Sports Analytics: Analyzing Player Performance and Market Value

Abdullah Khan, Damian Moore and Phillip Liang

Background: This project focuses on sports analytics, specifically analyzing soccer (association football) player performance, statistics, and market value. The aim is to investigate how a player’s on-field performance influences their market value and transfer fees. Soccer, being one of the most popular sports worldwide, has an extensive amount of player and game-related data, yet the relationship between these statistics and a player's market value is not fully understood.

Motivation: The motivation for this project stems from a passion for soccer and the potential to uncover valuable insights for a range of stakeholders, including professional soccer club managers, analysts, and betting markets. While data analytics has become prevalent in many major sports, soccer remains relatively untapped in terms of data-driven valuation and predictive analysis. In fact, Muller et al. have highlighted that “data-driven approaches to estimating market value have not yet caught on in professional football. Football has long lagged behind other major sports in the use of data analytics”. [1] This gap presents an opportunity to explore how player statistics and game performance are correlated with market values, an area that could provide clubs with a competitive advantage in player trades and team management. Existing research has focused on aspects like age and its impact on player valuation, with Knutson observing that forwards experience a decline in goal-scoring performance after age 30. [2] Building on this, our project aims to explore a broader range of performance metrics—such as overall rating, penalties, ball control, free kick accuracy, and acceleration—to determine their influence on player market value and transfer fees.

Purpose of Project: The findings of this study could serve an audience of professional soccer club managers and decision-makers who are interested in integrating data analytics into their player acquisition and retention strategies. One tenet of this project is that data-driven approaches must be compatible with existing scouting and management practices, which is why the study focuses on commonly available player performance metrics rather than proprietary data. The project is designed to be an example of how clubs can make their teams more competitive by leveraging data analytics, especially in soccer—a sport that has historically lagged behind others in this area. The data from this study might guide expectations for how player performance metrics influence market values and transfer fees. This could help clubs determine how many resources would be necessary to acquire certain players, what kind of return on investment they might expect, and how to identify undervalued talent in the market. Additionally, the insights could assist in planning for player development, contract negotiations, and long-term team strategy, especially in markets where data analytics is not yet widely adopted.

Data Sources

Name	Description	Size	Access
European Soccer Database	The European Soccer Database is a dataset that includes detailed information on soccer matches, players, teams, and leagues from several European countries spanning from 2008 to 2016. It contains over 25,000 matches with rich match events like goal types, possession percentages, corners, crosses, fouls, and cards. The dataset features data on more than 10,000 players, including personal details such as names, height, weight, and extensive player attributes like overall rating, potential, penalties, free kick accuracy, acceleration, and ball control—sourced from EA Sports' FIFA video game series. Structured across seven interconnected tables (Country, League, Match, Player, Player Attributes, Team, and Team Attributes), it allows for relational analysis between different facets of the game. Additionally, it includes team formations using X, and Y coordinates and betting odds from up to 10 providers.	**Records**: Over **25,000 match records** and data on more than **10,000 players**. **Storage Size**: Approximately **30 to 40 MB** in CSV format.	**Location**: The dataset can be downloaded from Kaggle at European Soccer Database. **Access Method**: You can download the dataset directly from the Kaggle website or by using the Kaggle API. To download using the Kaggle API, execute the following command in your Jupyter Notebook or terminal: `!kaggle datasets download -d abdelrhmanragab/european-soccer-database`
Football Data from Transfermarkt	This dataset provides a collection of soccer player information sourced from Transfermarkt, a leading football statistics website. It includes detailed data on thousands of players from around the world, covering key aspects such as player names, nationalities, positions, current clubs, and historical market values. Spanning from the year 2000 to the present, the dataset offers valuable insights into player performance metrics, market valuation trends, and career progressions. Available in CSV format, it is ideal for analyzing how player attributes correlate with their market values, exploring transfer market dynamics, and conducting various sports analytics studies.	**Records**: Over 160,000 player records **Storage Size**: Approximately 300 MB in CSV format	**Location**: The dataset can be downloaded from Kaggle at Football Data from Transfermarkt. **Access Method**: You can download the dataset directly from the Kaggle website or by using the Kaggle API. To download using the Kaggle API, execute the following command in your Jupyter Notebook or terminal: `!kaggle datasets download -d davidcariboo/player-scores`

Data Cleaning and Manipulation

Merging Player Data with Attributes: To create a comprehensive dataset, we merged `euro_player_df` and `euro_player_attr_df` on the `player_api_id`. This resulted in a unified DataFrame `euro_merged_df`, which includes player’ personal details and their performance attributes.

