A Brief Analysis of Soccer Data

Sabo Vlad-Andrei & Negru Bogdan Liviu

Faculty of Mathematics and Computer SCience, University of Bucharest,

25th of May, 2024.

Introduction

- The Premise: Analyzing data from FBREF spanning 2017-2021 and Transfermarkt in the summer of 2021.
- The Objective: Derive insights to predict player market values effectively.

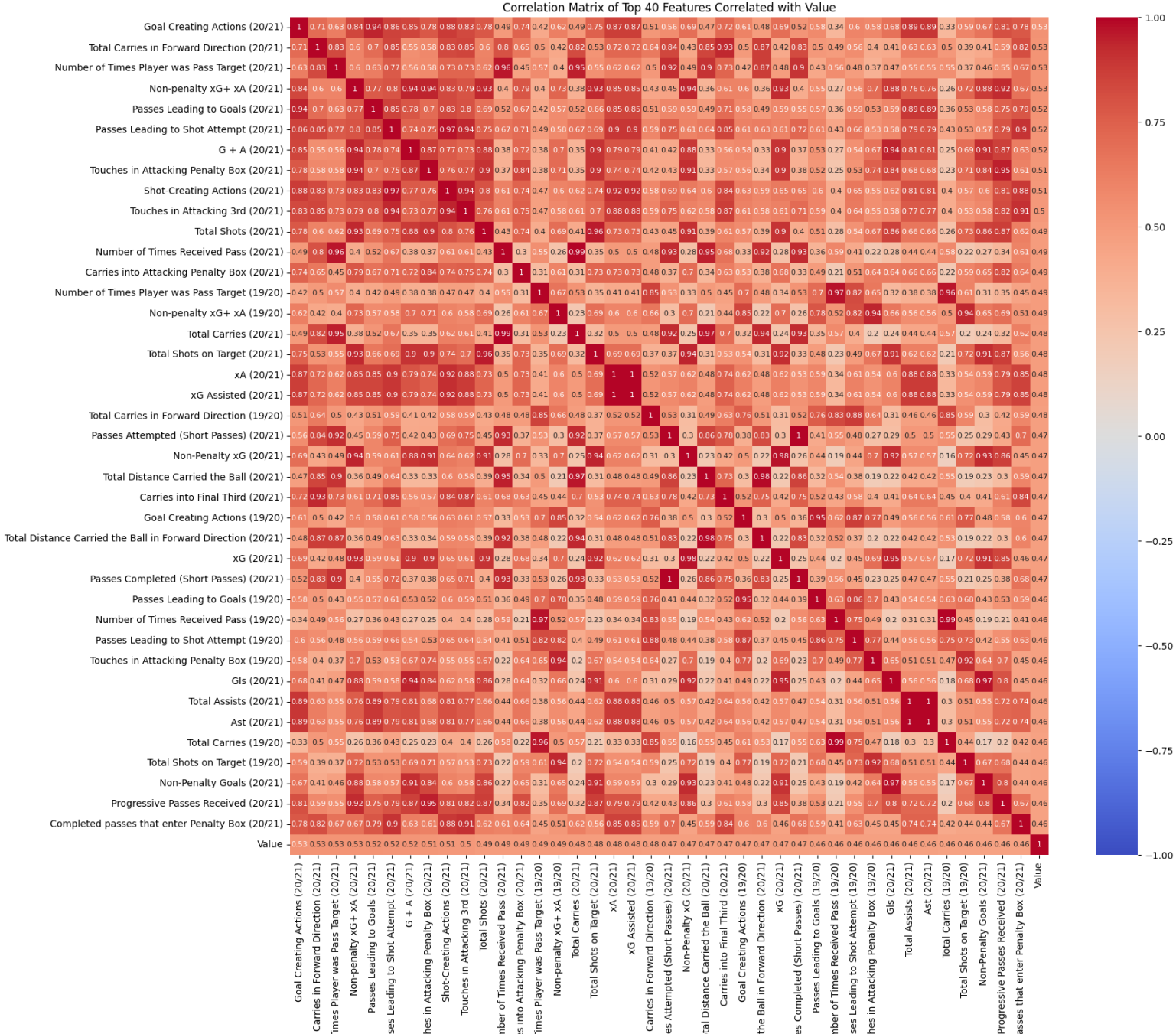
Data Overview

- Dataset includes player performance metrics and market values.
- Preprocessing involved handling missing or erroneus values and filtering relevant data.

