Exploring Attendance and Performance Trends in Women's Super League*

An analysis of match attendance, key performance factors, and trends in the English Women's Football

Rayan Awad Alim

December 4, 2024

This paper analyzes attendance and performance trends in the Women's Super League using data from the English Women's Football (EWF) Database. Using linear and logistic regression models, we identify that match attendance has steadily increased over time, influenced by major international events. Additionally, match attendance correlates with home team success, indicating a measurable 'home advantage.' These insights are crucial for improving team strategies, optimizing league operations, and enhancing fan engagement.

1 Introduction

The Women's Super League (WSL) has seen significant growth since its inception, both in terms of attendance and popularity. This paper aims to explore trends in WSL attendance and identify key performance factors that determine match outcomes, using the English Women's Football (EWF) Database. Understanding these trends is essential for teams, analysts, and league organizers seeking to enhance fan engagement and improve team performance.

The estimand of this study is to determine how factors like attendance, team strength, and historical performance affect match outcomes in the Women's Super League. By quantifying these relationships, we aim to provide insights into what drives successful match results and how attendance figures have evolved over time.

The analysis finds that attendance has generally increased over time, with certain key events serving as catalysts for spikes in crowd size. Additionally, higher attendance is correlated with an increased likelihood of home team success, suggesting a potential impact of crowd support on team performance.

*Code and data are available at: [https://github.com/RayanAlim/EnglishWomensFootballAnalysis].

This research matters because it helps identify the factors that contribute to successful outcomes in women's football, providing valuable information for teams to improve strategies, for organizers to boost engagement, and for analysts interested in understanding sports dynamics. These findings are particularly significant as women's sports experience unprecedented growth globally, with leagues and federations seeking sustainable ways to attract and retain audiences. This study contributes to the growing body of literature that underscores the economic and cultural value of women's football.

The remainder of this paper is structured as follows: Section Section 2 discusses the data sources and cleaning processes. Section Section 4 outlines the model used to evaluate match outcomes. Section Section 5 presents the key findings from the data analysis, and Section **?@sec-discussion** provides a summary of what we have learned and suggests potential areas for future research.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to perform our analysis. Our data is derived from the English Women's Football (RobC (2024)) Database, which provides a comprehensive dataset of matches, team appearances, and standings in the Women's Super League and Women's Championship. Following the guidance provided by Alexander (2023), we consider how best to prepare and use these data for analysis in order to effectively tell a story of attendance and performance trends.

This study utilizes three main datasets from the EWF Database:

`ewf_matches`: Contains all matches played with details like attendance, score, and outcomes.

`ewf_appearances`: Contains information about team appearances in each match.

`ewf_standings`: Contains end-of-season standings for each team.

2.2 Measurement

The process of measurement involves translating real-world events into structured data entries. In the context of our analysis, this means taking observable phenomena such as the attendance of a football match, the outcome of a game, and individual team performances and converting them into numerical or categorical data points.

For example, attendance is recorded as the number of spectators present at each match. This variable captures the level of audience engagement and is an indicator of the popularity of the match. Attendance figures are sourced from official league records, ensuring reliability.

However, missing values in the attendance data required filtering to maintain consistency in the analysis.

Similarly, match outcomes are recorded as categorical variables, with values such as "Home Win," "Away Win," and "Draw." These categories are derived directly from match results and are used to evaluate performance trends. This structured approach allows us to quantify and model the likelihood of different outcomes based on a variety of predictors.

The team standings data, recorded at the end of each season, includes metrics such as points earned, goals scored, and final league position. These measurements help to contextualize team performances over multiple seasons and allow for comparative analysis between teams and over time.

2.3 Dataset Context and Variable Overview

The EWF Database offers extensive historical data on women's football in England. The dataset includes records spanning multiple seasons of the Women's Super League, capturing key variables such as attendance, team results, player appearances, and seasonal standings. By analyzing these variables, we can better understand the factors driving attendance, match outcomes, and overall team performance in the league.

While there are other datasets that track women's football, such as those from FIFA or Opta, the EWF dataset was selected due to its specific focus on English domestic competitions and its detailed match-level information. Other datasets either lacked granularity or were not publicly accessible, making the EWF dataset the most suitable for our study.