Handling Missing Values: We assessed `euro_merged_df` for missing values using methods like `.isnull().sum()`. The dataset was relatively clean, with minimal missing values that did not significantly impact the analysis. Therefore, no imputation or removal of records was necessary.

Data Type Conversion: The `date` column, initially in string format, was converted to a DateTime object using `pd.to_datetime()`. This conversion was essential for any time-series analyses and ensured consistency in date-related operations.

Merging Player Information with Valuation Data: We merged `df_player` and `df_val` on `player_id` to create a comprehensive DataFrame `df_combined`. This DataFrame includes player details and their historical market valuations.

Handling Duplicates: We identified duplicate entries in `df_combined` due to multiple valuation records per player over time. To ensure data integrity, we removed duplicate entries based on `name`, keeping the most relevant records for our analysis.

Identifying Top Players: We sorted df_combined based on highest_market_value_in_eur to identify players with the highest peak market values. This step was crucial for focusing our analysis on players who have a significant impact on the market.

Extracting Valuation Histories: For the top five players identified, we extracted their valuation histories to analyze how their market values evolved over time. This involved filtering `df_combined` for each player and organizing their valuation data chronologically.

Sampling Data for Analysis: We also sampled approximately 2,000 players from `df_combined` to represent the broader player population in our analyses. This sample was used to compare general market value trends against those of the top players.

Filtering Specific Players: For initial exploratory analysis and to validate our approach, we filtered the dataset to focus on specific players, such as Lionel Messi. This involved selecting records from `euro_merged_df` where `player_name` matched the player of interest.

Grouping and Averaging: We grouped the data by `player_name` and resampled it annually using `resample('YE')`. This allowed us to calculate annual averages of performance metrics for each player, smoothing out short-term fluctuations and highlighting longer-term trends.

Data Merging

Aligning Data on Player Names: We recognized that both datasets contained player information, but the key identifiers differed (`player_api_id` vs. `player_id`). To merge the datasets, we aligned them on `player_name`, which was common to both.

Creating a Unified DataFrame: We merged the performance metrics from the European Soccer Database (`euro_merged_df`) with the market valuation data from the Transfermarkt dataset (`df_combined`) on `player_name`. This resulted in a unified DataFrame (`df_metrics`) containing both performance attributes and market values for each player.

Handling Discrepancies: During the merge, we addressed discrepancies such as players present in one dataset but not the other. Players without complete data across both datasets were excluded to maintain consistency in our analyses.

Resampling and Averaging Data: We resampled the combined data annually and calculated averages for both performance metrics and market values. This step reduced noise from short-term fluctuations and highlighted significant trends over time.

Calculating Statistical Metrics: We computed statistical measures such as mean and standard deviation for market values. This provided insights into the distribution of player market values and set the stage for further statistical analyses, like t-tests.

