Project Report

Data Analysis

Comprehensive Analysis For The Football Transfer Market

Groupe 14 :

Zakaria El Mrani
Eya Meski
Gaith Labidi
Halimata N'Diaye
Yasmine Ben yaala

---

ensIIE

Contents

List of Figures

Introduction

The football transfer market is a complex and dynamic ecosystem where player transfer fees are influenced by various factors, including market value, age, performance, club negotiations, and even market demand. Understanding the intricate relationships between these variables is crucial for analysts, clubs, and stakeholders aiming to make informed decisions. This study employs a combination of advanced analytical techniques—Factorial Correspondence Analysis (FCA), Canonical Correspondence Analysis (CCA), Principal Component Analysis (PCA), and Clustering—to provide insights into the drivers of transfer fees and evaluate the limitations of market value as a sole predictor.
In order to do this we used Analytical Tools which are :

- Factorial Correspondence Analysis (FCA): :
  FCA is a dimensionality reduction tool ideal for categorical data. It helps visualize relationships between categories of variables like transfer_fee and market_value_in_eur and perfermance and market_value.

- Canonical Correspondence Analysis (CCA):
  CCA explores the relationship between sets of variables, such as player attributes (e.g., age, performance) and outcomes (e.g., transfer fees). This method identifies the variables that significantly drive transfer fees and quantifies their contributions.

- Principal Component Analysis (PCA):
  PCA reduces the dimensionality of continuous variables, uncovering key components that explain the most variance in the dataset. By visualizing principal components, PCA identifies the primary factors influencing transfer fees and how they relate to one another.

- Clustering Analysis:
  Clustering methods (e.g., K-Means or Hierarchical Clustering) group players into segments based on features like age, performance, market value, and transfer fees. These clusters help uncover hidden patterns, such as identifying young players with high potential or overvalued players.

Chapter 1

Data Visualization

Introduction

The dataset, derived from multiple sources captures various aspects of player performance, match statistics, and transfer activities.Visualization techniques provide initial insights into the data structure and content.

1.1 Dataset Description

We have used the following files:

- **players.csv**: Player details such as positions, ages, and market values.
- **appearances.csv**: Player appearances with statistics like goals, assists, and minutes played.
- **transfers.csv**: Player transfer history, including fees and destination clubs.
- **games.csv**: Match details such as dates, teams, and outcomes.
- **clubs.csv**: Provides information about the clubs, including their identities and budgets.
- **club_games.csv**: Contains match records specific to each club, including the club's performance in the game.

Together, these files enable a comprehensive exploration of player performance, match dynamics, and market trends.

1.2 Data Visualization

Number of Columns in Each File

The chart below highlights the number of columns in each file.

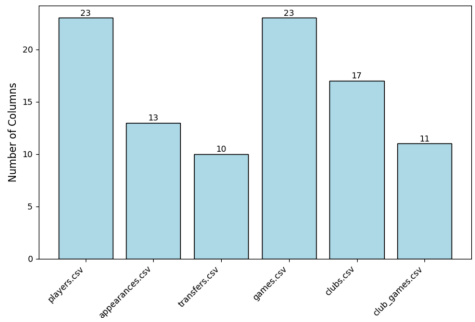

Figure 1.1: Number of Columns in Each File

The `players.csv` file, with the most columns (23), contains extensive player information. In contrast, the `transfers.csv` files have fewer columns, reflecting simpler structures.

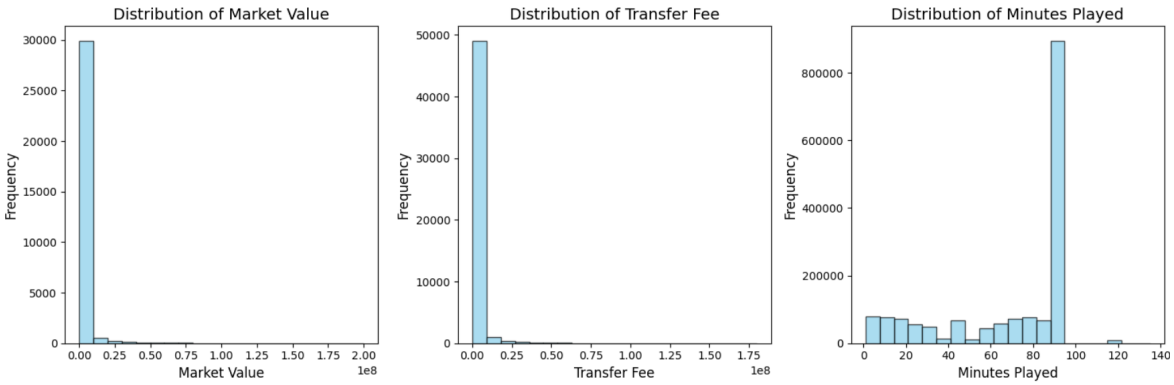

Figure 1.2: Histograms of Transfer fee, Minutes Played, and Market Value in EUR.