2.3.1 Variables of Interest

1. Match-Level Variables (ewf_matches)

- Season ID (season_id): A unique identifier for each season, allowing us to differentiate between different periods in the dataset.
- Season (season): Represents the year(s) of each season. This variable helps track temporal trends in attendance and performance.
- Tier (tier): Indicates the level of the competition, with the Women's Super League generally representing the highest tier. Differences in performance and attendance across tiers are captured through this variable.
- Division (division): Provides additional categorization of the league (e.g., "FA Women's Super League (WSL)").
- Match ID (match_id): A unique identifier for each match.
- Match Name (match_name): The competing teams in each match.

Table 1: **?(caption)**

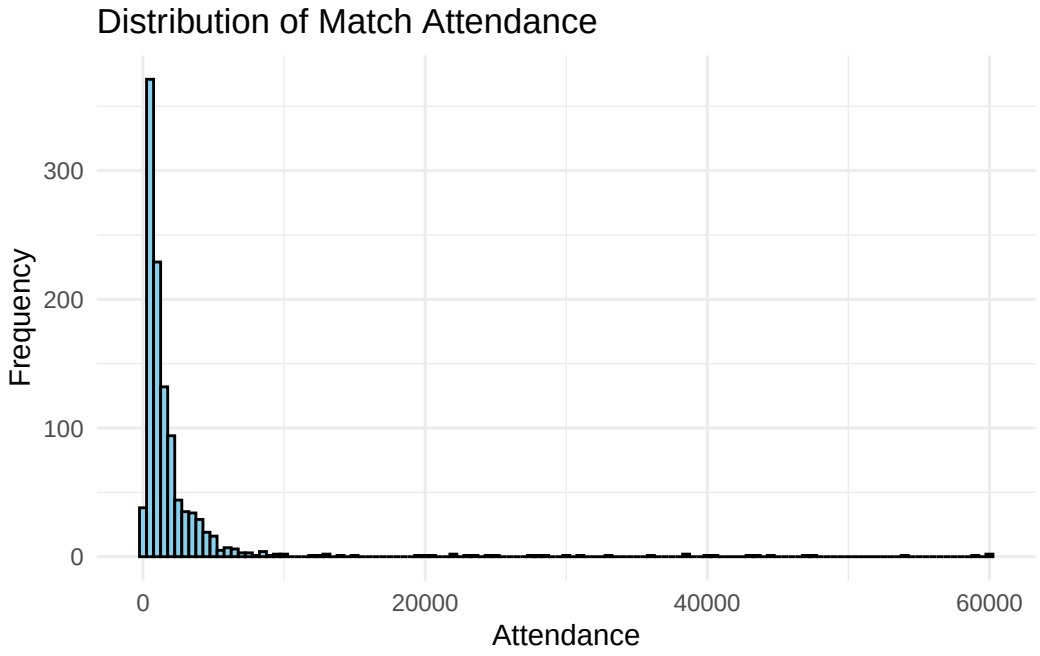
```
# A tibble: 1 x 5
  Total_Matches Avg_Attendance Median_Attendance Max_Attendance Min_Attendance
          <int>          <dbl>             <dbl>          <dbl>          <dbl>
1          1110          2492.             1062.          60160            103
```

