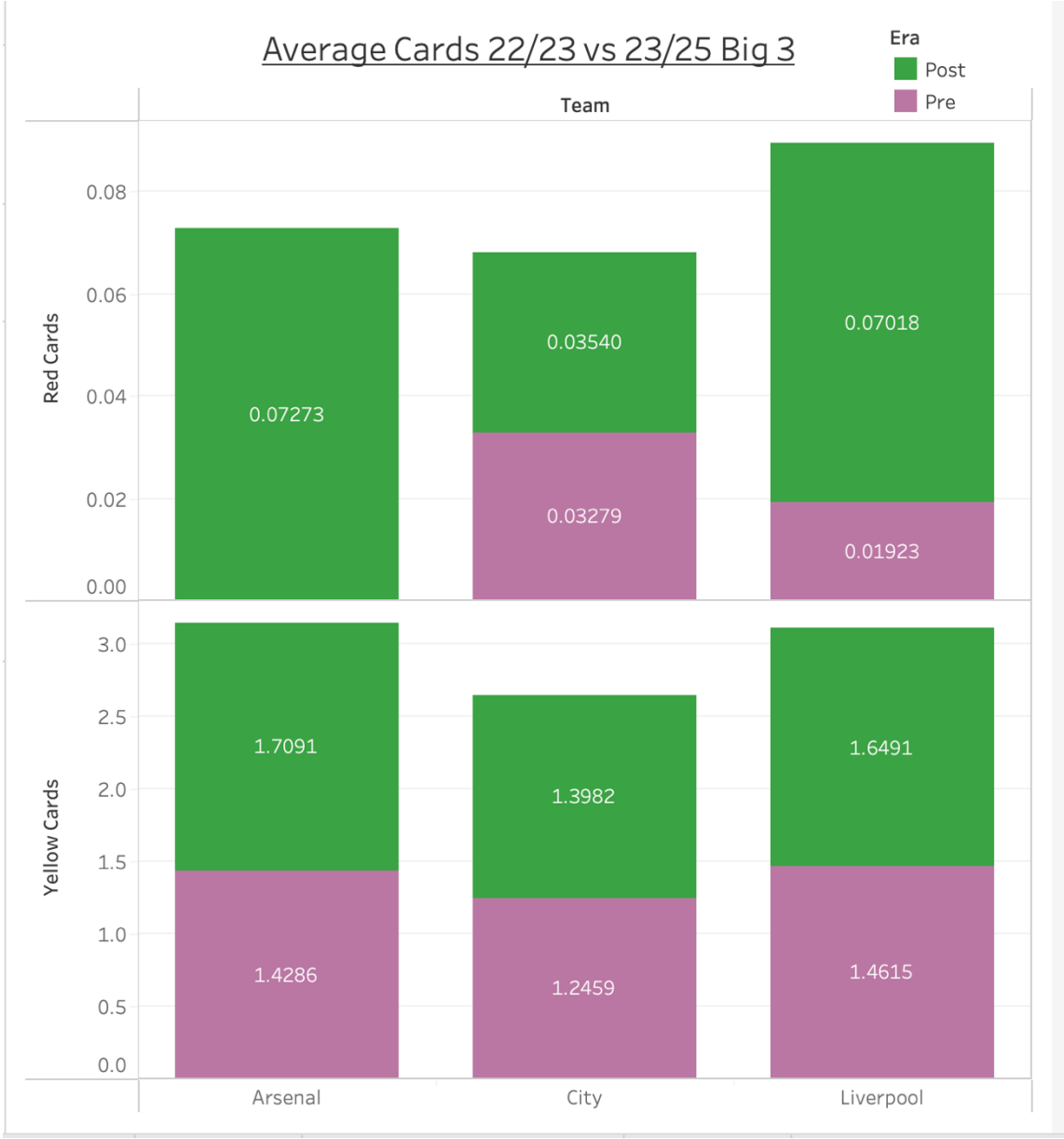
ARSENAL 2022-23 HAS LOWER POSSESSION THAN 2023-2025



