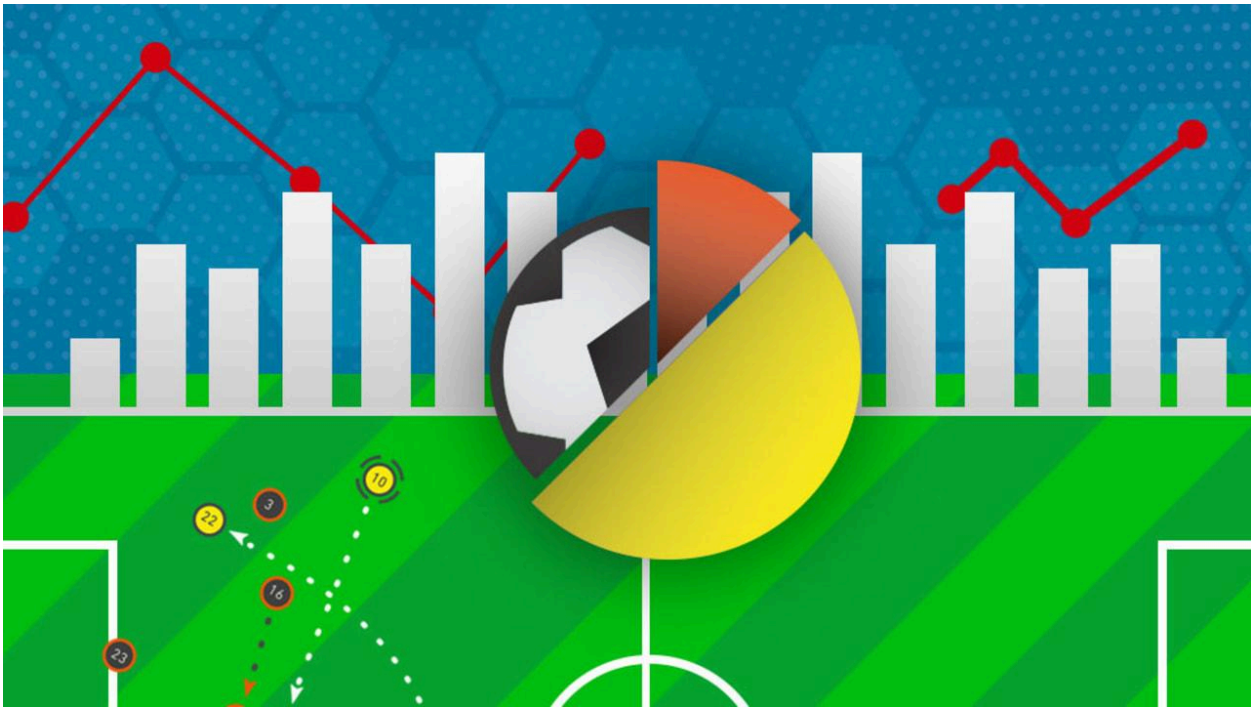
Next-Gen Soccer Analytics



Ethan Homler

MDA 720 Capstone

April 30, 2024

Table of Contents

Background / Objective

In the world of professional sports, especially in a widely watched sport like soccer, the edge afforded by data analytics cannot be overstated. With the development of advanced metrics and comprehensive data collection, the landscape of the sport has evolved into an arena where informed decisions greatly determine success on and off the pitch. This data-driven shift has not only changed how teams evaluate player performance on the pitch but also how they strategize their recruitment and identify talents that help bolster the squad.

This massive change in soccer presents teams with the opportunity to optimize performance metrics and player development. Recognizing this, the project here aims to create a dynamic analytics platform designed to give soccer teams in-depth insights into player statistics and their corresponding salaries. Sports analytics have proved to work in the professional world and it also comes as an economic benefit. For years, professional teams hire scouts to do research on players of their interest. While it benefits to watch prospective talents with their own eyes, teams spend a lot of money and resources on these scouts to find the right piece to their puzzle when they can get the information they need about players through their computer screen. Sports analytics is arguably a more cost-effective way for teams to study their squad player's performances and identify potential new suitors for their squads.

This project is an exploratory dive into the measurable elements of soccer, combining thorough data analysis with strategic management. The goal is to offer soccer clubs with advanced analytics tools, ensuring their decisions are grounded in data, not just relying on intuition.

Data Extraction/Collection/Scraping

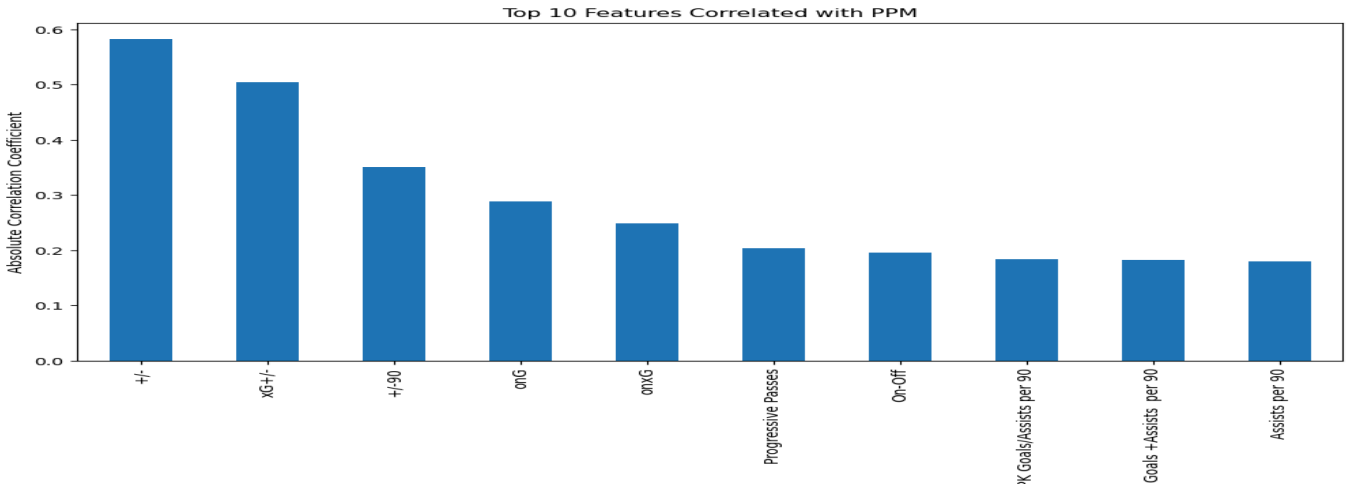
For the project, I wanted to manufacture a large database that contained as much useful, up-to-date statistics on players. Many platforms that have this information are owned by private companies. After thorough research, I landed on fbref.com. Their website contains valuable and updated open source data on professional players across all of Europe and the US. To create the database with in depth player statistics and salary data, I used https://fbref.com/en/comps/9/Premier-League-Stats. For the purpose of this project, I used a dataset from this website that contained the general statistics of all the English Premier League players of the 2023-2024 season. Some of the statistics included name, age, team, goals, assists,