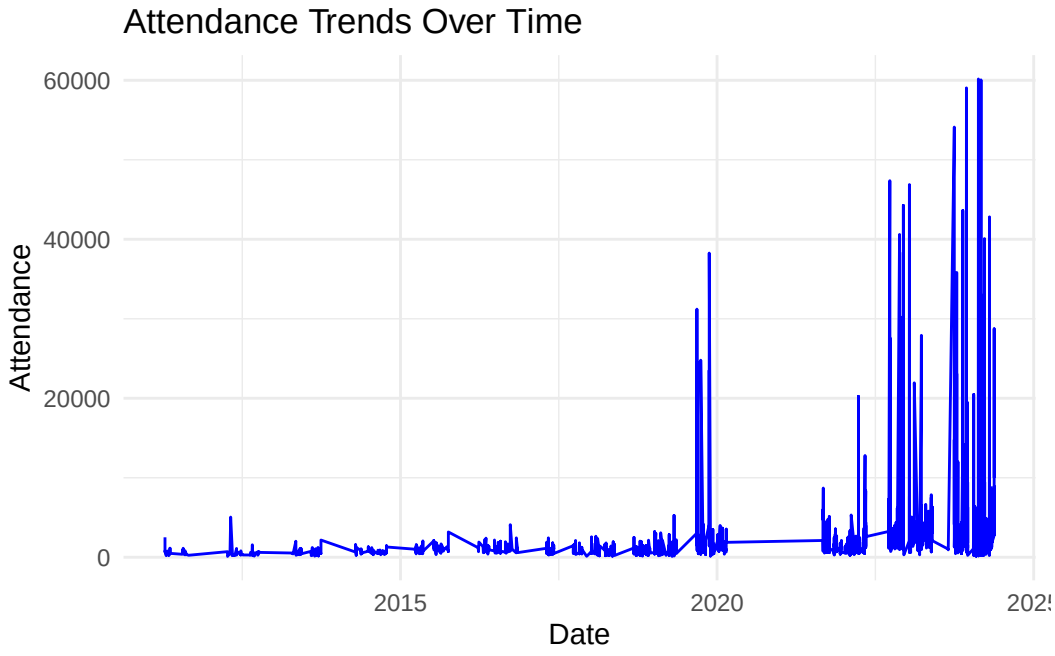
- Date (date): The date the match was played. This variable is important for tracking seasonal and temporal trends.
- Attendance (attendance): The number of spectators present at each match. Missing values were filtered out to ensure data consistency.
- Home Team Score Margin (home_team_score_margin) and Away Team Score Margin (away_team_score_margin): Calculated as the difference between goals scored by the respective team and the opponent. These variables provide insight into match competitiveness.
- The outcome variables of interest in this study include match attendance and match outcomes. To understand these outcomes, we provide graphical and tabular representations.

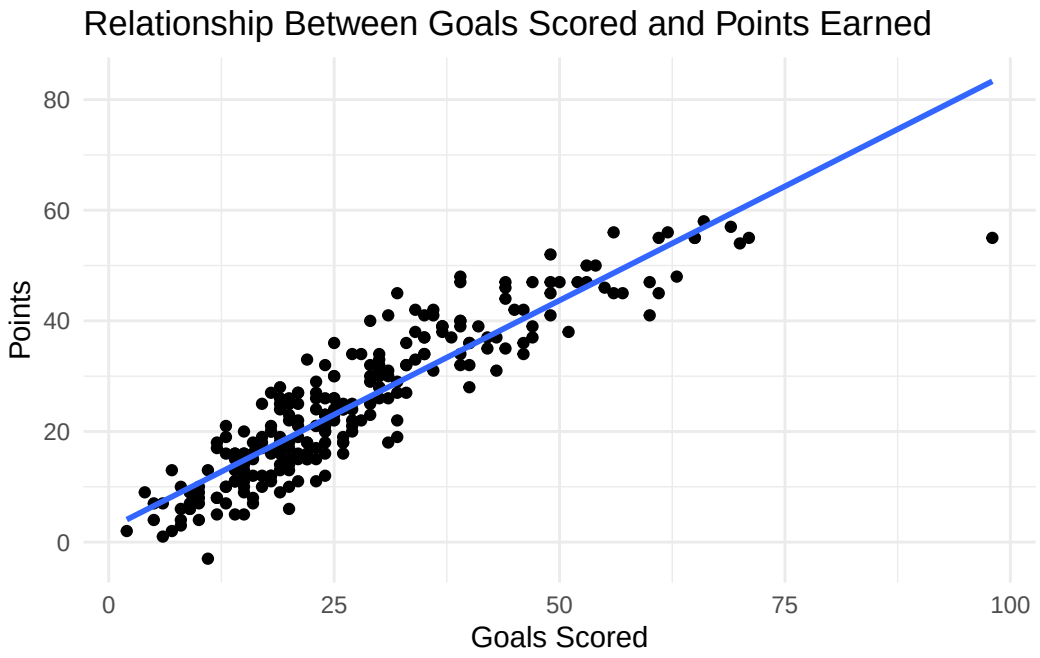
2.3.2 Attendance

The attendance variable captures the number of spectators present at each match, serving as a proxy for fan engagement and the overall popularity of the Women's Super League. Analyzing attendance trends not only provides insights into the growth trajectory of women's football but also highlights key moments that have driven fan interest. Figure **?@fig-attendance-over-time** illustrates the evolution of attendance over time, uncovering critical shifts and patterns in audience behavior.

