Final Project Report

What determine Soccer Players' Wages

The Case of the English Premier League (EPL)

by

Amin Kouh Bor

B.A., University of Yazd 2004

M.A., University of Shahid Beheshty, 2008

A Project Report Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Engineering

in the Department of Electrical and Computer Engineering

University of Victoria

Master's Final Project

What determine Soccer Players' Wages

The Case of the English Premier League (EPL)

by

Amin Kouh Bor

B.A., University of Yazd 2004

M.A., University of Shahid Beheshty, 2008

Supervisory Committee

Dr. Amirali Baniasadi, Supervisor

Department of Electrical and Computer Engineering

Dr. Kin Fun Li, Committee Member

Department of Electrical and Computer Engineering

Abstract

This study explores the determinants of wage structures in the English Premier League (EPL), one of the world's most financially influential soccer leagues. Examining both traditional performance metrics—such as goals scored, assists, and minutes played—and emerging non-performance factors, including nationality, marketability, and club revenue, this study seeks to identify how these variables affect wage disparities among EPL players. After data cleaning, utilizing a dataset of 3,522 player-seasons across 31 teams- that attended at least on EPL season- spanning seven EPL seasons, this research employs machine learning techniques to predict wages based on a comprehensive range of performance and demographic factors. The findings suggest that while performance metrics remain crucial, non-performance factors, particularly nationality and Age play an increasingly prominent role in wage determination, contributing to sustained wage inequalities. Additionally, advanced models indicate that club financial resources, player positional roles, and club's broadcast income significantly influence wage variability. This analysis provides an understanding Towards the economic and social forces shaping player wages and underscores the need for a balanced approach to wage determination to support both financial sustainability and equity in professional soccer.

Tables of Contents

Chapter1 - Research Introduction

1.1 Introduction

Soccer has evolved into a multi-billion-pound global entertainment industry, with the English Premier League (EPL) positioned at the forefront of this transformation. As one of the wealthiest soccer leagues globally, the EPL attracts top talent from around the world, which has contributed to significant wage growth among its players. The resulting wage disparities have sparked considerable interest among researchers, particularly as average salaries continue to rise, raising important questions regarding the underlying determinants of wage inequalities in the league.

Traditionally, player salaries have been influenced by performance metrics such as goals scored, assists, tackles made, clean sheets, and minutes played. However, recent research suggests that these factors only partially explain wage variations. Emerging determinants—including nationality, marketability, and subjective traits like physical attractiveness—are increasingly recognized as influential in shaping wage patterns within professional soccer.

Historically, a player's position on the field has also served as a crucial determinant of wages within the EPL and other top European leagues. Strikers and attacking midfielders, whose roles emphasize scoring and assisting, typically command higher salaries than defenders and goalkeepers. This positional wage disparity reflects the market's preference for players contributing directly to scoring, a valued asset for teams.

Performance-based metrics remain essential; for example, Yaldo and Shamir (2017), using machine learning to analyze the skills-wage relationship, found that attributes such as finishing ability, ball control, and positional awareness correlate with higher salaries. Nonetheless, as the EPL has expanded into a global entertainment brand, additional factors such as nationality are believed to play a significant role in wage determination. Players from South America, particularly from Brazil and Argentina, are expected to command wage premiums not only due to technical skills but also owing to their fan appeal, which boosts revenue through merchandise sales and increased game attendance.

1.2. The Influence of Marketability and Non-Performance-Based Factors

In addition to nationality, recent studies on wage structures in professional soccer have highlighted other off-field factors, notably the physical attractiveness of players, often referred to as the "beauty premium." Evidence suggests that players perceived as more attractive can command higher wages, even when their on-field performance is comparable to that of less attractive counterparts.

This trend emphasizes the growing importance of marketability in modern soccer, where each player's brand value contributes to the club's appeal and ability to attract sponsorships. Such marketability has become a critical lever in wage negotiations, as players' image and fan engagement increasingly influence salary discussions. Furthermore, team dynamics and negotiation strategies significantly shape wage structures within the EPL. Prestigious or historically successful teams, for instance, can leverage their reputation to attract skilled players at comparatively lower wages, using their status as a bargaining tool.

1.3 Research Question

In the complex context of wage determination within professional soccer, this paper examines the range of factors influencing player wages in the English Premier League (EPL), encompassing both traditional performance metrics and emerging non-performance variables. Specifically, this research seeks to address the following question: To what extent do non-performance-based factors, such as nationality, age, and club structure, contribute to wage inequality among players in the EPL?

1.4 Objectives and Scope of the Study

This study aims to comprehensively assess wage determinants in the EPL by analyzing both performance and non-performance factors, filling an existing knowledge gap in the field. Specifically, the objectives are to:

- Examine traditional performance metrics (e.g., goals scored, assists, minutes played) and determine their effect on player wages.
- Investigate non-performance-based factors, such as nationality, marketability, and physical attractiveness, to understand their contributions to wage disparities.
- Utilize machine learning techniques to accurately predict wages and identify key wage determinants.
- Provide insights into wage structure sustainability and equity within the EPL.

The scope of this study is limited to the EPL, yet it includes a robust sample of 3787 player-seasons across seven seasons, providing a broad perspective on wage determinants within one of the world's most influential soccer leagues.

1.5 Importance

This study builds on the work of Yaldo and Shamir (2017), who underscored the importance of performance metrics in determining player wages while also highlighting the rising influence of external factors. By examining both performance-based and non-performance-based variables, this research aims to provide a comprehensive understanding of wage determinants in the English Premier League. In doing so, it offers valuable insights into the economic and social forces shaping wage structures within one of the world's most lucrative sports leagues.

1.6 Thesis Structure

This thesis is organized into six main chapters. Chapter 1 introduces the study's context, research question, objectives, and importance. Chapter 2 reviews relevant literature, presenting key studies on both performance and non-performance wage determinants in soccer. Chapter 3 describes the dataset, data collection methods, and data preprocessing steps, followed by Chapter 4, which details the methodology, including machine learning techniques applied to analyze the data. Chapter 5 presents the findings and discusses the implications of wage determinants within the EPL. Finally, Chapter 6 offers a summary of the results, limitations of the study, and recommendations for future research.

1.7 Link to Broader Implications

Beyond the EPL, this research holds broader implications for understanding wage structures in professional sports, particularly as other major leagues grapple with similar issues surrounding performance metrics, marketability, and nationality. Insights from this study may inform wage policies and strategies for improving equity across leagues, while also providing a framework for machine learning applications in sports analytics. Furthermore, by examining socio-economic factors influencing player wages, this research contributes to the ongoing discussion around income inequality and labor dynamics in global sports industries.

Chapter 2 - Theoretical Framework and Literature

Over the past few decades, professional soccer has undergone extraordinary economic growth, evolving into a global industry with annual revenue exceeding US$33 billion. However, this financial expansion has not uniformly translated to financial integrity or operational efficiency within the sport. Scholars have extensively explored wage structures, transfer fees, and financial sustainability in soccer, identifying both challenges and opportunities in wage management and performance outcomes.

In their examination of soccer's financial vulnerabilities, Andrews and Harrington highlight player wages and transfer fees as major cost drivers for clubs, often exceeding 60 percent of club revenues. This imbalance creates financial strain, especially for elite clubs engaging in high-value transfers, which can jeopardize financial sustainability as clubs struggle to balance operational costs and income. Other significant operational costs—including personnel salaries, travel expenses, event management, and facility upkeep—exacerbate this financial burden.

Research on the English Premier League (EPL) reveals a complex relationship between spending, performance, and profitability. Burdekin and Franklin (2015) demonstrate that the EPL's financial landscape has shifted since its inception, with transfer spending largely driven by affluent ownership groups aiming for competitive success. While increased spending has elevated the league's top teams, smaller clubs have often been left financially disadvantaged. This study supports prior findings that while higher spending can boost performance, it frequently undermines profitability, particularly for clubs outside the league's top tier.

2.1 Wage Determination in the Labor Market

In labor markets, wages are typically determined by the interaction between labor supply and demand, where employers hire workers based on their marginal productivity and the revenue they generate. Ideally, wages would reach an equilibrium in a competitive market, balancing labor supply with demand. However, real-world labor markets deviate from this ideal, with wage determination influenced by factors like gender, race, education, occupation, and industry sector (e.g., nonprofit vs. for-profit). Wage disparities frequently stem from discrimination and job characteristics, such as oversight level or intrinsic motivation, which vary across sectors.

In the sports industry, additional factors like physical attractiveness influence earnings, particularly in spectator-oriented environments where public appeal plays a role (Berri et al., 2011; Dietl et al., 2020). Such findings highlight the range of non-performance factors affecting wage determination across various sectors, including professional sports.

2.2 Player Wage Determinants

While performance metrics are fundamental, player wage structures in soccer involve a nuanced mix of talent, popularity, and bargaining power. According to Lucifora and Simmons (2003), significant determinants of player wages include goals scored and assists, with wages exhibiting a convex relationship with these metrics, namely, small increases in a player's performance can lead to disproportionately large increases in their earnings. The "superstar effect" exemplifies this trend, as top forwards command wages far exceeding those of their teammates due to their goal-scoring abilities. Franck and Nüesch (2012) highlight player popularity—measured through media presence—as another crucial factor, enhancing marketability and boosting club revenue through merchandise sales and viewership.

Globalization and digital transformation have further amplified player visibility, intensifying competition for broadcasting rights and sponsorships, which, in turn, inflates wages (Kuyucu, 2020; Rowley et al., 2014). Wage disparities are also influenced by the "beauty premium," wherein attractive players, regardless of position, tend to earn more, with physical appeal enhancing marketability and sponsorship opportunities (Parshakov et al., 2024).

2.3 Globalization and Wage Disparities in Soccer

The Bosman ruling, which removed transfer fees and restrictions on foreign players, reshaped wage determination in European soccer by increasing competition for top talent and expanding the foreign player pool. This deregulation widened wage disparities within domestic leagues, particularly benefiting players from larger markets, though it did not disrupt overall competitive balance (Binder & Findlay, 2011). The continued global competition for high-performing players further drives wage inflation, challenging financial sustainability within clubs (Kesenne, 2007).