Distribution of Selected Variables

The following histograms illustrate the distributions of key variables from the dataset:

- Transfer fee: The distribution is highly skewed, with the majority of players valued at less than €25 million..
- Minutes Played: A sharp peak is observed around 90 minutes, suggesting standard full-game duration.
- Market Value in EUR: The distribution is heavily skewed, with most players valued below €25 million.

Conclusion

This chapter introduced the dataset and visualized its structure and key variables, providing a foundation for the detailed analyses in subsequent chapters.

Chapter 2

Overview of the Transfer Market Analysis

2.1 Objective

This analysis seeks to understand the dynamics between player market values and transfer fees in football, using a dataset (transfers.csv) of player transfers. The primary focus is to determine whether market value alone can sufficiently explain transfer fees or if other factors (e.g., player age, performance) significantly contribute.

2.2 Data Exploration and Problem Identification

The initial step involves examining the dataset to understand its structure and identify potential issues such as missing data, inconsistencies, or outliers. This includes:
-Checking the completeness of data in key fields like transfer_fee and market_value_in_eur.
-Comparing transfer fees with market values to compute a Fee-to-Value Ratio, which serves as an indicator of whether players are overpaid or underpaid

From this, we identify a core problem: market values do not always align with transfer fees, suggesting the presence of other influential factors.

2.3 Insights Into Overpaid and Underpaid Players

By analyzing the Fee-to-Value Ratio, we identify:

Overpaid Players

hose whose transfer fees significantly exceed their market value.

Underpaid Players

Those transferred for fees much lower than their market value.

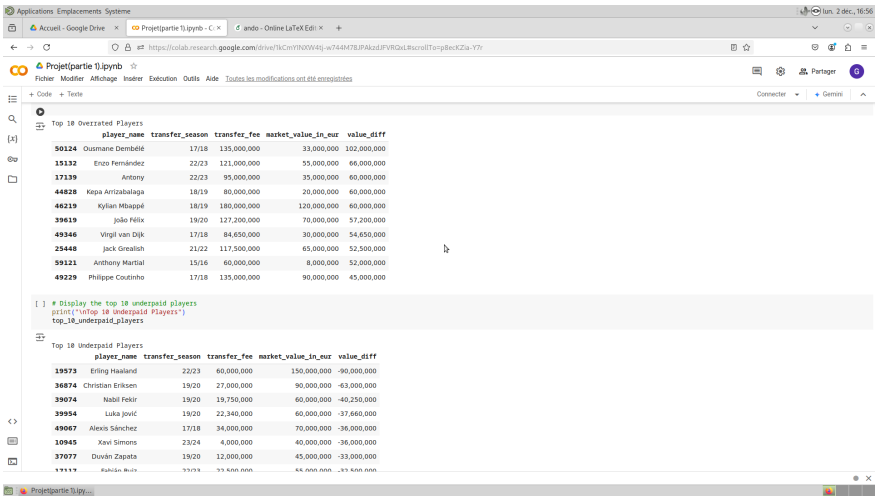

Figure 2.1: overpaid underpaid players

This step emphasizes real-world anomalies in the transfer market, often driven by factors such as hype, negotiation power, or market dynamics.

2.4 Factorial Correspondence Analysis (FCA)

Categorizing Data

- Both transfer_fee and market_value_in_eur are divided into categories (e.g., Low, Medium, High) for clearer analysis.
- A contingency table is created to count occurrences of each combination of fee and value categories.

Analyzing Relationships

- FCA identifies key dimensions explaining the variance in the data.
- the relationship between market value categories and transfer fee categories is visualized.

Findings

- The analysis reveals that market value explains only a part of the variance in transfer fees, with a weak correlation between the two.
- This suggests that market value, while important, is insufficient as the sole determinant of transfer fees.

