STATISTICAL ANALYSIS OF FIFA 2017 PLAYERS

STAT 250 – APPLIED STATISTICS

TERM PROJECT REPORT

DEPARTMENT OF STATISTICS OF

MIDDLE EAST TECHNICAL UNIVERSITY

Sefa Bulut - 2561116

Kemal Doğu Oktay - 2502250

Furkan Kazancıoğlu - 2502185

Rubar Akyıldız – 2561033

JUNE, 2024

Introduction

The purpose of this report is to analyze the data of players in the FIFA 2017 game. We do this analysis so that players who play the game can make better strategies. These players can be competitive players who just want to improve themselves, or they can be individuals who accept this game as a football simulation and try innovative tactics to apply it in real-life scenarios. We use exploratory data analysis, hypothesis testing, regression analysis, and ANOVA methods for analysis.

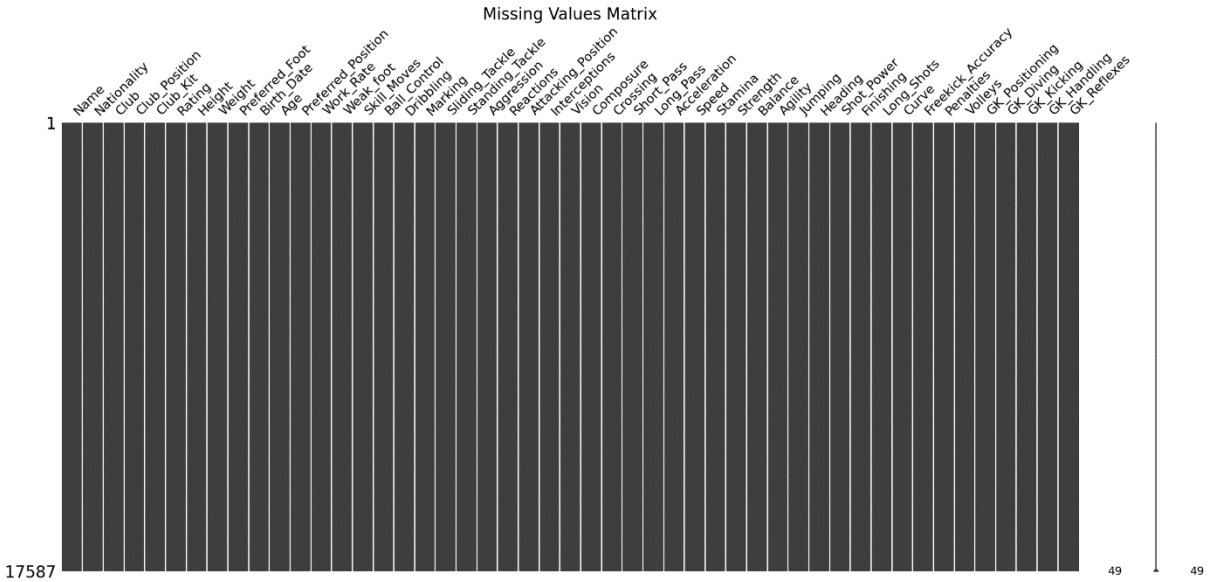
Using FIFA Game Data as Real-Life Data for Our Project

FIFA's data on players is valuable due to its correlation with real-life performance. Collecting this data involves professional teams rating players based on real-life information like goals, scoring opportunities, and free kick conversion rates. The importance of this data lies in its correlation with real-life performance, as we lack access to professional teams or a 30-year dataset. So we will assume that the scores out of 100 provided by these teams reflect real-life performance, and our research is built upon this assumption. The reliability of the datasets of FIFA games is also confirmed by Ronan Murphy with his article named "FIFA player ratings explained: How are the card number & stats decided?".