2.4 Position-Specific Wage Determination

Player wages are significantly influenced by position-specific responsibilities and visibility. Forwards, tasked with scoring goals, typically receive the highest wages, while defenders and goalkeepers, who have less visible roles, generally earn less. These positional wage disparities are particularly pronounced in leagues like US Major League Soccer, where forwards and midfielders may earn significantly more than defenders and goalkeepers (Onur, 2017; Ribeiro, 2019).

2.5 Broadcasting and Popularity

Beyond on-field performance, factors like popularity heavily influence wages in leagues such as the MLS. Star players often draw larger audiences, boosting team revenue. Clubs leverage their popularity to negotiate favorable wages, with franchise players sometimes exempt from salary caps, enabling them to earn more than their peers (Scarfe et al., 2021).

2.6 Nationality Effect

Nationality also impacts player wages in European soccer, with South American players, especially those from Brazil and Argentina, often commanding wage premiums. This premium, estimated between 11% and 15%, is partly attributed to the players' fan appeal, which can increase match attendance even outside South America (Farnell, 2024).

2.7 Club Wage Structure and First Team Performance

Research underscores the link between player wages and team performance, with teams investing more in player salaries generally achieving better results. Pellegrini et al. (2015) explore wage dispersion, finding that pay equity correlates with better performance among long-standing Serie A teams, while high wage disparity aligns with increased individual competitiveness.

2.8 Financial Fair Play and Wage Regulation

To address financial instability and wage inflation, UEFA introduced Financial Fair Play (FFP) regulations, promoting fiscal responsibility among clubs. However, FFP has also increased disparities between top-earning clubs and smaller clubs. This regulation underscores the ongoing challenge of balancing wage performance with financial sustainability.

Chapter3 - Data and Method

3.1 Dataset

This study employs exploratory machine learning methods to predict player wage levels using data from the English Premier League (EPL). The primary dataset includes information for 3786 plyers before, and 3,522 players after cleaning, across 31 teams over seven EPL seasons. Each player is represented by 130 feature columns and one label column, where the label indicates the player's wage.

3.2 Data Collection

Data were aggregated from several reputable sources. Player performance and profile data were obtained from FBref, a leading source for comprehensive soccer statistics. Team revenue data were sourced from official Premier League releases on premierleague.com and supplemented with financial insights from soccerplanet.com. Together, these sources provide a holistic view, capturing both individual player attributes and team-level financial contexts. Each record represents a player-season observation, including metrics like goals scored and minutes played, alongside player-specific details such as age, nationality, and position, as well as team financial variables. This comprehensive dataset is well-suited to exploring the factors influencing player wages across seasons, making it ideal for benchmarking and evaluating machine learning models within sports analytics.

3.3 Exploring the Dataset

The initial dataset contained 3,786 records. After a data cleaning process, which will be described later, 3,522 records remained. For analysis, the cleaned data was split into a training subset (80%) and a testing subset (20%), ensuring a balanced distribution of player positions in both sets. This resulted in 2,817 records for the training set and 705 for the test set. A stratified sampling approach, based on player position, was employed to maintain proportional representation across subsets. This layered approach is particularly valuable for model training, as it captures the distinct characteristics of each position and reflects the variation in player roles within the league. The stratified sampling enhances the model's generalizability across positions, thereby improving predictive accuracy for wages across different roles. This exploration process facilitates pattern discovery, feature selection refinement, and model tuning for optimal predictive performance.

3.4 Data Preprocessing

A clean, standardized, and encoded dataset is essential for accurate wage prediction over historical player data. Data preprocessing, which transforms raw data into an analytic-ready form, included the following key steps:

Removing Duplicate Records

Duplicate records introduce redundancy and can skew model performance. Thus, all duplicate player records were identified and removed, ensuring each player-season is represented only once.

Handling Missing Values

Missing data is a common issue, particularly in datasets with numerous variables. Rows with missing values in the 'Wage' column were removed to ensure data integrity for model training. For other features, the following imputation strategies were applied:

- Categorical Variables: Missing values in categorical features were filled with the most frequent value within each column, preserving the natural distribution.
- Numerical Variables: Missing values in numerical features were filled with the mean, thereby maintaining the overall trend without distorting the dataset.

Standardization and Scaling

Scaling was applied to numeric features to optimize model training. Two scaling methods were used:

- Min-Max Scaling: Normalized all numeric values within a range [0,1], beneficial for algorithms like K-Nearest Neighbors that require consistent value ranges.
- Standardization: Applied for models assuming normally distributed data, transforming features to have a mean of zero and a standard deviation of one. This method supports algorithms like linear regression and neural networks.

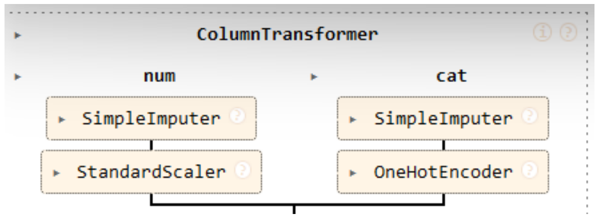
Encoding Categorical Variables

One-Hot Encoding was used to enable model interpretation of categorical information. This technique converts categorical variables into binary vectors (0s and 1s), ensuring that each category is represented without suggesting any ordinal relationship. Categorical features were expanded into binary columns, preserving categorical distinctions.

3.5 Pipeline Creation

A preprocessing pipeline was created using Simple Imputer and One Hot Encoder. This approach streamlined imputation, scaling, and encoding steps, enhancing both reproducibility and efficiency. Embedding these steps within a pipeline ensures consistent preprocessing, readily adaptable to any changes during model tuning.

Figure 3-1- Preprocessing Pipeline



Chapter4 - Data Analysis and Research Findings

In this chapter, I present a detailed analysis of the data, focusing on uncovering the factors that shape wage structures among players in the English Premier League (EPL). Through a combination of statistical tools and machine learning techniques, this chapter examines a range of variables—from on-field performance metrics to demographic and financial indicators—to predict player wages. By visualizing and analyzing the data, I aim to illustrate both obvious and nuanced patterns that emerge, shedding light on how various factors contribute to wage variability across the league.

The analysis begins with an exploratory review of the dataset, observing correlations and general trends between player attributes and their wages. Following this exploration, predictive models are applied to measure the influence of specific factors on wage levels, allowing a more precise understanding of each variable's role. Throughout this process, visualizations clarify complex relationships, such as the distribution of wages across different player positions and the impact of factors like nationality, position, and club revenue.

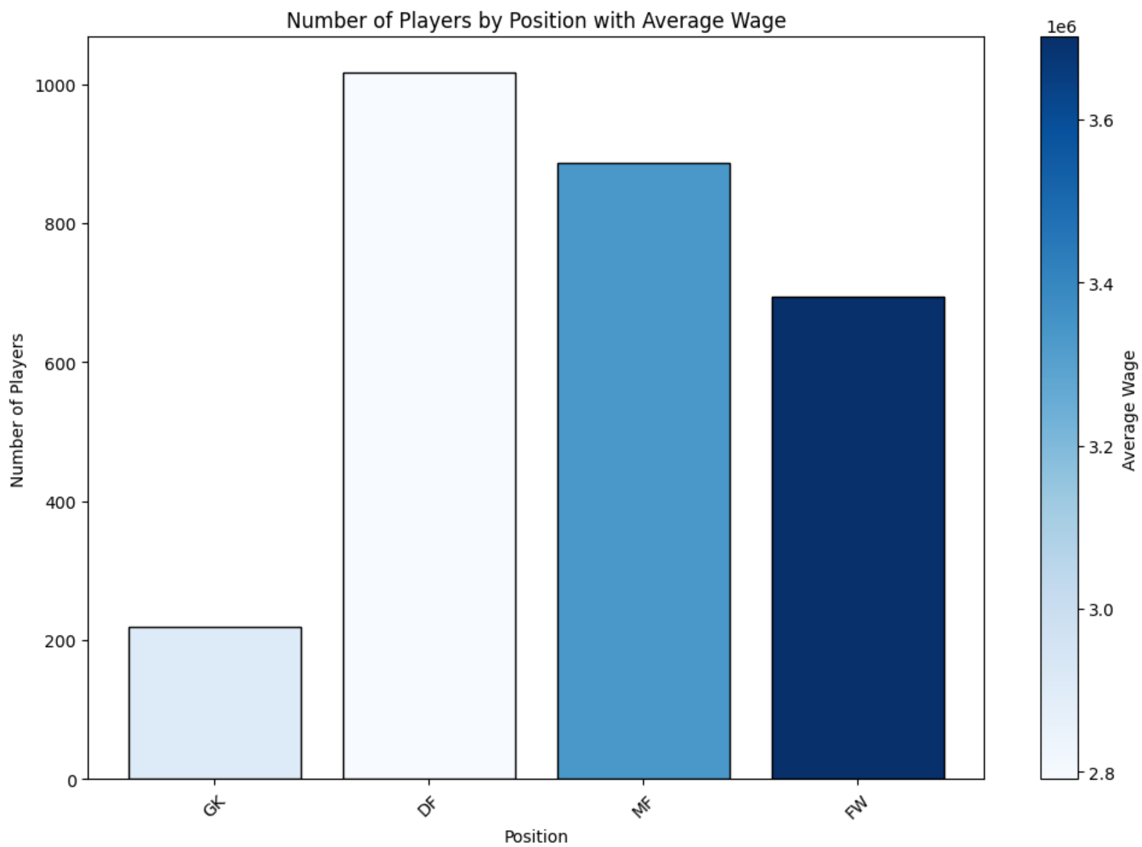
The insights gained from this analysis offer a comprehensive view of the economic and social elements that influence wage dynamics in professional soccer. These findings provide a basis for discussing broader implications for wage structures and financial sustainability in the EPL and similar leagues, setting the stage for further exploration in the chapters to come.

