Examining the Relationship between Player's Performance and Age in English Premier League

---

DATA-231 Final Project

Yongchan Lee

12/10/2024

Introduction

According to the Federation of Internationale de Football Association, as known as FIFA, more than half of the world population 5 billion people engaged with the FIFA World Cup Qatar 2022 (FIFA, 2023; UN, 2022). This statistic demonstrates the incredible worldwide interest in soccer, and it is clear that the World Cup plays a major role in the sport's immense popularity.

In the popularity of soccer, not only the World Cup but also national leagues play a significant role. Leagues with some of the largest fan bases are found in Europe, such as the English Premier League, Germany's Bundesliga, Spain's La Liga, France's Ligue 1, and Italy's Serie A. The English Premier League, for example, first began in 1992, giving it a long history of around 32 years. Over this period, various types of data have been collected, allowing analysts to predict league winners or outcomes of specific matches. Detailed player data has also been gathered, enabling analyses of players from multiple perspectives. A variety of variables are used to analyze players. However, the main question of this study is: Does a player's age affect their performance outcomes at the end of the season in the English Premier League?

According to one study on elite athletes, younger athletes performed better in terms of explosive power and sprint ability, while endurance improved with age (Allen, S. V., et al., 2015). Soccer, however, is a sport that requires multiple athletic abilities: explosive strength, good endurance, and quick decision-making. A study examined the relationship between age and both the physical and technical performances of soccer players; younger players excelled in physical performance, while older players performed better in technical aspects (Sal de Rellán-Guerra, A., et al.,2019). As soccer players age, they tend to show improvements in certain areas while declining in others. Given these findings, it seems challenging to conclude definitively that age has a consistent impact on player performance. Similarly, other studies indicate that age may not have a strong influence on overall player performance. Research focused on Premier League players found that only in the forward position did younger players show better performance, while age was not strongly related to performance in other positions (Jamil, M., & Kerruish, S. ,2020).

Defining "player performance" can be challenging; however, in this study, performance will be defined as numerical values Goals + Assists per 90 minutes, excluding penalty kicks (G+A-PK_90). The goal of this research is to create a predictive model for player performance, focusing primarily on how age impacts performance. In summary, G+A-PK_90 will serve as the response variable, and age will be the main explanatory variable. Other confounding variables are listed in *FigureA1*.

Method

The data is sourced from FBref which, established in 2000, provides statistics for various sports, including soccer, basketball, baseball, and golf, across 47 countries. They collected data through the manual tracking of football matches worldwide, powered by StatsBomb. The dataset contains 34 attributes, and only 13 attributes were selected and used for this study (FBref, 2024). The remaining 21 attributes were not selected because goals and assists were already included in the calculation formula. Including them would undermine the purpose of building the model, so they were excluded.

Detailed descriptions of the data are shown in *FigureA1*. This research includes data from all players who played more than 90 minutes in the 2023-2024 season. The population size is the same as the sample size, encompassing all Premier League players who played more than 90 minutes in the specified season. Although using multiple seasons' data through random sampling could define the population as all Premier League players, I chose a single season to avoid potential data dependence issues from repeated player appearances.

The dataset with 13 attributes had no missing values. However, important data cleaning steps were applied. Rows containing “GK” were removed. Goalkeepers are excluded from the Player Stats data since their primary role is to save goals, not to score or assist. One might think that defense is not related to assists and goals, but in modern soccer trends, defenders have become key contributors to assists. Therefore, they were included. Additionally, rows for players who did not play at least 90 minutes were removed, as at least 90 minutes is required to calculate the response variable *G+A-PK_90*. These adjustments reduced the sample size from 581 players to 456 players.

There were trivial modifications made to the attributes *Pos*, *Team*. For the variable *Pos*, there were originally 8 positions after removing the GK position. The values in the "Pos" variable consist only of "FW," "MF," and "DF." The Pos attribute contains pairs like (DF, MF), (MF, DF), (MF, FW), and (FW, MF), which may appear to be duplicates. However, these pairs represent different positions: the first position is the primary position, and the second position indicates a secondary position that the player also plays. Since the values contain commas, I have removed all commas from the Player Stats data; for example, the original form “DF,MF” changes to the “DFMF.”

For the variable *Team*, there are 20 unique values, and each team is stored under its full name. I have changed all teams' full names to their abbreviated names. For example, the original form “Manchester United” changes to the abbreviated form “MUN.”

The last important modification made to the attribute *Nation*. There was a variable named *Nation*, which contained 65 unique countries. However, since 17 countries have only one player and another 17 countries have fewer than four players, there was a high likelihood that unrepresented countries would appear in the test set after splitting the dataset into training and test sets. To address this, I created a new attribute *Continent* based on the attribute *Nation*. The attribute *Continent* contain 5 categories: Africa, Asia, Europe, North America, Oceania, and South America.

