Sports Salaries Analysis - Professional Soccer

University of Southern California

Summer 2021

Abstract

The focus of this paper will be to explore the relationship between salaries and performance of professional athletes. This will address the general assumption that players who demonstrate higher levels of performance are compensated comparatively. The research will be conducted on professional soccer, specifically on English Premier League athletes by identifying objective variables to evaluate performance and comparing it to their salary.

In order to investigate this relationship we will perform a series of statistical analysis (likely Analysis of Variance, Correlation and Regression Analysis). In the soccer field, we found on Kaggle a Dataset that contains all season-related metrics of player's performance. Specifically, we will look at important metrics per position (e.g. for a soccer forward, the most important metrics are goals and assists, for a Goalkeeper the goals conceded and clean sheets).

Our hypothesis is that there is a strong positive relationship between performance and salaries. We also think this relationship differs across sports and across positions. Although performance plays a significant role on the salaries of professional athletes, we think they cannot fully explain it's variation. Other factors like contracts negotiations, player's status, etc. also play an important role when determining an athlete's income.

1. Introduction

The ability to offer compensation in alignment with performance and perceived talent is a central component of being able to retain and hire employees across any organization. While professional athletic organizations are known to pay their athletes some of the highest salaries in the world depending on the sport, it would be assumed that this same logic would be followed where compensation matches performance to retain the best talent. The notable contracts that are signed by the most popular athletes are reported to the public and often top major news headlines with contract figures in the millions per year. More specifically, high grossing sports and organizations that have large fan bases are known to pay their athletes accordingly.

One example of this is professional soccer, which is generally regarded as one of the most popular sports in the world, where their athletes consistently occupy any charts that rank the highest paid athletes. With such large amounts of money being paid out to certain athletes, it would seem reasonable to expect that the level of compensation would be in alignment with the athlete's contributions to the organization. Implementing various forms of data visualization and analysis, this paper will examine whether the salaries are a true reflection of performance for professional soccer players.

This paper will adhere to the following structure: Section 2 will introduce previous works of literature that have researched the relationship between pay and performance in different areas of sports. This will provide general guidance into analysis methods and will offer useful insight into results that can be expected. Section 3 will outline the hypothesis of this paper which will be largely drawn from the literature offered in section 2. In section 4, the datasets will be described in detail including where it was taken from, the quality of the data, and which variables within the dataset were found irrelevant and ultimately excluded. Further, section 4 provides a general outline of the analysis methods that will be used for the research. Section 5 provides in-depth information regarding the analysis of the data both through discourse and visualization and section 6 concludes the overall findings of the research. The final section is 7, which offers suggestions for further research on this topic.

2. Literature Hypothesis and Quotations

Pay and Performance of Major League Baseball

Research has previously been conducted on the relationship between performance and player salaries within the Major League Baseball (MLB) organization in the work “Pay and Performance of Major League Baseball” by Gerald W. Scully. While the nature of the two sports are different, the fundamental approach and goals of the research are similar in finding how performance and pay are interrelated. Within his work, Scully looks to examine the relationship between Major League Baseball players salary and their respective performance, in comparison to each player’s estimated marginal revenue production. The general theory behind the marginal revenue production variable is that it measures the effect of individual player performance on the amount of revenue that player brings to the organization. Although this paper will exclude the created variable of marginal revenue production as team costs and total revenues are not available in the chosen dataset, emphasis will be placed on Scully’s method of selection and ultimate use of performance variables and his analysis surrounding them. Scully first split the analysis by positions, hitters and pitchers, as performance for each of these positions are measured by completely different variables.

For example, for hitters the performance measures used are batting average and slugging average, while the strikeout-to-walk ratio and innings pitched are used to evaluate pitching performance. The number of years in the league was used across both positions, as salaries generally increase the longer a player is in the league. These variables are then used to run regressions against the salaries reported within the dataset for both pitchers and hitters. From this, there was a high R-squared for each regression that was calculated, in which Scully (1974) concluded that: “Overall, 78 to 81 percent of the variation in ballplayer salaries is accounted for. The remaining variance may reasonably be attributed to the vagaries of the bargaining process. However, as can be readily seen, the salary process in major league baseball is quite deterministic. The concept that ballplayer salaries are related to performance seems reasonably well confirmed” (p. 927).

Team Payroll Versus Performance in Professional Sports

While research has been shown that salaries have some relationship on performance at the individual level, it would be useful to look at how it impacts performance at the organizational level. As each cumulative individual performance contributes to team success, determining whether this relationship exists at the organizational level would further confirm the impact of salaries on performance. Grant Shorin looks to examine this in his work, “Team Payroll Versus Performance in Professional Sports: Is Increased Spending Associated with Greater Success?”, where he studies the salaries paid out by the team over a 10-year period and evaluates this against the team’s overall success.

