Analyzing Player Metrics and Market Value in European Football

Abigail Chen, Maxwell Gerhart, and Sanjana Menon

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

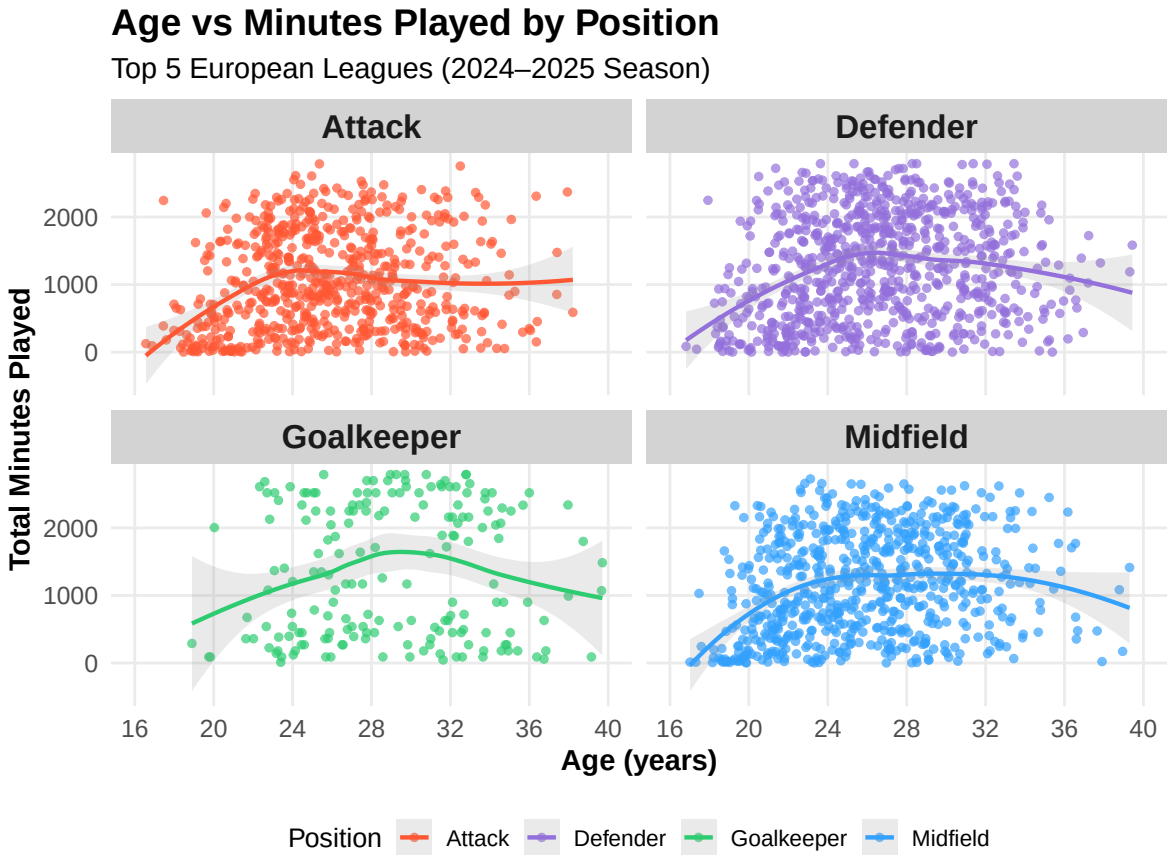


Figure 1: This scatter plot displays the relationship between player age and total minutes played during the 2024–2025 season across the top 5 European leagues. Faceted by position, it highlights positional differences in playing time patterns by age group, with trend lines indicating general age-performance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