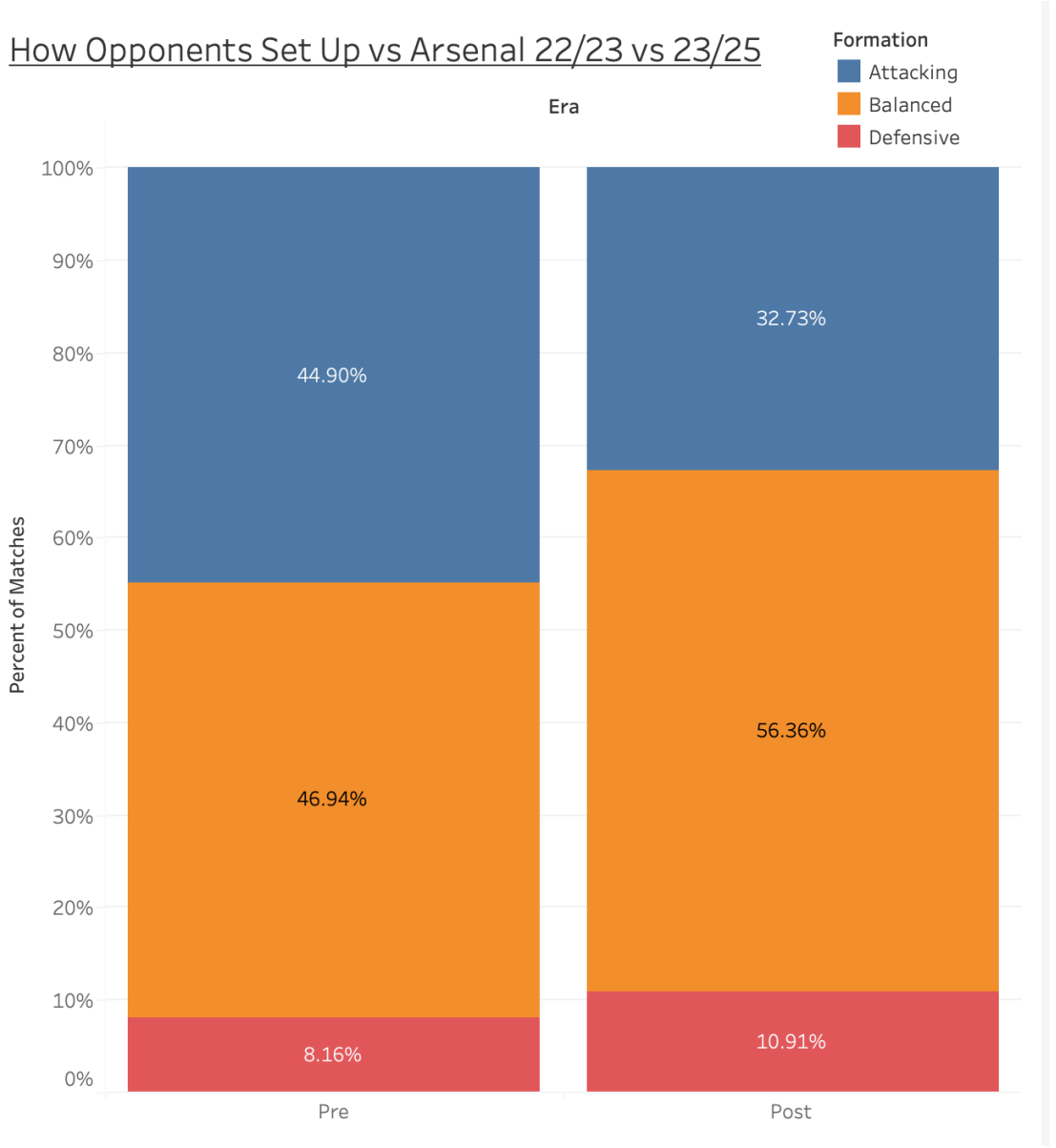
ARSENAL 2022-23 HAS HIGHER XG THAN 2023-2025