minutes/matches played, disciplinary issues, and expected performance/output. These are some of the essential numbers needed to evaluate a player's performance. However, the numbers from this dataset do not give a complete picture to a player's evaluation. To create a more holistic database. I utilized the tool of web scraping to gather more information about the players in the Premier League. I web scraped three more datasets from the same website, fbref.com. The three additional datasets I used were Premier League Player Miscellaneous Stats, Premier League Player Playing Time, and Premier League Player Wages.

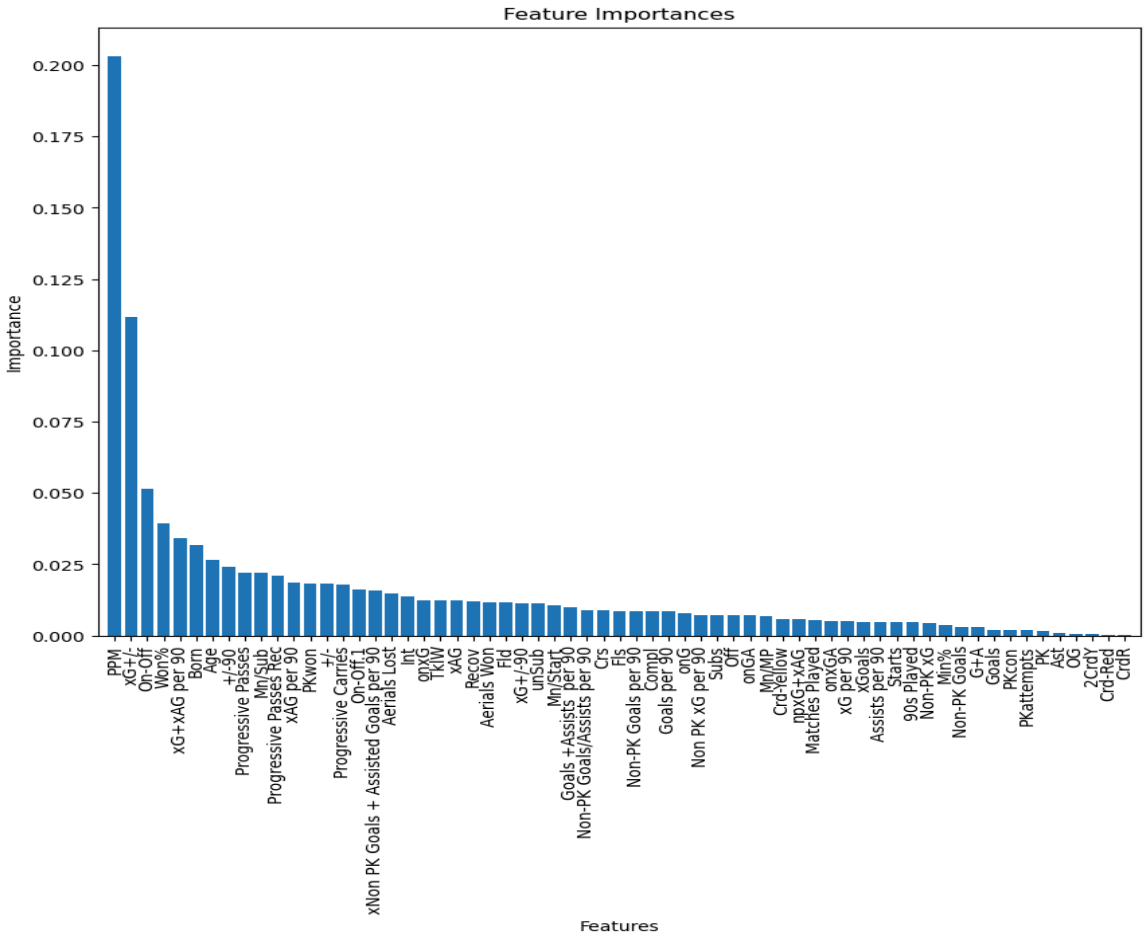
To combine the datasets together into one dataset, I first imported requests and BeautifulSoup into python, retrieved the urls, and then merged the data frames together by using the player name as the common identifier. The four datasets combined gave me an in-depth database with 555 players and 70 different features. Obviously combining four datasets together meant there was a lot of data cleaning to be done. To clean the data, several steps were performed. Duplicate columns (features) were identified and removed like age and team. Missing values were commonly found in the dataset mainly due to the fact that some players in the database haven't played any of the matches this season and did not record any stats. Depending on the context and significance of the data, missing values were either filled with a specified value, such as the mean or median, or the rows containing them were entirely dropped from the dataset. In addition, features that were deemed not essential to the player analysis database were filtered out, leaving only the features that served a purpose in evaluating players. The last part of the data cleaning process that was performed was converting some of the data types of several different features. Some of the features were converted from string to numeric due to the commas in place separating the thousands

General Data Exploration / Visualization

After the web scraping and data cleaning process began the exploration process of the large Premier League Player database that was constructed. Some of the libraries used in the exploration portion of the analysis were sklearn, matplotlib, and seaborn. One of the features of the dataset that stood out was PPM (Points Per Match). This statistic is essential for player analysis because it basically indicates how successful the team is when the player is on or off the field. A correlation matrix turned bar chart was created to find out the most important stats of the correlation with PPM.



