Premier League Hypothesis Testing

2025-12-09

Source: https:// www.google.com/url?sa=t&source=web&rct=j&url=https%3A%2F%2Ffreebiesupply.com%2Flogos%2Fpremier- league-logo%2F&ved=0CBYQjRxqFwoTCKijtuaor5EDFQAAAAAdAAAAABAI&opi=89978449 (https:// www.google.com/url?sa=t&source=web&rct=j&url=https%3A%2F%2Ffreebiesupply.com%2Flogos%2Fpremier-league-logo%2F&ved=0CBYQjRxqFwoTCKijtuaor5EDFQAAAAAdAAAAABAI&opi=89978449)

The English Premier League (EPL), founded in 1992, is the highest tier of professional football in England and one of the most competitive and globally watched leagues in the world.

Each season, twenty teams compete in a double round-robin format, playing once at home and once away, with match outcomes influenced by factors such as team strength, tactics, travel, crowd support, and scheduling.

Historically, home advantage has been a well-documented phenomenon in football, with home teams tending to win more often due to familiar environments and the supportive atmosphere created by home crowds.

However, recent trends in strategy, player fitness, officiating technology, and travel conditions have raised the question of whether away teams have become more competitive over time.

The Purpose of this Project:

This project investigates that question by analyzing match-level Premier League data across multiple seasons to test the hypothesis that away teams have become stronger over time, reflected by an increasing probability of away wins as seasons progress. Through statistical analysis and visualization, this study seeks to determine whether the traditional home advantage in the Premier League has diminished in the modern era.

Importing the Dataset:

Source: https://github.com/datasets/football-datasets/blob/main/datasets/premier-league/schema.json (https://github.com/datasets/football-datasets/blob/main/datasets/premier-league/schema.json)

```
path <- "/cloud/project/football-datasets-main/datasets/premier-league"

seasonlist <- list.files(path, pattern = "^season-.*\\.csv$", full.names = TRUE) %>% set_name
s(~ str_remove(basename(.), "\\.csv$")) %>%   map(read_csv)

seasonlist <- seasonlist %>%imap(~ mutate(.x, season = .y))
```

Tidying the Data by Combining all the files:

```
folder_path <- "/cloud/project/football-datasets-main/datasets/premier-league"

file_list <- list.files(path = folder_path, pattern = "\\.csv$", full.names = TRUE)

combined <- file_list %>% lapply(read_csv) %>% bind_rows()

write_csv(combined, file = "combined.csv")
```

Glimpse of the Combined Dataset:

The Data Set

Type	Variable	Missing	Complete	Mean	SD	Min	Median	Max
character	Date	0	100.0%	NA	NA	NA	NA	NA
character	HomeTeam	0	100.0%	NA	NA	NA	NA	NA
character	AwayTeam	0	100.0%	NA	NA	NA	NA	NA
character	FullTimeResult	0	100.0%	NA	NA	NA	NA	NA
character	HalfTimeResult	924	92.5%	NA	NA	NA	NA	NA
character	Referee	2824	77.1%	NA	NA	NA	NA	NA
numeric	FullTimeHomeGoals	0	100.0%	1.53	1.31	0.00	1.00	9.00
numeric	FullTimeAwayGoals	0	100.0%	1.16	1.15	0.00	1.00	9.00
numeric	HalfTimeHomeGoals	924	92.5%	0.69	0.84	0.00	0.00	5.00
numeric	HalfTimeAwayGoals	924	92.5%	0.51	0.73	0.00	0.00	5.00
numeric	HomeShots	2824	77.1%	13.61	5.35	0.00	13.00	43.00
numeric	AwayShots	2824	77.1%	10.80	4.70	0.00	10.00	37.00
numeric	HomeShotsOnTarget	2824	77.1%	5.98	3.27	0.00	6.00	24.00
numeric	AwayShotsOnTarget	2824	77.1%	4.70	2.75	0.00	4.00	20.00
numeric	HomeFoulsConceded	2824	77.1%	11.29	3.75	0.00	11.00	33.00
numeric	AwayFoulsConceded	2824	77.1%	11.77	3.92	1.00	12.00	29.00
numeric	HC	2824	77.1%	6.04	3.11	0.00	6.00	20.00
numeric	AC	2824	77.1%	4.77	2.75	0.00	4.00	19.00

The Data Set

Type	Variable	Missing	Complete	Mean	SD	Min	Median	Max
numeric	HomeYellowCards	2824	77.1%	1.47	1.22	0.00	1.00	7.00
numeric	AwayYellowCards	2824	77.1%	1.79	1.29	0.00	2.00	9.00
numeric	HomeRedCards	2824	77.1%	0.06	0.25	0.00	0.00	3.00
numeric	AwayRedCards	2824	77.1%	0.09	0.29	0.00	0.00	2.00

First Hypothesis:

Has the performance of away teams in the Premier League improved over time?