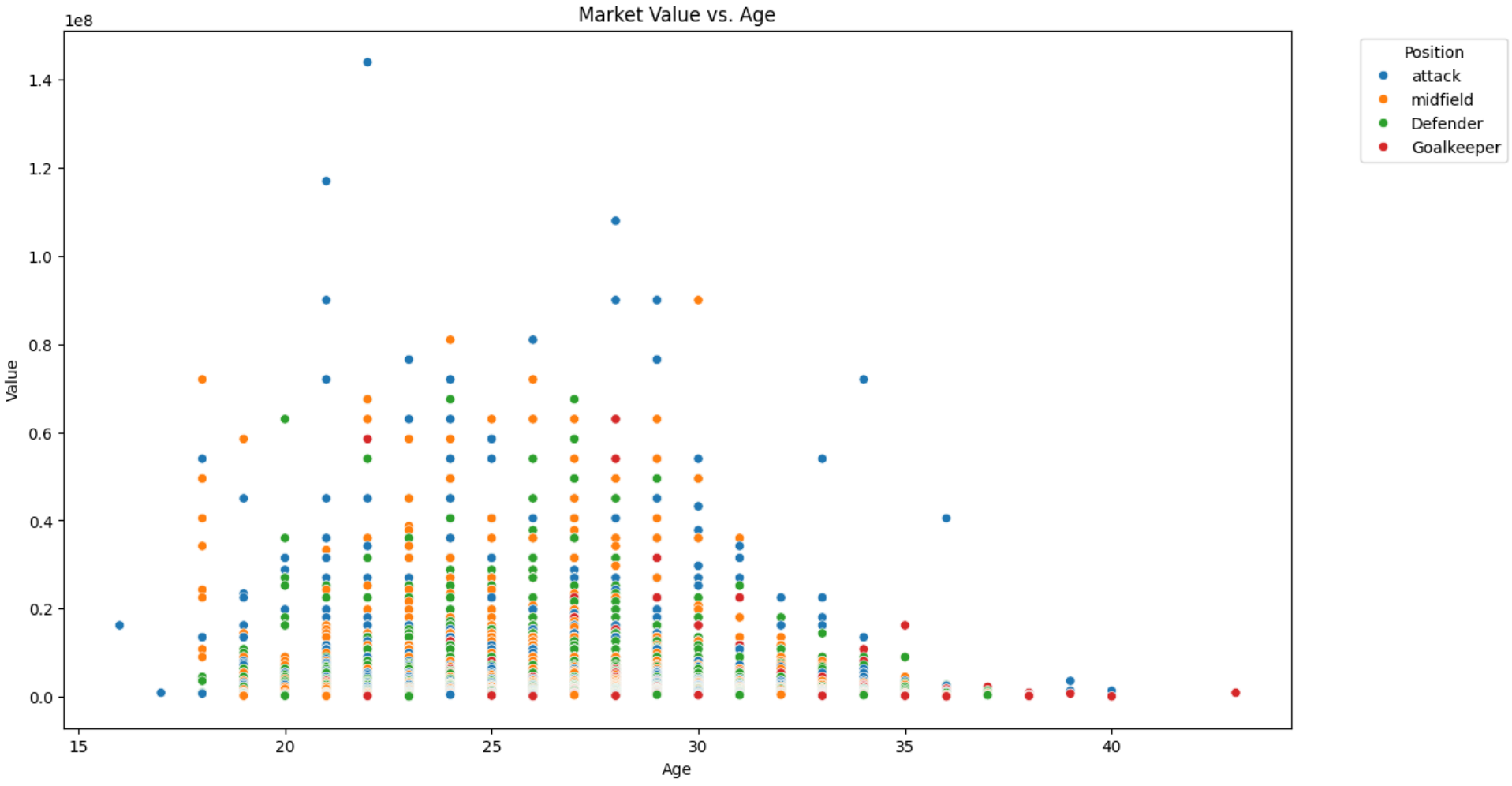
Exploratory Data Analysis - Insights

- Initial analysis focused on correlations between various factors such as nationality, league, position and related factors and the value.