Shorin researches teams within four different professional sports: MLB, NFL, NHL, and NBA. While evaluating organization performance, Shorin used variables such as win percentage and championships won. Regressions were run on all four sports for salary measured against win percentage, and the variable Simple Rating System, which was described as measuring how well a team performed in various key metrics against the league averages. In both of these variables, there was evidence that increases in payroll spending (how much players were paid) were associated with increases in win percentage, Shorin (2017) concludes: “Overall, all three regressions show a statistically significant relationship between payroll and regular season performance across all four leagues. This is consistent with previous studies on the MLB which found that higher spending was correlated with a better winning percentage” (p. 38). In regards to postseason performance, the results of the study varied by league as Shorin (2017) writes: “In sum, we see that when controlling for team-specific factors, there is a statistically significant relationship in the NBA and NHL between spending and a team’s odds of winning the championship, while no such relationship existed in the MLB and NFL. Each additional payroll standard deviation is associated with a twofold increase in the likelihood that a team wins a title” (p. 42). Despite this, both Scully’s study on individual performance and relative salary and Shorin’s study on organizational performance and total payroll point to some evidence that a relationship exists between salaries paid and relative performance.

3. Hypothesis

Based on the conclusion of the literature referenced above, there is evidence that suggests a relationship exists between performance and salaries paid in other professional sports. Gathering evidence from both the individual salaries and performance along with the organizational level only furthers this assumption This would make logical sense, as a player who has shown better performance can often demand higher compensation. This holds especially true when a player is talented enough to receive multiple offers from different teams, which can drive up their compensation even further. Drawing from this general assumption and from the insights presented in the literature, our hypothesis is as follows:

The null hypothesis is presented as:

Ho: The players who demonstrate higher levels of performance will obtain higher levels of salary for professional soccer players of the English Premier League.

The alternative hypothesis:

Ha: There is no relationship between levels of performance exhibited and salaries paid to professional soccer players of the English Premier League.

4. Data and Statistical Analysis

We found a very rich dataset, available on kaggle.com, containing a wide variety of important performance metrics of Professional Soccer players playing in the English Premier League season 2018-2019. The dataset contains observations of 304 soccer players but did not include the players salary. We found another publicly available dataset containing only the English Premier League Salaries and merged these two together for the purpose of our analysis. Since the performance metrics were from season 2018-2019, we used the players salary from the next season (2019-2020) since a good/bad performance now affects the compensation in the future. So if a player performed well in this season, his compensation should or would be affected in the next season. In total, the combined dataset had 49 dimensions (columns) but for our analysis, we ended up using just a few of them.

Data Quality

We started by leaving out columns like birthday, league, season, nationality and current club, since they were either not relevant for the purpose of our study or are all the same for every player. Second step of the preliminary analysis was to adjust the columns to their most appropriate type (e.g Position as an ordinal factor, etc.). Initial data review allowed us to identify that every important performance-related column was a numeric variable. There were no missing values in any of the columns, and most of the dimensions seem to have reasonable values, with the exception of 4 columns, namely:

- Rank_in_league_top_attackers
- Rank_in_league_top_midfielders
- Rank_in_league_top_defenders
- Rank_in_club_top_scorer

Figure 1 shows the distribution of the values of these 4 columns in the dataset across positions. From here we can already see some illogical values in the 4 columns, the first thing being that almost all positions have values in all rankings, where it would make sense that only 1 position should have a ranking value for its corresponding ranking column (E.g defenders should have a ranking in top defenders, the same for midfielders). Further on, we see Goalkeepers with a range of values from -1 to 400 in Top Attackers ranking, Defenders in the top 10 scorers and even Goalkeepers in the Top scorers ranking. For the sake of simplicity, and considering that this is not even an objective and pure performance metric, but a subjective ordinal ranking, whose methodology and source we do not know, we decide to drop these 4 columns due to Quality and Reliability issues.

