Tommaso Premoli N. Matricola: 34221A

THE ROLE OF THE MIDFIELDER IN EUROPEAN FOOTBALL: AN ANALYSIS OF DECISIVE PLAYERS IN THE TOP 5 EUROPEAN LEAGUES

The role of the midfielder in the sport of football is universally recognised as the most complete, requiring extraordinary versatility that includes defensive and offensive skills. This research aims to identify and analyse midfielders who have proven to be particularly decisive in their teams, in the highly competitive contexts of the top 5 European leagues: La Liga, Premier League, Serie A, Bundesliga and Ligue 1.

Using an approach based on data analysis and visualisation, the research will explore several key variables, including goals scored, assists, passing accuracy, defensive actions, and other relevant metrics to assess the overall impact of midfielders. The dataset will consist of the players' 90-minute statistics for the 2021-2022 season.

At the end of the research, we will have a clear understanding of which midfielders have been most decisive in their teams and how these contributions influence the overall performance of teams in Europe's top 5 leagues.

1. Introduction

The role of the midfielder football is universally recognised as the most complete one, as it requires an extraordinary versatility that includes defensive and offensive skills. The purpose of this research is to analyse the relevant football statistics of midfielders affiliated with teams ranked between first and tenth in the five major European leagues: Premier League, Serie A, La Liga, Bundesliga and Ligue 1. This investigation aims at analysing the variables of the available dataset in general and, then, identifying the players who had the greatest impact on their respective teams.

A group of datasets was used for the research, the set of which is called 'Big 5 European Leagues Player Stats 2022-23'. In total, there were nine datasets in .csv format. All information in them was obtained from the football statistics website FBref. Through experience and personal reworking, the most relevant variables for the analysis were selected and merged into one file. This dataset contains the 90-minute statistics of all players who played in the five major European leagues. The dataset was available on the Kaggle website and was updated weekly. Columns whose statistical variables were most superfluous for the purpose of the research were removed. The statistical variables most relevant for the analysis were kept:

- Age: player's age;
- MP: matches played;
- PPM (Points per Match): The average number of points scored by the team when the player is on the pitch;
- Gls: The total number of goals scored by the player;
- Ast: The total number of assists provided by the player;
- xG (Expected Goals): The expected number of goals for shots taken by the player;
- KP: The total number of key passes made by the player, that is, passes that lead to a shot attempt;
- SCA90 (Shot-Creating Actions per 90 minutes): The number of actions leading to a shot attempt per 90 minutes of play;
- Cmp_percent: The percentage of completed passes to total attempts;
- TB (Through Balls): The number of successful passes between opposing defences to create a shooting opportunity;
- PassLive (Live-Ball Passes Leading to a Goal): The number of completed passes leading directly to a goal;
- Tkl_percent: The success percentage in defensive challenges, calculated as the ratio of challenges won to the total number of challenges faced;
- Int: The total number of interceptions made by the player.

For research purposes, three variables were created: GA (sum of goals and assists); CmpTB (Sum of Cmp_percent and TB) and, finally, TkIn (Sum of Tkl_percent and Int).

2. Methodology

After having cleaned the dataset correctly to perform the analysis, the first step of the research is to analyse the individual variables taken into consideration to make a comparison between the 5 European competitions. For example, the average ages of the players, the distribution of midfielders among the 5 leagues and the average offensive support with goals by midfielders in the team per league will be examined.

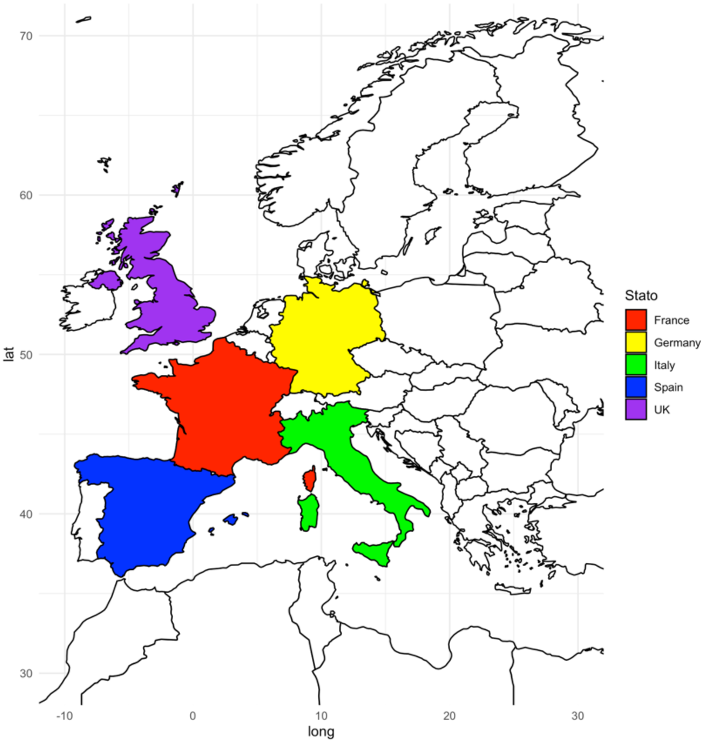
After that, the analysis moves on to the evaluation of the midfielders by analysing the variables to see which players were the most decisive during the 2022-22 season. By means of bubble charts and pie charts, some important variables were analysed.