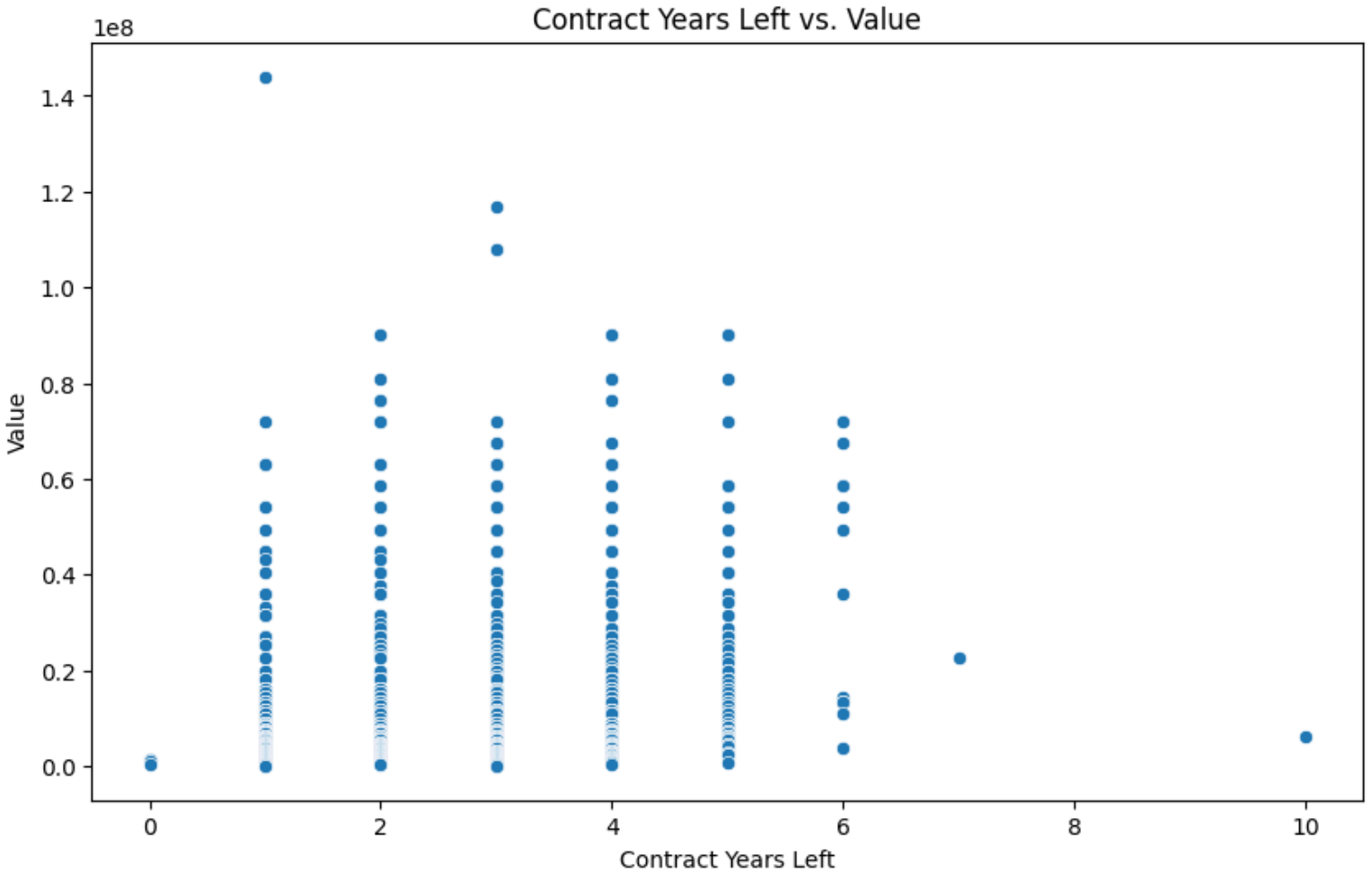
EDA - Initial Insights



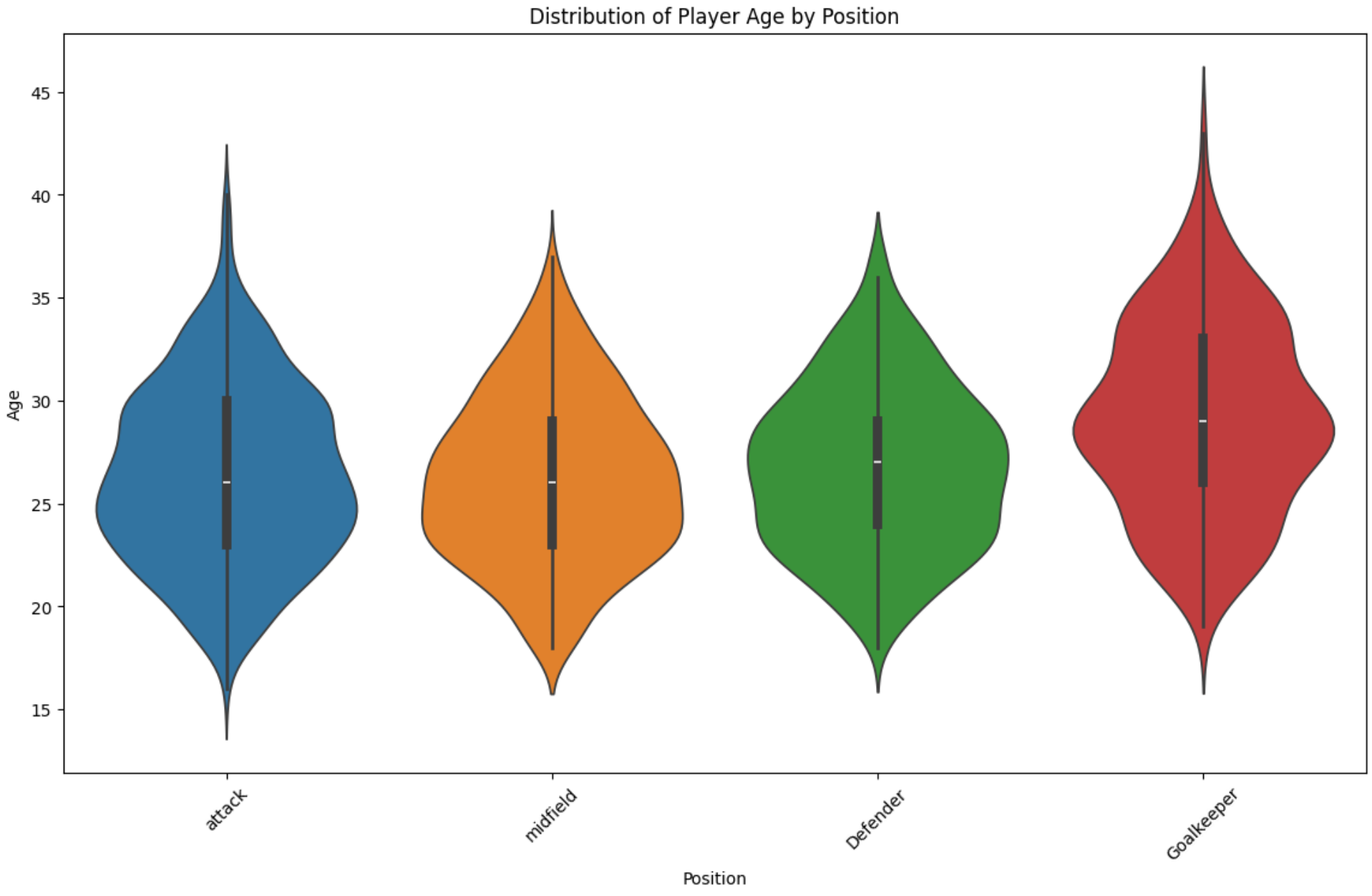