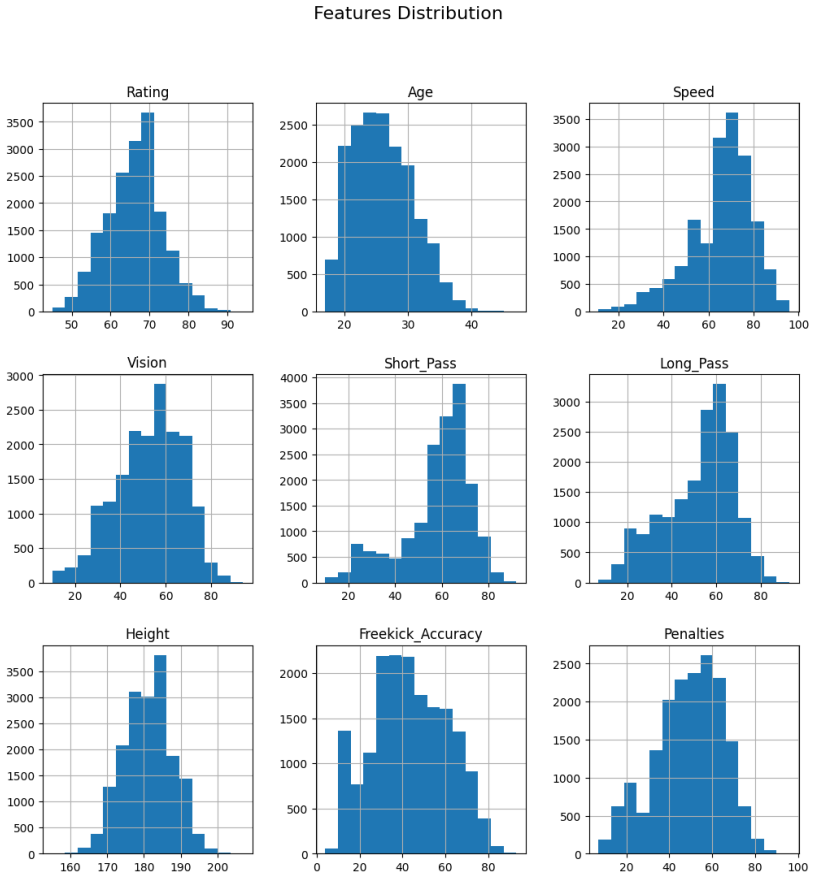
Data Source and Preparation

The raw dataset for this project is sourced from a github user named "dcupl" about FIFA 2017, including attributes such as player ratings, physical statistics, and skill levels as ordinal data. Data cleaning involves handling missing values, outliers.

Our slightly changed and cleaned FIFA 2017 dataset gives us information 17,587 football players, spanning 49 diverse features. It encompasses almost all of the notable players highlighting key metrics such as rating, age, nationality, and playing positions. Additionally, our dataset contains plenty of information for evaluating player abilities and comparing performance on multiple dimensions, because it includes so many valuable columns like ball control, dribbling, and several physical vc. qualities. All those data gives as a large space to uncover intricate patterns and relationships among abilities of the players, positions, and overall game vision, making it key to improving football simulation experiences and strategic gameplay analysis.



The case of the white areas among the blacks means if there was a null value, that would be it. As can be seen in the graph, we have no null values because our dataset has been thoroughly cleaned.



Since we have more than 17 thousand rows in our dataset, we performed our tests by approximating the normal distribution.As can be seen in the graph, our variables are approximately normally distributed. We choose those variables because we mostly used these 9 variables in our tests.

