POLITÉCNICA

Prediction of Football Statistics for bookmakers and football professionals

Data Mining – Final Project

Pablo SANCHEZ

Overview

1 - Introduction

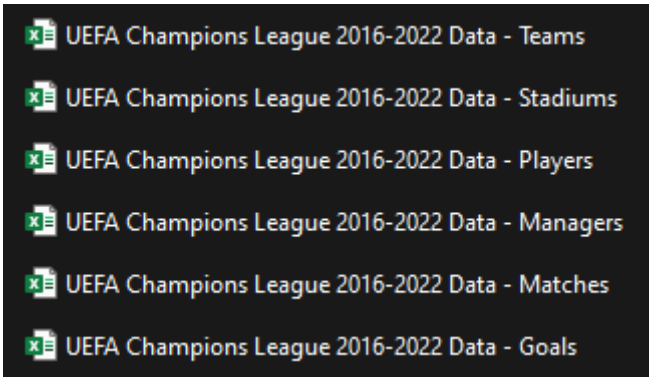
The studied dataset is a history of the most prestigious European football competition, UEFA Champions League (UCL), from 2016 to 2022. The UCL has been captivating audiences since its inception in 1955, evolving into one of the most widely viewed and eagerly anticipated football tournaments globally. This dataset, acquired from Kaggle (https://www.kaggle.com/datasets/cbxkgl/uefa-champions-league-2016-2022-data/data), dives into the various facets of the UCL, offering a detailed array of information, such as teams, players, goals, managers, matches, stadiums, and more. This dataset offers an opportunity to explore meaningful insights into the intricacies of team dynamics, player performances, managerial strategies, and the overall unfolding narrative of the UEFA Champions League.

With a robust volume of data - 2279 rows of goal-related information, 78 entries on team managers, 744 match records, 2769 player profiles, 86 stadium details, and 74 team specifications - this dataset promises to facilitate a comprehensive exploration and predictive modelling, to understand patterns within the UEFA Champions League landscape.

2 – Dataset Information

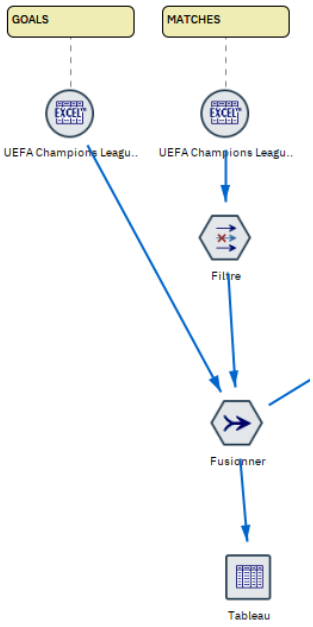
2.1 – Preliminary Work

A first step is to separate each sheet from “UEFA Champions League 2016-2022 Data.xlxs” into different .xlxs files:



Now, from these six files, we need to create one single dataset. To do so, we first create a merged table between Goals and Matches.

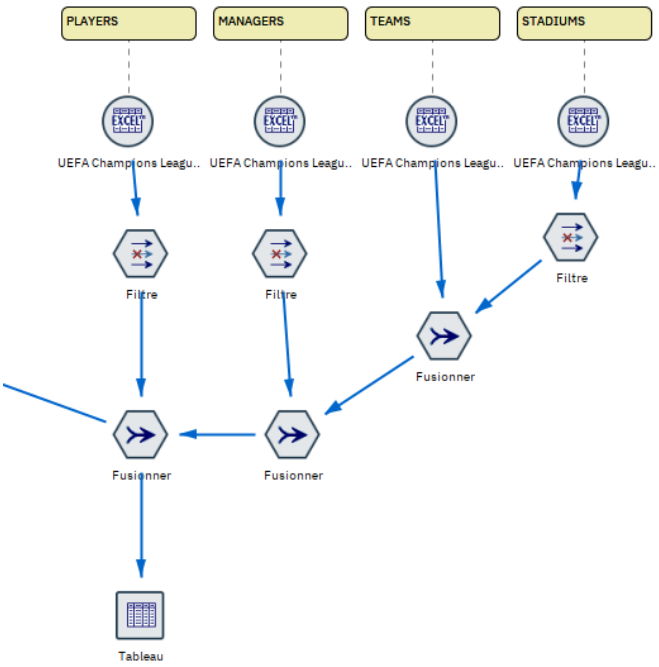
By reading through the tables, we can notice that in table “UEFA Champions League 2016-2022 Data – Matches.xlxs“, by sorting decreasingly Penalty shoot-out column, and knowing that the value can be 1 or 0, no game ended in a penalty shoot-out. Thus, we can get rid of this column by using a filter node, and then merge Goals and Matches, based on their common Match ID key and with inner join method.



Then, we merge all four other files, step by step. Firstly, we use a filter node to change Name into Home_Stadium in Stadium table. Then we merge Teams and Stadium, based on common keys Country and Home_Stadium.

From there, we can use Managers table, without forgetting to use a filter node to change Team into Team_Name and add Manager_ before all other fields (X -> Manager_X). We merge Manager table to the table we created by merging Teams and Stadiums, with inner join method and common key Team_Name.

Finally, we use Players table. We filter the fields, by adding Player_ before all fields except Player_ID and Jersey_Number (X -> Player_X). Also, information such as jersey number is irrelevant to our study, we can get rid of the column, with inner join method and common key Team_Name.

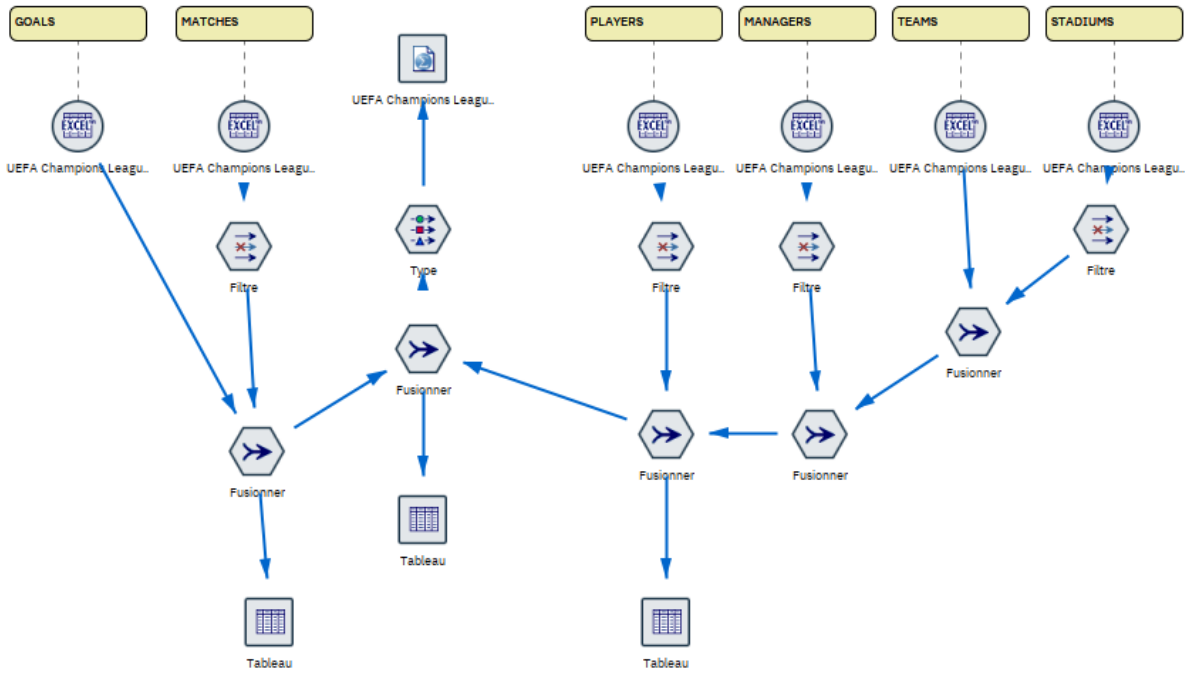


The last step is to create a .sav file out of these two tables. To do so, we merge the two created tables with external partial join, and using the following condition:

```
(PID = PLAYER_ID or ASSIST = PLAYER_ID) and (TEAM_NAME = HOME_TEAM or TEAM_NAME = AWAY_TEAM)
```

To create the file, we use a type node to automatically define field types, and a export statistics node, in which we write the path for where to create the new .sav file, and with a name of our choice (“UEFA Champions League 2016-2022 Data - V2”).

This new dataset displays information about all teams, with their stadium, their manager, and players. For each match the different teams played, if one of the players scored or assisted, additional information is given, from the match details to the goal details.



2.2 – Dataset Information

The newly created dataset displays the following information:

TEAM_NAME	**Team's name (Nominal)**
COUNTRY	Country where the team is located (Nominal)
HOME_STADIUM	Team stadium's name (Nominal)
CITY	City where the team is located (Nominal)
CAPACITY	Capacity of the team's stadium (Continuous)
MANAGER_FIRST_NAME	Teams manager's first name (Nominal)
MANAGER_LAST_NAME	Team manager's last name (Nominal)
MANAGER_NATIONALITY	Team manager's nationality (Nominal)
MANAGER_DOB	Team manager's date of birth (Continuous)
PLAYER_ID	Player's identifier (No type)
PLAYER_FIRST_NAME	Player's last name (No type)
PLAYER_LAST_NAME	Player's nationality (No type)
PLAYER_NATIONALITY	Player's date of birth (Nominal)
PLAYER_DOB	Player's team name (Continuous)
PLAYER_POSITION	Player's position on the field (Nominal)
PLAYER_HEIGHT	Player's height (Continuous)
PLAYER_WEIGHT	Player's weight (Continuous)
PLAYER_FOOT	Player's strong foot (Nominal)
MATCH_ID	Match identifier (No type)
SEASON	Season when the match happened (Nominal)
DATE_TIME	Date and time of the match (Nominal)
HOME_TEAM	Home team's name (Nominal)
AWAY_TEAM	Away team's name (Nominal)
STADIUM	Name of the stadium where the match happened (Nominal)
HOME_TEAM_SCORE	Home team's score (Continuous)
AWAY_TEAM_SCORE	Away team's score (Continuous)

ATTENDANCE	Attendance to the match (Continuous)
GOAL_ID	Goal identifier (No type)
PID	Player ID for goal scorer (No type)
DURATION	Minute of the game when the goal happened (Continuous)
ASSIST	Player ID for assist provider (No type)
GOAL_DESC	Description of the goal (Nominal)

If some players first and last name have no type, it can be explained for now because some players don't have a first or last name.

Limitations of our dataset:

This dataset stops in 2022, thus it can treat players or managers that are not active anymore. Another consequence is that some ne players or managers are not considered. Also, only one team is attributed to each player and manager. It means that for managers, we consider that they were managing their teams during the whole 2016-2022 period. For players, it means that a player may have scored for a team A in previous seasons, and now plays for team B.

These limitations affect the trustworthiness of our result. This study must not be taken in real life context but must belong to its proper context.

3 – Business understanding

3.1 – Business goal

The primary objective of this study is to use the dataset from the UEFA Champions League (UCL) history from 2016 to 2022 to provide valuable insights for bookmakers and betting enthusiasts.

In the realm of sports betting, accurately predicting match outcomes is of upmost importance. The business goal is to develop predictive models that can assist bookmakers in forecasting match results, identifying key factors influencing team performance, and ultimately improving the precision of odds offered in the betting market.

On another hand, this study could be used by football professionals, for example clubs' management, to identify key players, their aptitudes, or maybe link-up between players, as well as clubs' board, to identify players or managers with top performance, on a sport level, but also on a financial basis, with stadium attendance.

3.2 – Business Questions

To address the business goal, several key questions have been formulated:

Descriptive Analysis:

1. What is the number of matches played per team (giving information about team's participation in the competition)?

2. What is the distribution of wins, loss, and draw per team?

3. What is the average scoresheet for each manager?

4. What is the average of goals scored and goals conceded per team per match?

5. What is the distribution of wins, draws, loss for home team per team?

6. What is the distribution of wins, draws, loss for away team per team?

7. What is the distribution of goals scored per player (overall and average per match)?

8. What is the distribution of assists provided per player (overall and average per match)?

9. Does age of players affect their performance (wins, loss, draw, goals)?

10. What are players/managers nationalities?

11. What is the average height/weight for each position?

Diagnostic Analysis:

1. Are there specific players that significantly influence the likelihood of winning matches?
2. Are there specific player linkups (goal scorer and assist provider combination) that significantly influence the likelihood of scoring goals?
3. Do teams with winning history demonstrate a higher probability of winning matches?
4. Do teams with certain characteristics (e.g., city, stadium, stadium capacity, attendance, specific players/manager) demonstrate a higher probability of winning matches?

Predictive Analysis:

1. Can we build a model to predict match outcomes based on historical data, including team, manager, and player attributes, match context, and previous performances?

2. How accurately can we forecast goal differentials for individual matches, and what variables play a significant role in these predictions?

Prescriptive Analysis:

1. Based on the insights gained, what strategies can bookmakers employ to adjust odds and maximize profitability?

2. How can the predictive models assist club's management and board in identifying what can be considered key to winning matches?

3.3 – Data Mining Goal

The aim of this data mining work is to extract meaningful insights and patterns from “UEFA Champions League 2016-2022 Data”.

Leveraging advanced analytical techniques, the study aims to employ a combination of clustering analysis and correlation assessments among numerical variables. These techniques will provide a comprehensive understanding of the inherent structure and relationships within the data.

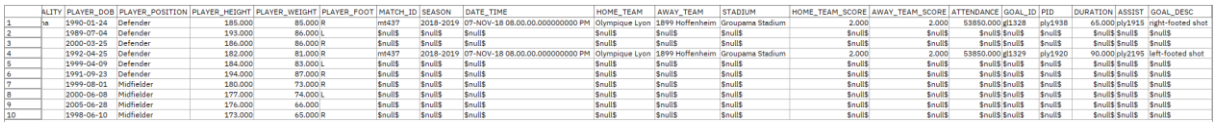
The goal is to uncover hidden patterns, identify influential factors, and develop predictive models using ensemble methods, linear regression, logistic regression, neural networks, or instance-based classification (Nearest Neighbour).

Through these techniques, the study seeks to equip bookmakers and football managing team with actionable insights for effective decision-making and strategic planning, on one side to set odds for betting purpose, on another side to improve football teams’ statistics.

4 – Data Understanding: Descriptive and Diagnostic analysis

4.1 – Missing and Null Values

This dataset is made so that we have null values. If a player didn't score, then all its information about match and goal details are null.



	ALITY	PLAYER_DOB	PLAYER_POSITION	PLAYER_HEIGHT	PLAYER_WEIGHT	PLAYER_FOOT	MATCH_ID	SEASON	DATE_TIME	HOME_TEAM	AWAY_TEAM	STADIUM	HOME_TEAM_SCORE	AWAY_TEAM_SCORE	ATTENDANCE	GOAL_ID	PID	DURATION	ASSIST	GOAL_DESC
1	ha	1990-01-24	Defender	185.000	85.000	R	mt437	2018-2019	07-NOV-18 08.00.00.000000000 PM	Olympique Lyon	1899 Hoffenheim	Groupama Stadium	2.000	2.000	53850.000	gl1328	ply1938	65.000	ply1915	right-footed shot
2		1989-07-04	Defender	193.000	86.000	L	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$
3		2000-03-25	Defender	186.000	86.000	R	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$
4		1992-04-25	Defender	182.000	81.000	R	mt437	2018-2019	07-NOV-18 08.00.00.000000000 PM	Olympique Lyon	1899 Hoffenheim	Groupama Stadium	2.000	2.000	53850.000	gl1329	ply1920	90.000	ply2195	left-footed shot
5		1999-04-09	Defender	184.000	83.000	L	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$
6		1991-09-23	Defender	194.000	87.000	R	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$
7		1999-08-01	Midfielder	180.000	73.000	R	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$
8		2000-06-08	Midfielder	177.000	74.000	L	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$
9		2005-06-28	Midfielder	176.000	66.000		$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$
10		1998-06-10	Midfielder	173.000	65.000	R	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$	$null$

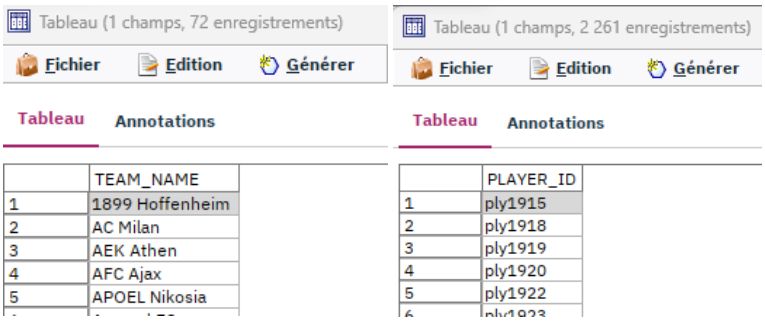
Because we want to keep a full dataset, and null values are part of the dataset, we won't erase null values, and just adapt the dataset to the questions we want to answer.

This dataset keeps track of all players from all teams and adds values (sometimes null) to describe a goal the player scored. It is important not to erase rows with null values if it's not for a specific purpose, because otherwise our dataset would be a goal scorer dataset and not a player's dataset.

We also have missing values, but only for player's names, thus we can adapt the content depending on our needs, for example by joining First Name and Last Name columns. There also are missing values for player's strong foot. We interpret this as ambidexterity.

4.2 – Outliers

A first read through our dataset, makes us understand that we are analysing a lot of teams, we have 72 different teams, for 2261 players:



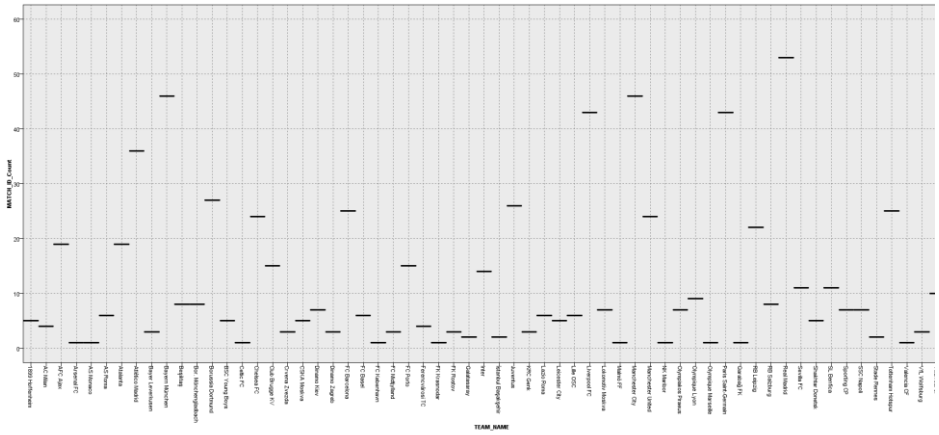
But, as we are looking for players with performances and statistics that are unusual. Therefore, we need to keep players to set standards, and players to beat those standards.

Erasing players because they score too many goals compared to the rest, in our case, would not make any sense.

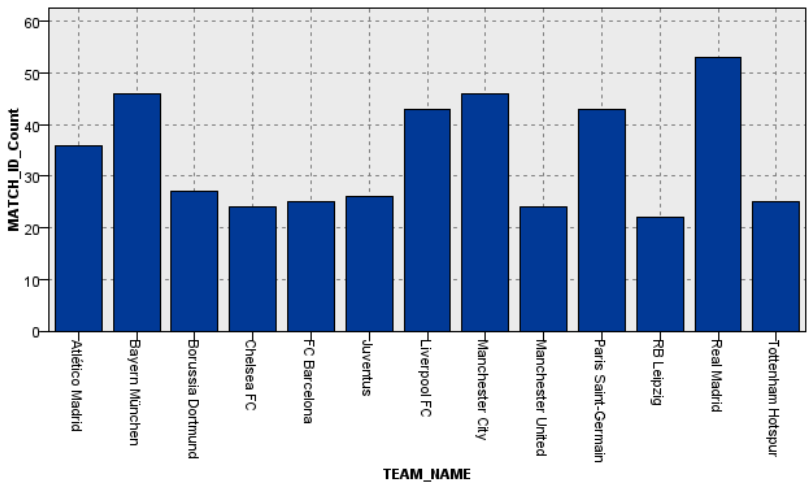
If in other studies, deleting outliers is appropriate, here we should keep every outliers, as they help us to settle predictions. For example, if the average participation of teams in UCL is 20 matches, and one teams that played 35 matches meets a team that played 1, then the odds are against the team that played less games.

4.3 – Analysis of distributions, graphs, and statistics

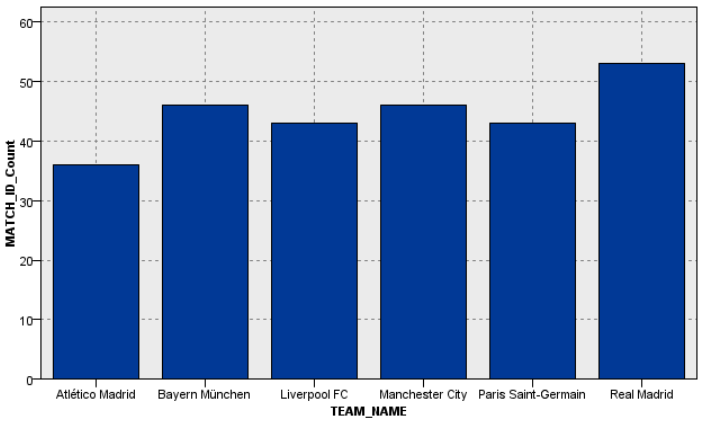
What is the number of matches played per team (giving information about team's participation in the competition)?



This grid shows the number of matches played per team. We can see that most teams played less than 20 games, but some teams are regularly involved in UCL games. By selecting only teams that played more than 20 games, we can gain visibility on this grid.



