DFB Pokal: Does it have its own rules?

— Seminar Report —
Visualization of Sports Data

Mohamed Elgabry, 269690

September 30, 2025

TU Dortmund University

Contents

List of Figures

1 Introduction & Motivation

The DFB-Pokal is one of the most unpredictable competitions in German football. Every year, stories emerge of small clubs challenging giants of the Bundesliga. A modest stadium, a team of semi-professionals, and suddenly they are playing against clubs with international stars and budgets many times larger. Sometimes, against all odds, the underdog wins. These "giant-killings" are part of what makes the Pokal so special, often summarized in the phrase "Der Pokal hat seine eigenen Gesetze" — the cup has its own rules.

What makes this competition fascinating is that it breaks with the usual order of football. In the league, the strongest teams tend to dominate over time. In the Pokal, everything can change in just one match. The imbalance between financial resources, player quality, and experience is clear, yet it can be overturned in ninety minutes. This tension between expectation and reality is at the heart of the tournament.

But the question is: how random are these surprises really? Are they just luck, or can we identify certain factors that make an upset more likely? For example, does the size of the division gap matter? Does it depend on the round of the competition? Has the frequency of upsets changed over the years? And finally, does the financial value of the teams — the "money factor" — actually decide outcomes, or does the Pokal ignore financial power?

The goal of this report is to move from stories to evidence. By analyzing match and player data, we want to understand whether the "own rules" of the Pokal can be explained by measurable factors. This does not remove the magic of the tournament — if anything, it helps us see more clearly what makes these surprises possible.

2 Problem Description and Objectives

2.1 Research Problem

Knockout tournaments like the DFB-Pokal are shaped by uncertainty. A single mistake, a moment of brilliance, or an unlucky bounce of the ball can change everything. Yet, when lower-division clubs defeat much stronger opponents, the question remains: **are these results completely random, or do certain structural factors make them more likely?**

Traditional explanations focus on chance or motivation — the idea that smaller clubs fight harder for their "game of the year." But these narratives do not fully explain why upsets occur so often and under such different conditions. To move beyond anecdotes, we need to ask whether measurable characteristics, such as the **gap in league divisions**, the

round of the competition, or the **financial strength of the teams**, systematically shape the probability of an upset.

2.2 Research Objectives

The aim of this study is to identify and quantify the main factors behind DFB-Pokal upsets. In particular, we focus on:

1. **Division gap** — does a larger structural difference between leagues reduce or increase the chances of a surprise?
2. **Round of competition** — are upsets more common in the early rounds, when stronger teams may underestimate opponents, or do they persist in later stages?
3. **Seasonal trend** — has the frequency of upsets changed over the years, and if so, in which direction?
4. **Financial value difference** — does the "money factor," measured as the gap in squad market values, strongly predict outcomes, or does the Pokal follow its "own rules" by making money less decisive?

By testing these objectives, the study seeks to uncover whether underdog victories follow patterns or whether they remain largely unpredictable.

2.3 Research Questions

To guide our analysis, we focus on the following questions:

- Does the **division gap** between teams significantly influence upset probability?
- Does the **round** of the tournament change the likelihood of surprises?
- Have **upset rates** increased or decreased over recent seasons?
- To what extent does the **money factor** (team value difference) explain outcomes, and where does it fail?

2.4 Scope and Limitations

This project focuses on DFB-Pokal matches between 2014 and 2025. Our dataset combines match results, team information, and squad values from Transfermarkt. While this allows us to analyze structural and financial differences, it does not capture tactical or psychological aspects of the game. In addition, market values are only a proxy for team strength, and may not reflect short-term factors such as injuries or form.

As a result, the study does not aim to build a perfect predictive model but rather to **explore which measurable factors correlate with upsets** and to what degree the Pokal's "own rules" can be grounded in data.

3 Data Source and Collection

3.1 Data Sources

All data used in this report was collected from **Transfermarkt.de**, one of the most comprehensive and reliable football databases available online. The scraping covered the period from the 2014/15 to 2024/25 seasons, ensuring that both recent and long-term tournament trends could be analyzed.