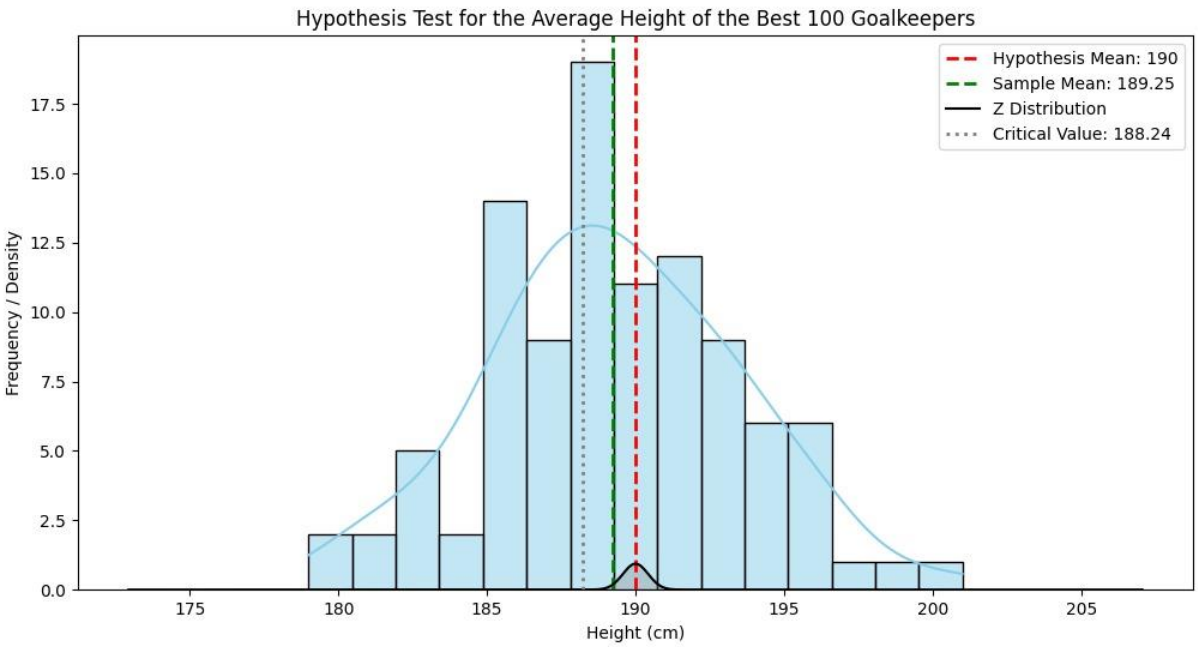
Analysis

1 - INFERENCES ABOUT MEAN (One sample hypothesis testing)

Research Question: Are the height of the sample of best 100 goalkeepers is greater than 190 cm?

For this test, we took sample of 100 highest overall rated goalkeepers. We formulated the null hypothesis (H0) that the average height is not greater than 190 cm against the alternative hypothesis (H1) that the average height is greater than 190 cm.

Given a significance level of 0.05, the **Z-value of -1.759** falls outside the rejection region, also the **p-value of 0.0784** is greater than the alpha level. Thus, we fail to reject the null hypothesis. Which means the analysis does not provide sufficient evidence to conclude that the average height of the top 100 goalkeepers is greater than 190 cm. Hence, we are free to state that it is not crucial to be a long person that is longer than 190 cm based on this sample



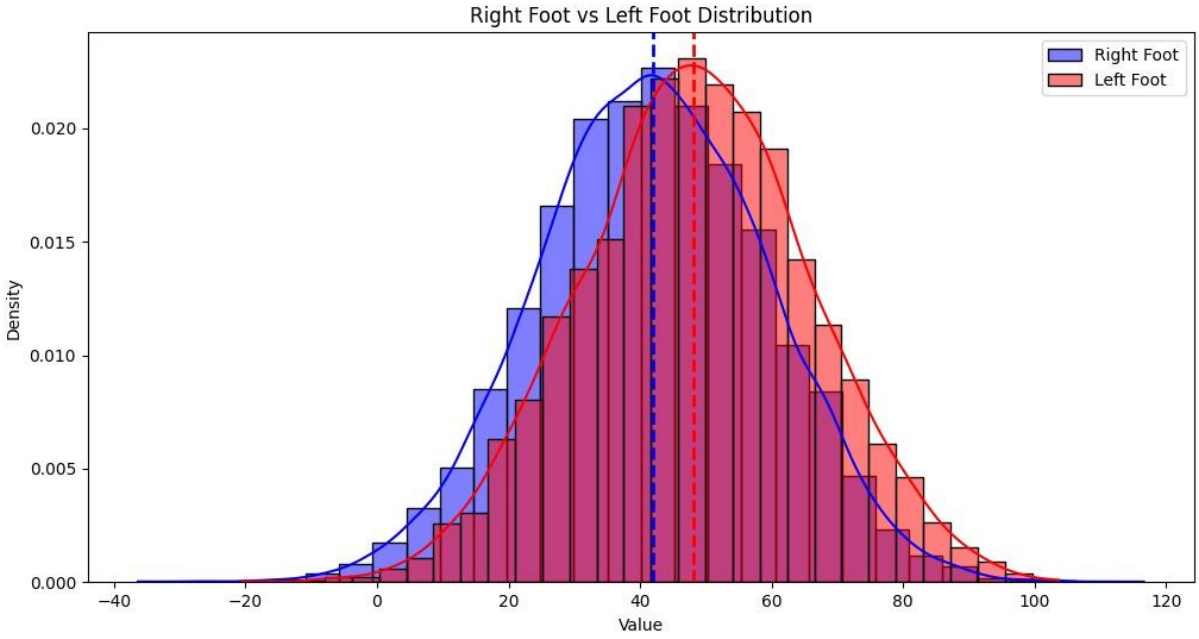
z-Statistic: -1.76
-1.7585395856749069

The results of a hypothesis test are displayed together with the height distribution of our sample goalkeepers in a histogram. The sample mean height of approximately 188 cm is shown by the green dashed line, while the red dashed line reflects the estimated mean height of 190 cm. For reference, we also draw a curve which represents the Z distribution curve and the critical value at 187 cm. Although heights appear to cluster around 190 cm, statistical analysis produced insufficient evidence to support the claim that the average height is higher. This result of the test implies that, in contrast to popular belief, the top goalkeepers in the game do not have average heights much higher than 190 cm, highlighting the significance of other characteristics in their play.

