Interactive Dashboard for Exploratory Analysis of Football Player Data from Transfermarkt

Nikki Kayastha
Department of Computer Science and Engineering
Kathmandu University
Dhulikhel, Kavre, Nepal

Siddhant Khadka
Department of Computer Science and Engineering
Kathmandu University
Dhulikhel, Kavre, Nepal

***Abstract*—This project presents an exploratory and interactive visualization of football player data, focusing on player market value, age, club association, and clustering insights. Utilizing the Streamlit framework along with Plotly for graphical representation, this project delivers a dashboard for filtering, comparison, and visualization of players across clubs and nations. Clustering methods such as KMeans and DBSCAN further assist in grouping players based on market potential. This system is designed for use by analysts, scouts, and enthusiasts aiming to derive insights from comprehensive player datasets.**

***Index Terms*—Football Analytics, Clustering, Market Value, Dashboard, Streamlit, Plotly KMeans, DBSCAN, Club Comparison.**

I. Introduction

Football scouting and club management increasingly rely on data analytics to inform decisions. This project is centered around building an interactive web-based dashboard using Python libraries including Streamlit and Plotly. It provides metrics on player age, market value, club distributions, and nationality filters, with clustering applied to player attributes to visualize potential groupings. The visual nature of the dashboard makes data-driven decisions easier and more accessible.

II. Dataset and Preprocessing

A.Dataset

The dataset used in this study, players.csv, is one of several CSV files scraped from Transfermarkt—a renowned football website that offers comprehensive statistics, market values, and transfer histories of professional players and clubs worldwide. It includes comprehensive attributes for 32,601 professional football players from various clubs and countries, covering demographic details, positional roles, physical attributes, contract information, and financial estimates such as market value. The dataset provides a valuable basis for analyzing player performance, market trends, and club strategies.

Additionally, the dataset is automatically updated once a week to ensure the information remains current and reflects recent transfers and market changes. Each row in the dataset represents a unique player and includes the following attributes:

- Player Id
- First Name
- Last Name
- Name
- Last Season
- Current Club Id
- Current Club Name
- Player Code
- Country Of Birth
- City Of Birth
- Country Of Citizenship
- Date Of Birth
- Position
- Sub Position
- Preferred Foot
- Height (cm)
- Contract Expiration Date
- Agent Name
- Image Url
- Url (Transfermarkt profile link)
- Current Club Domestic Competition Id
- Market Value (€)
- Highest Market Value (€)

1.Dataset Statistics

- Total players: 32,601
- Temporal coverage: Up to June 2025
- Geographic coverage: Global, Players from clubs across different countries
- Columns: 23
- Numerical columns: 17
- Categorical columns: 5

The dataset contained missing values, particularly in optional fields like agent_name and country_of_birth, which were handled during preprocessing. Nevertheless, the majority of core features were complete and ready for analysis.

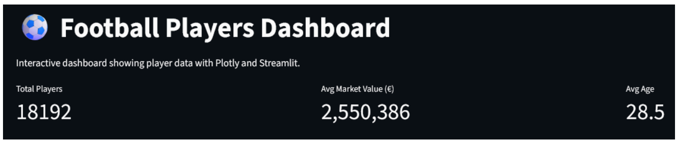

Fig1: Statistics of original dataset

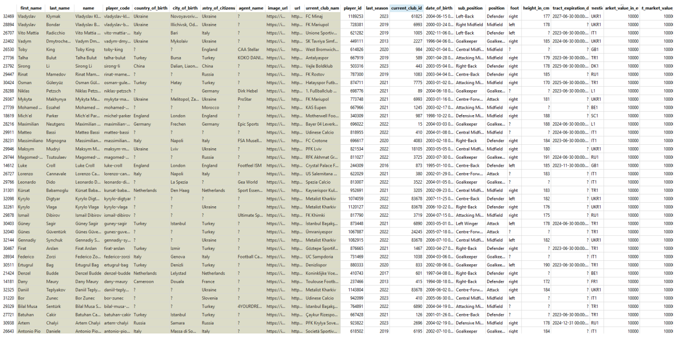

Fig2: Dataset Preview

Initial statistical summaries show that the average player height is around 183 cm, with some variation across positions (e.g., goalkeepers tend to be taller). Market value ranges from €25,000 to over €100 million, with players in top leagues and prime age (20–30) having significantly higher valuations.