Since the response variable G+A-PK_90 is numerical, and there are more than 1 explanatory variables, a multiple linear regression model will be used. Before modeling, the 456 samples were split into a training set and a testing set in a 0.75 : 0.25 ratio. This resulted in 342 samples for training and 114 for testing. The model was trained using the training set and then evaluated on the testing set.

Among the 7 models, shown in *Figure A11*, the final model was selected based on the four criteria: first, the adjusted R^2 value from the model trained by the training set; second, the correlation between the predicted and actual G+A-PK_90 values on the testing set; and third, the Mean Squared Error (MSE) between the predicted and actual G+A-PK_90 values on the testing set; and fourth, the number of parameter (simplicity) of the model.

The significance of each indicator in the model was assessed through individual t-tests, and the overall usefulness of the model was evaluated using an ANOVA table. The adjusted R^2 value provided insight into how much of the variability in G+A-PK_90 could be explained by the model. The Confidence Interval (CI) for each coefficient in the model presents the range of values within which we are 95% confident that the true coefficient lies. Finally, a VIF test was used to check for any multicollinearity among the variables in the model.

Results

The primary goal of the study is to find a relationship between players' age and their performance during the 23-24 Premier League season. In other words, the main objective is to create a model that predicts a player's G+A-PK_90 based on their match data. The response variable will be G+A-PK_90, while the explanatory variables will include Continent, Pos, Team, Age, MP, Starts, Min, CrdY, CrdR, PrgC, PrgP, and PrgG. The numerical variables five number summary and distributions are shown in *Figure A2* and *Figure A3*. Before modeling, I examined the relationships between each explanatory variable and the response variable.

Twelve explanatory variables were used to create seven models. As explained in the Method, the final model was selected based on four criteria: Adjusted R^2, Correlation between predicted and actual

G+A-PK_90, Mean Squared Error, and Simplicity. *Figure A11* shows that the model with the highest Adjusted R^2 is Model7, with a value of 0.5221. However, when comparing the correlation and the MSE values, Model7 demonstrates the lowest correlation and the highest MSE among the seven models.

The next highest Adjusted R^2 value is observed in Model3, which was generated through stepwise regression. When comparing the correlation and MSE values, Model3 shows the second-highest correlation and the lowest MSE values. While Model3 shows the second-highest correlation, Model6 shows the highest correlation and fourth-lowest MSE strong performance, but Model_3 was selected as the final model due to its simplicity. Model3 has 30 parameters compared to 38 parameters of Model6. In conclusion, Model3 was chosen as the final model due to its simplicity and competitive performances: Adjusted R^2, MSE, and correlation between actual and predicted value .

In the final model Model3, there are 4 explanatory variables: *Team*, *Pos*, *Age*, and *PrgG*. While variable Team has 20 categories, variable Pos has 8 cateogories. In the fitted model, however, there are 1 indicator for Age, 1 indicator for PrgG, 7 indicator for Pos, and 19 indicators for Team. Loss of one indicators "*ARS"* for *Team* and "*DF"* for *Pos* is caused because of the concept of a reference category, or dummy variable, and the reference category is presented as an intercept. The intercpet will be interpreted later in Results.

Variable	Indicator	Coefficient	Standard Error	P-value	95% Confidence Interval (CI)
Team	AVL	0.03910	0.06925	0.57274	(-0.09715589, 0.1753564)
	BHA	-0.10770	0.06426	0.09476	(-0.23414753, 0.0187446)
	BOU	-0.04267	0.06653	0.52177	(-0.17358194, 0.0882390)
	BRE	-0.02647	0.07147	0.71135	(-0.16710371, 0.1141580)
	BUR	-0.18838	0.06617	**0.00471**	(-0.31858672, -0.0581833)
	CHE	-0.05234	0.06704	0.43550	(-0.18425663, 0.0795627)
	CRY	-0.10310	0.07207	0.15357	(-0.24492827, 0.0387119)
	EVE	-0.19182	0.06815	**0.00520**	(-0.32592622, -0.0577138)
	FUL	-0.01057	0.07043	0.88076	(-0.14915818, 0.1280095)
	LIV	0.01295	0.07225	0.85779	(-0.12921675, 0.1551342)
	LUT	-0.08456	0.06585	0.20003	(-0.21413488, 0.0450039)
	MCI	0.05804	0.07194	0.42036	(-0.08350654, 0.1996048)
	MUN	-0.12942	0.07459	0.08370	(-0.27619568, 0.0173390)
	NEW	0.02327	0.06974	0.73880	(-0.11395552, 0.1605105)
	NFO	-0.10092	0.06542	0.12392	(-0.22965963, 0.0278003)
	SHU	-0.16426	0.06647	**0.01401**	(-0.29506458, -0.0334660)
	TOT	-0.01359	0.06607	0.83708	(-0.14360331, 0.1164073)
	WHU	-0.11296	0.06940	0.10462	(-0.24952503, 0.0235977)
	WOL	-0.12486	0.06656	0.06162	(-0.25583287, 0.0061128)

