UNIVERSITÀ DEGLI STUDI DI GENOVA

SCUOLA DI SCIENZE SOCIALI

DIPARTIMENTO DI ECONOMIA

Corso di laurea in Economics & Data Science

Elaborato scritto per la prova finale in

Statistical Models

An Empirical Analysis of Home Advantage and Betting Odds in Football

Docente di riferimento:

Prof. Fabio Rapallo

Candidato:

Hunor Bendeguz Kovács

Anno academico 2023-2024

Table of Contents

List of Figures

List of Tables

Abstract

Home advantage is a widely known, well-documented phenomenon. It means that teams across different sports have a higher chance of winning when they play in their own stadium, i.e. they perform consistently better at home than away. Post-match statistics and pre-match betting odds reflect these dynamics. The research questions are: *How has home advantage changed over time and across Europe's football leagues? To what extent are post-match statistics able to explain the actual match result? How accurate are the pre-match outcome probabilities based on betting odds?* Hypotheses: I expect that home advantage *decreases over time*, with *significant differences* between the leagues; that *only a handful* of post-match statistics *contain significant information* about the result; and that pre-match outcome probabilities based on betting odds have *high predictive power*. To begin with, home advantage is discussed in detail through an extensive literature review. The empirical analysis is conducted on a vast amount of match data (from 1888-89 to 2022-23). First, this thesis examines the phenomenon of home advantage, highlighting changes over time and variations across countries. Then, multinomial logistic regression models are built to find out how much statistically significant information do post-match statistics contain regarding the full time result. Finally, the predictive ability of pre-match betting odds provides the full picture. The main results of six different models in eight European football leagues are presented in a summarizing table. Results: home advantage has consistently decreased over time. The most important quantitative factor contributing towards home advantage is the average number of goals scored per game. Regarding the forecasting ability of post-match statistics and pre-match betting odds, six different models across eight European football leagues are compared through a common indicator (predictive accuracy) and are presented in an informative and easy-to-understand summarizing table (Table 6), one of the main achievements of this thesis. The special case of the half time result is discussed separately. Regarding the models, the predictive accuracy of post-match statistics proved to be quite high in each analysed league (around 60%). However, they can only predict posteriorly. Betting odds are finalised pre-match and are able to forecast future outcomes. They are notably accurate (up to 55%) in the case of the top four leagues, less so for lower divisions (below 44%). Lastly, a connection is established between the different parts of the thesis: the match-influencing effects of home advantage are incorporated in the betting odds, further demonstrating the influence of home advantage on match results in football.

Acknowledgements

First and foremost, I am grateful to God for the uncountable blessings I have received in my life, for His infinite love and daily guidance. Next, I am thankful to my parents and siblings for the countless moments shared, the upbringing, the love, the support, and I could go on and on.

This thesis would have not been possible without Professor Fabio Rapallo, whom I thank for his lessons, insights, patience, and availability. I would also like to thank Professor Maurizio Conti, the coordinator of the course, for his kind help with onboarding and bureaucracy. It was a pleasure to study at the University of Genoa, graduating in Economics and Data Science MSc.

I am especially thankful for my greatest ever academic mentor, Professor Zoltán Madari, who has supported, helped, taught, corrected, and advised me through 3 years of the bachelor's course and even beyond. It was thanks to him, that I took a liking to statistics in the first place. From him I received much more than mere knowledge or understanding: way of thinking. Besides that, he led me to numerous academic achievements, including winning first place at the 36th National Scientific Students' Associations Conference. He is undoubtedly the most helpful, impactful, and hardworking professor at the Corvinus University of Budapest, Hungary. Thanks to him, my Applied Economics BSc degree has real worth.

I would like to acknowledge two of my high school teachers. Ms. Ildikó Papp and Mr. György Molnár. The former passed on strong mathematical and analytical skills, while the latter transmitted to me the affection for Italy and the Italian language. On top of that, they are both excellent human beings, living by faith. I am proud of the four years of shared experiences.

Furthermore, I would also like to thank the whole Italian and international community of dormmates, fellow students, teammates, colleagues, professors, flatmates who welcomed me with open arms and were vital in these past two years. Without my love for football, this work would not have been possible either, I am thankful to God for every second.

Finally, I wholeheartedly thank my very few loyal and true friends, who have brought happiness, empathy, strength, resilience, mental, emotional, and spiritual support, encouragement, and affirmation amidst duties, problems, challenges, doubts, difficulties in my everyday personal life, and with whom I have shared a plethora of joyful and golden moments.

"Bless the Lord, O my soul; and all that is within me, bless His holy name. Bless the Lord, O my soul, and forget not all His benefits: (…)" (Psalms 103,1-5)

I. Introduction

I.1. Motivation

Football has played a central role all my life. In addition to actually playing the sport since I was 5 years old up until a quite high amateur level (Italian 5th division) and coaching different youth teams, plus mentoring young players for almost a decade now, I am really interested in the historical facts, trends, data, stats connected to it.

This led to a number of analyses conducted in this field, which most notably resulted in participating in the Scientific Students' Associations Conference during my bachelor's course at the Corvinus University of Budapest, Hungary. With my research paper – focusing on the potential effect of the teams' form on home advantage – I achieved 3rd place in my category in May of 2022. This result enabled me to present the same work as my bachelor's thesis, receiving excellent grade, and more importantly qualified me to the final, countrywide phase. This was the 36th National Scientific Students' Associations Conference, held in April 2023, where glory to God I achieved first place. Finally, in November 2024, a shortened and edited version of my original research paper has been published in the fourth edition of Litera Oeconomiae, which contains a selection of the best studies presented at the previously referred Conference.

These successful experiences gave me the motivation to continue research in football, home advantage and different statistics and dynamics connected to it. The final idea was formulated during that process: diving into home advantage much more meticulously both from a theoretical and an empirical point of view, and at the same time looking for possible connections with match statistics and the match outcome predictions based on official betting odds.

I.2. Topic

The goal of this study is triple. Observe and analyse in detail the phenomenon of home advantage, understanding its dynamics. Then check how much information official post-match statistics contain regarding the full time result. Finally, compare the findings of the first two steps with the forecasting ability of pre-match betting odds, thus obtaining a full picture.

To achieve these goals, firstly, I am going to present – both from a theoretical and an empirical aspect – the complicated phenomenon of home advantage, its possible causes, trends, changes over time and differences across leagues. Secondly, a multinomial logistic regression model will be built to quantify how the different match statistics describe the match outcome. Their ability and/or inability to posteriorly predict the match outcome will be discussed and put into context. Thirdly and finally, simple probability computations will show how and why the

official betting odds have a relatively high forecast accuracy and how they incorporate both the effect of home advantage and the expected values of the upcoming match statistics.

Home advantage is a widely known, well-documented phenomenon. Briefly it means that teams across different sports have a higher chance of winning when they play in their own stadium. In other words, they perform consistently better at home than away. (Neave & Wolfson, 2003) But what could possibly cause this? There are a number of evident and unexpected factors, but first of all it has to be stated that while home advantage is present in various sports, this study focuses on football, the most popular sport in the world. (Matheson, 2003) While home advantage in football is not a conjecture but a proven fact, the possible factors and the way they influence the match result are still not clear today. The supporters of the home team seem to play a natural role in contributing towards home advantage. They constitute the higher proportion of the total match attendance and aim to encourage and help their team to victory. In order to achieve this, they try to disturb the players of the away team, forcing errors. Moreover, the home crowd always pressures the referee to favour the home team in questionable situations. Accordingly, crowd effects have been long believed to be significant, up until the matches played behind closed doors due to the COVID-19 lockdowns, which caused some doubt. (Schwartz & Barsky, 1977) (Wunderlich, Weigelt, Rein, & Memmert, 2021) Travel effects (fatigue) of the away team also seem logical. Territoriality, however, is neither widely known, nor obvious, but still plays a significant role. It is a biological fact that real or perceived invasion in one's home or territory triggers a response both in humans and animals. Neave and Wolfson (2003) showed empirically that a significant rise can be detected in the testosterone level of the home teams' players before a home game. All these factors will be discussed in detail during the literature review part (Chapter II.). After this short introduction it is already crystal clear that home advantage might seem simple at first glance, the reality is the contrary. Moreover, because of the different pace of the development and evolvement of professional football around the world, home advantage can be significantly different across countries, leagues, and time as well.

A similar line of thinking can be made in the case of match predictions. First, it is important to divide the "predictive" accuracy of all information, data and statistics which are computed or available after the match and the forecasting ability of betting odds or different methods resulting in match probabilities calculated before the game. The first group is much more of a classification method (the multinomial logistic model, decision trees or random forest being prime examples) both figuratively and from a methodological point of view. Models based on

match statistics try to correctly guess the match result which can be immediately checked, as the match has already been played and the real result is already available. The second group on the other hand aims to actually forecast the outcome of the game by assigning probabilities to the three possible outcomes (H – home win, D – draw, A – away win). These probabilities are a result of mathematical models based on historical data, future expectations, uncertainty, incorporating information such as home advantage and match statistics from previous games as well. The models are obviously unavailable to the public as they constitute to the competitive advantage of different bookmakers and in the big picture, the profitability of the gambling industry. This study does not go into detail about betting strategies but instead focuses on the illustration and comparison of the predictive accuracy of the two groups, i.e. the post-match statistics and the pre-match betting odds, easily transferable to outcome probabilities.

In conclusion, based on the above-presented arguments, it is important, economically relevant, and statistically justified to analyse the phenomenon of home advantage, the forecasting ability and predictive accuracy of post-match statistics and pre-match betting odds in the most prominent leagues of the most popular sport on the planet. This leads to the research questions.

I.3. Research Questions

Therefore, the formulated research questions of my thesis are:

1. *How has home advantage changed over time and across Europe's football leagues?*
2. *To what extent are post-match statistics able to explain the actual match result?*
3. *How accurate are the pre-match outcome probabilities based on betting odds?*

I.4. Hypotheses

Based on my own research, expectations and most importantly on thoroughly elaborated relevant studies, papers and their findings, the hypotheses belonging to the previously stated research questions of my thesis are:

1. Home advantage *decreases over time*, with *significant differences* between the leagues.
2. *Only a handful* of post-match statistics *contain significant information* about the result.
3. Pre-match outcome probabilities based on betting odds have *high predictive power*.

II. Literature Review

Naturally, before diving into my analysis, it is necessary to give an overview and brief summary of the relevant and important studies, papers and past research conducted in this field. The aim was – besides presenting their methods and findings – to highlight the interesting parts and facts of the topic. To make this part easier to read, and to follow the structure of this work, this chapter is divided into two main sections: home advantage and betting odds, with a much higher weight and focus on the detailed analysis of home advantage, as this is one of the main goals.

II.1. Home Advantage

Obviously, a number of still relevant research has been conducted in this topic. Out of these I am going to summarize those, which are most related to the aforementioned research questions. The structure of the literature review part is following the aspects and logic of Pollard (2008). It goes without saying, that the works published after 2008 will be integrated into the same framework. First, I am going to state the basic facts about home advantage, then the papers and other references will be grouped by the most important potential explanatory factors.

II.1.1. Home Advantage in General and in Football

One of the first studies, that basically proved the conjecture, i.e. the existence of the phenomenon of home advantage in sports, was the work of Schwartz and Barsky (1977). Their research not only presented supporting evidence for home advantage in different organized sports, but also fine-tuned its definition. They found that in the USA home advantage is most notably present in the case of indoor sports (ice hockey, basketball) and is the least significant in the case of baseball and American football (both are played outdoors). According to the authors, the most important factor of home advantage is the more effective offensive (and not the defensive) actions in all sports. Furthermore, the location of the game (home or away match) is as strongly connected to the team's performance as the average quality of the team's players. Finally, the analysed data, the attendances, the relationship between attendance and performance and between attendance and the outcome of the match showed that home advantage is almost absolutely independent of the away team's fatigue (caused by the travel) and unfamiliarity with the pitch. We will see that these findings have been partly disaffirmed in later studies. However, their most important conclusion still holds strong today (it will be discussed in detail in Chapter II.1.2 below), i.e. that the home advantage is mostly due to the (social) support and encouragement of the home fans. (Schwartz & Barsky, 1977)

Narrowing down to football, it has been established long ago that home advantage plays a crucial role in determining the result. The existence of this phenomenon surely affects the behaviour of players, coaches, referees, fans, even the media, and how they approach the match. Surprisingly, even after decades of research, it is still unclear what the precise causes of home advantage are and how exactly their influence is exercised in practice. Therefore, it is important to summarize, compare the pro and contra evidence when we present these potential explanations, and analyse them in the light of the following four basic and widely acknowledged facts. One, home advantage has been in existence from the beginning of organised football (at least since as early as the end of the 19th century). Two, it is a global phenomenon with significant variations in different countries. Three, its influence has declined over the recent decades in Europe's strongest championships, the so-called top leagues, which are (as of 2023-24, ordered by UEFA coefficients): England - Premier League, Spain – La Liga, Italy – Serie A. Germany – Bundesliga, France – Ligue 1. (UEFA, 2024) Four, in general it is more dominantly present in football, than in any other sport. (Pollard, 2008)

Home advantage in football has been first analysed by Morris (1981). Shortly afterwards, he was followed by Dowie (1982) and Pollard (1986) who described in detail the phenomenon of home advantage in football (both qualitatively and quantitatively), also lining out the suspected major grounds of its existence. In the subchapters below I present the review of the relevant literature in the light of these identified causes to get a structured overview.

II.1.2. Crowd Effects

This is the most obvious factor in connection with home advantage, and naturally fans as well believe it to be dominant according to the study of Wolfson, Wakelin and Lewis (2005). The authors asked fans through online shared questionnaires about their view on their own contribution towards creating home advantage. Based on their answers, crowd support has a significantly greater effect on home advantage than familiarity, travel effects, territoriality, or referee bias. Fans felt themselves to be responsible and took credit for pushing their team to victory, for distracting opponents and believed themselves to be able to influence the referees to make decisions in favour of their team. However, they declined to be personally at fault for the disappointing results of their team. This is independent of sex, age, and eventual standing at the end of the season, however, on aggregate season ticket holders had more extreme feelings towards the question on the topic of responsibility. Furthermore, the answers implied that mechanisms such as feeling superior to rivals can incentivise fans to remain loyal to their team even in the case of disillusioning results. There are many examples with even stronger

connections between a club and its supporters. The attachment and belongingness can be so deeply ingrained into fans' identity, that they do not anymore have the option of abandoning their team, but instead feel a mutual rapport, which requires from both sides (the supporters and the club) to do everything possible to achieve success. (Wolfson, Wakelin, & Lewis, 2005)

However, a number of researchers found it quite difficult to precisely determine the ways and means by which crowd support exerts its effect. (Dowie, 1982) For example, the relationship with attendance is still ambiguous, as home advantage has been proven to exist even with very few supporters attending the game. (Pollard, 1986) (Pollard & Pollard, 2005)

Besides attendance, the density of the fans (gathered in a crowd or scattered in the stands), the intensity of the support (chants, singing, etc.) and the vicinity to the pitch (there are stadiums where the seats are really close to the sidelines and there are older, multiuse ones where for example the running track causes the fans to be farther away) are all factors that should be taken into account. Similarly, it is still an open question whether the primary effect of the supporters constitutes in giving advantage to the home team or a disadvantage for the away team, and whether this takes the form of a direct effect on the players or an indirect one through the influenced decisions of the referee (see Chapter II.1.5.). Nevill, Newell and Gale (1996) first proved the existence of home advantage in the 4-4 major English and Scottish leagues, then they found that the level of home advantage differs from division to division and these differences are significantly related to the average attendance of the given division. Their analysis on the importance of attendance was carried out focusing on two match deciding events (red cards and penalties). The authors detected that in general the home team was favoured in both aspects, but differently division by division. In divisions with high attendance, out of all players sent off during the season, only 30% were players of the home team. This is a relatively small proportion with respect to the divisions with low attendance (50%), where the home crowd seemingly was not able to create advantage in the case of red cards. The same goes for the penalties. The interpretation of these results is that the large home crowd was able to either drive the opposing player into careless behaviour (real fouls) or make the referee believe that the away team's player committed more fouls (perceived fouls). (Nevill, Newell, & Gale, 1996)

II.1.3. Travel Effects

Identically to the aforementioned crowd effects, there is no incontestable evidence for the travel effects, which presumably cause fatigue and other distractions worsening the performance of the away team. The length of the journey (distance travelled) has been analysed both in the case of domestic and abroad tours, but the conclusions are contradictory.

Clarke and Norman (1995) found a linear relationship between home advantage and the distance between the stadiums of the teams facing each other. The greater the distance, the bigger the advantage of the home team. Brown Jr. et al. (2002) came to a similar conclusion based on a decade of matches played by the qualified teams of the 1998 FIFA World Cup up until its official start. The longer the journey the away team had to take, the worse their performance. (FIFA is the common abbreviation of the International Association Football Federation from its original French name: *Fédération Internationale de Football Association*. From here on this study as well will refer to it as FIFA.) Goddard (2006) also addressed home advantage and more specifically the importance of geographical distance in his article about potential football game deciding factors. The author used the above-mentioned finding of the paper of Clarke and Norman (1995) and proved it empirically on 35 years of match data from the four major English divisions. He showed and illustrated (see Table 1) that the greater the distance, the worse the average performance of the away team. However, these results are far from conclusive. (Clarke & Norman, 1995) (Brown Jr, et al., 2002) (Goddard, 2006)

1. Table: Geographical Distance and Home Advantage 1970-2005

Distance between the pitches of the home and the away team (miles)				
Results	***<50***	***50–100***	***100–150***	***>150***
Home wins (%)	45.1	48.8	46.4	45.4
Draws (%)	28.5	26.7	27.6	27.5
Away wins (%)	21.0	24.5	25.9	27.2
Average goal per game				
Home team (%)	1.58	1.60	1.51	1.50
Away team (%)	0.97	1.06	1.08	1.10

Source: Own design based on the calculations of Goddard (2006)

Finally, Pollard, Silva and Medeiros (2008) quantified the effect of the away teams' journey on the match results using ordinal logistic regression. According to the results of their statistical analysis on 2326 games (five seasons between 2002-03 and 2006-07) in the Brazilian first division, the length of the journey has a significant, but small effect: every 1000 kilometres travelled means a 0.115 goal advantage for the home team. (Pollard, Silva, & Medeiros, 2008)

This further illustrates the inconclusiveness around travel effects. However, there is at least one consistent result among studies: home advantage decreases in the case of those local derbies where there is effectively no travel. (Pollard, 1986) For example, between 1996 and 2008 in the Turkish first division (Süper Lig) home teams acquired 61.5% of all points, but when two teams from Istanbul faced each other, this proportion is smaller (57.7%). Analogously, when two teams from distant cities faced off, the same proportion is quite high. (Seckin & Pollard, 2008)

II.1.4. Familiarity

When a team is playing at home, then they are familiar with the stadium, the environment, and the circumstances, which should mean an advantage for the home team. Researching this concept proved to be rather difficult, however the following thought-provoking findings all imply that familiarity is a believable factor contributing towards home advantage. Such advantage has been shown to be significant:

1) on artificial turf, (Barnett & Hilditch, 1993)
2) on unusually big or small football pitches, (Pollard, 1986) (Clarke & Norman, 1995)
3) when the home team was familiar with the type of the match ball. (Dosseville, 2007)

In England as of 1989 only four teams in the first four divisions played some of their home games on artificial turf. A proposal to block further construction of football pitches with artificial grass was submitted in the same year, claiming that home teams could gain further advantage. That is why Barnett and Hilditch (1993) examined the potential effect of artificial grass on home advantage. Based on data from 10 seasons of the first four divisions, they found that the advantage not only exists, but it is alarmingly high, wherefore the claim of the proposal regarding artificial surfaces was justified. (Barnett & Hilditch, 1993)

Regarding the second point it is necessary to provide the official dimensions of professional football pitches. The most common ones are 105 metres long and 68 metres wide, however these values can be freely changed in the interval set by FIFA in the official laws of the game. In the case of domestic matches, the clubs have more freedom, while in the case of international matches the rules are stricter both in terms of the length and the width. (FIFA, 2011)

2. Table: Official Dimensions of Professional Football Pitches

Most important different cases	**Width**		**Length**	
	Minimum	Maximum	Minimum	Maximum
Domestic matches	45	90	90	120
International matches	64	75	100	110
Most common	68		105	

Source: Own design based on the official laws of the game (FIFA, 2011)

While a difference of a few metres might sound negligible, in reality they have a huge influence on the tactics. For example, a smaller sized pitch is more suitable for tactics building on long balls, set plays (free kicks, corners), long throw-ins and goal kicks. Teams preferring short passes, build-up play, aiming for dominating possession play much more easily and comfortably (and thus often more successfully) on bigger sized (especially wider) pitches with more space.

