Hyperiondev

TASK

Exploratory Data Analysis on the FIFA-18 Data Set

Visit our website

Introduction

Data is growing very fast in today's world. It is not easy to process the data manually. Analysis and visualization of data programs allow for reaching even deeper understanding. The programming language Python, with its English commands and easy-to-follow syntax, offers an amazingly powerful open-source alternative to traditional techniques and applications [1].

Exploratory data analysis is an approach to see what the data can communicate to us away from the formal modeling or hypothesis testing task. EDA helps to analyze the data sets to summarize their statistical characteristics focusing on four key aspects, like, measures of central tendency (comprising of the mean, the mode and the median), measures of spread (comprising of standard deviation and variance), the shape of the distribution and the existence of outliers[1]. Exploratory Data Analysis (EDA) is an approach to summarize the data by taking their main characteristics and visualize it with proper representations [2]. Exploratory data analysis is performed on the FIFA-18 dataset.

This paper gives more insight to factors that influence player performance, Wage, age etc, and what is their relationship if it does exist. The fifa-18 dataset was analysed to give insight into football teams performance mostly based on Nationality. The analysis is helpful to football scouts, coaches, managers and clubs as it sheds light on countries to get best players from.

DATA CLEANING

The FFIA-18 dataset contains 17981 rows and 75 columns. The first step was to drop all the columns that were not needed. The dataset was reduced to 18 columns or attributes. The remaining attributes were chosen based on what scouts would generally look for in a player not necessarily based on players playing position.

The analysis assists in narrowing down countries where one is likely to get good players irrespective of specific playing position. The isnull() method and heatmap were used to find columns with missing values. The attribute

Club had 248 missing values and the records of the attribute were dropped. Figure 1 shows the heatmap with missing values before they are dropped.

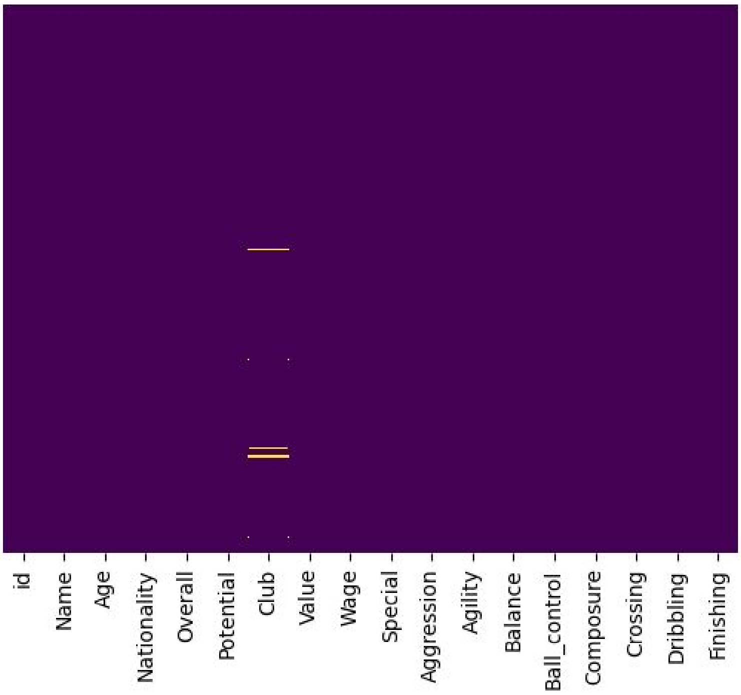

Figure 1. The heatmap of missing values in the dataset.

Figure 2 below shows the attributes of the dataset after the dropping missing values in the dataset. The following step was to convert the Wage and Value attributes format to a one that could be read by pandas. Attribute data types were checked to see if they are of the correct data type.

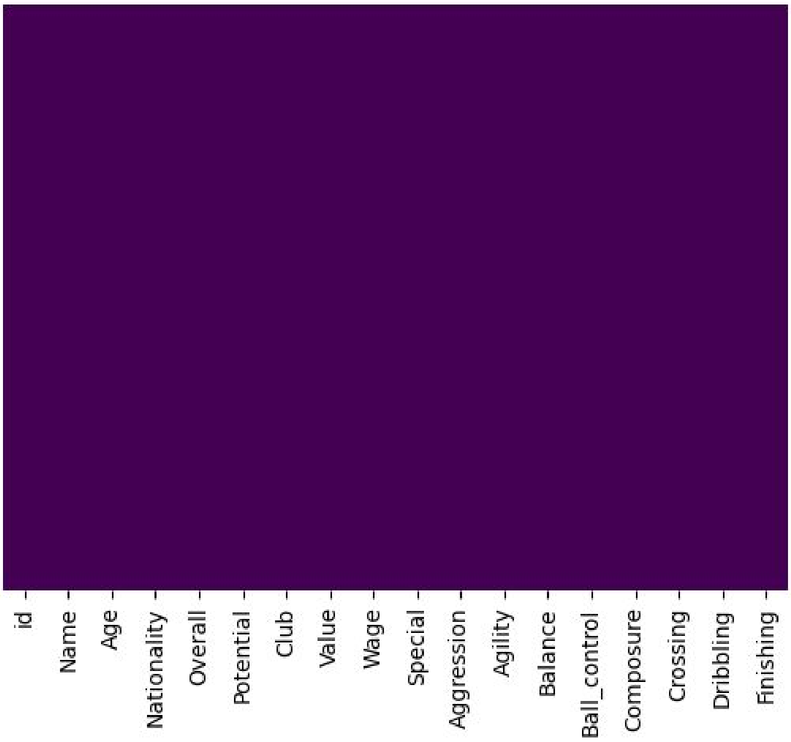

Figure 2. The heatmap of clean dataset before data type analysis.

Nine columns needed their data types corrected and resulted in the data set having missing values in those columns, as can be seen in figure 3.

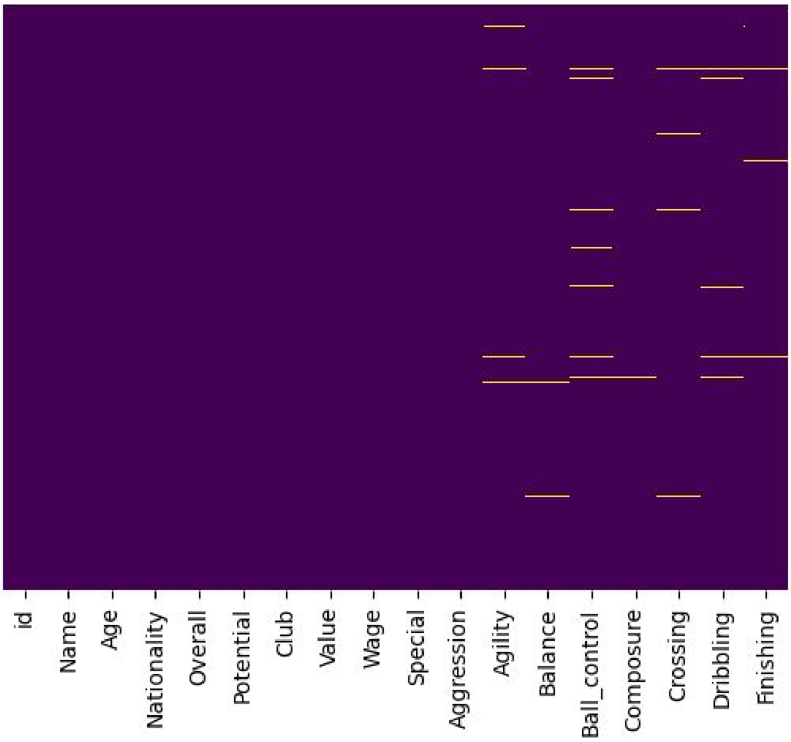

Figure 3. The heatmap of clean dataset after data type analysis.

The missing values were replaced with the mean of each column. Figure 4 shows the cleaned dataset heatmap with no missing values and correct attribute data types.