Figure1-1. Summary of variable Team in the final model

Figure 1-1 shows the summary of the first variable Team with its 19 indicators. Teams that positively influence G+A-PK_90 include AVL, LIV, MCI, and NEW, while teams with a negative impact are BHA, BOU, BRE, BUR, CHE, CRY, EVE, FUL, LUT, MUN, NFO, SHU, TOT, WHU, and WOL. Among the 19 indicators, the standard errors range between 0.065 and 0.075. In the t-test, only three indicators—BUR, EVE, and SHU—show a p-value smaller than 0.05. In context, holding Pos, Age, and PrgG constant, there is strong evidence that there is a significant relationship between indicators (BUR, EVE, SHU) and G+A-PK_90, while there is no evidence that there is a significant relationship between indicators (the rest of 16 indicators other than BUR, EVE, SHU) and G+A-PK_90.

The 95% confidence intervals for these three indicators are as follows: BUR (-0.3186, -0.0582), EVE (-0.3259, -0.0577), and SHU (-0.2951, -0.0335). In all three cases, both the lower and upper bounds of the confidence intervals are consistently negative. For the remaining 16 indicators, the 95% confidence intervals, as shown in *Figure 3-1*, exhibit a negative lower bound and a positive upper bound. In context, we are 95% confident that the true effect of BUR, EVE, and SHU on G+A-PK_90 is negative, suggesting that players in these teams tend to have a detrimental impact on non-penalty goals and assists per 90 minutes. However, for the other 16 indicators, the confidence intervals including zero indicate uncertainty about the direction of their effect.

Figure 1-2 shows the summary of the variable *Pos* with its 7 indicators. Positions that positively influence G+A-PK_90 include DFFW, FW, FWMF, MF, MFDF, MFFW, while a position with a negative impact is DFMF. Among the 7 indicators, the standard errors range between 0.034 and 0.0102. In the t-test, only three indicators— FW, FWMF, MF, MFFW—show a p-value smaller than 0.05. In context, holding Team, Age, and PrgG constant, there is strong evidence that there is a significant relationship between indicators (FW, FWMF, MF, MFFW) and G+A-PK_90, while there is no evidence to say that there is a significant relationship between indicators (DFFW, DFMF, MFDF) and G+A-PK_90.

Variable	Indicator	Coefficient	Standard Error	P-value	95% Confidence Interval (CI)
Pos	DFFW	0.0192	0.10186	0.84996	(-0.18114067, 0.2197131)
	DFMF	-0.01623	0.05431	0.76527	(-0.12311003, 0.0906454)
	FW	0.30555	0.03401	**< 2e^{-16}**	(0.23862446, 0.3724866)
	FWMF	0.30701	0.03816	**1.80e-14**	(0.23193288, 0.382105)
	MF	0.08077	0.02966	**0.00683**	(0.02240911, 0.1391342)
	MFDF	0.02827	0.07545	0.70808	(-0.12018287, 0.1767408)
	MFFW	0.22070	0.03853	**2.39e^{-8}**	(0.14489064, 0.2965149)

Figure1-2. Summary of variable Pos in the final model

The 95% confidence intervals for these four indicators are as follows: FW (0.23862, 0.37249), FWMF (0.23193, 0.38211), MF (0.00224, 0.13913), and MFFW (0.14489, 0.29651). In all four cases, both the lower and upper bounds of the confidence intervals are consistently positive. For the remaining 3 indicators, the 95% confidence intervals, as shown in *Figure 3-2*, exhibit negative lower bounds and positive upper bounds. In context, we are 95% confident that the true effect of FW, FWMF, MF, and MFFW on G+A-PK_90 is positive, suggesting that players in these positions tend to have a favorable impact on non-penalty goals and assists per 90 minutes. However, for the other three indicators, the confidence intervals including zero indicate uncertainty about the direction of their effect.

Figure 1-3 shows the summary of the variable *PrgG. PrgG* positively influences G+A-PK_90, with a coefficient of 0.00078. The standard error of *PrgG* is 0.00017, and the

Variable	Indicator	Coefficient	Standard Error	P-value	95% Confidence Interval (CI)
PrgG	PrgG	0.00078	0.00017	**7.62e^{-6}**	(0.00044661, 0.0011266)

Figure1-3. Summary of variable PrgG in the final model

indicator shows a p-value smaller than 0.05 in the t-test. In context, holding Team, Pos, and Age constant, there is strong evidence that there is a significant relationship between PrgG and G+A-PK_90.

The 95% confidence intervals for the indicator *PrgG* is between 0.000446 and 0.001127. In context, we are 95% confident that the true effect of Progressive Passes Received on G+A-PK_90 is positive, suggesting that players who receive more progressive passes tend to have a favorable impact on non-penalty goals and assists per 90 minutes.