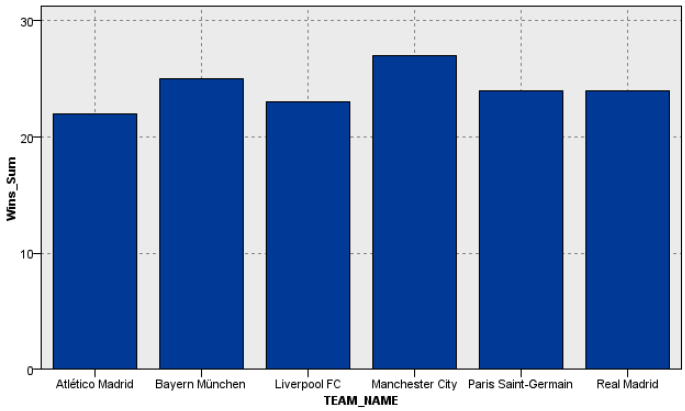
We can now identify clubs with most appearances in UCL. This teams are expected to participate more, thus to win more games. Among these teams, we can see that Atletico Madrid, Bayern Munchen, Liverpool, Manchester City, PSG and Real Madrid are the teams with most appearances, standing apart from the rest with more than 30 appearances. They can be considered the most experienced teams. Plus, knowing that group stage is 6 matches, and that the dataset is based on 6 seasons, we can say that teams with more than 36 (6 times 6) appearances should be out of group stage most years. Is this group, we would find the following teams (teams with more 36 or more appearances:



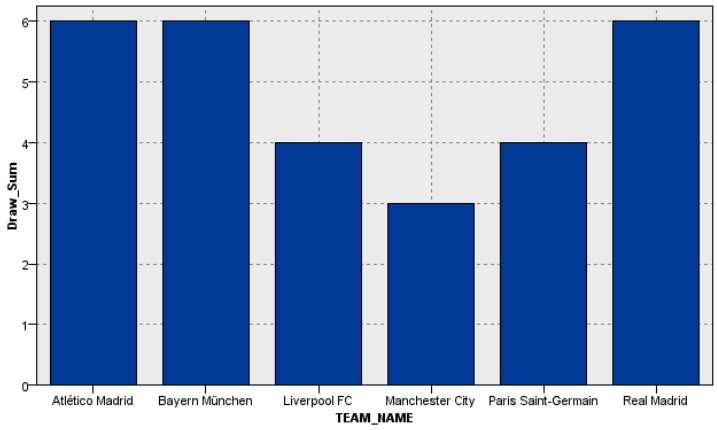
What is the distribution of wins, loss, and draw per team?

For this question, we will keep the select teams with more than 36 appearances, with the following Select node:

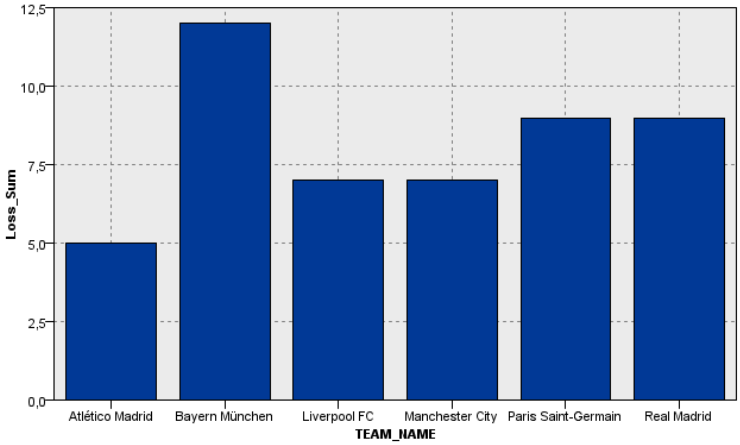
	TEAM_NAME	Draw_Sum	Loss_Sum	Wins_Sum
1	Atlético Madrid	6	5	22
2	Bayern München	6	12	25
3	Liverpool FC	4	7	23
4	Manchester City	3	7	27
5	Paris Saint-Germain	4	9	24
6	Real Madrid	6	9	24



All of the teams have more than 20 wins, but we can see that Manchester City and Bayern Munchen slightly stand out, meaning they wont more games.

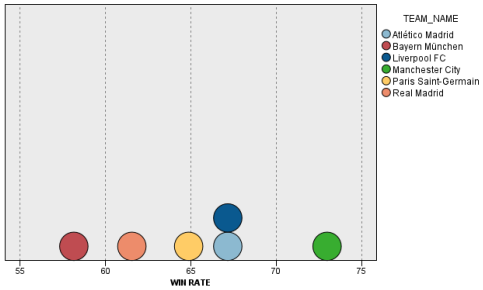


We can see that Manchester City is the team with less draws, when Atletico, Real Madrid, and Bayern have a common number of 6 draws. It means that City tends to either win or lose.

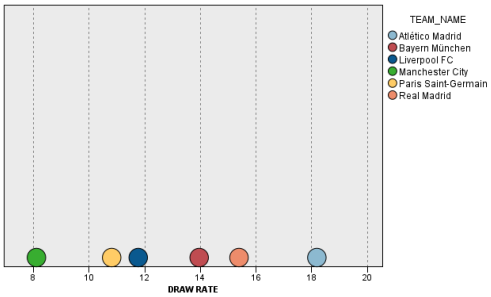


We can see that Atletico rarely loses, when Bayern loses a lot of matches, maybe its an indicator of the dynamics of those teams.

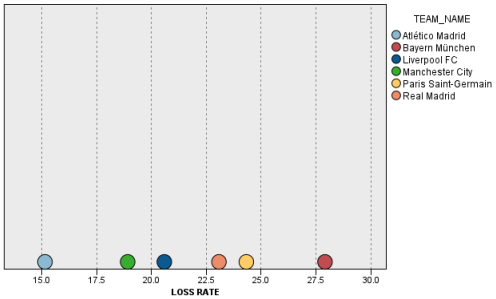
From these graphs, we can conclude that Manchester City is a team with impressive win rate, they rarely draw or lose. Bayern München played a lot of games, but rarely draw compared to the number of wins and losses. Finally, Atletico stands out for its low ratio of losses. These additional graphs confirm the described dynamics:



Manchester City has the best win rate (in %) by far, confirming their dominance, and Bayern wins more than half of their matches, but has the lowest win rate.



By far, Atletico has the biggest draw rate, which confirm our intuition, and confirm their defensive style of play. Manchester City indeed wins or loses as they draw only 8% of the time.



Finally, Atletico having the best loss rate (15%), shows that even if they rely on defence, it pays as they rarely lose. A more opened way of playing like Bayern make them the worst loss rate among those elite teams.

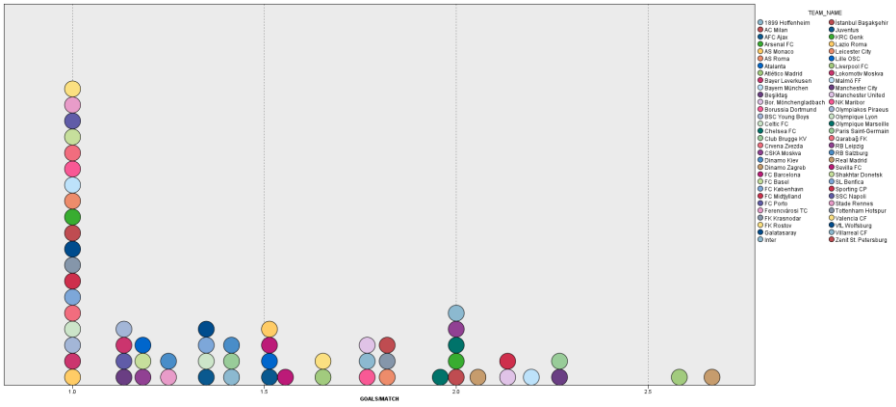
What is the average scoresheet for each team?

Again here, we will keep the select teams with more than 36 appearances, with the following Select node.

	TEAM_NAME	GOALS SCORED/MATCH	GOALS CONDEDED/MATCH
1	Atlético Madrid	1.639	1.817
2	Bayern München	2.196	1.842
3	Liverpool FC	2.581	2.055
4	Manchester City	2.283	2.178
5	Paris Saint-Germain	2.256	2.181
6	Real Madrid	2.057	2.245

If we only rely on the average of goals scored and goals conceded in UCL, Atletico and Real Madrid would be expected to lose their games, while Bayern, Liverpool, City and PSG would be expected to win their games.

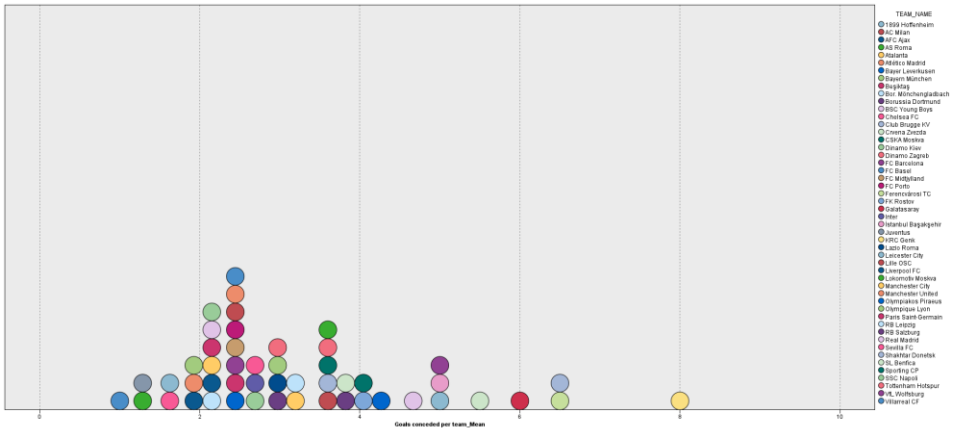
What is the average of goals scored and goals conceded per team per match?



The average number of goals scored per team is the one above. We can see that most teams with low number of participation score in average 1 goal per match. Meaning that even outsiders tend to score in UCL.

Two teams score more than 2.5 goals per match: Dinamo Zagreb and Liverpool FC. Behind them, with an average of 2 to 2.5 goals per match, we can find Manchester City, PSG, Real Madrid, or Bayern Munchen.

A high average of goals per game is interesting for clubs like Liverpool, City, Bayern, Real Madrid, or PSG, when we compare this number to the number of matches played, meaning that they perform well.

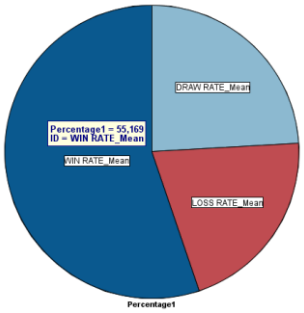


When looking at the average number of goals conceded per match, we can see that most teams concede approximately 2 goals. One team stands

apart, with an average of 8 conceded goals per match, its Genk. Thus, when playing against Genk, we can expect the opposition team to score a lot.

In the best defences, we find Juventus, AS Roma, Atletico, Chelsea, or surprisingly, FC Basel, that are less likely to concede goals.

What is the distribution of wins, draws, loss for home team per team?

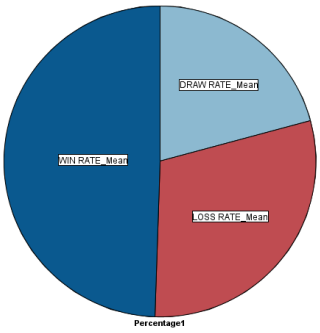


	WIN RATE_Mean	DRAW RATE_Mean	LOSS RATE_Mean
1	55.169	24.062	20.768

We can see that a team who is playing home has a huge advantage, with more than 1 chance over 2 to win the match. In 1 case over 4, the match ends in a draw. Finally, in 1 case over 5, the home team loses.

Statistics say that home team has way more chances to win, or at least 4 chances over 5 to earn at least 1 point from the match (3 = win, 1=draw, 0=loss).

What is the distribution of wins, draws, loss for away team per team?

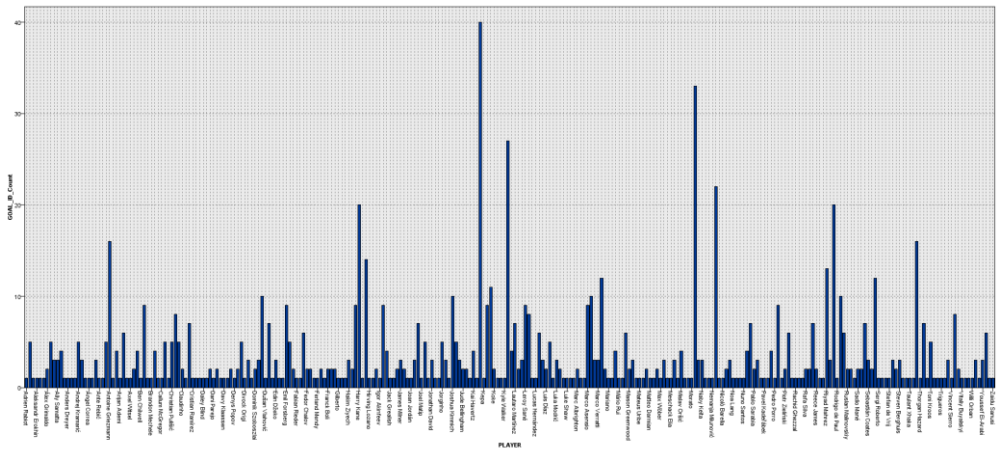


	WIN RATE_Mean	DRAW RATE_Mean	LOSS RATE_Mean
1	49.452	20.797	29.751

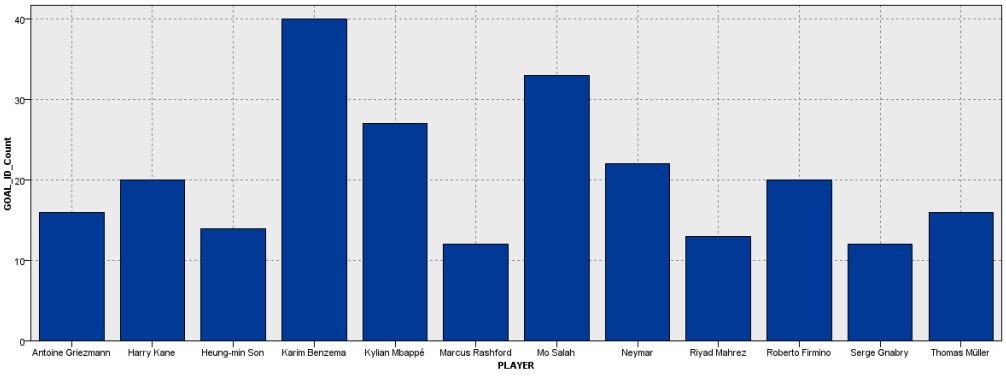
On the other hand, a team who plays away has slightly less than 1 chance over 2 to win, 1 chance over 5 to draw, or 1 chance over 3 to lose.

In general, statistics show that home team is advantaged, as the away team has more chances to lose, and less chances to win.

What is the distribution of goals scored per player (overall and average per match)?

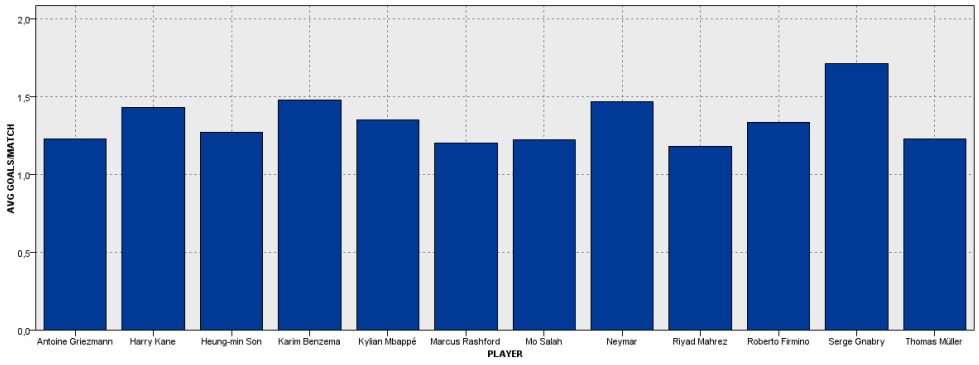


This is a record of all goal scorers, and their number of goals. Again, here we have way too many players to draw conclusions, so let's select players with more than 10 goals.



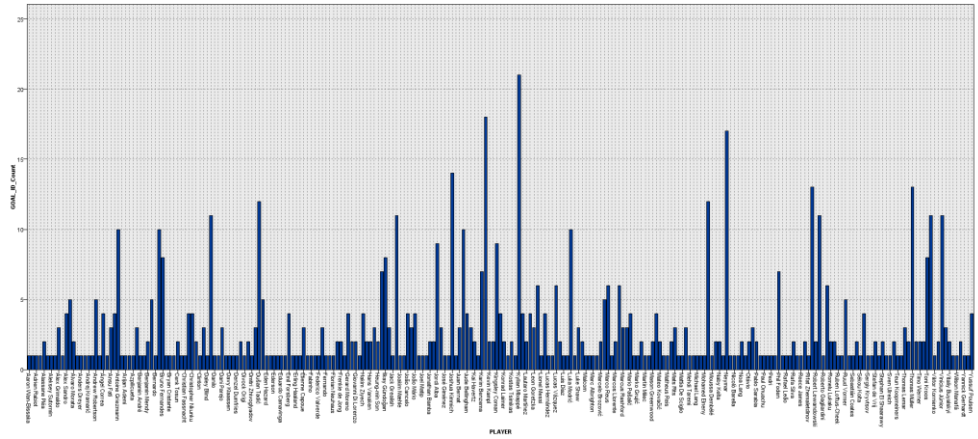
We can see that Karim Benzema and Mo Salah dominate by far this ranking, with 40 and 33 goals, followed by Kylian Mbappé, with 27 goals. Neymar completes the ranking of players above 20 goals, with 22 goals.

Now, to get the average number of goals per player per match, let's keep the list of players above 10 goals, to make it easier to read.

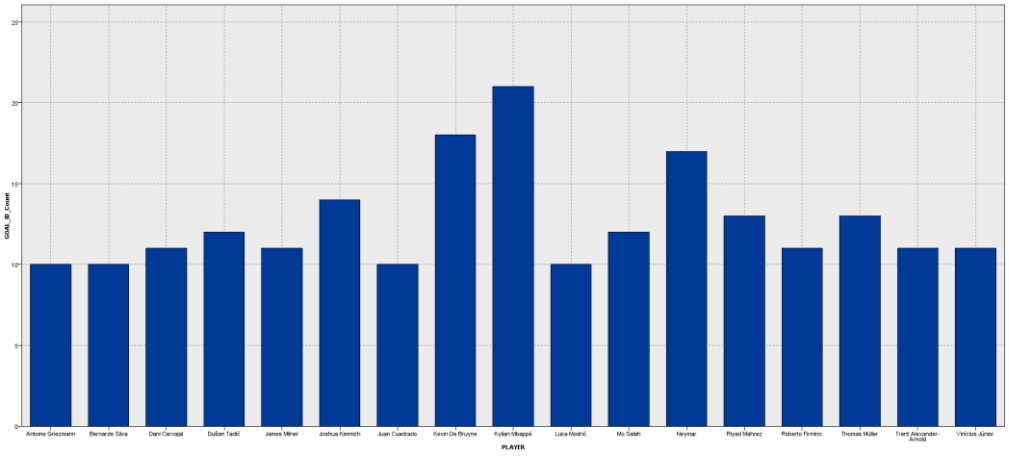


When looking at the average number of goals per match, we realized that all the players in this ranking statistically score at least one goal per match. Also, If Karim Benzema is in second place, underlining his reliability, we can see that Serge Gnabry has the best ratio in the UCL.

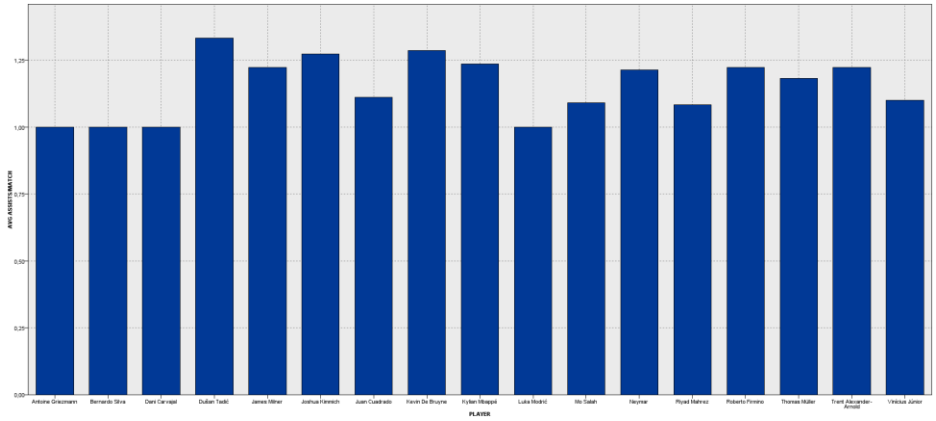
What is the distribution of assists provided per player (overall and average per match)?



This is a record of all assist providers, and their number of assists. Again to gain visibility, let's select only players with at least 10 assists.



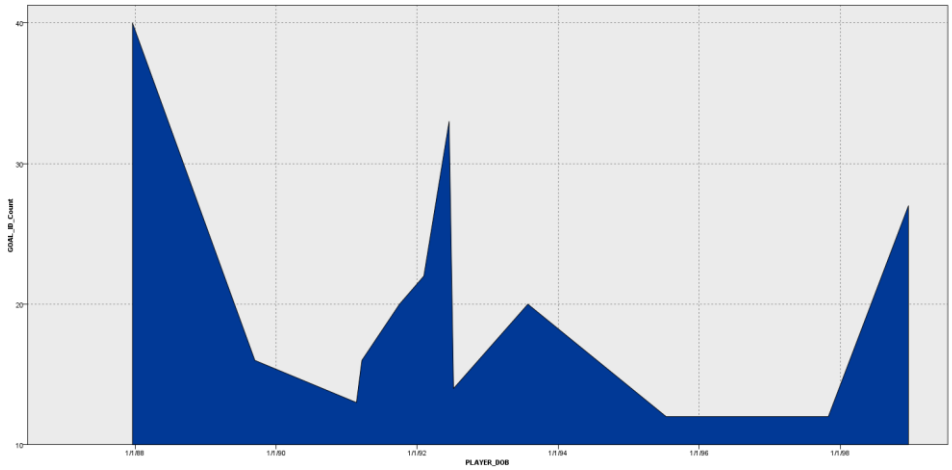
The top 3 is made of Mbappé, De Bruyne, and Neymar, that are both in teams which scores at least 2 goals per games according to statistics. Now let's see the average number of assists per match for the players in this ranking.



In terms of regularity, the top 3 is made of Tadic, De Bruyne, and Mbappé. Meaning that De Bruyne and Mbappé statistically show great ability to provide assist, as regularly provide, and have high overall number of passes.

Does age of players affect their performance (wins, loss, draw, goals)?

For this question, let's take again the ranking with most prolific goal scorers.



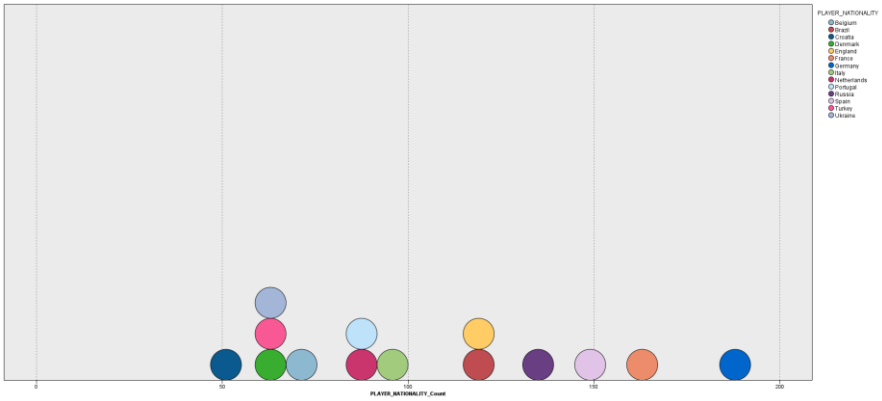
This is the number of goals scored per year of birth. We can see that the oldest player has most goals.