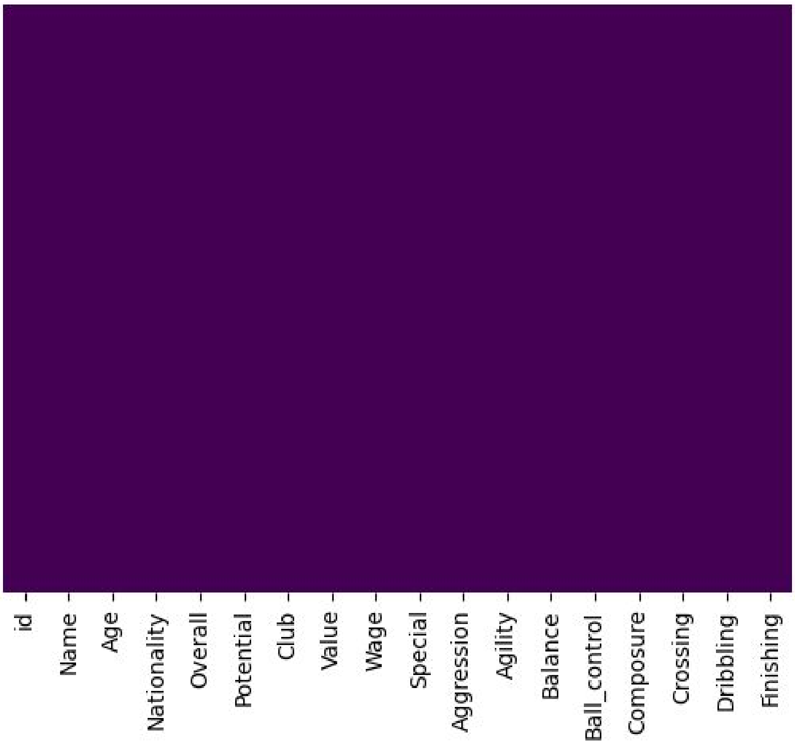

Figure 4. The heatmap of clean dataset with correct data types for attributes.

MISSING DATA

The 248 missing data values in the Club column had their records dropped while the missing data that resulted from changing data types was replaced by the average of each column.

DATA STORIES AND VISUALIZATIONS

The exploratory data analysis of the fifa-18 dataset, was to visualize the number of players playing in each country and list the top 30 countries with the most number of players per country. Figure 5 shows the top 30 countries with the most number of players. This shows that a large number of players play in England, Germany and Spain. This makes it due to these countries' large leagues.

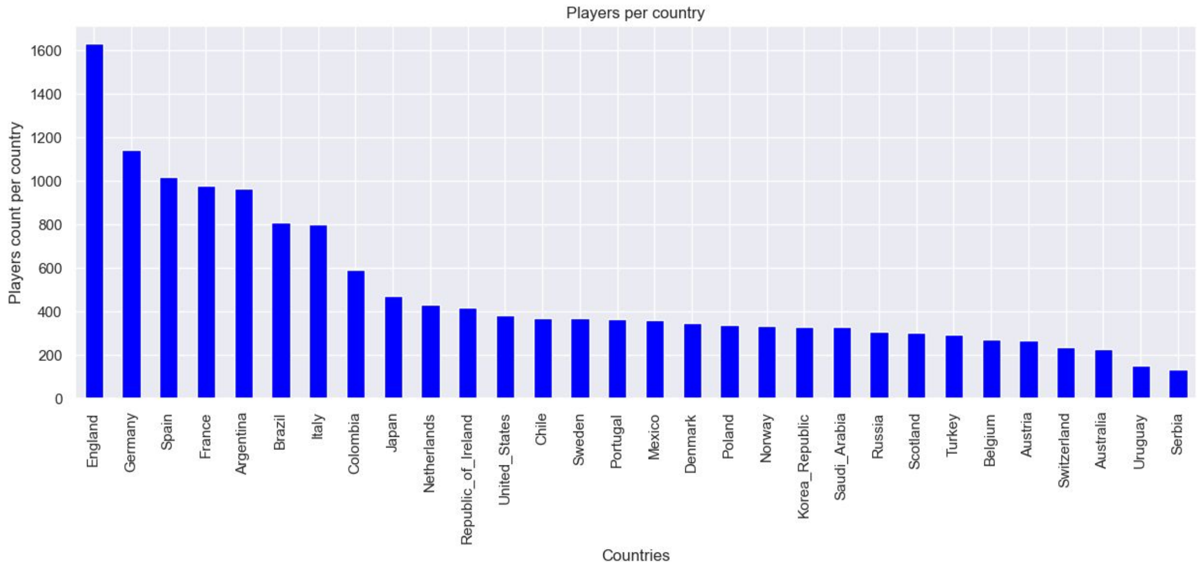

Figure 5. Top 30 countries with most players.

Figure 6, shows 30 countries with the lowest number of players per country. This information gives information about the country's league size and football popularity.

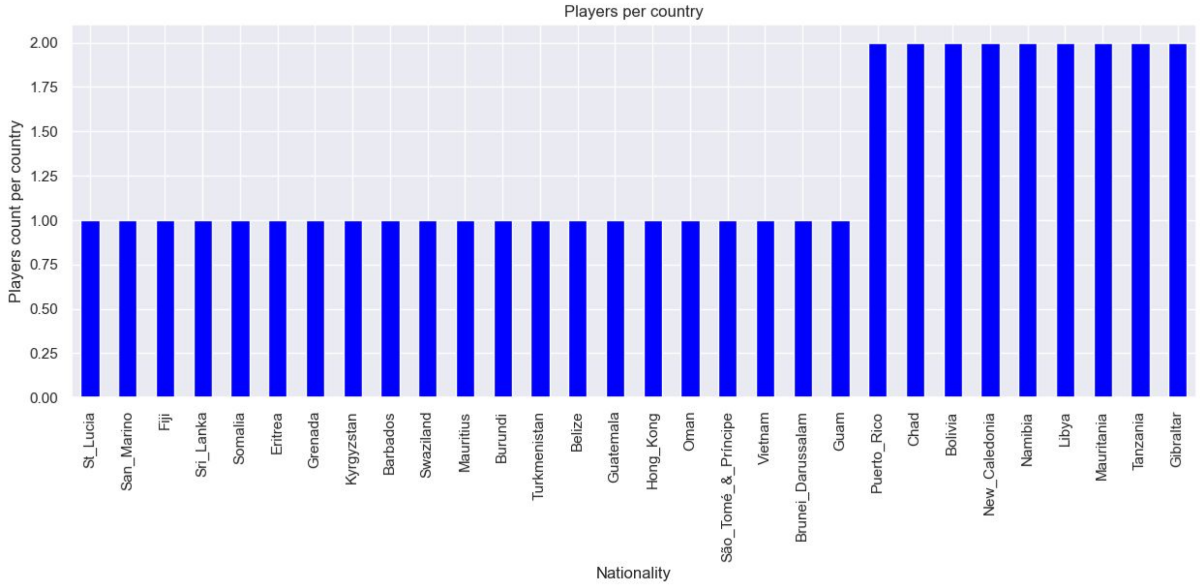

Figure 6. Top 30 countries with least amount of players.

Figure 7 and figure 8 give the top 50 clubs with the highest number of players and bottom 50 clubs with lowest number of players respectively.

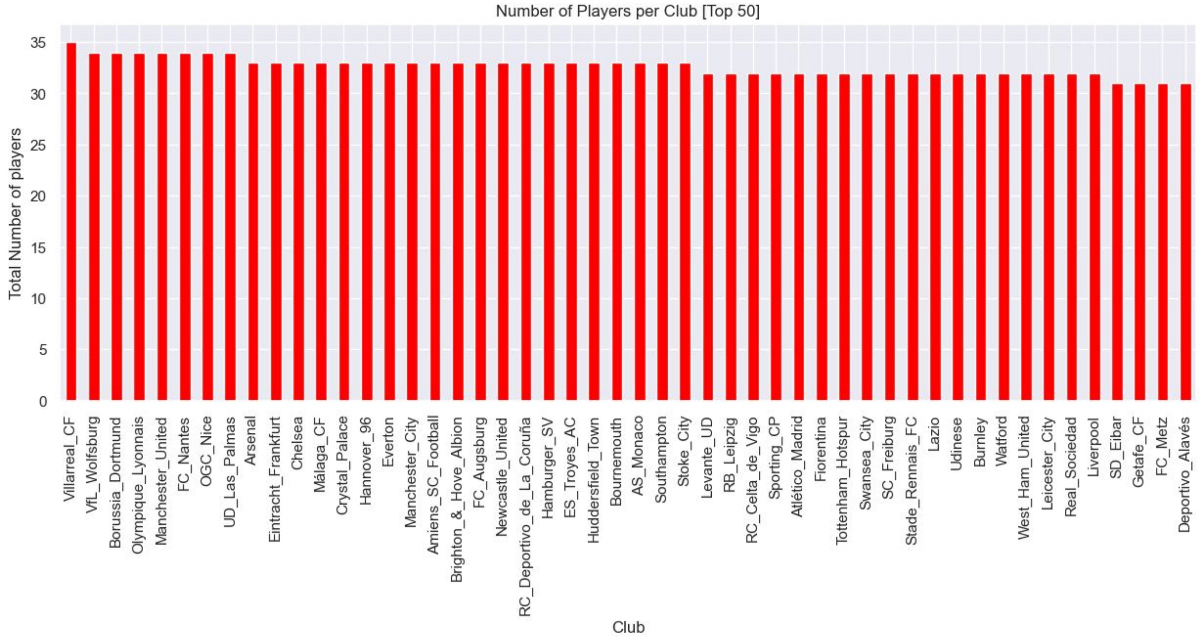

Figure 7. Top 50 clubs with the most number of players.