EDA - Market Value vs. Age



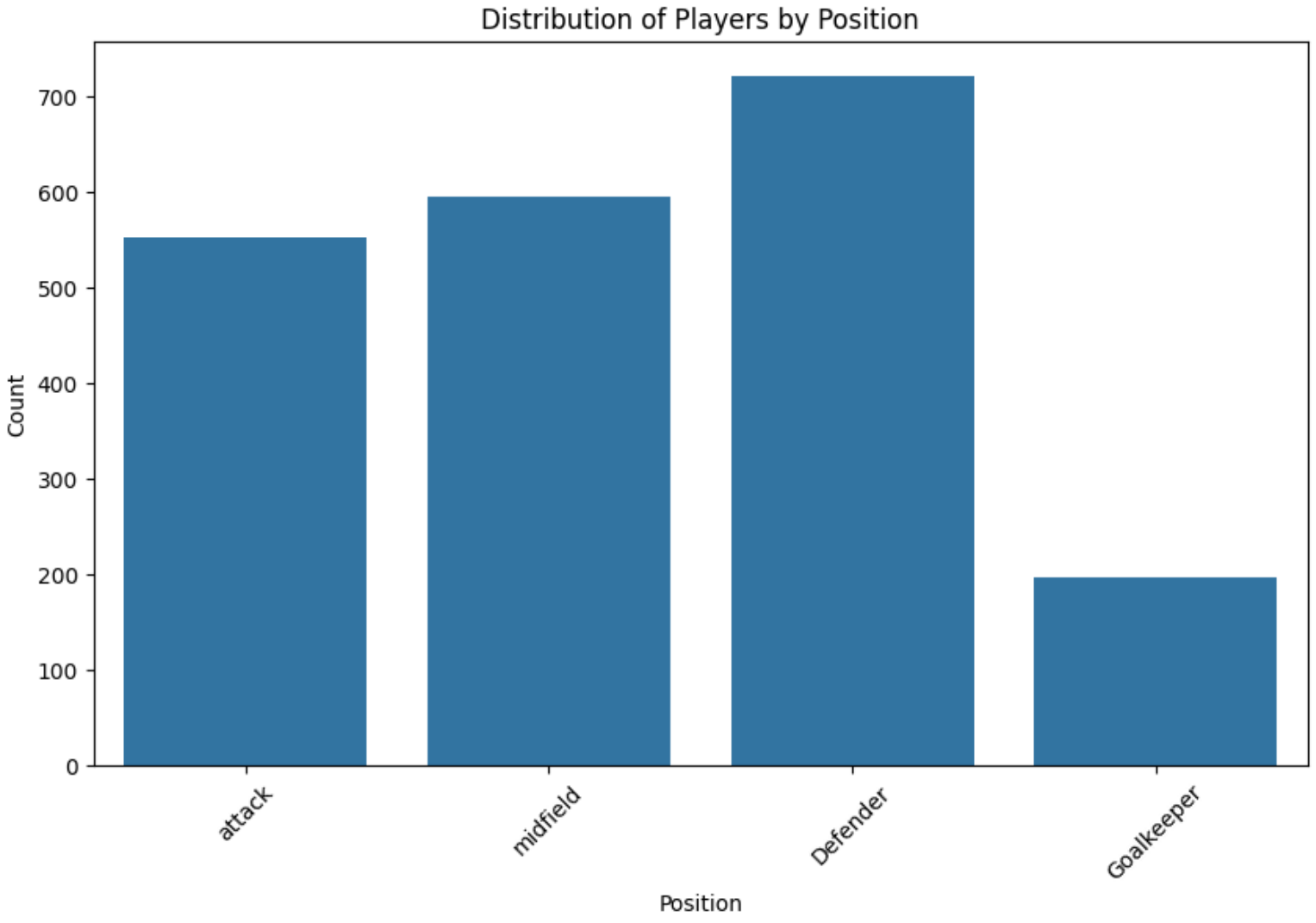
EDA - Contract Years Left and Value