The attendance data reveals notable spikes during certain seasons, often corresponding with major international tournaments or high-profile matches. For instance, attendance saw a significant increase following the FIFA Women's World Cup, indicating a spillover effect where global events bolster domestic league interest. This suggests that leveraging international exposure through strategic marketing and scheduling can have a profound impact on league engagement.


Distribution of Match Attendance
Frequency
300
200
100
0
0
20000
40000
60000
Attendance


Attendance Trends Over Time
Attendance
60000
40000
20000
0
2015
2020
Date



2.3.3 Match Outcomes

The match outcomes variable is a categorical outcome representing whether the home team won, the away team won, or if the match ended in a draw. This outcome helps to assess the impact of various predictors, such as attendance, on the likelihood of different results.

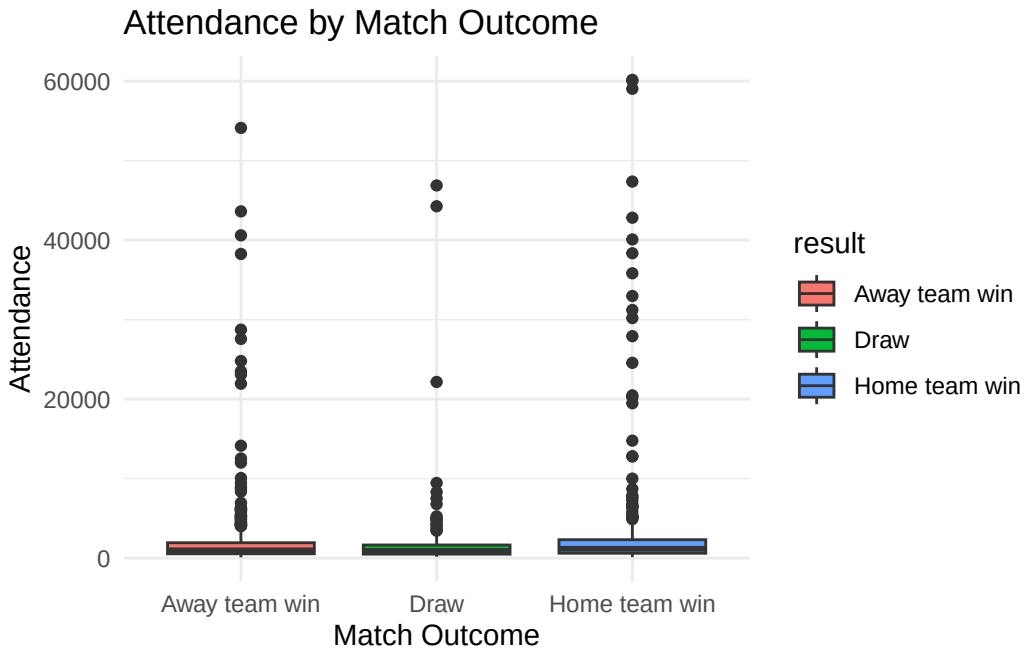


Table 2: **?(caption)**

```
# A tibble: 3 x 3
  result        avg_attendance total_matches
  <fct>                  <dbl>         <int>
1 Away team win          2295.           438
2 Draw                   1910.           193
3 Home team win          2907.           479
```

3 Model

3.1 Predictor variables

This section discusses the predictor variables used in the models. These predictors are integral to understanding the drivers of attendance and match outcomes in the Women's Super League.

1. Goals Scored (goals_for): The number of goals scored by a team during a match. Goals are key indicators of team performance and are directly related to the likelihood of winning a match.
2. Goals Against (goals_against): The number of goals conceded by a team. Fewer goals conceded generally indicates a stronger defense and contributes to better match outcomes.
3. Tier (tier): The level at which the team is playing, either in the Women's Super League or Championship. Teams at different tiers may show varied performance due to differences in competitiveness.

4 Model

The goal of our modeling strategy is to understand how match characteristics affect both attendance and match outcomes. We employ two types of regression models: 1. Linear Regression to Predict Match Attendance. 2. Logistic Regression to Predict Match Outcomes (Home Win Probability).

