Visualization of Football Players' Analytics

Alessio Vernarelli - Sapienza University of Rome
vernarelli.1946128@studenti.uniroma1.it - Visual Analytics Exam

***Abstract*— Football (soccer) is one of the most popular sports in the world, gathering approximately 3.5 billion fans (SportsBrowser, 2025 [1]), and it offers many possibilities for data-driven analysis. This study is inspired by fans' interest in visualizing statistics of players and discovering information from them, so it aims to provide everyone with the tools to do that autonomously. By acting on a dataset containing players from the top five European Leagues, along with more than 40 different statistics measured on the current season, as well as the estimated market value, it provides ways to compare different players between them, look for players similar to desired ones, express the relation between the measured numbers and the market value estimations, and so on. The idea is to support both casual use (e.g., personal curiosity as well as decisions in video games or amateur journalism about real-life situations) and professional use (e.g., by scouts or clubs themselves to integrate the work done on the field with the analysis of data), in a world in which everything moves towards the use of algorithms and softwares.**

***Index Terms*— Sport Analytics, Football Analytics, Players Statistics, Football Visualization, Players Similarities, Players Comparison, Football Market, Interactive Analysis.**

I. INTRODUCTION

Football is without a doubt one of the main topics in the world. It involves a lot of people in all of its aspects, fans interested in some teams or players, people who only watch the national team, professionals who work in the environment, and journalists who write news daily.

Each person involved in this world may find it interesting to analyze and use data to integrate their work or their opinions, but understanding something useful from tons of raw data is a too hard challenge, with no actual reason to be like that.

In situations like these, the visual analytics approach becomes one of the main tools to solve the problem: while reading numbers and acronyms without any context can be confusing, watching graphs immediately gives the answers someone is looking for.

This study aims to help people analyze players' performances during the season, to compare them, to compute average numbers on user-selected subsets of players, to understand relationships between performances and estimated market values, and so on.

Right now there are very few systems that allow one to do that, and the majority of free and diffused ones only offer single, static, independent visualizations.

This means that there are already ways to compare players and perform the previously stated tasks, but they are limited, they just return the statistics they are asked for, and this makes it difficult for the user to discover some insight other than what is printed on the screen.

The idea of this study is to overcome this analysis style and to give more power to the user, allowing him to define all the parameters on which (s)he wants to perform the analysis and to dynamically change them while watching also the graphs evolve according to them.

This makes it easy to see how some statistics may evolve when a variable changes, while at the moment one would need to search for the single graphs in those specific conditions and put them one near each other to try to look at differences.

Coordination and interactivity give a huge boost to football analysis potential, and this study is an example of an attempt to exploit this potential.

II. RELATED WORK

Even if I know really few applications doing this job properly online, there are a lot of studies and other papers published on this topic, trying to find solutions to allow a simple analysis.

I selected two papers that try to face the problem in the same way this study does, by selecting a player and visualizing statistics related to him, one paper that proposes something similar, but focusing on age, nationality and specific position of players instead of market value and another paper that instead starts from the same considerations of these studies, but then addresses a different problem, aiming to use statistics to predict the position of the player on the based on his values.

A. Infootmation: Visualize Soccer Player Statistics [2]

This study starts with the same goal as my proposal, as it says in the paper itself: "tools rarely focus on the interactive visualization of this data [...] Infootmation tries to fill in this lack of interactive visualizations in current tools".

However, it develops a solution differently; while it focuses on the particular detail of the shots taken by the player, storing and visualizing both the position on the field and the body part involved in the shot, my study aims to give a more general idea of the players' performances during a single season and which aspects influenced more the estimation of their market value.

The objectives of the two studies are similar, as well as the main intended user who is the casual football fan, but the explored solutions are very different, and they could easily be integrated to have a single system covering both points of view.

B. Soccer Scoop: Dynamic Visualizations for Soccer Statistical Analysis [3]

This study seems more similar to my proposal, because its goal is to represent core information about a single player and to allow a comparison between two players, but it has a crucial difference: it only provides two single and separate visualizations, without seeking coordination and interactivity, which may be a key point to discover new insights.