Figure 1-4 shows the summary of the intercept. The intercept is -0.033927, and this indicates the non-penalty goals and assists per 90 minutes for a Defensive player (DF) in team Arsenal (ARS), holding Age and PrgG constant. The value of G+A-PK_90 cannot be negative. However, the intercept is displayed as a negative value. This is not a significant issue because the minimum age of a player is 18, and as it will be explained next, the coefficient for Age is 0.00724. When the lowest age 17 is used as the Age value, the result is 0.123. Therefore, the predicted value of the model remains positive. The standard error is 0.0855906. In the t-test, the indicator shows a p-value of 0.69212. The 95% confidence intervals for the intercept is between -0.20233 and 0.134482, with the negative lower and the positive upper bounds. In context, for a Defensive player (DF) in the Arsenal (ARS) team, the non-penalty goals and assists per 90 minutes could range from a negative value to a positive value.

Variable	Indicator	Coefficient	Standard Error	P-value	95% Confidence Interval (CI)
	Intercpet	-0.0339237	0.0855906	0.69212	(-0.20232932 , 0.1344819)

Figure1-4. Summary of the intercept in the final model

Lastly, *Figure 1-5* shows the summary of the variable *Age. Age* positively influence G+A-PK_90, with a coefficient of 0.00724. The standard error is 0.00017, and the indicator shows a p-value smaller than 0.05 in the t-test. In context, holding Team, Pos, and PrgG constant, there is a strong evidence that there is a significant relationship between Age and G+A-PK_90.

Variable	Indicator	Coefficient	Standard Error	P-value	95% Confidence Interval (CI)
Age	Age	0.00724	0.00266	**0.00690**	(0.00200491, 0.0124924)

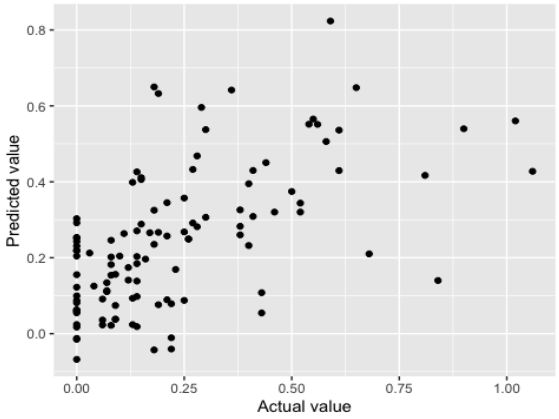
Figure1-5. Summary of variable Age in the final model

The 95% confidence intervals for the indicator *Age* is between 0.002 and 0.01249. In context, we are 95% confident that the true effect of Age on G+A-PK_90 is positive, suggesting that players who are older tend to have a favorable impact on non-penalty goals and assists per 90 minutes.

Based on the *Figure 2*, the final Model 3 performs 0.4644 on the adjusted R^2. This indicates that the 46.44% of variability in G+A-PK_90 is explained by the model based on the variables Pos, Age, PrgG, and Team. When comparing the actual G+A-PK_90 to the predicted G+A-PK_90 from the testing data, they are showing correlation of 0.588 and MSE of 0.037. The scatterplot of the actual G+A-PK_90 versus the predicted G+A-PK_90 is shown in *Figure 2*.

The usefulness of the final model can be assessed through the ANOVA table. For a multiple linear regression model to be considered useful, the p-values for all explanatory variables in the ANOVA table should be less than 0.05. As shown in *Figure A7*, all p-values are below 0.05, providing strong evidence that the model based on Pos, Team, Age, and PrgG is useful in predicting G+A-PK_90.

Figure2. Scatterplot of the actual and predicted G+A-PK_90



To assess multicollinearity in the model, a Variance Inflation Factor (VIF) test was conducted. The results of the VIF test are shown in *Figure A8*. The GVIF^(1/(2Df)) values are as follows: Pos is 1.052, PrgG is 1.203, Team is 1.0163, and Age is 1.065. Since all calculated values are less than 5, we can conclude that there is no redundancy among the four explanatory variables in predicting G+A-PK_90.

Linear regression requires six key conditions: linearity, zero-mean residuals, constant variance, normality, independence, and randomness. Among these, normality and constant variance of the model appear to be problematic. In *Figure A10*, the QQ-plot shows that the points deviate upward toward the right end rather than following the dotted line, raising questions about the condition of normality. Similarly, the Residuals vs. Fitted plot reveals that residuals on the lower-right side display a noticeable trend, suggesting an issue with constant variance. On the other hand, randomness and independence are sufficiently satisfied, as explained in the Method section. Additionally, linearity and zero-mean residuals do not appear to pose significant issues, as evidenced by *Figure A10*.

Discussion

The study began with the question: Is there a relationship between a soccer player's performance and their age? To answer this question quantitatively, it was necessary to analyze actual data to find correlation. The dataset used comprised the statistics of 456 players who played at least 90 minutes during the 2023-2024 English Premier League season. Player performance was measured using non-penalty goals and assists per 90 minutes (G+A-PK_90).

