Report - June 2024

Football Data Analysis Report

Federico Paschetta

Table of Contents

01. Introduction

In the rapidly evolving world of professional football, **data analysis** has become a crucial element in enhancing team performance, scouting talent, and strategizing game plans. This project aims to leverage advanced data analytics to gain deeper insights into the **performance** and **potential** of football players. By employing sophisticated statistical techniques and machine learning algorithms, we analyze a wide range of player metrics, from basic statistics like goals and assists to more complex variables such as pass completion rates, expected goals and progressive carries. The primary objective of this project is to create a comprehensive analytical framework that can help coaches, analysts, and scouts make informed decisions based on empirical evidence. This involves the collection and processing of large datasets and feature engineering to derive meaningful metrics.

Key activities in this project include:

01. **Data Collection**

Gathering extensive data from various sources including match reports, player tracking systems, and historical performance databases.

---

02. **Data Cleaning and Preprocessing**

Ensuring the data is accurate, consistent, and ready for analysis by handling missing values, outliers, and standardizing formats.

---

03. **Data Analysis**

Visualizing and summarizing the main characteristics of the data to identify patterns, trends, and anomalies.

---

04. **Feature Engineering**

Creating new features that can provide additional insights or improve the predictive power of our models.

---

05. **Model Development and Evaluation**

Building and testing various machine learning models to predict player performance and identifying the most effective ones through rigorous evaluation.

By the end of this project, we aim to deliver a robust analytical tool that not only enhances the understanding of player dynamics but also provides a competitive edge to football teams through data-driven insights. This report will outline the methodologies employed, the results obtained, and the implications of our findings for the future of football analytics.

02. Methodology

The methodology for this project encompasses several key steps, each critical to achieving accurate and meaningful insights from the data. These steps include data collection, data cleaning and preprocessing, exploratory data analysis, feature engineering, model development and evaluation.

01. Data Collection

Data collection is the foundational step in our analysis. We sourced data from various reputable sources, including:

- **Transfermarkt**
  Demographic data about players
- **FBref**
  Advanced player performance metrics

---

02. Data Cleaning and Preprocessing

Data cleaning and preprocessing ensure that the data is accurate, consistent, and ready for analysis. This step involves:

- **Handling missing values**
  Imputing or removing missing data points to prevent biases in the analysis.

- **Outliers Detection**
  Identifying and addressing anomalies that may skew the results.

- **Standardization and Normalization**
  Ensuring that all data points are in a consistent format and scale, facilitating accurate comparisons and analyses.

---

03. Exploratory Data Analysis (EDA)

EDA is crucial for understanding the main characteristics of the data. During this phase, we:

- **Visualize Data**
  Use charts, graphs, and heat maps to identify trends, patterns, and correlations.

- **Summarize Statistics**
  Calculate key statistics such as mean, median, standard deviation, and distribution to understand data variability and central tendencies.

- **Identify Relationships**
  Analyze relationships between different variables to uncover potential predictors and correlations between player performance attributes.

04. **Feature Engineering**

Feature engineering involves creating new variables that can enhance the predictive power of our models. This will be done using Exploratory Data Analysis and previous insights.

---

05. **Model Development and Evaluation**

In this phase, we build and evaluate machine learning models useful to learn insights like players similarities.

To reach the best results possible, the basic tool I used along this project has been Python (Jupyter Notebooks in particular).

03. Scraping

The primary objective of the web scraping task was to **automate** the **extraction** of up-to-date and comprehensive player data, including advanced performance metrics such as non-penalty goals per 90 minutes and expected-assists per 90 minutes, as well as demographic attributes like club affiliation, league participation, physical characteristics, and financial valuations. This automated approach ensures a scalable and efficient means of gathering large datasets, which are crucial for conducting in-depth analyses and developing predictive models.
The data for this project was sourced from two primary websites: Transfermarkt.com and FBref.com. Transfermarkt.com was selected for its comprehensive demographic and financial data, while FBref.com was chosen for its detailed player statistics.

01. Importing necessary libraries

At first I imported the following libraries, essential to conduct a proper analysis:

- **pandas**
  I used pandas to manage data easily.

- **selenium**
  I imported selenium to automate web browsing and interactions with websites. In particular, I used *webdriver* class and its functions.

- **time**
  I used time to add delays between webpage loadings and information retrievals.

---

02. Defining Driver and Scraping functions

I began the actual scraping process by creating a Firefox WebDriver using Selenium's built-in function. Then, I developed the following essential functions to retrieve data from the sources:

- **load_link_dicts**
  Loads dictionaries with league names and corresponding URLs from text files.

- **click_cookie_fbref**
  Finds and clicks the cookie acceptance button on FbRef if present.

- **go_to_tm**
  Redirects WebDriver to the Transfermarkt URL given and waits the page to be fully loaded.

- **get_teams_links**
  Extracts links to Transfermarkt team webpage from league page.

- **load_tm_teams_links_dict**
  Fills a dictionary with team links from Transfermarkt.

- **get_players_data**
  Extract player data from Transfermarkt.

- **go_to_fbr**
  Navigate to a FbRef URL.

- **get_fbr_teams_links**
  Extract FbRef team webpage links from FbRef league webpage.

- **load_fbr_teams_links_dict**
  Fills a dictionary with team links from FbRef.

- **get_fbr_players_links**
  Get FbRef player links from FbRef team webpage.

- **load_fbr_players_links**
  Load player links from FbRef.

- **get_fbr_players_info**
  Extract player statistics and informations from FbRef. player webpage link.

To obtain the correct information, I used Selenium's WebDriver functions find_element and find_elements. By calling these functions with a CSS or XPATH selector, it is possible to retrieve HTML elements from the webpage and, using the text attribute, access their displayed values.

03. **Main Loop for Data Extraction**

The actual scraping process starts with the initialisation of the data structures. This is followed by a main loop that scrapes data for each league. This initialisation involves creating dictionaries to store URLs for FbRef and Transfermarkt, along with the number of teams per league. During the main loop, these dictionaries are populated and used to navigate each league's website, extract team and player data, and consolidate the information into dataframes for analysis. The dataframes are then saved as CSV files to be used in the following steps of the analysis.

04. Cleaning

This chapter builds on the automated web scraping task, which provided a rich dataset from Transfermarkt.com and FBref.com. It focuses on the critical process of **data cleaning** and **preparation**. The raw data extracted includes extensive player statistics, demographic details, and financial valuations. We undertook a thorough cleaning process to ensure the data's usability for analysis and predictive modelling. This entailed importing the raw datasets, extracting relevant league names, and merging data from the two sources to create a unified and comprehensive dataset. These steps were taken to standardise the data, address any inconsistencies or missing values, and prepare it for robust statistical analysis and model development in the subsequent chapters.

01. Importing libraries and Reading Data

To begin the data cleaning process, I first import the following necessary Python libraries:

- **pandas**
  Useful for data manipulation.

- **numpy**
  Useful for numerical operations.

- **dateutil**
  Useful to handle date-related data and operations

Next, I imported the data files from Transfermarkt.com and FBref.com, which were stored in CSV format, and loaded them into pandas DataFrames for easier handling.

---

02. Datasets Description

Each CSV file from FBref contains data pertaining to a specific league and adheres to a uniform structure. This structure is also observed in the CSV files from Transfermarkt. The following section presents the structures of the respective DataFrames that have been scraped.

Transfermarkt Files

Feature	Description	Type
Player	Name and Surname of the player	string
Role	Role of the player on the field	string
Team	Team to which the player belongs	string
Birth	Birthdate of the player (Mon DD, YYYY)	string
Value	Market value of the player (in €)	string

FbRef Files - Demographical/Contractual Data

Feature	Description	Type
Player	Name and Surname of the player	string
Role	Role of the player on the field	string
Position	Specific positional information	string
Foot	Preferred foot of the player	string
Height	Height of the player (in cm)	string
Weight	Weight of the player (in kg)	string
Birth	Birthdate of the player (Month DD, YYYY)	string
Nationality	Player's nationality	string
Club	Club to which the player belongs	string
Wage	Player's weekly wage (in €)	string
Expiration	Expiration year of player's contract	string

FbRef Files - Goalkeepers Performance Data

Feature	Description	Type
PSxG-GA	Post-Shot expected Goals minus Goals Allowed	float
Goals Against	Total number of goals conceded	float
Save Percentage	Team to which the player belongs	float
PSxG/SoT	Post-shot expected goals per shot on target	float
Save% (Penalty Kicks)	Percentage of penalty kicks saved	float
Clean Sheet Percentage	Percentage of matches with a clean sheet	float
Touches	Total number of touches	float
Launch%	Percentage of goal kicks that were launched	float
Goal Kicks	Total number of goal kicks	float
Avg. Length of Goal Kicks	Average length of goal kicks	float
Crosses Stopped%	Percentage of crosses stopped	float
Def. Actions Outside Pen. Area	Defensive actions outside the penalty area	float
Avg. Distance of Def. Actions	Average distance of defensive actions	float

FbRef Files - Movement Players Performance Data

Feature	Description	Type
Non-PenaltyGoals	Number of goals scored excluding penalties	float
npxG:Non-PenaltyxG	Expected goals excluding penalties	float
ShotsTotal	Total number of shots taken	float
Assists	Number of assists	float
xAG:Exp.AssistedGoals	Expected assisted goals	float
npxG+xAG	Combined expected non-penalty goals and expected assisted goals	float
Shot-CreatingActions	Number of actions that lead to a shot	float
PassesAttempted	Total number of passes attempted	float
PassCompletion%	Percentage of successful passes	float
ProgressivePasses	Number of passes that move the ball significantly forward	float
ProgressiveCarries	Number of carries that move the ball significantly forward	float
SuccessfulTake-Ons	Number of successful dribbles or take-ons	float
Touches(AttPen)	Number of touches in the attacking penalty area	float
ProgressivePassesRec	Number of progressive passes received	float

Tackles	Number of tackles made	float
Interceptions	Number of interceptions	float
Blocks	Number of blocks made	float
Clearances	Number of clearances	float
AerialsWon	Number of aerial duels won	float

The datasets, which comprise the aforementioned structures for each league, are then read and subsequently merged into a single Transfermarkt and FBref dataset.

---

03. **Basic FbRef Dataset Cleaning**

I remove eventual dataset duplicates, which can be reached with the Player attribute. Then I remove redundant and inaccurate attributes in FbRef datasets (*Position* and *Role*), as they are already present and more accurate in Transfermarkt.

At this point I analyse the needs of each attribute in FbRef distinctly.

- ***Height***
  As *Height* column contain data formatted as strings (e.g. 180cm), I remove 'cm' letters and convert the remaining string (the correct value) to float.

- ***Weight***
  Weight, instead, wasn't available for each player, but only for

some. The ones not following the correct format (e.g. '66kg') are considered missing values and, so, are filled with None. 'kg' substring is removed from correct ones and each value is converted to float type.

- ***Club***

  In order to maintain consistency with Transfermarkt dataset, some club names are mapped to match other dataset values (e.g. 'Brighton' instead of 'Brighton & Hove Albion'). Moreover, FbRef includes also some players on loan, so I detect them and place that values to None.

- ***Nationality***

  Nationality column, similarly, has partial informations for some nationalities (e.g. 'Costa' is scraped instead of 'Costa Rica'). So, in the same way, once detected the inaccuracy, I use a map to match the imprecise informations to their correct values.

- ***Wage***

  Wage given on FbRef is weekly and uses ',' as thousand sign. I choose to transform the wage to yearly and then, removing the commas, I can convert the data to float (keeping only two decimal figures).

- ***Expiration***

  Regarding *Expiration* I just clean eventual scraping errors, which lead some data to have a point at the end of their value. so I apply a lambda function to remove an eventual point.

I look at all the columns in the whole dataset and some are useless because some players' webpages don't have the right statistics. I remove them.

04. **Merging Datasets and Additional Cleaning**

I merge the Transfermarkt and Fbref datasets using the Player column. The new merged data frame has some duplicate columns, like Role and *Birth*. I renamed the columns (e.g. *Role.x* and *Role.y*) so I can remove whichever one I want and rename it again.

Then I decide to add a derived column, *Age*, which can be obtained with 'relativedelta' function from 'DateUtil' package. With this method I can perform a subtraction between actual timestamp and player birth date, in order to get updateed age.

Finally I treated column *Value,* indicating player market value. The problem here is the presence of both players valuing milions of euros and players valuing thousands of euros. Those two types of values are formatted, as understandable, in different ways (e.g. '10.00m' and '750k'). Here I need to use a lambda function to be applied to the whole dataset: in the first case I just remove the letter 'm', in the second one I remove the letter 'k' and I divide the value to get the same measure unit. Then I convert everything to float.

---

05. **Dividing Datasets**

Now I divide the merged DataFrame in two:

- ***Goalkeepers***
  This DataFrame contains all players in merged DataFrame whose role is 'Goalkeeper'. In this way it will contain all goalkeepers demographical/contractual data and performance statistics.

- ***Movement Players***
  This contains all players in merged DataFrame whose role is not 'Goalkeeper' (so all Defenders, Midfielders and Forwards). As the previous one, it will contain all movement players data, both demographical/contractual and performance.

To have both DataFrames with correct columns, I delete goalkeepers performance attributes from movement players DataFrame and vice versa.

---

06. Handling Goalkeepers DataFrame

I start the cleaning of performance columns in goalkeepers DataFrame by removing rows having *GoalsAgainst* as null. I do this because those rows have null values across all other performance attributes, meaning that they didn't play enough games to provide data.

Then, investigating null values presence, I discover some null values in the attribute *Save%(PenaltyKicks)*, which can be possible as penalty kicks occur quite rarely. I inferred null values with the median which, in this case, is 0.

---

07. Handling Movement Players DataFrame

I apply the same first step to movement players DataFrame, removing rows having null performance data. All movement players performance columns depend on player's contribution on the pitch and not on events happening only sometimes (e.g. Penalty Kicks

Save Percentage), so it's not necessary to infer any value in any column.

In this DataFrame in *Role* column it's actually present the *Position* data for each player (e.g. 'Centre-Back', 'Left-Winger') and not the *Role* one ('Defender', 'Midfielder', 'Forward'): so, I rename the *Role* column to *Position* and, for each row, I use a map to get the *Role* information and store it in the proper column.

---

08. Integrating EA FC 24 Dataset

To investigate how well sports video games reproduce and analyze players characteristics, I decide then to integrate a different data source, EA FC 24 players statistics dataset. I chose the dataset at this link, because of its exhaustiveness. I start reading it and renaming *Position* column to *Fifa_Position* in order to differ from the actual player position. Later I remove the following attributes from EA FC 24 players dataset, due to its uselessness or their redundancies with other DataFrame:

- ***Unnamed: 0*** (DataFrame index)
- ***Nation***
- ***Club***
- ***Age***
- ***URL***
- ***GK***
- ***Gender***

Now I merge the EA FC 24 DataFrame firstly with movement players DataFrame and then with goalkeepers DataFrame, both using, as keys, columns *Name* from EA FC 24 DF and *Player* from the other ones. At this point I can end the cleaning process by saving final DataFrames to CSV in local, in order to be used later in the analysis step.

---

09. Final DataFrame Structure

As previously said, I saved 4 final and cleaned DataFrames, which will be listed below with their columns.

- ***goalkeepers_df.csv***

Feature	Description	Type
Player	Name and Surname of the player	string
Foot	Preferred foot of the player	string
Height	Height of the player (in cm)	float
Weight	Weight of the player (in kg)	float
Nationality	Player's nationality	string
Club	Club to which the player belongs	string
Wage	Player's yearly wage (in M of €)	float

Expiration	Expiration year of player's contract	int
PSxG-GA	Post-Shot expected Goals minus Goals Allowed	float
GoalsAgainst	Total number of goals conceded	float
SavePercentage	Team to which the player belongs	float
PSxG/SoT	Post-shot expected goals per shot on target	float
Save%(PenaltyKicks)	Percentage of penalty kicks saved	float
CleanSheet Percentage	Percentage of matches with a clean sheet	float
Touches	Total number of touches	float
Launch%	Percentage of goal kicks that were launched	float
GoalKicks	Total number of goal kicks	float
Avg.Lengthof GoalKicks	Average length of goal kicks	float
CrossesStopped%	Percentage of crosses stopped	float
Def.ActionsOutside Pen.Area	Defensive actions outside the penalty area	float
Avg.Distanceof Def.Actions	Average distance of defensive actions	float
Role	Role of the player on the field	string
Birth	Birthdate of the player (YYYY-MM-DD)	float
Value	Market value of the player (in M of €)	float
League	Domestic league where the player plays	string
Age	Actual age of the player	int

- ***players_df.csv***

Feature	Description	Type
Player	Name and Surname of the player	string
Foot	Preferred foot of the player	string
Height	Height of the player (in cm)	float
Weight	Weight of the player (in kg)	float
Nationality	Player's nationality	string
Club	Club to which the player belongs	string
Wage	Player's yearly wage (in M of €)	float
Expiration	Player's yearly wage (in M of €)	int
Non-PenaltyGoals	Number of goals scored excluding penalties	float
npxG:Non-PenaltyxG	Expected goals excluding penalties	float
ShotsTotal	Total number of shots taken	float
Assists	Number of assists	float
xAG:Exp.AssistedGoals	Expected assisted goals	float
npxG+xAG	Combined expected non-penalty goals and expected assisted goals	float
Shot-CreatingActions	Number of actions that lead to a shot	float

PassesAttempted	Total number of passes attempted	float
PassCompletion%	Percentage of successful passes	float
ProgressivePasses	Number of passes that move the ball significantly forward	float
ProgressiveCarries	Number of carries that move the ball significantly forward	float
SuccessfulTake-Ons	Number of successful dribbles or take-ons	float
Touches(AttPen)	Number of touches in the attacking penalty area	float
ProgressivePassesRec	Number of progressive passes received	float
Tackles	Number of tackles made	float
Interceptions	Number of interceptions	float
Blocks	Number of blocks made	float
Clearances	Number of clearances	float
AerialsWon	Number of aerial duels won	float
Position	Specific field position where the player plays	string
Birth	Birthdate of the player (YYYY-MM-DD)	string
Value	Market value of the player (in M of €)	float
League	Domestic league where the player plays	string
Age	Actual age of the player	int
Role	Role of the player on the field	string

- ***goalkeepers_fifa_df.csv***
- ***players_fifa_df.csv***

Both DataFrames start from the previously listed ones and add EA FC 24 statistics shown below.

Feature	Description	Type
FIFA_Position	Position assigned to the player on EA FC video game	string
Overall	Overall player value	float
Pace	EA FC 24 Pace attribute value	float
Shooting	EA FC 24 Shooting attribute value	float
Passing	EA FC 24 Passing attribute value	float
Dribbling	EA FC 24 Dribbling attribute value	float
Defending	EA FC 24 Defending attribute value	float
Physicality	EA FC 24 Physicality attribute value	float
Acceleration	EA FC 24 Acceleration attribute value	float
Sprint	EA FC 24 Sprint attribute value	float
Positioning	EA FC 24 Positioning attribute value	float
Finishing	EA FC 24 Finishing attribute value	float

Shot	EA FC 24 Shot attribute value	float
Long	EA FC 24 Long attribute value	float
Volleys	EA FC 24 Volleys attribute value	float
Penalties	EA FC 24 Penalties attribute value	float
Vision	EA FC 24 Vision attribute value	float
Crossing	EA FC 24 Crossing attribute value	float
Free	EA FC 24 Free attribute value	float
Curve	EA FC 24 Curve attribute value	float
Agility	EA FC 24 Agility attribute value	float
Balance	EA FC 24 Balance attribute value	float
Reactions	EA FC 24 Reactions attribute value	float
Ball	EA FC 24 Ball attribute value	float
Composure	EA FC 24 Composure attribute value	float
Interceptions	EA FC 24 Interceptions attribute value	float
Heading	EA FC 24 Heading attribute value	float
Def	EA FC 24 Defence attribute value	float
Standing	EA FC 24 Standing attribute value	float
Sliding	EA FC 24 Sliding attribute value	float

Jumping	EA FC 24 Jumping attribute value	float
Stamina	EA FC 24 Stamina attribute value	float
Strength	EA FC 24 Strength attribute value	float
Aggression	EA FC 24 Aggression attribute value	float
Att work rate	EA FC 24 Offensive work rate attribute value	string
Def work rate	EA FC 24 Defensive work rate attribute value	string
Preferred foot	Player preferred foot	string
Weak foot	Number of EA FC 24 player's weak foot stars (1 to 5)	float
Skill moves	Number of EA FC 24 player's skill moves stars (1 to 5)	float

05. Analysis

In this chapter, I focus on the **data analysis** phase, where we apply a range of analytical techniques to derive meaningful **insights** from the cleaned dataset. This phase is crucial for converting raw data into useful information, which helps in making informed decisions and addressing our research objectives. I aim to identify patterns, relationships, and trends within the data. These insights will provide a deeper understanding of the dataset and support strategic decision-making.

01. Importing libraries

To start the data analysis process, I import the following Python libraries:

- **pandas**
  Useful for data manipulation.
- **seaborn**
  Useful for data visualization
- **matplotlib**
  Useful for data visualization (alternative to seaborn)

- **sklearn**
  Useful for machine learning models and data pre-processing.

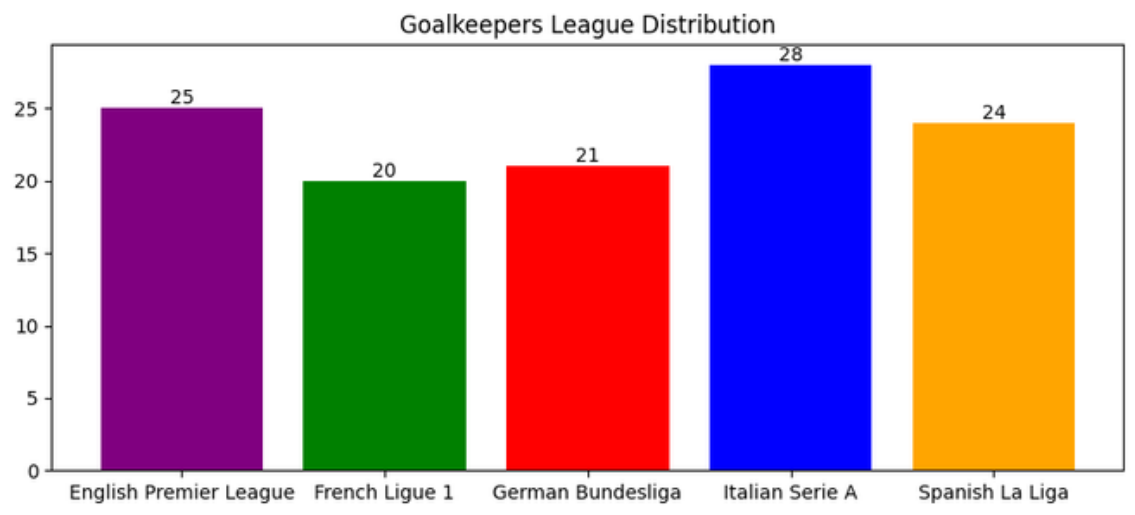
---

02. Distribution Analysis

I begin importing and reading the four DataFrames, *goalkeepers_df.csv, goalkeepers_fifa_df.csv, players_df.csv* and *players_fifa_df.csv.*
Then I start my analysis process visualizing the distribution of the different attributes along their ranges, in order to have a brief look at the DataFrames we're dealing with.

League

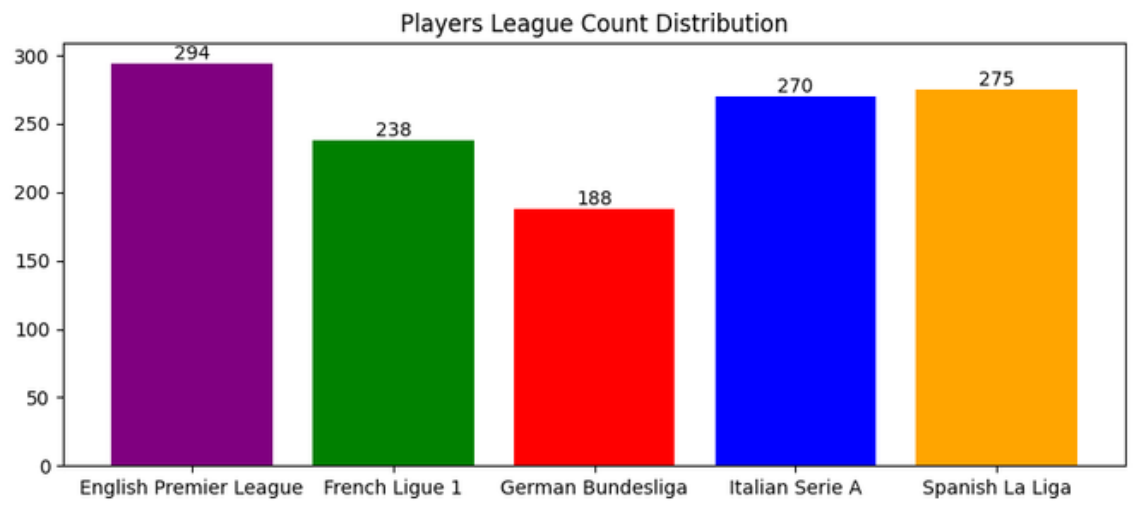
Goalkeepers



Goalkeepers League distribution shows a balance in the DataFrame. As easily guessed, it hosts a little more than one GK per team, so 'Italian, English and Spanish leagues are the most represented ones, due to the higher number of teams (20 against 18). Each team has, on average, **1.2 goalkeepers** in the dataset, which can sound weird as only one keeper can stay on the pitch simultaneously. This can occur primarily for two reasons: cohexistence of two equally

considered players (e.g. Carnesecchi-Musso in Atalanta) or injury to the starting goalkeeper, letting the substitute collect some appearances (e.g. Kepa-Lunin alternance during Courtois rehab).

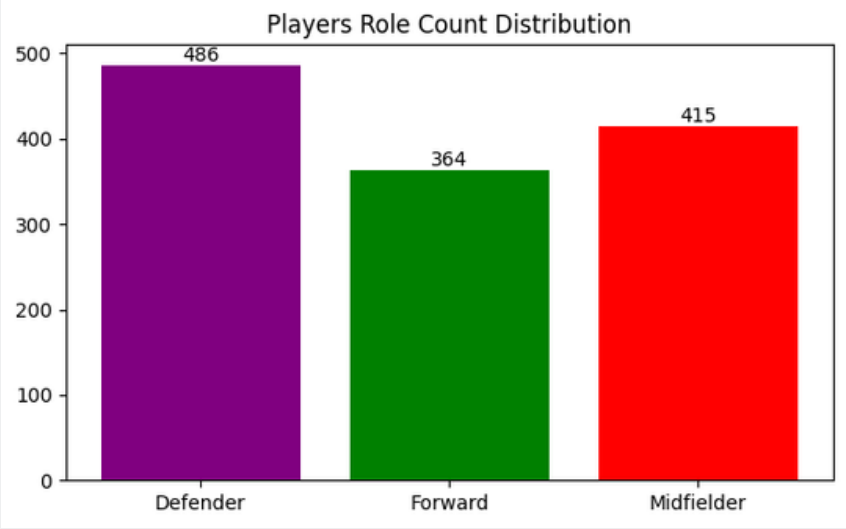
Movement Players



League distribution among movement players looks similar to the previous one, with the difference of Premier League being the most represented one, instead of Serie A. On average, there are **13.2 movement players** per team, with a peak of 14.7 in English domestic league and a low of 10.4 in German Bundesliga, making evident some problems in the scraping process or in the website.

Role

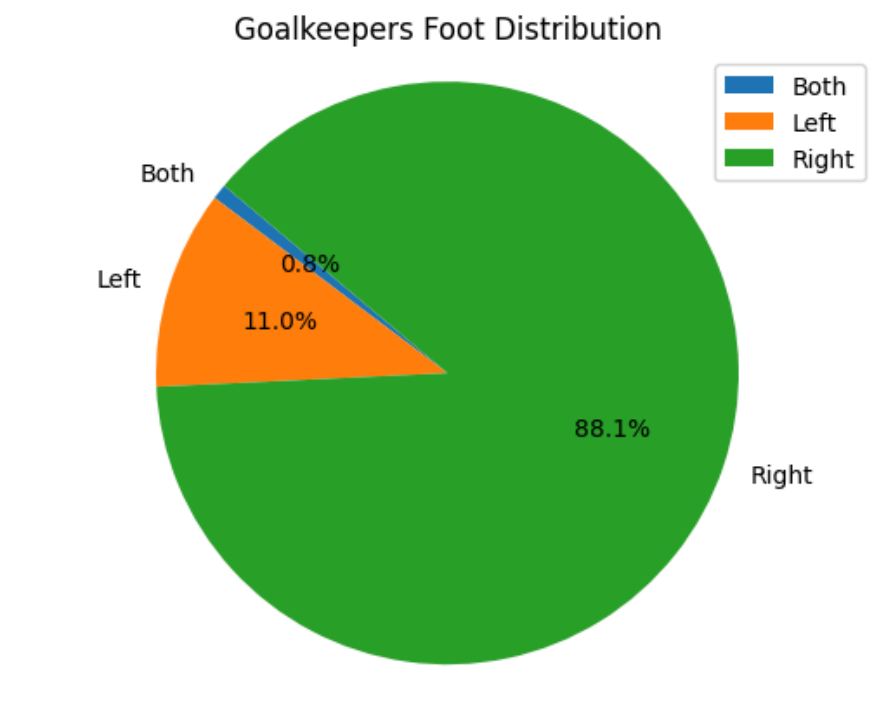
Movement Players



This graph underlines the usual role distribution in many teams, having the **38.4%** of **defenders**, **32.8%** of **midfielders** and only **28.7%** of **forwards**. This is linear with the lineups usually employed by the teams (e.g. 433, 442), having a lower number of forwards, than other roles . The highest number of defenders, then, is also caused by the classification of players, in particular of 'Wing-backs' who, usually, play wide on the midfield line but, in the DataFrame, are considered 'Left-backs' or 'Right-backs', adding up to defenders quota.

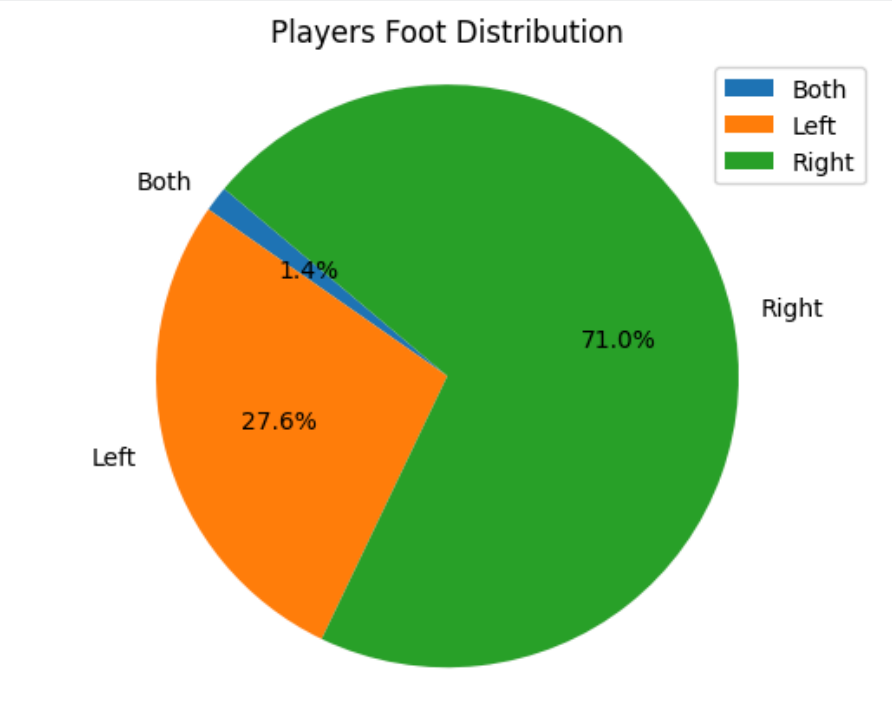
Foot

Goalkeepers



Goalkeepers preferred foot distribution looks similar to handedness population distribution, having approximately **10%** of **left-handed** (here it's 11% of left-footed) and **1%** of **ambidexterity** (here it's 0.8%). This results can show some insights about goalkeepers, but it's mandatory to keep in mind the limited size of the sample, with only 118 data.

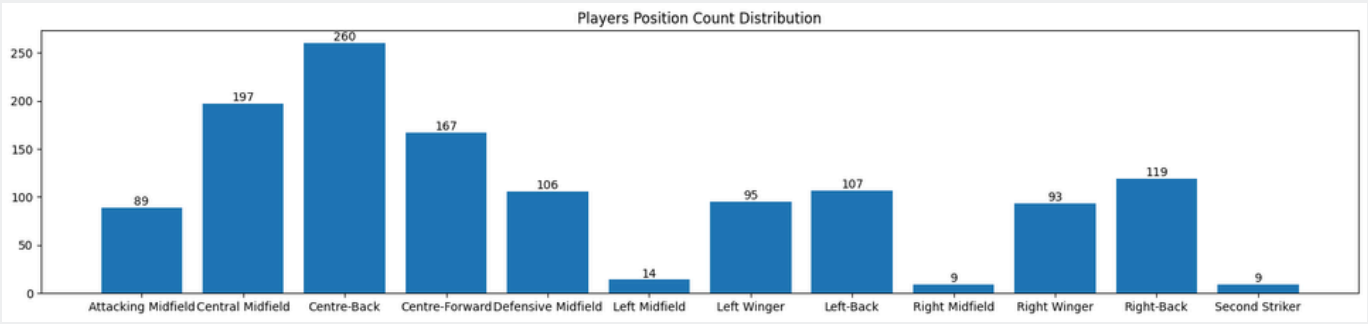
Movement Players



Preferred foot distribution among non-goalkeepers players deviates consistently from both population distribution and goalkeepers one. **Left-footed players** percentage grows from about 10% to more than **27%** and **ambidexterious players** grow from less than 1% to **1.4%**. Those data give some important informations in my opinion, telling that, at high level, left-footed players may be more valuable to the teams, due to their diversity. Given those data, some questions of different nature may arise, for example if this high percentage is caused only by their way of playing 'differently', using more majority's weak foot, or by neuro-related reasons (e.g. creativity hemisphere more developed).