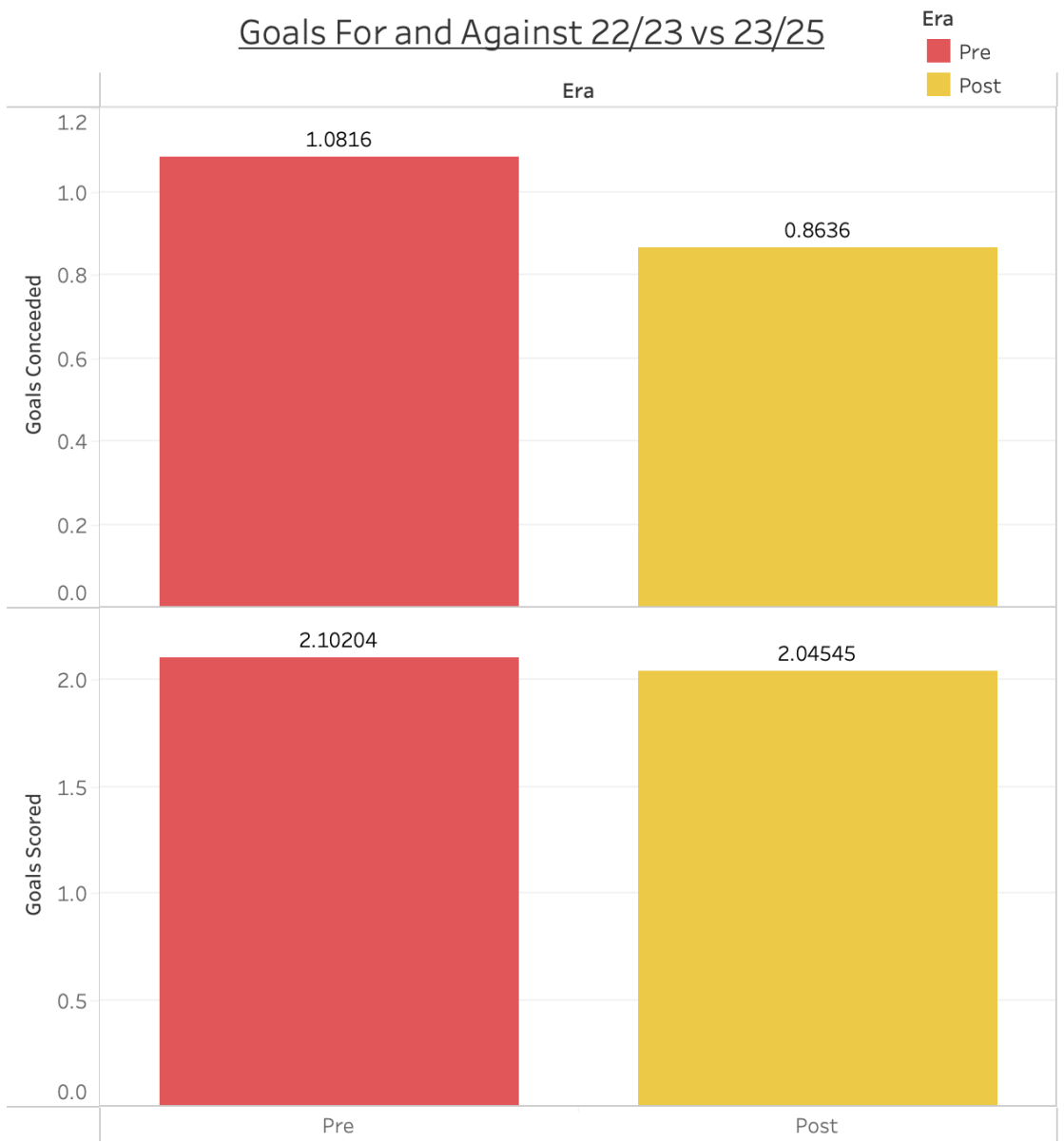
ARSENAL 2022-23 HAS LESS CARDS THAN 2023-2025



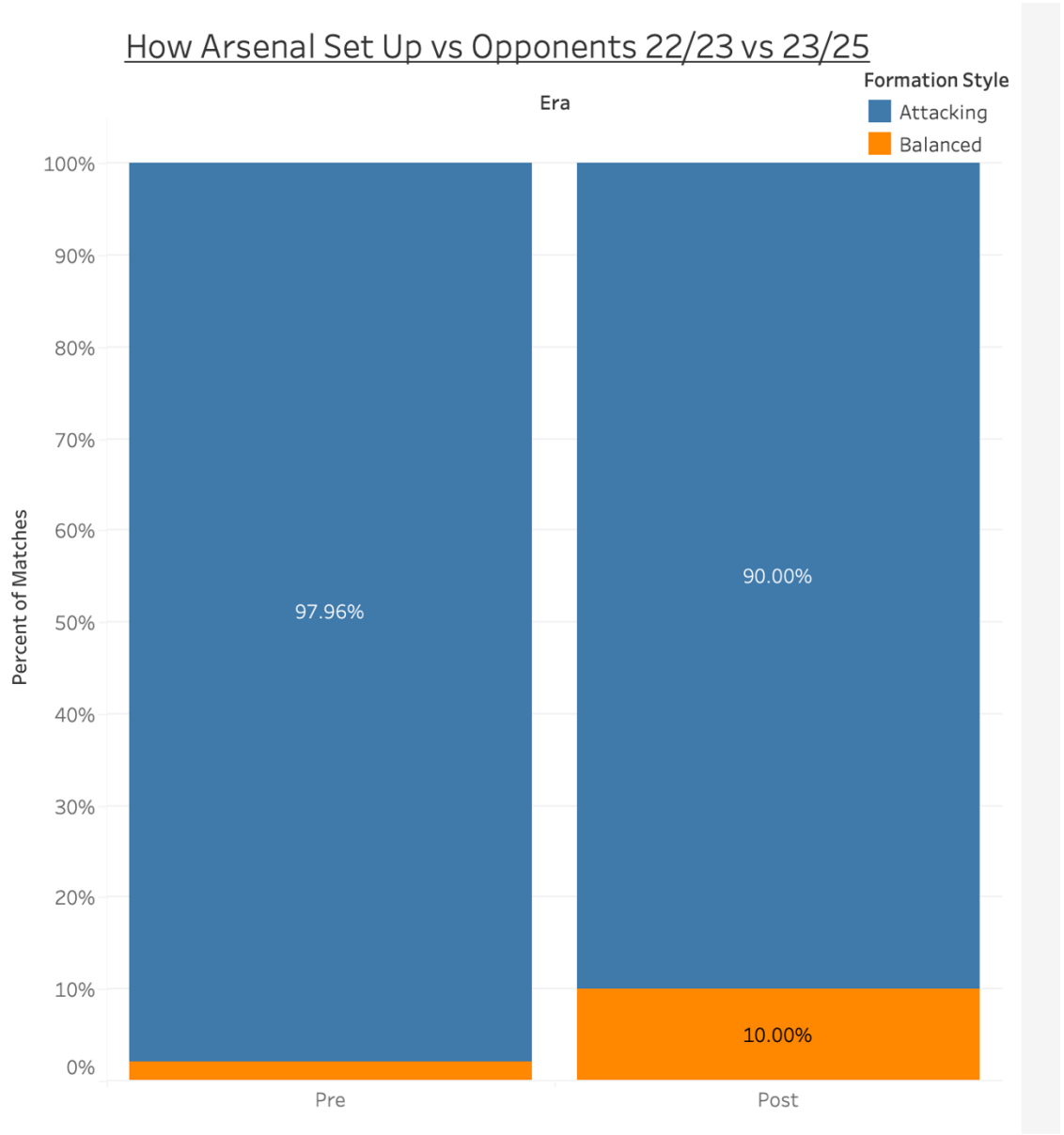
ARSENAL 2022-23 TEAMS PLAYED MORE ATTACKING AGAINST US