A multiple regression model was developed to predict G+A-PK_90 using 12 variables, including age and other categorical variables. After several rounds of model testing, the best model included only four variables: Age, Pos, PrgG, and Team. Based on the model's inference, it can be concluded that there is a positive relationship between a player's age and their performance.

There is a key consideration when interpreting this model. Age range should be considered. The findings indicate that a relationship exists between a player's age and their performance, with older players generally showing better performance. However, this does not imply that a 70-year-old player would outperform players in their 20s or 30s. The dataset is limited to players aged 17 to 38, and the model is built solely on this range. Therefore, making predictions or drawing conclusions for ages beyond this range would be inappropriate.

Possible confounding variables might include a player's physical attributes (height and weight) and their passing and shooting accuracy throughout the season. Since soccer is a physical sport, height and weight could be valuable additions for explaining performance more effectively. Passing accuracy and shooting accuracy, on the other hand, are indicative of a player's technical skills. It can be expected that players with higher accuracy in these areas are likely to demonstrate better performance.

Major limitation of this study is its small population size. Since the dataset includes only a single season of data from the English Premier League, it is difficult to conclude that the findings would be consistent with data from other seasons or leagues.

One more point I would like to discuss is the limitation of position. As I mentioned earlier, the reason for including defenders is that, in modern football, defenders also contribute significantly to attacks. However, there is a difference between defenders mainly contributing to assists, while midfielders and forwards are mainly contributing for both assists and goals. The limitation here is that, in creating a model to predict player performance, there is a slight unfairness in the performance predictions for all players.

The weakness of the model, as mentioned in the Results section, is that two of the six conditions—normality and constant variance—are not satisfied. This highlights the possibility that the inferences drawn from this model could be invalid. On the other hand, the strength of the model lies in its simplicity. Being able to predict a player's performance for an entire season using only four variables provides valuable information for future football-related studies.

This study utilized multiple linear regression to create a model predicting G+A-PK_90. However, more advanced algorithms are increasingly being used as regressors. As an extension of this research, it would be worthwhile to explore more complex yet powerful algorithms, such as Artificial Neural Networks or Random Forests, to develop models with better performance.

Bibliography

FIFA. (2023). *One month on: 5 billion engaged with the FIFA World Cup Qatar 2022™*. Inside FIFA. https://inside.fifa.com/tournaments/mens/worldcup/qatar2022/news/one-month-on-5-billion-engaged-with-the-fifa-world-cup-qatar-2022-tm

United Nations. (2022). *Population*. United Nations. https://www.un.org/en/global-issues/population

Allen, S. V., & Hopkins, W. G. (2015). Age of Peak Competitive Performance of Elite Athletes: A Systematic Review. Sports Medicine, 45(10), 1431–1441. https://doi.org/10.1007/s40279-015-0354-3

Sal de Rellán-Guerra, A., Rey, E., Kalén, A., & Lago-Peñas, C. (2019). Age-related physical and technical match performance changes in elite soccer players. Scandinavian Journal of Medicine & Science in Sports, 30(12), 2414–2424. https://doi.org/10.1111/sms.13463

Jamil, M., & Kerruish, S. (2020). At what age are English Premier League players at their most productive? A case study investigating the peak performance years of elite professional footballers. International Journal of Performance Analysis in Sport, 20(6), 1120–1133. https://doi.org/10.1080/24748668.2020.1833625

FBref. (n.d.). *2023-2024 Premier League Stats*. Retrieved December 6, 2024, from https://fbref.com/en/comps/9/2023-2024/stats/2023-2024-Premier-League-Stats

Figure A1 *Description of 13 variables, including both response and explanatory variable*

Variable name	Original definition	Units	Range or Levels	Rationale
G+A-PK_90	Total goals and assists minus penalty goals per 90 min. (A player who scores a goal is credited with a goal. A player who provided the final pass before the goal is credited with an assist.)	Number of G+A-PK per 90 min	0 – 1.19	Respose variable
Continent	Continent of the player is from	Continent	5 continents	Confounding variable
Pos	Position most commonly played by the player	Name of the abbreviated position	8 positions	Confounding variable
Team	The player's team	Name of team	20 teams	Confounding variable
Age	Age at season start (August 1st)	Age	17 – 38	Main explanatory variable
MP	Matches played by player or squad	Number of matches	2 – 38	Confounding variable
Starts	Game or games started by the player	Number of matches	0 – 38	Confounding variable
Min	Toal minutes played by the player	Minutes	92 – 3420	Confounding variable
CrdY	Number of yellow cards received by the player.	Number of yellow cards	0 – 13	Confounding variable
CrdR	Number of red cards received by the player.	Number of red cards	0 – 2	Confounding variable
PrgC	**Progressive Carries** Carries that move the ball towards the opponent's goal line at least 10 yards in the last six passes or any carry into the penalty area	Number of Carries	0 – 218	Confounding variable
PrgP	**Progressive Passes** Completed passes that move the ball towards the opponent's goal line at least 10 yards in the last six passes, or any completed pass into the penalty area	Number of Passes Made	0 -376	Confounding variable
PrgG	**Progressive Passes Received** Completed passes received that move the ball towars the opponent's goal line at least 10 yards in the last six passes, or any completed passes into the penalty area.	Number of Passes Received	0 - 508	Confounding variable

