Matheus Grover

Why is Arsenal Always Second

Motivation:

There is a clear problem at Arsenal. Our current manager has been here for five years, and we have won just one trophy out of a possible twenty in those five years. This manager Mikel Arteta joined us when the club was struggling, ending his first season in 8th place. The next season we placed 8th again, then 5th, then finally we got 2nd that year and for the first time in 6 years we qualified for the Champions League the most prestigious knockout tournament in the world. The next year we got 2nd again and to the quarter finals of the Champions League. This last year we got 2nd and this time to the semi-finals of the Champions League. So, every Arsenal fan knows our problem, we're consistently getting better but still haven't pushed over the line yet.

In the 2022/23 season Arsenal took everyone by surprise; we went from 5th place the season before to spending 248 days out of a 296 day season in 1st place before falling apart and placing 2nd. This was the record for the most days a team had been in first place without winning the league. I discussed this problem with my applied consultant; my consultant believes this collapse led to Arteta adapting a different style of play during the 2023-2025 seasons that hasn't been as successful as 2022-2023. After the collapse, Arteta's 2023–25 tactical shift underperformed, further hampered by referee bias and deeper opponent defenses.

Data and EDA:

To solve this problem, I gathered the data from Arsenal, Manchester City, and Liverpool. Manchester City who were 1st in the 22/23 and 23/24 and 3rd in 24/25 and Liverpool who were 5th in 22/23, 3rd in 23/24, and 1st in 24/25.

Two of the most important themes in this class have been: "Ask questions about the data before you see it" and "Understand your data before you do analysis"

The longest part of this project was setting up my dataset. I had data that was scrapped for a previous project that had all of Arsenals game stats for 23/24 and 24/25 and for this project I added 22/23 and then all of those seasons for Liverpool and Manchester City.

Before I put the data into Tableau, I wrote down 8 predictions on what the data looks like, these were general predictions about all three teams to get a better feel for what I was working with. For Homework1 I created a Tableau workbook answering these predictions. These questions helped me understand what I was looking for, decide what variables I needed, changes to make, cleaning that needed to happen before I started my EDA.

I began by importing Arsenal's match logs from the Current_Data tab of a Google Sheet, converted the date, time, xG, and xGA fields to proper types, and standardized each formation string so every shape followed the 4-3-3 style pattern. Using lubridate I parsed dates and created two derived fields: Season (e.g., 2023/24) and era that labels 2022/23 as Pre and all later seasons as Post. I recoded Result to a W-D-L factor, added a Points column (3, 1, 0), added in yellow and red card counts by left-joining the YR tab on the google sheet using date as the key.

Next, I assigned each opponent to a difficulty tier Top: 1-6, Good: 7-9, Mid: 10-12, Low: 13-17, Non-League: 18-20. In order to reduce bias, I decided those bin widths before averaging each teams finishing place for the last three seasons. Top 5 European leagues were weighted similar and other leagues were weighted lower. Then for later analysis I encoded them at an ordinal level, then classified both team and opponent formations as Defensive, Balanced, or Attacking and gave each a numeric level. The finished data set (Flag_df) captures every cleaned field/new column and is the sole input for the logistic regressions, decision trees, and k-means clustering that follow.

Now having completed a first sweep of EDA and fixes I was ready to ask myself 6 new questions specific to Arsenal's problem (located in the Appendix below) then conducted a second round of Tableau EDA to fully understand my data before I went into analysis. This part of the EDA was much more focused on what my analysis would be on.

Models:

Ready for analysis I decided to run logistic regressions, Decision Trees, and k-means clustering. The results from these models support that post 22/23 tactical shifts define Arsenal's recent problem.

Our baseline Model One is a logistic regression that includes only team, era, xG, and possession separates wins from non-wins with an AUC of 0.739. Model Two allows xG, possession, and all card variables to interact with era lifts the AUC to 0.759, implying that discipline from referees carries a different weight before and after Spring 2023 which was supported by Figure1 in the appendix where we can see Arsenal has more average red and yellow cards per game post Spring 2023.

