The Relative Age Effect in Competitive Football Leagues

A Statistical Analysis Investigating Trends and Relationships.

June 15th, 2023

Kabir Guron

Table of Contents

1.0 Introduction

The relative age effect (RAE) is a phenomenon where athletes—especially in their youth—appear to perform better than if they were born earlier in the year. Their success is cumulative through the **Matthew Effect**. The Matthew Effect is the notion that an initial subtle advantage or disadvantage can accumulate to become significant long term. It is proposed that youth athletes born earlier in the year have an advantage over other players in their respective cohort. A ten year old born in January is about 10% older than a ten year old born in December. Since most people experience a growth spurt between ages 8 and 13 due to puberty, an age advantage gives a noticeable athletic advantage. Consequently, it is possible to notice slight patterns where athletes are distributed earlier in the year, as study confirms the presence of RAE in Quebec's minor hockey leagues. The disparity is also pronounced in higher athletic competitions as seen in Figure 1 in the UEFA tournament, the most prestigious men's football tournament.

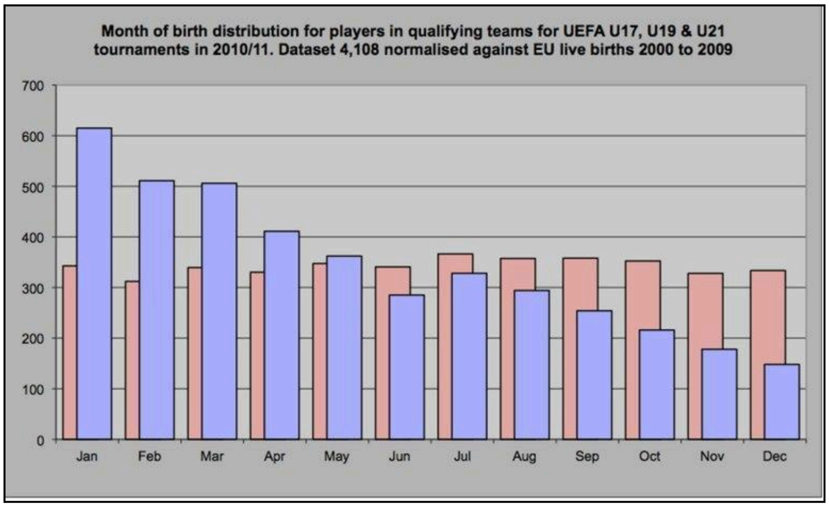

Figure 1. Distribution of birth months of the general population (red) against distribution of birth months for players qualified for UEFA tournaments (blue).

The objective of this report is to determine if the RAE is present in the English Premier League, mutually considered the most competitive football league in the world in 2023. It can be achieved through assessing player performances with respect to their birthdate.

1.1 Hypothesis

It is hypothesized that a player's performance is negatively correlated with the number of days that player was born after January 1st. It is also expected that birthdays are distributed towards the beginning months.

1.2 Methodology

Depending on the player's position, there should be an appropriate statistic that captures their performance. For example, defenders should not be expected to score as many goals as forwards as their roles are different. Generally, defenders perform tackles to regain possession of the ball. Table 1 shows the statistics collected by each player in the Premier League. It is recognized that no single statistic can truly summarize a player's performance, but the indicators chosen are representative of their role. Goalkeeper statistics were not included due to the complexity of their role; it would be too inaccurate to rank goalkeepers based on a single stat. For instance, excellent defenders can make a goalkeeper's clean sheets (number of matches without conceding a goal) appear inflated.

Table 1. Statistics gathered for Premier League players by position. Roles are stated for context.

