2023

FOOTBALL INSIGHTS UNLEASHED: A DATA-DRIVEN JOURNEY

Hai Nguyen

Hai Nguyen

Why FIFA 2023 Players Analytics?
As a devoted football fan and FIFA gamer, I'm intrigued by the intersection of football statistics and data analytics. This FIFA 2023 Players Analytics project is born from my aspiration to apply analytics to FIFA data, exploring deeper insights into team and player dynamics.

Personal Connection:
This project is not just technical for me; it's a fusion of my background, love for football, and enthusiasm for data analytics. I'm excited about uncovering hidden patterns and contributing to a deeper understanding of the game.

OUTLINE

- EXECUTIVE SUMMARY
- INTRODUCTION
- DATA COLLECTION
- EXPLORATORY DATA ANALYSIS FOR TEAMS
- EXPLORATORY DATA ANALYSIS FOR PLAYERS
- MACHINE LEARNING MODELS for PREDICTION
- CONCLUSION
- APPENDIX

EXECUTIVE SUMMARY

- **Objective:**
- Provide comprehensive insights into the performance, characteristics, and values of football teams and players in the 2023 FIFA Players Analytics project.
- **Methodology:**
- **Data Collection:**
  - Utilized the "teams_fifa23.csv" and “players_fifa23.csv" for analysis.
- **Exploratory Data Analysis (EDA):**
  - Conducted thorough EDA to uncover data patterns, trends, and outliers.
  - Examined key performance metrics, characteristics, and values of teams and players.
- **Machine Learning Models:**
  - Applied machine learning models for predictions based on the collected data.
  - Leveraged algorithms to analyze and forecast various football player performance aspects.
- **Data Integration:**
  - Integrated insights from EDA and machine learning models to present a holistic view of the football landscape.

EXECUTIVE SUMMARY

Key Findings:

- **Teams Analysis:**
  - Distribution of Team Ratings (Overall, Attack, Midfield, Defence)
  - Comparison of Average Team Performance by Top Leagues
  - Identification of Most Competitive Leagues Based on Domestic Prestige
  - Recognition of Top Clubs Based on International Prestige
  - Analysis of Top 20 Teams' Composition and Performance
- **Players Analysis:**
  - Exploration of Physical Characteristics and Composition of Players
  - Analysis of Player Values, Wages, and Nationality Distribution
  - Identification of Top Players and Clubs by Value and Wages
  - Understanding Attacking and Defensive Work Rate Pairs
  - Insight into Mean Attribute Values Across Football Positions
  - Prediction Models for Overall Player Ratings and Player Positions

EXECUTIVE SUMMARY

Implications:

- Valuable insights for team managers, club owners, and scouts in player recruitment.
- Informed decision-making for league organizers and sponsors.
- Enhanced understanding of player attributes for fans and analysts.

INTRODUCTION

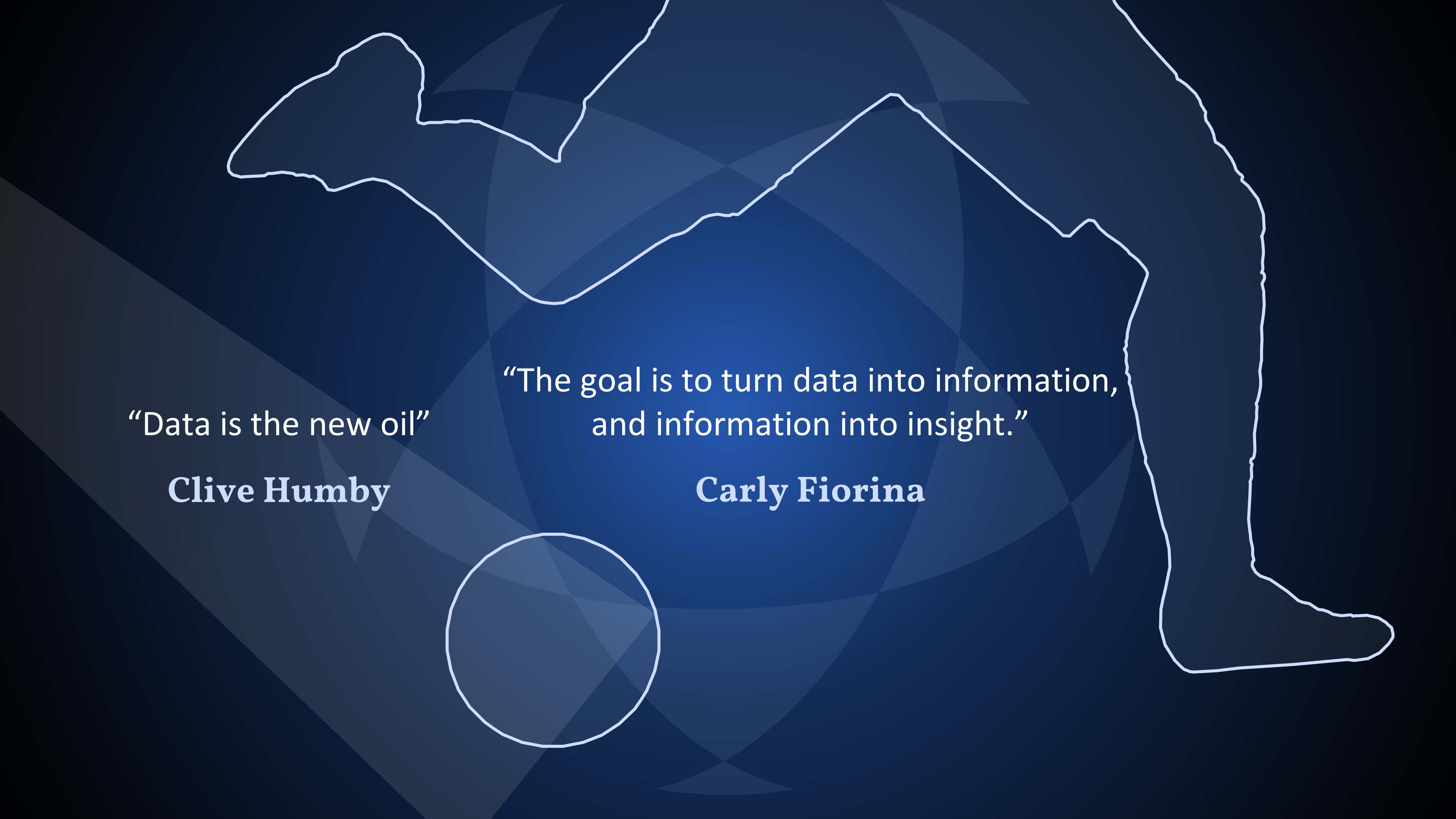
Background:
The world of football is not just a game; it’s a dynamic ecosystem where teams, players, and leagues continuously evolve. As I delve into the heart of this FIFA 2023 Players Analytics project, I embark on a journey to dissect and understand the intricate details that shape the landscape of global football.

Objective:
My primary objective is to unravel the hidden patterns and insights residing in the data, meticulously collected and curated from the expansive realm of FIFA. I aim to empower stakeholders with actionable information by employing advanced analytics and facilitating strategic decision-making in team management, player recruitment, and league dynamics.

INTRODUCTION

Significance:
The significance of this analysis extends beyond the pitch. From club owners seeking the next football sensation to fans craving a deeper understanding of their favorite teams, my insights bridge the gap between data and passion. This project isn't just about numbers; it's about unraveling the stories and strategies that define the beautiful game.

Join me on this journey through data, where every statistic tells a tale, and every trend reveals a facet of the footballing world waiting to be explored. Welcome to the FIFA 2023 PLAYERS ANALYTICS project – where data meets the pitch, and the game unfolds in numbers.



“Data is the new oil”

Clive Humby

“The goal is to turn data into information, and information into insight.”

Carly Fiorina

DATA COLLECTION

Source of Data:

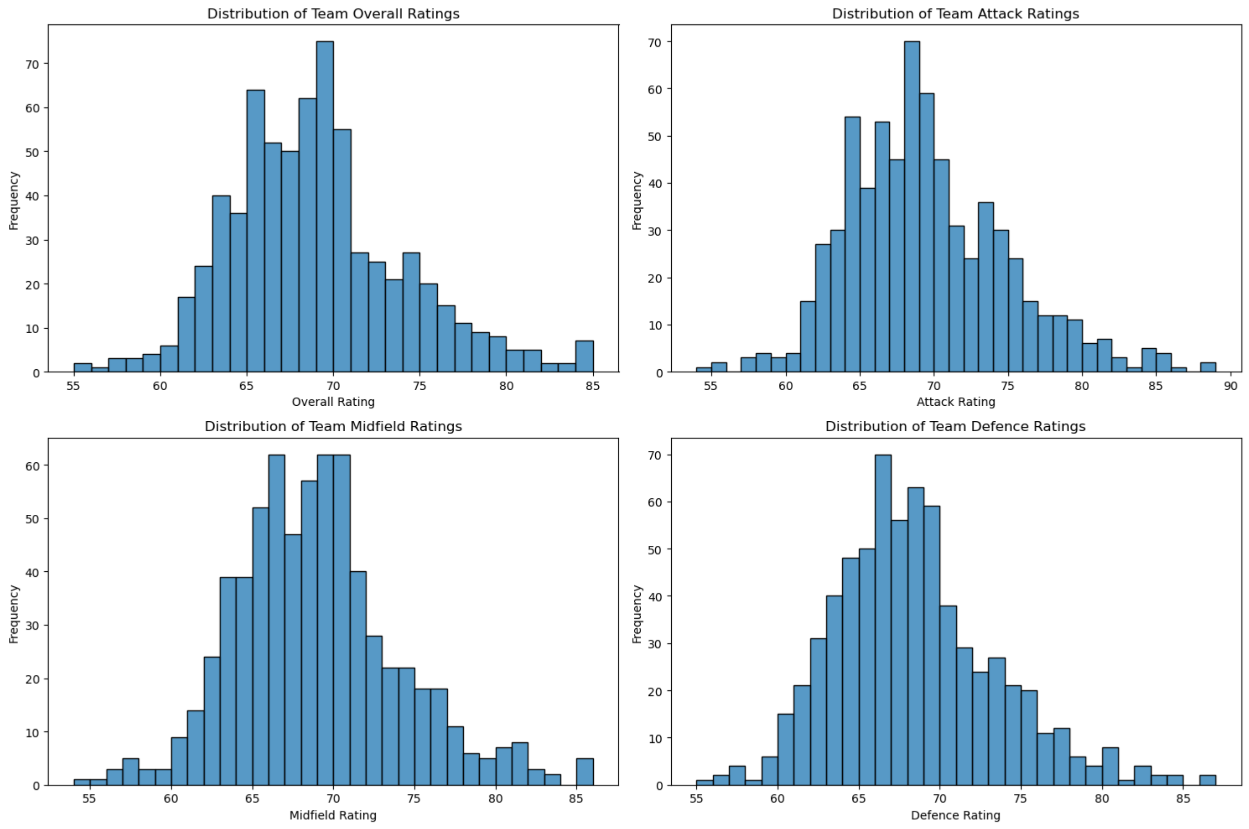
The foundation of my analysis lies in two robust datasets, meticulously curated from Kaggle, updated as of September 29, 2022. These datasets, named **players_fifa23.csv** and **teams_fifa23.csv**, serve as my primary reservoirs of information.

1. **players_fifa23.csv:**
   - This dataset encapsulates a comprehensive array of metadata, attributes, and ratings for a staggering 18,539 football players. Each entry is a detailed snapshot, offering information into the nuanced aspects of player performance and characteristics.
2. **teams_fifa23.csv:**
   - Complementing the player dataset, teams_fifa23.csv furnishes us with a panoramic view of team dynamics. Loaded with critical information, this dataset contains details on team attributes, and ratings, providing a holistic perspective on the collective strength of football clubs.

DATA COLLECTION

These datasets serve as the cornerstone of my exploration into the fascinating world of football analytics. The metadata, attributes, and ratings they contain pave the way for insightful analyses and predictions, propelling me into the heart of this 2023 FIFA Team and Players Analytics project.

EXPLORATORY DATA ANALYSIS for TEAMS



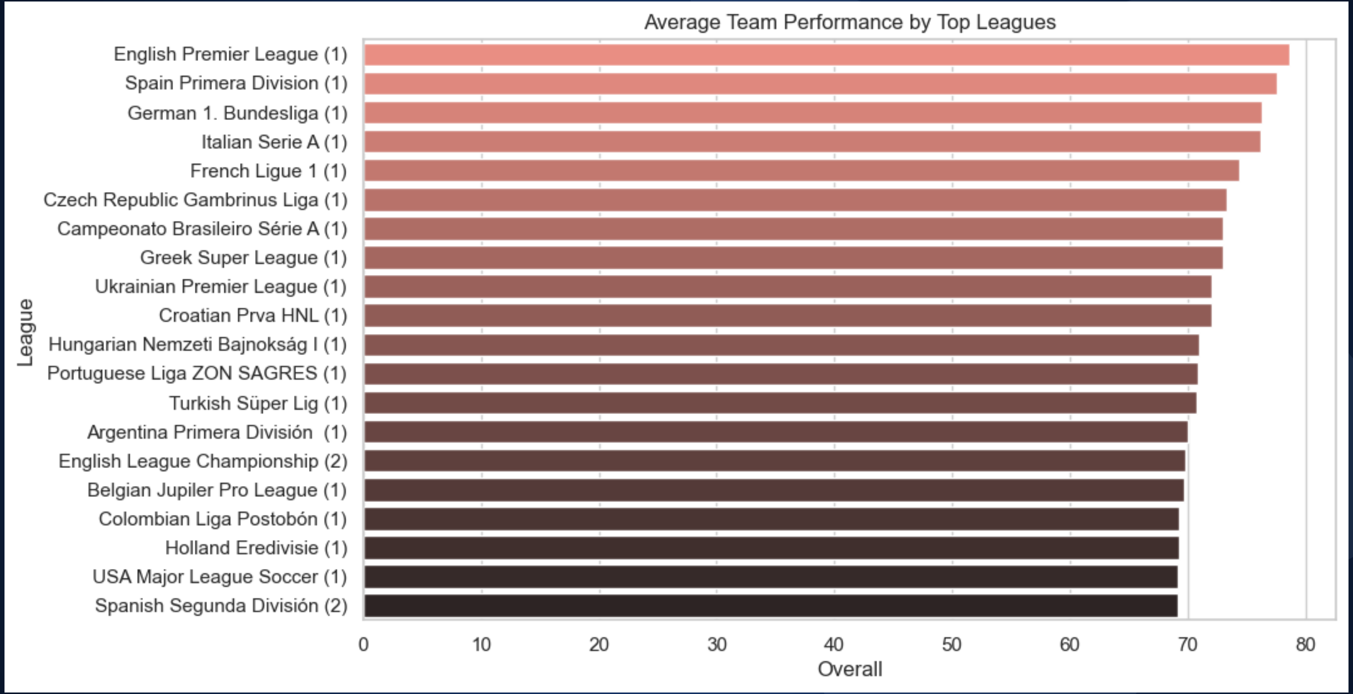
The histograms depicting the distribution of team ratings, including Overall, Attack, Midfield, and Defence, reveal a pattern consistent with a normal distribution.

EXPLORATORY DATA ANALYSIS for TEAMS

- **Key Insights:**

1. **Normal Distribution Pattern:** Teams exhibit a normal distribution, indicating most fall within an average rating range, a common statistical trend.
2. **Central Cluster (65-70 Points):** A concentration of teams in the 65-70 point range suggests a majority have moderate ratings, forming the peak of the curve.
3. **Frequency Decreases with Deviation:** Teams deviating from the central range show a decline in frequency, emphasizing that extremely high or low-rated teams are less common.
4. **Interpretation of Ratings:** Central range teams are likely balanced, high ratings signify strengths, and low ratings indicate potential areas for improvement.
5. **Competitiveness Implications:** The normal distribution suggests a competitive mid-range for most teams, fostering a balanced sports landscape.
6. **Strategic Considerations:** Teams can leverage this distribution insight for tactical decisions, recruitment, and overall team development.

EXPLORATORY DATA ANALYSIS for TEAMS



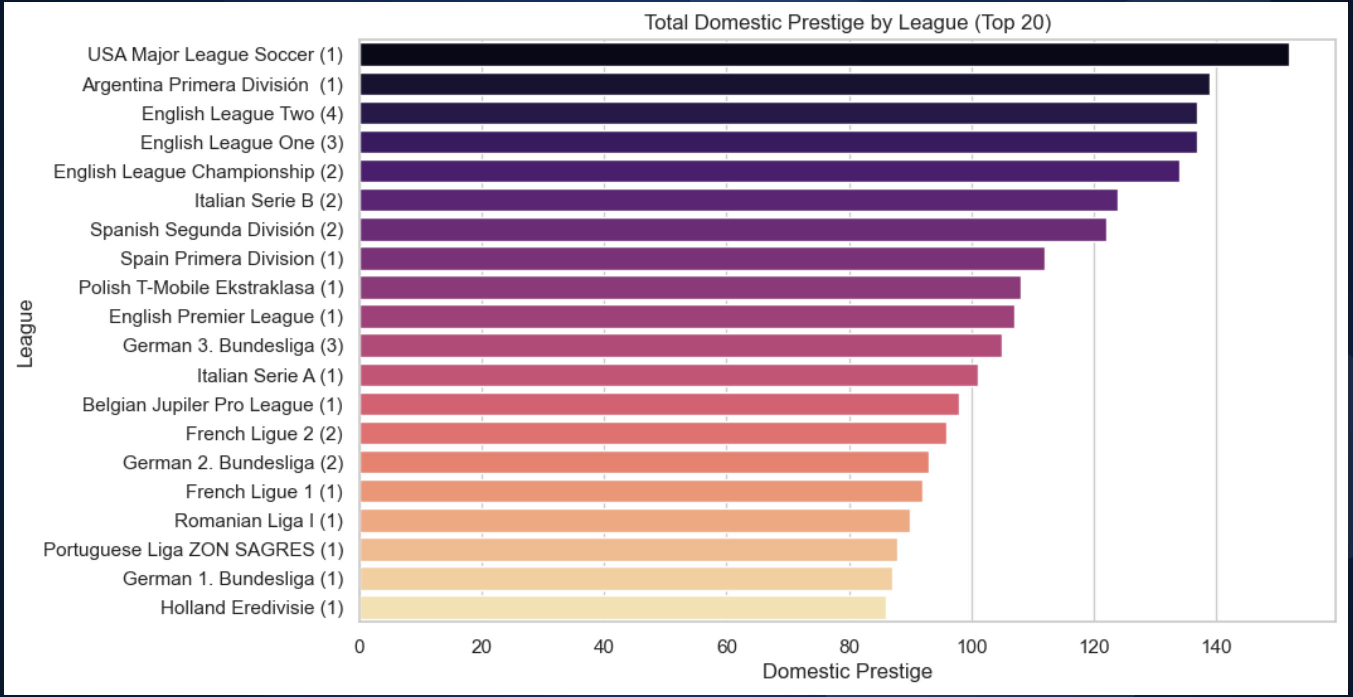
The bar chart provides a clear visual depiction of average team performance across various top leagues, allowing for a comparative analysis of their overall ratings.

EXPLORATORY DATA ANALYSIS for TEAMS

- **Key Observations from Bar Plot on Top Football Leagues:**

1. **Clear Comparative Analysis:** The bar plot enables a quick and intuitive comparison of team performance across different top football leagues, particularly in Overall ratings.
2. **Top Five Leagues Identified:** The English Premier League, Spain Primera Division, German 1. Bundesliga, Italian Serie A, and French Ligue 1 stand out as the top five leagues with the highest overall team performance.
3. **Implication of High Overall Performance:** Consistent high performance in these top five leagues suggests increased global attention, with higher overall performance contributing to competitive and entertaining football.
4. **Viewer Attraction and Global Appeal:** The attractiveness of these leagues aligns with the global appeal of football, indicating larger fan bases, increased viewership, and heightened interest worldwide.
5. **Factors Contributing to Attractiveness:** The appeal of these leagues may stem from factors such as team quality, competitive balance, star players, and match intensity, making them appealing for investments from stakeholders like broadcasters and sponsors.

EXPLORATORY DATA ANALYSIS for TEAMS



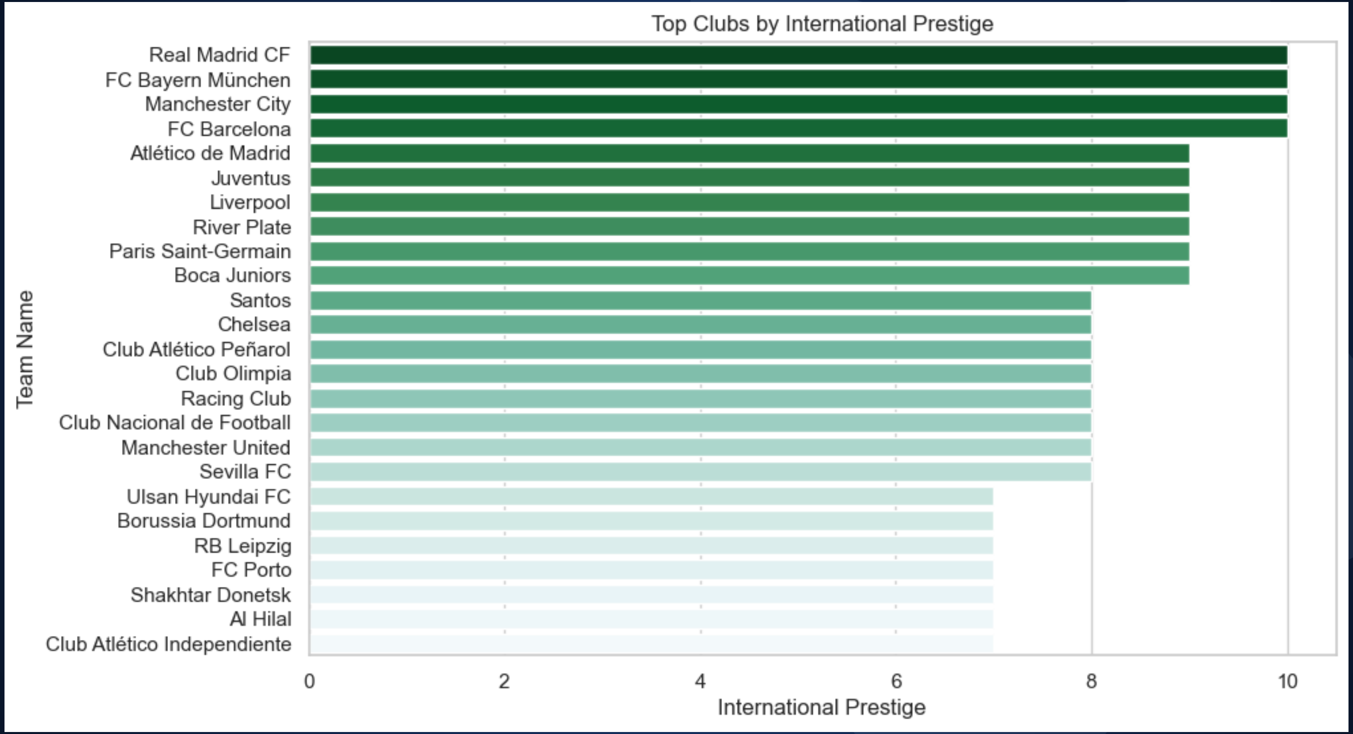
The bar chart provides a visual representation of the total DomesticPrestige of various leagues, enabling them to compare and identify the most competitive leagues based on the sum of DomesticPrestige metric.

EXPLORATORY DATA ANALYSIS for TEAMS

- **Key Insights from Cumulative Domestic Prestige Metric Chart:**

1. **Valuable Comparative Assessment:** The chart facilitates a comparative evaluation of league competitiveness based on their Domestic Prestige scores, offering insights to stakeholders.
2. **Surprising Leader - USA Major League Soccer:** The USA Major League Soccer achieves the highest Domestic Prestige score, surpassing 150. Factors contributing to this include the league's popularity, infrastructure investments, player development, and increasing global recognition.
3. **Leagues in 120-140 Range:** Following closely are leagues like Argentina Primera División, English League Two, English League One, English League Championship, Italian Serie B, and Spanish Segunda División, with Domestic Prestige scores between 120 and 140.
4. **Understanding Domestic Prestige Metric:** While Overall performance considers various on-field aspects, Domestic Prestige focuses on a league's reputation and recognition within its home country. This distinction explains variations in scores, where high Overall performers may not lead in Domestic Prestige.
5. **Global Recognition vs. Domestic Following:** Leagues like English Premier League, and Spain Primera Division may excel in Overall performance but might not lead in Domestic Prestige. Leagues like MLS and Argentina Primera División show high Domestic Prestige due to a significant domestic following and recognition.