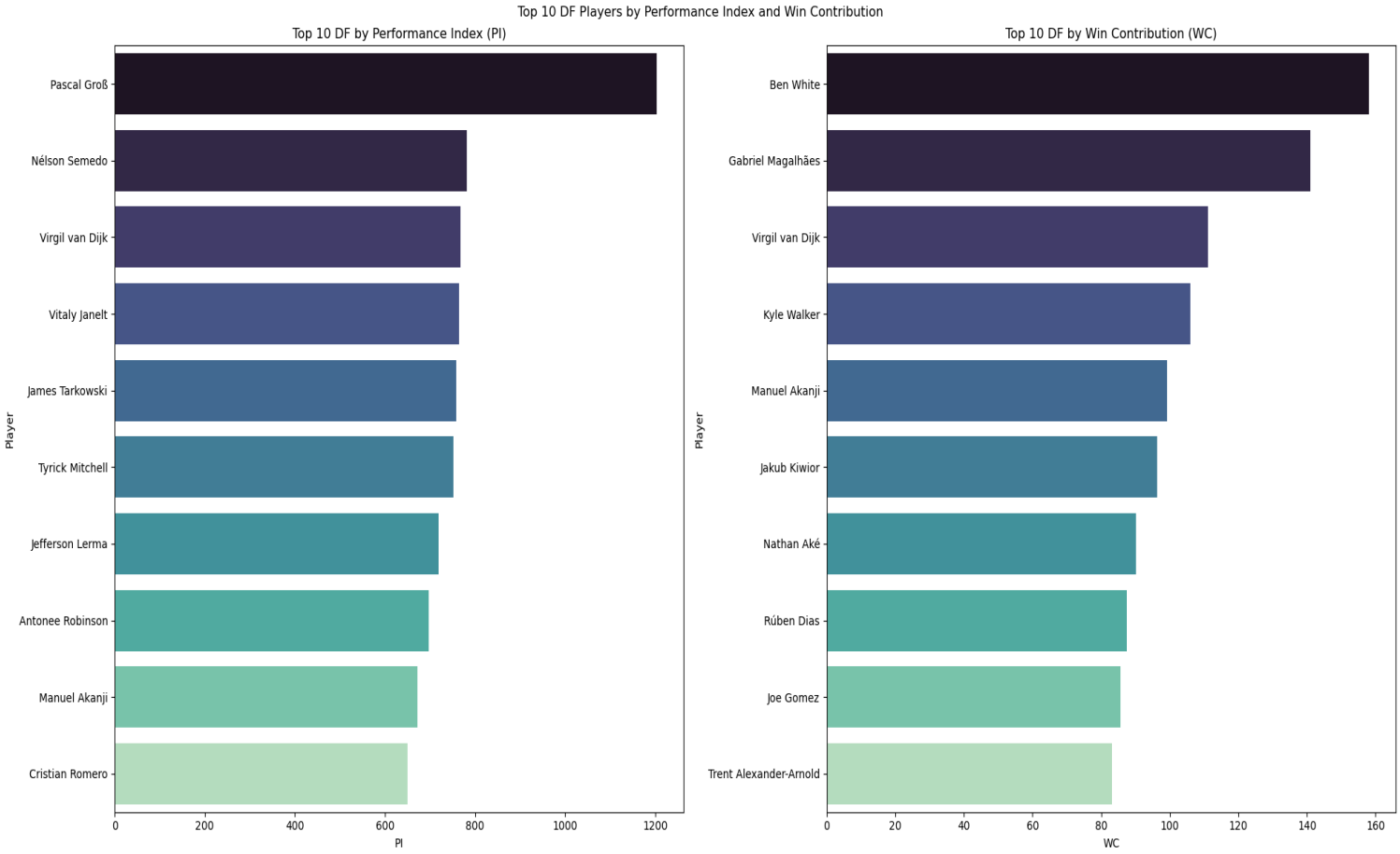
Based on the findings, the two highest correlated statistics to PPM are '+/-' (Goals scored minus goals allowed by the team while the player was on the pitch) and 'xG +/-' (Expected goals scored minus expected goals allowed by the team while the player was on the pitch per 90 minutes played.) Continuing with feature importances, one of the biggest reasons why teams are using data analytics is to understand player value and why they are worth what they are worth. Using the comprehensive database, another feature importance chart was created using 'Weekly Wages' as the target variable. This will show which stats give players their value and can help teams figure out how to value their own players.