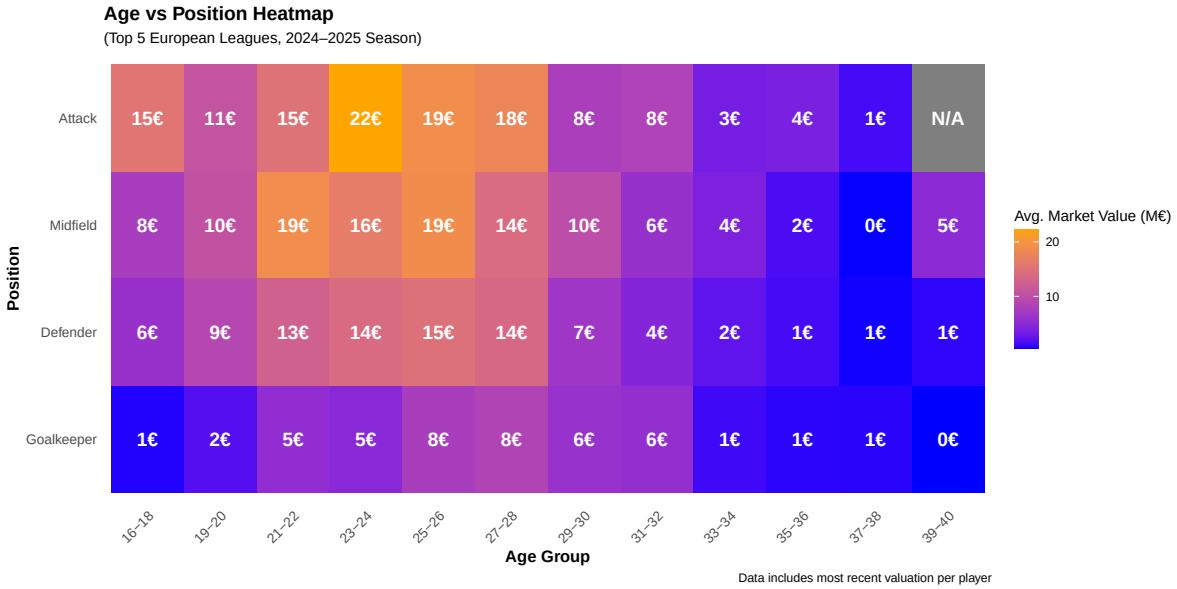


Figure 2: This heatmap visualizes average market value (in millions of euros) across different age groups and positions. Warmer colors represent higher values, with labels showing rounded figures. This allows easy comparison of how age and role impact player valuation in elite European football.

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.