Thirdly, players of the home teams can react quicker and more efficiently in match situations because they are already familiar with how the given type of match ball rolls, spins, slips, curves on the given surface (this brings us back to the first point again, as the same ball behaves differently on artificial turfs). Their previous experiences (trainings) provide them with higher probabilities to anticipate what will happen (where will the ball bounce etc.) (Dosseville, 2007)

On top of these three factors, there is some evidence for the home team gaining advantage from being familiar with the climate and/or the height above mean sea level (often shortened to sea level), but mainly only in extreme cases. (Pollard, Silva, & Medeiros, 2008) (Seckin & Pollard, 2008) The concept of sea level possibly affecting sport performance was most thoroughly studied by McSharry (2007). He found that sea level negatively affects performance physiology, which is illustrated in South America by the general underperformance of teams from low sea level cities playing away against teams from much higher sea level cities. The bigger the difference between the sea levels, the higher the goals scored (and the lower the goals conceded) by the teams with higher elevation. Every 1000 metres (sea level difference) increase the goal difference by 0.5 goals. These findings are obviously reflected in the home win ratios as well. In South America the home win ratio was 0.537 if the facing teams were from cities with the same average elevation. This value explodes to 0.825 in the extreme case when the home team's city lies almost 3700 metres above the city where the away team's is from (for example Bolivia versus Brazil) and vice versa. In summary, sea level provides a significant advantage for the teams from cities with higher elevation both in games played at lower and at higher sea level

(but only in international competitions). This means that teams from areas with lower sea level cannot acclimatize to the high sea level, decreasing the performance physiology. One very intriguing implication of this fact (beyond the direct effect on home advantage) is that for these international games with huge sea level difference, players will ultimately be selected into the squad not based on their general skills etc., but predominantly based on their resistance to altitude sickness. (McSharry, 2007)

A final note to familiarity. If we try to compare the effects discussed up until now, we can state that in general crowd and travel effects contribute less towards home advantage than the familiarity factor (which is more difficult to quantify). (Pollard, 1986) It was exactly this familiar environment and circumstances that ceased to exist because of the long break in football due to the Second World War, which caused the further decrease of home advantage in England and Italy in the years right after WW2. (Morris, 1981) (Pollard & Pollard, 2005) The same reasoning can be made and is the most likely explanation, when home advantage for a given team becomes smaller (lower home win ratio, lower goals per game ratio etc.) after they move to a new stadium.

II.1.5. Referee Bias

In contrast to the previous subchapters, there is no uncertainty here. It has been undeniably proven that the referees' decisions are unquestionably in favour of the home team. This has been shown over and over again in the past 50 years, first when analysing referee decisions, namely the frequency of penalty cards (yellow cards – cautions and red cards – dismissals) and penalty kicks awarded. (Lefebvre & Passer, 1974) (Nevill, Newell, & Gale, 1996) (Garicano, Palacios-Huerta, & Prendergast, 2005) (Thomas, Reeves, & Smith, 2006)

Lefebvre and Passer (1974) have shown on data (240 matches) from the 1973-74 season of the Belgian first division that the away team commit more fouls, more aggressive challenges, receive more penalty cards and more penalty kicks are awarded against them. However, they did not yet make the connection between home advantage and referee bias, or at least this notion does not appear explicitly in their paper. (Lefebvre & Passer, 1974) Thomas, Reeves and Smith (2006) arrived at similar conclusions on a larger sample from English football and they clearly stated and showed the effect on home advantage as well. (Thomas, Reeves, & Smith, 2006)

Maybe the most important and influential study in this topic is the one by Garicano, Palacios-Huerta and Prendergast (2005) who were studying how non-financial incentives affect behaviour. More specifically, they focused on bias and corruption caused by social pressure.

They present empirical evidence on professional football referees favouring the home team in order to satisfy the crowd in the stadium. The question of added time (additional minutes after the regular 90 minutes to compensate the time lost during the game due to extraordinary breaks) is the exclusive discretionary and decisional authority of the referee. The authors show that the referees systematically favour the home team through awarding shorter additional time (also called stoppage time) in the case of tight matches when the home team was winning and significantly longer additional time when the home team was losing. In the case of those matches where there was no question about the outcome when nearing full time (i.e. one team was winning by at least 2 or 3 goals) no such bias is detectable, the referees behave objectively. A further result is that the referees – consciously or subconsciously – adjust the level of their bias to the importance of the game. Meaning that referee bias in terms of added time is smaller in a second division game in the middle of the season between two mid-table teams than for example in a last round matchup between two title contenders. (Garicano, Palacios-Huerta, & Prendergast, 2005) Not surprisingly, similar bias was found in the Bundesliga. (Dohmen, 2005)

In the meantime, referee bias was proved under "laboratory" conditions as well. (Nevill, Balmer, & Williams, 2002) Then, further empirical studies were conducted, focusing again on referee decisions (yellow and red cards, penalties, added time) but now carefully controlling for the confounding variables (the problem of confounding appears when an omitted important and significant variable exerts its effect in the given model through an insignificant variable). (Sutter & Kocher, 2004) (Dohmen, 2005) (Boyko, Boyko, & Boyko, 2007) (Dawson, Dobson, Goddard, & Wilson, 2007) (Buraimo, Forrest, & Simmons, 2010)

Nevill, Balmer and Williams (2002) analysed the influence of the presence or the lack of crowd noise on referee decisions in the case of different tackles and challenges. The method of the study was to show the participating professional referees different match situations on video and ask for a decision after each one. One group was watching with crowd noise in the background the other was sitting in a completely silent room. The presence of crowd noise had crucial influence on the referees' decisions. Those who were watching with crowd noise were significantly more hesitant to call a foul and in fact they called significantly less (15.5%) fouls against the home team than those, who were watching the same situations in silence. Thus, crowd support influences referees to make decisions in favour of the home team. The main reason behind this dynamic is that referees try to avoid triggering boos, outcry, resentment and displeasing the home crowd in general with their decisions. (Nevill, Balmer, & Williams, 2002)

Dohmen (2005) showed that the bias exists also in terms of penalties awarded, but more importantly he showed that the composition of the crowd affects the scale and the direction of the bias. The intensity of social pressure – measured by the vicinity of the supporters to the football pitch – determines how strong of an influence is the crowd able to exert on the referee's decisions. (Dohmen, 2005)

Buriamo, Forrest and Simmons (2010) also showed that home teams clearly receive less yellow and red cards than away teams in the Premier League and in the Bundesliga. However, their results are more robust than those of previous studies, as they added match events that can change a team's behaviour and aggressivity level to their model. Therefore, they were able to control for the fact that at any given point in the match, the team that is losing is more likely to commit fouls and aggressive challenges, thus the probability of receiving a penalty card is higher independently of playing home or away. However, even for controlling this changing level of aggressiveness a significant bias was detected in favour of the home team in terms of penalty cards. Furthermore, the authors showed that in Germany those home teams, who have running track around the pitch (which causes the fans to be farther away) receive more yellow and red cards on average, than those home teams whose supporters are closer to the pitch. (Buraimo, Forrest, & Simmons, 2010)

To summarize, we can state that referee bias is evident. The most likely reason for that is the presence of supporters, more precisely the size of the crowd, the intensity of the pressure they can apply and their vicinity to the pitch, but a clear answer does not (yet) exist in this regard. There might be other, not yet listed factors in play as well that create the referee bias. For example similarities in the nationality or the cultural identity of the referees and the players can result in special treatment. (Messner & Schmid, 2007) On the one hand, the presence of the referee bias is clear, it has not yet been unquestionably shown whether this bias is primarily in favour of the home team or against the away team or both at the same time. While everybody expects objective and unbiased decisions from the referees, if we look closely at all the factors having to be taken into consideration during the decision-making process, entirely impartial decisions are highly unlikely if not impossible. What is apparent, that at the end of the day this referee bias (regardless of its forms or reasons) results in an advantage for the home team.

II.1.6. Territoriality and Special Tactics

It has been long known in the field of biology that a real or perceived invasion in one's home or territory triggers a response both in humans and animals. It seems to be reasonable that this phenomenon – called territoriality – also plays a role in forming the home teams' advantage in football. Morris (1981) was the first to present this idea. Empirical support arrived somewhat later, in the form of studies showing that a rise can be detected in the testosterone level of the home team's players before the game. Neave and Wolfson (2003) were the first to focus on possible hormonal explanations and found that among football players the salivary testosterone level is significantly higher before home games than before away games. The same goes for games against "arch-rivals" and against "regular opponents". So, the authors showed that the relationship between human competitions, rivalry, testosterone, territoriality, and dominance not only exists, but it is also strong and significant. More importantly for us, their results imply that territoriality is a crucial factor in home advantage across sports. (Neave & Wolfson, 2003)

Later, this has been illustrated by showing that home advantage is higher for teams whose home stadium is found in countries, cities, or remote regions with historical or current conflicts. (Pollard & Pollard, 2005) This might be caused by higher or more intense territoriality, implying that geographical location can also trigger territoriality, which in turn creates a greater advantage for the home team. Pollard (2006) showed that the level of home advantage is extremely different across Europe. The Balkan countries (especially the Bosnian and Albanian football leagues) have way higher than average home advantage, while North Europe (the Baltic states, Scandinavia, and the British Islands) is the total opposite. South America has a similarly high variance while the other continents are more stable, with no significant geographical differences. Focusing on Europe, with the help of a multiple regression model, the author found that geographical location, crowd, and travel effects explain 76.7% of the variance of home advantage across all countries(!). The observed huge difference between the football leagues of European countries can then be explained by the different levels of territoriality. (Pollard, 2006)

Home and away teams can prepare for and approach the given game differently also in terms of tactics and strategies. If the away team is aiming to play safe with a more defensive mindset, then this can provide a territorial and psychological advantage for the home team. (Pollard, 1986) (Page & Page, 2007) However, there is still no firm evidence that would directly connect the applied tactics with home advantage in football. What has been achieved is the clear demonstration of the difference in the match performance indicators between the home and

away team. This indirectly expresses the significant role of tactics and strategies. (Tucker, Mellalieu, James, & Taylor, 2005) (Seckin & Pollard, 2008)

Tucker et al. (2005) analysed the technical and tactical behaviour of teams in function of the location of the match (howe or away game). In home games the observed English professional football team made more successful passes, tackles, and won more aerial duels than in away games. In home games, in the final (attacking) third, they made more crosses, took more corner kicks, completed more successful passes and dribbles, attempted more aerial duels, and had more shots on target than in away games. In away games, in the defensive third, they took more goal kicks, had more ball recoveries, made more clearances, and won more aerial duels than in home games. These statistics all suggest that the location of the game has an influence on the strategies of the teams. Thus, it might be possible that the existence of home advantage influences the teams' choice of tactics and vice versa, the chosen and applied tactics also help to create the phenomenon of home advantage. But again, in this matter compelling empirical evidence is yet to be found. (Tucker, Mellalieu, James, & Taylor, 2005)

II.1.7. Rule Changes and Psychological Effects

Even though football is a simple game, rules and regulations often undergo important changes, which might also affect home advantage to some extent. To list some notable examples that have been mentioned as potentially influencing the changes in home advantage: raising the number of substitutions allowed during the game, extending the break between the two halves, introducing the back-pass rule and more severe punishment for dangerous (slide) tackles. However, as of today there is no clear result about how big of a role these changes played in the decrease of home advantage over time, if any. One topic with somewhat more consensus is the transition from the 2-points-for-a-win to the 3-points-for-a-win system. The new 3-1-0 (win-draw-loss) points system was first implemented in England in 1981, to incentivise attacking football through assigning bigger weights to wins thus decreasing the number of boring draws. The other leagues followed the English example inside 10-15 years. (Moschini, 2010) Maybe it is not a surprise, that researchers mostly agree that this transition did not influence home advantage at all. (Dowie, 1982) (Pollard, 1986) (Pollard & Pollard, 2005) The few opposing studies claim that it did indeed decrease home advantage as the new system provides more incentives for the away team to perform better, implement more offensive tactics, and play for the win instead of accepting the draw. (Jacklin, 2005) The last noteworthy rule change is the Bosman ruling from 1995, which made it easier for players to transfer from club to club also internationally. This facilitated the process which resulted in the situation we see these days,

where most teams in the strongest football leagues are filled with foreign players. Teams do not anymore consist of native-born, locally trained, homegrown talents like half a century ago. In contrast with the previous changes, this one most probably did indeed have an effect on home advantage, as the (social) bond between the players and their hometown and home fans has been weakened, further decreasing home advantage.

One cannot fail to mention the psychological effects. Players and coaches are obviously aware of the existence of home advantage, consequently, their (mental) attitude before and during the game is surely affected as well. A likely and interesting possibility is that while there exist actual reasons for the home advantage, their effects are magnified by the attitude, believes and perception of the players and the surrounding staff (coach, etc.), thus creating a self-preserving phenomenon: even if the actual factors causing home advantage would cease to exist, the players might still believe, that the home team has higher winning chances regardless. This would affect their attitude and behaviour, sustaining home advantage even without physical reasons behind it. (Pollard, 1986) (Pollard & Pollard, 2005)

The importance of mental (attitude), perception, psychological and physiological aspects was examined in detail by Neave and Wolfson (2004). According to them, being aware of the already presented factors (crowd, referees, travel etc.) can already affect the players' psychological and physiological responses even before these factors actually exercise their effects. Thus, on average, players are already in a mental disadvantage before an away game. (Neave & Wolfson, 2004) The connection between the psychological state of the players and home advantage has also been shown by Waters and Lovelle (2002), who analysed the players' perception and self-evaluations reflecting back on past games. The players recalled experiencing significantly higher individual and team confidence, and more positive pregame attitude in the case of home matches. They believed this was due to better physical and mental preparation, sleep, crowd effects, supporters, and referee bias. (Waters & Lovell, 2002)

A recent study conducted a similar quantitative experiment for the psychological state of coaches and/or managers, focusing on their pregame expectations, objectives, and tactical decisions in the function of the location of the match. The participating almost 300 coaches (with different expertise) were handed detailed information on an upcoming fictional game and were asked to make the tactical decisions. They were randomly divided into two groups, which differed only in the location of the game in question. Independently of their expertise, the coaches assigned to the home team had on average higher win expectations, set out more

offensive objectives and applied more braver tactics than those assigned to the away team. (Staufenbiel, Lobinger, & Strauss, 2015)

II.1.8. COVID-19

The pandemic caused all sports matches to be played behind closed doors. For football this meant the second half of the 2019-20 season and almost the entirety of the 2020-21 season. Despite the numerous negative consequences, this provided the unique opportunity to analyse professional football matches with no attendance, which was not possible before and (hopefully) will not be possible in the future either on this scale. In terms of home advantage crowd effects and referee bias were re-examined in detail. Wunderlich et. al. (2021) claim that supporters are not necessary for home advantage, which contradicts the majority of the findings presented in Chapter II.1.2. They found that during the lockdown period the home win ratio has not dropped significantly when compared to the previous 10 years (where the matches were played with supporters in attendance). Furthermore, they showed that without crowd support the home team dominates the game on a significantly lower level (measured in the number of shot attempts and shots on target). Plus, based on empirical data (number of fouls, yellow and red cards), they confirmed the theory presented in Chapter II.1.5., i.e. that the presence of the crowd forces the referee to be biased (often subconsciously) in favour of the home team and against the away team. Analysing the same indicators, they showed that the referee bias disappeared during the lockdown(!). In summary, the study contradicts the previous papers and the claim that crowd would be the main driver behind home advantage. This is supported by the interesting and important case of amateur football games, which are obviously played in front of only a small number of fans, and where home advantage is still considerably high. Therefore, the authors claim that home advantage is first and foremost influenced by factors not connected (neither directly nor indirectly) to the large number of supporters. Hence, territoriality (Chapter II.1.6.) and familiarity (Chapter II.1.4.) could play a much more important role than crowd effects. (Wunderlich, Weigelt, Rein, & Memmert, 2021)

However, McCarrick et. al. (2021) after conducting their own calculations, criticized the methodology of Wunderlich's study, and thus questioned the correctness of its findings, because it compares the lockdown period with a too long interval (all matches since 2010) instead of only using the 2019-20 season, which is perfect for such studies, as the first part of the season was played in the presence of supporters, while the second part behind closed doors. Analysing fifteen European leagues, McCarrick and his colleagues found that the goals scored and points won by the home team decreased significantly during the lockdowns, demonstrating the home

teams' worsened performance. They acknowledged and confirmed the results of Wunderlich and his colleagues regarding the disappearance of referee bias and the lower level of game dominance by the home team. They conclude that because these results show the significant drop in home advantage in the case of matches played behind closed doors, crowd effects do indeed contribute significantly toward home advantage (in accordance with previous findings presented in Chapter II.1.2. and II.1.5.). (McCarrick, Bilalic, Neave, & Wolfson, 2021)

II.1.9. Summary

One of the numerous difficulties with studying the potential factors contributing toward home advantage is that they most probably work contemporaneously, affecting each other as well. It is extremely complicated to research, find, isolate, and quantify these interactions. To close the literature review part dedicated to the detailed theoretical examination of home advantage, I would like to quote the most cited researcher in the field of home advantage, Richard Pollard, whose fundamental papers from 1986 and 2008 are both ending with the following, still relevant sentence: "Clearly, there is still much to be learnt about the complex mechanisms that cause home advantage, both in soccer and other sports. The topic remains a fruitful area of research for sports historians, sociologists, psychologists and statisticians alike." (Pollard, 2008, pp. 248)

II.2. Betting Odds

The gambling industry is a whole separate world, which has been studied by a variety of experts, let them be mathematicians, economists, policy makers or even historians. As stated in the introduction, for the topic (Chapter I.2.) of this thesis, the most important aspect is the forecasting ability, i.e. the predictive accuracy of the simple pre-match betting odds. Keeping that in mind, I am focusing on two main papers in this field.

Spann and Skiera (2009) compared the predictive accuracy of different methods, highlighting tipsters and betting odds. The incentive behind more accurate predictions is naturally the potential ability to earn consistent profits in the betting market. The authors conducted an empirical study using data from around 800 Bundesliga games (spanning three seasons). They found that betting odds have quite strong forecasting accuracy, represented by a hit rate of 52.93%, which is significantly outperforming the tipsters' different guessing strategies. An interesting implication is, that when the forecasts of these different methods can be combined, then it results in a substantially higher forecast accuracy, but still, none of the forecasts (not even their combinations) lead to systematic profits in betting markets. The reason for this are the taxes and fees, which were especially high in Germany at that time. (Spann & Skiera, 2009)

Wunderlich and Memmert (2018) also showed that betting odds often outperform mathematical models when the forecasting is sports related. They point out that an important problem is that the variables making up the betting odds are obviously not entirely accessible, thus in contrast to other predicting models (for example based on ratings or rankings) no clear measure of team-specific quality can be concluded from the betting odds, i.e. reverse engineering is basically impossible. The authors also investigated the idea of Spann and Skiera (2009), the approach of combining the mathematical methods and the information included in betting odds. They developed their own forecasting model based on the ELO rating system (known and used in chess, for example) and using betting odds as a source of information. Data from almost 15.000 top league matches (between 2007-08 and 2016-17) were used. The authors' own new combined model is a betting odds based ELO model, which clearly outperformed classic ELO models on the same data, thus demonstrating that pre-match betting odds contain more relevant information than the result of the match itself! (Wunderlich & Memmert, 2018)

III. Methodology

For each research question, for each goal of the study different methods and approaches are needed. This chapter contains the detailed theoretical presentation of these analytical tools but does not go into detail of their actual application on the data. This will be the role of the most interesting part, i.e. the chapters containing the empirical analyses (see Chapter V.).

To find the answers to the questions raised, I collected data. The through process of data collection, data management and data manipulation is presented in the next chapter (see Chapter IV.). A wide range of statistical and analytical tools, tests, models etc. will be used in order to visualize, understand and analyse what the empirical data suggests. All my analytical work is done by using the programming language called R. In the spirit of reproducibility of academic works, I also attach – alongside the thesis itself – the whole code used from the very first step to the very last line of code (see Appendix). Detailed comments are added in between the lines of the code; thus, it is quite easy to understand even for readers who might not be as familiar with the programming language R or statistical analysis in general.