Figure A2 *Distribution of response variable*

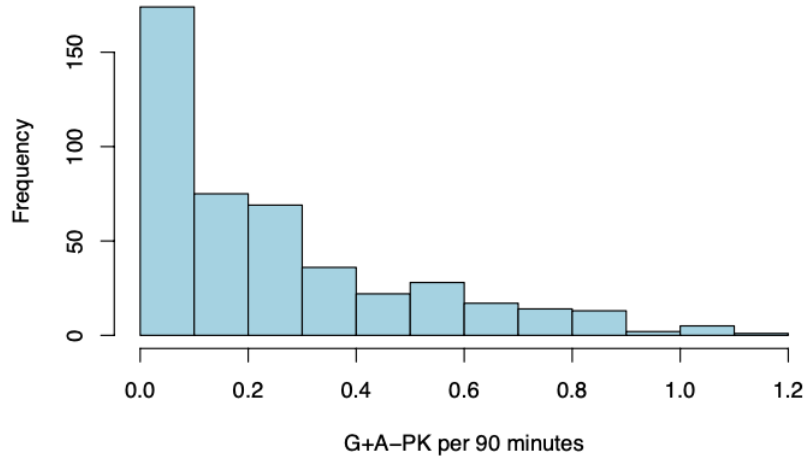


Figure A3 *Distribution of numerical variable*

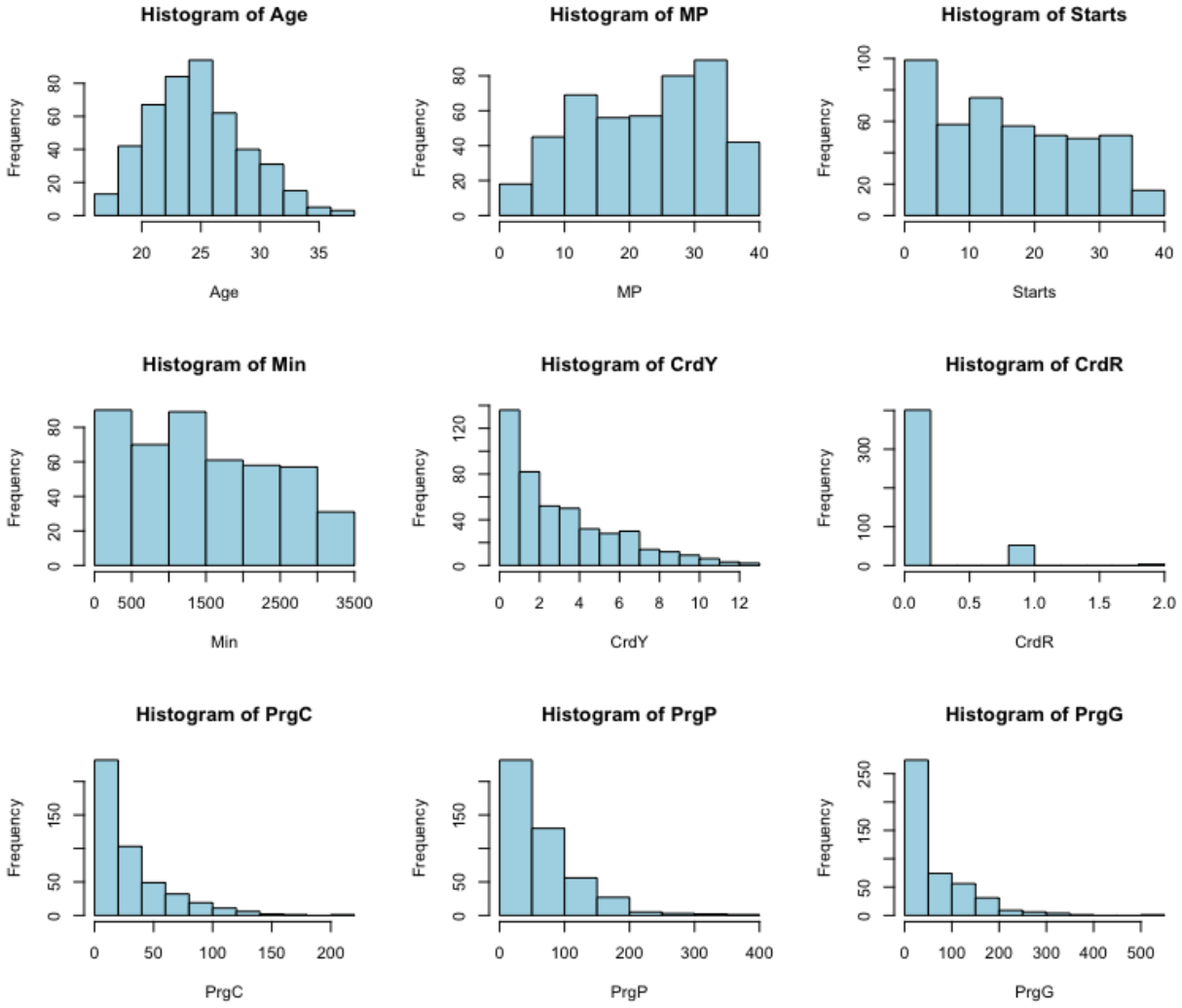


Figure A5 *Five number summaries of numerical variables*

	G+A-PK_90	Age	MP	Starts	Min	CrdY	CrdR	PrgC	PrgP	PrgG
Min.	0.000	17.00	2.0	0.00	92.0	0.000	0.000	0.00	0.00	0.00
Q1	0.050	22.00	14.0	7.00	696.2	1.000	0.000	9.00	22.00	12.75
Med.	0.180	25.00	24.0	15.00	1371.0	3.000	0.000	20.00	50.00	35.50
Mean	0.248	25.26	22.8	16.65	1491.0	3.458	0.1272	31.19	64.03	63.41
Q3	0.380	28.00	32.0	26.00	2216.0	5.000	0.000	43.25	86.00	93.50
Max.	1.190	38.00	38.0	38.00	3420.0	13.00	2.000	218.0	376.0	508.0

Figure A6 *Distribution of categorical explanatory variable*

