Competitiveness of the English Premier League

Geetanjali Deore (220315825)

Supervisor: Mark Butler

Aston University

Master of Science (Data Analytics)

September 2023

Abstract

The English Premier League (EPL) is one of the most popular football leagues globally, renowned for its competitiveness and unpredictable match outcomes. This dissertation analyses the competitive balance of EPL over the past 20 seasons using various statistical measures. The metrics used include the normalized standard deviation, concentration ratios, Herfindahl-Hirschman Index and Gini indices, Lorenz curves, and relative entropy. The data covers match statistics and team performance metrics across 7,980 games played between 2002-2003 and 2022-2023. The analyses revealed fluctuations in the league's competitiveness over time, with the 2010-2011 season emerging as the most balanced and the 2018-2019 season as the least balanced overall. While the normalized standard deviation and concentration ratios indicate declining competitiveness recently, other metrics, such as the Gini coefficient, point to a greater balance. The relative entropy measure showed consistent unpredictability in match outcomes over the observed period. Linear regression forecasts project a mix of improving and stable competitiveness based on past trends. This dissertation provides a rigorous quantitative examination of the dynamics of competitive balance in EPL. Multiple metrics offer a robust, multidimensional perspective on league evolution. These insights can inform critical decisions regarding broadcasting rights, player transfers, and league policies aimed at sustaining competitive equilibrium, spectator interest, and commercial growth.

Keywords: Competitive Balance, Normalized Standard Deviation, Concentration Ratio, Herfindahl-Hirschman, Gini Coefficient, Lorenz Curve, Relative Entropy

Acknowledgement

I want to express my deepest gratitude to my supervisor, Dr. Mark Butler, for his guidance and support throughout this research project. His vast expertise, endless patience, and invaluable mentorship have made it possible to complete this dissertation. I am incredibly grateful for all his insights and direction.

I also sincerely appreciate Aston University for providing me with the opportunity and resources to undertake this research. The dedication and assistance of the entire staff at Aston has enriched my research experience.

Most importantly, my parents deserve immense recognition. Their constant love and belief in me have been the bedrock guiding me through my academic journey. I also want to give special mention to my friends and colleagues, whose continuous encouragement, companionship, and feedback greatly enhanced this process.

The successful completion of this dissertation is due to its many contributions. I am eternally grateful to everyone who supported and believed in me during this rewarding research experience.

Contents

List of Figures

List of Tables

List of Abbreviations

HAGCR	Home Average Goal Conceded Ratio
AAGCR	Away Average Goal Conceded Ratio
NAGCR	Net Average Goal Conceded Ratio
NAMSI	National Measure of Season Imbalance
C5ICB	Index of Competitive Balance
HAGR	Home Average Goal Ratio
AAGR	Away Average Goal Ratio
NAGR	Net Average Goal Ratio
HICB	Herfindahl Index of Competitive Balance
HHI	Herfindahl-Hirschman Index
HHI*	Normalised Herfindahl-Hirschman Index
WPCT	Win Percentage
CBR	Competitive Balance Ratio
EPL	English Premier League
GDP	Gross Domestic Product
CSV	Comma Separated Values
NSD	Normalised Standard Deviation
SD	Standard Deviation
RE	Relative Entropy
C5	Five Club Concentration Ratio

Chapter 1

Introduction

1.1 Background of English Premier League

The English Premier League (EPL), since its inauguration in 1992, has firmly established itself as a focus in global football. It carries both socio-cultural significance and notable economic impact. Emerging from the framework of the old English First Division, the EPL was envisioned as a breakaway league, aiming to harness and elevate the English football experience. The EPL is composed of 20 elite teams that compete for glory over 38 matches each season. The league dynamics incorporate both promotion and relegation, closely intertwined with the lower tiers of the English Football League (EFL). Historically, clubs such as Manchester United, Arsenal, Chelsea, and Liverpool have been dominant forces. However, recent decades have witnessed a meteoric rise in teams such as Manchester City and the unexpected triumphs of teams such as Leicester City during the 2015-16 season.

The league's universal appeal has been amplified by the presence of iconic players from various parts of the world, each bringing a unique flair and dynamism. The EPL's history is marked by phases of dominance: the early 2000s relished the intense rivalry between Arsenal's 'Invincibles' and Sir Alex Ferguson's Manchester United. Subsequent years saw Chelsea under JosÃl' Mourinho and Manchester City under Pep Guardiola setting new standards and benchmarks in the league. The financial stature of the EPL has grown consistently, making it one of the most lucrative football leagues in the world.

1.2 Concept of competitiveness in sport industry

Competitive balance is vital for maintaining and strengthening fans' excitement during sports events. The element of unpredictability in match results is the foundation of this balance. The increased uncertainty in outcomes encourages greater fan engagement, which is evident in ticket sales and broadcast viewership numbers. Neale (1964) appropriately described this as the "League Standing Effect". A noticeable decline in competitive balance could lead to diminishing fan interest, initially in underperforming teams, and eventually spreading even to top-performing teams. The demise of the All American Football Conference between 1946 and 1949 is a testament to this, a downfall attributed to its lack of competitive balance by Fort and Quirk (1997).

The widespread appeal and acclaim of the Premier League is largely down to its competitive balance - how evenly matched are the teams in terms of performance. Competitiveness plays a vital role in the league's ongoing success. Rottenberg (1956) was a pioneer in introducing the concept of competitive balance, emphasising its importance in the economic analysis of professional sport. At its core, competitive balance examines the inequality of both match and championship results. It looks at the distribution of wins across teams in a single season, the consistency of team performance over consecutive seasons, and the accumulation of championship victories over longer periods, as detailed by Kringstad and Gerrard (2007). Owen, Ryan, and Weatherston (2007a) further explained the balance of competition as a comprehensive lens to assess the equality or disparity in the strength of competing teams.

While the relative standard deviation of win percentages is a commonly used metric for competitive balance, it's important to recognise its inherent limitations, especially in football. For instance, football leagues, unlike major US sports, often have a significant proportion of matches ending in draws. This was evident in various leagues, including the Premier League, during the 2018-19 season. Additionally, such metrics may not adequately represent the dominance dynamics within a league. To overcome these challenges, alternative methodologies based on standard economic theories, such as the concentration ratio (Hall and Tideman, 1967), have been advocated.

Zimbalist (2002) remarked on the multitude of methods available to assess competitive balance, likening it to the challenges of quantifying economic measures, such as money supply. It's paramount to differentiate between metrics assessing the relative strength of teams in a specific season and those evaluating the sustained dominance of certain teams over multiple seasons. This research offers an in-depth exploration of the Premier League's competitive dynamics, using a range of statistical methodologies to generate insights that can inform key decisions, from broadcasting negotiations to player recruitment strategies.

Chapter 2

Literature Review

The Premier League's worldwide popularity and financial success are clear to see. This review aims to summarize important insights from many studies about how evenly matched teams are in sports leagues, particularly in football.

2.1 Economic Impact and Significance of the Premier League

The Premier League's financial robustness is a testament to its dominant position in the global football landscape. According to a report by Deloitte (2023), the league achieved a remarkable milestone with revenues hitting Âč5.5 billion for the 2021/22 season, reflecting a 12% growth from the preceding year. This surge was predominantly fueled by record matchday and commercial earnings, marking the league's supremacy among Europe's elite football leagues.

However, a deeper look reveals a concentration of this wealth. The 'big six' clubs, which include Manchester United, Manchester City, Liverpool, Chelsea, Arsenal, and Tottenham, have significantly influenced these figures. Collectively, they were responsible for 66% of the total matchday income and 76% of the commercial earnings. Despite their financial stronghold, these clubs have shown varied performances on the field, presenting opportunities for other teams in the league to challenge their dominance. This competitive dynamic could further evolve with the introduction of new and more lucrative broadcast deals (Young, 2022).

On the flip side, there are challenges to consider when discussing the league's competitiveness. Clubs in the Premier League reported an all-time high wage bill of Âč3.6 billion in 2021/22. Interestingly, there is a pronounced link between the amount a club spends on wages and league position, suggesting that financial power plays a pivotal role in competitive outcomes. Moreover, despite its revenue success, the league has not been immune to financial strains, reporting pre-tax losses (Âč607 million) for the fourth consecutive year and accumulating a net debt of Âč2.7 billion. While the Premier League's investment in player transfers outstrips its

European counterparts, underpinned by its global appeal and attractive broadcasting agreements (Deloitte, 2023), the prevailing financial dominance of the ’big six’ clubs poses questions about the sustainability of competitive balance in the future.

2.2 Impact of Competitive Balance on Fan Engagement: A Detailed Review

The relationship between competitive balance and fan engagement, particularly fan attendance, has been a focal point of numerous academic investigations. This review seeks to elaborate some of the most notable findings in this field. Manasis, Ntzoufras, and Reade (2015) delved into the nuances of league dynamics and its implications for fan engagement. Their research placed significant emphasis on the concept of "ranking mobility" across different seasons. Their findings suggest that fluctuations in team standings from one season to the next play a crucial role in influencing fan attendance patterns.

Similarly, Brandes and Franck (2007) embarked on a comprehensive exploration of the European soccer leagues. Their objective was to uncover potential correlations between the equilibrium of team performances and fan turnout. Interestingly, while their research did identify some associations, it also suggested the possibility of other external factors influencing fan attendance, suggesting that competitive balance might not be the sole determinant. By shifting the focus to the French football league, Ligue 1, Andreff and Scelles (2015) made a striking observation. Their analysis revealed that fan attendance witnessed a noticeable surge during seasons characterized by closely contested competitions. This indicates that fans are more likely to flock to stadiums when the outcome of the league is uncertain and more teams are vying for the top spot.

Further reinforcing this narrative, two separate studies by Humphreys (2002) and Plumley, Ramchandani, and Wilson (2018) converged on a similar conclusion. Both studies underscored a tangible relationship between how evenly matched teams are (competitive balance) and the sustained interest of fans. Their findings indicate that maintaining a level playing field, where no single team dominates consistently, is pivotal for retaining and nurturing fan interest over the long term.

2.3 Metrics and Trends in Competitive Balance

The sports industry has witnessed a plethora of metrics being introduced to better understand and evaluate league dynamics. Alwell (2020), in their study, opted for the Gini coefficient as a tool to dissect the NBA’s competitive structure. The findings of this study highlight the potential benefits of a more balanced league in terms of fan engagement and revenue optimization. On a similar note, Inan (2018) turned to the C5CBI as a means to delve into the competitive intricacies of the Turkish Football Super League, shedding light on the dominance ranges exhibited by the league's top-tier clubs.

Deb (2021) broke the new ground by presenting a unique mathematical framework. Their research suggested that, under certain conditions, matches in premier European leagues could essentially result in equal outcomes for the competing teams. In a quest to understand dynamic competitive balance, DâĂŹOttaviano (2019) introduced the 'k index', setting it in contrast with the more static Herfindahl-Hirschman Index. Similarly, Michie and Oughton (2004) embarked on an analytical journey that focused on top-flight English football. Their approach was multifaceted, with the Herfindahl Index being one of the many tools used. Furthermore, Horowitz (1997) utilized the relative-entropy measure to trace the historical dynamics of Major League Baseball. In a bid to evolve traditional methodologies, Triguero-Ruiz and Cano (2018) proposed a refined metric, the HHI NORM, tailored for European soccer leagues.

The domain of competitive balance has also witnessed a surge in innovative methodologies aimed at refining our understanding. Nikolakaki et al. (2020) championed an efficient linear model, which they posited could mirror the performance outcomes of more intricate neural networks. Hall and Tideman (1967) contrasted a novel metric against established metrics, such as the concentration ratio and the Herfindahl index. Similarly, Manasis et al. (2013) introduced the Special Concentration Ratio (SCR I K) specifically tailored for European football dynamics. Evans (2014), meanwhile, undertook the mammoth task of consolidating various measures highlighted in existing literature, effectively differentiating between short-term team performance and long-term dominance.

The landscape of sports leagues is evolving and various studies have aimed to identify and interpret these changes. Ramchandani et al. (2018) undertook a comprehensive analysis, and while they found survival competitions in European leagues remaining largely consistent, there was a discernible decline in title competitions in certain leagues. Basini et al. (2023) pinpointed a transformative shift in the English Premier League, observing its metamorphosis into a two-tier league system around the dawn of the 21st century. Interestingly, Szymanski (2010) indicated that even with growing financial disparities among clubs, the core competitive balance in English soccer remained relatively stable. The extensive literature on the subject presents a holistic view of the significance, methodologies, impacts, and policy nuances related to competitive balance in Premier Leagues and other sporting events. Understanding these dynamics is instrumental in preserving the global appeal and financial viability of these leagues in the ever-evolving sports landscape.

2.4 Recommendations for Policies and Strategies

The quest for a balanced competitive environment in sports leagues also extends to policy recommendations. Fort and Quirk (1995) highlight the potential merits of enforceable salary caps in maintaining an equitable competitive landscape. Echoing the need for regulations, Plumley et al. (2022) emphasized the introduction of financial regulations to ensure fairness. Beck, Prinz, and Van Der Burg (2022) advocated for a more equitable distribution of broadcasting revenue coupled with progressive luxury taxes to level the playing field. Szymanski (2003) rounded off the discourse by emphasizing the paramount importance of integrating economic perspectives into the design and oversight of sports leagues.

Chapter 3

Problem Description

3.1 Objective

The primary aim of this dissertation is to analyse the competitive balance of English Premier League teams spanning the last 20 seasons. To achieve this, the study has been structured around three core objectives:

1. *Review and Selection of Measurement Methods:* Dive into the extensive literature on competitive balance to discern the most appropriate methods for gauging the EPL's competitiveness.
2. *Data Analysis of Past Two Decades:* Analyse data from the past 20 EPL seasons to determine shifts in competitive balance. This analysis will consider factors like league statistics, team performance metrics.
3. *Predicting Future Trends:* Based on the patterns observed in the last 20 seasons, this study attempts to forecast the league's competitive balance in the forthcoming seasons.

3.2 Structure of work

To perform the analysis and evaluate the trend following structure and path is followed:

- Collecting comprehensive data of the past 20 EPL seasons, including team performance metrics, and any other relevant datasets.
- Working on the literature on competitive balance in sports leagues for method selection.
- Selection of suitable statistical measures or methods for conducting detailed analyses and predicting the future trends.

Overall, this study's findings will significantly contribute to academics of competitive balance and help students, researchers, policymakers, and stakeholders in the English Premier League with actionable insights.

3.3 Ethical Considerations and Limitations

In this study, the data used are openly available to the public, focusing on the last 20 seasons of the English Premier League (EPL). It is important to note that if there is any missing or incomplete data from these seasons, it could affect the accuracy of our results. Predicting future trends in the EPL can also be tricky because unexpected events, like economic downturns or pandemics, can change things. With these challenges in mind, the research is carried out and apply methods in a thoughtful and thorough way, while also being aware of the study's limitations.

Chapter 4

Data and Methodology

4.1 Data Collection and Feature Extraction

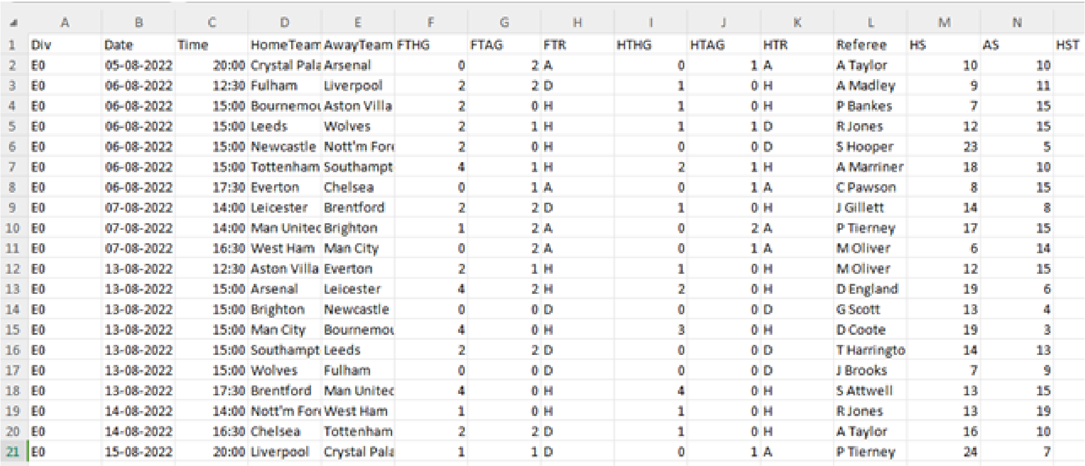
4.1.1 Data Collection and Preprocessing

The dataset utilized for this research project is a consolidated compilation sourced from Football-data.co.uk, comprising 20 individual files which is later merged for further analysis. The dataset is housed in a single CSV (Comma Separated Values) file, making it convenient for data manipulation and analysis within an Anaconda environment, compatible with packages such as TensorFlow and Pandas. The dataset contains a total of 7,980 rows (matches) and 23 columns (features or metrics associated with each match). The dataset has undergone rigorous data cleaning procedures to ensure its reliability and accuracy for analysis. This includes the removal of duplicate entries, as well as the handling of missing and null values to maintain the dataset's integrity. One of the primary motivations for selecting this specific dataset is its balanced size, offering a comprehensive range of data points that are neither too scant to undermine the model's accuracy nor too extensive to demand excessive computational resources. The dataset encapsulates various metrics related to the performance of teams in the English Premier League (EPL) over a span of 20 seasons.