	Goalkeeper	**Defender**	**Midfielder**	**Forward**
Expected Role	Only one on a team. Responsible for guarding the net. Allowed to contact the ball directly with their hands.	Responsible for staying-back during and providing resistance to the opposing team from scoring.	Partly responsible for defence and offence. Flexible role but primarily associated with helping forwards score goals.	Primarily responsible for scoring goals. A secondary role could be providing assists (a pass made to a player that scores shortly later).
Performance Indicator	N/A	**Tackle percentage:** ratio of successful tackles versus attempted tackles. A tackle is a player trying to take the ball away with physical contact.	**Big chances created per appearance:** for every appearance they make, the number of significant goal-scoring opportunities they have provided. For example, a pass that is exceptional and easily scorable for another player.	**Goals per appearance:** the number of goals scored on average per appearance.

Through a Python web scraping program, players were gathered individually and compiled on a spreadsheet automatically on the official Premier League database. A player was only added if (1) their position was a forward, midfielder, or defender, (2) they had at least one appearance to confirm activity, and (3) the program could successfully collect all data necessary. If the program had difficulty accessing an element, it skipped the iteration and moved onto the next player. These errors were random and unpredictable, so they do not affect the general trend of the sample taken. Additional data for the birthdates of players that played in the World Cup were collected as a control variable.

2.1 Single Variable Frequency Results

Gathering from a sample of 680 players that played in the World Cup, there is a visible preference for players born near the beginning of the year (Figure 2). The first three months have the highest frequency of players while the remaining months appear evenly distributed. This suggests that the relative age effect at a competitive international level still exists.

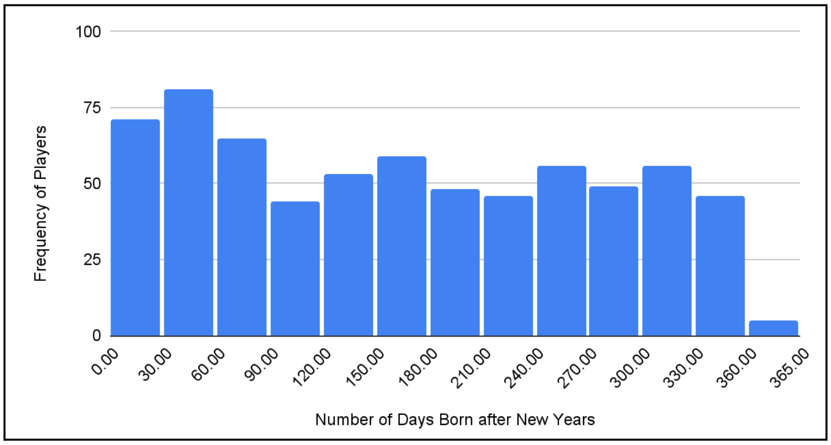

Figure 2. Histogram distributing number of days born after New Years for players in the 2022 World Cup.

Premier League players do not have as strong of a bias towards the beginning months as seen in Figure 3. The months with the greatest frequency are September, February, and January in that order.

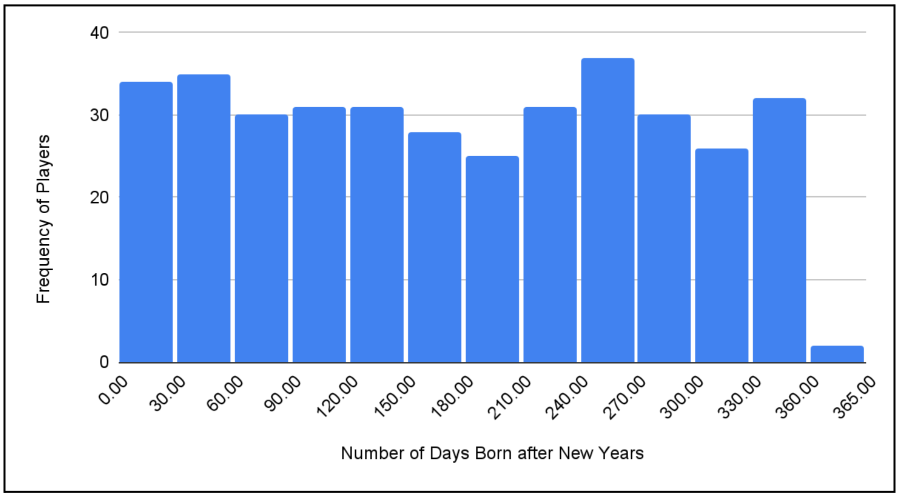

Figure 3. Histogram distributing number of days born after New Years for players in the Premier League (2023).