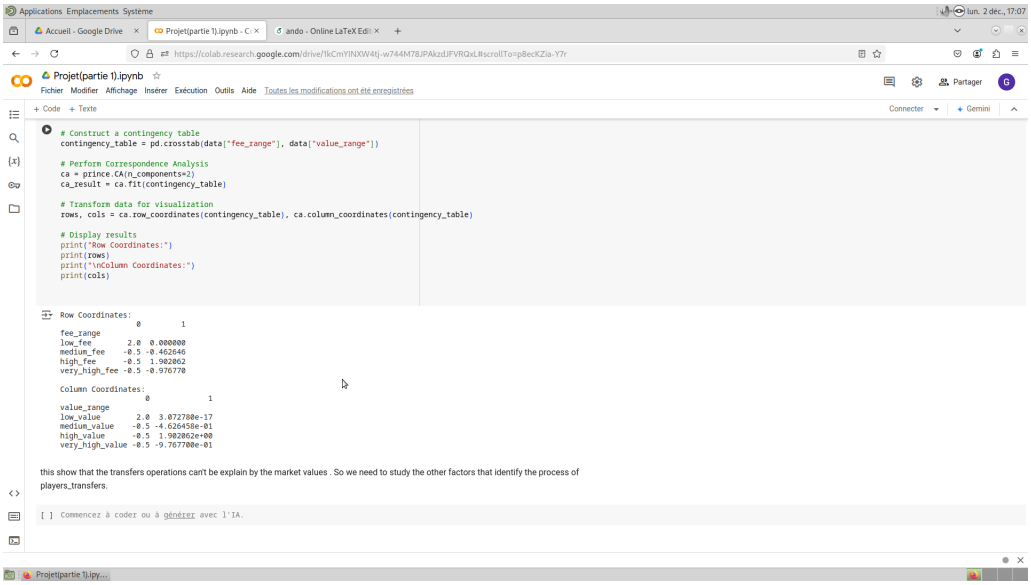

Figure 2.2: findings

Conclusion

The analysis demonstrates that market value alone cannot explain transfer fees, as evidenced by the FCA and Fee-to-Value Ratio results. Other influential factors—such as age, performance, and club strategies—must be integrated into valuation models.

Chapter 3

PCA Analysis and Player Performance Clustering

Context and Objectives

In this analysis, we aimed to understand the performance of football players based on key metrics, including **minutes played, goals scored, assists**, and **disciplinary actions** (yellow/red cards). Additionally, we sought to explore how these performances correlate with **market value**. To achieve this, we leveraged **Principal Component Analysis (PCA)** and **clustering techniques** to analyze and visualize the data, using multiple graphs for interpretation.

The dataset used included player appearances, performance metrics, and transfer fee data sourced from the `appearances.csv` and `players.csv` files. By combining these datasets, we obtained a comprehensive view of each player's performance and position, enabling a deeper understanding of performance trends and patterns.

3.1 PCA Explained Variance

Why this step?

PCA reduces high-dimensional data to fewer components, capturing the majority of the variance and simplifying analysis. By focusing on the most informative components, we retain critical insights while discarding redundant information.

Results

The explained variance analysis showed:

- **PC1 accounts for 23.7% of the variance**, representing the most critical axis for player performance.
- **PC2 explains 20.6%**, with other components contributing less.
- Together, the components account for **100% of the variance**, suggesting that all variables play a significant role in the overall data structure, even though PC1 and PC2 are more influential.

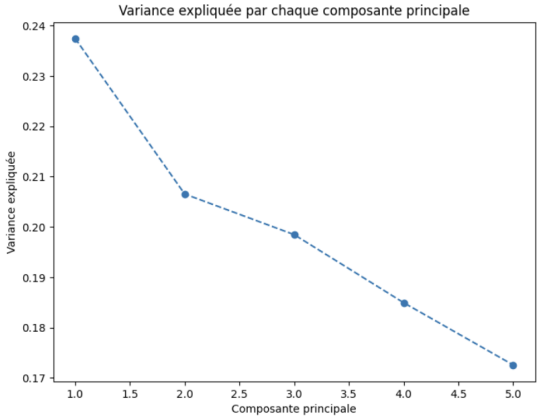

Figure 3.1: Explained Variance Analysis for PCA Components

This observation underscores the importance of each metric in shaping the dataset's variability, reflecting a balance in contributions across metrics.

3.2 Projection of Players by Position

Why this step?

Projecting the dataset onto the first two principal components (PC1 and PC2) allows us to visualize patterns and clusters in a simplified 2D space. By categorizing players based on their positions (e.g., forwards, defenders), we can assess whether playing roles correspond to distinct performance profiles.

Results

- **Forwards**: High contributions along PC1, reflecting offensive metrics such as goals and assists.
- **Midfielders**: Spread across both components, indicating versatility in contributions.
- **Defenders and goalkeepers**: Cluster differently, emphasizing their defensive and time-on-field metrics.