Below are columns description table 4.1 and an example of dataset sample instance (4.1):

4.1.2 Feature Extraction

Feature extraction on this dataset involves deriving key metrics that encapsulate the offensive and defensive capabilities of football teams in the English Premier League. These metrics, such as the Home Average Goal Ratio (HAGR) and Away Average Goal Conceded Ratio (AAGCR), provide a more granular understanding of team performances across seasons, enhancing the depth and precision of subsequent analyses.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Div	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST
2	E0	05-08-2022	20:00	Crystal Pala	Arsenal	0	2	A	0	1	A	A Taylor	10	10	
3	E0	06-08-2022	12:30	Fulham	Liverpool	2	2	D	1	0	H	A Madley	9	11	
4	E0	06-08-2022	15:00	Bournemou	Aston Villa	2	0	H	1	0	H	P Bankes	7	15	
5	E0	06-08-2022	15:00	Leeds	Wolves	2	1	H	1	1	D	R Jones	12	15	
6	E0	06-08-2022	15:00	Newcastle	Nott'm For	2	0	H	0	0	D	S Hooper	23	5	
7	E0	06-08-2022	15:00	Tottenham	Southampt	4	1	H	2	1	H	A Marriner	18	10	
8	E0	06-08-2022	17:30	Everton	Chelsea	0	1	A	0	1	A	C Pawson	8	15	
9	E0	07-08-2022	14:00	Leicester	Brentford	2	2	D	1	0	H	J Gillett	14	8	
10	E0	07-08-2022	14:00	Man Unitec	Brighton	1	2	A	0	2	A	P Tierney	17	15	
11	E0	07-08-2022	16:30	West Ham	Man City	0	2	A	0	1	A	M Oliver	6	14	
12	E0	13-08-2022	12:30	Aston Villa	Everton	2	1	H	1	0	H	M Oliver	12	15	
13	E0	13-08-2022	15:00	Arsenal	Leicester	4	2	H	2	0	H	D England	19	6	
14	E0	13-08-2022	15:00	Brighton	Newcastle	0	0	D	0	0	D	G Scott	13	4	
15	E0	13-08-2022	15:00	Man City	Bournemou	4	0	H	3	0	H	D Coote	19	3	
16	E0	13-08-2022	15:00	Southampt	Leeds	2	2	D	0	0	D	T Harringto	14	13	
17	E0	13-08-2022	15:00	Wolves	Fulham	0	0	D	0	0	D	J Brooks	7	9	
18	E0	13-08-2022	17:30	Brentford	Man Unitec	4	0	H	4	0	H	S Attwell	13	15	
19	E0	14-08-2022	14:00	Nott'm For	West Ham	1	0	H	1	0	H	R Jones	13	19	
20	E0	14-08-2022	16:30	Chelsea	Tottenham	2	2	D	1	0	H	A Taylor	16	10	
21	E0	15-08-2022	20:00	Liverpool	Crystal Pala	1	1	D	0	1	A	P Tierney	24	7	

- **Home Average Goal Ratio (HAGR):** This metric represents the average goal-scoring capability of a team when playing at home, relative to the overall league's average. A value greater than 1 suggests that the team scores more than the league average when playing at home, indicating a strong home offensive performance.

$$\text{HAGR} = \frac{\text{Average Home Goals Scored by Team}}{\text{League's Average Home Goals}}$$

- **Away Average Goal Ratio (AAGR):** This metric depicts the average goal-scoring capability of a team when playing away, in relation to the overall leagueâĂŹs average. A value greater than 1 indicates that the team, when playing away, scores more than the league average, suggesting a potent away offensive performance.

$$\text{AAGR} = \frac{\text{Average Away Goals Scored by Team}}{\text{League's Average Home Goals}}$$

- **Home Average Goal Conceded Ratio (HAGCR):** This metric provides insights into the average goals conceded by a team when playing at home, compared to the league's average. A value greater than 1 implies that the team concedes more goals at home than the league average, indicating potential defensive vulnerabilities when playing at home.

$$\text{HAGCR} = \frac{\text{Average Home Goals Conceded by Team}}{\text{League's Average Away Goals}}$$

- **Away Average Goal Conceded Ratio (AAGCR):** This metric represents the average goals conceded by a team when playing away, in relation to the league's average. A value greater than 1 suggests that the team concedes more goals when playing away than the league average, indicating potential defensive weaknesses in away matches.

$$\text{AAGCR} = \frac{\text{Average Away Goals Conceded by Team}}{\text{League's Average Home Goals}}$$

Column Name	**Meaning**
Season	The football season during which the match took place.
Date	The date of the match.
HomeTeam	The team playing at home.
AwayTeam	The visiting team.
FTHG	Full Time Home Goals - The number of goals scored by the home team by full time.
FTAG	Full Time Away Goals - The number of goals scored by the away team by full time.
FTR	Full Time Result - The result of the match at full time (H: Home win, A: Away win, D: Draw).
HTHG	Half Time Home Goals - The number of goals scored by the home team by half time.
HTAG	Half Time Away Goals - The number of goals scored by the away team by half time.
HTR	Half Time Result - The result of the match at half time.
Referee	The referee of the match.
HS	Home Shots - Total shots by the home team.
AS	Away Shots - Total shots by the away team.
HST	Home Shots on Target - Shots on target by the home team.
AST	Away Shots on Target - Shots on target by the away team.
HF	Home Fouls - Total fouls by the home team.
AF	Away Fouls - Total fouls by the away team.
HC	Home Corners - Total corners taken by the home team.
AC	Away Corners - Total corners taken by the away team.
HY	Home Yellow Cards - Total yellow cards received by the home team.
AY	Away Yellow Cards - Total yellow cards received by the away team.
HR	Home Red Cards - Total red cards received by the home team.
AR	Away Red Cards - Total red cards received by the away team.

- **Net Average Goal Ratio (NAGR):** A team's overall offensive strength by averaging its Home and Away goal ratios is captured by this metric.

$$\text{NAGR} = \frac{\text{HAGR} \ + \ \text{AAGR}}{2}$$

- **Net Average Goal Conceded Ratio (NAGCR):** This metric gauges a team's overall defensive strength, calculated as the average of the team's Home and Away goal conceded ratios.

$$\text{NAGCR} = \frac{\text{HAGCR} \ + \ \text{AAGCR}}{2}$$

These features will be used in various measures such as standard deviation, HHI, and CBR to look for competitiveness trends across 20 seasons of EPL and look for future trends.

4.2 Competitiveness Measures and Analysis

4.2.1 Standard Deviation

Standard deviation measures how spread out the numbers in a set are. A small standard deviation means most numbers are close to the average, while a large one shows they're spread out. This measurement helps understand the consistency, risk, or differences in data. It's used in many areas like finance, economics, and especially in sports data analysis. In sports like football and baseball, standard deviation helps understand how evenly matched teams in a league are. Studies by Michie and Oughton (2004) highlighted its use in seeing the differences in how teams perform in a season. Similarly, Horowitz (1997) used it to study the changing competitiveness in Major League Baseball.

a) Formula for Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Where:

- N is the total number of observations.
- x_i denotes each individual value.
- μ is the average of the data set.

Central to sports analytics is the metric of winning percentages' dispersion, a tool extensively harnessed by eminent researchers like Scully (1989) and Fort and Quirk (1997) . If $WPCT_{i,t}$ is the winning percentage of team i in season t, the league's competitive balance is gauged using:

$$\sigma_L = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(WPCT_{i,t} - 0.500)^2}$$

The 0.500 in the formula denotes an ideal situation where all teams have an equal probability of winning. The standard deviation of the winning percentage, denoted as σ_L, provides a reliable indicator of competitive balance for a particular season. However, when this measure is applied over multiple seasons, certain challenges arise. For example, two leagues might have similar σ_L values, but their team rankings could differ significantly. One league might see a single team consistently leading, while the other might have changing top teams.

Eckard (1998), Eckard (2001a), and Eckard (2001b) advanced approach of variance decomposition was introduced to address these fluctuations between seasons. Yet, this approach also presents its own set of difficulties. Additionally, various other metrics, including the Hirfindahl-Hirschman Indexes, Lorenz Curves, and Gini Coefficients, have been employed to evaluate the level of competition in sports.

b) Idealized Standard Deviation Fort and Quirk (1997) proposed $\frac{0.5}{\sqrt{N}}$ as the epitome of standard deviation, where N encapsulates the total games in a season. This proposition is predicated on the belief that every team stands an equal chance of clinching a victory, set at 50%. However, this measure has not been immune to critique, especially when empirical calculations yield values sub-1, challenging the tenets of an "ideal" balance.

c) Within-Team Standard Deviation This metric gauges a specific team's performance variability across seasons. Formally, for a particular team, it's given by:

$$\sigma_{T,i} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(WPCT_{i,t} - \overline{WPCT_i}\right)^2}$$

Where $WPCT_{i,t}$ is the team's average winning percentage over the seasons.

d) Within-Season Standard Deviation This evaluates the spread of team performances within a single season. The formula is given as:

$$\sigma_{N,i} = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(WPCT_{i,t} - 0.500\right)^2}$$

Here, 0.500 is the expected winning percentage in a balanced league. While metrics like standard deviation offer invaluable insights into competitive balance, it's crucial to understand their limitations and the broader context to derive meaningful interpretations.

f) National Measure of Seasonal Imbalance (NAMSI) Goossens (2005) proposed a normalised measure for the standard deviation of win percentages called the National Measure of Seasonal Imbalance (NAMSI). This measure incorporates both the maximum and minimum possible standard deviation values to produce a measure which has a maximum value of one (indicating a league with the maximum competitive imbalance) and a minimum value of zero (indicating a league with perfect competitive balance), regardless of the number of teams in the league. The formula for NAMSI is:

$$NAMSI = \frac{\sqrt{\sum_{i=1}^{n}\left(w_i - 0.5\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(w_{i_{max}} - 0.5\right)^2}}$$

Where:

- w_i is the win percentage of team i.
- $w_{i_{max}}$ is the maximum value of win percentage of team i (with perfect imbalance).
- N is the number of teams i in the league.

g) Normalised standard deviation (σ_{L^*}) Owen (2010) provides an equivalent normalised measure with a more general expression for the upper bound. The measure σ_{L^*} is calculated as:

$$\sigma_{L^*} = \frac{\sigma_{L'}}{\sigma_{L_{ub}}}$$

Where:

- $\sigma_{L'}$ is the standard deviation of win percentage static for a single season and is given by:

$$\sigma_{L'} = \sqrt{\sum_{i=1}^{N} \left(\frac{W_i}{G_i} - 0.5 \right)^2}$$

- $\sigma_{L_{ub}}$ is the upper bound of σ_L and is given by:

$$\sigma_{L_{ub}} = \sqrt{\frac{N+1}{12\,(N-1)}}$$

- W_i is the number of wins of team i.
- G_i is the number of games of team i.
- N is the number of team i in the league.

The Actual Observed Standard Deviation (σ_L) is a statistical measure that quantifies the amount of variation or dispersion in a set of values. In the context of sports analytics, σ_L is often used to measure the competitive balance within a league for a particular season. A high σ_L indicates a large disparity between team performances, suggesting that some teams are significantly outperforming others. Conversely, a low ASD suggests that teams are performing at similar levels, indicating a more balanced competition.

Owen (2010) also demonstrates that the normalised standard deviation measure, when applied to either 'win percentage' or 'absolute points' data, produces identical results to the equivalent ratio of observed standard deviation to the 'idealised' standard deviation (σ_R) Measure, if that is also normalised with respect to its upper bound.

This measure provides a relative scale between 0 and 1. An σ_{L^*} value close to 1 suggests a league with a high degree of competitive imbalance, while a value close to 0 indicates a league with near-perfect competitive balance. The normalization process ensures that the measure is independent of the number of teams in the league, allowing for meaningful comparisons across different settings. When comparing σ_L and σ_{L^*}, it's evident that σ_{L^*} offers a more standardized measure of competitive balance, independent of the league's size or match count. Thus, σ_{L^*} is more suitable for comparative studies across different leagues, seasons, or even sports.

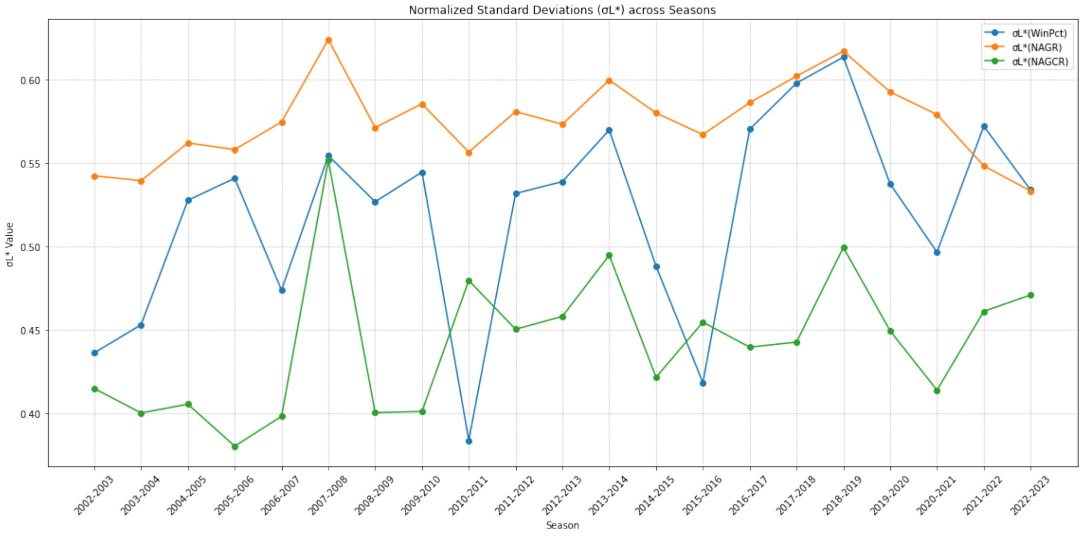
Implementation and Analysis

For every team across various seasons, the win percentage is deduced by analyzing both home and away match victories and goals scored and conceded, establishing it as a crucial metric to gauge a team's seasonal performance. Utilizing these win percentages, the standard deviation σ_L for each season is computed, serving as an immediate indicator of that season's competitive balance. To enhance this analysis, the standard deviation is normalized to its potential upper limit, resulting in σ_{L^*}. This metric, σ_{L^*}, offers a refined comparison of competitive balance: values close to 1 indicate a significant imbalance, while those nearing 0 depict a balanced competition.

The normalized standard deviation σ_{L^*} stands as a fundamental measure, historically based on the Win Percentage. Moving beyond conventional methods, this study adopts a comprehensive approach, considering not only the Win Percentage (σ_{L^*} for WPCT), but also incorporating two modern metrics: NAGR and NAGCR. These measures, meticulously analyzed across different seasons, are detailed in Table 4.2. Also, figure 4.2 provides a trend of the competitive balance dynamics in EPL leagues over the years.

The fluctuating σ_{L^*} values for Win Percentage reveal periods of both high and low competitive balance (Table 4.2). For instance, during the 2010-2011 season, the σ_{L^*} for WPCT was at a low of 0.383877, indicating a high degree of competitive balance. In contrast, the 2018-2019 season registered the highest σ_{L^*} for WPCT at 0.613625, signaling a season of competitive imbalance. These fluctuations confirm that the league's competitiveness is not static; it evolves over the years, reflecting various factors such as team performance, strategies, and external market conditions. For NAGR and NAGCR, the patterns in σ_{L^*} values are notably similar. The 2007-2008 season had the highest σ_{L^*} values for both NAGR (0.624184) and NAGCR (0.551739), indicating a season of high competitive imbalance. Interestingly, these metrics have shown a slight decline in recent years, suggesting a trend toward increased competitive balance.

Season	σ^*_L(WinPct)	σ^*_L(NAGR)	σ^*_L(NAGCR)
2002-2003	0.436292	0.542504	0.414942
2003-2004	0.453158	0.53957	0.4003
2004-2005	0.527781	0.562196	0.405531
2005-2006	0.541023	0.558235	0.380404
2006-2007	0.473894	0.57477	0.398208
2007-2008	0.554596	0.624184	0.551739
2008-2009	0.526905	0.571439	0.400509
2009-2010	0.544721	0.585642	0.40117
2010-2011	0.383877	0.556704	0.479722
2011-2012	0.531889	0.581008	0.450488
2012-2013	0.539017	0.573389	0.458194
2013-2014	0.569947	0.599917	0.494832
2014-2015	0.488291	0.580156	0.421638
2015-2016	0.418454	0.567225	0.454709
2016-2017	0.57041	0.586306	0.439727
2017-2018	0.598071	0.602274	0.442766
2018-2019	0.613625	0.617229	0.499497
2019-2020	0.537616	0.592791	0.44948
2020-2021	0.49671	0.579245	0.413951
2021-2022	0.572257	0.548444	0.461205
2022-2023	0.534011	0.533387	0.471074



4.2.2 Competitive Balance Ratio

Competitive balance is fundamental for sustaining and enhancing fan interest in sports. Over the years, various metrics have been devised to quantify this balance, with the Competitive Balance Ratio (CBR) being a significant measure among them (Neale, 1964). This metric transcends mere seasonal uncertainty, integrating championship uncertainty by accounting for the prolonged dominance or consistency of teams across multiple seasons. Although the standard deviation has been a conventional tool to measure seasonal uncertainty, it does not encompass the extended dominance of teams spanning several seasons. Addressing this gap, Humphreys (2003) and Eckard (2003) proposed a more dynamic metric capable of capturing both facets of uncertainty.

While Eckard (2003) introduced a methodology decomposing the variance of winning percentages into both a cumulative and a time-varying component, Humphreys further refined this concept. Drawing inspiration from Eckard's foundational work, Humphreys presented a simplified metric rooted in the same principles, coining it the Competitive Balance Ratio (CBR) Humphreys (2002).

CBR seamlessly integrates two pivotal components: the 'within-team-standard deviation' and the 'within-season-standard deviation' Humphreys (2002). The former encapsulates the variability of a specific team's performance over multiple seasons, while the latter portrays the performance dispersion of all teams during a single season. Mathematically, the CBR is derived as the ratio of these two standard deviations:

$$CBR = \frac{\sum_{i=1}^{n} SD_{wt,i}}{\sum_{s=1}^{S} SD_{ws,s}}$$

Where,

- $SD_{wt,i}$ represents the within team standard deviation for team i
- $SD_{ws,s}$ is the within season standard deviation for season s
- n is the total number of teams
- S is the total number of seasons

CBR values range between 0 and 1. A value of 0 means every team always gets the same rank every season, showing certain championship outcomes. Meanwhile, a value of 1 means every team has an equal chance to win any season, indicating unpredictable championship outcomes Humphreys (2002).

Using the CBR in European football leagues is tricky because of the promotion and relegation system. Not all teams stay in the top league for the whole time. Also, the teams in the top league can be different in each country, making things even more complicated Humphreys (2002). Still, the CBR is very useful for understanding how balanced the competition is. It's important for those in charge to know the difference between balance in a season and over many seasons. Changes they make might affect these two things differently.

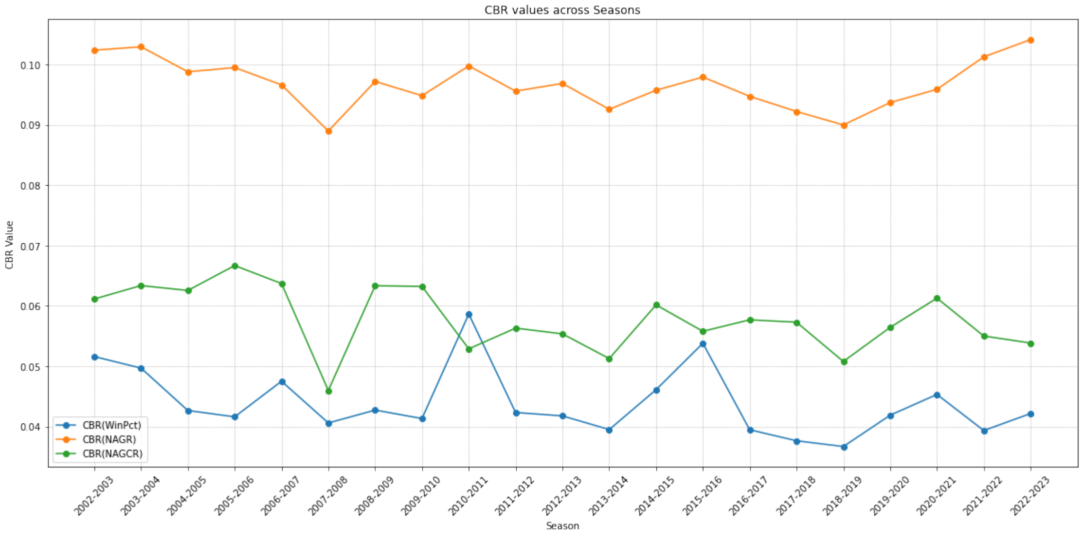
Researchers like Goossens (2005), Humphreys, and Eckard have talked about this kind of balance before. If the CBR value is closer to 1, it means different teams are winning in different seasons. If it's closer to 0, it means teams are pretty consistent in their wins over multiple seasons. In simple terms, our league's competitive balance has changed a lot over the years. Some seasons had teams that were very evenly matched, while others had clear top performers. This makes the league interesting because you never really know which team might shine in a particular season.