Regarding the changes in home advantage over time and across different leagues, the most important tools will be the generation of common indicators (for example goals per game ratio for the away team, etc.), data visualization, interpretation of the plots and graphs (which always has to include detailed comments and reasoning with respect to specific trends identified), and

finally comparative analysis both in a quantitative and a qualitative sense. To be more specific, I present the two most important indicators that will be used in Chapter V.1. These are the:

$$Home\ win\ ratio = \frac{Number\ of\ home\ wins}{Numbe\ of games\ played}\ (Note: by\ season)$$

$$Home\ goals\ per\ game\ ratio = \frac{All\ goals\ scored\ by\ the\ home\ team}{Number\ of games\ played}\ (Note: by\ season)$$

Obviously, the draw ratio, the away win ratio and the away goals per game ratio are calculated analogously. When plotting them over time, there are many methods to fit a trend on the data points. To find and fit a smooth trend I will use the generalized additive model (GAM) with integrated smoothness estimation. Without going into too much detail (this being an empirical not a theoretical work), the generalized additive model is a generalized linear model (GLM) in which the linear predictor is given by a user specified sum of smooth functions of the covariates plus a conventional parametric component of the linear predictor. (Wood, 2011)

For the analysis of the post-match statistics' explanatory information on the match result, the situation is totally clear. During the analysis part – keeping up with the modern academic approach in papers using data analysis, data science or any statistics related models – data visualization, easy-to-read, but still statistically relevant and correctly informative figures will play an important role. Descriptive statistics (assisted by boxplots to identify and illustrate skewed distributions etc.), correlation analysis (again, with insightful graphs) and tests (Pearson's chi-square test, Kruskal-Wallis test) will be carried out to get a first understanding of our empirical data. These will already give some implications about the potential significance and effect of the given variable. The most important part of the statistical analysis comes after that. I will thoroughly build (using stepwise model selection) multiple multinomial logistic regression models to identify which are the relevant factors and how do they contribute to the full time result. The goal is to achieve the best possible model. This means trying to maximize the predictive accuracy and the number of statistically significant explanatory variables, while carefully keeping an eye on the model complexity – model fit trade off. Here the Akaike and the Bayesian (or Schwarz) information criteria will be of assistance. The results, again, will be supported by summarizing plots (for example predicted probabilities as a function the given explanatory variable), representing the different dynamics. Let us go into more detail!

By now it seems evident, but still, it is important to highlight that the dependent variable is the full time result of the match, which is an unordered categorical variable with three values (H –

home win, D – draw, A – away win). This already anticipates the necessity of the multinomial logistic regression. But first of all, the separate testing of the dependent variable. In the case of the numeric independent variables the relationship, consequently, is between a categorical and a numeric variable. This can be tested by the Kruskal-Wallis rank sum test. (Hollander & Wolfe, 1973) The hypotheses are:

$$H_0^{Kruskal}: population\ medians\ are\ equal. The\ two\ variables\ are\ independent.$$

$$H_1^{Kruskal}: population\ medians\ are\ not\ equal. The\ two\ variables\ are\ not\ independent.$$

In the case of categorical independent variables, the to be tested relationship will be between two categorical variables. This is called association and can be tested for example by the Pearson's chi-squared test. (Pearson, 1900) The test statistic and the hypotheses are:

$$\chi^2_{Pearson} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(Observed\ frequency_{i,j} - Expected\ frequency_{i,j}\right)^2}{Expected\ frequency_{i,j}}$$

$$degrees\ of\ freedom\ (df) = (rows - 1) \cdot (coloumns - 1)$$

$$H_0^{Pearson}: The\ two\ variables\ in\ question\ are\ independent, not\ related.$$

$$H_1^{Pearson}: The\ two\ variables\ in\ question\ are\ related, not\ independent.$$

But what is exactly a multinomial logistic regression and why is it needed in this case? When the dependent variable is categorical with three or more categories, a multinomial logistic regression model has to be used. It has to be noted that if the dependent variable had only two categories the more known binary logistic regression would be perfect. They both belong to the family of generalized linear models (GLM) with discrete dependent variable, with the binary logistic regression being the special case. In the multinomial case there are two main types of the dependent variable: unordered (e.g. voting preferences) and ordered (exam grades, or anything measured on a Likert Scale). In my case the match result can take three discrete values with no clear order, thus I will consider the case of unordered multinomial logistic regression.

The multinomial model can be described briefly in the following way. Let us consider a dependent variable Y_i for the $i-th$ unit with k categories and probabilities $p_{i,1}, \dots, p_{i,k}$. Now a reference category has to be fixed, extending this way the binary model to the multinomial one. Usually, the last category k is chosen as the reference category. This results in:

$$\log\left(\frac{p_{i,1}}{p_{i,k}}\right) = \beta_{0,1} + \beta_{1,1}x_{1,i} + \beta_{2,1}x_{2,i} + \cdots$$

$$\log\left(\frac{p_{i,2}}{p_{i,k}}\right) = \beta_{0,2} + \beta_{1,2}x_{1,i} + \beta_{2,2}x_{2,i} + \cdots$$

$$\log\left(\frac{p_{i,k-1}}{p_{i,k}}\right) = \beta_{0,k-1} + \beta_{1,k-1}x_{1,i} + \beta_{2,k-1}x_{2,i} + \cdots$$

Just as in the binary logistic model, the coefficients can be interpreted with the help of the log-odds-ratios. For example, $\beta_{1,1}$ is the change in the log-odds of category 1 as opposed to category k for a 1 unit increase in the independent variable x_1. The estimated probabilities for the $i-th$ individual can also be retrieved for each category. (Evans & Rosenthal, 2004)

$$category\ 1 => \hat{p}_{i,1} = \frac{e^{\beta_{0,1}+\beta_{1,1}x_{1,i}+\beta_{2,1}x_{2,i}+\cdots}}{1 + (e^{\beta_{0,1}+\beta_{1,1}x_{1,i}+\beta_{2,1}x_{2,i}} + \cdots + e^{\beta_{0,k-1}+\beta_{1,k-1}x_{1,i}+\beta_{2,k-1}x_{2,i}+\cdots})}$$

$$category\ k-1\ => \hat{p}_{i,k-1} = \frac{e^{\beta_{0,k-1}+\beta_{1,k-1}x_{1,i}+\beta_{2,k-1}x_{2,i}+\cdots}}{1 + (e^{\beta_{0,1}+\beta_{1,1}x_{1,i}+\beta_{2,1}x_{2,i}} + \cdots + e^{\beta_{0,k-1}+\beta_{1,k-1}x_{1,i}+\beta_{2,k-1}x_{2,i}+\cdots})}$$

$$category\ k => \hat{p}_{i,k} = 1 - (\hat{p}_{i,1} + \cdots + \hat{p}_{i,k-1})$$

Then, the significance of the parameters can be easily checked by the Wald-tests or z-tests. Important data visualization techniques belonging to the results are the graphs of log-odds-ratios in function of the given independent variable, and the graphs of the predicted probabilities in function of the given independent variable. These have to be prepared individually. Regarding the problem of choosing between different multinomial logistic regression models, standard (classical) model selection techniques can be used. Starting with the full model is generally more reliable, thus I will use the backward stepwise algorithm, based on the Akaike and the Bayesian (also called Schwarz) information criteria (AIC and BIC). Both of them are measuring the goodness of the model, using penalty functions to favour smaller ones.

$$AIC = -2\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) + 2(p+1)$$

$$BIC = -2\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) + (p+1)\cdot \log n$$

Finally, the most useful indicator for me and for the second and the third research question will be the accuracy (ACC) of the different models. This can be calculated from the confusion matrix and measures the proportion of correct classifications (i.e. correctly predicted match result).

Concerning the third research question, the situation is much simpler. The reciprocal of the betting odds (in the European, i.e. decimal format) result in the probabilities assigned to the match outcomes by the bookmaker. Then, there exist different methods, theories and formulas to assign from these probabilities a single prediction to the given match. With small modifications I created my own rule, largely based on Sumpter's (2016) classification:

$$if\ \frac{1}{odds_H} > \frac{1}{odds_D}\ \&\ \frac{1}{odds_H} > \frac{1}{odds_A} \Leftrightarrow p_H > p_D\ \&\ p_H > p_A => Predicted\ outcome: H$$

$$if\ \frac{1}{odds_A} > \frac{1}{odds_D}\ \&\ \frac{1}{odds_A} > \frac{1}{odds_H} \Leftrightarrow p_A > p_D\ \&\ p_A > p_H => Predicted\ outcome: A$$

$$if\ \frac{1}{odds_A} = \frac{1}{odds_H} \Leftrightarrow\ p_H = p_A => Predicted\ outcome: D$$

After these steps the predictive accuracy can be easily calculated from the confusion matrix. The comparison of the different predictive accuracies has to be accompanied by a detailed qualitative reasoning and the conclusions have to be formulated taking into account both the previous findings from the literature review and the meanwhile achieved results of this thesis. Without further ado, it is time to look at and dive into the data.

IV. Data

IV.1. Source of the Data

This study has two main data sources. One is the "engsoccerdata" database, collection of data sets and R package complied by James P. Curley. He shared it for free public use on different websites and it can also be downloaded from the R archive. (Curley, 2016) The package contains historical match data from almost every European professional football league and was assembled from less structured sources. This will be the primary basis for the home advantage part of the work. The other one is the famous football-data.co.uk website, which offers easily downloadable csv files with match results, statistics, and betting odds. These files are mostly complete for the past 20 years. Thus, this source will play its role for the match prediction part.

IV.2. Presenting the Database

Therefore, for analysing the changes in home advantage over time, the following data sets will come in handy (complemented with the most recent seasons from the other source):

1) "england": containing every match result in the first four divisions in England from the 1888-89 season up to the 2015-16 season. Due to WWI and WWII, 11 seasons (from

1915-16 to 1917-18 and from 1939-40 to 1945-46) were played only in part or not at all. Because these seasons are incomplete or empty, they do not appear in the data base.

2) "germany": containing every match result from the German top flight (Bundesliga) from the 1963-1964 season up to the 2015-16 season. There are no missing seasons.
3) "italy": containing every match result from the first divisions in Italy (Serie A) from the 1929-30 season up to the 2015-16 season. Due to WWII, 2 seasons (1943-44 and 1944-45) were played only in part or not at all. Therefore, they do not appear in the data base.
4) "spain": containing every match result from the first division in Spain from the 1928-29 season up to the 2015-16 season. Due to the Spanish Civil War, 3 seasons (from 1936-37 to 1938-39) were played only in part or not at all. They do not appear in the data base. During WWII Spain was a neutral country, thus there were no interruptions.

Regarding the analysis of post-match statistics and pre-match betting odds more detailed data sets were needed. For the above listed leagues these are available "only" for the past 20 years, but in turn they contain much more information for the single matches. These are:

- Div = League Division
- Date = Match Date (dd/mm/yy)
- HomeTeam = Home Team
- AwayTeam = Away Team
- FTHG and HG = Full Time Home Team Goals
- FTAG and AG = Full Time Away Team Goals
- FTR and Res = Full Time Result (H = Home Win, D = Draw, A = Away Win)
- HTHG = Half Time Home Team Goals
- HTAG = Half Time Away Team Goals
- HTR = Half Time Result (H = Home Win, D = Draw, A = Away Win)
- HS = Home Team Shots
- AS = Away Team Shots
- HST = Home Team Shots on Target
- AST = Away Team Shots on Target
- HC = Home Team Corners
- AC = Away Team Corners
- HF = Home Team Fouls Committed
- AF = Away Team Fouls Committed
- HY = Home Team Yellow Cards

- AY = Away Team Yellow Cards
- HR = Home Team Red Cards
- AR = Away Team Red Cards
- B365H = Bet365 home win odds
- B365D = Bet365 draw odds
- B365A = Bet365 away win odds

Now, almost everything is ready and available for me to start the detailed analysis.

IV.3. Data Manipulation

As it usually happens, data cleaning, data management and data manipulation takes a significant amount of time and effort. This thesis was no exception. I carefully had to check for NAs (missing data), filter for them and then decide what to do with them. In some cases, there were blank lines added by mistake, these had to be removed obviously. Another situation was that for some Italian matches the result was awarded by the federation (because of some particular incident). These artificial 3-0 or 0-3 results would unnaturally influence the analysis, so I removed these matches completely, as the loss of data is still only marginal. Some new variables had to be created (e.g. the calculation of the home win ratio), and then the different datasets (pre-2015-16 with 2015-16 to 2022-23) had to be merged to create a single working data frame for each league. For the different plots and graphs sometimes special formatting was needed, which called for the creation of temporary variables and even data sets, which were of course removed after they played their role to not influence further analysis. Anyhow, the code contains every step from the original data in a clear and consistent way and is perfectly reproducible.

V. Analysis

The whole analysis was conducted on the data presented above (Chapter IV.) and keeping in mind every important aspect and approach mentioned above during the presentation of the topic, the goals, the exact research questions, and the detailed methodology of this study (Chapter I. and Chapter III. respectively).

V.1. Changes in Home Advantage over Time

In the literature review and during the detailed examination of home advantage it became clear that a lot of studies concluded that in general the effect of home advantage on the match result has decreased over time and these dynamics are significantly different geographically, i.e. between the different football leagues, divisions, championships both globally and in Europe. The potential theoretical reasons have also been presented and discussed. However, it is worth

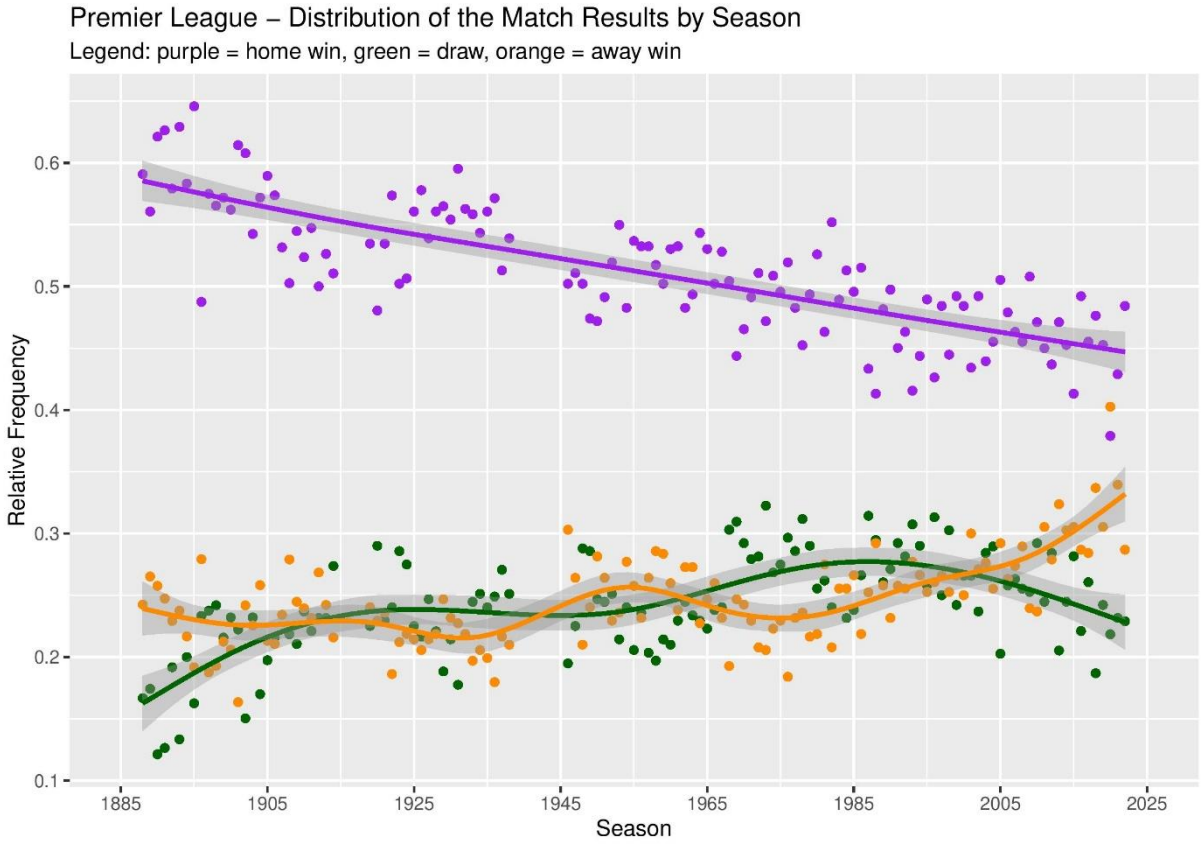
to examine and study empirically, how these changes came into play; what similarities and differences can be observed between the leagues and between the trends. Some parts of the subchapters below have been inspired by the disquisitions of Long (2019) and Kovács (2023).

V.1.1. Distribution and Trends of Match Results

First, I looked at the ratios of the home and away wins by season, focusing on how it changed over time in the classical top four domestic leagues of European club football. It has to be noted again that these (English, German, Italian and Spanish first divisions) are the four best domestic leagues in the world. The existence of home advantage is effectively illustrated through these ratios. Furthermore, they can provide information regarding the changes of football itself over time and across countries.

Let us begin with the English Premier League, which is highest level of the English football pyramid, and maybe the most competitive domestic league in the world. The first official season is the 1888-89 one. This makes it also the eldest football league of the world. The figure below (Figure 1) shows the ratios of the matches won by the home team, draws and the matches won by the away team for each and every season from 1888-89 up to 2022-23.

1. Figure: Distribution of the Match Results by Season in the Premier League

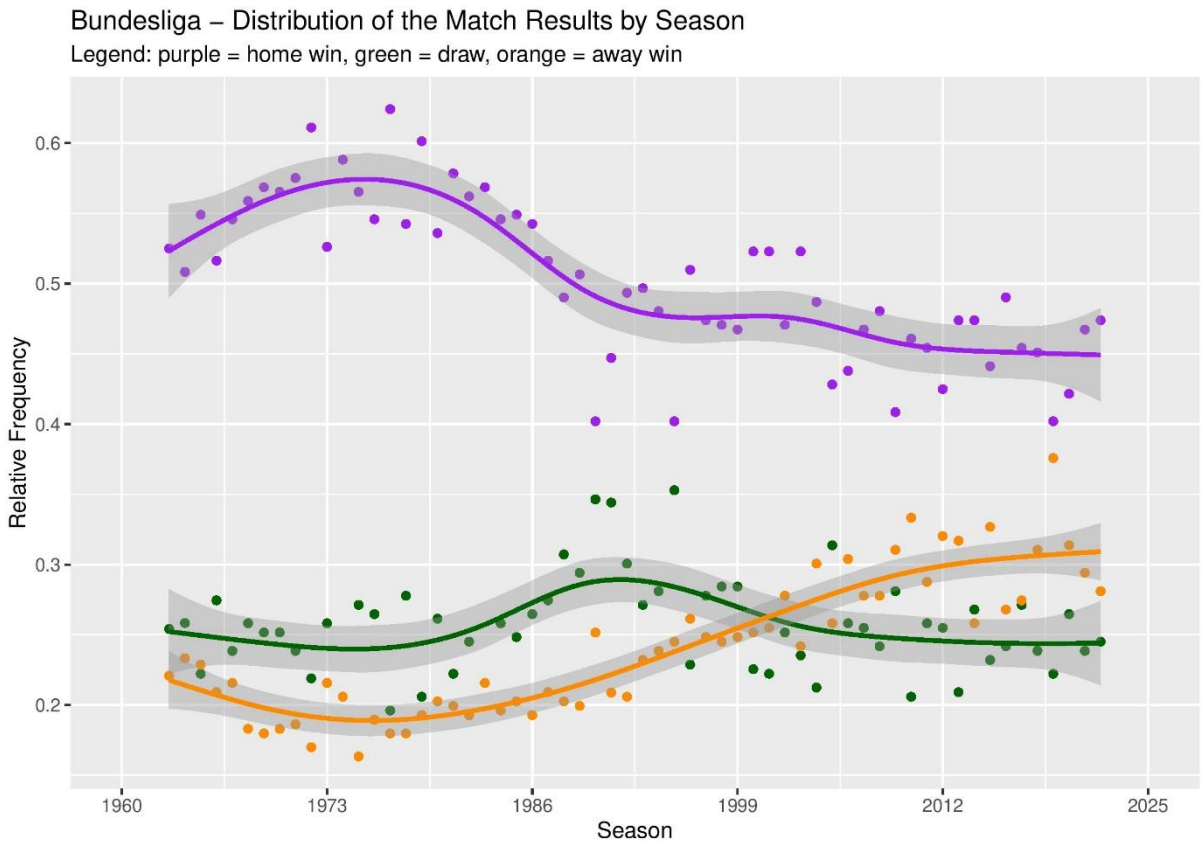


Source: Own calculations and own design based on the data presented in Chapter IV.

In the past 135 years (or 134 seasons; see Chapter IV.2 – seasons cancelled due to the World Wars) we can observe a declining trend in the home win ratio in the English Premier League. In the early stages around 59% of the games were won by the home team. The draw ratio and the away win ratio was roughly the same: the away team won approximately 24% of the games, while the remaining 17% of the games ended in a draw. These days the home team only wins ca. 45% of the matches, the away team is up to ca. 32% and the draw ratio also increased to ca. 23%. Consequently, Figure 1 indicates a clear, consistently decreasing trend in home advantage, at least as far as the English Premier League is concerned. (If we assume this rate of decrease and calculate with it, we get that the expected home win ratio will be 0% by 2400, which is obviously impossible but still shows how quickly and how much football – and more specifically the presence of home advantage – has changed in the past 130 years in England.)