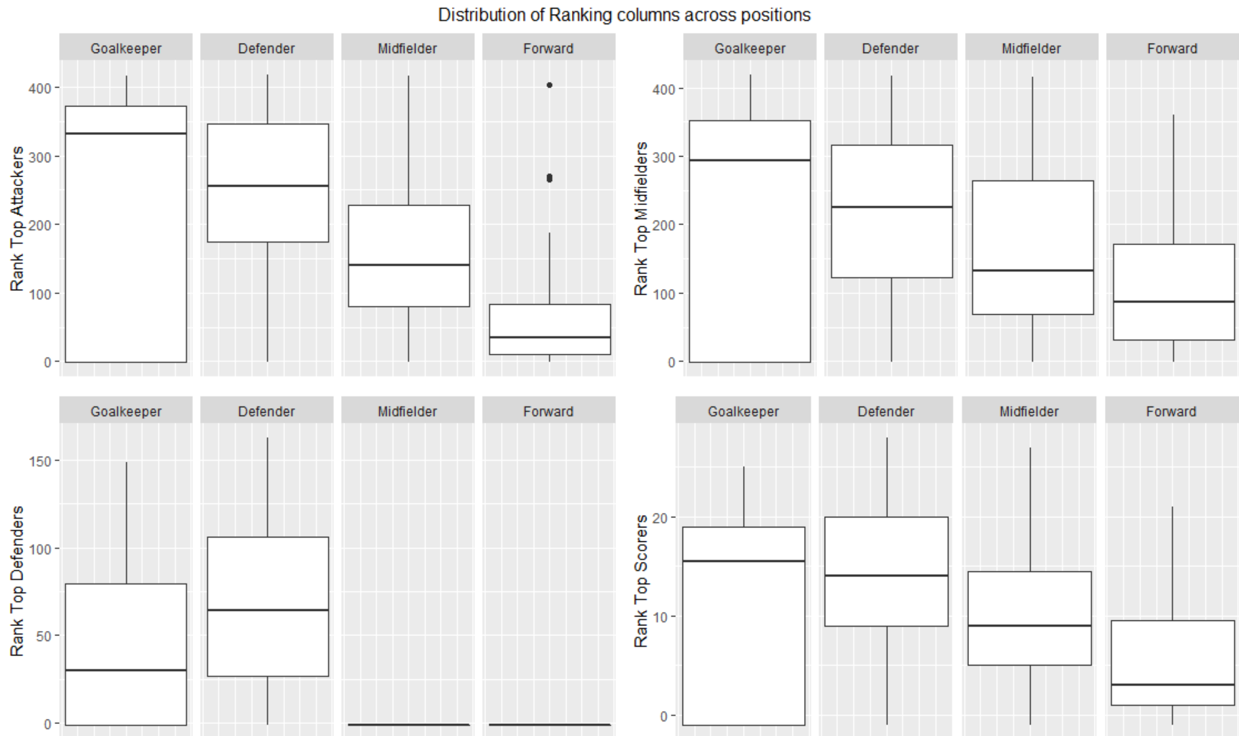

Figure 1: Distribution of Ranking columns across positions

Multicollinearity

A good practice before doing any statistical analysis and/or predictive analysis is to check if the dimensions of the dataset are correlated, in order not to have the information twice and to keep only the columns that are adding value and not redundant information. For this purpose a Correlation analysis was performed across all columns in the dataset. As expected we found a lot of strongly correlated columns, which needed to be removed. As shown in Figure 2 there are many positive strong correlations in the variables, suggesting that many of them are redundant and should be excluded.

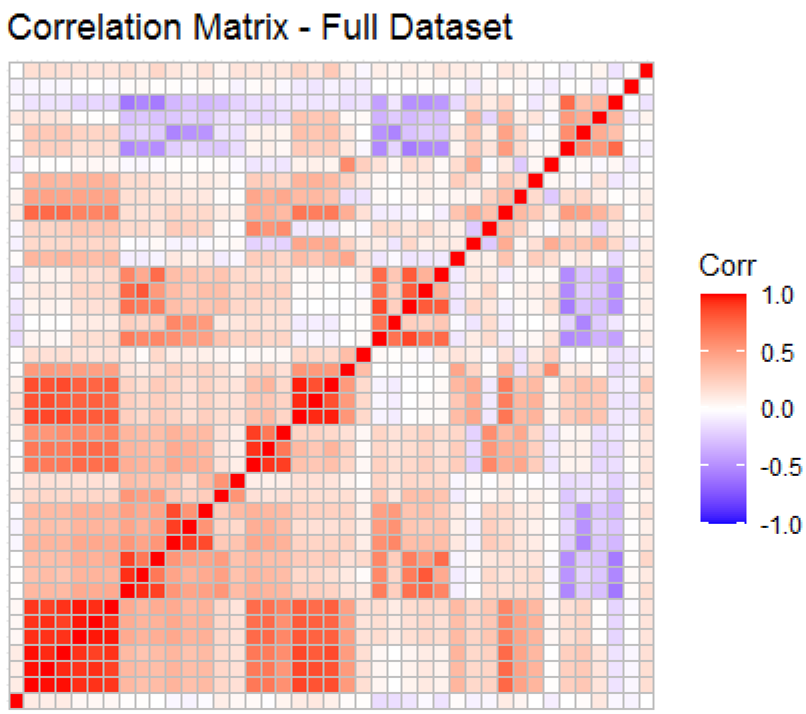

Figure 2: Correlation Matrix of the entire dataset

We first addressed the most direct one: Overall vs. Home and Away metrics. In the dataset there are several columns derived by others, in particular the 'Overall' and the 'Home' & 'Away'. Logically, the relationship holds for any metric:

$$Overall = Home + Away$$

These columns just discriminate a player's appearance based on if they play home or away in the game and consequently are highly correlated (Correlation average of 0.97 between overall and home and 0.96 between overall and away). For the purpose of this study, we don't care if the player has played in a home or in an away game, or if he has scored in a home or an away game, the important metric is that he has played and he has scored. Consequently we decided to drop all the 'components' (home and away) of the information and just keep the general column, namely the overall. The dropped columns were:

- Minutes played (home and away)
- Appearances (home and away)
- Goals (home and away)
- Assists (home and away)
- Clean sheets (home and away)
- Conceded (home and away)
- Goals per 90 (home and away)

This step reduced the size of the dataset considerably and allowed us to focus only on the ‘purest’ metrics.