2.2 Single Variable Probability Distribution Analysis

Minor data analysis can be performed to ensure that the data collected is fairly representative. Checking for a normal distribution resemblance can bring assurance that enough data was collected. The analysis will gather the statistics for defender tackle percentages as a single variable analysis. Observing Figure 4, it is clear that the distribution resembles a bell curve due to the peak appearing centralized on the mean along with its symmetrical nature.

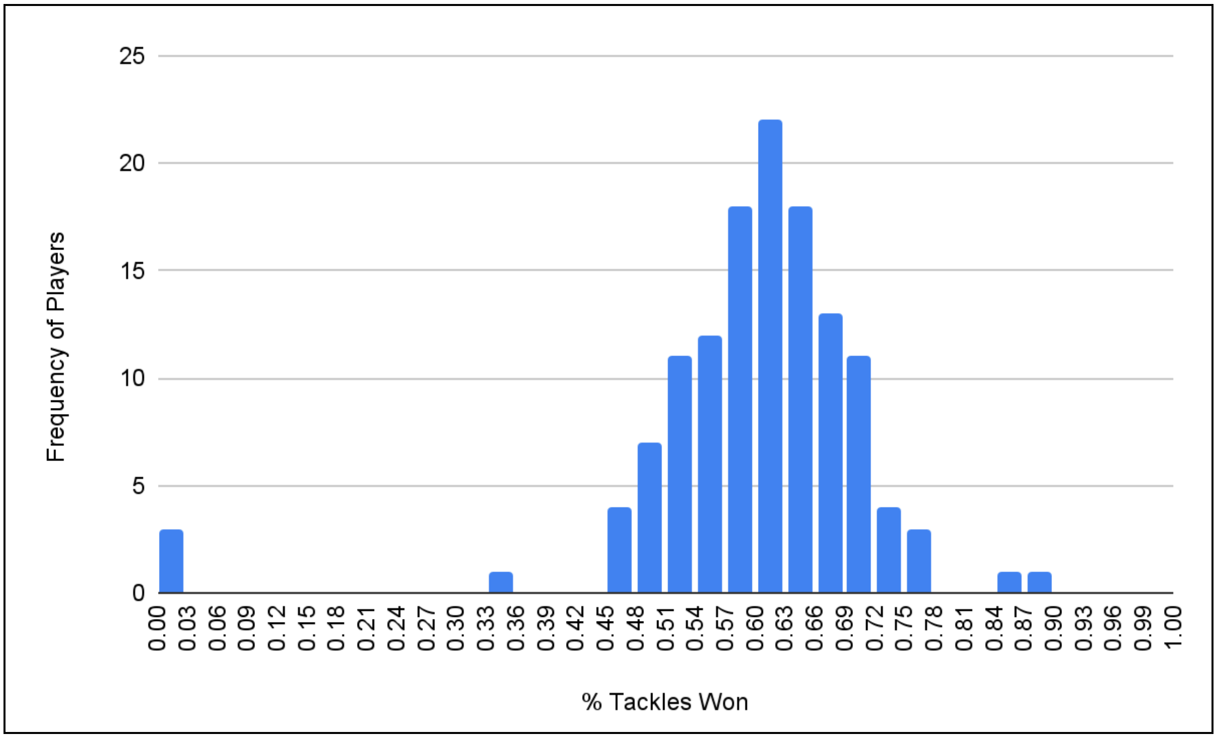

Figure 4. Histogram visualizing the distribution of tackles won.