2 - COMPARISONS OF MEANS (Two-sample hypothesis testing)

Research Question: Is there a difference in free-kick abilities between right-footed and left-footed players?

The null hypothesis (H0) suggested equal means, while the alternative hypothesis (H1) suggested differences. With a **p-value of roughly 0.00** and a **Z-statistic of -20.16**, the results showed a significant difference between the two groups. Even though, it is a common belief that left-footed players are better at free-kick, we was curious it is really true.



z-value: -20.16
p-value: 2.12e-90
2.1208753384176544e-90

The free-kick skills of right- and left-footed players were compared using a histogram, which is colored blue and red, respectively. The results, as we propose, demonstrated a remarkable difference among the two groups, with left-footed players having better

free-kick performance. Given that the statistical study demonstrated a distinct distinction between the two distributions, this conclusion may have implications for strategic choices made during gameplay or while making strategies when selecting players to take free kicks.

3 - INFERENCES ABOUT PROPORTIONS (One sample hypothesis testing)

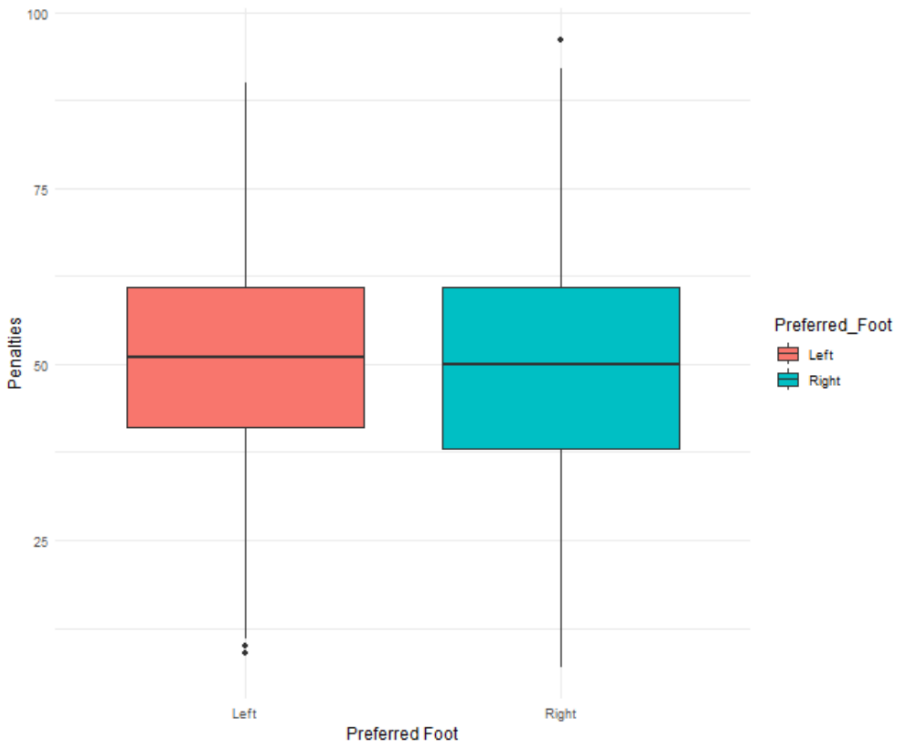
Research Question: Are at least 20% of Italian footballers centre-backs?

We memorise that they call Italians "defenders". Therefore, we conducted a test that the null hypothesis (H0) that 20% of Italian football players are center-backs (stoppers). The proportion being more than 20% was the alternative hypothesis (H1). We are unable to reject the null hypothesis because the test results reveal a **Z-statistic of -1.745** and a **p-value of 0.959**, neither of which are significantly greater than 0.05. Proving that the analysis's assertion that more than 20% of Italian football players are center backs cannot be proven by our information, or it really is not more than a myth.

4 - COMPARISONS OF PROPORTIONS (Two-sample hypothesis testing)

Research Question: Is there a difference in penalty abilities between left and right- footed players?