EXPLORATORY DATA ANALYSIS for TEAMS



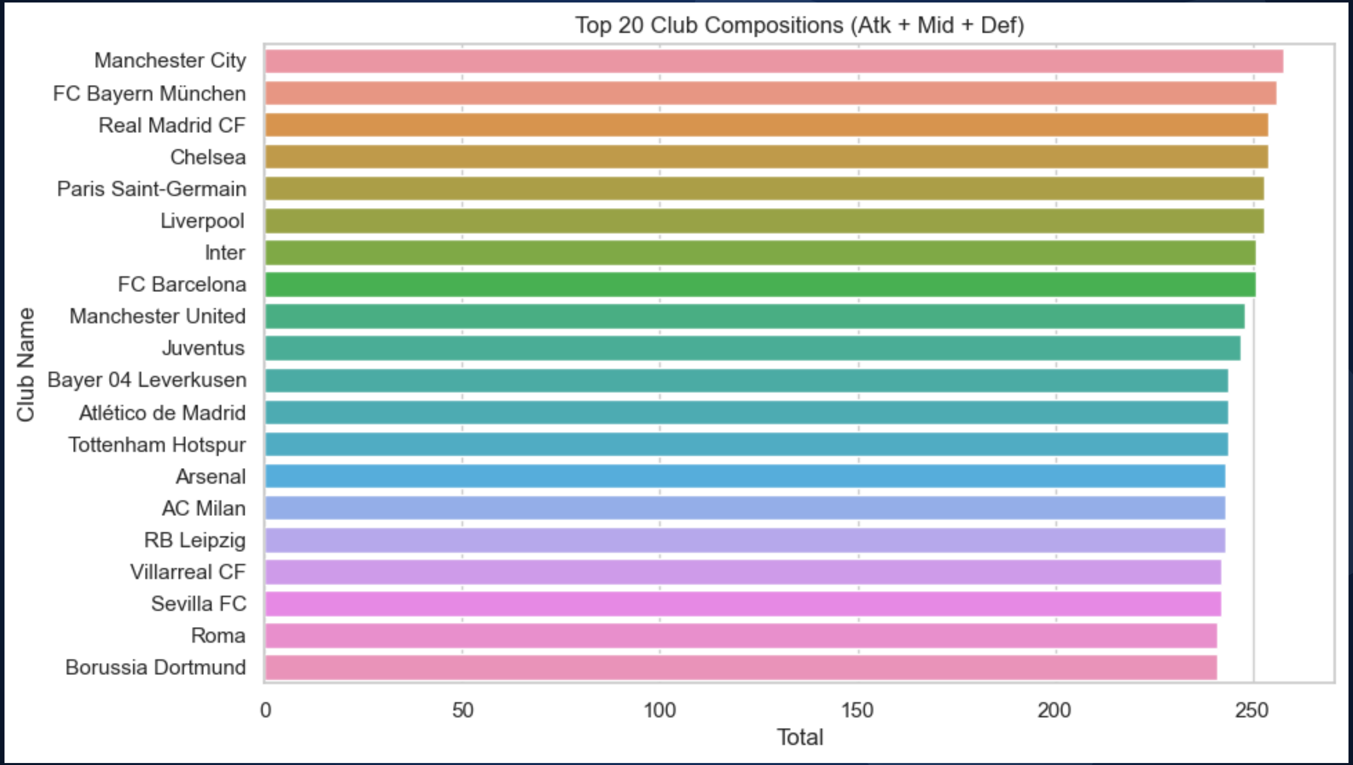
The bar chart represents the top clubs based on their International Prestige, enabling them to compare and identify the most successful teams in international competitions based on their IntPrestige metric.

EXPLORATORY DATA ANALYSIS for TEAMS

- **Key Insights from Top Clubs International Prestige Chart:**

1. **Valuable Resource for Stakeholders:** The chart provides stakeholders with a visual representation of top clubs based on International Prestige, allowing for direct comparison and identification of the most successful teams in international competitions.
2. **Top Clubs with Perfect Score (10):** Notably, Real Madrid CF, FC Barcelona, Atlético de Madrid (Spain Primera Division), FC Bayern München (German 1. Bundesliga), and Manchester City (English Premier League) achieve the highest International Prestige score of 10, indicating exceptional success on the global stage.
3. **Distinction Between Domestic and International Prestige:** The chart highlights an interesting distinction between Domestic Prestige, influenced by a league's reputation within its home country, and International Prestige, reflecting a club's global performance and recognition.
4. **Link Between Overall Performance and International Prestige:** The observed perfect scores for these clubs suggest a correlation between high Overall performance in their respective leagues (Spain Primera Division, German 1. Bundesliga, and English Premier League) and exceptional International Prestige.
5. **Consistency in Top European Leagues:** Clubs from top European leagues consistently perform well internationally, contributing to their impressive International Prestige scores.

EXPLORATORY DATA ANALYSIS for TEAMS



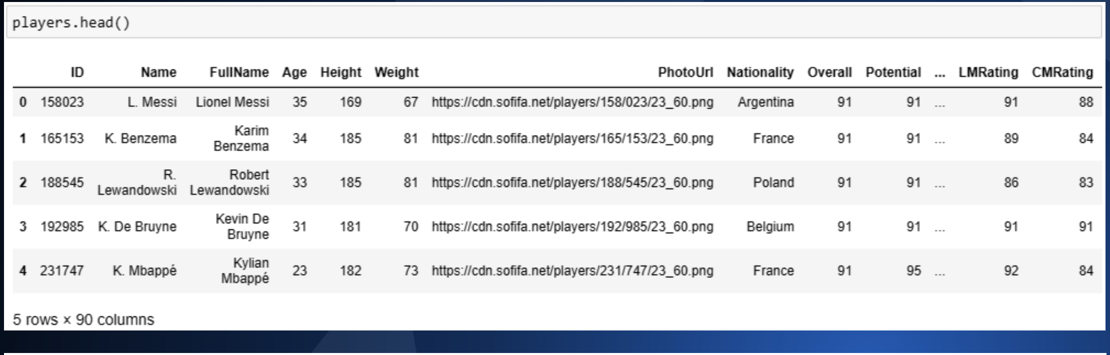
The bar chart shows the top 20 teams based on their composition of Attack, Midfield, and Defence, enabling them to compare and identify the teams with the strongest overall performance based on the Total metric.

EXPLORATORY DATA ANALYSIS for TEAMS

- **Key Insights from Top 20 Teams Composition Chart:**

1. **Valuable Comparative Analysis:** The chart visually presents the top 20 teams based on Attack, Midfield, and Defence composition, allowing stakeholders to identify teams with the strongest overall performance using the Total metric.
2. **Top 6 Clubs with Total Metric above 250:** Manchester City, FC Bayern München, Real Madrid CF, Chelsea, Paris Saint-Germain, and Liverpool emerge as top-performing clubs with Total metric scores surpassing 250, showcasing their excellence across attack, midfield, and defense.
3. **Consistency in Top Performance:** The presence and dominance of these clubs in the Total metric highlight their consistent excellence in various facets of the game, distinguishing them as hallmark performers.
4. **Strong Leagues Connection:** These top clubs predominantly hail from leagues that have consistently demonstrated high overall performance in previous charts. This connection underscores the correlation between strong leagues and the production of formidable clubs, often engaged in competitive and international play.
5. **Global Recognition:** The global recognition and following of clubs like Manchester City, FC Bayern München, Real Madrid CF, Chelsea, Paris Saint-Germain, and Liverpool are evident through their strong presence in both domestic and international competitions. Their performances contribute to high Total metric scores, elevating their status and appeal to a global football audience.

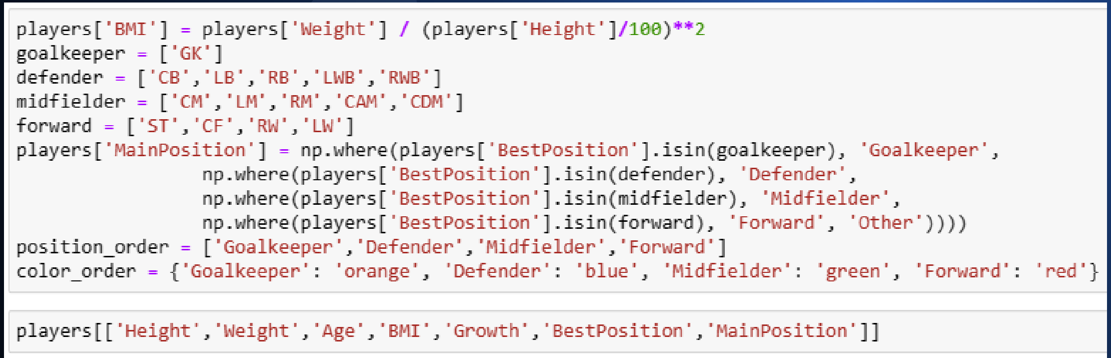
EXPLORATORY DATA ANALYSIS for PLAYERS



```
players.head()
```

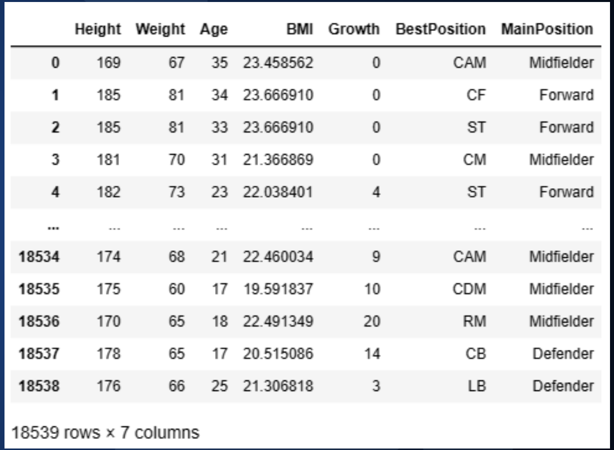
	ID	Name	FullName	Age	Height	Weight	PhotoUrl	Nationality	Overall	Potential	...	LMRating	CMRating
0	158023	L. Messi	Lionel Messi	35	169	67	https://cdn.sofifa.net/players/158/023/23_60.png	Argentina	91	91	...	91	88
1	165153	K. Benzema	Karim Benzema	34	185	81	https://cdn.sofifa.net/players/165/153/23_60.png	France	91	91	...	89	84
2	188545	R. Lewandowski	Robert Lewandowski	33	185	81	https://cdn.sofifa.net/players/188/545/23_60.png	Poland	91	91	...	86	83
3	192985	K. De Bruyne	Kevin De Bruyne	31	181	70	https://cdn.sofifa.net/players/192/985/23_60.png	Belgium	91	91	...	91	91
4	231747	K. Mbappé	Kylian Mbappé	23	182	73	https://cdn.sofifa.net/players/231/747/23_60.png	France	91	95	...	92	84

5 rows × 90 columns



```
players['BMI'] = players['Weight'] / (players['Height']/100)**2
goalkeeper = ['GK']
defender = ['CB','LB','RB','LWB','RWB']
midfielder = ['CM','LM','RM','CAM','CDM']
forward = ['ST','CF','RW','LW']
players['MainPosition'] = np.where(players['BestPosition'].isin(goalkeeper), 'Goalkeeper',
                 np.where(players['BestPosition'].isin(defender), 'Defender',
                 np.where(players['BestPosition'].isin(midfielder), 'Midfielder',
                 np.where(players['BestPosition'].isin(forward), 'Forward', 'Other'))))
position_order = ['Goalkeeper','Defender','Midfielder','Forward']
color_order = {'Goalkeeper': 'orange', 'Defender': 'blue', 'Midfielder': 'green', 'Forward': 'red'}
```

```
players[['Height','Weight','Age','BMI','Growth','BestPosition','MainPosition']]
```



	Height	Weight	Age	BMI	Growth	BestPosition	MainPosition
0	169	67	35	23.458562	0	CAM	Midfielder
1	185	81	34	23.666910	0	CF	Forward
2	185	81	33	23.666910	0	ST	Forward
3	181	70	31	21.366869	0	CM	Midfielder
4	182	73	23	22.038401	4	ST	Forward
...	...	...	...	...	...	...	...
18534	174	68	21	22.460034	9	CAM	Midfielder
18535	175	60	17	19.591837	10	CDM	Midfielder
18536	170	65	18	22.491349	20	RM	Midfielder
18537	178	65	17	20.515086	14	CB	Defender
18538	176	66	25	21.306818	3	LB	Defender

18539 rows × 7 columns

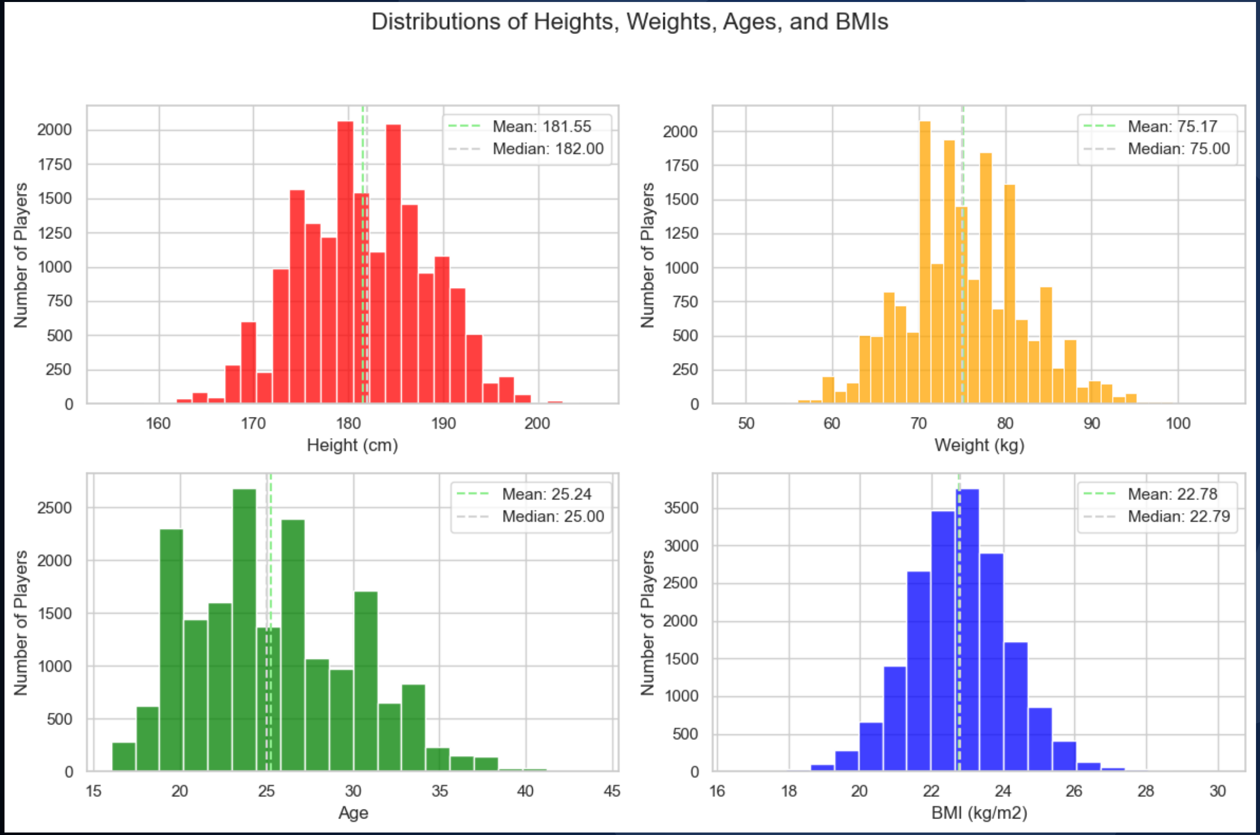
EXPLORATORY DATA ANALYSIS for PLAYERS

Using Python to manipulate the player dataset:

- First add a new column to the dataset called 'BMI', which calculates the Body Mass Index of each player based on their weight in kilograms and their height in meters.
- Next, categorize each player's "BestPosition" into one of four main positions: "Goalkeeper", "Defender", "Midfielder", or "Forward". The mapping of BestPosition to MainPosition is done using a series of nested numpy "where" statements, which assign each player's MainPosition based on their BestPosition.
- Then, creates two variables called 'position_order' and 'color_order'. The 'position_order' variable is a list that specifies the order in which the main positions should be displayed in the resulting dataframe. The 'color_order' variable is a dictionary that maps each main position to a color for use in visualizations.
- Finally, select a subset of columns from the players’ dataset and displays them, including 'Height', 'Weight', 'Age', 'BMI', 'Growth', 'BestPosition', and 'MainPosition'. This displays a dataframe that includes the new 'BMI' and 'MainPosition' columns that were added earlier.

► Overall, this appears to be preparing the players’ dataset for analysis and visualization, specifically by creating a new column for BMI, categorizing players by their main position, and selecting a subset of columns for display.

EXPLORATORY DATA ANALYSIS for PLAYERS



The histograms presented in the plot offer insights into the distribution of Height, Weight, Age, and BMI, shedding light on the physical characteristics of football players and providing a basis for comparison with the general population.

EXPLORATORY DATA ANALYSIS for PLAYERS

- **Height Histogram:** The Height histogram indicates a distribution that is slightly skewed to the left, with the mean (181.55 cm) being slightly lower than the median (182 cm). This suggests that the majority of football players in the dataset have above-average heights. The small difference between the mean and median suggests a relatively symmetric distribution.
- **Weight Histogram:** The Weight histogram, with a mean (75.17 kg) close to the median (75 kg), shows a distribution that is relatively symmetric and resembles a normal distribution. Football players in the dataset appear to have a moderate range of weights without significant skewness.
- **Age Histogram:** The Age histogram presents a distribution with a mean (25.24 years) close to the median (25 years), indicating a fairly symmetric distribution. The age of football players in the dataset seems to be concentrated within a relatively narrow range, possibly reflecting the typical age range of active football players.
- **BMI Histogram:** The BMI histogram is described as being most similar to a normal distribution, which is supported by the mean (22.78) and median (22.79) being very close. This suggests that BMI values for football players exhibit a typical and balanced distribution, which is often considered a sign of good physical condition.

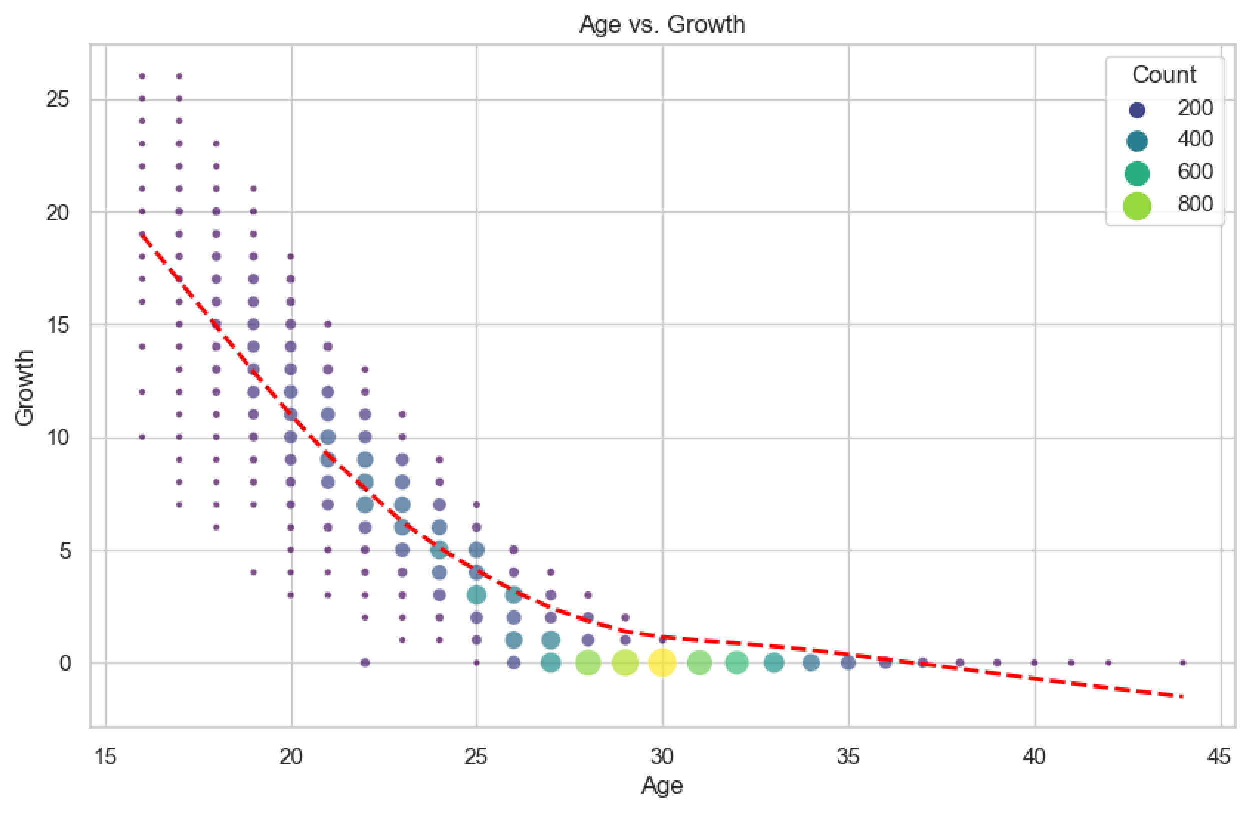
EXPLORATORY DATA ANALYSIS for PLAYERS

From these observations, we can get the following insights:

- Football players in the dataset tend to have above-average heights, which is consistent with the physical requirements of the sport.
- The weight distribution is relatively symmetric, indicating that football players come in a range of body sizes.
- The age distribution is centered around the typical age range for active football players, which is usually between late teens and early thirties.
- The BMI distribution resembling a normal distribution suggests that football players, on average, maintain a healthy weight relative to their height.

► Overall, these insights provide a glimpse into the physical characteristics of football players, highlighting the requirements of the sport and the importance of maintaining a balanced physical condition. In contrast to the general population, football players typically exhibit distinct characteristics owing to their specific physical demands and training regimens.

EXPLORATORY DATA ANALYSIS for PLAYERS



The scatter plot effectively illustrates the relationship between Age and Growth, with the size of the dots representing the frequency of data points and the trendline helping estimate the general direction of this relationship.

EXPLORATORY DATA ANALYSIS for PLAYERS

Several key insights can be drawn from this visualization:

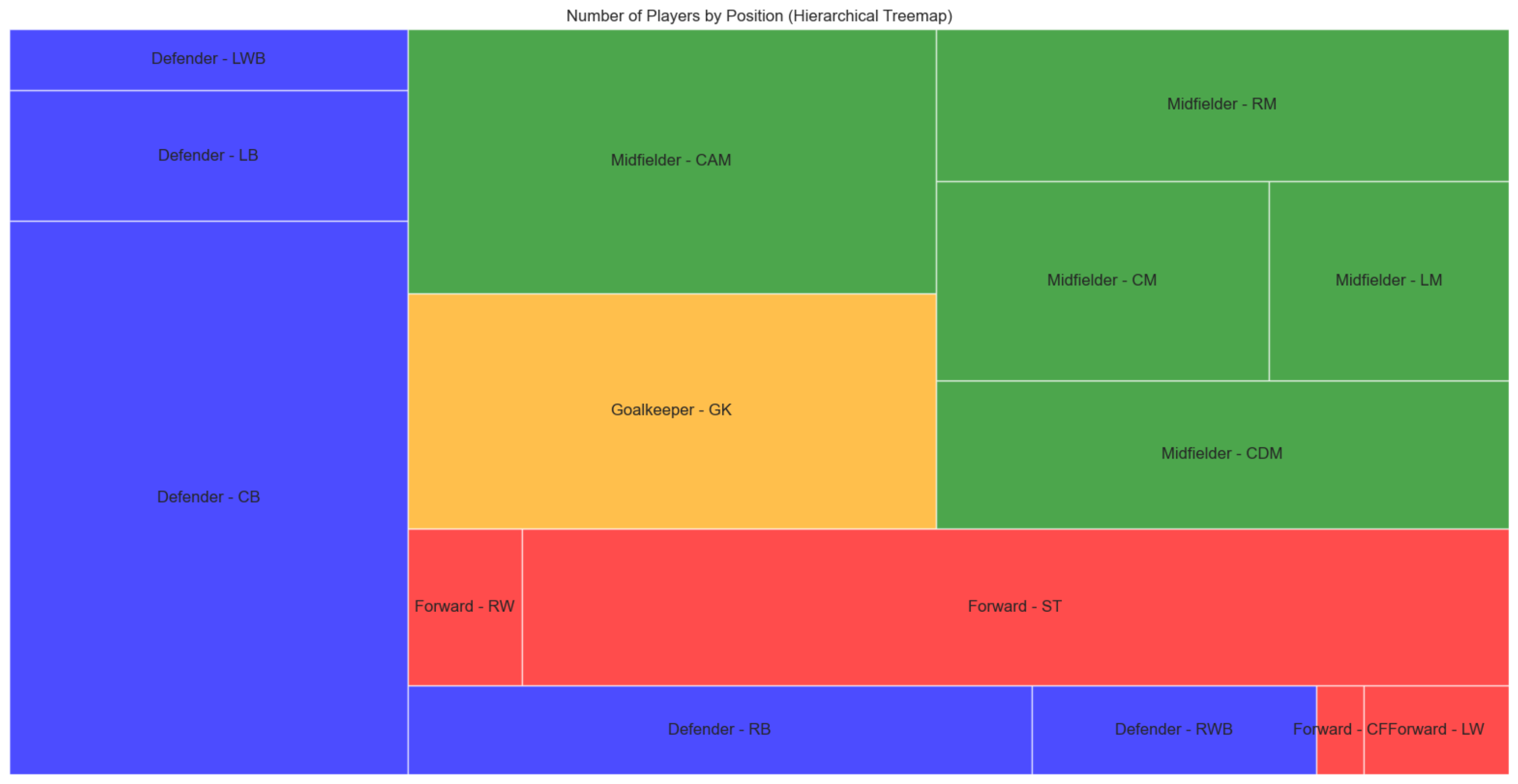
1. Age vs. Growth Trend: The plot reveals a negative correlation between Age and Growth. In other words, as players get older, their potential for growth decreases. This observation aligns with common knowledge in sports, where younger individuals often have more significant potential for improvement and development compared to older players.