Therefore, in football, youth give the opportunity to improve, but experience and efficiency comes with age.

What are players/managers nationalities?

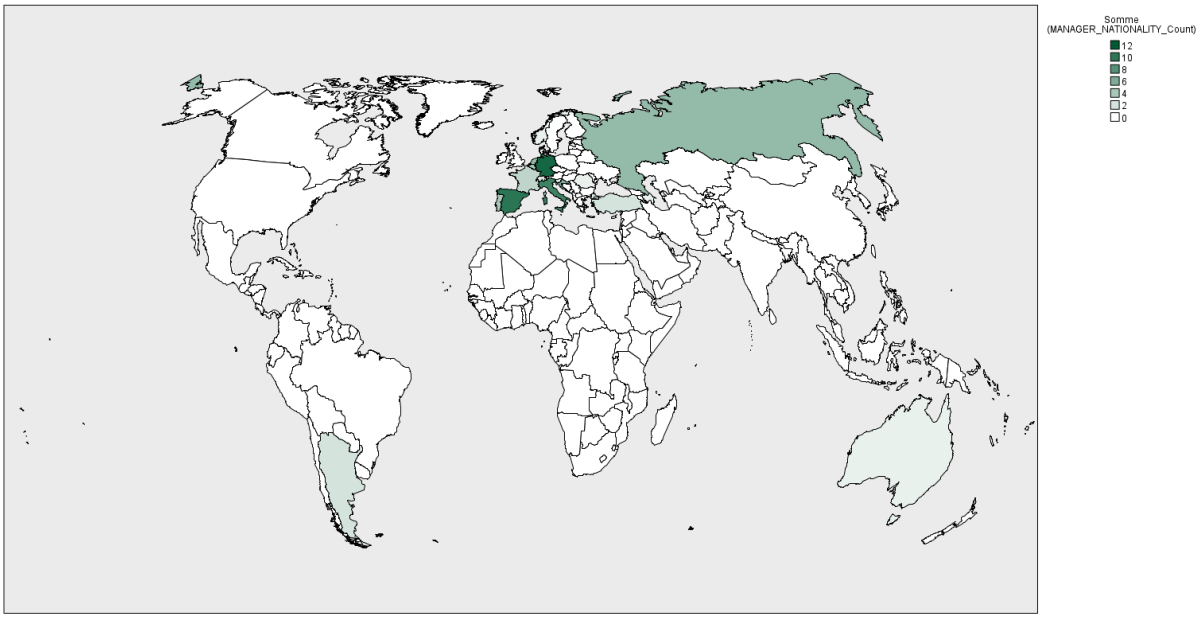
For players:

Most players come from Europe or South America. If we look closer at the represented countries, we have those countries, represented by more than 50 players:



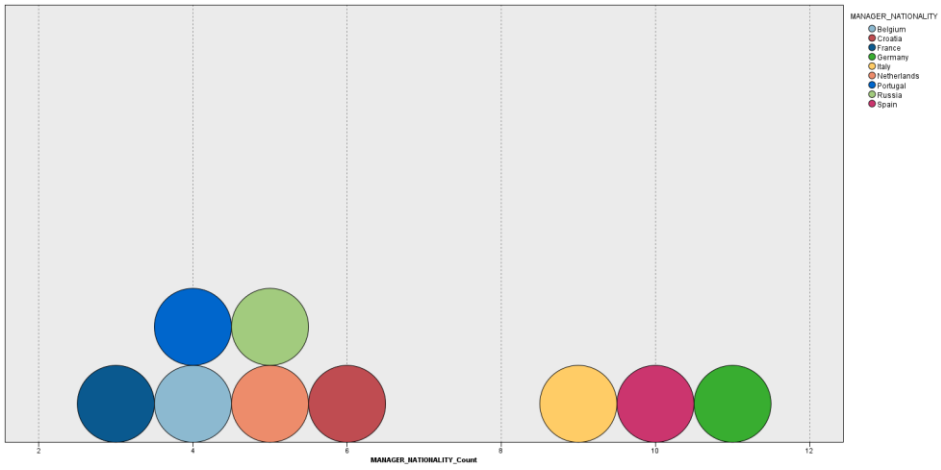
The most represented country is Germany, followed by France and Spain.

For Managers:



Most managers come from Europe, mostly Germany, Spain, and Italy.

This is confirmed by this graph, showing that only these two countries provide around 10 managers:



What is the average height/weight for each position?

For Goalkeepers:

	PLAYER_HEIGHT_Mean	PLAYER_WEIGHT_Mean
1	190.570	83.267

For Defenders:

	PLAYER_HEIGHT_Mean	PLAYER_WEIGHT_Mean
1	182.967	76.349

For Midfielders:

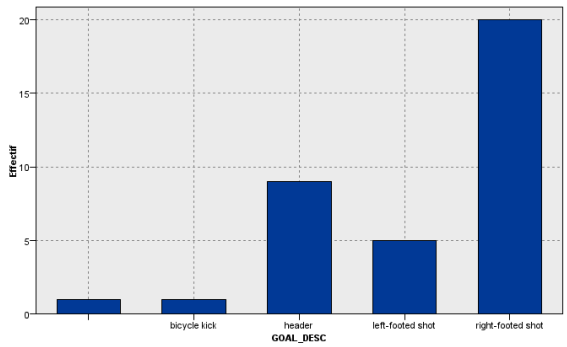
	PLAYER_HEIGHT_Mean	PLAYER_WEIGHT_Mean
1	179.506	72.570

For Forwards:

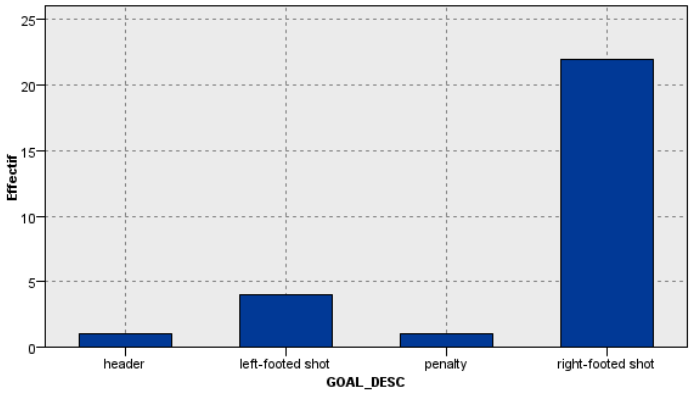
	PLAYER_HEIGHT_Mean	PLAYER_WEIGHT_Mean
1	180.473	74.795

From this, we can conclude that in UCL football, the tallest and heaviest player are goalkeepers by far, when field players are lights. Defenders and Forwards are approximatively the same size because they usually fight for balls in the air.

In a previous study, I compared Cristiano Ronaldo, all-time best UCL goal scorer, to Mbappé, rising star of football, on field and on statistics. Here is the conclusion that I made:



Ronaldo scores most goals with his strong foot (right), but 1 time over 5, he can shoot with weak foot, and almost 1 time over 4, he scores with his head, making him a complete player.



Most of Mbappé's goals are scored from his strong foot (right), meaning he is limited in his options.

	PLAYER_FIRST_NAME	PLAYER_LAST_NAME	PLAYER_HEIGHT_Mean	PLAYER_WEIGHT_Mean
1	Cristiano	Ronaldo	187.000	83.000
2	Kylian	Mbappé	178.000	73.000

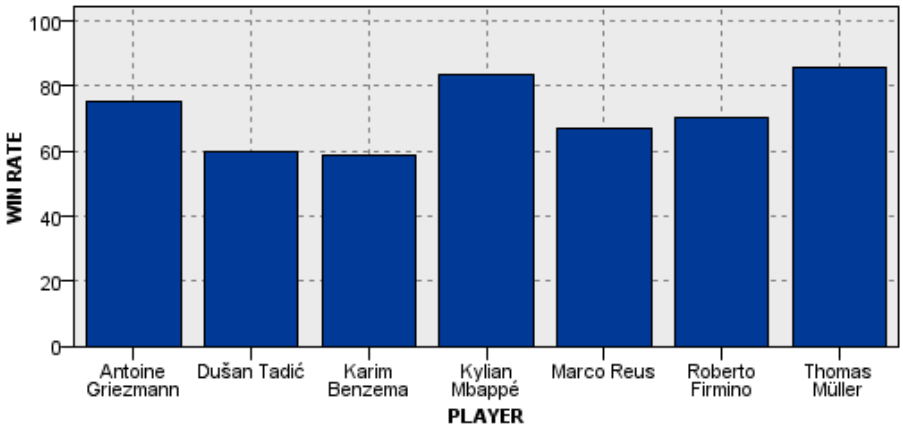
From height and weight, we also realize that it is easier for Ronaldo to play with his head, while Mbappé is way lighter, making him faster, thus less limited by the fact that he mostly scores with his strong foot. Both players had to adapt to their bodies.

Are there specific players that significantly influence the likelihood of winning matches?

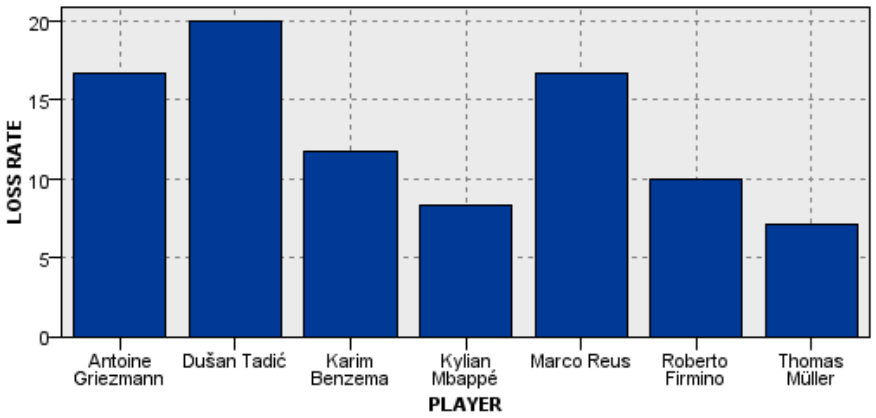
For this question, we keep players with at least 10 goals scored, or 10 assists provided. We end up with this list of 24 players:

	PLAYER_ID	PLAYER_FIRST_NAME	PLAYER_LAST_NAME	GOAL_Provided	GOAL_Scored
1	ply115	Antoine	Griezmann	10	16
2	ply135	Joshua	Kimmich	14	10
3	ply140	Kingsley	Coman	9	11
4	ply141	Serge	Gnabry	4	12
5	ply143	Thomas	Müller	13	16
6	ply1801	Harry	Kane	3	20
7	ply1805	Heung-min	Son	2	14
8	ply225	Marco	Reus	6	10
9	ply509	Juan	Cuadrado	10	3
10	ply564	Trent	Alexander-Arnold	11	1
11	ply581	James	Milner	11	3
12	ply589	Roberto	Firmino	11	20
13	ply590	Mo	Salah	12	33
14	ply60	Dušan	Tadić	12	10
15	ply652	Bernardo	Silva	10	9
16	ply653	Kevin De	Bruyne	18	9
17	ply664	Riyad	Mahrez	13	13
18	ply704	Marcus	Rashford	3	12
19	ply739	Kylian	Mbappé	21	27
20	ply741		Neymar	17	22
21	ply828	Luka	Modrić	10	3
22	ply831	Karim	Benzema	7	40
23	ply836		Rodrygo	6	10
24	ply837	Vinícius	Júnior	11	8

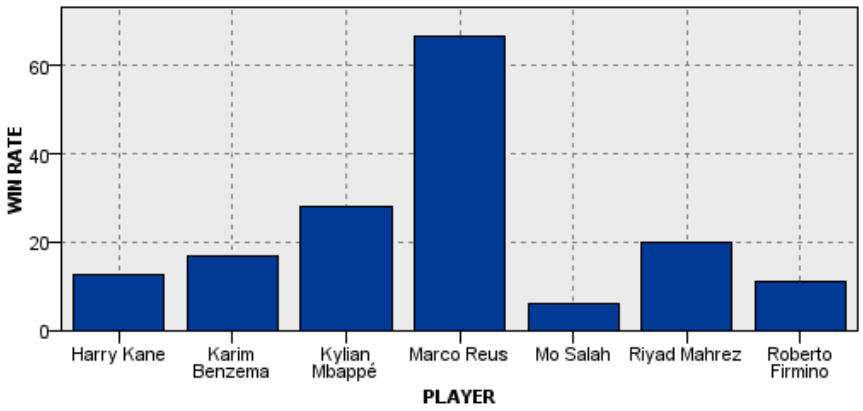
By eliminating null values, we can analyse statistics for the following players, starting with their win rate at home:



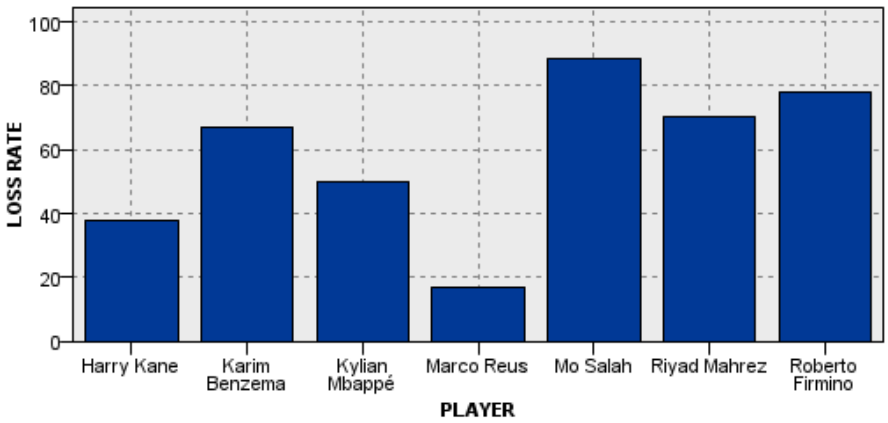
We can see that Mbappé, Muller, and Griezmann win approximately 4 matches over 5 when playing at home, while Benzema has the lowest score with 3 over 5.



Still at home, Dusan Tadic, Marco Reus, and Antoine Griezmann have the highest loss rate of this ranking, which make sense if we look at their clubs (AJAX, Dortmund, and Atletico).

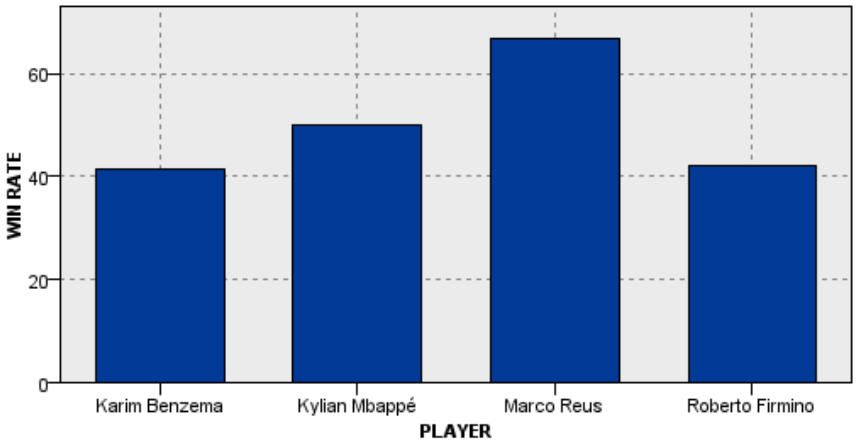


Now this is the win rate when playing away from home. We can see that Marco Reus over-performs, while Mbappé stands out, and most players wins at max 20% of their away games.

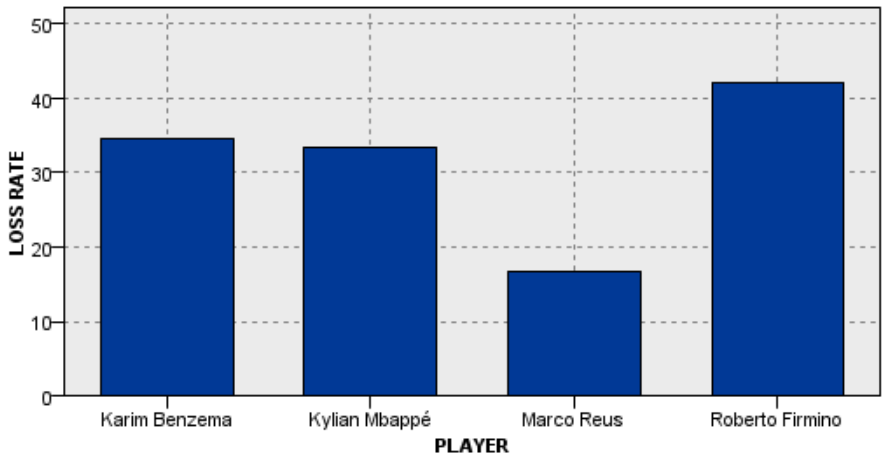


The loss rate away from home makes sense, as players with low win rate have highest loss rate, and players with high win rate have the lowest loss rate.

Now let's look at the general wins/loss percentages per player, home or away. Eliminating players with null values we find the following:



All four players supposedly win at least half or their match, and Marco Reus is the player with best win rate among players with at least10 passes and 10 goals.

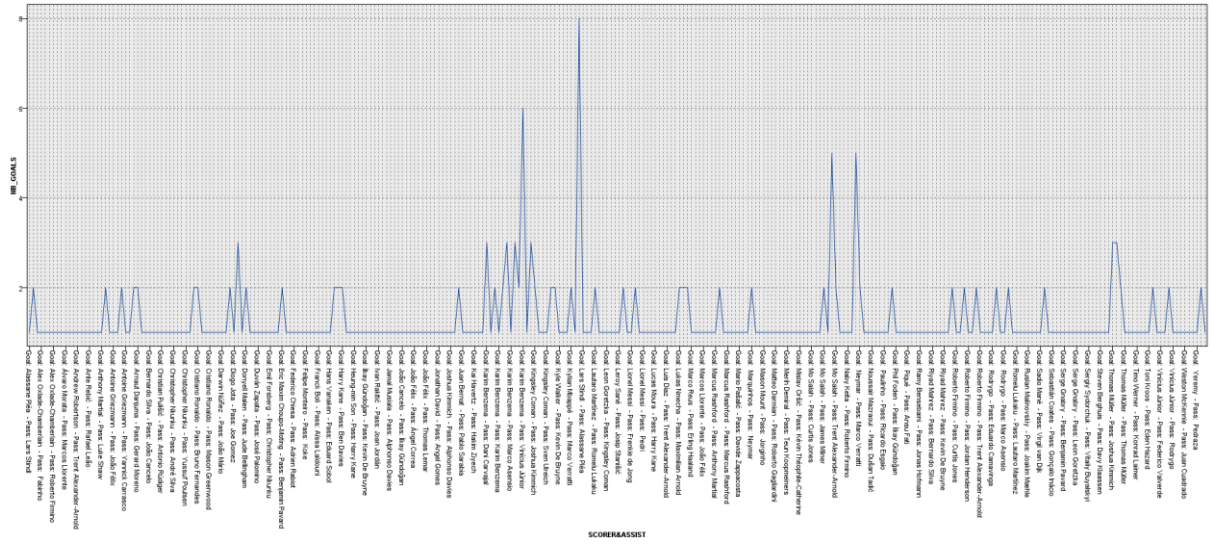


This graph shows that if he wins most of the time, Reus also rarely loses, while the others have more or less equal chances to win or lose.

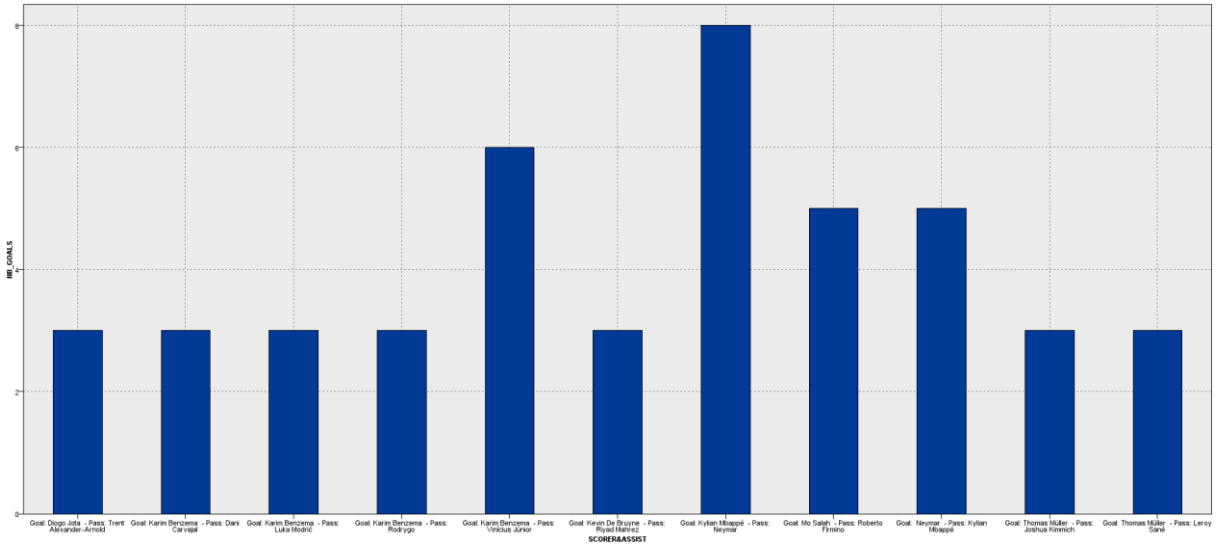
Individual indicators show that in overall, having Marco Reus in the team increase the chances of winning, and decrease the chances of losing. When playing at home, Mbappé and Müller are the most reliable players to win a match, while Marco Reus has high stakes to lose a match. But when playing away, Marco Reus has the highest chances to win, and Mo Salah the

highest chances to lose. We can say that Marco Reus is more reliable when playing away, while Mo Salah, Mbappé, or Müller are more comfortable when playing at home.

Are there specific player linkups (goal scorer and assist provider combination) that significantly influence the likelihood of scoring goals?



Here we have too many values, let's narrow down to at least 3 linkups.