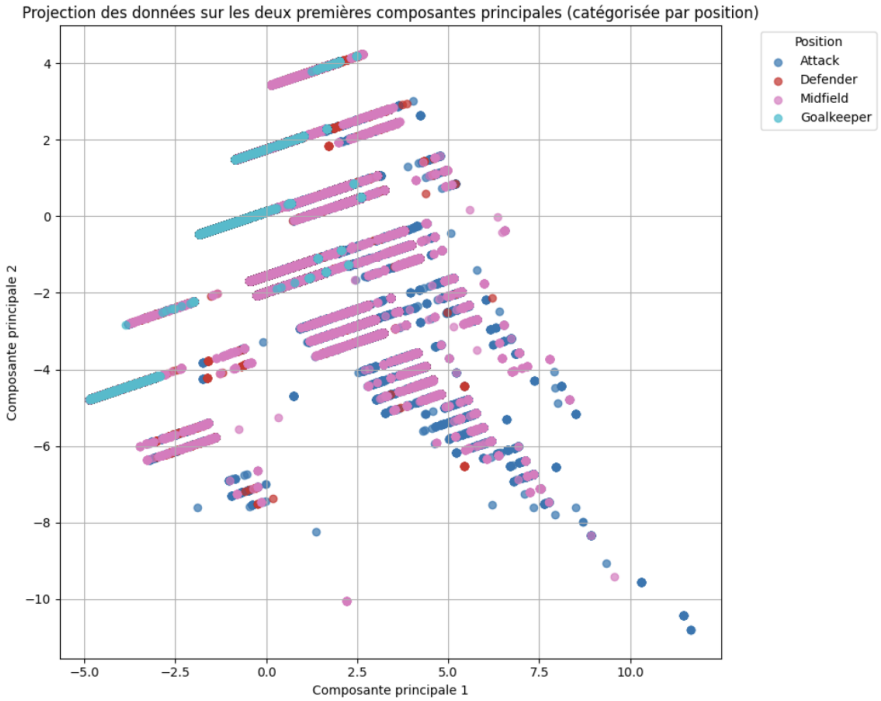

Figure 3.2: PCA Projections by Player Position

This step highlights the effectiveness of PCA in visualizing performance trends and supports the hypothesis that player roles influence performance patterns.

3.3 Player Clustering Using K-Means

Why this step?

While PCA provides insights into positional trends, it doesn't group players explicitly based on overall performance profiles. To address this, we applied **K-Means clustering** to group players with similar characteristics into performance-based clusters. This step aims to:

- Identify homogeneous groups of players.
- Provide actionable insights for team managers and scouts.

Results

- **Cluster 1**: High-performing players excelling in goals and assists.
- **Cluster 2**: Consistent contributors across metrics.
- **Cluster 3**: Defensive players with lower offensive contributions.

- **Cluster 4**: Players with limited playing time or underwhelming metrics.

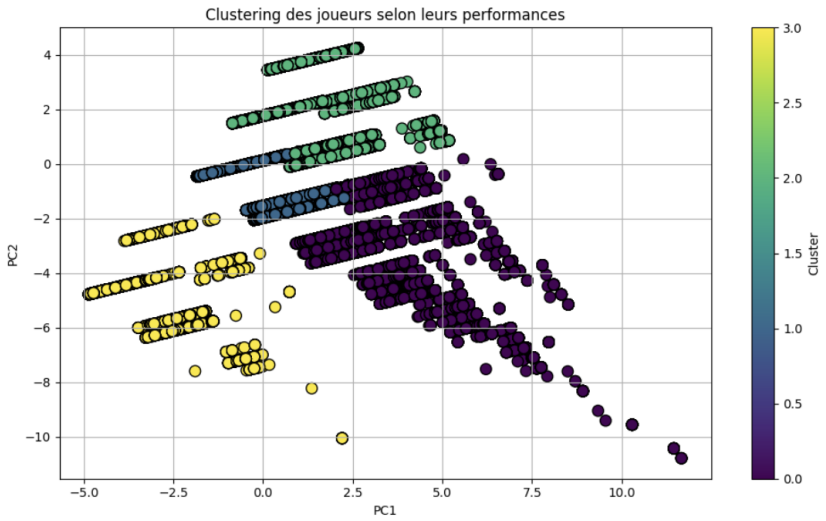

Figure 3.3: Player Clustering Using K-Means (k=4)

3.4 Relationship Between Performance and Transfer Fee

Why this step?

Performance metrics alone are insufficient without considering their financial implications. By analyzing the relationship between composite performance scores and transfer fee, we aim to:

- Understand how performance translates to monetary value.
- Identify undervalued or overvalued players for investment opportunities.

Results

The scatter plot of performance vs. transfer fee shows:

- A weak positive correlation: Higher-performing players typically have higher transfer fees.
- Notable outliers:
  - **Undervalued players**: High performance but low transfer fee.
  - **Overvalued players**: High transfer fee despite moderate performance.

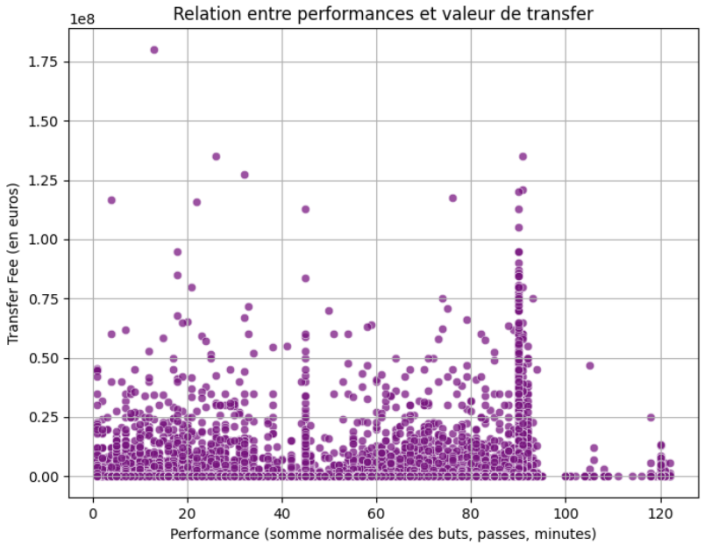

Figure 3.4: Performance vs. Transfer Fee

Conclusion

This study demonstrates the utility of PCA and clustering in simplifying performance data and revealing actionable insights:

- PCA highlighted positional trends and the balanced importance of all metrics in explaining variability.
- Clustering provided nuanced player groupings beyond official roles.
- Transfer Fee shows weak correlations with PCA components, indicating the need for additional techniques to capture external influences.