3.2 Data Structure and Content

Two main datasets were prepared from Transfermarkt data:

- **Match data** (dfb_matches.csv) — Contains match-level information, including match date, teams, final result, competition round, and team divisions. This dataset was primarily used for the win rate analysis across different factors such as division gap, round, and season trends.

- **Players data** (cleaned_players.csv) — Contains player- and team-level information, including aggregated squad market values at the time of each match. This dataset was merged with `dfb_matches.csv` to compute the financial value difference ("money factor") between teams.

3.3 Data Cleaning and Preprocessing

Data preparation was carried out in Python using Jupyter notebooks (`lower_div_win_rate.ipynb`, `team_value_diff1.ipynb`, `team_value_diff2.ipynb`). The key steps were as follows:

- **Win rate analysis (dfb_matches.csv)** — The dataset was already in match-level format. Only minor preprocessing was required, such as loading the file into `pandas`, filtering matches by division gap, round, or season, and calculating win rates for lower-division teams. The results were visualized with bar charts and line plots.

- **Handling missing player values (cleaned_players.csv)** — Player-level market values occasionally contained missing entries. These were imputed using the average value of all available players for the same team and season, ensuring no team had missing data for its squad valuation.

- **Team market value aggregation** — After filling missing values, player-level data were aggregated by team and season to compute the total squad market value. This provided the financial strength measure for each club.

- **Merging datasets** — The aggregated team values were merged into the match dataset (`dfb_matches.csv`) for both home and away teams. Matches missing value information for either side were dropped to maintain data consistency.

- **Derived features** — Several new features were constructed:
  - *Value difference*: the difference in total squad values between home and away teams (also log-transformed in later steps for scaling).
  - *Higher-value team*: indicator of which side had the stronger financial squad according to market values.
  - *Division comparison*: whether the home team, away team, or neither played in a higher division.
  - *Winner*: binary outcome variable (underdog win vs. favorite win).

The final dataset is match-level, combining structural information (division gap, round, season), financial strength (squad values), and outcomes, ready for descriptive analysis and regression modeling.

3.4 Limitations of Data Collection

While Transfermarkt provides one of the most complete publicly available datasets on football matches and squad values, some limitations remain:

- **Market values as proxies** — Player and team market values on Transfermarkt are estimates and reflect perceived market strength rather than objective performance. They may not capture short-term factors such as injuries, form, or tactical adjustments.

- **Incomplete historical data** — Earlier seasons occasionally lack complete player valuation information. Although missing values were filled using team-season averages, this introduces approximation into the dataset.

- **Scope of available data** — The dataset does not include detailed in-game performance metrics (e.g., expected goals, possession, or pressing statistics). As a result, the analysis is limited to structural factors (division gap, round, season) and financial strength.

- **Exclusion of certain matches** — Matches with missing team values for either side were dropped, which slightly reduces coverage across all seasons.

These limitations mean that the analysis highlights broad financial and structural patterns behind upsets, but cannot capture the full complexity of tactical or psychological influences on match outcomes.

4 Variables

4.1 Dependent Variable

The central outcome of interest is whether a lower-division team defeats a higher-division opponent in the DFB-Pokal. We define an **upset** as any match where the team from the lower league wins.

Formally, the dependent variable is a binary indicator:

- 1 = Upset (lower-division team wins)
- 0 = No upset (higher-division team wins)

Draws are not possible in the DFB-Pokal since matches proceed to extra time or penalties until a winner is determined.

4.2 Independent Variables

The independent variables capture the structural and financial characteristics of each match. They are grouped into two categories:

A. Structural Variables

- **Division gap** — The difference in league levels between the two teams (e.g., Bundesliga vs. 3. Liga corresponds to a gap of 2). This measures the size of the quality difference implied by the competition structure.
- **Round** — The stage of the competition (e.g., 1st round, 2nd round, quarterfinal). This captures whether upsets are more frequent in earlier rounds when stronger teams may rotate or underestimate opponents.
- **Season** — The season identifier, used to examine whether upset frequency has changed over time and to implement leave-one-season-out cross-validation.
- **Home advantage** — A binary indicator (1 = team from the higher division plays at home, 0 = otherwise). Cup rules often grant home advantage to lower-division teams, which may influence results.

