Stat 184 Final Project

Introduction

Football is the most popular sport in the world, where major tournaments like the FIFA World Cup and major leagues such as the English Premier League, Ligue 1, Serie A, La Liga and the Bundesliga attract billions of followers and viewers. Better known as soccer in the United States, football is a simple game where two teams compete to score goals by moving the ball into the opposing net, primarily using their feet. Contrasting the simplicity of the game is the dynamics behind the scenes, where clubs invest heavily in players after analyzing metrics like market value, match statistics, and positional roles. This leads us to our project, which examines how player attributes such as age, minutes played, position, and goal contributions as well as the team attribute of winning percentages relate to market valuation and ultimate team success across the top 5 leagues, particularly for the 2024-2025 season.

Research Questions

Our goal is to get a better understanding of how player and eventually, team performance metrics relate to both market valuation and team outcomes. In order to accomplish that, we ask: How does a player's age and position influence their minutes played and market value? Do players with higher goal contributions consistently have greater market worth? How do team averages such as player age, minutes played, and goal contributions correlate with win percentages across competitions?

Through visual analyses, including scatterplots, boxplot, heatmap, and summary tables, we aim to understand how measurable performance indicators influence both individual worth and collective outcomes in the sport.

Provenance of Our Data

For this project, we used four datasets to gather core performance indicators in football, all obtained from Transfermarkt's Football Data posted on Kaggle. Transfermarkt is a widely used football database that tracks player market values, match statistics, transfers, and club performance across global competitions. The dataset was created and maintained by David Cariboo, who used a web-scraping tool called transfermarkt-scraper along with Python scripts and SQL queries to extract, clean, and publish the data. The full data pipeline and source code are available on his GitHub. The purpose of his work is to keep these datasets up to date and publicly available on well-known data catalogs. The data is publicly available under the CC0 (Public Domain) license, with an expected update frequency of once per week. Our primary data source was Player Valuation while our secondary sources were Player Appearances, Player Statistics & Demographics and Game Statistics.

Primary Dataset

- **Player Valuation**: This was our main dataset that provided market value data for players over time. Key attributes were used consistently throughout our visualizations, which were player_id, date, and market_value_in_eur. The case is a single player.

Secondary Datasets

- **Player Appearances**: This dataset includes detailed match-level data on individual players. We used attributes like game_id, player_id, player_club_id, goals, assists, and minutes_played to analyze player contributions across matches. The case is a single player.
- **Player Statistics & Demographics**: This dataset includes personal and positional data for players. Our analysis focused on position, while also referencing identifiers like player_id and broader details such as date_of_birth for age related analysis. The case is a single player.
- **Game Statistics**: This dataset includes match-level team performance indicators. We used game_id, home_club_id, away_club_id, home_club_goals, and away_club_goals to calculate our new attribute, win percentages. Other fields, including competition_id, season, and round, offer useful structural context for game outcomes. The case is a single game.

Implementation of FAIR & CARE Principles

FAIR Principles

- **Findable**:
  The dataset is hosted on Kaggle, a widely-used platform that assigns a permanent identifier and includes metadata such as title, description, and file previews. These features ensure the dataset is easily located through web search and academic repositories.
- **Accessible**:
  Access is open with a Kaggle account, and the data can be downloaded in CSV format without technical barriers. This accessibility supports broad use across educational and research communities.
- **Interoperable**:
  Provided in a standard tabular structure (CSV), the dataset can be integrated into common data science workflows using tools like R, Python, and SQL. Column headers and values follow widely understood naming conventions, which supports semantic clarity.
- **Reusable**:
  The dataset includes basic metadata, source attribution, and a clear license (CC0: Public Domain Dedication), allowing for reuse without restriction. Users can confidently integrate and repurpose the data for academic, commercial, or exploratory purposes.

CARE Principles

- **Collective Benefit**:
  The dataset supports public interest research in sports analytics, education, and predictive modeling. By promoting open access to structured player data, it fosters innovation and insight in football analysis.
- **Authority to Control**:
  As the dataset aggregates public sports data without identifiable personal or community information, issues of consent and governance are minimal. However, transparency about the original data sources could further strengthen accountability.
- **Responsibility**:
  While the dataset itself poses low ethical risk, users must remain mindful of how derived insights are applied—particularly in contexts like player valuation or contract negotiation, where misuse could influence careers or reinforce bias.
- **Ethics**:
  The dataset avoids sensitive or protected attributes and reflects common practices in sports statistics. Ethical use demands acknowledgment of context and purpose, ensuring analyses do not misrepresent players or their performance.

Exploratory Data Analysis (EDA)