Figure A6-1. Distribution of Continent

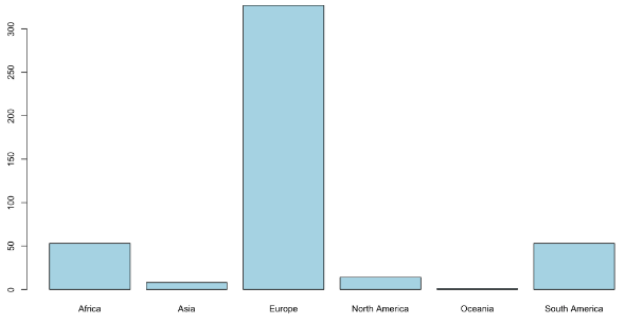


Figure A6-2. Distribution of Team

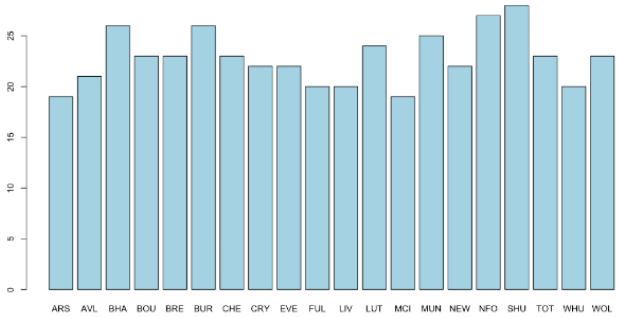


Figure A6-3. Distribution of Pos

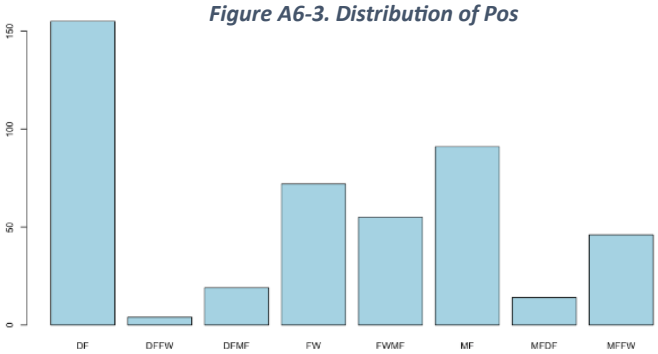


Figure A7 *ANOVA Table for the final model*

	DF	Sum Sq	Mean Sq	F-value	Pr (>F)
Pos	7	8.2455	1.17793	32.5047	**< 2.2e^{-16}**
PrgG	1	1.5346	1.53455	42.3455	**3.033e-10**
Team	19	1.6813	0.08849	2.4418	**0.0008267**
Age	1	0.2681	0.26808	7.3976	**0.0068958**
Residuals	313	11.3427	0.03624		

Figure A8 *VIF Test on the final model*

	GVIF	DF	GVIF^(1/(2*Df))
Pos	2.056010	7	1.052832
PrgG	1.447981	1	1.203321
Team	1.849619	19	1.016315
Age	1.134143	1	1.064962