Using a 90% confidence will indicate where most players should lie between in terms of percent tackles. A 90% confidence interval is appropriate since it considers the majority of players while excluding clear outliers seen. A higher confidence level will increase the interval unnecessarily. Table 2 undergoes the calculations necessary for this interval.

Table 2. Mean, standard deviation, and percent error calculations.

Statistic	**Value**	**Importance**
Mean	0.60	The average tackle percentage was 60%.
Standard deviation	0.13	The average player is 13% away from the mean.
Critical value	1.645	Z-Score for 90% confidence interval.
Number of Players (n)	130	Sample size of defenders.
Margin of error	0.0182	9/10 footballers have a tackle percentage 60% ±1.8%.

Therefore, every nine out of ten footballers have a tackle percentage from 58.2% to 61.8%. Additionally, it is possible to find the probability of finding a player between a given interval. Let the interval of interest be from 65% to 75%. If a player is in this interval, they lie above the mean, their performance can be considered slightly above average.

$$Z_{65\%} = \frac{x - \mu}{\sigma} = \frac{0.65 - 0.60}{0.13} = 0.38$$

$$Z_{75\%} = \frac{x - \mu}{\sigma} = \frac{0.75 - 0.60}{0.13} = 1.15$$

Using a Z-score chart,

$$P(65 < x < 75) = 0.8749 - 0.6480 = 0.2269 \approx 0.23 \qquad (1)$$

Therefore, there is a 23% that a given player is above average. It is a useful metric to gauge defender performance within the two variable analysis.

2.3 Two-Variable Statistical Analysis

To determine whether a correlation exists between a player's relative age and performance, a scatter plot can be created to observe any visible trends. Scatter plots can superimpose data for two variables which allows researchers to quickly identify general correlations, outliers, and regions of concentrated data. For linear correlations, a line of best fit can be created for better visualization.

The dependent variable will be different for each position. Although all players have appearances, some players are unable to score goals, create big chances, or perform tackles. It is either a consequence of a player’s performance or limited time on a field. Regardless, those data points should be included since players that perform better should often receive more playing time. Figures 5, 6, and 7 visualize player performance and birthdate by their respective positions.

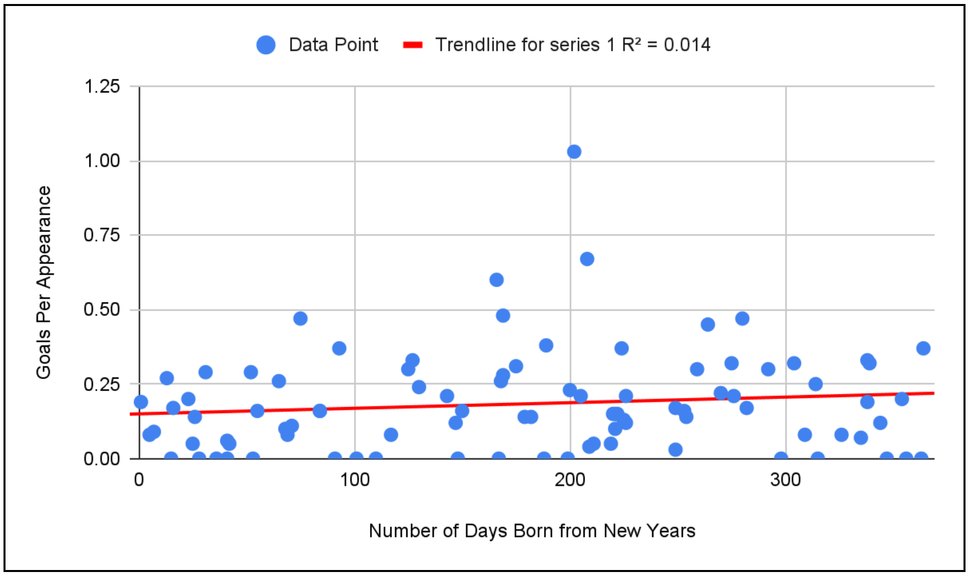

Figure 5. The relationship between a forward's number of goals per appearance and the number of days born after New Years. Correlation coefficient is +0.113.