Position

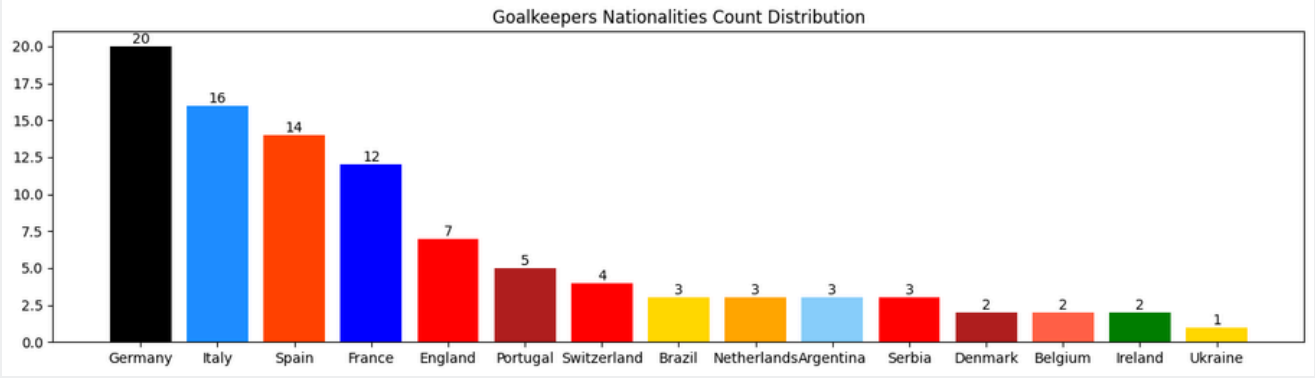
Movement Players



In the previous graph I show the count distribution on movement players *Position* attribute. It seems apparent that the **position more represented** are those that form the **backbone** of the team, the central position in all three field positions: **centre-back, central midfield and centre-forward**. That can happen because of the need of those positions (usually with more than one player) in quite all football lineup. Then, some of the positions listed above are questionable, even more so looking at the players covering them: some of them can be inserted in more than one category, leading to a low count of some positions (e.g. Frosinone Calcio's player Nadir Zortea cover the same lineup position than Torino's Raoul Bellanova, but the first is considered Right Midfielder, while the second Right-Back).

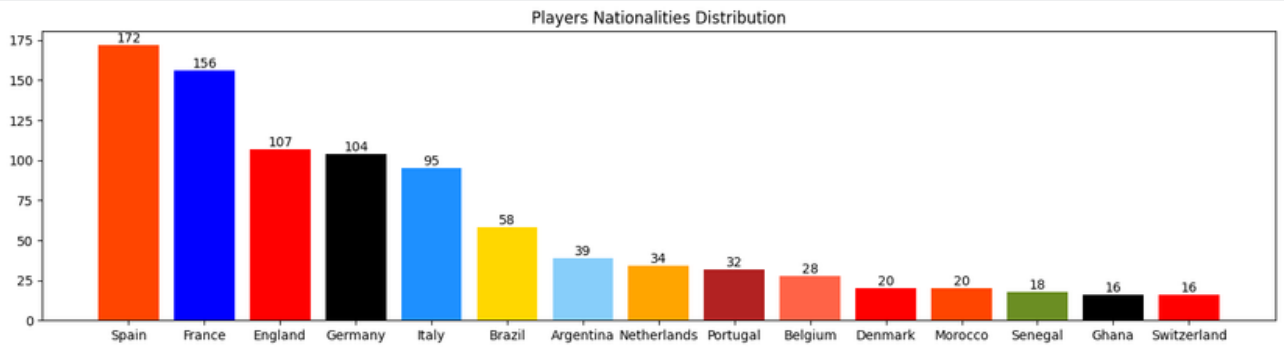
Nationality

Goalkeepers



Goalkeepers nationalities count barchart exhibits, as known, the **importance** of the **German goalkeepers school** came out in the last decade: FC Bayern Munich's Manuel Neuer and FC Barcelona's Marc-André Ter Stegen are two of the best goalkeepers in the world, but are only the tip of the iceberg. Of course the most represented nationalities are those the top 5 European leagues are (contributing to almost the 60% of keepers), but then there are many nations paired just below.

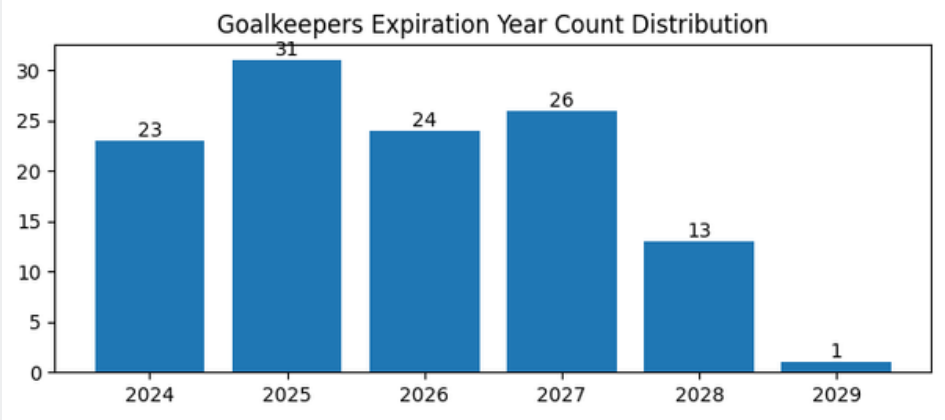
Movement Players



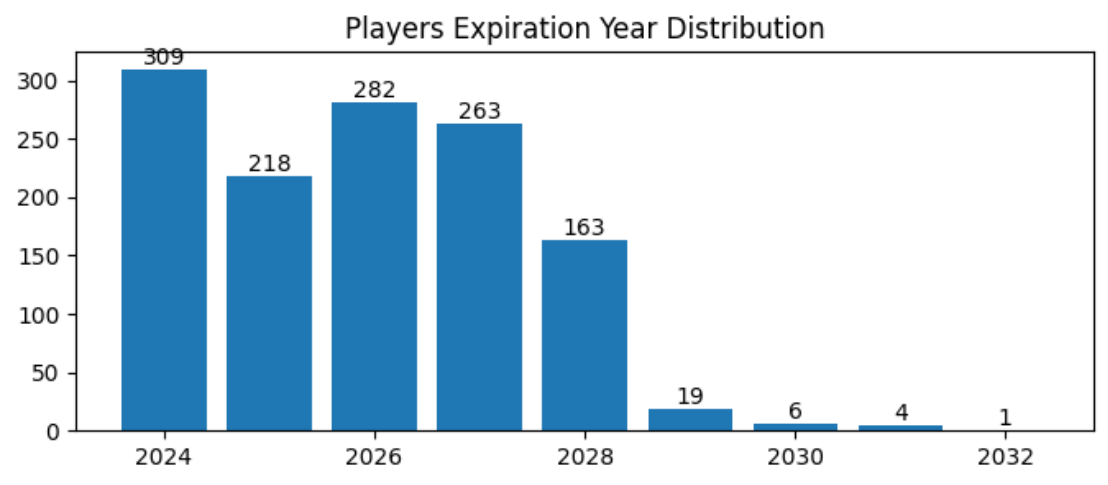
Movement players nationalities distribution has the same **top-5**, as easily imagined, but here, instead of 60%, they sum up to 50%, which can be seen as quite low. The **most represented nationalities** outside the top-5 are **Brazil** and **Argentina**, witnessing their huge importance in the football world. Then I can see that also African powers like Senegal, Morocco and Ghana rely heavily on top leagues and let Europe develop and grow their players, reason why African national teams perform better and better in world cups.

Nationality

Goalkeepers



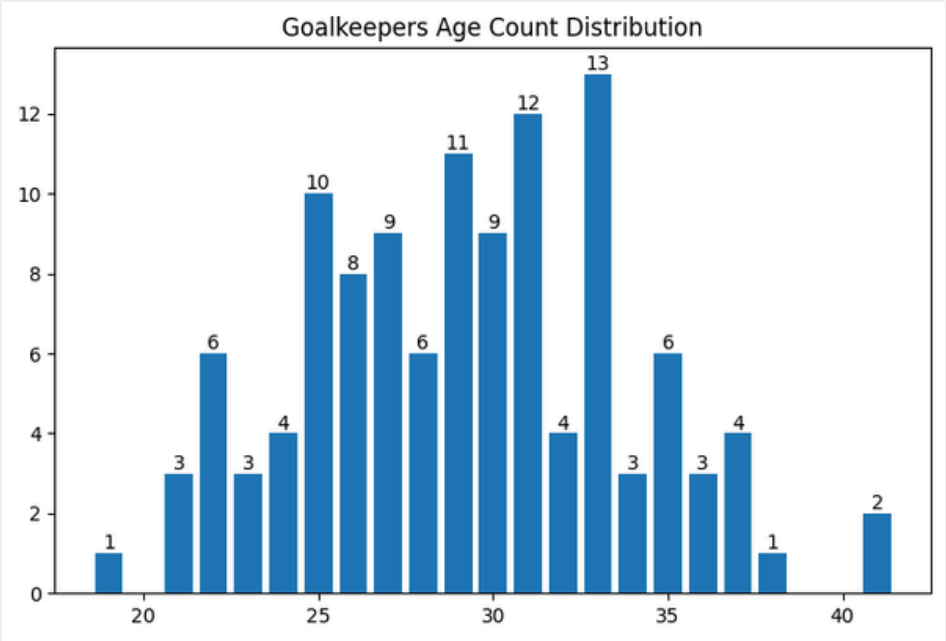
Movement Players



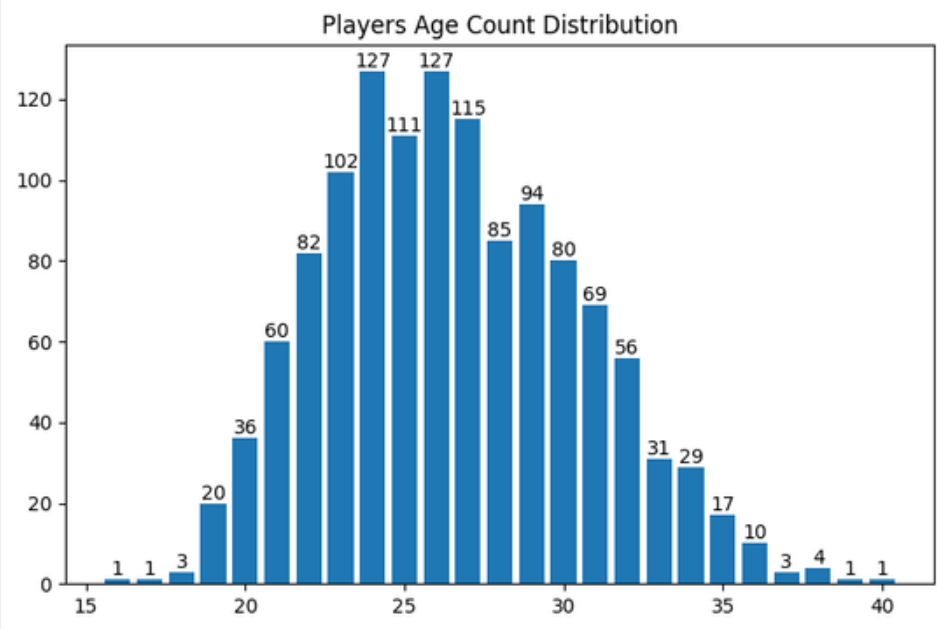
Contract **expiration year** distribution looks **nearly linear** in goalkeepers, while in movement players graph appears evident a **peak** in **2024**. This peak is not that higher than the rest, so I suppose it does not represent any relevant insight.

Age

Goalkeepers



Movement Players



Looking at the two previous barcharts, appear manifest the difference between goalkeepers' and movement players career longevity. While the first one has quite a randomic slope, which becomes constant summing couples of adjacent counts, indicating that usually goalkeepers start their careers at around 20 years old and retire at 38-40, movement players graph shows a curve, starting low earlier (at 16-18 years), then a peak at around 24-26 and later a slow decrease, until 36-37 years. Overall, **goalkeepers age mean** is **29.3 years old**.
Players in positions requiring more technical skills than physical, in fact, tend to have a longer career. Going further into details, the **age mean** per *Role* stays quite the same at **26 years old**: in all three roles central positions (where technique is more important) have higher mean age, but are balanced with lateral ones ('Wingers', 'Lateral Midfielders' and 'Full-Backs'), having a lower mean (e.g. 'Right-Winger' has 25.51, against 'Central Forwards' 27.1)

Height and Weight

Role	**Height** (in cm)	**Weight** (in kg)
Goalkeeper	189.7	81.9
Defender	183.3	75.3
Midfielder	180.4	72.2
Forward	180.8	72.9

As appears in the table above, **goalkeepers** are way **taller** and **heavier** than the rest of the players, due to their advantage in target coverage. For the movement players, this table simplifies too much, so I go deeper getting values for each position.

Position	**Height** (in cm)	**Weight** (in kg)
Centre-Back	187.3	78.6
Left-Back	178.3	71.2
Right-Back	179.3	71.9
Defensive Midfield	183.0	74.1
Central Midfield	180.0	72.2
Right Midfield	181.1	71.3

Left Midfield	179.9	71.0
Attacking Midfield	178.2	70.4
Left Winger	178.4	69.7
Right Winger	177.1	69.6
Second Striker	180.4	73.6
Central Forward	184.3	76.5

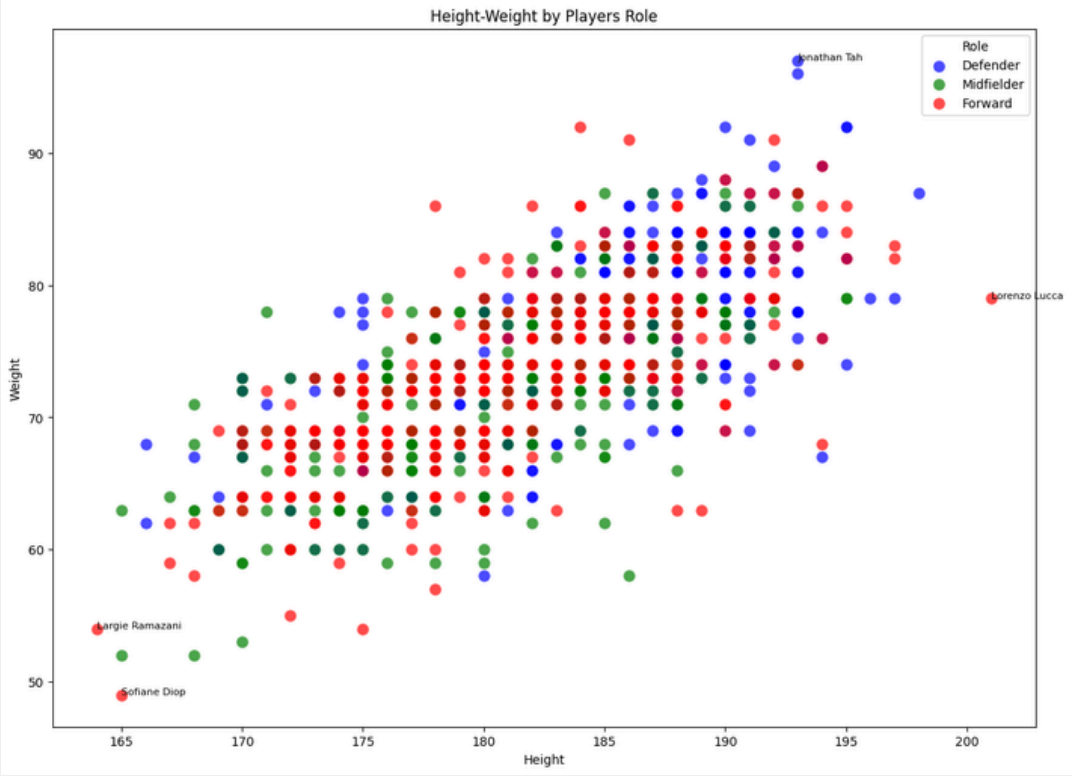
Similarly to what stated before, **height** and **weight** are **polarized** among role different positions For example 'Centre-Backs' are way taller and heavier than 'Left-Backs' and 'Rignt-Backs'. In the same way, 'Defensive Midfielders' compared to 'Lateral Midfielders' and 'Central Forwards' to 'Wingers'. That can be easily combined with the *Age* attribute to detect three kinds of positions:

- **Central positions, requiring height and weight**
  Centre-Backs, Defensive Midfielders and Central Forwards belong to this group. Those are positions where physical skills are required, so longilinear players are preferred. They tend to have a longer career as do not need to have particular agility and speed qualities, which fall down sooner.

- **Balanced positions**
  Positions like Lateral Midfielders or Second Strikers can have various "stereotypes", varying much in physical attributes. On average they tend to follow European typical height (180 cm).

- **Wingers and classical Full-Backs positions**
  |remaining positions, such as wingers, attacking midfielders and some really technical gifted full-backs (e.g. Roberto Carlos, Marcelo, Dani Alves) usually rely on their creativity and

genius with the ball and take advantage of their agility and speed. Often they are brevilinear, light and, as their rapidity decreases fast, they tend to have a shorter higher-level career.



In the chart above it's quite manifest the segmentation I described before: in the **central part** there is the nearly the totality of **midfielders**, while **forwards** and **defenders** compose the **outline** of the shape. Defenders are either tall and heavy or short and light, so do forwards. Then, looking at the scatter plot per role, those differences appear manifest.

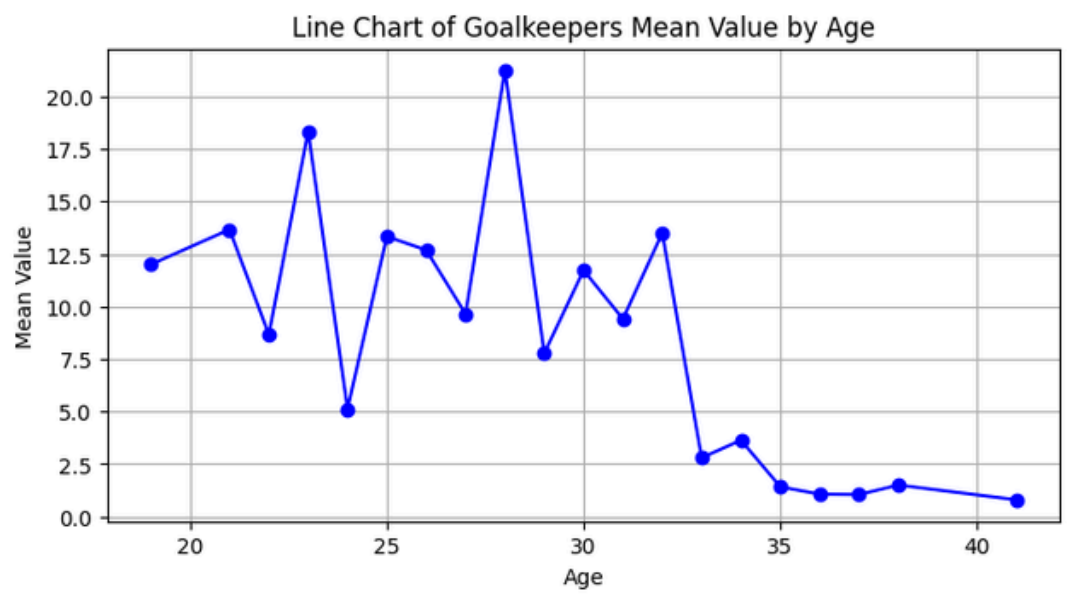
Defenders Midfielders Forwards

03. Correlation Analysis

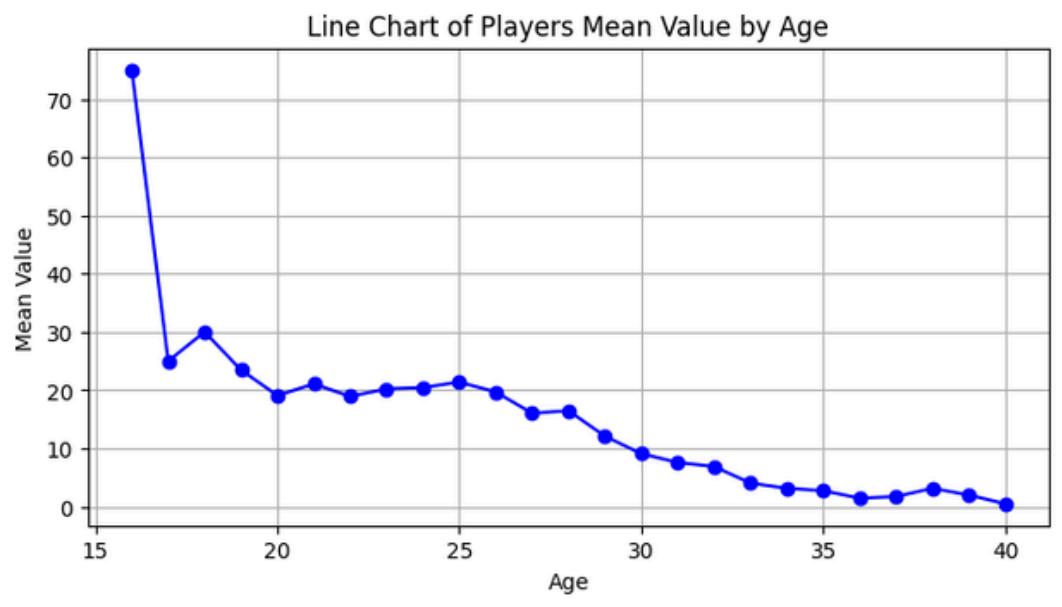
Then I start analyzing combination of attributes to see if they're related and in what manner. This type of analysis can lead to understand how football market works and to identify profitable occasions.

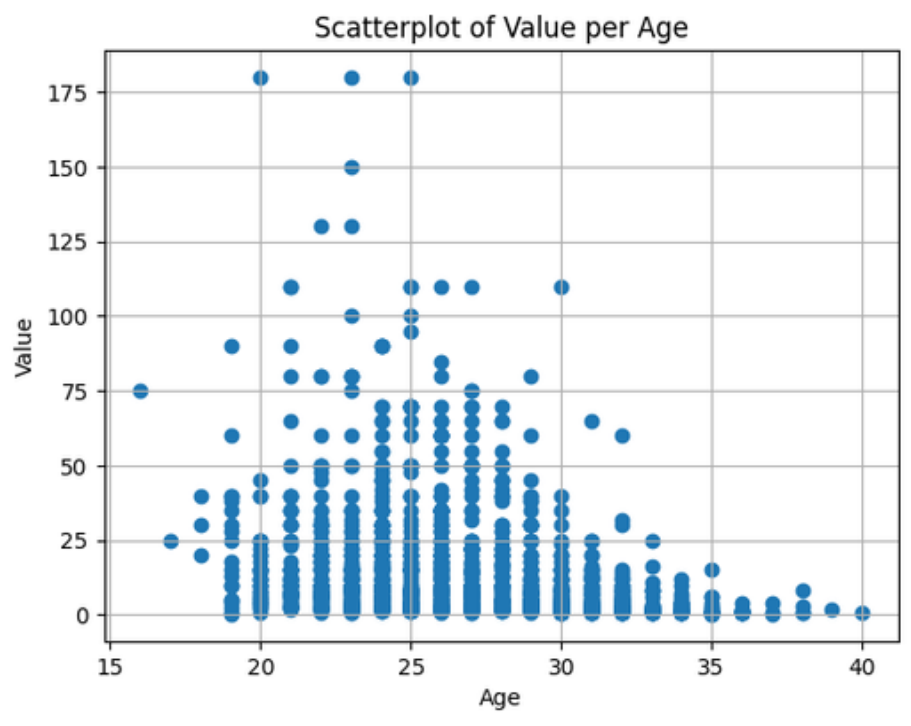
Value - Age

Goalkeepers



Movement Players

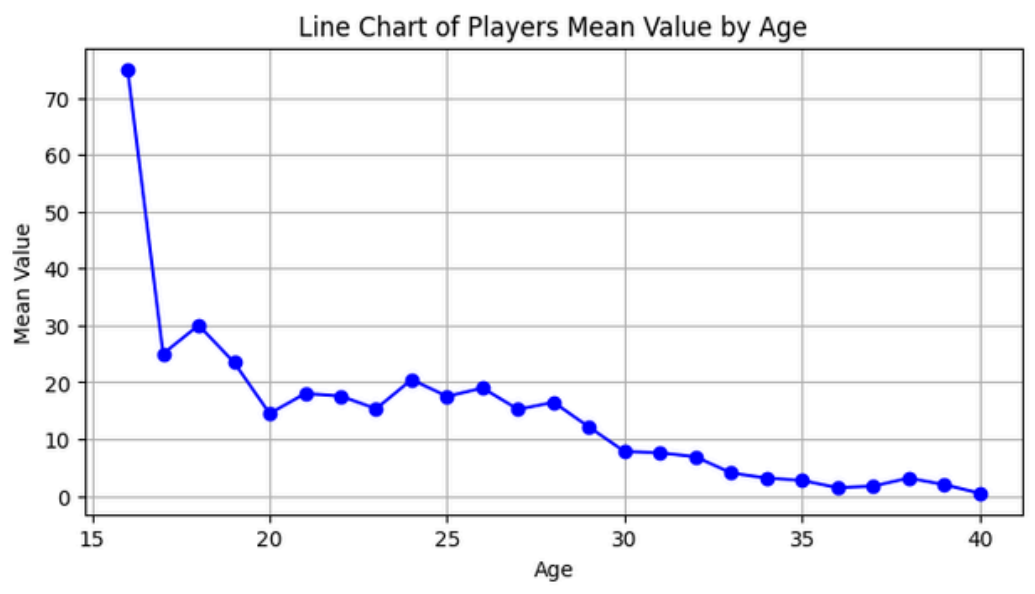




As evident above, there is a huge difference between goalkeepers' and movement players' line charts. **Goalkeepers** value looks swinging much more, due to the limited size of the sample, and doesn't seem to decrease constantly, but stays in the same range from **18** to **33 years old** and then collapses immediately. That confirms for another time the **longevity** of the goalkeepers careers, even from a value perspective.

On the other hand, movement players value looks a lot like a **downward parabola**, where the older the player is, the less valuable he is. The line chart starts with an outlier, because of the presence of only one 16 years old player in the DataFrame, Lamine Yamal, whose value is €75M and surely is an exception more than a norm. Apart from this, we can gain much insights from the graph, starting from the fact that, on average, value peak doesn't coexist with performance peak. Highest mean value is in the range between **18** and **24** years, while sources state that football players performance peak arrives between 25 and 27 years old. Obviously this discrepancy occurs due to the **projections** and **hopes** of growth a player shows, but, as stated before, the best performance doesn't always guarantee the highest value, at least in football.

Then, taking a look at the scatter plot just below, appears manifest that downward parabola trend can't be really taken for granted, because mean value, mostly for ages between 16 and 25, is affected a lot by outliers, who are top players in the world. Removing the 13 players valuing more than €100M, which I list below, the line chart becomes the following. We can state that younger players obviously tend to be overvalued (compared to olders) because of their future expectations and their longevity, but their value doesn't always respect the difference with older players' one in terms of performance, but maybe also in other terms (marketing, status, etc...)



Most Valuable Players (over €100M)

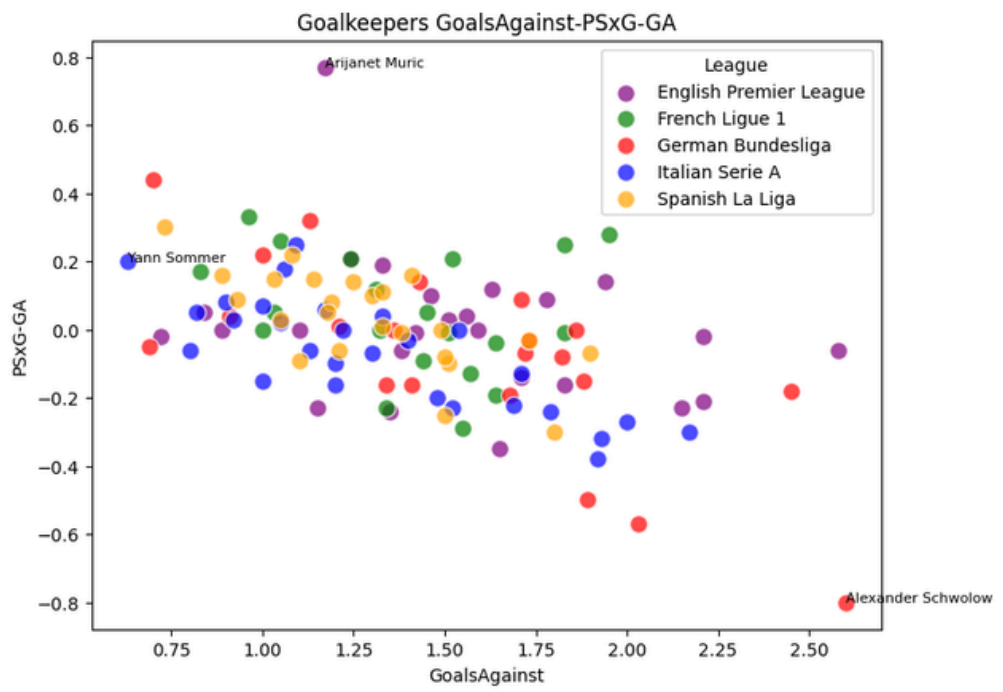
Player	Age	Value (in €M)
Jude Bellingham	20	180
Erling Haaland	23	180
Kilian Mbappé	25	180
Vinicius Junior	23	150
Bukayo Saka	22	130
Phil Foden	23	130

Jamal Musiala	21	110
Florian Wirtz	21	110
Declan Rice	25	110
Victor Osimhen	25	110
Lautaro Martinez	26	110
Rodri	27	110
Harry Kane	30	110

Goalkeepers

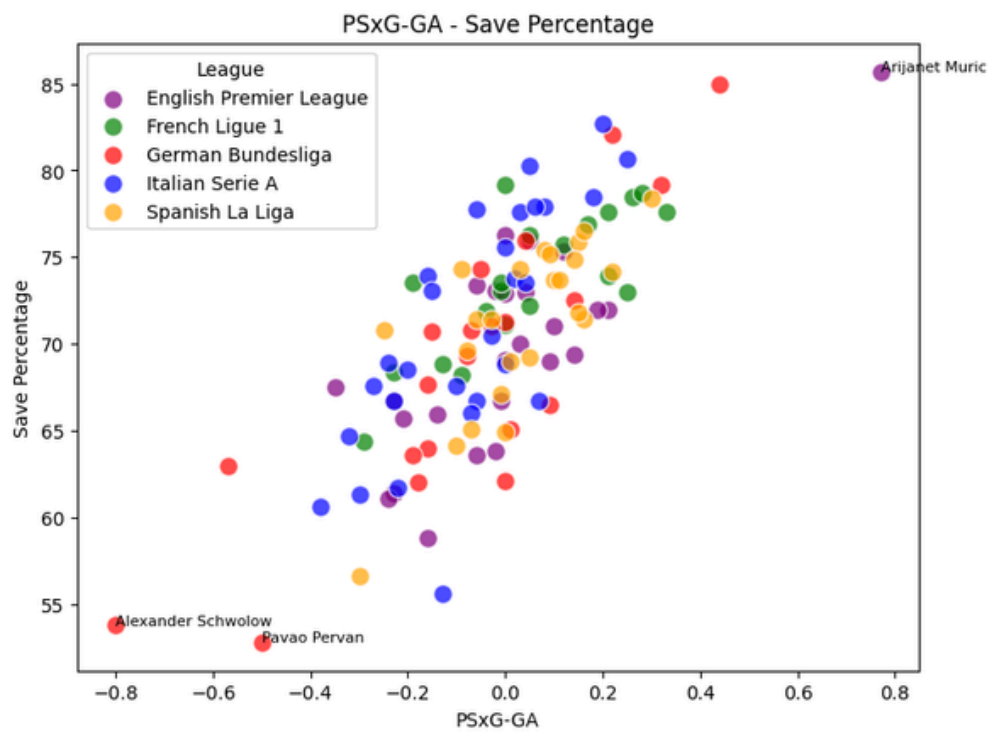
Now I will dive deeper in the analysis of correlation between goalkeepers advanced statistic measures. As goalkeepers cover the same role and position, points in scatter plots will have different colors based on leagues keepers play in.

GoalsAgainst - PSxG-GA



First couple of attributes are *GoalsAgainst* , which represents the number of goals conceded by the goalkeeper per 90', and *PSxG-GA*, acronym of 'Post Shot Expected Goals minus Goals Allowed', which is expected goals, based on how likely the goalkeeper is to save the shot, minus the value explained before, all per 90'. High *PSxG-GA* values suggest better luck or above average ability to save shots. As visible in the graph below, those two attributes tend to have a nearly inverse correlation, having a correlation coefficient of **-0.53**. That usually happens because the most skilled goalkeepers, of course, play in the best teams that, usually, have also the best defenders and, as a consequence, concede less goals. With the league highlighting, becomes evident the **difficulty** to **score** in **Serie A**, opposed to the **better ease** in **Premier League**, as blue dots tend to form a line, lower than the one formed by purple dots. An important outlier can be identified in **Arijanet Muric**, Burnley's goalkeeper, who looks really good (or lucky) at saving shots, having 0.77 of PSxG-GA.