B. Preprocessing Steps

Data preprocessing is a crucial step to ensure the quality, accuracy, and usability of the dataset before analysis. For this project, The following steps were performed in Orange Data Mining, as visualized in the workflow diagram:

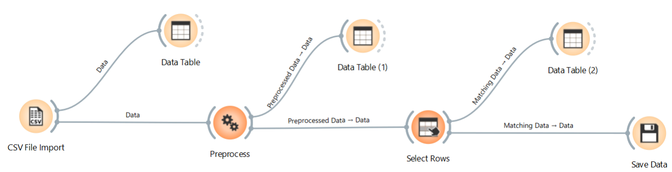

Fig3: Preprocessing workflow via Orange

1. Initial Preprocessing

The dataset was passed through the Preprocess widget to clean and prepare the data:

- Missing Values: Rows containing missing (NaN) values in key columns were dropped to maintain the integrity of the analysis.
- This helped reduce noise in the dataset and eliminated incomplete records that could bias or skew results.

2. Manual Data Filtering

After preprocessing, the data was further refined using the Select Rows widget.

- Unrealistic Heights: Five player entries had height values below 100 cm, which is not realistic for professional athletes. These rows were filtered out using a conditional rule.
- Invalid Contract Expiry Date: One player's contract expiration date was before the year 2000, which was deemed invalid for the context of this dataset. This row was also removed.
- This step ensured the dataset reflects valid and realistic player profiles.

3. Visual Confirmation and Saving the Cleaned Data

The output of both the Preprocess and Select Rows widgets was passed into Data Table widgets to visually inspect and confirm that the changes were correctly applied. The number of rows was reduced to 18,192 from the original 32,601, with most of the reduction resulting from missing values—such as the absence of agent information. Although removing such entries was not strictly necessary—for example, many young players do not have agents—it was done to ensure a cleaner and more optimal dataset for analysis. Finally, the cleaned and filtered dataset was saved using the Save Data widget for future analysis and modeling tasks.

III. Exploratory Data Analysis and Visualization

We developed an interactive dashboard using Streamlit to analyze professional football player data through dynamic, user-driven visualizations. This platform allows users to intuitively explore key player metrics such as age, market value, and positional attributes by applying custom filters based on nationality, position, and club affiliation. Through a suite of interactive charts, comparisons, and statistical summaries, the dashboard reveals underlying trends, highlights top-performing players, and supports insight-driven evaluation. It serves as a practical tool for analysts, scouts, and decision-makers seeking to better understand the structure and value distribution of modern football squads.

A. Histogram

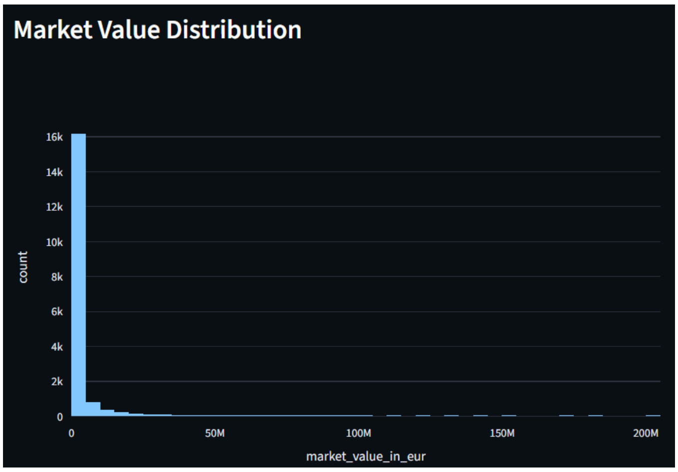

Fig4: Right-Skewed Distribution of Football Players' Market Values Highlighting Valuation Inequality

The histogram illustrates the overall distribution of football players' market values in euros. The x-axis represents the market value, ranging from very low amounts to over 200 million euros, while the y-axis shows the count of players falling within each value range. The distribution is highly right-skewed, indicating that the vast majority of players have relatively low market values, with a sharp peak in the lower value ranges.