  Future analyses could incorporate defensive metrics or longitudinal data to refine the models further, contributing to more robust player evaluation frameworks in sports analytics.

Chapter 4

Relationship between performance and transfer fee using FCA

Introduction

In this chapter, we explore the relationship between **player performance** and **transfer fee** using **Factorial Correspondence Analysis (FCA)**. This method allows us to examine the association between key performance metrics such as **minutes played, goals scored, assists**,and the transfer fee of players.

4.1 The Role of Performance Metrics in Determining Transfer Fee

To further investigate the determinants of transfer fees, we performed an FCA analysis between player performance metrics and their transfer fees. The objective was to evaluate whether a stronger relationship exists between these variables compared to the market value data.

The performance metrics considered in this analysis include:

- Number of goals scored
- Assists provided
- Minutes played

4.1.1 Methodology

The analysis followed a similar approach to the previous FCA study. Performance metrics and transfer fees were categorized and structured into a contingency table to evaluate their associations. Singular value decomposition (SVD) was applied to identify the primary dimensions of variance.

4.1.2 Findings

The scatter plots shown in Figures 4.1 and 4.2 and 4.3 represent the results of the FCA for the positions of Defender, Attack and Midfielder, respectively, across the years 2021, 2022, and 2023. Each point in the scatter plot corresponds to a category of either performance or tranfer fee. The blue points represent the performance categories, while the red points correspond to transfer fee categories.

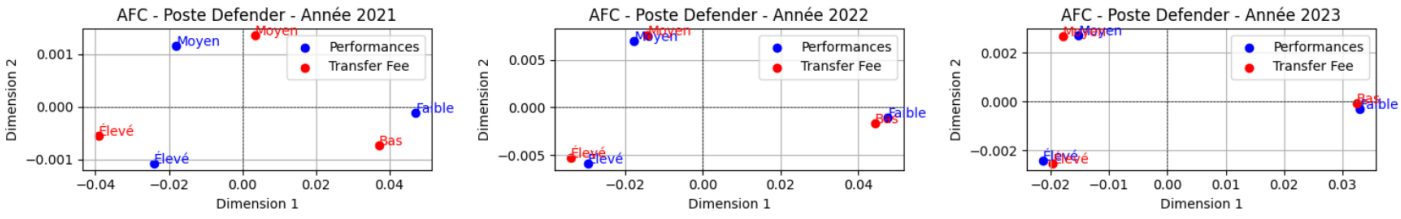

Figure 4.1: FCA - Position: Defender

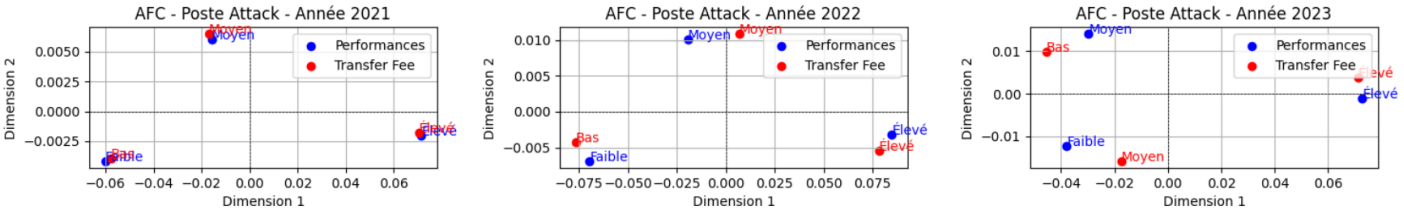

Figure 4.2: FCA - Position: Attack

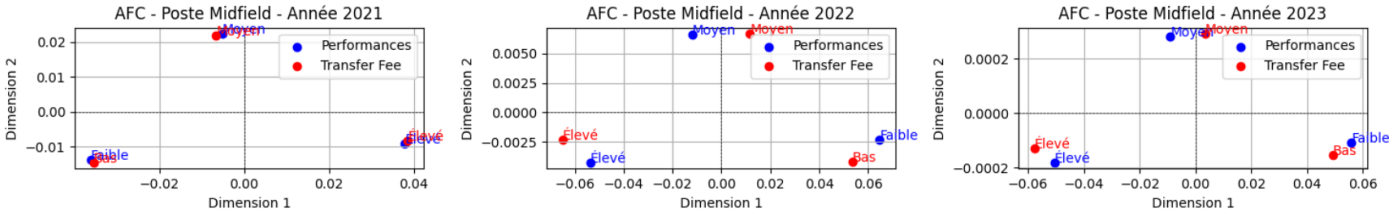

Figure 4.3: FCA - Position: Midfielder

The FCA results revealed the following key insights:

- A moderate association was observed between performance and transfer fee across all players.
- Players with consistently high performance are typically associated with significantly higher transfer fees, emphasizing the critical role of performance as a determining factor.
- The first dimension accounted for a larger proportion of the variance compared to the second, highlighting the greater influence of performance in the analysis.