ARSENAL 2022-23 CONCEDED AND SCORED MORE GOALS



FINALLY, ARSENAL SWITCH THEIR PLAYSTYLE AFTER 2022-23



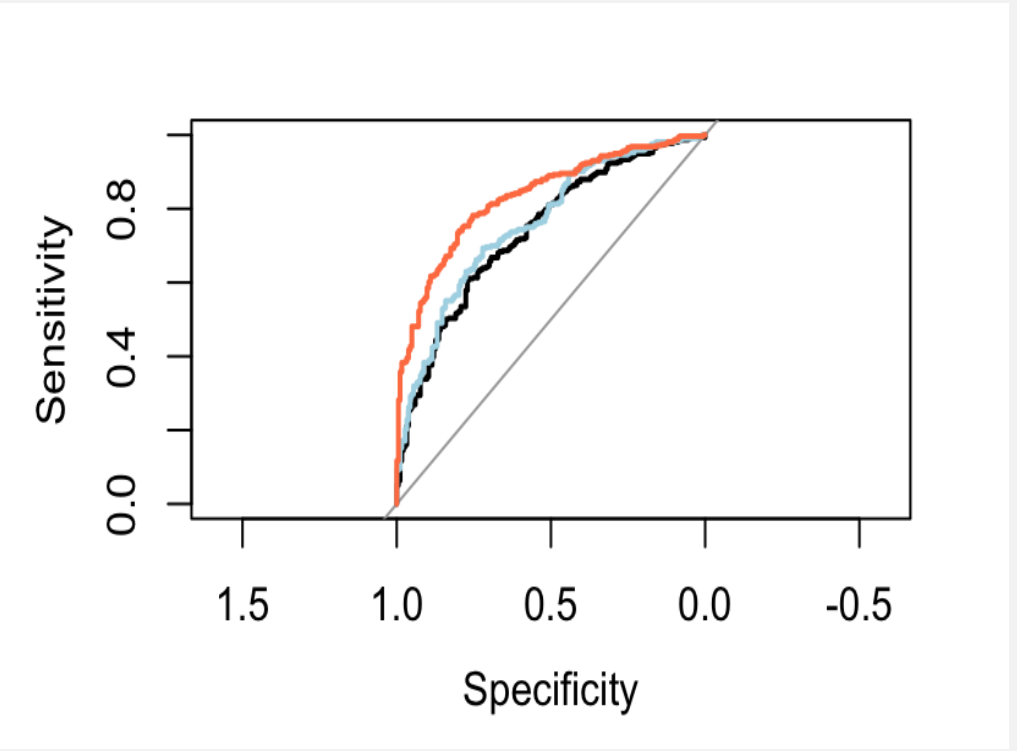
LOGISITIC REGRESSION

Model 1(Baseline): glm(Win ~ Team + era + xG + Poss)

Model 2 (Interaction) glm(Win ~ Team + era * (xG + Poss + CrdY + CrdR + Two_CrdY)

Model 3(Full Tactical Model): glm(Win ~ Team + era * (xG + Poss + formation_style + Opponent_difficulty + opp_formation_style + CrdY + CrdR + Two_CrdY)