As we move to the right, the bars decrease significantly in height, showing that only a small number of players are valued at tens or hundreds of millions of euros. This type of distribution is typical in football economics, where a handful of top-tier players command enormous fees, while the bulk of professionals are valued modestly. The histogram thus highlights the inequality in player valuation, emphasizing how rare elite valuations are in the broader population of football players.

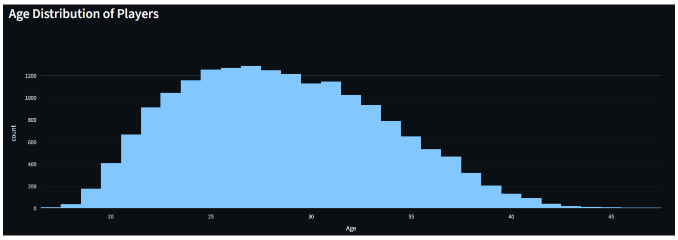

Fig5: Histogram showing age distribution of players

This histogram shows the frequency of football players across different age groups. The distribution forms a bell-shaped curve, indicating that most players are concentrated between the ages of 23 and 30, with the peak around 26–27 years old, suggesting this is the most common age range for professional players.

There are fewer younger players under 20 and older players above 35, with the number gradually tapering off at both extremes. This pattern reflects typical career trajectories in football, where players peak in performance during their mid-20s and decline in representation as they approach their late 30s and beyond.

B. Box Plot

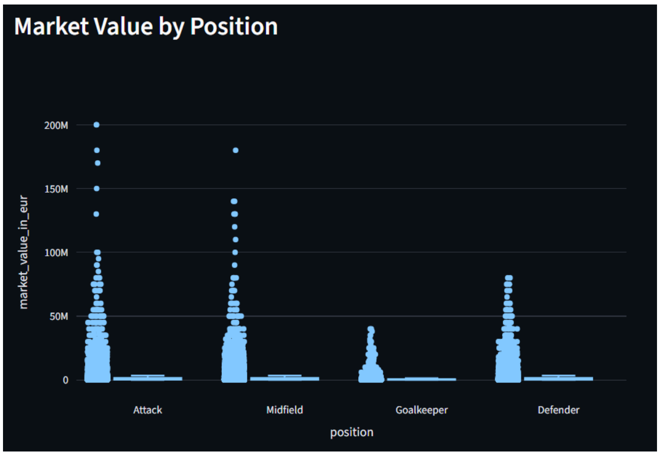

Fig6: Box Plot of Market Value by Position

The right-hand plot is a box plot that summarizes how football player market values vary across different positions: Attack, Midfield, Goalkeeper, and Defender. The line inside each box is the median market value. The "whiskers" extend to cover most of the remaining data, while any individual dots above the whiskers represent outliers, which are players with exceptionally high values.

From this plot, it's evident that Attackers tend to have the highest market values, including some extreme outliers exceeding 200M euros. Midfielders also have high values but slightly less so. Defenders and Goalkeepers generally have lower medians, indicating a more concentrated and lower valuation range. This box plot clearly demonstrates the economic prioritization of attacking and creative roles in modern football, where goal scorers and playmakers drive the highest fees.

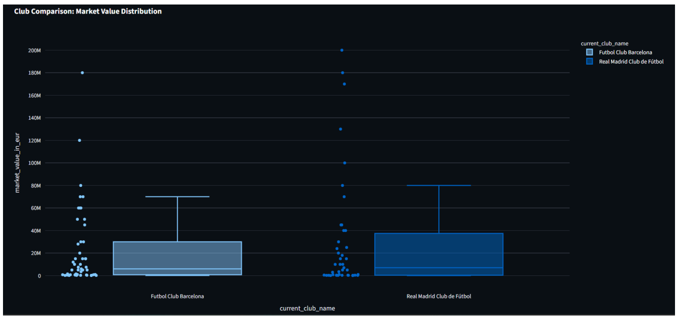

Fig7: Box Plot showing comparison of 2 clubs based on market value

This interactive box plot visualizes the market value distribution of players across selected football clubs, allowing users to compare multiple teams simultaneously. In this instance, Fútbol Club Barcelona and Real Madrid Club de Fútbol are shown. Each dot represents a player's market value in euros, while the box plot for each club shows the distribution—median, interquartile range, and potential outliers. The chart reveals that both clubs have players across a wide range of market values, but Real Madrid has a slightly higher median and broader spread.