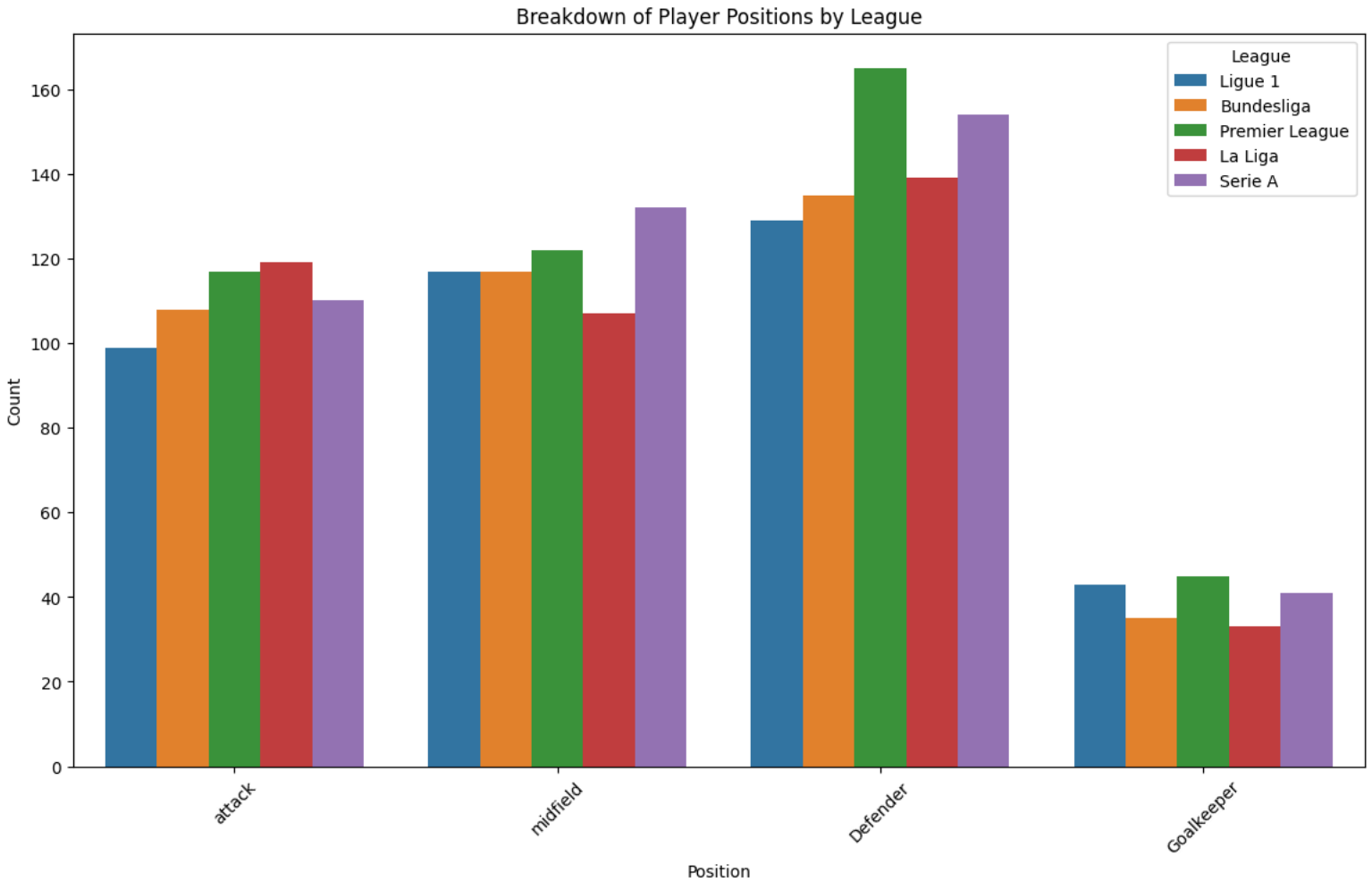
EDA - Distribution of Player Age by Position