B. Financial Variables

- **Team value difference** — Calculated as the difference in squad market values (home minus away) based on Transfermarkt estimates. This is the key explanatory variable for the "money factor." To reduce skewness, values were also log-transformed in later regression models.
- **Higher-value team** — A categorical variable indicating which side (home or away) had the higher squad value.

5 Methodology

5.1 Analytical Approach

The analysis in this project is descriptive and exploratory, supported by statistical hypothesis testing. The goal is not to build a predictive model, but rather to examine whether measurable structural and financial factors are systematically associated with the occurrence of upsets in the DFB-Pokal.

The workflow consisted of three main steps:

1. **Data preparation** — Match data from `dfb_matches.csv` and squad values from `cleaned_players.csv` were cleaned, merged, and enriched with derived variables such as division gap, season, and team value difference.
2. **Win rate analysis** — The proportion of lower-division victories was calculated across different categories, including division gap, round of the competition, and season. This provided an overview of structural patterns in upset frequency.
3. **Money factor analysis** — Squad market values were aggregated at the team-season level and merged into the match dataset. The differences in financial strength between competing teams were compared to actual outcomes, highlighting the role of the "money factor."

5.2 Descriptive and Exploratory Analysis

The descriptive analysis focused on four dimensions:

- **Division gap** — Win rates of lower-division teams were calculated for different division gaps to test whether upsets become less likely as the league distance increases.
- **Round of competition** — Win rates were compared across rounds to see if upsets are concentrated in the early stages of the tournament.
- **Seasonal trends** — Annual upset rates were plotted to examine whether the frequency of upsets has changed over time.
- **Money factor** — Squad market values were used to measure financial strength. By comparing value differences in matches with and without upsets, we explored whether financial disparity is a reliable predictor of outcomes.

5.3 Statistical Hypothesis Testing

To complement the descriptive results, several hypothesis tests were conducted:

- **Chi-square tests** — Examined the association between upset probability and (i) competition round, and (ii) division gap.
- **Mann–Whitney U and Welch t-test** — Compared the absolute value differences in upset versus non-upset matches.
- **Proportion test** — Tested whether higher-value teams win significantly more than 50% of matches when facing teams from nearby divisions.

5.4 Visualization

The results were presented primarily through visual summaries:

- Bar charts for win rates by division gap and competition round.
- Line plots for upset frequency across seasons.
- Boxplots for comparing team value differences between upsets and non-upsets.

These visualizations were used together with statistical tests to highlight the role of structural and financial factors in DFB-Pokal outcomes.

5.5 Methodological Limitations

The analysis is limited to descriptive statistics and hypothesis testing. No predictive regression models were applied, and the results are therefore associational rather than causal. Market values are only proxies for squad strength and do not capture tactical, psychological, or situational factors. In addition, sample sizes in later rounds and at large division gaps are small, which may reduce statistical power. Despite these limitations, the methodology provides a robust framework for identifying broad patterns in the occurrence of upsets.

6 Descriptive Statistics and Exploratory Analysis

This section presents descriptive analyses of upset frequencies in the DFB-Pokal, structured around the four research questions. Visualizations are used to illustrate the main patterns, and the results are later supported by formal hypothesis tests.

6.1 Does the money factor matter?

- Match Outcome

We begin with the financial perspective, examining whether differences in squad market values explain upset probabilities. Figure 1 summarizes match outcomes when teams