Further analysis in correlated metrics found that Minutes played and Appearances had a highly positive correlation (0.95). Since Minutes played englobes more information (A player could have played only 1 minute of one game and that is equivalent to 1 appearance) on the performance of the player, we decided to keep this column and remove Appearances.

We also found in the dataset other important correlations. Metrics like ‘Minutes_per_goal’, ‘Assists_per_90’ or ‘Min_per_Card’ are calculations derived from the original data which are ‘Goals’, ‘Assists’, ‘Cards’ and of course ‘Minutes_played’. Although we know that these columns are correlated, we decided to keep them for predictive modelling since they could be stronger predictors of salary than the original metrics.

Dataset Division

Since we were interested in studying the relationship between performance metrics and salaries, soccer, like many other sports, has different positions with different specific functions. Although the team itself works as one, in general Goalkeepers and Defenders are responsible for keeping the team from receiving goals, and Midfielders and Forwards are to some extent responsible for scoring the goals. In consequence, the performance in different positions cannot be measured by the same set of metrics, a Defender should not be evaluated for his ability to score goals, but for stopping opponents from scoring. Midfielders should be evaluated on their ability to create plays, provide assists and score goals more than defensive attributes. So each position has a set of metrics that best describes it’s performance. This is why we splitted the dataset in 4 subsets, namely: **Goalkeepers**, **Defenders**, **Midfielders** and **Forwards**to analyze in a much more appropriate way the relationship between performance and salary.

We later selected the most appropriate performance metrics for each position before running any further analysis. This step is to ensure that the analysis does not include any misleading or irrelevant information, for example: it does not make any sense to evaluate a defender for their amount of goals scored or an attacker for the amount of goals conceded because that is beyond their function. However there are some metrics that correspond to every position, like minutes played, cards received, among others. For each subset, we selected the most reasonable and appropriate columns.

Statistical Analysis

In order to analyze the relationship between salary and performance we performed a series of statistical analysis, which include:

- Correlation Analysis
- Analysis of Variance
- Multiple Linear Regression

The Correlation Analysis was carried out to see if there is a natural strong correlation between any of the performance metrics with Salary. If there is, then it would possibly mean that that metric is a strong predictor of salary. Similarly, strong negative correlated metrics with Salary would hint us to understand what measures negatively affect a player's salary. We carried out the correlation analysis in each position, considering the most important metrics.

We also performed an Analysis of Variance or ANOVA to test if there is a statistically significant difference in the means of salary in the 4 different positions, and if there is, then among which groups. This step was in order to better understand the salary column in the dataset and to see if there exists a significant difference between groups or positions, as is the case in many professional sports.

Lastly, we performed Multiple Linear Regressions to try to predict the player's salary based on their performance. This technique also helps to determine what predictors are statistically significant and which of them are the strongest. We performed 4 Linear regressions, one for each position, using the appropriate performance metrics. The output, or dependent variable in all cases was Annual_salary and the independent variables differed across positions, but included goals, assists, goals conceded, cards, minutes played, etc.

5. Results and Discussion

Initial Review

Initial analysis of the dataset shows no significant difference of the salary across positions. Figure 3 shows what it seems like a normal but left skewed distribution of salaries, where most of the goalkeepers, defenders, midfielders and forwards earn around 2,500-3,000 Thousand pounds a year (peak of the distributions) with a few outliers earning more than 10,000 Thousand Pounds a year.