To conclude, a regression analysis will be conducted in which the dependent variable will be the average number of points obtained by the team when the player is on the pitch (PPM). The objective is to determine which independent variable has a significant influence on the variability of points per game (PPM).

2.1. Data Cleaning

The following steps were taken to put the data in order. First of all, all rows that contained at least one null value (called na) were removed. After that, only midfielders who played for teams that ranked from first to tenth in their league during the 2022/23 season were taken into account. Thus, the following teams were eliminated from the dataset: Reims, Montpellier, Toulouse, Brest, Strasbourg, Nantes, Auxerre, Ajaccio, Troyes, Angers, Crystal Palace, Chelsea, Wolves, West Ham, Bournemouth, Nottingham Forest, Everton, Leicester City, Leeds United, Southampton, Monza, Udinese, Sassuolo, Empoli, Salernitana, Lecce, Hellas Verona, Spezia, Sampdoria, Rayo Vallecano, Sevilla, Celta Vigo, Cádiz, Getafe, Valencia, Almería, Valladolid, Espanyol, Elche, Köln, Hoffenheim, Werder Bremen, Bochum, Augsburg, Stuttgart, Schalke 04 and Hertha BSC. Finally, midfielders who played less than ten matches during the year were excluded from the selection, as their presence may not have had a significant impact on the matches. The final cleaned and corrected dataset contains 319 rows, one for each player.

A geographical map of the European continent shows the states in which the five major European championships are played. The colours that are shown on the map have been reproduced in the following analysis to better understand the results. As the legend indicates in red is France, in yellow Germany, in green Italy, in blue Spain and in purple the United Kingdom.

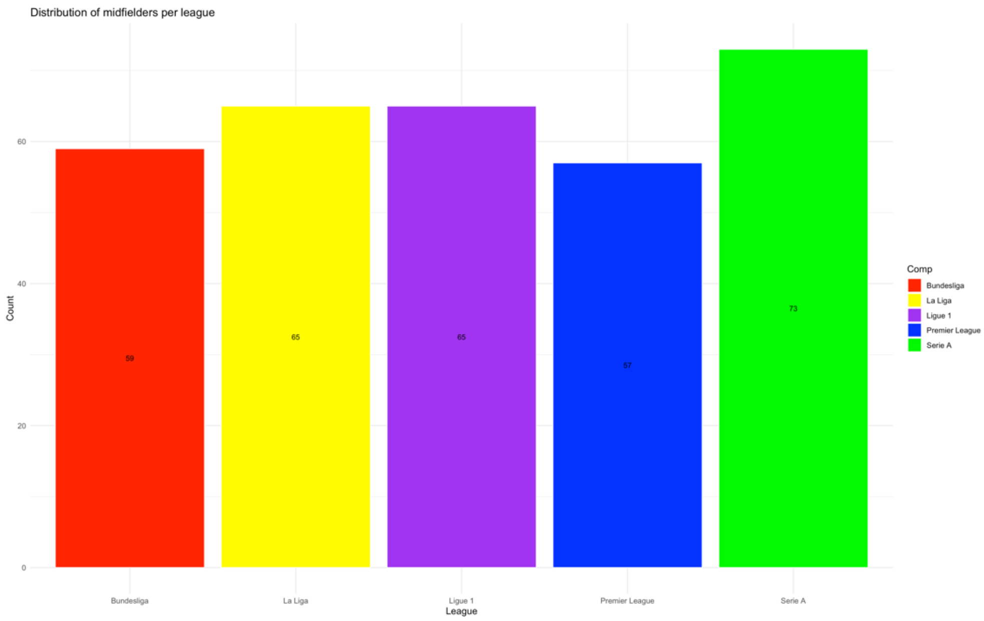

70
60
50
40
30
lat
-10
0
10
20
30
long
Stato
France
Germany
Italy
Spain
UK

3. Results

First of all, the dataset is analysed in general for greater understanding.