2. Age Concentration: The most significant concentration of larger dots is observed in the age range of approximately 28 to 32. The largest dot at the age of 30 with a Growth of 0 suggests that around this age, players typically reach their peak performance and are less likely to experience growth in their abilities. This age range is often considered the prime of a footballer's career.

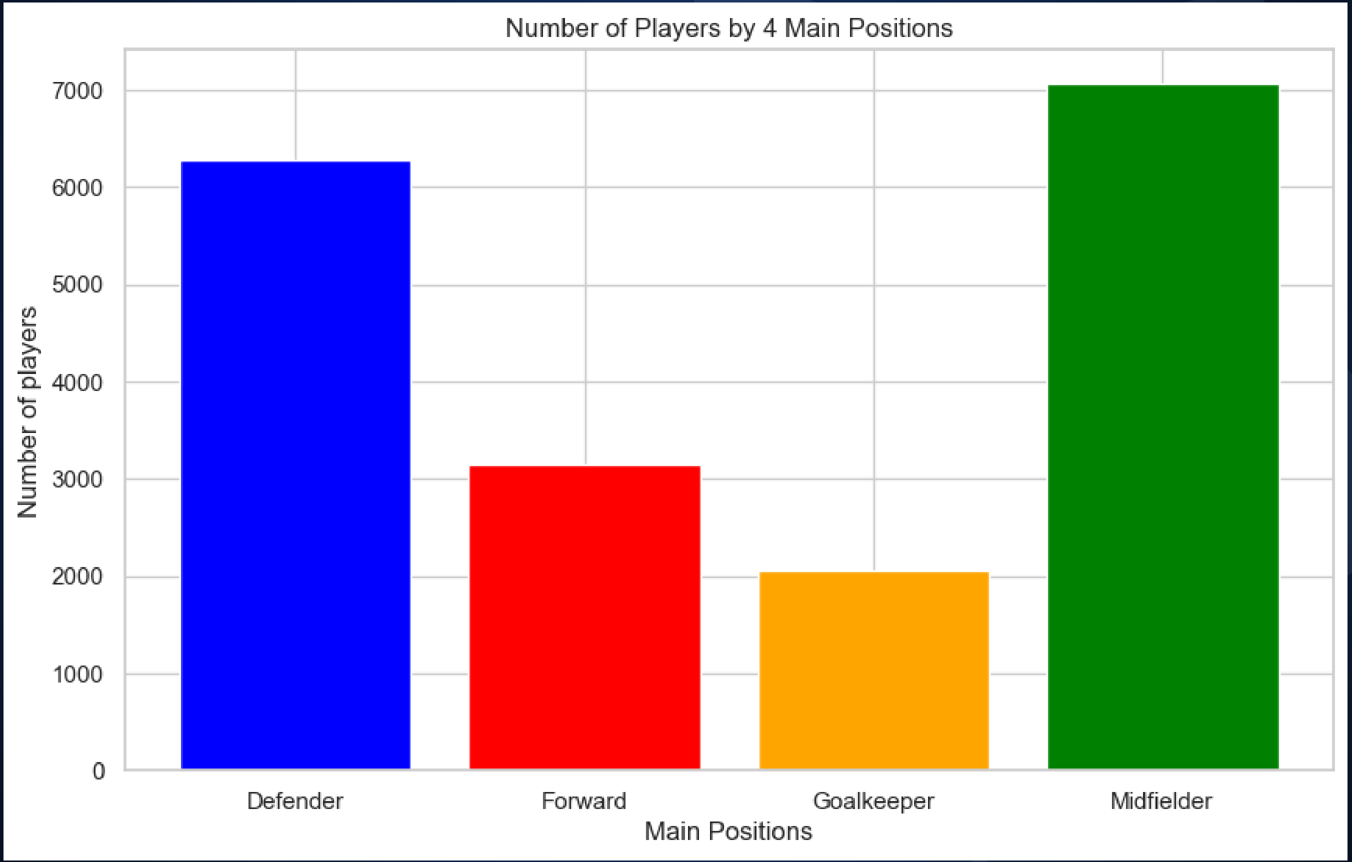
3. Ceiling After Age 30: An intriguing observation is the absence of Growth after the age of 30. This further reinforces the notion that, in the world of professional football, players tend to reach a performance plateau after entering their thirties. Their physical abilities may start to decline, and they may focus more on maintaining their current skill levels rather than experiencing growth.

► This scatter plot and its analysis provide valuable insights into the relationship between age and growth in the context of football players. It highlights the general trend of declining growth as players get older, with the most significant concentration of growth occurring during the late twenties and early thirties before reaching a plateau in performance after the age of 30. These findings align with the typical trajectory of a footballer's career and underscore the importance of youth development in the sport.

EXPLORATORY DATA ANALYSIS for PLAYERS



EXPLORATORY DATA ANALYSIS for PLAYERS



EXPLORATORY DATA ANALYSIS for PLAYERS

The two visualizations, the treemap chart, and the bar chart, provide valuable insights into the distribution of players by their main positions and their best positions. These observations can be related to modern football tactics and the evolution of the game.

1. Distribution of Players by Main Position:

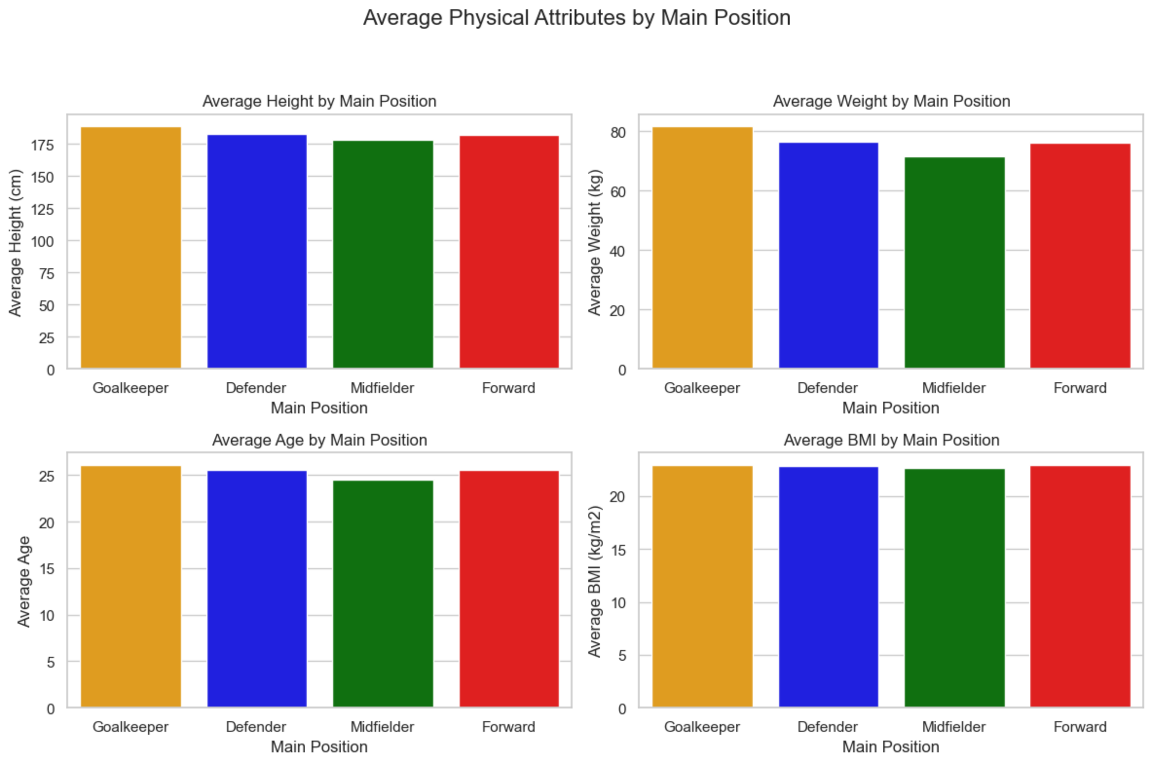
- **Midfielder (Over 7000 players):** The high number of players in the midfielder position reflects the increasing emphasis on midfield dominance in modern football. Midfielders play a crucial role in controlling the game, transitioning between defense and attack, and setting the tempo of play. Different subcategories (CAM, RM, CM, LM, CDM) within the midfield position allow for more specialized roles, aligning with the tactical diversity seen in modern football.
- **Defender (Over 6000 players):** Within the defender category, the majority of players are center-backs (CB). This reflects the importance of strong and reliable central defenders in modern football. Many coaches prefer to build their defense around a solid central pairing, which may explain the prevalence of CBs.
- **Forward (Over 3000 players):** The forward position includes various subcategories, but most players are categorized as strikers (ST). Strikers are responsible for scoring goals, and they remain a focal point in many teams' offensive strategies. The prevalence of STs reflects the ongoing emphasis on goal-scoring in the game.
- **Goalkeeper (Around 2000 players):** The goalkeeper position is unique because there's typically only one goalkeeper in each starting lineup. This naturally limits the number of players in this position, and it's also a specialized role that requires unique skills. Goalkeepers play a pivotal role in protecting the goal area, which is essential for a team's defensive strategy.

EXPLORATORY DATA ANALYSIS for PLAYERS

2. Modern Football Tactics and Player Roles:

- The prevalence of various midfield subcategories (CAM, RM, CM, LM, CDM) and center-backs (CB) highlights the tactical diversity in modern football. Teams often use multiple midfielders with distinct roles to control possession, press opponents, and create scoring opportunities. Center-backs are critical for defensive stability and building from the back.
- The focus on goal-scoring is evident with the predominance of strikers (ST). Coaches continue to prioritize players who can convert goal-scoring opportunities, and goal-scorers often receive significant attention in team tactics.
- The emphasis on technical and tactical aspects of the game is a shift from the past, where football might have been more focused on individual brilliance. Modern football values a collective approach, with players taking on specific roles and responsibilities within the team structure.
- In summary, these visualizations reflect the evolving tactics in modern football, with an emphasis on teamwork, specialized roles, and strategic diversity. Coaches and teams are adapting to these changes to maximize their performance and achieve success in a highly competitive environment.

EXPLORATORY DATA ANALYSIS for PLAYERS



The bar chart subplots depict the average values of four physical attributes (height, weight, age, and BMI) for soccer players in each of the four main positions (goalkeeper, defender, midfielder, and forward)

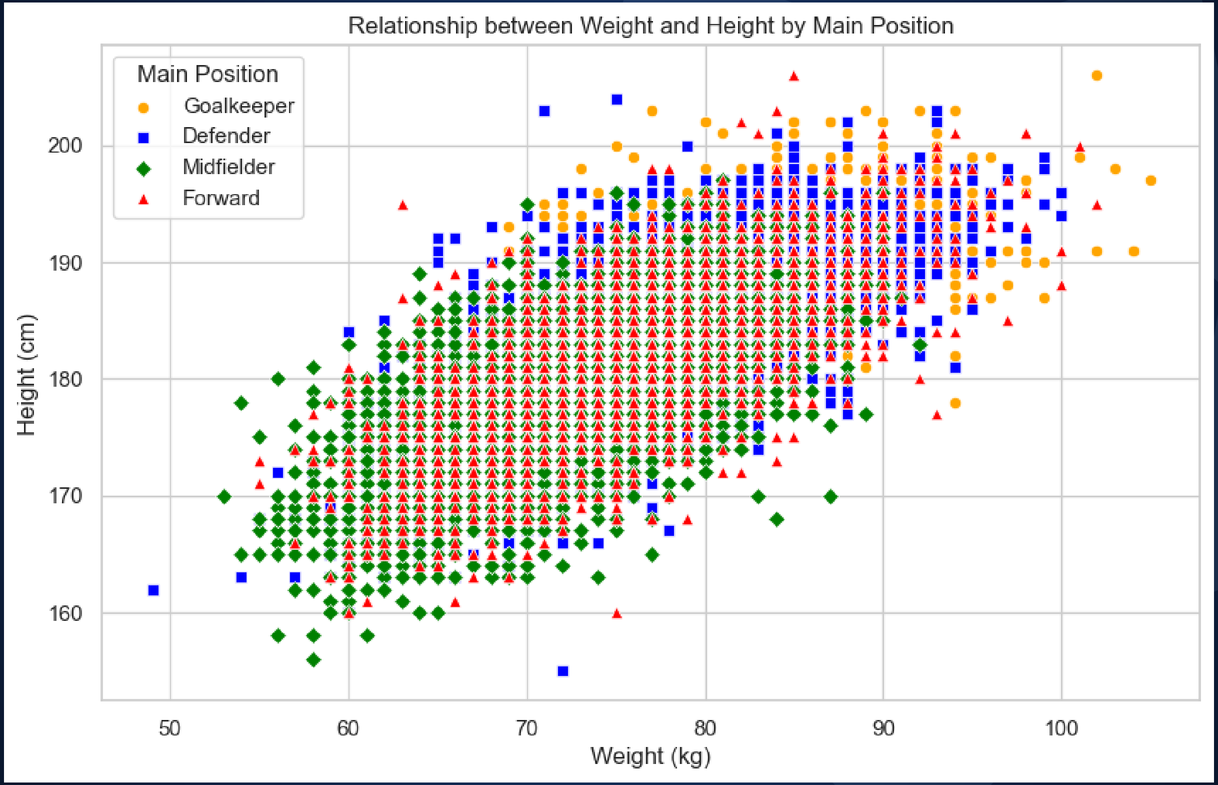
EXPLORATORY DATA ANALYSIS for PLAYERS

Several interesting insights:

1. **Midfielders:** Across all four physical attributes (height, weight, age, and BMI), midfielders consistently show the lowest average values. This suggests that midfielders, as a group, tend to be shorter, lighter, younger, and have a lower BMI compared to players in the other positions. This might be indicative of the role of midfielders, which often requires agility, speed, and ball control rather than physical stature.
2. **Goalkeepers:** On the contrary, goalkeepers consistently display the highest average values for all physical attributes. They tend to be taller, heavier, older, and have a higher BMI compared to players in other positions. This aligns with the requirements of the goalkeeper position, which necessitates a strong physical presence, reach, and experience.
3. **Defenders and Forwards:** The physical attributes of defenders and forwards are relatively similar. This indicates that, on average, players in these positions have physical characteristics that fall between those of midfielders and goalkeepers. Defenders and forwards may require a balance of physical attributes, including height, weight, age, and BMI, to fulfill their roles effectively.

► These insights highlight the varying physical demands of different soccer positions. Goalkeepers are expected to excel in physical attributes like height and weight, while midfielders prioritize agility and speed over stature. Defenders and forwards, positioned between these two extremes, often require a well-rounded combination of physical attributes to succeed. Coaches and teams use these insights to tailor their recruitment and training strategies, aligning them with the specific physical and tactical requirements of each position on the field.

EXPLORATORY DATA ANALYSIS for PLAYERS



The scatter plot revealing the relationship between Weight and Height by MainPosition

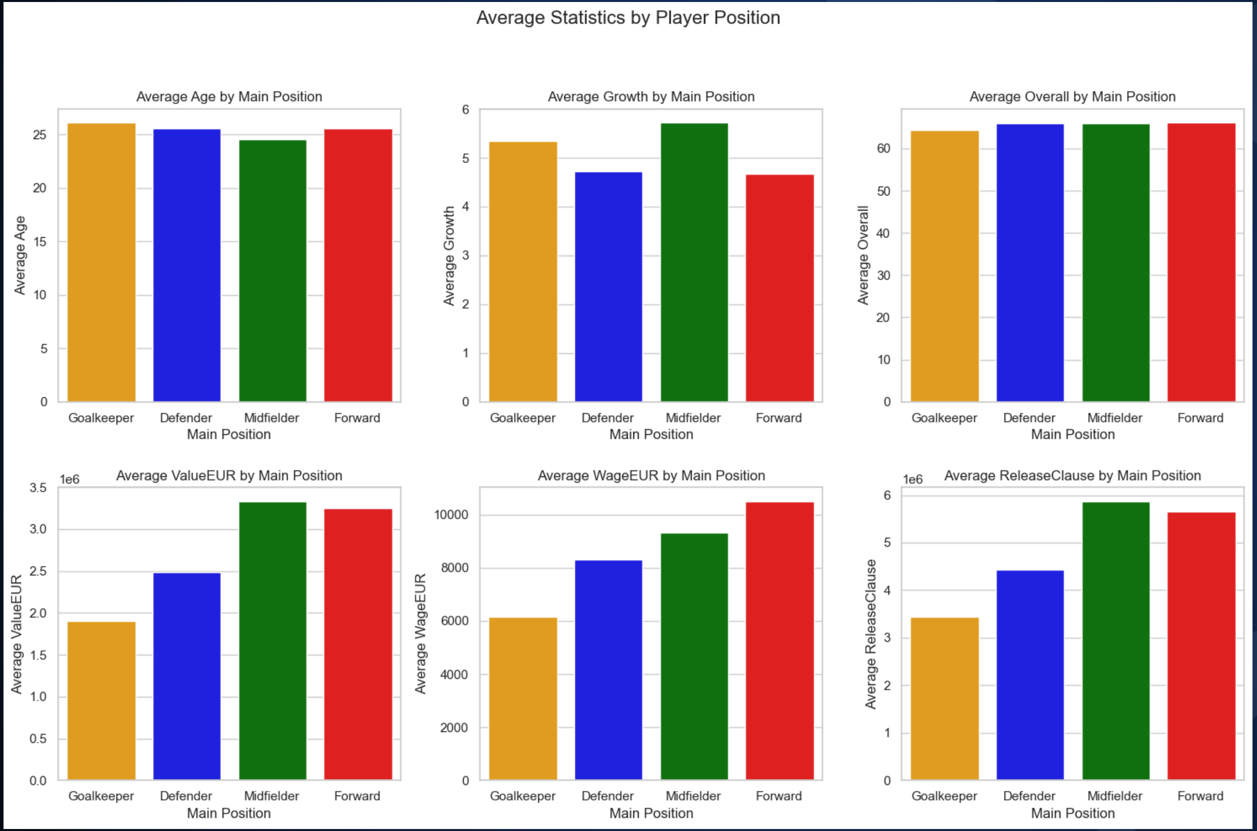
EXPLORATORY DATA ANALYSIS for PLAYERS

Several noteworthy insights:

1. **Weight and Height Concentration:** The majority of players in the dataset have weights ranging from 70 kg to 80 kg and heights between 175 cm to 185 cm. This concentration is likely indicative of the preferred range of physical attributes for soccer players, which allows for a balance between speed, agility, and strength.
2. **Midfielder Characteristics:** Midfielders stand out in this visualization as they tend to be shorter and lighter compared to players in other positions. The dense concentration of dots in the lower values for both weight and height suggests that midfielders are typically more compact in stature. This can be attributed to the role of midfielders, which often emphasizes quick ball control and agility.
3. **Other Positions:** In contrast, players in positions other than midfielders appear to be higher in both weight and height. This aligns with the physical demands of positions such as defenders, who need to win aerial duels and clearances, and forwards, who often play a more physical role in challenging defenders and goalkeepers.
4. **Physical Limits for Midfielders:** The absence of players in the Midfielder position above 200 cm in height and with a weight exceeding 100 kg highlights the practical physical limits for midfielders. These limits are likely due to the specific demands of the position, which require agility, ball control, and endurance. Extremely tall or heavy players may find it challenging to excel in midfield roles.

► These insights stress the significance of varied physical attributes across soccer positions. Coaches and scouts prioritize these traits for optimal player recruitment and lineup decisions. A balance in height, weight, speed, and skill is crucial for player success in their designated positions.

EXPLORATORY DATA ANALYSIS for PLAYERS



The chart showcasing average statistics by player position offers several notable insights into the distinctions between different positions in football.

EXPLORATORY DATA ANALYSIS for PLAYERS

1. Age:

Midfielders have the lowest average age, below 25. This suggests that midfielders tend to be younger compared to other positions. They often need to possess agility, stamina, and the ability to cover large areas of the field, which can be associated with a younger age group. In contrast, players in other positions tend to be slightly older.

2. Growth:

Goalkeepers and Midfielders have an average growth above 5, with Midfielders having the highest value. This implies that goalkeepers and midfielders are more likely to experience continued growth and improvement in their skills and performance. Forward and Defender players have growth below 5, suggesting they may reach their peak performance earlier in their careers.

3. Overall:

The average overall rating for all positions is around 65 points on a scale of 100. This suggests a relatively balanced distribution of player quality across different positions.

4. ValueEUR:

Midfielders and Forward players have the highest average values, above 3 million Euros. This can be attributed to the fact that these positions are typically more involved in goal-scoring opportunities and playmaking, making them more valuable in the transfer market. Defenders have an average value of around 2.5 million Euros, while Goalkeepers have the lowest average value, below 2 million Euros, likely due to the limited transfer market value of goalkeepers compared to outfield players.

EXPLORATORY DATA ANALYSIS for PLAYERS

3. WageEUR:

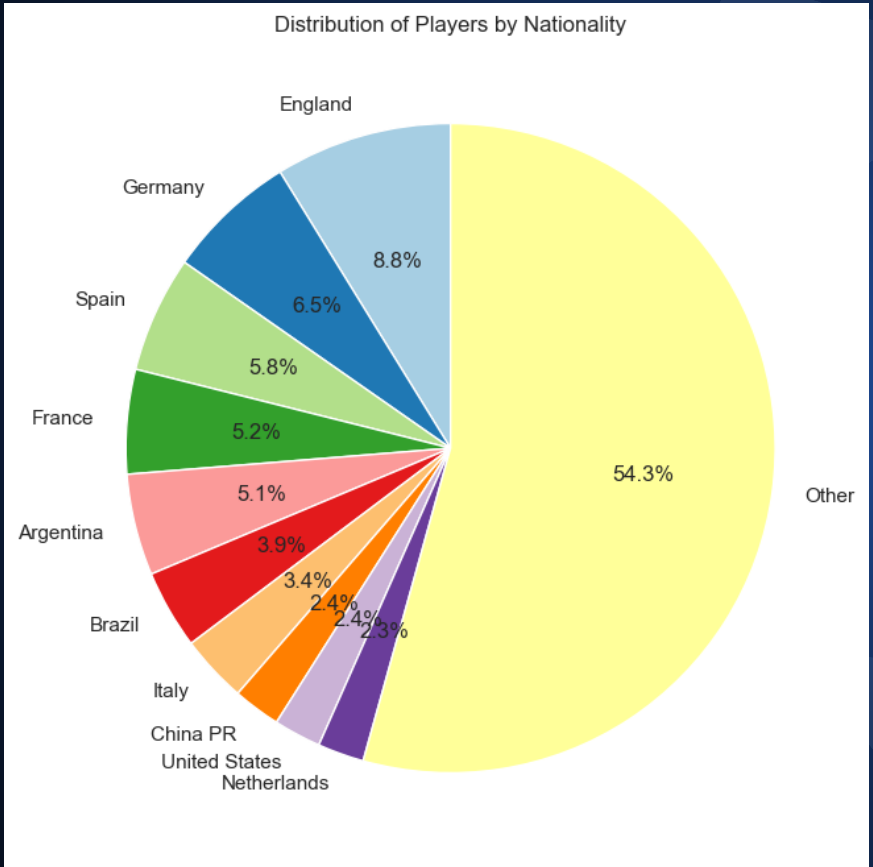
Forward players have the highest average wages, exceeding 10,000 Euros, possibly due to their goal-scoring responsibilities and marketability. Midfielders are next with around 9,000 Euros, followed by Defenders with 8,000 Euros. Goalkeepers have the lowest average wages, around 6,000 Euros, reflecting the traditionally lower earning potential of goalkeepers compared to outfield players.

4. ReleaseClause:

Midfielders have the highest average release clauses, nearly 6 million Euros, indicating their perceived value and the cost of acquiring them. Forward players follow with around 5.5 million Euros, while Defenders have an average release clause of around 4.5 million Euros. Goalkeepers have the lowest average release clauses, around 3.5 million Euros, reinforcing the trend of goalkeepers being less expensive in the transfer market.