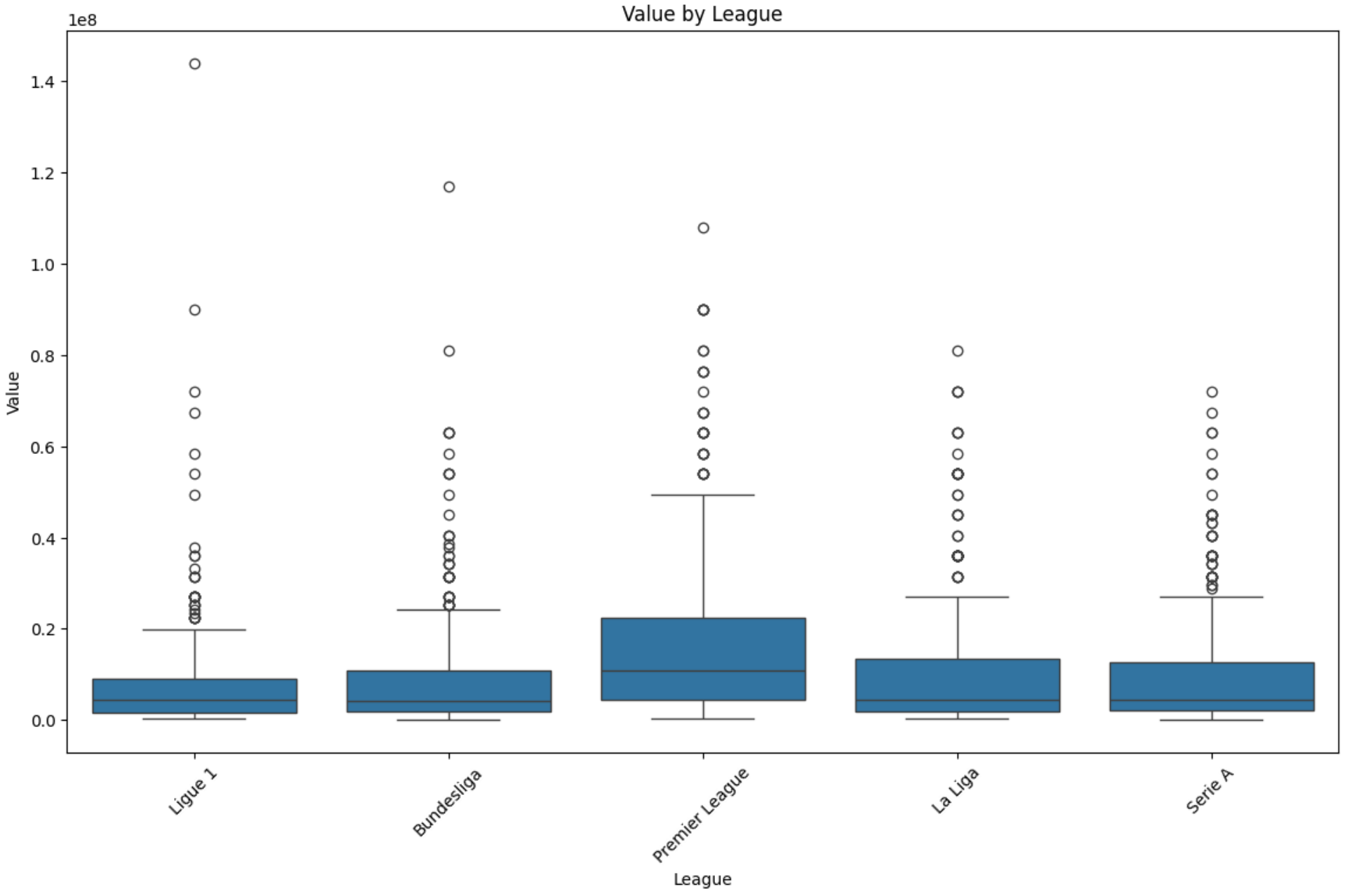
EDA - Distribution of Player Positions



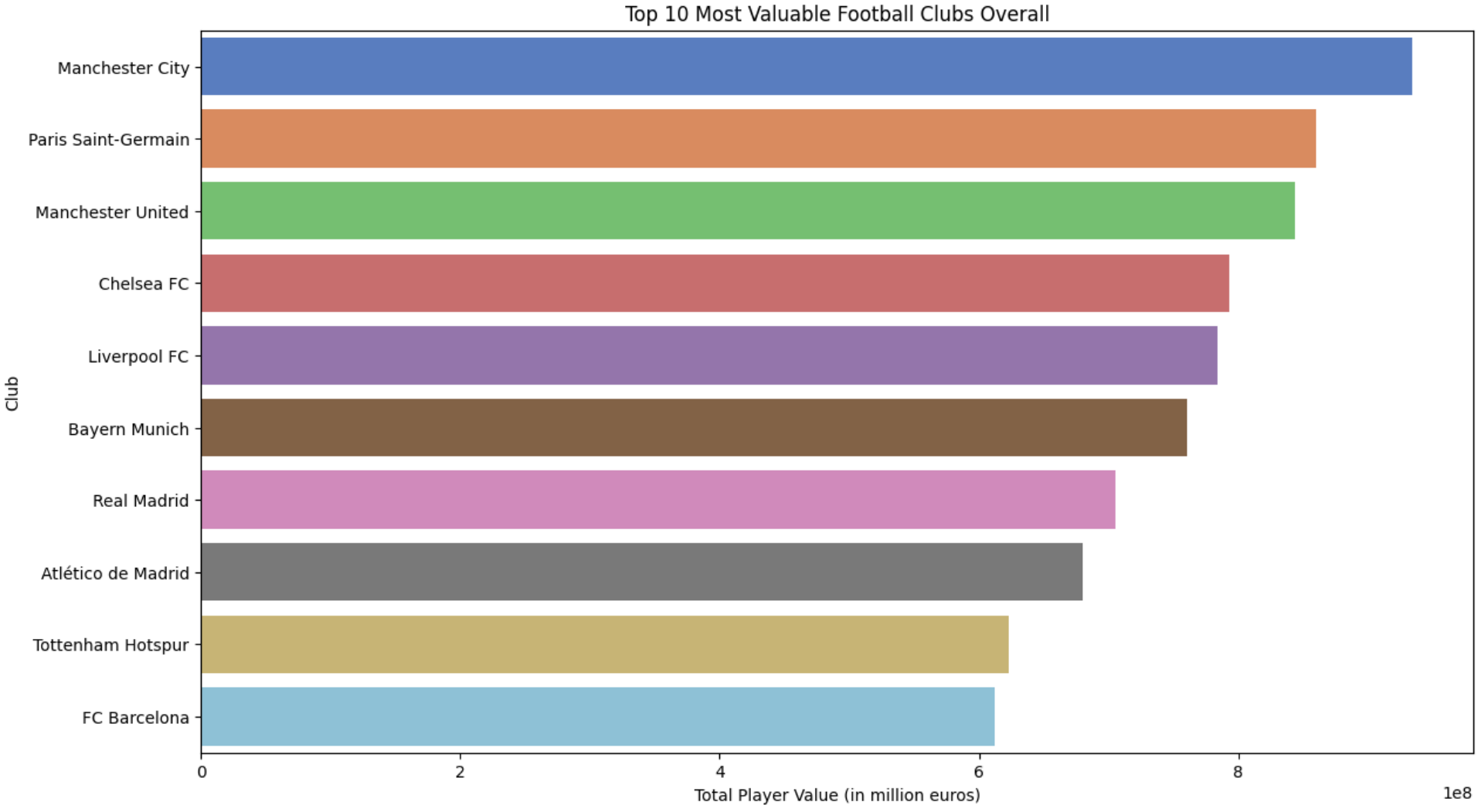
EDA - Position Breakdown by League



