Introduction:

In this project, we aim to analyze whether being the home team in a Premier League match influences referee decisions, specifically in the awarding of yellow and red cards. The key question is whether home advantage, driven by fan presence, impacts the number of cards given to home teams relative to away teams. To reduce bias from factors like the strength of the teams (underdogs often foul more), we focus on the ratio of yellow and red cards per foul as the main metric for analysis.

Methodology:

Data Collection:

The data for this project was sourced from Kaggle, which contains match statistics from Premier League games. The dataset was imported into PostgreSQL for initial querying and data handling.

Data Cleaning and Preparation

To ensure the dataset was ready for analysis, the following steps were performed:

Data Loading: The CSV file was loaded into R using the read.csv() function.

Basic Data Inspection: We checked the dataset for missing values and inspected the data using functions like head() to verify the data format.

Data Formatting: Columns like HomeTeamYellowCards, HomeTeamRedCards, AwayTeamYellowCards, and AwayTeamRedCards were in numeric format, so no conversions were needed. Rows with missing values in critical columns such as HomeTeamFouls and HomeTeamYellowCards were removed.

Feature Engineering: New variables, such as cards per foul ratio for home and away teams, were created to capture referee behavior more clearly. Percentages of home yellow and red cards relative to total cards were also calculated.

Data Aggregation: The data was grouped by referee and season to analyze referee tendencies over time.

Data Visualization: The data was reshaped using melt() for visualization in ggplot2, allowing us to plot trends in home card percentages across seasons.

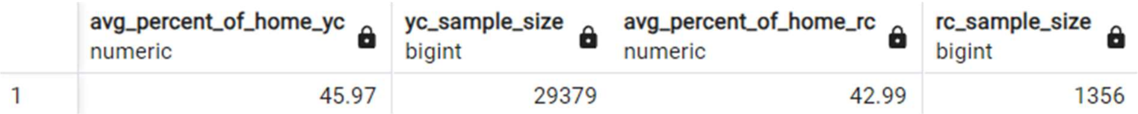
Tools Used:

PostgreSQL: For initial data management and querying.

R: For statistical analysis and data visualization.

Packages: ggplot2 for visualizations and reshape2 for reshaping the data for plotting.

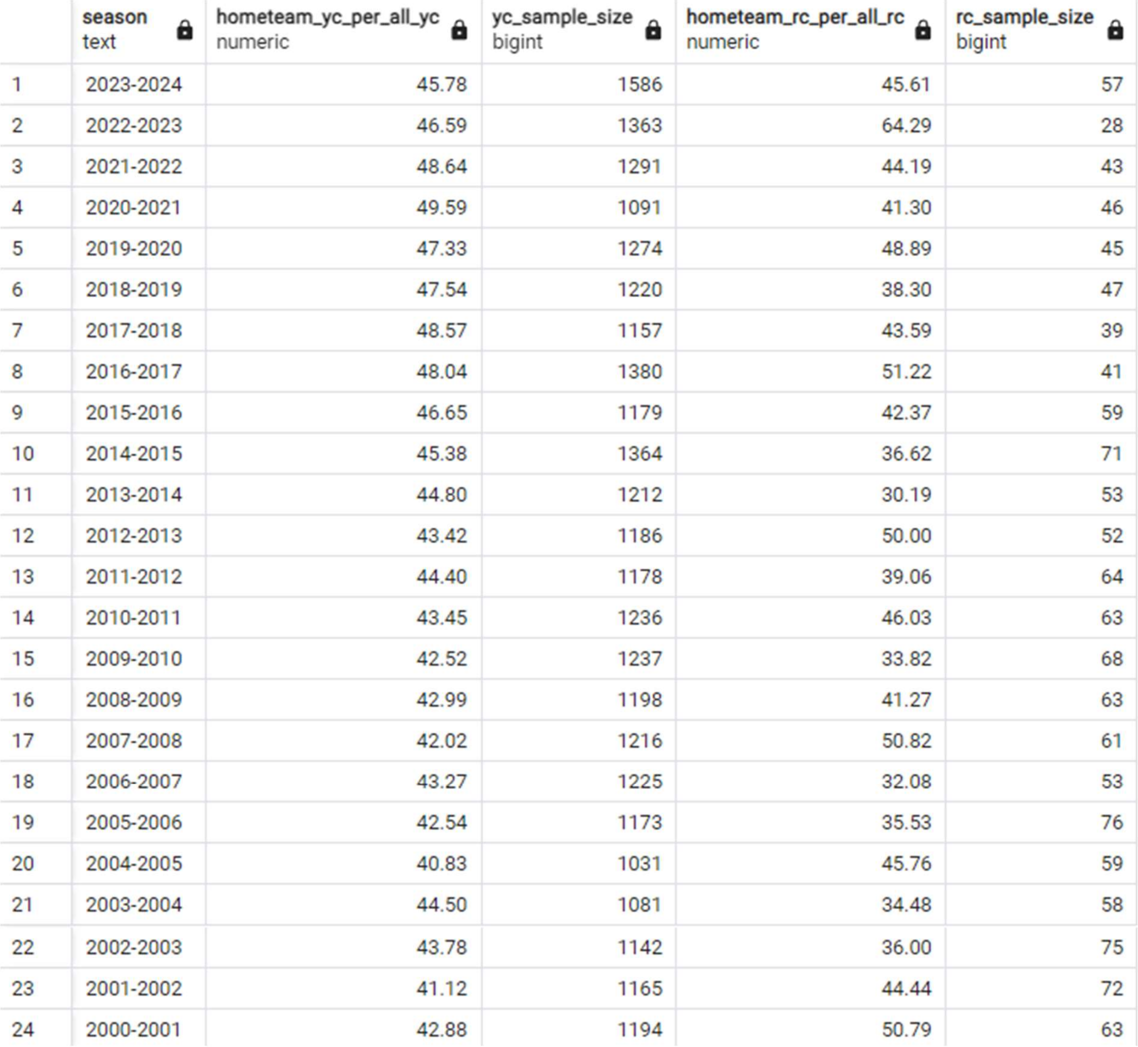
The ratio of home yellow cards to total yellow cards and home red cards to total red cards