► These insights reflect the unique roles and market values associated with each player position in football, with factors such as age, growth potential, overall performance, value, wage, and release clauses varying based on the specific demands of each position on the field.

EXPLORATORY DATA ANALYSIS for PLAYERS



The pie chart illustrating the distribution of players by nationality

EXPLORATORY DATA ANALYSIS for PLAYERS

Several key insights:

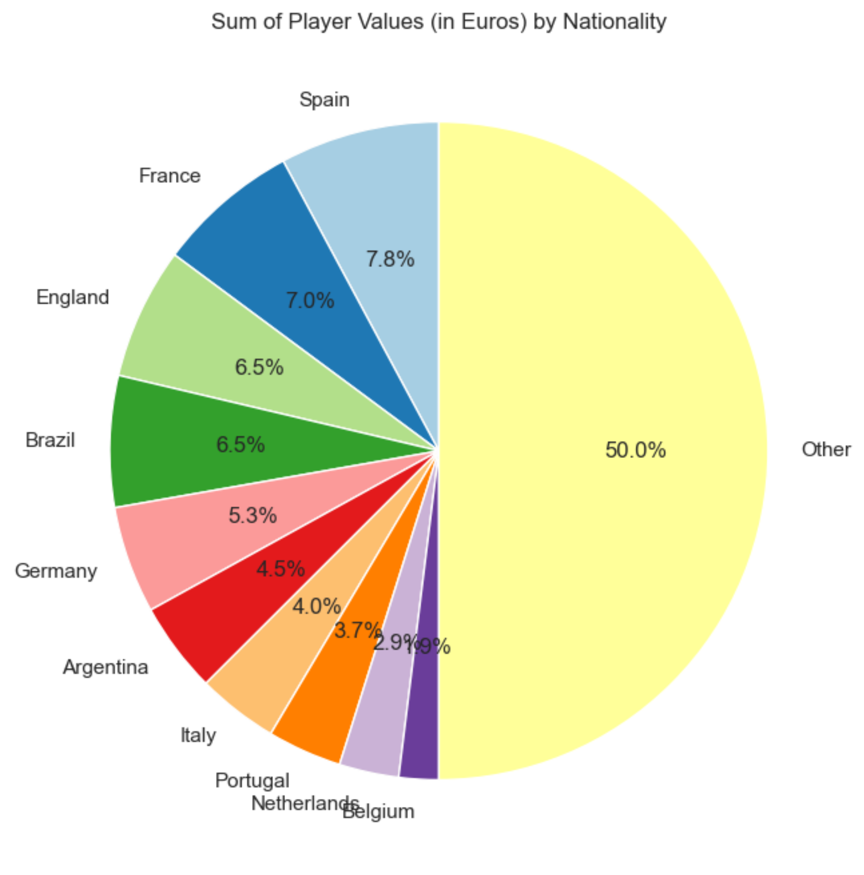
1. **European and South American Dominance:** The majority of players come from Europe and South America. England, Germany, Spain, France, Italy, the Netherlands, and other European countries collectively account for a substantial portion of the player population. Similarly, Argentina and Brazil, two prominent South American nations, also contribute significantly to the player base. This reflects the strong footballing traditions and talent pools in these regions.
2. **Global Appeal of Football:** The distribution of players from countries like China PR, the United States, and others highlights the global appeal of football. These countries have been investing in football development, and this is reflected in the presence of players from diverse backgrounds. The growing popularity of the sport in countries like China PR and the United States has contributed to their representation in the player population.
3. **Concentration in Traditional Football Powerhouses:** The representation of players from football powerhouses like England, Germany, Spain, and Brazil is significant. These countries have strong footballing infrastructures, well-established leagues, and a rich history of success in international competitions. As a result, they consistently produce a large number of talented players.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Diverse and Inclusive Nature of Football:** The "Others" category, comprising 54.3% of players, underscores the inclusive and diverse nature of football. It indicates that players come from various countries and backgrounds, making football a truly global sport that transcends boundaries and cultures.
5. **Economic and Social Factors:** The distribution of players can also be influenced by economic and social factors. Countries with well-funded youth development programs and strong footballing traditions are more likely to produce a higher number of players. Additionally, factors such as accessibility to coaching and infrastructure, as well as the passion for the sport, play a role in determining player representation from different countries.

► The pie chart highlights the global nature of football and the contributions of various nations to the sport. It reflects the historical prominence of Europe and South America in producing football talent while also showcasing the sport's increasing popularity and development in regions like Asia and North America. Football's universal appeal continues to make it one of the most widely played and watched sports worldwide.

EXPLORATORY DATA ANALYSIS for PLAYERS



The pie chart representing the sum of player values (in Euros) by nationality

EXPLORATORY DATA ANALYSIS for PLAYERS

Several notable insights:

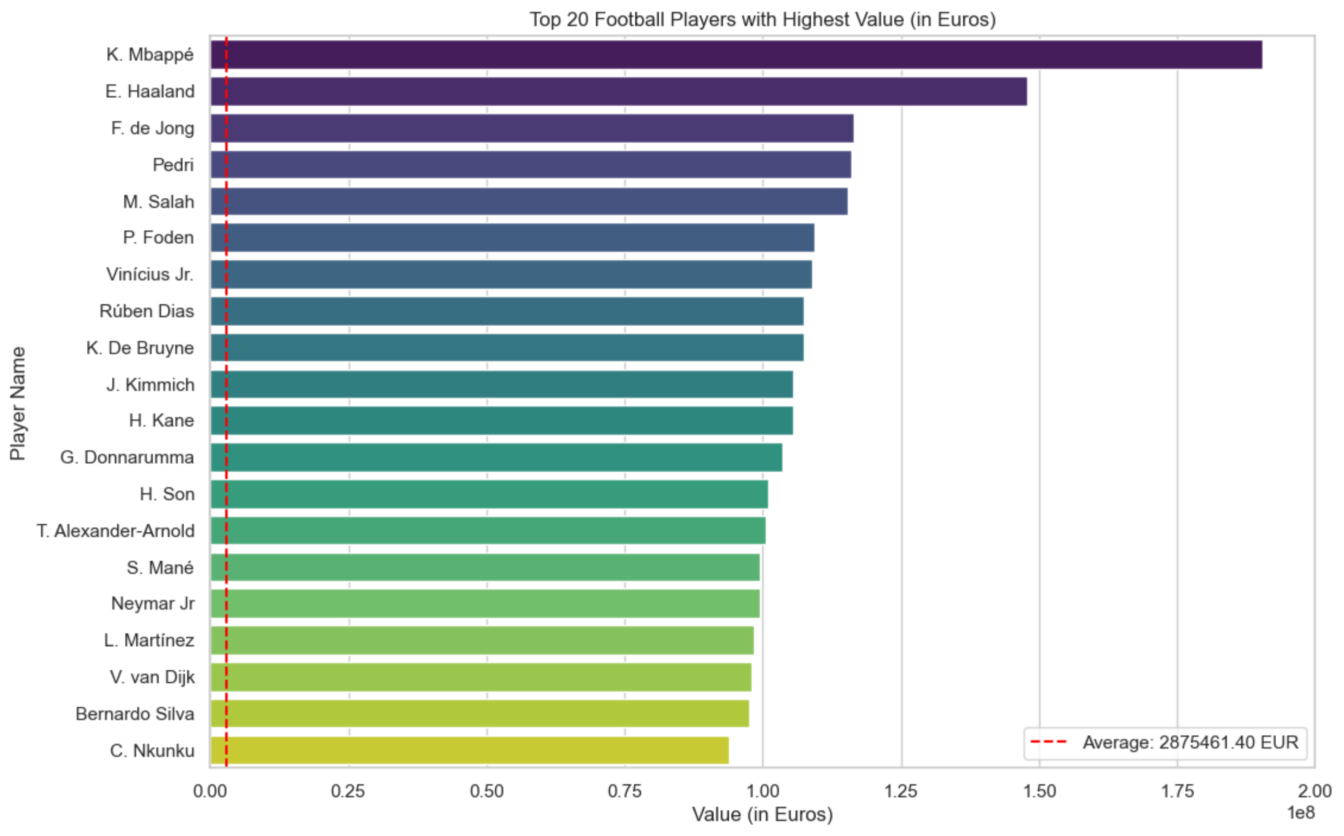
1. **Diversity of Nationalities:** The distribution of player values is diverse, with contributions from various nations. Spain, France, England, Brazil, Germany, Argentina, Italy, Portugal, the Netherlands, and Belgium all play significant roles in terms of player value.
2. **European and South American Influence:** European nations, such as Spain, France, England, Germany, Italy, Portugal, the Netherlands, and Belgium, continue to be prominent in the valuation of players. These countries have well-established footballing traditions and strong leagues, which contribute to the high total player values.
3. **South American Contribution:** Brazil and Argentina, two of the most well-known South American footballing nations, also feature prominently in player values. Their soccer academies have produced many highly-valued players who have made a significant impact on the global stage.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Economic Factors:** Economic factors play a significant role in player valuation. European leagues, such as those in Spain, England, France, and Germany, often have higher financial resources, which can lead to greater player valuations. The investment in football development and infrastructure in these countries also plays a role.
5. **Globalization of Football:** The presence of a "50.0%" category labeled as "Others" highlights the globalization of football. It demonstrates that players from various nations worldwide contribute to the total player values, reflecting the increasing reach and popularity of the sport on a global scale.

► This pie chart reveals the distribution of player values by nationality, emphasizing the influence of European and South American nations in the valuation of players. It underscores the importance of strong footballing traditions, well-funded leagues, and economic factors in determining player values while also reflecting the global and inclusive nature of the sport.

EXPLORATORY DATA ANALYSIS for PLAYERS



The bar chart displaying the top 20 football players with the highest value (in Euros)

EXPLORATORY DATA ANALYSIS for PLAYERS

Several insights into the footballing world:

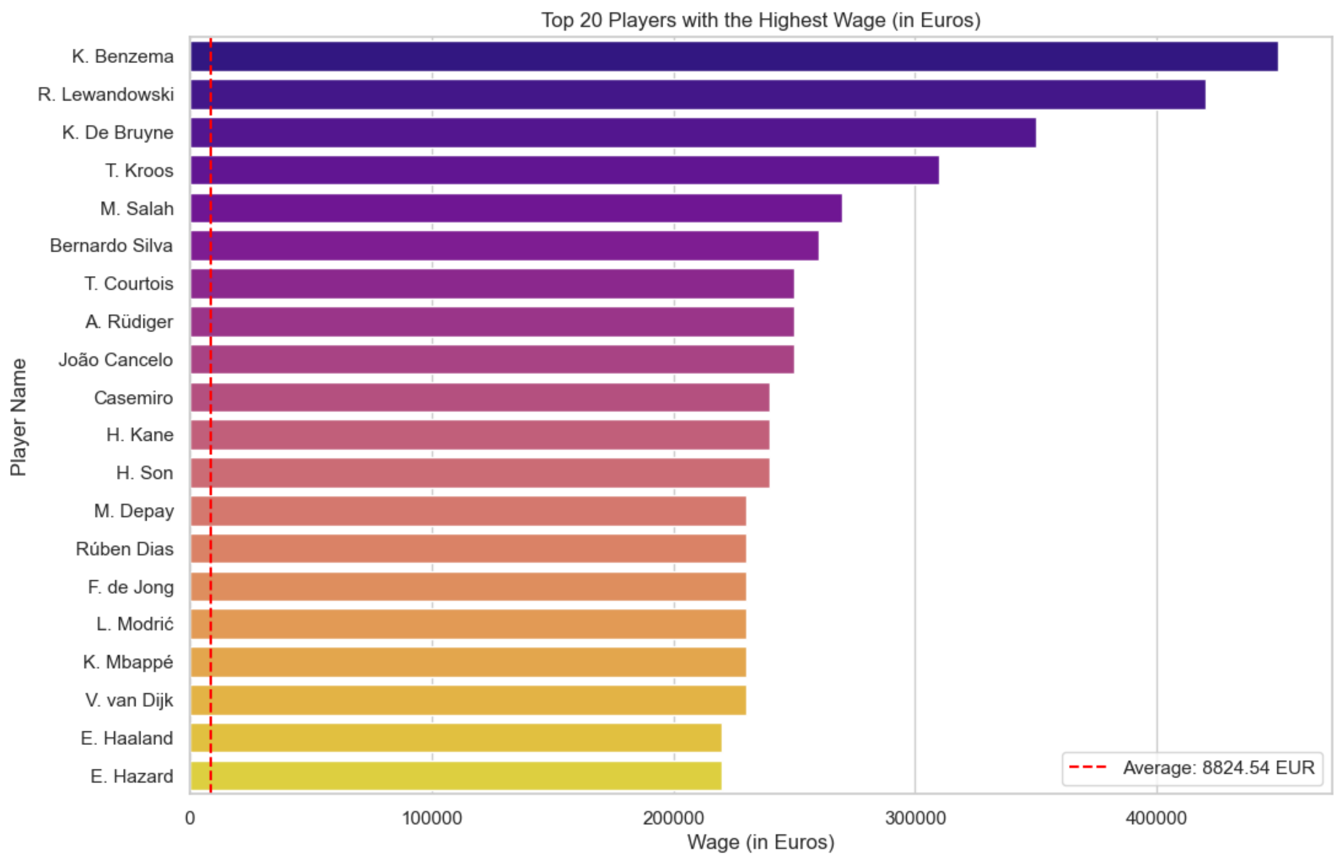
1. **K. Mbappé's Exceptional Value:** Kylian Mbappé, with a value of 190.5 million Euros, stands out as the most valuable player. His high valuation reflects his extraordinary talent and performances. Mbappé is known for his incredible speed, dribbling ability, and goal-scoring prowess. He began his professional career at AS Monaco, where he helped the team win Ligue 1. He subsequently joined Paris Saint-Germain (PSG) and has continued to impress on both the domestic and international stages. His performances for France in the FIFA World Cup have solidified his status as one of the world's top players.
2. **Emerging Talent Erling Haaland:** Erling Haaland's value of 148 million Euros places him as one of the most valuable young talents in the world. The Norwegian striker has attracted significant attention due to his exceptional goal-scoring ability and physical attributes. Haaland began his professional career at Molde FK and later joined RB Salzburg before moving to Borussia Dortmund and then Manchester City. His performances in the Bundesliga, the Premier League, and the UEFA Champions League have earned him recognition as one of the top prospects in world football.
3. **High Valuations for Elite Players:** The presence of several players with values around 100 million Euros indicates the robust market for top football talent. The top players have about 50 to 100 times more value than the average players (237,654,610 EUR). These players are typically established stars who have consistently performed at the highest levels for their respective clubs and national teams.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Market Dynamics and Transfers:** Player valuations are influenced by a variety of factors, including club performance, international achievements, and transfer fees. High-profile transfers, such as those involving Mbappé and Haaland, often lead to substantial increases in a player's value. These transfers can be influenced by a player's contract situation, market demand, and the financial resources of the buying club.
5. **Historical Significance:** The chart reflects the continual evolution of the football market, with player values increasing over time. In recent years, player valuations have reached unprecedented levels, and clubs are willing to invest substantial sums to secure the services of top talent.

► This chart showcases the high values associated with top football players, including established stars like Kylian Mbappé and emerging talents like Erling Haaland. The increasing financial investment in the sport and the competitive market for elite players have contributed to the high player valuations observed in the chart.

EXPLORATORY DATA ANALYSIS for PLAYERS



The bar graph displaying the top 20 players with the highest wages in Euros

EXPLORATORY DATA ANALYSIS for PLAYERS

Several notable insights:

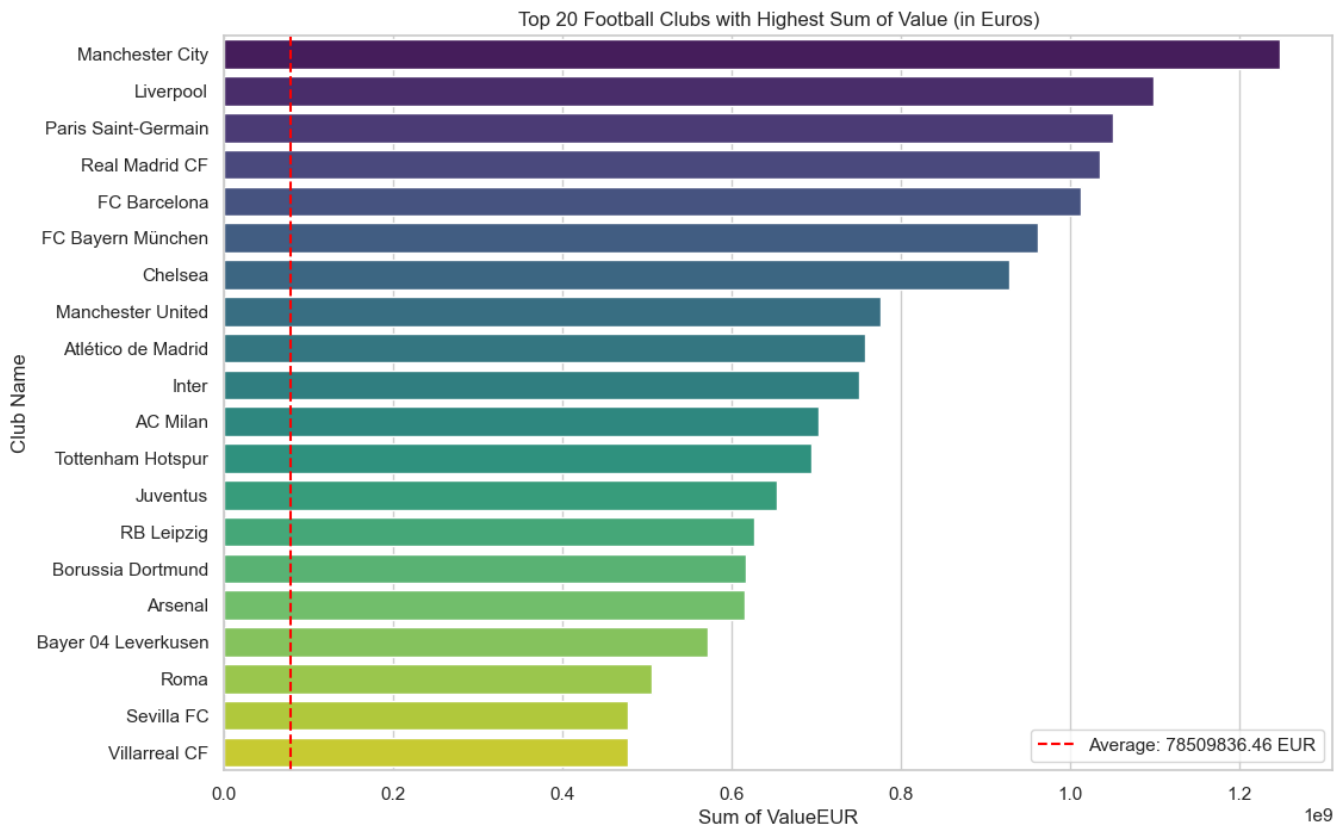
1. **Top Earners:** K. Benzema and R. Lewandowski are the highest-paid players, with wages of approximately 450,000 Euros and 420,000 Euros, respectively. Both players have enjoyed outstanding careers, consistently performing at the highest level, which has led to lucrative contracts and endorsements.
2. **Wage Disparity:** The large gap between the wages of the top players and the rest of the top 20 players is striking. Players like K. Benzema and R. Lewandowski earn significantly more than the rest of the list. This wage disparity is a reflection of the unique value and marketability that elite players bring to their clubs and sponsors.
3. **Lowest Paid Players:** E. Haaland and E. Hazard, while among the top 20 highest earners, have comparatively lower wages of approximately 220,000 Euros. This demonstrates that even within the top echelon of players, there can be variations in wages based on factors like individual performance, contract negotiations, and the financial resources of their clubs.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Wage Gap with Average Players:** The observation that the average wage of a player is around 8824.54 Euros highlights the significant wage gap between the top players and the average players in the game. Top players earn approximately 20 to 40 times more than the average players. This wage differential is driven by the unique skills, marketability, and performance of elite players, which make them highly sought after and justify their higher wages.
5. **Market Dynamics:** The significant variation in wages within the top 20 players underscores the complex market dynamics of football. Player wages are influenced by various factors, including club finances, competition for talent, individual performance, and endorsements. These factors result in variations in player earnings even among the top earners.

► The graph highlights the substantial wage disparities among football players, with the top earners like K. Benzema and R. Lewandowski enjoying significantly higher wages than their peers. The fact that top players earn many times more than the average player reflects the value and market demand associated with the very best talents in the sport.

EXPLORATORY DATA ANALYSIS for PLAYERS



The bar chart displaying the top 20 football clubs with the highest sum of value in Euros

EXPLORATORY DATA ANALYSIS for PLAYERS

Several key insights:

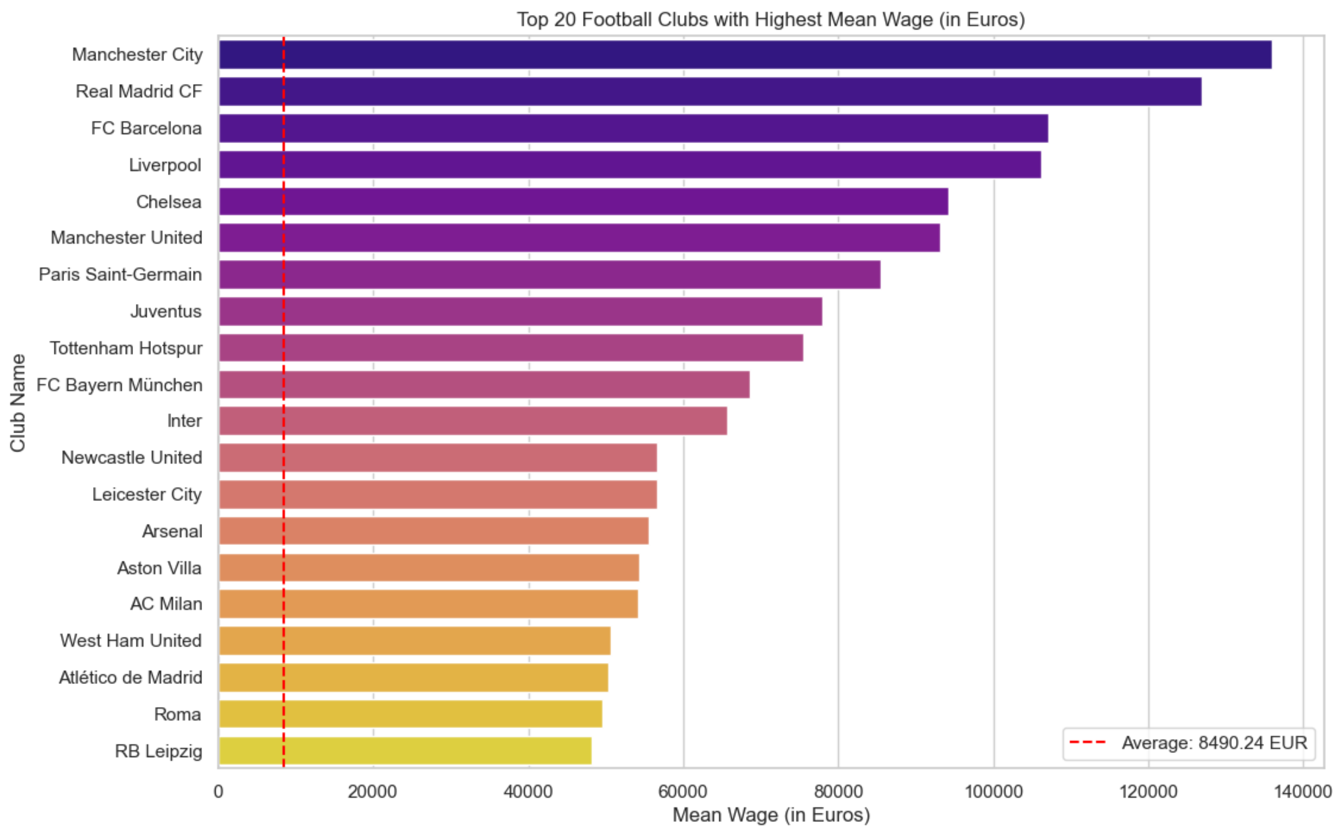
1. **Financial Dominance of Top Clubs:** Manchester City, Liverpool, and Paris Saint-Germain occupy the top positions with the highest club values, demonstrating their financial strength in the football world. These clubs have consistently achieved success on the pitch and have leveraged their strong financial positions to compete at the highest level.
2. **Variance in Club Values:** The chart illustrates a wide variance in club values within the top 20. Clubs like Manchester City, Liverpool, and Paris Saint-Germain boast values exceeding 1 billion Euros, while others, such as Roma, Sevilla FC, and Villarreal CF, have significantly lower values of approximately 500 million Euros. This variation is influenced by factors such as revenue streams, player assets, commercial partnerships, and financial stability.
3. **Significant Gap with Average Clubs:** The chart emphasizes the substantial gap between the values of the top clubs and the average club, which is indicated at around 78,509,836 Euros. This gap highlights the competitive advantage enjoyed by the top clubs in terms of attracting top talent, expanding their global fan base, and generating substantial revenue. The financial dominance of these top clubs reflects their ability to make major investments in player acquisitions, infrastructure, and marketing.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Market Dynamics and Investment:** Club values are subject to market dynamics, including transfer fees, sponsorship deals, and broadcasting contracts. The top clubs consistently perform well both on and off the field, contributing to their high valuations. Their financial resources enable them to invest in world-class facilities, scout top talents, and strengthen their brand globally.
5. **Financial Impact on Competitiveness:** The financial disparities among clubs can have a significant impact on the competitive landscape of football. Top clubs can maintain their dominance through sustained investments and player acquisitions, while others may struggle to compete at the highest level due to financial constraints.