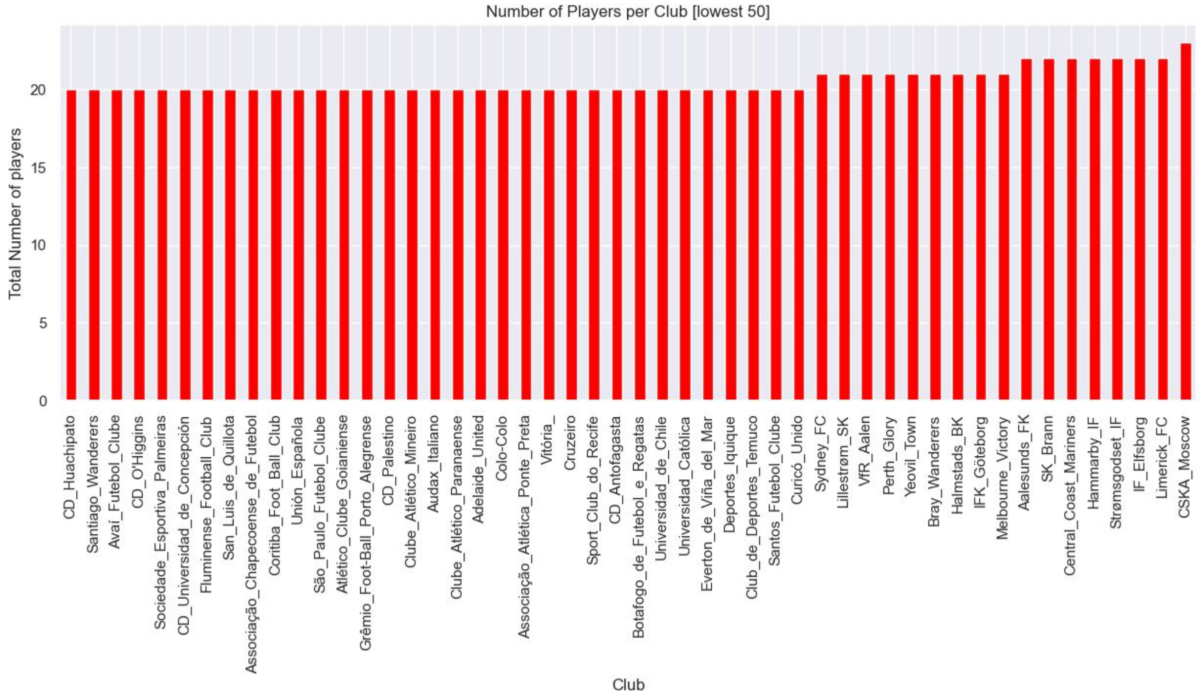

Figure 8. Bottom 50 clubs with the least number of players.

The FIFA-18 dataset has 647 different or unique clubs. Figure 9 shows Age distribution of the player dataset while figure 10 show Overall player distribution.

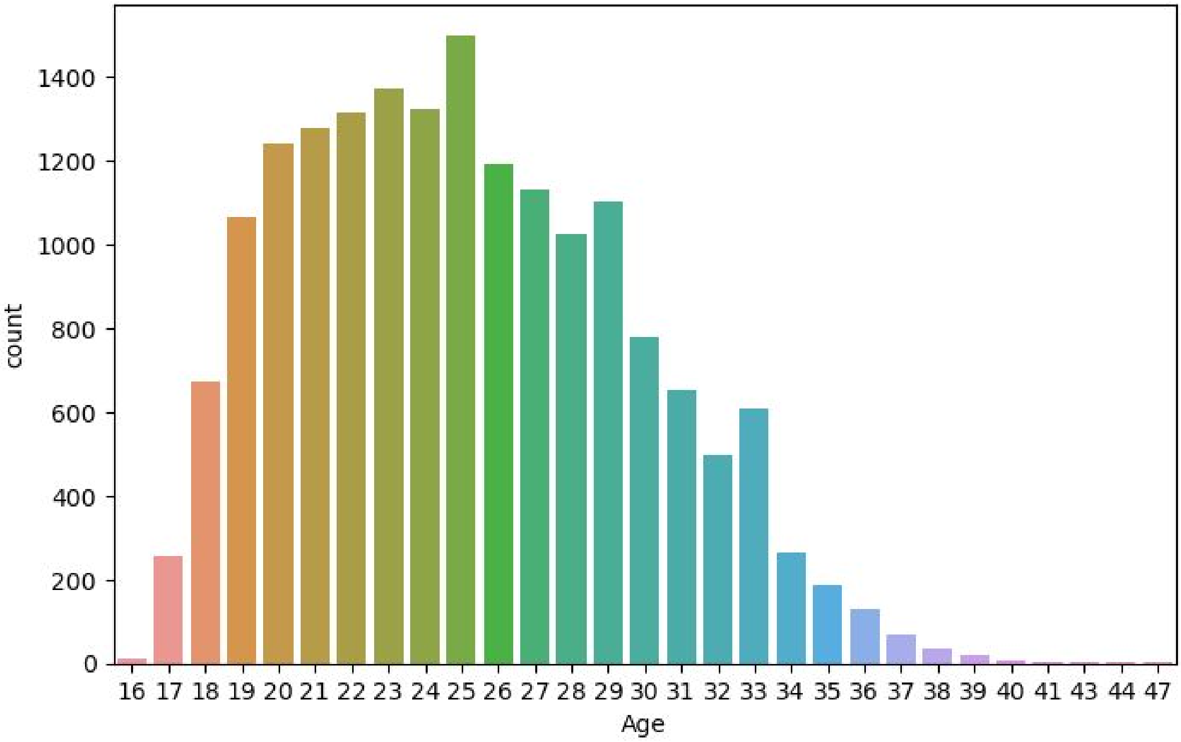

Figure 9. Age distribution of FIFA-18 dataset.

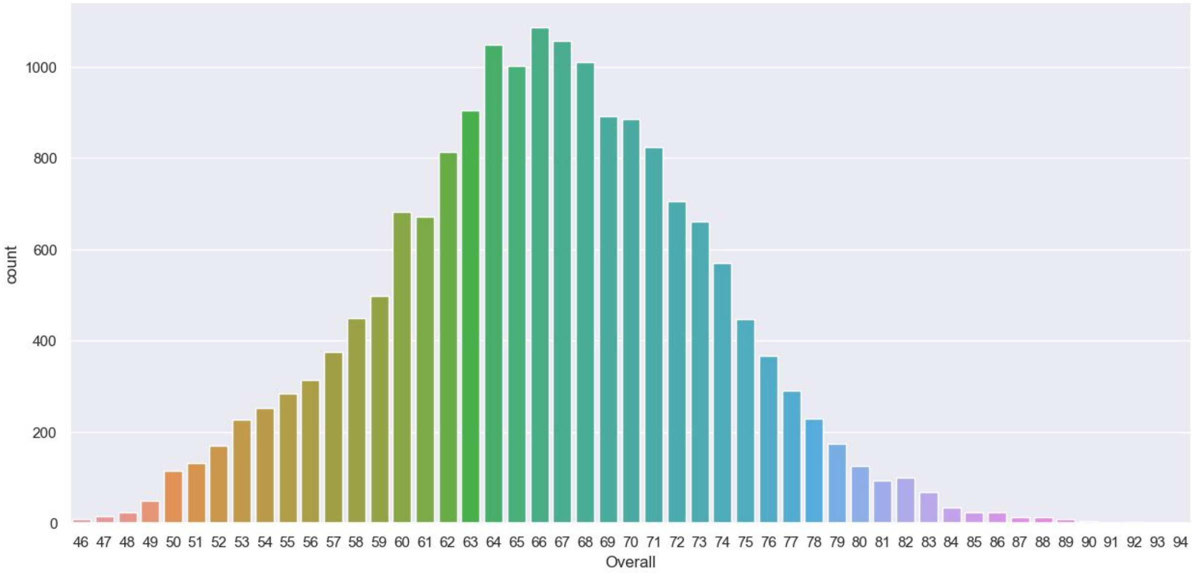

Figure 10. Overall distribution of FIFA-18 dataset.

Figure 11 shows the probability density function of the age attribute in the dataset, while figure 12 shows the wage distribution plot. The dataset has 647 different football clubs out of the 164 countries. This would mean on average each country in the dataset has at least four clubs.