I looked at the same ratios – plotting them over time – in the case of the other leagues as well. Figure 2 shows the ratios of the matches won by the home team, draws and the matches won by the away team from 1963-64 until 2022-23 in the top division of Germany.

2. Figure: Distribution of the Match Results by Season in the Bundesliga

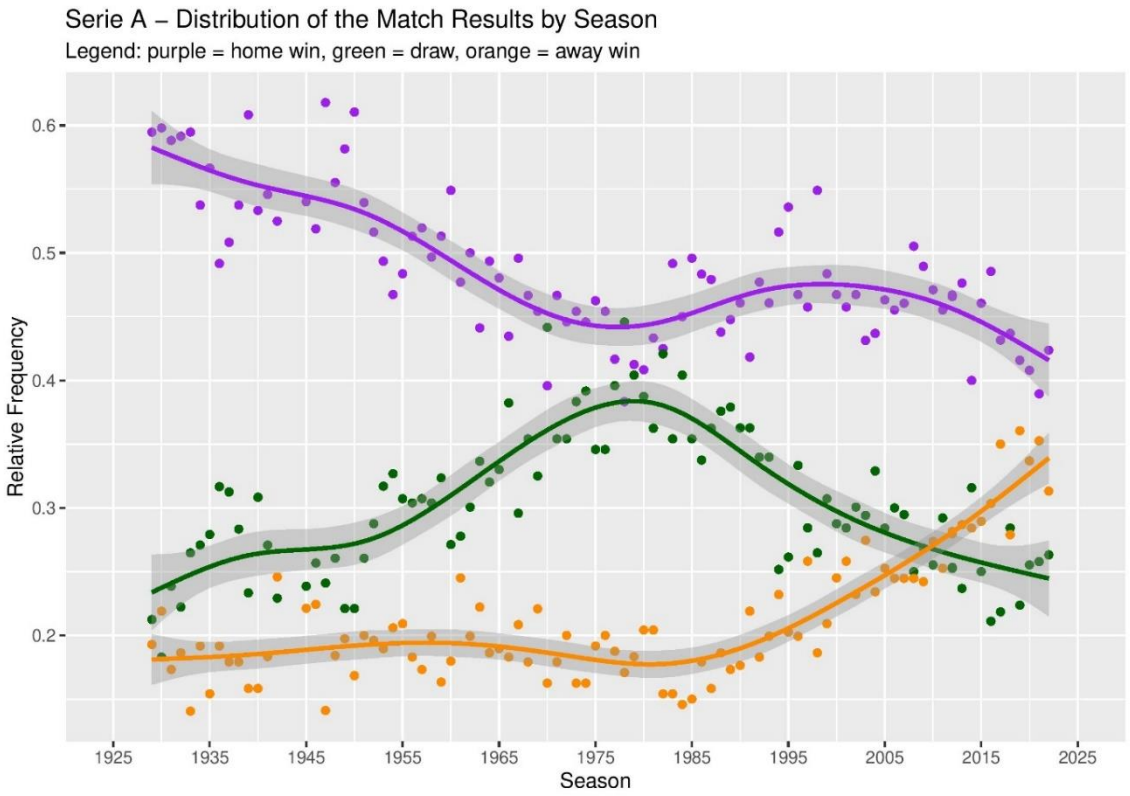


Source: Own calculations and own design based on the data presented in Chapter IV.

Here, the decrease in the home win ratio is less obvious. However, this is also due to the fact (among others) that – in contrast with England's first division – here the first data are from the 1963-64 season, i.e., from almost 80 years later. The growth of the away win ratio on the other hand is much more eye-catching. In the 60s in Germany the home team won 50-55% of the matches, 25% of the games were drawn and the away team found a way to win only in 20% of the cases. Fast forward to 2022-23 and the away team won almost 30% of the games, the draw ratio is roughly the same and the home win rate is a bit higher than 45%. What does this tell us? Well, that home advantage is still significantly present in the Bundesliga. This can be explained by the fact, that the German top division consistently has by far the highest average attendance (from 2008-09 until COVID-19 restrictions constantly over 40.000) season after season among all(!) football leagues in the world. (Lange, 2021) Thanks to the low ticket prices (the lowest amongst the top leagues) (Matthews, 2011) and the supporters-centric policies of the clubs, the German Bundesliga has on average the highest number of supporters present at the stadium out of not only the European top leagues, but every domestic league in the world. (Burke, 2017) This obviously also implicates big home crowds. On matchdays, these factors combined create the notoriously exceptional (home) atmosphere, which the Bundesliga is understandably extremely proud of. As it has been presented during the analysis of crowd effects in Chapter II.2, the high number of home fans is one of the most important explanatory variables of home advantage. Thus, the still relatively high home win ratio (and consequently the home advantage) in the Bundesliga, can be most probably explained by the significant size of the home crowd.

The Italian top division (Serie A) comes with tremendously interesting surprises and trends between 1929-1930 and 2022-23; see Figure 3. Two seasons (1943-44 and 1944-45) are missing due to the Second World War.

3. Figure: Distribution of the Match Results by Season in the Serie A



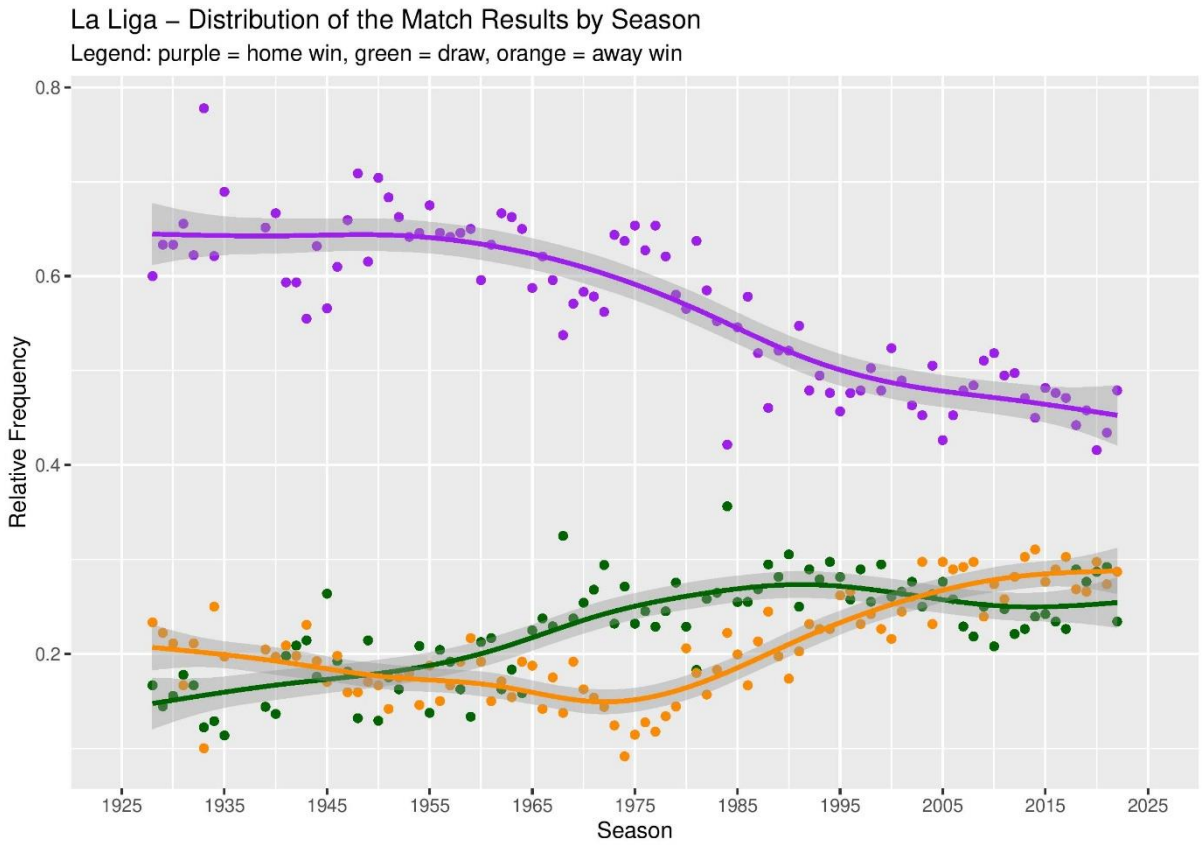
Source: Own calculations and own design based on the data presented in Chapter IV.

The changes that happened in Serie A are significantly different to what we have seen so far. To begin with, the huge number of draws and the strange trend of the draw ratio is striking. In the early 1930s the draw ratio was moving around the already-seen 25% level, but in the second part of the decade it rose above 30%. After a mild stop in the 40s and early 50s (back to 25%), the late 50s launched a big increase, keeping its momentum through the 60s (generally above 30%) arriving at the famous defensive football of the 70s and 80s resulting in extremely high draw ratios (around 40%!). The greatest example is the 1978-79 season where the home team won 38%, the away side won just 17% of the matches, leaving 45% of the games to end in a draw, which is an astonishing number. The draw ratio only started to decrease in the late 90s, thanks to the finally constantly improving performance of the away teams. This caused the convergence of the draw ratio back to the "more normal" 25% level (already familiar from England and Germany), which has been a mainly fixed value for the most recent 20 years of Italian top tier football. To sum up the information indicated by Figure 3, the changes in the win ratios have been very chaotic in the Italian Serie A, producing huge extreme values in all three cases. We have not met these "swings" neither in Germany (nor in Spain – as we will see) and especially not in England, where the trends are much more balanced (lower variance). The

reason for the dominance of draws can be found in the famous, but today slightly less known defensive tactics called “catenaccio”, a major characteristic of Italian football from the 60s until the end of the 90s. “Catenaccio” is a tactical system in football, which emphasizes defence with strict man-to-man marking and a “sweeper” defender and/or defensive midfielder. The primary objective (maybe strangely) is not to score, but to not let the opponent find a goal. If both teams play defensive football like this, then in most cases the number of shots on target and so the number of goals will be lower than average, resulting in the increased probability of a draw. As the evolution of football and football tactics made the “catenaccio” obsolete, out-dated, the number of goals per game rose, as we will see in the next subchapter (Chapter V.1.2.).

The generally steady rate of home advantage with sudden drops due the modern football might be best seen in the case of the Spanish first division (Figure 4), as the shape of the home win ratio trend shows it between 1928-29 and 2022-23. There are three seasons missing (1936-37, 1937-38, 1938-39), the event to blame is the Spanish Civil War.

4. Figure: Distribution of the Match Results by Season in the La Liga



Source: Own calculations and own design based on the data presented in Chapter IV.

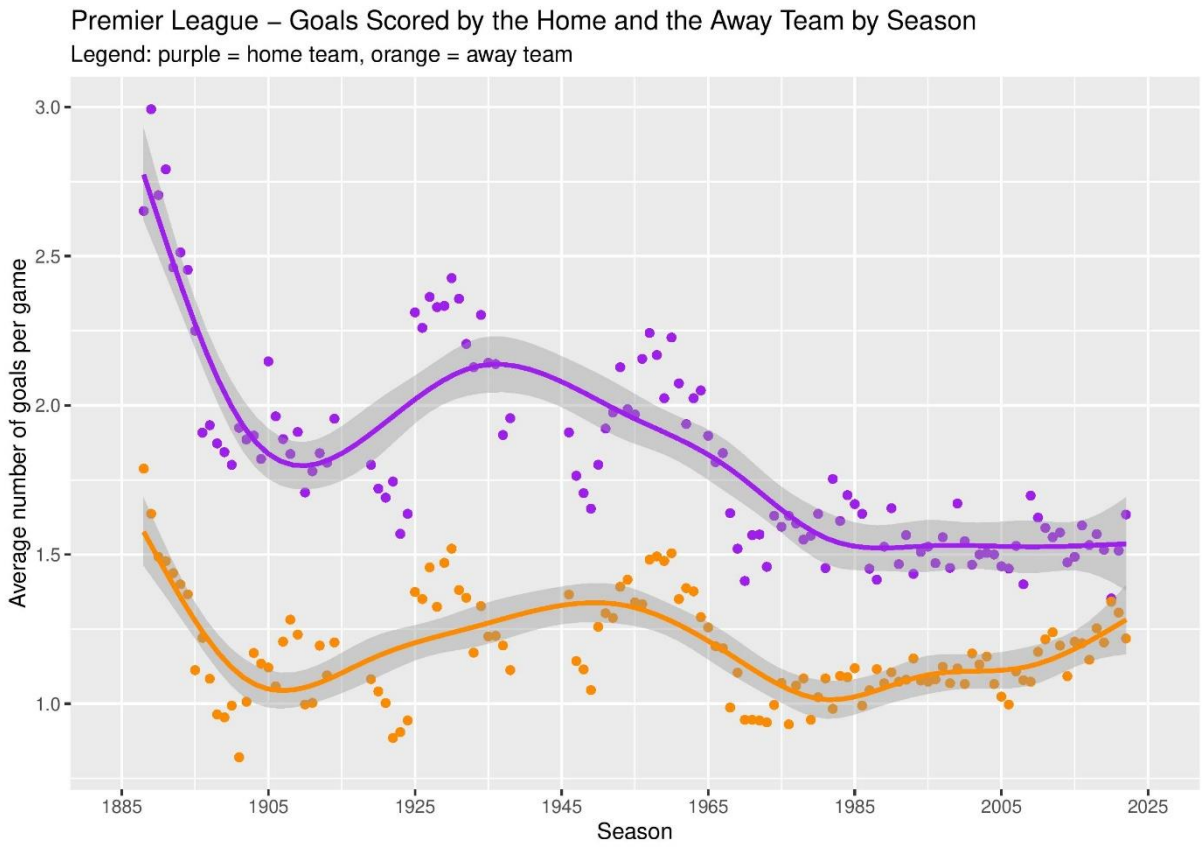
The home win ratio stayed steadily around the high initial value (above 60% with a few slightly lower exceptions) for long decades (almost for half a century). The decrease of home advantage

only started in the 80s, when away teams started to perform better (increasing away win and draw ratios), causing the home wine ratio to finally fall under 50% (a previously unthinkable level) starting in the late 90s. However, the gap (while undoubtably narrower) is still clearly visible in 2022-23 too: home teams win 48%, away teams win 29% of the games, while 23% end in a draw.

V.1.2. Distribution and Trends of the Average Number of Goals per Game

In the game of football out of all the result determining factors the number of goals is obviously the most important one. Therefore, it is worth to have a look at the goals scored by the home and the away teams, which could even explain some part of the trends seen in the previous chapter, i.e. the dramatic shift of the home win ratio. The figure below (Figure 5) shows the average number of goals per game scored by the home and the away team by season in the English Premier League.

5. Figure: Goals Scored by the Home and the Away Team by Season in the Premier League



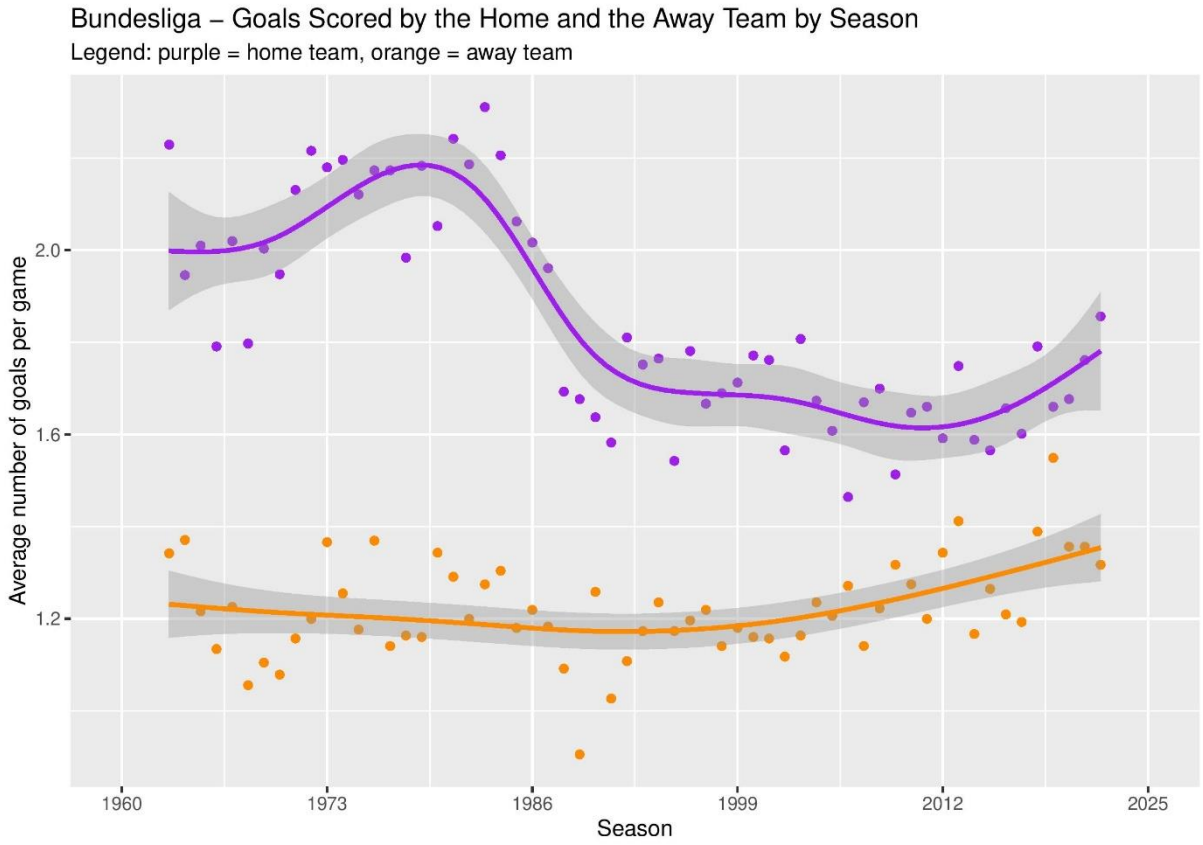
Source: Own calculations and own design based on the data presented in Chapter IV.

It is evident straight away, that both the home and the away teams' number of goals per game has been quite volatile over the more than 130 years (seasons). Similarly to the win ratios, the average goal per game indicator has decreased in the analysed period. On the one hand, in

contrast to the 2.5 goals/game during the late 1890s, home teams these days only score around 1.5 goals per game. On the other hand, the away teams' goals/game ratio has fluctuated between 1.0 and 1.3, which is a much more stable with respect to the volatile changes observed in the case of home teams. This implies that finding the net (i.e. playing high scoring attacking football) – like in the early days of the league – has become more and more difficult for the home teams as the years passed. Consequently, the home win ratio has diminished significantly, meaning that in the Premier League home teams do not anymore win ca. 60%, but only around 45% of the games.

Figure 6 plots the same indicator over time in the Bundesliga, Figure 7 in the Serie A and finally Figure 8 in La Liga.

6. Figure: Goals Scored by the Home and the Away Team by Season in the Bundesliga

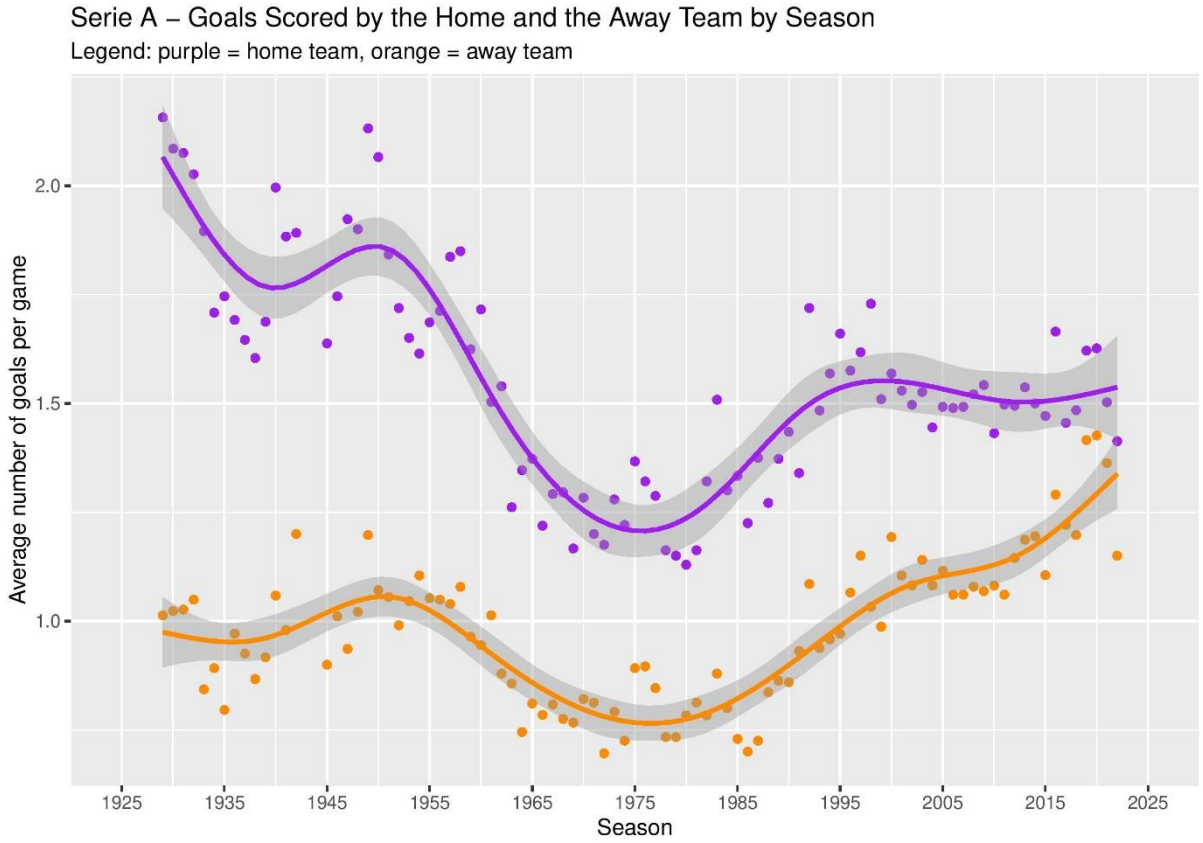


Source: Own calculations and own design based on the data presented in Chapter IV.