In this experiment, we looked at whether left- and right-footed players' ability to take penalties differs significantly from free kicks. However, we're going to prove it differently this time. We took a sample of players with penalty scores greater than 70 in the testing of penalty-taking capability. The alternative hypothesis (H1) indicated a difference ($p1-p2\neq0$), while the null hypothesis (H0) claimed there is no difference in the proportions of successful penalty-takers between the two groups ($p1-p2=0$). The analysis generated a **p-value of 0.828** and a **Z-statistic of -0.217**. We are unable to reject the null hypothesis because the p-value is significantly higher than the usual significance level of 0.05. This suggests that, different from free-kicks left and right footed players do not differ statistically significantly in their ability to take penalties. After that, we decided to build a plot to see if it really true that players do not differ.



As the test yields, both groups show similar distributions with overlapping ranges. So, it is a myth that a player's left- or right-footedness affects their ability to accept penalties, regardless of the preferences or common beliefs regarding gameplay. Unlike the free-kicks test, it is free to state that other factors should be considered when choosing players for penalty situations.

5 - MULTIPLE LINEAR REGRESSION

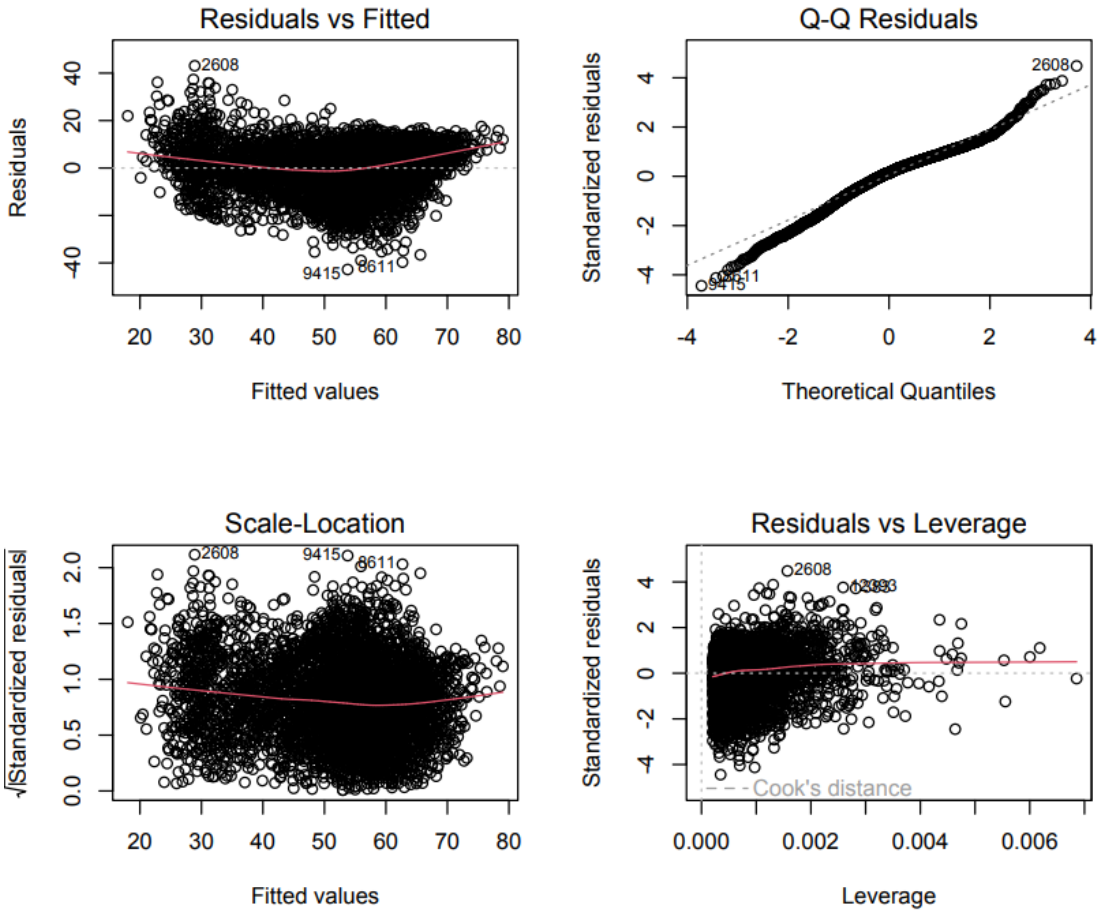
Research Question: How does players' ages, long pass abilities, short pass abilities game vision of the players?