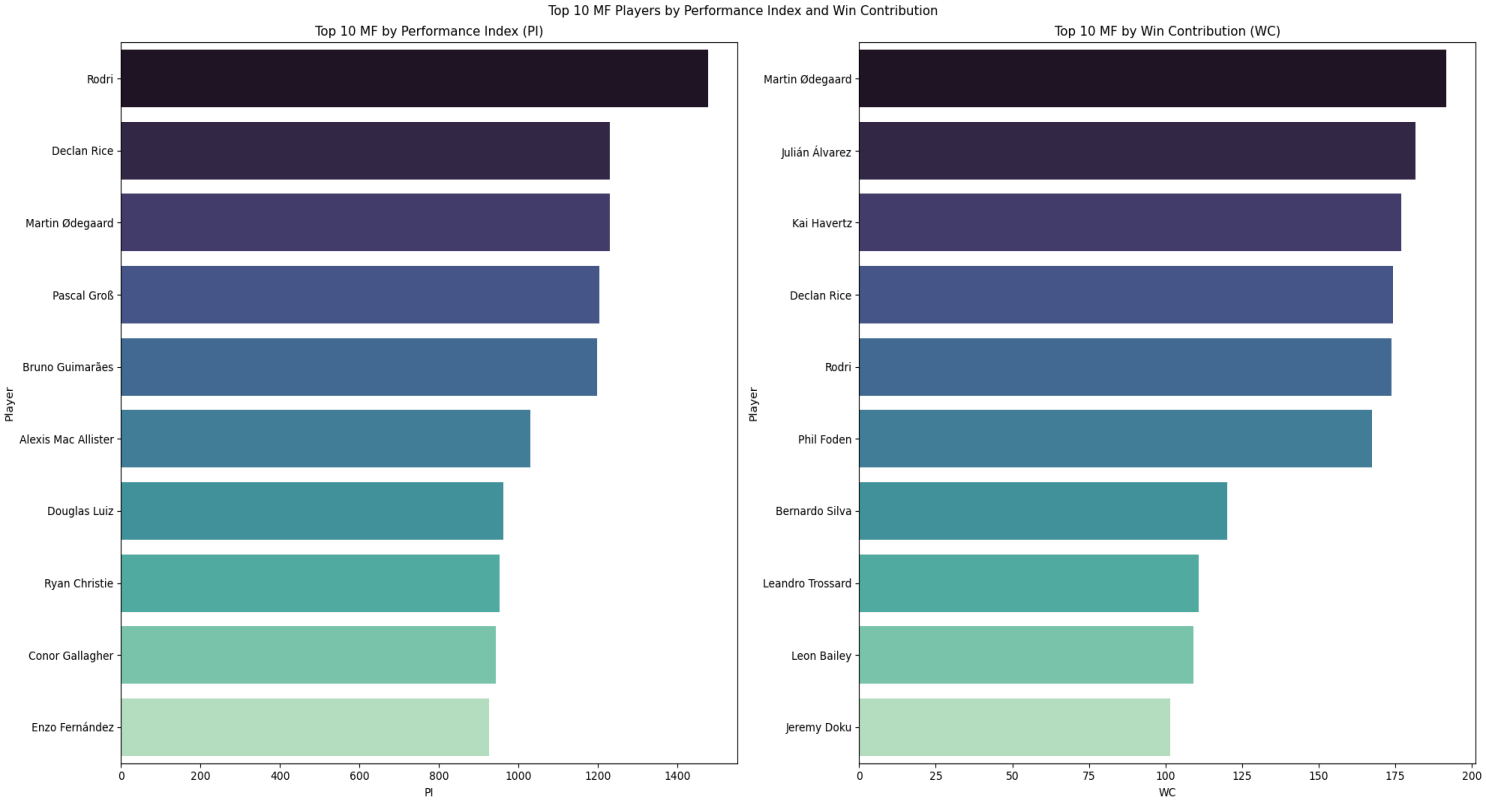


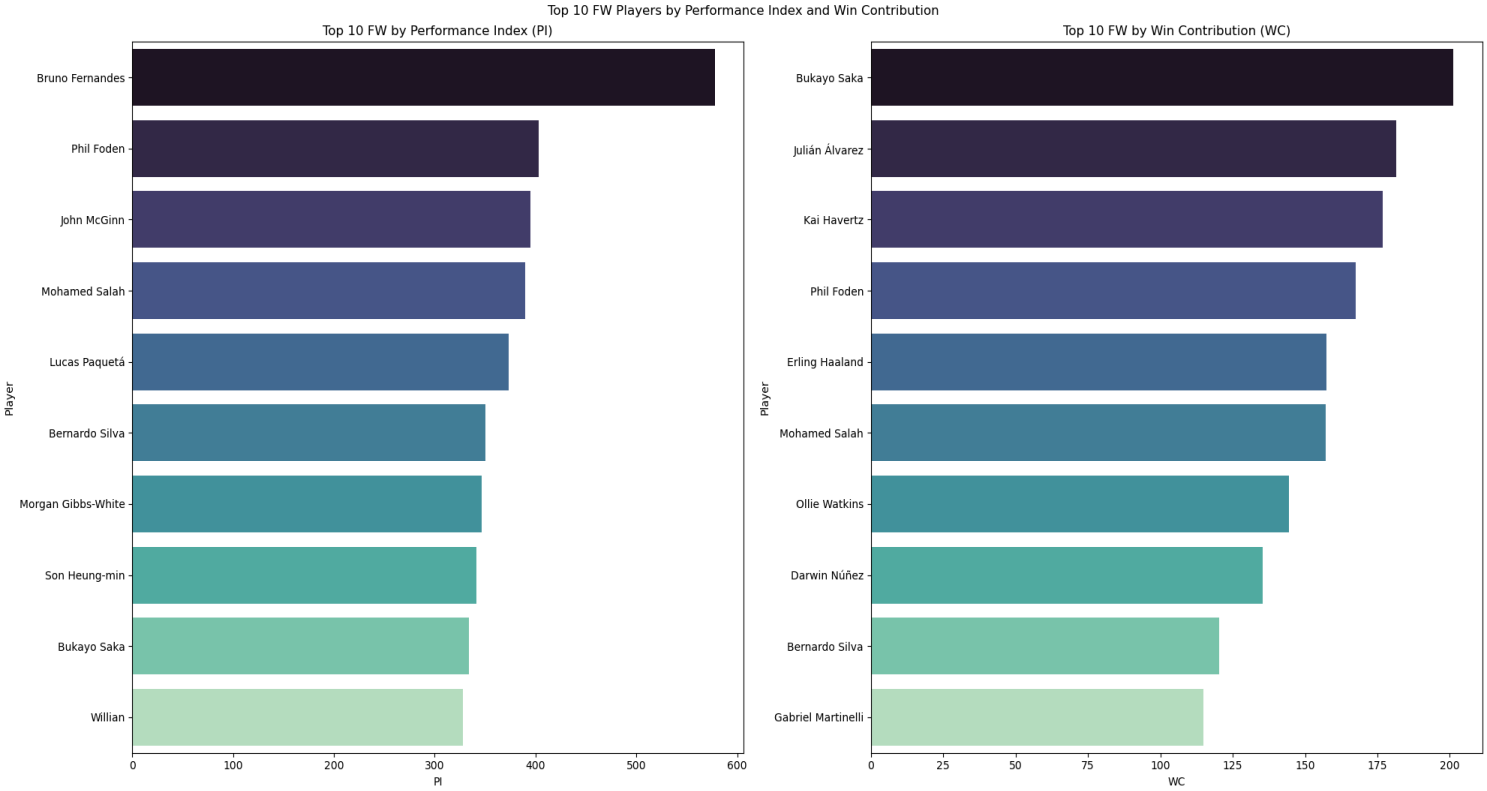
The findings show that 'xG +/-', 'PPM' and 'On-Off ' (Net goals per 90 minutes by the team while the player was on the pitch minus net goals allowed per 90 minutes by the team while the player was off the pitch) are the strongest features correlated to weekly wages. Based on the

results, while individual contributions are important for a player's value, team success seems to be the prominent factor in determining the worth of a player.