The away teams' goals/game ratio is more consistent in the case of the Bundesliga, as well. On the whole, dispersion around the 1.2 value was a strong characteristic during the league's 60 seasons. It is important to note, that we are witnessing a moderate rise in the recent seasons, reaching 1.30-1.35, and only falling once (in 2017-18) below the 1.2 threshold in the past 8 seasons. However, the home teams' case is much different. Up until the end of the 80s the home

teams had a staggering (especially in comparison to the Serie A for example, as we will see later) goals per game ratio – almost always over 2.0(!) – which perfectly aligns with the high (around 55%) home win ratio of the 70s and 80s, observed and discussed in the previous chapter. After this time period the average goals per game ratio for the home team drops to the 1.6-1.8 region. The use of the term "drop" is obviously quite harsh as these averages are still rather high, and indeed, thanks to the prolificacy of both sides, Bundesliga can boast about averaging more goals per game – around 2.7-2.8, but often over (!) 3.0 – than any other top league. (OneFootball, 2023)

7. Figure: Goals Scored by the Home and the Away Team by Season in the Serie A

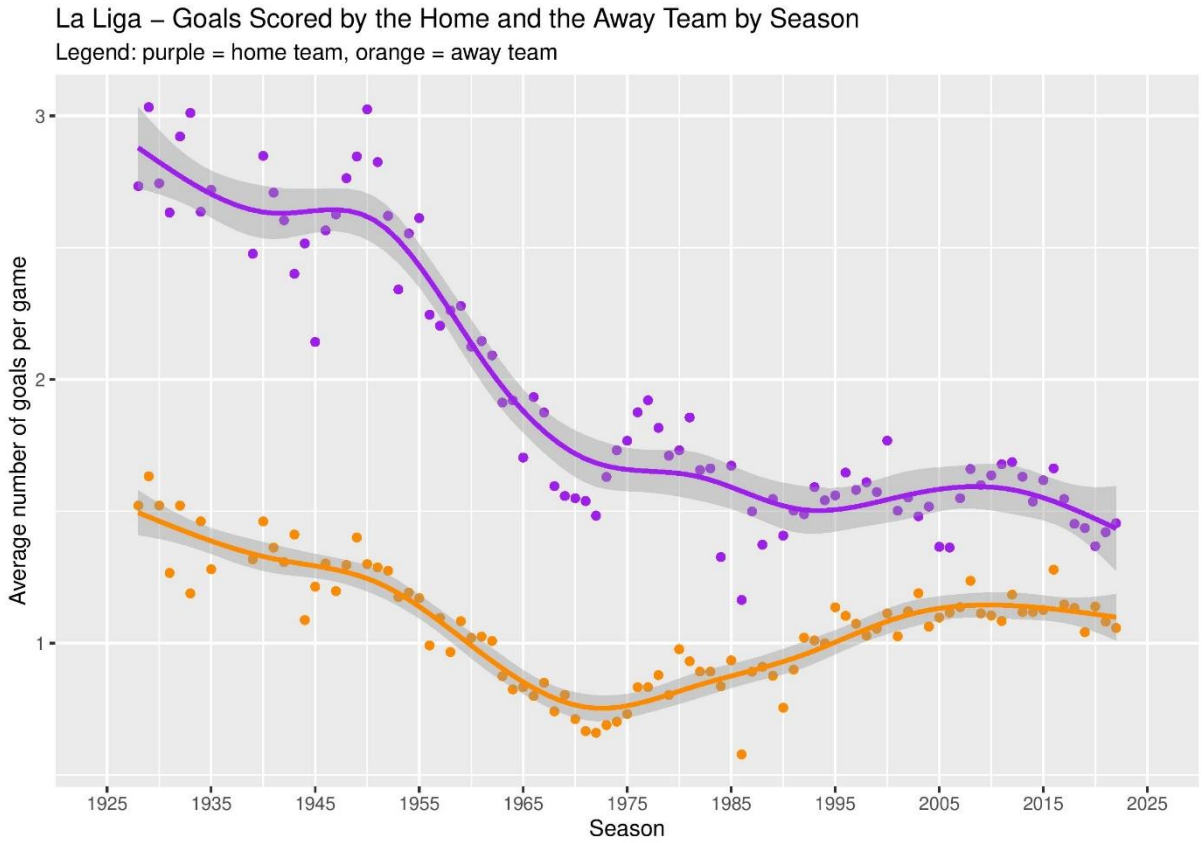


Source: Own calculations and own design based on the data presented in Chapter IV.

In the Italian top flight (Serie A), the trends of the goals scored per game beautifully show the defensive tactics of the 70s and reinforces the observations made in the previous subchapter (Chapter V.1.1.) when discussing the "catenaccio". This approach, strategy caused that in the 1972-73 season, the home teams only scored 1.18 goals per game, while the away teams only managed to find the net 0.7 times per game on average. Putting these together, we can see that in that season the average number of goals per game was way below 2.0, meaning that the most frequent results were 0-0, 1-0, 0-1, which is quite absurd! Obviously, it goes hand in hand with

the low home win ratio presented in the previous subchapter. After the 1970s, as the "catenaccio" style of play was slowly becoming outdated, and more offensive tactics were becoming more and more popular and applied by managers, the number of goals scored rose both at home and away. The away teams' attacking performance (measured in their number of goals scored) improved more quickly – really closing up on the home teams – resulting in the huge jump in the away win ratio.

8. Figure: Goals Scored by the Home and the Away Team by Season in the La Liga



Source: Own calculations and own design based on the data presented in Chapter IV.

The plot (Figure 8) above leads us to observe some quite interesting trends in the Spanish first division (La Liga). The dispersion of the data here is huge, the average goals/game has not been moving in such a narrow lane as we have seen before in the case of the other three top leagues. In the beginning, from the 20s until the middle of the 50s the home teams averaged way more than 2.5 goals per game. Interestingly enough, as time passed on, not only the home teams began to score less goals but the away teams, as well! In other words, the goal difference between the teams facing each other has not diminished. The goals/game indicator has declined until the dramatically low level of 0.66 for the away side, meaning that they were very rarely able to score more than once in a game. This immense "goalscoring inability" of the away teams

in La Liga caused the enormously high home win ratio – and therefore home advantage – even during the 70s. The peak of this extreme home advantage was the 1976-76 season, when 65% of the games were won by the home team, thanks to (among other factors) the 1.77 goals scored by the home team on average with the away teams only managing to reach 0.73 goals per game on average. However, from the 80s, similarly to the Serie A, the performance of the away side has improved greatly, scoring more and more goals, while the home side got stuck at around the 1.6 value. Therefore, the dominance of the home side has diminished gradually.

V.1.3. Differences Between the Leagues

Summarizing the first part of the analysis, i.e. the changes and characteristics of home advantage over time, we can say with confidence, based on detailed examination of huge amount of data in the European football leagues, that the home advantage's effect on the match result has clearly declined over time. As to what concerns the leagues individually, the following statements can be made. Similarly to England, we have observed declining trends in the home win ratio in the Spanish, German and Italian first divisions as well. However, based on the informative figures there are a number of intriguing differences which are worth to be discussed. The Spanish first division (La Liga) stands out in the degree of change in the home win ratio. In the 1950s the home team won almost 70% of the matches. This extremely high home win ratio plummeted to a 45% value by the 2000s. This degree of change is unique. The peculiarity of the Italian first division (Serie A) is that it has the lowest home win ratio among the analysed leagues: from the initial level of circa 60% in the 1920s, it has been diminishing ever since – obviously not constantly. The biggest rate of decline happened during the 30-year period between 1950 and 1980, reaching a low point in the 1978-89 season, where the home win ratio was at a remarkably low 38%! The following two decades brought a short period of increase due to the ultra-defensive tactics (the infamous "catenaccio" – presented above in detail) becoming obsolete, thanks to which the home win ratio rose above 50% on some occasions. The recent years came with the rapid improvement of the away teams' attacking performance (higher goals per game ratio) which kick-started another decreasing period in the home win ratio, that can be seen today. As regards the German first division (Bundesliga), it is fundamental to highlight the role of the home crowd in influencing the match result, while also noting that the amount of data available was the smallest in this case (the first retrievable season is the 1963-64 season). As we have seen, the fact, that the Bundesliga consistently has the highest attendance among all football leagues in the world, can greatly explain home advantage in Germany through the psychological effects exerted on the facing teams and the referees by

the large home crowd, which in the end result in the quantitative and observable factors, i.e. the high number of goals scored per game by the home team and the high home win ratio. These empirical results align with previous findings in the scientific literature (see Chapter II.1.2. – Crowd Effects).

If we would like to compare the leagues based on their home win ratio instead of focusing on the leagues' characteristics separately, we can make the following statements. In the first part of the 20th century, there was on average a ca. 10 percentage points gap between the home win ratios of the analysed leagues (ca. an interval of 55%-65% on average). Today, this gap can be found at the 45%-50% mark (so a gap of only 5 percentage points), which means that not only did the home win ratio decline in all four analysed leagues separately, but ultimately the difference between these home win ratios decreased as well, which means that the leagues sort of converged to each other. Obviously, during the over 120 years long period there were times when this difference between the leagues was huge. For example, in the 1975-76 season, when in La Liga 65% of the games were won by the home team, while in the Serie A only 46% of the games. The Bundesliga and the Premier League lay between these two extremes with a home win ratio of 57% and 50% respectively. This extraordinary gap of almost 20 percentage points illustrates how massive was the difference between the leagues at that time. The intensity of the leagues might provide a plausible explanation for this high dispersion. In this time range (i.e. around the 1975-76 season) the Serie A and the Premier League were categorically competitive leagues, meaning that there were always 4-5 top teams fighting for the league title, while La Liga was a two-horse race and the Bundesliga had, in effect, only one favourite. La Liga saw the same two teams arriving in the first and second place year after year with the others trailing several points behind them. The above referred clear favourite in the Bundesliga was Borussia Mönchengladbach in the 1969-70, 1970-71 and the 1974-75, 1975-76, 1976-77 seasons with Bayern Munich running ahead of the pack in the three seasons in-between. Coming back to the general trends observed, the difference between the leagues' home win ratios has decreased over time, with the home win ratios decreasing separately in all four leagues as well. By the 1990s the home win ratio was at ca. 50% in all four leagues and since then it has dropped and mainly remained under this value with only a 5 percentage points difference among the leagues.

We have to note, that the below 50% values mean that for the home team the probability of drawing or losing the game is higher than the probability of winning it. This could lead to false conclusions, like: "while there is still a higher win probability when playing at home than when playing away, the home advantage in a strictly statistical sense does not exist, as for the home

team the probability of not winning is higher than the probability of winning". One has to be very careful with such statements; and the reason why these statements cannot actually be made is because they criminally underrate (or worse, they do not even take into account) the role and the value of the draw. Furthermore, if we formulate the cited statement from a slightly different perspective, we get the following: "the home advantage is an existing phenomenon (even from a strictly statistical point of view), as for the home team the probability of not losing the game is much higher than the probability of losing it". I have not made any changes to the statement except the main focus: what is considered success? Only victory counts or avoiding defeat as well? To close this argument, we have to highlight the draws again. They do matter, and yes, they do play a great role in the phenomenon of home advantage as well. Why? Firstly, and banally, because the 1 point for a draw is still higher than the 0 point for the loss. Secondly, and more importantly, because an underdog (on paper less favourite) team can often achieve a draw which feels like a win against their – on paper – more favourite opponent, in many instances thanks to the home crowd, environment, atmosphere, i.e. thanks to the home advantage.

V.2. Match Statistics and Betting Odds

Turning to the other two topics, goals and research questions of this thesis, the multinomial logistic regression has to be preceded by an extensive descriptive statistics and correlation analysis variable by variable otherwise the model outputs would be lacking context.

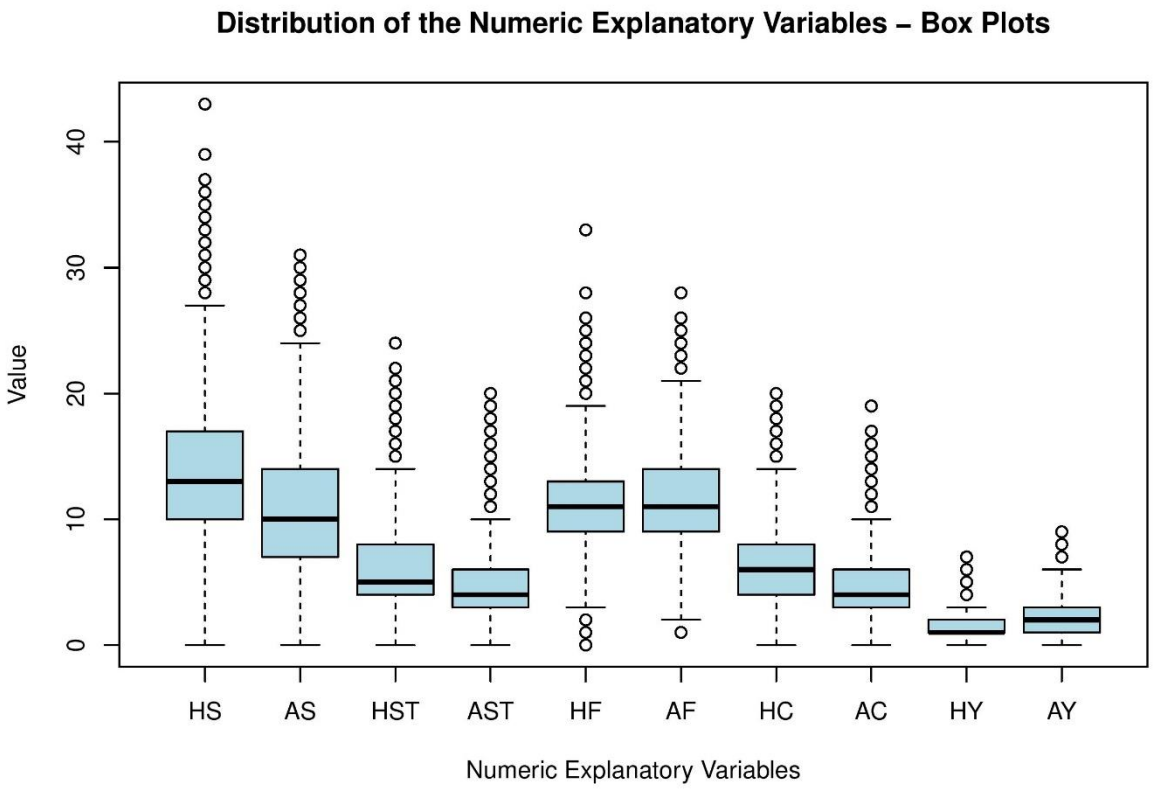
V.2.1. Preliminary Data Inspection and Descriptive Statistics

Again, let us start with the English Premier League (from 2004-05 to 2022-23). The finalised data set for this part contains the full time result (FTR) as the future dependent variable, the half time result (HTR), the 12 post-match statistics that will serve as explanatory variables (HS, AS, HST, AST, HF, AF, HC, AC, HY, AY, HR, AR – see Chapter IV.2. for the detailed descriptions, but in short: shots, shots on target, fouls, corners, yellow and red cards for the home and the away team respectively) and the three pre-match betting odds of the online bookmaker Bet365 (for the match outcome). For the analysis of the post-match statistics the betting odds are not taken into account, they will be dealt with separately (as described clearly in the topic and methodology parts – Chapter I.2. and Chapter III.), and the accuracy of the different models will be compared afterwards.

First of all, the FTR and the HTR variables have to be set as factors (categorical variables). Their unordered levels are A – away win, D – draw, H – home win respectively, with the reference level being A, away win. Next, the red card variables (HR, AR) have only a few

distinct values (0, 1, 2, 3 and 0, 1, 2 respectively), which means that they are not "real" numeric variables; thus for the upcoming statistical models and methods they are better as factors (categorical variables). Their ordered levels are 0-1-2-3 and 0-1-2 respectively, i.e. the reference level being 0 red cards in both cases. Now the data is prepared for analysis. For the numeric variables boxplots are the best way the get a first grasp of the distributions (see Figure 9).

9. Figure: Descriptive Statistics – Box Plots of Post-Match Statistics in the Premier League



Source: Own calculations and own design based on the data presented in Chapter IV.

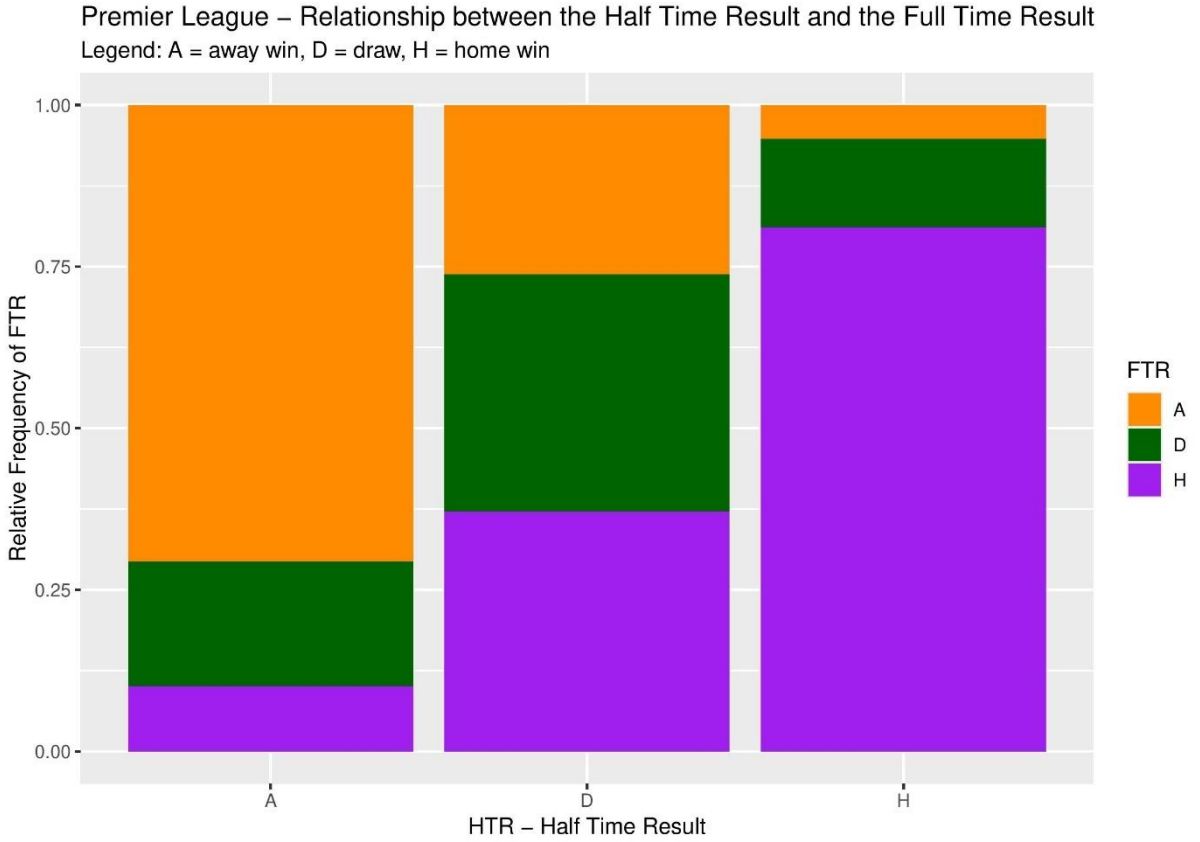
Not surprisingly, each variable has a strongly skewed distribution (with a right tail). More often than not, taking the natural logarithm helps in these cases, however this is a bit more particular case as taking the logarithms cannot be justified here, because they have no economic/statistic/real-world meaning or interpretation. The logarithmized variables would make absolutely no sense, thus while they might give better results on paper, they will have no actual meaning. So, I acknowledge the not so problematic right-skewed distributions, and that they might somewhat worsen the multinomial logistic regression estimates but for the sake of meaningful interpretation I chose to work with these distributions.