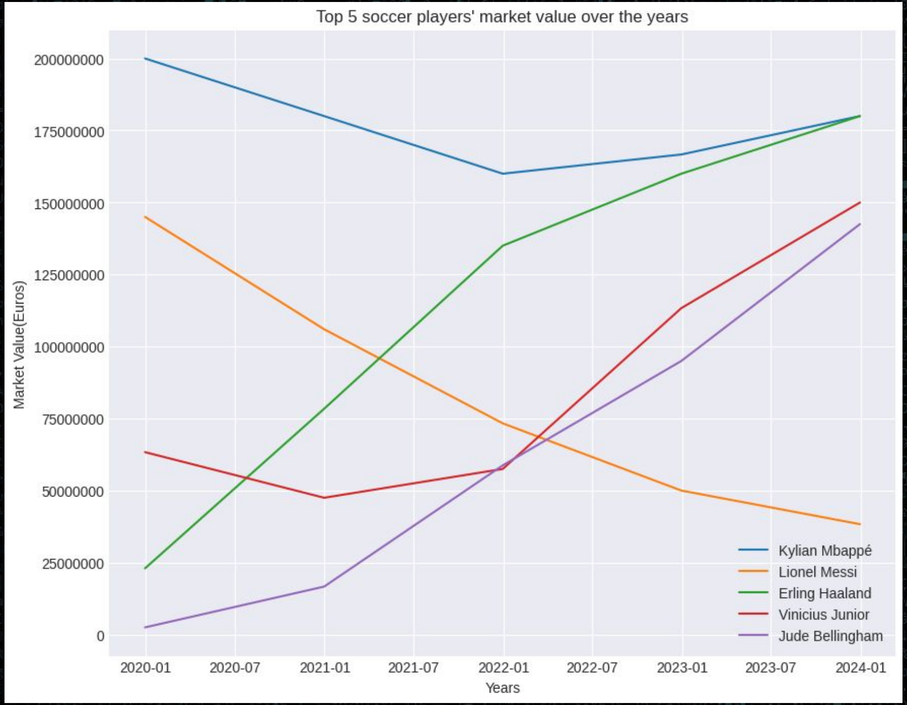
Identifying Groups for Comparison: We segmented players into groups, such as the top 100 players based on highest market values and the general player population. This allowed us to perform comparative analyses and understand differences between elite players and the broader pool.

Aggregating Market Values by Country: We grouped players by their `country_of_citizenship` and aggregated their highest market values. This provided total market value figures for each country, highlighting the global distribution of soccer talent.

Standardizing Country Names: To ensure compatibility with geospatial mapping tools, we standardized country names. This involved correcting inconsistencies and aligning country names with recognized international standards.

Merging with Geospatial Data: We integrated the country-level data with geospatial information, such as ISO country codes. This prepared the dataset for creating choropleth maps and other geospatial visualizations.

Analysis: Top 5 Soccer Players’ Market Value over the Years



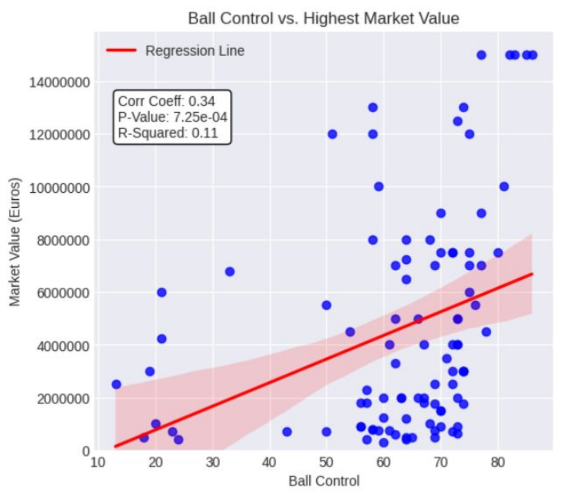
This line chart provides a visualization for the market values of the top 5 soccer players from 2019-2024. It is focused on the players who have had the highest peak market values historically in order to see the recent trends in their valuation histories. The market values are averaged yearly and plotted for each player to see their values over time.
The mentioned players were Kylian Mbappé, Lionel Messi, Erling Haaland, Vinicius Junior, and Jude Bellingham. Mbappé shows a slight dip at the beginning of 2022 while Lionel Messi's market valuation steadily declines; this could be due to him entering his 23rd season and his age. Despite this, he still held a very respectable market value of about 38 million Euros in 2024, at the age of 37! At that age, most other players typically retire or decline heavily in performance. The average soccer player does not have a valuation remotely close to that number. In fact, the peak market value of the average player was calculated to be about $4 million. This shows that top players are well above the competition.