The first analysis concerns the distribution of the players under consideration per league by creating a histogram. In this way, it is possible to know the number of midfielders per league that are present in the dataset.

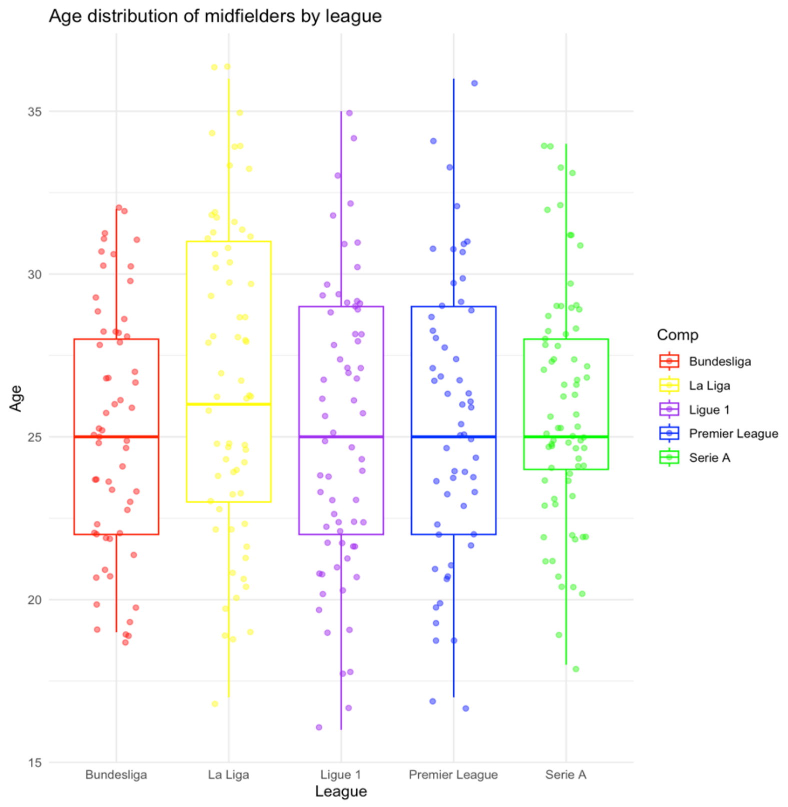
```
> counts <- table(X2022_23_Midfielders_Stats$Comp)
> data <- data.frame(Comp = names(counts), Count = as.numeric(counts))

> histogram <- ggplot(data, aes(x = Comp, y = Count, fill = Comp)) +
       geom_bar(stat = "identity", color = "white") +
       geom_text(aes(label = Count), position = position_stack(vjust = 0.5), size =
3) +
       labs(title = "Distribution of midfielders per league", x = "League", y =
"Count") +
       scale_fill_manual(values = c("Bundesliga" = "red", "Premier League" =
"blue", "La Liga" = "yellow", "Serie A" = "green", "Ligue 1" = "purple")) +
       theme_minimal()
> print(histogram)
```



It can be seen that in the dataset there is a majority of midfielders from Serie A with 73, followed by La Liga and Ligue 1 both with 65, the Bundesliga with 59, and finally the Premier League 57.

In the second phase of the analysis, we focus on determining the average age of the midfielders, classifying them according to their respective league of origin. The graph that was used is a series of boxplots. One for each competition, which is distinguished by the previously mentioned colour. Outliers are also eliminated. This means that outliers will not be shown in the graph.

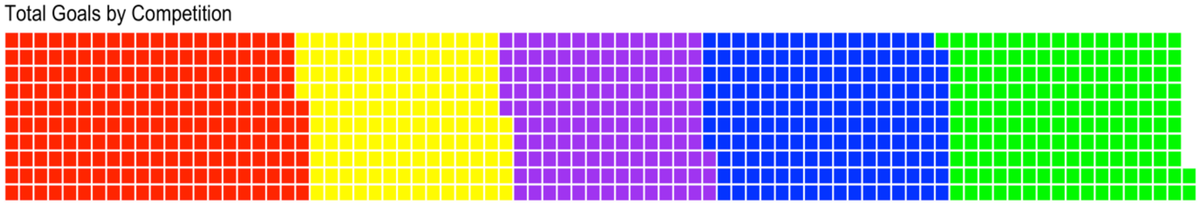
```
> ggplot(X2022_23_Midfielders_Stats, aes(x = Comp, y = Age, color = Comp)) +
geom_boxplot(outlier.shape = NA) +
geom_jitter(position = position_jitter(0.2), alpha = 0.5) +
scale_color_manual(values = c("Bundesliga" = "RED", "Premier League" = "BLUE", "La
Liga" = "YELLOW", "Serie A" = "GREEN", "Ligue 1" = "PURPLE")) +
labs(title = "Age distribution of midfielders by league", x = "League", y = "Age")
+
theme_minimal()
```



The average age is almost similar among all midfielders. There is a difference in the Spanish league where the average age is significantly higher. It is possible to notice less variability in the data in the Italian league than in the others.