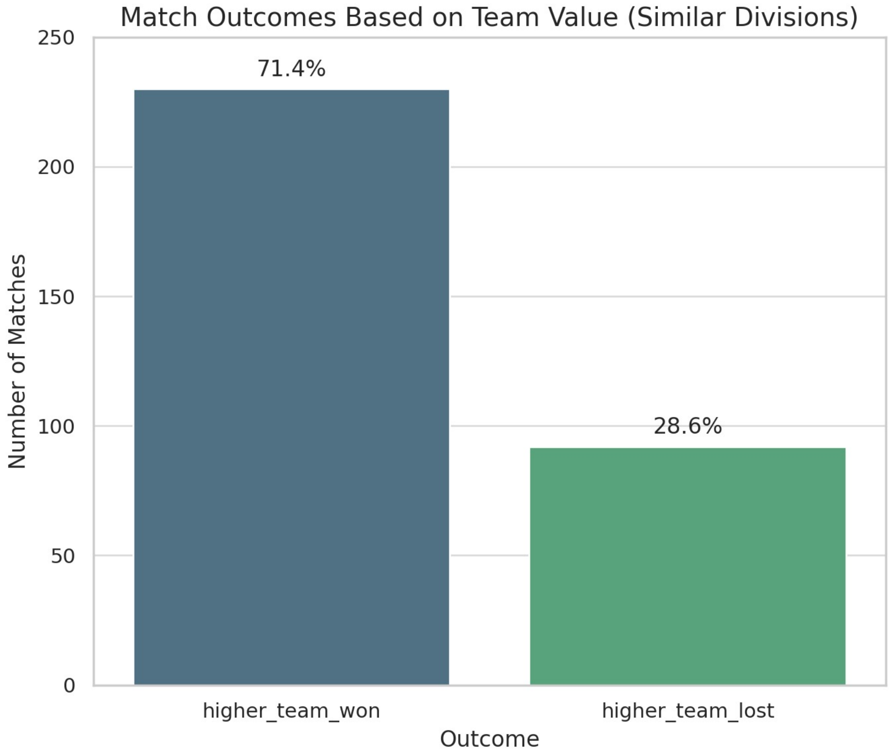

Figure 1: Match outcomes based on squad value differences (similar divisions).

from similar divisions (difference of at most one) faced each other. The richer team won 71.4% of these matches, while the poorer team managed to win only 28.6%.

To assess whether this pattern is statistically significant, we performed a one-sample proportion test. The null hypothesis assumed that the richer team would win 50% of matches by chance. The test result ($z = 8.51$, $p < 0.001$) shows that the richer team's win rate is significantly higher than 50%, with a 95% Wilson confidence interval of [66.3%, 76.1%].

In addition, Mann–Whitney U and Welch t-tests compared absolute squad value gaps between upset and non-upset matches. Both tests were significant ($p < 0.01$), showing that upsets occur when the financial gap is systematically smaller.

Taken together, these results confirm that the money factor plays a decisive role: large value gaps strongly favor the richer team, while smaller gaps open the door for underdog victories.

- Team Value Difference

Figure 2 shows When the higher-division team won (right side), the value difference was typically much larger, showing that financial superiority often translated into victory. When the higher-division team lost (left side), the gaps were noticeably smaller, suggesting that upsets occur mainly when the underdog was not far behind in financial strength.

In simple terms: large money gaps make victories for the favorite almost certain, while smaller gaps leave space for surprises.