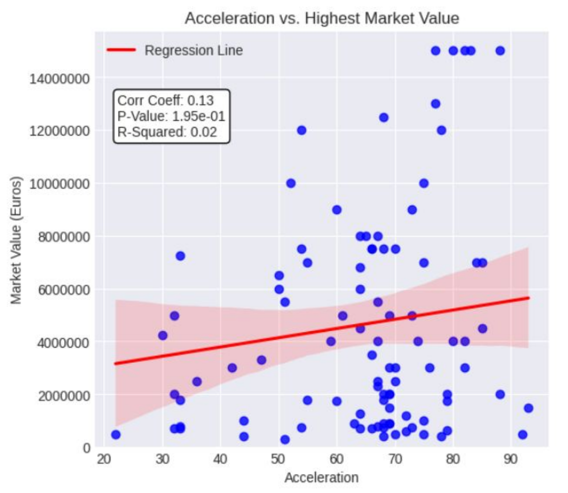
The rising stars are Haaland, Junior, and Bellingham. They are all young players(20-25 year-olds), entering their prime years, and will likely be the upcoming soccer stars for years to come. The visualization effectively reveals hidden dynamics in the soccer market and predicts the players the market will likely be focused on in the future. It also hints at the influence of factors like age on valuation.

Analysis: Attribute Scatter Plots

Scatter Plots

The following series of scatter plots visualize the relationship between a player's performance metric and their highest recorded market value. Each dot represents a player, with their performance metric on the x-axis and their market value (in Euros) on the y-axis. The red regression line shows the trend between the variables, and the shaded area around the line represents the confidence interval(95%). The boxes provide statistics such as correlation coefficients, P-values, and R-squared coefficients. Five performance metrics were examined: overall rating, penalties, free kick accuracy, acceleration, and ball control. We chose to remove the outliers and select a random sample of about 2000 players for the regression(100 visible in the visualization).





General observations

The positive slope of the regression line suggests that there is an upward trend: as a player's metric increases, their market value tends to increase. There is a very wide range of values for players with similar metrics, indicating that while the given metric is a contributing factor, other variables also influence market value.

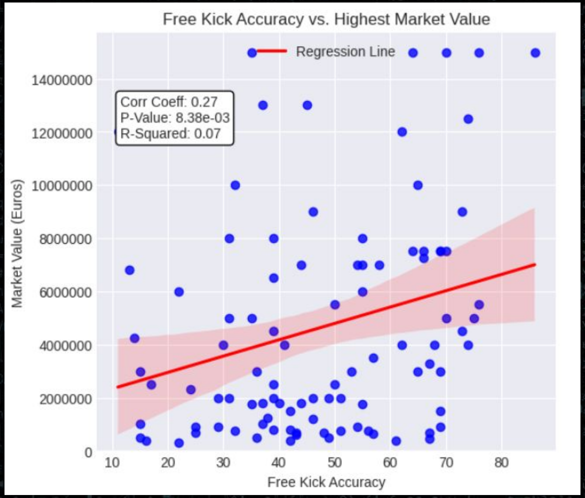
Ball Control

A higher ball control weakly correlates with a higher market value, with a correlation coefficient of about 0.3. The points in this plot appear to be distributed differently than in later plots. A few players have a ball control between 10 and 50, while the vast majority of players cluster around a ball control of mid 50s to mid 80s. Though weak, the effect is significant with a p-value under 0.05.

Acceleration

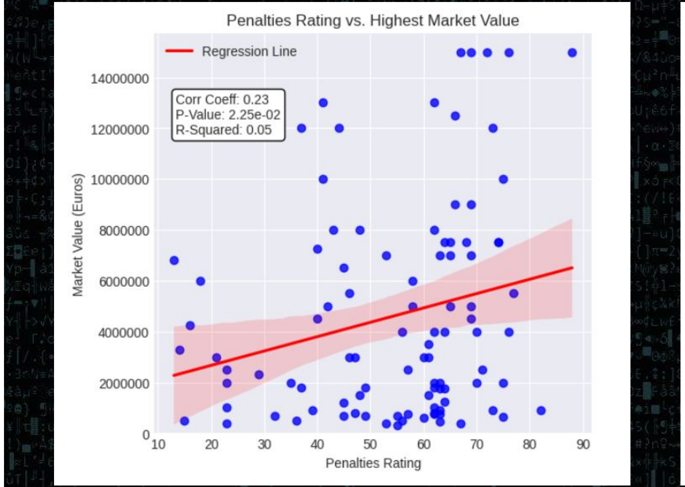
Acceleration has a very weak positive correlation with market value, with a correlation coefficient of about 0.1. There is a very wide range of values for players with similar acceleration, especially for players with an acceleration between 60 and 80. The p-value is not significant either(over 0.05). Thus, it appears that a player's market value is not really affected by how well the player accelerates.