► The chart underscores the significant financial disparities among football clubs, with a select few like Manchester City, Liverpool, and Paris Saint-Germain commanding the highest values. These clubs' financial strength positions them as major players in the global football landscape, allowing them to remain competitive and attract top talent.

EXPLORATORY DATA ANALYSIS for PLAYERS



The horizontal bar chart presenting the top 20 football clubs with the highest mean wage in Euros

EXPLORATORY DATA ANALYSIS for PLAYERS

Several key insights:

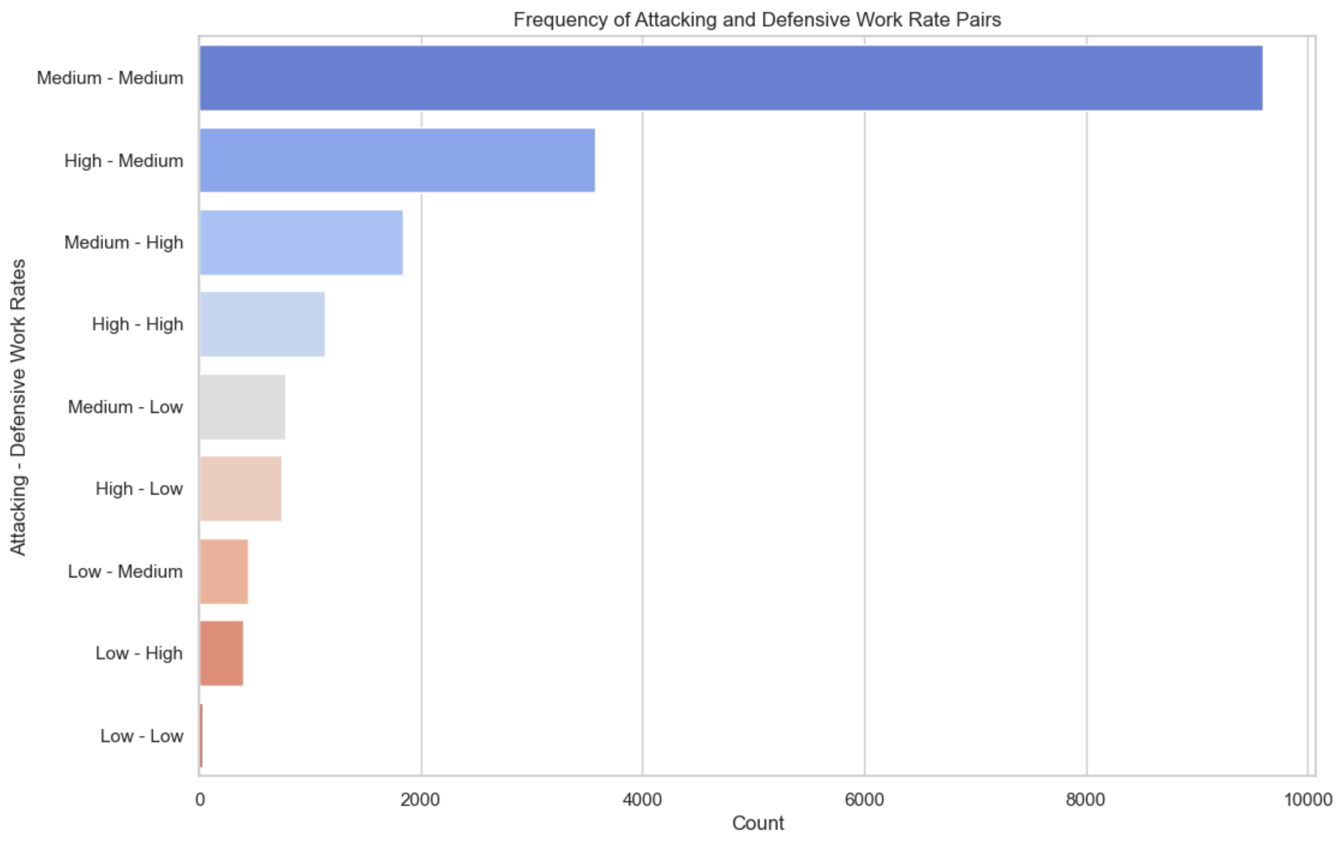
1. **Manchester City's Wage Dominance:** Manchester City leads the list with the highest mean wage, nearly 140,000 Euros. This reflects the club's significant financial resources and its ability to attract top talent by offering competitive wages. Manchester City's wage expenditure is indicative of its commitment to building a highly competitive squad.
2. **Variance in Mean Wages:** The chart illustrates a wide variance in mean wages within the top 20 clubs. While Manchester City commands the highest wage, clubs like RB Leipzig have significantly lower mean wages, around 50,000 Euros. This variance is influenced by factors such as club revenue, financial management, and the competitive landscape in their respective leagues.
3. **Significant Wage Gap with Average Clubs:** The chart highlights the substantial wage gap between the top clubs and the average club, with the average wage of a club indicated at approximately 8,490.24 Euros. This gap emphasizes the competitive advantage enjoyed by the top clubs in terms of attracting and retaining top talent. Top clubs pay their players about 4 to 14 times more than the average clubs, underlining the financial powerhouses' ability to offer competitive wages.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Impact on Player Recruitment and Performance:** High mean wages among top clubs are often associated with their ability to attract star players and build successful squads. These clubs can afford to invest in top talent, which contributes to their on-field success and competitiveness.
5. **Market Dynamics and Financial Resources:** Mean wages are influenced by club financial resources, league competitiveness, and market dynamics. Clubs in more lucrative leagues or with strong commercial partnerships often have a financial advantage in offering competitive wages. Successful clubs can generate additional revenue through prize money, sponsorships, and merchandise sales, further enhancing their ability to offer higher wages.

► The chart highlights the significant wage disparities among football clubs, with top clubs like Manchester City offering substantially higher mean wages compared to the average club. These financial powerhouses' capacity to pay top salaries contributes to their competitiveness and allows them to attract and retain top players, ultimately shaping their success on the field.

EXPLORATORY DATA ANALYSIS for PLAYERS



The bar graph displaying the frequency of attacking and defensive work rate pairs

EXPLORATORY DATA ANALYSIS for PLAYERS

Several insights into the playing styles of footballers:

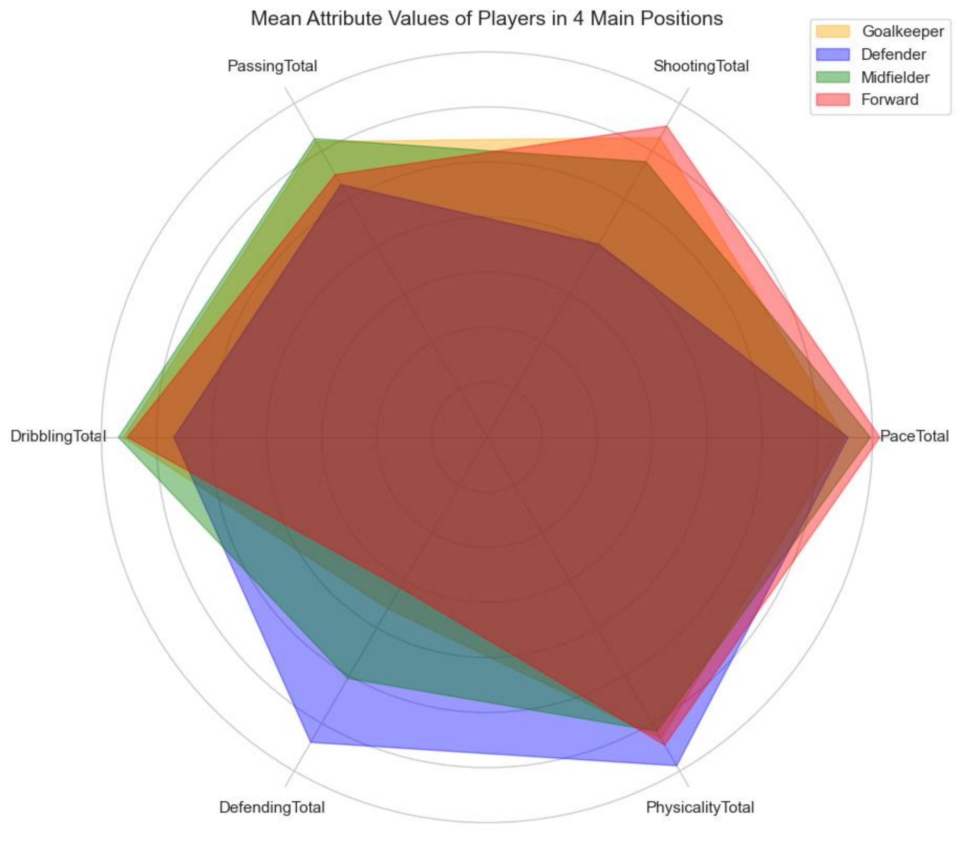
1. **Dominance of Medium-Medium Pair:** The most prevalent work rate pair is medium-medium, with an approximate frequency of 9600. This indicates that a significant number of footballers exhibit a balanced approach, contributing both to attacking and defensive aspects of the game. Players with medium-medium work rates are versatile and can engage in both offensive and defensive duties, making them valuable assets on the field.
2. **Variety in Work Rate Combinations:** The graph showcases a range of work rate pairs, including high-medium, medium-high, and high-high pairs. These combinations suggest diverse playing styles among footballers, with some prioritizing offensive contributions, some emphasizing defensive responsibilities, and others striking a balance between the two. The presence of different work rate pairs reflects the variety of roles and playing philosophies in modern football.
3. **Low-Low Pair as the Least Frequent:** The low-low work rate pair has the lowest frequency. This suggests that footballers with a minimal inclination toward both attacking and defensive duties are less common. Players with low-low work rates may be more specialized in a specific aspect of the game or may have a more limited involvement in both attacking and defensive phases.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Influence on Team Tactics and Strategy:** The distribution of work rate pairs can affect team tactics and strategy. Teams with a majority of medium-medium work rate pairs may adopt a balanced and adaptable style of play. Conversely, teams with a higher frequency of high-high pairs may prioritize an aggressive, attacking approach, while those with low-low pairs might focus on defensive solidity and specialization.
5. **Player Versatility and Adaptability:** The prevalence of medium-medium pairs suggests that versatility and adaptability are highly valued traits in modern football. Players who can seamlessly transition between attacking and defensive roles contribute to team flexibility and can fulfill multiple positions depending on tactical requirements.

► The bar graph reveals the frequency distribution of attacking and defensive work rate pairs, showcasing the prevalence of the medium-medium pair and the diversity of playing styles among footballers. The insights from this graph contribute to understanding the dynamics of team composition, player roles, and strategic approaches in the ever-evolving landscape of football.

EXPLORATORY DATA ANALYSIS for PLAYERS



The radar chart comparing mean attribute values across four main football positions — Goalkeeper, Defender, Midfielder, and Forward —

EXPLORATORY DATA ANALYSIS for PLAYERS

Valuable insights into the distinct skill sets of players in each position:

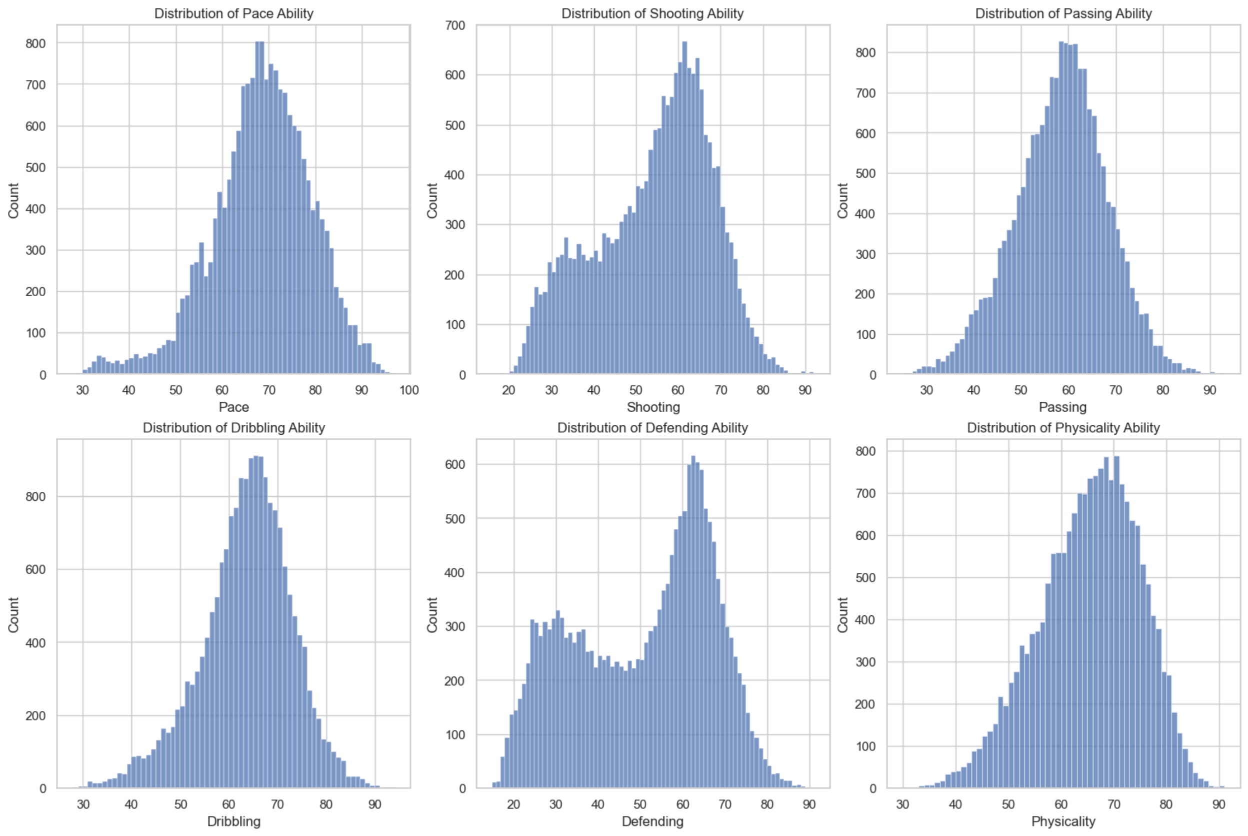
1. **Midfielders' Uniform Attributes:** The observation that midfielders have the most uniform attributes in all aspects suggests that players in this position are well-rounded and contribute consistently across various soccer abilities. This versatility is a key characteristic, allowing midfielders to influence the game both defensively and offensively.
2. **Defenders' Strength in Defending and Physicality:** Defenders stand out with the highest mean attribute values in defending and physicality. This reinforces the traditional role of defenders as players who excel in stopping opponents and engaging in physical battles. Their defensive prowess contributes to the overall stability of the team's backline.
3. **Forwards' Emphasis on Shooting and Pace:** Forwards showcase the highest mean attribute values in shooting and pace abilities. This aligns with the expectations for players in the forward position, where scoring goals and utilizing speed to create goal-scoring opportunities are crucial. The emphasis on shooting and pace reflects the offensive nature of the forward role.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Trade-Off Between Abilities:** The radar chart highlights a trade-off between different abilities, indicating that no single position dominates all axes. This reflects the specialization and unique roles associated with each position. For instance, goalkeepers may have lower mean attribute values in pace, emphasizing their focus on shot-stopping and goalkeeping skills. Similarly, forwards may have lower mean attribute values in defending, as their primary role is goal-scoring rather than defensive duties.
5. **Strategic Implications for Team Composition:** Coaches and team managers can use the insights from the radar chart to strategically compose their teams. Understanding the strengths and weaknesses associated with each position helps in player selection, formation planning, and overall team strategy.

► The radar chart provides a comprehensive view of mean attribute values across key soccer abilities for different player positions. The trade-offs and specialized skill sets associated with each position underscore the complexity and strategic considerations involved in team composition and player roles in football.

EXPLORATORY DATA ANALYSIS for PLAYERS



EXPLORATORY DATA ANALYSIS for PLAYERS

The histograms depicting the distribution of football abilities, such as pace, shooting, passing, dribbling, defending, and physicality, provide valuable insights into the variations and characteristics of player skills:

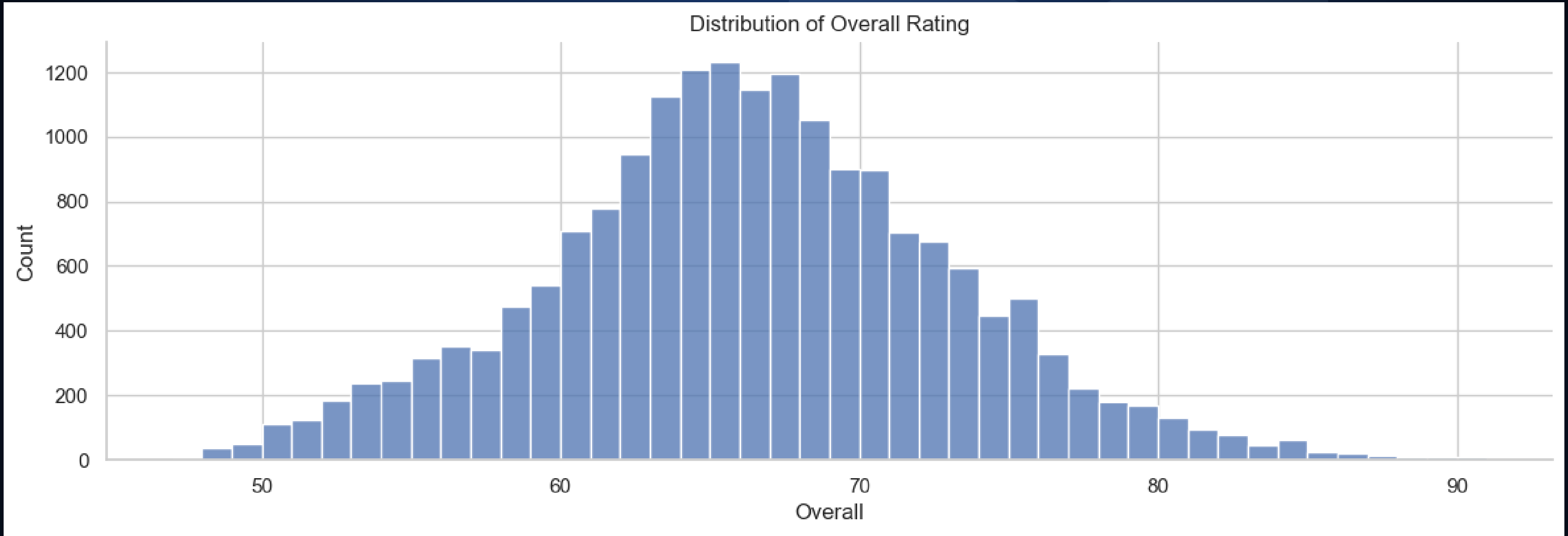
1. **Normal Distribution Patterns:** The approximately bell-shaped histograms across different abilities indicate a normal distribution in each case. This suggests that, on average, most players have scores around the mean, while fewer players exhibit either very high or very low scores. The normal distribution is a common pattern in statistical data and implies a balanced distribution of abilities among football players.
2. **Symmetry and Skewness Differences:** The variations in shapes and peaks among the histograms highlight differences in the distribution patterns of specific abilities. For instance, the passing histogram is described as more symmetrical with a single peak, indicating that passing abilities are more consistent and evenly distributed among players. In contrast, the defending histogram is noted as more skewed to the right with two peaks, suggesting that defending abilities exhibit greater variability and may be influenced by distinct subgroups of players.
3. **Bimodal Nature of Defending Ability:** The presence of two peaks in the defending histogram suggests a bimodal distribution. This indicates the existence of two distinct groups of players with different levels of defending ability. The bimodal nature may be attributed to specialized roles within the defending position, such as center-backs and full-backs, each having unique defensive attributes and contributing to the observed distribution.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Implications for Player Evaluation and Team Strategy:** Coaches and team analysts can use these insights to better understand the distribution of abilities within their squads. Recognizing the normal distribution patterns and variations in skewness can inform player evaluation processes, helping teams identify areas of strength and potential areas for improvement. It also aids in developing strategic approaches that leverage the diverse skill sets present within the team.
5. **Focus on Consistency and Variability:** The observations about passing ability being more consistent and normally distributed, while defending ability is more variable and bimodal, highlight the importance of considering the nature of each ability. Teams may prioritize consistent and evenly distributed abilities in certain positions while acknowledging and strategizing around the variability and specialization observed in other positions.

► The histograms provide a visual representation of the distribution of football abilities, offering insights into the normality, symmetry, skewness, and bimodal nature of these distributions. This information can be valuable for teams and coaches in player assessment, team composition, and strategic decision-making in the world of football.

EXPLORATORY DATA ANALYSIS for PLAYERS



The histogram displaying the distribution of overall ratings among football players.

EXPLORATORY DATA ANALYSIS for PLAYERS

Several noteworthy insights:

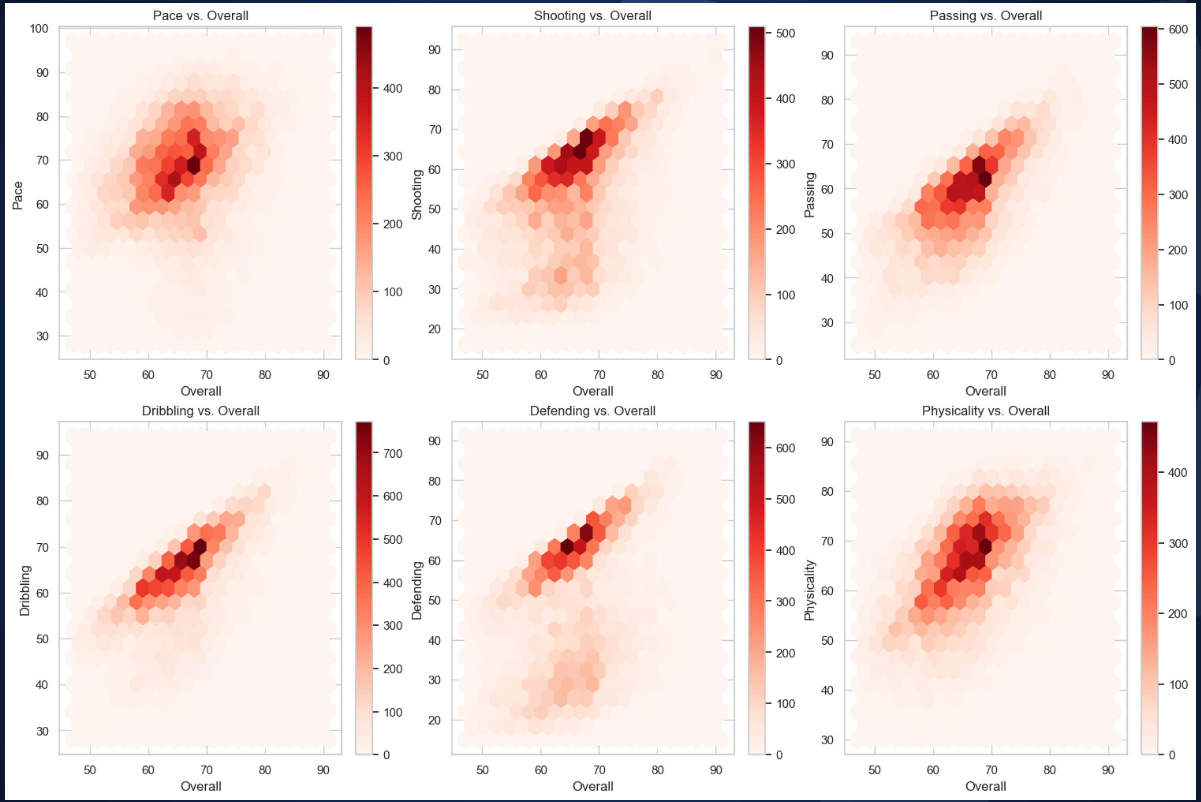
1. **Normal Distribution and Central Tendency:** The observation that the data appears to be normally distributed with a mean of around 66 and a standard deviation of around 10 suggests that the majority of soccer players exhibit average skills and abilities. The normal distribution is a common pattern in datasets, indicating a balanced distribution around the mean.
2. **Concentration of Ratings:** The concentration of ratings between 60 and 75 signifies that most players fall within this range, emphasizing the prevalence of players with average skills. This range likely represents a broad segment of players who contribute consistently to their teams without necessarily standing out as exceptional or below average.
3. **Majority of Players Have Average Skills:** The fact that the majority of ratings fall within the 60 to 75 range reinforces the idea that most soccer players possess average skills and abilities. This aligns with the expectation that a substantial portion of the player population contributes at a competitive and reliable level.
4. **Outliers Indicate Exceptional and Poor Performances:** The presence of outliers on both the lower and higher ends of the spectrum suggests that there are players with exceptional as well as below-average skills and abilities. These outliers could represent star players with outstanding performances or players who may be struggling to meet the average standards.