Model 1 AUC: 0.739 **Model 2 AUC:** 0.759 **Model 3 AUC:** 0.832

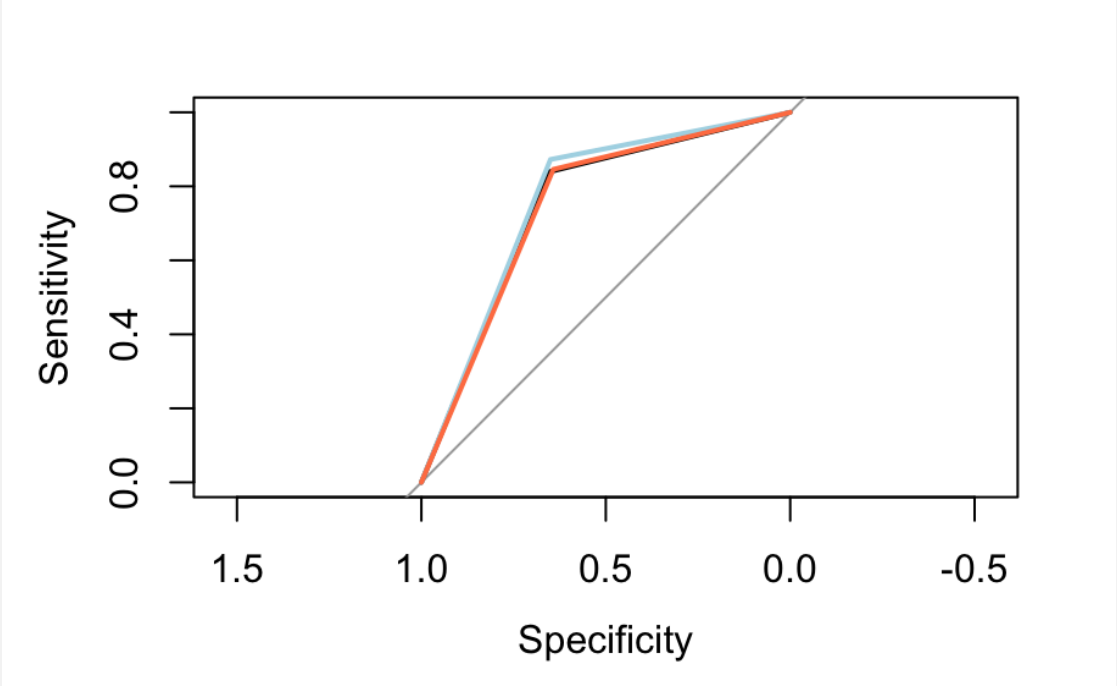


MODEL 3:

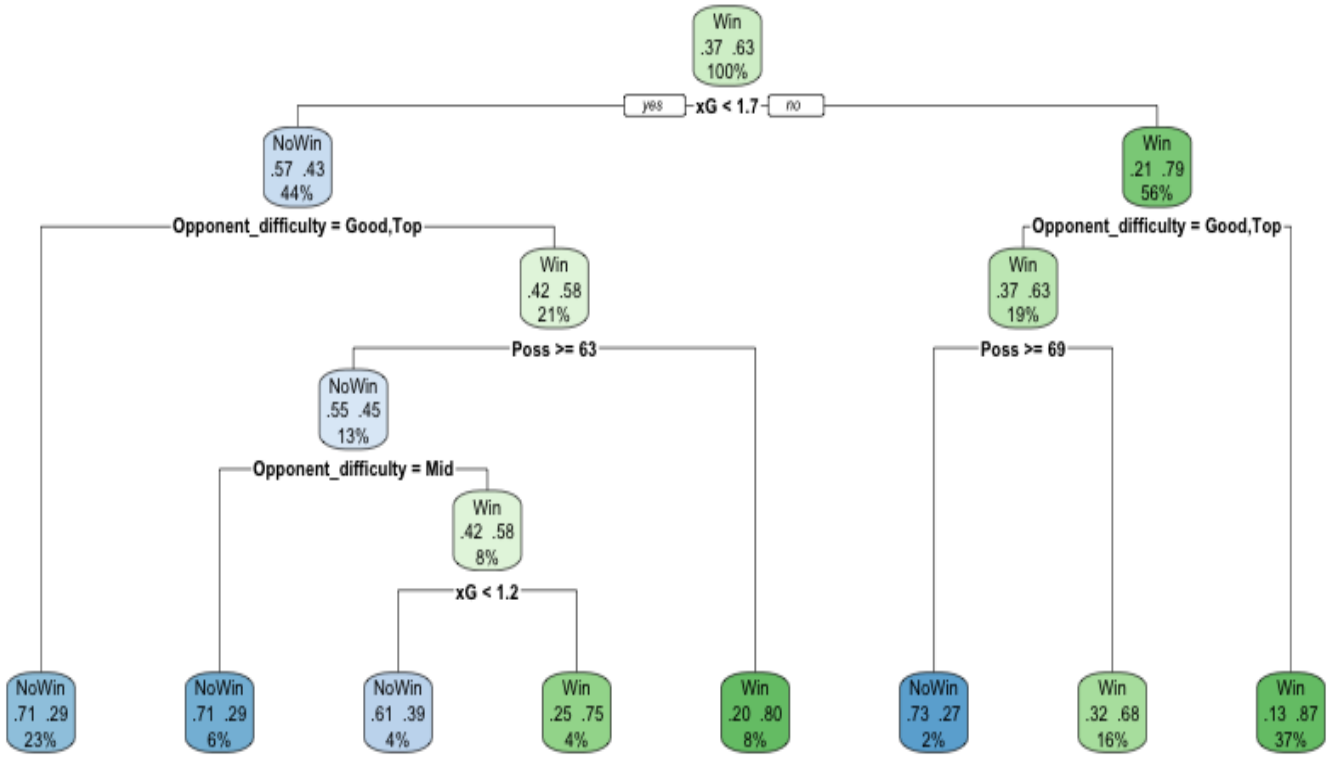
- The most interesting coefficients were:
- An extra expected goal raises win odds in the Post era, Beta = 0.958 and Odds Ratio = 2.6. However, the boost was larger in the Pre era with a Beta = 1.48 and OR = 4.4).
- Possession no longer matters after 2023 the Beta is almost equal to 0, whereas it modestly lowered win odds in 2022-23.
- Before 2023, each extra yellow made a win *less* likely. After 2023, each yellow made a win *more* likely.
- In 2022–23 playing against a defensive-minded team hurt Arsenal's win odds. After 2023, it *helped*, a defensive opponent tripled their chances of winning.

