VISUALIZATION OF FOOTBALL ANALYTICS

VERNARELLI ALESSIO
1946128

GOAL OF THE PROJECT

Aim of the exam and

what I tried to implement

GOAL

The exam aims to develop a project that, starting from a dataset, uses interactive visualizations and analytics to extract information from the data easily.

I chose to analyze a dataset of **football statistics** with the goal of providing users with the means to compare players, identify similarities between them, and extract insights such as what makes a player more valuable.
The idea comes from the fact that I, as a football fan myself, often want some way to visualize the performances of some player, either with reference to real market news or to help me make decisions in football video-games.
However, a tool like this may also be useful for professionals working in the field, since scouting is an area in football that is evolving towards the use of data and algorithms day after day.

DATA STRUCTURE

Definition of the dataset on which

I created my visualizations

DATASET

As for the dataset, I downloaded two of them from Kaggle:

- one only containing the **statistics** of the players from the top 5 European Leagues (https://www.kaggle.com/datasets/hubertsidorowicz/football-players-stats-2024-2025?select=players_data_light-2024_2025.csv);
- the second one from Transfermarkt, the most famous football website, containing the **estimated market value** of all the players (https://www.kaggle.com/datasets/davidcariboo/player-scores?select=players.csv).

kaggle

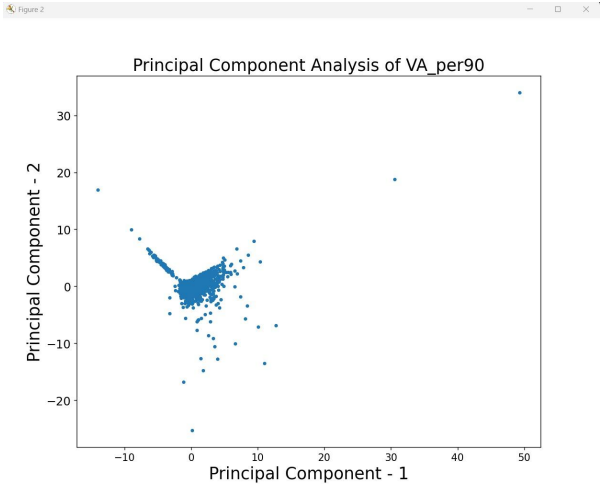
PREPROCESSING

1. First of all, I **removed** all the **attributes** I didn't need, because there were more than 70 in total.
2. I **joined** the two datasets on the name of the players, making sure to keep only those players in the top five leagues. To do that I needed to use a **similarity score** to find matches, because some players were stored with a shorter name, some with special characters of other languages and so on.
3. At this point, I needed to manually manage all those cases in which two **different** players had the **same** name and surname, deciding who to keep in the dataset and who to drop.

4. Then, I decided to convert all the numerical attributes in the '**per 90 minutes**' version of the statistic, using the number of minutes played. In this way the numbers are already comparable in some way, being distributed over the same period of time.
5. As a last thing, I used a dimensionality reduction technique, **Principal Component Analysis (PCA),** to find the components on which the data were varying the most, and I saved them in 2 additional columns 'x' and 'y'.

At this point the dataset was ready and I was able to exploit PCA to draw a scatterplot even if the data had many dimensions.

A S = # tuples * # dimensions = 2744 * 48 = 131712.



RESULT

The final result is a dataset where each tuple represents a player with **descriptive attributes** like its ID, Name, Nationality, Position, Team, Competition, Age, Matches Played, Minutes, etc.
And a series of attributes representing the **statistics** of that player in the current season like Goals Scored, Assists, Yellow/Red Cards, Errors, etc., all converted in the per 90 minutes version, as well as his **market value**.

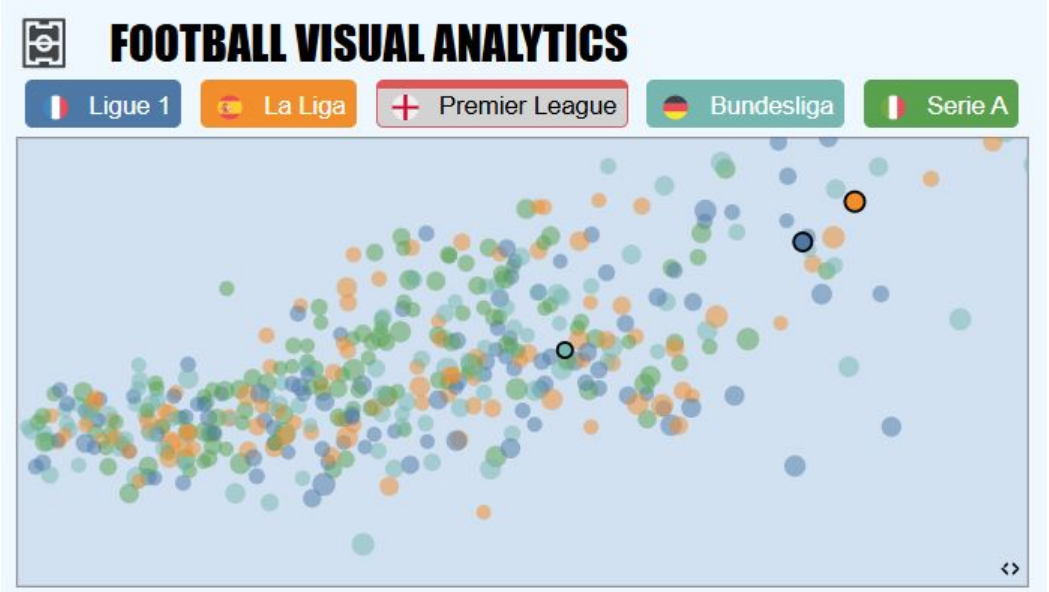
The same player will appear two times if he changed team during the season, to consider his performances in relation to the team.