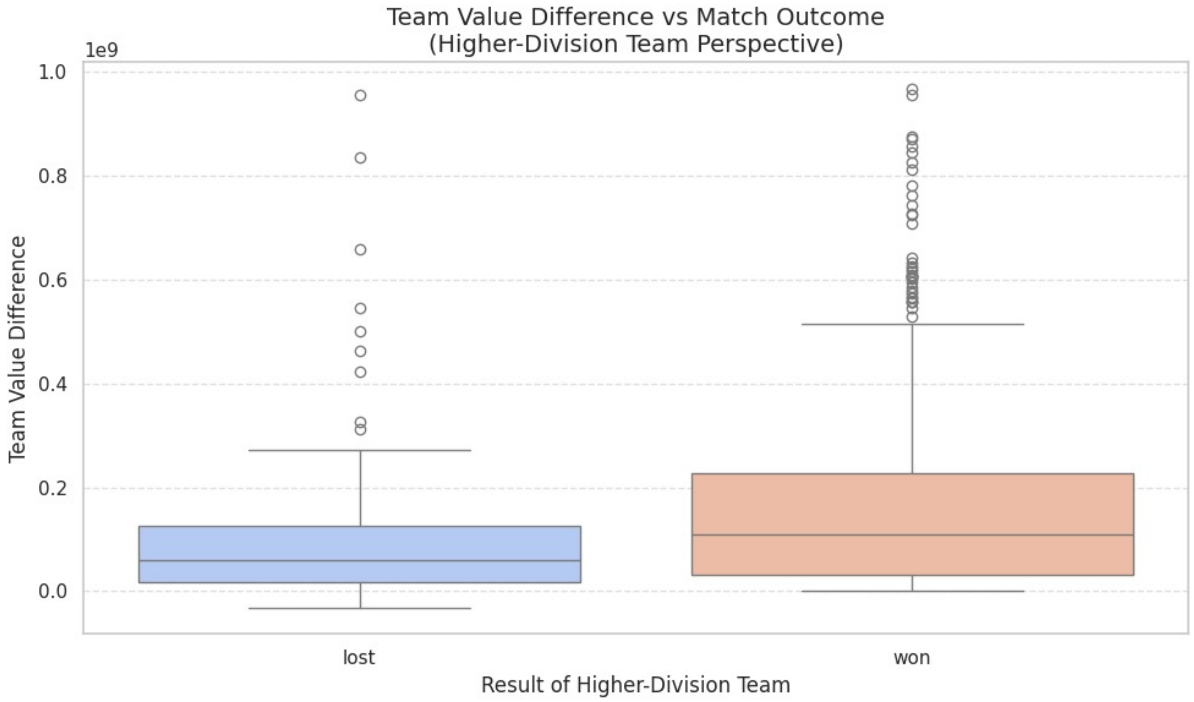

Figure 2: Team value difference vs. match outcome from the perspective of the higher-division team.

The hypothesis tests support this visual impression. Both the Mann–Whitney U test ($p < 0.001$) and Welch's t-test ($p = 0.0035$) show that the financial gaps in upset matches are significantly smaller than in matches won by the higher-division team. This confirms that large value differences strongly protect favorites, while smaller gaps increase the likelihood of an upset.

6.2 Do structural factors matter?

- Tournament Round

Figure 3 shows how the win rate of lower-division teams varies by round. In the first round, the upset rate is relatively low (17.4%), but it increases in the second (29.6%) and third round (28.3%). From the quarterfinals onwards, the win rate declines again, reaching 0% in the final.

In simple terms: upsets are more common in the middle rounds, but much less likely in the late stages of the tournament.

The hypothesis tests confirm this non-uniform distribution. A chi-square test shows that win rates differ significantly across rounds ($p < 0.001$). This suggests that the stage of the competition plays a crucial role in determining the likelihood of an upset.