Playing Time, Market Value & Position

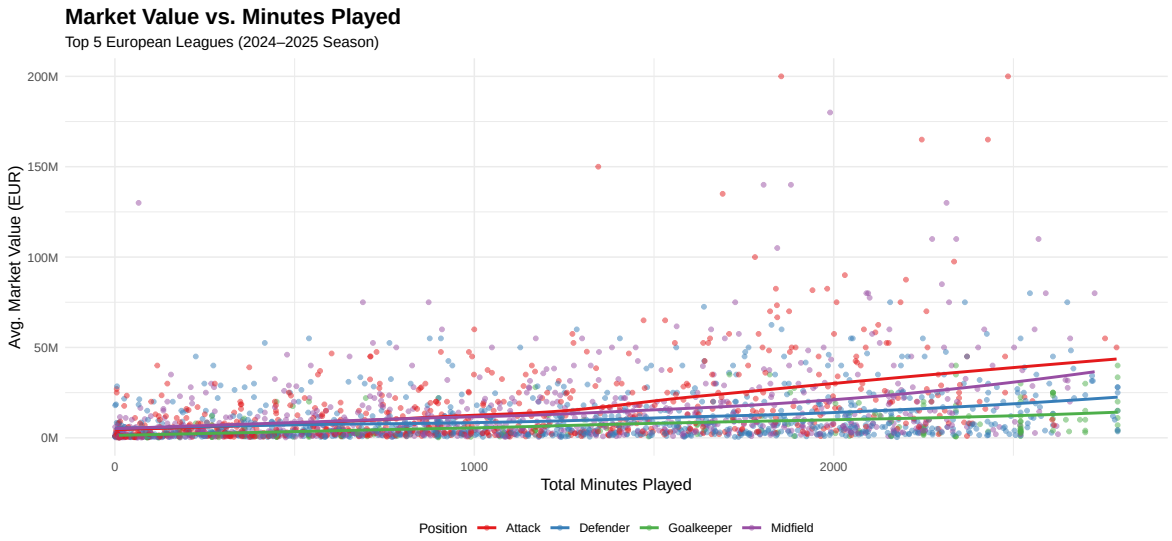


Figure 3: This scatterplot displays the relationship between total minutes played and average market value (in millions of Euros) for players in the top 5 European leagues during the 2024-2025 season. Each point represents a player, colored by their position: Attack (red), Defender (blue), Goalkeeper (green), and Midfield (purple). Trend lines for each position group illustrate the general tendency in market value relative to playing time.

Insight

Generally, players who accumulate more playing time tend to have a higher market value, as indicated by the upward sloping trend lines across all positions. However, the rate of increase in market value with minutes played appears to be steeper for attacking and midfield players compared to defenders and goalkeepers. This suggests that for outfield positions, consistent playing time is more strongly correlated with higher valuation. There are also notable outliers, particularly among attacking and midfield players, who command very high market values even with varying amounts of playing time, potentially reflecting exceptional talent or potential.

Goals, Market Value & Position

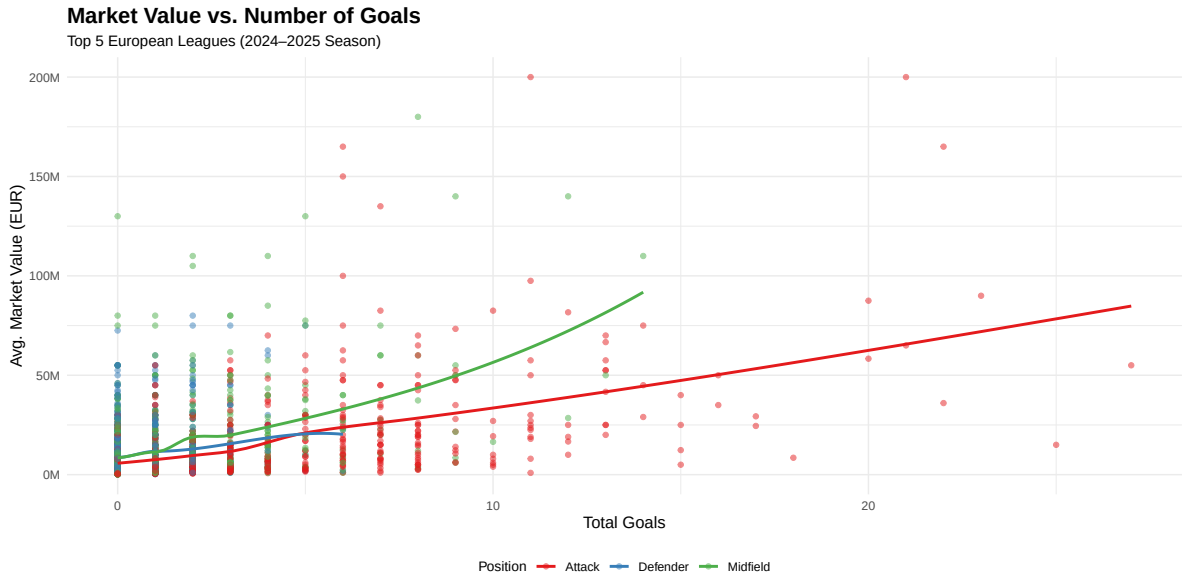


Figure 4: This scatterplot illustrates the correlation between total goals scored and average market value (in millions of Euros) for players in the top 5 European leagues during the 2024-2025 season. Each data point represents a player, distinguished by their position: Attack (red), Defender (blue), and Midfield (green). Trend lines for each position category highlight the general pattern in market value relative to the number of goals scored.