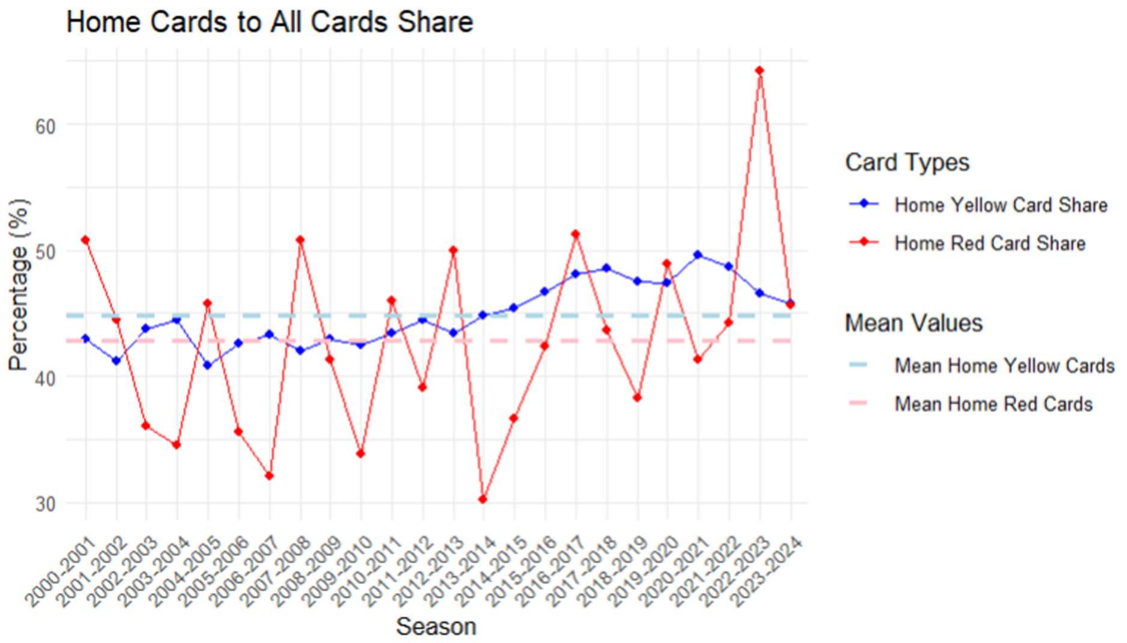
	avg_percent_of_home_yc numeric	yc_sample_size bigint	avg_percent_of_home_rc numeric	rc_sample_size bigint
1	45.97	29379	42.99	1356

As we can see, less than 46% of yellow cards per foul were shown to the home team. A sample size of almost 30,000 yellow cards is large enough to suggest that there is merit to our hypothesis. This trend is even more pronounced with red cards, as nearly 43% were shown to the home team. While the sample size for red cards is smaller, the proportion still indicates that being the host does have an impact.