Implementation and Analysis

Let's break down the league's competitive balance using the Competitive Balance Ratio (CBR). This ratio helps us understand how evenly matched the teams are over different seasons. The CBR for Win Percentage has seen fluctuations over the years. The highest balance was observed in the 2010-2011 season with a value of 0.058641, indicating significant variations in team performances from one season to the next (Table 4.3). Conversely, the 2018-2019 season recorded the lowest value at 0.036685, suggesting more consistent team performances across multiple seasons. A higher value shows there was a lot of change in how teams performed from one season to the next.

Season	CBR(WinPct)	CBR(NAGR)	CBR(NAGCR)
2002-2003	0.051596	0.102359	0.061124
2003-2004	0.049676	0.102915	0.06336
2004-2005	0.042652	0.098773	0.062543
2005-2006	0.041608	0.099474	0.066674
2006-2007	0.047502	0.096612	0.063693
2007-2008	0.04059	0.088964	0.045969
2008-2009	0.042723	0.097176	0.063327
2009-2010	0.041326	0.094819	0.063223
2010-2011	0.058641	0.099748	0.05287
2011-2012	0.042323	0.095575	0.056301
2012-2013	0.041763	0.096845	0.055355
2013-2014	0.039497	0.092563	0.051256
2014-2015	0.046102	0.095716	0.060154
2015-2016	0.053796	0.097898	0.055779
2016-2017	0.039465	0.094712	0.057679
2017-2018	0.037639	0.092201	0.057283
2018-2019	0.036685	0.089967	0.050777
2019-2020	0.041872	0.093675	0.056428
2020-2021	0.04532	0.095866	0.061271
2021-2022	0.039337	0.10125	0.054993
2022-2023	0.042155	0.104108	0.053841

For the NAGR measure (CBR(NAGR)), the balance also changed from season to season. It went from around 0.0889 in the 2007-2008 season to about 0.1041 in the 2022-2023 season. This change could be because of different team strategies, players joining or leaving teams, or other factors that affected how competitive the league was. For the NAGCR measure (CBR(NAGCR)), the balance ranged from about 0.0459 in the 2007-2008 season to 0.0667 in the 2005-2006 season. This suggests that game results and how matches played out also changed from season to season.



Referring to the graph (Figure 4.3), the Competitive Balance Ratio (CBR) based on Win Percentage indicates a minor uptick during the 2010-2011 season before settling around its mean in the following years, maintaining general stability. The CBR(NAGR) metric reveals a gradual decline from 2002 until 2018, with a modest recovery, particularly noticeable in 2022-2023, and a significant dip during 2007-2008. In contrast, the CBR(NAGCR) demonstrates periodic fluctuations, with a pronounced peak in the 2005-2006 season. Overall, while there have been certain variations, the competitive balance in the EPL, as gauged by these metrics, has remained fairly steady over the past two decades. Notably, the 2010-2011 season might have seen teams with comparable win percentages, leading to heightened competitive balance. The diminishing trend in CBR(NAGR) could suggest a transformation in the intrinsic NAGR values, hinting at changing dynamics within the league.

4.2.3 Normalised Herfindahl-Hirshmann Index and Herfindahl Index of Competitive Balance

The Herfindahl-Hirschman Index (HHI) is a measure originally developed to gauge concentration within industries. By calculating the sum of the squares of the market shares for each firm within an industry, it provides insights into the distribution of market power (Davies, 1979). Mathematically, it's represented as:

$$\text{HHI} = \sum_{i=1}^{N} s_i^2$$

Where,
s_i = Market share of the i^{th} firm
N = Total number of firms in the industry

In an ideal scenario where every firm has an equal market share, the HHI would be $\frac{1}{N}$ (Hart, 1975). Conversely, in a situation of a single firm monopolizing the industry, the HHI would be 1. However, a significant limitation of the HHI is its sensitivity to the number of firms in an industry (Davies, 1979). To counteract this limitation and make the HHI more versatile for various contexts, a normalized version, HHI^*, has been introduced:

$$\text{HHI}^* = \frac{\text{HHI} - \frac{1}{N}}{1 - \frac{1}{N}}$$

This measure, which ranges between 0 (indicating perfect balance) and 1 (indicating maximum imbalance), can be especially useful in settings with varying numbers of entities (Hart, 1975). When the HHI is adapted to the realm of sports, specifically football, 'firms' are analogized as 'teams', and 'market share' becomes the percentage of total points won by a team in a season.

While the C5 ratio focuses on the dominance of the top 5 clubs, it doesn't capture internal disparities among these top clubs or the remaining teams. In contrast, the HHI offers a broader perspective, considering each team's share of points in a season (Michie and Oughton, 2004) . To allow for comparisons between leagues with different sizes, Owen, Ryan, and Weatherston (2007a) introduced a normalized measure of competitive balance. They formulated a measure that ranges from 0 (indicating perfect balance) to 1 (indicating utmost imbalance).

The HHI's sensitivity to the number of teams led to the evolution of the Herfindahl Index of Competitive Balance (HICB). This index normalizes the HHI by relating it to the value expected in a perfectly balanced league:

$$\text{HICB} = \frac{\text{HHI}}{\frac{1}{N}} \times 100$$

In this equation, $\frac{1}{N}$ represents the HHI value for a perfectly balanced league. Thus, for any league, irrespective of its size, the HICB would ideally be 100. Any deviation from this benchmark indicates shifts in competitive balance (Owen, Ryan, and Weatherston, 2007b). The HHI and HICB, backed by decades of research and application (Davies, 1979; Hart, 1975; Owen, Ryan, and Weatherston, 2007b), offer comprehensive insights into the competitive dynamics within leagues. They transcend mere measures of dominance, capturing the nuances and intricacies of point distribution among all participating entities.

Implementation and Analysis

The competitiveness of the English Premier League across various seasons is analysed using three key metrics: the Herfindahl-Hirschman Index (HHI), the normalized HHI (HHI^*), and the Herfindahl Index of Competitive Balance (HICB). These metrics are computed for Points, Net Average Goal Ratio (NAGR), and Net Average Goal Cost Ratio (NAGCR) for each season. The HHI gauges the concentration of performance among teams, with HHI^* providing a normalized perspective, and HICB offering a comparative measure against an ideally balanced league.

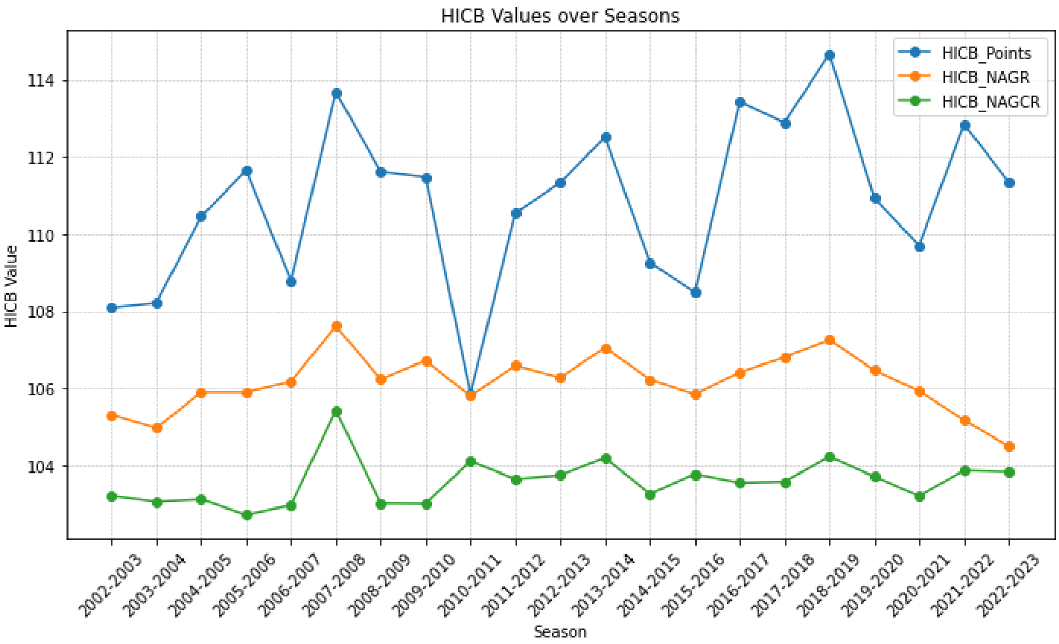
Season	HHI*(Points)	HHI*(NAGR)	HHI*(NAGCR)
2002-2003	0.004259	0.002799	0.001692
2003-2004	0.004326	0.002615	0.00161
2004-2005	0.005502	0.003105	0.001644
2005-2006	0.00613	0.003105	0.001432
2006-2007	0.004628	0.00325	0.001565
2007-2008	0.007202	0.004007	0.00286
2008-2009	0.006114	0.003283	0.00159
2009-2010	0.006041	0.003537	0.001585
2010-2011	0.003085	0.003051	0.002166
2011-2012	0.005549	0.003469	0.001918
2012-2013	0.005963	0.003303	0.001972
2013-2014	0.006587	0.003714	0.002211
2014-2015	0.004877	0.003277	0.001716
2015-2016	0.004467	0.003076	0.001988
2016-2017	0.007068	0.003376	0.001867
2017-2018	0.006783	0.003582	0.001884
2018-2019	0.007724	0.00382	0.002226
2019-2020	0.005753	0.003405	0.001954
2020-2021	0.005105	0.003121	0.001688
2021-2022	0.006763	0.002724	0.002039
2022-2023	0.005968	0.002366	0.002018

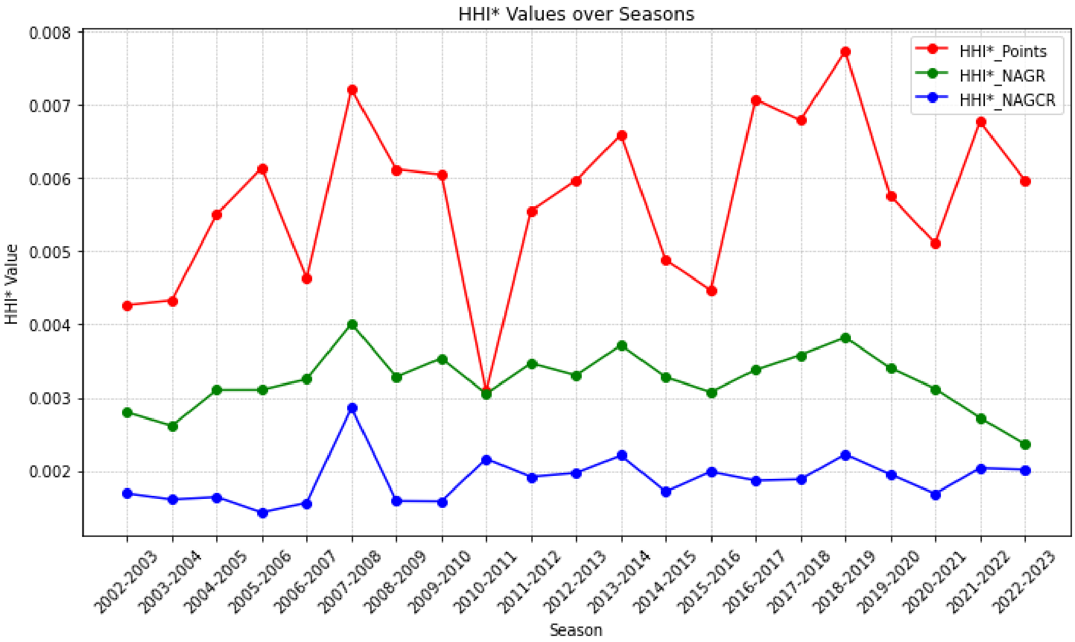
The normalized HHI (HHI^*) further refines the competitive balance insights. For instance, while the 2010-2011 season shows a decline in HHI for points, its corresponding HHI^* value offers a clearer picture of the league's competitiveness when accounting for the number of teams. The variations in HHI^* values are relatively minor, indicating a stable competitive balance over the years.The HICB values consistently exceed 100 across all metrics, indicating a competitive imbalance relative to an ideal scenario. For example, the graph 4.4 show that seasons like 2007-2008 and 2016-2017 have slightly elevated HICB values, hinting at a reduced competitive balance during those periods.

However, the overall trend is fairly consistent, showcasing the league's resilience in maintaining competitive equilibrium. The graph depicts a relatively stable competitive balance over the years based on points, with slight elevations in certain seasons (e.g., 2007-2008). A consistent trend is

Season	HICB(Points)	HICB(NAGR)	HICB(NAGCR)
2002-2003	108.092517	105.317303	103.21439
2003-2004	108.219158	104.9681	103.059112
2004-2005	110.453389	105.898646	103.123133
2005-2006	111.647311	105.899729	102.721307
2006-2007	108.793439	106.175636	102.97317
2007-2008	113.683432	107.613623	105.433758
2008-2009	111.616594	106.237301	103.020591
2009-2010	111.478105	106.720148	103.011777
2010-2011	105.861031	105.796768	104.114646
2011-2012	110.542789	106.590269	103.644933
2012-2013	111.328947	106.275577	103.746505
2013-2014	112.515561	107.056897	104.200512
2014-2015	109.26566	106.227248	103.260672
2015-2016	108.487671	105.843688	103.77806
2016-2017	113.428317	106.414628	103.547779
2017-2018	112.887464	106.805098	103.579518
2018-2019	114.674897	107.257403	104.229534
2019-2020	110.931035	106.469333	103.712185
2020-2021	109.699807	105.929583	103.206728
2021-2022	112.85041	105.175773	103.873341
2022-2023	111.339284	104.49566	103.834917

observed for NAGR, mirroring the competitive balance trends seen in points. The 2007-2008 season again stands out as a period of reduced competitive balance. The NAGCR trend closely aligns with the other two metrics, further solidifying the observations from the HICB of Points and NAGR graphs.





4.2.4 The Five-Club Concentration Ratio (C5) and Index of Competitive Balance(C5ICB)

The English Premier League (EPL) is a complex amalgamation of football skills, business interests, and spirited competition. Given its importance, it becomes academically relevant to study its competitive landscape. Notably, the Five-Club Concentration Ratio (C5) and the Index of Competitive Balance (C5ICB) have been recognized as essential metrics in this regard (Davies, 1979; Hart, 1975). Introduced by Michie and Oughton (2004), these instruments offer profound perspectives on how success is distributed within the league (Michie and Oughton, 2004).

The literature offers two distinct applications of this concept, each with differing benchmark statistics. One approach, suggested by Nikolakaki et al. (2020), measures the concentration of the top teams against the maximum points they could theoretically garner, known as the ’attainable’ concentration ratio. Another approach, termed the ‘C5 ratio’ by Michie and Oughton (2004), compares this concentration to the total points won by all teams in the league. Both metrics can be applied to evaluate ’open’ or ’closed’ leagues, with considerations for inter-league and temporal comparisons.

Research by Michie and Oughton (2004) showed that from 1947 to 2004, the dominance of the top clubs in English football grew by 6.4% between 1989 and 2004. This suggests that the top clubs became more powerful during this period.

The Five-Club Concentration Ratio (C5) in the EPL

The C5 ratio, a concept from business studies, measures the dominance of the top players in any industry. In the context of the English Premier League (EPL), it looks at how the top five football clubs perform compared to all the teams in the league. Here’s how it’s calculated:

$$C5 \text{ Ratio} = \frac{\text{Total points won by the top five clubs}}{\text{Total number of points won by all clubs}}$$

For a 20-team league like the EPL, the C5 ratio can be as low as 0.25 (if all teams perform equally) and as high as 0.55 (if the top five teams win all their games). Based on the data from 2002 to 2023 (Table 5), we see that the dominance of the top five teams has varied. For instance, in the 2007-2008 season, the dominance was high with a C5 ratio of 0.380769. On the other hand, in the 2010-2011 season, the league was more balanced with a C5 ratio of 0.342080.

The Index of Competitive Balance (C5ICB)

The C5 ratio tells us about the dominance of the top 5 teams, but it doesn’t consider how big the league is. On the other hand, the C5ICB takes the league size into account and compares the C5 ratio to a perfect balance scenario (Michie and Oughton, 2004). In simple terms, C5ICB is calculated as:

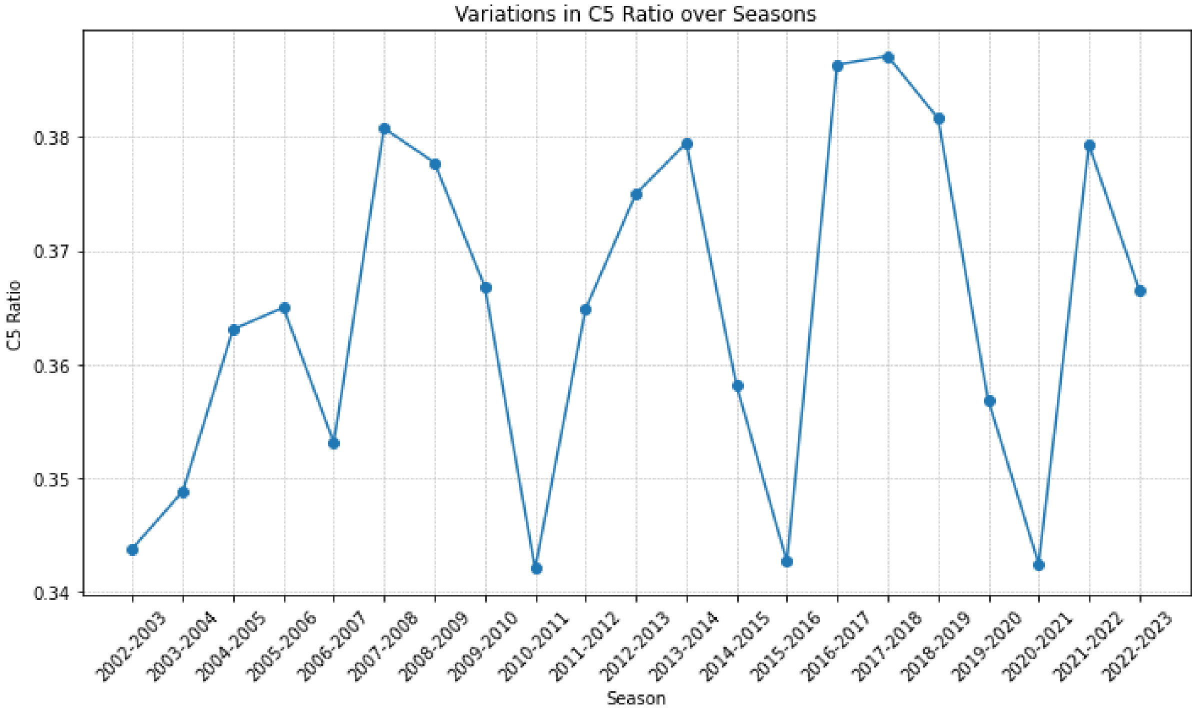
$$C5ICB = \left(\frac{C5 \text{ ratio}}{\frac{5}{N}}\right) \times 100$$

Here, N is the total number of teams in the league. If the C5ICB is 100, the league has perfect balance. If it’s more than 100, it means less balance. Using the C5 ratio and C5ICB can help understand how competitive a sports league is. For example, research by Michie and Oughton (2004) showed that from 1989 to 2004, the English Premier League became 6.4% less balanced. When we compare this to other European football leagues, we see differences. Italy’s main league became much less competitive after 1992. Germany's league became a bit less balanced over 10 years. But France and Spain saw their competitiveness go up and down over time (Michie and Oughton, 2004).