Analysis: Attribute Scatter Plots



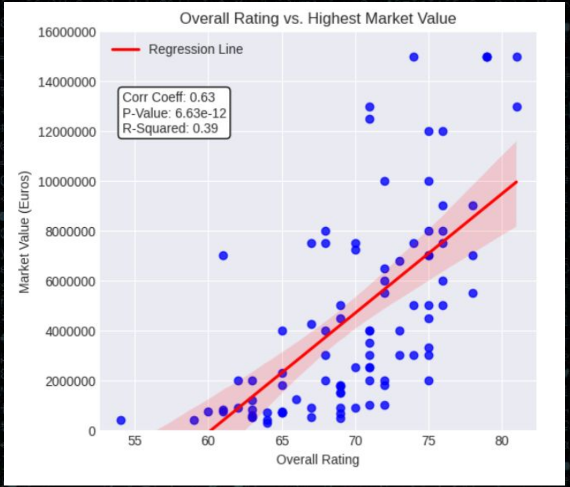
Free Kick Accuracy

The points are distributed in a similar way as in the acceleration plot. A higher free kick accuracy weakly correlates with a higher market value, and the correlation coefficient is about 0.3. There is a very wide range of values for players with similar accuracies, such as players with accuracies between 40 and 70. Thus, it appears that a player's market value is not really affected by how good the player is at taking free kicks.



Penalties Rating

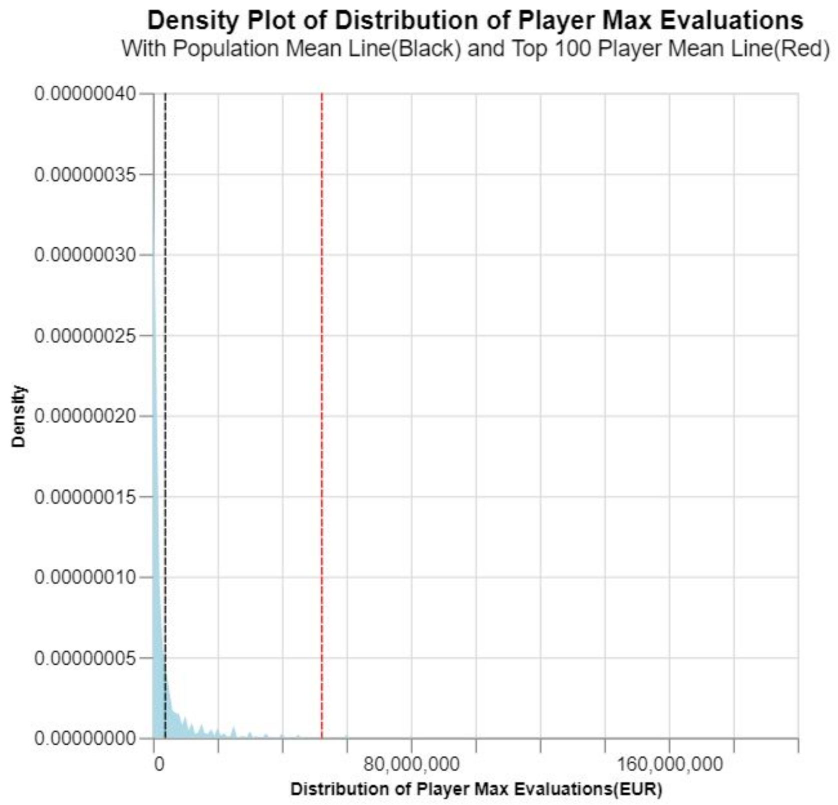
The general upward trend indicates that a higher penalty rating weakly correlates with a higher market value, and the correlation coefficient is about 0.2. Players with a 70 to 80 penalty rating do not tend to have a significantly higher market value than players with a 40 to 50 rating. Thus, it appears that a player's market value is not really affected by how good the player is at taking penalty kicks.