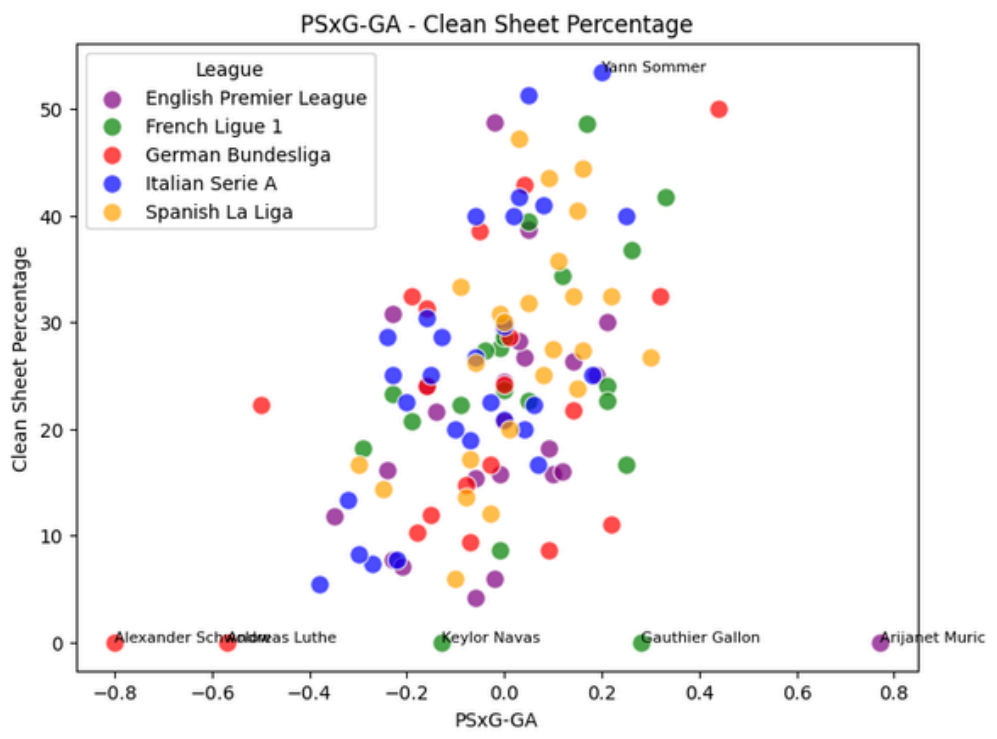
PSxG-GA - SavePercentage



I carry on my analysis combining *PSxG-GA* again, as it should provide informations about goalkeeper effectiveness, and *SavePercentage*, which is a basic information about this role. As manifest, scatter plot looks like following a very steep line. A goalkeeper who has good

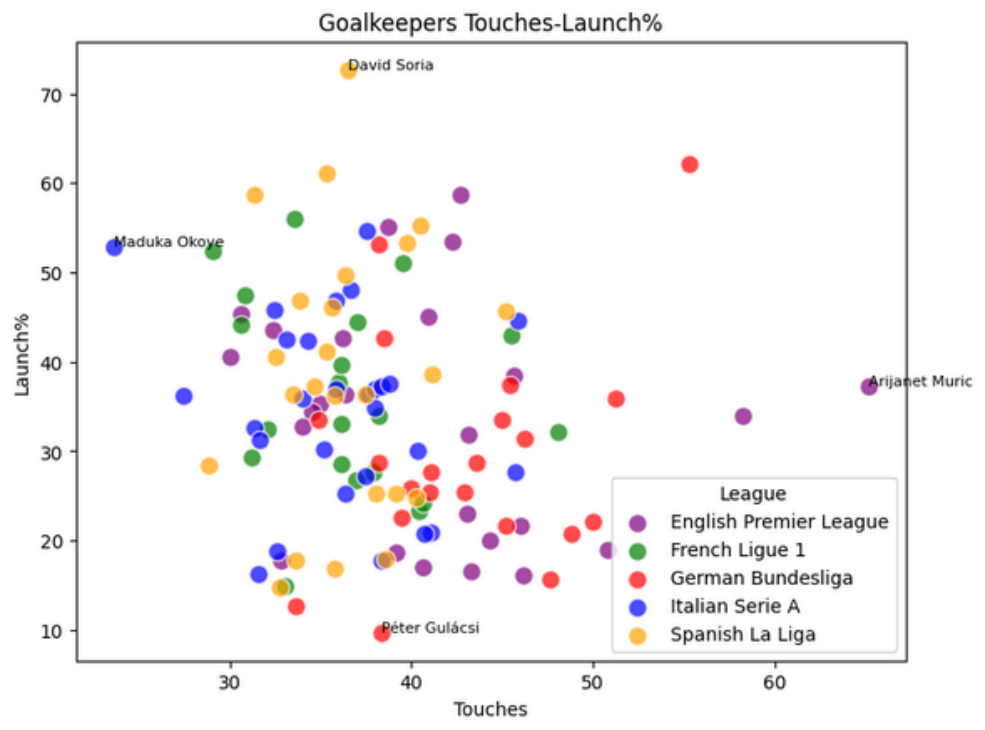
saving skills stops a higher percentage of shots: that looks obvious. Here the outlier is **Arijanet Muric** too, being not only an outlier in *PSxG-GA,* as saw earlier, but also in SavePercentage, having more than **85%** of saves.

PSxG-GA - CleanSheetPercentage



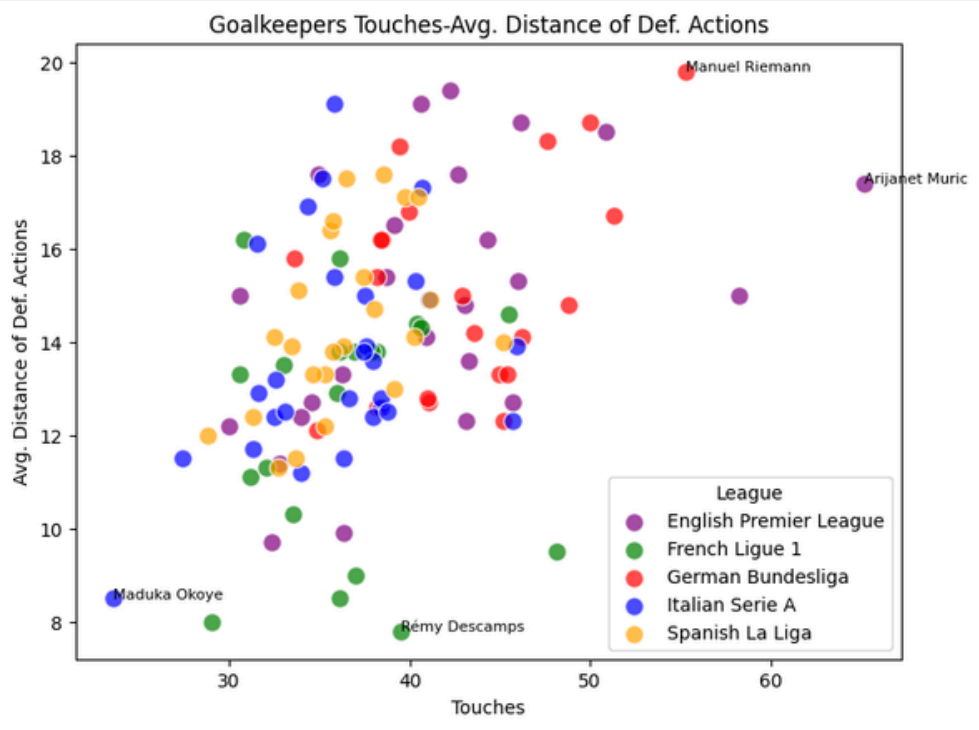
The graph above shows the scatter plot based on *PSxG-GA* attribute on the x-axis and the *CleanSheet%* on the y-axis. The huge cloud in the center indicates a very **weak positive correlation**: there is surely not a goalkeeper with low *PSxG-GA* and high *CleanSheet%*, except Wolfsburg's Pavao Pervan, having really bad *PSxG-GA* and average *CleanSheet%*, a usual combination for backup goalies. There are, indeed, some players having low *CleanSheet%* and high *PSxG-GA.* like, Arijanet Muric. Best keepers resulting from the plot are Inter Milan's **Yann Sommer** and Leverkusen's **Matej Kovar**, having good values in both attributes.

Touches - Launch%



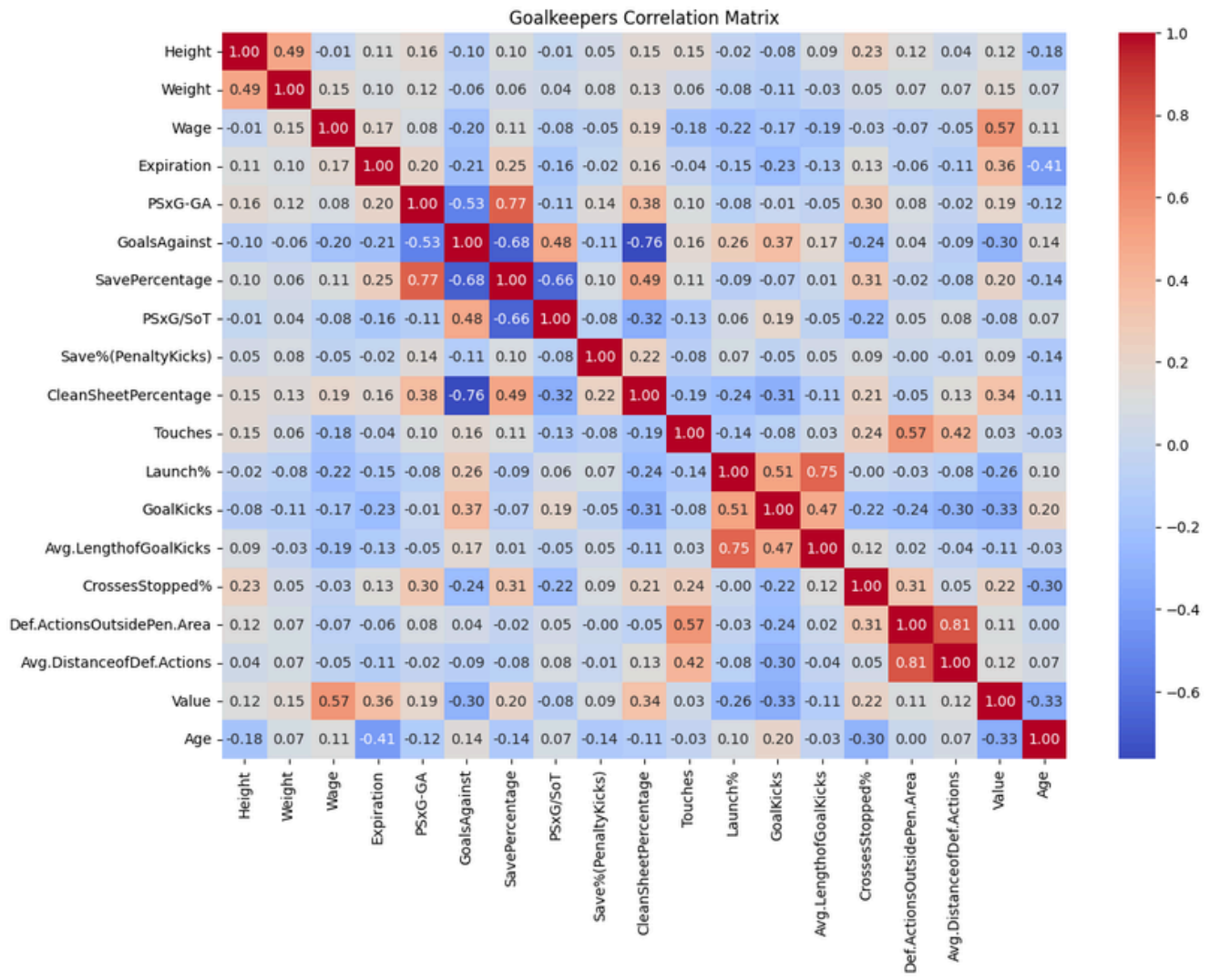
The scatter plot listed above investigates the correlation between *Touches*, the number of times the goalie touches the ball, and *Launch%*, the percentage of passes longer than 40 yards. As evident, there is **no correlation** of any type between the two columns, differently than what was expected. Usually teams using building-up tactics benefit a lot of goalkeepers as active and starting part of their game, but a player like Ederson, starting goalkeeper of Guardiola's Manchester City having 65.4% mean ball possession, has **18.7%** of *Launch%, but* only **39.17** *Touches* per 90'. From the graph Serie A goalies look like touching the ball less than other leagues, while Bundesliga ones being more active with the ball. Udinese's **Maduka Okoye** is the goalkeeper touching the ball the least, opposed to Burnley's **Arijanet Muric** being, by far, the player contributing with more touches.
On the other hand, Getafe's **David Soria** launches the 72.7% of the ball he touches, while RB Leipzig's **Peter Gulacsi** has only 38.4% of *Launches%.*

Touches - Avg. Distance of Def. Actions



The graph above represents a scatter plot in order to inspect correlation between *Touches* and *Average Distance of Defensive Actions*. It shows a weak direct correlation (value of 0.57): as expected, the **more** the goalkeeper tends to **touch** the ball during team build-up phase, the **further** he will **stay** from the post. Players with most extreme stats are the previously cited Okoye and Muric, as well as Bochum's **Manuel Riemann**, goalie playing the furthest on average, and Nantes' backup keeper **Remy Descamps**, who is the one playing the nearest to the goal. Looking at the color clusters, we can state that Ligue 1 goalkeepers are the most 'conservative' (12.3 yards of *Avg.DistanceofDef.Actions*), while Bundesliga and Premier League have the most 'aggressive' ones, playing respectively 15.2 and 14.8 yards far from the goal on average.

Correlation Matrix



Above I put goalkeepers attributes' correlation matrix. With the help of this it is possible to investigate definitely association between each columns. Looking with attention at it, it's possible to identify some couples with high correlation level (I chose 0.5 or -0.5 as threshold): let's focus on those.

Wage - Value

Wage and *Value* attributes have a correlation of 0.57. That's explainable as, theoretically, best players are the most valuable and so usually are the one earning the most.

- **PSxG-GA - GoalsAgainst**

  As already commented, *PSxG-GA* and *GoalsAgainst* have a correlation of -0.53, meaning that the more the goalies are skilled at saving goals, the less goals they concede.

- **PSxG-GA - SavePercentage**

  Correlation between *PSxG-GA* and *SavePercentage* stands at 0.77, being a strong direct correlation. As can easily be guessed, goalkeeper saving ability and percentage of saves grow and decrease together.

- **GoalsAgainst - SavePercentage**

  GoalsAgainst and *SavePercentage* attributes have a correlation of -0.76, being one of the strongest. That's easily imaginable as *GoalsAgainst* takes part in *SavePercentage* equation, as an inverse term: the fewer goals are conceded (*GoalsAgainst*), the higher the *SavePercentage is.*

- **GoalsAgainst - CleanSheetPercentage**

  Similarly to the one above, *GoalsAgainst* and *CleanSheetPercentage* have a strong inverse correlation, of -0.76. Clean sheet (and so its percentage) has always been used as a performance measure of good defense and goalkeepers, and the better a defense is, the fewer goals are conceded.

- **Touches - Def.ActionsOutsidePen.Area**

  Touches and *Def.ActionsOutsidePen.Area* have a 0.57 correlation. This can be caused by the growing active presence of goalkeepers in building maneuver too: this usually leads goalkeepers to touch the ball more and cover a more advanced position on the court in both phases.

- **Launch% - GoalKicks**

  Launch% and *GoalKicks* have a correlation of 0.51. That can be addressable to the new way of building up the game the teams do. The game starts from the goalkeeper not with a long launch anymore, but with a possession phase. That culminates with the defender beating the goal kick back to the goalkeeper. This focus on possession changed the way goalies play and led to statistic correlations like this.

- **Launch% - Avg.LengthofGoalKicks**

  Similarly to the one above, *Launch%* and *Avg.LengthofGoalKicks* correlates with 0.75. The cause can be addressed to the same phenomenon, as teams using build-up with active goalkeeper launch less and less, while goalies who tend to launch more have developed more and more skill and power, kicking the ball further.

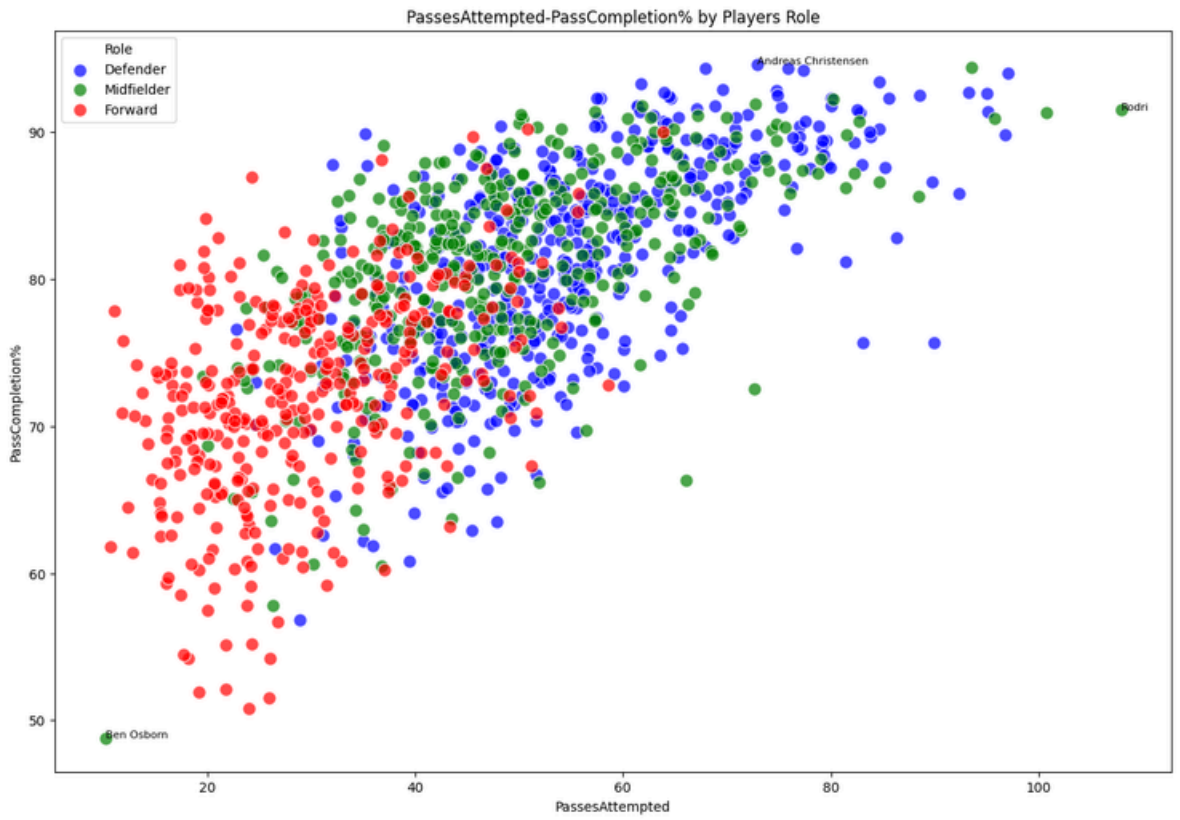
- **Def.ActionsOutsidePen.Area - Avg.DistanceofDef.Actions**

  Lastly, *Def.ActionsOutsidePen.Area* and *Avg.DistanceofDef.Actions* have a very strong correlation, with 0.81. Those attributes are easily relatable, as, usually, the more a goalkeeper gets out of the area, the further he will go on average.

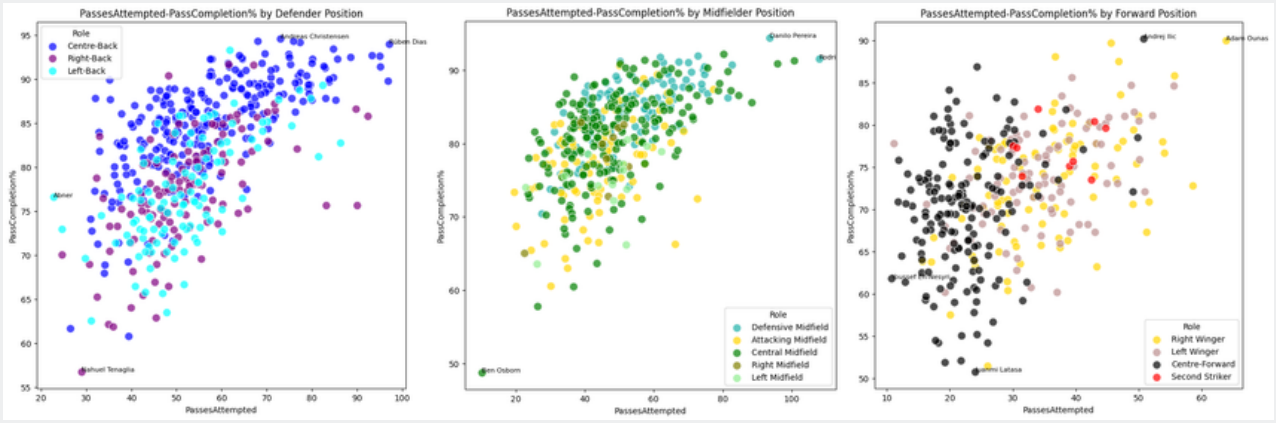
Movement Players

At this time, I conduct a deeper correlation analysis taking into account only movement players and their position.

PassesAttempted - PassCompletion%



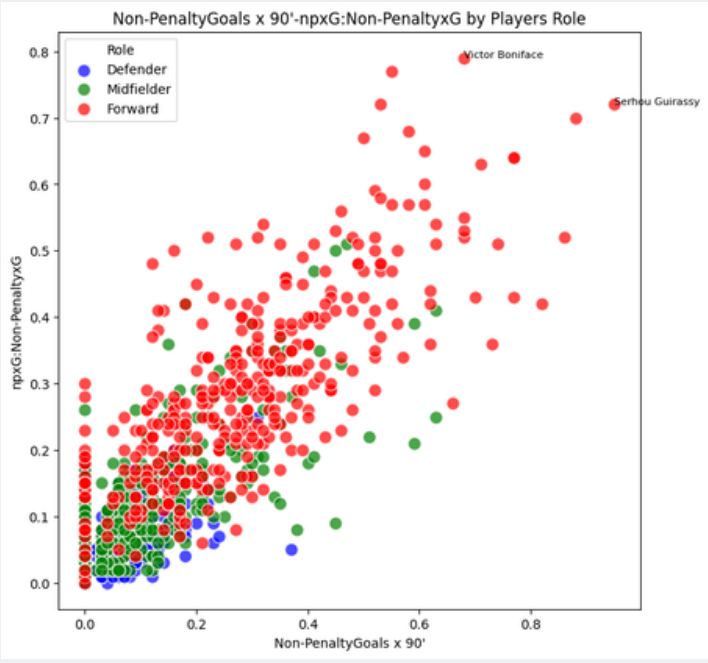
The scatterplot above investigates the relationship between *PassesAttempted* and *PassCompletion%*, highlighting players by their role (blue for defenders, green for midfielders and red for forwards). It is manifest the inferiority of forwards in terms of statistics, as they cover the lowest and leftiest part of the graph. Then there is a mixed mass in the middle, while defenders look to have the best pass accuracy. Most accurate player in the top 5 leagues is Barcelona's **Andreas Christensen**, with **94.6%** of passes completed, while the one attempting the most is, by far, Manchester City's **Rodri,** with **107.93** passes per 90' with **91.5%** of accuracy. The two variables appears quite **directly correlated**, as shown by its **0.69** factor, meaning that the players who attempt more passes are, usually, the ones who complete them the most.



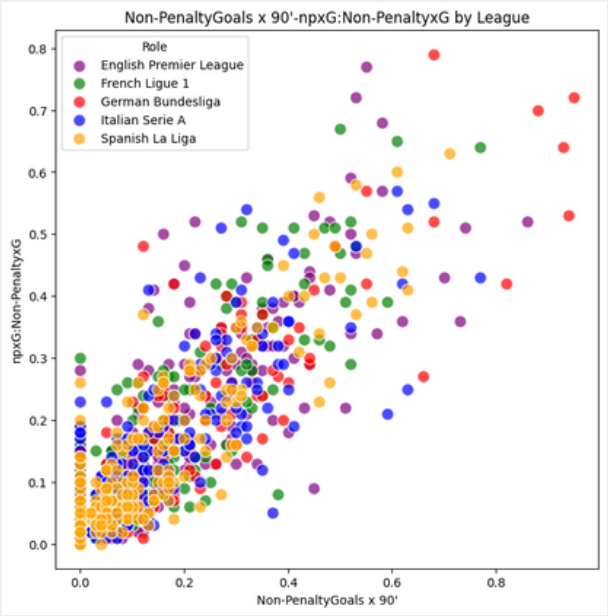
Defenders Midfielders Forwards

These single role *PassesAttempted-PassCompletion%* scatterplots show some important insights about roles and positions differences. While in defenders and midfielders graphs the two variables look more or less correlated, in **forwards'** one they look **unrelated**. In this one it's visible a cluster on the left side, featuring few *PassesAttempted* per 90', belonging to centre-forward players. The rest of the graph looks quite casual.
Midfielders scatter plot, on the other hand, divides players, as attacking midfielders look the ones attempting fewer passes and the least accurate, central midfielders look behaving averagely, while defensive ones attempt the most passes and are the most accurate. Defenders, similarly, are divided between centre-backs, the most prolific and accurate, and full-backs, lower and leftier in the graph.

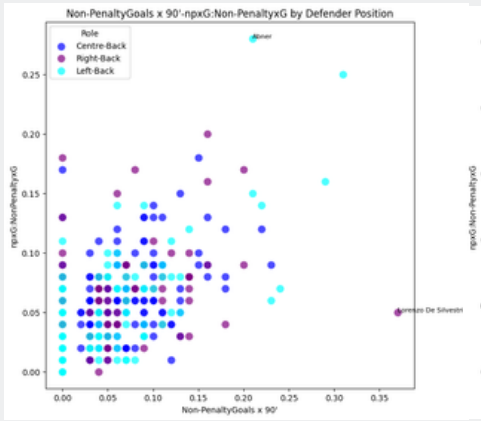
Non-PenaltyGoals - Non-PenaltyxG



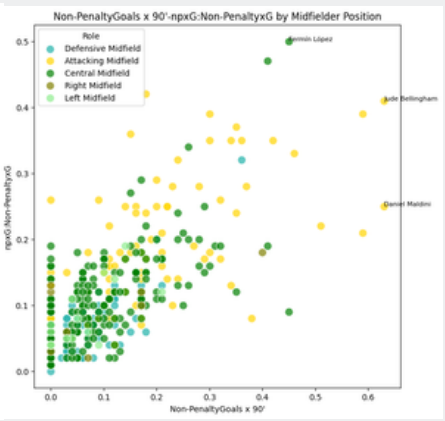
The graph above lists *Non-PenaltyGoals* juxtaposed to *Non-PenaltyxG*, in order to detect the relationship between goals and expected goals. As visible, the graph is quite useless, due to the major prevalence of forwards. This reveals a truth that can't be forgotten: **forwards are the ones who score**. Overall, the two most 'extreme' players look to be Bayer Leverkusen's **Victor Boniface** and Stuttgart's **Serhou Guirassy**, having respectively the highest value of *npxG*, at 0.79, and *npG*, at 0.95.



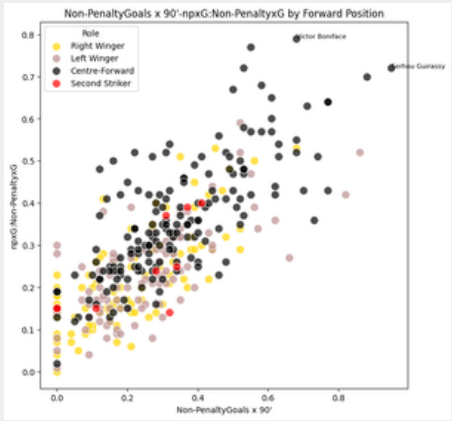
After looking at the previous results, I investigate the distribution of leagues for the same attributes. **Most prolific** players play in **Bundesliga**, while the league where are there the **highest number** of goal **opportunities** is **Premier League**. Those facts look quite realistic as Premier League is considered the most entertaining one, due to its offensive play style, while this season Bundesliga's top scorers have the best statistics among top 5 leagues.



Defenders



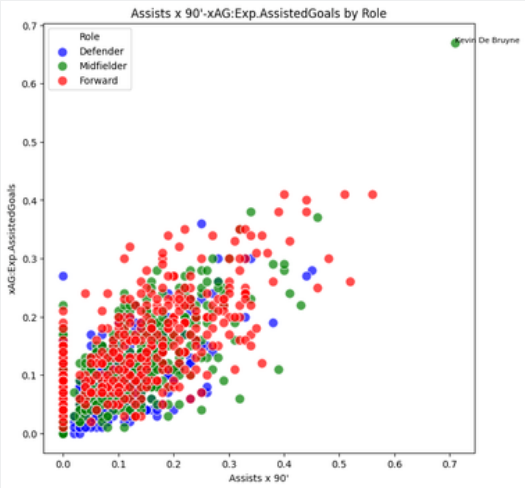
Midfielders



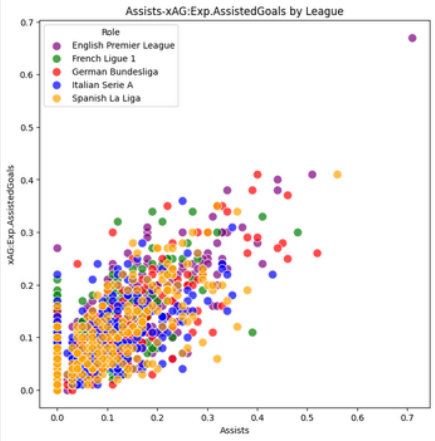
Forwards

The graphs above show different distributions of positions regarding *NonPenaltyGoals* and *NonPenaltyxG*. In defenders plot, as expected, centre-backs cover the lowest and leftiest part, while the outline is made by full-backs (especially wing-backs). It happens similarly in midfielders graph as defensive midfielders are the less involved ones in offensive chores, while central midfielders cover the central part of the graph and attacking midfielders fill the rest of the graph. Among the top performers I can list Granada's **Faitout Mouassa** and Real Betis' **Abner** as top defenders for *NonPenaltyxG* and Bologna's **De Silvestri** for *NonPenaltyGoals*, but the all the previous cited players don't have that much data to provide a safe outcome. Soon after them there is **Grimaldo** who, indeed, have a brilliant season ensuring for him.
Regarding midfielders, of course the most dominating one appears to be Real Madrid's **Jude Bellingham**, with more than 1 goals every 180 minutes. Then, Monza's Daniel Maldini looks to be a worthy contender, with exactly the same *NonPenaltyGoals*, but having less than 600 minutes played, I can't consider the statistic reliable. On the other hand, Barcelona's **Fermin Lopez** is the best midfielder for *NpxG*, followed by Manchester City's **Kevin De Bruyne**.
Scoring, as known, is the kingdom of forwards and, in particular, centre-forwards. They cover the top-righ part of the graph, having most performing players the ones already cited.
Metz's **Asoro** and Cadiz's **Darwin Machis** have surprising performance, having both *NonPenaltyGoals* above 0.20 and *NpxG* really close to 0. Then Liverpool's **Diogo Jota** and Bayern Munich's **Matthys Tel** have incredibly *NonPenaltyGoals* (0.86 and 0.82) even being left-wingers.

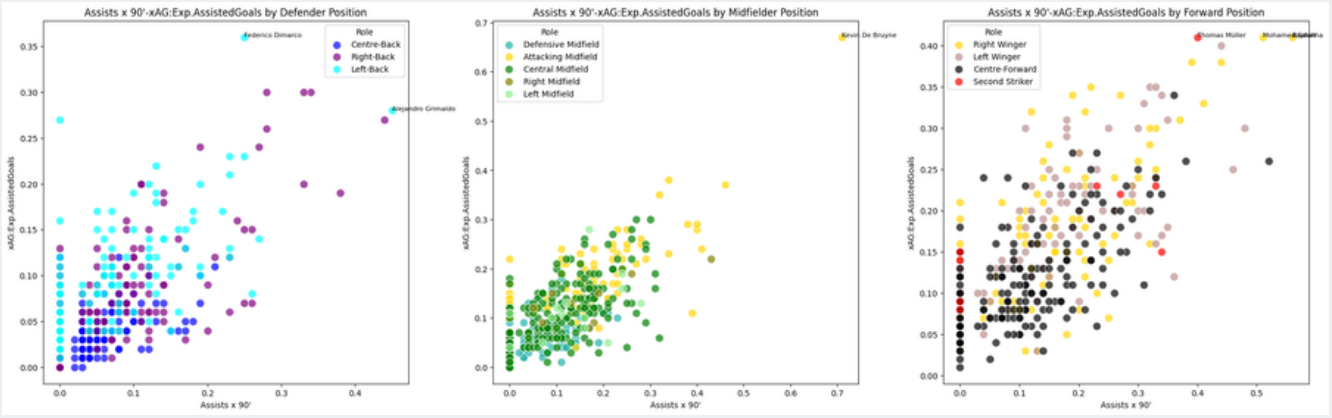
Assists - xAG: Exp. AssistedGoals



Scatteplot listed above investigates correlation between *Assists* and *xAG* (expected assisted goals). As visible there, defenders occupy the bottom-left part of the chart, while the remaining portion is filled by midfielders and forwards. The main outcome I can get from the plot is the **unfair advantage** Manchester City has in **Kevin De Bruyne**, who looks playing a completely different sport from the data.



This graph shows the same scatterplot but highlighting the *League* attribute. Based on the colors of the dots, I can state that the **leagues** with **best assistmen** are **Bundesliga** and **Premier League**, being only Barcelona's Raphinha and PSG's Ousmane Dembelé and Bradley Barcola ones not belonging to the previous cited leagues in the top 15 with best combinations.