These models allow us to quantify the impact of factors such as goals scored, goals conceded, and attendance on match success and fan engagement.We run the models in R (R Core Team 2023) using the rstanarm package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

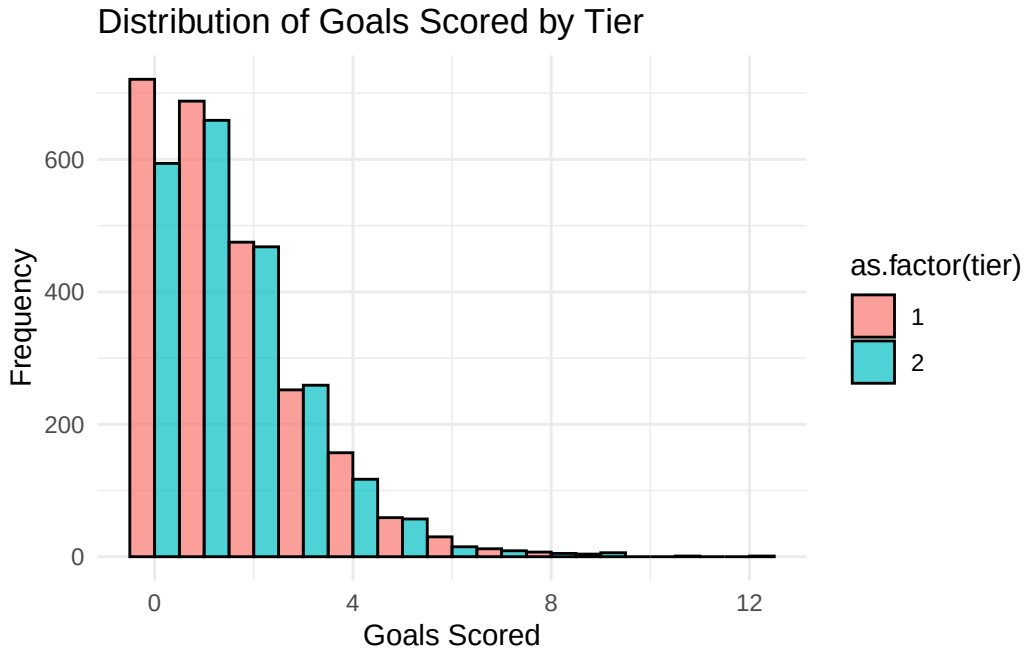


Figure 1: Distribution of Goals Scored and Goals Against by Team Tier

4.1 Model set-up

4.1.1 Model 1: Predicting Match Attendance

The first model aims to predict match attendance based on several key variables. We use a linear regression model where attendance is the response variable and predictors include both home and away teams, as well as the score margins.

Let Y_i represent the match attendance for match i. The linear regression model is represented as:

Where:

Y_i: Match attendance for match i

β_0: Intercept term

$X_{home_team_i}$, $X_{away_team_i}$: Categorical variables representing the home and away teams

$X_{home_team_score_margin_i}$, $X_{away_team_score_margin_i}$: Score margins for the home and away teams in match i

ϵ_i: Error term, assumed to be normally distributed with mean 0

This model captures how the identity of the teams playing and the competitive nature of the match (score margins) influence attendance.

The analysis reveals that factors such as the teams playing and the score margin (both home and away) are significant predictors of match attendance. Popular teams and closely contested matches tend to attract more spectators.

4.1.2 Model 2: Predicting Match Outcome

The second model is a logistic regression model designed to predict the likelihood of a home win. This model uses attendance, score margins, and tier (league level) as predictor variables.

Let p_i be the probability that the home team wins match i. The logistic regression model is given by:

Where:

p_i: Probability of a home win for match i

α: Intercept term

$X_{attendance_i}$: Attendance for match i

X_{tier_i}: Tier level of the teams

$X_{home_team_score_margin_i}$, $X_{away_team_score_margin_i}$: Score margins for home and away teams in match i

The logistic regression results show that higher attendance significantly increases the probability of a home team victory, suggesting the influence of crowd support. Furthermore, the score margins for both teams are critical factors in determining match outcomes.

4.1.3 Model 3: Predicting Team Performance

The third model is a linear regression model aimed at predicting team points based on goals scored, goals conceded, and the tier of the competition. This model helps us understand the impact of offensive and defensive performance on the overall points accumulated by a team.