VISUALIZATIONS

Chosen visualizations to represent the data

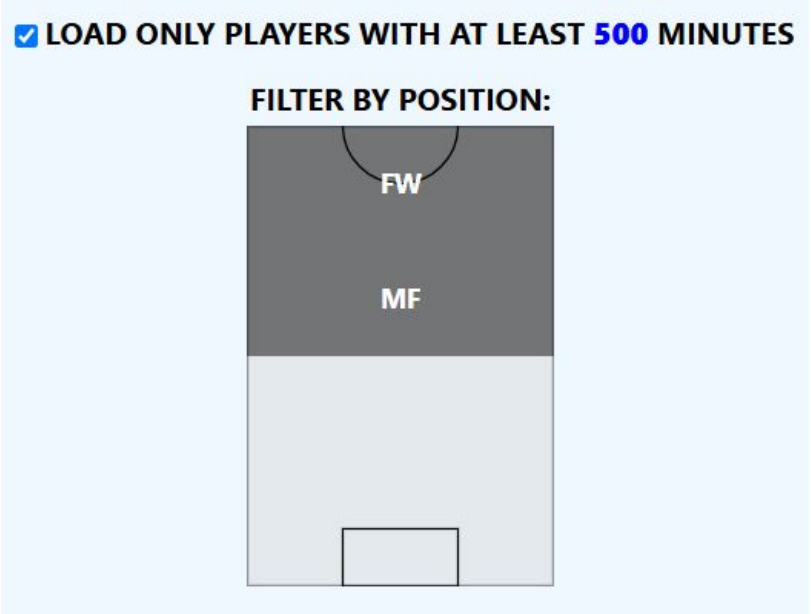
SCATTERPLOT

The **scatterplot** is created by drawing the points on the principal components returned by **PCA**, in such a way that they can be represented on the two dimensions holding the most possible information.



FIELD FILTER

I needed a way to filter the players on their **position** and I thought about this solution: the position is encoded exactly as the respective area occupied on the **field**, so that it is intuitive for the users to know what to do to achieve their task.



SIMILAR PLAYERS

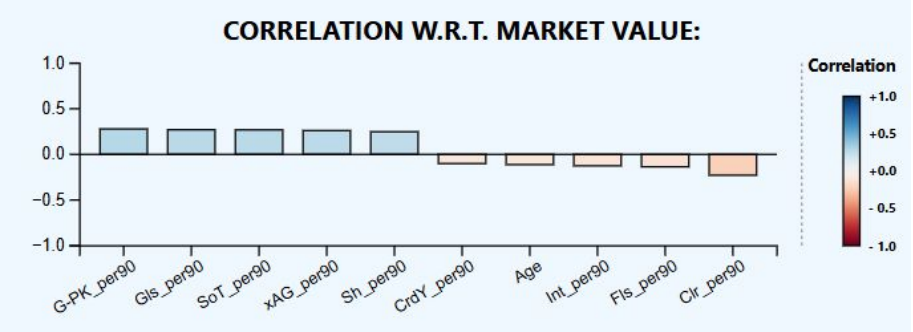
To show similar players I created some **cards** containing the information of the players returned by the software.
The idea is to show the key information that may interest the user while keeping them **sorted** on the market value.

BOXPLOTS

I used boxplots to easily convey the **distribution** of the currently selected subset of players on the **market value** for each league. Boxplots are among the best visualization techniques to represent **ranges**.