The largest interaction model, Model Three, builds off Model Two and adds an interaction of formation style and opponent difficulty with era, which pushes AUC to 0.832.

Since Model 3 has the highest AUC score of the three models, I pulled their coefficients to interpret. All of the following coefficients were highly significant. The most interesting coefficients were the following: 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 =

1.481 and OR = 4.4). Possession no longer matters after 2023 the Beta is almost equal to 0, where it slightly lowered win odds in 2022-23. For yellow cards before 2023, each extra yellow made a win less likely. After 2023, each yellow made a win more likely. Which is interesting, it could be related to the fact that more yellow cards are common in games with low blocks/high defending which are typically worse teams. For an Opponent team being defensive, 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. I think this suggests Arteta's Spring 2023 overhaul flipped Arsenal's approach to defensive opponents: where compact defensive teams once cut Arsenal's win probability by about a third, the new tactics now make those same setups one of Arsenal's biggest advantages, tripling their chances of victory when opponents sit deep.

My Decision trees reach the same conclusions in a more visual way. Across all seasons the first split is xG < 1.7; low-chance matches dominate the left branch and rarely end well unless possession exceeds the mid 60s against a mid-tier opponent. When only looking at Post 22/23 matches, xG remains the first rule, but the tree shows that Arsenal must also clear a possession threshold of about 58 percent to avoid a loss. In Pre 22/23 matches, the primary branch is if the team is a top/good team and if is we are more likely to lose. Figures 2-4 Below.

Finally, I tested the hunch from the regression that cards pile up whenever Arsenal are dragged into low-block contests. Running k-means (k = 2) on expected goals and yellow-card counts cleanly separated the data into two match archetypes. Open Games cluster where xG is high and fouls are scarce; Low-Block games sit in the opposite corner with low xG and many cards. Wins concentrate almost entirely inside the Open-Game cloud, while the Low-Block points line up with the loss leaves in the decision trees. The picture reinforces the earlier finding: yellow cards only help when they appear alongside plenty of chances. Figure 5.

Arsenal's three consecutive 2nd place finishes tracks back to the tactical shifts made after the 2022-23 collapse. Is it possible to get back to the higher xG, more attacking, less targeted football that we played in 22/23?

Appendix:

Questions answered by the Era EDA Tableau Workbook:

-Did Arsenal 22/23 have lower possession than 23-25?

-Did Arsenal 22/23 have higher expected goals than 23-25?

-Did Arsenal 22/23 have less cards than 23-25?

-Did teams play more attacking against Arsenal in 22/23

-Did Arsenal 22/23 concede more goals than 23-25?

-Did Arsenal 22/23 score more goals than 23-25?