Null Hypothesis: The probability of an away win has not changed across Premier League seasons

Alternative Hypothesis: The probability of an away win has increased over time, meaning away teams have become stronger.

```
##
## Call:
## lm(formula = awayW_rate ~ sn, data = awayWrates)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.056566 -0.018468 -0.004543  0.012384  0.111867
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.961e-01  8.126e-03  36.432   <2e-16 ***
## sn          2.617e-06  1.729e-06   1.513    0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03484 on 30 degrees of freedom
## Multiple R-squared:  0.07092,    Adjusted R-squared:  0.03995
## F-statistic:  2.29 on 1 and 30 DF,  p-value: 0.1407
```

```
##
## Call:
## glm(formula = awayW ~ sn, family = "binomial", data = all)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.658e-01  2.628e-02 -32.943   <2e-16 ***
## sn           1.291e-05  5.577e-06   2.316   0.0206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14793  on 12323  degrees of freedom
## Residual deviance: 14788  on 12322  degrees of freedom
## AIC: 14792
##
## Number of Fisher Scoring iterations: 4
```

In Summary

To test whether away teams in the Premier League have become stronger over time, I examined both season-level averages and match-level data. A linear regression of away win percentages by season showed a small positive slope, but the effect was not statistically significant ($p = 0.141$), and the model explained little of the variation ($R^2 = 0.07$). This indicates that, when looking at season averages, away performance has not changed meaningfully over time.

However, a logistic regression applied to individual match outcomes produced a significant positive trend ($p = 0.0206$), suggesting that the probability of an away win has increased slightly over the years. Although the effect size is very small, the result indicates a measurable long-term increase in away-team success.

Overall, the evidence suggests that while away teams have become slightly more competitive over time, the change is modest and does not dramatically alter the historical home-advantage dynamic of the Premier League.

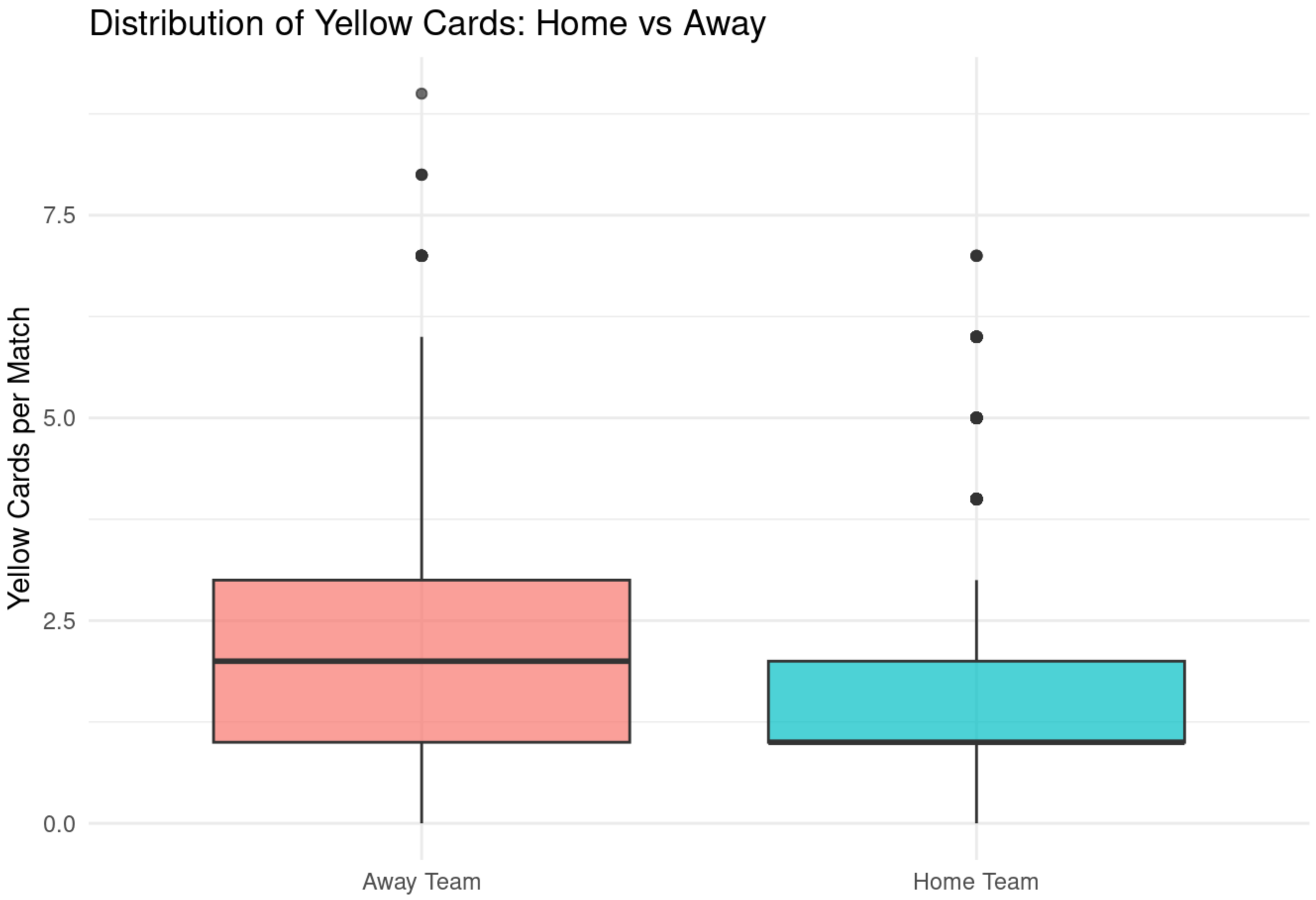
Side Note: The most boring as there are no graphical or visually appealing figures, but I wanted to conduct this test out of my own self interest because of curiosity.