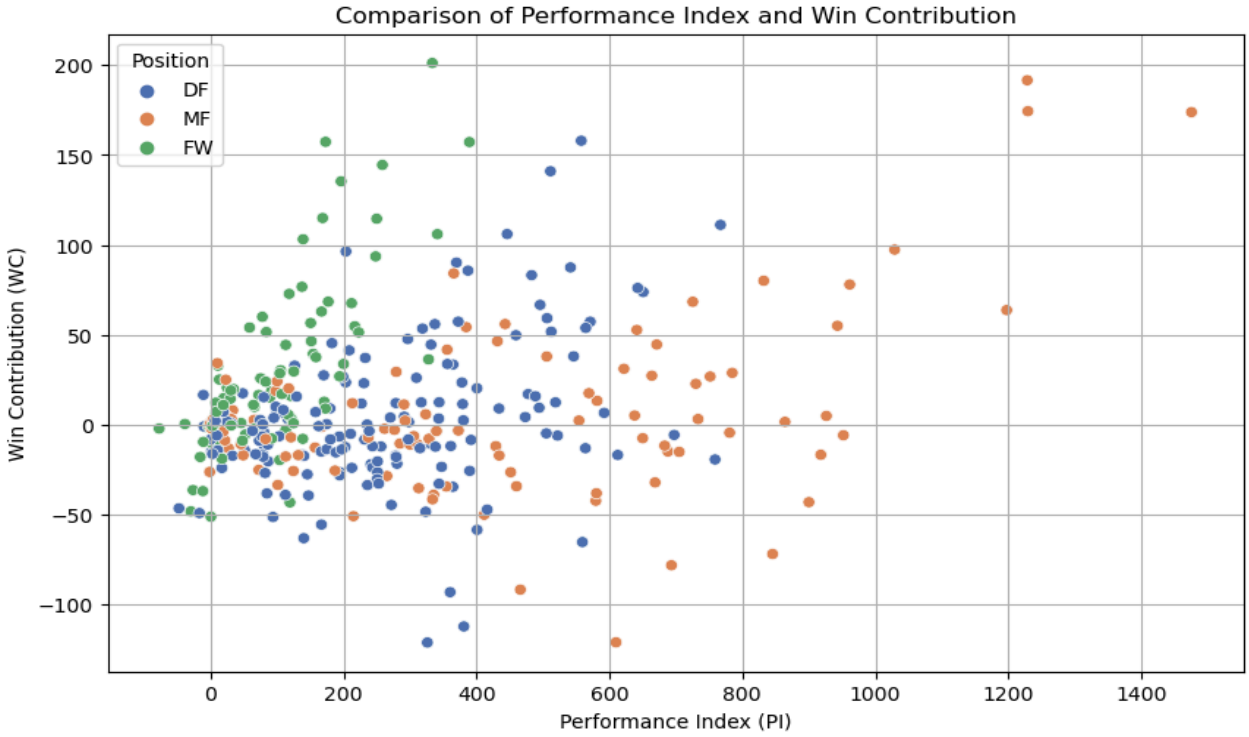
While having a large database with lots of statistics can be useful in itself, understanding the sport and the importance of the features can help unlock so much more in data. To make this database stand out, two different metrics were created: Player Performance Index (PI) and Win Contribution (WC). The Performance Index was constructed by normalizing, combining several key performance metrics for each position meaning a player's index for based on his player specific position, and then weighting the importance of each metric. What this does is combine some of the most important metrics for each position and gives it a score. For forwards, it combined goals, assists, xG +/- 90 (highly correlated to value), and progressive passes. For midfielders it was goals, assists, xG +/- 90, progressive passes, recoveries, and tackles won. Finally for defenders, the metrics were tackles won, interceptions, progressive passes, aerial duels won, and xG +/- 90. Win Contribution (WC) is a metric designed to calculate the direct and indirect contributions of players to their team's performance. It combines various metrics that influence game outcomes into a single, comprehensive measure. This holistic approach provides a unique assessment of a player's impact on their team's success. To calculate, it uses plus/minus score, on-off impact, xG plus/minus, goals, and assists—each weighted to reflect its importance in influencing game outcomes. To put it into action, a visual was made of the top 10 players in each position based on their Performance Index and Win Contribution.




Top 10 MF Players by Performance Index and Win Contribution
Top 10 MF by Performance Index (PI)
Rodri
Declan Rice
Martin Ødegaard
Pascal Groß
Bruno Guimarães
Alexis Mac Allister
Douglas Luiz
Ryan Christie
Conor Gallagher
Enzo Fernández
Player
0
200
400
600
800
1000
1200
1400
PI
Top 10 MF by Win Contribution (WC)
Martin Ødegaard
Julián Álvarez
Kai Havertz
Declan Rice
Rodri
Phil Foden
Bernardo Silva
Leandro Trossard
Leon Bailey
Jeremy Doku
Player
0
25
50
75
100
125
150
175
200
WC

