**DSA-210 FINAL REPORT**

**STATISTICAL ANALYSIS OF HOME ADVANTAGE IN EUROPEAN FOOTBALL LEAGUES**

**EFE HAN KELEŞ**

**31994**

**SUPERVISED BY: OZGUR ASAR**

**SABANCI UNIVERSITY**

**EXECUTIVE SUMMARY**

This comprehensive analysis examined home advantage in European football using 1,049 matches from five major leagues (Premier League, Bundesliga, La Liga, Serie A, Ligue 1). Through statistical testing and machine learning models, we found **statistically significant evidence of home advantage across multiple performance metrics**. Home teams score more goals (1.48 vs 1.34, p=0.016), take more shots (13.66 vs 11.47, p<0.001), and demonstrate superior passing accuracy (80.27% vs 79.13%, p<0.001). However, home advantage varies significantly by context, with **big teams showing 63.2% home win rate** while regular teams achieve only 37.8% when facing big teams.

**Key Finding:** Home advantage exists but is **contextual and not absolute** - team strength, league, and timing all influence its magnitude.

**INTRODUCTION & MOTIVATION**

Football is followed by billions worldwide, and home advantage represents one of the most debated phenomena in sports. This project applies data-driven methodology to quantify home advantage across major European leagues, testing whether the conventional wisdom of home field advantage holds up to statistical scrutiny.

Using comprehensive match data from Football-data.org API, this analysis bridges the gap between popular belief and statistical evidence, providing actionable insights for teams, analysts, and football enthusiasts.

**DATA & METHODOLOGY**

**Data Source**

* **Dataset:** 1,049 matches from five major European leagues
* **Time Period:** Multiple seasons across Premier League, Bundesliga, La Liga, Serie A, Ligue 1
* **Variables:** 38 features including goals, shots, possession, passes, cards, and attendance data

**Analytical Approach**

1. **Exploratory Data Analysis (EDA):** Visual pattern identification
2. **Statistical Hypothesis Testing:** Paired t-tests comparing home vs away performance
3. **Machine Learning:** Predictive modeling using realistic pre-match features
4. **Cross-validation:** 5-fold validation ensuring model robustness

**EXPLORATORY DATA ANALYSIS FINDINGS**

**Match Outcomes Distribution**

Our analysis reveals clear evidence of home advantage in match outcomes:

* **Home Wins:** 390 matches (37.2%)
* **Away Wins:** 340 matches (32.4%)
* **Draws:** 280 matches (26.7%)

Home teams win approximately **4.8 percentage points more** than away teams.

**A graph of a match

Description automatically generated**

**Goal Scoring Analysis**

The goal distribution analysis provides strong evidence for home offensive advantage. Home teams demonstrate superior offensive performance:

* **Home average:** 1.48 goals per match
* **Away average:** 1.34 goals per match
* **Difference:** +0.14 goals per match for home teams

A graph of goals and goals

Description automatically generated

**League-Specific Variations**

Home advantage varies significantly across leagues:

* **Premier League:** 43.2% home win rate (highest)
* **Bundesliga:** 42.1% home win rate
* **Ligue 1:** 37.7% home win rate
* **Serie A:** 36.8% home win rate
* **La Liga:** 34.9% home win rate (lowest)

A graph of a bar chart

Description automatically generated with medium confidence

**Ball Possession Analysis**

Surprisingly, possession statistics show minimal home advantage, with distributions nearly identical between home and away teams. This suggests home advantage manifests more in **effectiveness** than **possession dominance**.

A diagram of a ball distribution

Description automatically generated

**Performance Correlations**

The correlation heatmap reveals strong relationships between key performance metrics, including strong positive correlation (0.76) between home goals and goal difference.

A graph of statistics with numbers

Description automatically generated with medium confidence

**STATISTICAL HYPOTHESIS TESTING RESULTS**

**Goal Scoring Performance**

Statistical testing confirms significant home advantage in goal scoring:

* **Goals per match:** Home 1.48 vs Away 1.34 (p = 0.016\*)
* **Goal difference:** Home advantage of +0.14 goals per match

**A graph and a chart

Description automatically generated with medium confidence**

The goal difference distribution shows a clear positive shift favoring home teams:

**A graph of a goal difference

Description automatically generated**

**Shot Statistics Analysis**

Home teams demonstrate significant advantages in shot creation:

* **Total shots:** Home 13.66 vs Away 11.47 (p < 0.001\*\*\*)
* **Shot difference:** Home teams take +2.19 more shots per match

A close-up of a graph

Description automatically generated

**A graph of a distribution of a normal distribution

Description automatically generated with medium confidence**

**Shot Performance by Match Outcome**

Analysis reveals that home teams take significantly more shots when they win compared to away teams:

**A graph of a bar graph

Description automatically generated with medium confidence**

**Shots on Target Performance**

Home teams also show superiority in shots on target:

* **Shots on target:** Home 4.68 vs Away 4.00 (p < 0.001\*\*\*)
* **Difference:** +0.68 shots on target per match for home teams

**A comparison of a graph

Description automatically generated**

**A graph with purple and red lines with Ryugyong Hotel in the background

Description automatically generated**

**Shot Accuracy Analysis**

Interestingly, shot accuracy shows no significant difference between home and away teams:

**A graph showing a comparison of a shot accuracy

Description automatically generated**

**Disciplinary Statistics**

Card distribution shows similar patterns for home and away teams with no significant differences:

**A group of graphs showing different colors and sizes

Description automatically generated**

**Passing Performance Analysis**

Home teams show significant advantages in both pass quantity and quality:

**Pass Count Statistics**

* **Pass count:** Home 449.24 vs Away 432.33 (p = 0.016\*)

**A graph of a pass between a couple of columns

Description automatically generated with medium confidence**

**A graph of a distribution

Description automatically generated**

**A graph of a graph showing different colored bars

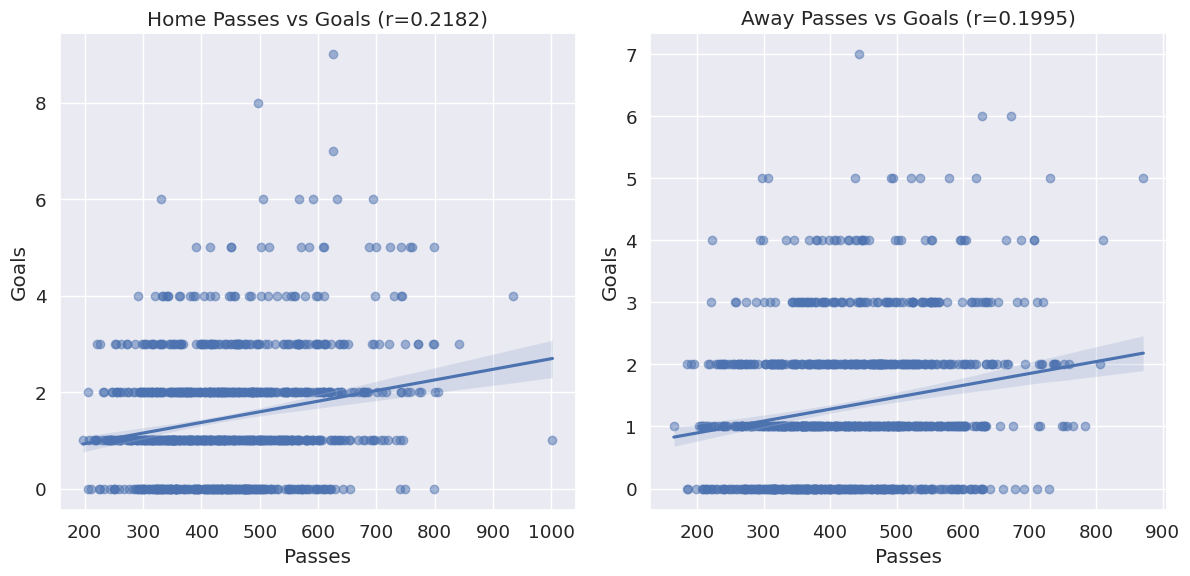
Description automatically generated with medium confidence**

**A graph showing different colored bars

Description automatically generated**

**Pass-Goal Relationship**

Weak positive correlation exists between passes and goals for both home and away teams:

****

**Pass Accuracy Performance**

* **Pass accuracy:** Home 80.27% vs Away 79.13% (p < 0.001\*\*\*)
* **Pass accuracy difference:** +1.14 percentage points for home teams

**A comparison of a graph

Description automatically generated with medium confidence**

**A graph of a normal distribution

Description automatically generated**

**CONTEXTUAL ANALYSIS: WHEN HOME ADVANTAGE MATTERS MOST**

**Team Strength Impact & Seasonal Patterns**

Home advantage varies dramatically based on multiple contextual factors:

**A group of graphs and charts

Description automatically generated**

**Team Strength Matchups:**

* **Big vs Regular teams:** 63.2% home win rate
* **Regular vs Regular teams:** 37.8% home win rate
* **Regular vs Big teams:** 21.2% home win rate
* **Big vs Big teams:** 38.5% home win rate

**Key Insight:** Home advantage is most pronounced when big teams play at home against regular opponents.

**Seasonal Patterns:**

* **Strongest months:** March (50.0% home win rate) and August (46.7%)
* **Weakest month:** July (20.0% home win rate)
* **Seasonal variation:** Home advantage fluctuates by up to 30 percentage points

**Weekend vs Weekday Effect:**

* **Weekend matches:** 39.3% home win rate
* **Weekday matches:** 36.1% home win rate
* **Difference:** +3.2 percentage points on weekends