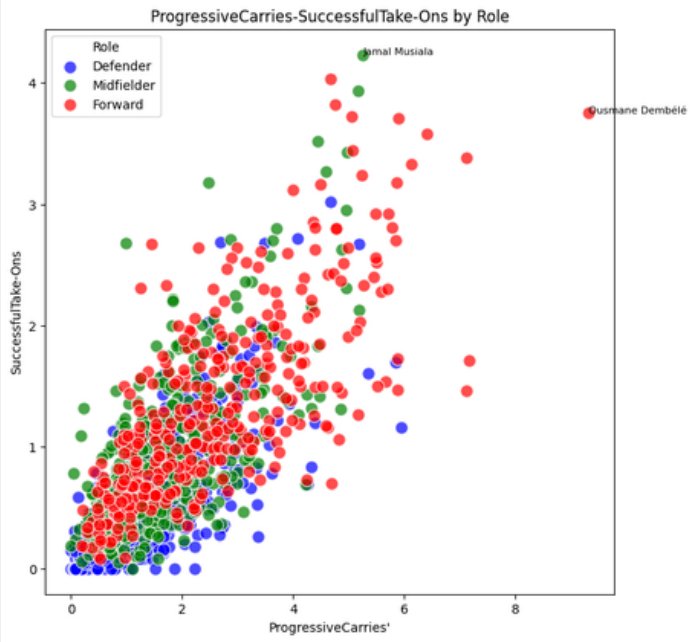
Defenders Midfielders Forwards

Defenders positions in-depth analysis regarding assists reveals, as expected, a major offensive contribution from full-backs than centre-backs, being Bournemouth's **James Hill** the only central defender overcoming **0.20 *Assists*** and Inter Milan's **Alessandro Bastoni** the other one having consistently more than **0.10 *xAG***. Then, there is a

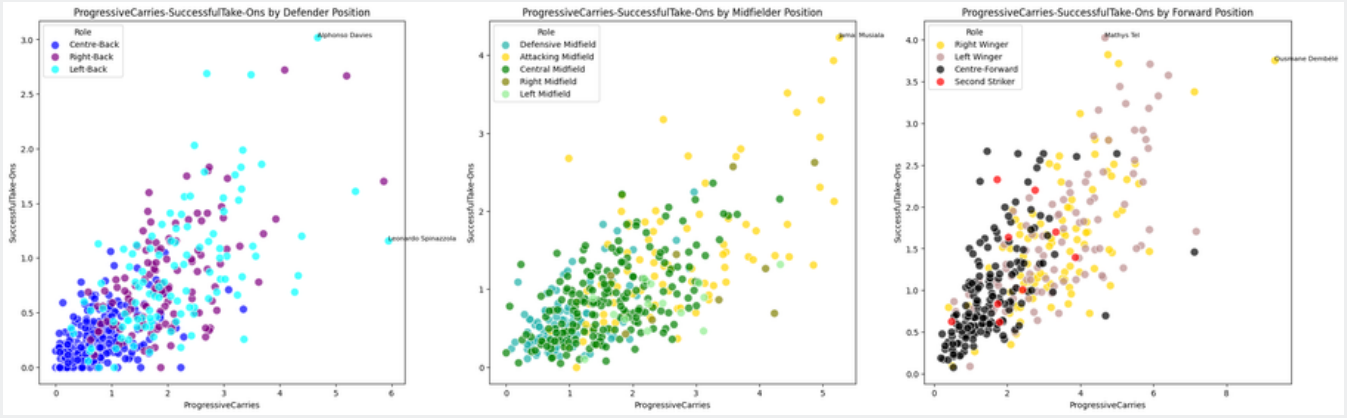
bunch of full-backs (mostly wing-backs) having surprisingly statistics as defenders. Best assistmen figure to be Leverkusen's **Grimaldo**, Stuttgart's **Pascal Stenzel**, Liverpool's **Trent-Alexander Arnold** (who often plays more as midfielder than defender), Newcastle's **Keiran Trippier** and Girona's **Yan Couto**. *xAG* leader, indeed, is Inter Milan's **Federico Dimarco**, having an outstanding **0.36**, but gaining only 0.25 assists per 90'.
Regarding midfielders, other than previous cited De Bruyne's dominance, there is an obviouos prevalence of attacking midfielders as assistmen compared to other positions, as 10 out of the 14 midfielders with more than 0.3 cover that position. Brest's **Kamory Doumbia**, Getafe's **Ilaix Moriba** and Aston Villa's **Youri Tielemans** (all central midfielders) overcome 0.3 *Assists* threshold, but have really low *xAG*, questioning their real assist merits. The other non attacking midfielder, then, is Frosinone's **Nadir Zortea**, with a surprising 0.43 *Assists* and a consistent *0.22 xAG.* Between top attacking midfielders, other than De Bruyne, only Borussia Dortmund's **Julian Brandt**, Leverkusen's **Jonas Hofmann** and Tottenham's **Giovani Lo Celso** have more than 0.30 *xAG.*
Central-forwards, as seen before, are the most prolific players on the field, so they tend not to have incredible passing and assisting statistics, with the exception of Leverkusen' **Victor Bonifac**e (0.52 *Assists* per 90', but only 0.26 *xAG*). Wingers and second strikers, instead, have the best statistics: Liverpool's **Mohamed Salah**, Leverkusen's **Leon Bailey**, Barcelona's **Raphinha** and Bayern Munich's **Thomas Muller** all have more than 0.40 in both *Assists* and *xAG*, but other wingers like Crystal Palace's **Michael Olise**, PSG's **Bradley Barcola** and Bayern Munich's **Matthys Tel** have more than 0.40 in *Assists* but lower values in *xAG*.

Overall, the two variables look **directly correlated**, as 0.76 correlation factor confirms.

ProgressiveCarries - SuccessfulTakeOns

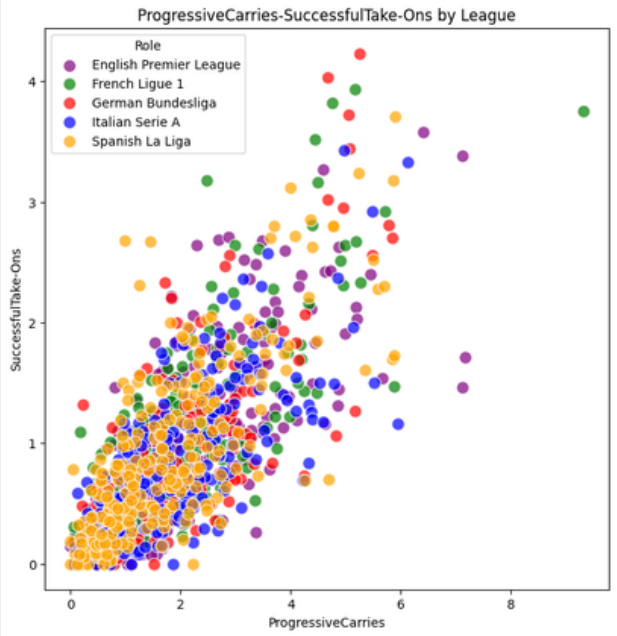


Graph here above exhibits a scatterplot having *ProgressiveCarries* and *SuccessfulTake-Ons* as axis. *ProgressiveCarries* denote carries that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes or any carry in the penalty area. *SuccessfulTake-Ons* indicates number of opponents taken on successfully by dribbling past them. With the role differentiation based on colors, it's possible to identify the great prevalence of forwards on high values of x-axis, with a mean higher than 2.5, against midfielders' 1.74 and defenders' 1.35, while there is a good combination of midfielders and forwards along *SuccessfulTake-Ons*. Regarding this attribute, defenders have a mean of 0.52, midfielders 0.93 and forwards 1.31.



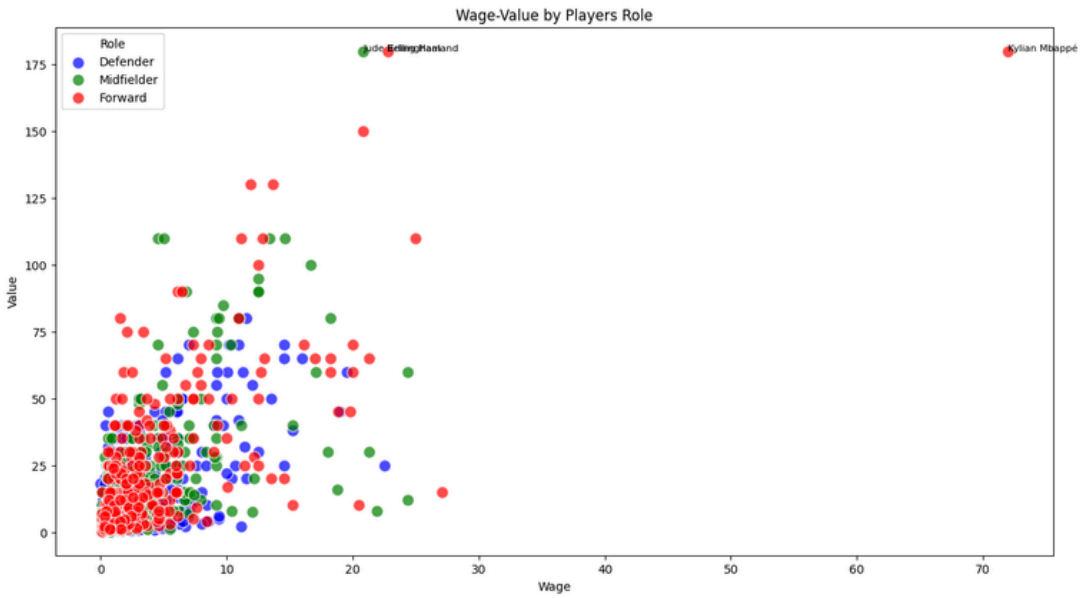
Defenders Midfielders Forwards

Similarly to previous attributes, defenders plot sees a prevalence of **full-backs** at high values, midfielders one sees a prevalence of **attacking midfielders** and forwards one **wingers**. Leftiest graph, indeed, have lowest values in centre-backs, as they don't tend to carry the ball and take on opponents. Roma's **Leonardo Spinazzola** is the defender carrying the ball the most, while Bayern Munich's **Alphonso Davies** is the one taking on the most players. Other players among the top in those attributes are Brighton's Valentin Barco, Wolverhampton's Rayan Ait-Nouri, Barcelona's Joao Cancelo and Alejandro Balde and Girona's Yan Couto.
Looking at midfielders plot, attacking midfielders look having higher values in both axis, with means of 2.7 in *ProgressiveCarries* and 1.45 in *SuccessfulTake-Ons*, compared to 1.62 and 0.87 for central midfielders and 0.99 and 0.61 for defensive midfielders. Right and left midfielders have higher values even than attacking midfielders, but they are only represented by few players. Top performer in both columns is Bayern Munich's **Jamal Musiala**, followed by Lyon's Rayan Cherki, Rennes' Désiré Doué, Brighton's Julio Enciso and Leverkusen's Florian Wirtz. Sassuolo's Cristian Volpato too has incredible statistics, but with a limited play time.
It's important also to note that, among the players having more than 4.0 at least in one of the given attributes, only 4 players don't cover attacking midfielders position: I'm talking about Roma's Nicola Zalewski, who's considered left-midfielder, and three right-midfielders, Bournemouth's Romain Faivre, Inter Milan's Tajon Buchanan and Torino's Raoul Bellanova.
Forward plot, in the same way, follows the graph regarding *Assists*, with wingers dominating the top-right part of the figure. Here, too, Barcelona's **Ousmane Dembelé** and Bayern Munich's **Matthys Tel** are among top performers, then there are Manchester City's Jack Grealish, Tottenham's Timo Werner (only centre-forward among top players), Chelsea's Noni Madueke and Bournemouth's Luis Sinisterra (this last one with only 645 minutes played) for *ProgressiveCarries* column, and, in addition to the above mentioned, Lille's Edon Zhegrova, Bayern Munich's Leroy Sané, and Athletic Bilbao's Nico Williams, really good in *SuccessfulTake-Ons* statistic.

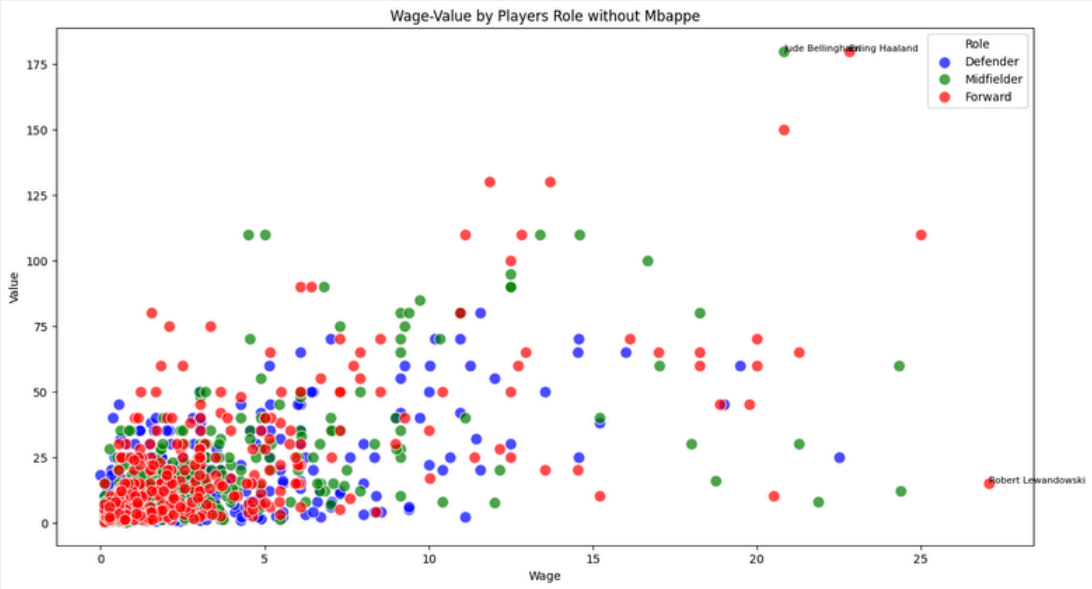


This graph looks for differences in game styles between leagues and, actually, French and English leagues appear to be the ones with the best dribblers. Both have also highest means in *ProgressiveCarries* and *SuccessfulTake-Ons*, with 1.92 and 0.90 for Premier League and 1.88 and 0.97 for Ligue 1, compared with 1.79 and 0.87 for Bundesliga, 1.79 and 0.79 for Serie A and 1.72 and 0.89 for La Liga. With those statistics, **Premier League** and **Ligue 1** look to prefer more **dribbling** and **individual plays**, while the others favor more team-oriented play styles.

Wage - Value

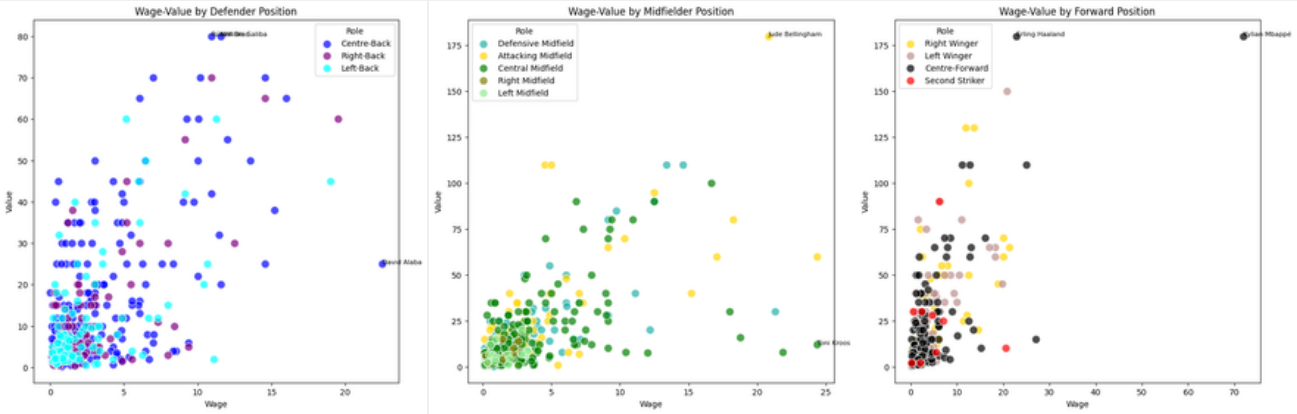


Now let's move away from game statistics to player value and wage informations. The chart above illustrates *Wage* variable on the x-axis and *Value* on the y-axis, both in terms of millions of €. This graph doesn't look that interpretable due to the shrink of the axis in order to keep PSG's **Mbappé** visible. So I recreate the scatterplot removing his data.



This graph conveys some curious information: initially it is important to note that the attribute **most correlated** with value is ***Wage***, with a factor of **0.66**, but keeping only players worth more than 50 million this collapses to 0.49, while it grows again to 0.63 by changing this threshold to 75 million. This shows that those two variables look loosely correlated for players whose *Value* is between 30th and 80th quantile, while it's strongly correlated for players outside of this interval. According to sources, the players who provide more value to their club the most compared to their salary are **Florian Wirtz** of Leverkusen, **Kvicha Kvaratskhelia** of Napoli, **Evan Ferguson** of Brighton, and **Lamine Yamal** of Barcelona. On the other hand, footballer's with very expensive contracts but with not valueing much, are Real Madrid's Luka Modric, Toni Kroos and David Alaba, Barcelona's Robert Lewandowski and Bayern Munich's Thomas Muller. Of course, this information should be taken with caution, because these are experienced players near the end of their careers, earning a lot on expiring contracts made when they were shining stars of their clubs.
It is also important to note the different distribution of the players based on their role: it's way harder for a defender to be valued more

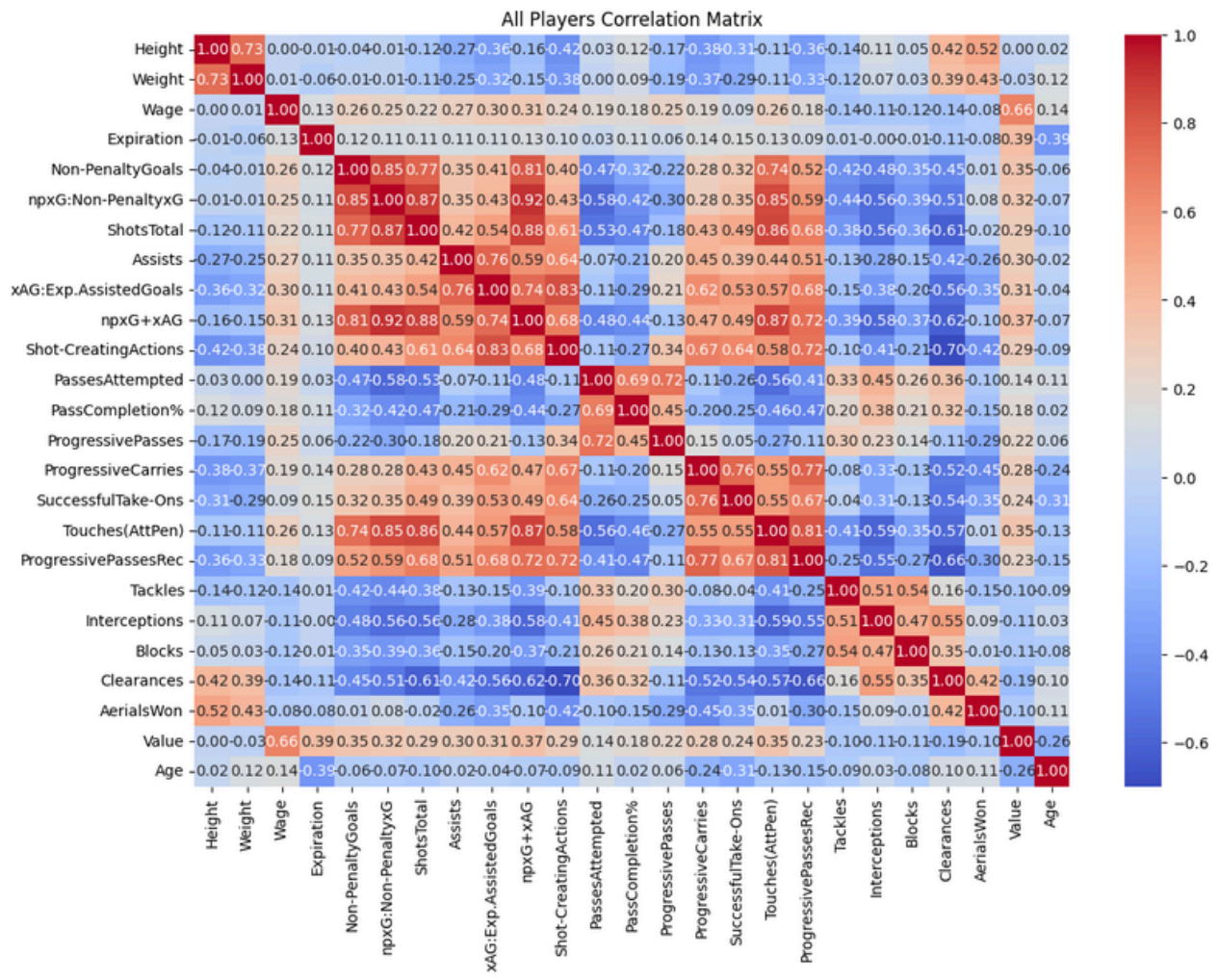
(and to gain more) than for a midfielder and, of course, a forward. This is marginal for low value players, but becomes more and more evident if I consider only high value players. €12.3M and €2.7M are respectively mean *Value* and *Wage* regarding defenders only, it grows to €16.4M and €3.0M for midfielders and €19.9M and €3.7M regarding forwards. Keeping only rows with *Value* higher than €50M, those figures become €65.3M and €11.1M for defenders, €87.1M and €11.4M for midfielders and €85.8M and €13.5M for forwards.



Defenders Midfielders Forwards

Above scatterplots give insights about *Wage* and *Value* in different positions. Leftmost graph shows that position doesn't affect much previously cited feature, having positions similar means, also taking into consideration only most valuable defenders.
Midfielders scatter plot, instead, looks a bit different, due to way higher attacking midfielders attributes data. Considering all players, indeed, they have *Value* mean of €20.7M, opposed to central midfielders' €15.5M and defensive ones' €16.4M. Taking into consideration only players whose *Value* is higher than €50M, those data grow to €92.2,M €81.1M and €88.3M, while *Wage* mean values are €13.5M, €9.6M and €10.7M.
Forwards plot, lastly, doesn't provide enough insights as doesn't appear to be any discrepancy. Looking at the means, central forward ones are lower due to higher amount of data, while second strikers look to be more valuable, but their sample is really small. Applying same filter as before, all means reach between €80M and €85M, except second strikers, as said before.

Correlation Matrix



I put movement players attributes' correlation matrix. With it, I can investigate correlation between each columns. I will focus on variables strongly associated (values higher than 0.5 and lower than -0.5) not already analyzed. Then I will go deeper in the analysis, listing single role correlation matrix.

Height - Weight

Height and *Weight* values are associated with a 0.73 correlation value. This, obviously, doesn't give any football statistical information, as those are only biological data.

NonPenaltyGoals - ShotsTotal

NonPenaltyGoals and *ShotsTotal* features have a 0.77 correlation, due to the necessity of shots to score a goal. As

obvious, the more a player shoot, the more he will probably score and, so, the higher *NonPenaltyGoals* value he will have.

- **NonPenaltyGoals - npxG+xAG**
  Correlation between *NonPenaltyGoals* and *npxG+xAG* is really strong, with a 0.81 factor. That, as really intuitable, is caused by the presence of *npxG* in the second attribute sum. *npG* is much correlated with *npxG* and this latter one is in the equation, so, for transitive property, they are strongly correlated.

- **NonPenaltyGoals - Touches(AttPen)**
  Touches(AttPen) indicates number of touches in attacking penalty area. Due to this meaning, it's easy to realize why it's correlated with *NonPenaltyGoals* with a 0.72 correlation factor: the more a footballer stays in offensive penalty area, the more he should score.

- **NonPenaltyGoals - ProgressivePassesRec**
  ProgressivePassesRec, instead, means the number of passes received with a ball progress of at least 10 yards. Here, too, the correlation is explainable, due to the high number of goals scored after verticalizations. Correlation is not that incredibly strong, with a 0.52 factor.

- **NonPenaltyxG -ShotsTotal**
  NonPenaltyxG and *ShotsTotal* correlate with a huge 0.87 factor. As visible, quite all variables correlated with *NonPenaltyGoals* are correlated with *NonPenaltyxG* too,. Here, similarly to the correlation with *NonPenaltyGoals*, it's possible to state that the more shots a player does, the more *xG* he cumulates and so the higher value he gets.

- **NonPenaltyxG -npxG+xAG**
  Correlation between *NonPenaltyxG* and *npxG+xAG* is the highest in correlation matrix, with a 0.92 coefficient. That's

immediately understandable, as second variable contains first one in its formula, so they are surely directly correlated.

- **NonPenaltyxG - PassesAttempted**

  Then I notice an unusual inverse correlation between *NonPenaltyxG* and *PassesAttempted* with a -0.58 factor. That can be justified with the discrepancy of 'tasks' players need to perform on the field. Of course players who attempt more passes tend to be good assistmen but not good scorers as, obviously, if they pass the ball often, they can't shoot much.

- **NonPenaltyxG - Touches(AttPen)**

  Likewise with *NonPenaltyGoals*, more touches in offensive penalty area mean more shots and increase in xG. These two attributes are correlated with a 0.85 coefficient.

- **NonPenaltyxG - ProgressivePassesRec**

  The same type of argument can be done for *ProgressivePassesRec*. Received passes don't necessarily mean shots, so correlation is looser, 0.59.

- **NonPenaltyxG -Interceptions**

  NonPenaltyxG and *Interceptions* due to the offensive nature of *xG* opposed to the defensive nature of interception action. As those belong to opposite player roles, it's fair to have a inverse correlation coefficient of -0.56 in this combination.

- **NonPenaltyxG -Clearances**

  Same type of speech can be given if *Clearances* variable is put instead of *Interceptions*. Clearance can be considered a bit less proper of defensive players as it can happen in corner kicks situations too and can be made by whichever player. So their correlation coefficient stands at -0.51.

- **ShotsTotal - Exp.AssistedGoals(xAG)**

  These two variables surprisingly correlate with a 0.54 factor. The reason behind this can be found in the identity of roles topping both attributes charts. Offensive players tend to have so way higher values in the two variables that the correlation crashes down to 0.17 if it's considered only forwards dataset.

- **ShotsTotal - npxG+xAG**

  Likewise quite all correlation seen before, *ShotsTotal* imply a contribution to *npxG* as this value can be seen as a summation of all shots, each multiplied per a different factor. Given that, the contribution appears quite manifest, with a 0.88 coefficient.

- **ShotsTotal - ShotCreatingActions**

  Similarly to the previous one, this combination of variables implies a direct contribution of the first in the second variable value. This leads to a direct correlation of 0.61.

- **ShotsTotal - PassesAttempted**

  As indicated priorly, finishers usually don't participate much in building-up phase, where many passes are attempted. That's why those two variables are inversely correlated, with a correlation factor of -0.53.

- **ShotsTotal - Touches (AttPen)**

  ShotsTotal and *Touches* correlate with a 0.86 coefficient, for the reason listed earlier, as forwards tend both to touch the ball in offensive penalty area much more than other players and to shoot the ball more often.

- **ShotsTotal - ProgressivePassesRec**

  Of course here too, the more a player receives progressive passes, the closer he is to penalty area, the more he will tend to shoot. These have a 0.68 correlation.

- **ShotsTotal - Interceptions**

  Likewise previous cases, *ShotsTotal* is a clearly offensive attribute, so he has an inverse correlation with defensive variables like *Interceptions*, in this case of -0.56.

- **ShotsTotal - Clearances**

  In the same way as the one above, *ShotsTotal* and *Clearances* correlate with -0.61.

- **Assists - npxG+xAG**

  Assists is strongly correlated with *xAG* and this one is an addend in second column equation: this results in a 0.59 correlation.

- **Assists - ShotCreatingActions**

  Here, too, *ShotCreatingActions* definition includes volume of passes as addends, so some of the players who create more *ShotCreatingActions* serve more *Assists* too, with a 0.64 factor.

- **Assists - ProgressivePassesRec**

  This correlation looks a bit counterintuitive, but usually *Assists* are passes done close to the goal, sometimes after receiving long passes from behind. That can be noticed as these two attributes correlate with a 0.51 factor.

- **Exp.AssistedGoals (xAG) - npxG+xAG**

  Likewise the case of npxG - npxG+xAG, xAG contributes directly to npxG+xAG value, so it must be directly correlated, in this case with a 0.74 value.

- **Exp.AssistedGoals (xAG) - ShotCreatingActions**

  Here, too, passes (and so *Assists*) contribute to *ShotCreatingActions* variable, so *xAG* automatically is correlated, with 0.83 coefficient.

- **Exp.AssistedGoals (xAG) - ProgressiveCarries**
  xAG and *ProgressiveCarries* are correlated, as sometimes dangerous passes follow players long carries. They correlate quite loosely, as 0.62 coefficient indicates.

- **Exp.AssistedGoals (xAG) - SuccessfulTake-Ons**
  As imaginable, often a *ProgressiveCarries* imply taking on defenders and so the *xAG* has high correlation with both, in this case resulting in a 0.53 factor.

- **Exp.AssistedGoals (xAG) - Touches(AttPen)**
  All attributes related to goals are very likely to be correlated with *Touches* as goal actions usually happen in penalty area. Here the coefficient is 0.57.

- **Exp.AssistedGoals (xAG) - ProgressivePassesRec**
  Similarly to *Assists*, *xAG* correlates with *ProgressivePassesRec* with 0.68, as usually assistmen follow team building-up phase with *ProgressivePasses* too.

- **Exp.AssistedGoals (xAG) - Clearances**
  In the same way as quite all other offensive statistics, there is a -0.56 inverse correlation with Clearances.

- **npxG+xAG - ShotCreatingActions**
  These two attributes have a direct correlation of 0.68 due to the simillar meaning of them, as *ShotCreatingActions* include passes and shots (and so *npxG* and *xAG*),

- **npxG+xAG - Touches(AttPen)**
  npxG+xAG and *Touches* have a really strong direct correlation (0.87) . This, of course, follow both *npxG* and *xAG* correlations with *Touches.*

- **npxG+xAG - ProgressivePassesRec**

  npxG+xAG correlates with *ProgressivePassesRec* as both its addend do. The factor is 0.72.

- **npxG+xAG - Interceptions**

  Likewise previous cases, offensive metric *npxG+xAG* has -0.58 inverse correlation with *Interceptions.*

- **npxG+xAG - Clearances**

  In the same way as above, *npxG+xAG* has -0.62 correlation with *Clearances*.

- **ShotCreatingActions - ProgressiveCarries**

  Due to the many highly correlated variables in common, *ShotCreatingActions* and *ProgressiveCarries* have a 0.67 correlation, as *ProgressiveCarries* often lead to a goal action.

- **ShotCreatingActions - SuccessfulTake-Ons**

  Here too, a take-on often lead to a shot creating action, making those variables correlated with a 0.68 coefficient.

- **ShotCreatingActions - Touches(AttPen)**

  As quite all previous offensive attributes, *ShotCreatingActions* is correlated with *Touches* as the prevalence of shot actions happen in goal area. Correlation factor is 0.58.

- **ShotCreatingActions - ProgressivePassesRec**

  ShotCreatingActions and *ProgressivePassesRec* have a 0.72 correlation due to usual concurrence of a shot creating action following a progressive pass received.

- **ShotCreatingActions - Interceptions**

  Here, too, offensive attribute *ShotCreatingActions* is inversely correlated with defensive attribute *Interceptions* with a -0.70 coefficient.

- **PassesAttempted - ProgressivePasses**

  PassesAttempted and *ProgressivePasses* have a 0.72 correlation. This can occur since *ProgressivePasses* are just a particular type of passes, and so they add to *PassesAttempted* total.

- **PassesAttempted - Touches(AttPen)**

  These two variables have a -0.56 correlation, as playmaker midfielders, who play the highest number of passes, tend to stay far from the penalty area.

- **ProgressiveCarries - Touches(AttPen)**

  ProgressiveCarries and *Touches* correlate with a 0.55 coefficient: that looks caused from a concurrence, as attacking midfielders and wingers are the players with higher *ProgressiveCarries* values, and, as offensive players, with most *Touches* too. Given that, if I filter the DataFrame to have only forwards, these variables' correlation crashes down to 0.28

- **ProgressiveCarries - ProgressivePassesRec**

  0.77 correlation between these two variables can be traced back to a similar reason than the previous correlation. Offensive players have highest values in both variables, so the correlation is caused by the concomitance of the two attributes.

- **ProgressiveCarries - Clearances**

  ProgressiveCarries and *Clearances* share a -0.52 correlation, likewise all previous correlations between offensive attributes and variables like *Interceptions* and *Clearances*.

- **SuccessfulTake-Ons - Touches(AttPen)**

  Correlation between *SuccessfulTake-Ons* and *Touches* stands at 0.55, for same reason as *ProgressiveCarries*: offensive players (excluded centre-forwards) dominate

SuccessfulTake-Ons charts and, of course *Touches* ones. Then a small portion of take-ons happens in penalty area, contributing to both attributes. Centre-forwards mitigate this correlation since they don't take-on that much.

- **SuccessfulTake-Ons - ProgressivePassesRec**

  ProgressivePassesRec too is a statistic widely dominated by offensive players. So, as before, the concurrence lead to a 0.67 correlation.

- **SuccessfulTake-Ons - Clearances**

  SuccessfulTake-Ons and *Clearances* have a -0.54 inverse correlation. The reason behind that can be found, as previously, in the opposition of game phases tasks require, as first variable is explicitly offensive, while second one is defensive.

- **Touches(AttPen) - ProgressivePassesRec**

  Correlation between *Touches* and *ProgressivePassesRec* stands at 0.81. Many progressive passes are received (or are followed by touches) in penalty area, so there are often concurrences of both variables. Anyway, dominance of forwards in both statistics leads to direct correlation.

- **Touches(AttPen) - Interceptions**

  Touches and *Interceptions* share an inverse correlation of -0.59. Forwards have most touches in penalty area, while defenders have the most interceptions, so the correlation is justified.

- **Touches(AttPen) - Clearances**

  In exactly the same way *Touches* and *Clearances* have a -0.57 correlation.

- **ProgressivePassesRec - Interceptions**
  Similarly to what stated before, *ProgressivePassesRec* and *Interceptions* share a correlation of -0.55, due to the different game phases where the two actions occur.

- **ProgressivePassesRec - Clearances**
  Likewise, same thing happen with *Clearances* attribute, with a -0.66 correlation coefficient.