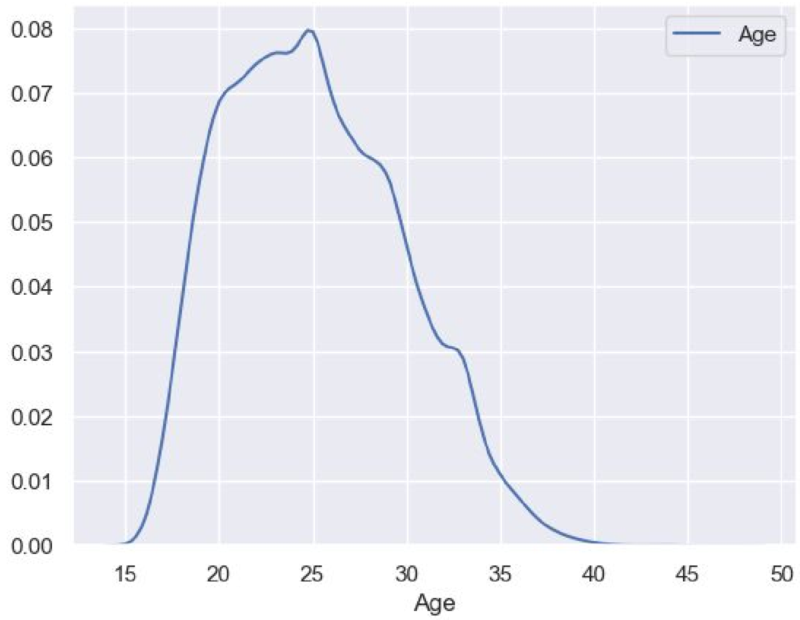

Figure 11. PDF diagram of the age attribute.

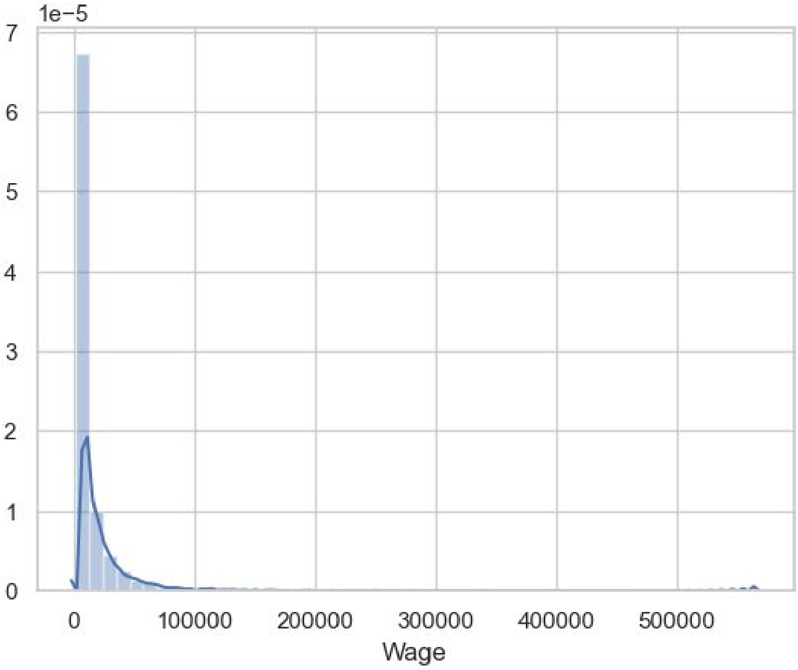

Figure 12. PDF diagram of the wage attribute.

The group-by method was implemented on the Name, Age, Nationality and Club attributes. Figure 13 displays 50 rows of countries' age distribution using a bar plot.

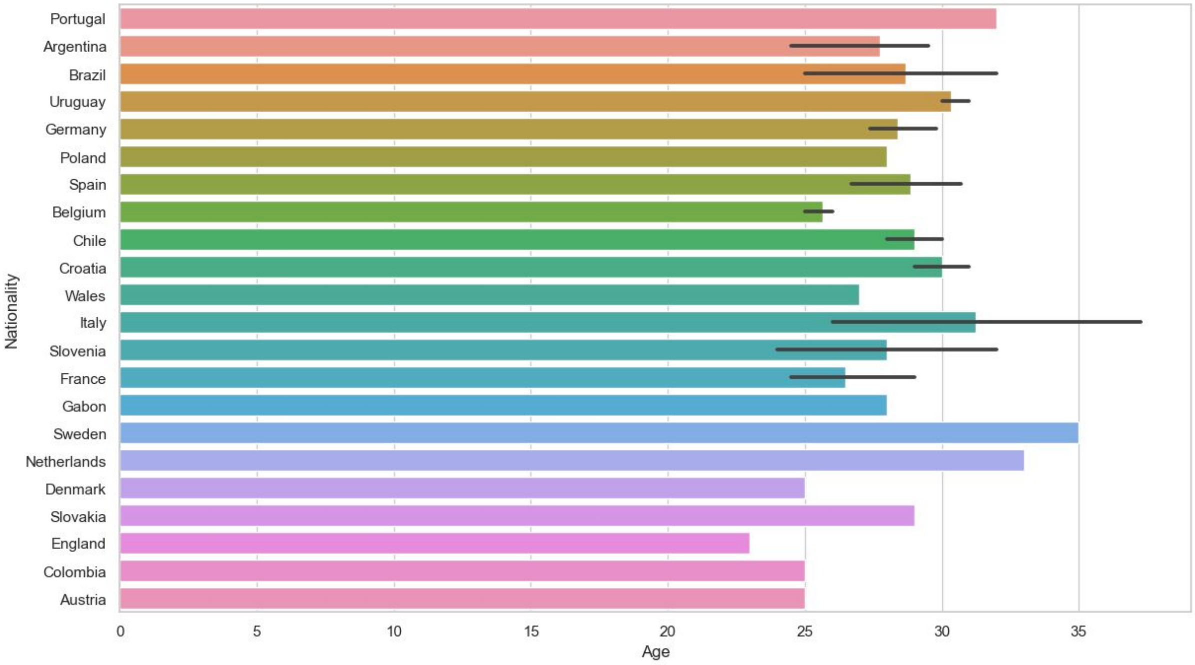

Figure 13. Countries age distribution based on 50 records.

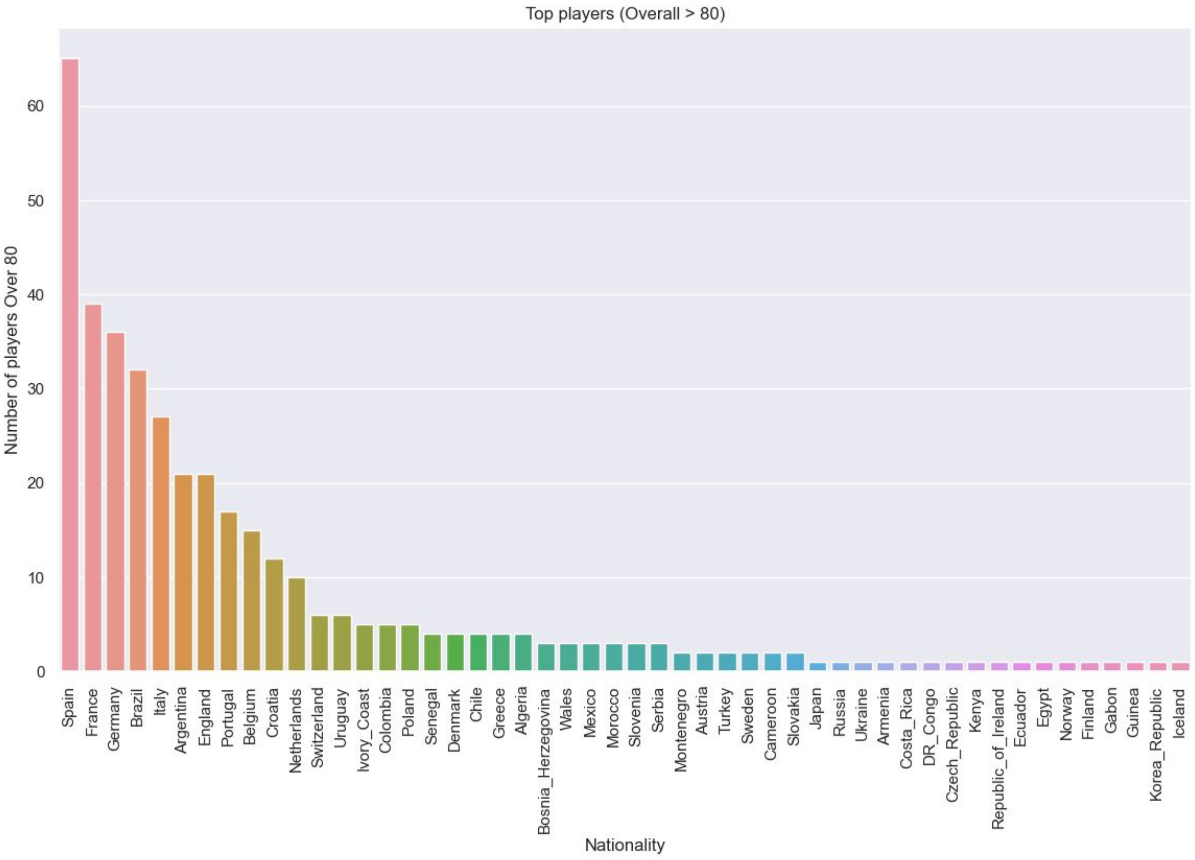

Figure 14. Players overall greater than 80 per country.