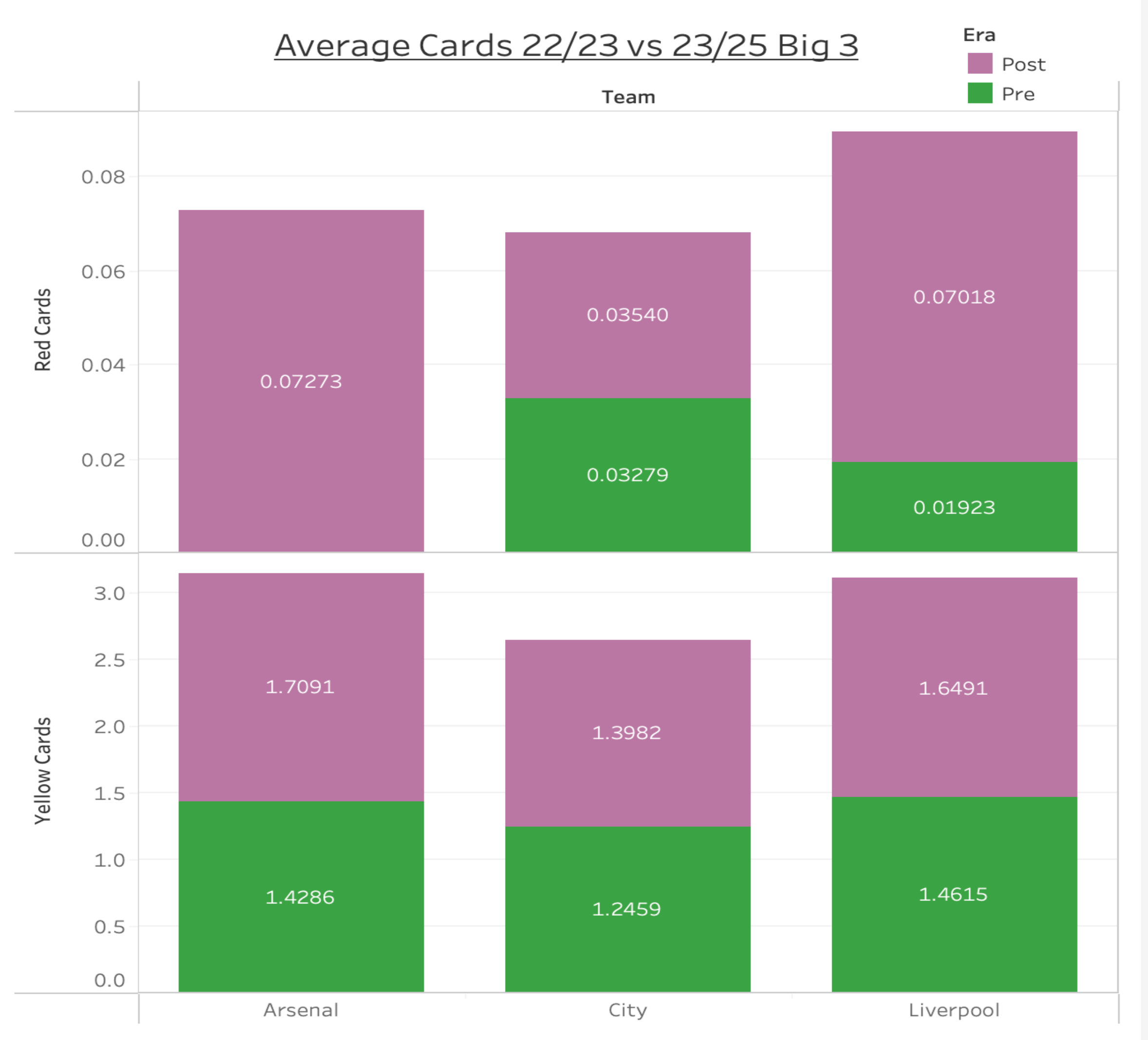

Figure 1: Average Cards in 22/23 vs 23/25

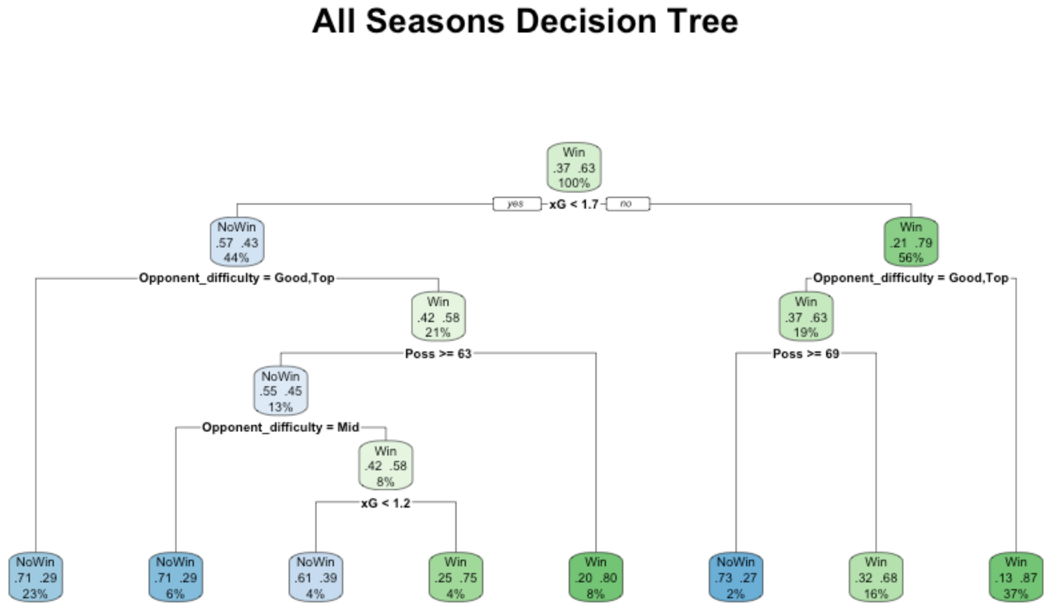