4.1. Visualization

To initiate data analysis, the dataset, stored in an Excel file, was imported and loaded into the working environment. Visualization of player distribution by position, including average wages, was then conducted. Color intensities represent wage levels, highlighting wage disparities across player roles and revealing which positions command higher or lower average wages.

To further illustrate our analysis of wage distribution by player position, we have also included a boxplot of wages by position, ordered by the average wage for each role. This will help us see the dispersion and central tendency of wages in each category of position, variation, and outliers that might exist. We can easily see which roles generally have higher or lower wages because the positions are sorted according to average wage, making clear the comparison of how wages vary between player roles.

Figure4-1- Distribution of Players by Position with Corresponding Average Wage Levels



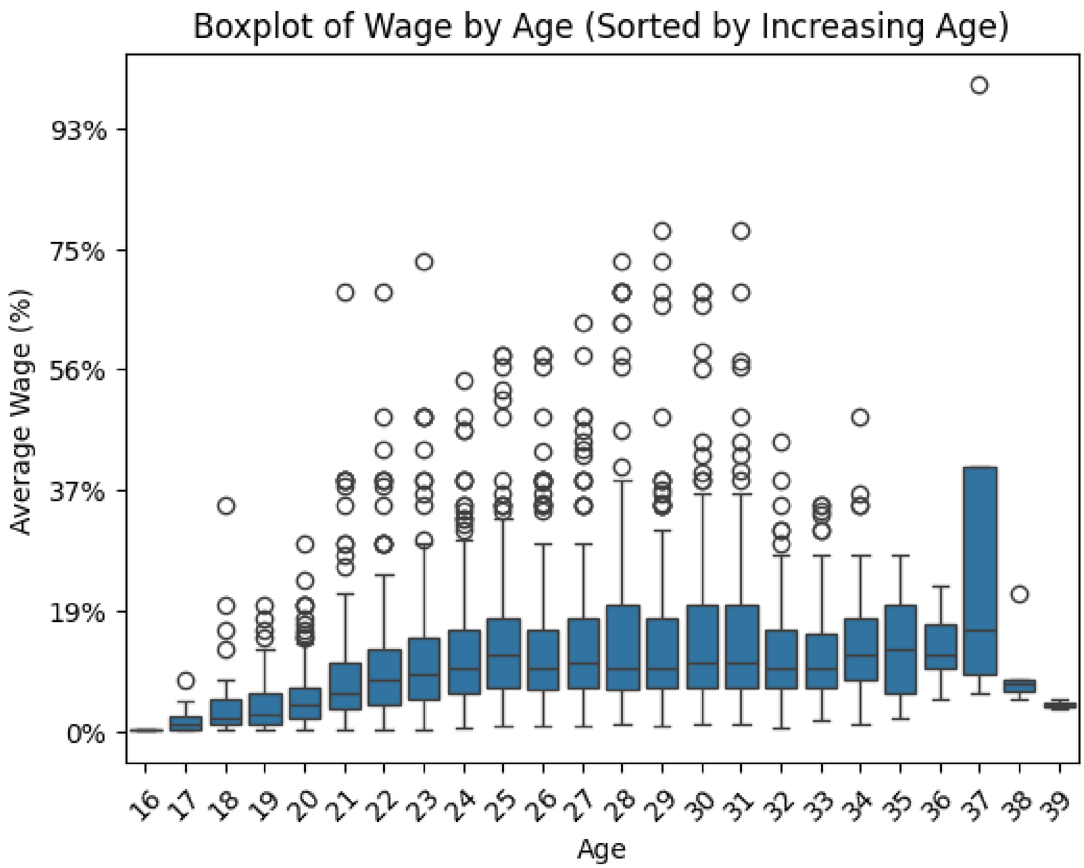
Above, the bar chart "Number of Players by Position with Average Wage" helps to visualize the distribution of players across different positions, using color intensity to show the average wage for each position. The x-axis categorizes positions, such as DF, FW, GK, and MF, while the y-axis denotes the number of players. Darker the blue in the color map, higher the average wages, as represented on the right-hand side color bar ranging from lower to higher wage value. This provides an intuitive visualization of the variation in wages across positions and, correspondingly, of the distribution of the players, and thus any comparison of patterns across different roles is rather easy.

Figure4-2- Box plot of average player wage by position

The boxplot analysis of player wages by position, expressed as percentages of the maximum wage, reveals relatively similar median wages across different positions (Defender, Goalkeeper, Midfielder, and Forward), indicating a consistent central wage tendency. However, there is notable variability in wage distributions, with the Forward position showing a higher prevalence of outliers, suggesting that top earners in this role receive substantially higher wages compared to their peers. This pattern implies that while the average wages are comparable, the Forward position, likely due to its crucial impact on match outcomes and marketability, includes individuals commanding significantly higher salaries.

Then, we have plotted the same for different age groups:

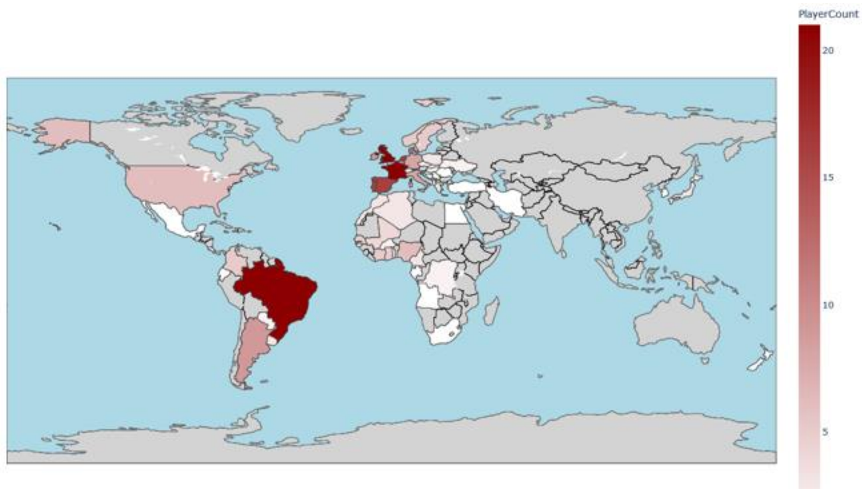
Figure4-3- Box plot of average player wage by age



This boxplot displays the distribution of players' wages across different ages, sorted from youngest to oldest to highlight trends. The chart shows that wages generally increase with age, peaking around the late 20s to early 30s, typically around age 30, after which wages tend to plateau or decrease. This suggests that players in their late 20s and early 30s often earn higher wages on average, likely due to accumulated experience, growing reputation, and better negotiation leverage. However, after age 35, wages show a notable decline. This shift is attributed to the physical toll on players' bodies, where the effects of aging outweigh the benefits of experience, leading to diminished performance and, consequently, reduced wages.

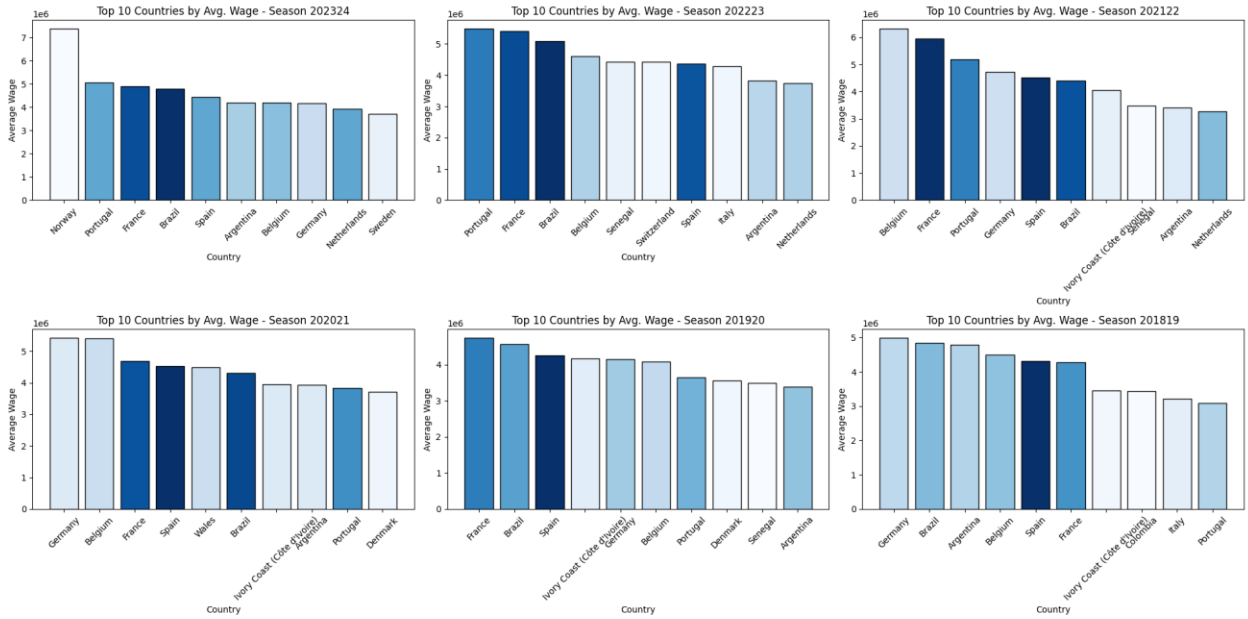
We created a choropleth map showing the geographic distribution by country of the players in the EPL 2023/24 Season. First, we filtered the dataset for the 2023/24 Season. Then we grouped the data by Country, then counted the number of players for each. Each country is shaded from white to dark red, with darker tones indicating higher player counts. Country borders and adjusted map dimensions enhance clarity, creating a comprehensive visual of the global origins of EPL players for the season.