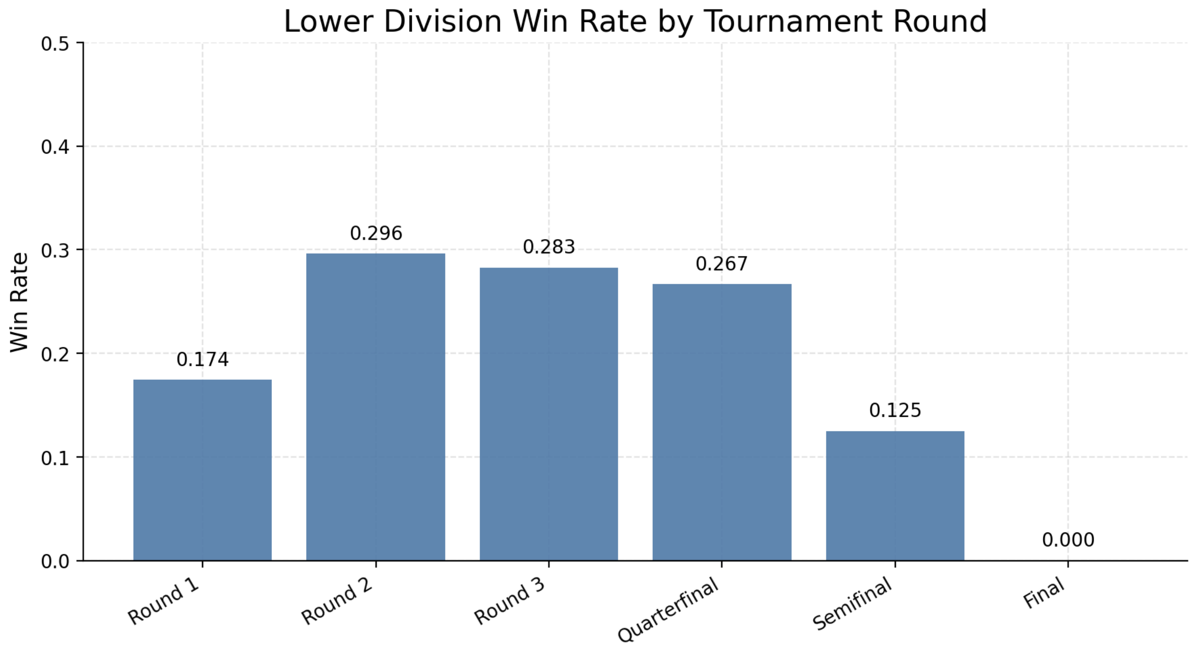

Figure 3: Lower-division win rate by tournament round.

- Division Gap

Figure 4 shows how the probability of an upset depends on the division gap between the two teams. When the gap is only one division, the underdog win rate is highest (26.2%). It drops slightly when the gap is two divisions (23.0%), and becomes very low when the gap is three divisions (9.1%).

In simple terms: the further apart the teams are in league level, the less likely it is for the lower-division side to cause an upset.

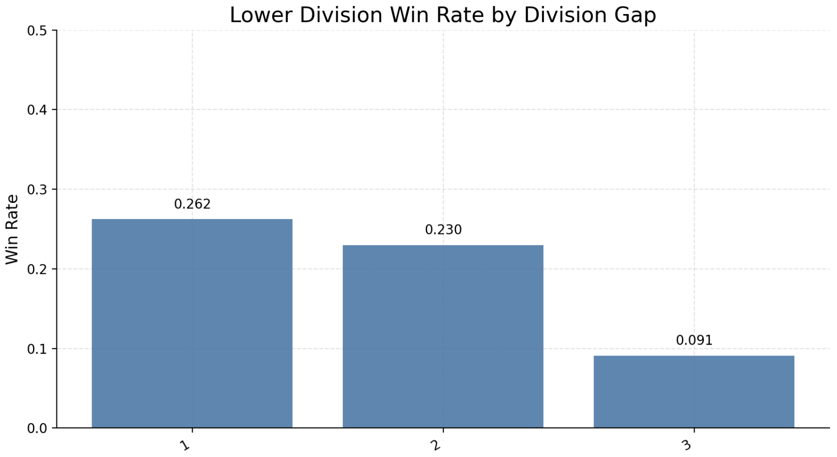

Figure 4: Lower-division win rate by division gap.

The hypothesis tests confirm this relationship. A chi-square test of independence finds a significant association between division gap and match outcome ($p < 0.001$), showing that larger division gaps strongly reduce upset probabilities.

- Seasonal Trend

Figure 5 shows how the win rate of lower-division teams against higher-division opponents evolved across seasons. The values fluctuate between 0.17 and 0.29, without a clear upward or downward long-term trend. While some seasons (e.g., 2014 and 2023) show relatively high upset rates, others (e.g., 2017 and 2022) are noticeably lower.

In simple terms: the chances of underdogs winning do not steadily increase or decrease over time but instead vary from season to season.