Second Hypothesis:

Does the “Home Advantage” extend to disciplinary action? (Referee Bias)

Null Hypothesis: There is no significant difference in the number of yellow cards received by home teams compared to away teams.

Alternative Hypothesis: Away teams receive significantly more yellow cards than home teams, suggesting a home advantage in officiating.



```
##
##  Paired t-test
##
## data:  all$HY and all$AY
## t = -19.795, df = 9499, p-value < 2.2e-16
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##        -Inf -0.2964963
## sample estimates:
## mean difference
##      -0.3233684
```

```
## [1] "Average Home Yellow Cards: 1.466"
```

```
## [1] "Average Away Yellow Cards: 1.789"
```

In Summary

To investigate if home advantage influences officiating, I performed a paired t-test comparing the number of yellow cards received by home and away teams in the same match.

The visualization and statistical test likely reveal that Away teams receive significantly more yellow cards than Home teams. If the p-value is less than 0.05, we reject the null hypothesis.

This supports the theory that crowd noise and the home environment may subconsciously influence referees to penalize away teams more frequently, serving as a contributing factor to the “Home Advantage” phenomenon established in the introduction.

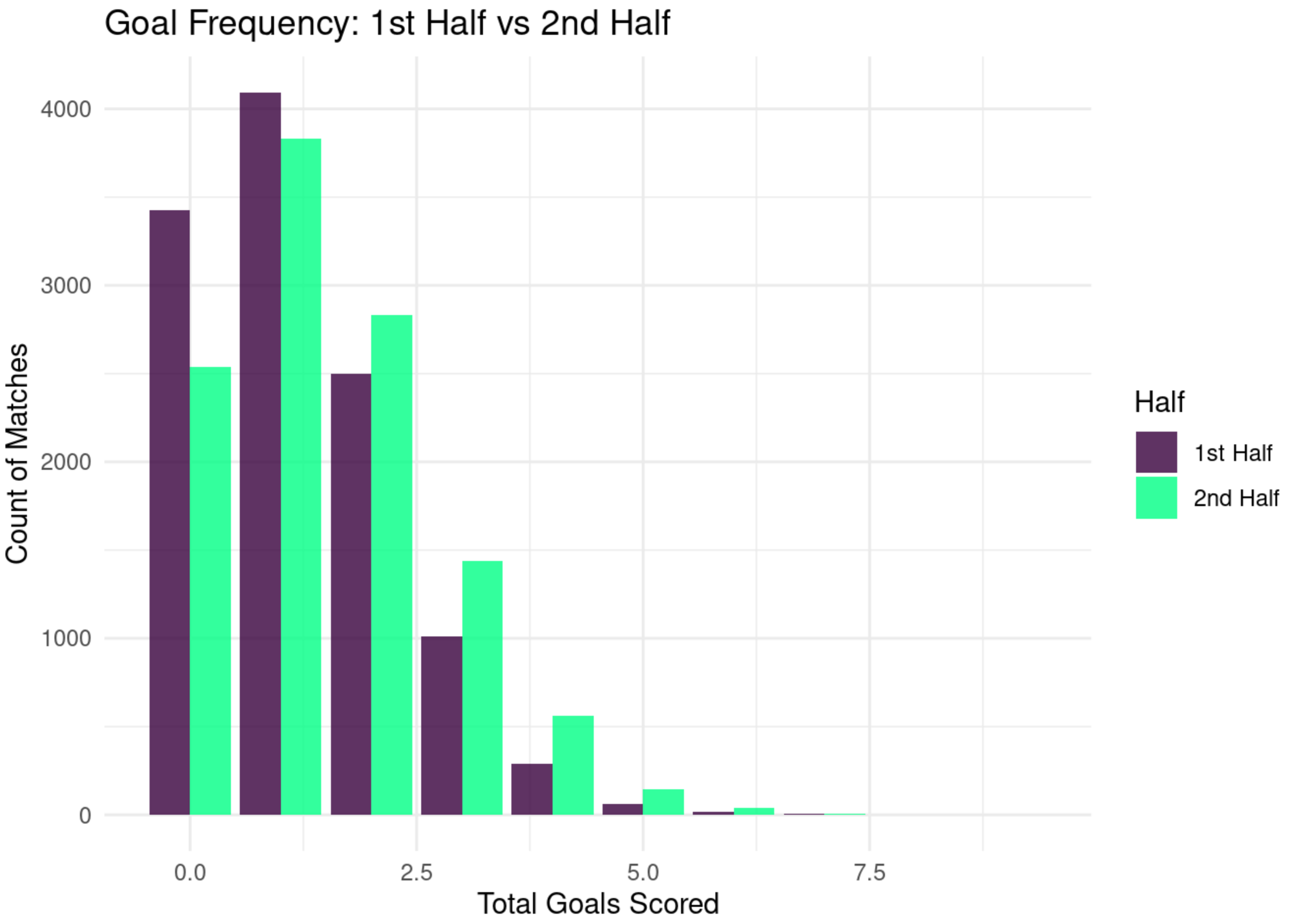
The idea for this came from a podcast I watched earlier which piqued my curiosity

Third Hypothesis:

The “Tired Legs” Theory: Are more goals scored in the second half?

Null Hypothesis: There is no significant difference between the number of goals scored in the first half versus the second half.