4-4- Distribution of Players Across Countries: A Heatmap



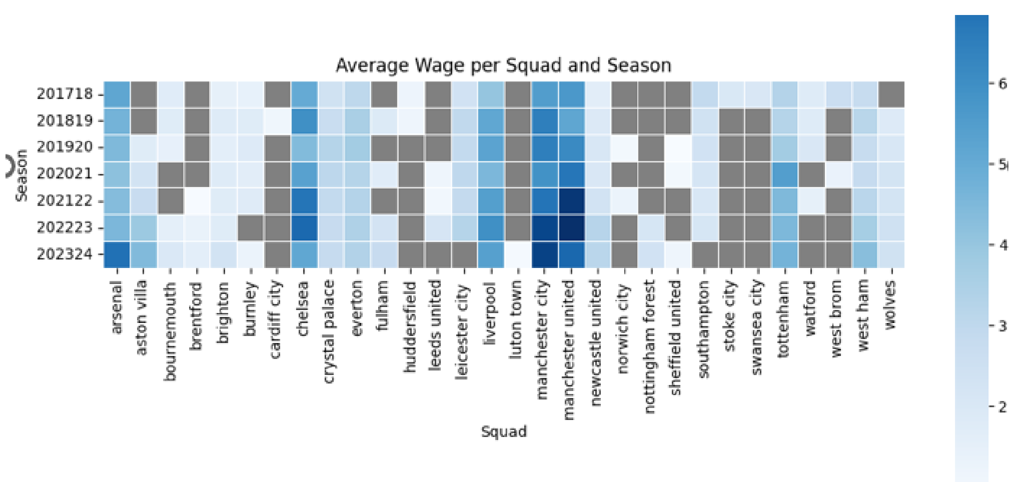
To look at recent trends in wages across countries, we generated bar plots of the top 10 countries by average wage for each of the six most recent seasons. For each season, we filtered the data to calculate the average wage per country and plotted the top 10 countries in individual subplots. The output is a 2x3 grid of bar plots, one for each season. This makes the comparisons between seasons easier. The bar lengths represent the average wage for each country, while the color intensity corresponds to the total number of players from each country. This visualization highlights wage disparities and trends in the EPL across various countries over recent years.

4-5- Matrix plot of players counts and wage by country



Next, to visualize the trend of wage overtime, I created a heatmap of an average player wage for each EPL squad across different seasons. Using a pivot table, I calculated the mean of wage for every combination of squad-season. The x-axis represents squads, the y-axis represents seasons, and gray color in a cell depicts that a particular squad is not active during that respective season so that clarity becomes prominent. Active cells are in blue; the darker, the higher the average wage. This heat map pictorially shows the trend and variation in player wages of squads over several seasons and thus helps draw inferences about the wage dynamics in the league.

4-6- Heatmap of players average wage by EPL clubs



4.2. Feature Selection:

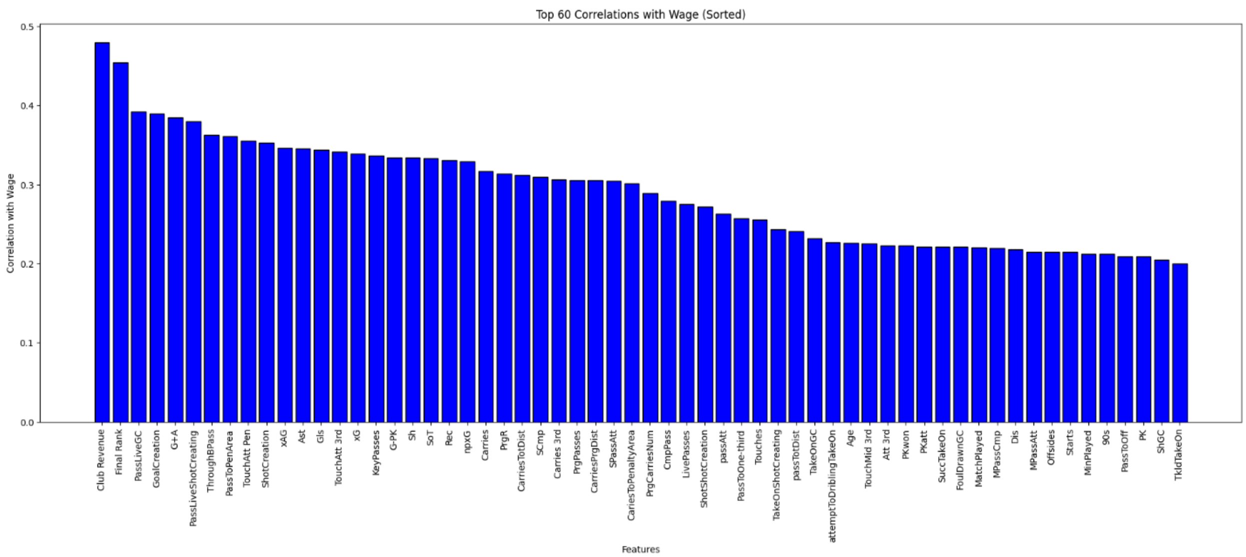
It worth mentioning that PCA destroys the original features and create new ones in new directions. So, we used other feature selection methods to highlight the features that most correlated with player wages, identified through three key selection methods: Lasso regression, variance, and simple correlation with wage. Lasso regression underscores G-xG (goals minus expected goals) and np (non-penalty goals minus expected goals) as strong indicators, suggesting the influence of scoring metrics on wage. Variance analysis focuses on financial and distance-related performance features, with Club Revenue showing the highest variance, indicating that club wealth is a critical factor in wage variability, followed by passing and carry distances like passTotDist and Deviated_Wage. Finally, the correlation analysis ranks Club Revenue as the most positively correlated feature with wage (0.4792), followed by metrics tied to offensive contributions, such as Goal Creation and ThroughBPass (through-ball passes). This combined view demonstrates that both club financial status and individual performance in goal-related and progressive play strongly influence player wage.

Tabl1- Feature selectin methods and selected features

Lasso		Variance		Correlation	
Feature	**Value**	**Feature**	**Value**	**Feature**	**Value**
G-xG	990160.73	Club Revenue	112600000	Club Revenue	0.4792
np:G-xG	837259.21	passTotDist	91550000	GoalCreation	0.3892
DefGCA	583363.69	Deviated_Wage	26520000	ThroughBPass	0.3622
PKatt	328305.11	PrgDist	16860000	PassToPenArea	0.3609
PK	318901.86	AvgWage_Team_Season	8215000	xAG	0.3464
PKwon	194497.14	SDVWage_Team_Season	6440000	Ast	0.3455
OG	180417.04	CarriesTotDist	5264000	Gls	0.3437
2CrdY	178340.22	CarriesPrgDist	1636000	PrgR	0.3137
PassDeadGC	177836.98	Touches	518200	CariesToPenaltyArea	0.3015
Age	177640.41	passAtt	392600	PrgCarriesNum	0.2892
DefShotCreation	163891.76	LivePasses	331200	CmpPass	0.2794
CrdR	160922.09	CmpPass	275000	ShotShotCreation	0.2721
npxG	156754.93	Rec	241000	PassToOne-third	0.2576
FoulDrawnGC	155285.98	Carries	178400	TakeOnShotCreating	0.2432
Gls	150911.32	TouchMid 3rd	149200	TakeOnGC	0.2321
xAG	150055.15	TouchDef 3rd	106000	attemptToDribling	0.2268
TakeOnGC	112970.11	SPassAtt	73860	Age	0.2257
Ast	94133.85	MPassAtt	71910	Att 3rd	0.2232
xG	92113.84	SCmp	60870	PKwon	0.2226
Pos2?	87537.46	MPassCmp	58420	FoulDrawnGC	0.2212

Following is the bar chart now, showing the top 60 features that have the highest correlation with the players' wages. The chart shows every feature on the x-axis while the y-axis represents the values of its correlation with wage after sorting the top 60 values from the sorted data of correlations.

4-7- Bar plot of the selected features (based on the correlation)



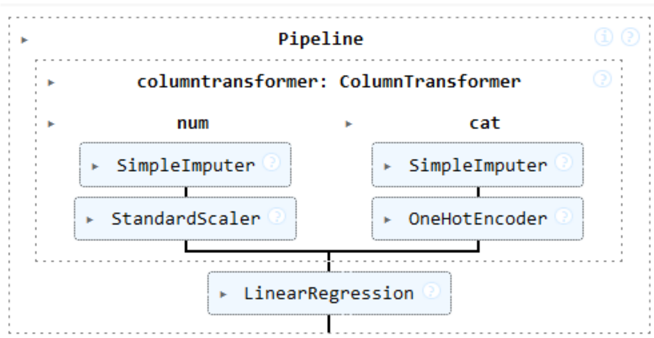
This plot helps to identify the most influential features associated with wage, which are Club revenue, Final rank, and passing power during the game. These show which player attributes are closely related to the wage level. On the other side, penalty kick conceded is negatively correlated with the player’s wage.

4.3. Regression Models and Specification Analysis

In this section, we are going to apply a set of regression models to investigate the relations between player features and their respective wage levels. We apply different specifications and model types to find out which factors strongly influence player wages and check whether these relations hold across different models. In turn, this will enable us to compare and, hence, validate our findings, observe nonlinearities that may exist, and, finally, capture complexities in the data set. From these models, the derived insight will afford a comprehensive understanding of the main predictors of player wages and provide a basis for a sound predictive framework.

First, a pipeline was created to fasten this process of modeling and to ensure the outputs were consistent. This pipeline takes in raw input data, applies all the necessary preprocessing steps-clean, select features, and encode data-followed by the application of the regression models. This pipeline also gains efficiency from the automation of these steps, given that it guarantees uniformly prepared data for each regression model to underpin comparability of the results between different model specifications.