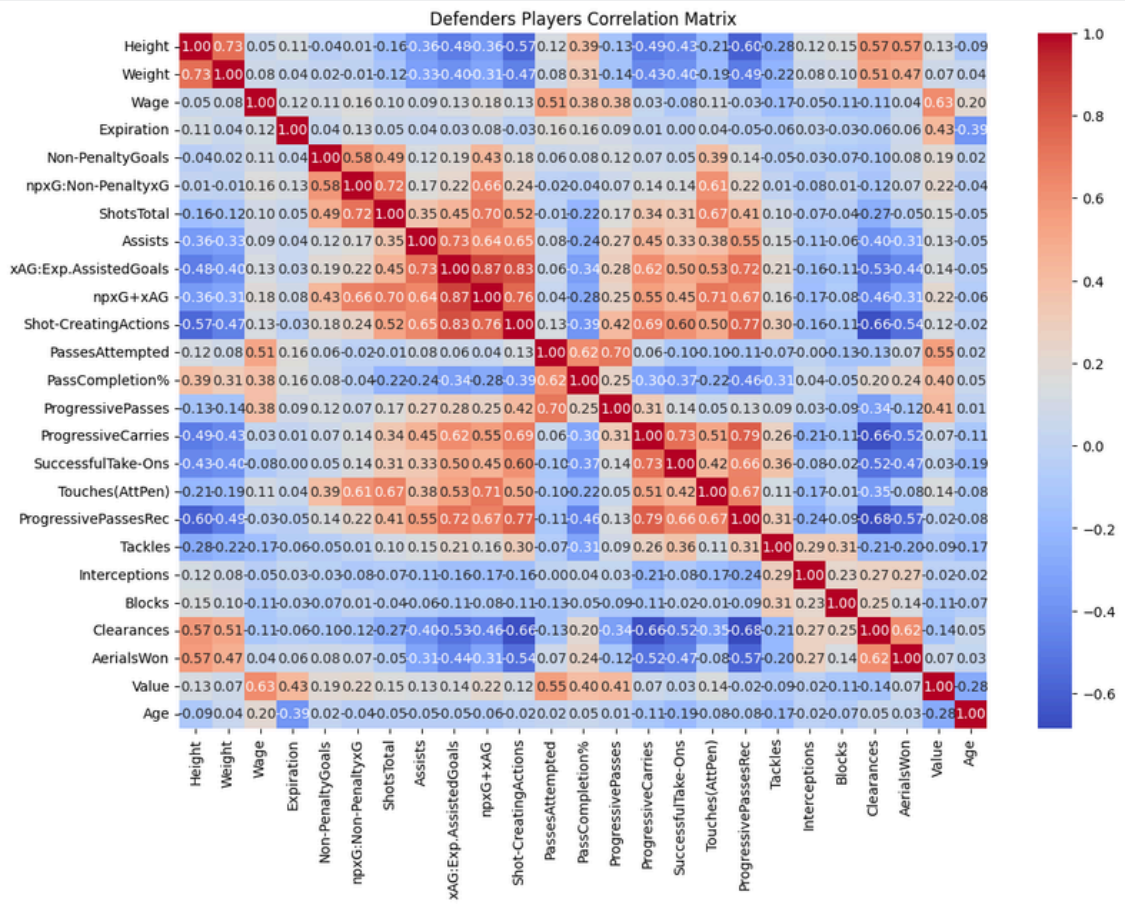
- **Tackles - Interceptions**
  Tackles and *Interceptions* have a 0.51 direct correlation. That can be caused by the presence of quite high values in midfielders and defenders and, in parallel, low values in both attributes for attackers.

- **Tackles - Blocks**
  Same thing happens for *Tackles* and *Blocks* variables, with a 0.54 factor. Both attributes regard pretty defensive plays, so these correlations shouldn't surprise. Defenders' correlation matrix, indeed, make this correlation crash down to 0.31.

- **Interceptions - Clearances**
  Following previous defensive attributes, *Interceptions* and *Clearances* correlate with a 0.55 factor.

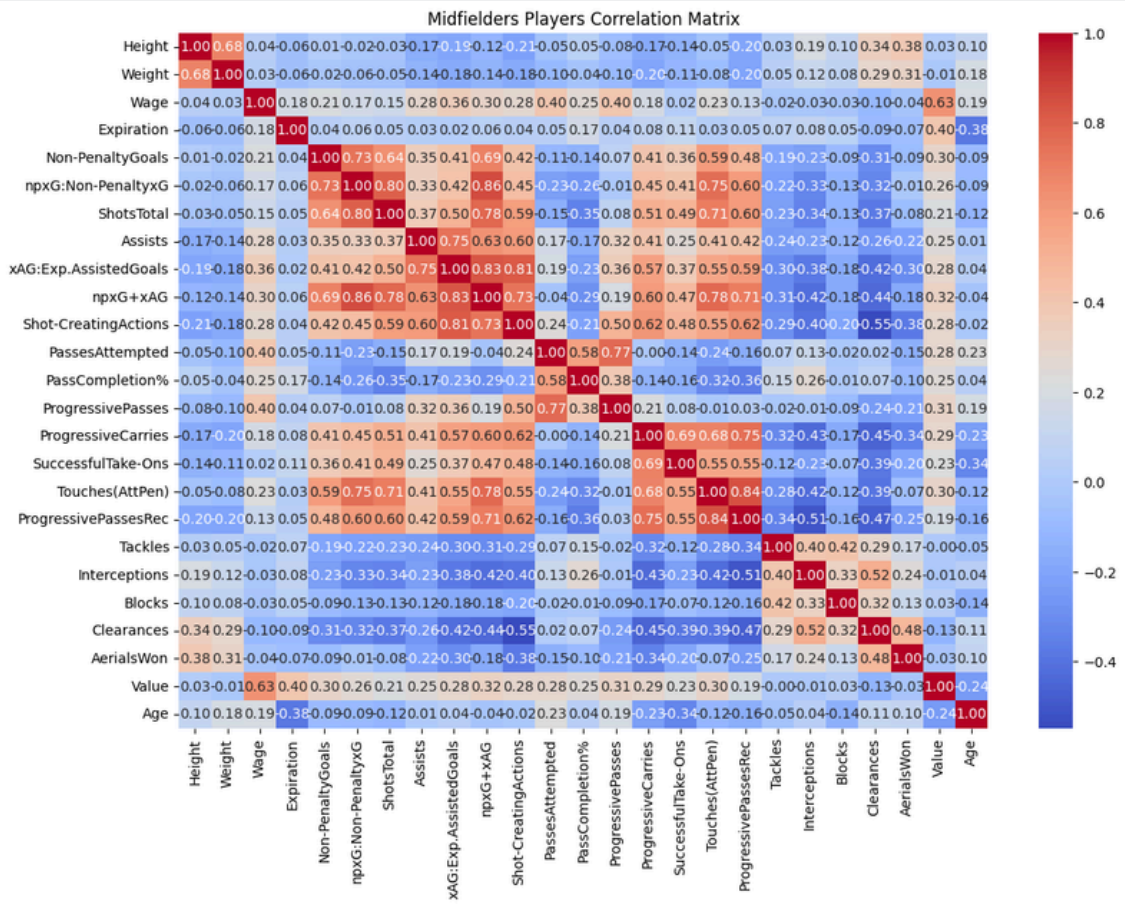
Now, let's have a brief look at single role correlation matrices, highlighting main differences with general one.

Defenders



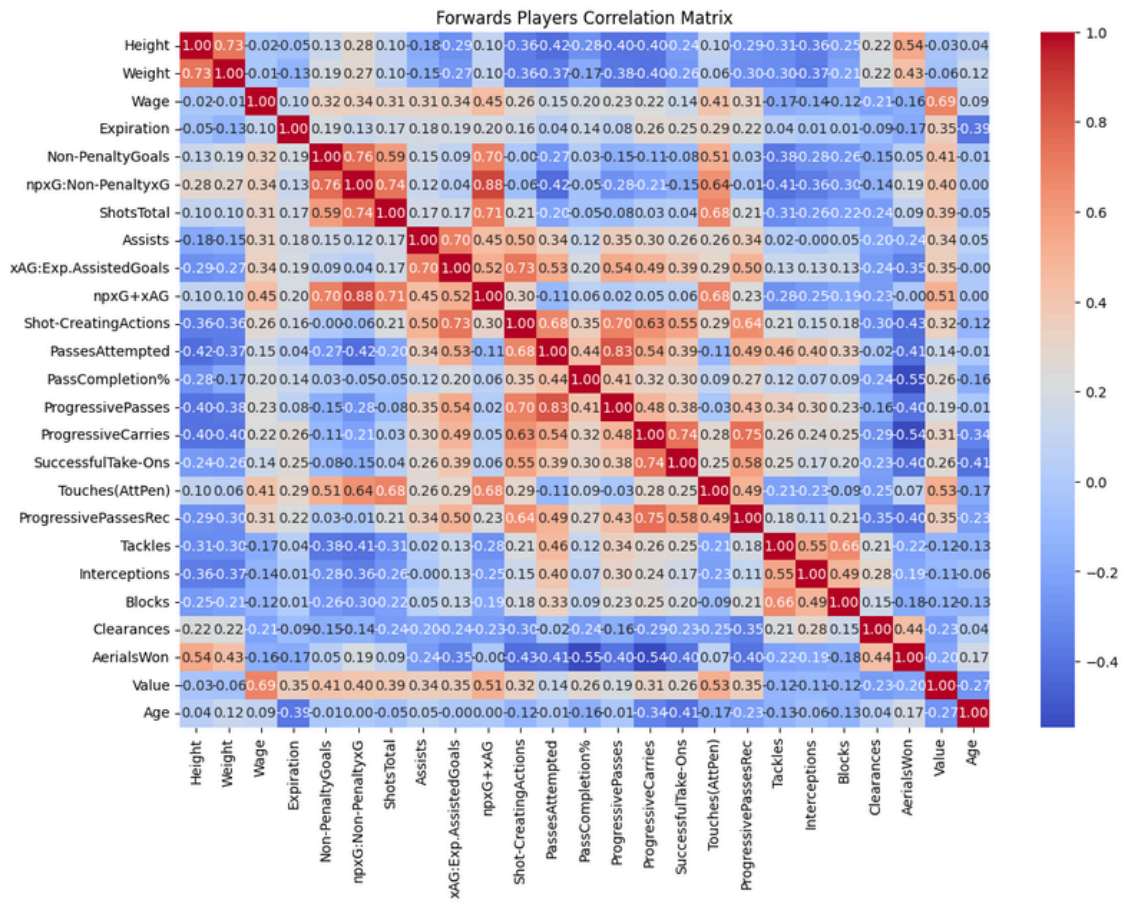
The matrix, overall, looks having much softer colors, since it deals with more similar players than the whole DataFrame. Remaining direct correlations are ones between offensive variables regarding assists and goals and those between carries and take-ons. Then, some inverse correlations between offensive features such as *ShotCreatingActions* and defensive ones like *Clearances* or *Interceptions* stay high.

Midfielders



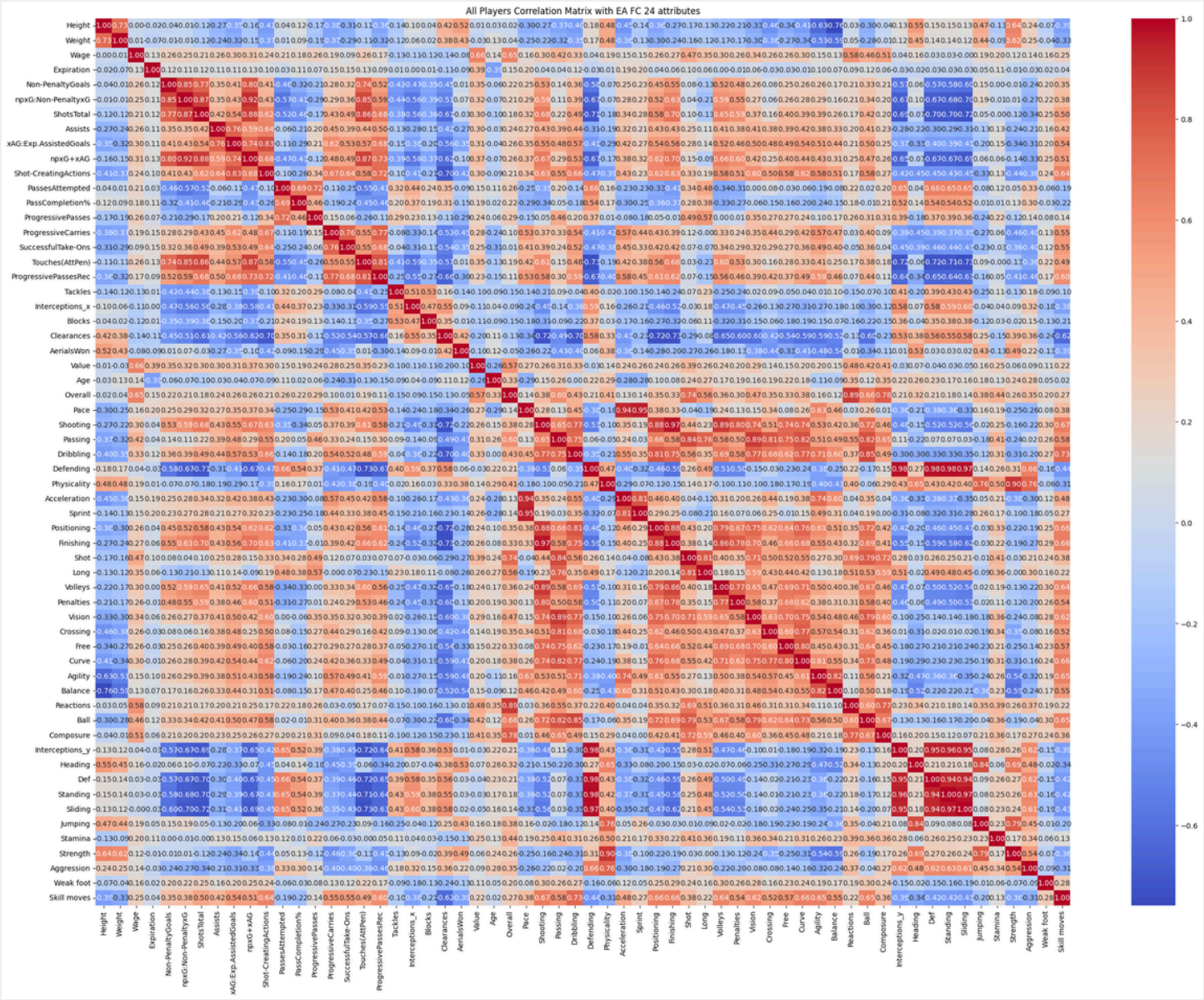
Midfielders' matrix looks similar to previous one regarding heat zones, but, overall, is more polarized, having more variables highly correlated (directly or inversely) and less with values near 0. Highest values of correlation remains, more or less, in same zones, with few minor changes.

Forwards



Lastly, correlation matrix of only forwards DataFrame changes a bit from previous ones, as scoring and assist attributes become way less correlated and quite all variables other than that share correlations with coefficients minor than 0. This table is maybe caused by the greater similarity among players than other roles. Surprisingly, forwards have a 0.55 correlation between *Touches* and *Value*, which is a really different data compared to others correlation matrices.

EA FC Correlation Matrix



Then I attach correlation matrix of all players after join with attributes from EA Sports FC 24. As visible from the color zones in the matrix above, there aren't huge correlations between real matches attributes and EA FC ones: quite all dark red squares are in the upper left side or in the lower right part, meaning that high coefficient correlations belong to couples of EA FC variables or real life matches. In the lower left part, instead, there are some inverse correlations between real life offensive statistics (like *npG, npxG, ShotsTotal*) and in-game defensive attributes (like *Defense, Standing, Sliding, Jumping*).

Then, there are other correlations, providing some EA FC 24 attributes evaluation criteria.
Correctly, variables from real-life data with highest correlations with *Overall* are *Value* and *Wage*, which, theoretically, should describe the best players.
Looking at 6 main players attributes (*Pace, Shooting, Dribbling, Passing, Defending, Physicality*), some curious data emerges:

- **Pace**
  Variables most correlated with *Pace* are *SuccessfulTake-Ons* (0.53) and *ProgressivePassesRec* (0.53), that can look reliable since, as seen before, those are statistics where wingers and full-backs (on average, the fastest players) top all other positions.

- **Shooting**
  Shooting has highest correlations with *ShotsTotal* (0.68) and, quite surprisingly, with *npxG+xAG* (0.67), while *npG* is only 0.53 and *npxG* is 0.59.

- **Dribbling**
  Dribbling, on the other hand, has two highest (or, in this case, lowest) coefficients with *Shot-CreatingActions* (0.66) and, with surprise, *Height* (-0.40). This looks quite unusual, as, until now, physical attributes didn't correlate much with statistics.

- **Passing**
  Passing EA FC statistics, based on correlations, doesn't look too reliable: most correlated variables are *Shot-CreatingActions* (0.55) and *Clearances* (inverse correlation with -0.49 coefficient). *PassesCompletion%* and *ProgressivePasses* have respectively only 0.05 and 0.46.

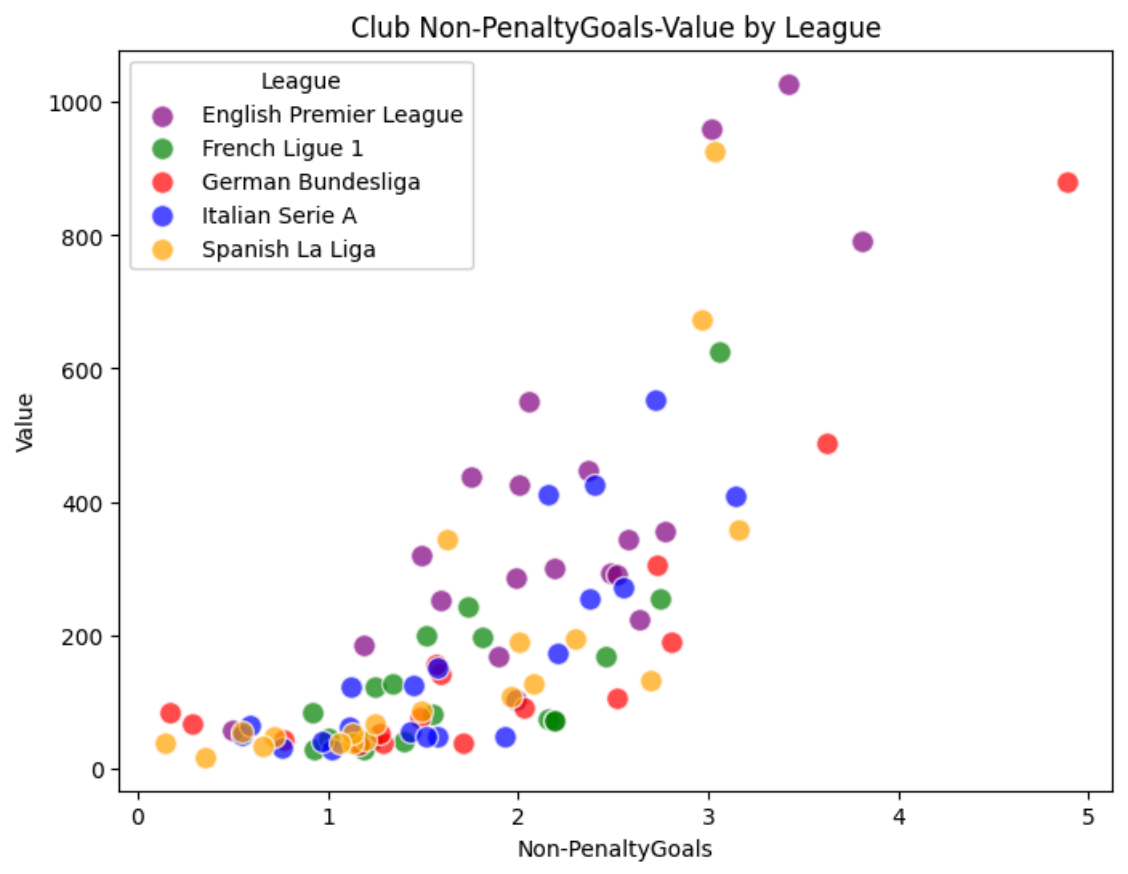
- **Defending**
  Defending attribute strong inverse correlations with offensive attributes, such as *ShotsTotal* (0.71), *Touches(AttPen)* (0.73) and *npxG* (0.77), while defensive stats have lower correlations (i.e. *Blocks* with 0.37 and *Clearances* with 0.58)

- **Physicality**

  Physicality, finally, has not any really strong correlation, having strongest ones with *Height* and *Weight*, both at 0.48.

Clubs Aggregate

- **Wage - Value**



Clubs *NonPenaltyGoals - Value* plot above looks following a hyperbolic function. Analyzing the graph, it appears manifest that, **in order to score more goals, it's necessary to have a more valuable team** and, from a certain point on (in this case I would say from 2 *npG*), it's mandatory to invest huge sums in the team to increase *npG*.

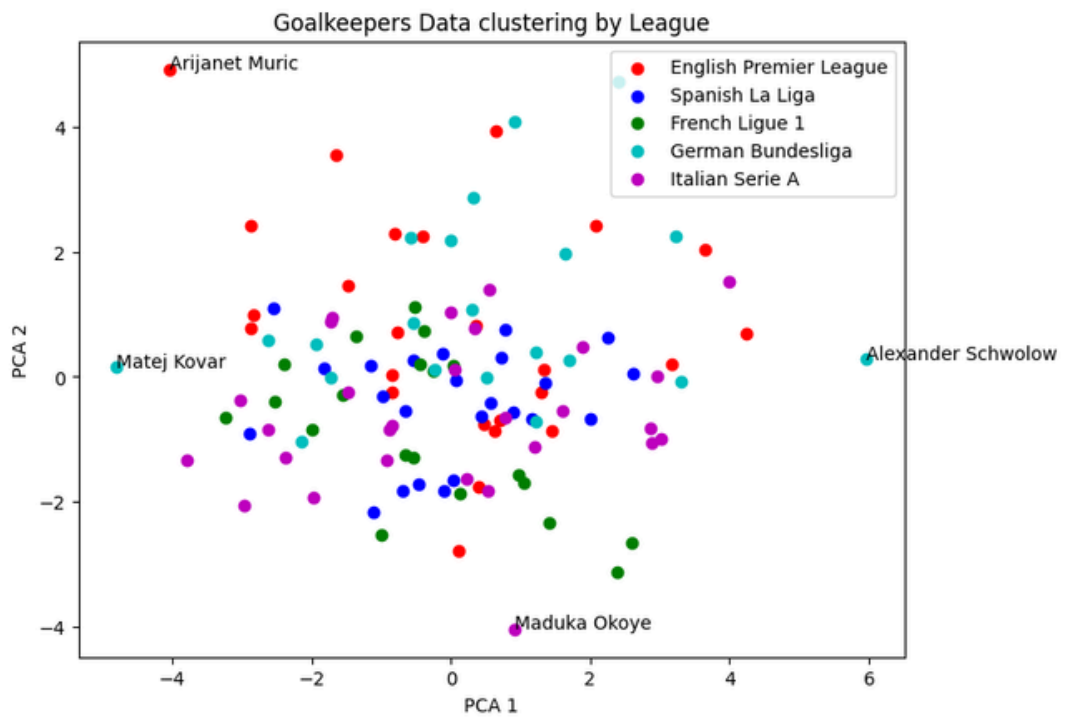
Regarding leagues distribution, it's easy to notice the **superiority** of **English Premier League** over other top four leagues, at least in terms of team value and show (goals): purple dots appear all in quite high on y-axis and shifted to the right in on x-axis.
Among top teams in *npG* there are **Bayern Munich**, leading with 4.89, followed by **Liverpool** and **Manchester City**; only other club with team value over €800m is Real Madrid with little over 3 in *Non-PenaltyGoals*.
It's important to state that npG is obtained summing up all players npG in a certain team: result obtained so doesn't match with team season goals, but with the sum of player goals per 90'. AC Milan, for example, has 3.14 *npG*, against only 2.00 goals per game scored in official matches, due to players like Luka Jovic or Noah Okafor, both with more than 0.6 *Non-PenaltyGoals* per 90' but, actually, only 6 goals scored in the whole year.

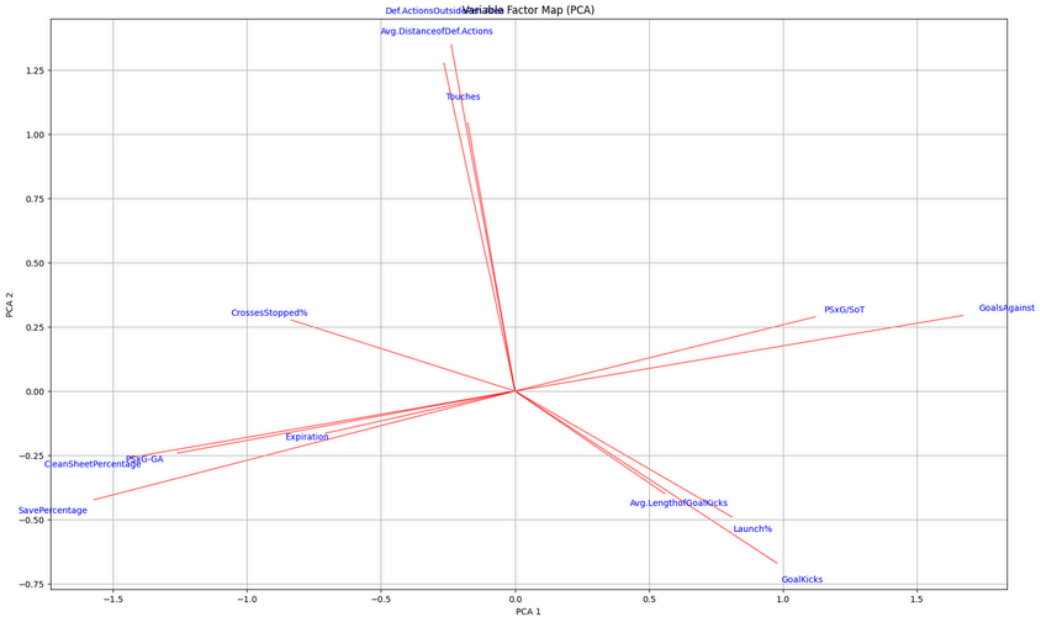
---

04. Clustering Analysis

Goalkeepers

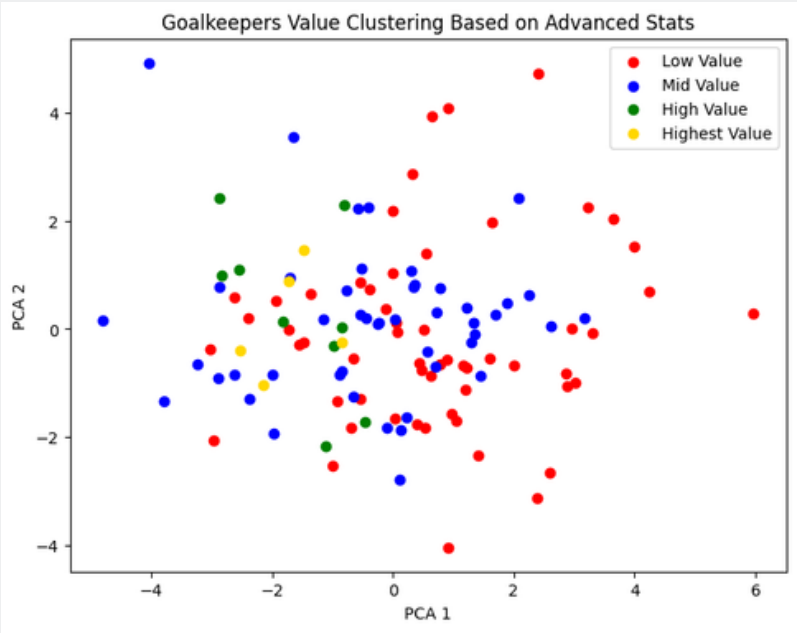
League Clustering



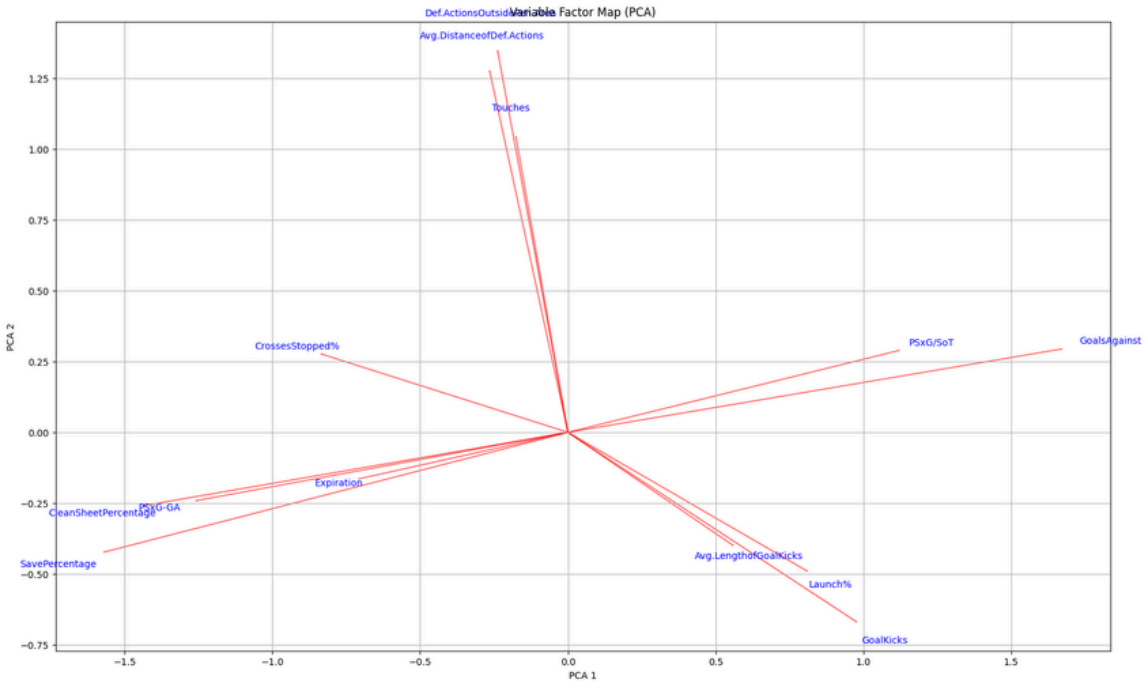


Taking a look at the goalkeeper league clustering, it seems quite impossible to precisely divide goalkeepers into clusters. Premier League goalkeepers are very scattered and do not provide much information, Ligue 1, Serie A and La Liga goalkeepers tend to be less involved in the build-up phase, touching the ball less often and staying closer to the goal. Ligue 1 goalkeepers, in particular, also tend to kick the ball more and further than players in other leagues.
On the other hand, **Bundesliga** goalkeepers and some **Premier League** goalkeepers (who play for teams that use the whole team in the build-up) tend to play **further away** from the goal, touch the ball more often, and kick the ball less often during goal kicks.

Value Clustering



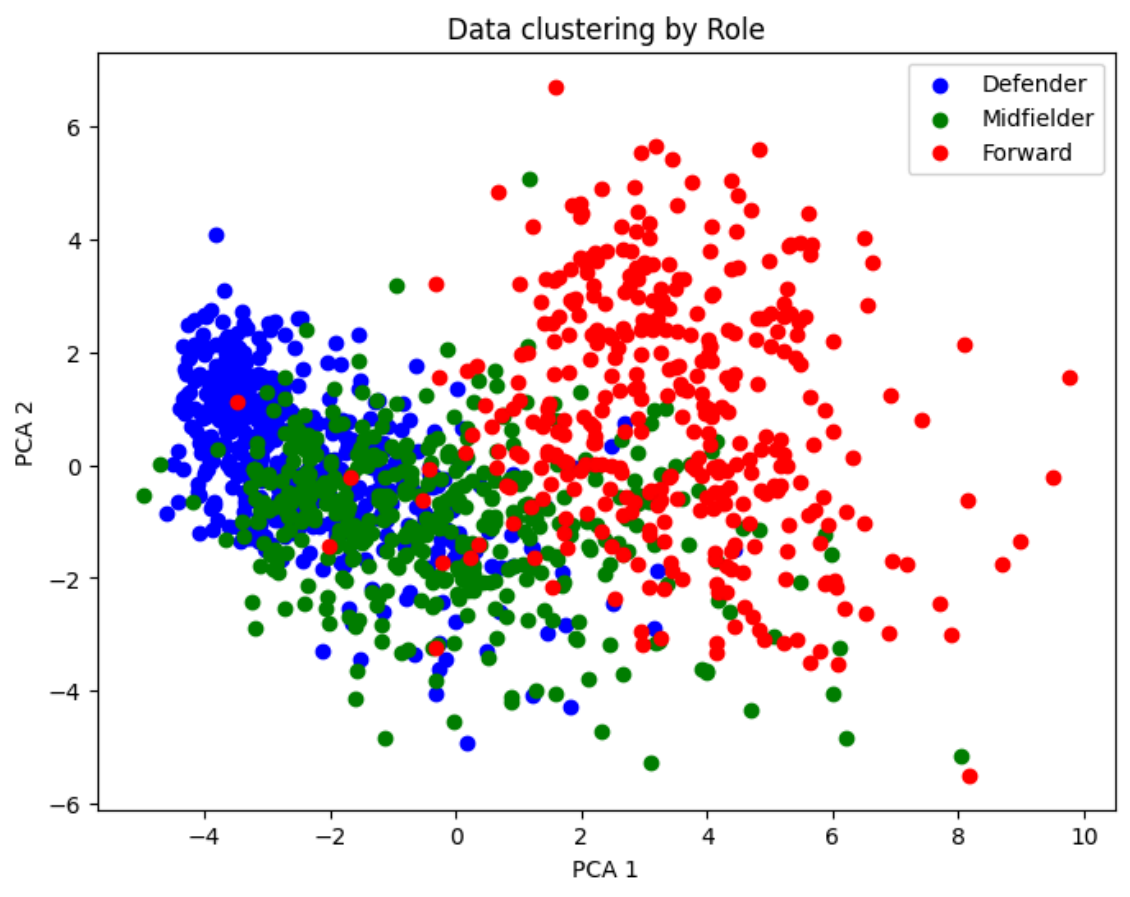
Then I move to goalkeepers value clustering, in order to identify most 'valuable' feature a goalkeeper can have. I consider 'Low Value' those players whose value is minor than median value (50% of goalkeepers worth less), 'Mid Value' those whose value is between 0.5 and 0.9 percentiles (keepers in top half in value, but out of top 10%) and 'High Value' top 10% of goalkeepers. I highlight in yellow most valuable 1% of goalkeepers in DataFrame.
Looking at the previous clustering plot, high value goalies appear on the left of the graph, but not at the edge, while majority of less valuable stay on the right.