Figure 14 above shows the number of players per country in the dataset whose overall is greater than 80.

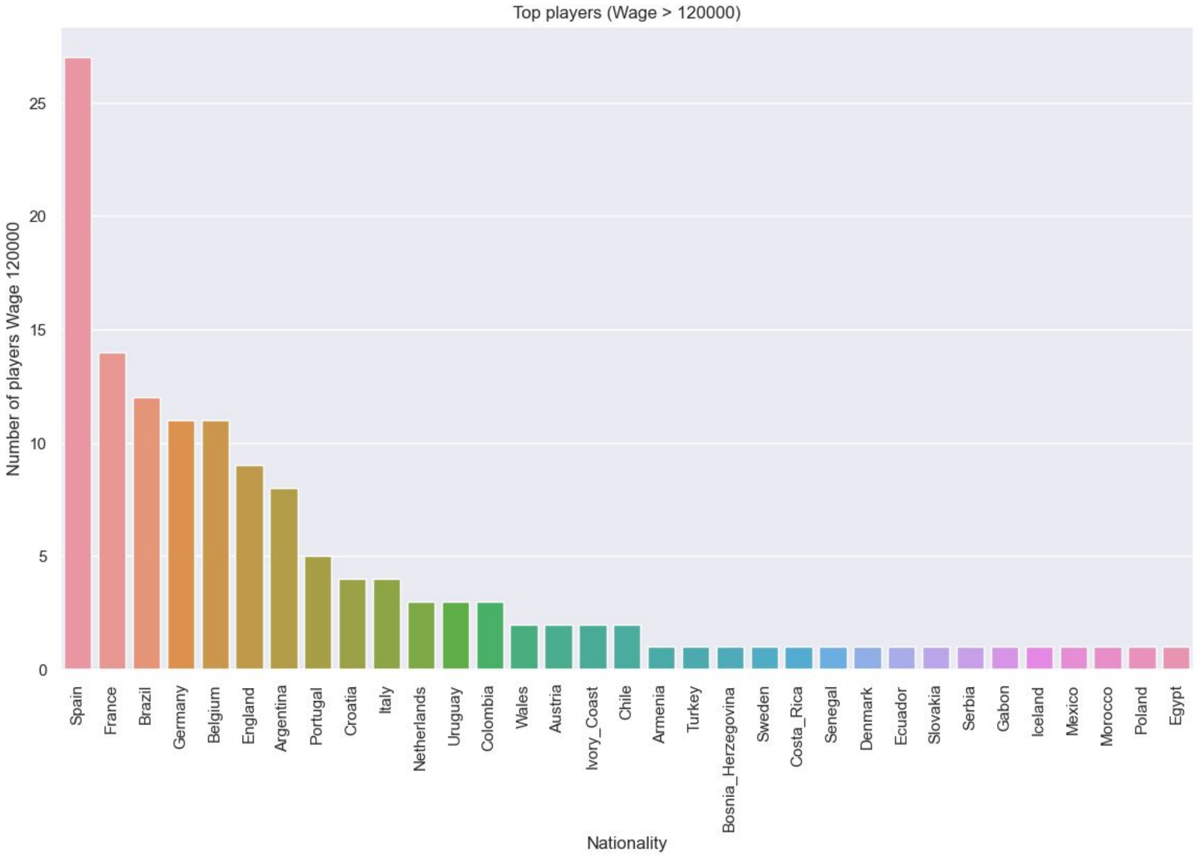

Figure 15. Players wage greater than 12 0000 per country.

Figure 15 shows countries which have the most paid players. Span, France, Brazil, Germany and England have the highest players with wages greater than 120 000. Figure 16 and figure 17 are the boxplot for 6 countries with players that have the best Overall ratings and Potential respectively.

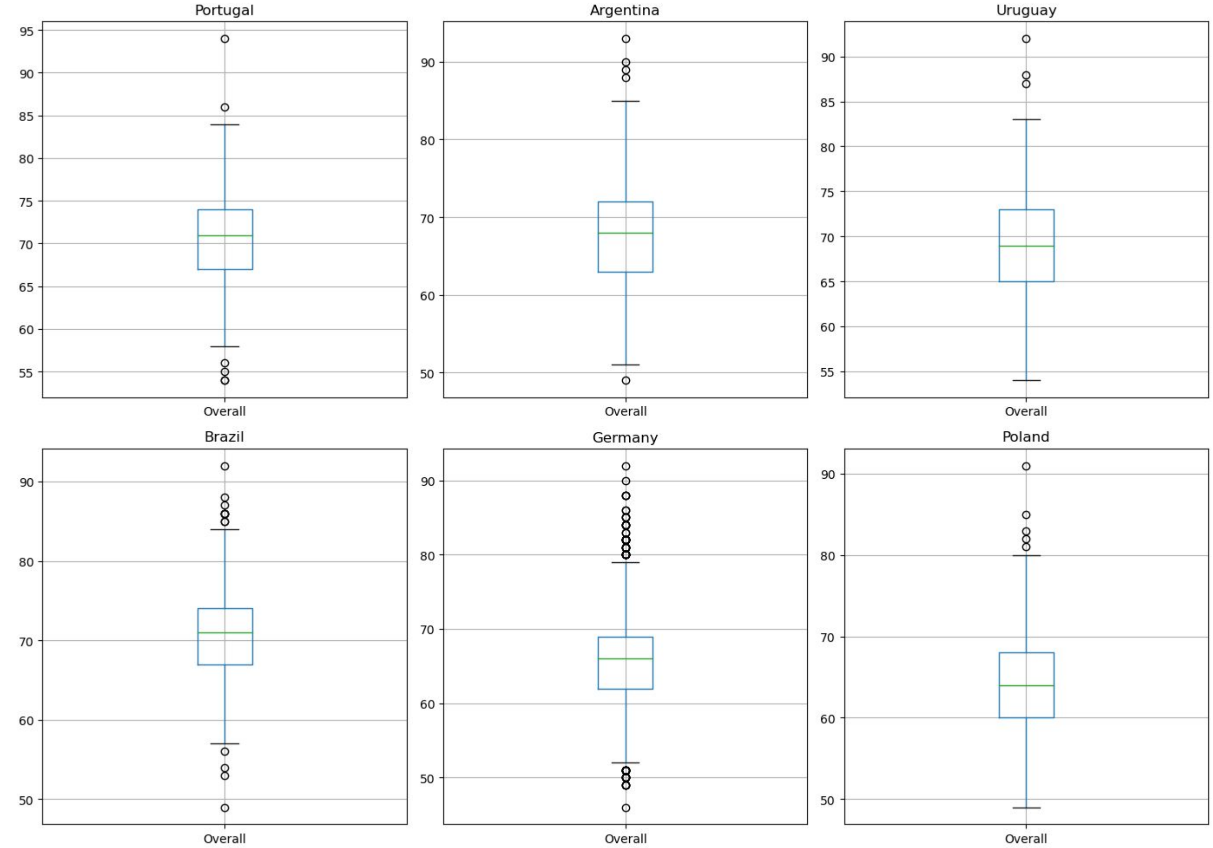

Figure 16. Top 6 Overall player ratings based on their country performance.