The intended user is a football manager who wants to sign new players and needs a tool to help with his decisions; my study also aims to support something like that, but it is more general and not entirely focused on this idea.

Visualizations are original and really well developed, they are born from the same necessities that led to this study and they represent a different result coming from a similar way to look at the same problem.

C. A Data Science Approach to Football Team Player Selection [4]

Again, this paper is about solutions to analyze different players and compare them to solve decision-making problems of anyone who needs to choose between players for any reason.

However, something is different in this case too. The study looks at the players not through their single statistics during the season, but it divides them into groups on the basis of their age, nationality, and positions, and assigns them just a single value representing the overall rating. At this point, similarities and comparisons between players are computed on these parameters.

While my solution agrees with this idea on the computation of similar players, where similarity is calculated with a sort of performance score on the overall statistics of the players, my software is more oriented to give importance to the single statistics, taking into account that one may be interested only in some characteristics of a player without caring about other ones.

Clustering players on age, position, and nationality is a perfect choice when the application is used to aid in building a team, because there are constraints to respect. But as for my goal, it was better to let the user define the values to use to restrict the search, still remaining on quantitative attributes.

However, the selection of players by position is still present in some way also in my study, because some of the visualizations consider only players sharing at least one role with the selected one, even if I just consider the macro-position (the role) and not the specific area of the field covered by the player.

D. In-game behavior analysis of football players using machine learning techniques based on player statistics [5]

As previously stated, this paper starts from the same considerations that led to this and the other presented studies, but takes them to a completely different destination.

Starting from the need for visualizations to analyze sports data, it uses machine learning to recognize the position of players from their statistics by analyzing the patterns and the most related ones.

Even if the goal of the study is completely different from the one of this paper, I decided to add it to show how a similar idea and a similar initial dataset may lead to different solutions developed by different people.

In summary, visual analytics for sports, particularly for football, is something that offers a large number of possibilities, even unexplored ones. The great number of people interested in the matter and the various kinds of proposed solutions justify the continuous research for new techniques and new ways to use already existing ones.

The key point of this paper is to propose the application of coordinated and interacting visualizations in the analysis of the statistics of players to support curiosity and decision making. The ultimate goal is to make everything clear for the users just by looking at the graphs, without the need to deeply understand the numbers and the math behind them, while giving them the power to lead the analysis in any direction.

III. DATA

To perform any kind of statistical analysis, one of the main problems is to gather a large quantity of data. Fortunately, it is not difficult to find football data on the web.

A. Dataset

This study is based on two datasets downloaded from Kaggle, "Football Players Stats (2024-2025)", a dataset of 2854 rows and 165 attributes, storing all the players from the top 5 European Leagues with almost any kind of statistic measured in the current season, and "Football Data from Transfermarkt", a dataset collecting the market values for all the players registered in the system of Transfermarkt, the most famous website about football and market values specifically.

B. Preprocessing

My idea was to have a unique dataset with all the players from the most important leagues, their stats, and the respective market values, but there was no dataset already doing that, so I decided to take different sources and merge them in some way. To do that I took the dataset about the top five leagues and I dropped all the columns I was not interested in. I performed a left join with the other dataset to have the market value of each player, only keeping the rows corresponding to those players in the top five leagues and assigning a null value to eventual tuples not in the second dataset, because I didn't want to lose any player.

However, this was already a problem: datasets created by different people are likely to have some differences even for the same values. Maybe some players are called with a nickname instead of their first name, maybe foreign players with special characters in their names, not in our alphabet, are written in different ways, and so on. To overcome the issue I kept all the successful joins and then I scanned again the remaining tuples of the first dataset, searching this time not

for a natural join on the common attribute, but calculating the similarity score between two names using fuzzy matching. In this way, I took the market value from the joined tuples where the names were at least 90% similar and I put it in the corresponding missing values created by the left join.