Since the interface supports selecting more than two clubs, this tool is useful for benchmarking market value trends and identifying value outliers among elite clubs.

C.Scatter Plot

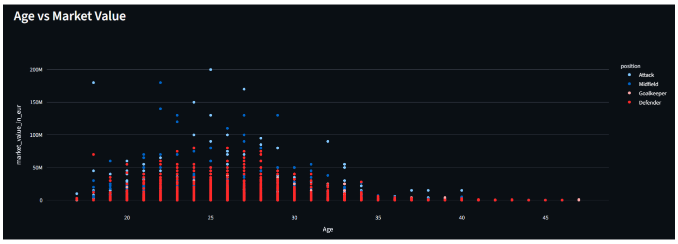

Fig8: Scatter Plot showing Age vs Market value

The scatter plot shows a clear trend where football players typically reach their highest market values between ages 22 and 28, which is considered their prime playing period. After this age range, player values begin to decline steadily, with a noticeable drop beyond age 30, reflecting reduced demand and resale potential for older players.

The plot also reveals that attacking and midfield players generally hold higher market values compared to defenders and goalkeepers, who tend to cluster at lower values regardless of age. Interestingly, some exceptionally talented young players under 20 already command very high valuations, signaling strong market confidence in their potential. Overall, the plot illustrates how both age and playing position significantly influence a player's market value, with offensive roles being more financially rewarded and defensive roles valued lower but often sustaining longer careers.

D. Pie Chart

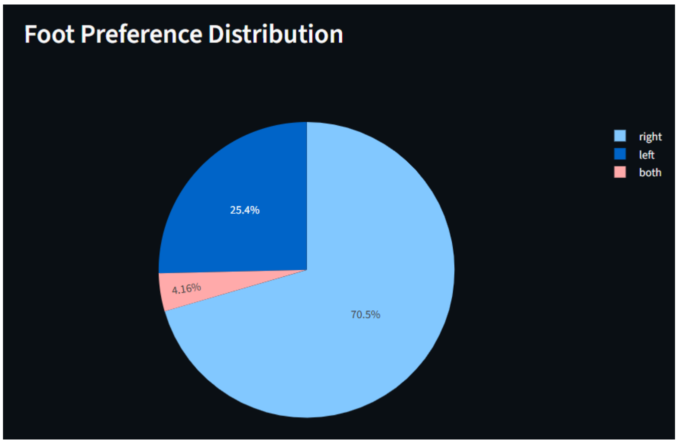

Fig9: Pie Chart for foot preference

The pie chart titled "Foot Preference Distribution" displays how football players are distributed based on their dominant foot. It shows that a significant majority of players, 70.5%, are right-footed, while 25.4% are left-footed, and only a small minority, 4.16%, are proficient with both feet.

This uneven distribution highlights the dominance of right-footed players in professional football, which may be due to natural handedness trends and training practices. The relatively low percentage of two-footed players suggests that being ambidextrous is rare, even at elite levels, though it can offer tactical versatility and value on the pitch.

E.Bar Chart

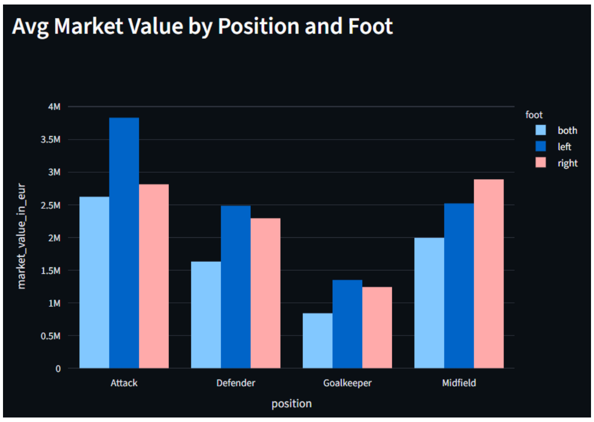

Fig10: Bar Chart for Position and Foot

The bar chart on the right compares the average market value of players based on both their position and foot preference. Notably, left-footed players consistently show higher average market values in the Attacker and Defender categories, with left-footed attackers reaching nearly 4 million euros, the highest among all groups. Right-footed midfielders also have high values, outperforming both left- and two-footed peers in that position. Interestingly, players who use both feet, despite their versatility, tend to have lower average values across all positions, especially among goalkeepers.