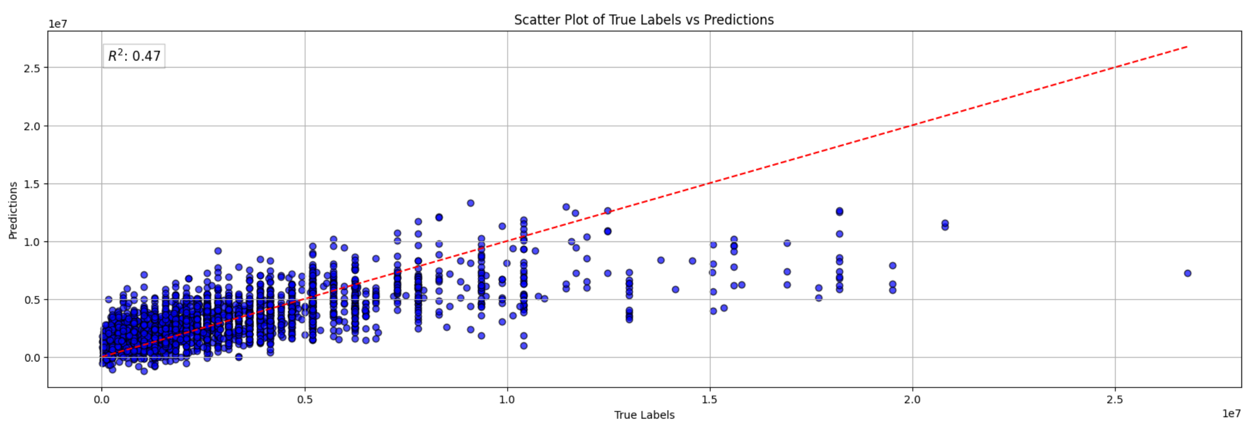
4-8- Preprocessing and regression pipeline



4.2.1 Simple Linear Regression Model

The simple linear regression model serves as a foundational approach to examine the relationship between player features and wage. By assuming a linear relationship, this model estimates the direct effect of each feature on wages while providing interpretable coefficients that show the marginal impact of each predictor. Although simple regression is limited in capturing complex interactions and non-linearities, it serves as a valuable baseline for understanding the fundamental associations in the data and provides a straightforward benchmark for comparison with more sophisticated models.

Figure 4-9- Predicted vs. Actual Wages with Fitted Line – Simple Regression

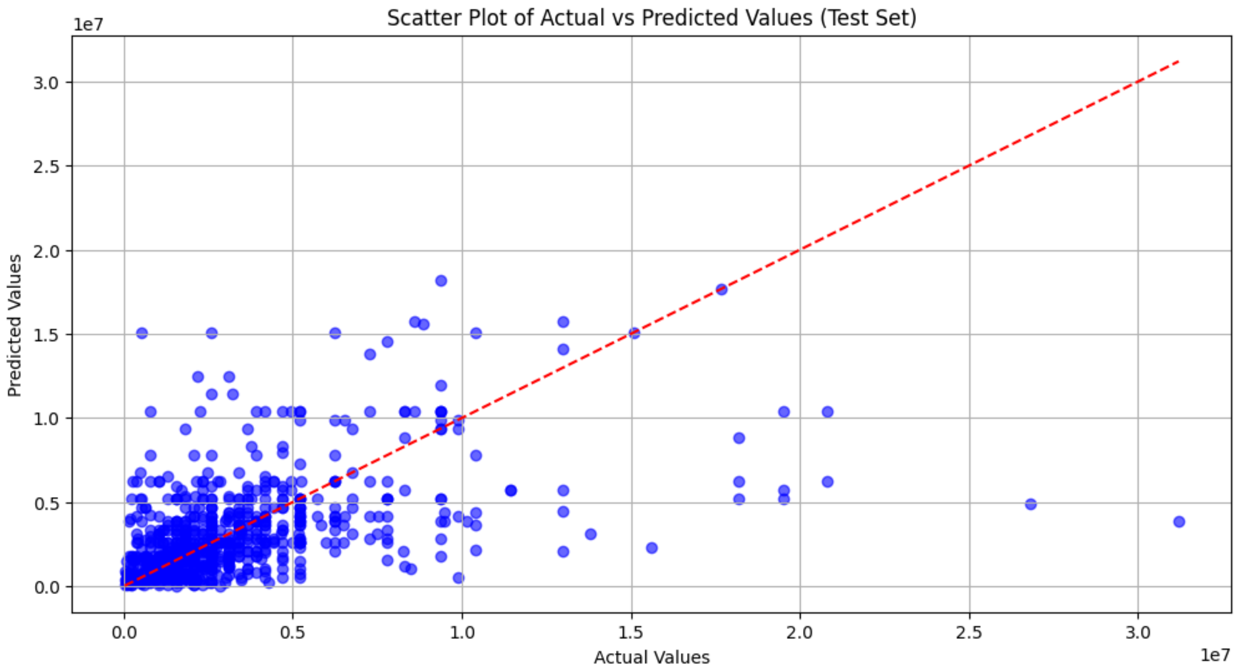


This scatter plot displays the predicted wage values versus the actual wage values from the simple regression model. Each blue dot represents a player, with the x-axis indicating the true wage and the y-axis showing the model's prediction. The red dashed line represents the ideal scenario where predictions perfectly match actual values, i.e. a 45-degree line. The R square value of 0.47 indicates that approximately 47% of the variance in wage is explained by the model, suggesting a moderate fit. While there is a positive correlation between true and predicted wages, the spread of points around the line, especially at higher wage levels, indicates that the model has limitations in accurately predicting wages, likely due to non-linear patterns and interactions not captured by a simple linear model.

4.2.2. Decision Tree Regressor

The decision tree regressor is a non-linear model that uses a tree-like structure to segment the data based on feature values, resulting in distinct wage predictions for different groups of players. Unlike linear regression, decision trees can capture complex interactions between variables, making them well-suited to handle non-linear relationships. This model recursively splits the data into branches, enabling it to uncover patterns in the data that a linear approach might miss. However, decision trees can be prone to overfitting, which is something we address by tuning hyperparameters and testing with cross-validation.

Figure 4-10- Predicted vs. Actual Wages with Fitted Line – Decision Tree Regressor



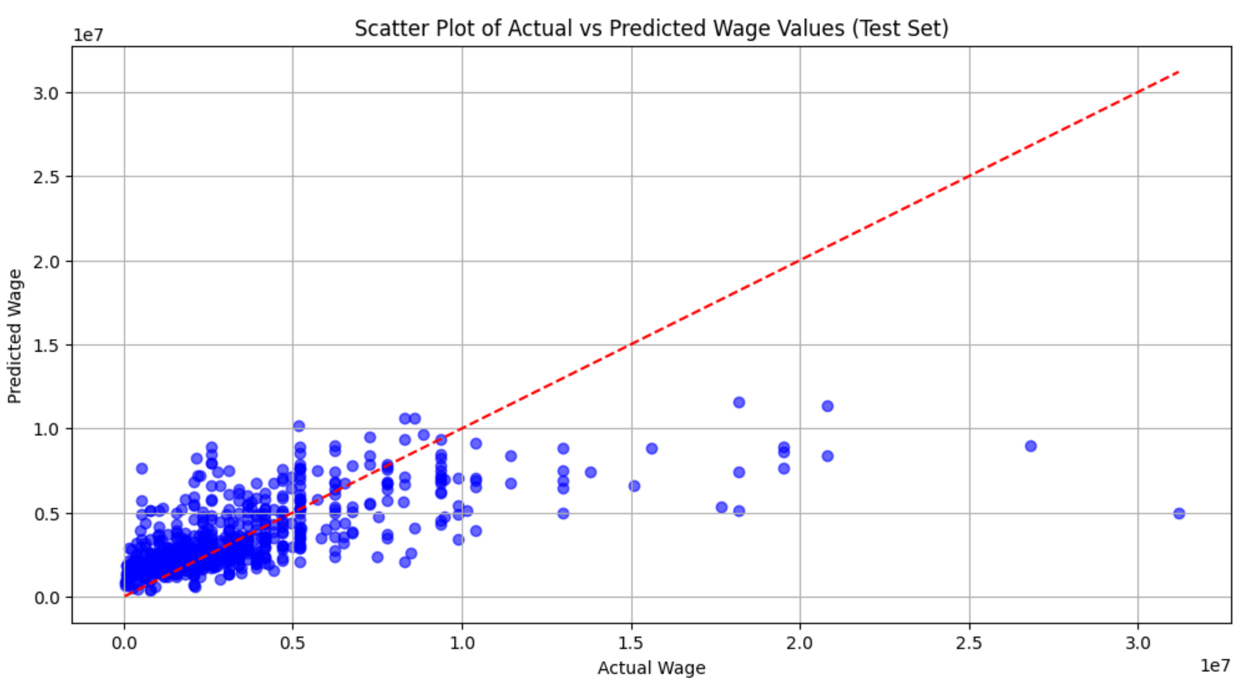
The results from our decision tree regressor on the training set show a Root Mean Squared Error (RMSE) of 0 and an R2 score of 1, indicating a perfect fit to the training data. While this might initially seem ideal, it actually suggests that the model has likely overfit to the training set, capturing noise and specific patterns that do not generalize well. When evaluated on the test set, the model yields a much higher RMSE of 3,269,969.63 and a low R2 score of 0.08, demonstrating poor predictive performance on unseen data. This discrepancy confirms overfitting and highlights the need to explore alternative models that may better generalize to new data. We will now proceed to test other models to improve predictive accuracy and reduce overfitting.

4.2.3. Random Forest Regression

Building on the decision tree model, random forest regression is an ensemble method that combines multiple decision trees to improve predictive accuracy and stability. By averaging the results from numerous trees, random forests mitigate the risk of overfitting associated with individual trees and provide a more robust estimate of wages. This method also incorporates randomness in the selection of features and data samples for each tree, enhancing generalizability and offering a clearer view of the features that consistently influence player wages. Random forest regression is a powerful tool for capturing non-linear relationships while maintaining high predictive accuracy.