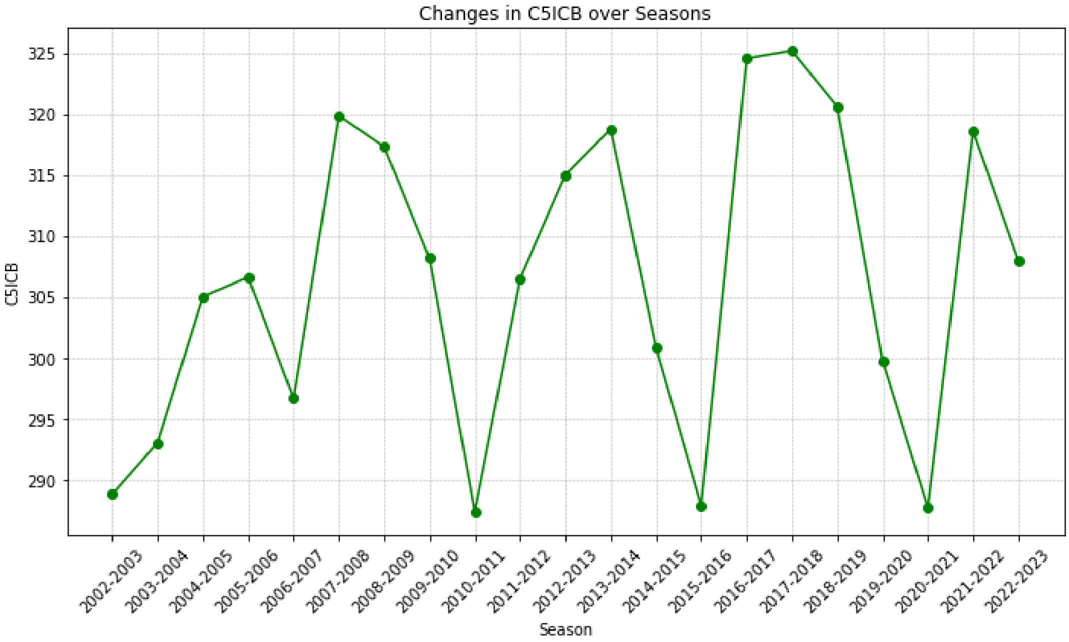
Implementation and Analysis

Utilizing the provided dataset, we calculated two primary metrics: the C5 Ratio and the C5 Index of Competitive Balance for the English Premier League seasons. The methodologies and interpretations are detailed below:

1. C5 Ratio: The C5 Ratio measures the proportion of points secured by the top five teams relative to the entire league. It provides an insight into the dominance of these top teams within the league for a particular season. A higher C5 ratio indicates a more pronounced dominance of the top 5 teams in the league Michie and Oughton, 2004. From our calculations (refer Table 4.6), the C5 ratios for the observed seasons span a range from approximately 0.34 to 0.39. This suggests that the top five teams garnered between 34% and 39% of the total points available in each season.



2. C5ICB: The C5 Index of Competitive Balance is a standardized metric, offering a percentage-based representation of the C5 ratio in relation to an ideally competitive league. An index value of 100 represents perfect competitive balance, while values exceeding 100 denote varying degrees of imbalance Michie and Oughton, 2004. Our dataset reveals C5ICB values ranging from about 288 to 325 (refer Table 4.6), indicating periodic fluctuations in the competitive balance of the league.



Notable seasons such as 2005-2006, 2009-2010, 2014-2015 and 2019-2020 display C5ICB values on the lower end, suggesting enhanced competitiveness during these periods. Conversely, the 2006-2007, 2012-2013, 2016-2016 and 2017-2018 seasons exhibited elevated C5ICB values, pointing towards increased dominance by the top five teams. While the league has witnessed oscillations in its competitive balance, there isn't a definitive trend towards either increasing or decreasing competitiveness in the recent years. Such variations infer the dynamic nature of the league, with different teams having their moments of prominence across seasons.

Projecting the league's competitive trajectory based solely on past data presents challenges. Nevertheless, recent seasons suggest a tendency towards maintaining a semblance of competitive equilibrium. External influences, such as regulatory changes, revenue distribution models, and global events, might shape future trends. In summation, over the evaluated periods, the English Premier League showcased a diverse competitive landscape without any dominant group persistently overshadowing the league across seasons. The unpredictability of outcomes adds to the league's appeal, making it a compelling spectacle for audiences and stakeholders alike.

Season	C5 Ratio	C5ICB
2002-2003	0.34381	288.8
2003-2004	0.348837	293.023256
2004-2005	0.363107	305.009709
2005-2006	0.365005	306.603951
2006-2007	0.353167	296.660269
2007-2008	0.380769	319.846154
2008-2009	0.377756	317.315436
2009-2010	0.366858	308.16092
2010-2011	0.34208	287.346939
2011-2012	0.364852	306.475645
2012-2013	0.375	315.0
2013-2014	0.379473	318.757062
2014-2015	0.358166	300.859599
2015-2016	0.342691	287.8606
2016-2017	0.386364	324.545455
2017-2018	0.387128	325.18732
2018-2019	0.381665	320.59869
2019-2020	0.35687	299.770992
2020-2021	0.342479	287.682119
2021-2022	0.379278	318.593156
2022-2023	0.366572	307.920228

4.2.5 Gini Coefficients and Lorenz Curve

The Lorenz curve and Gini coefficient are well-established tools for measuring inequality, often used in economics to study income or wealth distribution. In the context of sports economics, these tools can provide insights into competitive balance within a league over multiple seasons. Quirk and Fort (1992) were among the first to apply these methods to assess the distribution of championships among teams in U.S. sports leagues. They plotted the Lorenz curve by ranking teams based on their titles-per-year ratio and then used the Gini coefficient to quantify the degree of inequality Quirk and Fort (1992).

In football, the Lorenz curve has been adapted to represent the distribution of league points among teams in a single season, rather than long-term championships. This seasonal approach was introduced to provide a more dynamic measure of competitive balance, without being affected by changes in the number of teams (Szymanski and Kuypers, 1999). Teams are ranked from the lowest to the highest number of points, and the Lorenz curve is plotted based on the cumulative percentage share of points across all teams. The Gini coefficient is then calculated using the formula:

$$G_D = \sum_{i=1}^{n} (x_i y_{i+1} - x_{i+1} y_i)$$

Where x_i and y_i represent the cumulative percentages of the number of teams and titles, respectively, won by teams up to rank i.

Szymanski and Kuypers (1999) applied the Lorenz curve to evaluate competitive balance in top European football leagues, concluding that the English league exhibited the least concentration of championships among clubs. Goossens (2005) extended this analysis to eleven European football leagues over a 42-year period, also reporting the weighted Gini coefficients for each league Goossens (2005). The Lorenz curve and Gini coefficient have limitations when applied to 'open' leagues or leagues with varying numbers of teams over different periods. Quirk and Fort (1992) and Goossens (2005) addressed this by only including teams that have been in the league for at least ten years and by weighting the teams based on the number of seasons they are included in the competition.

The Lorenz curve and Gini coefficient offer valuable perspectives on competitive balance in football leagues, complementing other indicators like the C5 ratio and the H-index. They provide a robust framework for assessing both seasonal and long-term competitive balance, albeit with some limitations that need to be carefully addressed in the analysis.

Implementation and Analysis

To chart the Lorenz curve and compute the Gini coefficient, the following methodology was employed for each season:

- Teams were ranked based on their accumulated points for the season.
- Cumulative percentages of points and teams were then calculated from this ranked list.
- The Lorenz curve was plotted, taking the cumulative percentage of teams on the x-axis and the cumulative percentage of points on the y-axis.
- The Gini coefficient for each season was determined by computing the area between the Lorenz curve and the line of equality (45Âř line).

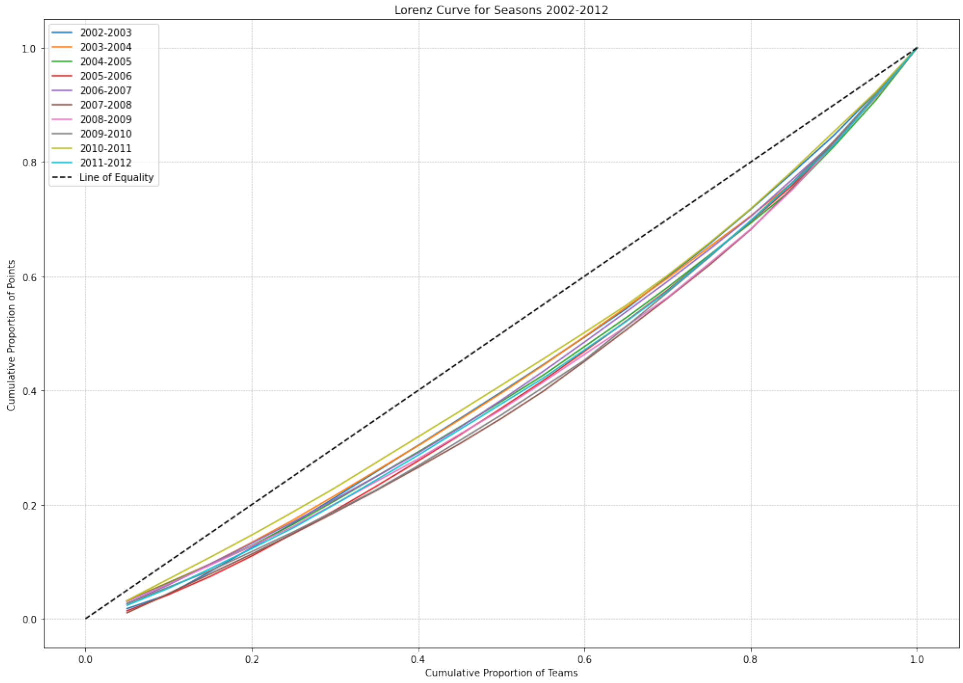
The Gini coefficient provides a scalar value to gauge inequality. In our context, a Gini coefficient of 0 would mean perfect equality, where every team has garnered the same number of points. Conversely, a Gini coefficient of 1 would suggest maximal inequality, indicating that one team has secured all the points, leaving none for the rest (Table 6).

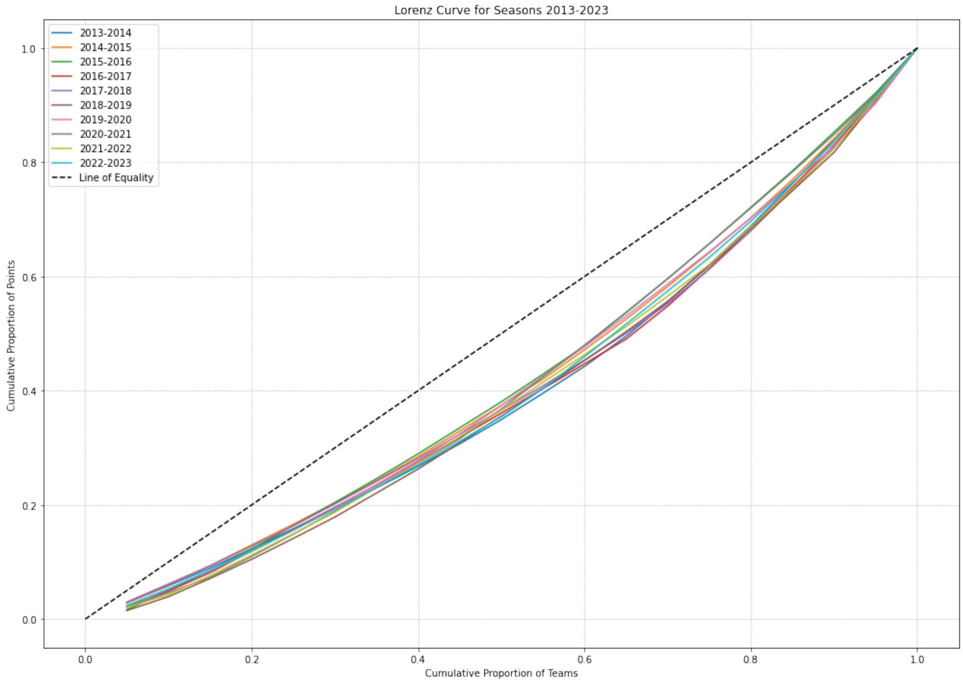
From the tabulated Gini coefficients (Table 4.7), it is discernible that the values hover around 0.60. This suggests a certain degree of competitive balance in the league. While the league hasn't achieved perfect equality, it's commendable that there hasn't been overarching dominance by any single team across the seasons. The Lorenz curves, as illustrated in figure 4.7 and 4.8, provide a visual representation of the point distribution among the EPL teams for each season. Notably, the closer the curve is to the diagonal line of equality, the more evenly distributed the points are among the teams.

Season	GiniCoefficient
2002-2003	0.622385
2003-2004	0.632525
2004-2005	0.626881
2005-2006	0.607365
2006-2007	0.625998
2007-2008	0.603196
2008-2009	0.621094
2009-2010	0.609931
2010-2011	0.640887
2011-2012	0.618892
2012-2013	0.6171
2013-2014	0.61167
2014-2015	0.621769
2015-2016	0.616616
2016-2017	0.611844
2017-2018	0.621425
2018-2019	0.600635
2019-2020	0.615252
2020-2021	0.606872
2021-2022	0.60862
2022-2023	0.61093

From the two decades observation of Figure 4.7 and Figure 4.8: The early 2000s presented a blend of competitive balance. While seasons like 2002-2003 and 2004-2005 showcased Lorenz curves nearly akin to the line of perfect equality, seasons such as 2005-2006 and 2010-2011 deviated further, hinting at the dominance of certain teams. The latter decade, on the other hand, predominantly displayed Lorenz curves hugging closely to the line of perfect equality, suggesting a more consistent competitive balance in the league.

Across the observed span, there's a perceptible shift towards heightened competitive balance in the EPL. The early 2000s bore witness to distinct variations in competitiveness. In contrast, the subsequent years, particularly the late 2010s and early 2020s, have been characterized by an increasing trend of balance. This shift might be attributed to a myriad of factors, including an equitable revenue distribution, strategic scouting, and the influx of adept managers (Szymanski and Kuypers, 1999). The Lorenz curves and Gini coefficients offer a lucid representation of the EPL's trajectory in terms of competitive balance over the years. The insights gleaned underscore the league's commitment to fostering an environment where any team stands a chance, enhancing the unpredictability, thrill, and viewer engagement of the matches.


Lorenz Curve for Seasons 2002-2012
2002-2003
2003-2004
2004-2005
2005-2006
2006-2007
2007-2008
2008-2009
2009-2010
2010-2011
2011-2012
Line of Equality
Cumulative Proportion of Points
Cumulative Proportion of Teams
0.0
0.2
0.4
0.6
0.8
1.0


Lorenz Curve for Seasons 2013-2023
2013-2014
2014-2015
2015-2016
2016-2017
2017-2018
2018-2019
2019-2020
2020-2021
2021-2022
2022-2023
Line of Equality
Cumulative Proportion of Points
Cumulative Proportion of Teams
0.0
0.2
0.4
0.6
0.8
1.0

4.2.6 Relative Entropy

Relative entropy, also known as the Kullback-Leibler divergence, measures the difference between two probability distributions. Its value is non-negative, and it ranges from 0 to ∞. Relative entropy, originating from information theory, has been employed as a tool to measure competitive balance within sports leagues. However, when relative entropy is used in the context of measuring competitive balance in sports leagues, as in the case presented earlier, its interpretation and range might be different. In such cases, it's defined in a specific way to capture the degree of uncertainty about which team might have won a randomly-selected game relative to the maximum possible uncertainty. Here, the range is generally between a value close to 1 (indicating a highly competitive league) and a value that depends on the number of teams in the league (indicating a less competitive league). Specifically, Horowitz (1997) utilized this measure to assess seasonal competitive balance in Major League Baseball. The essence of the relative entropy measure is to gauge the "degree of uncertainty" about which team might have triumphed in a randomly chosen game, relative to the highest possible uncertainty Horowitz (1997). The formula for calculating relative entropy, denoted by RE, is:

$$RE = \frac{E}{E_{\text{max}}}$$

Where E represents actual entropy which is calculated as:

$$E = -\sum p_i \log_2 p_i$$

With p_i denoting the proportional of total victories in the league season for team i and

$$E_{\text{max}} = -\log_2\left(\frac{1}{N}\right)$$

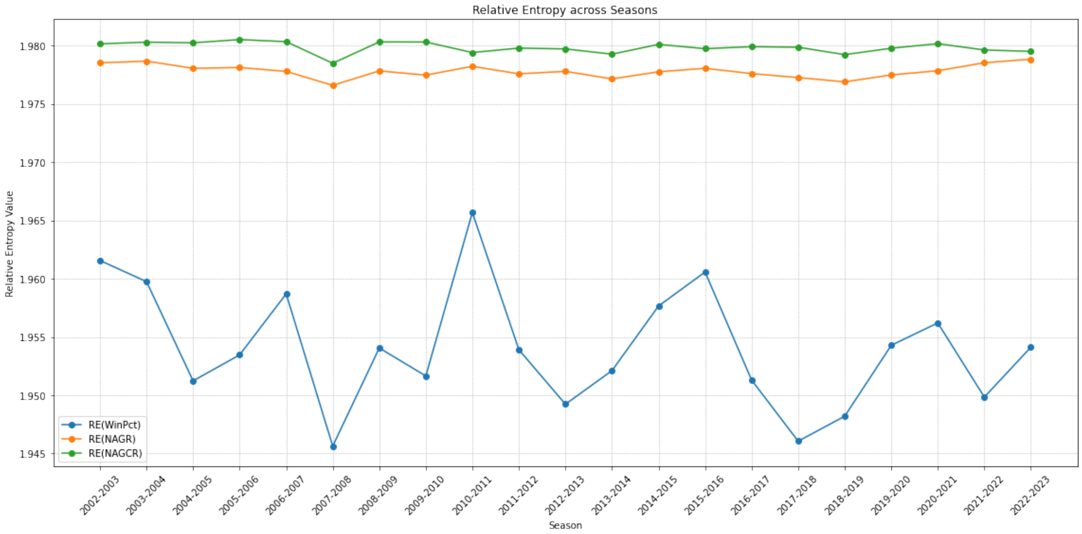
N being the total number of teams in a league.