Figure 2: All Seasons Decision Tree

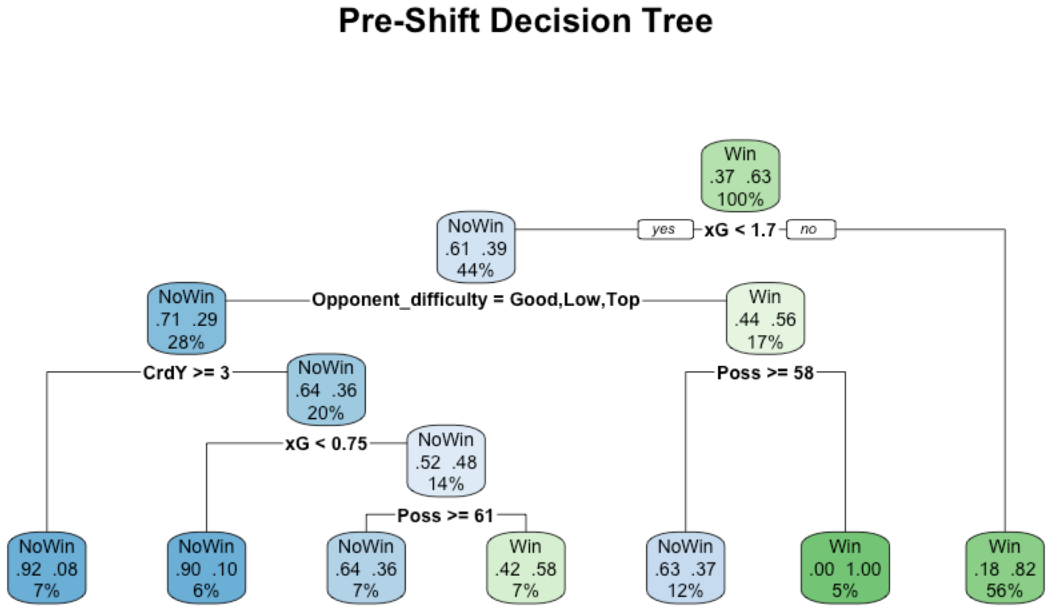

Figure 3: Pre Decision Tree

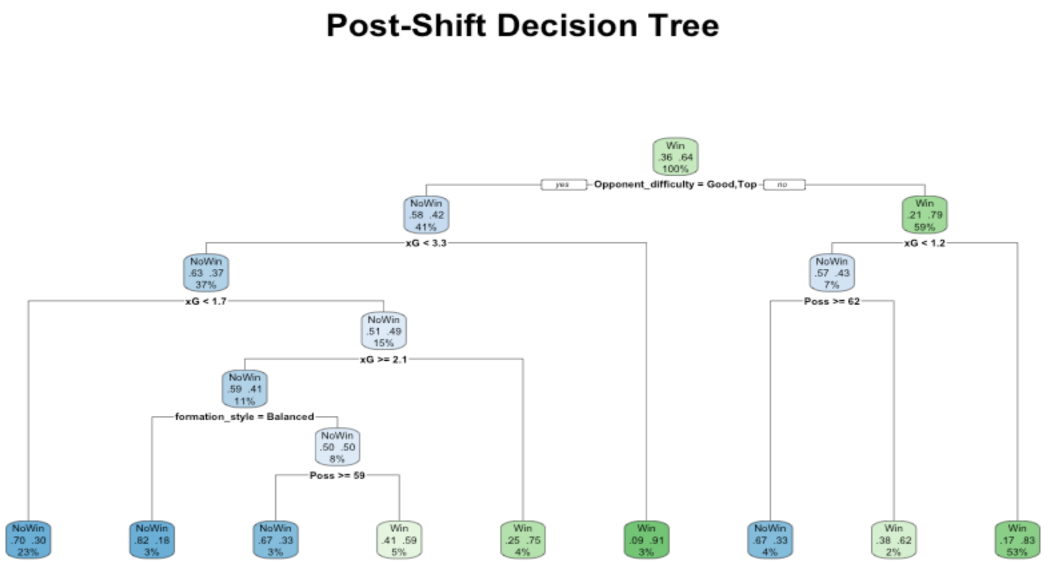