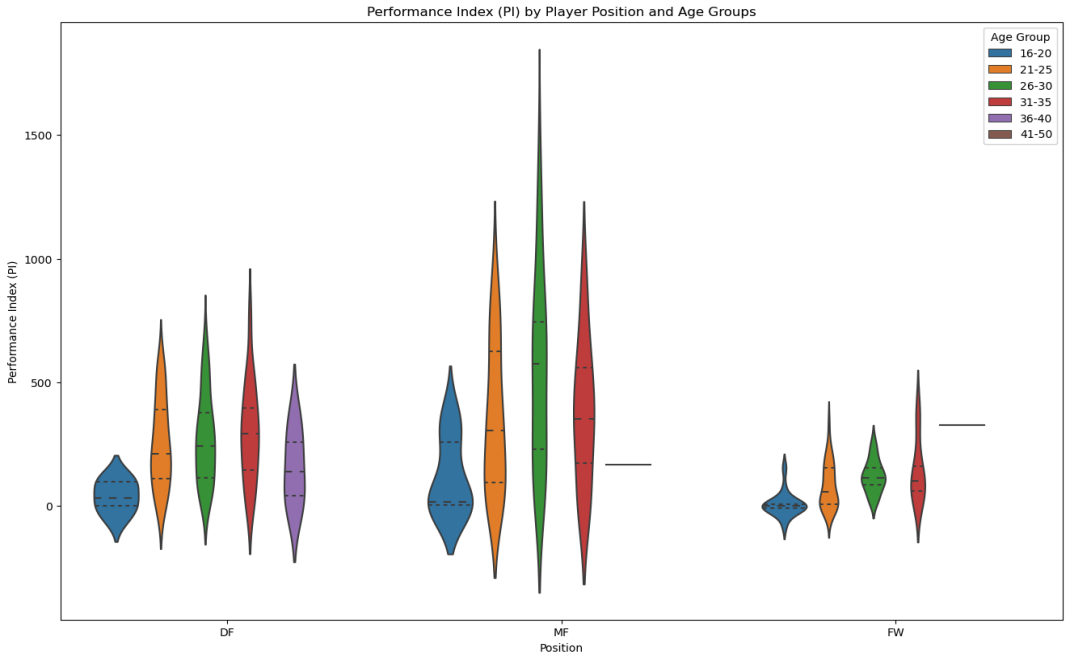

Top 10 FW Players by Performance Index and Win Contribution
Top 10 FW by Performance Index (PI)
Bruno Fernandes
Phil Foden
John McGinn
Mohamed Salah
Lucas Paquetá
Bernardo Silva
Morgan Gibbs-White
Son Heung-min
Bukayo Saka
Willian
Player
0
100
200
300
400
500
600
PI
Top 10 FW by Win Contribution (WC)
Bukayo Saka
Julián Álvarez
Kai Havertz
Phil Foden
Erling Haaland
Mohamed Salah
Ollie Watkins
Darwin Núñez
Bernardo Silva
Gabriel Martinelli
Player
0
25
50
75
100
125
150
175
200
WC

After completing the top 10 charts for each position for PI and WC, an analysis of the relationship between PI and WC was created with players categorized by position.



Based on the results of the scatter plot shown above, there is a trend where higher PI corresponds with higher WC, suggesting that players with better performance metrics tend to contribute more positively to their teams' success. In terms of positions, forwards often have higher PI scores, reflecting their roles in scoring and assisting, crucial for high PI ratings. Their WC scores are also generally positive but show significant variability. Midfielders show to be spread across the whole range of both the PI and WC scale and defenders show a general cluster in the lower to middle range of the PI scale and display a wide range of WC, from slightly negative to moderately positive.

Continuing on with the analysis, a violin plot was created that captured the relationship with Performance Index by Player Position and age groups. This plot branches off the scatter plot from before as it now is for teams to see what the general optimal performance ages are for each position. In terms of this plot, the ages were put into groups: 16-20, 21-25, 26-30, 31-35, 36-40, and 41-50.

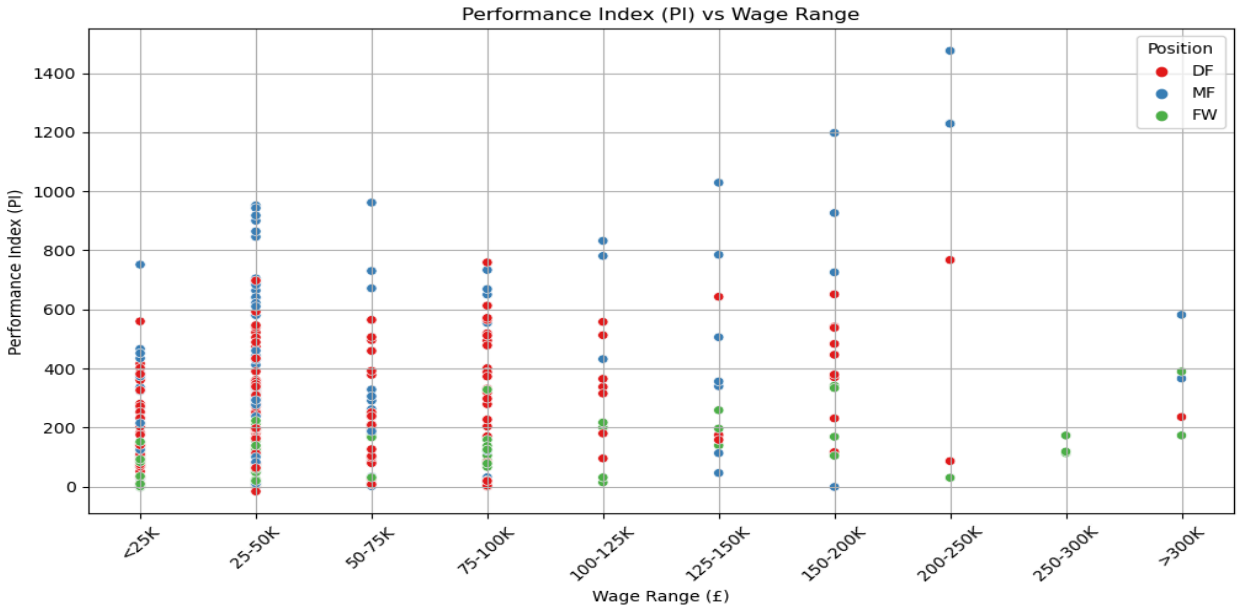


Based on the results of the violin plot, young players (16-20) across all positions generally have lower median PIs, which might be expected due to their lesser experience and adjustment period in professional soccer. Peak performance tends to fall in the 21-25 and 26-30 age groups. These players tend to show higher median PIs. In terms of declining performance indices, there is a visible decline in performance starting from the 31-35 age group for all positions, with narrower distributions and lower medians as age increases. This decline is steeper for forwards, suggesting that their peak performance period might be shorter compared to midfielders and defenders.