EXPLORATORY DATA ANALYSIS for PLAYERS

5. **Skill Disparities Across the Spectrum:** The spread of ratings across the entire spectrum implies that there is considerable variability in skill levels within the soccer player population. While most players fall within the average range, the outliers highlight the existence of players who significantly excel or lag behind in their overall abilities.
6. **Team Composition Considerations:** Teams and coaches can use this distribution information to assess the overall skill composition of their squads. Understanding the prevalence of average players, as well as the presence of exceptional and less skilled players, can inform team selection, player development strategies, and overall team dynamics.

► The histogram provides a visual representation of the distribution of overall ratings among soccer players, showcasing a normal distribution with a concentration of ratings in the average range. The outliers at both extremes draw attention to players with exceptional or below-average skills, adding a layer of complexity to the overall skill landscape within the soccer player population.

EXPLORATORY DATA ANALYSIS for PLAYERS



The hexbin plots provide insights into the relationship between overall ratings and specific skill ratings for various football abilities.

EXPLORATORY DATA ANALYSIS for PLAYERS

Several key observations:

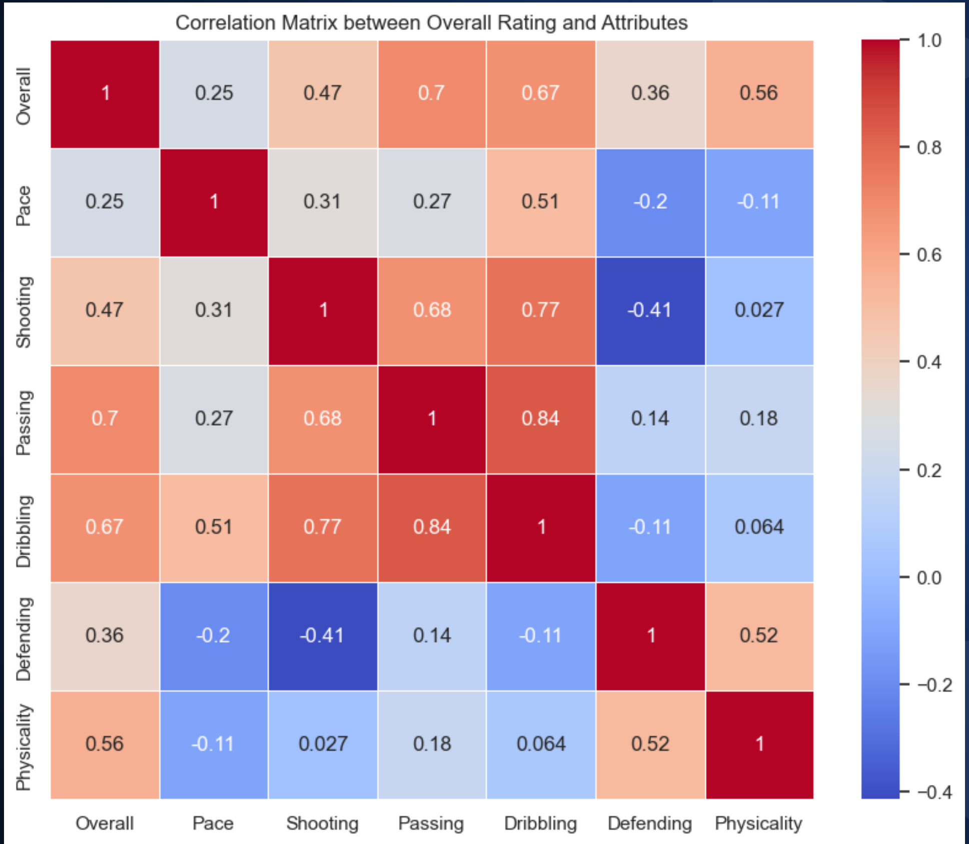
1. **Positive Correlation Across Skills:** The consistent positive correlation between overall ratings and specific skill ratings across all six football abilities indicates that, on average, as a player's overall rating increases, their proficiency in individual skills also tends to increase. This aligns with the expectation that higher-rated players excel across multiple aspects of the game.
2. **Varied Strengths of Correlation:** The variation in the strength of correlation among skills suggests that some skills have a more pronounced impact on the overall rating than others. The strong correlation between overall rating and shooting skill implies that shooting prowess significantly contributes to a player's overall effectiveness on the field. In contrast, the weaker correlation with pace suggests that pace may have a less decisive influence on the overall rating.
3. **Diversity in Player Profiles:** The wide range of variation in specific skill ratings for players with similar overall ratings highlights the diversity in player profiles. For instance, players with an overall rating of 80 can exhibit considerable variation in their passing skill ratings, ranging from 60 to 90. This diversity indicates that players with similar overall ratings can have distinct strengths and weaknesses in specific areas of the game.

EXPLORATORY DATA ANALYSIS for PLAYERS

4. **Individualized Player Strengths and Weaknesses:** The existence of players with similar overall ratings but differing skill ratings emphasizes that each player possesses unique strengths and weaknesses. Some players may excel in passing, while others may prioritize shooting or other skills. This individualization contributes to the richness and complexity of player profiles in football.
5. **Identification of Player Types:** The hexbin plots provide a tool for identifying different player types based on their specific skill ratings. For example, players clustered along the upper end of the shooting skill rating spectrum may be identified as prolific goal scorers. In contrast, the scattered points in the pace skill rating plot suggest that pace may be a less defining factor in distinguishing player types.
6. **Implications for Team Strategy and Player Recruitment:** Teams and managers can leverage these insights to formulate strategies that align with the specific strengths and weaknesses of their players. Understanding the nuanced relationships between overall ratings and individual skills allows for targeted recruitment and tactical planning based on the team's desired style of play.

► The hexbin plots provide a nuanced view of the relationship between overall ratings and specific skill ratings, revealing both common trends and individual variations among players. This information aids in understanding player diversity, identifying key contributors to overall effectiveness, and tailoring team strategies to capitalize on the unique strengths of individual players.

EXPLORATORY DATA ANALYSIS for PLAYERS



Correlation Matrix between Overall Rating and Attributes

	Overall	Pace	Shooting	Passing	Dribbling	Defending	Physicality
Overall	1	0.25	0.47	0.7	0.67	0.36	0.56
Pace	0.25	1	0.31	0.27	0.51	-0.2	-0.11
Shooting	0.47	0.31	1	0.68	0.77	-0.41	0.027
Passing	0.7	0.27	0.68	1	0.84	0.14	0.18
Dribbling	0.67	0.51	0.77	0.84	1	-0.11	0.064
Defending	0.36	-0.2	-0.41	0.14	-0.11	1	0.52
Physicality	0.56	-0.11	0.027	0.18	0.064	0.52	1

The correlation matrix sheds light on the relationship between overall ratings and various attributes of football players.

EXPLORATORY DATA ANALYSIS for PLAYERS

Valuable insights into the interplay of skills and their impact on the overall assessment:

1. **Positive Correlation Between Dribbling and Passing:** The highest positive correlation (0.84) between dribbling and passing ratings suggests a strong association between these two skills. Players with advanced dribbling abilities are likely to also exhibit high passing skills. This correlation aligns with the idea that players adept at ball control and maneuvering are often effective playmakers.
2. **Negative Correlation Between Defending and Shooting:** The lowest negative correlation (-0.41) between defending and shooting ratings indicates an inverse relationship. Players excelling in defensive skills tend to have lower shooting ratings. This suggests a trade-off between defensive prowess and goal-scoring abilities, emphasizing the specialization often seen in players fulfilling defensive roles.
3. **Attributes with No or Weak Correlation:** The observation that pace exhibits a weak correlation with overall ratings implies that this skill may not significantly influence a player's overall rating. In practical terms, this means that a player's pace prowess alone may not be a decisive factor in determining their overall effectiveness on the field.
4. **Significance of Dribbling and Passing in Overall Rating:** The strong positive correlation between dribbling and passing ratings suggests that players excelling in ball control and distribution are likely to receive higher overall ratings. This emphasizes the importance of technical skills and playmaking abilities in contributing to a player's overall effectiveness.

EXPLORATORY DATA ANALYSIS for PLAYERS

5. **Strategic Considerations for Player Recruitment and Team Formation:** Teams and managers can use these correlation insights strategically. For instance, when recruiting players, they may prioritize individuals with a balanced combination of dribbling and passing skills. The inverse correlation between defending and shooting ratings could inform decisions about player roles and tactical formations.
6. **Understanding Attribute Impact on Overall Rating:** The correlation matrix provides a nuanced understanding of how specific attributes contribute to a player's overall rating. Recognizing which skills have a strong correlation, weak correlation, or no correlation helps teams assess player strengths and weaknesses more comprehensively.

► The correlation matrix serves as a valuable tool for deciphering the relationships between overall ratings and various attributes in soccer players. The identified correlations provide actionable insights for teams and managers in player recruitment, tactical planning, and understanding the nuanced dynamics of player skills in the context of overall performance.

MACHINE LEARNING MODELS for PREDICTION

Predict Overall Rating

Predict Best Position

MACHINE LEARNING MODELS for PREDICTION

Feature Selection Overview for Machine Learning Models

In the pursuit of effective machine learning models for football player analysis, a streamlined selection process focused on attributes containing the term 'Total' was employed. These chosen features provide a consolidated and comprehensive view of player skills. Below is a concise overview of the selected attributes:

1. Best Position:

- *Definition:* Player's optimal field position.
- *Significance:* Informs team formation and tactical decisions.

2. Overall:

- *Definition:* Comprehensive player rating.
- *Significance:* Summarizes overall performance succinctly.

3. Pace:

- *Definition:* Player's speed and agility.
- *Significance:* Impacts offensive and defensive strategies.

4. Shooting:

- *Definition:* Player's goal-scoring proficiency.
- *Significance:* Crucial for team success.

5. Passing:

- *Definition:* Player's accuracy and effectiveness in passing.
- *Significance:* Key for ball movement and possession.

6. Dribbling:

- *Definition:* Player's ability to maneuver with the ball.
- *Significance:* Influences offensive play.

7. Defending:

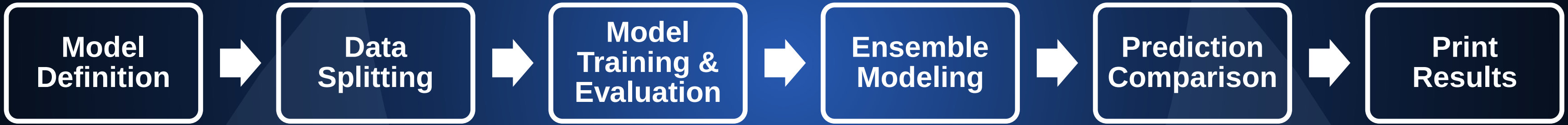
- *Definition:* Player's defensive capabilities.
- *Significance:* Critical for team resilience.

8. Physicality:

- *Definition:* Player's physical strength and stamina.
- *Significance:* Insights into endurance and on-field effectiveness.

MACHINE LEARNING MODELS for PREDICTION

Machine-learning workflow for regression analysis



Model Definition

Data Splitting

Model Training & Evaluation

Ensemble Modeling

Prediction Comparison

Print Results

MACHINE LEARNING MODELS for PREDICTION

The detailed outline of a machine-learning workflow for regression analysis:

1. Model Definition:

- A list is created, comprising dictionaries for different regression models such as Linear Regression, K-Nearest Neighbor Regression, Decision Tree Regression, Random Forest Regression, Neural Network Regression, Gradient Boosting Regression, AdaBoost Regression, XGBoost Regression, LightGBM Regression, and Support Vector Machines.
- Each dictionary includes the model's name, the estimator instance, and a set of hyperparameters to be tuned using GridSearchCV.

2. Data Splitting:

- The dataset is split into training and testing sets.
- Relevant features (X) and the target variable (y) are defined.

3. Model Training and Evaluation:

For each model:

- GridSearchCV is employed to tune hyperparameters using cross-validation (cv=5).
- The best estimator, best hyperparameters, and R2 score on the test set are printed.

MACHINE LEARNING MODELS for PREDICTION

4. Ensemble Modeling:

- A list of the best estimators from the trained models is created.
- A Voting Regressor is instantiated using the best estimators.
- The Voting Regressor is fitted on the training data and evaluated on the test data.
- A Stacking Regressor is instantiated using the best estimators with a final estimator as Linear Regression.
- The Stacking Regressor is fitted on the training data and evaluated on the test data.

5. Prediction Comparison:

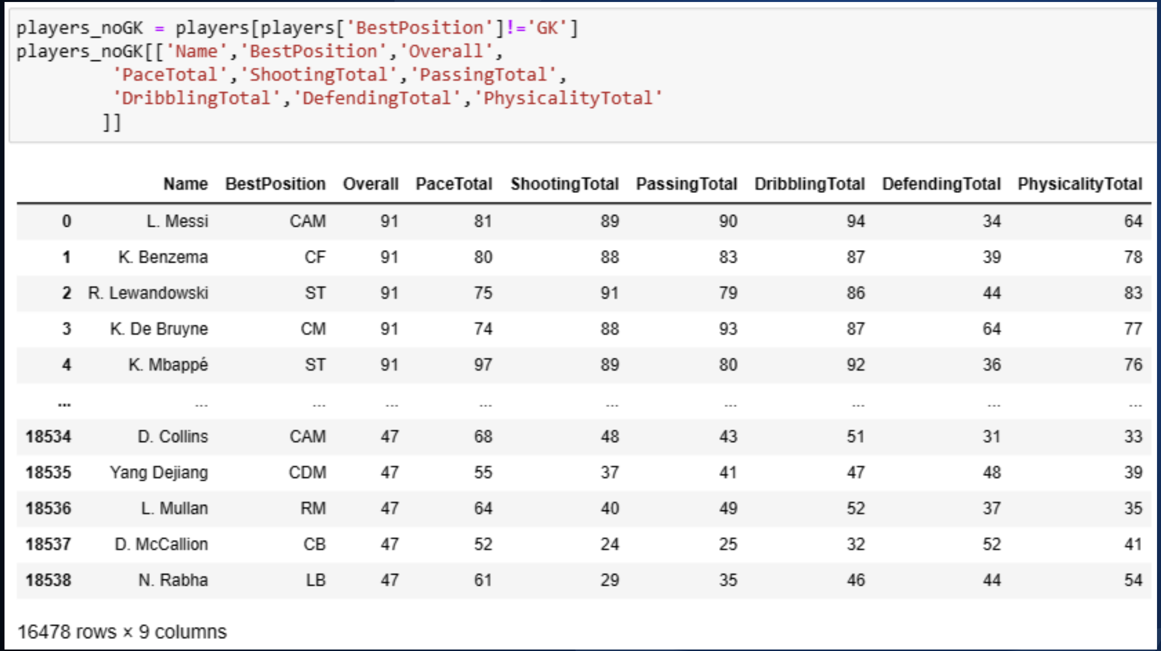
- Predictions are made using each model, the Voting Regressor, and the Stacking Regressor on a subset of the test data (first 20 rows).
- Predicted results are stored in a dataframe along with the actual values.

6. Print Results:

- R2 scores for the Voting Regressor and Stacking Regressor on the test data are printed.
- The dataframe containing predicted and actual values is printed.

MACHINE LEARNING MODELS for PREDICTION

Predict Overall Ratings for Players excluding Goalkeepers



```
players_noGK = players[players['BestPosition']!='GK']
players_noGK[['Name','BestPosition','Overall',
          'PaceTotal','ShootingTotal','PassingTotal',
          'DribblingTotal','DefendingTotal','PhysicalityTotal'
         ]]
```

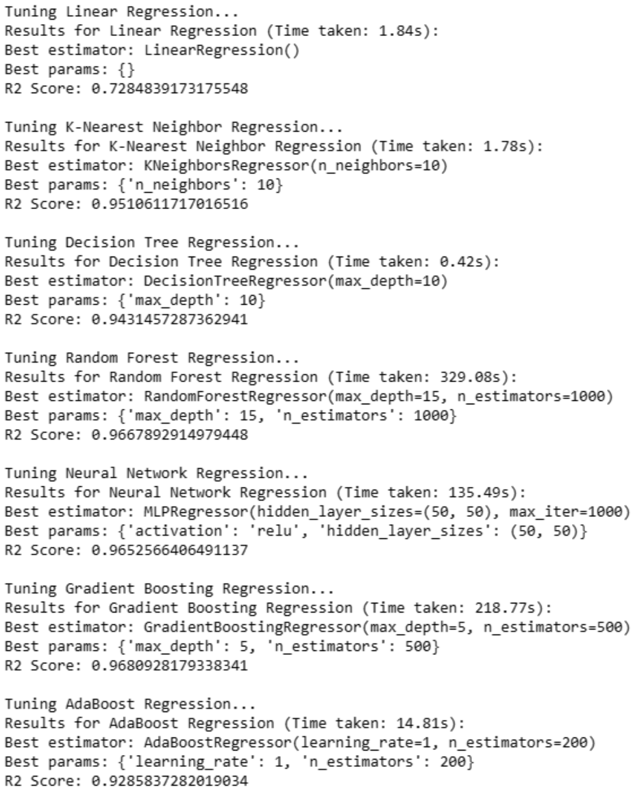
	Name	BestPosition	Overall	PaceTotal	ShootingTotal	PassingTotal	DribblingTotal	DefendingTotal	PhysicalityTotal
0	L. Messi	CAM	91	81	89	90	94	34	64
1	K. Benzema	CF	91	80	88	83	87	39	78
2	R. Lewandowski	ST	91	75	91	79	86	44	83
3	K. De Bruyne	CM	91	74	88	93	87	64	77
4	K. Mbappé	ST	91	97	89	80	92	36	76
...	...	...	...	...	...	...	...	...	...
18534	D. Collins	CAM	47	68	48	43	51	31	33
18535	Yang Dejiang	CDM	47	55	37	41	47	48	39
18536	L. Mullan	RM	47	64	40	49	52	37	35
18537	D. McCallion	CB	47	52	24	25	32	52	41
18538	N. Rabha	LB	47	61	29	35	46	44	54

16478 rows × 9 columns

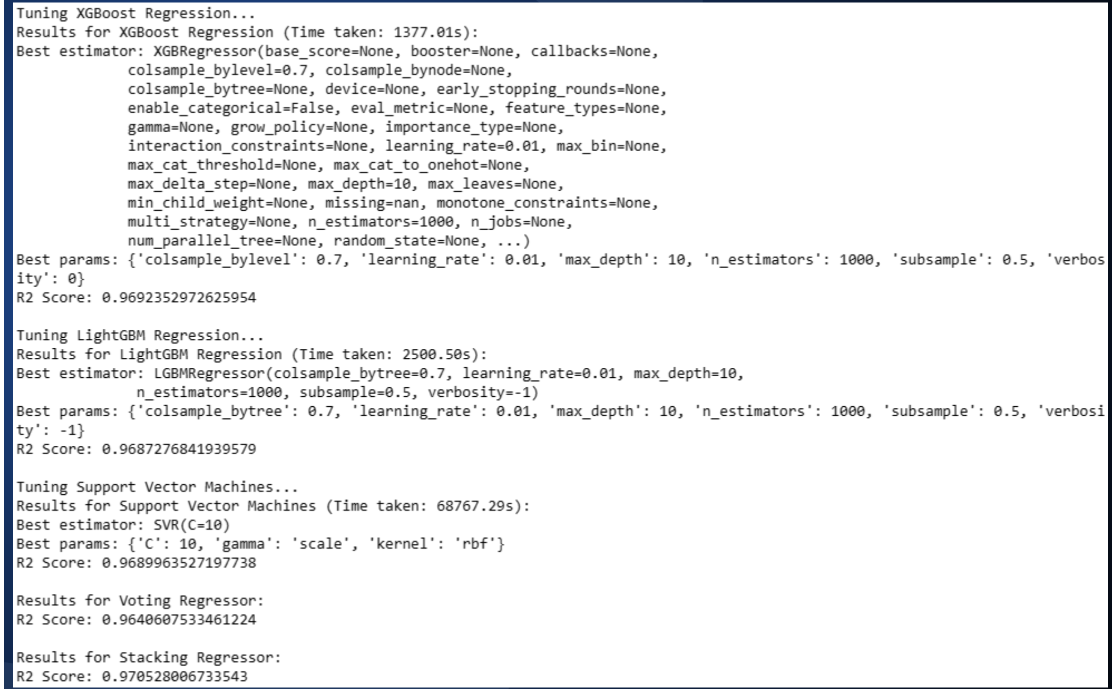
We manipulate the DataFrame players_noGK and show the tabular dataset, offering football player statistics, excluding goalkeeper data.

MACHINE LEARNING MODELS for PREDICTION

Predict Overall Ratings for Players excluding Goalkeepers



```
Tuning Linear Regression...
Results for Linear Regression (Time taken: 1.84s):
Best estimator: LinearRegression()
Best params: {}
R2 Score: 0.7284839173175548

Tuning K-Nearest Neighbor Regression...
Results for K-Nearest Neighbor Regression (Time taken: 1.78s):
Best estimator: KNeighborsRegressor(n_neighbors=10)
Best params: {'n_neighbors': 10}
R2 Score: 0.9510611717016516

Tuning Decision Tree Regression...
Results for Decision Tree Regression (Time taken: 0.42s):
Best estimator: DecisionTreeRegressor(max_depth=10)
Best params: {'max_depth': 10}
R2 Score: 0.9431457287362941

Tuning Random Forest Regression...
Results for Random Forest Regression (Time taken: 329.08s):
Best estimator: RandomForestRegressor(max_depth=15, n_estimators=1000)
Best params: {'max_depth': 15, 'n_estimators': 1000}
R2 Score: 0.9667892914979448

Tuning Neural Network Regression...
Results for Neural Network Regression (Time taken: 135.49s):
Best estimator: MLPRegressor(hidden_layer_sizes=(50, 50), max_iter=1000)
Best params: {'activation': 'relu', 'hidden_layer_sizes': (50, 50)}
R2 Score: 0.9652566406491137

Tuning Gradient Boosting Regression...
Results for Gradient Boosting Regression (Time taken: 218.77s):
Best estimator: GradientBoostingRegressor(max_depth=5, n_estimators=500)
Best params: {'max_depth': 5, 'n_estimators': 500}
R2 Score: 0.9680928179338341

Tuning AdaBoost Regression...
Results for AdaBoost Regression (Time taken: 14.81s):
Best estimator: AdaBoostRegressor(learning_rate=1, n_estimators=200)
Best params: {'learning_rate': 1, 'n_estimators': 200}
R2 Score: 0.9285837282019034
```



```
Tuning XGBoost Regression...
Results for XGBoost Regression (Time taken: 1377.01s):
Best estimator: XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=0.7, colsample_bynode=None,
             colsample_bytree=None, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=0.01, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=10, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=1000, n_jobs=None,
             num_parallel_tree=None, random_state=None, ...)
Best params: {'colsample_bylevel': 0.7, 'learning_rate': 0.01, 'max_depth': 10, 'n_estimators': 1000, 'subsample': 0.5, 'verbos
ity': 0}
R2 Score: 0.9692352972625954

Tuning LightGBM Regression...
Results for LightGBM Regression (Time taken: 2500.50s):
Best estimator: LGBMRegressor(colsample_bytree=0.7, learning_rate=0.01, max_depth=10,
              n_estimators=1000, subsample=0.5, verbosity=-1)
Best params: {'colsample_bytree': 0.7, 'learning_rate': 0.01, 'max_depth': 10, 'n_estimators': 1000, 'subsample': 0.5, 'verbosi
ty': -1}
R2 Score: 0.9687276841939579

Tuning Support Vector Machines...
Results for Support Vector Machines (Time taken: 68767.29s):
Best estimator: SVR(C=10)
Best params: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
R2 Score: 0.9689963527197738

Results for Voting Regressor:
R2 Score: 0.9640607533461224

Results for Stacking Regressor:
R2 Score: 0.970528006733543
```

MACHINE LEARNING MODELS for PREDICTION

Predict Overall Ratings for Players excluding Goalkeepers

1. Top Performers:

- **Stacking Regressor (R2: 0.9705):** Demonstrates superior predictive capabilities, showcasing the power of ensemble learning.
- **XGBoost Regression (R2: 0.9692):** Excellent performance with fine-tuned hyperparameters, a top choice for accurate predictions.