V.2.2. Relationship Between the Dependent and the Independent Variables

Let us go one by one. FTR and HTR are two categorical variables, so their relationship is called association. This can be tested for example by the Pearson's chi-squared test (for the detailed

description of the test see the methodology part – Chapter III.). Here H_0 is rejected, FTR clearly depends on HTR. Let us illustrate this relationship (Figure 10):

10. Figure: Relationship Between the Half Time and Full Time Result in the Premier League



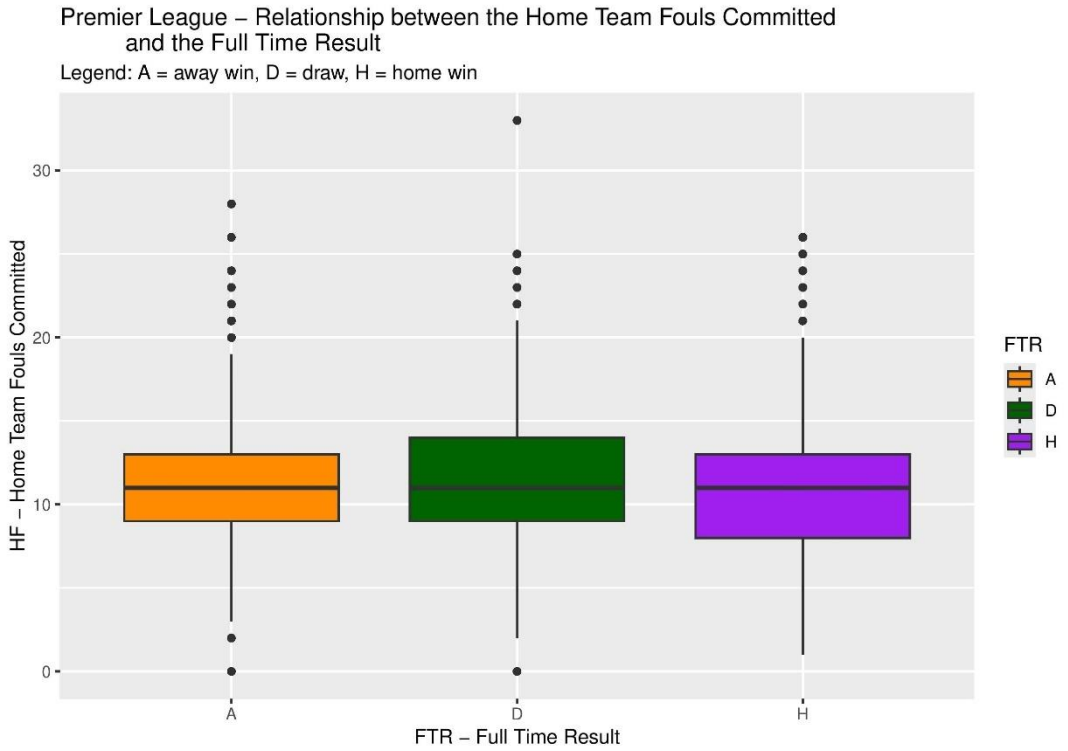
Source: Own calculations and own design based on the data presented in Chapter IV.

Figure 10 clearly shows the strong relationship between HTR and FTR. If the home team is leading after the first half, they will also win the game with around 80% probability. Similar (but weaker) statement can be formed for the away team, and not unexpectedly a draw after 45 minutes basically bears no information about the full time result.

FTR is categorical, and HS is numeric, so the Kruskal-Wallis test is needed to test their relationship (for the detailed description of the test see the methodology part – Chapter III.). Here H_0 is rejected, FTR clearly depends on HS. Same goes for AS, HST, AST, HR and AR as well. FTR is not independent form HC, AC and HY either, but their relationship is less strong. On the other hand, in the case of HF, AF and AY the H_0 of the Kruskal-Wallis test cannot be rejected, meaning that FTR is independent from these three variables individually.

Let us visualize and comment the case of HF (Figure 11) and AST (Figure 12).

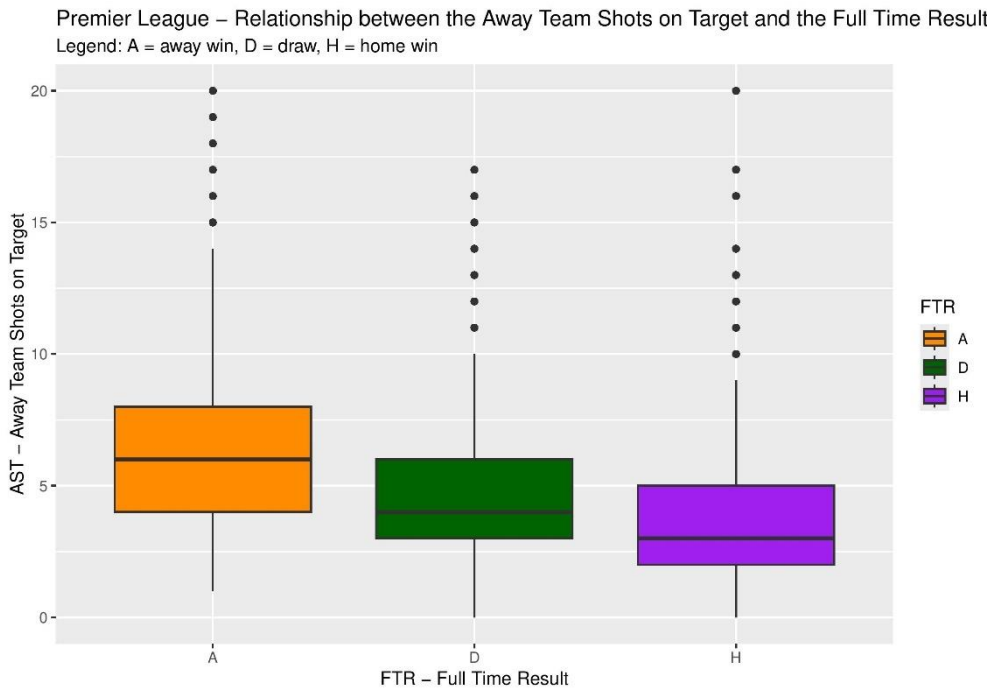
11. Figure: Demonstrated Independence Between the FTR and the HF in the Premier League



Source: Own calculations and own design based on the data presented in Chapter IV.

It is clear as daylight, that the medians are indeed the same across the three possible match outcomes, i.e. that the H_0 of the Kruskal-Wallis test holds, meaning that the number of fouls committed by the home team in itself does not have any information on the result of the game.

12. Figure: Relationship Between the FTR and the AST in the Premier League



Source: Own calculations and own design based on the data presented in Chapter IV.

The shots on target by the away team on the other hand, clearly correlate with the final result of the match. The higher the number of the AST, the higher the probability of the game finishing with an away win. If the away team fails to hit the target at least two times, it becomes basically guaranteed that the home team does not lose that match (see Figure 12).

In summary, when looking at the variables individually, most of them have a statistically significant relationship with FTR. The three variables which FTR seems not to depend on are HF, AF, AY, i.e. all connected to "normal" fouls and their penalization.

V.2.3. Multinomial Logistic Regression – Model Selection and Prediction

Following the guidelines set and described in the methodology part (Chapter III.), the starting model is the full model, i.e. the one that contains every explanatory variable. This is also the reference model for the backward stepwise model selection. Both the AIC and the BIC method result in the same model which will be my final multinomial logit model. After computing the Wald-tests (z-tests) and the p-values, the following output can be presented (Table 3).

3. Table: Multinomial Logistic Regression Models in the Premier League

	Dependent variable (FTR – reference category: A)			
	First model		Final model	
Variables	FTR – D	FTR – H	FTR – D	FTR – H
HTRD	1.571*** (0.082)	2.219*** (0.104)	1.571*** (0.081)	2.208*** (0.104)
HTRH	2.226*** (0.126)	4.586*** (0.134)	2.226*** (0.125)	4.565*** (0.133)
HS	0.005 (0.010)	0.006 (0.011)		
AS	-0.003 (0.012)	-0.019 (0.013)		
HST	0.122*** (0.017)	0.286*** (0.018)	0.127*** (0.014)	0.296*** (0.015)
AST	0.190*** (0.019)	-0.311*** (0.020)	0.193*** (0.014)	0.332*** (0.016)
HF	0.007 (0.011)	-0.010 (0.012)		
AF	0.023** (0.010)	0.020* (0.011)	0.024** (0.010)	0.020* (0.011)
HC	-0.010 (0.014)	-0.071*** (0.015)	-0.008 (0.013)	-0.065*** (0.014)
AC	0.044*** (0.015)	0.078*** (0.016)	0.042*** (0.015)	0.070*** (0.015)
HY	0.008 (0.032)	-0.063* (0.035)	0.014 (0.030)	-0.077** (0.033)
AY	0.082*** (0.031)	0.034 (0.033)	0.083*** (0.030)	-0.077** (0.033)
HR1	-0.451*** (0.143)	-1.147*** (0.173)	-0.451*** 0.142)	-1.166*** (0.173)
HR2	-2.637** (1.064)	-13.934*** (0.00)	-2.632** (1.062)	-13.685*** (0.00)

HR3	-11.373*** (0.00)	-11.363*** (0.00)	-13.724*** (0.00)	-13.438*** (0.00)
AR1	0.567*** (0.154)	0.796*** (0.160)	0.570*** (0.154)	0.810*** (0.160)
AR2	13.508*** (0.390)	15.829*** (0.390)	12.270*** (0.390)	14.590*** (0.390
Constant	-1.550*** (0.251)	-1.905*** (0.273)	-1.471*** (0.190)	-2.062*** (0.212)
Akaike	11078.020	11078.020	11071.980	11071.980
Bayesian	11325.86	11325.86	11278.52	11278.52

Note: *p<0.1; **p<0.05; ***p<0.01

Source: Own calculations and own design based on the data presented in Chapter IV.

Based on the significance of the variables, the values of the Akaike (AIC) and the Bayesian (BIC) information criteria (see their detailed description in the methodology part – Chapter III.), the final model is definitely better than the first (full) model. The HS and the AS were excluded in the process, most probably because they are strongly correlated with the HST and AST variables, which are better predictors for FTR. Their removal helps the model specification. Also, HF is excluded from the final model, but this is not a surprise, because – as we have seen above in Chapter V.2.2. – FTR did not even depend on HF on its own. The vast majority of the coefficients are convincingly significant and have logical interpretations, which are straightforward. Thus, instead of analysing each, I will give the meaning of two interesting variables HST and HR. When the home team shots on target increase by 1, the chance (probability) of the full time result being a draw with respect to being an away win increases:

$$by\ a\ factor\ of: e^{\widehat{\beta}_{HST}{}^{D}} = e^{0.127} = 1.135417,$$

$$i.e.\ by: \left(e^{\widehat{\beta}_{HST}{}^{D}} - 1\right) \cdot 100\% = (e^{0.127} - 1) \cdot 100\% = 0.135417 \cdot 100\% = 13.5417\%.$$

Parallelly, when the home team shots on target increase by 1, the chance (probability) of the full time result being a home win with respect to being an away win increases:

$$by\ a\ factor\ of: e^{\widehat{\beta}_{HST}{}^{H}} = e^{0.296} = 1.34447,$$

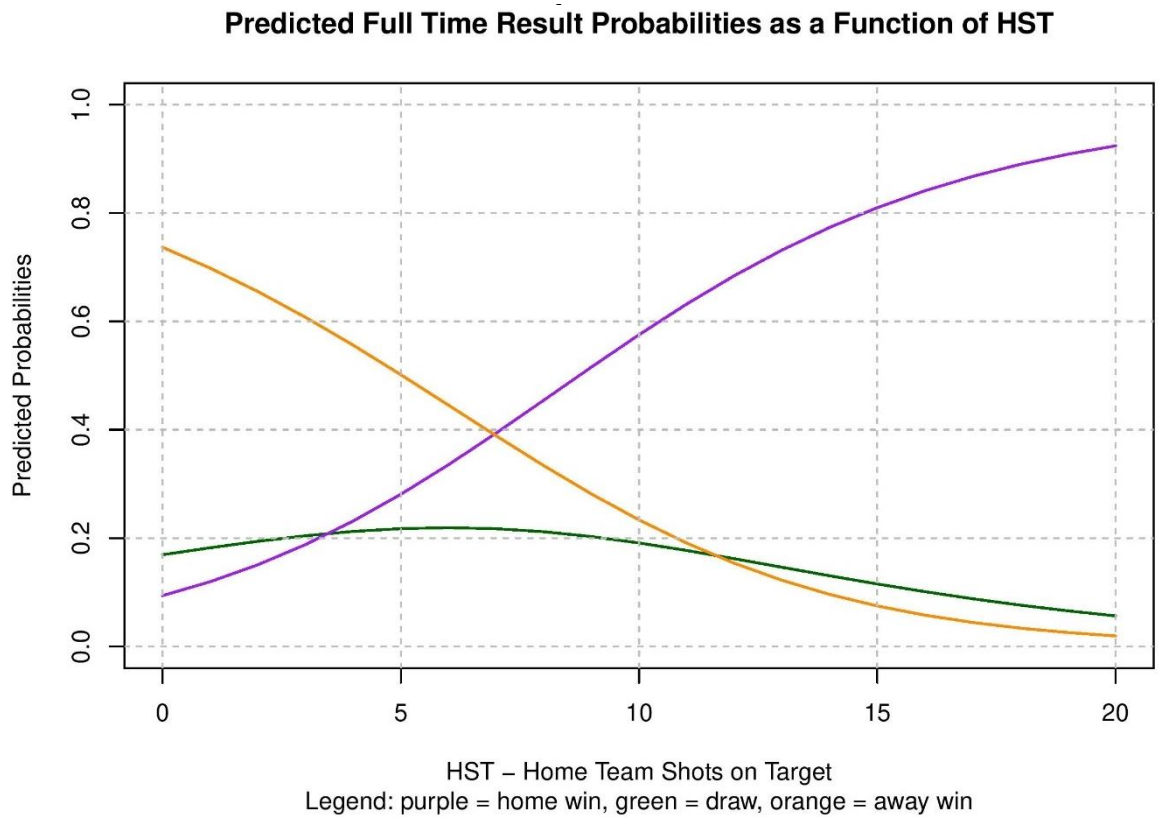
$$i.e.\ by: \left(e^{\widehat{\beta}_{HST}{}^{H}} - 1\right) \cdot 100\% = (e^{0.296} - 1) \cdot 100\% = 0.34447 \cdot 100\% = 34.447\%.$$

Thus, with the home team shots increasing the full time result is more likely to be a draw than an away win, and even more likely to be a home win than an away win, so more likely to be a home win than a draw as well (see below).

$$if\ HST\ increases\ by\ 1, then\ p_H > p_D > p_A$$

At this point, this argument can be completed by plotting the predicted probabilities for the home team shots on target variable. This type of plot illustrates all the interpretations made above and is one of the most famous and useful characteristics of the multinomial logistic regression models (see Figure 13).

13. Figure: Predicted Full Time Result Probabilities as a Function of the Home Team Shots



Source: Own calculations and own design based on the data presented in Chapter IV.

The already described dynamics and trends are easy to see on Figure 13. With zero home shots on target, away teams have roughly 75% chance of winning. Note, that the home win probability is still not zero. Why? Is it possible for a team to win with zero shots on target? Yes, because own goals exist. As the home team is able to raise their number of shots on target, the probability of a home win is increasing and the probability of an away win is decreasing. The trend of the draw probability is really compelling. With only 0-3 shots on target by the home team, the probability of a draw is higher than the probability of a home win. Meaning that home teams in the Premier League should be able to hit the target at least 3 times in a game to even have a considerable chance of winning. The next point of intersection is at around 7 home shots on target, where the home win and away win probability are equal, just below 40%. It is great news, that this predicted, estimated result is confirmed by the original empirical data. Without any models, the break-even point between the home win and the away win probability is

7.080193, which is almost exactly the same as what the plot indicates. Not surprisingly, this is also the maximum point of the predicted draw probabilities, after that it starts to decrease until it reaches the last point of intersection, at 12 home shots on target. After this point the predicted probability of an away win is even lower than the predicted probability of a draw. Finally, if the home team is able to have 20 attempts on target, they are almost sure to win the match, with a predicted probability of approximately 90%.

Let us look at the home team red cards now. When the number of red cards given to the home team increases from 0 to 1, the chance (probability) of the full time result being a draw with respect to being an away win decreases:

$$by\ a\ factor\ of: e^{\widehat{\beta}_{HR1}{}^{D}} = e^{-0.451} = 0.63699,$$

$$i.e.\ by: \left(1 - e^{\widehat{\beta}_{HR1}{}^{D}}\right) \cdot 100\% = (1 - e^{-0.451}) \cdot 100\% = 0.363 \cdot 100\% = 36.3\%.$$

Parallelly, when the home team red cards increase from 0 to 1, the chance (probability) of the full time result being a home win with respect to being an away win decreases:

$$by\ a\ factor\ of: e^{\widehat{\beta}_{HR1}{}^{H}} = e^{-1.166} = 0.31161,$$

$$i.e.\ by: \left(1 - e^{\widehat{\beta}_{HR1}{}^{H}}\right) \cdot 100\% = (1 - e^{-1.166}) \cdot 100\% = 0.68839 \cdot 100\% = 68.839\%.$$

Thus, with the home team red cards increasing from 0 to 1, the full time result is less likely to be a draw than an away win, and even less likely to be a home win than an away win, so less likely to be home win than a draw as well (see below).

$$if\ HR\ increases\ from\ 0\ to\ 1, then\ p_H < p_D < p_A$$

Every other coefficient can be interpreted analogously. I want to highlight one more intriguing effect, the case of HY. From Table 3 it is clear that $\hat{\beta}_{HY}{}^{D}$ is not significant, meaning that an increase in the yellow cards given to the home team has no significant effect on the probability of the full time result being a draw with respect to being an away win. However, $\hat{\beta}_{HY}{}^{H}$ is significant and negative, meaning that that an increase in the yellow cards given to the home team significantly decreases the chance (probability) of the full time result being a home win with respect to being an away win. This means that an increase in the home yellow cards results in a certain decrease in the probability of a home win but does not have any information whether the full time result is a draw or an away win.

Now, having reached, analysed and interpreted the final multinomial logistic regression model, I can calculate the model's predicted match outcome probabilities and the specific outcome categories. This can be represented in a confusion matrix, which I prepare both for the first (full) model and the final (best) model (see Table 4). Then, the accuracies can be computed and compared as well, as discussed thoroughly in the already cited Chapter I.2. and Chapter III.

4. Table: Confusion Matrices of the Models – Predicted Outcomes in the Premier League

First (full) model					**Final (best) model**				
		predicted FTR					predicted FTR		
		A	D	H			A	D	H
actual FTR	A	1519	320	292	actual FTR	A	1517	312	302
	D	507	507	758		D	506	513	753
	H	244	354	2719		H	244	361	2712
Accuracy (ACC) = 0.6572					Accuracy (ACC) = 0.6568				

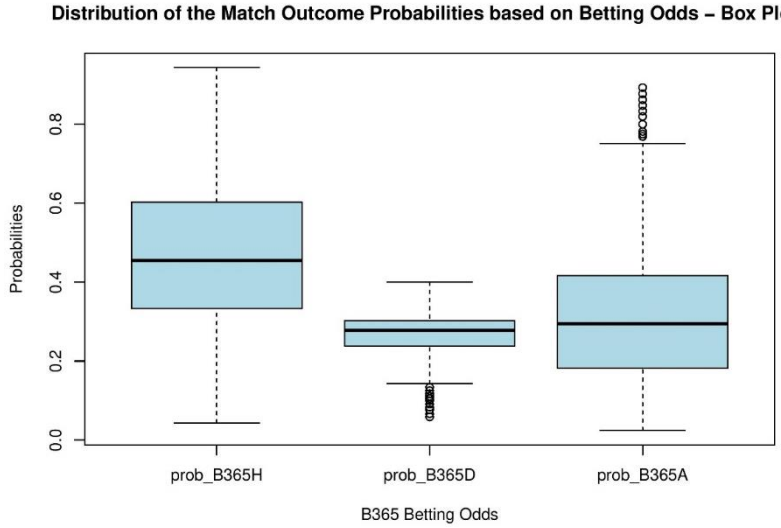
Source: Own calculations and own design based on the data presented in Chapter IV.