Conclusion

Player performance metrics show a significant association with transfer fee. Metrics like **minutes played, goals, and assists** clearly influence valuation. This highlights the importance of prioritizing performance-based evaluations. The next chapter will explore the relationship between **club performance** and **transfer activity** to broaden the analysis.

Chapter 5

Relationship between performance and transfer fee using CCA

Objective

The objective of this study is to analyze the relationship between football players' performance and their transfer value or fee, to better understand the factors influencing the transfer market. To deepen this analysis, we employed Canonical Correlation Analysis (CCA), a statistical method designed to identify the maximum linear relationships between explanatory variables (performance indicators) and target variables (transfer value and fee).

This approach aims to refine the insights identified during chapter 4 and provide a more detailed understanding of the determinants of the transfer market.

Methodology

Data processing was carried out on the transfer.csv, players.csv, appareances.csv and clubs.csv data sets to collect the most interesting data for the study of club performance in relation to transfer fees. This resulted in the following variables for the study:

Explanatory Variables (X):

- `total_goals`: Total goals scored
- `total_assists`: Total goals assists
- `championship_level`: League difficulty of the club where the player came from
- `average_minutes_played`: average minutes played (shows that the coach is satisfied with the players' performances
- `age`: age of the player when the transfer was made `position_encoded`: Players' position

Dependent Variable (Y):

- `transfer_fee`: Transfer fee of players
- `market_value`: Market value of the player at the time of the transfer

To ensure consistency across the dataset, all variables were standardized, bringing them to a common scale.

The first step in the analysis was data cleaning. Missing or inconsistent values were addressed to ensure the quality and reliability of the dataset.

For the variables total_goals and total_assists, I considered only those goals and assists achieved within 4 years prior to the transfer. This timeframe was chosen to focus on the player's most recent performances, as older statistics may not accurately reflect their current form or market value at the time of the transfer.

Additionally, the position variable was transformed into numerical values for simplicity: 1 for forwards and 2 for midfielders. This allowed for better handling of the categorical variable in the analysis. Similarly, the championship_level variable was encoded as 2 for strong leagues, 1 for average leagues, and 0 for weak leagues, to reflect the relative difficulty of the league in which the player competed.

Finally, I limited the analysis to forwards and midfielders, as the primary performance indicators available in the dataset are goals and assists . This decision was made because, for other positions such as defenders or goalkeepers, assists and goals do not reflect their performances. Therefore, restricting the analysis to attacking and midfield positions ensures a more accurate representation of performance relative to the available metrics.

Results

Correlations Between Initial Variables

The correlation matrix shows significant relationships:

- Transfer_fee and Market_value_in_eur: The correlation is 0.87, indicating a very strong relationship between transfer fee and market value. This suggests that market value is a good predictor of a player's transfer price.
- The correlation between total_assists and market_value_in_eur is stronger (0.50) compared to the correlation between total_assists and transfer_fee (0.38). This suggests that total_assists has a more significant relationship with a player's market value than with the transfer fee. The weaker correlation between total_assists and transfer_fee (0.38) implies that other variables, beyond just performance, have a more substantial impact on the transfer fee (for example contractual terms, and negotiating power, as well as external market factors such as club interests). Similarly for total_goals
- The correlations between championship_level and performances variables are weak, despite that, the league difficulty may still have an indirect effect on the market value or transfer fee, which is often the case for players coming from prestigious leagues

Canonical Values and Contributions

Contributions of Explanatory Variables (X):
CV1:

- `position_encoded (-0.408)`: Position is a significant contributor to this first component, indicating that a player's position is strongly related to their performance and transfer.
- `championship_level (-0.649)`: The championship level is the primary negative contributor, showing that the difficulty of the league is a determining factor in this dimension.
- `age_at_transfer (0.115`: Age contributes positively, indicating that younger players are more valued in this context.

CV2:

- `position_encoded (-0.709)`: Once again, position strongly influences this component, but it has a greater contribution here than in CV1.
- `championship_level (0.290)`: The difficulty of the championship plays a secondary but positive role, suggesting contextual effects on player valuation.

Contributions of Transfer Variables (Y): CV1:

- `transfer_fee`: The transfer fee contributes weakly but positively to this component, indicating a slight correlation with player performance and context.
- `market_value`: The market value slightly influences this dimension but with a weak negative contribution.

CV2: The contributions are also weak and close to zero.
Interpretation:

- `Position and Championship level`: These two variables play a central role in the first canonical components, showing that they strongly influence the relationship between performance and transfer. The position is particularly important, which reflect differences in valuation between midfielders and forwards (may be due to popularity).
- `Age`: Player age contributes positively, reflecting the overvaluation of young players with strong future potential.