A perfectly balanced league would yield a relative entropy value of one, which would occur when the actual uncertainty aligns with the maximum possible uncertainty. Conversely, the minimum value, suggesting the least balance in the league, is contingent on the number of teams. For instance, in a league comprised of 20 teams, the minimum value is roughly 0.93. Interestingly, if the number of teams in the league decreases, this potential range for the measure expands, and vice versa. Thus, a league with only 10 teams would have a minimum value of 0.89.

Horowitz (1997) approach is adaptable to 'open' leagues, as it assesses the league based on the participating teams at a given time. However, a limitation arises when attempting to compare leagues with differing team counts, as the scale of the measure hinges on the team quantity. In his study, Horowitz (1997) computed the values of RE for the American and National baseball leagues from 1903 to 1995. These values were pivotal in discerning the trend in competitive advantage across both leagues over time. Additionally, his regression methodology facilitated the examination of factors that might impact the overarching trend in a league's competitive balance, either in the short or long run.

Implementation and Analysis

In this study, we looked at how unpredictable the English Premier League is by using a measure called Relative Entropy (RE). We focused on three main areas: win rates, goal ratios, and cumulative goal ratios. You can check out the details in Table 4.8 and see the trends in Figure 4.9.



The unpredictability in the English Premier League, as measured by the RE values for win percentages, remains relatively consistent over time, oscillating between 1.945 and 1.966. A closer inspection reveals that the 2010-2011 season exhibited the highest unpredictability with an RE value of approximately 1.966. Conversely, the 2007-2008 season manifested slightly more predictability, registering an RE value close to 1.946. Diving deeper into the goal ratios, the variance in unpredictability is minimal, with RE values ranging from 1.976 to 1.979 across seasons. Such consistent RE values suggest that teams have demonstrated relative uniformity in their performances when evaluated based on goals scored and conceded. Likewise, the cumulative goal ratios further underscore this consistency, as their unpredictability metrics consistently lie between 1.978 and 1.980. In essence, these metrics collectively attest to the sustained competitive balance and consistency of teams in the league throughout the observed seasons.

In simple terms, the English Premier League has been pretty steady in terms of competition. While some seasons have had a few surprises, no single team has been able to dominate or lag behind for a long time. This makes the league exciting to follow, just like many experts have

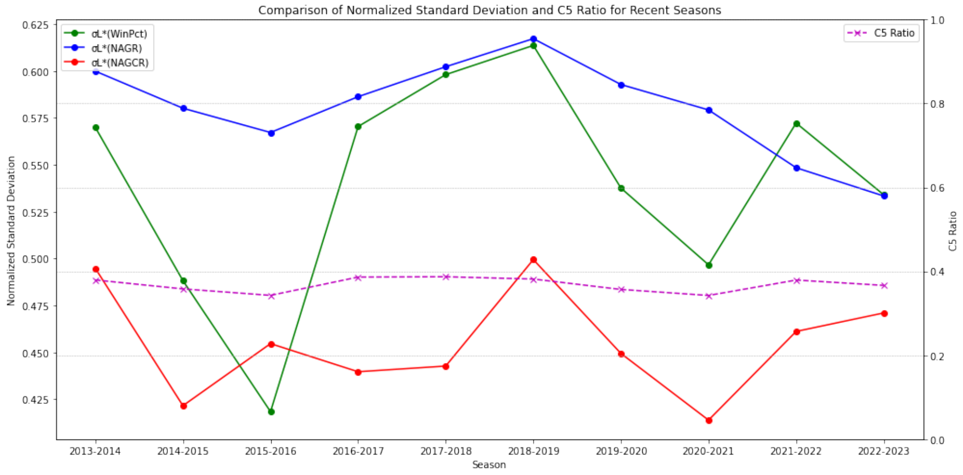
Season	**RE(WinPct)**	**RE(NAGR)**	**RE(NAGCR)**
2002-2003	1.961558	1.978516	1.980143
2003-2004	1.959745	1.978666	1.980294
2004-2005	1.951234	1.978045	1.980236
2005-2006	1.953463	1.978120	1.980514
2006-2007	1.958698	1.977782	1.980334
2007-2008	1.945636	1.976585	1.978482
2008-2009	1.954054	1.977825	1.980313
2009-2010	1.951653	1.977462	1.980304
2010-2011	1.965678	1.978219	1.979401
2011-2012	1.953904	1.977573	1.979783
2012-2013	1.949229	1.977783	1.979712
2013-2014	1.952112	1.977143	1.979261
2014-2015	1.957655	1.977748	1.980099
2015-2016	1.960571	1.978051	1.979733
2016-2017	1.951302	1.977593	1.979903
2017-2018	1.946060	1.977255	1.979857
2018-2019	1.948201	1.976886	1.979219
2019-2020	1.954285	1.977481	1.979772
2020-2021	1.956204	1.977840	1.980161
2021-2022	1.949839	1.978525	1.979622
2022-2023	1.954123	1.978833	1.979500

pointed out before.

Chapter 5

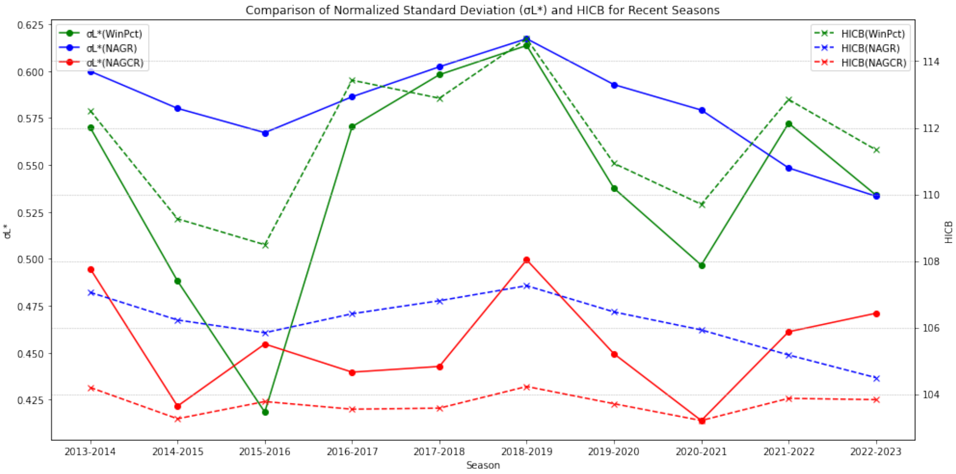
Results and Discussions

For recent changes in competitiveness, SD and Concentration Ratios are compared from the year 2013-2023. Over the recent seasons, the σ_{L^*} for WinPct has had slight fluctuations. From 2013 to 2023, the graph depicting the Normalized Standard Deviation (σ_{L^*}) and the C5 Ratio provides significant insights. Between 2013 and 2016, there was stability in the normalized standard deviation for win percentages, suggesting consistent performance disparities among teams. However, a peak in 2017 followed by a decline until 2019 implies a fluctuation in team performance levels. By 2020, a recovery is observed, leveling off by 2023, indicating a return to competitive balance. Interestingly, the trends for NAGR and NAGCR closely mirror those of win percentages, suggesting these metrics are interconnected.



The C5 Ratio, representing the dominance of the top five teams, steadily rose from 2013 to 2018. This indicates a phase of increased dominance by a select few teams, possibly due to advantageous gameplay or strategies. However, a dip in 2019-2020 suggests a decrease in this dominance, while stabilization from 2021 onwards indicates a consistent performance distribution. Such patterns might be influenced by factors like team management shifts, player transfers, or league rule changes.

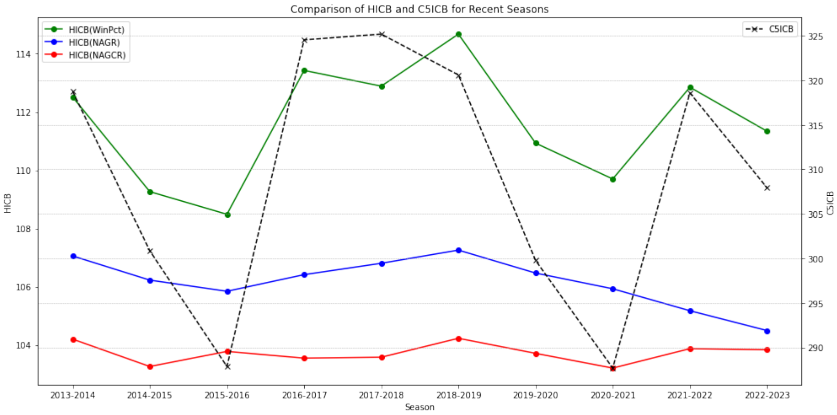
It's intuitive to expect a negative correlation between the Normalised standard deviation and HICB. When σ_{L^*} is high (indicating competitive balance), HICB should be low (indicating a spread out metric across teams).



As observed in Figure 5.2, when σ_{L^*} increases, suggesting better competitive balance, HICB decreases, indicating that the metric's concentration among teams is reducing. This is particularly evident in recent seasons, suggesting that the league's competitive balance has improved in recent years.From 2013 to 2018, the English Premier League exhibited dynamic shifts in team performance. The 2013-2014 season marked a decline in performance disparities, though dominance by a few teams was evident, as indicated by the rise in the HICB metrics. This trend of a few teams dominating continued until 2015-2016. However, by 2016-2017, performance disparities peaked, indicating some teams significantly outshining others. Yet, by the end of 2017-2018, a noticeable shift occurred, with an evident decline in disparities, suggesting a more evenly matched league.

The subsequent years, from 2018 to 2023, showcased a more stable league performance. The 2018-2019 season was particularly stable, with both normalized standard deviations and HICB metrics reflecting a balanced competitive environment. While there were minor fluctuations in subsequent years, by 2022-2023, both metrics stabilized, hinting at a competitive equilibrium. This decade-long overview underscores a transition from pronounced disparities and dominance to a more balanced competitive landscape by 2023.

In the graph (Figure 5.3), we're comparing HICB (Herfindahl Index Corrected for Bias) values for WinPct, NAGR, and NAGCR against C5ICB. Starting from the 2013-2014 season and culminating in 2022-2023, there's a noticeable decline in the HICB for WinPct and the C5ICB values. This descending trajectory indicates that the competitive balance in the league has been improving. This suggests a decreased dominance by a few teams over the course of these recent seasons. Overall for the seasons spanning 2013-2014 to 2022-2023, both HICB and C5ICB metrics signal a growth in competitive balance within the league. While HICB offers a comprehensive view by factoring in all teams' performances, C5ICB hones in on the dominance



exerted by the top 5 teams. The consistent downward trend in these metrics over recent seasons highlights a promising move towards a more competitively balanced league.

Based on the overall analysis of various measures to find the competitive balance of league, The 2017-2018 and 2018-2019 seasons frequently emerges as the most competitive across various measures, while the 2010-2011 season often appears as the least competitive (Refer Table 8). This suggests that there has been significant variation in league competitiveness over the years. The current trend for the σ_{L*}(WinPct) and HICB(WinPct) metrics indicates declining competitiveness over the past five years. In contrast, the C5 and C5ICB measures, which examine the distribution of points among the top teams, show increasing competitiveness. This could imply a closing gap between top-tier teams and the rest of the league, making the top spots more contested.

Measure	Most Competitive Year	Least Competitive Year	Current Trend (past 5 years)	Future Trend (Linear Regression)
σ_{L*}(WinPct)	2018-2019	2010-2011	Declining	Declining
Gini Coefficient (Points)	2018-2019	2010-2011	Stable	Stable
HICB (WinPct)	2018-2019	2010-2011	Declining	Declining
C5 (WinPct)	2010-2011	2017-2018	Increasing	Increasing
C5ICB (WinPct)	2010-2011	2017-2018	Increasing	Increasing
HHI* (WinPct)	2010-2011	2018-2019	Declining	Stable
HICB (Points)	2010-2011	2018-2019	Stable	Increasing
Relative Entropy (WinPct)	2010-2011	2007-2008	Stable	Stable
Gini Coefficient (Points)	2018-2019	2010-2011	Stable	Stable

The use of linear regression provides a simple and interpretable model to project future trends based on past data. While it assumes that the past linear trend will continue into the future, it's worth noting that real-world factors (e.g., changes in league regulations, financial disparities, or global events like pandemics) could disrupt these trends. The future trends, as projected by linear regression, indicate a mix of stability and change. For instance, metrics like the Gini Coefficient and Relative Entropy predict stable competitiveness. However, the HICB for points predicts an increase, suggesting a potential future where teams might be more evenly matched in terms of points earned.

Chapter 6

Conclusion

The English Premier League's competitiveness has had its ups and downs between 2013 and 2023. At first, things were pretty stable. Teams were more or less evenly matched until 2016. But then, in 2017, some teams started to pull away from the pack. Things got back to normal by 2023, with teams being more evenly matched again. When we looked at other data, like HICB, it showed the same thing: the league is becoming more balanced over time.

The top five teams were really dominant up to 2018 but then started to face more competition. All these changes could be due to many things like new players, changes in team management, or even new rules in the league. The numbers (HICB and C5ICB) we used to measure competitiveness show that the league is getting more balanced. This means that more teams have a chance to do well, making the league more exciting for fans. When we looked at different years, it was clear that the 2017-2018 and 2018-2019 seasons were the most competitive. On the other hand, the 2010-2011 season was not very competitive at all. Over the past five years, some measures like normalised standard deviation (WinPct) and HICB showed that the league was becoming less competitive. But other measures like C5 and C5ICB said the opposite. This could mean that while some top teams are still dominant, the rest of the teams are catching up, making the race for the top spots more exciting.

Looking ahead, we used linear regression to guess what might happen in the future. While these are just educated guesses, they show that we could expect more of the same in terms of competitiveness. But real-world events, like changes in league rules or even global situations, could shake things up. Overall, the Premier League has been getting more balanced over the last decade. This is great for fans, as more balanced competition usually means more exciting games. Future work could focus on improving the methods we use to measure competitiveness, maybe by taking into account more real-world factors that could affect game outcomes.

Bibliography

Alwell, K. (2020). “Analyzing competitive balance in professional sport”. In.

Andreff, W. and N. Scelles (2015). “Walter C. Neale 50 years after: Beyond competitive balance, the league standing effect tested with French football data”. In: *Journal of Sports Economics* 16.8, pp. 819–834.

Basini, F. et al. (2023). “Assessing competitive balance in the English Premier League for over forty seasons using a stochastic block model”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 186.3, pp. 530–556.

Beck, H., A. Prinz, and T. Van Der Burg (2022). “The league system, competitive balance, and the future of European football”. In: *Managing Sport and Leisure*, pp. 1–24.

Brandes, L. and E. Franck (2007). “Who made who? An empirical analysis of competitive balance in European soccer leagues”. In: *Eastern Economic Journal* 33.3, pp. 379–403.

Davies, S. (1979). “Choosing between concentration indices: The iso-concentration curve”. In: *Economica*, pp. 67–75.

Deb, S. (2021). “A mathematical take on the competitive balance of a football league”. In: *arXiv preprint arXiv:2102.09288*.

Deloitte (2023). *Annual Review of Football Finance 2023*. Manchester: Deloitte’s Sports Business Group.

DâĂŹOttaviano, F. (2019). “A metric to measure dynamic competitive balance with respect to prize concentration”. In: *Journal of Sports Analytics* 5.3, pp. 191–204.

Eckard, E.W. (1998). “The NCAA cartel and competitive balance in college football”. In: *Review of Industrial Organization* 13, pp. 347–369.

— (2001a). “BaseballâĂŹs blue ribbon economic report: Solutions in search of a problem”. In: *Journal of Sports Economics* 2.3, pp. 213–227.

— (2001b). “The origin of the reserve clause: Owner collusion versus âĂIJpublic interestâĂİ”. In: *Journal of Sports Economics* 2.2, pp. 113–130.

— (2003). “The ANOVA-based competitive balance measure: A defense”. In: *Journal of Sports Economics* 4.1, pp. 74–80.

Evans, R. (2014). “A review of measures of competitive balance in the âĂIJanalysis of competitive balanceâĂİ literature”. In.

Fort, R. and J. Quirk (1995). “Cross-subsidization, incentives, and outcomes in professional team sports leagues”. In: *Journal of Economic literature* 33.3, pp. 1265–1299.

— (1997). “Introducing a competitive economic environment into professional sports”. In: pp. 3–26.

Goossens, K. (2005). “Competitive balance in European football: Comparison by adapting measures: National measure of seasonal imbalance and top 3”. In.

Hall, M. and N. Tideman (1967). “Measures of Concentration”. In: *Journal of the American Statistical Association* 62.317, pp. 162–168.

Hart, P.E. (1975). “Moment distributions in economics: an exposition”. In: *Journal of the Royal Statistical Society: Series A (General)* 138.3, pp. 423–434.

Horowitz, I. (1997). “The increasing competitive balance in Major League Baseball”. In: *Review of Industrial Organization* 12, pp. 373–387.

Humphreys, B.R. (2002). “Alternative measures of competitive balance in sports leagues”. In: *Journal of Sports Economics* 3.2, pp. 133–148.

— (2003). “Comments on âĂIJThinking about competitive balanceâĂİ”. In: *Journal of Sports Economics* 4.4, pp. 284–287.

Inan, T. (2018). “30 Years Trend of Competitive Balance in Turkish Football Super League”. In: *Journal of Education and Training Studies* 6.1, pp. 63–69.

Kringstad, M. and B. Gerrard (2007). “Beyond competitive balance”. In: pp. 149–172.

Manasis, V., I. Ntzoufras, and J. Reade (2015). “Measuring competitive balance and uncertainty of outcome hypothesis in European football”. In: *arXiv preprint arXiv:1507.00634*.