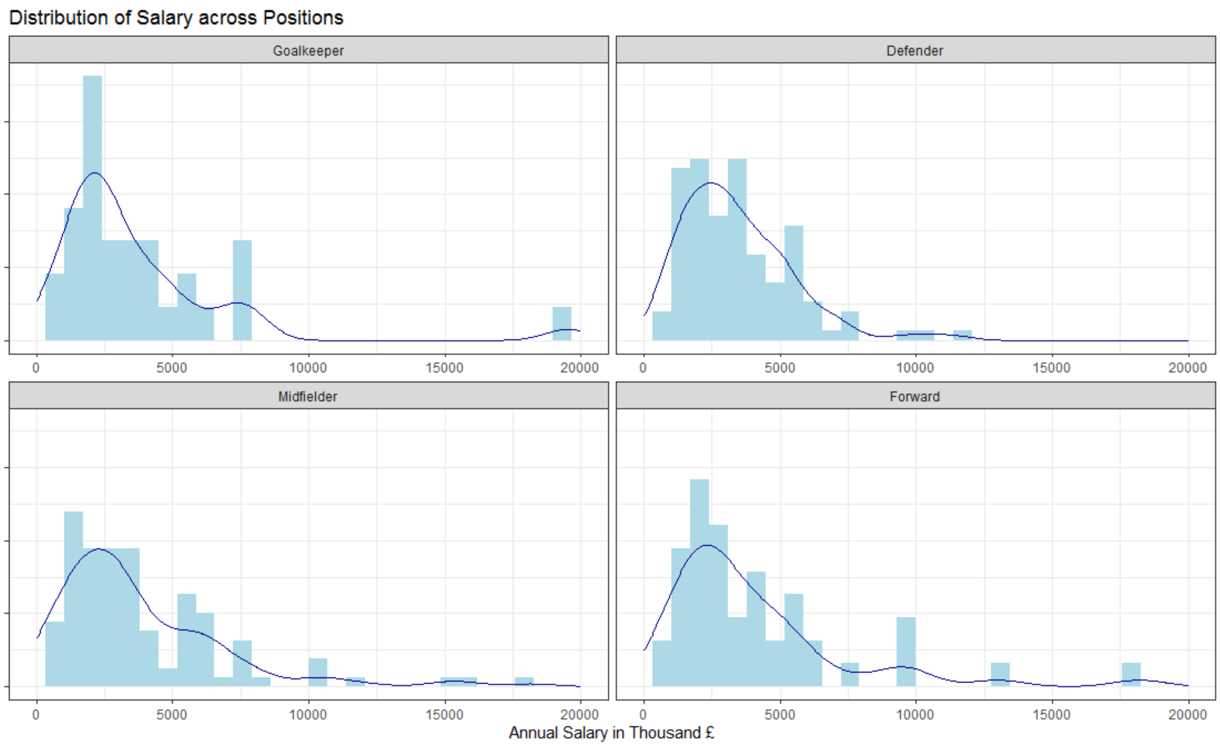

Figure 3: Distribution of Salaries across Positions

From the boxplot graph in Figure 4 we can also see that there is no significant difference between the distributions. Goalkeepers seem to earn in general less than the other positions and midfielders and attackers have a wider range of salaries than the rest. However, no significant conclusion can be made from this section.

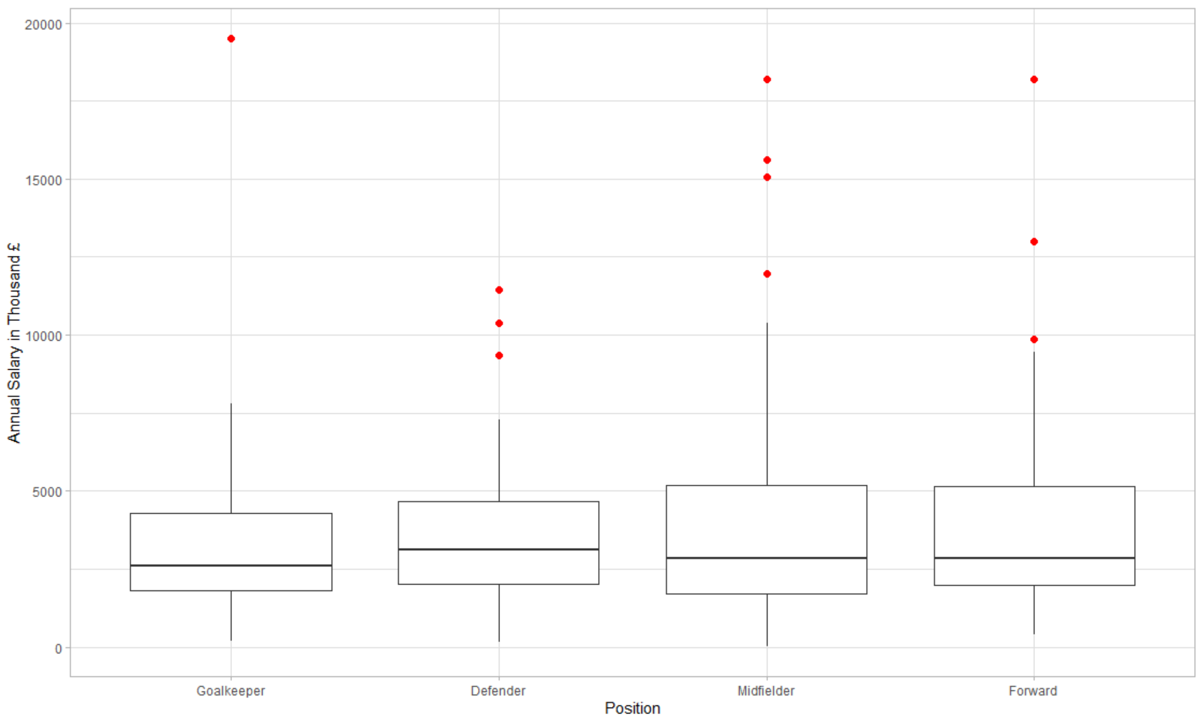

Figure 4: Boxplot of distribution of salary across positions

Correlation Analysis

Taking the most important performance metrics from every position and running the correlation analysis of each metric with salary, yield the following results:

Among the **goalkeepers**, we find that although the signs of the correlation are logical, they are not strong enough to state that they could be good predictors of salary. Age seems to be the variable with the strongest correlation but is also not too strong and age is not an actual metric of performance.