- `Performance (goals and assists)`:These variables do not have significant contributions in the first components, which may seem counterintuitive. This could be explained by their correlation already being integrated into contextual indicators such as position or championship difficulty.
- `Target Variables (transters)`: The contributions of the financial variables (transfer fee and market value) are weak in the first dimensions, indicating non-linear relationships or the influence of other unmodeled factors (such as external market conditions, team strategies, player potential, or even the negotiation process)

Visualization of Canonical Relationships

- **Heatmap of Canonical Correlations:** The heatmap indicates low or null values for cross-canonical components, suggesting a clear separation between the contributions of X and Y variables.
- **Scatter Plot of Canonical Variates:** Limited dispersion is observed between the first canonical variate of performance variables and transfer fees, indicating a moderate relationship.

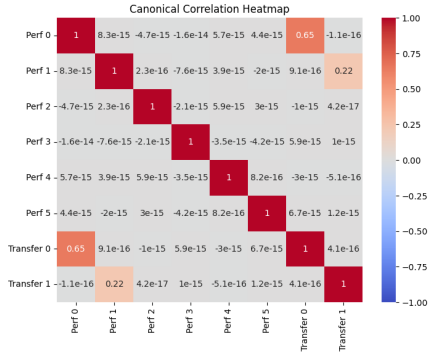

Figure 5.1: Correlation Circles for Canonical Components

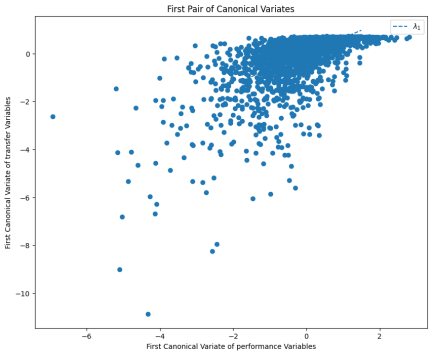

Figure 5.2: Two first Canonical Components Relation Plot

- **Correlation Circles:** The two correlation circles show how performance variables (left) and transfer-related variables (right) are related along canonical variates:
  - Left Circle (Performance): Position, goals, and assists are negatively correlated with the first variate, while age and minutes played have weak associations with both variates.
  - Right Circle (Transfer Data): Market value and transfer fee are negatively correlated with each other, with market value slightly influenced by performance and other factors like position and championship level.

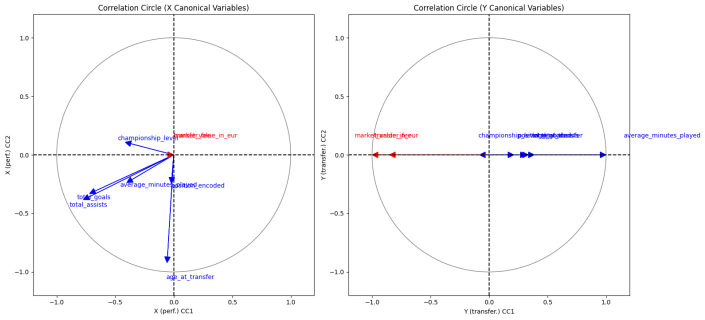

Figure 5.3: Correlation Circles for Canonical Components.

Conclusion

The Canonical Correlation Analysis (CCA) reveals several key insights into the relationship between football players' performance and their transfer value or fee:

- **Position** and **Championship Level** are the most influential factors in explaining the relationship between performance and transfer fees. Players' positions (especially forwards vs. midfielders) and the strength of the leagues they come from significantly affect how players are valued in the transfer market.
- **Age** also plays a notable role, with younger players generally being more highly valued, likely due to their future potential.
- **Performance indicators** like goals and assists do not significantly contribute to the first canonical components, suggesting their effects are already integrated into other variables such as position or league difficulty. This highlights the complex, non-linear nature of player valuation.
- Despite the correlation between total goals/assists and market value, the actual **transfer fee** shows a weaker correlation with performance metrics, suggesting that other factors, such as player age, market conditions, or negotiation power, also play a significant role.
- This is further confirmed by the weak contributions of transfer fee and market value to the first canonical variate.

In conclusion, while player performance (goals, assists) is important, **contextual factors** like position, league strength, and age appear to have a more substantial impact on both market value and transfer fees. This analysis underscores the complexity of the transfer market and the multifaceted nature of player valuation, which goes beyond individual performance metrics.

Chapter 6

Relationship between club performance and transfer fee

Objective

The goal of this study is to explore the relationships between club performance variables (such as total goals scored, number of wins, and median position) and transfer-related variables (such as transfer fees). Canonical Correlation Analysis (CCA) was applied to identify significant relationships between these two sets of variables.