As observed in Figure 5, Erling Haaland is a significant exception to the data (202 days, 1.03 goals per appearance). He achieved the highest record for most goals in the Premier League this season (36 goals). Harry Kane (208 days, 0.67 goals per appearance) is another talent in the Premier League. Note that both players are European and born in the month of July.

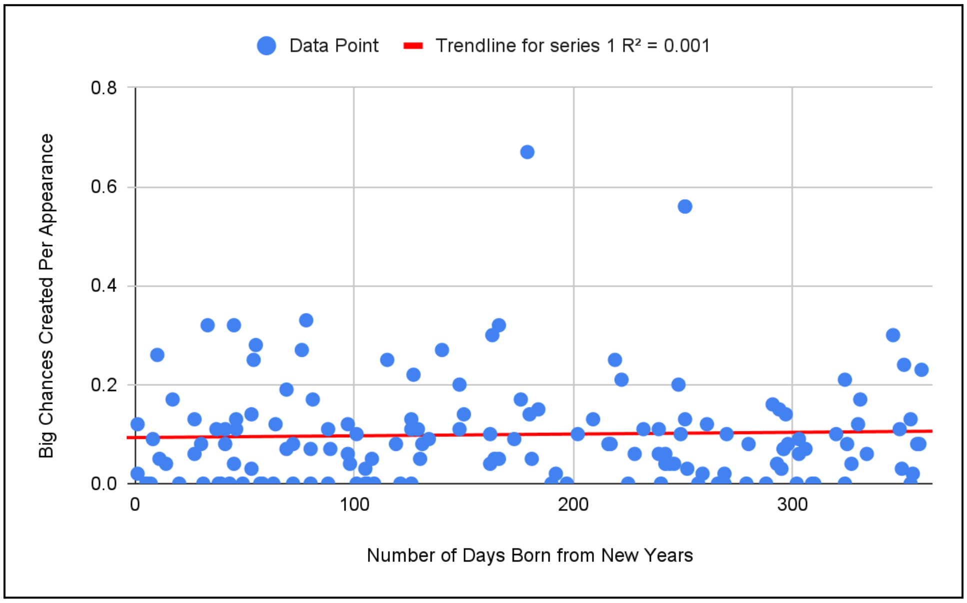

Figure 6. The relationship between a midfielder's number of big chances created per appearance and the number of days born after New Years. The correlation coefficient is +0.035.

In Figure 6, the three midfield outliers with impressive averages are players Kevin De Bruyne (179 days, 0.67 chances per appearance), Lewis Hall (251 days, 0.56 chances per appearance) and Bruno Fernandes

(251 days, 0.56 chances per appearance). De Bruyne is currently regarded as the best midfielder in the world. Bruno Fernandes is a solid Portuguese player with a fair reputation. Lewis Hall is lesser known with only 9 appearances this season, so there are insufficient appearances to consider him a top-class player. Note that De Bruyne is European and born in late June and Bruno Fernandes was born in September.