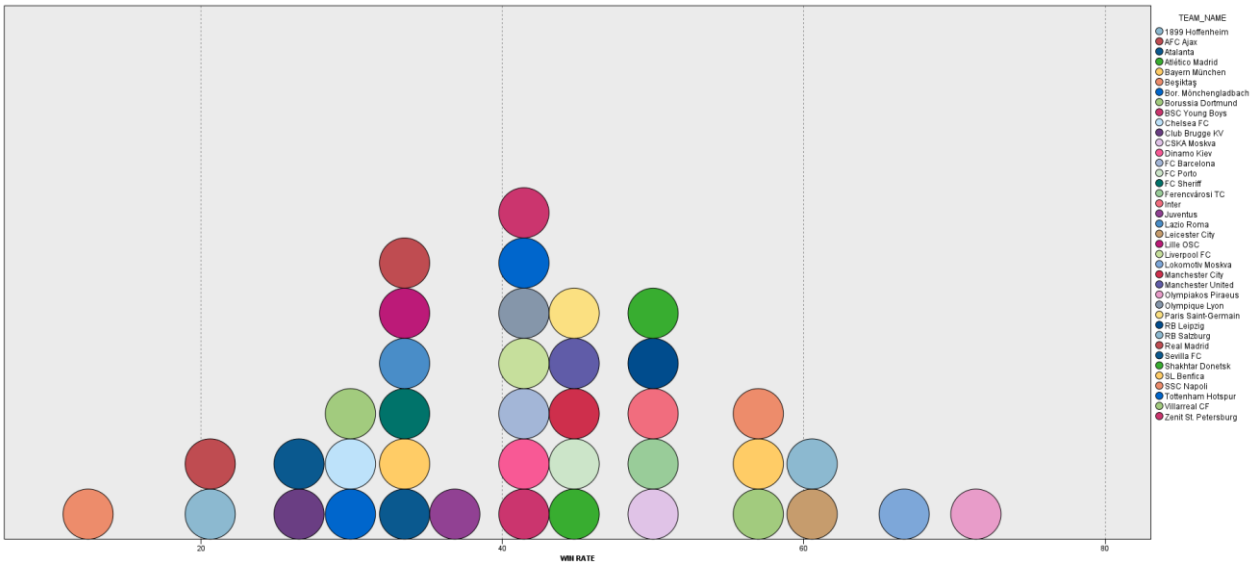
Here we can see that four linkups stand out: Firmino to Salah and Mbappé to Neymar with 5 goals each, Vinicius to Benzema with 6 goals, and Neymar to Mbappé with 8 goals.

Thus, during a game, we can statistically expect Vinicius to deliver an assist to Benzema, Firmino to deliver an assist to Salah, but most probably, Mbappé or Neymar delivering an assist to each other.

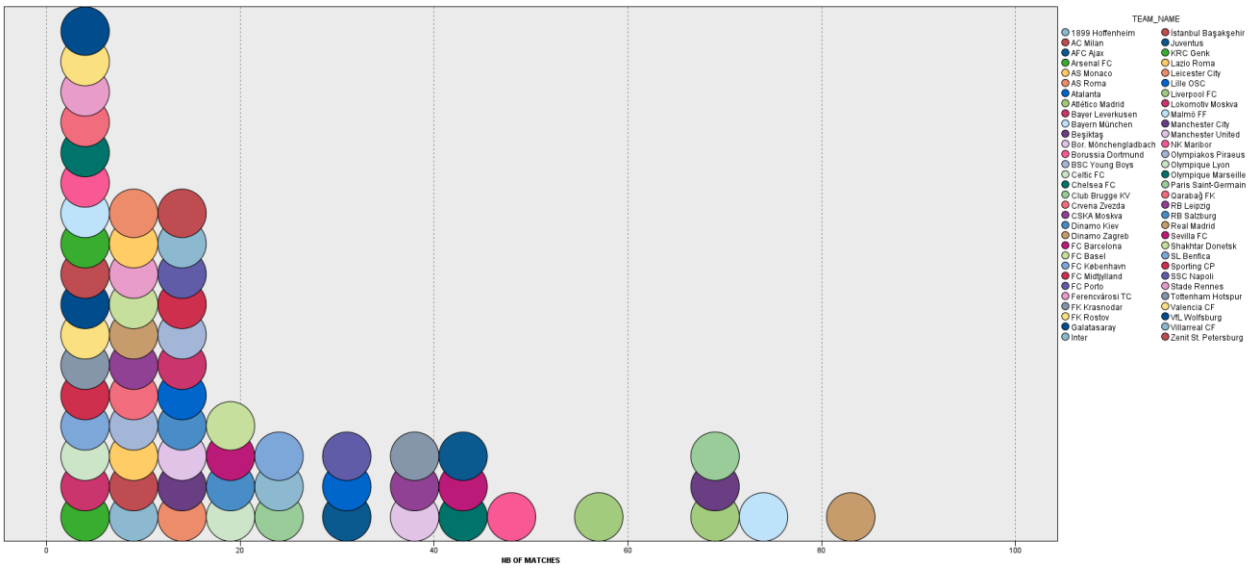
Do teams with winning history demonstrate a higher probability of winning matches?

This question can be formulated as “Do teams with most appearances have a better winning rate?”. Therefore, let’s compare appearances and win rates for all teams.

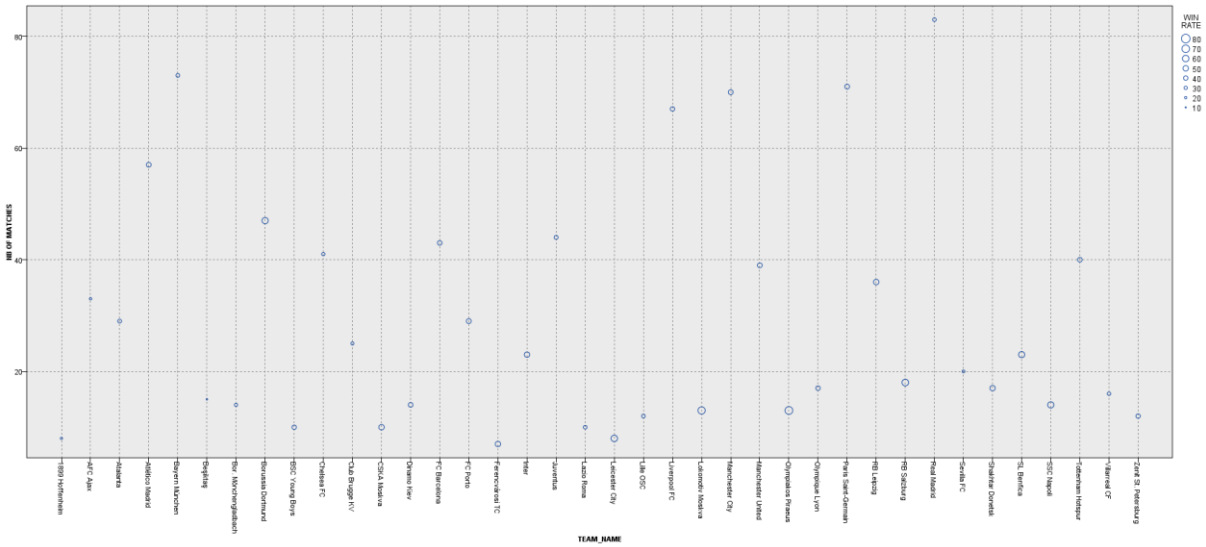
This is the win rate, home or away, for all teams:



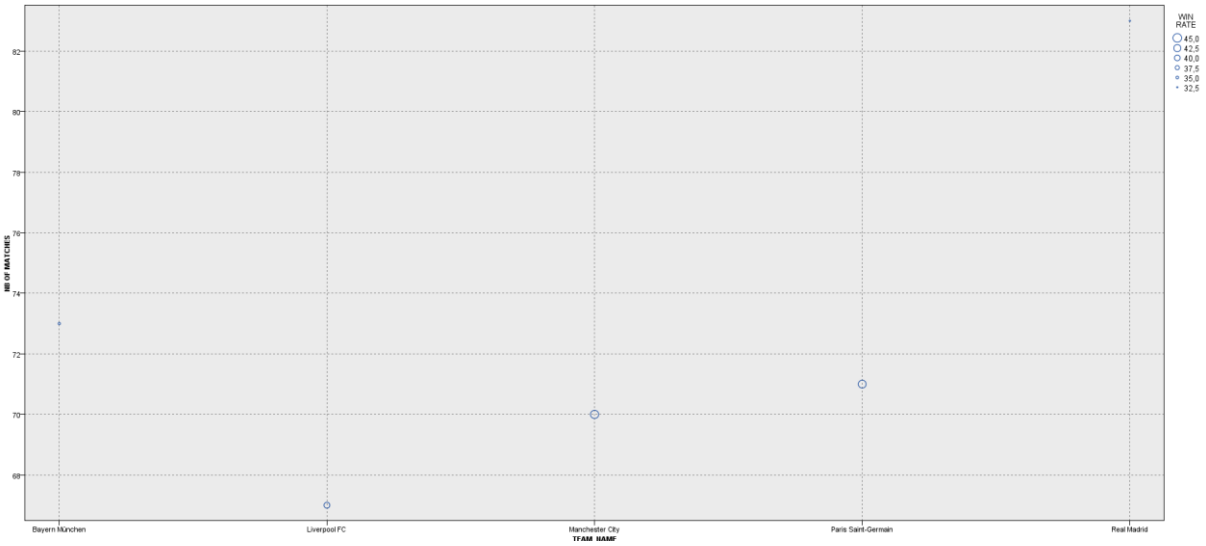
This is the number of matches played for all teams:



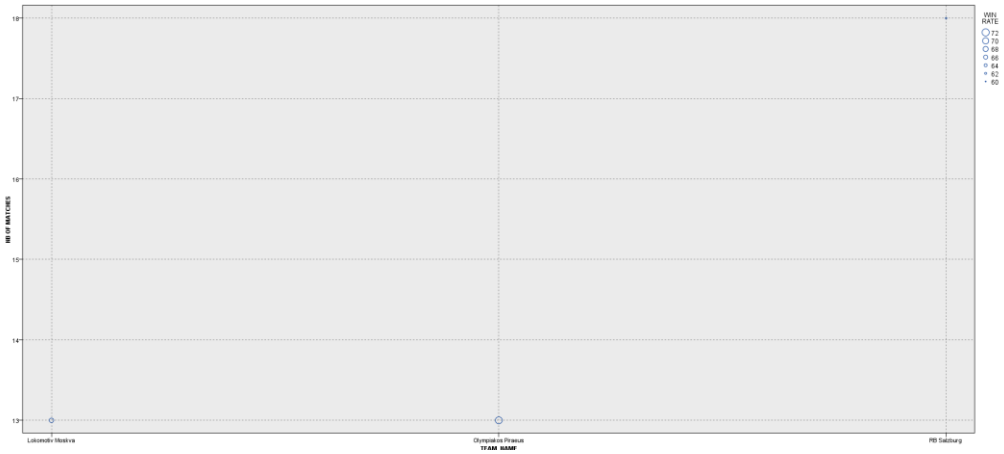
Now let's compare win rate to number of matches played:



Based on this graph, we can notice that specific teams stand out for their number of games played (more than 60%). Those 5 teams get between 30-to-45%-win rate:



But the 3 teams with best win rate have very low participation:



We can conclude that the win rate is not affected by the history of the club in UCL. But this statistic can be counterbalanced by the fact that the more you play games, the more your win rate is lowered.

The only conclusion we can draw from this question is that UCL is a balanced competition, where even outsiders can win against big names.

4.4 – Clustering Analysis

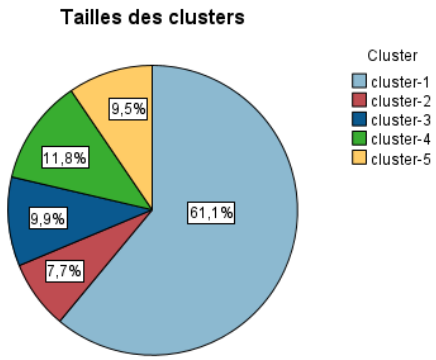
Our dataset is too wide to find relevant clusters. Therefore, separate it again.

We want to identify first types of players:

Champ	Mesure	Valeurs	Manquant	Vérifier	Rôle
TEAM_NAME	Nominal	"1899 Hoffe...		Aucun	Entrée
COUNTRY	Nominal	Austria,Azer...		Aucun	Entrée
HOME_STADIUM	Nominal	"Agia Sophia ...		Aucun	Entrée
CITY	Nominal	Amsterdam,...		Aucun	Entrée
CAPACITY	Continu	[4735.0,993...		Aucun	Entrée
MANAGER_FIR...	Nominal	"",Albert,Ale...		Aucun	Entrée
MANAGER_LAS...	Nominal	Abascal,Alle...		Aucun	Entrée
MANAGER_NA...	Nominal	Argentina,Au...		Aucun	Entrée
MANAGER_DOB	Continu	[1945-07-29...		Aucun	Entrée
PLAYER_ID	Sans type			Aucun	Aucun
PLAYER_FIRST...	Sans type			Aucun	Aucun
PLAYER_LAST_...	Sans type			Aucun	Aucun
PLAYER_NATI...	Nominal	Albania,Alger...		Aucun	Entrée
PLAYER_DOB	Continu	[1953-04-06...		Aucun	Entrée
PLAYER_POSIT...	Nominal	"02/2022","...		Aucun	Entrée
PLAYER_HEIGHT	Continu	[162.0,202.0]		Aucun	Entrée
PLAYER_WEIG...	Continu	[54.0,100.0]		Aucun	Entrée
PLAYER_FOOT	Nominal	"",L,R		Aucun	Entrée

This is the fields we're going to study. From using auto-clustering node, we know that with a silhouette of 0.313, K-means with 5 clusters is the most adapted clustering method.

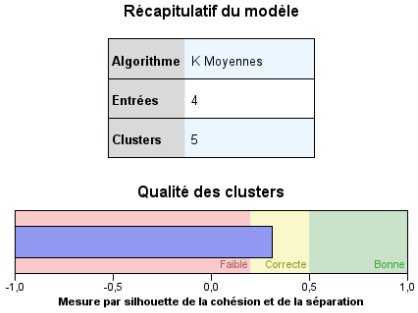
Utiliser ?	Graphique	Modèle	Durée de création (min)	Silhouette	Nombre de clusters	Plus petit cluster (N)	Plus petite cluster (%)	Plus grand cluster (N)	Plus grand cluster (%)	Plus petite/ Plus grande	Importance
☑		k moyenne 1	< 1	0,313	5	177	7	1407	61	0,126	0,0
☐		Kohonen 1	< 1	0,196	7	1	0	589	25	0,002	0,0
☐		TwoStep 1	< 1	0,044	2	813	42	1119	57	0,727	0,0



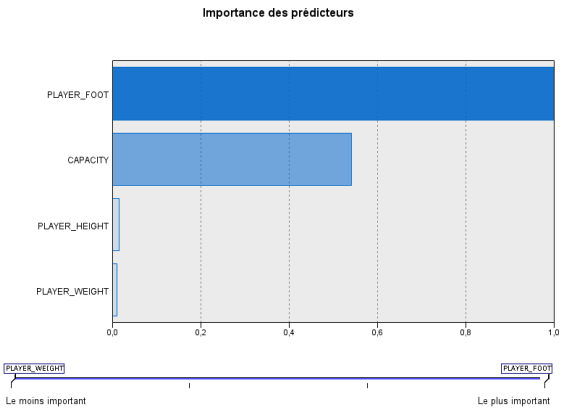
We can see that we have one major cluster, with 61,1% of the population, while the 4 others share approximately 10% each.

Taille du cluster le plus petit	177 (7,7%)
Taille du cluster le plus grand	1407 (61,1%)
Rapport des tailles : Cluster le plus grand par rapport au cluster le plus petit	7,95

The smallest cluster as a population of 177, while the biggest one has population of 1407, which is highly inequal.



The cluster quality is “correct”, but we can see that it is closer to “weak” than “good”.



In creating the clusters, the most important predictor was the player's foot, while his weight and height have almost no importance.

Here are the different clusters:

Cluster	cluster-1	cluster-4	cluster-3	cluster-5	cluster-2
Libellé					
Description					
Taille	61,1% (1407)	11,8% (271)	9,9% (229)	9,5% (219)	7,7% (177)
Entrées	PLAYER_FOOT R (100,0%)	PLAYER_FOOT L (100,0%)	PLAYER_FOOT L (100,0%)	PLAYER_FOOT (100,0%)	PLAYER_FOOT (100,0%)
	CAPACITY 45 404,47	CAPACITY 30 153,37	CAPACITY 62 419,95	CAPACITY 27 671,89	CAPACITY 65 165,55
	PLAYER_HEIGHT 183,16	PLAYER_HEIGHT 181,70	PLAYER_HEIGHT 181,00	PLAYER_HEIGHT 182,24	PLAYER_HEIGHT 182,54
	PLAYER_WEIGHT 76,29	PLAYER_WEIGHT 74,90	PLAYER_WEIGHT 74,21	PLAYER_WEIGHT 75,33	PLAYER_WEIGHT 75,50

We can define 5 profiles of players:

- Cluster 1 (61,1% of the population):
  - Player is right footed
  - He plays in a stadium of 45 404,47 seats
  - He is 183,16 cm tall
  - He weights 76,29 kg
- Cluster 4 (11,8% of the population):
  - Player is left footed
  - He plays in a stadium of 30 153,37 seats
  - He is 181,70 cm tall
  - He weights 74,90 kg

- Cluster 3 (9,9% of the population):
    - Player is left footed
    - He plays in a stadium of 62 419,95 seats
    - He is 181,00 cm tall
    - He weights 75,33 kg
- Cluster 5 (9,5% of the population):
    - Player is ambidextrous
    - He plays in a stadium of 27 671,89 seats
    - He is 182,24 cm tall
    - He weights 76,29 kg
- Cluster 2 (7,7% of the population):
    - Player is ambidextrous
    - He plays in a stadium of 65 165,55 seats
    - He is 182,54 cm tall
    - He weights 75,50 kg

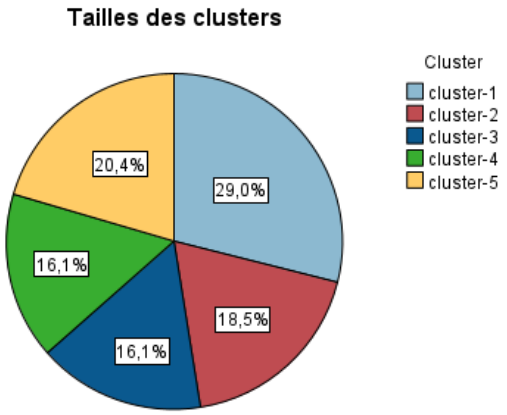
Despite similarities between clusters, we can conclude that 6 players over 10 are right footed, that all players are about 182 cm tall and 75 kg, and that only 16% (approximately) of the players play in stadiums of more than 60 000 seats.

Then types of goals:

Champ	Mesure	Valeurs	Manquant	Vérifier	Rôle
MATCH_ID	Sans type			Aucun	Aucun
SEASON	Nominal	"2016-2017"...		Aucun	Entrée
DATE_TIME	Nominal	"01-DEC-20 ...		Aucun	Entrée
HOME_TEAM	Nominal	"1899 Hoffe...		Aucun	Entrée
AWAY_TEAM	Nominal	"1899 Hoffe...		Aucun	Entrée
STADIUM	Nominal	"Alfredo Di S...		Aucun	Entrée
HOME_TEAM_...	Continu	[0.0,8.0]		Aucun	Entrée
AWAY_TEAM_...	Continu	[0.0,8.0]		Aucun	Entrée
ATTENDANCE	Continu	[0.0,98299.0]		Aucun	Entrée
GOAL_ID	Sans type			Aucun	Aucun
PID	Sans type			Aucun	Aucun
DURATION	Continu	[0.0,120.0]		Aucun	Entrée
ASSIST	Sans type			Aucun	Aucun
GOAL_DESC	Nominal	"","back heel...		Aucun	Entrée

This is the fields we're going to study. Again, from using auto-clustering node, we know that with a silhouette of 0.122, K-means with 5 clusters is the most adapted clustering method.

Utiliser ?	Graphique	Modèle	Durée de création (min)	Silhouette	Nombre de clusters	Plus petit cluster (N)	Plus petite cluster (%)	Plus grand cluster (N)	Plus grand cluster (%)	Plus petite/ Plus grande	Importance
☑		k moyenne 1	< 1	0,122	5	366	16	660	28	0,555	0,0
☐		Kohonen 1	< 1	0,012	10	3	0	638	27	0,005	0,0
☐		TwoStep 1	< 1	0,006	3	649	28	852	37	0,762	0,0



Clusters sizes here are more balanced, as they go from 16,1% to 29%.

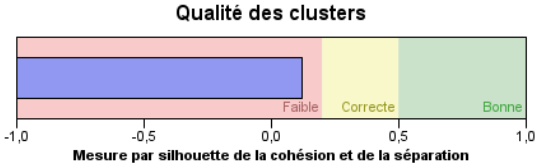
Taille du cluster le plus petit	366 (16,1%)
Taille du cluster le plus grand	660 (29%)
Rapport des tailles : Cluster le plus grand par rapport au cluster le plus petit	1,80

Population is also better split, with a smallest cluster of population 366 and biggest cluster of population 660.

But, despite the best clustering method and number of clusters, we can see that the quality of the clusters is “weak”, there is no need to go further.

Récapitulatif du modèle

Algorithme	K Moyennes
Entrées	6
Clusters	5

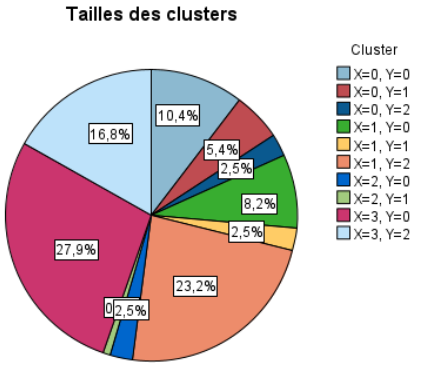


We can try to cluster goal scorers:

Champ	Mesure	Valeurs	Manquant	Vérifier	Rôle
PLAYER_ID	Sans type			Aucun	Aucun
PLAYER	Sans type			Aucun	Aucun
PLAYER_DOB	Continu	[1984-09-28...		Aucun	Entrée
PLAYER_FOOT	Nominal	"",L,R		Aucun	Entrée
PLAYER_HEIGHT	Continu	[165.0,196.0]		Aucun	Entrée
PLAYER_NATI...	Nominal	Albania,Alger...		Aucun	Entrée
PLAYER_POSIT...	Nominal	Defender,Fo...		Aucun	Entrée
PLAYER_WEIG...	Continu	[56.0,99.0]		Aucun	Entrée
MATCH_ID_Co...	Continu	[1,27]		Aucun	Entrée
GOAL_ID_Count	Continu	[1,40]		Aucun	Entrée

Again, from using auto-clustering node, we know that with a silhouette of 0.414, Kohonen clustering with 10 clusters is the most adapted clustering method.

Utiliser ?	Graphique	Modèle	Durée de création (min)	Silhouette	Nombre de clusters	Plus petit cluster (N)	Plus petite cluster (%)	Plus grand cluster (N)	Plus grand cluster (%)	Plus petite/ Plus grande	Importance
☑		Kohonen 1	< 1	0,414	10	2	0	78	27	0,026	0,0
☐		k moyenne 1	< 1	0,339	5	8	2	126	45	0,063	0,0
☐		TwoStep 1	< 1	0,280	2	108	38	172	61	0,628	0,0



We see that clusters sizes go from 0,7% to 27,9%, with two main clusters at 27,9 and 23,2%, a third one at 16,8%, and all others under 10%.