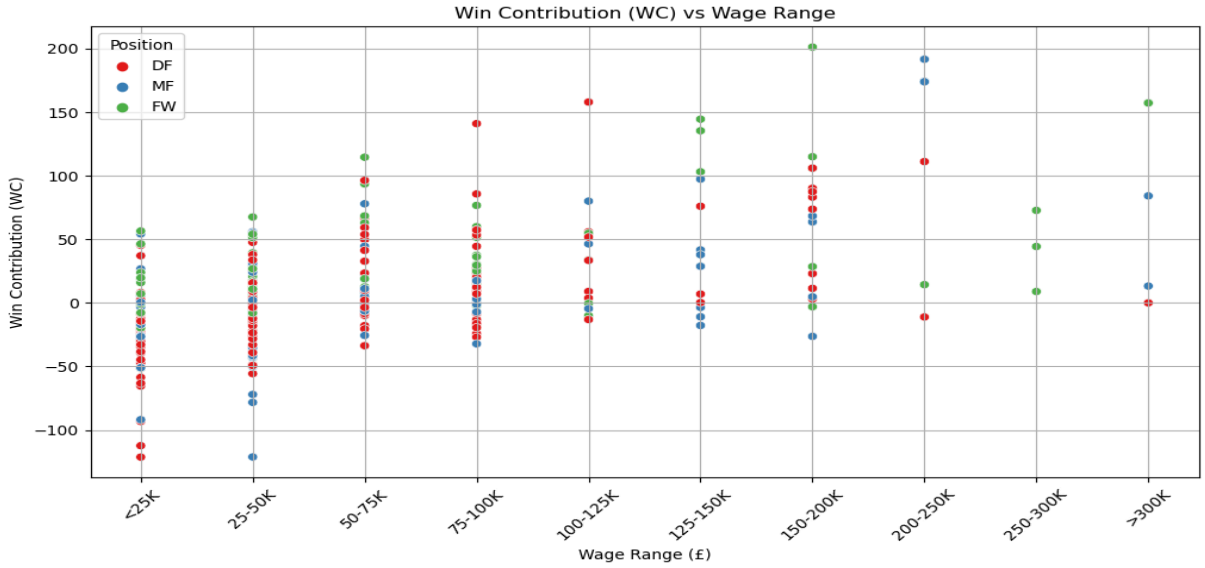
Performance and Wage Data Analysis

After conducting an initial exploratory analysis to understand the performance statistics and their relationships with each other, the project shifted focus towards integrating wage data with performance metrics. Including the financial aspect in player analysis is essential for data-driven decision-making in professional soccer teams. This approach allows teams to evaluate the cost-effectiveness of players in relation to their on-field contributions, aligning financial investment with player performance. This performance x wage analysis not only helps in optimizing team budgets but also provides insights into the economic dynamics of player contracts, enabling teams to make informed decisions about player retention, acquisitions, and salary negotiations.

The first part in this analysis was creating a scatter plot that showed the relationship between Performance Index and Weekly Wage as well as Win Contribution and Weekly Wage.

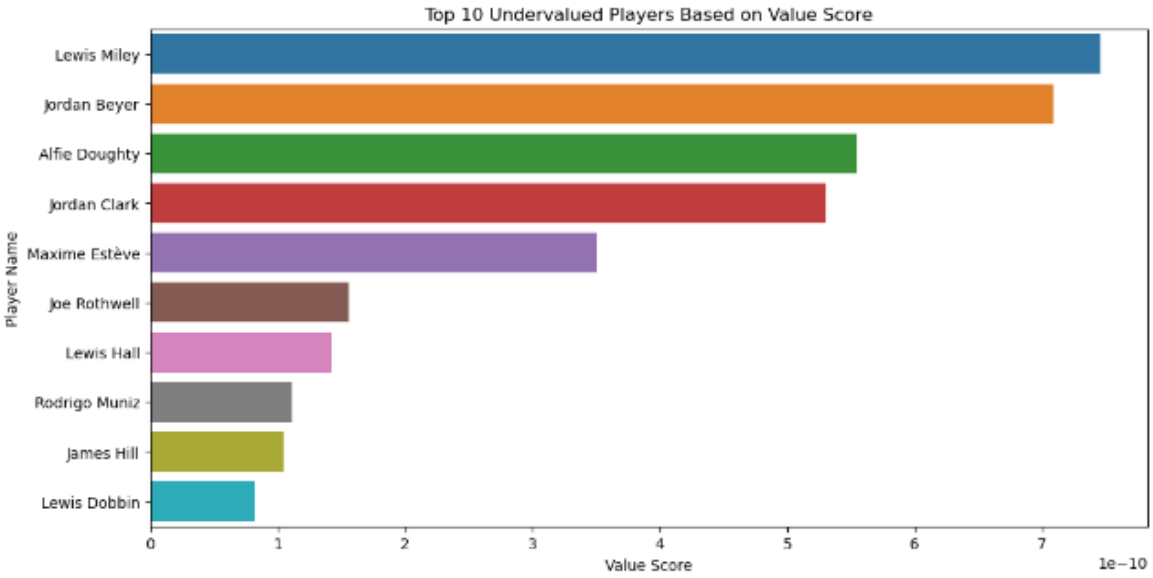


The result of the plot shows a diverse spread of performance indices across wage brackets for defenders, midfielders, and forwards. There appears to be a trend where players with higher wages generally have higher PIs, although there are notable exceptions, particularly in lower wage brackets where some players exhibit high PIs. This could be beneficial for teams who are looking to identify cost-effective, high performing players. Defenders tend to have a lower range of performance index values across all wage ranges, suggesting a more consistent valuation based on performance. The overall results show higher wages do not always correlate with greater impact on the field.