Age, Playing Time & Position

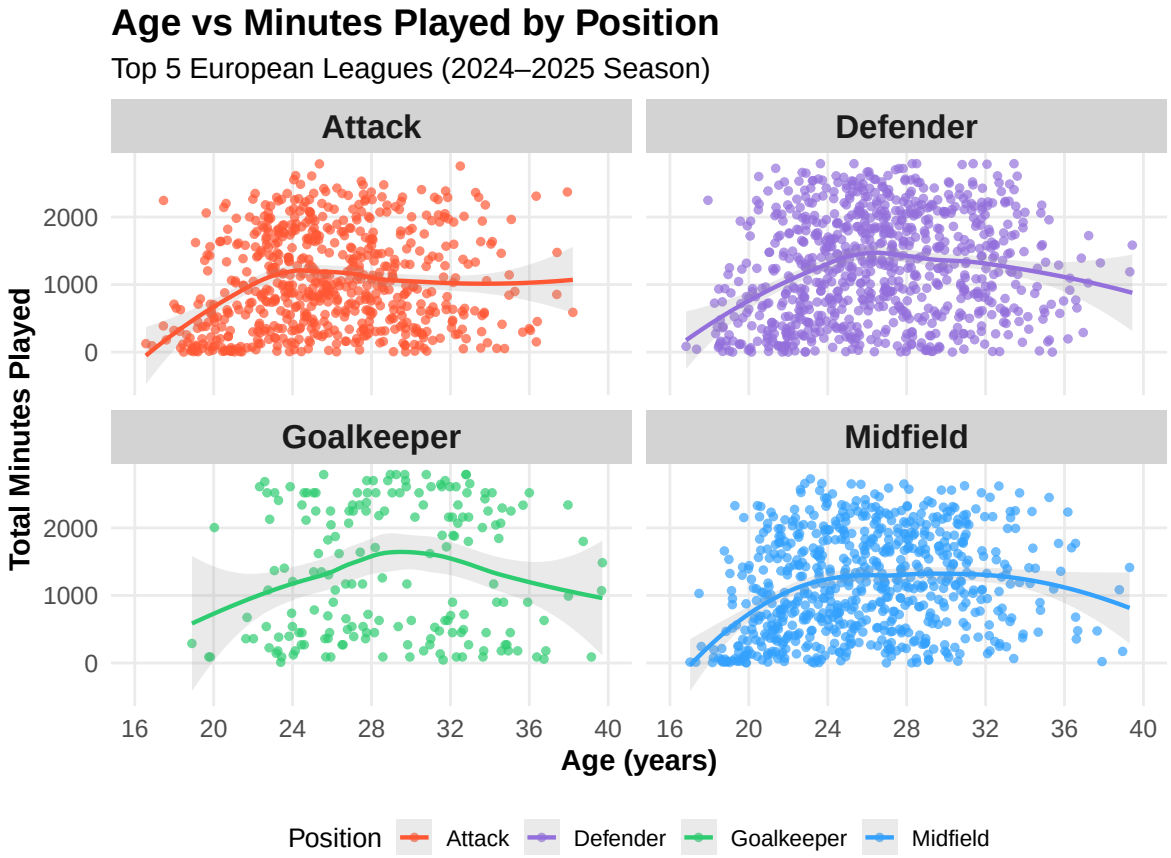

Figure 1: This scatter plot displays the relationship between player age and total minutes played during the 2024–2025 season across the top 5 European leagues. Faceted by position, it highlights positional differences in playing time patterns by age group, with trend lines indicating general age-performance curves.

Insight

This visual allows us to analyze the peak age based on minutes played for each position. From the scatter plot, we can see that attackers tend to peak earliest, around ages 24 to 26, while goalkeepers peak later, around ages 29 to 30. Midfielders and defenders are more evenly spread, peaking between 25 and 28.

It's important to keep this in mind when comparing market value across different age groups and positions.

Age, Position & Market Value

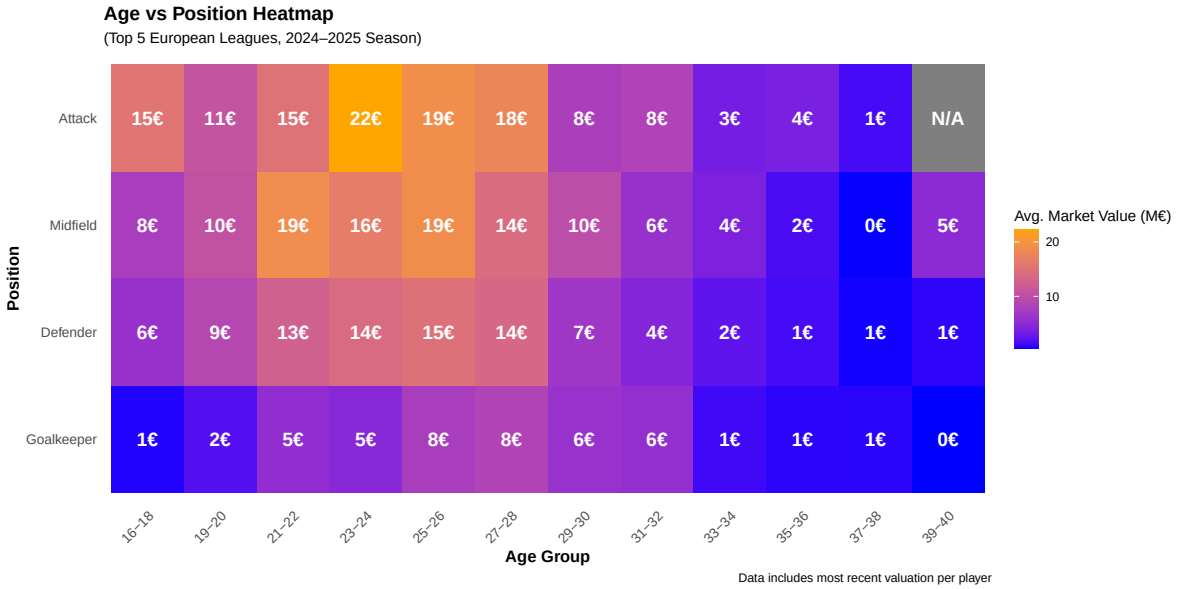

Figure 2: This heatmap visualizes average market value (in millions of euros) across different age groups and positions. Warmer colors represent higher values, with labels showing rounded figures. This allows easy comparison of how age and role impact player valuation in elite European football.

Insight

We see that players tend to be more valuable in positions higher up the pitch. Additionally, there is a clear peak in market value just before the typical peak in minutes played. This suggests that clubs invest more in players approaching their prime, anticipating strong performance and longevity. The alignment between expected trends and the data reinforces the reliability of this insight.