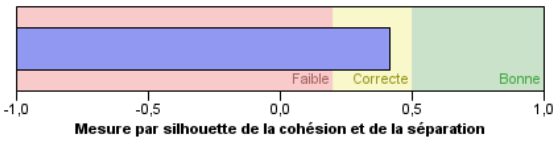
Taille du cluster le plus petit	2 (0,7%)
Taille du cluster le plus grand	78 (27,9%)
Rapport des tailles : Cluster le plus grand par rapport au cluster le plus petit	39,00

Cluster size is unbalanced, the smallest has only 2 individuals, and the biggest one has 78 individuals.

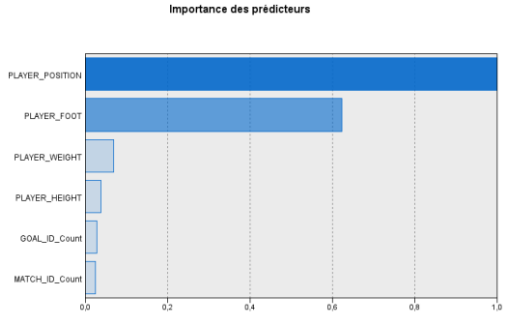
Récapitulatif du modèle

Algorithme	Kohonen
Entrées	6
Clusters	10

Qualité des clusters



The quality of the clustering method is “correct” and tends towards “good”.



Most important predictors are Position and Strong Foot, the rest is shared between height, weight, number of goals and number of matches.

Taille	27,9% (78)	23,2% (65)	16,8% (47)	10,4% (29)	8,2% (23)	5,4% (15)	2,5% (7)	2,5% (7)	2,5% (7)	0,7% (2)

Entrées										
	PLAYER_POSITION Forward (100,0%)	PLAYER_POSITION Midfielder (100,0%)	PLAYER_POSITION Defender (100,0%)	PLAYER_POSITION Defender (100,0%)	PLAYER_POSITION Forward (100,0%)	PLAYER_POSITION Midfielder (100,0%)	PLAYER_POSITION Midfielder (100,0%)	PLAYER_POSITION Midfielder (100,0%)	PLAYER_POSITION Forward (100,0%)	PLAYER_POSITION Goalkeeper (100,0%)
	PLAYER_FOOT R (100,0%)	PLAYER_FOOT R (100,0%)	PLAYER_FOOT R (95,7%)	PLAYER_FOOT L (100,0%)	PLAYER_FOOT L (100,0%)	PLAYER_FOOT L (100,0%)	PLAYER_FOOT R (100,0%)	PLAYER_FOOT (100,0%)	PLAYER_FOOT (100,0%)	PLAYER_FOOT R (100,0%)
	PLAYER_WEIGHT 75,71	PLAYER_WEIGHT 74,38	PLAYER_WEIGHT 79,72	PLAYER_WEIGHT 75,17	PLAYER_WEIGHT 74,74	PLAYER_WEIGHT 74,20	PLAYER_WEIGHT 60,57	PLAYER_WEIGHT 70,86	PLAYER_WEIGHT 72,71	PLAYER_WEIGHT 83,50
	PLAYER_HEIGHT 181,08	PLAYER_HEIGHT 181,11	PLAYER_HEIGHT 185,02	PLAYER_HEIGHT 180,38	PLAYER_HEIGHT 180,65	PLAYER_HEIGHT 181,27	PLAYER_HEIGHT 169,71	PLAYER_HEIGHT 178,71	PLAYER_HEIGHT 181,43	PLAYER_HEIGHT 186,00
	GOAL_ID_Count 5,06	GOAL_ID_Count 2,38	GOAL_ID_Count 1,77	GOAL_ID_Count 1,55	GOAL_ID_Count 5,70	GOAL_ID_Count 3,07	GOAL_ID_Count 1,71	GOAL_ID_Count 1,43	GOAL_ID_Count 6,29	GOAL_ID_Count 1,00
	MATCH_ID_Count 3,97	MATCH_ID_Count 2,20	MATCH_ID_Count 1,70	MATCH_ID_Count 1,48	MATCH_ID_Count 4,74	MATCH_ID_Count 3,00	MATCH_ID_Count 1,71	MATCH_ID_Count 1,43	MATCH_ID_Count 5,29	MATCH_ID_Count 1,00

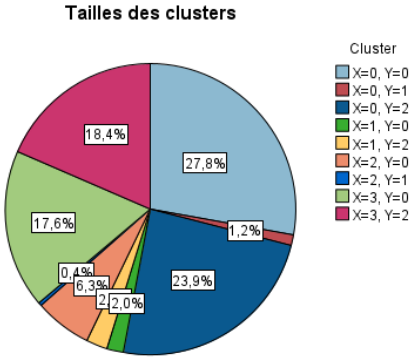
From those 10 clusters, we can see that 2 goalkeepers already scored 1 goal in UCL. Almost 40% of goal scorers are Forwards, 32% are Midfielders, and 26% are Defenders. Distribution of Left and Right footed players is almost qual for Defenders, where most Forwards and Midfielders are Right footed players. Forwards are 180 cm high, same for midfielders, and Defenders are a bit taller. Goal count for Forwards is around 5 goals, around 2 for Midfielders, and around 1,5 for Defenders. Finally, Forwards tends to have scored more matches (around 4,5), Midfielders scored in around 2,5 matches, and 1,5 for Defenders.

We can try to cluster assist providers:

Champ	Mesure	Valeurs	Manquant	Vérifier	Rôle
PLAYER_ID	Sans type			Aucun	Aucun
PLAYER	Sans type			Aucun	Aucun
PLAYER_DOB	Continu	[1984-06-28...		Aucun	Entrée
PLAYER_FOOT	Nominal	"",L,R		Aucun	Entrée
PLAYER_HEIGHT	Continu	[165.0,201.0]		Aucun	Entrée
PLAYER_NATI...	Nominal	Albania,Alger...		Aucun	Entrée
PLAYER_POSIT...	Nominal	Defender,Fo...		Aucun	Entrée
PLAYER_WEIG...	Continu	[56.0,93.0]		Aucun	Entrée
MATCH_ID_Co...	Continu	[1,17]		Aucun	Entrée
GOAL_ID_Count	Continu	[1,21]		Aucun	Entrée

Again, from using auto-clustering node, we know that with a silhouette of 0.363, Kohonen clustering with 9 clusters is the most adapted clustering method.

Utiliser ?	Graphique	Modèle	Durée de création (min)	Silhouette	Nombre de clusters	Plus petit cluster (N)	Plus petite cluster (%)	Plus grand cluster (N)	Plus grand cluster (%)	Plus petite/ Plus grande	Importance
☑		Kohonen 1	< 1	0,363	9	1	0	71	27	0,014	0,0
☐		k moyenne 1	< 1	0,323	5	9	3	75	29	0,12	0,0
☐		TwoStep 1	< 1	0,213	2	111	43	144	56	0,771	0,0



Most assist providers are shared between 4 clusters, from 27% of population to 17%, and the rest is divided in clusters of less than 6% of population.

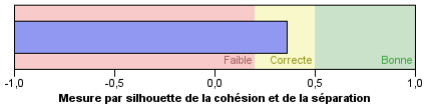
Taille du cluster le plus petit	1 (0,4%)
Taille du cluster le plus grand	71 (27,8%)
Rapport des tailles : Cluster le plus grand par rapport au cluster le plus petit	71,00

Here again, cluster sizes are unbalanced, with a smallest cluster of 1 individual, and a biggest cluster of 71%.

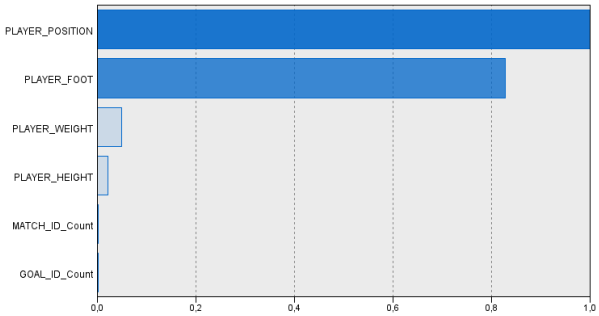
Récapitulatif du modèle

Algorithme	Kohonen
Entrées	6
Clusters	9

Qualité des clusters



Cluster quality is “correct”, with a perfect balance between “weak” and “good”.



Again, most important predictors are Position and Strong Foot, the rest is shared between height, weight, number of goals and number of matches, Although, strong foot and position almost share the same importance.

Taille									
	27,8% (71)	23,9% (61)	18,4% (47)	17,6% (45)	6,3% (16)	2,4% (6)	2,0% (5)	1,2% (3)	0,4% (1)

Entrées									
	PLAYER_POSITION Midfielder (100,0%)	PLAYER_POSITION Forward (100,0%)	PLAYER_POSITION Defender (100,0%)	PLAYER_POSITION Defender (60,0%)	PLAYER_POSITION Midfielder (100,0%)	PLAYER_POSITION Forward (100,0%)	PLAYER_POSITION Midfielder (100,0%)	PLAYER_POSITION Goalkeeper (100,0%)	PLAYER_POSITION Defender (100,0%)
	PLAYER_FOOT R (100,0%)	PLAYER_FOOT R (100,0%)	PLAYER_FOOT R (100,0%)	PLAYER_FOOT L (100,0%)	PLAYER_FOOT L (100,0%)	PLAYER_FOOT (100,0%)	PLAYER_FOOT (100,0%)	PLAYER_FOOT R (100,0%)	PLAYER_FOOT (100,0%)
	PLAYER_WEIGHT 72,83	PLAYER_WEIGHT 75,79	PLAYER_WEIGHT 78,53	PLAYER_WEIGHT 75,00	PLAYER_WEIGHT 72,88	PLAYER_WEIGHT 72,33	PLAYER_WEIGHT 69,80	PLAYER_WEIGHT 89,67	PLAYER_WEIGHT 77,00
	PLAYER_HEIGHT 179,80	PLAYER_HEIGHT 181,49	PLAYER_HEIGHT 183,72	PLAYER_HEIGHT 180,22	PLAYER_HEIGHT 179,44	PLAYER_HEIGHT 181,00	PLAYER_HEIGHT 179,80	PLAYER_HEIGHT 191,67	PLAYER_HEIGHT 186,00
	MATCH_ID_Count 2,58	MATCH_ID_Count 3,10	MATCH_ID_Count 1,91	MATCH_ID_Count 2,73	MATCH_ID_Count 2,44	MATCH_ID_Count 2,00	MATCH_ID_Count 3,60	MATCH_ID_Count 1,00	MATCH_ID_Count 1,00
	GOAL_ID_Count 2,80	GOAL_ID_Count 3,38	GOAL_ID_Count 2,04	GOAL_ID_Count 3,04	GOAL_ID_Count 2,44	GOAL_ID_Count 2,00	GOAL_ID_Count 3,60	GOAL_ID_Count 1,00	GOAL_ID_Count 1,00

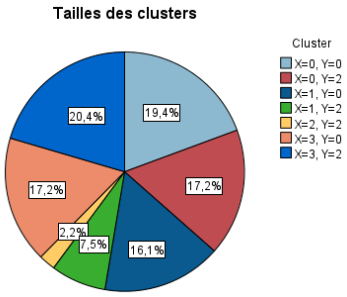
From those 9 clusters, we can see that 3 goalkeepers already provided 1 assist in UCL. Almost 36% of assist providers are Midfielders, 26% are Forwards, and 36% are Defenders. Distribution of Left and Right footed players is almost qual for Defenders, where most Forwards and Midfielders are Right footed players. Conclusions about players size is the same as for goal scorers. Matches with assist provided count for Forwards and Midfielders is around 2,5 matches, and around 2 for Defenders (but left footed defenders provide assist more often than right footed defenders). Finally, Forwards tends to provide around 2,8 assists, Midfielders around 2,6 assists, and defenders 2,5 (again left footed defenders provide more than right footed defenders).

Finally, we can try to cluster teams with their average number of goals scored and conceded per match:

Champ	Mesure	Valeurs	Manquant	Vérifier	Rôle
TEAM_NAME	Nominal	"Atlético Ma...		Aucun	Entrée
CAPACITY	Continu	[48712.0,81...		Aucun	Entrée
CITY	Nominal	Liverpool,Ma...		Aucun	Entrée
COUNTRY	Nominal	England,Fran...		Aucun	Entrée
HOME_STADIUM	Nominal	"Allianz Aren...		Aucun	Entrée
MANAGER_DOB	Continu	[1959-06-10...		Aucun	Entrée
MANAGER_FIR...	Nominal	Carlo,Christo...		Aucun	Entrée
MANAGER_LAS...	Nominal	Ancelotti,Gal...		Aucun	Entrée
MANAGER_NA...	Nominal	Argentina,Fr...		Aucun	Entrée
PLAYER_DOB	Continu	[1985-09-09...		Aucun	Entrée
PLAYER_FIRST...	Nominal	"",Alex,Alph...		Aucun	Aucun
PLAYER_FOOT	Nominal	"",L,R		Aucun	Entrée
PLAYER_HEIGHT	Continu	[165.0,195.0]		Aucun	Entrée
PLAYER_ID	Nominal	ply103,ply10...		Aucun	Aucun
PLAYER_LAST_...	Nominal	Aké,Alaba,Al...		Aucun	Aucun
PLAYER_NATI...	Nominal	Algeria,Arge...		Aucun	Entrée
PLAYER_POSIT...	Nominal	Defender,Fo...		Aucun	Entrée
PLAYER_WEIG...	Continu	[60.0,95.0]		Aucun	Entrée
GOALS SCORE...	Continu	[1.0,2.0]		Aucun	Entrée
GOALS CONDE...	Continu	[1.81666666...		Aucun	Entrée

Again, from using auto-clustering node, we know that with a silhouette of 0.549, Kohonen clustering with 7 clusters is the most adapted clustering method.

Utiliser ?	Graphique	Modèle	Durée de création (min)	Silhouette	Nombre de clusters	Plus petit cluster (N)	Plus petite cluster (%)	Plus grand cluster (N)	Plus grand cluster (%)	Plus petite/ Plus grande	Importance
☑		Kohonen 1	< 1	0,549	7	2	2	19	20	0,105	0,0
☐		k moyen...	< 1	0,504	5	9	9	35	37	0,257	0,0
☐		TwoStep 1	< 1	0,292	3	25	26	35	37	0,714	0,0



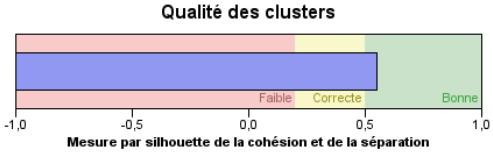
Most teams are separated between 5 clusters, from 20% of population to 16%, and the rest is divided in clusters of less than 7,5% of population.

Taille du cluster le plus petit	2 (2,2%)
Taille du cluster le plus grand	19 (20,4%)
Rapport des tailles : Cluster le plus grand par rapport au cluster le plus petit	9,50

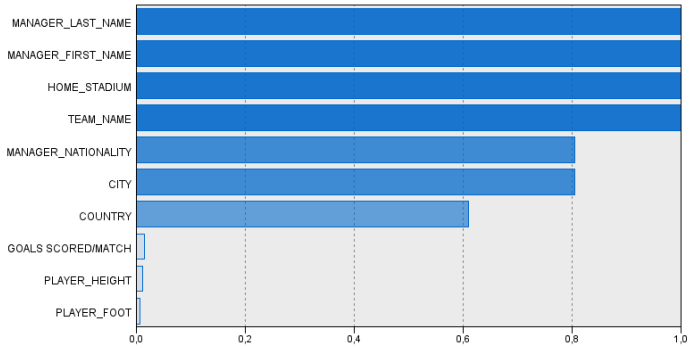
Again, distribution is unbalanced, with smallest cluster of 2 teams, and biggest cluster of 19 teams.

Récapitulatif du modèle

Algorithme	Kohonen
Entrées	14
Clusters	7



Cluster quality is “good”.



Among the predicators, the most important are the manager, the home stadium, and the team's name itself. Then manager's nationality, and the city and country of the club also have importance.

Taille	20,4% (19)	19,4% (18)	17,2% (16)	17,2% (16)	16,1% (15)	7,5% (7)	2,2% (2)
Entrées	HOME_STADIUM Anfield (100,0%)	HOME_STADIUM	HOME_STADIUM	HOME_STADIUM	HOME_STADIUM	HOME_STADIUM	HOME_STADIUM
	MANAGER_FIRST_NAME	MANAGER_FIRST_NAME	MANAGER_FIRST_NAME	MANAGER_FIRST_NAME	MANAGER_FIRST_NAME	MANAGER_FIRST_NAME	MANAGER_FIRST_NAME
	MANAGER_LAST_NAME	MANAGER_LAST_NAME	MANAGER_LAST_NAME	MANAGER_LAST_NAME	MANAGER_LAST_NAME	MANAGER_LAST_NAME	MANAGER_LAST_NAME
	TEAM_NAME	TEAM_NAME	TEAM_NAME	TEAM_NAME	TEAM_NAME	TEAM_NAME	TEAM_NAME
	CITY Liverpool (100,0%)	CITY Madrid (100,0%)	CITY Manchester (100,0%)	CITY München (100,0%)	CITY Madrid (100,0%)	CITY Paris (100,0%)	CITY Paris (100,0%)

MANAGER_NATIONALITY	MANAGER_NATIONALITY	MANAGER_NATIONALITY	MANAGER_NATIONALITY	MANAGER_NATIONALITY	MANAGER_NATIONALITY	MANAGER_NATIONALITY
COUNTRY England (100,0%)	COUNTRY Spain (100,0%)	COUNTRY England (100,0%)	COUNTRY Germany (100,0%)	COUNTRY Spain (100,0%)	COUNTRY France (100,0%)	COUNTRY France (100,0%)
GOALS SCORED/MATCH	GOALS SCORED/MATCH	GOALS SCORED/MATCH	GOALS SCORED/MATCH	GOALS SCORED/MATCH	GOALS SCORED/MATCH	GOALS SCORED/MATCH
PLAYER_HEIGHT 182,32	PLAYER_HEIGHT 179,39	PLAYER_HEIGHT 182,81	PLAYER_HEIGHT 184,19	PLAYER_HEIGHT 181,53	PLAYER_HEIGHT 176,86	PLAYER_HEIGHT 176,50
PLAYER_FOOT R (84,2%)	PLAYER_FOOT R (72,2%)	PLAYER_FOOT L (56,2%)	PLAYER_FOOT R (75,0%)	PLAYER_FOOT R (66,7%)	PLAYER_FOOT L (57,1%)	PLAYER_FOOT R (100,0%)
PLAYER_WEIGHT 75,53	PLAYER_WEIGHT 74,06	PLAYER_WEIGHT 76,25	PLAYER_WEIGHT 79,12	PLAYER_WEIGHT 75,53	PLAYER_WEIGHT 72,57	PLAYER_WEIGHT 70,50
PLAYER_POSITION Defender (36,8%)	PLAYER_POSITION Defender (38,9%)	PLAYER_POSITION Defender (43,8%)	PLAYER_POSITION Forward (37,5%)	PLAYER_POSITION Midfielder (40,0%)	PLAYER_POSITION Defender (57,1%)	PLAYER_POSITION Forward (100,0%)
CAPACITY 54 074,00	CAPACITY 81 044,00	CAPACITY 55 097,00	CAPACITY 75 024,00	CAPACITY 68 000,00	CAPACITY 48 712,00	CAPACITY 48 712,00
GOALS CONDEDED/MATCH	GOALS CONDEDED/MATCH	GOALS CONDEDED/MATCH	GOALS CONDEDED/MATCH	GOALS CONDEDED/MATCH	GOALS CONDEDED/MATCH	GOALS CONDEDED/MATCH

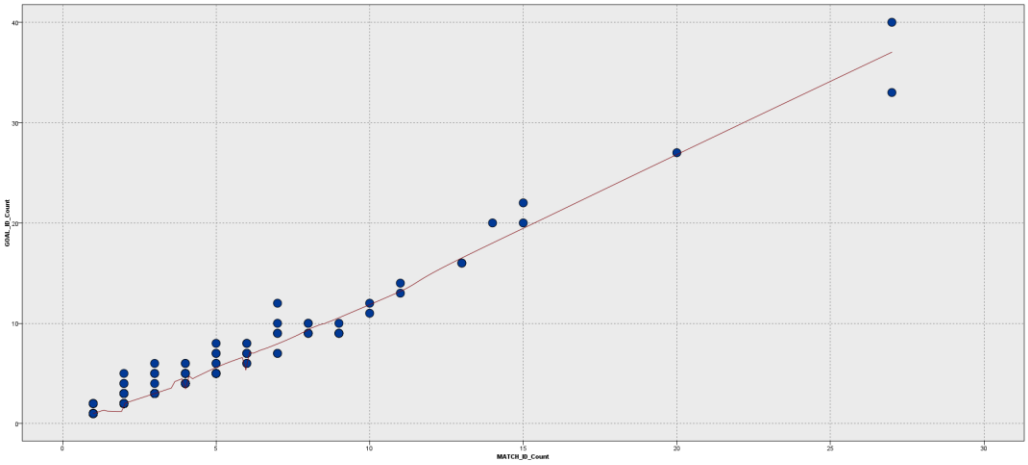
Clusters are built around actual teams, with their manager and home stadium.

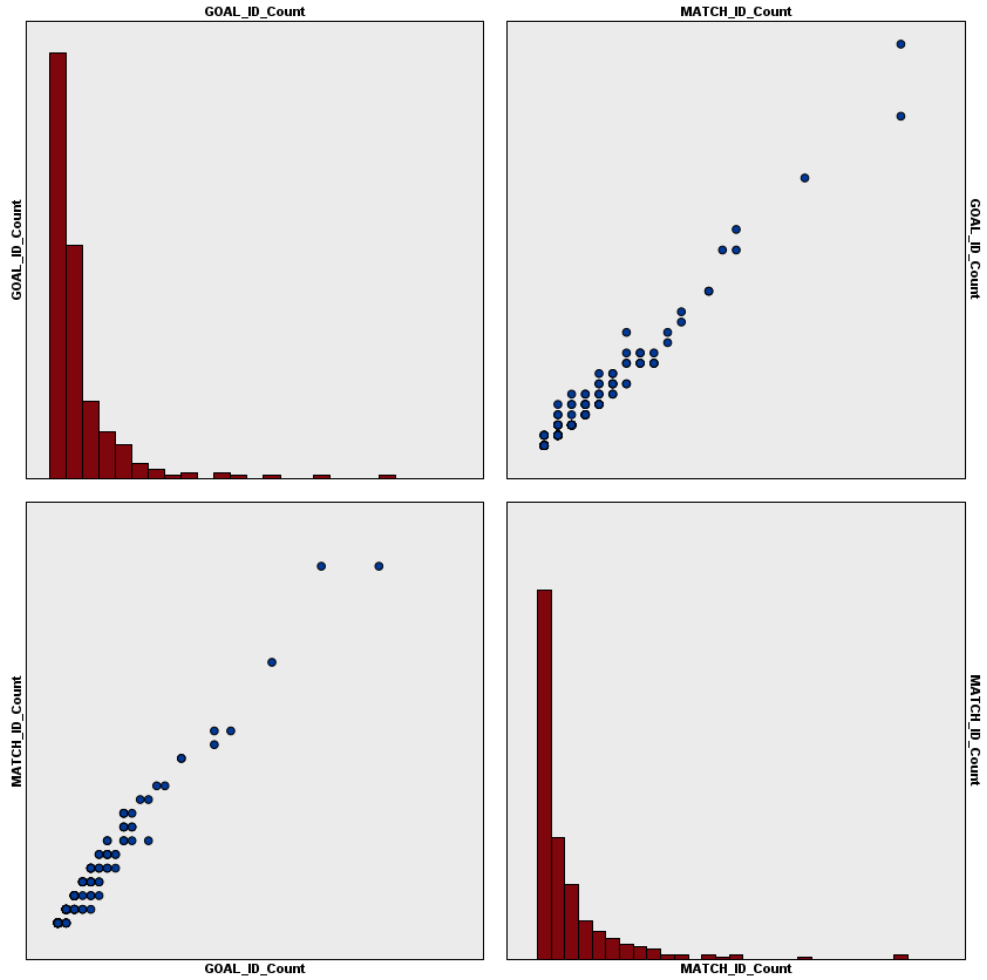
4.5 – RFM Analysis

As the study gives no access to recency, frequency, and monetary data, it is not possible to conduct RFM analysis.