Overall Rating

Player ratings in the mid-60s and 70s show significant variations in their highest market value, possibly attributed to several external factors like age, player position, injuries, or recent performance. In spite of that, the general upward trend reinforces the idea that a higher overall rating correlates with a higher market value, and the correlation coefficient is about 0.6. For the few players with player ratings above 80, all of their highest market values are above 10 million Euros, supporting this correlation.

Analysis: Market Value Distribution of Soccer Players



Using the maximum valuation from each player's history, we created a distribution to see how player market values compared when they were at the peak in their careers.

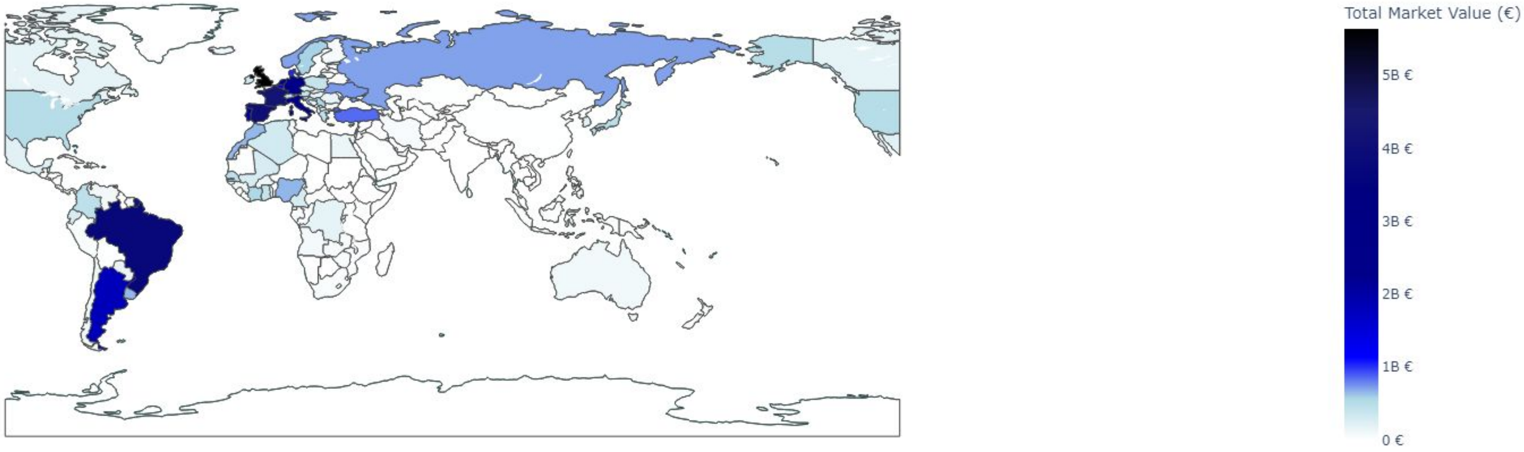
The x-axis represents the maximum market valuation of players in Euros, while the y-axis indicates the density.

The plot highlights a highly skewed distribution, where most players' market values cluster near the lower end of the spectrum.

- The blue graph represents the max value density distribution of a random sample of 4000 players.
- The black dashed line represents the sample population mean, which is situated far to the left meaning the average market value is relatively low.
- The red dotted line represents the mean market value of the top 100 players, and it is higher up the x-axis when compared to the black line representing the average player.
- The t-test when comparing the population mean and top 100 player's max valuations is just above -312.24, while the p-value is basically zero. This indicates a very significant difference between the two groups that cannot be explained by chance.

This reflects on the disparity in value that occurs between the average and top players, and how rare it is to make it to the top of the sport, pointing to the sport's competitive nature.

Visualization: Global Distribution of Soccer Market Value



Global Distribution of Football Player Market Value and Player Citizenship

In order to see where the top valued players have historically come from, we created a choropleth map that shows the total market value distribution. The darker shades of blue indicate the countries with a higher total market value of players, while lighter shades represent lower values. The map shows the highest concentration of market value in South America and Europe, particularly in countries like Brazil, Argentina, Spain, the United Kingdom, France, Germany, and Italy. These regions are well-known for their deep soccer heritage, and this visualization clearly represents that. This pattern highlights Europe's dominance in professional leagues and international transfers.