This was still not enough as some more cleaning was necessary. During a season a player may change team in the transfer window and this leads to having multiple tuples for the same player but with different stats and different teams. I decided to keep those rows separated because a new environment, a new coach, a new team may be key points in a player's performance, but to keep a reference to such rows I changed the attribute containing the ID of each row to have the same value if the player is the same.

As for the last cleaning step, I needed to do something by hand. There are cases in which different players have the same name and the same surname even if they are different players, especially in lower leagues. This event happening during the join created some tuples about players not existing in reality, so I had to check duplicates and manage them by hand, deciding who to keep and who to delete.

In the end, I decided to perform another pass to prepare data for visualization. Since the numbers of players are not comparable when they refer to different periods, I converted all the numerical statistics in the 'per 90 minutes' version, using the number of minutes played in the season stored in one of the attributes. This allows for immediate comparison of numbers avoiding the risk of comparing different measures and producing misleading results.

The final dataset consists of a table of 2744 rows and 48 attributes, describing players, some information about them, their statistics per 90 minutes, and their market value.

IV. VISUALIZATIONS

The idea of the project is to divide the page into 6 sections through a 2*3 grid. Approximately, each of the six sections will contain a visualization.

The key point of the project is that any time something is selected or a filter is modified, all the graphs are recomputed and updated to dynamically show the changes. This is the interaction that allows us to easily understand what makes the difference between various situations and which parameters influence the values.

A. Scatterplot: Dimensionality Reduction

First of all, the dataset is preprocessed again to create two additional columns, containing the values for the two principal components of PCA (Principal Component Analysis). This is a dimensionality reduction technique that, given a bag of multidimensional points (the players), finds a way to represent them on two dimensions, in such a way that they can be visualized on a scatterplot.

In particular, PCA normalizes all the numerical features, because it needs all of them to have the same scale, and then takes the linear combination of the attributes with the highest variance, followed by the second highest variance which is uncorrelated to the first component. Those will be the dimensions on which the data will be projected.

The scatterplot allows to zoom in on the points, recomputing each time the scales to visualize the correct proportions, and the user can click on the points on it to select at most three players. When a fourth player is selected, the oldest selection is dropped.

Points are also encoded through colors, which reflect the league in which the player plays (and that can also be selected to filter the data by leagues) and through size, as bigger points represent players with a higher market value.

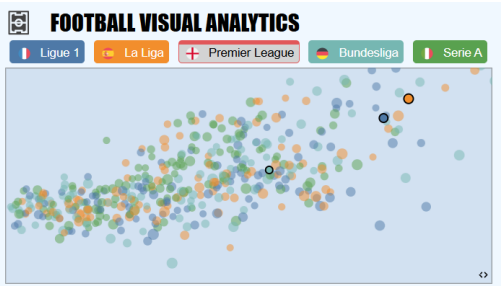


Fig. 1. PCA Scatterplot + League Filter

B. Field Filter

This visualization is the drawing of an actual football field with four clickable bands, one for each role (FW - Forward, MF - Midfielder, DF - Defender, GK - Goalkeeper). By clicking on each band, the user can add that role to the set of the selected ones and only the players belonging to the selected roles will be shown in the scatterplot and will be available for analysis.

There is also the option to ignore all the players with less than 500 minutes, because they may add weight in the computation of the averages or of the proportions of values while having all the statistics inevitably low.

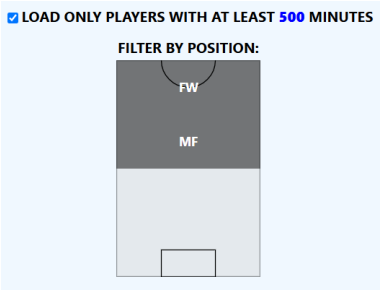


Fig. 2. Position Filter

C. Similar Players

The user can also directly search for a player through a search bar instead of looking for the right point in the scatterplot. The system tries to auto-complete the name of the player while (s)he is typing based on the names stored in the dataset and, once a player is selected, it computes the 10 most similar players, ordered by market value, using cosine similarity on the set of attributes.