4.6 – Correlation between numerical variables

- Between Number of matches and Number of goals for goal scorers:





MATCH_ID_Count

Statistiques

Effectif	280
Moyenne	2.825
Min	1
Max	27
Plage	26
Variance	12.123
Ecart type	3.482
Erreur standard de la moyenne	0.208

GOAL_ID_Count

Statistiques

Effectif	280
Moyenne	3.296
Min	1
Max	40
Plage	39
Variance	21.335
Ecart type	4.619
Erreur standard de la moyenne	0.276

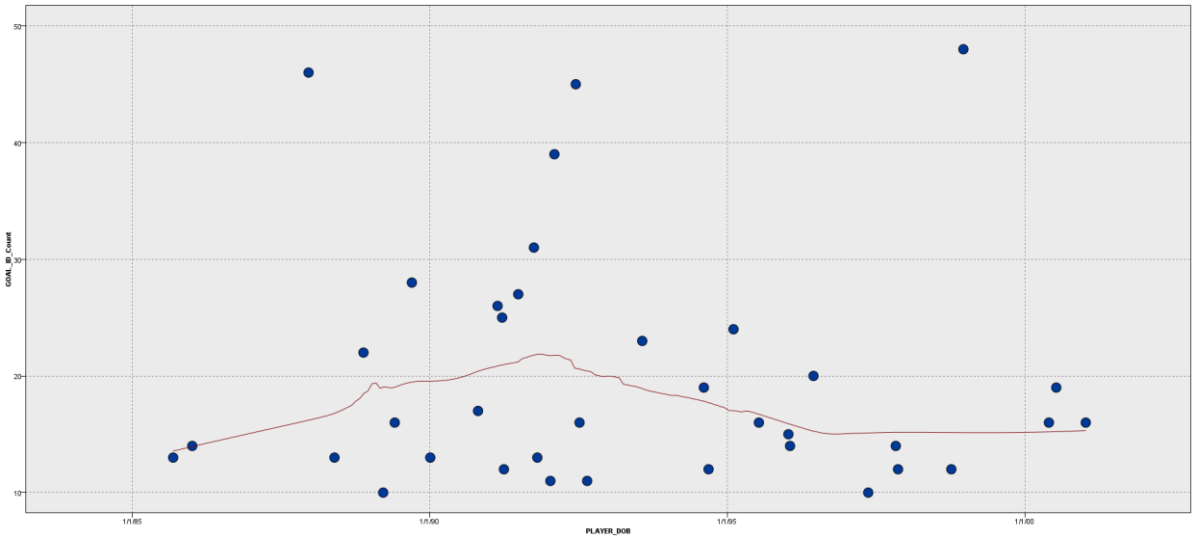
Corrélations de Pearson

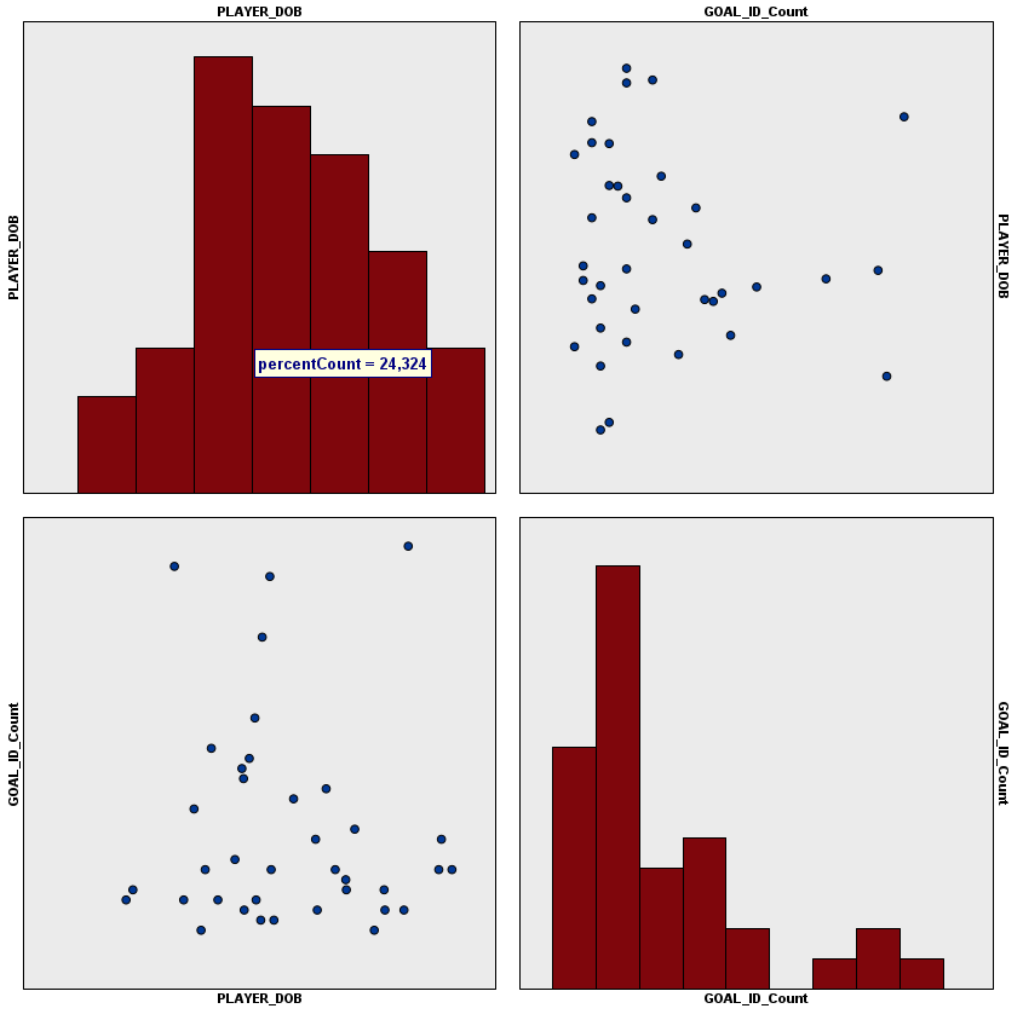
	MATCH_ID_Count	GOAL_ID_Count
MATCH_ID_Count	1.000/Parfait	0.986/Elevé
GOAL_ID_Count	0.986/Elevé	1.000/Parfait

There is a “Strong” (0.986) correlation between number of games played and number of scored goals for goal scorers. Meaning that a player with a lot of goals is expected to have played a lot of matches, and a goal scorer with a lot of matches is expected to have scored a lot of goals.

- Between Date of Birth and Number of goals for player:

For this question, we select players with at least 10 goals.

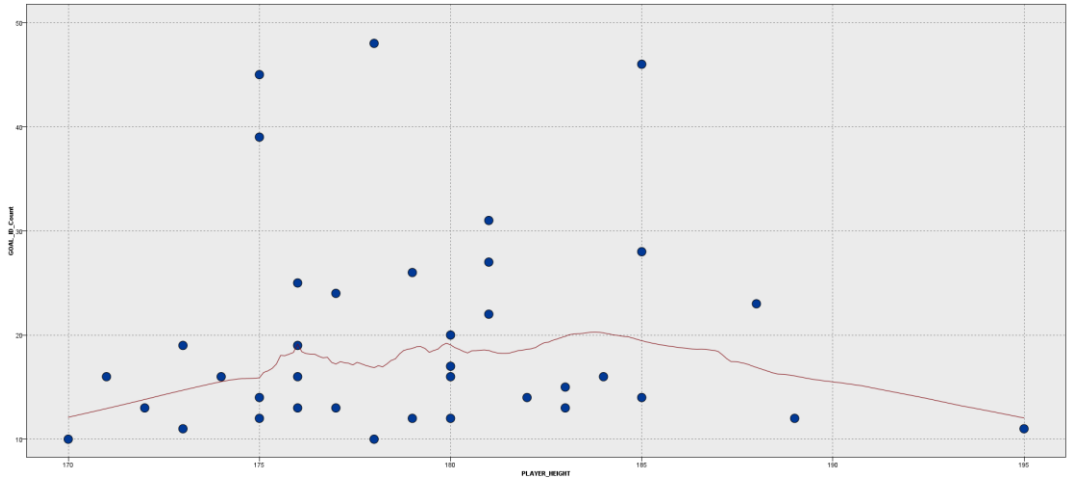


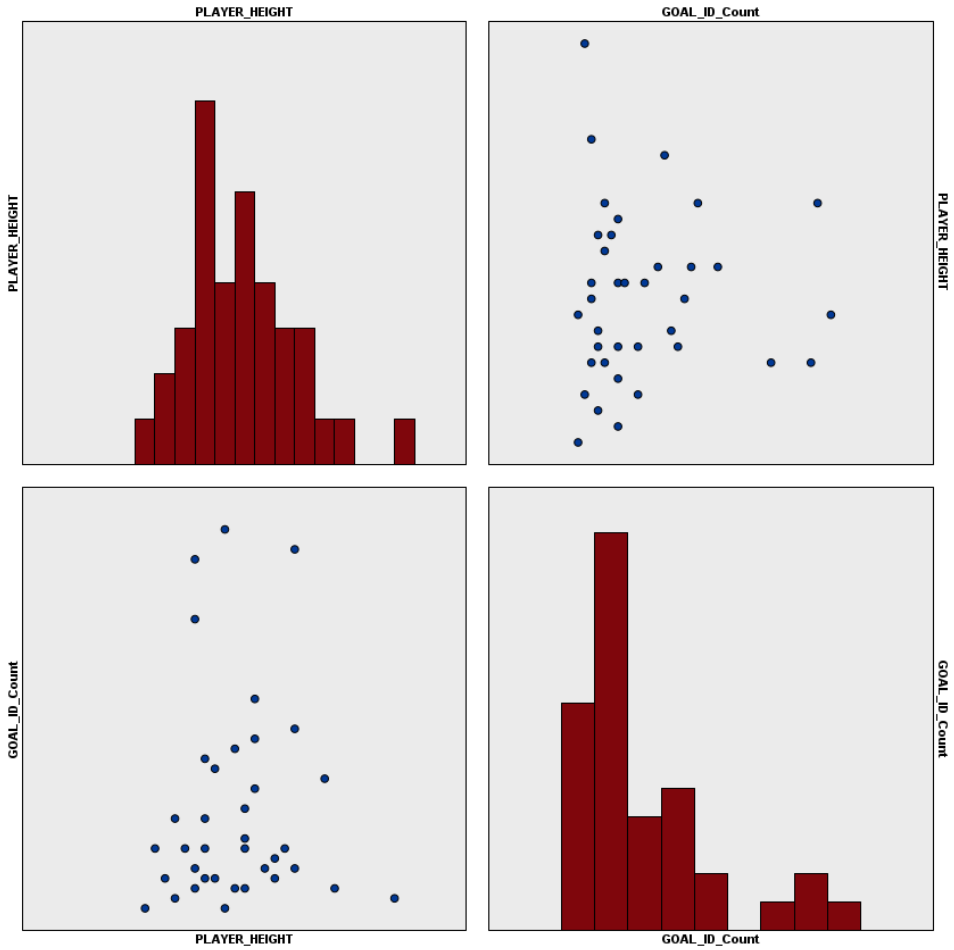


There doesn't seem to be any specific correlation between age and number of goals scored in the top scorers of UCL.

- Between Height and Number of goals for player:

Again, we reduce the dataset to players with more than 10 goals.





PLAYER_HEIGHT

Statistiques

Effectif	37
Moyenne	179.108
Min	170.000
Max	195.000
Plage	25.000
Variance	28.544
Ecart type	5.343
Erreur standard de la moyenne	0.878

GOAL_ID_Count

Statistiques

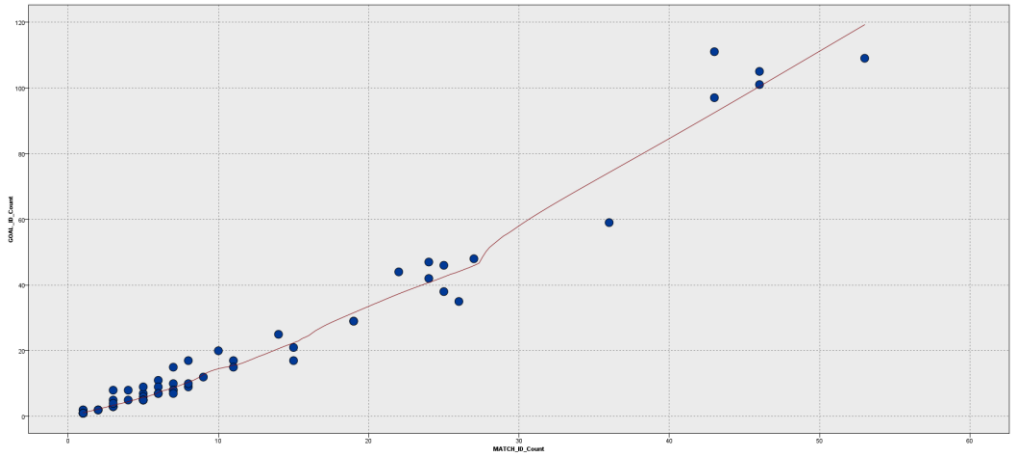
Effectif	37
Moyenne	19.946
Min	10
Max	48
Plage	38
Variance	105.775
Ecart type	10.285
Erreur standard de la moyenne	1.691

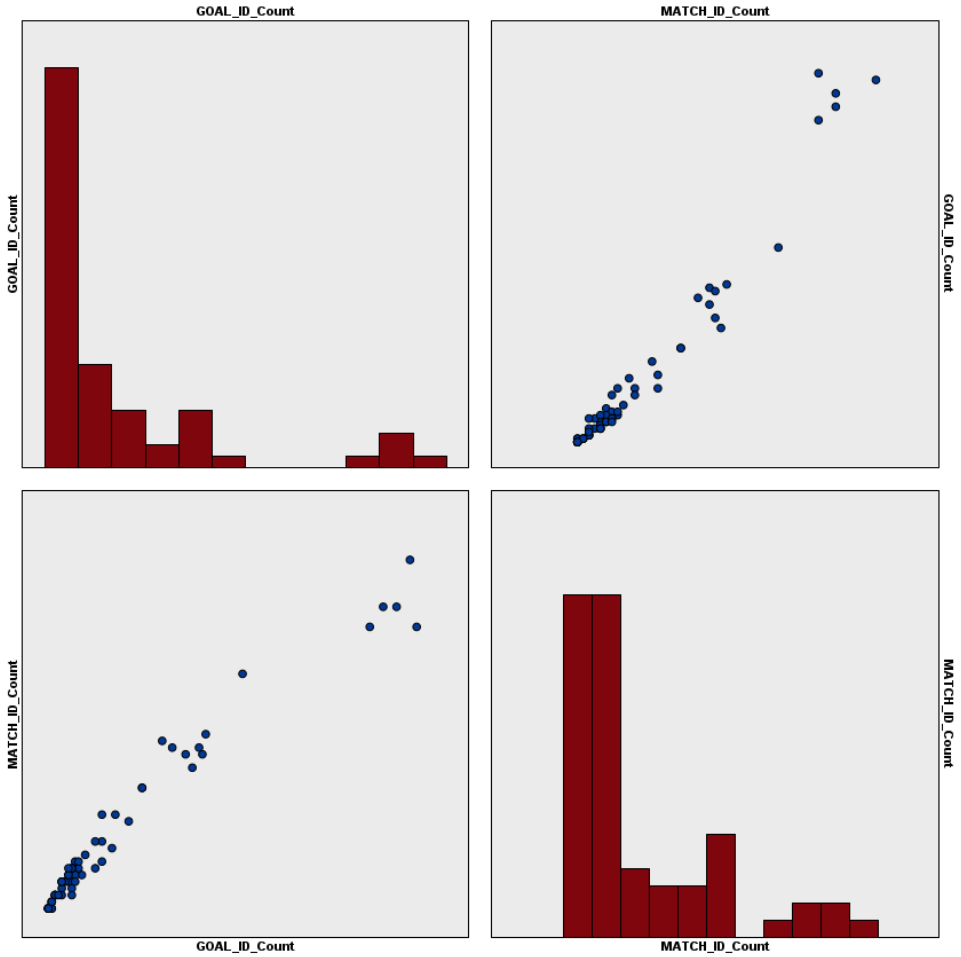
Corrélations de Pearson

	PLAYER_HEIGHT	GOAL_ID_Count
PLAYER_HEIGHT	1.000/Parfait	0.025/Faible
GOAL_ID_Count	0.025/Faible	1.000/Parfait

There is only a “Weak” (0.025) correlation between player’s height and his ability to score goals.

- Between Number of matches and Number of goals for goal scoring teams:





MATCH_ID_Count

Statistiques

Effectif	62
Moyenne	11.629
Min	1
Max	53
Plage	52
Variance	173.450
Ecart type	13.170
Erreur standard de la moyenne	1.673

GOAL_ID_Count

Statistiques

Effectif	62
Moyenne	20.677
Min	1
Max	111
Plage	110
Variance	830.714
Ecart type	28.822
Erreur standard de la moyenne	3.660

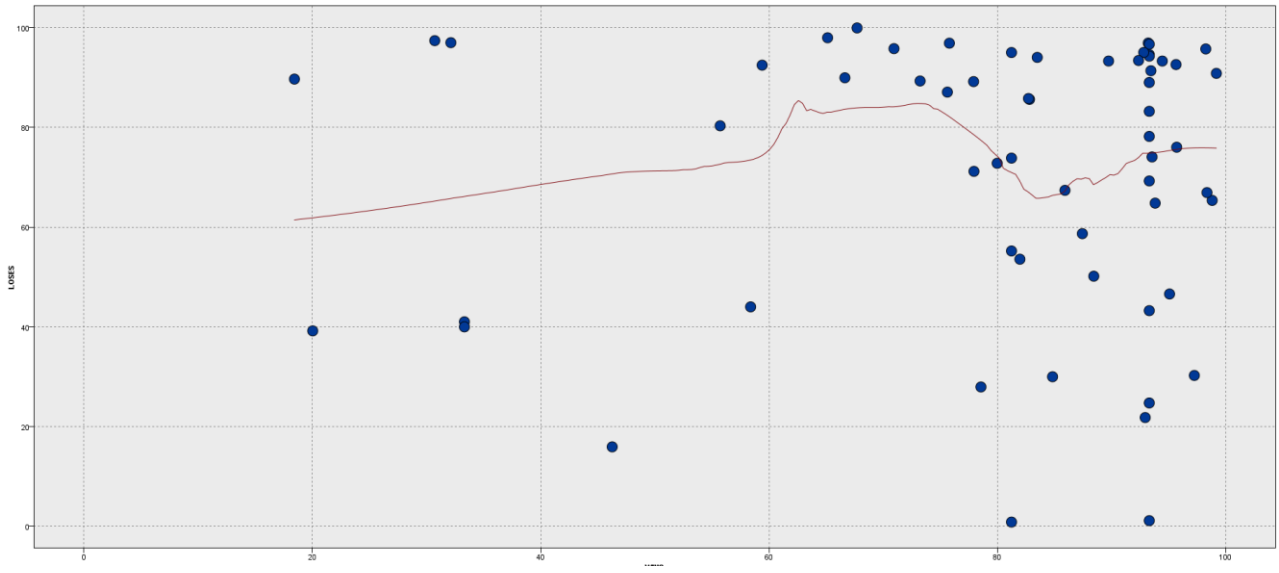
Corrélations de Pearson

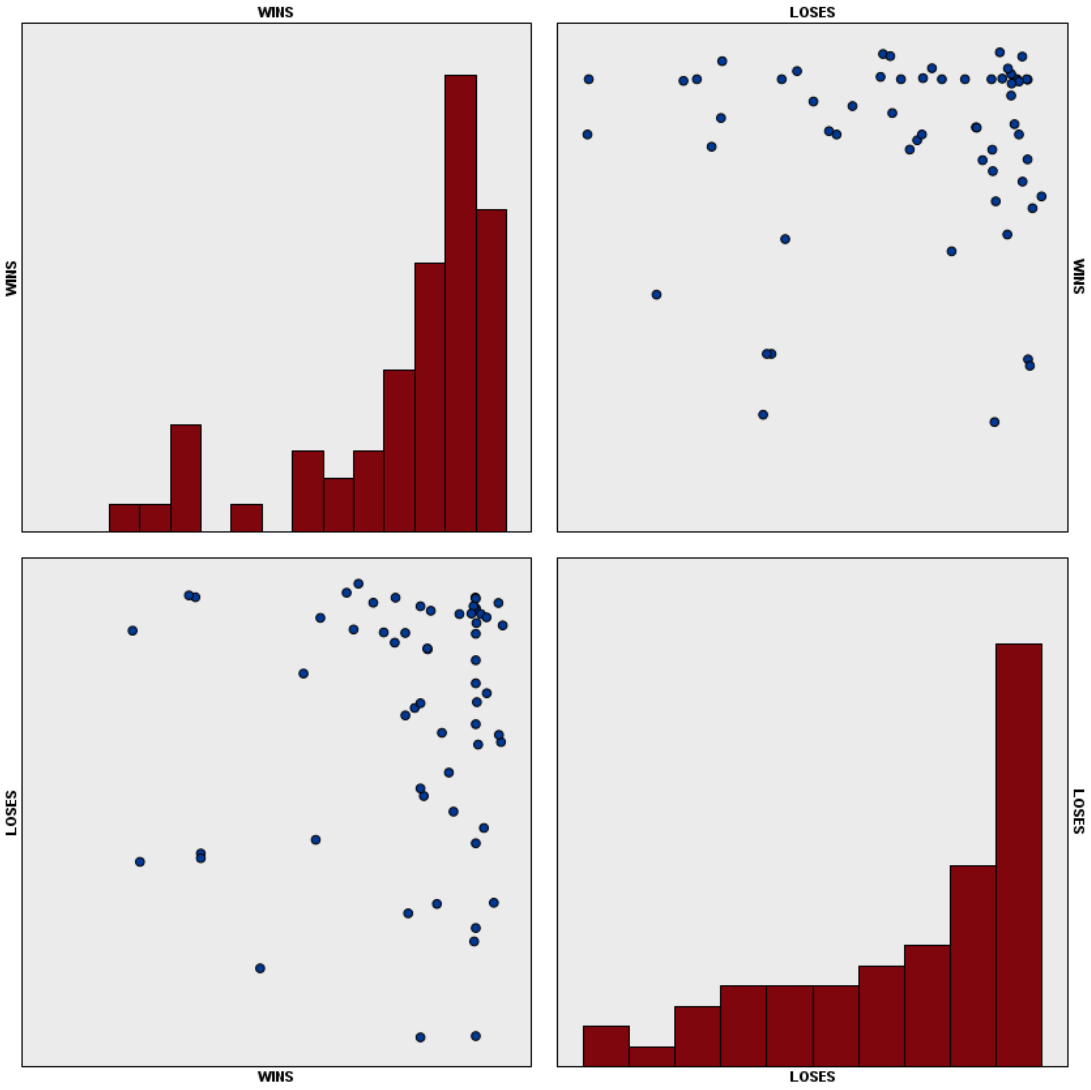
	MATCH_ID_Count	GOAL_ID_Count
MATCH_ID_Count	1.000/Parfait	0.980/Elevé
GOAL_ID_Count	0.980/Elevé	1.000/Parfait

There is a “Strong” (0.98) correlation between number of games played and number of scored goals for goal scoring team. Meaning that a team historically scoring a lot of goals is expected to have played a lot of matches, and a team historically playing a lot of matches is expected to have scored a lot of goals.