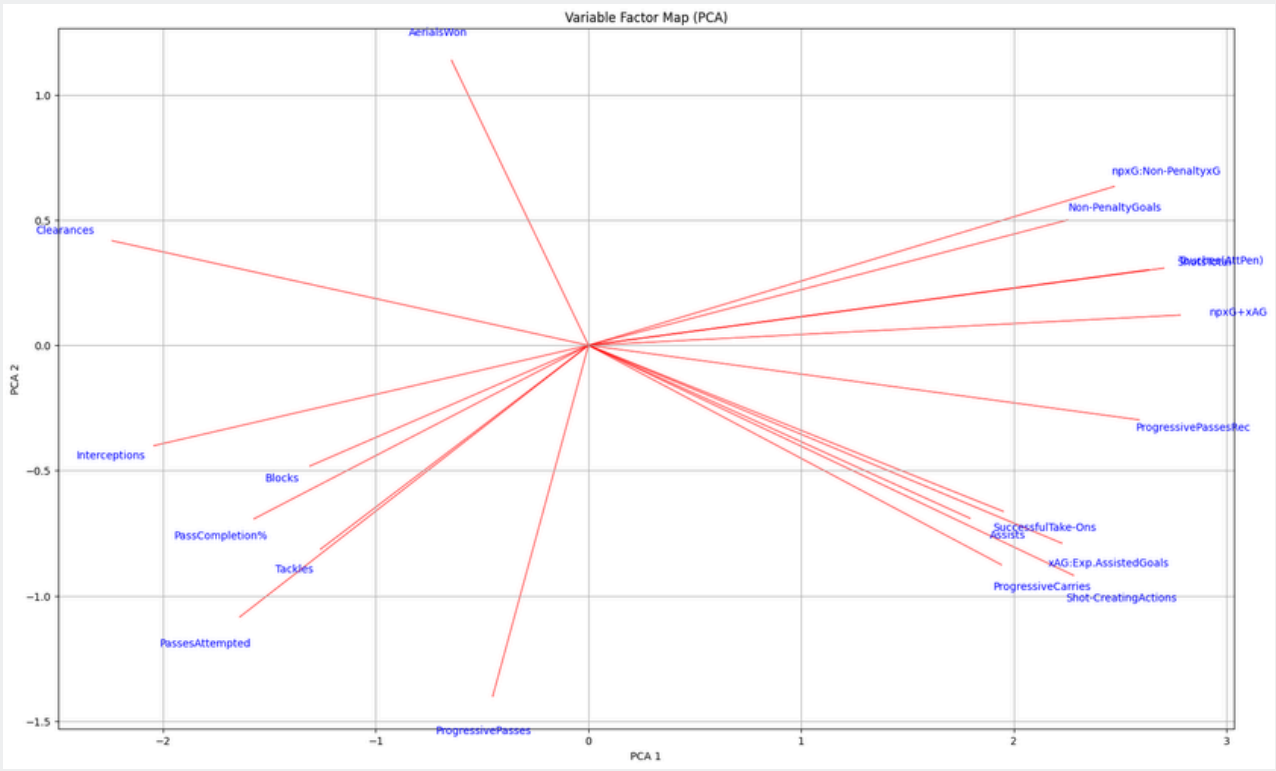
Most important features appear above in the Variable Factor Plot: obviously *GoalsAgainst* is a determinant attribute to discriminate best goalkeepers, while *SavePercentage* and *CleanSheetPercentage* appear to be important to raise keepers value.

Movement Players

Role Clustering

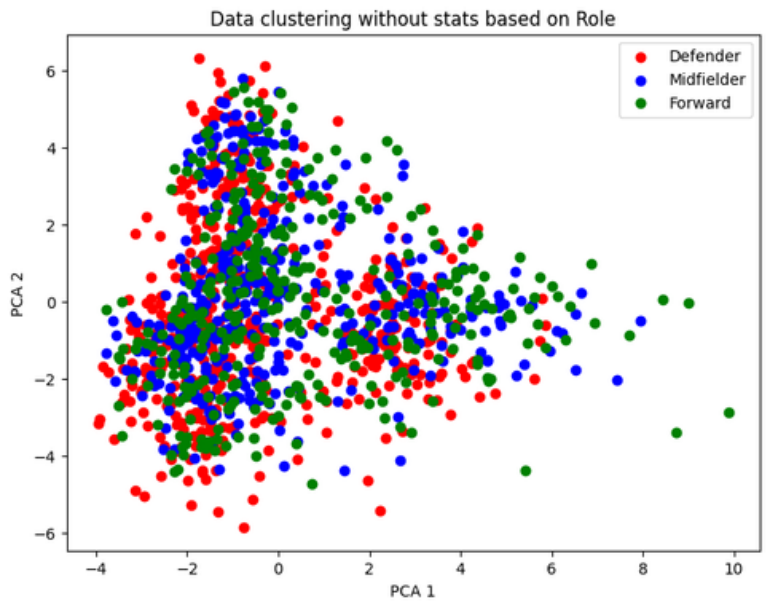


Here I first removed demographic/contractual variables, in order to have only statistics to deal with. I reduced dimensionality using PCA, to have only 2 axis and to be able to depict all informations. As visible, it's quite easy to visualize the three clusters. defenders on the upper-left part, midfielder in the lower part and forwards in the upper-right part of the plot. Forwards look quite divided from the rest of the dots, while defenders and midfielders don't have a sharp separation. Of course, then, there are some dots in other clusters, due to the their distinct nature. For example, Lille's Andrej Ilic, due to his limited play time, has *Clearances* value of 3.05, way higher than forwards 0.7 mean. This, of course, makes him appear more as a defender than as a forward. Similarly, **Federico Dimarco,** left-back of Inter Milan, has 0.5 in *npxG+xAG*, which is 5 times higher than 0.1 mean value among defenders: this, instead, determines Dimarco's **offensive ability** as wing-back and correctly places him closer to wingers than to defenders.

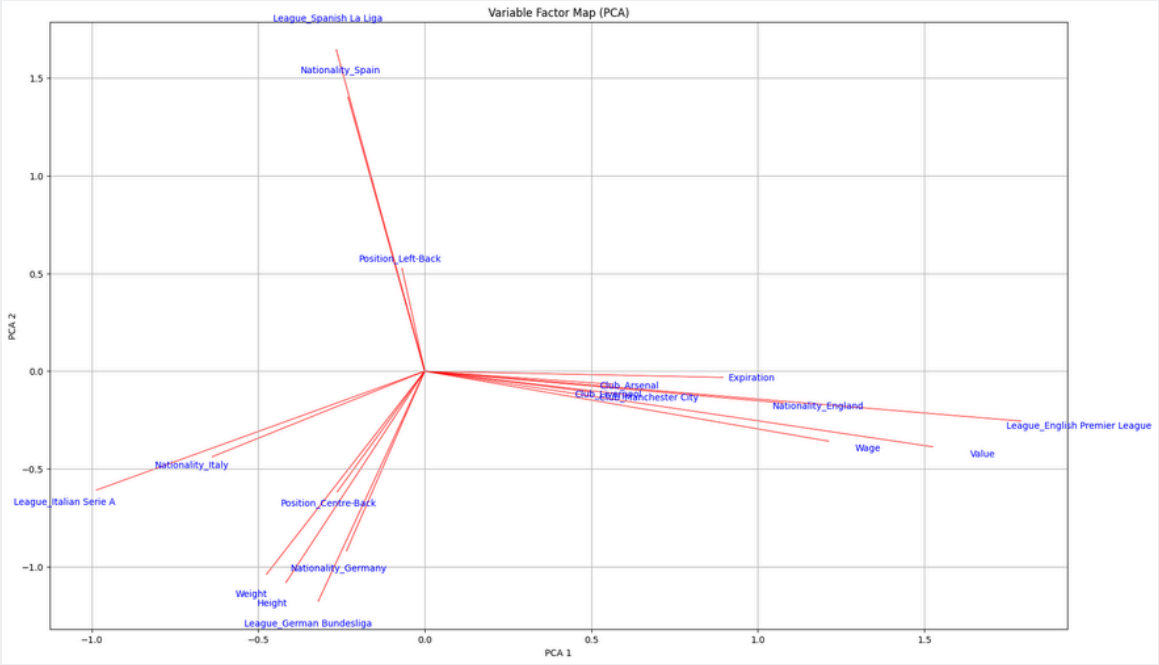


As appears from the Variable Factor Plot above, **roles** are correctly **divided** by values in characteristic statistics, such as *Clearances* and *Interceptions* for defenders, *Non-PenaltyGoals, Touches* and *ShotsTotal* for forwards. Most determining variables for midfielders is *ProgressivePasses*, even if they look covering the gap between defenders and forwards as they do on the field.

I proceed my investigation creating again PCA and KMeans clustering but relying only on demographical and contractual informations, instead of advanced stats like before.

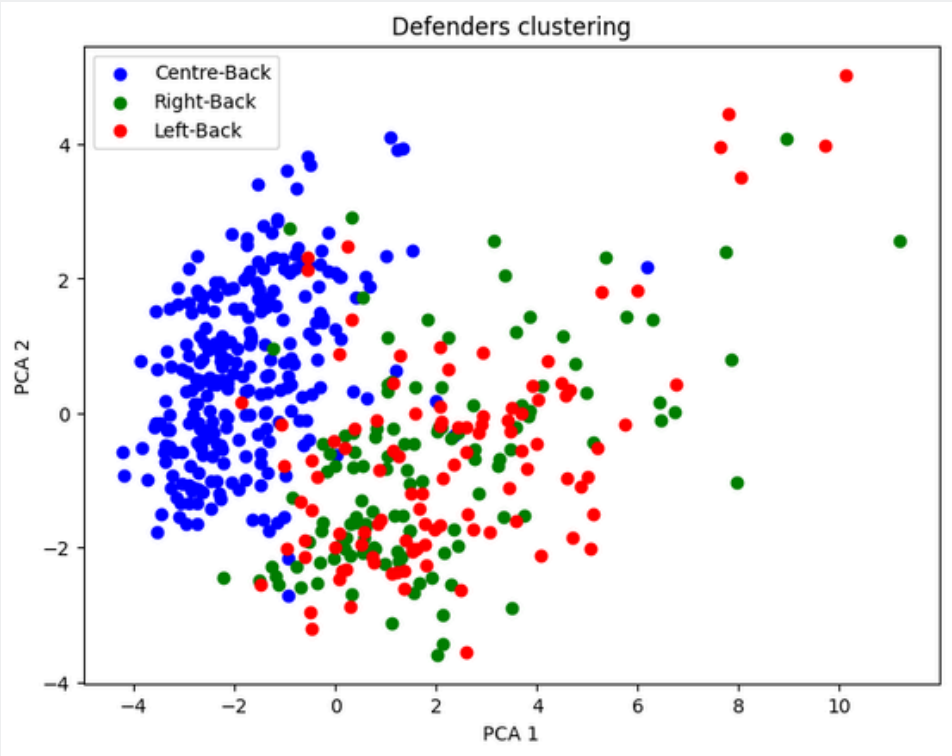


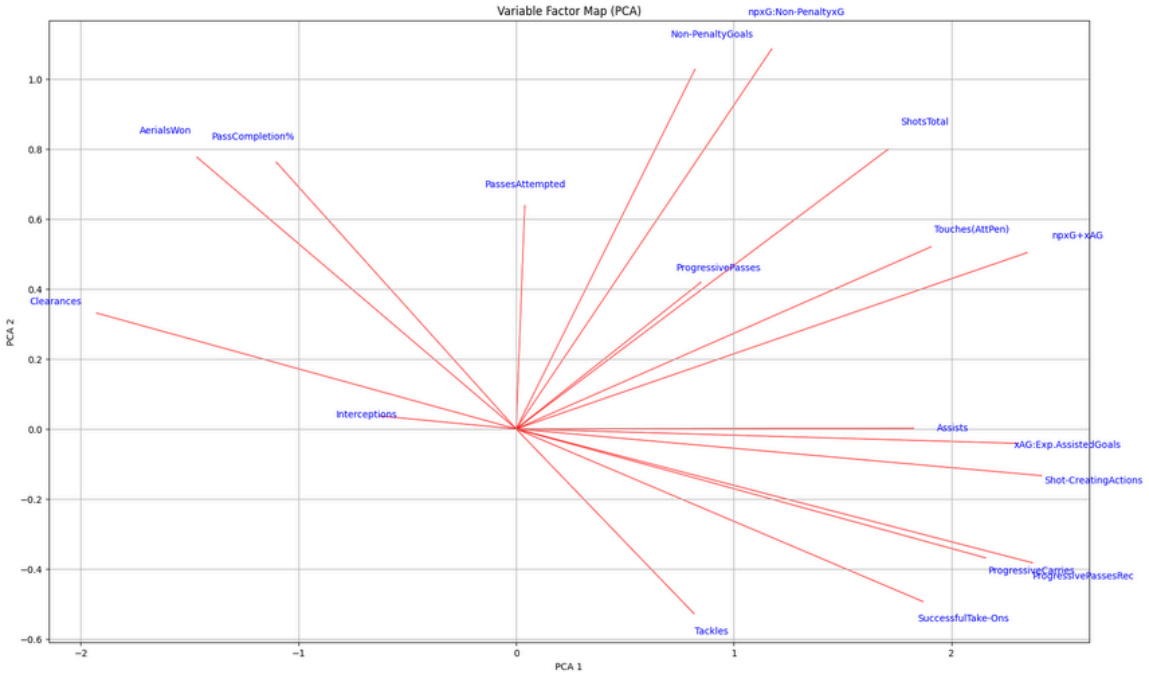
As easily guessed, clustering don't work out for the intended purpose. There are no defined clusters, but, instead, a huge cloud of points of different colors.



Here I post the Variable Factor Map of bidimensional PCA done before. Its utility, after KMeans clustering results obtained above, is questionable but, for the sake of completeness I provide it.

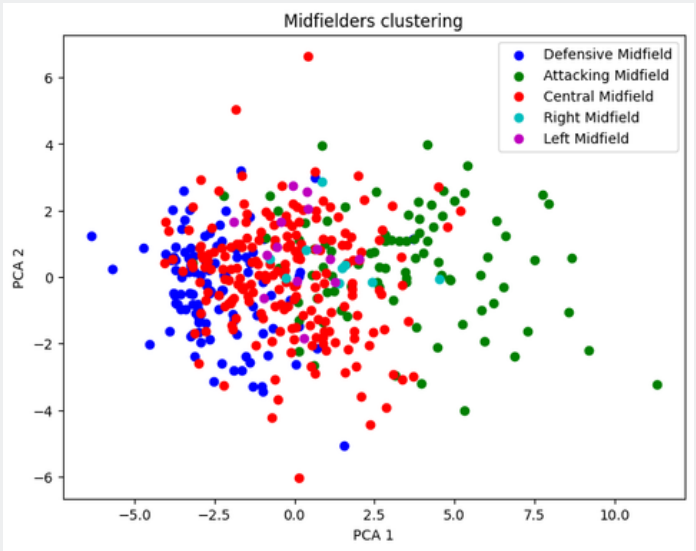
Defenders Clustering



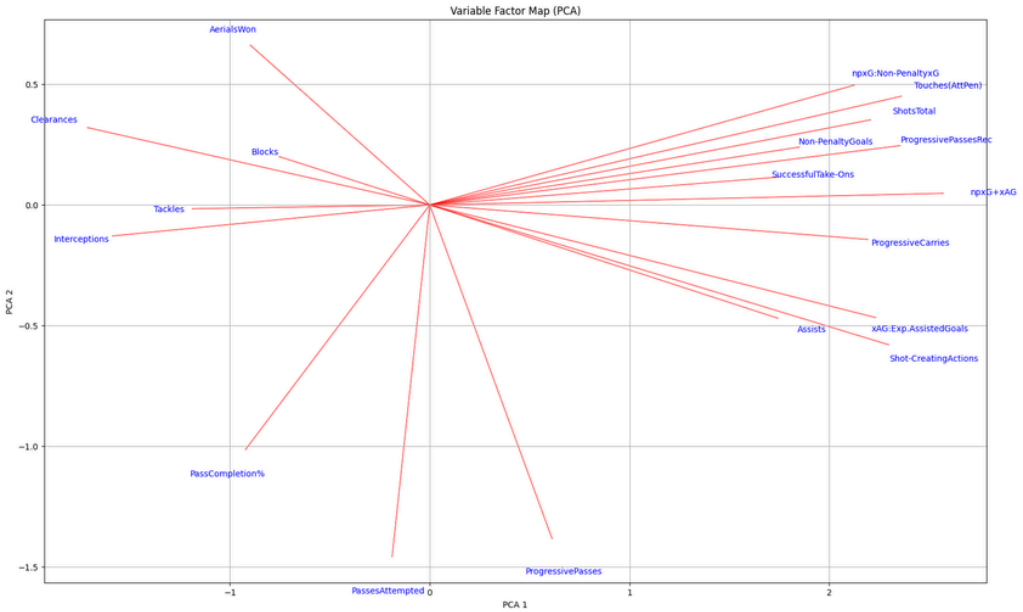


Defenders clustering look doing its task perfectly, i**dentifying** precisely **two different clusters,** one regarding **centre-backs**, the blue one on the left, and a cloud of green and red dots on the right, referring to **full-backs.** Looking at Variable Factor Map I can detect main variables separating the two roles: *Clearances, AerialsWon* and *PassCompletion%* are the most describing ones for centre-backs, while *xAG, Shot-CreatingActions* and *ProgressivePassesRec* describe full-backs the most.
Then, a mini-cluster inside full-backs can be identified, probably composed by really offensive wing-backs, as *npxG, npG* and *ShotsTotal* are the most determining attributes. Players like Inter Milan's Federico Dimarco or Lyon's Clinton Mata.

Midfielders Clustering

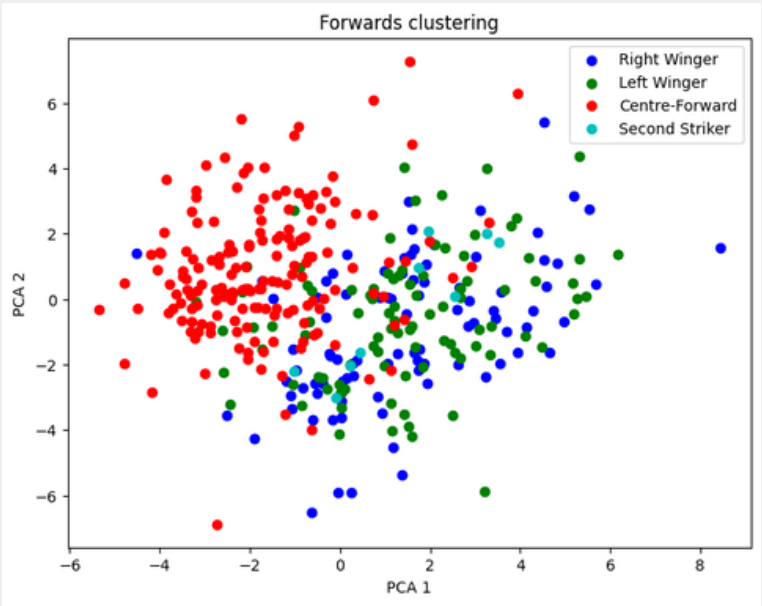


Midfielders clustering, similarly, identifies **three main clusters**, but they don't clearly separate like in defenders plot. This graph, instead, looks much like role clustering plot, with more defensive players (there defenders, here defensive midfielders) on the left and more offensive (there forwards, here attacking midfielders) on the right.

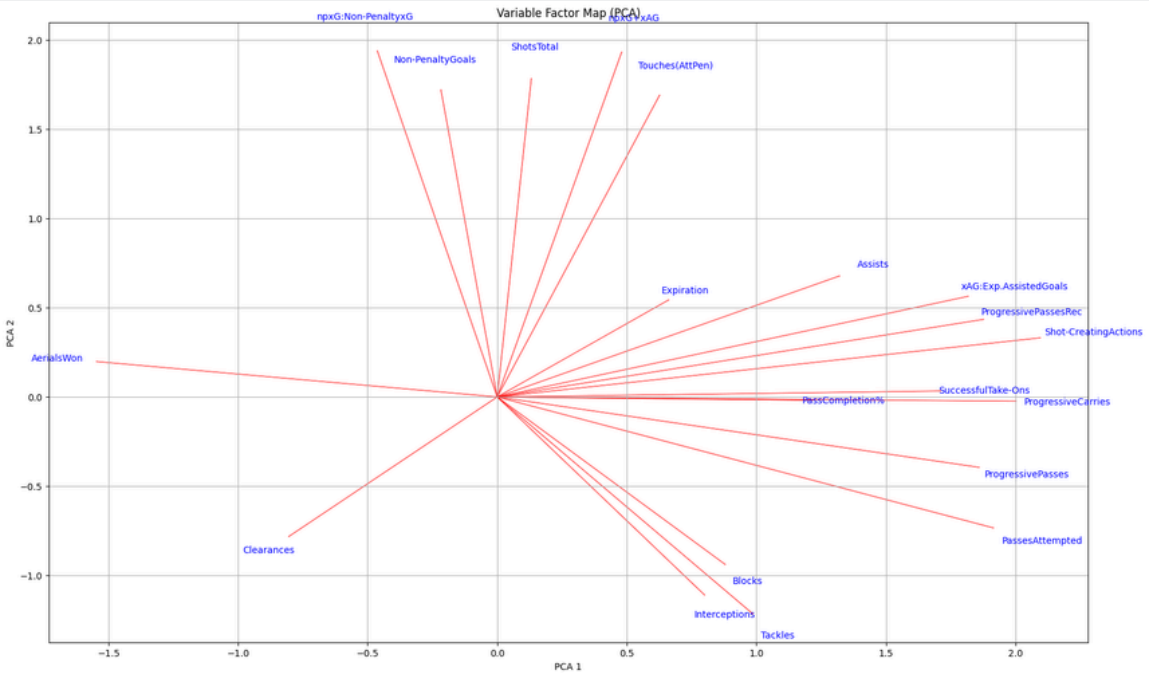


Most characterising variables are, of course, *Clearances* and *Interceptions* for left dots (defensive midfielders) and *Touches(AttPen)*, *npxG+xAG* and *Shot-CreatingActions* for dots on the right (attacking midfielders). Central, right and left midfielders cover the space in the middle, as they usually have a mix of both defensive and offensive qualities.

Forwards Clustering

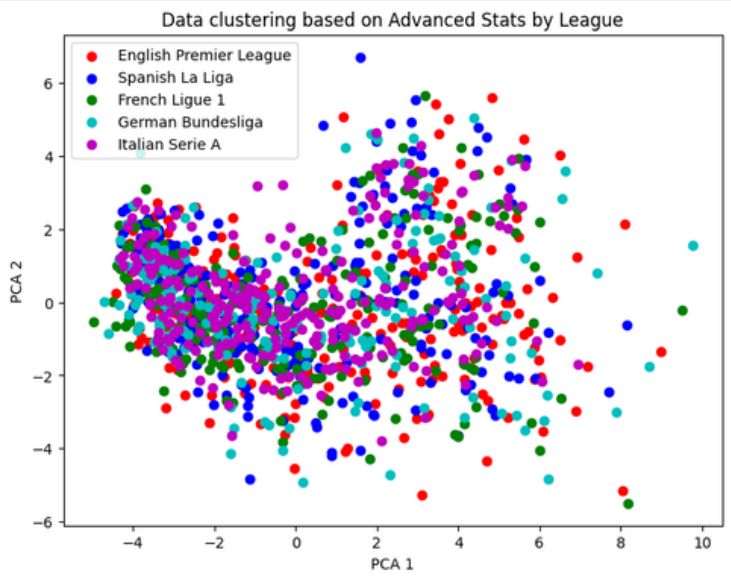


Forwards, lastly, **separate** the data in **two clusters** quite different. On the left part of the figure there are red dots, belonging to centre-forwards. The remaining part of the graph is made by scattered dots, due to the variety of quality second strikers and wingers have.

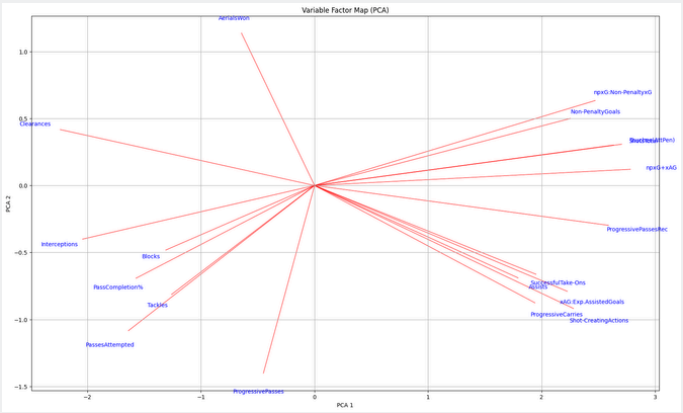


From the Variable Factor Plot above, main differences between centre-forwards and the rest of forwards lie in team play and building-up partecipation: *PassesAttempted*, *ProgressiveCarries* and *ProgressivePassesRec* are most proper variables of second-strikers and wingers, while *AerialsWon* and *npxG* are centre-forwards' ones. These results mean that **centre-forwards** tend to **play** a lot of **head games** and **banks** and are, by far, teams main **scorers**. On the other hand, **second strikers** and **wingers**, tend to **keep** the ball way more, to **pass** and **carry** the ball more, to make more **assists** and **take-ons**.

League Clustering



I investigate relation between players and leagues, after correlation, using PCA and KMeans clustering but, as visible above, the result is not that useful to understand much. Dots cloud is scattered along the whole graph and doesn't follow any relationship.

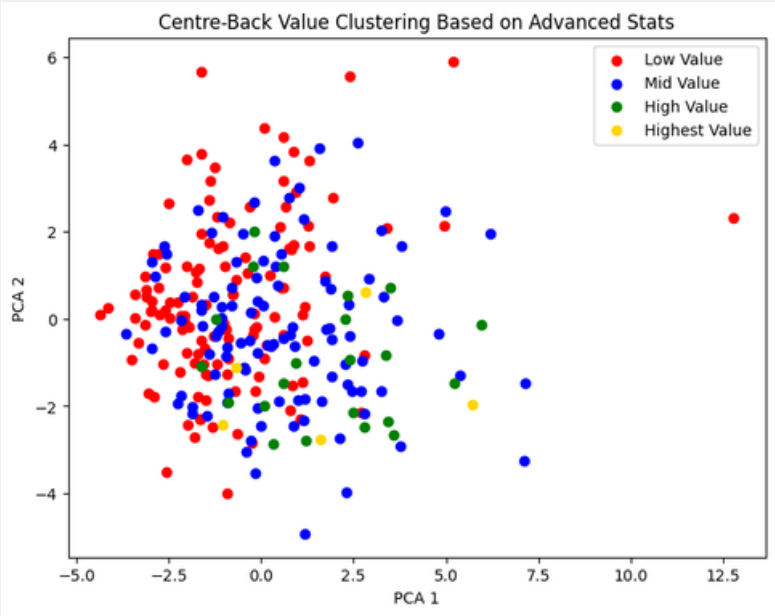


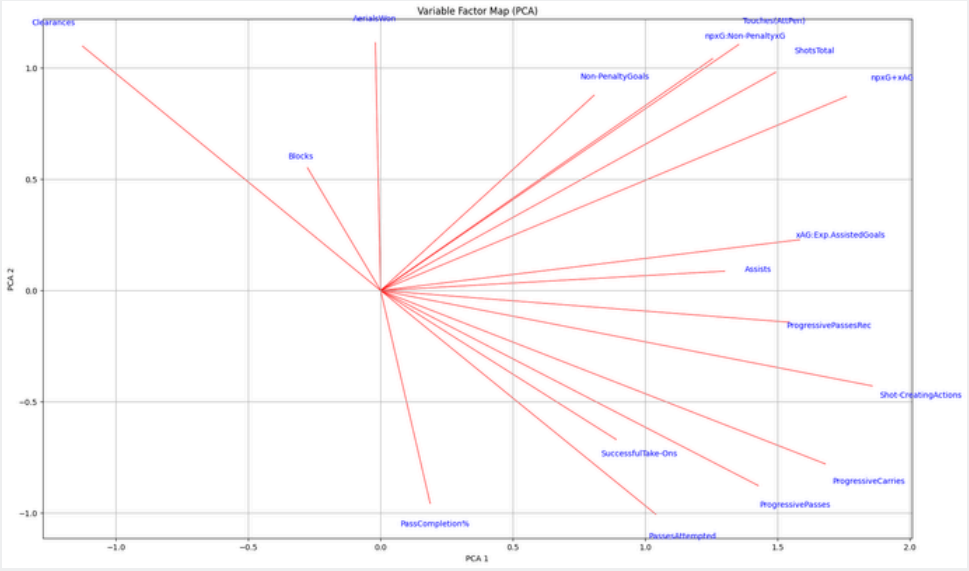
Likewise previously, I list Variable Factor Map in order to comprehend better the figure above.

Value Clustering

Now I do PCA and then KMeans for each position in players' DataFrame and, for each position, I highlight the points using four colors, associated to their own value: I use red for players whose value is under 50th percentile (low), blue for players whose value is between 50th and 90th percentile (medium) and red for top 10% of players (high value). I use yellow for top 1% (highest value). Left-midfielders, right-midfielders and second strikers aren't dealt due to limited size of their DataFrame

Centre-Backs

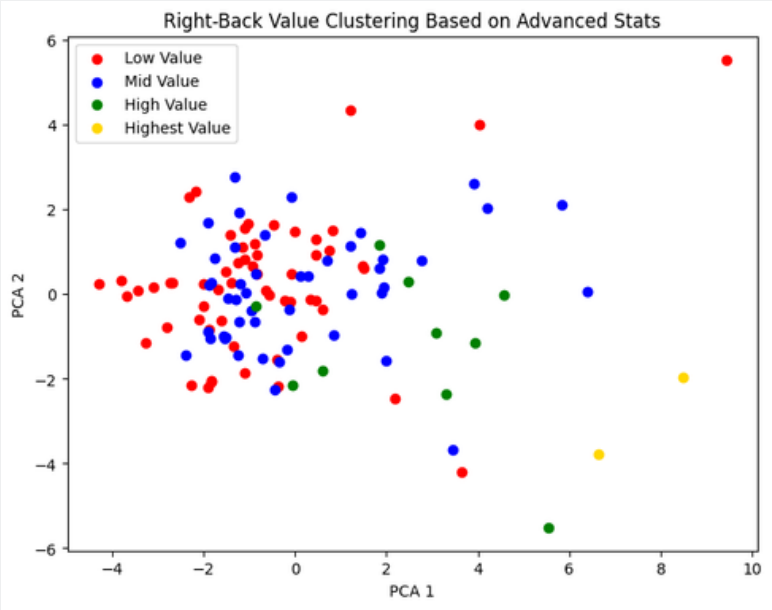


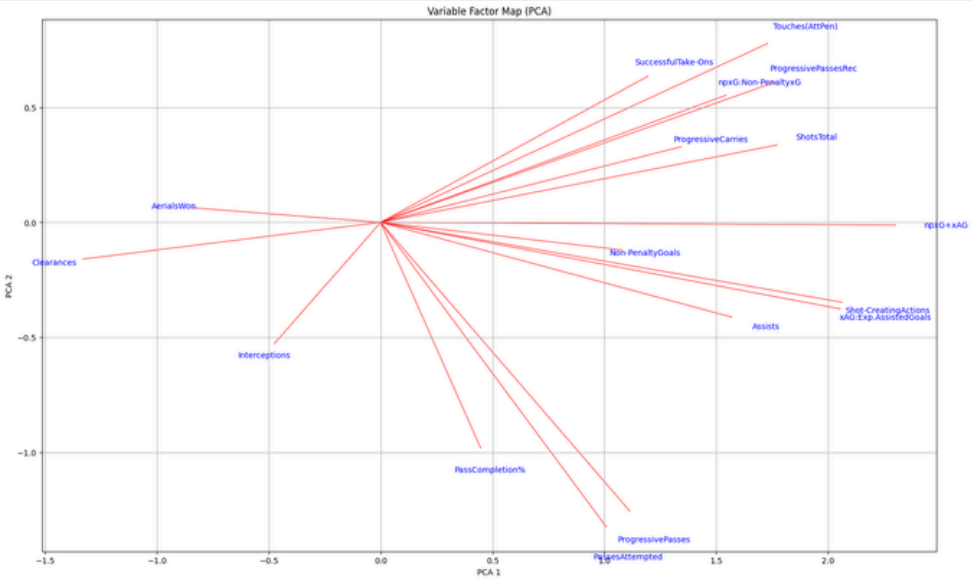


Centre-backs don't distribute in precise clusters, but, following the graph above, centre-backs on the left look less worth, while players with high value stay in the bottom right. Looking at the Variable Factor

Plot, players able to **pass the ball accurately** and to make **assists**, on average, **worth more**, but the presence of some yellow dots in the middle of the cloud witness that centre-back value doesn't depend on only few qualities, but on a **complex combination** of many.