Manasis, V. et al. (2013). “Quantification of competitive balance in European football: development of specially designed indices”. In: *IMA Journal of Management mathematics* 24.3, pp. 363–375.

Michie, J. and C. Oughton (2004). “Competitive balance in football: Trends and effects”. In.

Neale, W.C. (1964). “The peculiar economics of professional sports”. In: *The Quarterly Journal of Economics* 78.1, pp. 1–14.

Nikolakaki, S.M. et al. (2020). “Competitive balance in team sports games”. In: *2020 IEEE Conference on Games (CoG)*. IEEE, pp. 526–533.

Owen, P.D. (2010). “Limitations of the relative standard deviation of win percentages for measuring competitive balance in sports leagues”. In: *Economics Letters* 109.1, pp. 38–41.

Owen, P.D., M. Ryan, and C.R. Weatherston (2007a). “Measuring competitive balance in professional team sports using the Herfindahl-Hirschman index”. In: *Review of Industrial Organization* 31, pp. 289–302.

— (2007b). “Measuring competitive balance in professional team sports using the Herfindahl-Hirschman index”. In: *Review of Industrial Organization* 31, pp. 289–302.

Plumley, D., G. Ramchandani, and R. Wilson (2018). “Mind the gap: an analysis of competitive balance in the English football league system”. In: *International Journal of Sport Management and Marketing* 18.5, pp. 357–375.

Plumley, D. et al. (2022). “Looking forward, glancing back; competitive balance and the EPL”. In: *Soccer and society* 23.4-5, pp. 466–481.

Quirk, J. and R. Fort (1992). *Pay Dirt, the Business of Professional Team Sports*. Princeton: Princeton University Press.

Ramchandani, G. et al. (2018). “A longitudinal and comparative analysis of competitive balance in five European football leagues”. In: *Team Performance Management: An International Journal* 24.5/6, pp. 265–282.

Rottenberg, S. (1956). “The baseball players’ labor market”. In: *Journal of political economy* 64.3, pp. 242–258.

Scully, G.W. (1989). *The business of major league baseball.*

Szymanski, S. (2003). “The economic design of sporting contests”. In: *Journal of economic literature* 41.4, pp. 1137–1187.

— (2010). *Income Inequality, Competitive Balance and the Attractiveness of Team Sports: Some Evidence and a Natural Experiment from English Soccer*. Palgrave Macmillan UK.

Szymanski, S. and T. Kuypers (1999). *Winners and losers: The business strategy of football Harmonds worth.*

Triguero-Ruiz, F. and A. ÃĄvila Cano (2018). “Measuring competitive balance in the major european soccer leagues”. In: *Journal of Physical Education and Sport* 18.3, pp. 1335–1340.

Young, Ernst & (2022). *Premier League Economic and social impact*. London.

Zimbalist, A.S. (2002). “Competitive balance in sports leagues: An introduction”. In: *Journal of Sports Economics* 3.2, pp. 111–121.

Chapter 7

Appendix

Reading and Preprocessing of Data

```
[ ]: #Importing required libraries
     import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import warnings
     warnings.filterwarnings('ignore')
```

```
[ ]: # Define a list of season names
     seasons = ['2002-2003','2003-2004', '2005-2006',␣
      ↪'2006-2007', '2007-2008', '2008-2009',
                '2009-2010', '2010-2011', '2011-2012',␣
      ↪'2012-2013', '2013-2014', '2014-2015', '2015-2016',
                '2016-2017', '2017-2018', '2018-2019',␣
      ↪'2019-2020', '2020-2021', '2021-2022', '2022-2023']



     # Create an empty list to store the processed DataFrames
     season_dataframes = []

     # Loop through each season
     for season in seasons:
         # Read the CSV file for the current season
         filename = f'Season {season}.csv'
         data = pd.read_csv(filename)
         print(data.shape)
```

```
    # Select the required features and copy the DataFrame
    data_selected = data[['Date','HomeTeam', 'AwayTeam',␣
 ↪'FTHG', 'FTAG', 'FTR', 'HTHG', 'HTAG', 'HTR',
                         'HS', 'AS', 'HST', 'AST', 'HF',␣
 ↪'AF', 'HC', 'AC', 'HY', 'AY', 'HR', 'AR']].copy()

    # Add the "Season" column with the current season value␣
 ↪to the DataFrame using .loc
    data_selected.loc[:, 'Season'] = season

    # Reorder the columns to have "Season" as the first␣
 ↪column
    data_selected =␣
 ↪data_selected[['Season','Date','HomeTeam', 'AwayTeam',␣
 ↪'FTHG', 'FTAG', 'FTR', 'HTHG', 'HTAG',
                                   'HTR', 'HS', 'AS', 'HST',␣
 ↪'AST', 'HF', 'AF', 'HC', 'AC', 'HY', 'AY', 'HR', 'AR']]

    # Append the processed DataFrame to the list of␣
 ↪season_dataframes
    season_dataframes.append(data_selected)


S_0405=pd.read_csv('Season 2004-2005.csv',␣
 ↪encoding="ISO-8859-1",  usecols=range(0, 57))
data1 = S_0405[['Date','HomeTeam', 'AwayTeam', 'FTHG',␣
 ↪'FTAG', 'FTR', 'HTHG', 'HTAG', 'HTR', 'HS', 'AS', 'HST',␣
 ↪'AST', 'HF', 'AF', 'HC', 'AC', 'HY', 'AY', 'HR', 'AR']].
 ↪copy()
data1.loc[:, 'Season'] = '2004-2005'
data1 = data1[['Season','Date','HomeTeam', 'AwayTeam',␣
 ↪'FTHG', 'FTAG', 'FTR', 'HTHG', 'HTAG', 'HTR', 'HS', 'AS',␣
 ↪'HST', 'AST', 'HF', 'AF', 'HC', 'AC', 'HY', 'AY', 'HR',␣
 ↪'AR']]
season_dataframes.append(data1)
```

```
# Append the "2004-2005" DataFrame to the list of␣
 ↪season_dataframes
print(season_dataframes)
```

[ ]:

```
# Sort the DataFrame by the 'Date' column in ascending order
combined_season_data = combined_season_data.
 ↪sort_values(by='Season')

# Print the sorted DataFrame
print(combined_season_data)

#Checking datatypes of dataframe
combined_season_data.info()

#Checking size of dataframe
combined_season_data.shape

#Checking Null values
pd.isna(combined_season_data).sum()
```

Feature Extraction

[ ]:

```
# Calculate average home goals and away goals scored for␣
 ↪each team
avg_home_goals_scored = combined_season_data.
 ↪groupby('HomeTeam')['FTHG'].mean()
avg_away_goals_scored = combined_season_data.
 ↪groupby('AwayTeam')['FTAG'].mean()

# Calculate average home goals and away goals conceded for␣
 ↪each team
avg_home_goals_conceded = combined_season_data.
 ↪groupby('HomeTeam')['FTAG'].mean()
avg_away_goals_conceded = combined_season_data.
 ↪groupby('AwayTeam')['FTHG'].mean()

# Calculate league's average goals for home and away
league_avg_home_goals = combined_season_data['FTHG'].mean()
league_avg_away_goals = combined_season_data['FTAG'].mean()
```

```
# Compute Home and Away goal ratio and goal conceeded ratio␣
↪for each match
combined_season_data['HAGR'] =␣
↪combined_season_data['HomeTeam'].apply(lambda x:␣
↪avg_home_goals_scored[x] / league_avg_home_goals)
combined_season_data['AAGR'] =␣
↪combined_season_data['AwayTeam'].apply(lambda x:␣
↪avg_away_goals_scored[x] / league_avg_away_goals)

combined_season_data['HAGCR'] =␣
↪combined_season_data['HomeTeam'].apply(lambda x:␣
↪avg_home_goals_conceded[x] / league_avg_away_goals)
combined_season_data['AAGCR'] =␣
↪combined_season_data['AwayTeam'].apply(lambda x:␣
↪avg_away_goals_conceded[x] / league_avg_home_goals)

combined_season_data['NAGR'] = (combined_season_data['HAGR']␣
↪+ combined_season_data['AAGR']) / 2
combined_season_data['NAGCR'] =␣
↪(combined_season_data['HAGCR'] +␣
↪combined_season_data['AAGCR']) / 2

# Extract relevant columns
combined_season_featureextracted =␣
↪combined_season_data[['Season', 'HomeTeam','HAGR','AAGR'␣
↪,'HAGCR','AAGCR','NAGR', 'NAGCR']].drop_duplicates()

combined_season_featureextracted
```

[ ]:

```
# save dataframe into csv file in system
combined_season_data.to_csv('C:/Users/HP/Desktop/
↪DissertationPrograms/Final_Work/combined_season_data.csv')
```

Standard Deviation

```
[ ]: # Load the CSV file into a DataFrame
     combined_season_data = pd.read_csv('combined_season_data.
      ↪csv')

     # Display the first few rows of the DataFrame
     combined_season_data.head()
```

```
[ ]: df = combined_season_data.copy()

     # Calculate win percentage for each team for each game
     df['HomeWin'] = (df['FTR'] == 'H').astype(int)
     df['AwayWin'] = (df['FTR'] == 'A').astype(int)

     # Calculating win percentages for home and away games␣
      ↪separately
     home_stats = df.groupby(['Season', 'HomeTeam']).
      ↪agg(HomeGames=('HomeTeam', 'size'), HomeWins=('HomeWin',␣
      ↪'sum')).reset_index()
     away_stats = df.groupby(['Season', 'AwayTeam']).
      ↪agg(AwayGames=('AwayTeam', 'size'), AwayWins=('AwayWin',␣
      ↪'sum')).reset_index()

     # Merging home and away statistics based on season and team
     team_stats = pd.merge(home_stats, away_stats,␣
      ↪left_on=['Season', 'HomeTeam'], right_on=['Season',␣
      ↪'AwayTeam'])

     # Calculating total games played, total wins, and win␣
      ↪percentage for each team for each season
     team_stats['TotalGames'] = team_stats['HomeGames'] +␣
      ↪team_stats['AwayGames']
     team_stats['TotalWins'] = team_stats['HomeWins'] +␣
      ↪team_stats['AwayWins']
     team_stats['WPCT'] = team_stats['TotalWins'] /␣
      ↪team_stats['TotalGames']

     # Combine NAGR and NAGCR columns from the original dataframe
```

```
team_stats = pd.merge(team_stats, df[['Season', 'HomeTeam',␣
 ↪'NAGR', 'NAGCR']].drop_duplicates(), on=['Season',␣
 ↪'HomeTeam'])

# Calculate ÏČL for all seasons
ÏČL_WinPct = team_stats.groupby('Season')['WPCT'].std().
 ↪reset_index().rename(columns={'WPCT': 'ÏČL(WinPct)'})
ÏČL_NAGR = team_stats.groupby('Season')['NAGR'].std().
 ↪reset_index().rename(columns={'NAGR': 'ÏČL(NAGR)'})
ÏČL_NAGCR = team_stats.groupby('Season')['NAGCR'].std().
 ↪reset_index().rename(columns={'NAGCR': 'ÏČL(NAGCR)'})

# Merging the ÏČL values for all metrics into one dataframe
sd_table = pd.merge(ÏČL_WinPct, ÏČL_NAGR, on='Season')
sd_table = pd.merge(sd_table, ÏČL_NAGCR, on='Season')

sd_table
```

```
[ ]: plt.figure(figsize=(16, 8))

     # Plotting ÏČL for each metric across seasons
     plt.plot(sd_table['Season'], sd_table['ÏČL(WinPct)'],␣
      ↪label='ÏČL(WinPct)', marker='o')
     plt.plot(sd_table['Season'], sd_table['ÏČL(NAGR)'],␣
      ↪label='ÏČL(NAGR)', marker='o')
     plt.plot(sd_table['Season'], sd_table['ÏČL(NAGCR)'],␣
      ↪label='ÏČL(NAGCR)', marker='o')

     # Setting labels, title, and legend
     plt.xlabel('Season')
     plt.ylabel('ÏČL Value')
     plt.title('ÏČL Values Across Seasons for Different Metrics')
     plt.legend()
     plt.xticks(rotation=45)
     plt.tight_layout()

     plt.show()
```

[ ]:

```
# Using the maximum number of teams in the dataset as N for␣
↪all calculations
N = max(team_stats.groupby('Season')['HomeTeam'].nunique())

ASD_ub = np.sqrt((N + 1) / (12 * (N - 1)))

# Calculating ASD* for each metric
ASD_star_WinPct = sd_table['ÏČL(WinPct)'] / ASD_ub
ASD_star_NAGR = sd_table['ÏČL(NAGR)'] / ASD_ub
ASD_star_NAGCR = sd_table['ÏČL(NAGCR)'] / ASD_ub

# Adding the calculated values to the dataframe
sd_table['ÏČL*(WinPct)'] = ASD_star_WinPct
sd_table['ÏČL*(NAGR)'] = ASD_star_NAGR
sd_table['ÏČL*(NAGCR)'] = ASD_star_NAGCR

# Keeping only the relevant columns for clarity
sd_normalized_table = sd_table[['Season', 'ÏČL*(WinPct)',␣
↪'ÏČL*(NAGR)', 'ÏČL*(NAGCR)']]

sd_normalized_table
```

[ ]:

```
plt.figure(figsize=(16, 8))

# Plotting ÏČL for each metric across seasons
plt.plot(sd_normalized_table['Season'],␣
↪sd_normalized_table['ÏČL*(WinPct)'], label='ÏČL*(WinPct)',␣
↪marker='o')
plt.plot(sd_normalized_table['Season'],␣
↪sd_normalized_table['ÏČL*(NAGR)'], label='ÏČL*(NAGR)',␣
↪marker='o')
plt.plot(sd_normalized_table['Season'],␣
↪sd_normalized_table['ÏČL*(NAGCR)'], label='ÏČL*(NAGCR)',␣
↪marker='o')

# Setting labels, title, and legend
plt.xlabel('Season')
plt.ylabel('ÏČL* Value')
plt.title('Normalized Standard Deviations (ÏČL⋆) across␣
 ↪Seasons')
plt.legend()
plt.xticks(rotation=45)
plt.grid(True, which='both', linestyle='--', linewidth=0.5)
plt.tight_layout()

plt.show()
```

[ ]:

```
# Forcasting of Trends

from sklearn.linear_model import LinearRegression

# 1. Most Competitive Year
most_competitive_year =␣
 ↪sd_normalized_table['Season'][sd_normalized_table['ÏČL⋆(WinPct)'].
 ↪idxmax()]

# 2. Least Competitive Year
least_competitive_year =␣
 ↪sd_normalized_table['Season'][sd_normalized_table['ÏČL⋆(WinPct)'].
 ↪idxmin()]

# 3. Current Trend of Competitiveness Balance
recent_years = 5  # Last 5 seasons
recent_trend = sd_normalized_table['ÏČL⋆(WinPct)'].
 ↪tail(recent_years).values
if recent_trend[-1] > recent_trend[0]:
    current_trend = "Increasing"
elif recent_trend[-1] < recent_trend[0]:
    current_trend = "Declining"
else:
    current_trend = "Stable"

# 4. Future Trend of Competitiveness Balance using linear␣
 ↪regression
X = np.array(range(len(sd_normalized_table) - recent_years,␣
 ↪len(sd_normalized_table))).reshape(-1, 1)
```

```
y = sd_normalized_table['ÏČL⋆(WinPct)'].tail(recent_years).
 ↪values
model = LinearRegression().fit(X, y)
predicted_value = model.predict([[len(sd_normalized_table)]])

if predicted_value > y[-1]:
    future_trend = "Increasing"
elif predicted_value < y[-1]:
    future_trend = "Declining"
else:
    future_trend = "Stable"

most_competitive_year, least_competitive_year,␣
 ↪current_trend, future_trend
```

CBR

```
[ ]: # Calculate within-season-standard deviation for each metric␣
      ↪for each season
     SDws_WinPct = team_stats.groupby('Season')['WPCT'].std()
     SDws_NAGR = team_stats.groupby('Season')['NAGR'].std()
     SDws_NAGCR = team_stats.groupby('Season')['NAGCR'].std()

     # Calculate average win percentage for each team across all␣
      ↪seasons
     team_avg_WinPct = team_stats.groupby('HomeTeam')['WPCT'].
      ↪mean()

     # Calculate within-team-standard deviation for Win␣
      ↪Percentage for each team
     SDwt_WinPct = team_stats.groupby('HomeTeam')['WPCT'].
      ↪apply(lambda x: ((x - team_avg_WinPct[x.name]) ** 2).sum()␣
      ↪/ len(x))

     # Calculate average NAGR for each team across all seasons
     team_avg_NAGR = team_stats.groupby('HomeTeam')['NAGR'].mean()
```

```
# Calculate within-team-standard deviation for NAGR for each␣
 ↪team
SDwt_NAGR = team_stats.groupby('HomeTeam')['NAGR'].
 ↪apply(lambda x: ((x - team_avg_NAGR[x.name]) ** 2).sum() /␣
 ↪len(x))

# Calculate average NAGCR for each team across all seasons
team_avg_NAGCR = team_stats.groupby('HomeTeam')['NAGCR'].
 ↪mean()

# Calculate within-team-standard deviation for NAGCR for␣
 ↪each team
SDwt_NAGCR = team_stats.groupby('HomeTeam')['NAGCR'].
 ↪apply(lambda x: ((x - team_avg_NAGCR[x.name]) ** 2).sum() /
 ↪ len(x))

# Calculate CBR for each season for each metric
CBR_WinPct = SDwt_WinPct.mean() / SDws_WinPct
CBR_NAGR = SDwt_NAGR.mean() / SDws_NAGR
CBR_NAGCR = SDwt_NAGCR.mean() / SDws_NAGCR

# Create a dataframe to store the results
cbr_per_season = pd.DataFrame({
    'Season': SDws_WinPct.index,
    'CBR(WinPct)': CBR_WinPct.values,
    'CBR(NAGR)': CBR_NAGR.values,
    'CBR(NAGCR)': CBR_NAGCR.values
})

cbr_per_season
```

[ ]:
```
plt.figure(figsize=(16, 8))

# Plotting ÏČL for each metric across seasons
plt.plot(cbr_per_season['Season'],␣
 ↪cbr_per_season['CBR(WinPct)'], label='CBR(WinPct)',␣
 ↪marker='o')
```

```
plt.plot(cbr_per_season['Season'],␣
 ↪cbr_per_season['CBR(NAGR)'], label='CBR(NAGR)', marker='o')
plt.plot(cbr_per_season['Season'],␣
 ↪cbr_per_season['CBR(NAGCR)'], label='CBR(NAGCR)',␣
 ↪marker='o')

# Setting labels, title, and legend
plt.xlabel('Season')
plt.ylabel('CBR Value')
plt.title('CBR values across Seasons')
plt.legend()
plt.xticks(rotation=45)
plt.grid(True, which='both', linestyle='--', linewidth=0.5)
plt.tight_layout()

plt.show()
```