- Between Home Win and Attendance over Capacity:

This question aims to find a correlation between whether playing in front of a fully filled with crowd stadium influences a team in winning a game at home.





WINS

Statistiques

Effectif	60
Moyenne	79.125
Min	18.435
Max	99.181
Plage	80.745
Variance	435.251
Ecart type	20.863
Erreur standard de la moyenne	2.693

LOSES

Statistiques

Effectif	60
Moyenne	71.131
Min	0.861
Max	99.937
Plage	99.076
Variance	740.536
Ecart type	27.213
Erreur standard de la moyenne	3.513

Corrélations de Pearson

	WINS	LOSES
WINS	1.000/Parfait	0.044/Faible
LOSES	0.044/Faible	1.000/Parfait

We can see that it is only a “Weak” correlation between attendance and the outcome of a match. Although, it is important to notice that most stadiums are often at almost max capacity (when looking at the dots). Thus, we can’t draw specific conclusions.

From the correlation between numerical variable study, we can see that age and size are not relevant when talking about a player’s ability to score goals. What truly matter’s is if he played a lot of matches in which he only scored.

A team who historically scored and played a lot, is expected to score, and play again.

Finally, attendance doesn’t seem to have a strong impact on the outcome of the game, even if this is not a definitive conclusion, as most stadium are at almost max capacity for UCL games.

4.7 – ANOVA & Chi-squared test

In this database, we don’t have any categorical field. Therefore, we can’t do ANOVA or Chi-squared tests. For example, we could have done an ANOVA test by creating a column “OUTCOME” that would be “win”, “draw”, or “loss”, and compare it to the number of goals, or the stadium capacity, but in one case it’s useless as the outcome is a known direct consequence of the score, and we already studied the effect of attendance.

5 – Modelling: Predictive analysis

For this part, we will use 3 different datasets from the clustering we did.

Goal scorers, with player details, number of goals, and number of matches in which they scored:

Champ	Mesure	Valeurs	Manquant	Vérifier	Rôle
PLAYER_ID	Sans type			Aucun	Aucun
PLAYER	Sans type			Aucun	Aucun
PLAYER_DOB	Continu	[1984-09-28...		Aucun	Entrée
PLAYER_FOOT	Nominal	"",L,R		Aucun	Entrée
PLAYER_HEIGHT	Continu	[165.0,196.0]		Aucun	Entrée
PLAYER_NATI...	Nominal	USA,Mali,Ira...		Aucun	Entrée
PLAYER_POSIT...	Nominal	Forward,Midf...		Aucun	Entrée
PLAYER_WEIG...	Continu	[56.0,99.0]		Aucun	Entrée
MATCH_ID_Co...	Continu	[1.0,27.0]		Aucun	Entrée
GOAL_ID_Count	Indicateur	22.0/1.0		Aucun	Cible

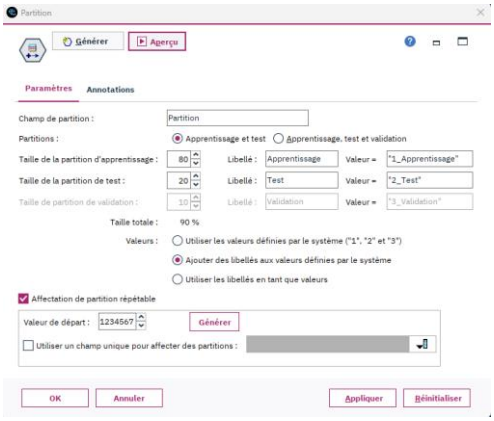
Assist providers, with player details, number of assists, and number of matches they assisted in:

Champ	Mesure	Valeurs	Manquant	Vérifier	Rôle
PLAYER_ID	Sans type			Aucun	Aucun
PLAYER	Sans type			Aucun	Aucun
PLAYER_DOB	Continu	[1984-06-28...		Aucun	Entrée
PLAYER_FOOT	Nominal	"",L,R		Aucun	Entrée
PLAYER_HEIGHT	Continu	[165.0,201.0]		Aucun	Entrée
PLAYER_NATI...	Nominal	USA,Mali,Ira...		Aucun	Entrée
PLAYER_POSIT...	Nominal	Forward,Midf...		Aucun	Entrée
PLAYER_WEIG...	Continu	[56.0,93.0]		Aucun	Entrée
MATCH_ID_Co...	Continu	[1.0,17.0]		Aucun	Entrée
GOAL_ID_Count	Indicateur	17.0/1.0		Aucun	Cible

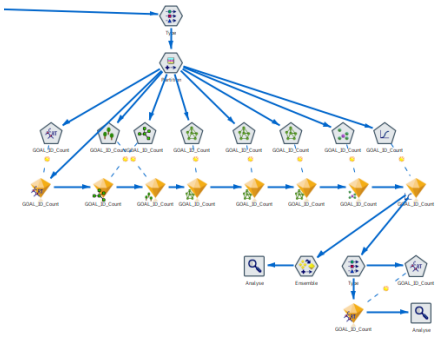
Teams with their average number of goals scored and conceded per match (switching target, thus 2 different set of data):

Champ	Mesure	Valeurs	Manquant	Vérifier	Rôle
TEAM_NAME	Nominal	"Bayern Mün...		Aucun	Entrée
CAPACITY	Continu	[48712.0,81...		Aucun	Entrée
CITY	Nominal	Paris,Madrid,...		Aucun	Entrée
COUNTRY	Nominal	Spain,France...		Aucun	Entrée
HOME_STADIUM	Nominal	Anfield,"Allia...		Aucun	Entrée
MANAGER_DOB	Continu	[1959-06-10...		Aucun	Entrée
MANAGER_FIR...	Nominal	Pep,Diego,Ca...		Aucun	Entrée
MANAGER_LAS...	Nominal	Klopp,Simeo...		Aucun	Entrée
MANAGER_NA...	Nominal	Spain,Italy,Fr...		Aucun	Entrée
PLAYER_DOB	Continu	[1985-09-09...		Aucun	Entrée
PLAYER_FIRST...	Nominal	"",Mo,Joe,Lu...		Aucun	Aucun
PLAYER_FOOT	Nominal	"",L,R		Aucun	Entrée
PLAYER_HEIGHT	Continu	[165.0,195.0]		Aucun	Entrée
PLAYER_ID	Nominal	ply94,ply96,...		Aucun	Aucun
PLAYER_LAST_...	Nominal	Gündoğan,Jo...		Aucun	Aucun
PLAYER_NATI...	Nominal	Spain,Egypt,I...		Aucun	Entrée
PLAYER_POSIT...	Nominal	Forward,Midf...		Aucun	Entrée
PLAYER_WEIG...	Continu	[60.0,95.0]		Aucun	Entrée
GOALS_SCORE...	Indicateur	1.45454545...		Aucun	Cible
GOALS_CONDE...	Indicateur	2.24456521...		Aucun	Cible

All 4 datasets are separated in 80% training data and 20% testing data.



5.1 – Ensemble Method



- Goal Scorers:

Résultats du champ de sortie GOAL_ID_Count

Comparaison de $XF-GOAL_ID_Count avec GOAL_ID_Count

'Partition'	1_Apprentissage		2_Test	
Correct	108	50,94 %	29	42,65 %
Incorrect	104	49,06 %	39	57,35 %
Total	212		68	

In the training partition, the ensemble model made 108 correct predictions (50.94% accuracy) and 104 incorrect predictions (49.06%). In the test partition, it made 29 correct predictions (42.65% accuracy) and 39 incorrect predictions (57.35%).

The ensemble model achieves moderate accuracy on the training partition, correctly predicting around 51% of the outcomes. However, the drop in accuracy to 42.65% on the test set raises concerns about the model's ability to generalize to new, unseen data.

Accord entre $R-GOAL_ID_Count $R1-GOAL_ID_Count $R2-GOAL_ID_Count $N-G

'Partition'	1_Apprentissage		2_Test	
Accord	85	40,09 %	21	30,88 %
Désaccord	127	59,91 %	47	69,12 %
Total	212		68	

Comparaison de Accord avec GOAL_ID_Count

'Partition'	1_Apprentissage		2_Test	
Correct	83	97,65 %	19	90,48 %
Incorrect	2	2,35 %	2	9,52 %
Total	85		21	

In the training partition, the ensemble model agreed on predictions for 85 cases (40.09%) and disagreed for 127 cases (59.91%). In the test partition, it agreed on predictions for 21 cases (30.88%) and disagreed for 47 cases (69.12%).

The C&R tree nodes within the ensemble model made 83 correct predictions (97.65%) and 2 incorrect predictions (2.35%) in the training partition. In the test partition, it made 19 correct predictions (90.48%) and 2 incorrect predictions (9.52%).

The ensemble model with C&R tree nodes exhibits lower overall accuracy (agreement) in both the training and test partitions compared to the model with ensemble nodes. However, the C&R tree nodes within the ensemble show high accuracy on both the training and test sets, indicating their strong predictive performance.

- Assist providers:

Résultats du champ de sortie GOAL_ID_Count

Comparaison de $XF-GOAL_ID_Count avec GOAL_ID_Count

'Partition'	1_Apprentissage		2_Test	
Correct	95	49,22 %	31	50 %
Incorrect	98	50,78 %	31	50 %
Total	193		62	

In the training partition, the ensemble model made 95 correct predictions (49.22% accuracy) and 98 incorrect predictions (50.78%). In the test partition, it made 31 correct predictions (50%) and 31 incorrect predictions (50%).

The ensemble model achieves around 50% accuracy on both the training and test partitions, indicating that its predictions are not consistently better than random chance. The balanced accuracy suggests that the model is making predictions that are roughly as accurate as random chance, both on the training and test sets.

Accord entre $R-GOAL_ID_Count $R1-GOAL_ID_Count $R2-GOAL_ID_Count $N-

'Partition'	1_Apprentissage		2_Test	
Accord	70	36,27 %	26	41,94 %
Désaccord	123	63,73 %	36	58,06 %
Total	193		62	

Comparaison de Accord avec GOAL_ID_Count

'Partition'	1_Apprentissage		2_Test	
Correct	59	84,29 %	21	80,77 %
Incorrect	11	15,71 %	5	19,23 %
Total	70		26	

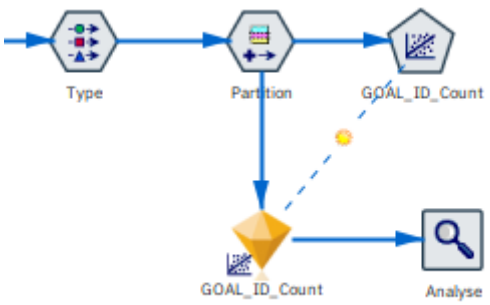
In the training partition, the ensemble model agreed on predictions for 70 cases (36.27%) and disagreed for 123 cases (63.73%). In the test partition, it agreed on predictions for 26 cases (41.94%) and disagreed for 36 cases (58.06%).

The C&R tree nodes within the ensemble made 59 correct predictions (84.29%) and 11 incorrect predictions (15.71%) in the training partition. In the test partition, it made 21 correct predictions (80.77%) and 5 incorrect predictions (19.23%).

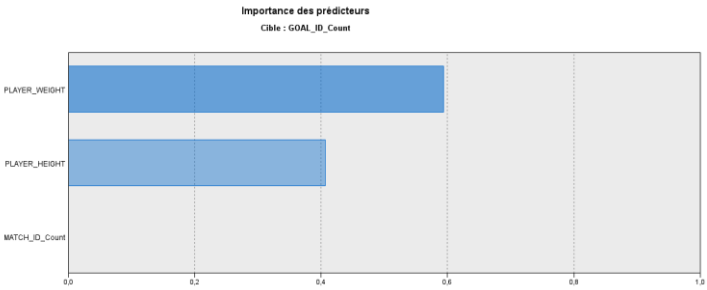
The ensemble model with C&R tree nodes exhibits lower overall accuracy (agreement) in both the training and test partitions compared to the model with ensemble nodes. However, the C&R tree nodes within the ensemble show higher accuracy, especially in the training set, indicating strong predictive performance from these nodes.

In summary, the application of ensemble methods to predict the number of assists provided by football players yields mixed results across different configurations. While ensemble methods offer versatility, their effectiveness is contingent on thoughtful configuration and ongoing refinement to ensure robust performance across various datasets and prediction tasks.

5.2 – Linear Regression



- Goal Scorers:



When predicting the number of goals for goal scorers, the most important predictors are a player height and weight.

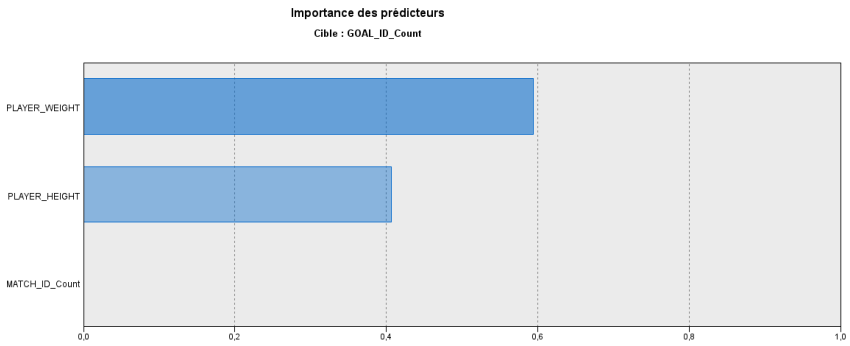
Résultats du champ de sortie GOAL_ID_Count

Comparaison de $E-GOAL_ID_Count avec GOAL_ID_Count

'Partition'	1_Apprentissage	2_Test
Nombre minimal d'erreurs	-1,941	-2,737
Nombre maximal d'erreurs	4,259	2,75
Nombre moyen d'erreurs	0,0	-0,022
Erreur absolue moyenne	0,429	0,626
Ecart type	0,679	1,013
Corrélation linéaire	0,986	0,987
Occurrences	212	68

The linear regression model is performing well in predicting the number of goals scored for goal scorers. The mean error close to zero suggests that, on average, the model does not exhibit a systematic bias in its predictions. The very high linear correlation values indicate an excellent fit between predicted and actual values, suggesting that the model captures the linear relationship well. The standard deviation values are relatively low, suggesting that the model's predictions are consistent on the training set. However, there is an increase in standard deviation for the test set, indicating a bit more variability in predictions for unseen data. The model generalizes well to the test partition, as indicated by the small mean error and high linear correlation.

- Assist providers:



When predicting the number of assists for assist providers, the most important predictors are a player height and weight.

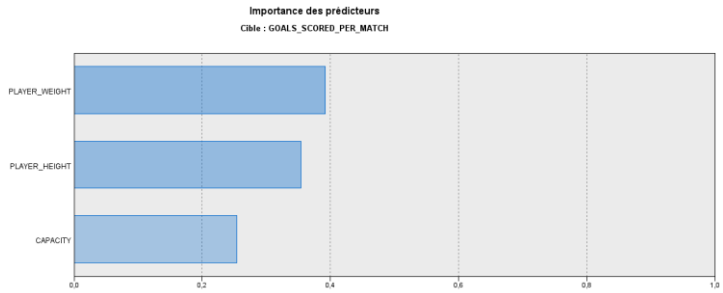
Résultats du champ de sortie GOAL_ID_Count

Comparaison de $E-GOAL_ID_Count avec GOAL_ID_Count

'Partition'	1_Apprentissage	2_Test
Nombre minimal d'erreurs	-3,327	-2,85
Nombre maximal d'erreurs	1,829	1,014
Nombre moyen d'erreurs	-0,163	-0,25
Erreur absolue moyenne	0,449	0,434
Ecart type	0,698	0,692
Corrélation linéaire	0,988	0,992
Occurrences	193	62

The linear regression model is performing well in predicting the number of assists provided by assist providers. The mean error close to zero indicates that, on average, the model makes unbiased predictions. The very high linear correlation values suggest an excellent fit between predicted and actual values, indicating that the model captures the linear relationship well. The low standard deviation values suggest that the model's predictions are consistent, with minimal variability in errors. The model generalizes well to the test partition, as indicated by the small mean error, low standard deviation, and high linear correlation.

- Goals scored per team:



When predicting the number of goals scored per team, the important predictors are the team's player height and weight, and stadium capacity of the match they are playing.

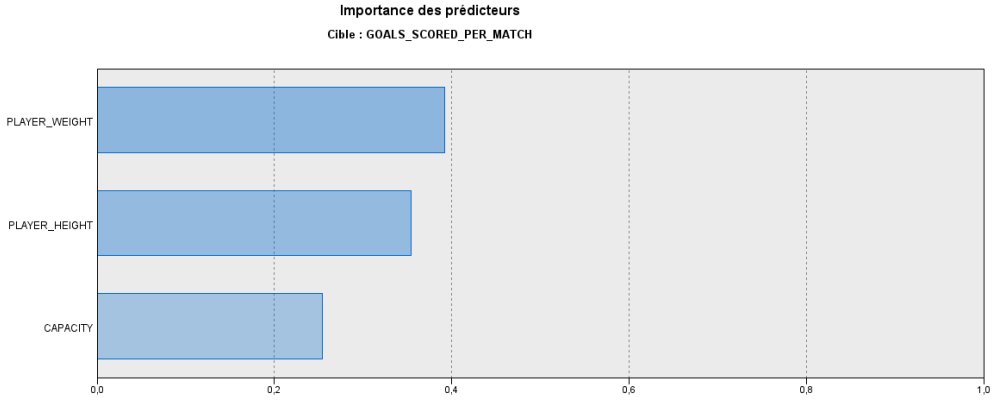
Résultats du champ de sortie GOALS_SCORED_PER_MATCH

Comparaison de $E-GOALS_SCORED_PER_MATCH avec GOALS_SCORED_PER_MATCH

'Partition'	1_Apprentissage	2_Test
Nombre minimal d'erreurs	-0,333	-0,284
Nombre maximal d'erreurs	0,759	0,893
Nombre moyen d'erreurs	-0,0	0,037
Erreur absolue moyenne	0,182	0,212
Ecart type	0,241	0,305
Corrélation linéaire	0,342	0,131
Occurrences	69	24

The linear regression model is making unbiased predictions, as indicated by the mean error values close to zero. The mean absolute error values are relatively low, suggesting that, on average, the model's predictions are accurate in terms of the magnitude of goals scored. The low standard deviation values indicate that the model's predictions are consistent. The linear correlation values are relatively weak, especially for the test partition, suggesting that the model's fit to the data might not be as strong as desired. The model may not capture the variability in goals scored well. The model's performance on the test set, as indicated by occurrences and other metrics, might be limited due to the relatively small number of data points.

- Goals conceded per team:



Again, when predicting the number of goals conceded per team, the important predictors are the team's player height and weight, and stadium capacity of the match they are playing.

Résultats du champ de sortie GOALS_SCORED_PER_MATCH

Comparaison de $E-GOALS_SCORED_PER_MATCH avec GOALS_SCORED_PER_MATCH

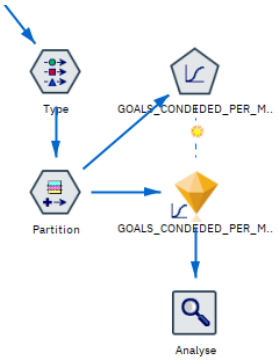
'Partition'	1_Apprentissage	2_Test
Nombre minimal d'erreurs	-0,333	-0,284
Nombre maximal d'erreurs	0,759	0,893
Nombre moyen d'erreurs	-0,0	0,037
Erreur absolue moyenne	0,182	0,212
Ecart type	0,241	0,305
Corrélation linéaire	0,342	0,131
Occurrences	69	24

Again, the linear regression model is making unbiased predictions, as indicated by the mean error values close to zero. The mean absolute error values are relatively low, suggesting that, on average, the model's predictions are accurate in terms of the magnitude of goals conceded. The low standard deviation values indicate that the model's predictions are consistent. The linear correlation values are relatively weak, especially for the test partition,

suggesting that the model's fit to the data might not be as strong as desired. The model may not capture the variability in goals conceded well. The model's performance on the test set, as indicated by occurrences and other metrics, might be limited due to the relatively small number of data points.

In summary, the linear regression models have demonstrated strong performance in predicting goals scored by goal scorers and assists provided by players. But some caution is warranted when predicting goals scored and goals conceded by football teams.

5.3 – Logistic Regression



- Goal Scorers:

Résultats du champ de sortie GOAL_ID_Count

Comparaison de $L-GOAL_ID_Count avec GOAL_ID_Count

'Partition'	1_Apprentissage		2_Test	
Correct	182	85,85 %	44	64,71 %
Incorrect	30	14,15 %	24	35,29 %
Total	212		68	

The model shows relatively high accuracy in the training partition, correctly classifying goals scored above or below a certain threshold 85.85% of the time. However, the drop in accuracy to 64.71% on the test set suggests potential challenges in generalizing to new, unseen data.