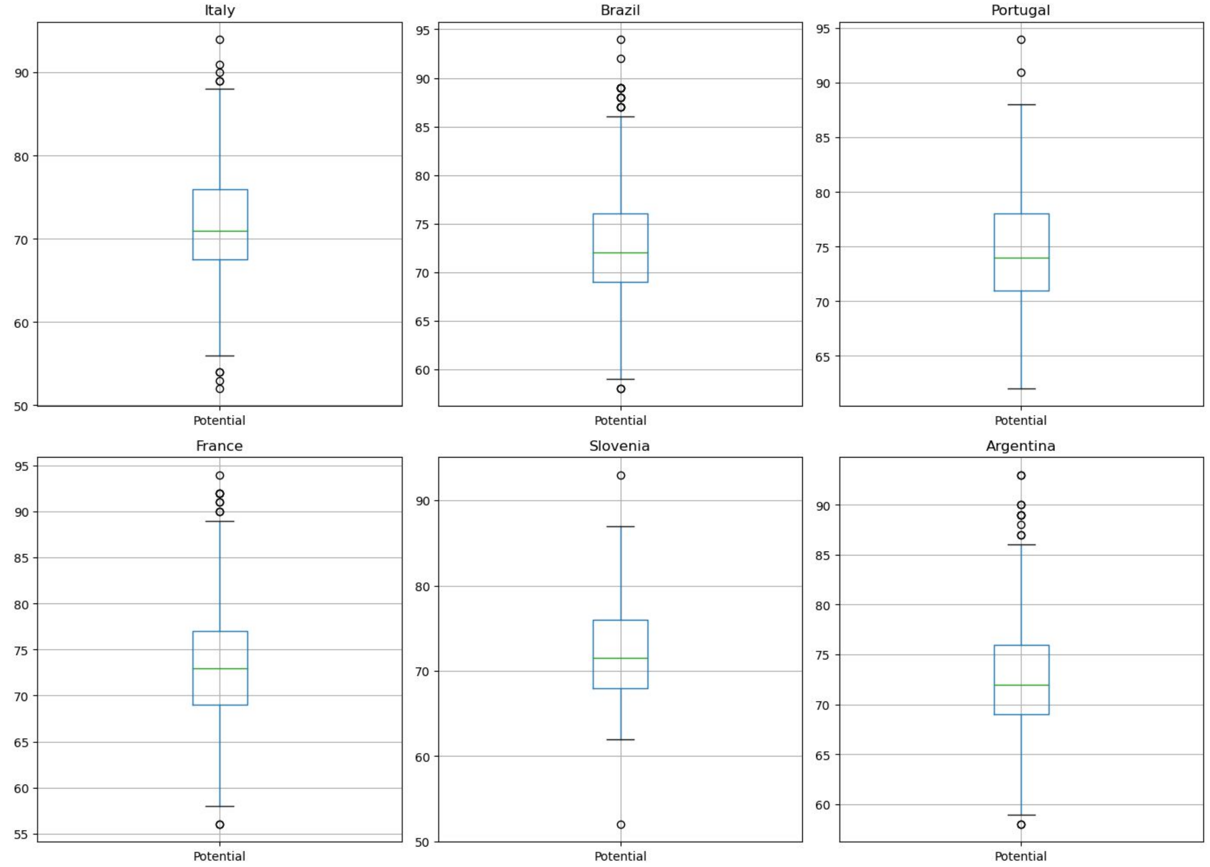

Figure 17. Top 6 Potential player ratings based on their country performance.

Figure 18 and figure 19 display top 6 players ball control and balance ratings based on their countries performance.

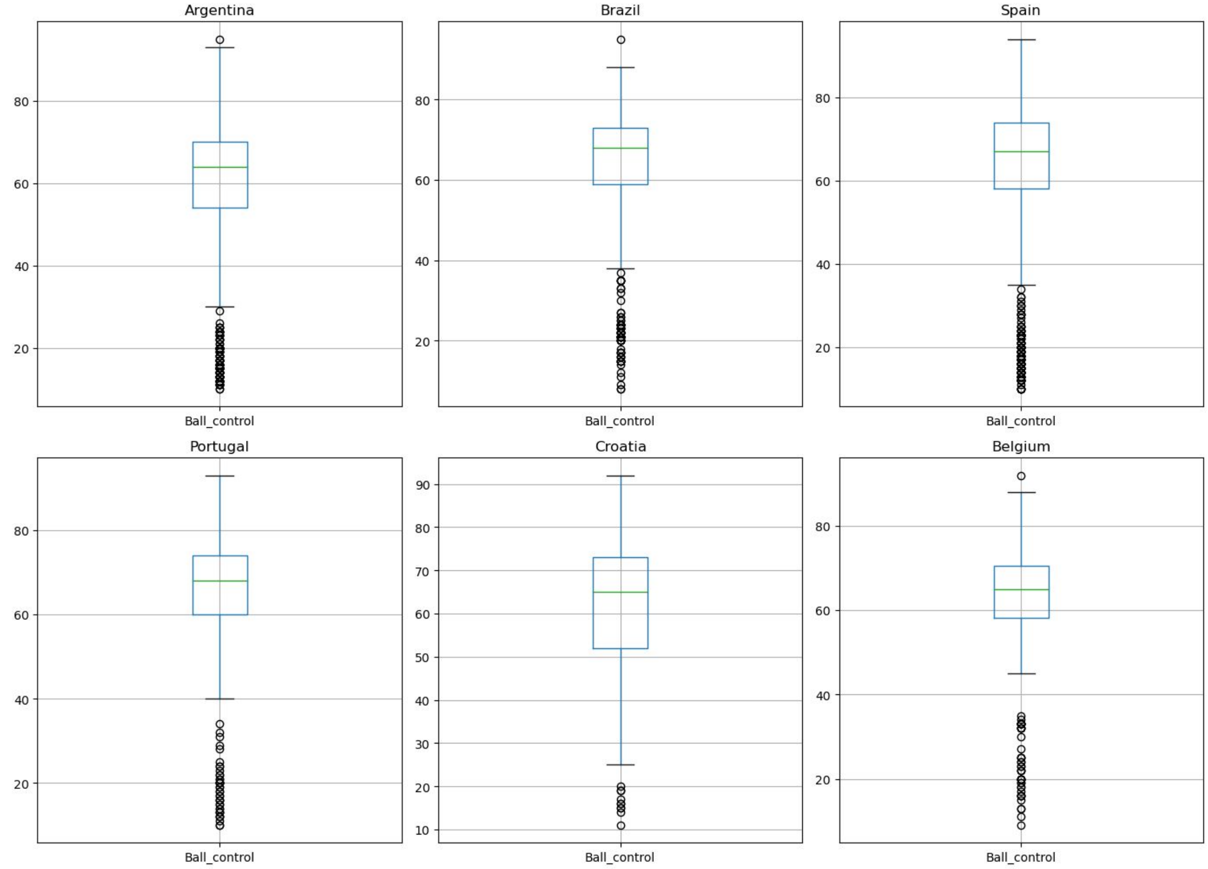

Figure 18. Top 6 Ball control player ratings based on their country performance.

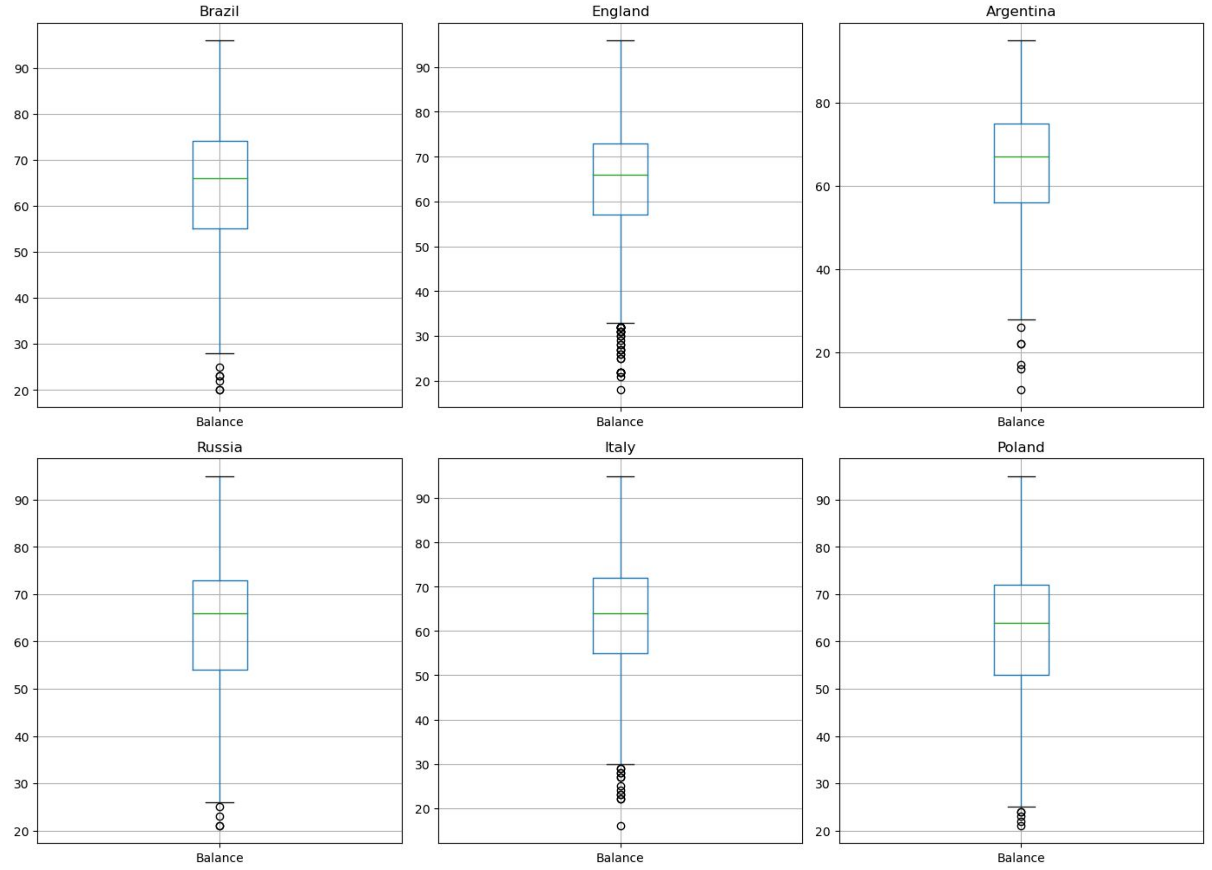

Figure 19. Top 6 Balance player ratings based on their country performance.

Figure 20 and figure 21 shows top 6 players finishing and composure ratings based on their countries performance.