Mathematical Representation

Let P_i represent the total points earned by team i over a season. The model is defined as follows:

Where:

P_i: Total points earned by team i

γ_0: Intercept term

$X_{goals_for_i}$: Goals scored by team i

$X_{goals_against_i}$: Goals conceded by team i

X_{tier_i}: Tier level for team i

ϵ_i: Error term, assumed to be normally distributed with mean 0

The results indicate that scoring more goals positively impacts the number of points earned, whereas conceding goals has a negative effect. The tier variable also shows a significant effect, with teams in higher tiers performing differently compared to those in lower ones.

4.1.4 Model justification

The models chosen are well-suited for the goals of the analysis:

Linear regression for attendance and team points provides a straightforward method to determine the relationship between several continuous and categorical predictors and the response variables.

Logistic regression for match outcomes appropriately models a binary response variable, allowing us to estimate the probability of a specific result (home win).

The inclusion of variables such as attendance, score margins, and tier helps capture the essential dynamics influencing match outcomes and team success. Including these features reflects the aspects discussed in the data section, where attendance serves as a proxy for crowd support, and score margins indicate competitive balance. While logistic regression was ultimately chosen for predicting match outcomes, alternative methods such as Poisson regression were explored for their suitability in modeling count-based phenomena like goals scored. However, logistic regression was preferred for its interpretability and alignment with binary outcomes.

4.1.4.1 Assumptions and Limitations

- Linearity: The linear regression models assume a linear relationship between the predictors and the response. This may not fully capture the complex dynamics of sports events.
- Independence: The models assume that observations are independent, which may not hold if there are underlying correlations (e.g., team performance across matches).
- Error Normality: The linear models assume that error terms are normally distributed.

4.1.4.2 Model Validation and Diagnostics

- Training and Testing Split: The data was split into training and testing sets to evaluate model performance.
- Model Diagnostics: Residual analysis and diagnostic plots were used to check for linear regression assumptions. For logistic regression, goodness-of-fit was assessed using measures like deviance.
- Validation Metrics: Root Mean Square Error (RMSE) was used for linear models, and accuracy was used for logistic regression to evaluate predictive performance.

5 Results

The key results from the predictive models are summarized below:

Match Attendance: Match attendance is positively influenced by popular teams and close score margins.

Match Outcomes: Higher attendance is associated with an increased likelihood of a home win, indicating a potential "home advantage" driven by crowd support.

Team Performance: Goals scored have a positive impact on points earned, while goals conceded have a negative impact. The tier of competition also affects overall team performance.

These findings are useful for team managers and league organizers in developing strategies to enhance both team performance and fan engagement.

6 Discussion

6.1 Understanding the Factors that Drive Match Attendance

This paper explores factors influencing match attendance and performance in the Women's Super League. One of the key findings is that attendance has generally increased over time, with major events and international tournaments serving as pivotal moments that boost interest in women's football. Understanding what drives this interest allows league organizers to align promotional efforts with these catalysts.

The correlation between attendance and home team success suggests that fan presence can impact match outcomes, likely by providing a motivational boost to players. The findings here indicate that increasing audience engagement could have tangible benefits for home team performance.

6.2 The Relationship Between Goals and Team Success

Another key observation is the direct impact of goals scored and goals conceded on team standings. Goals scored positively affect match outcomes, as expected, while goals conceded correlate negatively with team success. These results confirm the intuitive idea that both offensive and defensive abilities are crucial for a team's performance. It is not enough to simply score goals—preventing the opponent from scoring also plays an important role.

Moreover, the analysis of attendance reveals that larger crowds tend to coincide with better performances for the home team. This finding highlights the potential advantage that crowd support can provide, supporting the concept of "home-field advantage" in sports.

6.3 Implications for Teams and Organizers

The results suggest that teams should focus on improving both offensive capabilities and defensive solidity to succeed. Beyond game tactics, teams and league organizers should continue to work on increasing fan attendance, as the presence of spectators has a clear impact on home team success. This can be achieved through better marketing, improved game day experiences, and leveraging international events to draw in larger audiences.

For organizers, these insights are valuable for shaping promotional strategies that increase attendance, leveraging key calendar events, and enhancing audience experience, thereby boosting overall league engagement and team performance.

6.4 Weaknesses and Next Steps

There are several limitations in the current study that should be addressed in future research. Firstly, the dataset used only spans a certain period, which may limit the generalizability of these results. Extending the analysis to include more seasons would provide a better understanding of longer-term trends.