Performance Metrics	Correlation with annual_salary
age	0.144
minutes_played_overall	0.015
clean_sheets_overall	0.051
conceded_overall	-0.055
cards_per_90_overall	-0.045

Table 1: Correlation between performance in **goalkeepers** and salary

Among the **defenders**, we find similar patterns as in the goalkeepers: most of the signs of the correlations make sense (e.g. minutes_played and goals impact positively on salary) but the values are not significant. The correlations are too low to state that they are potentially good predictors of salary. However, we see a hint with the signs, that at least most of them make sense in terms of logic. Assists and goals_involved_per_90_overall however are not logically correlated with salary, but the value is also close to 0.

Performance Metrics	Correlation with annual_salary
age	0.070
minutes_played_overall	0.017
goals_overall	0.108
assists_overall	-0.031
clean_sheets_overall	0.149
conceded_overall	-0.088
cards_per_90_overall	-0.085
goals_involved_per_90_overall	-0.059
min_per_match	0.070

Table 2: Correlation between performance in **defenders** and salary

Among the **midfielders**, we start to find more illogical values and relationships. Examples are negative correlations between minutes played, goals, assists with salary. Although these negative correlations do not make sense, the magnitude is also not so significant (a relatively low value, like most of the correlations). These results start to question our hypothesis and previous findings in goalkeepers and defenders. However the same pattern of low magnitude in correlation remains. Age is the strongest predictor with -0.2.

Performance Metrics	Correlation with annual_salary
age	-0.200
minutes_played_overall	-0.158
goals_overall	-0.137
assists_overall	-0.079
penalty_goals	-0.058

penalty_misses	-0.059
cards_per_90_overall	-0.159
goals_involved_per_90_overall	-0.103
min_per_match	-0.183

Table 3: Correlation between performance in **midfielders** and salary

Lastly, among the **forwards** we see again similar values and signs as in the defenders and goalkeepers. In general, relatively low magnitude of correlations but with logical signs. The variable assists is the highest correlation, which can be interpreted as reasonable since assists is an important performance indicator.

Performance Metrics	Correlation with annual_salary
age	0.005
minutes_played_overall	0.087
goals_overall	0.058
assists_overall	0.157
penalty_goals	0.108
penalty_misses	-0.166
cards_per_90_overall	0.038
goals_involved_per_90_overall	0.001
min_per_match	0.115

Table 4: Correlation between performance in **forwards** and salary

In general, from the correlation analysis across all positions, we can state that there is not a single significant relationship between any of the variables and salary, since the values do not exceed 0.20 in magnitud (low correlation). There are logical correlations between some performance indicators and salary but they are not high enough to be significant, however there are some correlations. Indicators like minutes played, clean sheets, minutes per match and assists are relatively important metrics in goalkeepers, defenders, midfielders and forwards respectively.

We also took what we considered the most important performance indicator for each position and plotted it against the salary in a scatterplot, to see if visually there is some kind of relevant pattern.

For goalkeepers, we selected the number of clean sheets, for defenders the amount of goals conceded, for midfielders the assists and for attackers the amount of goals scored. Results are shown in the figure below. We would expect (for Goalkeepers, Midfielders and Forwards) to see that the higher the metric, the higher the salary, since the indicators chosen are all positive indicators. For Defenders we would expect to see a negative relationship. However, results point out that visually, there is no clear sign of any of the described relationships. In the Defenders scatterplot, there seems to be a slight shift of the points to the right when the numbers of goals conceded increase. For the Forwards, one could argue as well that to some extent the salaries of top scorers are higher. However none of these relationships are clear.