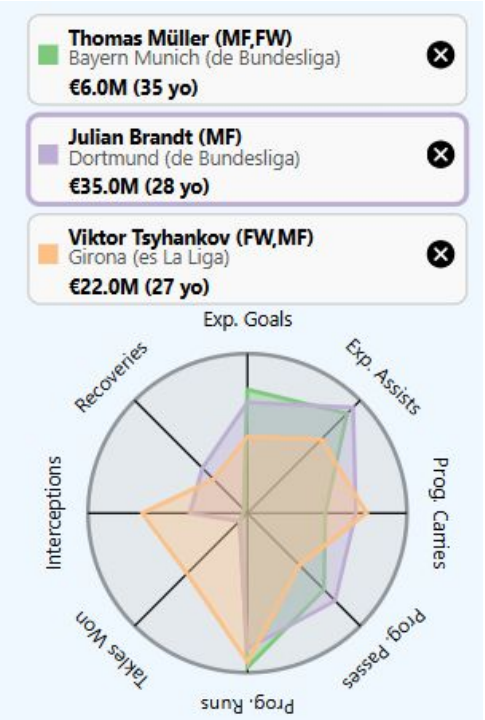
VERTICAL BARGRAM

To represent the **correlation** of statistics with the **market value** I picked a bar diagram with vertical bars, all starting from a neutral middle zone where the correlation is 0. From that point, bars go up if there exists a **positive correlation**, showing a connection between the attributes, while they go down if the correlation is **negative**.



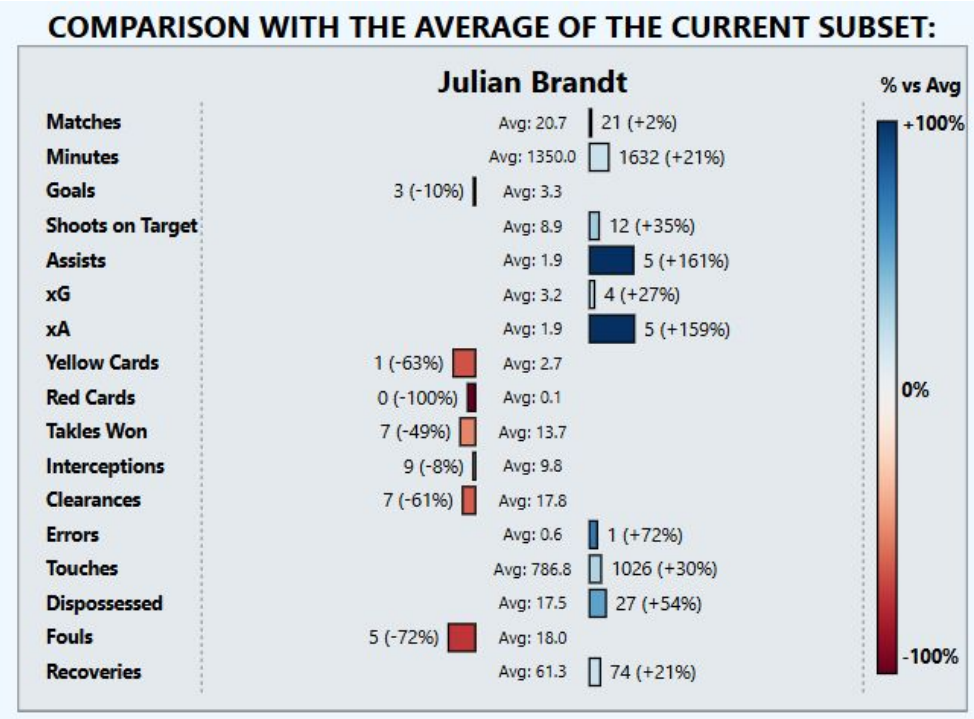
RADAR CHART

This is one of the most common visualizations in the field of football statistics, but it is essential. A **radar** or **star chart** doesn't represent the absolute value of an attribute, but the **proportion** of that attribute in relation to the other ones in the subset. This means that a spike towards one of the stats just indicates that the value is higher w.r.t. the others, not that that player has a high value for that stat.



HORIZONTAL BARGRAM

To compare the **absolute value** of some statistics of the player with reference to the average, I picked a **horizontal bargram**. Since each bar represents a stat with a different scale of values it is not possible to actually compare bar lengths, so I opted again for the idea of bars going towards the **positive** or **negative direction** highlighting the difference.



CONCLUSION

FINAL CONSIDERATIONS

These visualizations allow the user to extract insights about similar players in many ways, by directly searching for a player and retrieving the similar ones computing cosine similarity on the values in the dataset, by selecting nearby points in the scatterplot, by comparing shapes in the radar chart, and by looking at the numbers in the comparison area.

They also allow to make an in-depth analysis on what typically influences the market value of the players since almost any visualization considers that (and everything influences the results of Pearson correlation in the bargram for single players).

I chose to implement this project because I'm really interested in the topic, I always see friends and people in general try to use statistics when talking about football, but correctly using visualizations take this concept to another completely different level; the power graphs have in explaining data - also to people who look at this only as an hobby and not as a work - is wonderful, and I feel like there are too few tools doing this right now, even if this topic is one of the most diffused in the world.

Also, the few free software programs I know that do something similar only use a series of single, independent, and static visualizations, while the interactivity and coordination between different visualizations can really help in the discovery of unexpected facts.

VISUAL ANALYTICS 2024-2025