2. Balanced Performance:

- **Voting Regressor (R2: 0.9641):** Achieves a balanced performance by combining diverse models, a reliable choice without overfitting.
- **Gradient Boosting Regression (R2: 0.9681):** Strong predictive capability with an optimal learning rate of 1 and 200 estimators.

3. Model Complexity:

- **Random Forest Regression (R2: 0.9668):** Effectively captures nonlinear relationships with an optimal depth of 15 and 1000 estimators.

4. Computational Intensity:

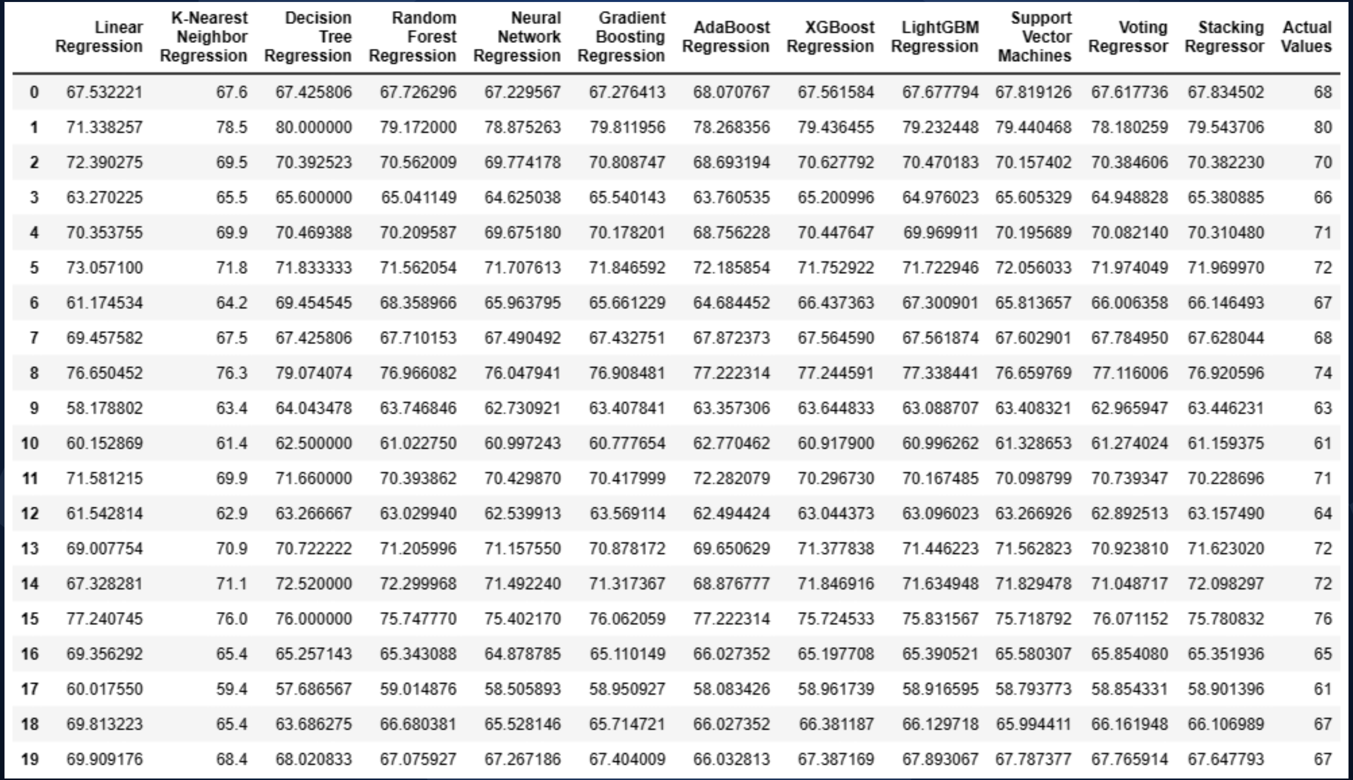
- **Support Vector Machines (R2: 0.9690):** High predictive capability, but notable for longer tuning time (68767.29s).

5. Ensemble Advantage:

- Leveraging ensemble methods enhances overall model performance, showcasing the synergy of diverse algorithms.

MACHINE LEARNING MODELS for PREDICTION

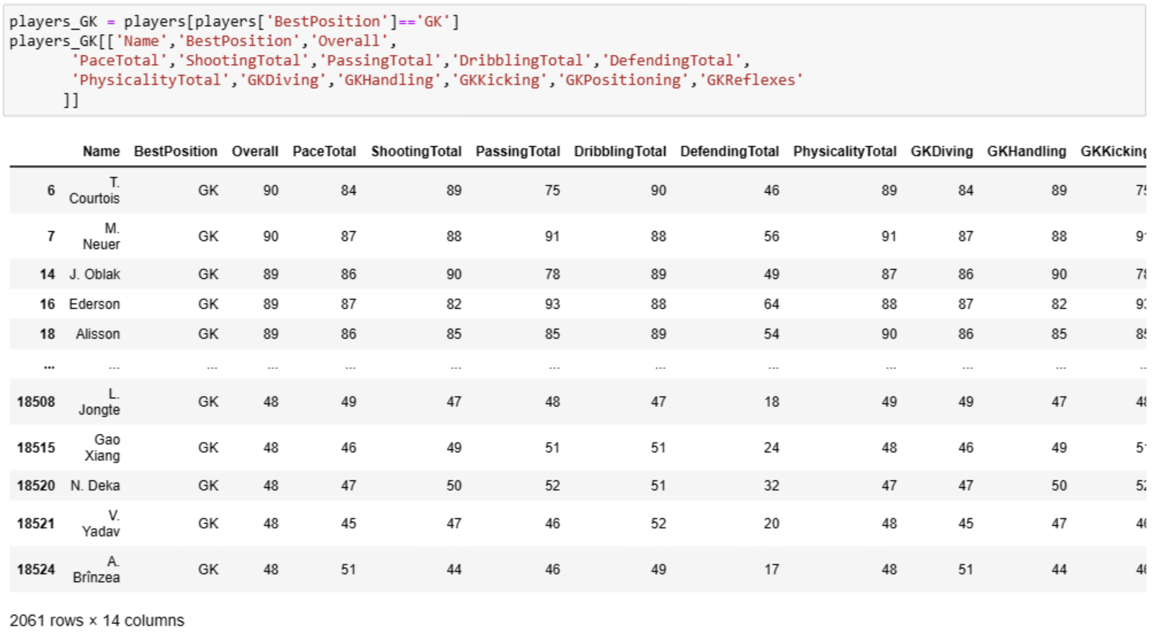
Predict Overall Ratings for Players excluding Goalkeepers



	Linear Regression	K-Nearest Neighbor Regression	Decision Tree Regression	Random Forest Regression	Neural Network Regression	Gradient Boosting Regression	AdaBoost Regression	XGBoost Regression	LightGBM Regression	Support Vector Machines	Voting Regressor	Stacking Regressor	Actual Values
0	67.532221	67.6	67.425806	67.726296	67.229567	67.276413	68.070767	67.561584	67.677794	67.819126	67.617736	67.834502	68
1	71.338257	78.5	80.000000	79.172000	78.875263	79.811956	78.268356	79.436455	79.232448	79.440468	78.180259	79.543706	80
2	72.390275	69.5	70.392523	70.562009	69.774178	70.808747	68.693194	70.627792	70.470183	70.157402	70.384606	70.382230	70
3	63.270225	65.5	65.600000	65.041149	64.625038	65.540143	63.760535	65.200996	64.976023	65.605329	64.948828	65.380885	66
4	70.353755	69.9	70.469388	70.209587	69.675180	70.178201	68.756228	70.447647	69.969911	70.195689	70.082140	70.310480	71
5	73.057100	71.8	71.833333	71.562054	71.707613	71.846592	72.185854	71.752922	71.722946	72.056033	71.974049	71.969970	72
6	61.174534	64.2	69.454545	68.358966	65.963795	65.661229	64.684452	66.437363	67.300901	65.813657	66.006358	66.146493	67
7	69.457582	67.5	67.425806	67.710153	67.490492	67.432751	67.872373	67.564590	67.561874	67.602901	67.784950	67.628044	68
8	76.650452	76.3	79.074074	76.966082	76.047941	76.908481	77.222314	77.244591	77.338441	76.659769	77.116006	76.920596	74
9	58.178802	63.4	64.043478	63.746846	62.730921	63.407841	63.357306	63.644833	63.088707	63.408321	62.965947	63.446231	63
10	60.152869	61.4	62.500000	61.022750	60.997243	60.777654	62.770462	60.917900	60.996262	61.328653	61.274024	61.159375	61
11	71.581215	69.9	71.660000	70.393862	70.429870	70.417999	72.282079	70.296730	70.167485	70.098799	70.739347	70.228696	71
12	61.542814	62.9	63.266667	63.029940	62.539913	63.569114	62.494424	63.044373	63.096023	63.266926	62.892513	63.157490	64
13	69.007754	70.9	70.722222	71.205996	71.157550	70.878172	69.650629	71.377838	71.446223	71.562823	70.923810	71.623020	72
14	67.328281	71.1	72.520000	72.299968	71.492240	71.317367	68.876777	71.846916	71.634948	71.829478	71.048717	72.098297	72
15	77.240745	76.0	76.000000	75.747770	75.402170	76.062059	77.222314	75.724533	75.831567	75.718792	76.071152	75.780832	76
16	69.356292	65.4	65.257143	65.343088	64.878785	65.110149	66.027352	65.197708	65.390521	65.580307	65.854080	65.351936	65
17	60.017550	59.4	57.686567	59.014876	58.505893	58.950927	58.083426	58.961739	58.916595	58.793773	58.854331	58.901396	61
18	69.813223	65.4	63.686275	66.680381	65.528146	65.714721	66.027352	66.381187	66.129718	65.994411	66.161948	66.106989	67
19	69.909176	68.4	68.020833	67.075927	67.267186	67.404009	66.032813	67.387169	67.893067	67.787377	67.765914	67.647793	67

MACHINE LEARNING MODELS for PREDICTION

Predict Overall Ratings for Goalkeepers



```
players_GK = players[players['BestPosition']=='GK']
players_GK[['Name','BestPosition','Overall',
       'PaceTotal','ShootingTotal','PassingTotal','DribblingTotal','DefendingTotal',
       'PhysicalityTotal','GKDiving','GKHandling','GKKicking','GKPositioning','GKReflexes'
      ]]
```

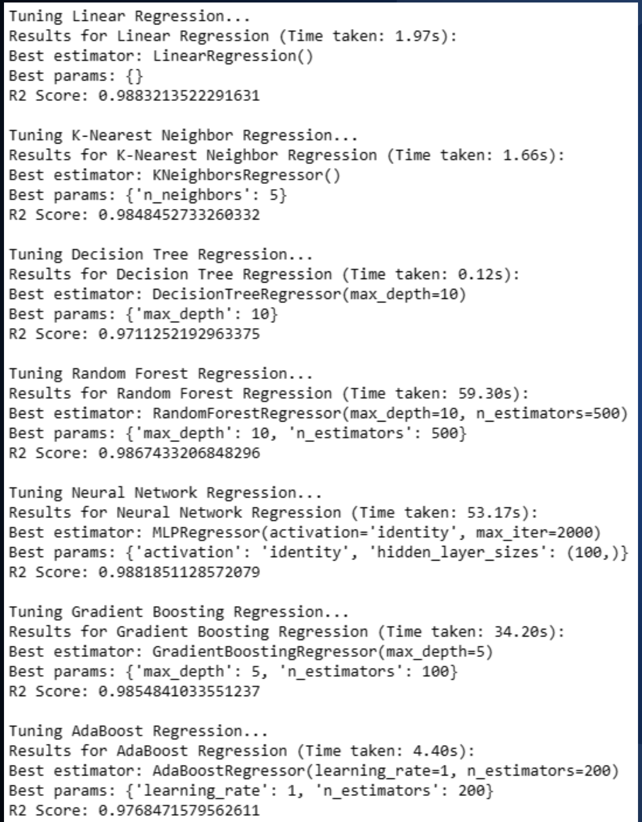
	Name	BestPosition	Overall	PaceTotal	ShootingTotal	PassingTotal	DribblingTotal	DefendingTotal	PhysicalityTotal	GKDiving	GKHandling	GKKicking
6	T. Courtois	GK	90	84	89	75	90	46	89	84	89	75
7	M. Neuer	GK	90	87	88	91	88	56	91	87	88	91
14	J. Oblak	GK	89	86	90	78	89	49	87	86	90	78
16	Ederson	GK	89	87	82	93	88	64	88	87	82	93
18	Alisson	GK	89	86	85	85	89	54	90	86	85	85
...	...	...	...	...	...	...	...	...	...	...	...	...
18508	L. Jongte	GK	48	49	47	48	47	18	49	49	47	48
18515	Gao Xiang	GK	48	46	49	51	51	24	48	46	49	51
18520	N. Deka	GK	48	47	50	52	51	32	47	47	50	52
18521	V. Yadav	GK	48	45	47	46	52	20	48	45	47	46
18524	A. Brînzea	GK	48	51	44	46	49	17	48	51	44	46

2061 rows × 14 columns

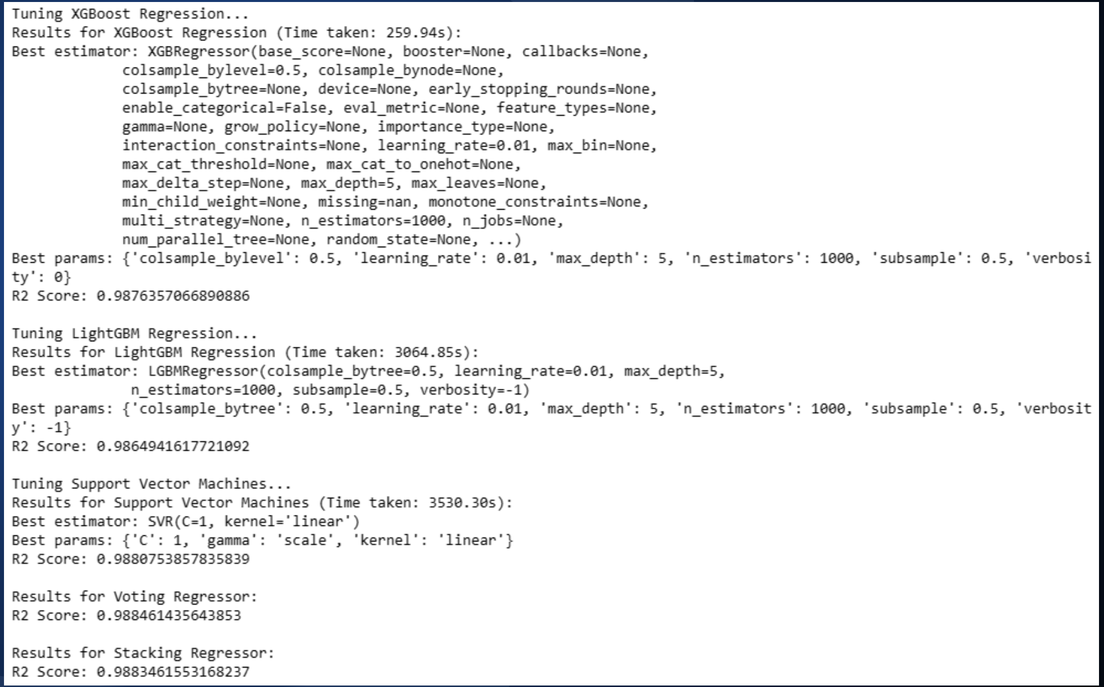
We manipulate the DataFrame players_GK and show the tabular dataset, offering football player statistics for goalkeepers.

MACHINE LEARNING MODELS for PREDICTION

Predict Overall Ratings for Goalkeepers



```
Tuning Linear Regression...
Results for Linear Regression (Time taken: 1.97s):
Best estimator: LinearRegression()
Best params: {}
R2 Score: 0.9883213522291631

Tuning K-Nearest Neighbor Regression...
Results for K-Nearest Neighbor Regression (Time taken: 1.66s):
Best estimator: KNeighborsRegressor()
Best params: {'n_neighbors': 5}
R2 Score: 0.9848452733260332

Tuning Decision Tree Regression...
Results for Decision Tree Regression (Time taken: 0.12s):
Best estimator: DecisionTreeRegressor(max_depth=10)
Best params: {'max_depth': 10}
R2 Score: 0.9711252192963375

Tuning Random Forest Regression...
Results for Random Forest Regression (Time taken: 59.30s):
Best estimator: RandomForestRegressor(max_depth=10, n_estimators=500)
Best params: {'max_depth': 10, 'n_estimators': 500}
R2 Score: 0.9867433206848296

Tuning Neural Network Regression...
Results for Neural Network Regression (Time taken: 53.17s):
Best estimator: MLPRegressor(activation='identity', max_iter=2000)
Best params: {'activation': 'identity', 'hidden_layer_sizes': (100,)}
R2 Score: 0.9881851128572079

Tuning Gradient Boosting Regression...
Results for Gradient Boosting Regression (Time taken: 34.20s):
Best estimator: GradientBoostingRegressor(max_depth=5)
Best params: {'max_depth': 5, 'n_estimators': 100}
R2 Score: 0.9854841033551237

Tuning AdaBoost Regression...
Results for AdaBoost Regression (Time taken: 4.40s):
Best estimator: AdaBoostRegressor(learning_rate=1, n_estimators=200)
Best params: {'learning_rate': 1, 'n_estimators': 200}
R2 Score: 0.9768471579562611
```



```
Tuning XGBoost Regression...
Results for XGBoost Regression (Time taken: 259.94s):
Best estimator: XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=0.5, colsample_bynode=None,
             colsample_bytree=None, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=0.01, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=5, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=1000, n_jobs=None,
             num_parallel_tree=None, random_state=None, ...)
Best params: {'colsample_bylevel': 0.5, 'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 1000, 'subsample': 0.5, 'verbosi
ty': 0}
R2 Score: 0.9876357066890886

Tuning LightGBM Regression...
Results for LightGBM Regression (Time taken: 3064.85s):
Best estimator: LGBMRegressor(colsample_bytree=0.5, learning_rate=0.01, max_depth=5,
              n_estimators=1000, subsample=0.5, verbosity=-1)
Best params: {'colsample_bytree': 0.5, 'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 1000, 'subsample': 0.5, 'verbosit
y': -1}
R2 Score: 0.9864941617721092

Tuning Support Vector Machines...
Results for Support Vector Machines (Time taken: 3530.30s):
Best estimator: SVR(C=1, kernel='linear')
Best params: {'C': 1, 'gamma': 'scale', 'kernel': 'linear'}
R2 Score: 0.9880753857835839

Results for Voting Regressor:
R2 Score: 0.988461435643853

Results for Stacking Regressor:
R2 Score: 0.9883461553168237
```

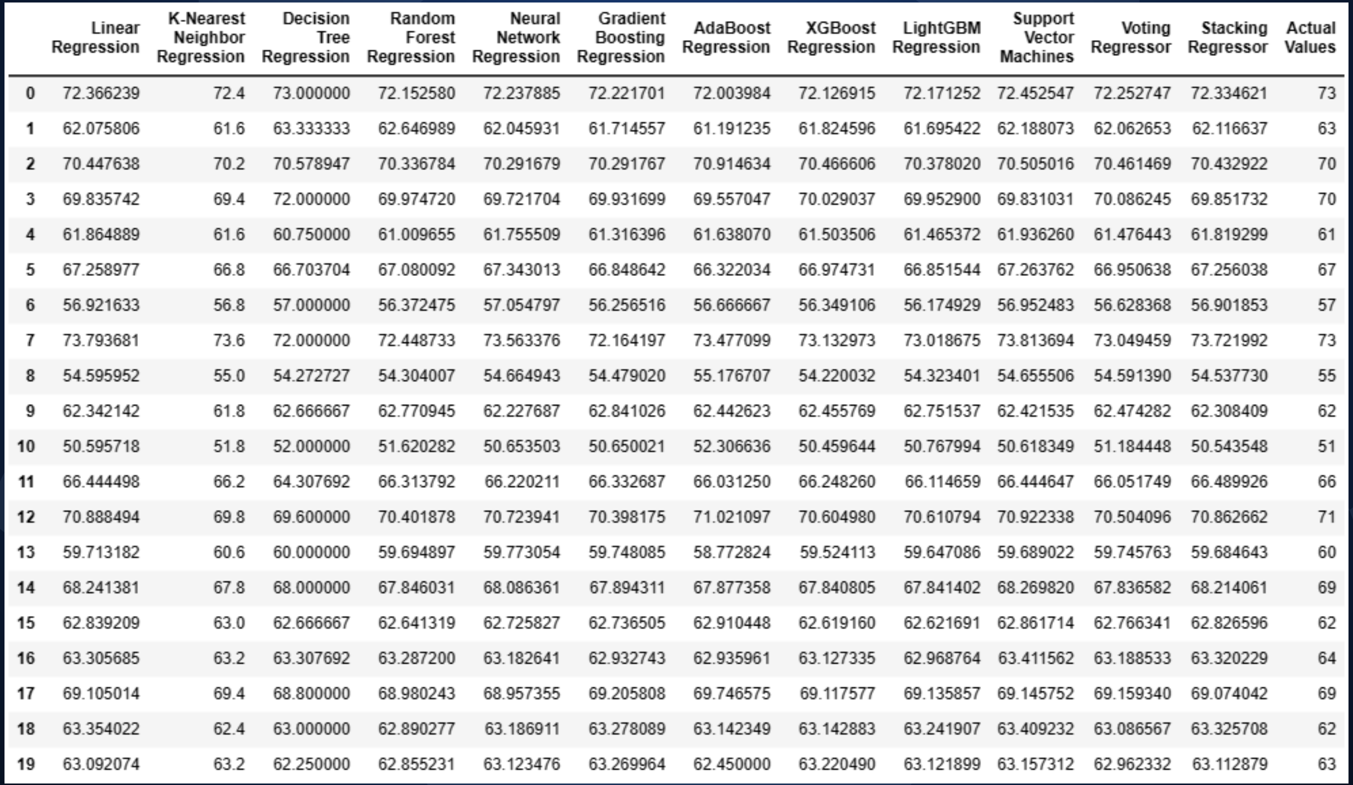
MACHINE LEARNING MODELS for PREDICTION

Predict Overall Ratings for Goalkeepers

1. **Top Performers:**
   - **Linear Regression (R2: 0.9883):** Surprisingly high R2 suggests effectiveness in capturing patterns.
   - **Voting Regressor (R2: 0.9885):** Ensemble method excels, providing high combined predictive performance.
2. **Consistent Performance:**
   - **SVM (R2: 0.9881):** Exceptional predictive capabilities with a linear kernel.
   - **Neural Network (R2: 0.9882):** Competitive performance, particularly with identity activation.
3. **Ensemble Advantage:**
   - **Stacking Regressor (R2: 0.9883):** Diverse model integration enhances overall reliability.
   - **Voting Regressor (R2: 0.9885):** Collective strength leads to superior predictive capability.
4. **Model Robustness:**
   - **Random Forest (R2: 0.9867):** Robust performance with moderate training time.
   - **LightGBM (R2: 0.9865):** Strong predictive ability with optimized hyperparameters.
5. **Fine-Tuned Hyperparameters:**
   - **XGBoost (R2: 0.9876):** Fine-tuned parameters contribute to enhanced accuracy.
   - **LightGBM (R2: 0.9865):** Optimal hyperparameters boost effectiveness.
6. **Computational Considerations:**
   - **SVM (Time: 3530.30s):** Longest tuning time indicates computational intensity, justified by high R2.