Insight

This graph suggests a positive correlation between the number of goals scored and a player's market value, particularly seen in the upward trend lines for attacking and midfield positions. Goalkeepers were omitted due to their lack of goals. For attacking players, the market value tends to increase more significantly with each goal, showing that a higher premium is placed on better goal-scoring ability. Midfielders also show a positive trend, though less steep than attackers. On the other hand, defenders show a relatively flat trend line with fewer goals, indicating that their market value is less directly influenced by the number of goals they score. This suggests that factors other than goal-scoring, such as defensive skills, may contribute more to their valuation.

Goal Distribution by Position

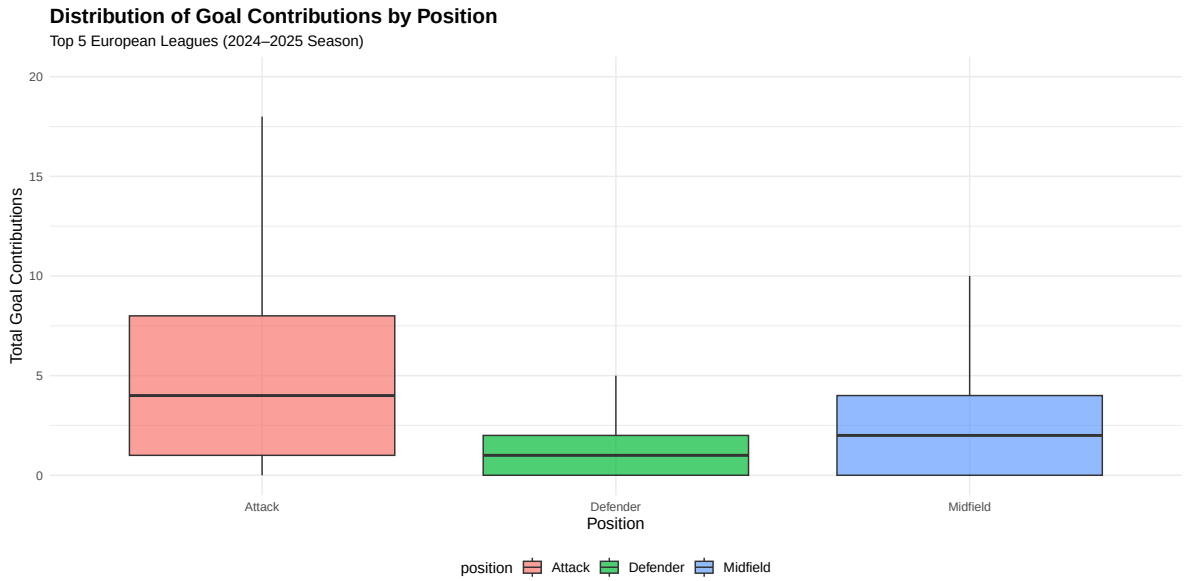


Figure 5: This boxplot displays the distribution of total goal contributions across different player positions in the top 5 European leagues during the 2024-2025 season. The boxes represent the interquartile range (IQR), the horizontal line inside each box indicates the median, and the whiskers extend to show the variability outside the lower and upper quartiles. Outliers, if any, would be plotted as individual points beyond the whiskers.

Insight

The distribution of goal contributions varies significantly by position. Attacking players exhibit the highest median number of goal contributions and the widest spread. This shows that while many contribute a moderate number of goals and assists, there is a strong tendency to achieve considerably higher totals. Midfielders show a lower median and a narrower IQR compared to attackers, meaning they have a more concentrated range of goal contributions. Defenders generally have the lowest median number of goal contributions, with a very tight IQR close to zero, which is expected given their primary role is not attacking.

The presence of some outliers for defenders and midfielders also display instances where players in these positions have made surprisingly high attacking contributions.