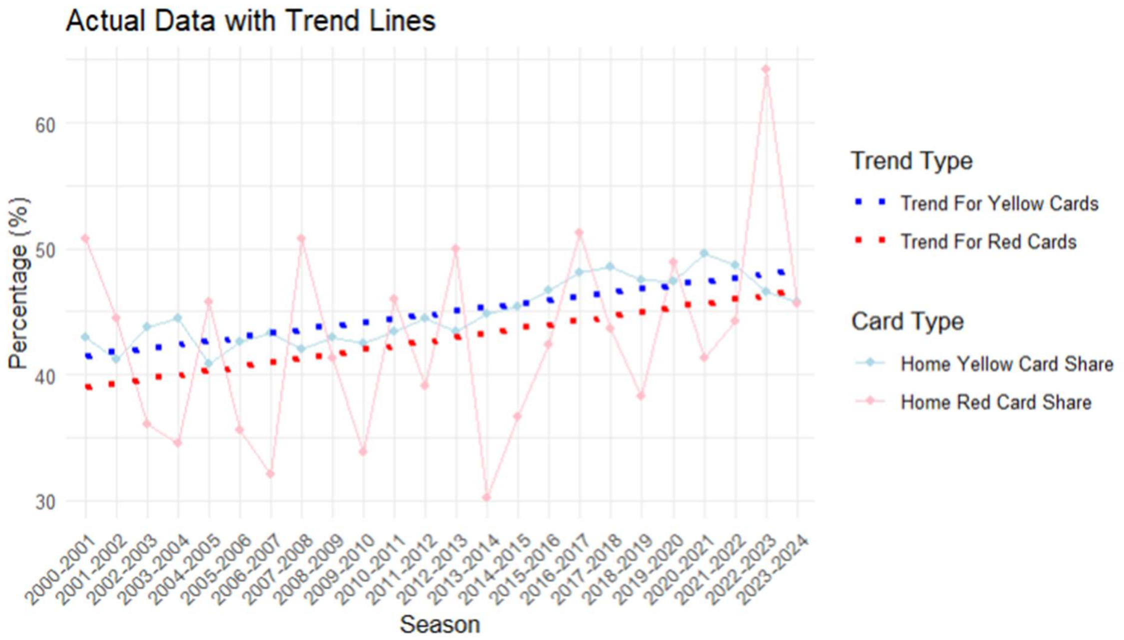
The ratio of home yellow cards to total yellow cards and home red cards to total red cards, broken down by season



	season text	hometeam_yc_per_all_yc numeric	yc_sample_size bigint	hometeam_rc_per_all_rc numeric	rc_sample_size bigint
1	2023-2024	45.78	1586	45.61	57
2	2022-2023	46.59	1363	64.29	28
3	2021-2022	48.64	1291	44.19	43
4	2020-2021	49.59	1091	41.30	46
5	2019-2020	47.33	1274	48.89	45
6	2018-2019	47.54	1220	38.30	47
7	2017-2018	48.57	1157	43.59	39
8	2016-2017	48.04	1380	51.22	41
9	2015-2016	46.65	1179	42.37	59
10	2014-2015	45.38	1364	36.62	71
11	2013-2014	44.80	1212	30.19	53
12	2012-2013	43.42	1186	50.00	52
13	2011-2012	44.40	1178	39.06	64
14	2010-2011	43.45	1236	46.03	63
15	2009-2010	42.52	1237	33.82	68
16	2008-2009	42.99	1198	41.27	63
17	2007-2008	42.02	1216	50.82	61
18	2006-2007	43.27	1225	32.08	53
19	2005-2006	42.54	1173	35.53	76
20	2004-2005	40.83	1031	45.76	59
21	2003-2004	44.50	1081	34.48	58
22	2002-2003	43.78	1142	36.00	75
23	2001-2002	41.12	1165	44.44	72
24	2000-2001	42.88	1194	50.79	63



We can observe a large variance in the red card data, likely due to the small sample size. As noted earlier, both yellow and red card averages are below 50%. However, there seems to be an improvement over time, with the data showing less bias toward the home team. Let’s examine this trend more closely.



In our analysis, we applied linear regression models to understand trends in the awarding of yellow and red cards to home teams over multiple football seasons. The models were developed using the lm() function in R, a standard approach for fitting linear models. The visual representation of the trend suggests that away teams are being penalized less harshly over time compared to home teams, which may indicate that referees are becoming better at managing the pressure from home fans.

Conclusion:

Our analysis found that referees tend to penalize away teams more than home teams, with home teams receiving less than 46% of yellow cards per foul and only 43% of red cards. The sample size of around 30,000 yellow cards is large enough to support this observation. However, we also observed that the trend is changing, as away teams seem to be penalized less harshly in more recent seasons, suggesting that referees may be adapting to the pressure of home crowds.

Limitations: The analysis could be further refined by incorporating additional variables such as specific referee tendencies or game context (e.g., scoreline, match importance). Additionally, the smaller sample size for red cards means that conclusions about red card bias are less robust than for yellow cards.