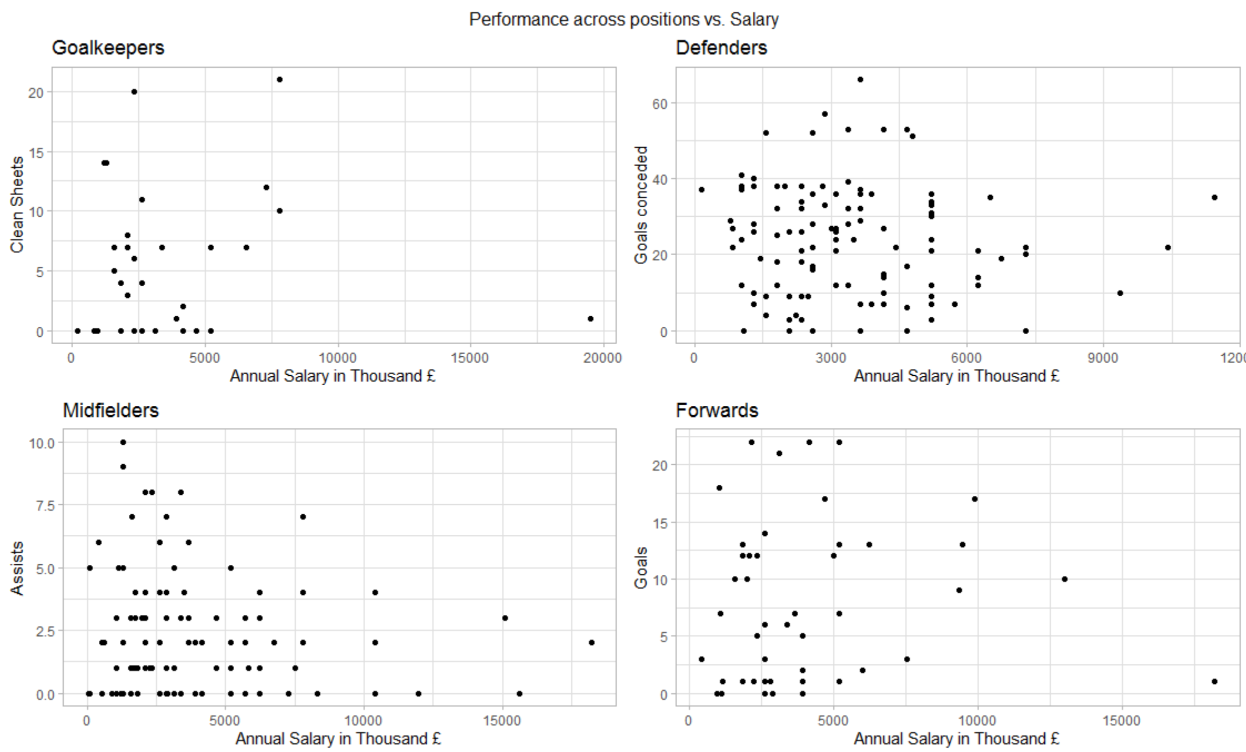

Figure 5: Scatterplot of most important performance metrics vs. Salary across positions

Analysis of Variance

The analysis of variance between salary and positions resulted in a F-value of 0.65 and a P-value of 0.59, which is considerably higher than 0.05. With these results accepted the null hypothesis that there is no statistically significant difference between the variances of salaries across the positions at a 5% confidence interval. This statement agrees with our previous results of the visual analysis of the distribution of salaries. Additionally, since we have 4 groups in the independent variable position, we performed post hoc tests for analysis of variance to see if there were any statistically significant differences between any of the pairs of groups involved. We performed a Tukey multiple comparisons of means and results are shown in the following figure:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = annual_salary ~ position, data = df)