Top 10 Players by Market Value, Position, Goals, Appearances & Minutes Per Appearances

Table 1: Top 10 Players by Average Market Value (2024–2025 Season)

Player Name	Position	Goals	Appearances	Minutes Played	Avg. Minutes per Appearance	Avg. Market Value (EUR, Millions)
Vinicius Junior	Attack	11	24	1854	77.25000	200
Erling Haaland	Attack	21	28	2485	88.75000	200
Jude Bellingham	Midfield	8	24	1990	82.91667	180
Kylian Mbappé	Attack	22	28	2429	86.75000	165
Lamine Yamal	Attack	6	27	2245	83.14815	165
Bukayo Saka	Attack	6	18	1345	74.72222	150
Jamal Musiala	Midfield	12	25	1805	72.20000	140
Florian Wirtz	Midfield	9	25	1881	75.24000	140
Phil Foden	Attack	7	25	1691	67.64000	135
Rodri	Midfield	0	2	66	33.00000	130

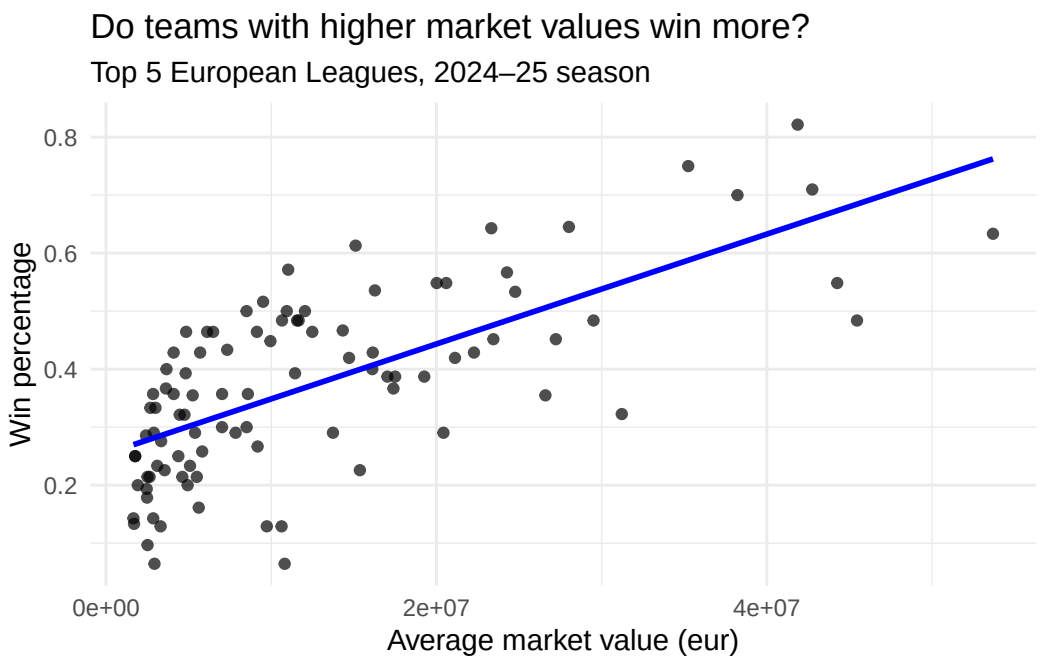
Figure 6: This table presents the top 10 players by average market value for the 2024-2025 season. It includes key performance indicators such as position, goals scored, total appearances, total minutes played, average minutes per appearance, and average market value in millions of Euros.

Insight

The cumulative table highlights that attacking players generally command the highest market values, occupying six of the top ten spots. Vinicius Junior and Erling Haaland share the highest average market value at €200 million. Even though goals scored appear to be a significant factor in valuation (as seen with Haaland and Mbappé), other factors such as overall performance, and position also have a crucial role, as a slight majority of the players with the higher market value are attackers. However, strong midfielders also share high market values like the highly valued midfielders like Jude Bellingham and Jamal Musiala. Interestingly, Rodri, a midfielder with a high market value, has played significantly fewer minutes and scored

no goals in the appearances listed, suggesting his valuation is likely based on his crucial role and impact beyond direct goal contributions.

Market Value & Win Percentages



Top 15 Teams by Win Percentages and Player Metrics

Conclusion

This project examined how player attributes influence market value and team success in top European leagues. We found that attackers peak earlier and hold higher market values, while goal contributions and minutes played are strong indicators of valuation. Team-level data showed that clubs with higher average player values tend to win more. Overall, the analysis highlights how performance metrics and positional roles shape both individual worth and team outcomes in modern football.