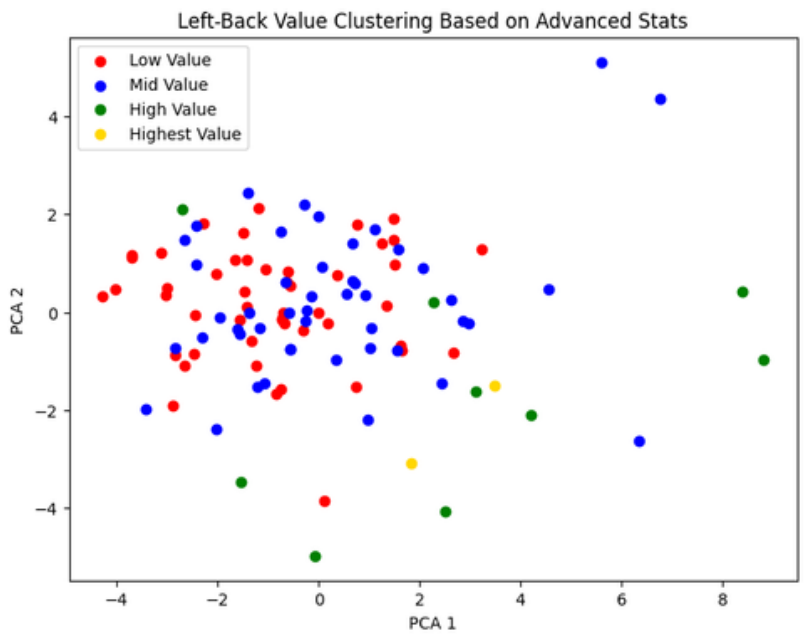
Right-Backs

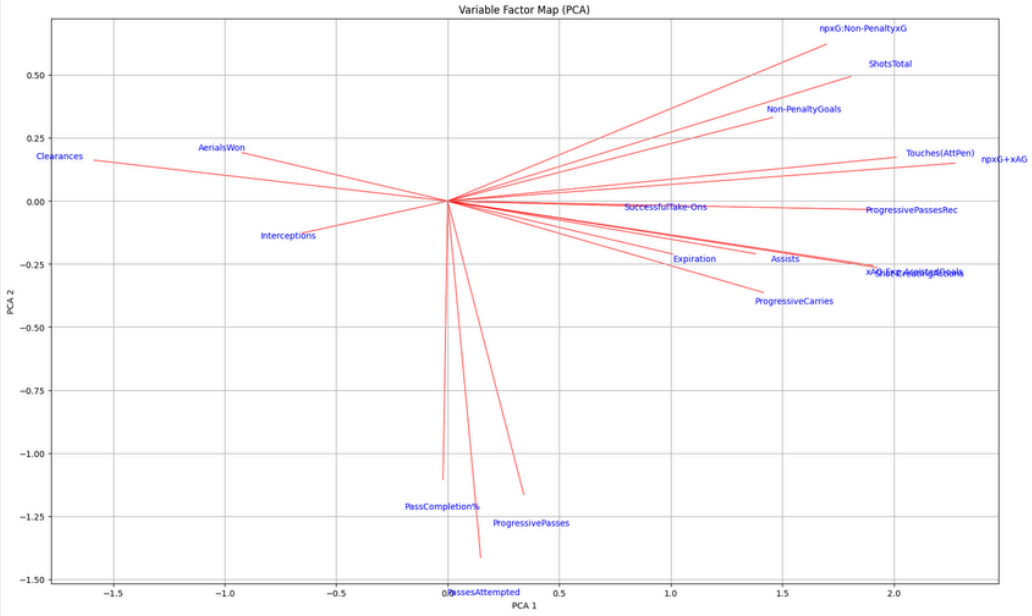




Right-backs look being worth worthier and worthier going low-right in the figure. Except for a red dot, who should be Pascal Stenzel of Stuttgart, all worthiest players belong to the bottom right-hand corner. Most determining variables are ***ProgressivePasses, PassesAttempted, xAG*** and ***ShotCreatingActions***.

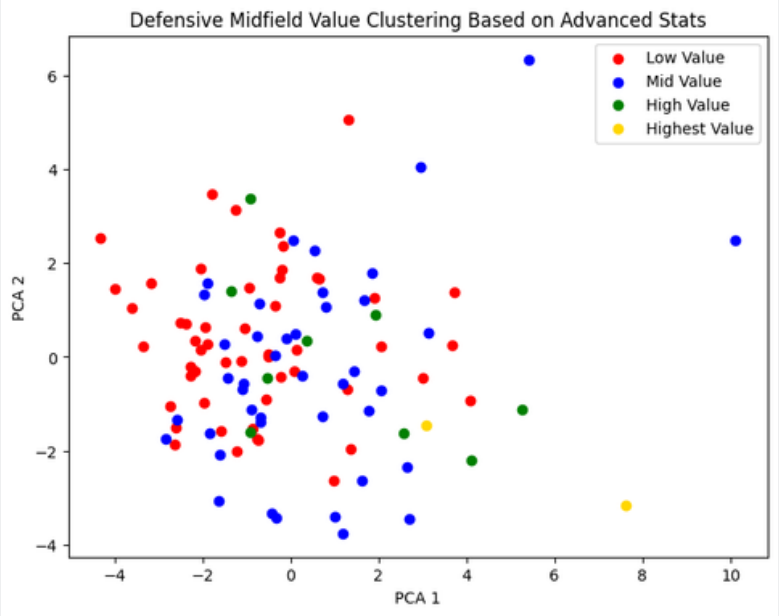
Left-Backs





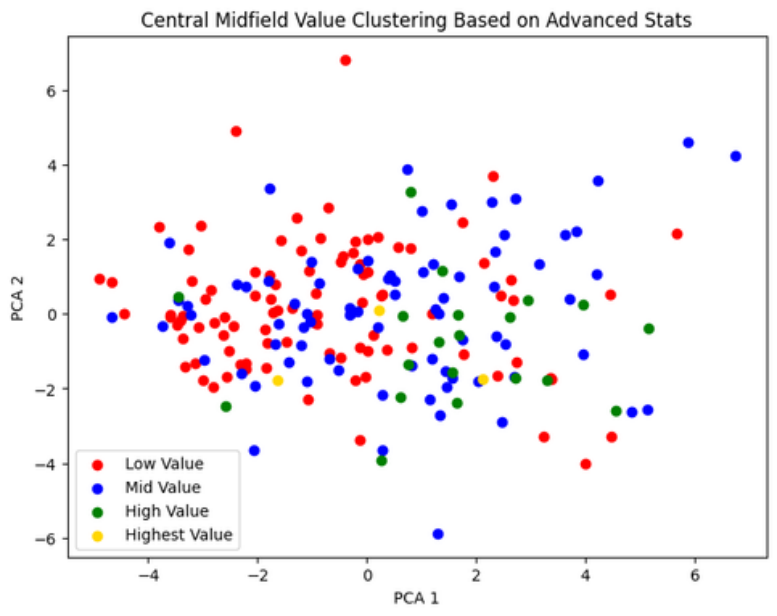
Left-Backs clustering follows the same structure as right-backs, having highest value players near the low corner on the right. Most important columns, of course, are the same as before: the identity of distribution among full-backs can be detected from defenders KMeans cluster listed above, where left-backs and right-backs blend uniformly.

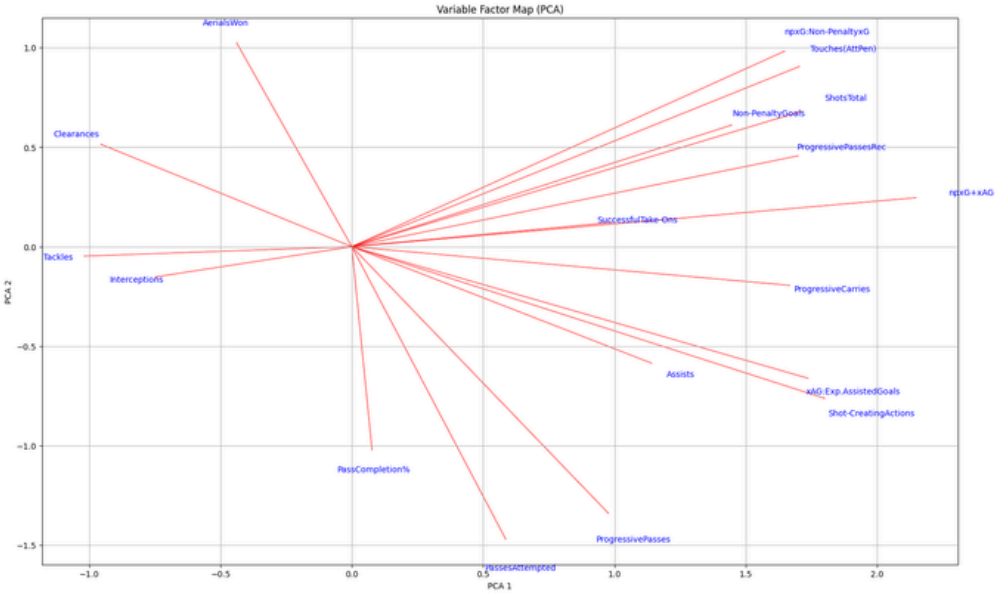
Defensive Midfielders



The defensive midfielder, on the other hand, doesn't look very clustered, and none of the information we can glean from the graph above looks very reliable. We can hypothesize a direction of **increasing value** from the **left** to the **lower right** part of the figure. This is due to the fact that most of the red dots are on the left side, but there are some green dots in the middle., like Real Madrid's Aurelien Tchouameni and Everton's Amadou Onana. Overall, I can state that among defensive midfielders qualities in variables like ***Shot-CreatingActions***, ***ProgressiveCarries*** or ***ProgressivePasses*** are really **appreciated** and take players to the next level in term of value. Then there are some players whose defense is so good that their value is justified quite just by it.

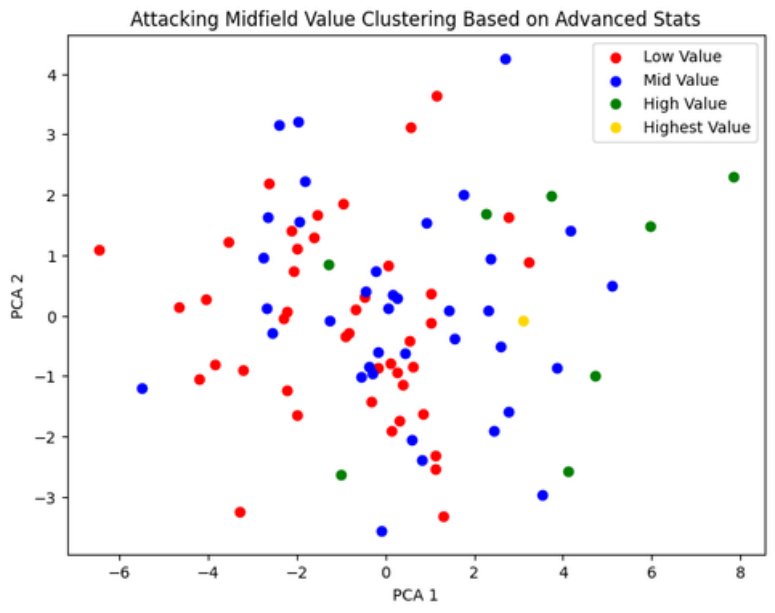
Central Midfielders

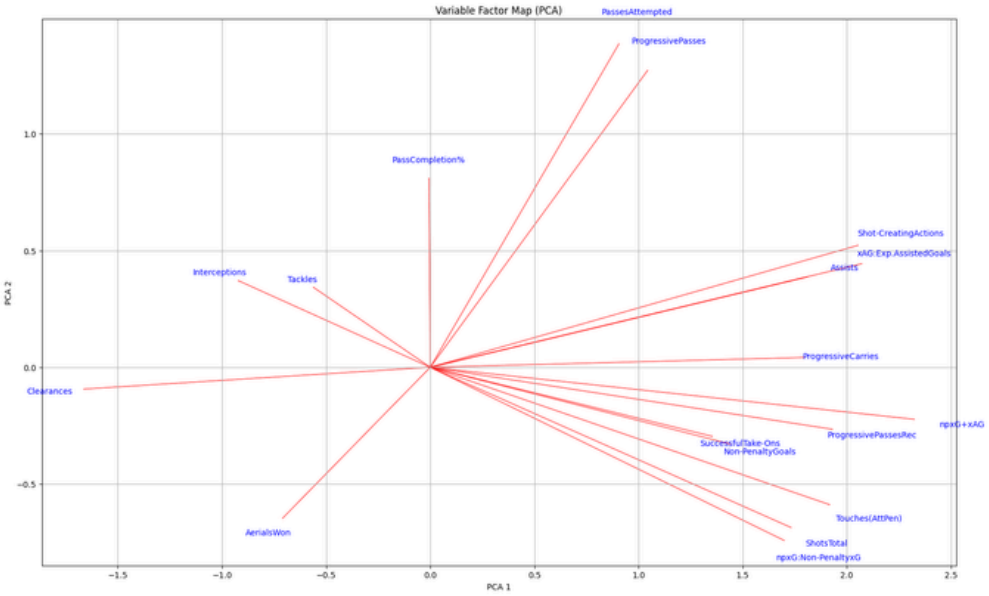




I follow on with central midfielders, whose division too doesn't look that clear. Red dots appear mostly on the left side of the figure, then blue points outline the cloud of points. Green points are placed mostly at the bottom-right part of the cloud, but, as yellow points witness, there isn't any variable which drive the players value growth. Probably there are **many central midfielders** with **better defensive qualities** and, so, lower value, while the **more offensive** the players get, the **more valuable** they become. Elite midfielders owe their value to their completeness, as their central position in PCA shows.

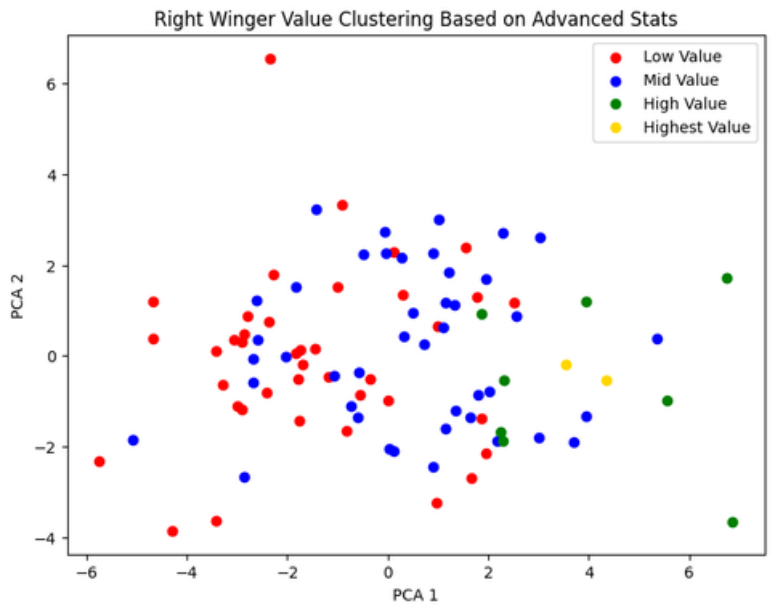
Attacking Midfielders

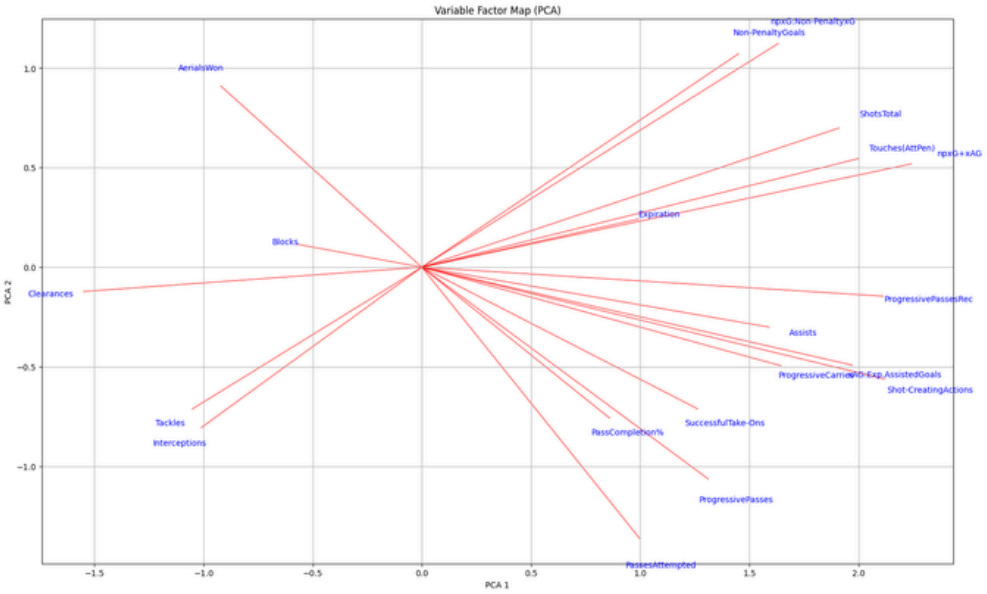




Similarly to previous position, attacking midfielders don't divide in clusters easily. Looking at colors in KMeans plot, green dots, except few, are all located on the right side of the figure, giving the most relevance to *npxG+xAG, Shot-CreatingActions* and *xAG* attributes. Left high value players away, it's hard to detect differences between low and medium value players as they are completely blended.

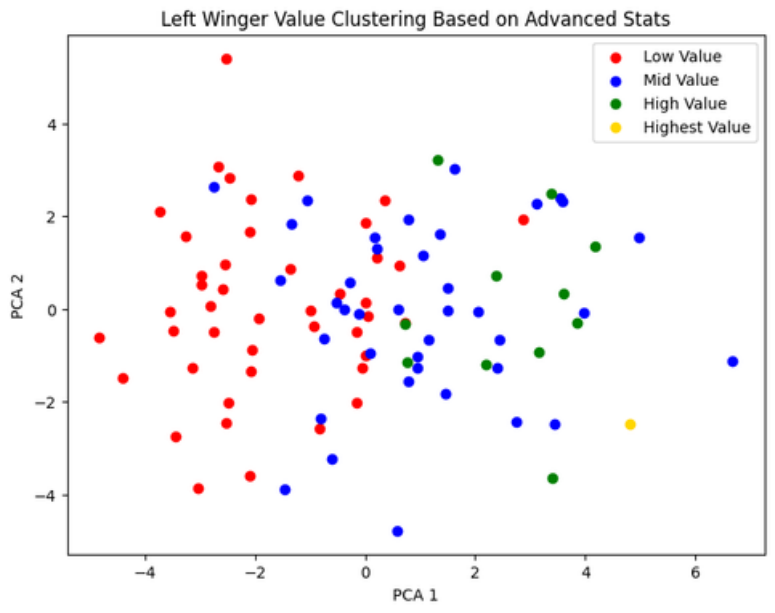
Right-Wingers

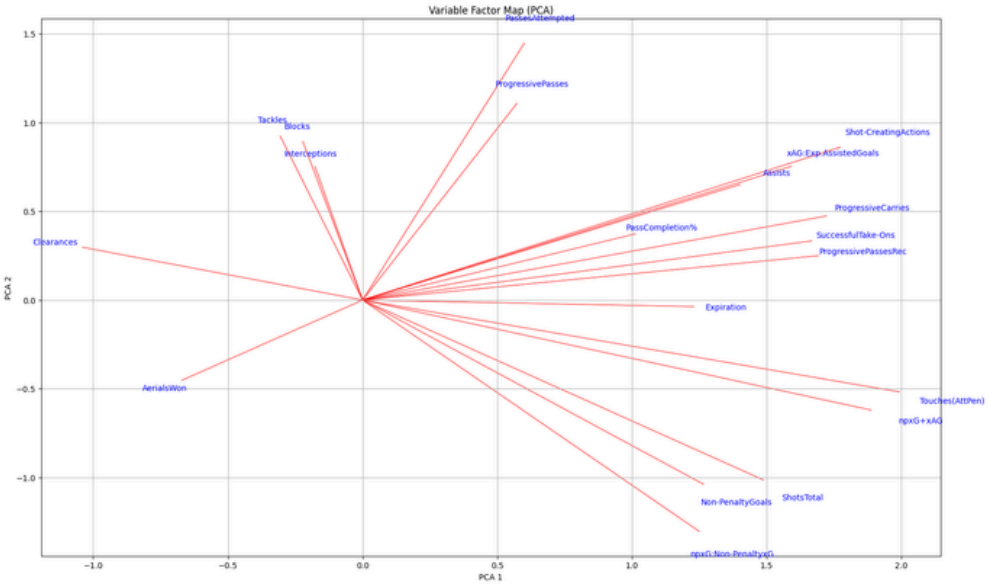




Right wingers, instead, appear quite **precisely divided**: the leftmost area contains the major part of low value left-wingers, while medium value players stay in the middle. On the right there are quite only worth wingers and top 1% in DataFrame. Taking in consideration this analysis, **defense** looks a **conviction** for left-wingers: defensive variables like *Clearances*, *Blocks*, *Tackles* shift the dots to the left the most, while ***ProgressivePassesRec***, ***npxG+xAG*** and ***Shot-CreatingActions*** contribute the most to top players.

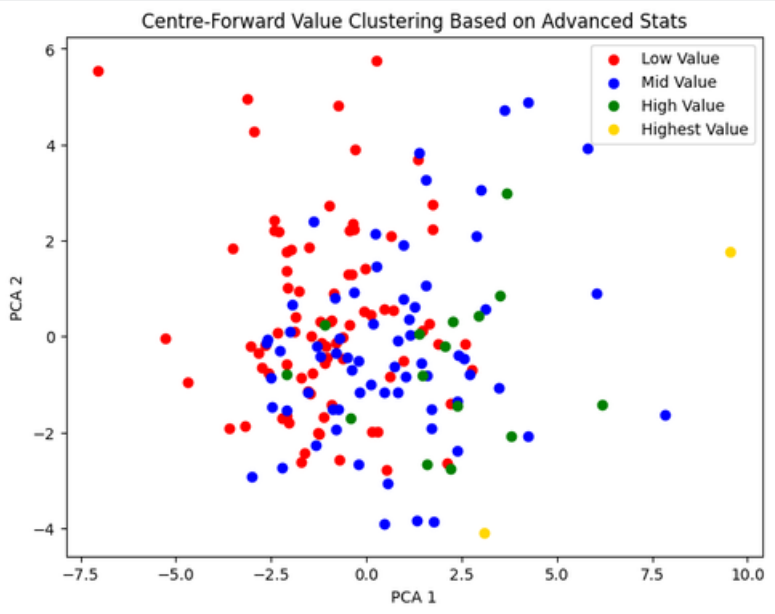
Left-Wingers

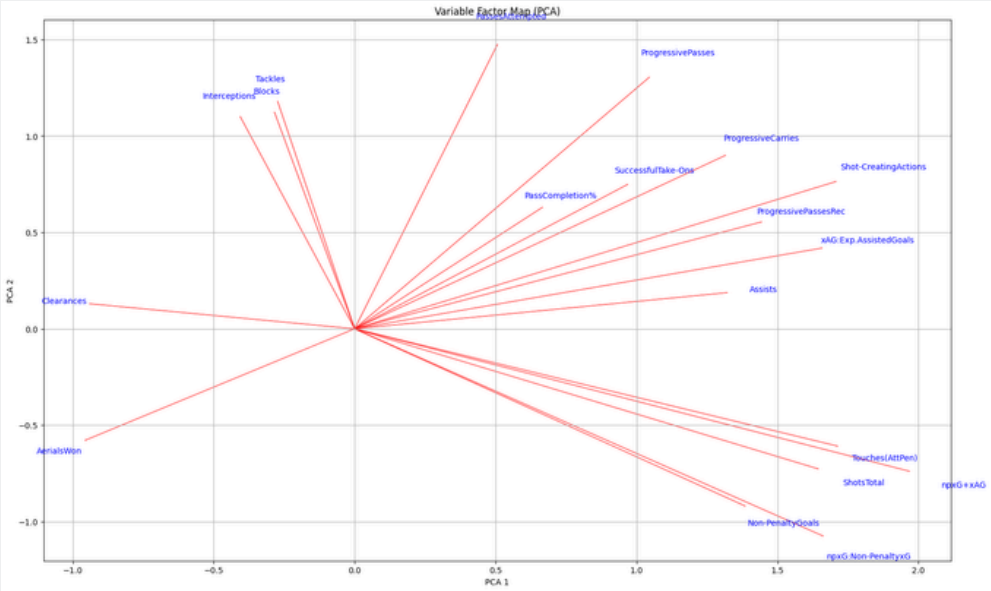




Likewise right-backs with left-backs, right-wingers share quite the same PCA and KMeans with left-wingers. Clusters look really similar and decisive variables stay more or less the same.

Centre-Forwards



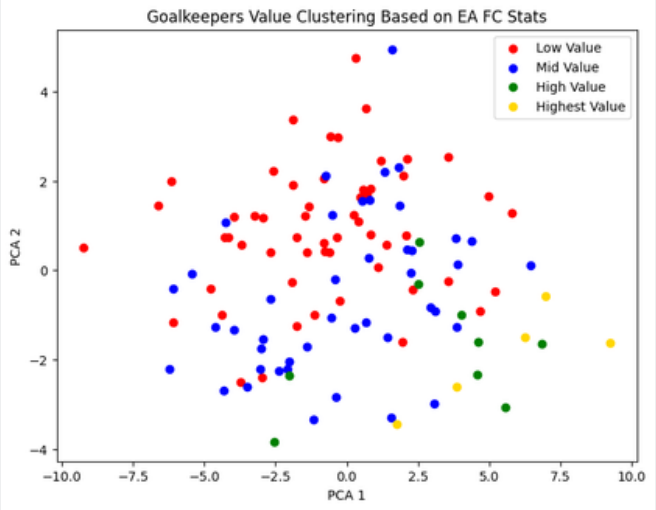


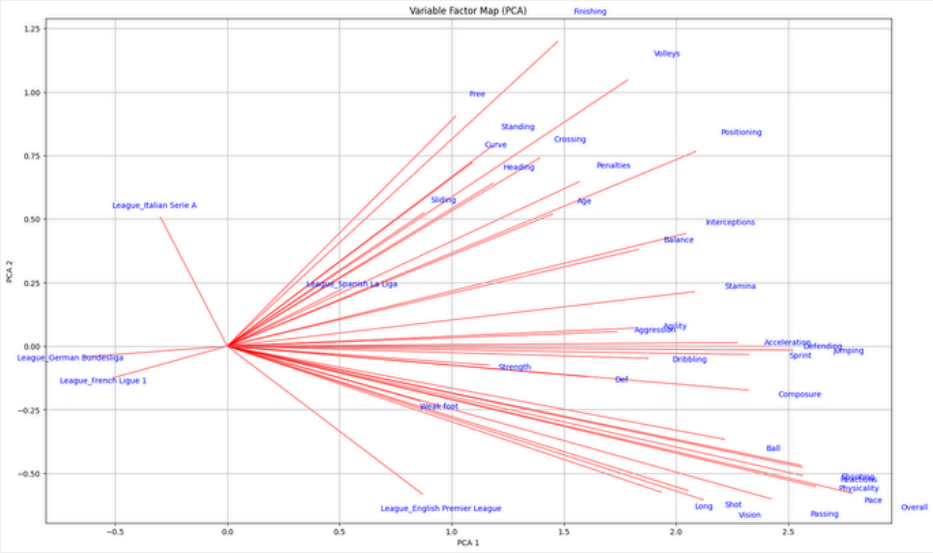
Centre-forwards, lastly, don't have a clear separation in clusters, but appear to be two poles. The upper left corner seems to house less worth strikers, while lower right area should belong to best centre-forwards. In reality this is not 100% valid, since top 2 strikers are placed one in the middle of x-axis, at the bottom edge of the figure, and the other in the middle of y-axis, at the rightmost edge. Variables determining worth players seem to be ***Touches, ShotsTotal, npxG*** and ***npG,*** while less valuable centre-forwards have higher values in *Clearances, Interceptions* and *Blocks.* These facts state that centre-forwards are still **valued** on **how much they score**, more than anything else.

EA Sports FC 24 Attributes Clustering

I end cluster analysis doing same thing I did before, highlighting value scales, but taking into account not advanced stats anymore, but only EA FC 24 attributes. I will do attach results for four main roles (goalkeepers, defenders, midfielders and forwards), but on the script there are plots of single positions too.

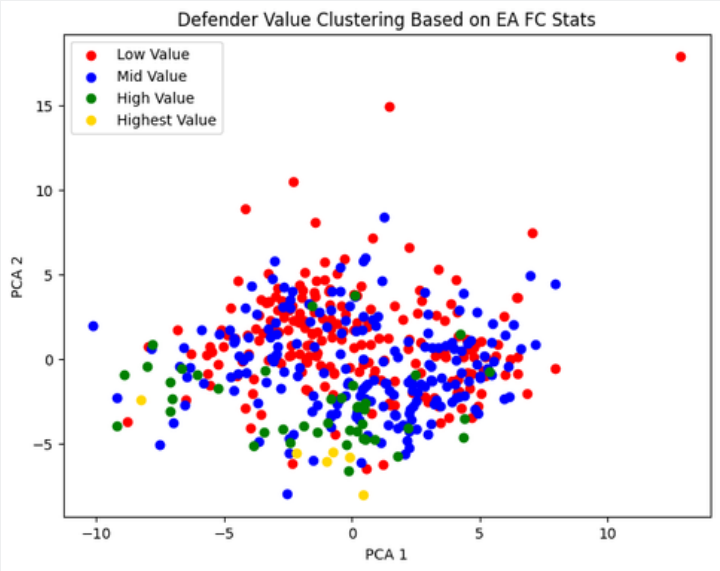
Goalkeepers

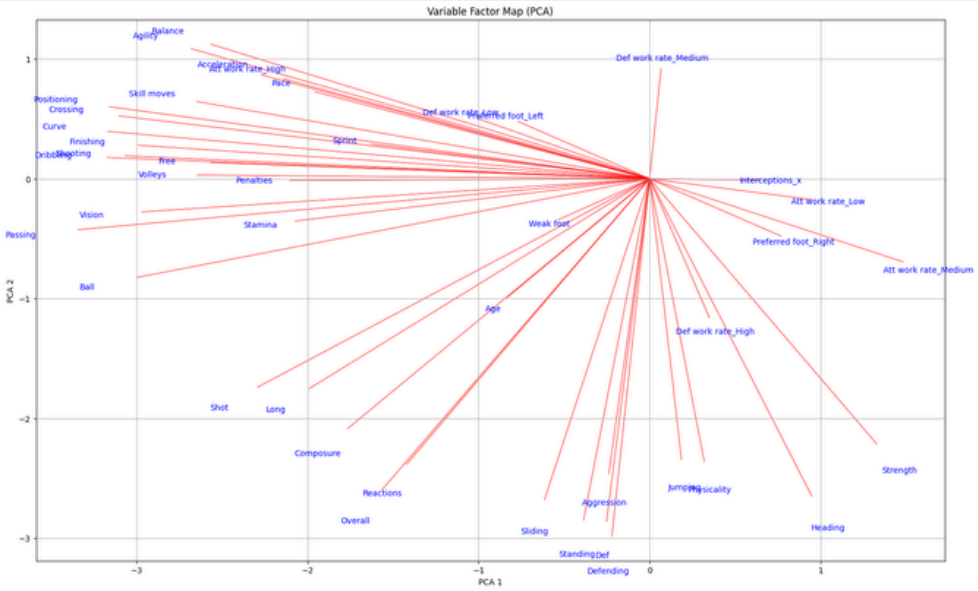




Goalkeepers value in scatter plot above looks **growing** in the **down-right** direction, since, following an imaginary arrow, at first there are red dots, then blue and finally green and yellow points. Main determining attributes look to be *Overall*, *Pace*, *Physicality* and *Passing*, while *Finishing* seems to be least important feature.

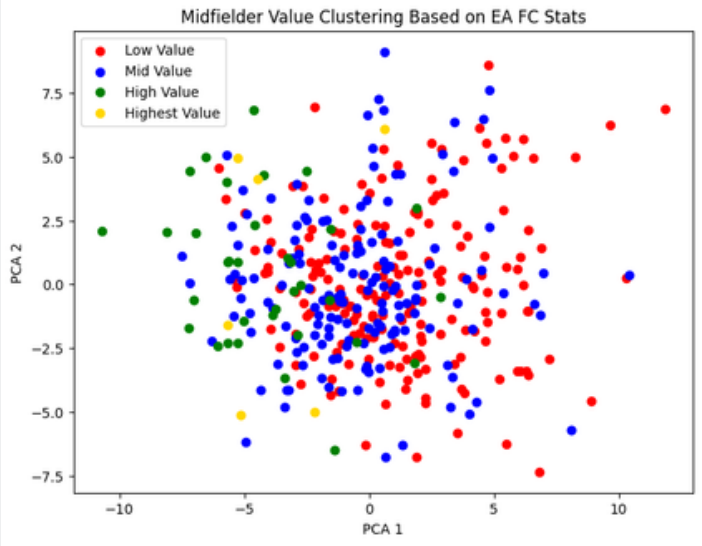
Defenders

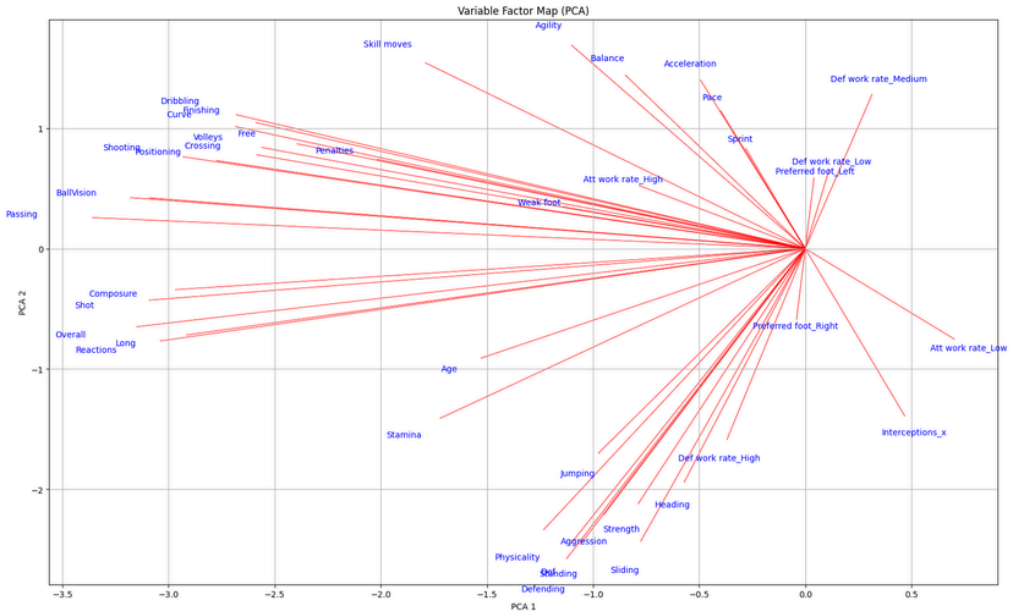




Defenders are less divided than goalkeepers and it's harder to identify precisely attributes related with value. Green and yellow dots belong to the lowest part of the figure, while some of the red points stay in the top half. Looking at Variable Factor Plot, of course ***Overall*** and ***Defending*** seem to be **most important** features to raise value followed by *StandingDef, Sliding* and *Reactions*. On the other hand, 'Medium' category in *Def work rate* appears to be the most common characteristic among less worth defenders.