The list will be composed of players different from himself with similar statistics and sharing at least a role with him.

Fig. 3. Similar Players

D. Market Value and Correlations

This section is divided again into 2 visualizations.

First of all, there are box plots that are used to show the distribution of the players concerning the market value in the currently selected leagues. They are useful to see which league has the majority of players having a higher or lower value, which league has the outliers with the highest market values at all, and so on. The color encoding follows the scatterplot one.

Fig. 4. Box plots

The box plots are followed by a visualization created through the use of Crossfilter [8]: it is a horizontal bar built on the range of the market values on the overall dataset, which allows to select a sub-range and to pan the selection or to enlarge or reduce it.

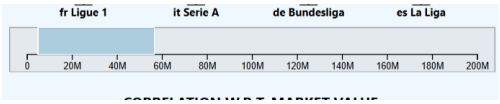


Fig. 5. Market Value Crossfilter

The last visualization of this area is the Pearson's correlation of the statistics over the current selected subset with the market value. It is a bargram starting from a middle horizontal line on correlation zero, with bars going up for positive correlations and down for negative correlations. In the graph, only the top 5 and the bottom 5 statistics are represented to have a clear visualization.

This is the way the user can understand which features are causing such market values to be estimated. The color is assigned through a d3.js scale interpolating red and blue. Dark red is assigned to statistics strongly negatively correlated (-1), while dark blue is assigned to statistics strongly positively correlated (1). Of course, when the subset contains a decent number of players, it will be difficult to have some attribute spiking high on the others, so colors will be quite faded.

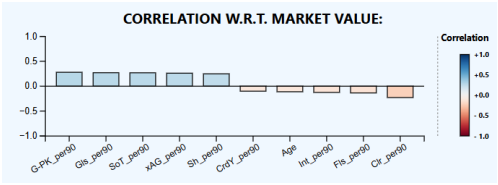


Fig. 6. Correlation of Statistics w.r.t. Market Value

E. Radar Chart

The radar chart is one of the two main visualizations focusing on visual encoding players' statistics. This one aims to compare different players on a subset of pre-selected statistics, which change on the basis of the players' position: if a goalkeeper is selected along with movement players, they will all be compared on goalkeeper stats, and so on.

It is important to note that a star chart is not supposed to represent the absolute value of the statistic for the represented players, but it computes the proportion of that attribute for that player in relation to the other players in the subset. This means that a big shape covering the entire radar tells that the player has numbers which are higher with respect to the other considered players, but may still be low if interpreted following common sense or another point of view.

On top of the radar chart, the cards of the selected players are shown with the color used in the chart. These cards are clickable and allow switching the current player among the selected ones, where the current player is the one used for the visualizations involving a single player, like the list of similarities. The players can also be removed from the list.

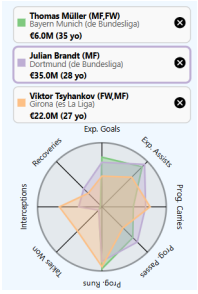


Fig. 7. Star Chart comparing different Players

F. Stats Comparison

Such a current player is then also used for the last visualization. In this case, another set of attributes of the player is compared, but this time to the average values of the currently selected subset.

The way in which this visualization works is the following: each attribute is taken singularly and the respective values of both the player and the subset are converted into the true value over the whole season, not the per 90 minutes version. At this point, the system computes the mean and writes it in the middle, ready to do the same work as the correlation bargram, but vertically. However, the length of the bars is calculated after normalizing each row on the range of values of that attribute. The idea is that bars are not supposed to be compared between them, because the weight of the values would be too different (think about the number of minutes of a player compared with the number of goals). Instead, the information it wants to convey can be easily grasped from the direction towards which the bar is going: a bar going right represents a value bigger than the average one, while a bar going left represents the contrary.