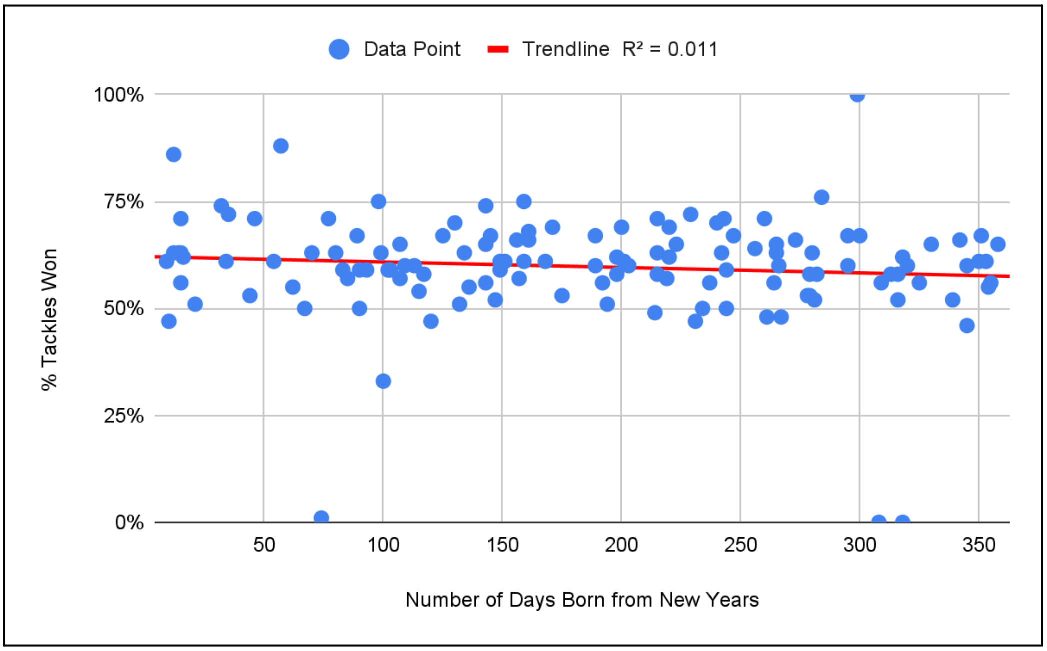

Figure 7. The relationship between a defender's tackle percentage and the number of days born after New Years. The correlation coefficient is –0.103.

High performing defenders (classified by having a tackle percentage between 65% and 75%) are evenly distributed across all months as seen in Figure 7. As seen with a 100% tackle percentage, Odeluga Offiah is considered to be an outlier. It is found that he only had two appearances, so his statistics are likely inflated. He would require more game appearances to be considered as well performing. After investigating, the outliers found in defender performance are not considered top-performing athletes. For example, Juan Larios has a slide tackle percentage of 86% but a market value of approximately £3.7 million. For comparison, the average Premier League player is worth £20 million. This suggests that the defender performance indicator does not best represent defender performance.

Since $R^2 < 0.1$ for all scatter plots, there is a weak correlation across all data. The independent variable (relative age) fails to explain variance in the dependent variable (performance). No consistent relationship appears to exist between any of the independent and dependent variables. There is no general correlation between player performance and data of birth.

3.1 Reverse Relative Age Effect

The British school system works differently than Canada's with respect to enrollment age. In a Canadian school, students in a given school grade will all be born in the same year. However, British schools accept students that will turn the same age in a school year. For example, if a child turns 4 between the dates September 1st 2022 to August 31st 2023, they will enroll in school full-time in September 2023. As a result, students born in September will be more biologically developed than students in July with respect to a cohort of students attending the same grade level.

Once accounting for the British school system, it explains why the month of September (approximately 240 days to 270 days) has the greatest frequency of European players (Figure 3). Instead of January, September students are the most physically developed in their cohort. Therefore, the RAE likely occurs starting September instead of January in Europe. It is still important to consider that not all the players in the Premier League were born in Europe, so their school system may differ across regions.

The **reverse relative age effect** (also known as the underdog effect) is a phenomenon where individuals who were at a developmental disadvantage at a young age due to the RAE are among the most talented individuals in the world. It suggests in order for an individual to stand-out despite their growth disadvantage, their raw performance must compensate. Eventually, these compensated individuals grow to become indifferent from their cohort of players physically, putting them at an advantage indirectly. The players Erling Haaland, Harry Kane, and Kevin De Bruyne are all European top performing athletes born near the end of their school year. In the data set, these players are significant outliers that lie several standard deviations above the mean. Table 3 shows the calculated Z-scores for each athlete.

Table 3. The Z-score for statistical outliers in player performance.