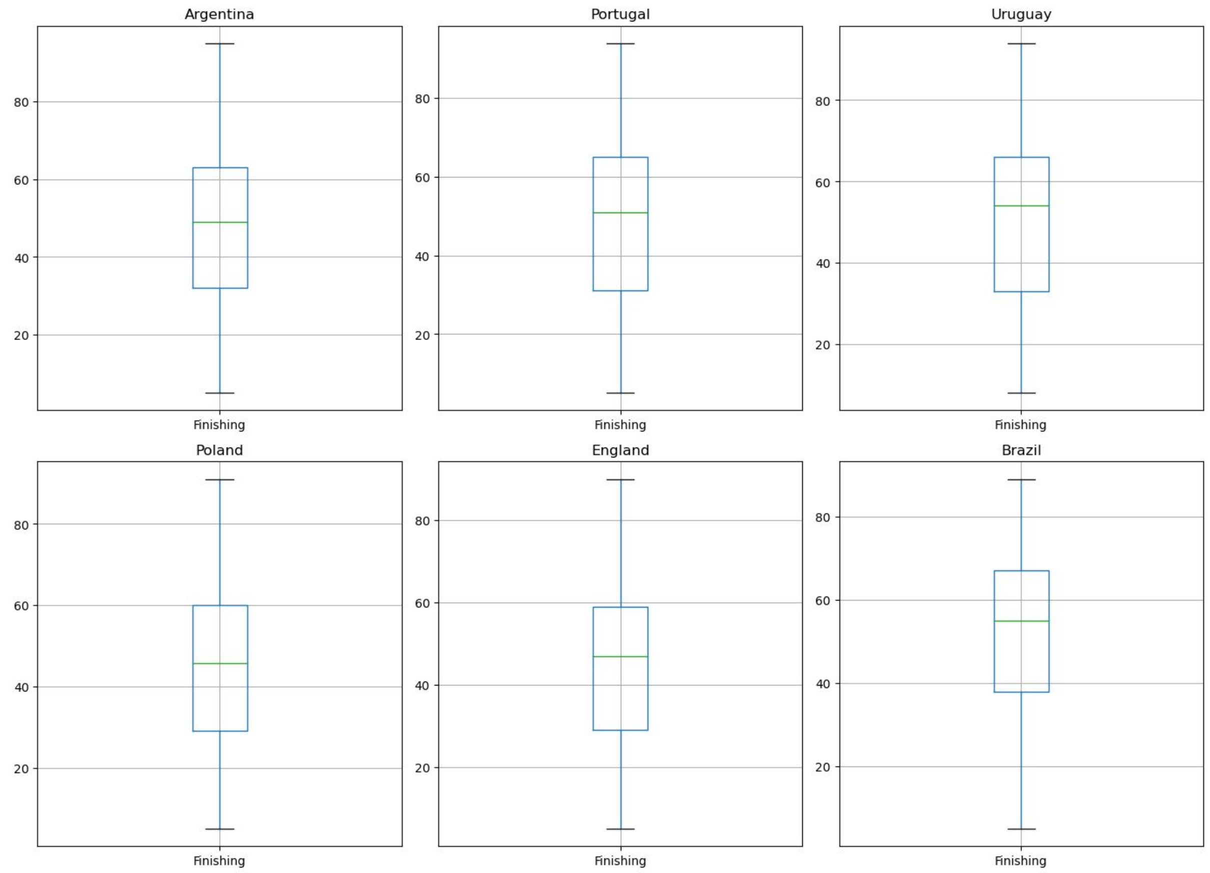

Figure 20. Top 6 finishing player ratings based on their country performance.

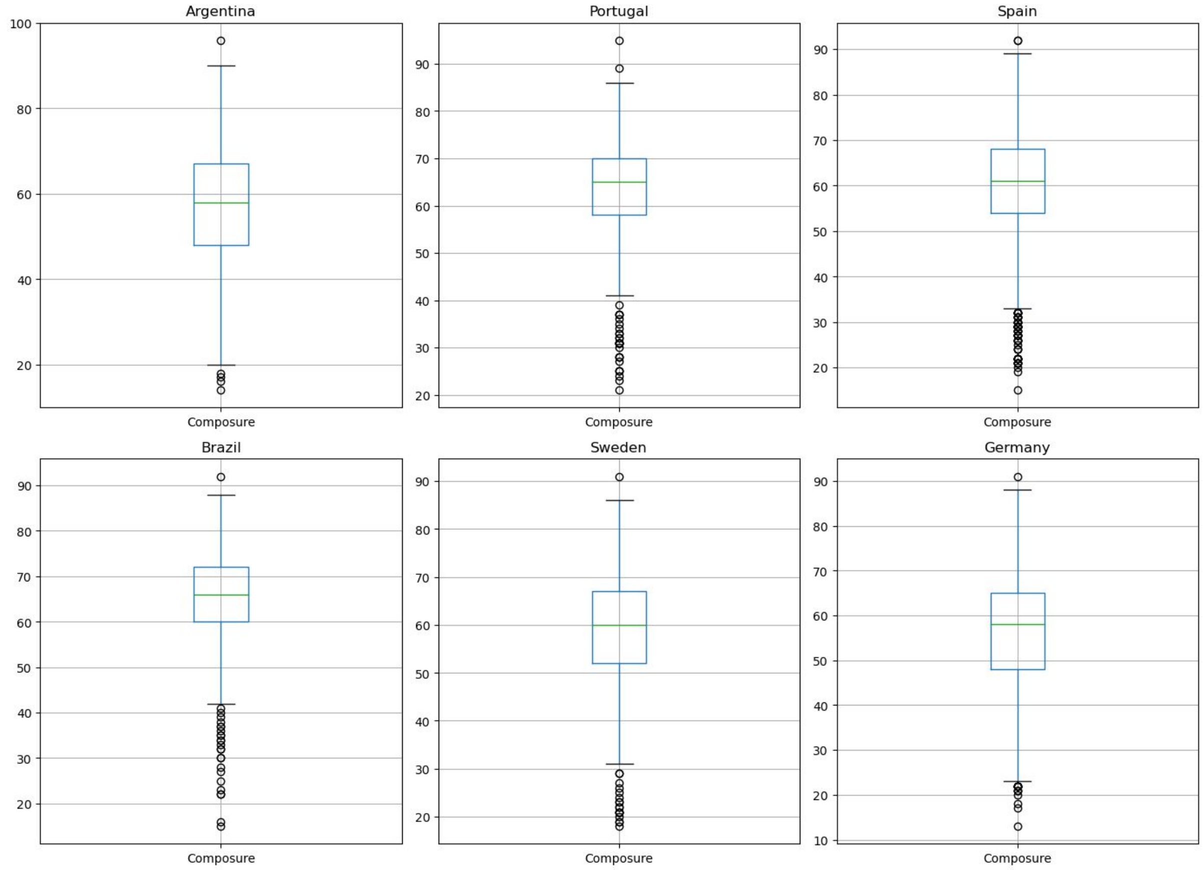

Figure 21. Top 6 composure player ratings based on their country performance.

Figure 22 shows a bar plot of the first 50 rows, 22 countries with their balance and can be observed that balance is one attribute that is essential in top player performance.

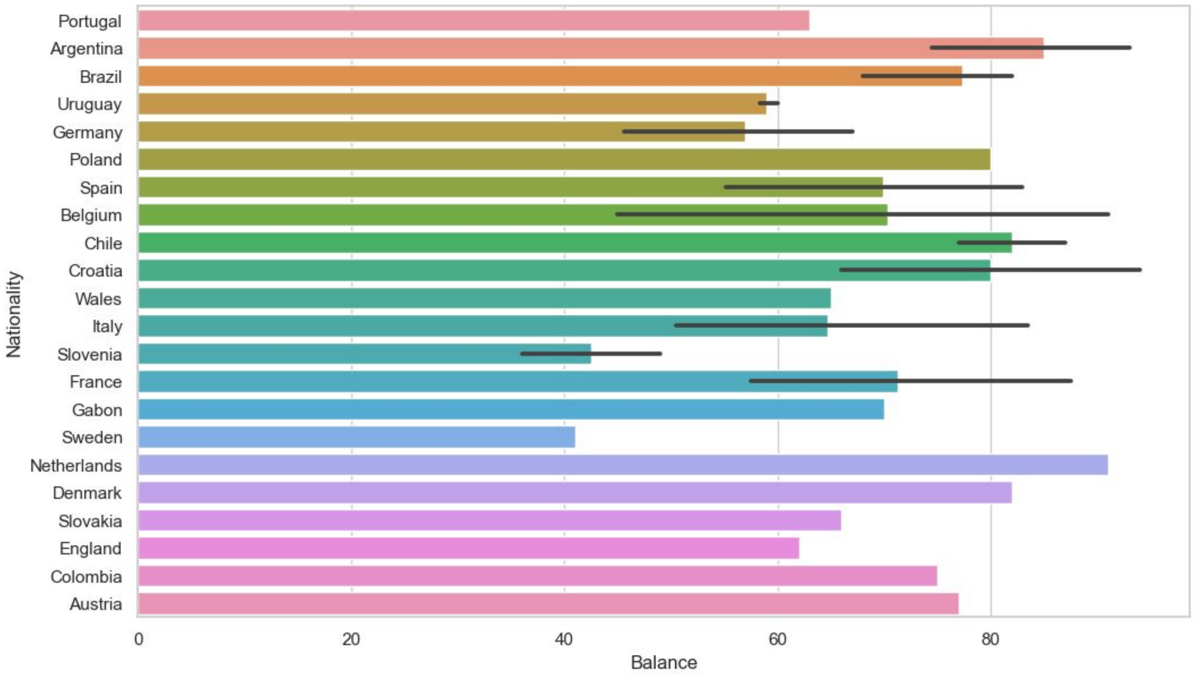

Figure 22. Overview of countries' balance performance.