The length and the color encode just how much this value differs from the average one and it is computed in percentage. If a player has four goals and the average is two, the value is the double and then it is 100% bigger, so the bar reaches the respective length (limited by a maximum computed in such a way to leave free space to add the label with the numeric value) and the color will be dark blue and vice-versa.

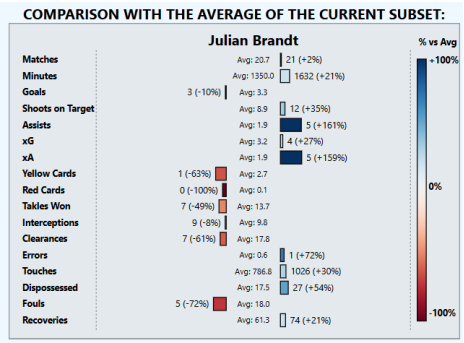
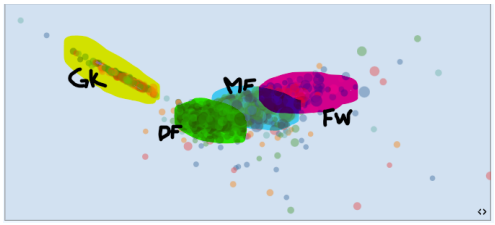


Fig. 8. Comparison of a Player w.r.t. Average

V. RESULTS

Using the systems makes it easy to discover useful information, both known and unexpected ones.

First of all, we can easily see that the PCA is working because players in the same positions, and then likely with similar stats, are always near each other. Even if I didn't use t-SNE which actually focuses on dividing the data into clusters, we can recognize patterns, with the goalkeeper one being the most evident isolated on the left.



Speaking about goalkeepers, age is the attribute with the most negative correlation: this is something that may seem strange for someone not really into football, but it reflects a popular belief which sees goalkeepers becoming better with experience and having a longer career with respect to other positions.

In contrast, selecting only the range with the highest market values, we can see that age is really positively correlated. Nowadays the prices on the market are always growing and young stars such as Lamine Yamal, Vinicius Junior and Erling Haaland have a huge market value.

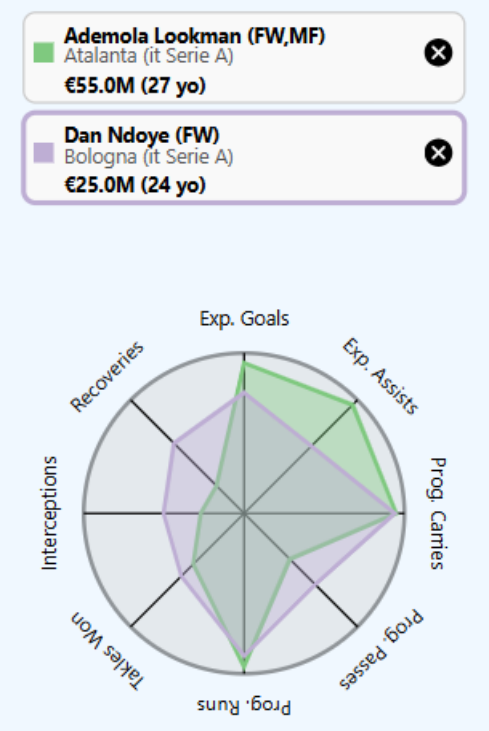
If we decide to select defenders instead, we can see that the two most correlated attributes are Touches and Progressive Passes. This is coherent with the recent typical play style of teams which involves building from the back. It is common to see teams that refuse to launch the ball away and instead try to create some potential offensive action by playing the ball from the defense in their own area.

Looking at the box plots when no filter is applied, we can see that the distribution of Premier League players has a higher median market value and the 75% of data is in a very small range near it, showing how Premier League is the league with the higher market values on average, even if the league with most players above 150M is La Liga.

Italy and Ligue 1 have sensibly lower outliers, either because players in these leagues are undervalued or because the level is in fact lower than the other leagues.