[ ]:

```
# For CBR, a lower value indicates higher competitiveness.

# 1. Most Competitive Year
most_competitive_year_cbr =␣
 ↪cbr_per_season['Season'][cbr_per_season['CBR(WinPct)'].
 ↪idxmin()]

# 2. Least Competitive Year
least_competitive_year_cbr =␣
 ↪cbr_per_season['Season'][cbr_per_season['CBR(WinPct)'].
 ↪idxmax()]

# 3. Current Trend of Competitiveness Balance
recent_trend_cbr = cbr_per_season['CBR(WinPct)'].
 ↪tail(recent_years).values
if recent_trend_cbr[-1] < recent_trend_cbr[0]:
    current_trend_cbr = "Increasing"
elif recent_trend_cbr[-1] > recent_trend_cbr[0]:
    current_trend_cbr = "Declining"
else:
    current_trend_cbr = "Stable"
```

```
# 4. Future Trend of Competitiveness Balance
X_cbr = np.array(range(len(cbr_per_season) - recent_years,␣
 ↪len(cbr_per_season))).reshape(-1, 1)
y_cbr = cbr_per_season['CBR(WinPct)'].tail(recent_years).
 ↪values
model_cbr = LinearRegression().fit(X_cbr, y_cbr)
predicted_value_cbr = model_cbr.
 ↪predict([[len(cbr_per_season)]])

if predicted_value_cbr < y_cbr[-1]:
    future_trend_cbr = "Increasing"
elif predicted_value_cbr > y_cbr[-1]:
    future_trend_cbr = "Declining"
else:
    future_trend_cbr = "Stable"

most_competitive_year_cbr, least_competitive_year_cbr,␣
 ↪current_trend_cbr, future_trend_cbr
```

HHI* and HICB

[ ]:

```
data = combined_season_data.copy()

# Creating a function to determine points for a match
def determine_points(result):
    if result == 'H':
        return 3
    elif result == 'D':
        return 1
    else:
        return 0

# Calculating points for home and away teams
data['HomePoints'] = data['FTR'].apply(lambda x:␣
 ↪determine_points(x))
data['AwayPoints'] = data['FTR'].apply(lambda x:␣
 ↪determine_points('A' if x == 'H' else ('H' if x == 'A'␣
 ↪else 'D')))
```

```
# Grouping by season and team to get the total points for␣
 ↪each team in each season
home_points = data.groupby(['Season',␣
 ↪'HomeTeam'])['HomePoints', 'HAGR', 'HAGCR'].sum().
 ↪reset_index()
away_points = data.groupby(['Season',␣
 ↪'AwayTeam'])['AwayPoints', 'AAGR', 'AAGCR'].sum().
 ↪reset_index()

# Renaming columns for merging
home_points.rename(columns={'HomeTeam': 'Team', 'HomePoints':
 ↪ 'Points', 'HAGR': 'NAGR', 'HAGCR': 'NAGCR'}, inplace=True)
away_points.rename(columns={'AwayTeam': 'Team', 'AwayPoints':
 ↪ 'Points', 'AAGR': 'NAGR', 'AAGCR': 'NAGCR'}, inplace=True)

# Merging home and away data
total_points = pd.concat([home_points, away_points]).
 ↪groupby(['Season', 'Team']).sum().reset_index()

total_points.head()


# Defining the function to calculate HHI, HHI*, and HICB
def calculate_hhi_values(season_data, column):
    N = len(season_data)
    total = season_data[column].sum()

    # Calculate the squared market shares for each team
    season_data['Squared_Share'] = (season_data[column] /␣
 ↪total) ** 2

    # HHI calculation
    hhi = season_data['Squared_Share'].sum()

    # HHI* calculation
    hhi_star = (hhi - (1 / N)) / (1 - (1 / N))
    # HICB calculation
    hicb = (hhi / (1 / N)) * 100

    return hhi, hhi_star, hicb

# Lists to store the results
seasons = []
hhi_points, hhi_nagr, hhi_nagcr = [], [], []
hhi_star_points, hhi_star_nagr, hhi_star_nagcr = [], [], []
hicb_points, hicb_nagr, hicb_nagcr = [], [], []

for season, season_data in total_points.groupby('Season'):
    seasons.append(season)

    hhi_p, hhi_star_p, hicb_p =␣
 ↪calculate_hhi_values(season_data, 'Points')
    hhi_n, hhi_star_n, hicb_n =␣
 ↪calculate_hhi_values(season_data, 'NAGR')
    hhi_c, hhi_star_c, hicb_c =␣
 ↪calculate_hhi_values(season_data, 'NAGCR')

    hhi_points.append(hhi_p)
    hhi_star_points.append(hhi_star_p)
    hicb_points.append(hicb_p)

    hhi_nagr.append(hhi_n)
    hhi_star_nagr.append(hhi_star_n)
    hicb_nagr.append(hicb_n)

    hhi_nagcr.append(hhi_c)
    hhi_star_nagcr.append(hhi_star_c)
    hicb_nagcr.append(hicb_c)

# Combining results into a DataFrame
results_df = pd.DataFrame({
    'Season': seasons,
    'HHI_Points': hhi_points,
    'HHI_NAGR': hhi_nagr,
    'HHI_NAGCR': hhi_nagcr,
    'HHI*_Points': hhi_star_points,
    'HHI*_NAGR': hhi_star_nagr,
    'HHI*_NAGCR': hhi_star_nagcr,
    'HICB_Points': hicb_points,
    'HICB_NAGR': hicb_nagr,
    'HICB_NAGCR': hicb_nagcr
})

results_df
```

[ ]:

```
# Adjust the figure size to better fit vertically stacked plots
plt.figure(figsize=(10, 12))

# Plotting HICB values
plt.subplot(2, 1, 1)
plt.plot(results_df['Season'], results_df['HICB_Points'], marker='o', linestyle='-', label='HICB_Points')
#plt.plot(results_df['Season'], results_df['HICB_NAGR'], marker='o', linestyle='-', label='HICB_NAGR')
#plt.plot(results_df['Season'], results_df['HICB_NAGCR'], marker='o', linestyle='-', label='HICB_NAGCR')
plt.xticks(rotation=45)
plt.title('HICB Values over Seasons')
plt.ylabel('HICB Value')
plt.xlabel('Season')
plt.legend()
plt.grid(True, which='both', linestyle='--', linewidth=0.5)

# Plotting HHI* values
plt.subplot(2, 1, 2)
plt.plot(results_df['Season'], results_df['HHI*_Points'], marker='o', linestyle='-', color='red', label='HHI*_Points')
plt.plot(results_df['Season'], results_df['HHI*_NAGR'], marker='o', linestyle='-', color='green', label='HHI*_NAGR')
plt.plot(results_df['Season'], results_df['HHI*_NAGCR'],␣
 ↪marker='o', linestyle='-', color='blue',␣
 ↪label='HHI*_NAGCR')
plt.xticks(rotation=45)
plt.title('HHI* Values over Seasons')
plt.ylabel('HHI* Value')
plt.xlabel('Season')
plt.legend()
plt.grid(True, which='both', linestyle='--', linewidth=0.5)

# Adjust layout
plt.tight_layout()
plt.show()
```

```
[ ]: from sklearn.linear_model import LinearRegression

     # Most and least competitive year based on HHI* for Points
     most_competitive_year_hhi_star =␣
      ↪results_df['Season'][results_df['HHI*_Points'].idxmin()]
     least_competitive_year_hhi_star =␣
      ↪results_df['Season'][results_df['HHI*_Points'].idxmax()]

     # Most and least competitive year based on HICB for Points
     most_competitive_year_hicb =␣
      ↪results_df['Season'][results_df['HICB_Points'].idxmin()]
     least_competitive_year_hicb =␣
      ↪results_df['Season'][results_df['HICB_Points'].idxmax()]

     # Current trend of competitiveness balance for HHI* for␣
      ↪Points
     recent_hhi_star_trend = results_df['HHI*_Points'].tail(5).
      ↪pct_change().mean()
     current_trend_hhi_star =␣
      ↪determine_trend(recent_hhi_star_trend)

     # Current trend of competitiveness balance for HICB for␣
      ↪Points
```

```
recent_hicb_trend = results_df['HICB_Points'].tail(5).
 ↪pct_change().mean()
current_trend_hicb = determine_trend(recent_hicb_trend)

# Forecasting future trend using Linear Regression
X = np.array(range(len(results_df))).reshape(-1, 1)

# For HHI*
model_hhi_star = LinearRegression().fit(X,␣
 ↪results_df['HHI*_Points'])
forecast_hhi_star = model_hhi_star.predict(X[-5:])  # Last 5␣
 ↪seasons
future_trend_hhi_star = determine_trend(np.mean(np.
 ↪diff(forecast_hhi_star)))

# For HICB
model_hicb = LinearRegression().fit(X,␣
 ↪results_df['HICB_Points'])
forecast_hicb = model_hicb.predict(X[-5:])  # Last 5 seasons
future_trend_hicb = determine_trend(np.mean(np.
 ↪diff(forecast_hicb)))

most_competitive_year_hhi_star,␣
 ↪least_competitive_year_hhi_star, current_trend_hhi_star,␣
 ↪future_trend_hhi_star, most_competitive_year_hicb,␣
 ↪least_competitive_year_hicb, current_trend_hicb,␣
 ↪future_trend_hicb
```

C5 and C5ICB

```
[ ]: df = combined_season_data.copy()
     # Calculating points for home and away teams in each match
     df['HomePoints'] = df.apply(lambda row: 3 if row['FTR'] ==␣
      ↪'H' else 1 if row['FTR'] == 'D' else 0, axis=1)
     df['AwayPoints'] = df.apply(lambda row: 3 if row['FTR'] ==␣
      ↪'A' else 1 if row['FTR'] == 'D' else 0, axis=1)

     # Creating a dataframe for home teams
```

```
home_df = df[['Season', 'HomeTeam', 'HomePoints']]
home_df.columns = ['Season', 'Team', 'Points']

# Creating a dataframe for away teams
away_df = df[['Season', 'AwayTeam', 'AwayPoints']]
away_df.columns = ['Season', 'Team', 'Points']

# Combining home and away dataframes
combined_df = pd.concat([home_df, away_df])

# Grouping by season and team to get the sum of points for␣
 ↪each team in each season
season_team_points = combined_df.groupby(['Season',␣
 ↪'Team'])['Points'].sum().reset_index()

# Sorting values by season and points to get top teams
season_team_points_sorted = season_team_points.
 ↪sort_values(by=['Season', 'Points'], ascending=[True,␣
 ↪False])

# Getting the top 5 teams for each season
top_5_teams_each_season = season_team_points_sorted.
 ↪groupby('Season').head(5)

# Calculating total points by top 5 teams each season
top_5_points = top_5_teams_each_season.
 ↪groupby('Season')['Points'].sum()

# Calculating total points by all teams each season
total_points = season_team_points.
 ↪groupby('Season')['Points'].sum()

# Calculating C5 Ratio
c5_ratio = top_5_points / total_points

# Calculating C5ICB
N = season_team_points['Team'].nunique()
c5_icb = (c5_ratio / (5/N)) * 100
```

```
# Combining the results
c5_results = pd.concat([c5_ratio, c5_icb], axis=1)
c5_results.columns = ['C5 Ratio', 'C5ICB']
c5_results.reset_index(inplace=True)
c5_results
```

```
[ ]: import matplotlib.pyplot as plt

     # Plotting the variations in C5 Ratio and C5ICB over the␣
      ↪seasons

     plt.figure(figsize=(15,6))

     # Plotting C5 Ratio
     plt.subplot(1, 2, 1)
     plt.plot(c5_results['Season'], c5_results['C5 Ratio'],␣
      ↪marker='o', linestyle='-')
     plt.xticks(rotation=45)
     plt.title('Variations in C5 Ratio over Seasons')
     plt.ylabel('C5 Ratio')
     plt.xlabel('Season')
     plt.grid(True, which='both', linestyle='--', linewidth=0.5)
     # Plotting C5ICB
     plt.subplot(1, 2, 2)
     plt.plot(c5_results['Season'], c5_results['C5ICB'],␣
      ↪marker='o', linestyle='-', color='green')
     plt.xticks(rotation=45)
     plt.title('Changes in C5ICB over Seasons')
     plt.ylabel('C5ICB')
     plt.xlabel('Season')
     plt.grid(True, which='both', linestyle='--', linewidth=0.5)
     # Adjust layout
     plt.tight_layout()
     plt.show()
```

```
[ ]: from sklearn.linear_model import LinearRegression
```

```
# For C5 Ratio: Lower values indicate more competitiveness.
# For C5ICB: Lower values indicate more competitiveness.

# 1. Most Competitive Year for C5 Ratio
most_competitive_year_c5 =␣
 ↪c5_results['Season'][c5_results['C5 Ratio'].idxmin()]

# Least Competitive Year for C5 Ratio
least_competitive_year_c5 =␣
 ↪c5_results['Season'][c5_results['C5 Ratio'].idxmax()]

# Current Trend of Competitiveness Balance for C5 Ratio
recent_years = 5  # Last 5 seasons
recent_trend_c5 = c5_results['C5 Ratio'].tail(recent_years).
 ↪values
if recent_trend_c5[-1] < recent_trend_c5[0]:
    current_trend_c5 = "Increasing"
elif recent_trend_c5[-1] > recent_trend_c5[0]:
    current_trend_c5 = "Declining"
else:
    current_trend_c5 = "Stable"

# Future Trend of Competitiveness Balance for C5 Ratio
X_c5 = np.array(range(len(c5_results) - recent_years,␣
 ↪len(c5_results))).reshape(-1, 1)
y_c5 = c5_results['C5 Ratio'].tail(recent_years).values
model_c5 = LinearRegression().fit(X_c5, y_c5)
predicted_value_c5 = model_c5.predict([[len(c5_results)]])

if predicted_value_c5 < y_c5[-1]:
    future_trend_c5 = "Increasing"
elif predicted_value_c5 > y_c5[-1]:
    future_trend_c5 = "Declining"
else:
    future_trend_c5 = "Stable"

# 1. Most Competitive Year for C5ICB
```

```
most_competitive_year_c5icb =␣
 ↪c5_results['Season'][c5_results['C5ICB'].idxmin()]

# Least Competitive Year for C5ICB
least_competitive_year_c5icb =␣
 ↪c5_results['Season'][c5_results['C5ICB'].idxmax()]

# Current Trend of Competitiveness Balance for C5ICB
recent_trend_c5icb = c5_results['C5ICB'].tail(recent_years).
 ↪values
if recent_trend_c5icb[-1] < recent_trend_c5icb[0]:
    current_trend_c5icb = "Increasing"
elif recent_trend_c5icb[-1] > recent_trend_c5icb[0]:
    current_trend_c5icb = "Declining"
else:
    current_trend_c5icb = "Stable"

# Future Trend of Competitiveness Balance for C5ICB
X_c5icb = np.array(range(len(c5_results) - recent_years,␣
 ↪len(c5_results))).reshape(-1, 1)
y_c5icb = c5_results['C5ICB'].tail(recent_years).values
model_c5icb = LinearRegression().fit(X_c5icb, y_c5icb)
predicted_value_c5icb = model_c5icb.
 ↪predict([[len(c5_results)]])

if predicted_value_c5icb < y_c5icb[-1]:
    future_trend_c5icb = "Increasing"
elif predicted_value_c5icb > y_c5icb[-1]:
    future_trend_c5icb = "Declining"
else:
    future_trend_c5icb = "Stable"

(most_competitive_year_c5, least_competitive_year_c5,␣
 ↪current_trend_c5, future_trend_c5,
 most_competitive_year_c5icb, least_competitive_year_c5icb,␣
 ↪current_trend_c5icb, future_trend_c5icb)
```

Gini Coefficient and Lorenz Curve

[ ]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Load the data
df = combined_season_data.copy()

# Calculate points for each match
df['HomePts'] = df['FTR'].apply(lambda x: 3 if x == 'H' else␣
 ↪(1 if x == 'D' else 0))
df['AwayPts'] = df['FTR'].apply(lambda x: 3 if x == 'A' else␣
 ↪(1 if x == 'D' else 0))

# Aggregate points for each team per season
home_pts = df.groupby(['Season', 'HomeTeam']).agg({'HomePts':
 ↪ 'sum'}).reset_index()
away_pts = df.groupby(['Season', 'AwayTeam']).agg({'AwayPts':
 ↪ 'sum'}).reset_index()
team_pts = pd.merge(home_pts, away_pts, left_on=['Season',␣
 ↪'HomeTeam'], right_on=['Season', 'AwayTeam'])
team_pts['TotalPts'] = team_pts['HomePts'] +␣
 ↪team_pts['AwayPts']
team_pts = team_pts[['Season', 'HomeTeam', 'TotalPts']]

# Function to calculate Gini coefficient
def gini_coefficient(x):
    n = len(x)
    s = 0
    for i in range(n):
        s += (i+1) * x[i]
    return 1 - 2 * (s / sum(x) - (n + 1) / 2) / n

# Calculate Lorenz curve and Gini coefficient for each season
gini_values = []

for season in team_pts['Season'].unique():
    season_data = team_pts[team_pts['Season'] == season].
↪sort_values(by='TotalPts')
    season_data['CumulativePts'] = season_data['TotalPts'].
↪cumsum()
    season_data['CumulativePct'] =␣
↪season_data['CumulativePts'] / season_data['TotalPts'].
↪sum()
    season_data['TeamPct'] = (season_data.index + 1) /␣
↪len(season_data)

    # Compute Gini coefficient
    gini = gini_coefficient(season_data['CumulativePct'].
↪values)
    gini_values.append((season, gini))

gini_df = pd.DataFrame(gini_values, columns=['Season',␣
↪'GiniCoefficient'])

gini_df
```

[ ]:

```
# Filter the data for the seasons 2002-2012
filtered_seasons = ["2002-2003", "2003-2004", "2004-2005",␣
↪"2005-2006", "2006-2007", "2007-2008",
                    "2008-2009", "2009-2010", "2010-2011",␣
↪"2011-2012"]

filtered_data =␣
↪combined_season_data[combined_season_data['Season'].
↪isin(filtered_seasons)]

# Redefine the function to compute Lorenz curve
def compute_lorenz_curve(points):
    sorted_points = np.sort(points)
    total_points = sorted_points.sum()
    cumulative_points = np.cumsum(sorted_points) /␣
↪total_points
    cumulative_teams = np.arange(1, len(sorted_points) + 1) /
↪ len(sorted_points)
```

```
    return cumulative_teams, cumulative_points

# Calculate the total points for each team per season
def calculate_team_points(df):
    # 3 points for a win, 1 point for a draw
    df['HomePts'] = np.where(df['FTR'] == 'H', 3, np.
 ↪where(df['FTR'] == 'D', 1, 0))
    df['AwayPts'] = np.where(df['FTR'] == 'A', 3, np.
 ↪where(df['FTR'] == 'D', 1, 0))