Table 4 indicates that the final model is indeed better than the first model, as it has achieved a better model specification, better AIC and BIC values, more significant variables, while only losing a marginal amount of predictive accuracy. This means that post-match statistics are able to correctly posteriorly classify the full time result in 65.68% of the cases. However, it has to be noted that this relatively high value is mostly thanks to the strong relationship between the half time result and the full time result. In fact, if HTR is presumed to be unknown, and the same model selection process is run without it, the accuracies are "only": 0.5745 and 0.5731 respectively. In spite of that, the multinomial logistic regression model building process concluded with success, the coefficients are interpretable and contain significant information on the match result, the predicted probabilities are not only logical and informative, but they are also supported by the original data, and finally the posterior classification accuracy is strong.

V.2.4. Match Outcome Prediction with Betting Odds

Turning to comparing these results with the predictive accuracy of the pre-match betting odds, the same steps (descriptive statistics, examining direct statistical relationships) can be repeated. But first, following the steps and the analytical framework outlined in the methodology part (see Chapter III.), the betting odds have to be transformed first to probabilities. Recall, that the reciprocals of the betting odds result in the probabilities assigned to the match outcomes by the bookmaker.

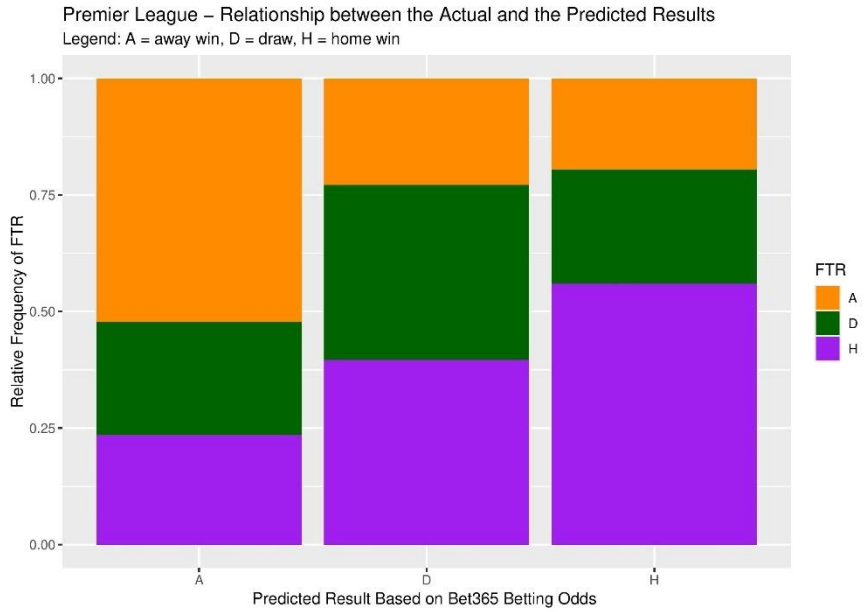
14. Figure: Box Plots of Outcome Probabilities based on Betting Odds in the Premier League



Source: Own calculations and own design based on the data presented in Chapter IV.

Some outliers and the relatively small interquartile range have to be noted, but otherwise the distributions of the probabilities are comfortable to work with. And in fact, they have to be worked with, as these probabilities have to be transformed further, now to predicted results. So, using these probabilities, a single prediction is assigned to each match, based on a chosen rule, which in my case is the somewhat modified classification rule of Sumpter (2016). After having obtained them, the relationship between the predicted results and the actual results (2 categorical variables, so it is an association) can be tested (Pearson's chi squared test) and illustrated (Figure 15) the same way as for example in the case of the association between HTR and FTR (see Chapter V.2.2. above). H_0 is rejected, FTR depends on the predicted results.

15. Figure: Relationship Between the Actual and the Predicted Results in the Premier League



Source: Own calculations and own design based on the data presented in Chapter IV.

Figure 15 clearly shows that the match outcome predictions based on the pre-match betting odds are clearly influencing (or correctly forecasting) the actual full time result. The relationship is strongest (the forecasts are most accurate) in the case of a home win, i.e. home wins seem to be the easiest to predict ("safe bet"), which provides further evidence for home advantage, and the influence of home advantage not only on the actual final result of the game, but already on the pre-match expectations (represented by the betting odds), as well.

Finally, the predictions and their accuracy can be represented in a confusion matrix (Table 5).

5. Table: Actual and Predicted Results based on Betting Odds in the Premier League

Confusion Matrix – Predictions based on Pre-Match Betting Odds				
		predicted FTR		
		A	D	H
actual FTR	A	1144	11	976
	D	533	18	1221
	H	513	19	2785
Accuracy (ACC) = 0.5467				

Source: Own calculations and own design based on the data presented in Chapter IV.

The result is astonishing. The predictions based on the pre-match betting odds are accurate in almost 55% of the cases, which is almost as good as the posterior predictions (classification) of the post-match statistics (when HTR is presumed to be unknown). This confirms and demonstrates the findings of previous studies conducted in this field (presented in the literature review part; see Chapter II.2.), i.e. that pre-match betting odds contain more relevant information than post-match statistics and even the result of the match itself. (Wunderlich & Memmert, 2018) This indicates how fine-tuned and well-constructed the mathematical-statistical models of the bookmakers are, used to compute exact probabilities and to "price" the games, i.e. to compute the betting odds with a precision of several decimals. No wonder that these models are held in complete secrecy and security. Finally, it also shows that relevance of the efficient-market hypothesis, which in this case means that these pricing (betting odds computing) models most likely do indeed contain indicators of home advantage. Vice versa, home advantage is so strong and robust even today, that it even affects betting odds estimates.

V.2.5. Comparative Analysis – Differences Between the Models and Across Europe

After having analysed in detail the Premier League, the same lengthy and meticulous but needed and precise method can be repeated for the other top leagues (Bundesliga, Serie A, La Liga), plus for some further football leagues for comparison purposes, where I was able to retrieve the same variables (France – Ligue 1, England's 2nd, 3rd and 4th divisions – Championship, League 1 and League 2). This means that the whole process (from Chapter V.2.1. to Chapter V.2.4.) from the preliminary data inspection, descriptive statistics, box plots, analysis of the individual statistical relationships between the explanatory variables and the dependent variable (FTR), the multinomial logistic regression model building, selection and prediction, to the match outcome prediction based on betting odds is applicable analogously. In line with the arguments presented in the chapters above during the analysis, the best way to present the main results (i.e. the predictive accuracies) is a compact table, which contains the six major models in the case of each league. These are the full multinomial logistic model, the final (after the model selection process) multinomial logistic model, the same when the half time result is assumed to be unknown, the HTR in itself (to demonstrate its strong relationship with FTR), and finally and most importantly the predictions based on the probabilities computed from the pre-match betting odds. Behold Table 6.

6. Table: Summary – Predictive Accuracies of Different Models Across European Leagues

	Models					
Leagues	Post-match statistics (full)	Post-match statistics (final)	HTR assumed unknown (full)	HTR assumed unknown (final)	Half time result alone	Pre-match betting odds
Premier League	0.6572	0.6568	0.5745	0.5731	0.6026	0.5467
Bundesliga	0.6278	0.6273	0.5044	0.5027	0.6008	0.5127
Serie A	0.6561	0.6565	0.5946	0.5924	0.5811	0.5447
La Liga	0.6719	0.6727	0.6067	0.6077	0.5917	0.5400
Ligue 1	0.6437	0.6437	0.5733	0.5733	0.5873	0.5082
Championship	0.6236	0.6228	0.5375	0.5372	0.5875	0.4610
League 1	0.6278	0.6272	0.5392	0.5392	0.5843	0.4716
League 2	0.6239	0.6252	0.5428	0.5423	0.5873	0.4489

Source: Own calculations and own design based on the data presented in Chapter IV

Table 6 shows that the final multinomial logistic regression models proved to be better than the full ones in each league. They achieved smaller models, better AIC and BIC values and more significant variables, while only losing a marginal amount of predictive accuracy, or even improving it in a few cases (Serie A, La Liga, League 2). The same goes for the multinomial logistic regression models when the half time result was assumed to be unknown. About HTR it can be stated that it is the variable that has the strongest relationship with FTR, and that is why it is able to have a good predictive accuracy on its own as well. The motive for that is that while match statistics are post-match variables and betting odds are pre-match, HTR is mid-match. Finally, the extreme precision of pre-match betting odds is well-illustrated for the top leagues (lower divisions have higher volatility, they are less predictable by nature, therefore also betting odds perform poorer in those cases). In the case of the top leagues, their forecast accuracy competes with the classification power of match statistics and HTR. One could say, well, the accuracies of predictions based on betting odds are still smaller. While that is true, let us highlight the most important distinction between them. These models based on match statistics can only predict posteriorly, which is not true forecasting (what will happen in the future), but classification (what happened in the past). The three simple betting odds values, which are publicly available way before the start of the game, are computed pre-match, are based on only anteriorly available information and statistics, and refers to the future, i.e. predicts the outcome of an upcoming match (true forecasting). This indicates that betting odds in all probability incorporates the relevant, historically proven match-influencing information of home advantage, which – though it has decreased – still exists and exerts its effect today as well. This further demonstrates the importance of home advantage.

VI. Results

This study presented the topic in detail, giving context to the world football, home advantage and betting odds through an extensive literature review. After a detailed theoretical description of the methodology, a range of statistical methods and tools were used to analyse the data, that was thoroughly collected, managed, cleaned, merged and presented beforehand.

The analysis of home advantage included examining its changes over time and its different characteristics in the top four leagues (England, Germany, Italy, Spain). Based on data collected from the earliest seasons possible (in England this is the 1888-89 season) up until the modern days (2022-23 season) I have made several calculations, constructed indicators, (e.g. home win ratios, average number of goals scored per game), prepared informative graphics, from which a decreasing trend can be observed in the effect of home advantage on the match result. This

decreasing trend has been consistently present since the beginning of organised football, and it seems to have stabilised. The most important quantitative factor contributing towards home advantage is the average number of goals scored per game, and its changes over time. This work has discussed several other possible quantitative and qualitative explanations (based on an extensive literature review connected to this empirical study) from environmental factors, through the geographical location, the psychological dynamics, the travel fatigue, the referee bias, the crowd effects and the team tactics, to the effects of the rules and the rule changes.

In the second part (post-match statistics and pre-match betting odds), descriptive statistics and tests of statistical relationships led to model building, where stepwise model selection resulted in optimal multinomial logistic regression models. Betting odds were first transformed to probabilities and then to actual predictions. The special case of the half time result was examined and discussed separately. Finally, the main results (predictive accuracies of the six different models and across eight European football leagues) were presented an in informative and easy-to-understand summarizing table (Table 6). While the predictive accuracy of post-match statistics proved to be higher, they can only predict posteriorly. Betting odds are finalised pre-match and are able to forecast future outcomes. The match-influencing effects of home advantage are incorporated in these odds and predictions, further demonstrating its relevance.

VII. Conclusion

It is safe to say that this study contains relevant information about the dynamics of home advantage, the accuracies of match result prediction models and the importance of betting odds. With the help of a thorough statistical analysis the theoretical parts were successfully put into context by looking at home advantage and betting odds through real-world empirical data.

It is important to reiterate the most essential conclusions of this work. The best way to do this is to recall the research questions (Chapter I.3.), the hypotheses formulated beforehand (Chapter I.4.) and then to summarize the most important results of the analysis, which provide the basis for evaluating the hypotheses, thus answering the research questions. The conclusions are:

1. Home advantage has indeed decreased over time, with significant differences between the leagues. Thus, the first hypothesis is confirmed.
2. Several different post-match statistics contain statistically significant information on the match result. Thus, the second hypothesis is partly rejected.
3. Pre-match betting odds do indeed have high predictive accuracy. Thus, the third hypothesis is also confirmed.

VIII. Limitations, Bias, Future Studies

This work is naturally far from covering all aspects of home advantage, match prediction and betting odds, there is definitely room for further calculations, more detailed analysis.

The betting odds estimates have some limitations regarding the data available. Another question is the applicability of the accuracy, as there also exist more advanced indicators for classification problems that measure the goodness of prediction in a more sophisticated way. The population could be broadened by including further football leagues. Same goes for the time horizon. Also, the outliers left in the data sets (see Chapter V.2.1.) might cause some bias. Same goes for the transformation of betting odds into predicted match result probabilities, because betting odds contain the profit margin of the bookmakers, thus the sum of the predicted probabilities exceeds 100%. This does not directly make the results biased, as the rule of assigning the actual prediction to each match is based on nominal comparisons, but in some not decisive cases (matches with no clear favourites) the predicted outcome might be biased. Again, there exist more complicated formulas that try to eliminate the profit margin from the betting odds, but these are difficult to apply, as the margins change over time.

Regarding future studies, there are a number of ways to start. Firstly, more advanced statistical tools could be used to build more complicated and sophisticated models. Secondly, the relevance of the topic makes it worth to conduct the study again in 5-10 years (even with the same methodology), putting special emphasis on the potential match-results-changing effects of COVID-19, the lockdowns and the games played behind closed doors. Thirdly, it would be interesting to analyse the possible influence and significance of recent rule changes in addition to those seen in Chapter II.1.7. These are:

1. the abolition of the away goals rule in 2021 (GOAL, 2022) and
2. the introduction of the video assistant referee (VAR) system.

The latter allows the referee to review questionable situations before making a decision. It was introduced in 2018 by the International Football Association Board (IFAB). The first major tournament where it was used was the 2018 World Cup. (IFAB, 2018) After the 2019-20 season most leagues adapted it, which theoretically should eventually lead to more impartial decisions by the referee, but at this point it is too early to say. Future studies are needed here. I, myself would be interested to revisit these datasets in the future as well.

References

Barnett, V., & Hilditch, S. (1993). The effect of an artificial pitch surface on home team performance in football (soccer). *Journal of the Royal Statistical Society: Series A (Statistics in Society), 156*(1), 39-50.

Boyko, R. H., Boyko, A. R., & Boyko, M. G. (2007). Referee bias contributes to home advantage in English Premiership football. *Journal of Sports Sciences, 25*(11), 1185-1194.

Brown Jr, et al. (2002). World Cup Soccer Home Advantage. *Journal of Sport Behavior, 25*(2), 134-145.

Buraimo, B., Forrest, D., & Simmons, R. (2010). The 12th man?: refereeing bias in English and German soccer. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*(2), 431-449.

Burke, D. (2017, October 24). *Top 10 best attended leagues in the world revealed*. Retrieved March 16, 2022, from https://onefootball.com/en/news/top-10-best-attended-leagues-in-the-world-revealed-17133211.

Clarke, S. R., & Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer. *Journal of the Royal Statistical Society: Series D (The Statistician), 44*(4), 509-521.

Curley, J. P. (2016). *engsoccerdata: English Soccer Data 1871-2016. R package version 0.1.5*. Retrieved október 18., 2021, from https://github.com/jalapic/engsoccerdata.

Dawson, P., Dobson, S., Goddard, J., & Wilson, J. (2007). Are football referees really biased and inconsistent?: evidence on the incidence of disciplinary sanction in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(1), 231-250.

Dohmen, T. J. (2005). Social pressure influences decisions of individuals: Evidence from the behavior of football referees. *Elérhető: SSRN 725541*.

Dosseville, F. E. (2007). Influence of ball type on home advantage in French professional soccer. *Perceptual and Motor Skills, 104*(2), 347-351.

Dowie, J. (1982). Why Spain should win the World Cup. *New Scientist, 94*(10), 693-695.

Evans, M. J., & Rosenthal, J. S. (2004). *Probability and Statistics: The science of uncertainty.* Toronto: Macmillan.

FIFA. (2011). *Laws of the Game 2011/2012.* Zürich: Fédération Internationale de Football Association.

Garicano, L., Palacios-Huerta, I., & Prendergast, C. (2005). Favoritism under social pressure. *Review of Economics and Statistics, 87*(2), 208-216.

GOAL. (2022, March 8). *Away goals rule: What is it & why did UEFA drop it from Champions League & Europa League?* Retrieved April 4, 2022, from https://www.goal.com/en-us/news/away-goals-rule-what-is-it-will-uefa-drop-from-champions/c8h61pd7sm916zulqe11ug53.

Goddard, J. (2006). Who wins the football? *Significance, 3*(1), 16-19.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric Statistical Methods.* New York: John Wiley & Sons.

IFAB. (2018, March 3.). *Historic step for greater fairness in football*. Retrieved February 3., 2022, from https://www.theifab.com/news/historic-step-for-greater-fairness-in-football/.

Jacklin, P. B. (2005). Temporal changes in home advantage in English football since the Second World War: What explains improved away performance? *Journal of Sports Sciences, 23*(7), 669-679.

Kovács, H. B. (2023). A hazai pálya előnye a labdarúgásban. In Antalík, I. et. al., *Litera Oeconomiae IV.* (pp. 291-315). Komárom: SJE-GIK.

Lange, D. (2021, November 18). *Average attendance games of the German football Bundesliga from 1990/91 to 2020/21*. Retrieved March 16, 2022, from https://www.statista.com/statistics/282974/average-per-game-attendance-german-football-bundesliga/.

Lefebvre, L. M., & Passer, M. W. (1974). The effects of game location and importance on aggression in team sport. *International Journal of Sport Psychology, 5*(2), 102-110.

Long, T. (2019, March 21.). *Home advantage over the year in European football*. Retrieved February 25., 2022, from https://rpubs.com/longtr99/homevsaway.

Matheson, V. A. (2003). *European football: a survey of the literature.* Williamstown: Williams College, Department of Economics.

Matthews, L. (2011, June 11). *Bundesliga breaks historical attendance record*. Retrieved March 16, 2022, from https://www.goal.com/en/news/15/german-football/2011/06/11/2527959/bundesliga-breaks-historical-attendance-record.

McCarrick, D., Bilalic, M., Neave, N., & Wolfson, S. (2021). Home advantage during the COVID-19 pandemic: Analyses of European football leagues. *Psychology of Sport and Exercise, 56*, 102013.

McSharry, P. E. (2007). Altitude and athletic performance: statistical analysis using football results. *British Medical Journal, 335*(7633), 1278-1281.

Messner, C., & Schmid, B. (2007). Über die Schwierigkeit, unparteiische Entscheidungen zu fällen. *Zeitschrift für Sozialpsychologie, 38*(2), 105-110.

Morris, D. (1981). *The Soccer Tribe.* London: Jonathan Cape.

Moschini, G. (2010). Incentives and outcomes in a strategic setting: the 3-points-for-a-win system in soccer. *Economic Inquiry, 48*(1), 65-79.

Neave, N., & Wolfson, S. (2003). Testosterone, territoriality, and the 'home advantage'. *Physiology & Behavior, 78*(2), 269-275.

Neave, N., & Wolfson, S. (2004). The Home Advantage: Psychological and Physiological Factors in Soccer. In D. Lavallee, J. Thatcher, & M. V. Jones, *Coping and Emotion in Sport* (pp. 131–148). New York: Nova Science Publishers.

Nevill, A. M., Balmer, N. J., & Williams, A. M. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise, 3*(4), 261-272.

Nevill, A. M., Newell, S. M., & Gale, S. (1996). Factors associated with home advantage in English and Scottish soccer matches. *Journal of Sports Sciences, 14*(2), 181-186.

OneFootball. (2023, October 19). *Bundesliga leads the way in goals per game in Europe*. Retrieved February 13, 2024, from https://onefootball.com/en/news/bundesliga-leads-the-way-in-goals-per-game-in-europe-38413094.

Page, L., & Page, K. (2007). The second leg home advantage: Evidence from European football cup competitions. *Journal of Sports Sciences, 25*(14), 1547-1556.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50*(302), 157-175.

Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of Sports Sciences, 4*(3), 237-248.

Pollard, R. (2006). Worldwide regional variations in home advantage in association football. *Journal of Sports Sciences, 24*(3), 231-240.

Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal, 1*(1), 12-14.

Pollard, R., & Pollard, G. (2005). Home Advantage in Soccer: A Review of its Existence and Causes. *International Journal of Soccer and Science, 3*(1), 28-38.

Pollard, R., & Pollard, G. (2005). Long-term trends in home advantage in professional team sports in North America and England (1876–2003). *Journal of Sports Sciences, 23*(4), 337-350.

Pollard, R., Silva, C. D., & Medeiros, N. C. (2008). Home advantage in football in Brazil: differences between teams and the effects of distance traveled. *Revista Brasileira de Futebol (The Brazilian Journal of Soccer Science), 1*(1), 3-10.