Filtering position by position it is clear that attackers are always the players with the highest market values since their outliers in the box plots reach the maximum height, while defenders and goalkeepers ones are really low. It is very difficult for a defender to be valued highly.

An example of use of the application may be the following: suppose a user is playing a football video game and wants to buy Ademola Lookman (50M), but it is too expensive for his budget. He decides to search for Lookman on the application and to search for similar players in Serie A with a market value of 30M at most. One of the most similar players returned by the system is Dan Ndoye (25M), who has, in fact, a similar shape in the star chart and has similar stats with respect to the average, even if they are obviously lower.



The user can also reason in general terms. We can see that the statistics of Dan Ndoye are good when only Serie A is selected as a league, but the bars become smaller and lighter if we use Premier League as a reference.

This shows how each country has its own play style which should participate heavily in decision making. The average of the same offensive stats used to analyze an attacker is much higher in England, which has a league in which it is easier to score with teams playing openly.

Italy instead is famous for the defense and then the statistics of an attacker in a country in which usually the number of scored goals is low, are better in proportion.

VI. CONCLUSIONS

A. Future Enhancements

Nothing is perfect and every project can be improved, especially by taking inspiration from the other papers on the same subject and further developing already present ideas.

Some additions that could be made to enhance the system are: support to more leagues, considering also second or third divisions and additional countries; more precise system of dividing player by position, considering not only the role but directly the area of the field in which he plays; support over more seasons, with the possibility to look at the aggregated data or to pick a particular season over the others.

Maybe a nice way to implement machine learning techniques would be to try to predict the market values of some player given his statistics and considering the already present ones.

An interesting idea, even if not strictly correlated and probably not integrable in this project, would be to monitor and represent also live statistics during a match, visualizing analytics about teams and players' performances, and comparing already terminated matches between them.

B. Conclusion

Football (soccer) is one of the most popular sports in the world, gathering approximately 3.5 billion fans. Fans like to dream, to create hypothetical scenarios, to discuss about news and opinions and all these things are better when supported by data.

This study wants to give people the means to easily analyze and discover new things about football players and it wants to do it through interactive visualization and cooperative graphs.

It uses dimensionality reduction to represent multidimensional data on a 2D scatterplot and allows the user to interact with it to create a series of visualizations. This study wants to compute the correlation of each statistic of a player with his market value, in such a way the user can understand why a player is valued more than others; it allows to compare players on different statistics and to analyze how a single player is performing among the average of his role, considering a subset selected by the user. It also wants to give alternatives to professionals or to fans playing videogames when considering a player, through the computation of similar profiles given their statistics.

In conclusion, this study wants to give everyone a tool that can be used to have fun with interactions and graphs and to grasp correct and meaningful insights from what today is a passion and a work for a large number of people in the world.

REFERENCES

[1] James Buttler, Top 10 Most Popular Sports In The World July 2025, https://sportsbrowser.net/most-popular-sports/.
[2] Dubois, Yann. "Infootmation: Visualize Soccer Player Statistics.".
[3] Rusu, Adrian, et al. "Dynamic visualizations for soccer statistical analysis.". 2010 14th International conference information visualisation. IEEE, 2010.
[4] Rajesh, P., Mansoor Alam, and Mansour Tahernezhadi. "A data science approach to football team player selection." 2020 IEEE international conference on electro information technology (EIT). IEEE, 2020.
[5] García-Aliaga, Abraham, et al. "In-game behaviour analysis of football players using machine learning techniques based on player statistics." International Journal of Sports Science & Coaching 16.1 (2021): 148-157.
[6] Hubert Sidorowicz, Football Players Stats (2024-2025), https://www.kaggle.com/datasets/hubertsidorowicz/football-players-stats-2024-2025?select=players_data_light-2024_2025.csv
[7] David Cariboo, Football Data from Transfermarkt, https://www.kaggle.com/datasets/davidcariboo/player-scores?select=players.csv.
[8] Crossfilter, https://crossfilter.github.io/crossfilter/.