The best result achieved on the training set yielded an R2 score of 0.47, indicating significant room for improvement. We then applied the trained model to the test set, and here are the results:

Figure 4-11- Predicted vs. Actual Wages with Fitted Line – Random Forest Regression

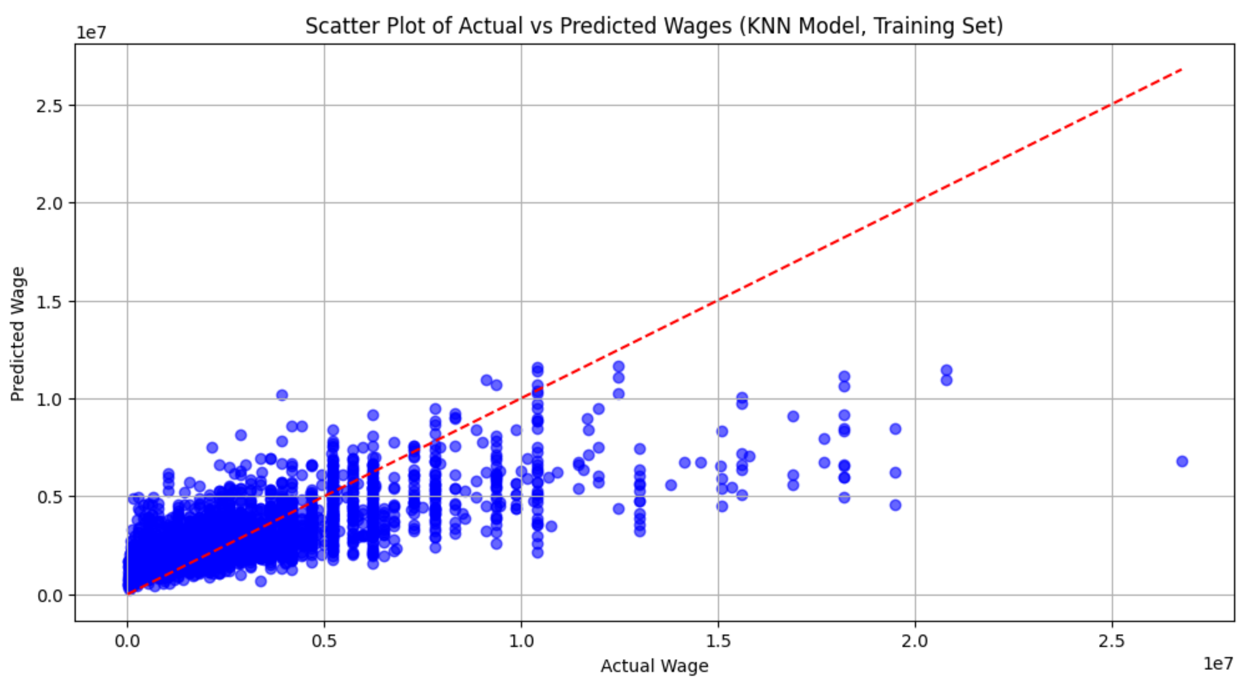


The results from the Random Forest model on the test set indicate a Root Mean Squared Error (RMSE) of 2,558,903.87 and an R2 score of 0.44. The RMSE value reflects the average prediction error in monetary units, suggesting that, on average, the model's wage predictions deviate by about 2.56 million from the actual values. The R2 score of 0.44 implies that the model explains approximately 44% of the variance in wage within the test set. While the Random Forest model performs better than the overfitted Decision Tree model, as indicated by the improved R2R^2R2, there remains considerable unexplained variance, suggesting that additional or more complex features might be needed to enhance predictive accuracy. The scatter plot shows that the model generally follows the trend of actual wages but struggles with higher wage values, indicating the possibility for more improvement.

4.2.4. K-Nearest Neighbors (KNN) Regression

K-Nearest Neighbors (KNN) regression is an instance-based learning method that estimates a player's wage based on the wages of the k most similar players in the dataset. Instead of learning explicit parameters, KNN relies on proximity in the feature space to make predictions, making it highly adaptable to various data structures. This model is particularly useful for capturing local patterns in the data and does not assume any specific functional form between features and wage. However, KNN can be sensitive to the choice of k and may perform less effectively with high-dimensional data, necessitating careful tuning to optimize its performance.

Figure 4-12- Predicted vs. Actual Wages with Fitted Line – KNN Regression

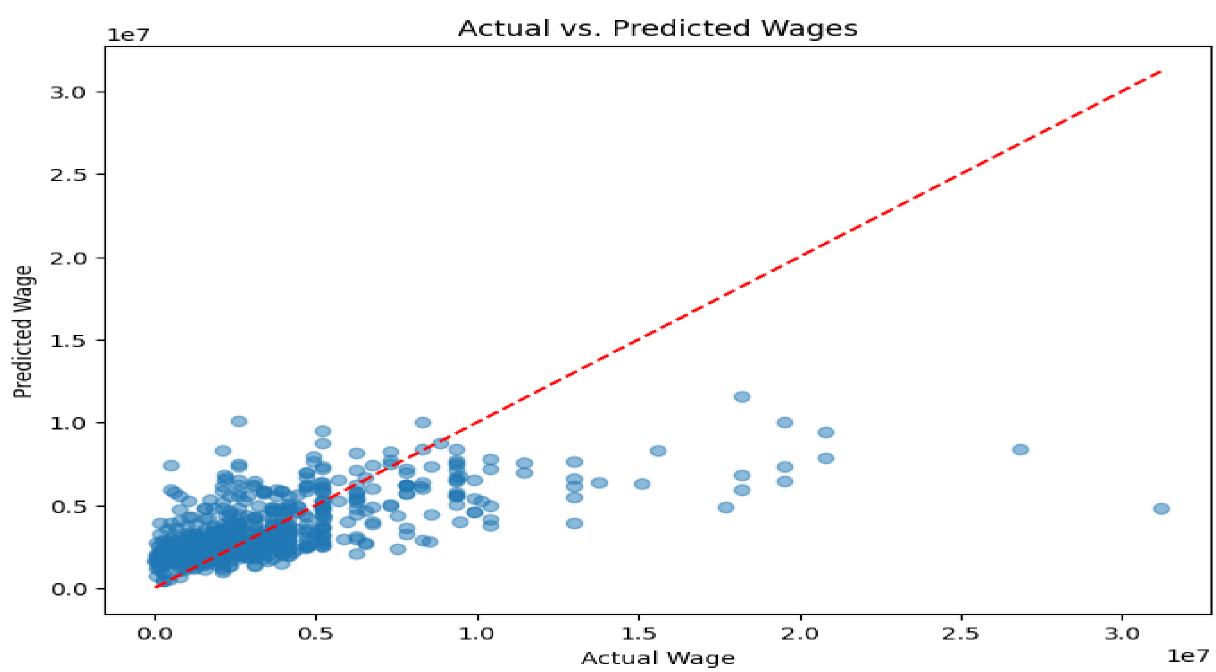


The scatter plot of actual vs. predicted wages for the training set reveals a clear indication of underfitting by the K-Nearest Neighbors (KNN) model. The predictions show substantial deviation from the ideal line, with a wide spread of points around it, especially at higher wage levels. This underfitting suggests that the model is too simplistic and fails to capture the underlying complexity of the wage data within the training set. After experimenting with various values of k, the best k based on RMSE was found to be 9, with a mean RMSE of 2,356,165.35 and a mean R2 of 0.3297. These metrics indicate limited predictive power, as the model explains only about 33% of the variance in wage, emphasizing that the high dimensionality of the features may be challenging for the KNN model to account for effectively at this value of k=9.

4.2.5. Randomized Search

This model leverages a randomized search for hyperparameter tuning to optimize a machine learning model for predicting player wages. Using the best model found during the random search (rnd_search.best_estimator_), it is then applied to the test set (X_test) to make predictions on player wages. The model's performance is evaluated using two key metrics: Root Mean Squared Error (RMSE) and R^2 score. RMSE provides a measure of prediction accuracy by indicating the average error in wage predictions, while R^2 reflects the model's ability to capture the variance in the actual wage data. These evaluations give insight into the model's effectiveness and predictive quality on unseen data.

Figure 4-13- Predicted vs. Actual Wages with Fitted Line – Randomized Search Regression

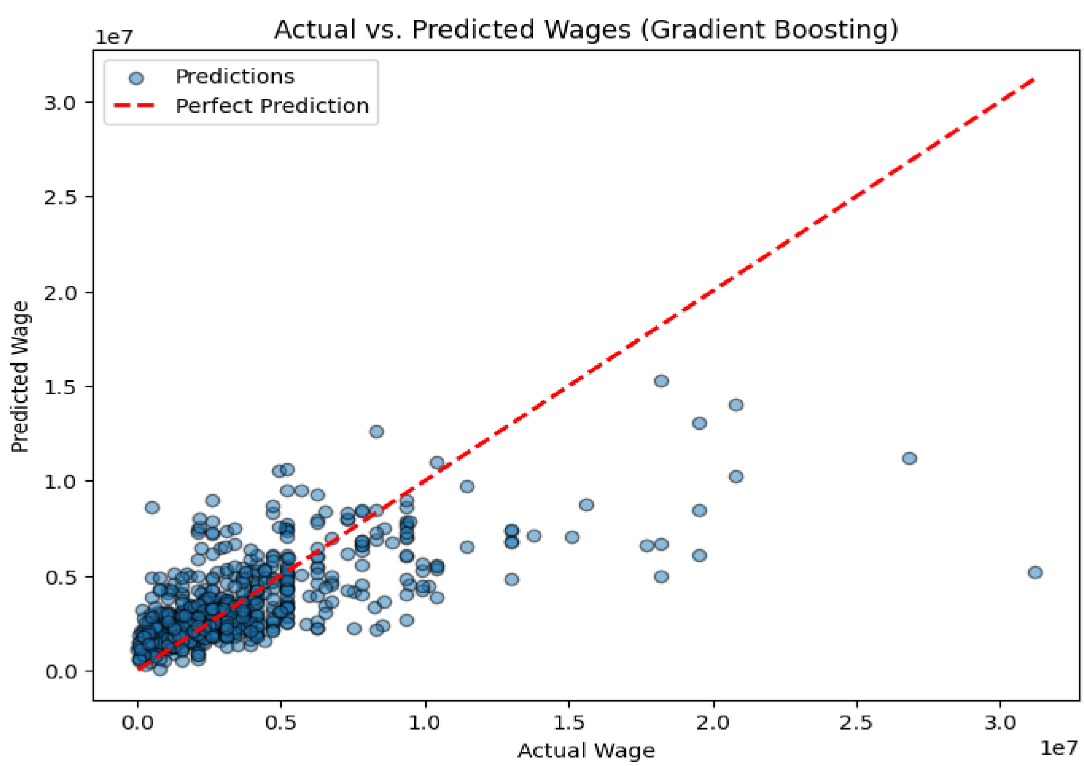


The model shows moderate predictive effectiveness, with a Final RMSE of 2.6 million, indicating that, on average, predictions deviate significantly from actual wage values. An R^2 of 0.40 reveals that the model explains only 40% of the variance in wages within the test set, leaving substantial variability unaccounted for. However, this model outperformed all previous models in predictive accuracy and demonstrated less overfitting, suggesting a more balanced fit between training and test data. This indicates progress in capturing wage patterns, although further improvements could still enhance precision and overall model performance.