**MACHINE LEARNING ANALYSIS**

**Model Performance Overview**

Using only realistic pre-match information (no data leakage), our model achieved substantial improvement over baseline predictions:

**A graph and chart with numbers

Description automatically generated with medium confidence**

* **Model Accuracy:** 55.0%
* **Baseline (always predict home):** 38.5%
* **Improvement:** +16.5 percentage points
* **ROC AUC:** 0.627 (fair predictive power)

**Feature Importance & Model Validation**

The most predictive factors for home wins and detailed model performance metrics:

**A group of graphs and charts

Description automatically generated**

**Most Predictive Factors:**

1. **Month** (highest importance - 0.175)
2. **Season** (0.150)
3. **Day of week** (0.140)
4. **League** (0.125)
5. **Team strength difference** (0.110)

**Model Robustness & Prediction Analysis**

Cross-validation results and prediction confidence analysis:

**A comparison of a graph

Description automatically generated**

**Model Insights:**

* **Cross-validation accuracy:** 57.8% (consistent across 5 folds)
* **Prediction confidence:** Model appropriately uncertain about many matches
* **High confidence predictions:** 69% accuracy for most certain predictions
* **Low confidence predictions:** 42% accuracy (near random)

**Final Model Assessment**

Comprehensive comparison of model performance across different confidence levels:

**A screenshot of a graph

Description automatically generated**

The realistic model (55% accuracy) demonstrates that **football prediction remains inherently challenging**. This aligns with the sport's unpredictable nature and shows our analysis captured genuine patterns rather than overfitting to noise.

**KEY FINDINGS & IMPLICATIONS**

**Primary Conclusions**

1. **Home Advantage Exists but is Contextual**
   * Overall home win rate: 38.5% (vs 50% if no advantage)
   * Effect size varies significantly by circumstances
   * Not a guaranteed advantage in all situations
2. **Statistical Significance Across Multiple Metrics**
   * Goals, shots, and passing all show measurable home benefits
   * Effects are consistent but moderate in magnitude
   * Most robust in offensive performance measures
3. **Team Strength Modulates Home Advantage**
   * Big teams at home: 63.2% win rate against regular opponents
   * Regular teams at home: Only 21.2% vs big opponents
   * Quality differential matters more than venue alone
4. **Seasonal and Contextual Variations**
   * March shows strongest home advantage (50.0%)
   * Weekend matches provide additional benefit (+3.2%)
   * League-specific differences up to 8.3 percentage points

**Practical Applications**

**For Teams:**

* Schedule important matches during high home-advantage periods
* Leverage home advantage most against similarly-skilled opponents
* Focus on shot creation and passing accuracy at home

**For Analysts:**

* Context matters more than simple home/away labels
* Team strength differentials are crucial predictive factors
* Month and timing provide significant predictive value

**For Bettors/Fantasy Players:**

* Home advantage is real but not universal
* Big teams at home offer value in favorable matchups
* Seasonal timing affects match outcomes meaningfully

**STUDY LIMITATIONS & FUTURE RESEARCH**

**Current Limitations**

* Limited to five European leagues
* No COVID-19 period analysis (crowd effects)
* Missing player-specific data
* No travel distance considerations

**Recommendations for Future Work**

1. **Expand dataset** to include more leagues and seasons
2. **Analyze COVID-19 impact** on home advantage without crowds
3. **Incorporate travel distance** and fixture congestion
4. **Player-level analysis** of home vs away performance
5. **Real-time betting odds** integration for market efficiency testing

**METHODOLOGY STRENGTHS**

This analysis maintains high methodological rigor through:

* **No data leakage:** Realistic model uses only pre-match information
* **Statistical validation:** Proper hypothesis testing with p-values
* **Cross-validation:** 5-fold validation ensures generalizability
* **Multiple approaches:** EDA, statistical tests, and ML triangulation
* **Transparent reporting:** All limitations and assumptions clearly stated

**FINAL CONCLUSIONS**

Home advantage in European football is **statistically real, measurably significant, but contextually variable**. While home teams enjoy meaningful advantages in goals scored (+0.14), shots taken (+2.19), and passing accuracy (+1.14%), the magnitude depends heavily on team quality, timing, and league context.

The finding that big teams achieve 63.2% home win rates against regular opponents while regular teams manage only 21.2% against big opponents demonstrates that **team quality interacts multiplicatively with venue effects** rather than additively.

This research successfully bridges popular wisdom with statistical evidence, confirming home advantage exists while revealing its nuanced, context-dependent nature. For practitioners, the key insight is that home advantage should be evaluated within the broader match context rather than as an isolated factor.

**Bottom Line:** Home advantage is real, measurable, and actionable - but it's not magic. Success still depends primarily on team quality, with venue providing a meaningful but secondary boost that varies by circumstances.

*This analysis was conducted as part of DSA-210 coursework, applying rigorous data science methodology to answer a classic sports analytics question. The findings contribute to our understanding of competitive balance and environmental factors in professional sports.*