Figure 4: Post Seasons Decision Tree

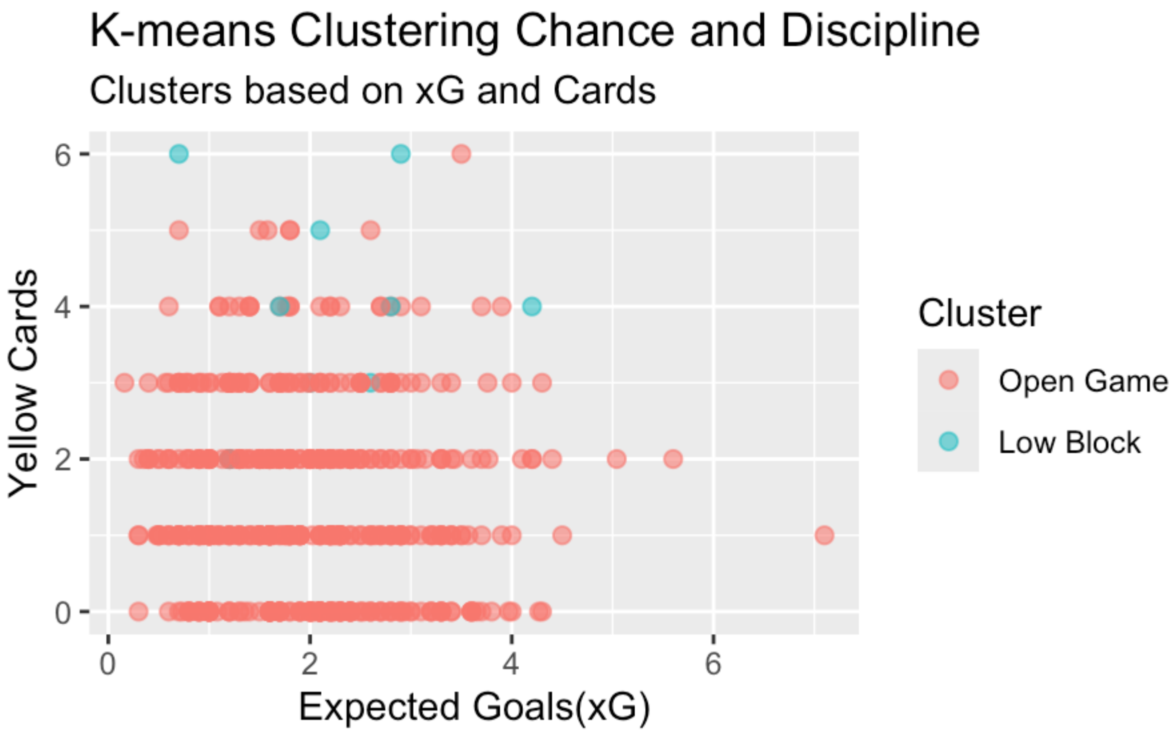

Figure 5: K-mean Clustering Chance and Discipline