This suggests that while two-footed ability is tactically beneficial, it may not translate directly into higher market valuation, whereas being a left-footed player—a scarce commodity—may carry a premium, especially in attack and defense.

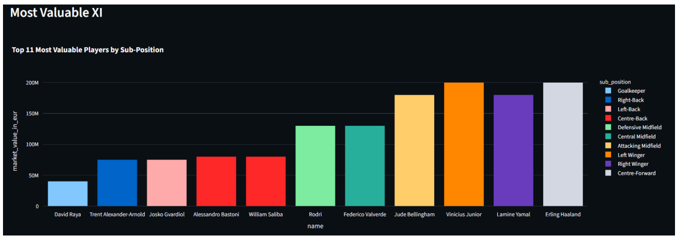

Fig11: Bar Chart showing most valuable player in each of the 11 sub positions

The bar chart titled "Most Valuable XI" highlights the highest market value player in each football sub-position, effectively forming the most valuable starting lineup across all areas of the pitch. Each bar represents a different sub-position, with the player holding the highest market value in that role.

Erling Haaland (Centre-Forward) and Vinicius Junior (Left Winger) top the chart with the highest values at €200 million, followed by Lamine Yamal (Right Winger) and Jude Bellingham (Attacking Midfield) at €180 million. Midfielders like Rodri and Federico Valverde, as well as defenders such as Trent Alexander-Arnold, Josko Gvardiol, Alessandro Bastoni, and William Saliba, are also included as the most valuable in their respective positions. David Raya represents the Goalkeeper role with the lowest market value among the XI, emphasizing the disparity in market valuation across different areas of the field.

F. Table

Top 10 Most Valuable Clubs

	current_club_name	market_value_in_eur
258	Manchester City Football Club	1,370,900,000.0000
323	Real Madrid Club de Fútbol	1,307,375,000.0000
186	Futbol Club Barcelona	1,101,050,000.0000
42	Arsenal Football Club	1,076,500,000.0000
253	Liverpool Football Club	1,052,250,000.0000
89	Chelsea Football Club	1,009,725,000.0000
302	Paris Saint-Germain Football Club	1,008,650,000.0000
129	FC Bayern München	900,475,000.0000
389	Tottenham Hotspur Football Club	870,950,000.0000
259	Manchester United Football Club	755,725,000.0000

Fig12: Table for Top 10 most valuable clubs

The table shows the top 10 most valuable football clubs by player market value, led by Manchester City (€1.37 billion) and Real Madrid (€1.31 billion). FC Barcelona, Arsenal, and Liverpool also exceed €1 billion. This reflects trends where strong spending and development drive high market values and competitiveness.

Top Players by Market Value

	name	position	Age	market_value_in_eur	current_club_name
11182	Erling Haaland	Attack	25	200000000	Manchester City Football Club
10015	Vinicius Junior	Attack	25	200000000	Real Madrid Club de Fútbol
17788	Lamine Yamal	Attack	18	180000000	Futbol Club Barcelona
14194	Jude Bellingham	Midfield	22	180000000	Real Madrid Club de Fútbol
9150	Kylian Mbappé	Attack	27	170000000	Real Madrid Club de Fútbol
11497	Bukayo Saka	Attack	24	150000000	Arsenal Football Club
14455	Florian Wirtz	Midfield	22	140000000	Bayer 04 Leverkusen Fußball
14161	Jamal Musiala	Midfield	22	140000000	FC Bayern München
10958	Phil Foden	Attack	25	130000000	Manchester City Football Club
9669	Rodri	Midfield	29	130000000	Manchester City Football Club

Fig13: Top 10 Football Players by Market Value: Age and Club Insights

This table shows the top 10 most valuable football players, led by Erling Haaland and Vinicius Junior at €200 million each. Most players are in their early to mid-20s, indicating a trend toward investing in young, high-potential athletes. Real Madrid and Manchester City dominate with several players, reflecting their financial power and focus on top talent across Europe's top leagues.