Schwartz, B., & Barsky, S. F. (1977). The home advantage. *Social Forces, 55*(3), 641-661.

Seckin, A., & Pollard, R. (2008). Home advantage in Turkish professional soccer. *Perceptual and Motor Skills, 107*(1), 51-54.

Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting, 28*(1), 55-72.

Staufenbiel, K., Lobinger, B., & Strauss, B. (2015). Home advantage in soccer–A matter of expectations, goal setting and tactical decisions of coaches? *Journal of Sports Sciences, 33*(18), 1932-1941.

Sumpter, D. (2016). *Soccermatics: Mathematical Adventures in the Beautiful Game.* England: Bloomsbury Sigma.

Sutter, M., & Kocher, M. G. (2004). Favoritism of agents–the case of referees' home bias. *Journal of Economic Psychology, 25*(4), 461-469.

Thomas, S., Reeves, C., & Smith, A. (2006). English soccer teams' aggressive behavior when playing away from home. *Perceptual and Motor Skills, 102*(2), 317-320.

Tucker, W., Mellalieu, D. S., James, N., & Taylor, B. J. (2005). Game location effects in professional soccer: A case study. *International Journal of Performance Analysis in Sport, 5*(2), 23-35.

UEFA. (2024, February 19). *Country coefficients*. Retrieved February 19, 2024, from https://www.uefa.com/memberassociations/uefarankings/country/#/yr/2024.

Waters, A., & Lovell, G. (2002). An examination of the homefield advantage in a professional English soccer team from a psychological standpoint. *Football Studies, 5*(1), 46-59.

Wolfson, S., Wakelin, D., & Lewis, M. (2005). Football supporters' perceptions of their role in the home advantage. *Journal of Sports Sciences, 23*(4), 365-374.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 3-36.

Wunderlich, F., & Memmert, D. (2018). The betting odds rating system: Using soccer forecasts to forecast soccer. *PloS one, 13*(6), e0198668.

Wunderlich, F., Weigelt, M., Rein, R., & Memmert, D. (2021). How does spectator presence affect football? Home advantage remains in European top-class football matches played without spectators during the COVID-19 pandemic. *Plos one, 16*(3), e0248590.

Appendix

R-script of the thesis:

```
# Data for Master's thesis
setwd("C:/Users/Bendeguz/Desktop/Genova/Master's thesis/Data")
getwd()

# Packages
library(tidyverse)
library(psych)
library(stringr)
library(dplyr)
library(ggplot2)
library(nnet)
library(stargazer)
library(car)

# Analysis - Part 1 - "Changes in home advantage over time"
# Part 1.1 - Distribution and trends of match results
# England - Premier League
# Data Manipulation
getwd()
setwd("C:/Users/Bendeguz/Desktop/Genova/Master's thesis/Data/England_PremierLeague")
getwd()

PremierLeague <-
  list.files(pattern = "*.csv") %>%
  map_df(~read_csv(.))

PremierLeague <- PremierLeague[,1:26]

# check for NAs
which(is.na(PremierLeague))

# remove NAs (only one line => probably 1 added blank line by mistake)
PremierLeague <- PremierLeague[-which(is.na(PremierLeague)),]

PremierLeague <- PremierLeague %>%
  mutate(Month = str_sub(Date, 4, 5)) %>%
  mutate(Year = str_sub(Date, start = -2))
PremierLeague$Month <- as.numeric(PremierLeague$Month)
PremierLeague$Year <- as.numeric(PremierLeague$Year)
PremierLeague <- PremierLeague %>%
  mutate(Season = case_when(Month >= 8 ~ Year,
                            Month < 8 ~ Year - 1) + 2000) %>%
  group_by(Season) %>%
  mutate(homewins = sum(FTR == "H")) %>%
  mutate(draws = sum(FTR == "D")) %>%
  mutate(awaywins = sum(FTR == "A")) %>%
  mutate(ngames = homewins + draws + awaywins) %>%
  mutate(homeratio = homewins / ngames) %>%
  mutate(drawratio = draws / ngames) %>%
  mutate(awayratio = awaywins / ngames)

PremierLeague_results <- PremierLeague %>%
  select(Season, homewins, draws, awaywins, ngames,
         homeratio, drawratio, awayratio) %>%
  distinct()

load("C:/Users/Bendeguz/Desktop/Genova/Master's thesis/Data/england.rda")

PremierLeague_old <- england %>%
  filter(tier == 1, division == 1) %>%
  group_by(Season) %>%
  mutate(homewins = sum(result == "H")) %>%
  mutate(draws = sum(result == "D")) %>%
  mutate(awaywins = sum(result == "A")) %>%
  mutate(ngames = homewins + draws + awaywins) %>%
  mutate(homeratio = homewins / ngames) %>%
  mutate(drawratio = draws / ngames) %>%
  mutate(awayratio = awaywins / ngames)

PremierLeague_old_results <- PremierLeague_old %>%
  select(Season, homewins, draws, awaywins, ngames,
         homeratio, drawratio, awayratio) %>%
```

```
  distinct()

PremierLeague_results_plot <- full_join(PremierLeague_old_results, PremierLeague_results)

# plot
ggplot(data = PremierLeague_results_plot, aes(x = Season)) +
  geom_point(aes(y = homeratio), color = "purple") +
  geom_point(aes(y = drawratio), color = "darkgreen") +
  geom_point(aes(y = awayratio), color = "darkorange") +
  ylab("Relative Frequency") +
  xlab("Season") +
  scale_x_continuous(limits = c(1885, 2025), breaks = seq(1885, 2025, by = 20)) +
  geom_smooth(aes(y = homeratio), method = "gam", se = TRUE, color = "purple") +
  geom_smooth(aes(y = drawratio), method = "gam", se = TRUE, color = "darkgreen") +
  geom_smooth(aes(y = awayratio), method = "gam", se = TRUE, color ="darkorange") +
  ggtitle("Premier League - Distribution of the Match Results by Season",
          subtitle = "Legend: purple = home win, green = draw, orange = away win")

remove(PremierLeague_old_results, PremierLeague_results, PremierLeague_results_plot, england)

# Part 1.2 - Distribution and trends of the average number of goals per game
# England - Premier League
# Data Manipulation
PremierLeague <- PremierLeague %>%
  group_by(Season) %>%
  mutate(hgoals = sum(FTHG)) %>%
  mutate(vgoals = sum(FTAG)) %>%
  mutate(homegoals = hgoals / ngames) %>%
  mutate(awaygoals = vgoals / ngames)

PremierLeague_goals <- PremierLeague %>%
  select(Season, ngames, hgoals, vgoals, homegoals, awaygoals) %>%
  distinct()

PremierLeague_old <- PremierLeague_old %>%
  group_by(Season) %>%
  mutate(hgoals = sum(hgoal)) %>%
  mutate(vgoals = sum(vgoal)) %>%
  mutate(homegoals = hgoals / ngames) %>%
  mutate(awaygoals = vgoals / ngames)

PremierLeague_old_goals <- PremierLeague_old %>%
  select(Season, ngames, hgoals, vgoals, homegoals, awaygoals) %>%
  distinct()

PremierLeague_goals_plot <- full_join(PremierLeague_old_goals, PremierLeague_goals)

# plot
ggplot(data = PremierLeague_goals_plot, aes(x = Season)) +
  geom_point(aes(y = homegoals), color = "purple") +
  geom_point(aes(y = awaygoals), color = "darkorange") +
  ylab("Average number of goals per game") +
  xlab("Season") +
  scale_x_continuous(limits = c(1885, 2025), breaks = seq(1885, 2025, by = 20)) +
  geom_smooth(aes(y = homegoals), method = "gam", se = TRUE, color = "purple") +
  geom_smooth(aes(y = awaygoals), method = "gam", se = TRUE, color ="darkorange") +
  ggtitle("Premier League - Goals Scored by the Home and the Away Team by Season",
          subtitle = "Legend: purple = home team, orange = away team")

remove(PremierLeague_old_goals, PremierLeague_goals, PremierLeague_goals_plot,
PremierLeague_old)

# Analysis - Part 2 - "Multinomial model for match outcome, comparison with betting odds"
# Original data with match statistics
# England - Premier League: 04/05-22/23
PremierLeague <- PremierLeague[, 1:26]
PremierLeague <- PremierLeague[, -c(1:6,8:9,11)]
str(PremierLeague)

# set the match outcomes variables as factors
# levels: away win, draw, home win in that order, i.e. reference level: away win
PremierLeague$FTR <- as.factor(PremierLeague$FTR)
PremierLeague$HTR <- as.factor(PremierLeague$HTR)
levels(PremierLeague$FTR)
levels(PremierLeague$HTR)

# the red cards have few distinct values values: not real numeric variables => set as factors
```

```
# levels: 0-1-2-3 and 0-1-2 respectively, i.e. reference level: 0 red cards
unique(PremierLeague$HR)
unique(PremierLeague$AR)
PremierLeague$HR <- as.factor(PremierLeague$HR)
PremierLeague$AR <- as.factor(PremierLeague$AR)
levels(PremierLeague$HR)
levels(PremierLeague$AR)

# Descriptive statistics:
# inspect the database, the distributions etc.
str(PremierLeague)
describe(PremierLeague[, -c(1:2,13:17)])
summary(PremierLeague)
boxplot(PremierLeague[, c(3:12)],
        main = "Distribution of the Numeric Explanatory Variables - Box Plots",
        xlab = "Numeric Explanatory Variables",
        ylab = "Value",
        col = "lightblue")

# right-skewed distributions, but not we can work with them
# logarithmization here does not make any sense
# Half Time Result (HTR) and betting odds will be analysed separately as well.
# FTR is the dependent variable

# Relationship between the explanatory variables and the dependent variable:
# 1. Full Time Result and Half Time Result (2 categorical variables - association):
# H0: independence
xtabs(~ FTR + HTR, data = PremierLeague)
chisq.test(xtabs(~ FTR + HTR, data = PremierLeague))

# H0 is rejected. FTR clearly depends on HTR
# Let's illustrate this relationship:
ggplot(data = PremierLeague, aes(x = HTR, fill = FTR)) +
  geom_bar(position = "fill") +
  ggtitle("Premier League - Relationship between the Half Time Result and the FTR",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("Relative Frequency of FTR") +
  xlab("HTR - Half Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# 2. Full Time Result and Home Team Shots (1 categorical and 1 numeric variable: Kruskal test)
# H0: independence
kruskal.test(FTR ~ HS, data = PremierLeague)

# H0 is rejected. FTR clearly depends on HS.
# Let's illustrate this relationship:
ggplot(PremierLeague, aes(x = FTR, y = HS, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Home Team Shots and the FTR",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("HS - Home Team Shots") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# Same process for the other numeric variables
# 3. Full Time Result and Away Team Shots
kruskal.test(FTR ~ AS, data = PremierLeague)

# H0 is rejected. FTR clearly depends on AS.
# Let's illustrate this relationship:
ggplot(PremierLeague, aes(x = FTR, y = AS, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Away Team Shots and the FTR,
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("AS - Away Team Shots") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# 4. Full Time Result and Home Team Shots on Target
kruskal.test(FTR ~ HST, data = PremierLeague)

# H0 is rejected. FTR clearly depends on HST.
# Let's illustrate this relationship:
ggplot(PremierLeague, aes(x = FTR, y = HST, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Home Team Shots on Target and the FTR",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
```

```
  ylab("HST - Home Team Shots on Target") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# 5. Full Time Result and Away Team Shots on Target
kruskal.test(FTR ~ AST, data = PremierLeague)

# H0 is rejected. FTR clearly depends on AST.
# Let's illustrate this relationship:
ggplot(PremierLeague, aes(x = FTR, y = AST, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Away Team Shots on Target and the Full
Time Result ",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("AST - Away Team Shots on Target") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# 6. Full Time Result and Home Team Fouls Committed
kruskal.test(FTR ~ HF, data = PremierLeague)

# H0 cannot be rejected. FTR clearly does not depend on HF
# Let's illustrate this independence:
ggplot(PremierLeague, aes(x = FTR, y = HF, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Home Team Fouls Committed
          and the Full Time Result",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("HF - Home Team Fouls Committed") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# we can see that the medians are indeed the same across the three possible match outcomes
# 7. Full Time Result and Away Team Fouls Committed
kruskal.test(FTR ~ AF, data = PremierLeague)

# H0 cannot be rejected. FTR clearly does not depend on AF
# Let's illustrate this independence:
ggplot(PremierLeague, aes(x = FTR, y = AF, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Away Team Fouls Committed and the Full
Time Result ",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("AF - Away Team Fouls Committed") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# we can see that the medians are indeed the same across the three possible match outcomes
# 8. Full Time Result and Home Team Corners
kruskal.test(FTR ~ HC, data = PremierLeague)

# H0 is rejected. FTR clearly depends on HC.
# Let's illustrate this relationship:
ggplot(PremierLeague, aes(x = FTR, y = HC, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Home Team Corners and the Full Time
Result ",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("HC - Home Team Corners") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# not that strong of a relationship as for example the shots (on target)
# 9. Full Time Result and Away Team Corners
kruskal.test(FTR ~ AC, data = PremierLeague)

# H0 is rejected. FTR clearly depends on AC.
# Let's illustrate this relationship:
ggplot(PremierLeague, aes(x = FTR, y = AC, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Away Team Corners and the Full Time
Result ",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("AC - Away Team Corners") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))
```

```
# Same. Not that strong of a relationship as for example the shots (on target)
# 10. Full Time Result and Home Team Yellow Cards
kruskal.test(FTR ~ HY, data = PremierLeague)

# H0 is rejected. FTR clearly depends on HY.
# Let's illustrate this relationship:
ggplot(PremierLeague, aes(x = FTR, y = HY, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Home Team Yellow Cards and the Full Time
Result ",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("HY - Home Team Yellow Cards") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))
# Not really interpretable as yellow cards are close to being a categorical variable.
# 11. Full Time Result and Away Team Yellow Cards
kruskal.test(FTR ~ AY, data = PremierLeague)

# H0 cannot be rejected. FTR clearly does not depend on AY.
# Let's illustrate this independence:
ggplot(PremierLeague, aes(x = FTR, y = AY, fill = FTR)) +
  geom_boxplot() +
  ggtitle("Premier League - Relationship between the Away Team Yellow Cards and the Full Time
Result ",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("AY - Away Team Yellow Cards") +
  xlab("FTR - Full Time Result") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# we can see that the medians are indeed the same across the three possible match outcomes
# 12. Full Time Result and Home Team Red Cards (2 categorical variables - association):
# H0: independence
xtabs(~ FTR + HR, data = PremierLeague)
chisq.test(xtabs(~ FTR + HR, data = PremierLeague))

# H0 is rejected. FTR clearly depends on HR
# Let's illustrate this relationship:
ggplot(data = PremierLeague, aes(x = HR, fill = FTR)) +
  geom_bar(position = "fill") +
  ggtitle("Premier League - Relationship between the Home Team Red Cards and the Full Time
Result ",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("Relative Frequency of FTR") +
  xlab("HR - Home Team Red Cards") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))

# 13. Full Time Result and Away Team Red Cards (2 categorical variables - association):
# H0: independence
xtabs(~ FTR + AR, data = PremierLeague)
chisq.test(xtabs(~ FTR + AR, data = PremierLeague))

# H0 is rejected. FTR clearly depends on AR
# Let's illustrate this relationship:
ggplot(data = PremierLeague, aes(x = AR, fill = FTR)) +
  geom_bar(position = "fill") +
  ggtitle("Premier League - Relationship between the Away Team Red Cards and the Full Time
Result ",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("Relative Frequency of FTR") +
  xlab("AR - Away Team Red Cards") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))
# summary: when looking at the variables individually, FTR does not depend on HF, AF, AY
# Multinomial logistic regression - model building, model selection and prediction
model0 = multinom(FTR ~ HTR + HS + AS + HST + AST + HF + AF + HC + AC + HY + AY + HR + AR,
                  data = PremierLeague)
summary(model0)
# backward stepwise model selection with AIC and with BIC:
mod.bA = step(model0)
summary(mod.bA)
n = dim(PremierLeague)[1]
mod.bB = step(model0, k = log(n))
summary(mod.bB)
# they result in the same model which will be my final multinomial logit model
model1 = multinom(FTR ~ HTR + HST + AST + AF + HC + AC + HY + AY + HR + AR,
                  data = PremierLeague)
summary(model1)
```

```
# Computation of the Wald-tests (z-tests) and the p-values
z = summary(model1)$coefficients / summary(model1)$standard.errors
p = (1 - pnorm(abs(z), 0, 1)) * 2
# predicted match outcome probabilities and predicted match outcomes
pred <- predict(model0, PremierLeague, type = "probs")
pred_cat <- predict(model0, PremierLeague)
table(PremierLeague$FTR, pred_cat)
accuracy0 <- mean(PremierLeague$FTR == pred_cat)
pred <- predict(model1, PremierLeague, type = "probs")
pred_cat <- predict(model1, PremierLeague)
table(PremierLeague$FTR, pred_cat)
accuracy1 <- mean(PremierLeague$FTR == pred_cat)
c(accuracy0, accuracy1)
AIC(model0, model1)
BIC(model0, model1)
stargazer(model0, model1, type = "text")

# plot of the predicted probabilities for the most interesting variable: HST
HST.d = seq(0, 20, 1)
beta0.1 = summary(model1)$coefficients[1, 1]
beta4.1 = summary(model1)$coefficients[1, 4]
beta0.2 = summary(model1)$coefficients[2, 1]
beta4.2 = summary(model1)$coefficients[2, 4]
p1 = exp(beta0.1 + beta4.1 * HST.d) / (1 + exp(beta0.1 + beta4.1 * HST.d) + exp(beta0.2 +
beta4.2 * HST.d))
p2 = exp(beta0.2 + beta4.2 * HST.d) / (1 + exp(beta0.1 + beta4.1 * HST.d) + exp(beta0.2 +
beta4.2 * HST.d))
p3 = 1 - p1 - p2
plot(HST.d, p1, type = "l", col = "darkgreen", ylim = c(0, 1), ylab = "Predicted
Probabilities",
     main = "Predicted Full Time Result Probabilities as a Function of HST",
     sub = "Legend: purple = home win, green = draw, orange = away win",
     xlab = "HST - Home Team Shots on Target", lwd = 1.5)
lines(HST.d, p2, col = "purple", lwd = 1.5)
lines(HST.d, p3, col = "darkorange", lwd = 1.5)
grid(nx = NULL, ny = NULL, lty = 2, col = "gray", lwd = 1)
# breaking point is almost exactly the mean of HST when FTR = H
meanHST <- PremierLeague %>%
  filter(FTR=="H")
mean(meanHST$HST)
# Comparison with the prediction accuracy of Half Time Result alone:
accuracy2 <- mean(PremierLeague$FTR == PremierLeague$HTR)
# Comparison with the prediction accuracy of betting odds:
describe(PremierLeague[, 15:17])
summary(PremierLeague[, 15:17])
boxplot(PremierLeague[, 15:17],
        main = "Distribution of the Betting Odds - Box Plots",
        xlab = "Bet365 Betting Odds",
        ylab = "Value",
        col = "lightblue")
PremierLeague <- PremierLeague %>%
  mutate(prob_B365H = 1 / B365H) %>%
  mutate(prob_B365D = 1 / B365D) %>%
  mutate(prob_B365A = 1 / B365A)
boxplot(PremierLeague[, 18:20],
        main = "Distribution of the Match Outcome Probabilities based on Betting Odds”
        xlab = "B365 Betting Odds",
        ylab = "Probabilities",
        col = "lightblue")
# Forecast with betting odds
PremierLeague <- PremierLeague %>%
  mutate(pred_result = case_when(prob_B365H > prob_B365D & prob_B365H > prob_B365A ~ "H",
                                 prob_B365A > prob_B365D & prob_B365A > prob_B365H ~ "A",
                                 prob_B365A == prob_B365H ~ "D"))
# Full Time Result and Predicted Results (2 categorical variables - association):
xtabs(~ FTR + pred_result, data = PremierLeague)
chisq.test(xtabs(~ FTR + pred_result, data = PremierLeague))
accuracy3 <- mean(PremierLeague$FTR == PremierLeague$pred_result)
# H0 is rejected. FTR clearly depends on the Predicted Results based on betting odds
# Let's illustrate this relationship:
ggplot(data = PremierLeague, aes(x = pred_result, fill = FTR)) +
  geom_bar(position = "fill") +
  ggtitle("Premier League - Relationship between the Actual and the Predicted Results",
          subtitle = "Legend: A = away win, D = draw, H = home win") +
  ylab("Relative Frequency of FTR") +
  xlab("Predicted Result Based on Bet365 Betting Odds ") +
  scale_fill_manual("FTR", values = c("A" = "darkorange", "D" = "darkgreen", "H" = "purple"))
```