Code Appendix

```r
# Load required libraries
library(readr)
library(ggplot2)
library(dplyr)

# Load data
appearances <- read_csv("C:/Users/maxwe/OneDrive - The Pennsylvania State University/Downloa

players <- read_csv("C:/Users/maxwe/OneDrive - The Pennsylvania State University/Downloads/da

player_valuations <- read_csv("C:/Users/maxwe/OneDrive - The Pennsylvania State University/D

# Define top 5 European football leagues
top5_leagues <- c("GB1", "FR1", "IT1", "ES1", "L1")  # Premier League, Ligue 1, Serie A, La L

# Define season date range
season_start <- as.Date("2024-08-01")
season_end <- as.Date("2025-06-30")

# Clean and filter appearances data
appearances_top5 <- appearances %>%
  filter(competition_id %in% top5_leagues) %>%
  mutate(date = as.Date(date))

# Clean and filter player valuations data
player_valuations_top5 <- player_valuations %>%
  filter(player_club_domestic_competition_id %in% top5_leagues) %>%
  mutate(date = as.Date(date))

# Prepare player data with standardized birth date
players <- players %>%
  mutate(birth_date = as.Date(substr(date_of_birth, 1, 10)))  # Extract date part

# Filter valuations for current season and calculate player age
valuations_season <- player_valuations_top5 %>%
  filter(between(date, season_start, season_end)) %>%
  left_join(players %>% select(player_id, birth_date), by = "player_id") %>%
  mutate(age = as.numeric(difftime(date, birth_date, units = "days")) / 365.25)

# Calculate minutes played per player for current season
player_season_stats <- appearances_top5 %>%
  filter(between(date, season_start, season_end)) %>%
```

```r
  group_by(player_id) %>%
  summarize(
    total_minutes = sum(minutes_played, na.rm = TRUE),
    matches_played = n(),
    avg_minutes_per_match = total_minutes / matches_played,
    .groups = "drop"
  )

# Merge most recent valuation, minutes played, and player position
player_analysis <- valuations_season %>%
  group_by(player_id) %>%
  arrange(desc(date)) %>%
  slice(1) %>%
  ungroup() %>%
  left_join(player_season_stats, by = "player_id") %>%
  left_join(players %>% select(player_id, position), by = "player_id")

# Create age bins and rounded age for analysis
player_analysis <- player_analysis %>%
  mutate(
    age_bin = cut(age, breaks = seq(16, 40, by = 2), include.lowest = TRUE),
    age_rounded = floor(age)
  )

# Create the scatter plot with facets for each position
ggplot(player_analysis, aes(x = age, y = total_minutes)) +
  geom_point(aes(color = position), alpha = 0.7) +
  geom_smooth(method = "loess", se = TRUE, aes(color = position), alpha = 0.2) +
  facet_wrap(~ position, ncol = 2) +
  scale_color_manual(values = c(
    "Attack" = "#FF5733",
    "Midfield" = "#33A1FD",
    "Defender" = "#9370DB",
    "Goalkeeper" = "#2ECC71"
  )) +
  scale_x_continuous(breaks = seq(16, 40, by = 4), limits = c(16, 40)) +
  labs(
    title = "Age vs Minutes Played by Position",
    subtitle = "Top 5 European Leagues (2024-2025 Season)",
    x = "Age (years)",
    y = "Total Minutes Played",
    color = "Position"
```

```
  ) +
  theme_minimal(base_size = 14) +
  theme(
    strip.background = element_rect(fill = "lightgray", color = NA),
    strip.text = element_text(face = "bold", size = 16),
    axis.title = element_text(face = "bold", size = 14),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 13),
    legend.text = element_text(size = 11),
    plot.title = element_text(size = 18, face = "bold"),
    plot.subtitle = element_text(size = 14),
    legend.position = "bottom",
    panel.grid.minor = element_blank()
  )
```