G. Cluster (Scatter Plot)

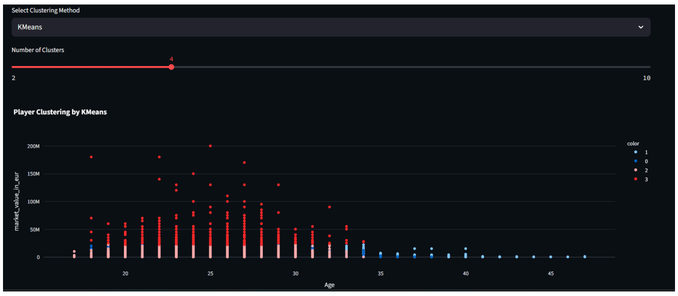

Fig14: Interactive Clustering of Football Players by Age and Market Value

This scatter plot reveals how players are distributed across these clusters- cluster 2 (red) dominates younger, high-value players (aged ~18–30), whereas clusters 0 and 1 (blue shades) contain mostly older players with lower market values. The slider at the top allows adjustment of the number of clusters, and the chart dynamically updates to reflect the grouping patterns accordingly. This helps identify distinct player segments, such as young stars, veterans, and mid-career athletes.

Fig15: DBSCAN Clustering of Players by Market Value

This image presents a DBSCAN clustering analysis of football players based on their market value, where clustering parameters such as epsilon and minimum samples are adjustable. The scatter plot at the bottom shows player distribution by market value, with each dot representing a player. Two colors are used: one for cluster 0 (light blue), indicating players grouped based on density, and another for noise points (-1) (darker blue), representing outliers that do not fit into any cluster.

The majority of players fall into a single cluster, reflecting similar market value ranges, while a noticeable number of high-value players are classified as outliers due to their significant deviation from the typical value range. This visualization effectively highlights both the core cluster of average-valued players and the distinctiveness of top-tier, highly valued individuals.

IV. Data Mining Tasks

In this project, we primarily focus on Exploratory Data Analysis (EDA) and interactive visualization to understand the climate dataset:

- Automatic Data Initialization:
  Instead of relying on user-uploaded data, the system automatically loads a built-in dataset containing attributes like player age, current club, and market valuation. A snapshot of the initial few records is shown for user orientation.
- Graphical View of Player Attributes:
  Dynamic visual elements like charts, categorical bar plots, and circular graphs are employed to present distributions of features such as playing position, nationality, team representation, etc.
- Trends Across Clubs and Career Stages
  Age-wise and club-wise visual comparisons using bar charts highlight patterns in player development and club investment. This reveals how market value and player demographics shift across clubs and age brackets
- Value Spread and Feature Linkages:
  This project uses visual tools like histograms and boxplots to explore how player attributes (e.g., age, market value) are distributed.
- Data Summary Export Tools:
  To facilitate further exploration, users can download analytical outputs such as statistical summaries, filtered data visualizations for offline use or documentation

V. Future Works

- Incorporate match stats like goals, assists, minutes played.
- Build valuation prediction models using regression.
- Extend the dashboard to support league-wide and country-wise breakdowns.

VI. Conclusion

The football dashboard effectively integrates data preprocessing, visualization, and clustering into an easy-to-use Streamlit interface. It supports stakeholders in evaluating players, benchmarking clubs, and discovering patterns across nationalities and positions. With added real-world stats, it can evolve into a full-fledged scouting tool.

Acknowledgment

We are deeply grateful to Dr. Rajani Chulyadyo of the Department of Computer Science and Engineering at Kathmandu University. Her continuous encouragement and insightful advice were invaluable throughout this project. Her expertise in data mining and visualization greatly influenced our approach and understanding. We sincerely appreciate the chance to work and learn under her guidance.

References

[1] Davidcariboo. (n.d.). Player Scores Dataset [Dataset]. Kaggle. https://www.kaggle.com/datasets/davidcariboo/player-scores/data?select=players.csv

[2] Streamlit Inc. (n.d.). Streamlit [Computer software]. https://streamlit.io/

[3] Plotly Technologies Inc. (n.d.). Plotly Python Open Source Graphing Library [Computer software]. https://plotly.com/python/

[4] Scikit-learn Developers. (n.d.). Scikit-learn: Machine Learning in Python [Computer software]. https://scikit-learn.org/

[5] Bioinformatics Laboratory, University of Ljubljana. (n.d.). Orange Data Mining Toolbox [Computer software]. https://orangedatamining.com/