We made a multiple linear regression model that predicts a player's game vision with relevant columns, and the fitted line is represented by the equation:

Game Vision = 5.721 + 0.257 * Age + 0.523 * Long Pass + 0.187 * Short Pass

Based on these coefficients, the game vision score rises by 0.257 units for every year of age, meaning that older players, or we can call veterans have slightly more accurate game vision. The greatest effect is caused by long pass ability, which raises game vision by 0.523 units for every unit increase, while short pass ability only raises game vision by 0.187 units. In short, the model reveals that all age and passing skills, especially long passing, significantly enhance a player's game vision.

Assumptions About Regression:



- Normality of Residuals:

The Shapiro-Wilk normality test shows a W value of 0.983 with a **p-value < 2.2e-16**, indicating that the residuals are normally distributed.

- Independence of Errors:

The Durbin-Watson test returns a **p-value of 0.262**, suggesting that the errors in the model are independent since p-value is greater than the 0.05.

- Homoscedasticity:

The residuals vs. fitted values plot reveals a horizontal spread, while the scale-location plot confirms that the residuals are uniformly distributed across the fitted values.

- Linearity:

Since there is no pattern in the residuals' spread, the residuals vs. fitted plot implies a linear relationship between predictors and the outcome variable.

- Influential Points:

The residuals vs. leverage plot identifies some influential points that may affect the regression model.

- Multicollinearity:

VIF:	Crossing	Short_Pass	Long_Pass
	3.125615	6.834926	5.604339

Variance inflation factors (VIFs) for the predictors are below 10, indicating no multicollinearity issues in the model.

6 - ONE-WAY AND MULTIPLE COMPARISONS

<u>Research Question</u>: Is there a difference between the average overall ratings of football players of different nationalities?

Our null hypothesis (H0) is all the differences in means are equal to zero, while the alternative hypothesis (H1) is at least one of them is different.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Nationality	159	152209	957.3	22.86	<2e-16 ***
Residuals	17427	729894	41.9		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The ANOVA results show an **<u>F-statistic of 22.86</u>** with a **<u>p-value less than 2.2e-16</u>**, indicating significant differences between the groups. Since our p value is less than the alpha level (0.05), there is enough evidence to reject the null hypothesis (H0).

Comparisons;

$Nationality

	diff
Albania-Afghanistan	1.136486e+01
Algeria-Afghanistan	1.712000e+01
Angola-Afghanistan	1.550000e+01
Antigua & Barbuda-Afghanistan	8.000000e+00
Argentina-Afghanistan	1.303236e+01
Armenia-Afghanistan	1.600000e+01
Aruba-Afghanistan	1.450000e+01
Australia-Afghanistan	7.935897e+00
Austria-Afghanistan	9.984962e+00
Azerbaijan-Afghanistan	1.162500e+01

However, since there are so many nationalities in the dataset, it is impractical to display all comparisons. Thus, we made the decision to include a table that briefly illustrates the mean rating differences amongst a few chosen nationalities. For example, the significant 11.4-point difference between the players from Albania and Afghanistan indicates that Albania is producing better players in terms of education and training than Afghanistan. Due to the large number of comparisons, we provide only a sample of 10 from the results, highlighting that the analysis reveals meaningful rating differences among various nationalities.

Conclusion

In conclusion, our FIFA 2017 dataset analysis reveals significant differences in player attributes and performance, with left-footed players showing superior free-kick skills and a strong correlation between passing skills and game vision. Despite stereotypes, only a small proportion of Italian players are center-backs, challenging traditional views. Age and passing abilities significantly impact game vision, with older players with better skills exhibiting enhanced game awareness. The absence of significant differences in penalty-taking abilities between left-footed and right-footed players suggests both can be equally effective in penalty scenarios. These findings can aid gamers and strategists in making informed decisions and exploring innovative tactics.

References

EA Sports FIFA 2017 Raw Data: https://github.com/dcupl-demos/fifaplayers/tree/main/dcupl/models

Our Cleaned EA Sports FIFA 2017 Data: https://github.com/sefabltt/STAT-250-TERM-PROJECT/blob/main/fifaplayers_cleaned_data.csv

FIFA player ratings explanation: https://www.goal.com/en/news/fifa-player-ratings-explained-how-are-the-card-number--stats-decided/1hszd2fgr7wgf1n2b2yjdpgynu