Methodology

Data processing was carried out on the transfer.csv, games.csv, club_games.csv and clubs.csv datasets in order to collect the most interesting data for the study of club performance in relation to transfer fees. This resulted in the following variables for the study:

Explanatory Variables (X):

- `total_goals`: Total goals scored
- `wins`: Number of wins
- `median_position`: Median position in the league

Dependent Variable (Y):

- `transfer_fee`: Transfer fee of players

Canonical Correlation Analysis

Canonical correlation analysis was performed to generate canonical variates, which represent optimal linear combinations of the input and output variables. The contributions of the variables were calculated to interpret their respective impacts.

Results

Correlations Between Initial Variables

The correlation matrix shows significant relationships:

- A strong positive correlation between `total_goals` and `wins` (`0.91`).
- A negative correlation between `median_position` and the performance variables.
- A weak correlation between `transfer_fee` and performance variables (approximately `0.27`).

Canonical Values and Contributions

Contributions of Explanatory Variables (X):

- `total_goals`: Minor contribution to the first canonical component (`0.045`).
- `wins`: Significant contribution to the second canonical component (`0.342`).
- `median_position`: Major impact on the third canonical component (`0.235`).

Contributions of Transfer Variables (Y):

- `transfer_fee`: Primarily correlated with the first canonical variate (singular value 1: `0.2706`).

Visualization of Canonical Relationships

- **Heatmap of Canonical Correlations:** The heatmap indicates low or null values for cross-canonical components, suggesting a clear separation between the contributions of X and Y variables.
- **Scatter Plot of Canonical Variates:** Limited dispersion is observed between the first canonical variate of performance variables and transfer fees, indicating a moderate relationship.

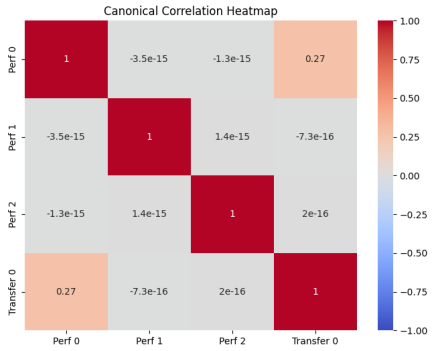

Figure 6.1: Correlation Circles for Canonical Components

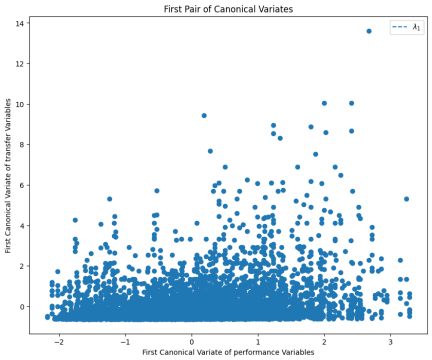

Figure 6.2: Two first Canonical Components Relation Plot

- **Correlation Circles:**
  - The correlation circles reveal a strong contribution of `wins` and `total_goals` to the first canonical variate.
  - The influence of `transfer_fee` on the studied variates appears minimal.

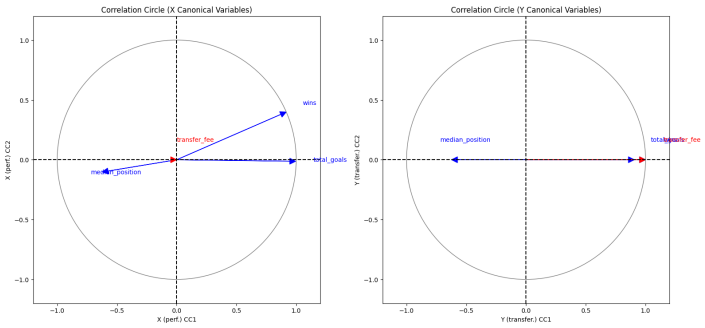

Figure 6.3: Correlation Circles for Canonical Components.

Conclusion

The results highlight specific relationships between clubs performance and transfer fees. However, the direct impact of variables such as `median_position` and `transfer_fee` appears limited.

This study identified links between clubs performance and transfer fees. Although moderate relationships were observed, the results suggest that unconsidered factors may have a greater influence on transfer fees.

Conclusion

In this project, we analyzed the relationships between various factors influencing player transfer fees, focusing on three main chapters: the relationship between transfer value and market value, between player performance and transfer fee, and the role of club performance in determining transfer fees.

- Transfer Value vs. Market Value: The analysis in this chapter showed that there is no strong relationship between transfer value and market value. This indicates that market value does not predict a player's transfer fee.
- Player Performance vs. Transfer Fee: In this chapter, we observed a moderate to strong ralation between player performance and transfer fee. Players with higher and more consistent performance tend to command higher transfer fees, emphasizing that performance is a crucial factor in the transfer market.
- Club Performance vs. Transfer Fee: The chapter highlighted the association between the overall performance of clubs and their players' transfer fees. Successful clubs tend to attract higher transfer fees for their players, reinforcing the idea that the success of a club can positively influence the value of its players in the transfer market.

Overall, the findings suggest that player performance and club performance are key drivers of transfer fees, while market value is not sufficient to explain the variability in transfer fees. These insights provide valuable information for clubs, agents, and analysts in understanding the complex dynamics of the player transfer market.