4.2.6. Gradient Boosting model

The Gradient Boosting model is an ensemble learning method that builds a series of decision trees, where each new tree aims to correct errors made by the previous ones. This model iteratively minimizes the error through gradient descent, adjusting its predictions based on residuals (errors) from prior iterations. By using a combination of parameters—100 trees (n_estimators=100), a moderate learning rate of 0.1, and a maximum tree depth of 3—the model strikes a balance between learning speed and complexity, potentially reducing overfitting while achieving high predictive power.

Figure 4-14- Predicted vs. Actual Wages with Fitted Line – Gradient Boosting Model



In the training set, the Gradient Boosting model achieved an RMSE of 1.53 million and an R^2 of 0.72, which suggests a substantial improvement in accuracy compared to the previous model's RMSE of 2.63 million and R^2 of 0.40. The higher R^2 (0.72) indicates that Gradient Boosting explains a much larger portion of wage variance within the training data, capturing more predictive patterns. This model's lower RMSE also reflects greater precision in wage prediction, suggesting it is significantly more effective in reducing error. Additionally, given Gradient Boosting's design to minimize overfitting, it could likely generalize better on unseen data, further setting it apart from the previous model's more limited performance.

4.4. Model Evaluation and Comparison

This section should provide a structured comparison of the models based on key metrics like RMSE, R^2, and other performance criteria. You can summarize the strengths and weaknesses of each model, discuss the trade-offs, and ultimately justify the final model selection based on your findings.

Figure 4-15- Model Selection Bar Chart

The evaluation of various models based on their Train and Test R^2 scores highlights Gradient Boosting as the most suitable choice for this task. Starting with Linear Regression, its low Train and Test R^2 scores (0.47 and 0.0, respectively) indicate that it struggles to capture variance in the data, showing limited predictive capability. The Decision Tree model, while achieving a perfect Train R^2 score of 1.0, significantly drops to 0.07 on the test set, demonstrating severe overfitting. This model memorizes the training data but fails to generalize to new data, making it unreliable for robust predictions.

Randomized Search performs better, with a high Train R^2 of 0.92 and a Test R^2 of 0.4, though the disparity between these scores suggests some overfitting. Random Forest provides a more balanced performance, with Train and Test R^2 scores of 0.48 and 0.44, indicating that it generalizes moderately well without significant overfitting. However, its overall predictive power remains modest.

Among the models, Gradient Boosting achieves the highest Test R^2 score of 0.45 and a Train R^2 of 0.72, reflecting a good balance between capturing patterns in the training data and maintaining generalization to unseen data. This balanced performance makes Gradient Boosting the preferred model for this task, as it offers a sophisticated approach with both reasonable complexity and predictive reliability.

Chapter 5 - Conclusion and Discussion

In conclusion, this research reinforces the multifaceted nature of wage determination in professional soccer. While on-field performance is foundational, the interplay of nationality, market demand, and marketability significantly impacts salary structures. Players from established soccering nations, those with high market appeal, and younger players entering prominent leagues often experience wage premiums that are only partially attributable to their athletic abilities. These findings underscore the importance of integrating objective, data-driven approaches to wage determination that consider the full range of player contributions while mitigating non-performance biases. By advancing toward a more transparent, analytics-based system, the soccer industry can foster more equitable wage practices and optimize player valuations in line with both market demands and sporting merit.

This study provides an in-depth examination of wage determinants in professional soccer, with a particular focus on how clubs in elite leagues balance quantifiable performance metrics with market-driven factors such as nationality, marketability, and systemic biases. While player performance remains a central component in wage determination, this research highlights the complexity and variability that characterize salary structures in global soccer.

5.1. Key Findings and Comparative Insights

This study reaffirms the strong correlation between player wages and performance metrics such as goals, assists, and defensive actions. The best model identified for wage prediction is Gradient Boosting, which demonstrated 72% explanatory power in the training set and 45% in the test set. Based on feature selection and the performance of the best-fit model, factors like last year's team rank and the previous season's broadcasting income are the most significant contributors to player wages. Following these, possessive skills such as long passes and other performance actions, including free kicks that ultimately create goal opportunities, play a crucial role. Additionally, the number of goals scored, and assists performed by players are key determinants influencing their wages for the following season.

Some clubs tend to invest more in acquiring players, driven by crowd expectations, the availability of a high budget, and other potential reasons. For example, as indicated by the heat map and efficiency scatter plot, Manchester United consistently pays more than other clubs to hire players with comparable performance levels. Nationality is another important factor impacting player wages, as evidenced by bar-matrix plots. Players from Brazil and Argentina often receive higher wages due to their attractiveness to spectators, even when their on-field performance is similar to that of their peers. This observation is further supported by regression results. Finally, age plays a significant role in wage determination, as suggested by bar plots and regression models. There is a slight upward trend in average wages from ages 16 to 37, which can be attributed to the increase in experience, strategic thinking, negotiation power, reputation, and other factors.

5.2. Discussion

Performance Metrics as a Core Wage Determinant: Consistent with the findings of Yaldo and Shamir (2017), this study underscores the correlation between player wages and performance metrics, including goals, assists, and defensive actions. These indicators are crucial, particularly in leagues such as the English Premier League, where statistical performance often directly influences contract negotiations and transfer market valuations. Ribeiro and Lima (2019) further support this trend, noting that inter-league transfers and a player's historical performance consistency can lead to wage adjustments, both positive and negative, based on the reputation and standards of different leagues. The reliance on advanced analytics by clubs signifies a growing shift towards performance-based valuations, although our study indicates that this focus is moderated by several non-performance-related factors.

Nationality Premium and Market Demands: Our findings align with Farnell et al. (2024) and Kuyucu (2021), who identified a "nationality premium" in European soccer leagues, where players from prominent soccering nations such as Brazil, Argentina, and select European countries are over-represented in higher wage brackets. This trend is likely driven by fan demand, media coverage, and the potential for higher merchandise sales, suggesting that clubs often view nationality as a market asset beyond a player's pure athletic contribution. Such preferences may inadvertently penalize players from less celebrated soccering nations, who, despite possessing equivalent skills, may face wage disparities. This highlights an inherent market bias that underscores the importance of nationality in the global soccer industry's valuation framework.

superstar and Marketability Effects: While our current study did not incorporate player reputation or brand appeal directly, literature by Lucifora and Simmons (2003) and Onur Celik and Meltem Ince-Yenilmez (2017) illustrates how players with a strong fan following or "superstar appeal" command significantly higher wages due to their capacity to generate club revenue through ticket sales, sponsorships, and merchandise. These "superstar effects" imply that marketability can sometimes outweigh pure performance, potentially distorting wage structures in favor of players who, while possibly less statistically impactful, possess substantial brand appeal. To address this factor in future research, we recommend including metrics like players' web search frequency and social media following as proxies for player reputation to better understand how this variable may affect wages alongside on-field performance.

Ethnic and Racial Considerations in Wage Differentials: Insights from Ma and Haugen (2022) underscore how ethnicity and race can affect labor market segmentation, even in sports settings that often profess meritocratic principles. This study observed wage disparities associated with players from ethnic minority backgrounds, who frequently face pay disadvantages despite having comparable on-field performance metrics. These findings resonate with Courey's (2020) work in non-sports labor markets, which illustrates that such biases, though subtle, can have substantial long-term impacts on wage and career trajectories for affected players. While these non-performance factors may seem minor in isolation, they cumulatively reflect underlying structural issues within soccer's wage determination system.

Dynamic Role of Age and Career Trajectory: Ribeiro and Lima (2019) and Scarfe et al. (2021) discuss how wage progression across a player's career can reflect age-related factors, with younger players potentially undervalued relative to their future potential, and older players sometimes receiving wages based on reputation rather than current output. Our study finds that age, combined with career trajectory and injury history, can significantly affect wages, particularly in leagues that favor young talent pipelines, such as the Bundesliga. The implications suggest that age-based wage decisions may not always reflect present-day performance but rather a composite of historical achievements and future expectations.

5.3. Implications and Future Research

The findings suggest that while performance metrics remain central to wage determination, the current system is intricately shaped by a blend of market forces, club-specific demands, and subjective valuation elements. This complex combination implies that purely performance-based valuations may not fully capture a player's perceived value to the club. Notably, we were able to explain nearly half of the variation in player wages in the English Premier League (EPL), which is a significant achievement given the complexity of wage determinants. Clubs are encouraged to integrate more advanced data analytics, not only to assess player potential and contribution more accurately but also to minimize biases associated with nationality, age, and ethnicity. Enhancing transparency in wage structures could foster more equitable pay practices across leagues, benefitting players and clubs alike.

Our research highlights opportunities for further investigation, particularly around the longitudinal impacts of wage disparities based on nationality, age, and marketability on players' career development. To extend this analysis, we recommend incorporating metrics that capture player reputation, such as social media followers and search engine trends, as these proxies could provide valuable insights into a player's marketability and popularity. Reputation, which is a significant determinant, was not fully explored in our current study due to data limitations. Another important factor is negotiation power, often controlled by player agents. This aspect poses challenges as it is not easily accessible, even when a model has considerable explanatory power. Additionally, player attractiveness and appearance could influence wages, factors that were not included in this work. Integrating these elements could further enrich and expand the present study, offering a more comprehensive understanding of wage determination.