REGRESSION TREES

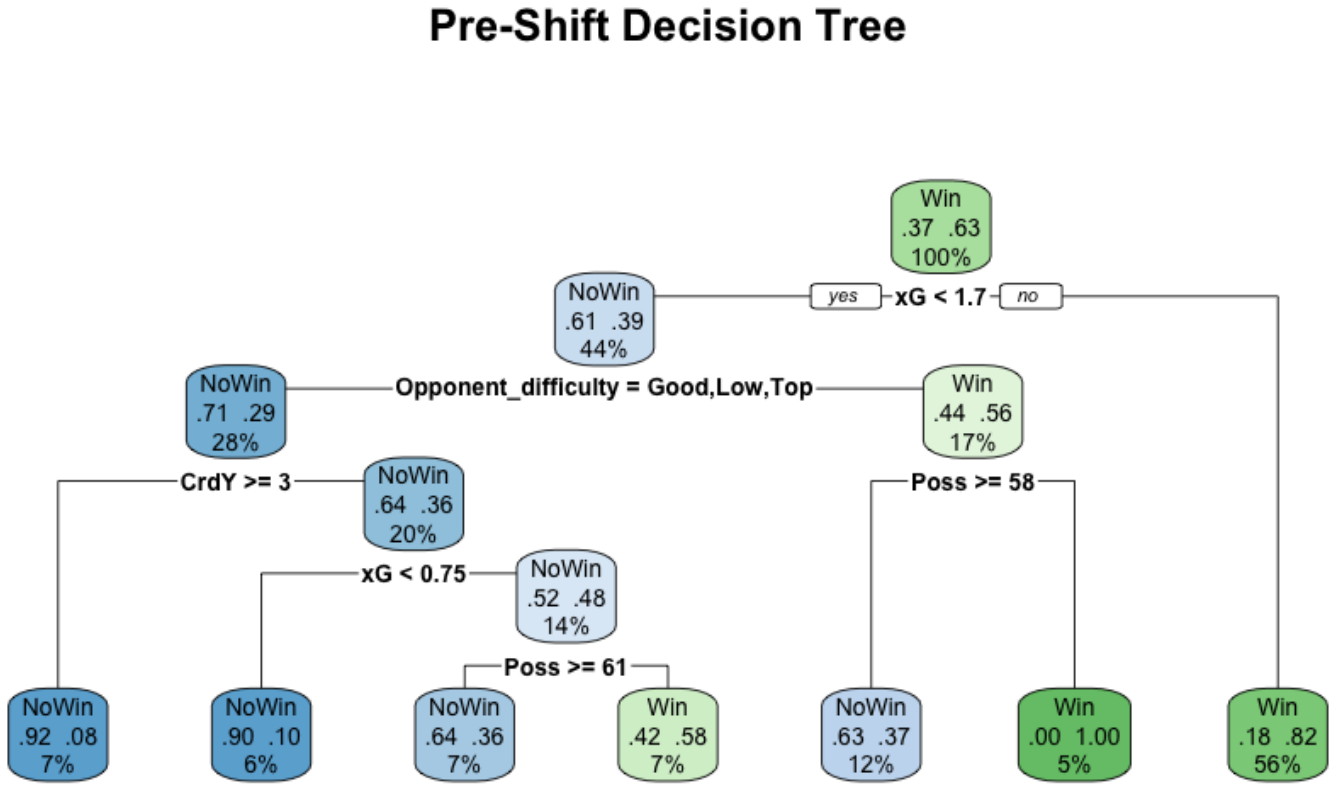
- **ACL_Fit_Tree_1** <- rpart(Win ~ era + Team + xG + Poss + CrdY + CrdR + Two_CrdY + formation_style + Opponent_difficulty + opp_formation_style, data = Flag_df, method = "class")
- **Tree_Pre** <- rpart(Win ~ xG + Poss + formation_style + Opponent_difficulty + CrdY, data = filter(Flag_df, era == "Pre"), method = "class")
- **Tree_Post** <- rpart(Win ~ xG + Poss + formation_style + Opponent_difficulty + CrdY, data = filter(Flag_df, era == "Post"), method = "class")
- **All Seasons AOC:** 0.7472, **Pre 22/23 AOC:** 0.7613, **Post 22/23 AOC:** 0.744