MACHINE LEARNING MODELS for PREDICTION

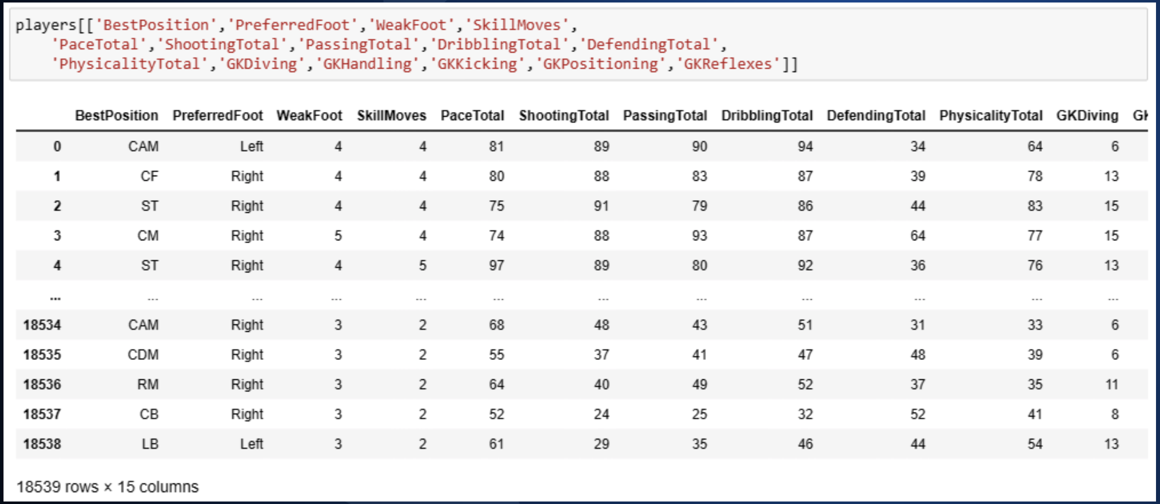
Predict Overall Ratings for Goalkeepers



	Linear Regression	K-Nearest Neighbor Regression	Decision Tree Regression	Random Forest Regression	Neural Network Regression	Gradient Boosting Regression	AdaBoost Regression	XGBoost Regression	LightGBM Regression	Support Vector Machines	Voting Regressor	Stacking Regressor	Actual Values
0	72.366239	72.4	73.000000	72.152580	72.237885	72.221701	72.003984	72.126915	72.171252	72.452547	72.252747	72.334621	73
1	62.075806	61.6	63.333333	62.646989	62.045931	61.714557	61.191235	61.824596	61.695422	62.188073	62.062653	62.116637	63
2	70.447638	70.2	70.578947	70.336784	70.291679	70.291767	70.914634	70.466606	70.378020	70.505016	70.461469	70.432922	70
3	69.835742	69.4	72.000000	69.974720	69.721704	69.931699	69.557047	70.029037	69.952900	69.831031	70.086245	69.851732	70
4	61.864889	61.6	60.750000	61.009655	61.755509	61.316396	61.638070	61.503506	61.465372	61.936260	61.476443	61.819299	61
5	67.258977	66.8	66.703704	67.080092	67.343013	66.848642	66.322034	66.974731	66.851544	67.263762	66.950638	67.256038	67
6	56.921633	56.8	57.000000	56.372475	57.054797	56.256516	56.666667	56.349106	56.174929	56.952483	56.628368	56.901853	57
7	73.793681	73.6	72.000000	72.448733	73.563376	72.164197	73.477099	73.132973	73.018675	73.813694	73.049459	73.721992	73
8	54.595952	55.0	54.272727	54.304007	54.664943	54.479020	55.176707	54.220032	54.323401	54.655506	54.591390	54.537730	55
9	62.342142	61.8	62.666667	62.770945	62.227687	62.841026	62.442623	62.455769	62.751537	62.421535	62.474282	62.308409	62
10	50.595718	51.8	52.000000	51.620282	50.653503	50.650021	52.306636	50.459644	50.767994	50.618349	51.184448	50.543548	51
11	66.444498	66.2	64.307692	66.313792	66.220211	66.332687	66.031250	66.248260	66.114659	66.444647	66.051749	66.489926	66
12	70.888494	69.8	69.600000	70.401878	70.723941	70.398175	71.021097	70.604980	70.610794	70.922338	70.504096	70.862662	71
13	59.713182	60.6	60.000000	59.694897	59.773054	59.748085	58.772824	59.524113	59.647086	59.689022	59.745763	59.684643	60
14	68.241381	67.8	68.000000	67.846031	68.086361	67.894311	67.877358	67.840805	67.841402	68.269820	67.836582	68.214061	69
15	62.839209	63.0	62.666667	62.641319	62.725827	62.736505	62.910448	62.619160	62.621691	62.861714	62.766341	62.826596	62
16	63.305685	63.2	63.307692	63.287200	63.182641	62.932743	62.935961	63.127335	62.968764	63.411562	63.188533	63.320229	64
17	69.105014	69.4	68.800000	68.980243	68.957355	69.205808	69.746575	69.117577	69.135857	69.145752	69.159340	69.074042	69
18	63.354022	62.4	63.000000	62.890277	63.186911	63.278089	63.142349	63.142883	63.241907	63.409232	63.086567	63.325708	62
19	63.092074	63.2	62.250000	62.855231	63.123476	63.269964	62.450000	63.220490	63.121899	63.157312	62.962332	63.112879	63

MACHINE LEARNING MODELS for PREDICTION

Predict Main Positions for Players



```
players[['BestPosition','PreferredFoot','WeakFoot','SkillMoves',
    'PaceTotal','ShootingTotal','PassingTotal','DribblingTotal','DefendingTotal',
    'PhysicalityTotal','GKDiving','GKHandling','GKKicking','GKPositioning','GKReflexes']]
```

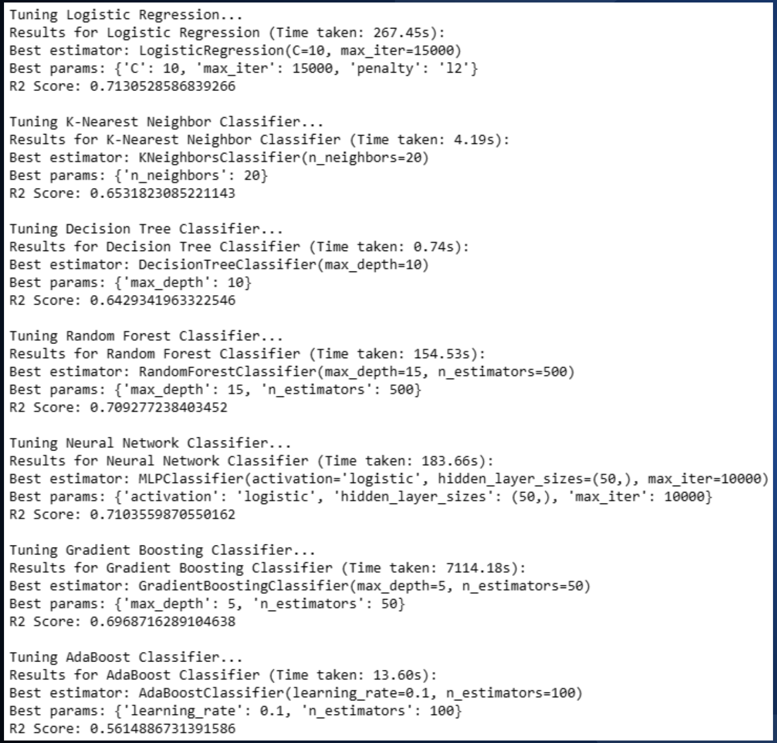
	BestPosition	PreferredFoot	WeakFoot	SkillMoves	PaceTotal	ShootingTotal	PassingTotal	DribblingTotal	DefendingTotal	PhysicalityTotal	GKDiving
0	CAM	Left	4	4	81	89	90	94	34	64	6
1	CF	Right	4	4	80	88	83	87	39	78	13
2	ST	Right	4	4	75	91	79	86	44	83	15
3	CM	Right	5	4	74	88	93	87	64	77	15
4	ST	Right	4	5	97	89	80	92	36	76	13
...	...	...	...	...	...	...	...	...	...	...	...
18534	CAM	Right	3	2	68	48	43	51	31	33	6
18535	CDM	Right	3	2	55	37	41	47	48	39	6
18536	RM	Right	3	2	64	40	49	52	37	35	11
18537	CB	Right	3	2	52	24	25	32	52	41	8
18538	LB	Left	3	2	61	29	35	46	44	54	13

18539 rows × 15 columns

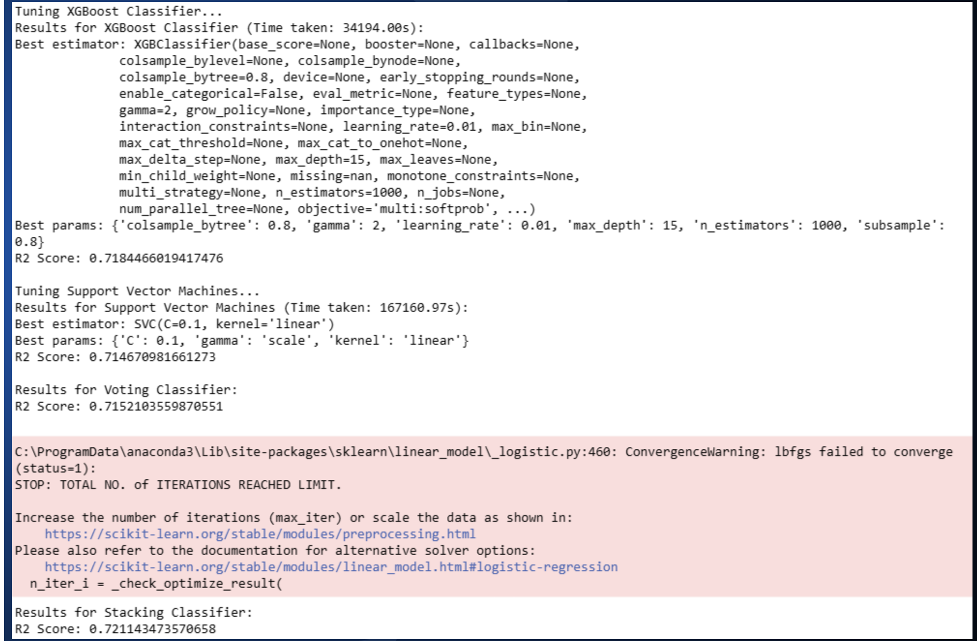
To manipulate the DataFrame 'players' and include all main features used for predicting the main position of a player, we would typically select relevant columns. After applying machine learning models, we will compare predictions to the actual values (BestPosition).

MACHINE LEARNING MODELS for PREDICTION

Predict Main Positions for Players



```
Tuning Logistic Regression...
Results for Logistic Regression (Time taken: 267.45s):
Best estimator: LogisticRegression(C=10, max_iter=15000)
Best params: {'C': 10, 'max_iter': 15000, 'penalty': 'l2'}
R2 Score: 0.7130528586839266

Tuning K-Nearest Neighbor Classifier...
Results for K-Nearest Neighbor Classifier (Time taken: 4.19s):
Best estimator: KNeighborsClassifier(n_neighbors=20)
Best params: {'n_neighbors': 20}
R2 Score: 0.6531823085221143

Tuning Decision Tree Classifier...
Results for Decision Tree Classifier (Time taken: 0.74s):
Best estimator: DecisionTreeClassifier(max_depth=10)
Best params: {'max_depth': 10}
R2 Score: 0.6429341963322546

Tuning Random Forest Classifier...
Results for Random Forest Classifier (Time taken: 154.53s):
Best estimator: RandomForestClassifier(max_depth=15, n_estimators=500)
Best params: {'max_depth': 15, 'n_estimators': 500}
R2 Score: 0.709277238403452

Tuning Neural Network Classifier...
Results for Neural Network Classifier (Time taken: 183.66s):
Best estimator: MLPClassifier(activation='logistic', hidden_layer_sizes=(50,), max_iter=10000)
Best params: {'activation': 'logistic', 'hidden_layer_sizes': (50,), 'max_iter': 10000}
R2 Score: 0.7103559870550162

Tuning Gradient Boosting Classifier...
Results for Gradient Boosting Classifier (Time taken: 7114.18s):
Best estimator: GradientBoostingClassifier(max_depth=5, n_estimators=50)
Best params: {'max_depth': 5, 'n_estimators': 50}
R2 Score: 0.6968716289104638

Tuning AdaBoost Classifier...
Results for AdaBoost Classifier (Time taken: 13.60s):
Best estimator: AdaBoostClassifier(learning_rate=0.1, n_estimators=100)
Best params: {'learning_rate': 0.1, 'n_estimators': 100}
R2 Score: 0.5614886731391586
```



```
Tuning XGBoost Classifier...
Results for XGBoost Classifier (Time taken: 34194.00s):
Best estimator: XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=0.8, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=2, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.01, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=15, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=1000, n_jobs=None,
              num_parallel_tree=None, objective='multi:softprob', ...)
Best params: {'colsample_bytree': 0.8, 'gamma': 2, 'learning_rate': 0.01, 'max_depth': 15, 'n_estimators': 1000, 'subsample':
0.8}
R2 Score: 0.7184466019417476

Tuning Support Vector Machines...
Results for Support Vector Machines (Time taken: 167160.97s):
Best estimator: SVC(C=0.1, kernel='linear')
Best params: {'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}
R2 Score: 0.714670981661273

Results for Voting Classifier:
R2 Score: 0.7152103559870551

C:\ProgramData\anaconda3\Lib\site-packages\sklearn\linear_model\_logistic.py:460: ConvergenceWarning: lbfgs failed to converge
(status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(

Results for Stacking Classifier:
R2 Score: 0.721143473570658
```

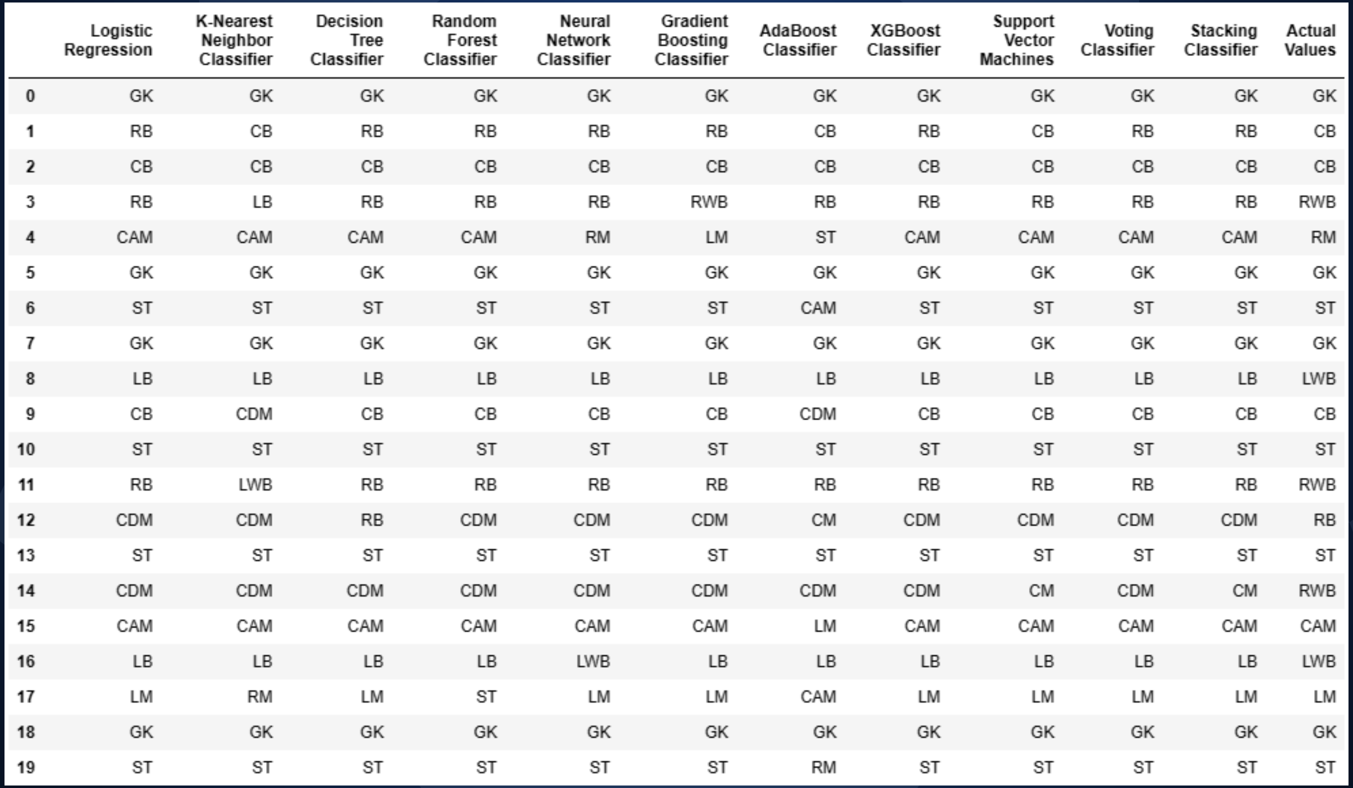
MACHINE LEARNING MODELS for PREDICTION

Predict Main Positions for Players

1. **Top Performing Classifiers:**
   - **Stacking Classifier (R2: 0.7211):** Outperforms individual models, indicating synergy in ensemble methods.
   - **XGBoost Classifier (R2: 0.7184):** Strong performer with optimized hyperparameters.
2. **Model Comparison:**
   - **SVM (R2: 0.7147):** Linear kernel demonstrates competitive performance with reduced computational complexity.
   - **Random Forest Classifier (R2: 0.7093):** Robust model with respectable accuracy and moderate training time.
3. **Fine-Tuned Hyperparameters:**
   - **XGBoost (R2: 0.7184):** Optimal parameters (colsample_bytree=0.8, gamma=2) contribute to enhanced accuracy.
   - **Logistic Regression (R2: 0.7131):** High regularization (C=10) and extensive iterations (max_iter=15000) improve performance.
4. **Computational Trade-offs:**
   - **XGBoost Classifier (Time: 34194.00s):** Longer training time suggests computational intensity, but justified by high R2.
   - **SVM (Time: 167160.97s):** Extended tuning time indicates computational complexity; linear kernel balances performance and efficiency.
5. **Ensemble Strength:**
   - **Stacking Classifier (R2: 0.7211):** Demonstrates the power of combining diverse models for improved predictive capability.
   - **Voting Classifier (R2: 0.7152):** Collective strength of individual models contributes to a solid overall performance.

MACHINE LEARNING MODELS for PREDICTION

Predict Main Positions for Players



	Logistic Regression	K-Nearest Neighbor Classifier	Decision Tree Classifier	Random Forest Classifier	Neural Network Classifier	Gradient Boosting Classifier	AdaBoost Classifier	XGBoost Classifier	Support Vector Machines	Voting Classifier	Stacking Classifier	Actual Values
0	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK
1	RB	CB	RB	RB	RB	RB	CB	RB	CB	RB	RB	CB
2	CB	CB	CB	CB	CB	CB	CB	CB	CB	CB	CB	CB
3	RB	LB	RB	RB	RB	RWB	RB	RB	RB	RB	RB	RWB
4	CAM	CAM	CAM	CAM	RM	LM	ST	CAM	CAM	CAM	CAM	RM
5	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK
6	ST	ST	ST	ST	ST	ST	CAM	ST	ST	ST	ST	ST
7	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK
8	LB	LB	LB	LB	LB	LB	LB	LB	LB	LB	LB	LWB
9	CB	CDM	CB	CB	CB	CB	CDM	CB	CB	CB	CB	CB
10	ST	ST	ST	ST	ST	ST	ST	ST	ST	ST	ST	ST
11	RB	LWB	RB	RB	RB	RB	RB	RB	RB	RB	RB	RWB
12	CDM	CDM	RB	CDM	CDM	CDM	CM	CDM	CDM	CDM	CDM	RB
13	ST	ST	ST	ST	ST	ST	ST	ST	ST	ST	ST	ST
14	CDM	CDM	CDM	CDM	CDM	CDM	CDM	CDM	CM	CDM	CM	RWB
15	CAM	CAM	CAM	CAM	CAM	CAM	LM	CAM	CAM	CAM	CAM	CAM
16	LB	LB	LB	LB	LWB	LB	LB	LB	LB	LB	LB	LWB
17	LM	RM	LM	ST	LM	LM	CAM	LM	LM	LM	LM	LM
18	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK	GK
19	ST	ST	ST	ST	ST	ST	RM	ST	ST	ST	ST	ST

MACHINE LEARNING MODELS for PREDICTION

Summary

1. Feature Selection:

- **Key Attributes:** Identified crucial player attributes ('Total' skills) for robust machine learning models.
- **Focused Features:** BestPosition, Overall, Pace, Shooting, Passing, Dribbling, Defending, Physicality, Goalkeeper Skills, Preferred Foot, and Weak Foot.

2. Machine-learning Workflow:

- **Regression Analysis:** Successfully applied machine-learning workflow for regression to predict Overall Ratings for Players.
- **Classification Tasks:** Extended workflow to predict Main Positions for Players.

3. Model Performance:

- **Top Performers:** Stacking Classifier and XGBoost Classifier showcased superior predictive capabilities.
- **Robust Models:** Random Forest and Neural Network demonstrated commendable accuracy with manageable computational demands.

4. Hyperparameter Optimization:

- **XGBoost Excellence:** Optimized hyperparameters (e.g., colsample_bytree=0.8, gamma=2) significantly enhanced model accuracy.
- **Logistic Regression Precision:** High regularization (C=10) and extensive iterations (max_iter=15000) improved performance.

MACHINE LEARNING MODELS for PREDICTION

Summary

5. **Computational Trade-offs:**
   - **XGBoost Intensity:** Longer training time justified by superior accuracy.
   - **SVM Efficiency:** Linear kernel balances competitive performance with reduced computational complexity.
6. **Ensemble Strengths:**
   - **Synergetic Performance:** Stacking Classifier highlighted the strength of combining diverse models for improved predictions.
   - **Collective Power:** Voting Classifier demonstrated collective strength, contributing to solid overall performance.

► **Key Takeaway:** A meticulously curated machine-learning pipeline, thoughtful feature selection, and hyperparameter tuning lead to robust models with varying strengths. Balancing computational efficiency and model accuracy is crucial, and ensemble methods prove effective in achieving superior predictive capabilities.

CONCLUSION

Team Analyses

- Our exploration into team dynamics unveiled patterns in league performances, emphasizing the dominance of the English Premier League, La Liga, Bundesliga, Serie A, and Ligue 1. Disparities in team values and wages underscored financial hierarchies, shedding light on the competitive landscape of football.

Player Analyses

- Diving into player attributes revealed intricate position-specific skills through charts. The normal distribution of overall player ratings highlighted the spectrum of talent, with outliers signifying exceptional or subpar skills, adding diversity to the sport.

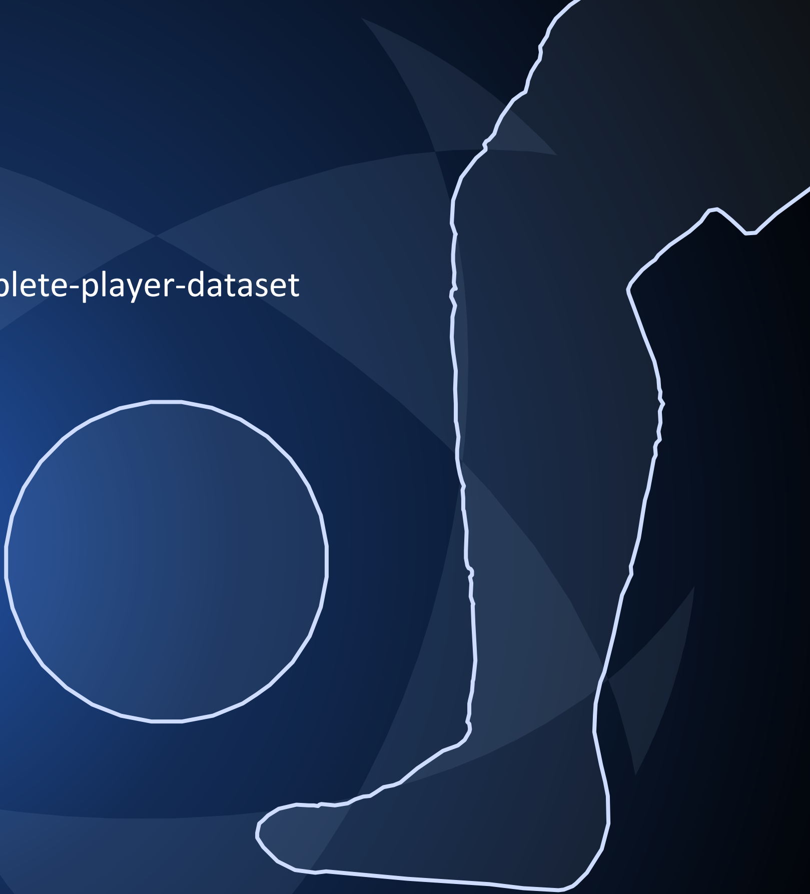
Machine Learning for Prediction

- In the realm of machine learning, we identified key attributes for player prediction, focusing on 'Total' skills. Regression and classification tasks, powered by models like XGBoost and Stacking Classifier, showcased robust performance. Hyperparameter optimization and ensemble methods proved pivotal, striking a balance between computational efficiency and predictive accuracy.

► Our data-driven journey through team, player analyses, and machine learning underscored the intricate dynamics of football. The fusion of analytics and predictive modeling opens new frontiers for understanding and enhancing the beautiful game.

APPENDIX

- **Data Sources**
- https://www.kaggle.com/datasets/cashncarry/fifa-23-complete-player-dataset
- **Tools Used**
- Jupyter Notebooks
- Python Libraries
  - Pandas
  - Matplotlib
  - Seaborn
  - Squarify
  - Scikit-learn
  - XGBoost
  - LightGBM
- **Code Repository**
- https://github.com/hai-t-nguyen/FIFA-2023-Players-Analytics



THANK YOU!