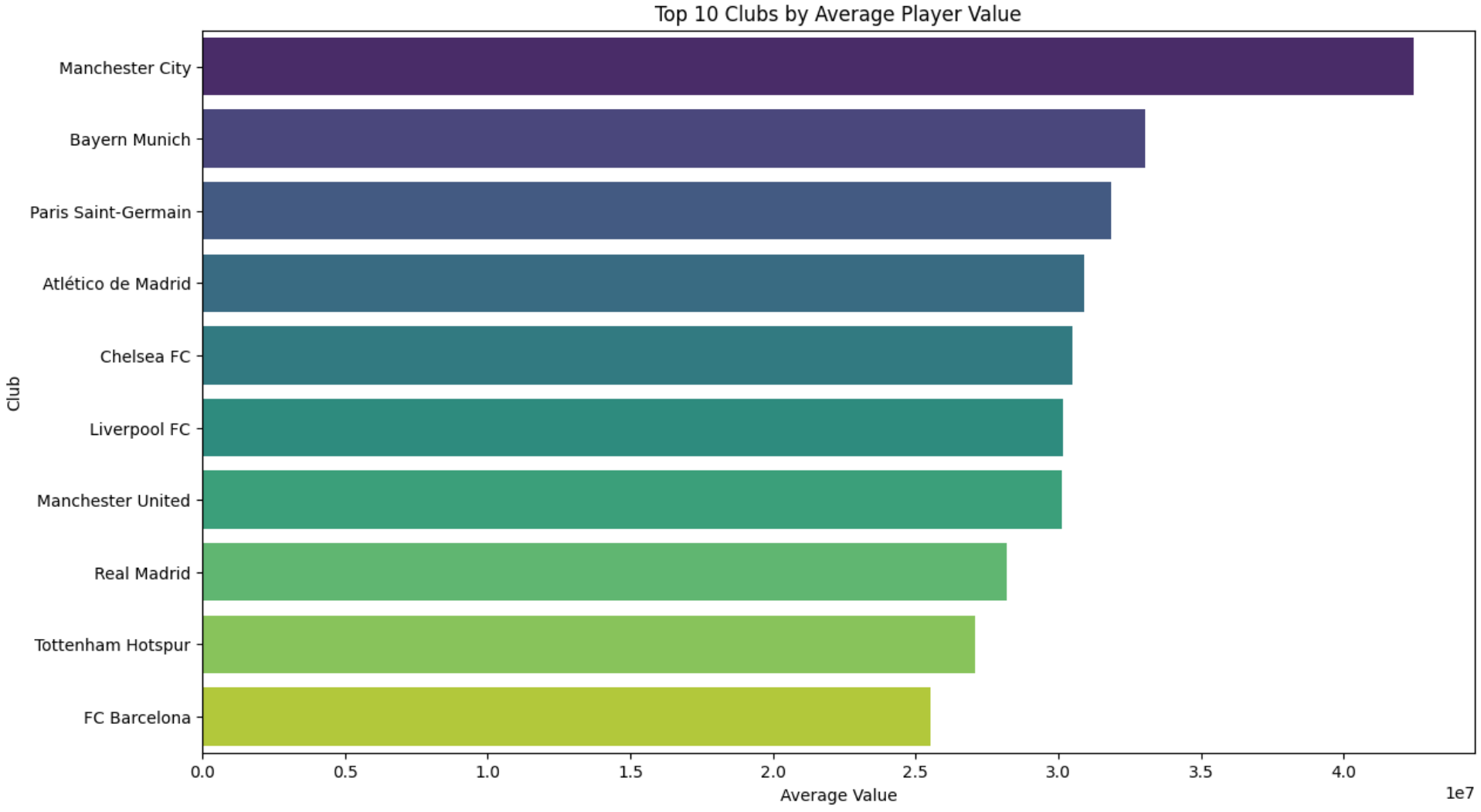
EDA - Market Value by League



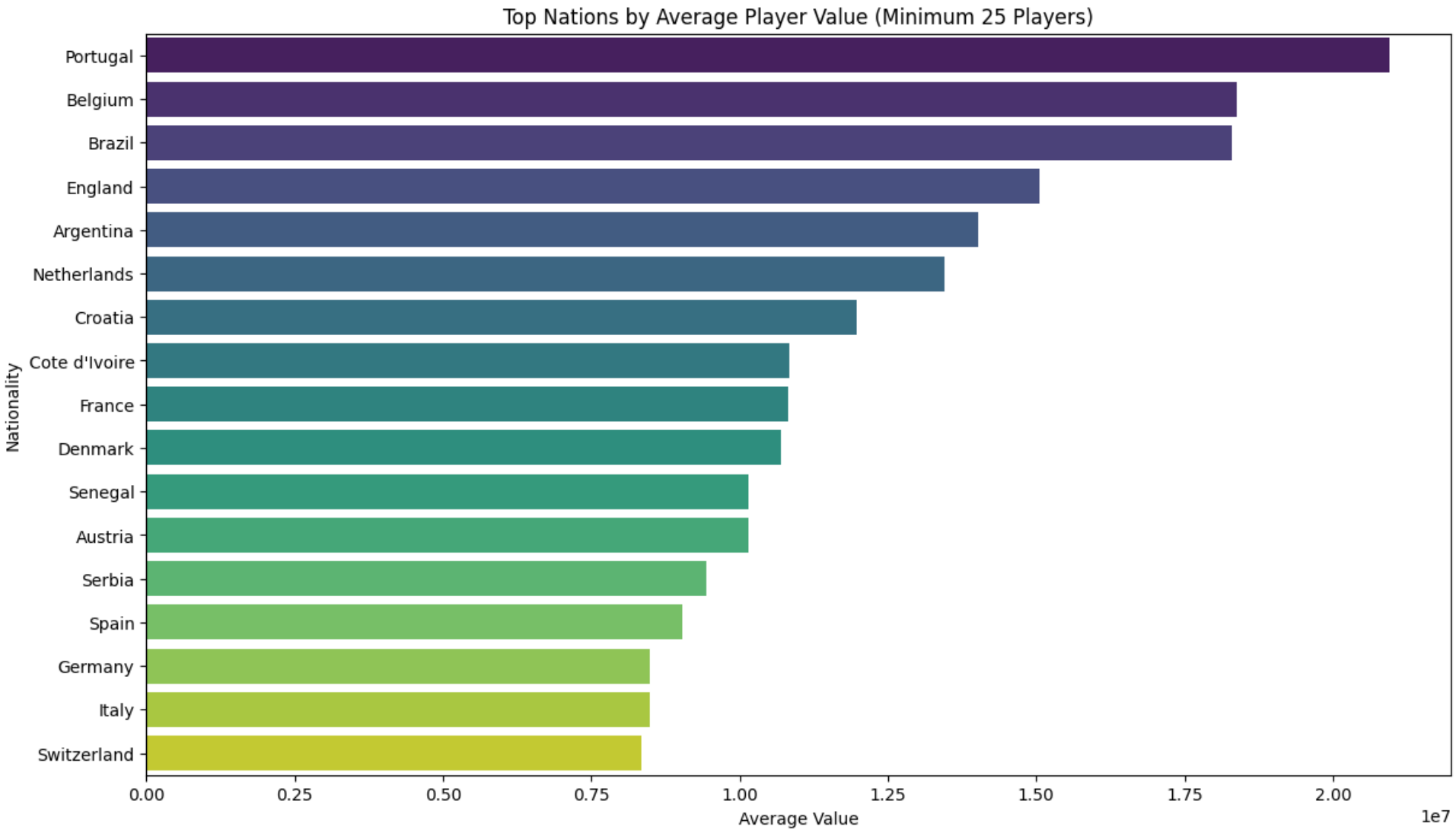
EDA - Top 10 Most Valuable Football Clubs