Figure A10 *Residual plot of the final model*

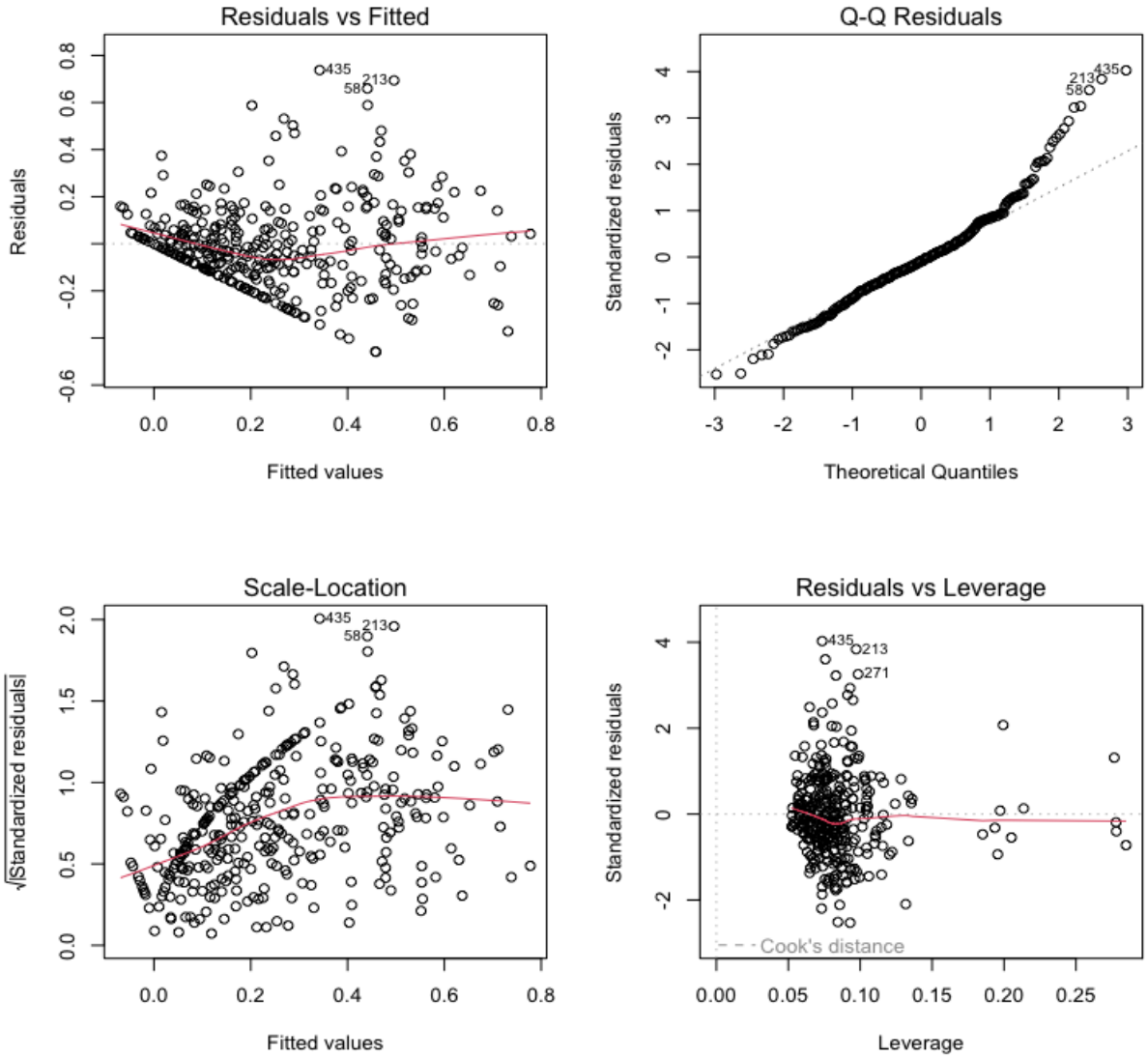


Figure A11. Comparison between the models based on the four criterions

Model	Y ~ X	Adjusted R^2	Correlation (Predicted and Actual)	Mean Squared Error (MSE)	Number of parameters
Model1	G.A.PK_90 ~ Continent+ Team + Pos + Age + MP + Starts + Min + CrdY + CrdR + PrgC + PrgP + PrgG	0.4583	0.5830567	0.03834	42
Model2	sqrt(G.A.PK_90) ~ Continent+ Team + Pos + Age + MP + Starts + Min + CrdY + CrdR + PrgC + PrgP + PrgG	0.4429	0.5572097	0.045291	42
Model3	G.A.PK_90 ~ Team + Pos + Age + PrgG	0.4644	0.5883499	0.037793	30
Model4	sqrt(G.A.PK_90) ~ Team + Pos + Age + PrgG	0.4409	0.5761768	0.045254	30
Model5	G.A.PK_90 ~ Pos + PrgG + Team + Age:Continent + Age^2 + Pos:Min	0.4618	0.5872166	0.0384418	44
Model6	sqrt(G.A.PK_90) ~ Pos + PrgG + Team + Age^2+ Pos:PrgC	0.4474	0.5901457	0.03977249	38
Model7	sqrt(G.A.PK_90) ~ Pos + PrgG + Team + $PrgG^2$ + Age + $PrgC^2$ + MP^2 + Pos:Team + Pos:PrgG	0.5221	0.4670345	0.07603796	123