Figure 23 and figure 24 display the heatmap correlation matrices of the dataset and the multiple plots for the various attributes distribution plots respectively.

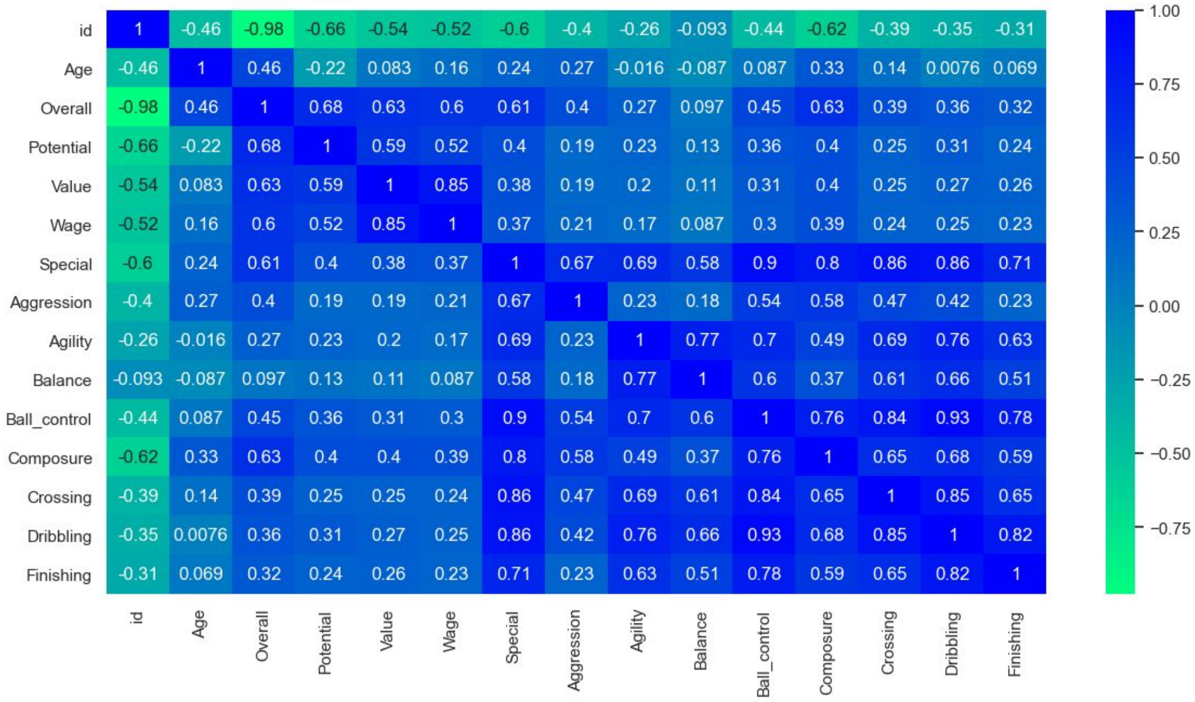

Figure 23. Heatmap of the FIFA-18 data set and its concentration of values.

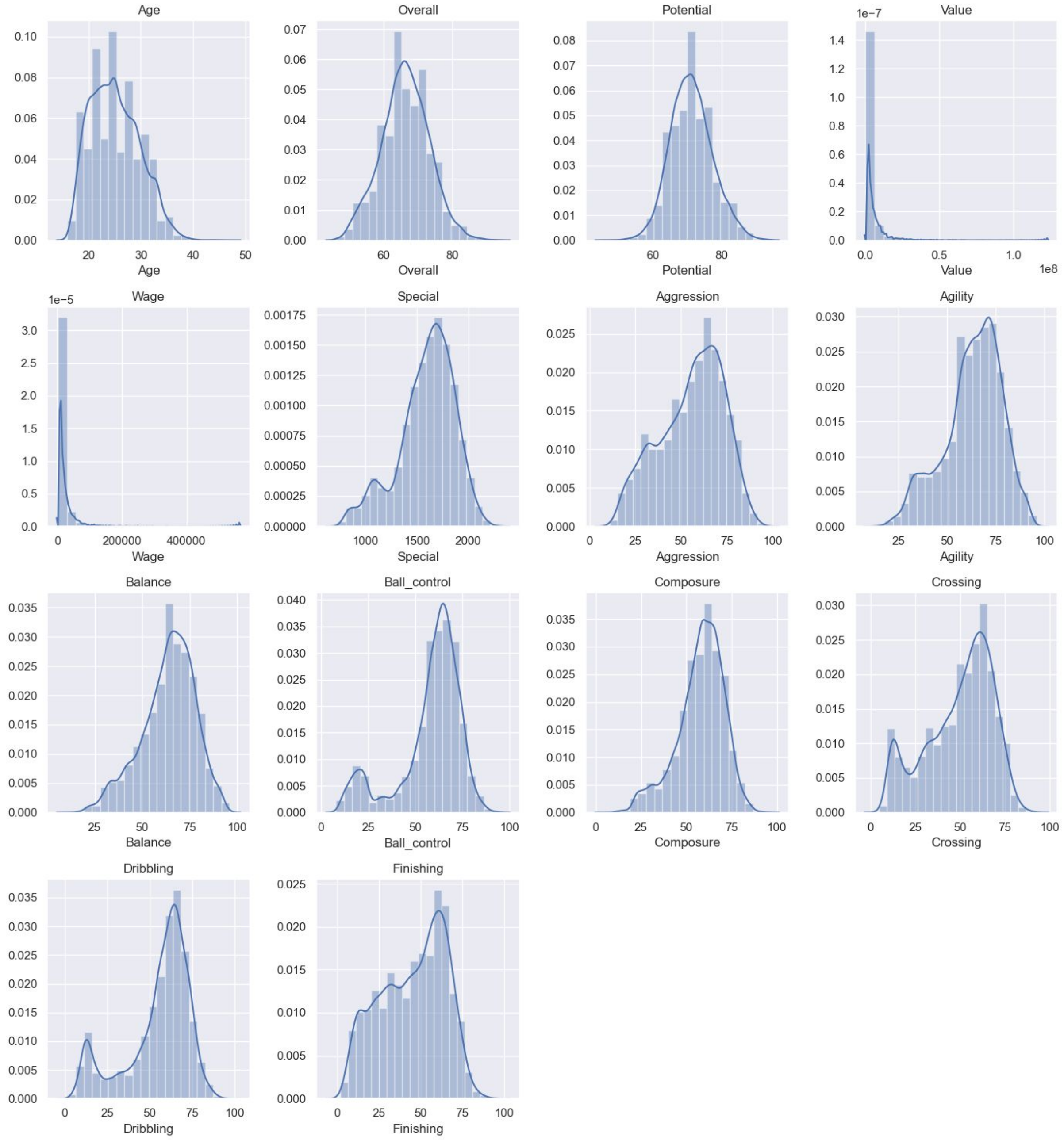

Figure 24. Distribution of modified FIFA-18 dataset attributes.

CONCLUSION

The data analysis gives a small but important picture of the attributes that play an essential role in making a top player. This analysis can be used by a football scout to narrow down player search based on countries. The box plots illustrate that the best players' overall ratings are not by coincidence as their countries focus on proper player development. A scout will select players based on performance and club's budget. The player performance has an effect on how much the club spends on the individual player.

REFERENCES

[1] Sahoo, K., Samal, A.K., Pramanik, J. and Pani, S.K., ‘Exploratory Data Analysis using Python’, *International Journal of Innovative Technology and Exploring Engineering (IJITEE).vol. 8, no. 12, pp 4728- 4735.*

[2] Restori, M, *What is Exploratory Data Analysis,* viewed 17 December 2020, < https://chartio.com/learn/data-analytics/what-is-exploratory-data-analysis/#:~:text=In%20data%20mining%2C%20Exploratory%20Data,us%20before%20the%20modeling%20task.>

THIS REPORT WAS WRITTEN BY: LEHLOHONOLO VICTOR SEBAETSE