ALL SEASONS DECISION TREE

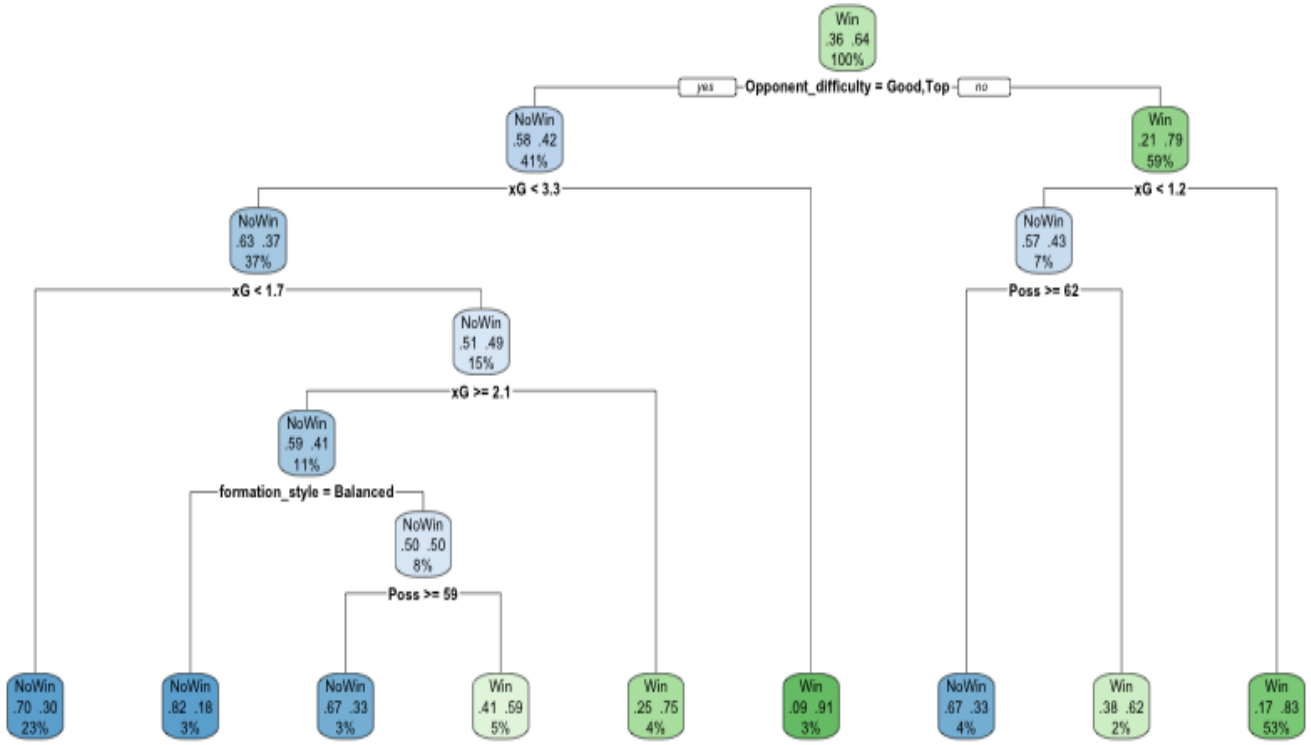


2022/23 DECISION TREE

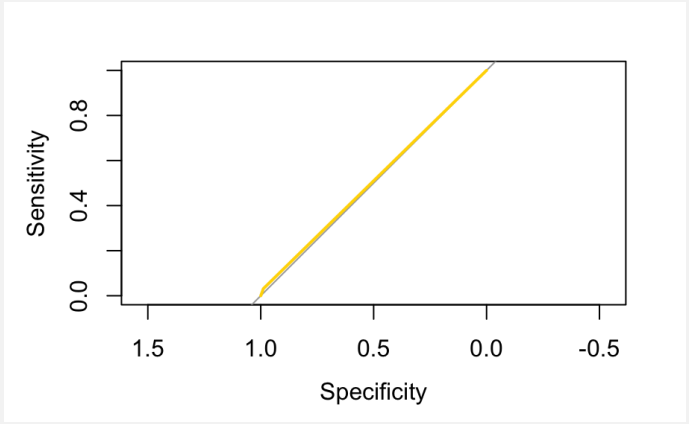
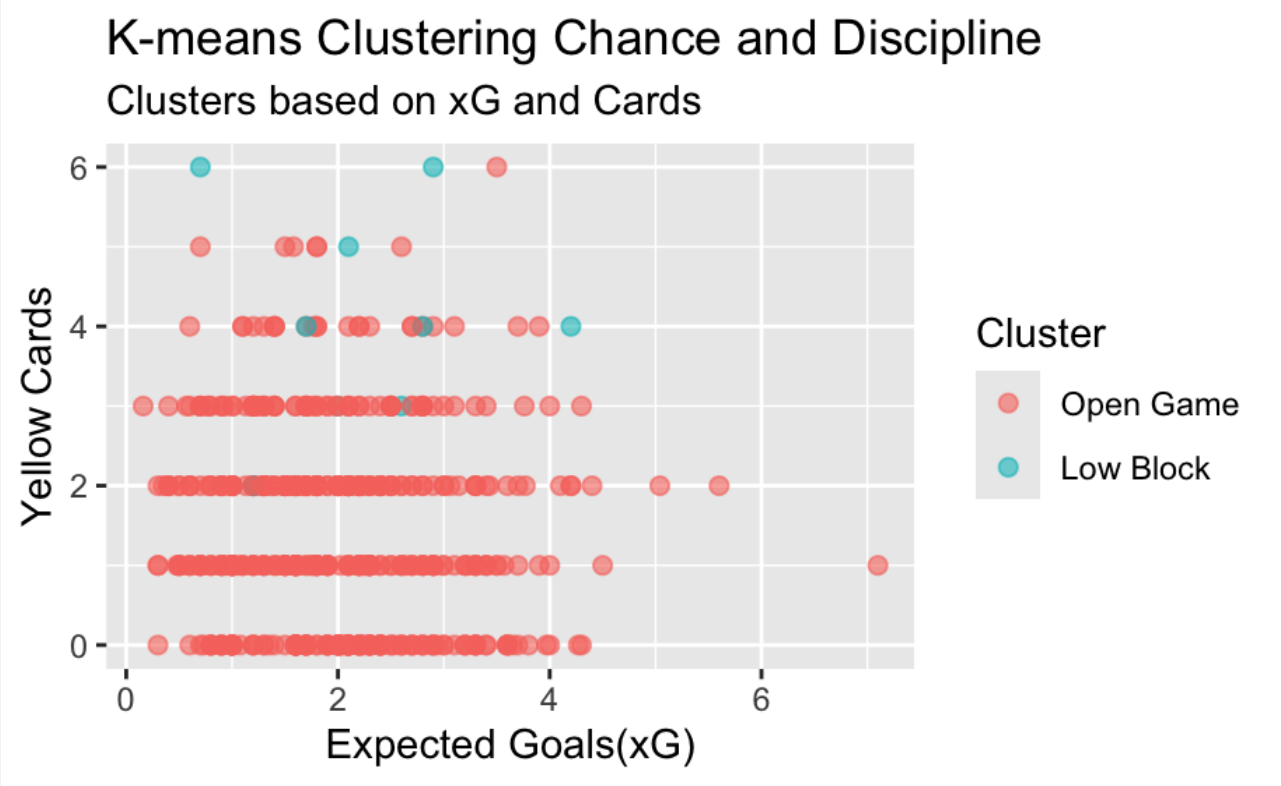


2023/24 & 2024/25 DECISION TREE

Post-Shift Decision Tree



K-MEANS



- **Two clear styles emerge**
- **Open Game** High xG, few yellows -> proactive, chance-heavy football
- **Low Block** Low xG, many yellows -> tight, foul-heavy contests
- **Quick take-away**
- When Arsenal create chances *and* stay disciplined, their matches cluster in the Open Game group, signaling attacking control.
- When chance creation stalls and fouls rise, matches shift to the Low Block group, showing a more defensive, stop-start pattern.