- Assist providers:

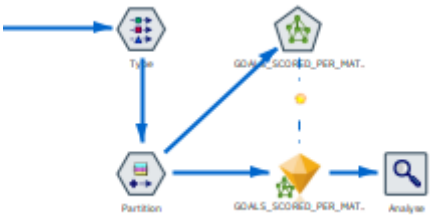
Résultats du champ de sortie GOAL_ID_Count

Comparaison de $L-GOAL_ID_Count avec GOAL_ID_Count

'Partition'	1_Apprentissage		2_Test	
Correct	95	49,22 %	31	50 %
Incorrect	98	50,78 %	31	50 %
Total	193		62	

The logistic regression model does not appear to perform well in distinguishing between players likely to provide a certain number of assists and those who are not. The accuracy is around 50% in both the training and test partitions, suggesting that the model's predictions are not better than random chance.

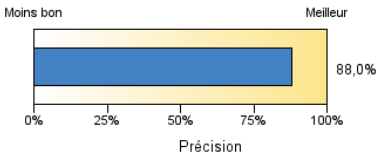
5.4 – Neural Networks



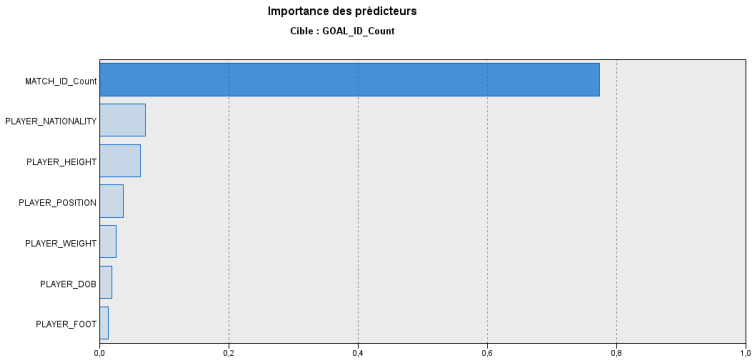
- Goal Scorers:

Récapitulatif du modèle

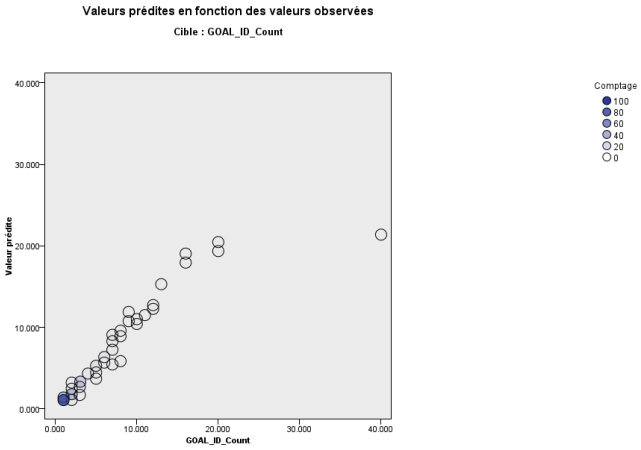
Cible	GOAL_ID_Count
Modèle	Perceptron multicouche
Règles d'arrêt utilisées	L'erreur ne peut être plus réduite
Neurones de première couche masquée	2

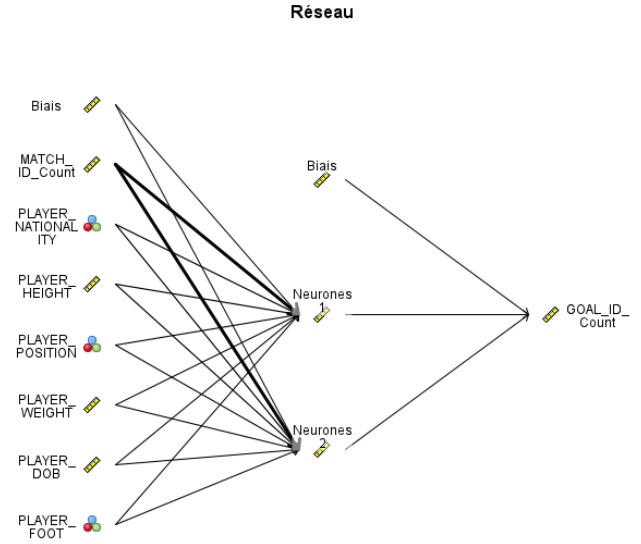


We can see that the model is globally good for predicting the number of goals scored by a goal scorer from 2016 to 2022.



When predicting the number of goals, the most important predictor is the number of matches played.





Résultats du champ de sortie GOAL_ID_Count

Comparaison de $N-GOAL_ID_Count avec GOAL_ID_Count

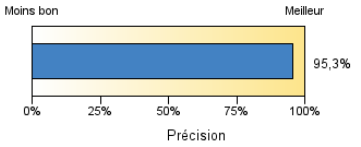
'Partition'	1_Apprentissage	2_Test
Nombre minimal d'erreurs	-3,026	-2,415
Nombre maximal d'erreurs	18,638	5,826
Nombre moyen d'erreurs	-0,023	0,245
Erreur absolue moyenne	0,441	0,818
Ecart type	1,434	1,343
Corrélation linéaire	0,939	0,961
Occurrences	212	68

The negative minimum error and low mean absolute error suggest that, on average, this model is performing well in predicting the number of goals for a goal scorer. The linear correlation values close to 1 indicate a strong positive relationship between the predicted and actual values. The standard deviation values suggest that the model's predictions are relatively consistent (less than 1,5 goals). Also, when comparing values from training and testing, we can see that the models perform well on unseen data (less errors, less standard deviation, higher linear correlation).

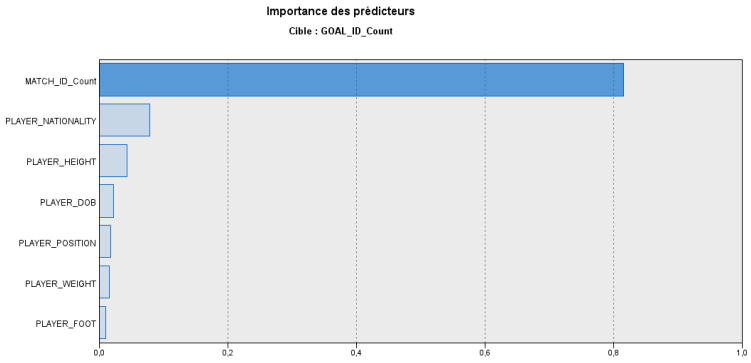
- Assist providers:

Récapitulatif du modèle

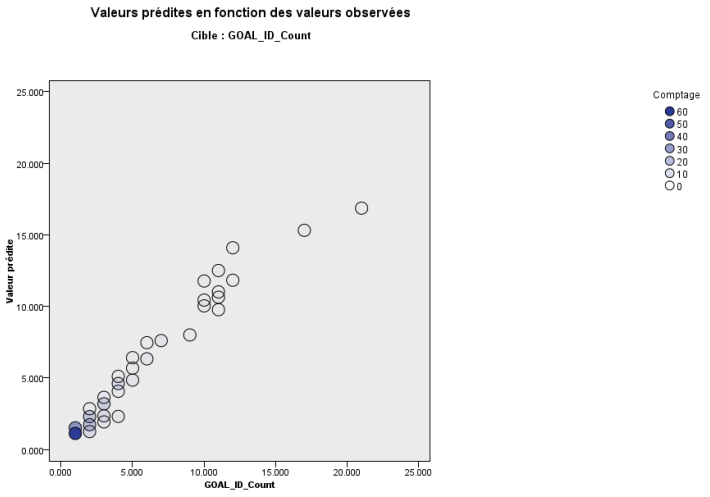
Cible	GOAL_ID_Count
Modèle	Perceptron multicouche
Règles d'arrêt utilisées	L'erreur ne peut être plus réduite
Neurones de première couche masquée	2



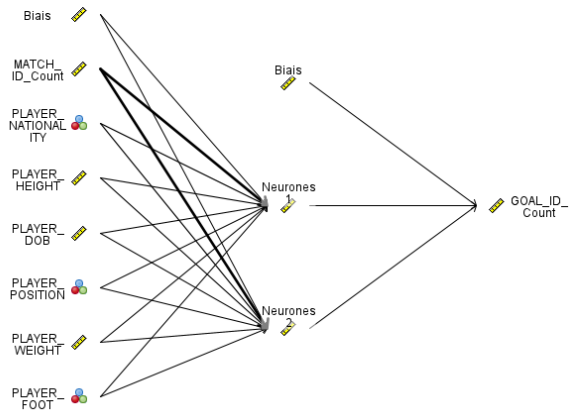
We can see that the model is almost perfect for predicting the number of assists provided by assist providers from 2016 to 2022.



When predicting the number of goals, the most important predictor is the number of matches played.



Réseau



Résultats du champ de sortie GOAL_ID_Count

Comparaison de $N-GOAL_ID_Count avec GOAL_ID_Count

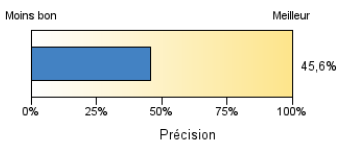
'Partition'	1_Apprentissage	2_Test
Nombre minimal d'erreurs	-2,099	-3,93
Nombre maximal d'erreurs	4,131	2,853
Nombre moyen d'erreurs	-0,107	-0,212
Erreur absolue moyenne	0,449	0,6
Ecart type	0,645	0,959
Corrélation linéaire	0,977	0,97
Occurrences	193	62

The negative minimum error and low mean absolute error suggest that, on average, this model is performing well in predicting the number of assists. The very high linear correlation values indicate an excellent fit between predicted and actual values. The standard deviation values are relatively low, suggesting that the model's predictions are consistent (less than 1 assist). The partition-wise comparison helps to see that the models perform well on unseen data (less errors, similar linear correlation).

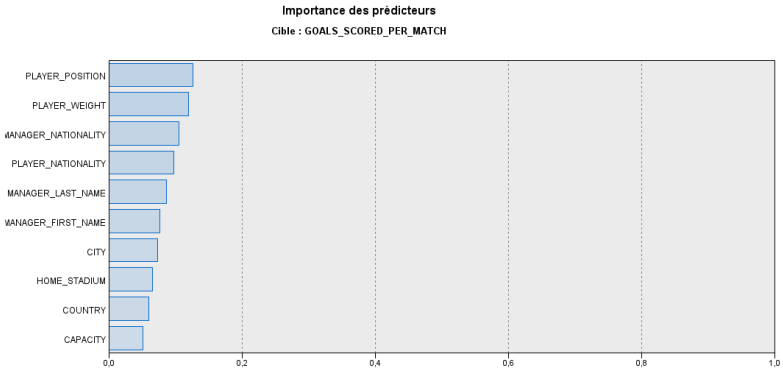
- Goals scored per team:

Récapitulatif du modèle

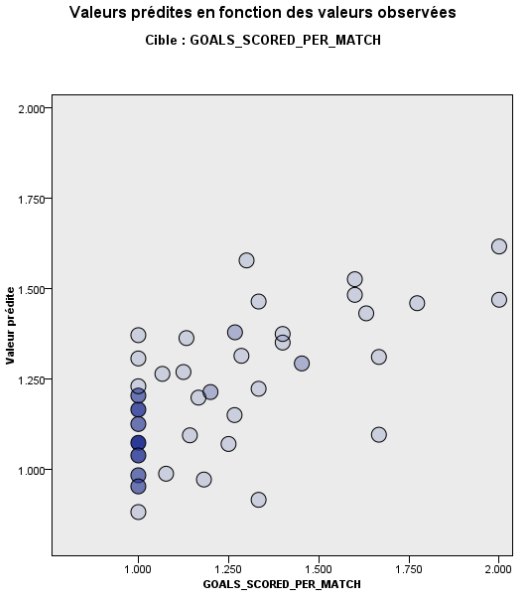
Cible	GOALS_SCORED_PER_MATCH
Modèle	Perceptron multicouche
Règles d'arrêt utilisées	L'erreur ne peut être plus réduite
Neurones de première couche masquée	9

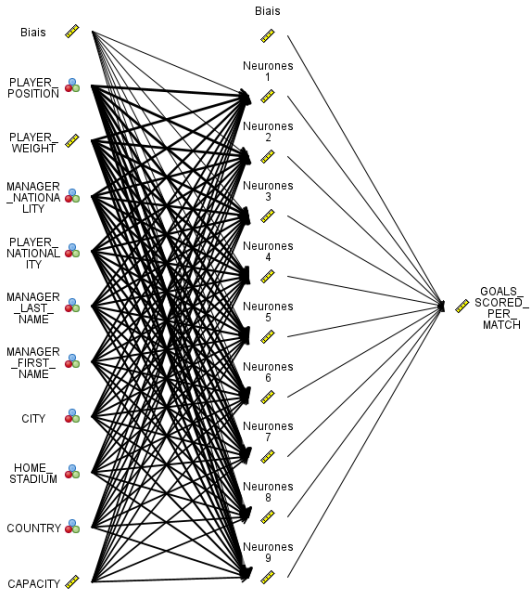


We can see that the model is globally not good for predicting the overall number of goals scored by a team.



When predicting the number of goals scored per team, all predictors have more or less the same importance.





Résultats du champ de sortie GOALS_SCORED_PER_MATCH

Comparaison de $N-GOALS_SCORED_PER_MATCH avec GOALS_SCORED_PER_MATCH

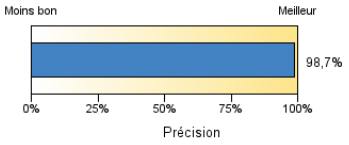
'Partition'	1_Apprentissage	2_Test
Nombre minimal d'erreurs	-0,372	-0,337
Nombre maximal d'erreurs	0,571	0,761
Nombre moyen d'erreurs	-0,008	0,036
Erreur absolue moyenne	0,143	0,167
Ecart type	0,189	0,262
Corrélation linéaire	0,676	0,432
Occurrences	69	24

The mean error shows that the model performs well on predicting the number of goals scored for a UCL team from 2016 to 2022. Standard deviation is also very low (0,1 goal). But, when comparing values from training and testing, we can see that the models perform bad on unseen data (more standard deviation, less linear correlation).

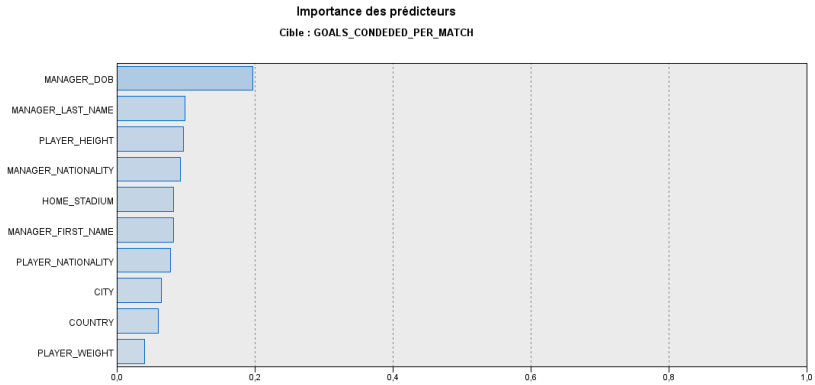
- Goals conceded per team:

Récapitulatif du modèle

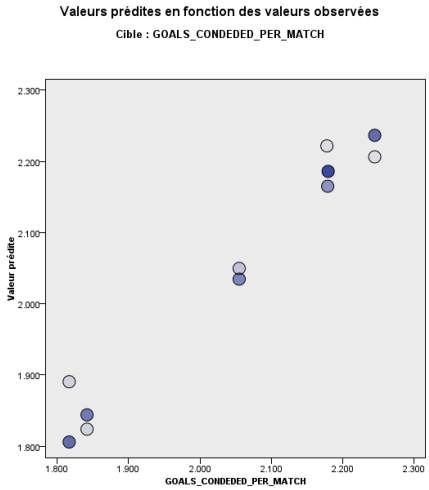
Cible	GOALS_CONDEDED_PER_MATCH
Modèle	Perceptron multicouche
Règles d'arrêt utilisées	L'erreur ne peut être plus réduite
Neurones de première couche masquée	8

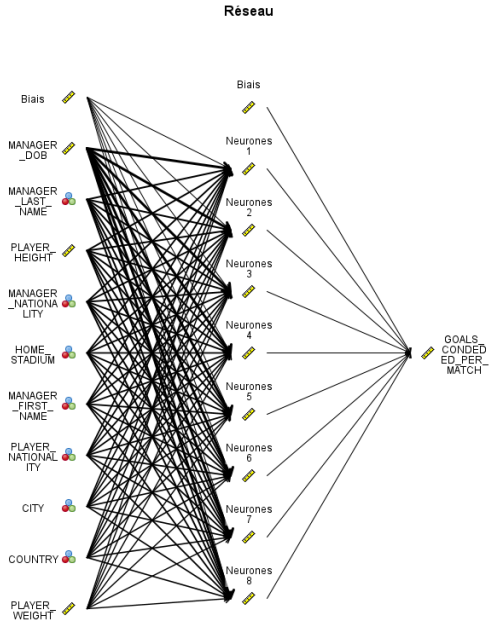


We can see that the model is almost perfect for predicting the number of goals conceded per teams from 2016 to 2022.



When predicting the number of goals conceded per team, all predictors have more or less the same importance.





Résultats du champ de sortie GOALS_CONDEDED_PER_MATCH

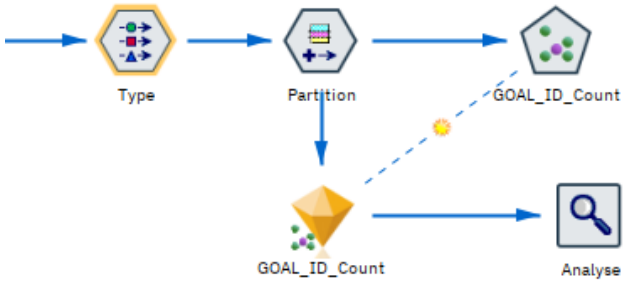
Comparaison de $N-GOALS_CONDEDED_PER_MATCH avec GOALS_CONDEDED_PER_MATCH

'Partition'	1_Apprentissage	2_Test
Nombre minimal d'erreurs	-0,08	-0,083
Nombre maximal d'erreurs	0,038	0,063
Nombre moyen d'erreurs	0,003	0,006
Erreur absolue moyenne	0,014	0,028
Ecart type	0,019	0,035
Corrélation linéaire	0,993	0,978
Occurrences	69	24

The mean error shows that the model almost never fails on predicting the number of goals conceded for a UCL team from 2016 to 2022. Standard deviation is also very low (0,035 goal in testing). Linear correlation is almost at 1 (0,97 in testing) indicating a strong positive relationship between the predicted and actual values. And when comparing values from training and testing, we can see that the models perform well on unseen data (most indicators are similar).

In conclusion, the neural network models are effective in predicting individual player performance metrics (goals and assists) and team defensive metrics (goals conceded).

5.5 – Instance-Based Classifier: Nearest Neighbour



- Goal Scorers:

Résultats du champ de sortie GOAL_ID_Count

Comparaison de $KNN-GOAL_ID_Count avec GOAL_ID_Count

'Partition'	1_Apprentissage		2_Test	
Correct	91	47,15 %	30	48,39 %
Incorrect	102	52,85 %	32	51,61 %
Total	193		62	

The KNN model's performance in predicting the number of goals scored by goal scorers seems limited. The accuracy on both the training and test partitions is below 50%, suggesting that the model's predictions are not consistently better than random chance. The drop in accuracy from the training to the test set indicates potential challenges in the model's ability to generalize to new, unseen data.

- Assist providers:

Résultats du champ de sortie GOAL_ID_Count

Comparaison de $KNN-GOAL_ID_Count avec GOAL_ID_Count

'Partition'	1_Apprentissage		2_Test	
Correct	102	48,11 %	28	41,18 %
Incorrect	110	51,89 %	40	58,82 %
Total	212		68	

The KNN model's performance in predicting the number of assists provided by assist providers again seems limited. The accuracy on both the training and test partitions is around 50%, indicating that the model's predictions are not consistently better than random chance. The close

match between training and test accuracy suggests that the model might generalize reasonably well to new, unseen data.

- Goals scored per team:

Résultats du champ de sortie GOALS_SCORED_PER_MATCH

Comparaison de $KNN-GOALS_SCORED_PER_MATCH avec GOALS_SCORED_PER_MATCH

'Partition'	1_Apprentissage		2_Test	
Correct	1	1,45 %	0	0 %
Incorrect	68	98,55 %	24	100 %
Total	69		24	

The KNN model performs very poorly in predicting the number of goals scored by football teams. The accuracy is close to zero in both the training and test partitions, indicating that the model is not effectively capturing the relationships between the predictors and the target variable. The extremely low accuracy on the test set suggests that the model is not generalizing well to new, unseen data. It is likely that the model is overfitting to the training data, failing to identify meaningful patterns that could be applied to other datasets.

In summary, the K-nearest neighbours (KNN) models shows limitations in predictive performance. KNN models face challenges in achieving accuracy consistently better than random chance.

6 – Evaluation

Can we build a model to predict match outcomes based on historical data, including team, manager, and player attributes, match context, and previous performances?

The ensemble method, linear regression, and neural network models exhibit varying degrees of success in predicting individual player performance metrics (goals and assists) and team defensive metrics (goals conceded). The ensemble method provides moderate accuracy, while the linear regression models, particularly in predicting goals and assists, perform well. The neural network models demonstrate effectiveness in predicting both individual and team performance metrics.

Business Action: Use predictions from ensemble method and neural networks for individual player performance predictions.

How accurately can we forecast goal differentials for individual matches, and what variables play a significant role in these predictions?

Linear regression models for predicting goals and assists by individual players show promise, suggesting that these models can be employed to forecast goal differentials. Neural networks, especially for team defensive metrics, also display strong predictive performance.

Business Action: Use linear regression models for predicting goal differentials, focusing on individual player metrics. Consider neural network models for comprehensive team defensive predictions.

Based on the insights gained, what strategies can bookmakers employ to adjust odds and maximize profitability?

The ensemble method, despite moderate accuracy, can provide insights into adjusting odds for individual player performance. Linear regression and neural network models offer valuable input for refining odds related to goal differentials and team defensive metrics.

Business Action: Collaborate with data scientists to integrate ensemble and linear regression models to adjust odds. Regularly update models to enhance predictive capabilities.

How can the predictive models assist club's management and board in identifying what can be considered key to winning matches?

Linear regression and neural network models emerge as effective tools for identifying key performance metrics related to winning matches, such as goals scored by individual players and team defensive capabilities.

Business Action: Develop performance strategies based on insights from linear regression models for individual players and neural networks for team metrics.

Overall Business Actions:

- Leverage Ensemble and Neural Networks: Incorporate insights from ensemble models for odds adjustment and utilize neural network predictions for team performance metrics.

- Continuous Model Refinement: Regularly update and refine predictive models, considering the strengths and limitations of each model type.

- Player-Specific Strategies: Develop strategies to enhance individual player performance based on predictions from linear regression models.

- Collaborative Approach: Establish ongoing collaboration between bookmakers or club management, and data scientists to enhance predictive capabilities and optimize strategies.