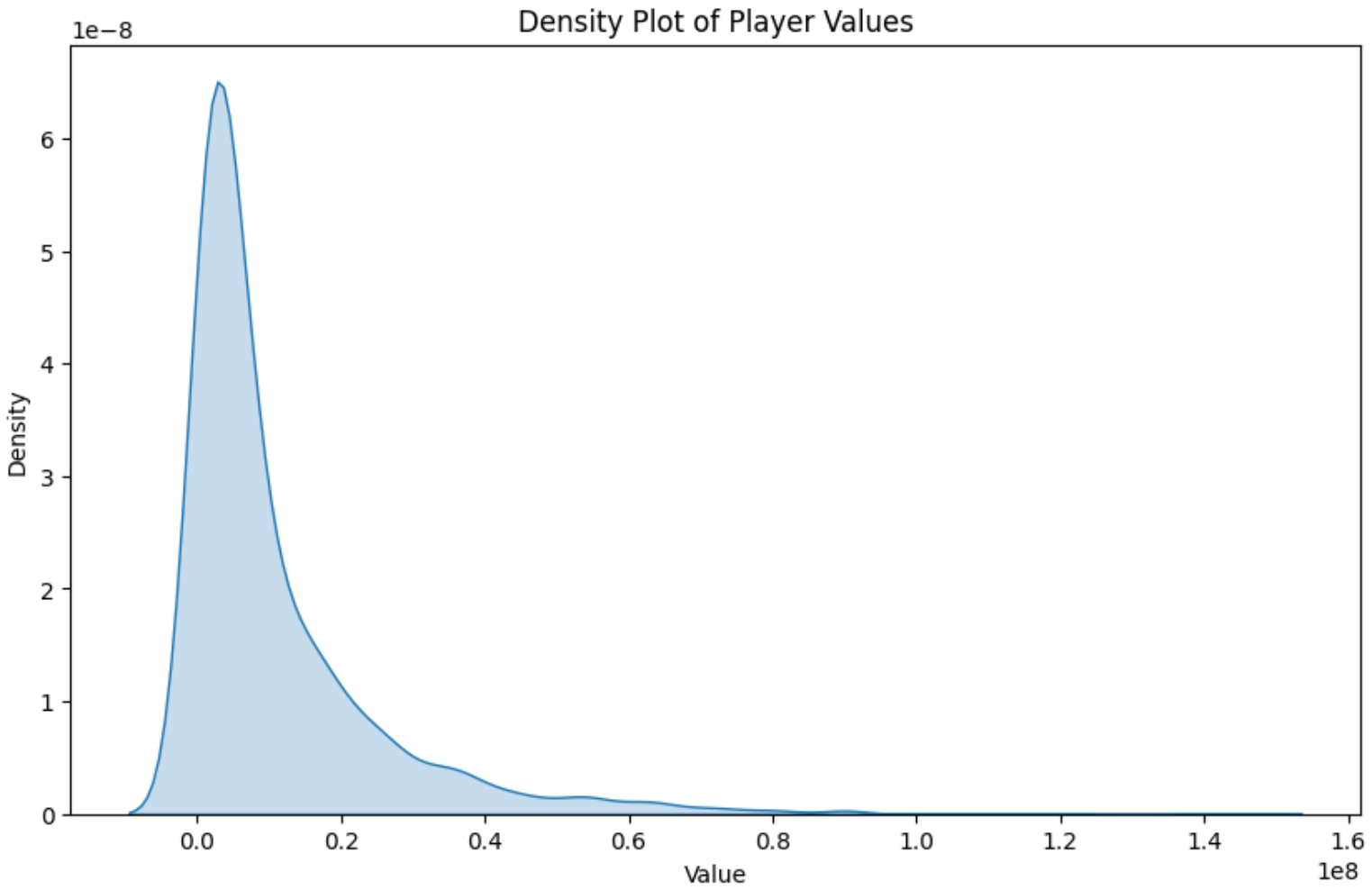
Top 10 Clubs by Average Player Value



EDA - Market Value Distribution by Nationality



EDA - Distribution of Player Values



Predictive Modeling

- Extra Data Preprocessing:
  - Search for relevant features related to value
  - Search for relevant statistics for each of the three outfield position (removed goalkeepers for this experiment)
  - Concatenate the features and use regression models to attempt to predict player values
- Models used: Linear Regression, Ridge Regression, and Random Forest.

Model Performance with Filled Missing Values

Model	**MSE** ($\times 10^{13}$)	**R-squared**
Linear Regression	8096.89	0.505
Ridge Regression	8096.24	0.505
Random Forest Regression	7858.02	0.520

Table: Comparison of model performance with filled missing values.

Model Performance with Dropped Missing Values

Model	**MSE** ($\times 10^{13}$)	**R-squared**
Linear Regression	12262.24	0.605
Ridge Regression	12251.34	0.605
Random Forest Regression	11940.89	0.615

Table: Comparison of model performance with dropped missing values.

Conclusions

- The analysis demonstrates significant variations in model performance based on data handling strategies (filling vs. dropping missing values).
- Random Forest Regression consistently outperformed Linear and Ridge Regression across different data preprocessing methods, indicating its robustness in handling complex patterns in football data.
- Insights from visual and statistical analyses provide a foundation for more targeted player valuation and potential investment strategies in football clubs.

Future Work and Improvements

- Future work could explore additional variables, more sophisticated modeling techniques, and larger datasets to refine the predictions and insights.
- Improvements could be the exploration of additional relationships and inclusion of extra plots, as well as using different models.

Thank You

Thank you for your attention.

Questions and comments are welcome!