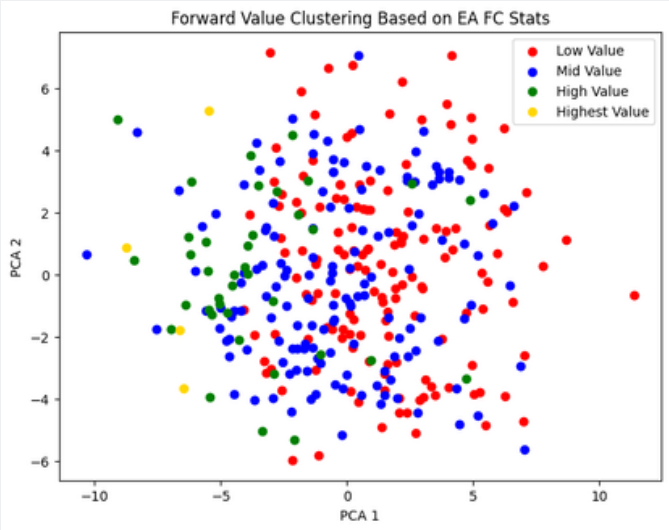
Midfielders

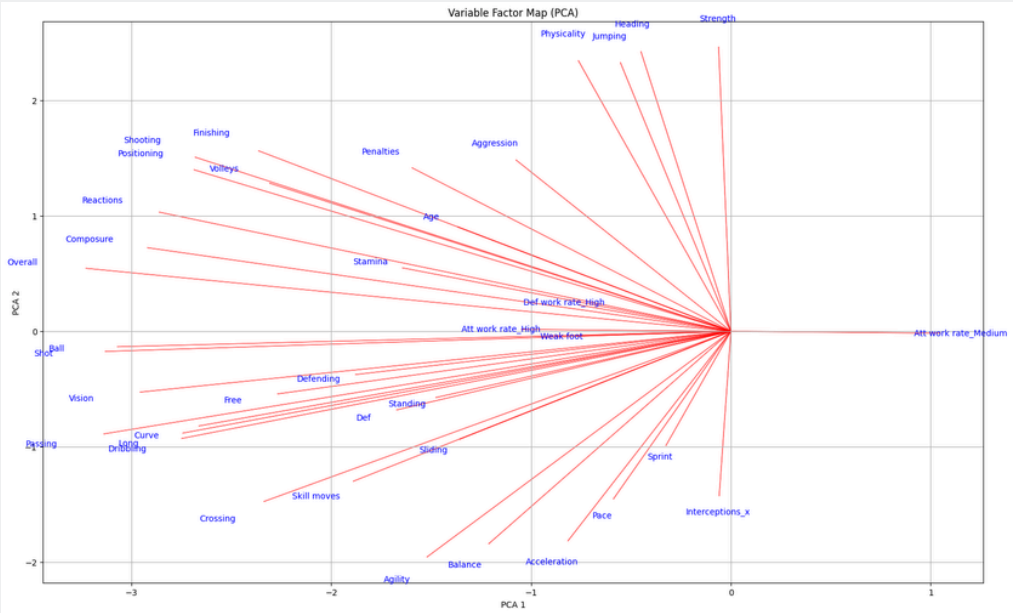




Midfielders, as visible above, don't arrange themselves regularly, but they seem to form a halo. On the right there are mainly low value players and central part is filled with a mix of blue and red dots. Green and yellow points form left edge along the whole y-axis. Following this cluster, ***Overall*** and ***Passing*** are the features high value players have the highest. Then, of course, midfielders divide themselves on their main skills: more defensive player are low in the figure, while offensive midfielders stay at the top.

Forwards





Forwards look **divided** in **three** value-based main **clusters**: from the right to the left, I can encounter low, medium and high value players, with the first two sharing some graph zones. As usual, ***Overall*** is the **most useful** variable to determine worth players, followed by *Passing, Shot, Ball* and *Vision*. Likewise defenders plot, *Att work rate* category 'Medium' tends to be common in low value players.

06. Modeling

After this deep analysis, I dive into the modeling part, in order to get knowledge from data I just analyzed. I find two problems solvable with my players DataFrame:

- Identifying playstyles for each position and associating them to actual footballers.
- Finding ideal substitute of a certain player, in order to lead market choices.

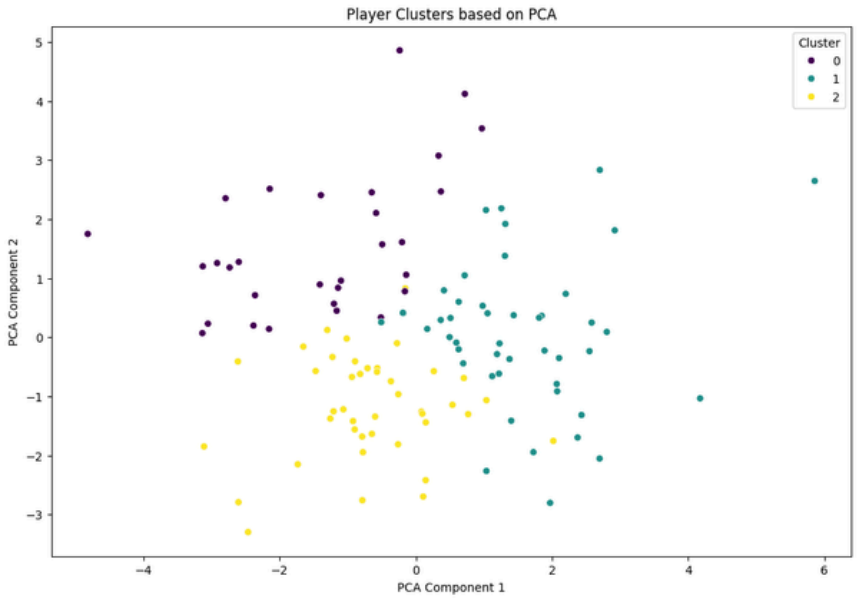
For each problem, I develop a script with an approach to solve it. I will analyze in detail *playstyles.ipynb* and *find_similars.ipynb* files.

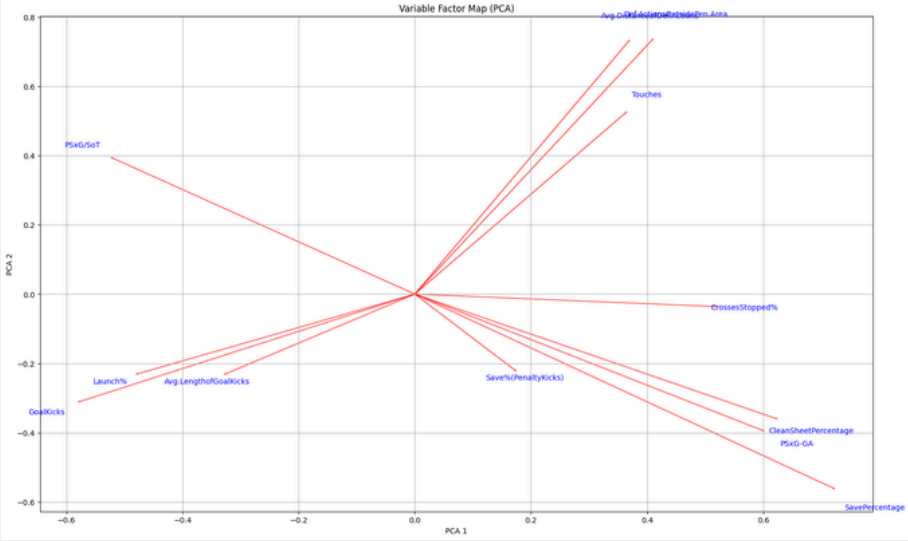
01. playstyles.ipynb

I start the task looking for known play styles for each position on the field. I use a link for goalkeepers, one for centre-backs and then another resource for full-backs, midifelders and forwards. Then I save the playstyles in a different file and I create the partial DataFrames (one for each position). For each position, I use again PCA and KMeans to create the plot and, given the number of clusters desired, associate each point to the correct cluster.

Goalkeepers

Goalkeepers, according to the link chosen, can be divided in **Line,** keeper who tends to stay much near the goal and are excellent at saving shots (i.e. Atletico Madrid's Jan Oblak), **Sweeper**, who positions himself higher up the pitch, sweeps up during defensive phase and excels in 1-on-1 situations (for example Napoli's Alex Meret), and **Ball Keeper**, who has great vision and passing ability and participates during building-up phase, like Manchester City's Ederson.



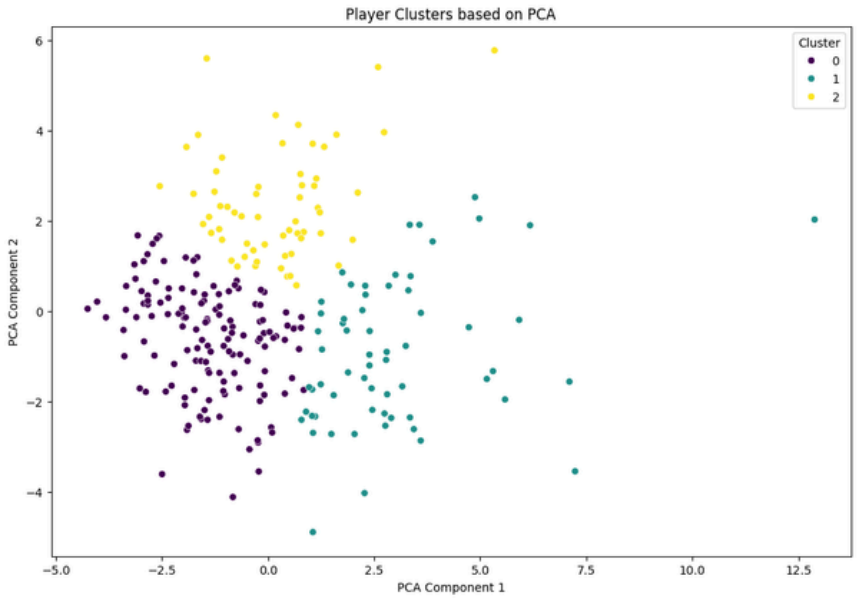


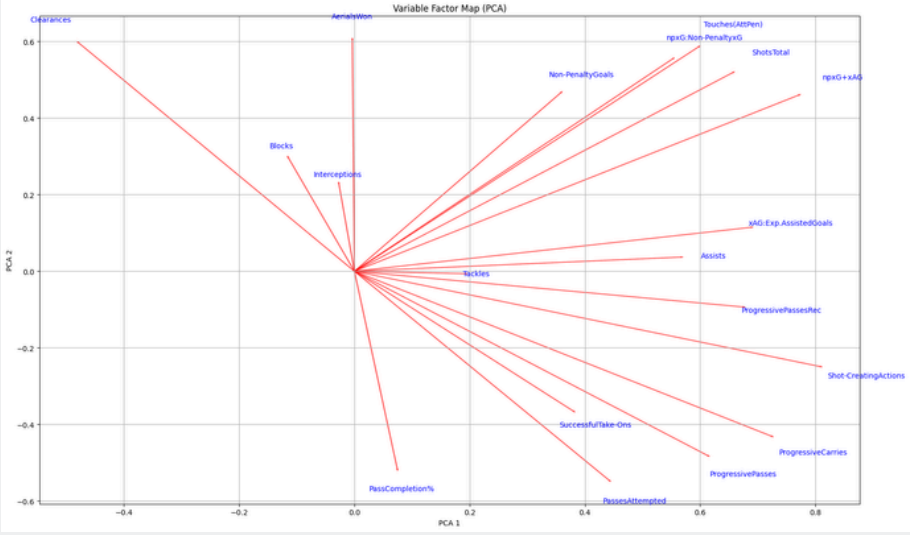
Following previous descriptions, **Ball Player** cluster emerges clearly as green cluster (number 1), due to *Touches* attribute and *Launch%* arrow pointing in the opposite direction. Then

I select cluster 2 (yellow one) as **Line**, since *SavePercentage* pointed down; I associate purple cluster (number 0) with **Sweeper** as it was the remaining one. Silhouette score is 0.12.

Centre-Backs

Centre-Backs play styles are three, according to this link: **Builder**, defenders who are able to benefit from possession so they have good technique and passing (i.e. Inter Milan's Alessandro Bastoni), **Destroyer**, players whose main goal is stopping the opponent with their incredible tackling and pressing, like Real Madrid's Eder Militao. Last play style is **ExtraFront**, composed by defenders joining often the attack, really good at aerial game, like Juventus' Bremer.

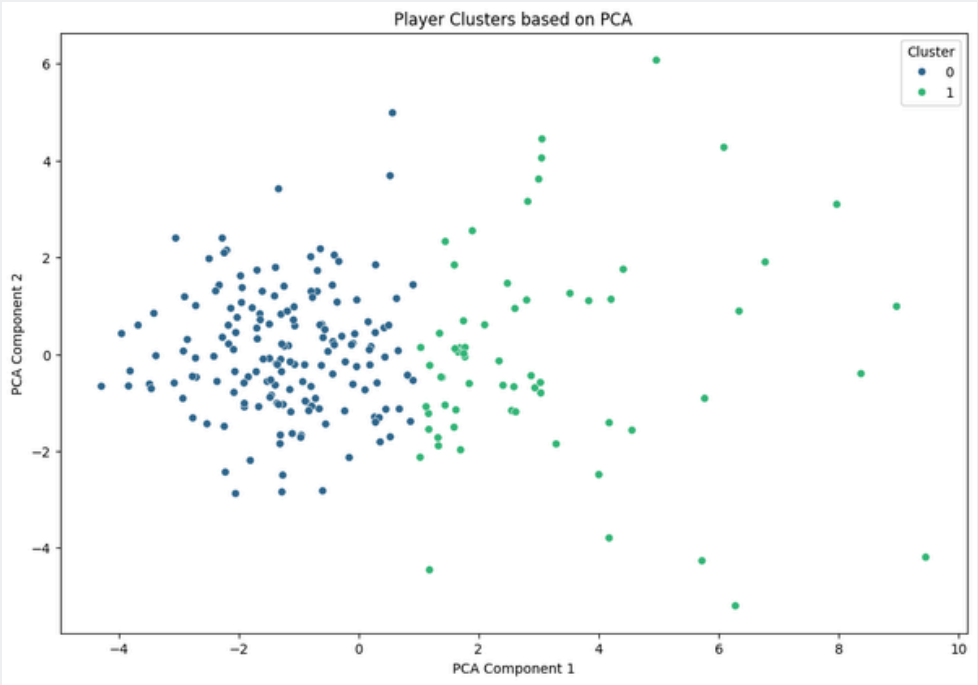


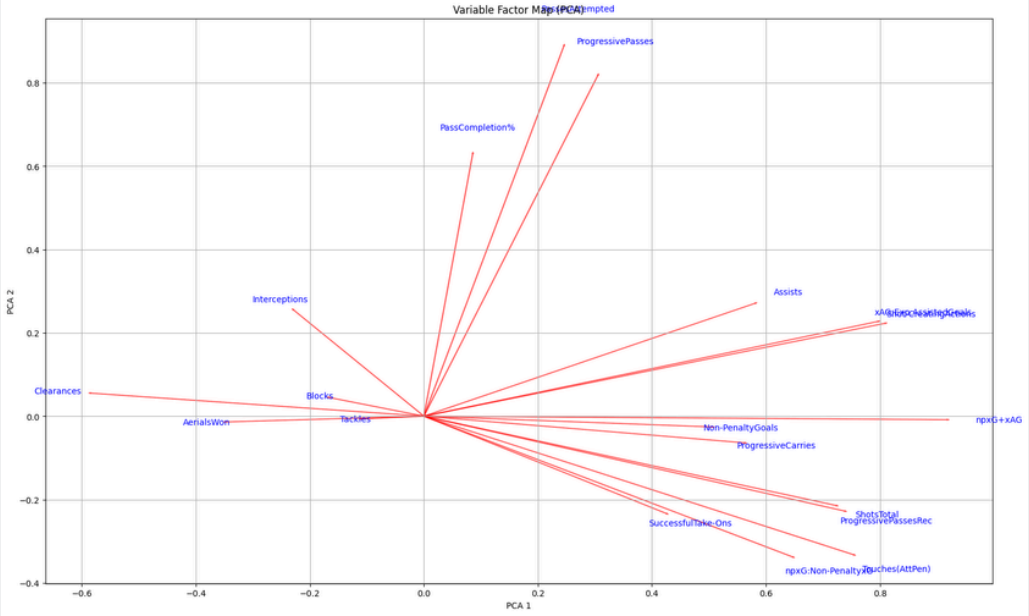


From PCA plot above, I can easily associate green cluster (number 1) with **Builder**, due to the relevance of *ProgressiveCarries, ProgressivePasses* and *PassesAttempted* to points position. Then I identify cluster 2 (yellow one) with **ExtraFront**, as *Touches, npxG* and *AerialsWon* have higher values in yellow dots. Lastly, purple cluster is **Destroyer** since it's the remaining one. Silhouette score is 0.15, so clustering can be considered decent.

Full-Backs

Full-backs (both left-backs and right-backs) are identified by only two play styles: **Defensive**, full-backs who tend to focus more on defensive phase, engage more in defensive duels and play less crosses, like PSG's Lucas Hernandez. On the other hand, **Wing-backs** get forward and contribute more to the attack with crosses and shots (i.e. Theo Hernandez of Milan).

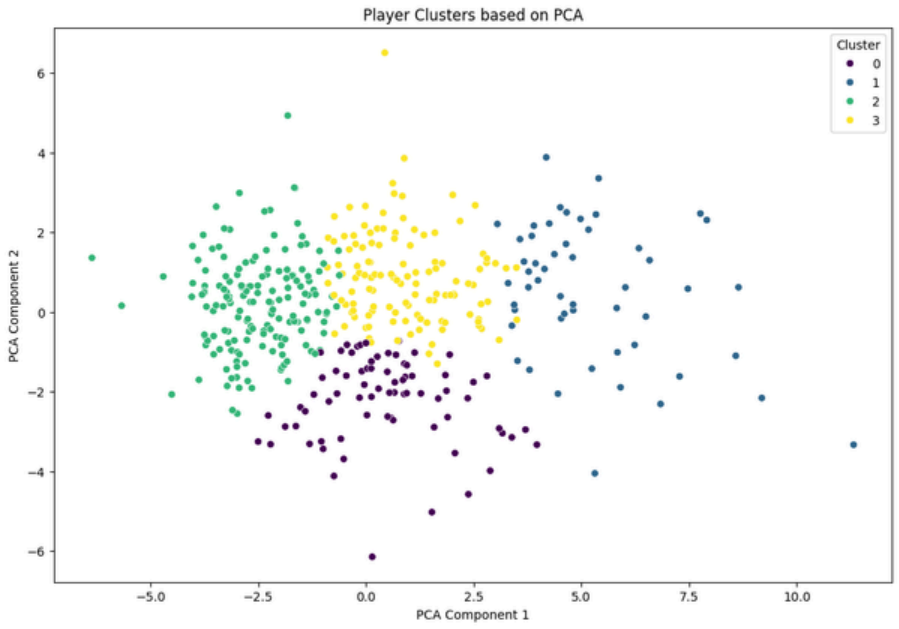


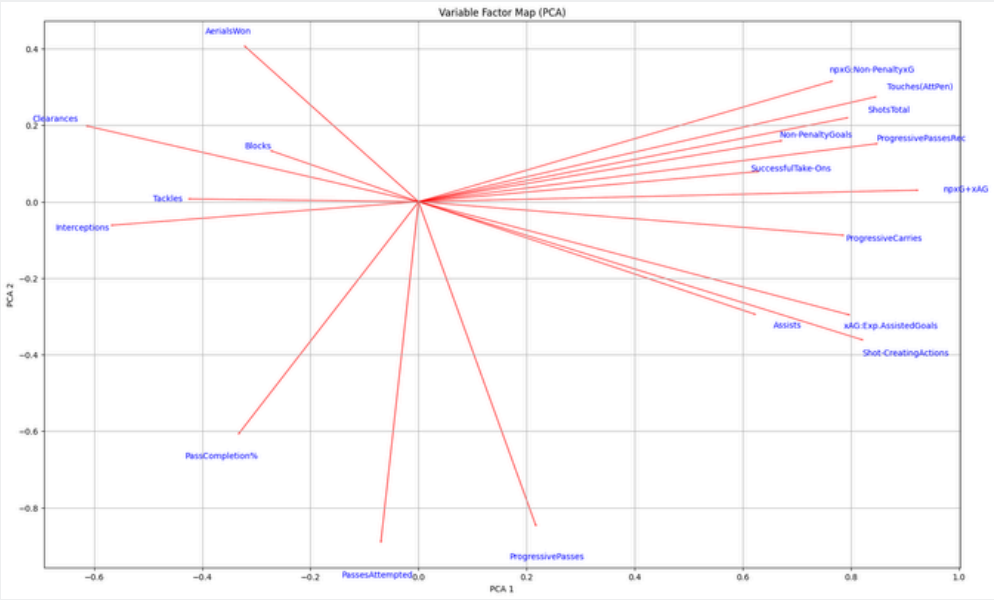


From the graphs displayed above appears manifest the difference: cluster 1 (green) can be identified in **Wing-backs**, due to offensive variables arrows pointing to the right, while in cluster 2 (blue) there are **Defensive** full-backs since all defensive variables like *Interceptions, Clearances, Tackles* and *Blocks* grow going to the left. Silhouette score is 0.24, so dots aren't divided badly.

Midfielders

Midfielders, instead, are divided not only by their play style but also by their position. According to the source, midfielders are divided in four play styles: **Creative**, proper of midfielders who progress the ball with passing and play more in the offense, like Leverkusen's Florian Wirtz, **Playmaker**, play style of players who tend to focus more for others building the phase with lot of passes (i.e. Atletico Madrid's Koke), **Box-to-box**, whose players run much along the whole field, engage in many defensive duels and then get forward, like Juventus' Adrien Rabiot, and **Ball-Winning**, proper of midfielders whose main concern is recovering possession for their team (i.e. Real Madrid's Aurelien Tchouameni).

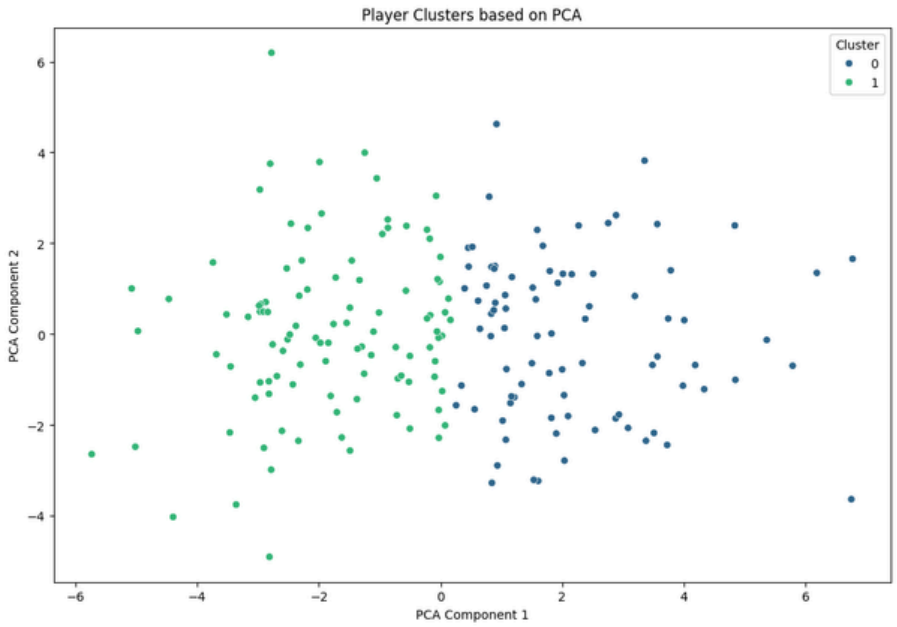


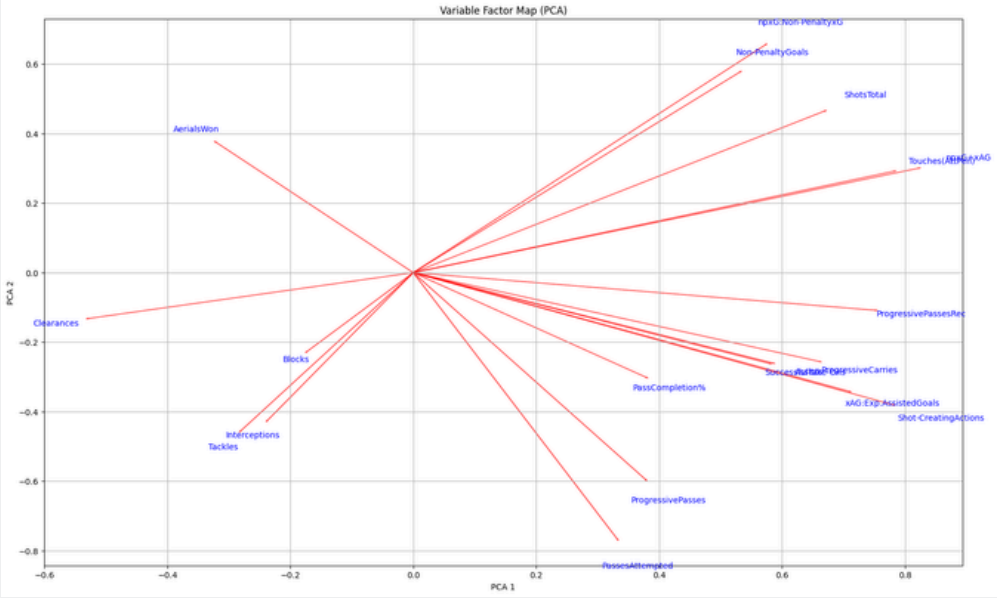


Points in PCA cluster quite well as displayed, even if silhouette score is only 0.16. I associate **Playmaker** label to cluster 0 (purple), as many passing attributes arrows point to the right. Cluster 2 (green) belongs to **BallWinning** midfielders, since defensive variables like *Interceptions*, *Tackles* and *Clearances* point to this cluster. Then I labeled yellow cluster (number 3) as **BoxToBox** due to the relevance of *PassCompletion%*, *ProgressivePasses* and *PassesAttempted* have in this cluster. Remaining dots are marked as **Creative** and, correctly, stay in the middle as they are really versatile midfielders.

Wingers

Then I do the same process with wingers, in this case both left and right wingers merged since they share the same play styles that, according to the link chosen, are just two: **Inverted** and **Wide**. First play style is proper of players who start wide and tend to go inside the field, they play more shots and fewer crosses and passes (i.e. Arsenal's Bukayo Saka). 'Wide' wingers, on the other hand, focus on creating chances for their team via crosses and passes, but take fewer shots, like Arnaut Danjuma of Everton.

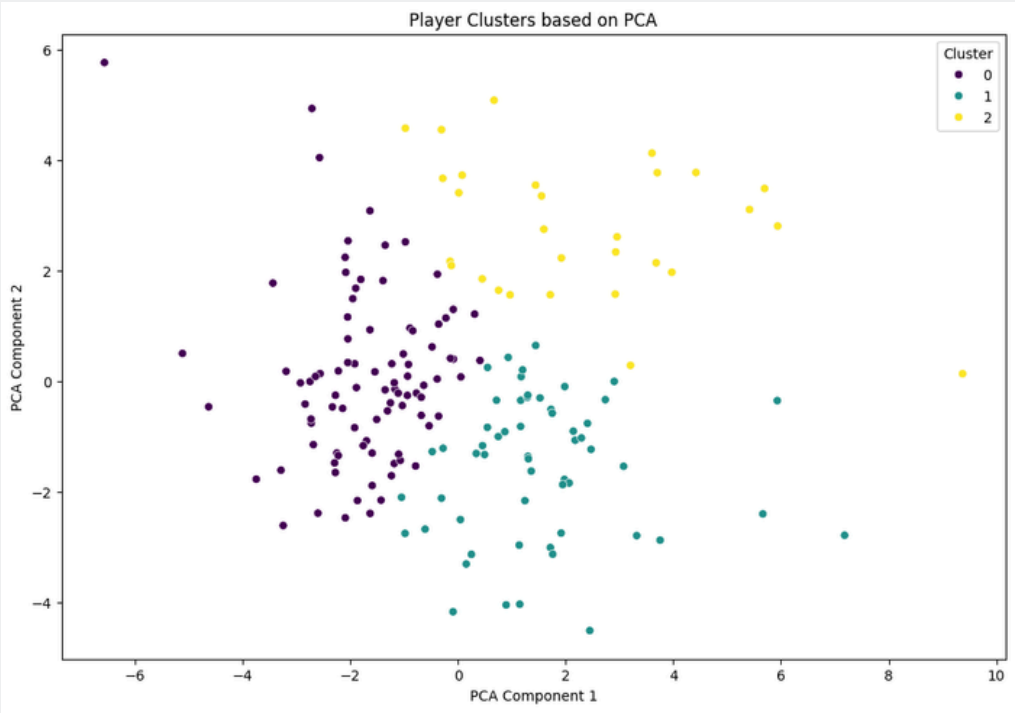


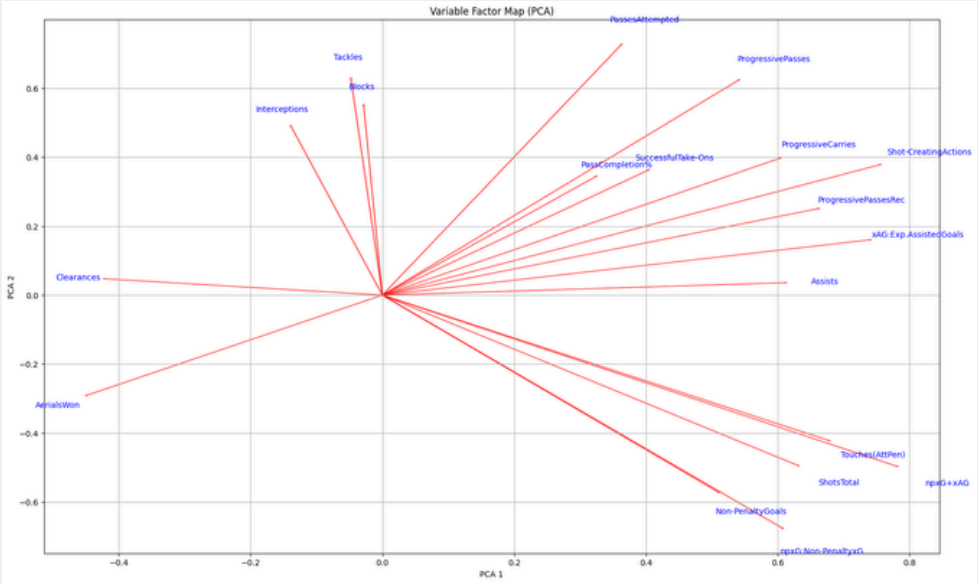


Looking at the Variable Factor Plot, I selected cluster 0 (blue) as **Inverted**, since it has players with higher values in attributes like *ShotsTotal, npG* or *npxG*. Then I associated the other cluster (green) with **Wide** play style.

Forwards

Lastly, I deal with forwards, which the source divides in 3: **False9**, forwards who use to go down in the middle of the field to participate in offense building and play more passes and fewer shots than other forwards (i.e. Arsenal's Gabriel Jesus); **Finisher** forwards are players who receive the ball near the penalty area and score goals rather than creating occasions for teammates, like Manchester City's Erling Haaland. **Target** players stay high up the pitch to receive passes and play banks: they engage in a lot of aerial duels and make many headers. (i.e. Tammy Abraham of Roma).





3 clusters separate quite well, but with a 0.16 silhouette score, low due to the scattered nature of the points. I identify **Target** label with cluster 0 (purple) since *AerialsWon* separates points on the left side of x-axis, **Finisher** label with green cluster (number 1) as *ShotsTotal, npG* and *npxG* arrows in Variable Factor Plot point to down-right. Remaining cluster (yellow) is associated with **False9**, as *PassesAttempted* and *ProgressivePasses* arrows witness.

After these analyses, I label each row in the dataset with the correct play style in the *Style* column.

02. **find_similars.ipynb**

find_similars.ipynb is a really simple script aiming to find most similar players to a given one. I use cosine distance as similarity measure to find closest players to the selected one. The script starts with libraries import, in particular I import the following:

- pandas, to deal with DataFrames
- sklearn, to preprcess and scale DataFrame attributes
- scipy, in particular *spatial* module, to calculate cosine distance
- numpy, to make calculations easier

Then I read all datasets I have, *goalkeepers_df.csv*, *goalkeepers_fifa_df.csv*, *players_df.csv* and *players_fifa_df.csv.* Regarding EA Sports FC 24 related DataFrames, I remove all variables but EA FC attributes. Main function is called *find_similar_players* and, given a DataFrame and the name of a player (or eventually its index), removes variables not helpful with the investigation, like *Wage, Age* and *Value*, scales remaining variables and computes, for each row in the resulting DataFrame, its cosine similarity measure. At the end returns a DataFrame with the *top_n* (if not specified, 5) similar players to the one given as parameter.

Results of this tool can't be measured but they seem pretty good, since after looking for players similar to Erling Haaland, I get the following result:

Player	Club	Age
Dusan Vlahovic	Juventus	24
Artem Dovbyk	Girona	26
Robert Lewandowski	Barcelona	35
Alvaro Morata	Atletico de Madrid	31
Joselu	Real Madrid	34

07. Conclusion

This report went deeply into the world of football data analysis, offering a thorough examination of various aspects of the game, including player and team performance, market trends, and league comparisons. My analysis provided significant insights that can be leveraged by clubs, managers, analysts, and other stakeholders to enhance their understanding and strategic approaches within the sport.

The integration of detailed data analysis into football operations is not just a trend but a necessity in the modern era of the sport. The insights gained from this analysis offer a strategic advantage that can lead to improved decision-making, enhanced performance, and sustained success on and off the field. As football continues to evolve, the role of data will become increasingly central, providing the foundation for innovation and excellence in the sport.
We hope that the findings and recommendations presented in this report will serve as valuable resources for clubs, managers, and analysts, guiding them towards more informed and effective strategies. By embracing data-driven approaches, the football community can look forward to a future where the beautiful game continues to thrive and captivate audiences around the world.