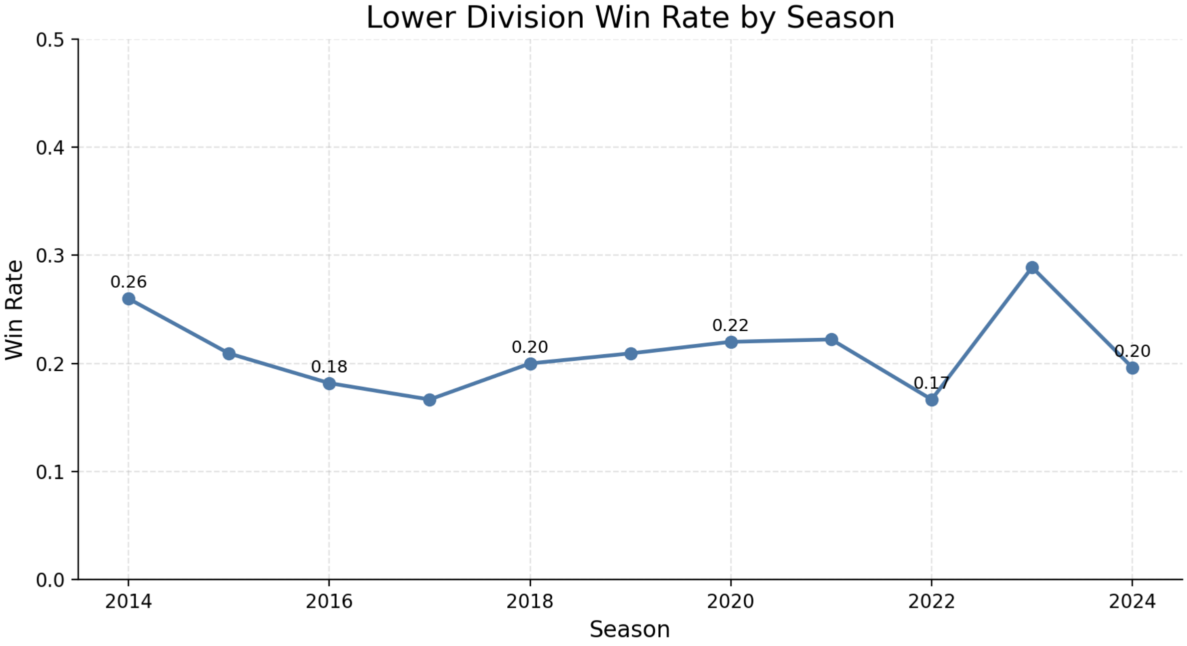

Figure 5: Lower-division win rate across DFB-Pokal seasons.

The hypothesis test results indicate that there is no significant trend over time. A logistic regression with season as predictor yields a non-significant coefficient ($p = 0.41$), confirming that the likelihood of upsets has not systematically changed across the observed period.

6.3 Key Insights

From the descriptive and exploratory analyses, several clear patterns emerged:

- **Financial strength matters.** When the higher-division team had a much larger squad value, upsets were rare. Smaller financial gaps, however, left room for surprises.

- **Structural factors play a role.** Upset probabilities decrease as the division gap widens, and they vary with the stage of the tournament: underdogs succeed more often in the middle rounds, but rarely in the final stages.

- **No systematic trend over time.** Although some seasons recorded unusually high upset rates, overall the likelihood of upsets has remained fairly stable across the observed period.

In short, the "magic of the cup" is not entirely random: it thrives when financial differences are small, division gaps are narrow, and the tournament context is favorable to the underdog.

7 Discussion and Conclusion

This study explored whether structural and financial factors explain the occurrence of upsets in the DFB-Pokal. The results suggest that while surprises remain central to the competition, they are not entirely random.

Starting with the **team value difference**, we found that the larger the financial gap between squads, the lower the chance of a lower-division victory. When values were closer, underdogs sometimes prevailed, but with wide gaps the probability dropped sharply, confirming the role of resources in shaping outcomes.

The **division gap** showed a similar pattern. Upsets were still achievable when only one tier separated the teams, but became rare with two or more divisions, reflecting structural differences in playing level and squad depth.

The **tournament round** also mattered. In the early stages, smaller clubs were more likely to surprise stronger opponents, while in later rounds the upset rate declined as top-tier clubs approached the competition more seriously.

These descriptive trends were supported by **hypothesis testing**, which confirmed significant effects for all three factors. Yet, the persistence of unexpected wins highlights why the Pokal remains special: inequalities shape the probabilities, but never remove the possibility of surprise.

In conclusion, the study shows that **team value difference, division gap, and tournament round are key drivers of upset likelihood**. At the same time, what keeps the DFB-Pokal exciting is that it never follows these rules completely. Even when the numbers suggest that an underdog has little chance, football still finds a way to surprise us. This mix of clear patterns and unexpected moments is exactly what makes the competition so special and why it continues to capture the imagination of fans every year.

Bibliography

- Transfermarkt. *Player, team, and match statistics database..* https://www.transfermarkt.com.

- Deutscher Fußball-Bund (DFB). *Official DFB-Pokal Competition Rules and Match Reports.* . https://www.dfb.de.