This map could potentially be a reference for managers to scout for new talent based on where money has been invested historically.

Summary

We have identified that there is a significant difference between the top soccer players when compared to the average player.

- First, the higher-valued players perform better than the average player in most metrics. However, there isn't a lot of disparity and the effect is not significant in many cases.
- Second, they tend to maintain an above-average market value, even into their declining years past the age of 30 in some cases.
- Third, top young players entering their primes can peak in the market value of well above 150 million Euros in the near future.

In general, we found a weak positive correlation between player statistics and market values. Overall rating has a stronger correlation with market values than ball control, acceleration, free kick accuracy, and penalties. Players with a higher peak market value do not tend to perform significantly better than weaker players. It is possible that external factors like age, player position, injuries, or recent performances are playing a role here, especially player position. Players with a relatively high peak market value and low penalties and free kick accuracy, for instance, would more likely be defenders or goalkeepers than midfielders or forwards. The latter category would be stronger in those metrics, given that they are offensive traits. On the other hand, there can be a goalkeeper with a high overall rating who is well-paid, and metrics like acceleration or ball control are less relevant when assessing the performance of a goalkeeper.

Most soccer players' market values cluster near the lower end of the peak market values, such as between 0 and 10 million.
The average peak market value of the top 100 players is well above that of most players, indicating that it is exceptionally rare to make it to the list of top 100 soccer players.
This points to the significantly competitive nature of the sport.
There is a significant gap between the mean of top and average market values which cannot be explained due to chance alone.

Europe and South America have the highest aggregate player market values in the world, indicating their deep soccer heritage and history of success. These regions are likely where the market will be focused in the future, as well as where the top talents are likely to be found.

Statement of Work and References

This team project was highly collaborative, with every team member contributing to brainstorming, analyzing, cleaning, and visualizing the data. We frequently submitted project updates to our mentor and had weekly check ups via messaging in Slack. Questions and concerns were addressed in a timely manner via Slack as well. A more specific breakdown of tasks is shown below.

Abdullah Khan

Abdullah led the efforts in data retrieval and cleaning. He leveraged the Kaggle API for efficient data acquisition and was instrumental in building the datasets used for most of the visualizations. Abdullah created the histograms/density plots and organized the tools and workflows within the Google Colab notebook, ensuring seamless collaboration among team members. He also contributed significantly to the markdown formatting and documentation within the Jupyter Notebook, providing clear explanations of the processes and findings. Additionally, Abdullah oversaw code development and performed thorough proofreading to maintain code quality and consistency throughout the project.

Damian Moore

Damian initiated the project by developing the initial draft of the project proposal, including the background and initial hypotheses. He scheduled meetings with data vendors to explore access to various data sources. Damian was responsible for building the choropleth map visualization and contributed to the development of line plots. His prior experience working with sports datasets in a professional capacity provided valuable domain knowledge to the team. Damian also developed narratives and performed analysis for the visualizations, offering insights into the data and contributing to the overall understanding of the findings.

Phillip Liang

Phillip brought expert domain knowledge to the project and was instrumental in building the scatter plots used in the analysis. He provided expert analysis of outliers and overall market value trends, offering unique perspectives on the data. Phillip contributed innovative ideas on methods to analyze the data and developed comprehensive narratives for the visualizations and analyses. He also played a key role in proofreading the written documentation and ensuring proper formatting, enhancing the clarity and professionalism of the final report.

Possible Future Improvement in Collaboration

In the future, it may be more effective to assess the strengths and weaknesses of each team member before allocating tasks to every member. For the Google Colab environment, multiple users cannot edit code at the same time, leading to issues in saving the project for the user who joined second. Thus, it may be better to let each other know who is going to be on the environment ahead of time and communicate that clearly.

References

[1] Muller, O. et al. Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operational Research. May 2017, https://repub.eur.nl/pub/120757/Repub_120757_O-A.pdf

[2] Knutson, T. Age and Value in the Transfer Market. StatsBomb. July 2013, https://statsbomb.com/articles/soccer/age-and-value-in-the-transfer-market/