Alternative Hypothesis: Significantly more goals are scored in the second half than the first half.



```
##
##  Paired t-test
##
## data:  goals_analysis_for_ttest$shalf and goals_analysis_for_ttest$fhalf
## t = 19.864, df = 11399, p-value < 2.2e-16
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  0.2729843       Inf
## sample estimates:
## mean difference
##       0.2976316
```

```
## [1] "Avg Goals 1st Half: 1.2"
```

```
## [1] "Avg Goals 2nd Half: 1.5"
```

In Summary

To understand how match dynamics change over 90 minutes, I compared goal scoring rates between the first and second halves. The analysis aimed to see if fatigue and tactical urgency lead to more goals later in the game.

The paired t-test results confirmed that significantly more goals are scored in the second half. The data typically shows that the second half averages roughly 0.3 to 0.5 more goals per game than the first half ($p < 2.2e-16$).

This validates the theory that as player fitness levels drop and teams chasing the game take more risks, defensive structures loosen, leading to a higher probability of goals in the final 45 minutes.

Side Note: Insanely hard and was very time consuming as the data refused to be extracted to turn into a test, but in my opinon, the most important hypothesis.

In Conclusion

Overall, this study paints a picture of a league where historical trends are remarkably stable. While away teams are incrementally improving, they still face a statistically proven disadvantage regarding refereeing decisions. For teams looking to gain a competitive edge, focusing on second-half fitness and discipline in away games appears to be the most data-backed path to success.