The second plot examines the Win Contribution across the same wage brackets and positions. Similar to the PI, there is a noticeable variation in WC among players of higher wage brackets. This indicates that higher wages do not always correlate with a proportionately higher impact on team success. You can also see again the midfield and forward players dominate the high wage brackets.

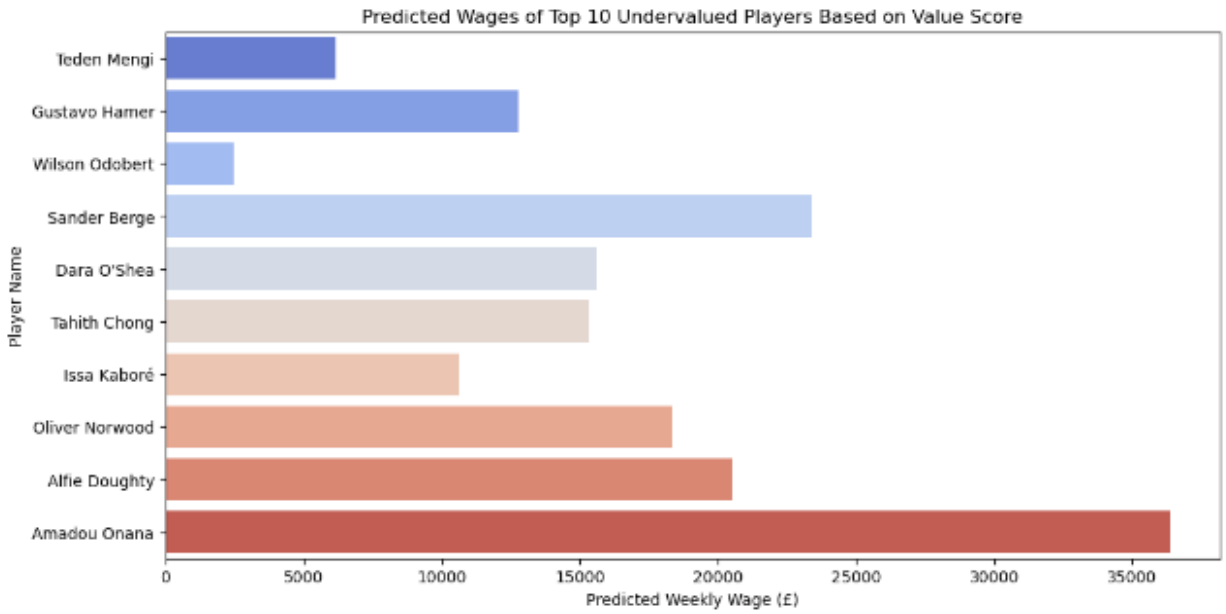
Another part of this analysis was creating a formula to help teams identify cost-efficient talent based on the stats provided and calculations created. To achieve this I created a new metric called 'value score' which took a player's PI and divided it by the weekly wage. This metric basically finds the top players who are performing at levels above what their weekly wage suggests they'd be playing at.



This is the resulting visual of the top 10 undervalued players in the premier league this season. Following this, a wage prediction function was created where a player's name can be manually entered and their predicted wage would be calculated. This gives teams the ability to properly pay players based on their performance metrics. For instance, Lewis Miley's predicted wage was calculated since he's the top undervalued player. The prediction was made from using a train/test split and using sklearn - linear regressor.

```
The predicted weekly wage for Lewis Miley is £33,871.60
```

After creating a prediction function, a visual of predicted wages of the top 10 undervalued players based on their value score was created.



The ability to predict the wages of undervalued players is a true asset for professional soccer teams. By accurately forecasting the potential wages of players who are currently undervalued, teams can secure top talent at a lower cost before their market value aligns with their on-field contributions. This strategic management not only allows teams to strengthen their roster within existing budget constraints but also provides a competitive edge by optimizing resource allocation. Also, identifying and investing in undervalued players can lead to major long-term benefits, like more leverage power in wage negotiations, higher returns on player sales, and the development of a dynamic team that can outperform teams with more money in their pockets.

Conclusion

This capstone project has successfully proven the impact of combining data analysis with advanced analytical tools to enhance decision-making processes within professional soccer teams. Throughout the project, a combined in depth database including performance metrics, playing time, and wages was scraped, compiled and analyzed, revealing insights into the relationships between player performance, their contributions to team success, and their financial valuation. The analysis began by constructing a raw dataset through web scraping, followed by rigorous data cleaning to eliminate any redundancy. Exploratory data analysis provided a deeper understanding of the performance metrics that influence both player and team performance. The

introduction of calculated metrics such as the Performance Index (PI) and Win Contribution (WC) further developed the analysis, allowing for a more detailed assessment of player impact beyond basic stats. A large portion of the project was integrating player wage data with performance metrics to assess the true value of players. The combination of performance metrics and wages sometimes showed a disproportionate relationship between wages and on-field contributions, which proved lower market teams can gain a competitive advantage by identifying undervalued talents. The predictive model developed to estimate player wages based on their performance metrics represents a forward-thinking approach to player management in soccer. This project has shown the use of advanced analytics in soccer, providing tools and methodologies that can be adapted and expanded upon by scouts and managers worldwide.

Bibliography/References/Works Cited

Premier League Standard Stats URL: https://fbref.com/en/comps/9/stats/Premier-League-Stats

Premier League Wage Stats URL: https://fbref.com/en/comps/9/wages/Premier-League-Wages

Premier League Playing Time Stats URL:
https://fbref.com/en/comps/9/playingtime/Premier-League-Stats

Europe's Top 5 Leagues Miscellaneous Stats URL:
https://fbref.com/en/comps/Big5/misc/players/Big-5-European-Leagues-Stats