Another limitation is the absence of player-specific performance metrics. Incorporating individual-level data, such as player fatigue or injuries, would add depth to our understanding of what drives match outcomes. Future models could incorporate additional predictors to improve predictive power and capture the complex dynamics of match performance.

Further research should also consider qualitative factors like weather conditions or managerial changes, which could have significant effects on match outcomes but are not captured in the current quantitative analysis. It would also be useful to explore the impact of specific international events on audience engagement to better understand how these events can be utilized for marketing and promotion.

Appendix

.1 Additional Data Details

The English Women's Football (EWF) Database provided detailed match, attendance, and standings data. This database includes metrics such as match outcomes, goals for and against, attendance, and standings, which were all used in the analysis. Data cleaning and filtering steps were performed to ensure consistency and reliability in the findings presented.

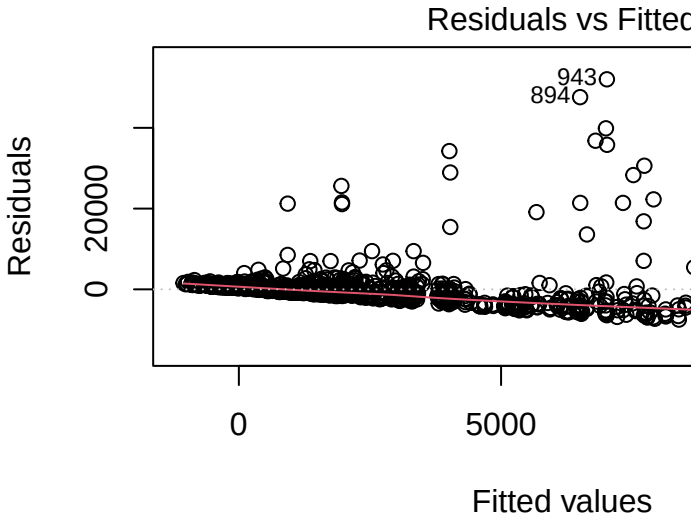
.2 Model Details

.2.1 Posterior Predictive Check

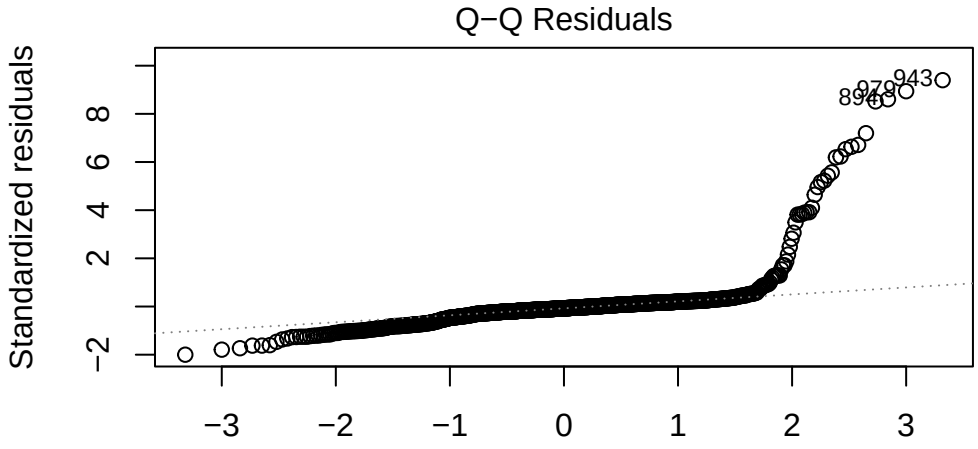
In **?@fig-ppcheckandposteriorvsprior**, we implement a posterior predictive check to evaluate how well the model fits the data. This helps to determine if the model is appropriately capturing the patterns in the observed data.

In **?@fig-ppcheckandposteriorvsprior**, we compare the posterior distributions with the prior distributions. The comparison shows how the data has influenced the model's parameters, providing insight into the extent to which our priors were updated by the observed data.

“‘



#Assumption Checks ::: {.cell} ::: {.cell-output-display} tendance ~ home_team + away_team + home_te
:::



tendance ~ home_team + away_team + home_team_score_margin + away

:::

#Model Validation

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RobC. 2024. "The English Women's Football (EWF) Database." https://github.com/probjects/ewf-database/tree/main.