$position
                           diff        lwr     upr     p adj
Defender-Goalkeeper   -279307.9 -1778596.4 1219981 0.9631979
Midfielder-Goalkeeper  106465.5 -1386976.8 1599908 0.9977793
Forward-Goalkeeper     353458.4 -1366615.4 2073532 0.9515188
Midfielder-Defender    385773.4  -608480.9 1380028 0.7481140
Forward-Defender       632766.3  -677510.7 1943043 0.5970342
Forward-Midfielder     246992.9 -1056590.4 1550576 0.9613962
```

Figure 6: Results of the Tukey multiple comparisons of salary means across positions

As shown in the figure, none of the p-adj values is below 0.05, meaning that the variance in salary in each pair of positions is not statistically significantly different than the other pair.

Multiple Linear Regression

The results of fitting multiple linear models to predict salary by the most important metrics across positions is shown in the following table:

Position	**Linear Model**	**R2**	**Significant Variables**	**Linear Model with significant variables**	**R2**
Goalkeepers	Salary = age + minutes_played + clean_sheets + conceded + min_per_match + cards_per_90	0.26	conceded, min_per_match	Salary = Conceded + min_per_match	0.16
Defenders	Salary = age + minutes_played + clean_sheets + conceded + assists +	0.10	None	-	-

	goals + goals_involved_per_90 + min_per_match + cards_per_90				
Midfielders	Salary = age + minutes_played + assists + goals + penalty_goals + penalty_misses + goals_involved_per_90 + min_per_match + cards_per_90	0.10	Age	Salary = Age	0.04
Forwards	Salary = age + minutes_played + assists + goals + penalty_goals + penalty_misses + goals_involved_per_90 + min_per_match + cards_per_90	0.12	None	-	-

Table 5: Results of Multiple Linear Regressions

The predictive power of the models are in general weak since the R2 do not exceed 0.26, being the goalkeepers the best predicted group among the four. Defenders, Midfielders and Forwards models present a much weaker R2, meaning that a low percent of the variance in salaries in these positions can be explained by the performance metrics.

In general, most of the models also present just a few statistically significant predictors, if none. For Goalkeepers, the only statistically significant predictors are goals conceded and minutes per match, while in Midfielders is age. Among Defenders and Forwards, no predictor variable is statistically significant.

These results strongly question our hypothesis and the expected results, since they are a proof that at least in this dataset and with this analysis, performance alone cannot explain a player's salary.

Limitations and Weaknesses

From the analysis of our Dataset and the results presented, we can state that our approach and data has certain limitations and weaknesses that could have affected our final results and driven us to the conclusion that salary and performance are not so strongly related in Soccer. These limitations are listed below:

- Limited sport, limited league: We have only performed our analysis on one particular Sport and one particular league. In order to drive more robust and trustworthy insights, several sports and several leagues within these sports should be analyzed simultaneously and with the same approach.
- Short time observations: We have only performed our analysis on the effect of one season performance on the salary in the next season. In professional sports, a player can have a really bad or really good season but that does not mean that they will necessarily get a raise or a reduction in the next season. Usually salary variations are a result of either long-time good/bad performances and/or club transfers (which are related to good/bad performances). A good idea would be to analyze the performance metrics of several seasons and compare them to salaries.
- Time gap: As it was mentioned before, the performance of one player in a current season may not directly affect his salary next season. Sometimes these changes take years and can even be affected by other outside factors.
- Complementary metrics: Although we have found really good performance metrics in a very rich dataset, we believe that maybe these metrics do not fully capture a players performance. Soccer is a complex sport and trainers and headhunters don't always look at just goals and assists, but other important metrics like tackles won, plays made, pass efficiency, among others. We believe that this sort of metrics could be extremely beneficial for our analysis
- Just performance vs salary: Our approach was to analyze the effect of performance on a player's salary. This relationship may not be as direct as we thought, so a good approach would be to analyze more directly related metrics like performance vs. market value or even perhaps performance vs. minutes played next season.
- Outside factors: As we mentioned before, sometimes a player's salary does not only depend on his performance, but also on many outside factors like the player's contract at the time he arrived at the club, the negotiations he made, the status of the player, the status of the club, etc.
- Limited Models and Statistical Analysis: Although the methods and models that we used are also widely used, we consider that they are mostly used in a preliminary approach and sometimes present several weaknesses. We believe that by using more sophisticated predictive models like Random Forest or Neural Networks, a deeper and more significant part of the relationship between performance and salary could be captured.

6. Conclusion

Despite the evidence presented in the literature referenced above, the same relationship between performance and salaries paid does not appear to exist within the data analyzed for professional soccer players of the English Premier League. Contrasting with Scully's research of Major League Baseball, one explanation for the varying results could be that the nature of the variables examined were ultimately just different. For example, Scully's study used career statistics to measure performance, whereas this study used statistics that were taken from only one season. Compared to Shorin's study of organizational performance and total salaries paid, the differing results could be explained by the fact that the comparison of individual salaries and performance does not translate in a similar way to the organizational level. In both studies, soccer was not a sport that was specifically studied which could potentially reveal evidence that this relationship just does not exist within this sport and in this particular league analyzed.

The data, approach and analysis conclude in the fact that we cannot state there is a significant relationship between higher performance and higher salaries in English Professional Soccer, thus the Ho or Null Hypothesis is accepted.

7. Further Research

The limitations and weaknesses of our data, methodology and results leave room open for further and more meticulous research on several lines. First of all, a broader set of sports and leagues should be analyzed in order to arrive at more meaningful and robust conclusions. Second, a longer set of observations should be included in order to address the fact of time differences and time gaps between cause and consequence.Third, we believe that not only the main important metrics should be included in each analysis, but those that are not so easily captured and usually analyzed as well. These sorts of metrics tell a lot about the player and we believe they have a strong effect on the perception and consequently salary of the player. Lastly, we encourage to analyze other outcome variables as well like a players Market value and to include several outside factors like player contracts, negotiations and status somehow in the analysis. Consequently, we believe that a stronger and more sophisticated set of models and statistical analysis methodologies could be used.

References

- Scully, G. W. (1974a). Pay and Performance in Major League Baseball. *The American Economic Review*, *64*(6), 915–930. https://www.jstor.org/stable/1815242
- Shorin, G. & Duke University. (2017). *Team Payroll Versus Performance in Professional Sports: Is Increased Spending Associated with Greater Success?* https://sites.duke.edu/djepapers/files/2017/06/grantshorin-dje.pdf