Future studies could explore the correlation between wage adjustments and a player's career trajectory, examining whether advanced tracking metrics such as stamina, mental resilience, and adaptability influence wage growth over time. As the soccer industry increasingly adopts digital tracking and data analytics, the potential for a more performance-centered and less biased wage determination system is substantial.

References

Adam, S., Joyce, R., & Xu, X. (2019). Labour's proposed income tax rises for high-income individuals. IFS Briefing Note BN265, Institution for Fiscal Studies.

Adler, M. (1985). Stardom and Talent. American Economic Review, 75(1), 208–212.

Alvaredo, F., Atkinson, A. B., Piketty, T., & Saez, E. (2013). The Top 1 Percent in International and Historical Perspective. Journal of Economic Perspectives, 27(3), 3–20.

Arellano, M., & Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. The Review of Economic Studies, 58(2), 277–297.

Bakija, J., Cole, A., & Heim, B. (2012). Jobs and Income Growth of Top Earners and the Causes of Changing Income Inequality: Evidence from U.S. Tax Return Data. Department of Economics Working Papers 2010-22, Department of Economics, Williams College.

Berk, R. (1990). A primer on robust regression. In J. Fox & J. S. Long (Eds.), Modern Methods of Data Analysis (Chap. 7, pp. 292–324). Sage.

Brandes, L., Franck, E., & Nüesch, S. (2008). Local Heroes and Superstars: An Empirical Analysis of Star Attraction in German Soccer. Journal of Sports Economics, 9(3), 266–286.

Bryson, A., Rossi, G., & Simmons, R. (2014). The Migrant Wage Premium in Professional Soccer: A Superstar Effect? Kyklos, 67(1), 12–28.

Carrieri, V., Principe, F., & Raitano, M. (2018). What makes you 'super-rich'? New evidence from an analysis of soccer players' wages. Oxford Economic Papers, 70(4), 950–973.

Coates, D., Frick, B., & Jewell, T. (2016). Superstar Salaries and Soccer Success: The Impact of Designated Players in Major League Soccer. Journal of Sports Economics, 17(7), 716–735.

Célérier, C., & Vallée, B. (2019). Returns to Talent and the Finance Wage Premium. The Review of Financial Studies, 32(10), 4005–4040.

Ferrari, L., & Rueda, M. (2015). 2015 MLS collective bargaining agreement: Moving in the Right Direction. Global Sports Law and Taxation, 6(2), 6–11.

Franck, E., & Nüesch, S. (2012). Talent And/Or Popularity: What Does It Take To Be A Superstar? Economic Inquiry, 50(1), 202–216.

Franzini, M., Granaglia, E., & Raitano, M. (2016). Extreme Inequalities in Contemporary Capitalism. Springer.

Frick, B. (2007). The Soccer Players' Labor Market: Empirical Evidence from the Major European Leagues. Scottish Journal of Political Economy, 54(3), 422–446.

Hamlen Jr, W. A. (1991). Superstardom in Popular Music: Empirical Evidence. Review of Economics and Statistics, 729–733.

Hausman, J. A., & Leonard, G. K. (1997). Superstars in the National Basketball Association: Economic Value and Policy. Journal of Labor Economics, 15(4), 586–624.

Hoffman, K., & Opitz, C. (2019). Talent and publicity as determinants of superstar incomes: empirical evidence from the motion picture industry. Applied Economics, 51(13), 1383–1395.

Jane, W.-J. (2016). The Effect of Star Quality on Attendance Demand: The Case of the National Basketball Association. Journal of Sports Economics, 17(4), 396–417.

Jewell, R. T. (2017). The Effect of Marquee Players on Sports Demand: The Case of U.S. Major League Soccer. Journal of Sports Economics, 18(3), 239–252.

Kaplan, S. N., & Rauh, J. (2013). It’s the Market: The Broad-Based Rise in the Return to Top Talent. Journal of Economic Perspectives, 27(3), 35–56.

Acronyms

1- Standard Statistics

var name	Description
mp	Match played by player
Starts	Games started by player
Min	Minutes played
90s	Minutes played divided by 90
Gls	Goals Scored by player
Ast	Goals assist by player
G+A	Gls + Ast
G-PK	Non- penalty goals scored
PK	Penalty kicks made by player(Scored)
Pkatt	Penalty kicks attempted by player(kicked)
CrdY	Yellow Cards
CrdR	Red Cards
XG	Expexted Goals including penalty kicks, not penalty shootouts
npXG	Non-Penalty Expected gals
xAG	Expected assisted goals: All passes that lead to scoring chance
prgC	progressive Carries towards the opponent's goal line at least 10 yards
prgP	progressive passes
prgR	progressive passes Received

2- Shoot Statistics

sh	Shots
sot	Shot on target
Dist	Average distance of shots from goal of player in yards
FK	Shots from free kicks
PK	Penalty kicks
G-xG	Goals scored by player - expected goals
np:G-xG	non- penalty: Goals scored by player - expected goals

3- Passing Performance

cmp	Passes cmpleted
passAtt	Passes attempted
cmp%	passes completed / passes attempted
passTotDist	Total distances that passes travelled in any direction (in yards)
passprgDist	Total yards distances that pass travelled toward the opponent's goal
Scmp	Short Passes cmpleted
SpassAtt	Short Passes attempted
Scmp%	Short passes completed / Short passes attempted
SpassTotDist	Total distances that Short passes travelled in any direction (in yards)
SpassPrgDist	Total distances that Short passes travelledtoward the opponent's goal (in yards)
MPasscmp	Medium Passes cmpleted
MPassAtt	Medium Passes attempted
MPasscmp%	Medium passes completed / Medium passes attempted
MPassTotDist	Total distances that Medium passes travelled in any direction (in yards)
MPassprgDist	Total distances that Medium passes travelledtoward the opponent's goal (in yards)
LPasscmp	Long Passes cmpleted
LPassAtt	Long Passes attempted
LPasscmp%	Long passes completed / Long passes attempted
LPassTotDist	Total distances that Long passes travelled in any direction (in yards)
LPassprgDist	Total yard distances that Long passes travelled toward the opponent's goal
KP	Key Passes: Passes that directly leads to shot
one-third	Passes to final one third
PPA	Passes to penalty area (Completed)
CrsPA	Crosses into Penalty Area (Completed)
prgP	Progressive Passes
Live	Live Ball Passes
Dead	Dead Ball Passes including kick-off, throw in, corner kick, goal kicks
TB	through balls: completed pass sent between defenders into open space
Sw	Switches: balls that travels more than 40 yards the width of the pitch
Crs	Crosses
Ti	Throw ins
CK	Crner Kick
In	Inswinging Corner kick
Out	Outswinging corner kick
Str	Striaght corner kick
cmp	pass result: completed
off	pass result:offside
block	pass result: blocked by opponents

4. Goal and Shot Creation

SCA	Shot Creation Action: dribling leading to foul, pass leading to shot, etc.
SCA90	SCA per 90 minutes
PassLive SCA	Complted live ball pass leading to shot attempt
PassDeadSCA	Complted dead ball pass leading to shot attempt
TOSCA	Successful take-on that leads to shot attempt
SHSCA	Shot that lead to another shot attempt
FldSCA	Foul drawn that lead to another shot attempt
DefSCA	Defensive action that lead to a shot attempt
GCA	goal Creation Action: dribling leading to foul, pass leading to goal, etc.
GCA90	GCA per 90 minutes
PassLive GCA	Complted live ball pass leading to goal
PassDeadGCA	Complted dead ball pass leading to goal
TOGCA	Successful take-on that leads to goal
SHGCA	shot that lead to a goal
FldGCA	Foul drawn that lead to another goal
DefGCA	Defensive action that lead to a goal

5. Defensive Performance

Tkl	Number of player tackled
TklW	Number of tackels that player won the possession of the all
Def3rd	Tackes in the deffensive 1/3
Min3rd	Tackes in the middle 1/3
Att3rd	Tackes in the attack 1/3
ChalAtt	Deffesnsive attempt against the dribbling player: successful and Unsuccessful
ChalTkl	Successful ChalAtt
ChalLost	Unsuccessful ChalAtt
Block	Total Number of Blocks by standing in tha pass of the ball
ShBlock	Shot blocks
PassBlock	Pass Blocks
Int	Interceptions
Clr	Clearances
Err	Errors leading to opponent's shot

5. Touches

Touches	Total Touches: taking the ball, dribling and passing or shoting are counted all 1 touch
TouchDefPen	Total Touches in deffensive penalty Area
TouchDef3rd	Total Touches in deffensive 3rd Area
TouchMid3rd	Total Touches in the middle 3rd Area
Touchatt3rd	Total Touches in offensive 3rd Area
Touch AttPen	Total Touches in offensive penalty Area
Take-on Att	attempt to past the defender by dribbling
Success Take-on	successful attempt of Take-on
Tkld Take-on	Unsuccessful Take on which is tackled by opponents.
Carries	Number of the time players control the ball with their feet
CarriesTotDist	Total distance of carrying the ball by player in yards in any direction
CarriesPrgDist	Total distance of carrying the ball by player in yards towards the opponent's goal line.
PrgC	Number of times a player carries the ball towards the opponent's goal line at least for 10 yards.
Carries 3rd	Number of times a player carries the ball to opponent's deffensive 1/3 area.
CPA	Number of times a player carries the ball to opponent's penalty area.
Mis	Miscontrols: number of times player failed to control the ball
Dis	Dispossessed
Rec	Number of passes received
PrgR	Number of progressive passes recevived

Other Statistics

CrdY	Number of Yellow Cards
CrdR	Number of Red cards
2CrdY	Number of 2nd yellow cards in a game
Fls	Folus committed
Fld	Fouls Drawn
Off	Number of offsides
Crs	Number of crosses
Int	Number of Interceptions
Tklw	Number of Tackels won
PKWon	Number of Penalty kicks won
Pkcon	Number of Penalty kicks conceded
OG	Own Goals
Recov	Ball Recoveries
Awon	Aerials Won
Alost	Aerials Lost