Player	Month and Day Born	Position	Performance Indicator	Mean	Player Data	Z-Score
Erling Haaland	July 21st	Forward	Goals per Appearance	0.18	1.03	5.00
Harry Kane	July 28th	Forward	Goals per Appearance	0.18	0.67	2.88
Kevin De Bruyne	June 28th	Midfielder	Big Chances per Appearance	0.10	0.67	5.18

Since the major top-performing athletes in the Premier league are born in the months prior to September, the reverse RAE is noticeable. Consequently, this supports the notion of the regular RAE too.

3.2 Significance

The main implication of the RAE is that players born in certain months aiming to pursue professional football are set at a disadvantage inherently. Club managers are more likely to make misinformed decisions regarding player selection since relatively younger players will appear less fit compared to their cohort of peers. Namely, Harry Kane at eight years old was kicked out of Arsenal's youth academy for this reason. In a statement from Liam Brady about Harry Kane, the Arsenal Academy director at the time, he admitted "He [Harry Kane] was a bit chubby, he wasn't very athletic but we made a mistake." His decision was likely skewed due to the RAE. Future coaches and managers could lose-out on potential superstars by making discriminatory comparisons. A possible solution to the RAE is delaying player selection since it allows players to reach their growth spurt and physical potential. If waiting is not feasible, grouping like players based on their weight and height could be more effective and representative. Overall, European coaches and managers should be wary of the RAE and aim to minimize its effects.

4.0 Conclusion

The analysis refutes the hypothesis but supports the general reasoning for the relative age effect. The single variable analysis suggested that the relative age effect is prevalent at an international level (World Cup) and is likely present in the Premier League. In the Premier League, there is a higher frequency of players born in September, where players are most physically developed within their respective grade level. The two variable analysis shows that there is no general correlation between player performance and the month that they were born. Outlier players that performed well above average support the existence of the reverse relative age effect too. The finding suggests that relative age discrimination may still exist in the Premier League. For this reason, managers and coaches should be cautious when selecting players at a young age.

Appendix

The following link is a spreadsheet that contains all the raw data collected.

https://docs.google.com/spreadsheets/d/1rpKWIOoCvNCyP1RgzsiAd3v4addoHBNmkMG2RtEsISc/edit?usp=sharing

Citations

Football transfers EN. (n.d.). *Juan Larios Lopez Transfer News, history, market value (xTV) & Career stats*. Transfer News, History, Market Value (xTV) & Career Stats. https://www.footballtransfers.com/en/players/juan-larios-lopez

Lemoyne, J., Huard Pelletier, V., Trudeau, F., & Grondin, S. (2021, January 15). *Relative age effect in Canadian hockey: Prevalence, perceived competence and performance*. Frontiers. https://www.frontiersin.org/articles/10.3389/fspor.2021.622590/full

Matthew effect. Matthew Effect - an overview | ScienceDirect Topics. (n.d.). https://www.sciencedirect.com/topics/psychology/matthew-effect#:~:text=The%20Matthew%20Effect%20refers%20to%20a%20pattern%20in%20which%20those,between%20the%20advantaged%20and%20disadvantaged.

Published by Statista Research Department, & 8, D. (2022, December 8). *European football player market value by league 2021*. Statista. https://www.statista.com/statistics/722417/football-player-market-value-europe/#:~:text=The%20Premier%20League%2C%20often%20touted,just%20over%2020%20million%20euros.

Timwig. (2021, February 22). *Why athletes' birthdays affect who goes pro - and who becomes a Star*. FiveThirtyEight. https://fivethirtyeight.com/features/why-athletes-birthdays-affect-who-goes-pro-and-who-becomes-a-star/

Valente, A. (2018, February 13). *Harry Kane was released by arsenal because he was "a bit chubby", says Liam Brady*. Sky Sports. https://www.skysports.com/football/news/11095/11248766/harry-kane-was-released-by-arsenal-because-he-was-a-bit-chubby-says-liam-brady