    # Aggregate points for each team
    home_points = df.groupby(['Season',␣
 ↪'HomeTeam'])['HomePts'].sum().reset_index()
    away_points = df.groupby(['Season',␣
 ↪'AwayTeam'])['AwayPts'].sum().reset_index()

    home_points.rename(columns={'HomeTeam': 'Team',␣
 ↪'HomePts': 'Points'}, inplace=True)
    away_points.rename(columns={'AwayTeam': 'Team',␣
 ↪'AwayPts': 'Points'}, inplace=True)

    total_points = pd.concat([home_points, away_points],␣
 ↪axis=0)
    total_points = total_points.groupby(['Season',␣
 ↪'Team'])['Points'].sum().reset_index()

    return total_points

team_points = calculate_team_points(combined_season_data)

# Filter for seasons 2002-2012
team_points_filtered = team_points[team_points['Season'].
 ↪isin(filtered_seasons)]

# Compute the Lorenz curve values for the filtered seasons␣
 ↪using the points data
lorenz_curves = {}
for season in filtered_seasons:
    season_data =␣
↪team_points_filtered[team_points_filtered['Season'] ==␣
↪season]
    lorenz_x, lorenz_y =␣
↪compute_lorenz_curve(season_data['Points'])
    lorenz_curves[season] = (lorenz_x, lorenz_y)

# Plotting
plt.figure(figsize=(14, 10))
for season, (lx, ly) in lorenz_curves.items():
    plt.plot(lx, ly, label=season)

# Plotting the line of equality
plt.plot([0, 1], [0, 1], color='black', linestyle='--',␣
↪label="Line of Equality")

plt.title('Lorenz Curve for Seasons 2002-2012')
plt.xlabel('Cumulative Proportion of Teams')
plt.ylabel('Cumulative Proportion of Points')
plt.legend()
plt.grid(True, which='both', linestyle='--', linewidth=0.5)
plt.tight_layout()
plt.show()
```

[ ]:

```
# Filter the data for the seasons 2002-2012
filtered_seasons = ['2013-2014', '2014-2015', '2015-2016',␣
↪'2016-2017', '2017-2018', '2018-2019', '2019-2020',␣
↪'2020-2021', '2021-2022', '2022-2023']



filtered_data =␣
↪combined_season_data[combined_season_data['Season'].
↪isin(filtered_seasons)]

# define the function to compute Lorenz curve
def compute_lorenz_curve(points):
    sorted_points = np.sort(points)
    total_points = sorted_points.sum()
    cumulative_points = np.cumsum(sorted_points) /␣
↪total_points
    cumulative_teams = np.arange(1, len(sorted_points) + 1) /
↪ len(sorted_points)
    return cumulative_teams, cumulative_points

# Calculate the total points for each team per season
def calculate_team_points(df):
    # 3 points for a win, 1 point for a draw
    df['HomePts'] = np.where(df['FTR'] == 'H', 3, np.
↪where(df['FTR'] == 'D', 1, 0))
    df['AwayPts'] = np.where(df['FTR'] == 'A', 3, np.
↪where(df['FTR'] == 'D', 1, 0))

    # Aggregate points for each team
    home_points = df.groupby(['Season',␣
↪'HomeTeam'])['HomePts'].sum().reset_index()
    away_points = df.groupby(['Season',␣
↪'AwayTeam'])['AwayPts'].sum().reset_index()

    home_points.rename(columns={'HomeTeam': 'Team',␣
↪'HomePts': 'Points'}, inplace=True)
    away_points.rename(columns={'AwayTeam': 'Team',␣
↪'AwayPts': 'Points'}, inplace=True)

    total_points = pd.concat([home_points, away_points],␣
↪axis=0)
    total_points = total_points.groupby(['Season',␣
↪'Team'])['Points'].sum().reset_index()

    return total_points

team_points = calculate_team_points(combined_season_data)

# Filter for seasons 2002-2012
team_points_filtered = team_points[team_points['Season'].
↪isin(filtered_seasons)]
```

```
# Compute the Lorenz curve values for the filtered seasons␣
 ↪using the points data
lorenz_curves = {}
for season in filtered_seasons:
    season_data =␣
 ↪team_points_filtered[team_points_filtered['Season'] ==␣
 ↪season]
    lorenz_x, lorenz_y =␣
 ↪compute_lorenz_curve(season_data['Points'])
    lorenz_curves[season] = (lorenz_x, lorenz_y)

# Plotting
plt.figure(figsize=(14, 10))
for season, (lx, ly) in lorenz_curves.items():
    plt.plot(lx, ly, label=season)

# Plotting the line of equality
plt.plot([0, 1], [0, 1], color='black', linestyle='--',␣
 ↪label="Line of Equality")

plt.title('Lorenz Curve for Seasons 2013-2023')
plt.xlabel('Cumulative Proportion of Teams')
plt.ylabel('Cumulative Proportion of Points')
plt.legend()
plt.grid(True, which='both', linestyle='--', linewidth=0.5)
plt.tight_layout()
plt.show()
```

[ ]:

```
# Most and least competitive year based on Gini Coefficient
most_competitive_year_gini =␣
 ↪gini_df['Season'][gini_df['GiniCoefficient'].idxmin()]
least_competitive_year_gini =␣
 ↪gini_df['Season'][gini_df['GiniCoefficient'].idxmax()]

# Current trend of competitiveness balance for Gini␣
 ↪Coefficient
# Calculate the percentage change over the last 5 years to␣
 ↪determine the trend
```

```
recent_gini_trend = gini_df['GiniCoefficient'].tail(5).
 ↪pct_change().mean()

# Predicting future trend using linear regression for Gini␣
 ↪Coefficient
from sklearn.linear_model import LinearRegression

X = np.array(range(len(gini_df))).reshape(-1, 1)  # Using␣
 ↪the index as a time variable
y = gini_df['GiniCoefficient'].values
reg = LinearRegression().fit(X, y)
future_gini_trend_slope = reg.coef_[0]

most_competitive_year_gini, least_competitive_year_gini,␣
 ↪recent_gini_trend, future_gini_trend_slope
```

Relative Entropy

```
[ ]: # Calculate win percentage for each team for each game
     df['HomeWin'] = (df['FTR'] == 'H').astype(int)
     df['AwayWin'] = (df['FTR'] == 'A').astype(int)

     # Calculating win percentages for home and away games␣
      ↪separately
     home_stats = df.groupby(['Season', 'HomeTeam']).
      ↪agg(HomeGames=('HomeTeam', 'size'), HomeWins=('HomeWin',␣
      ↪'sum')).reset_index()
     away_stats = df.groupby(['Season', 'AwayTeam']).
      ↪agg(AwayGames=('AwayTeam', 'size'), AwayWins=('AwayWin',␣
      ↪'sum')).reset_index()

     # Merging home and away statistics based on season and team
     team_stats = pd.merge(home_stats, away_stats,␣
      ↪left_on=['Season', 'HomeTeam'], right_on=['Season',␣
      ↪'AwayTeam'])

     # Calculating total games played, total wins, and win␣
      ↪percentage for each team for each season
```

```
team_stats['TotalGames'] = team_stats['HomeGames'] +␣
 ↪team_stats['AwayGames']
team_stats['TotalWins'] = team_stats['HomeWins'] +␣
 ↪team_stats['AwayWins']
team_stats['WPCT'] = team_stats['TotalWins'] /␣
 ↪team_stats['TotalGames']

# Combine NAGR and NAGCR columns from the original dataframe
team_stats = pd.merge(team_stats, df[['Season', 'HomeTeam',␣
 ↪'NAGR', 'NAGCR']].drop_duplicates(), on=['Season',␣
 ↪'HomeTeam'])
```

[ ]:

```
df = combined_season_data.copy()

# Calculating Relative Entropy for Win Percentage, NAGR, and␣
 ↪NAGCR

def calculate_relative_entropy(df, column):
    """Calculate the relative entropy for a given column."""
    # Calculate pi (proportion of total wins or metric value␣
 ↪in the league season for team i)
    df['pi'] = df[column] / df.groupby('Season')[column].
 ↪transform('sum')

    # Calculate E (actual entropy)
    df['E'] = -df['pi'] * np.log2(df['pi'])
    E_values = df.groupby('Season')['E'].sum().reset_index()

    # Calculate Emax (maximum possible entropy)
    N = df.groupby('Season')['HomeTeam'].nunique().
 ↪reset_index(name='N')
    Emax_values = -np.log2(1/N['N'])

    # Calculate Relative Entropy R
    R = E_values['E'] / Emax_values

    return R
```

```
# Calculate Relative Entropy for each metric
R_WinPct = calculate_relative_entropy(team_stats, 'WPCT')
R_NAGR = calculate_relative_entropy(team_stats, 'NAGR')
R_NAGCR = calculate_relative_entropy(team_stats, 'NAGCR')

# Combine the results into a DataFrame
relative_entropy_df = pd.DataFrame({
    'Season': team_stats['Season'].unique(),
    'RE(WinPct)': R_WinPct,
    'RE(NAGR)': R_NAGR,
    'RE(NAGCR)': R_NAGCR
})

relative_entropy_df
```

```
[ ]: # Plotting Relative Entropy values for each metric across␣
      ↪seasons
     plt.figure(figsize=(16, 8))

     plt.plot(relative_entropy_df['Season'],␣
      ↪relative_entropy_df['RE(WinPct)'], label='RE(WinPct)',␣
      ↪marker='o')
     plt.plot(relative_entropy_df['Season'],␣
      ↪relative_entropy_df['RE(NAGR)'], label='RE(NAGR)',␣
      ↪marker='o')
     plt.plot(relative_entropy_df['Season'],␣
      ↪relative_entropy_df['RE(NAGCR)'], label='RE(NAGCR)',␣
      ↪marker='o')

     # Setting labels, title, and legend
     plt.xlabel('Season')
     plt.ylabel('Relative Entropy Value')
     plt.title('Relative Entropy across Seasons')
     plt.legend()
     plt.xticks(rotation=45)
     plt.grid(True, which='both', linestyle='--', linewidth=0.5)
     plt.tight_layout()
     plt.show()
```

```
[ ]: from sklearn.linear_model import LinearRegression

     # Assigning trend labels
     def determine_trend(value):
         if value > 0.01:  # Threshold of 1% used to determine if␣
      ↪value is effectively increasing
             return "Increasing"
         elif value < -0.01:  # Threshold of -1% used to␣
      ↪determine if value is effectively decreasing
             return "Declining"
         else:
             return "Stable"

     # Most and least competitive year based on Relative Entropy␣
      ↪for WinPct
     most_competitive_year_RE =␣
      ↪relative_entropy_df['Season'][relative_entropy_df['RE(WinPct)'].
      ↪idxmax()]
     least_competitive_year_RE =␣
      ↪relative_entropy_df['Season'][relative_entropy_df['RE(WinPct)'].
      ↪idxmin()]

     # Current trend of competitiveness balance for Relative␣
      ↪Entropy for WinPct
     recent_RE_trend = relative_entropy_df['RE(WinPct)'].tail(5).
      ↪pct_change().mean()

     # Determining the trend
     current_trend_RE = determine_trend(recent_RE_trend)

     # Modifying the code to forecast using linear regression for␣
      ↪Relative Entropy for WinPct

     # Setting up the data for linear regression
     X = np.array(range(len(relative_entropy_df))).reshape(-1, 1)

     # Fitting linear regression to the data
```

```
model_RE = LinearRegression().fit(X,␣
 ↪relative_entropy_df['RE(WinPct)'])
forecast_RE = model_RE.predict(X[-5:])  # Forecast for the␣
 ↪last 5 seasons

# Using the mean percentage change of the forecast to␣
 ↪predict future trend
future_trend_RE_linear = determine_trend(np.mean(np.
 ↪diff(forecast_RE)))

most_competitive_year_RE, least_competitive_year_RE,␣
 ↪current_trend_RE, future_trend_RE_linear
```

Comparative Analysis

NSD and Contentration Ratio

```
[ ]: # 1. Standard Deviation & Normalized Standard Deviation:
     recent_sd_normalized = sd_normalized_table.tail(10)

     # 2. Concentration Ratios:
     recent_c5_results = c5_results.tail(10)

     # 3. League Table Analysis:
     recent_seasons = combined_season_data['Season'].unique()[-10:
      ↪]

     # Calculate total points for each team in the recent seasons
     recent_team_points = team_points[team_points['Season'].
      ↪isin(recent_seasons)]

     # Sort to get the top performing teams
     recent_team_points_sorted = recent_team_points.
      ↪sort_values(by=['Season', 'Points'], ascending=[True,␣
      ↪False])

     # Getting the top 5 teams for each recent season
     top_5_teams_recent_seasons = recent_team_points_sorted.
      ↪groupby('Season').head(5)
```

```
[ ]: # Setting up the plots
fig, ax1 = plt.subplots(figsize=(14, 7))

# Twin the axes for two different y-axes
ax2 = ax1.twinx()

# Plotting Normalized Standard Deviations on ax1
ax1.plot(recent_sd_normalized['Season'],␣
 ↪recent_sd_normalized['ÏČL⋆(WinPct)'], 'g-', marker='o',␣
 ↪label='ÏČL⋆(WinPct)')
ax1.plot(recent_sd_normalized['Season'],␣
 ↪recent_sd_normalized['ÏČL⋆(NAGR)'], 'b-', marker='o',␣
 ↪label='ÏČL⋆(NAGR)')
ax1.plot(recent_sd_normalized['Season'],␣
 ↪recent_sd_normalized['ÏČL⋆(NAGCR)'], 'r-', marker='o',␣
 ↪label='ÏČL⋆(NAGCR)')

# Plotting C5 Ratio on ax2
ax2.plot(recent_c5_results['Season'], recent_c5_results['C5␣
 ↪Ratio'], 'm--', marker='x', label='C5 Ratio')
ax2.set_ylim(0, 1)  # Setting the limit for C5 Ratio for␣
 ↪clarity

# Setting labels and titles
ax1.set_xlabel('Season')
ax1.set_ylabel('Normalized Standard Deviation',␣
 ↪color='black')
ax2.set_ylabel('C5 Ratio', color='black')
ax1.set_title('Comparison of Normalized Standard Deviation␣
 ↪and C5 Ratio for Recent Seasons')

# Display legends
ax1.legend(loc='upper left')
ax2.legend(loc='upper right')

plt.grid(True, which='both', linestyle='--', linewidth=0.5)
plt.tight_layout()
plt.show()
```

NSD and HICB

```
[ ]: # Extracting the recent data for normalized standard␣
      ↪deviation and HICB from the previously computed dataframes
     recent_sd_normalized_subset =␣
      ↪recent_sd_normalized[['Season', 'ÏČL∗(WinPct)',␣
      ↪'ÏČL∗(NAGR)', 'ÏČL∗(NAGCR)']]
     recent_hicb_subset = recent_hicb[['Season', 'HICB_Points',␣
      ↪'HICB_NAGR', 'HICB_NAGCR']]

     # Merging the datasets for comparison
     sd_hicb_comparison = pd.merge(recent_sd_normalized_subset,␣
      ↪recent_hicb_subset, on='Season')

     # Setting up the plots
     fig, ax1 = plt.subplots(figsize=(14, 7))

     # Plotting normalized standard deviation values on ax1
     ax1.plot(sd_hicb_comparison['Season'],␣
      ↪sd_hicb_comparison['ÏČL∗(WinPct)'], 'g-', marker='o',␣
      ↪label='ÏČL∗(WinPct)')
     ax1.plot(sd_hicb_comparison['Season'],␣
      ↪sd_hicb_comparison['ÏČL∗(NAGR)'], 'b-', marker='o',␣
      ↪label='ÏČL∗(NAGR)')
     ax1.plot(sd_hicb_comparison['Season'],␣
      ↪sd_hicb_comparison['ÏČL∗(NAGCR)'], 'r-', marker='o',␣
      ↪label='ÏČL∗(NAGCR)')

     # Twin the axes for two different y-axes
     ax2 = ax1.twinx()

     # Plotting HICB on ax2
     ax2.plot(sd_hicb_comparison['Season'],␣
      ↪sd_hicb_comparison['HICB_Points'], 'g--', marker='x',␣
      ↪label='HICB(WinPct)')
     ax2.plot(sd_hicb_comparison['Season'],␣
      ↪sd_hicb_comparison['HICB_NAGR'], 'b--', marker='x',␣
      ↪label='HICB(NAGR)')
```

```
ax2.plot(sd_hicb_comparison['Season'],␣
 ↪sd_hicb_comparison['HICB_NAGCR'], 'r--', marker='x',␣
 ↪label='HICB(NAGCR)')

# Setting labels and titles
ax1.set_xlabel('Season')
ax1.set_ylabel('ÏČL*', color='black')
ax2.set_ylabel('HICB', color='black')
ax1.set_title('Comparison of Normalized Standard Deviation␣
 ↪(ÏČL*) and HICB for Recent Seasons')

# Display legends
ax1.legend(loc='upper left')
ax2.legend(loc='upper right')

plt.grid(True, which='both', linestyle='--', linewidth=0.5)
plt.tight_layout()
plt.show()
```

HICB and C5ICB

```
[ ]: # Extracting the recent data for HICB and C5ICB from the␣
      ↪previously computed dataframes
     hicb_c5icb_comparison = pd.merge(recent_hicb,␣
      ↪recent_c5_results, on='Season')

     # Setting up the plots
     fig, ax1 = plt.subplots(figsize=(14, 7))

     # Plotting HICB values on ax1
     ax1.plot(hicb_c5icb_comparison['Season'],␣
      ↪hicb_c5icb_comparison['HICB_Points'], 'g-', marker='o',␣
      ↪label='HICB(WinPct)')
     ax1.plot(hicb_c5icb_comparison['Season'],␣
      ↪hicb_c5icb_comparison['HICB_NAGR'], 'b-', marker='o',␣
      ↪label='HICB(NAGR)')
     ax1.plot(hicb_c5icb_comparison['Season'],␣
      ↪hicb_c5icb_comparison['HICB_NAGCR'], 'r-', marker='o',␣
      ↪label='HICB(NAGCR)')
```

```
# Twin the axes for two different y-axes
ax2 = ax1.twinx()

# Plotting C5ICB on ax2
ax2.plot(hicb_c5icb_comparison['Season'],␣
 ↪hicb_c5icb_comparison['C5ICB'], 'k--', marker='x',␣
 ↪label='C5ICB')

# Setting labels and titles
ax1.set_xlabel('Season')
ax1.set_ylabel('HICB', color='black')
ax2.set_ylabel('C5ICB', color='black')
ax1.set_title('Comparison of HICB and C5ICB for Recent␣
 ↪Seasons')

# Display legends
ax1.legend(loc='upper left')
ax2.legend(loc='upper right')

plt.grid(True, which='both', linestyle='--', linewidth=0.5)
plt.tight_layout()
plt.show()
```