The next step in the analysis of the dataset consists of the elaboration of a waffle chart in order to explore the distribution of goals scored by midfielders in relation to the total, divided by each football league. In red is shown the Bundesliga, in yellow the La Liga, in purple the Ligue 1, in blue the Premier League and in green the Serie A.

```
> num_gol_per_comp <- aggregate(Gls ~ Comp, data = X2022_23_Midfielders_Stats, FUN
= sum)
> waffle(
     num_goal_per_league$Gls,
     rows = 10,
     colors = c("Bundesliga" = "red", "Premier League" = "blue", "La Liga" =
"yellow", "Serie A" = "green", "Ligue 1" = "purple"),
     title = list(
         label = "Total Goals by Competition",
         size = 0.8
     )
```

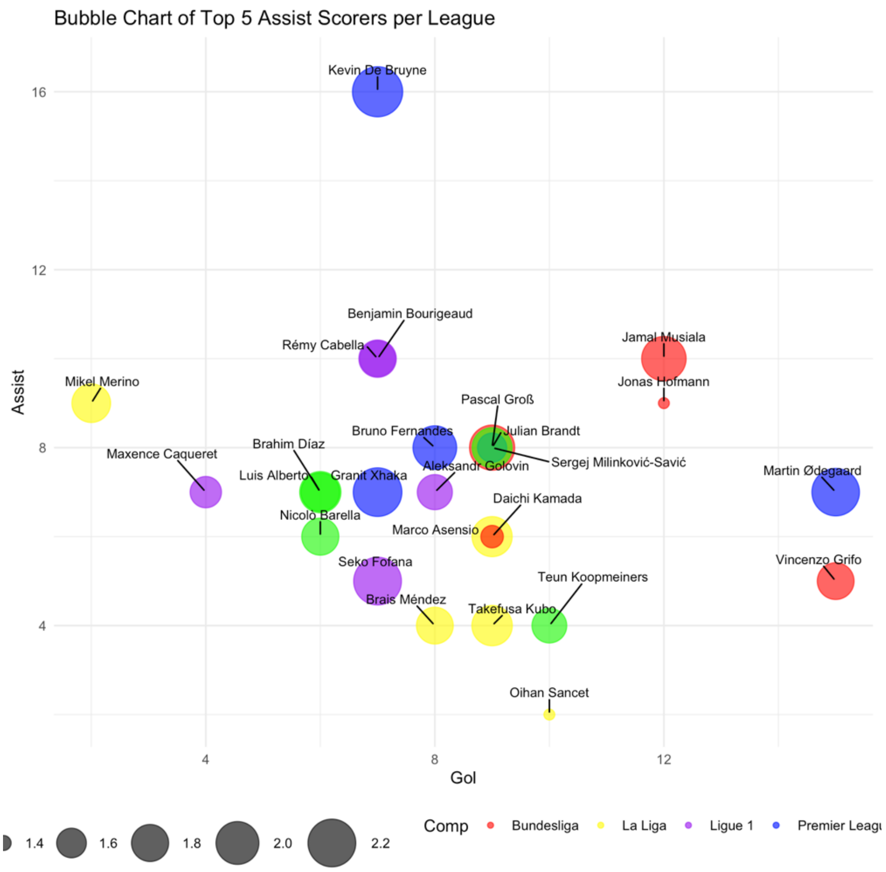


The Bundesliga (25.37%) stands out with the highest percentage, highlighting a significant offensive contribution by midfielders in that context. The Premier League (20.44%) follows with a significant percentage, suggesting an equally relevant role for midfielders in the league's offensive dynamics. Serie A (20.07%), while maintaining a considerable percentage, is slightly below the Premier League, while La Liga (17.12%) and Ligue 1 (17.00%) show lower percentages, indicating a relative involvement of midfielders in the finalisation phase in these leagues.

After a preliminary analysis of the relevant variables, we focused on identifying the most influential midfielders from an offensive, defensive and tactical point of view. For this detailed investigation, a bubble chart was used, with specific variables on the abscissas and ordinates and the size of the dots associated with the PPM variable (points per game scored by the team when the player was on the pitch). This visual approach provides an immediate overview of the midfielders' key performances.

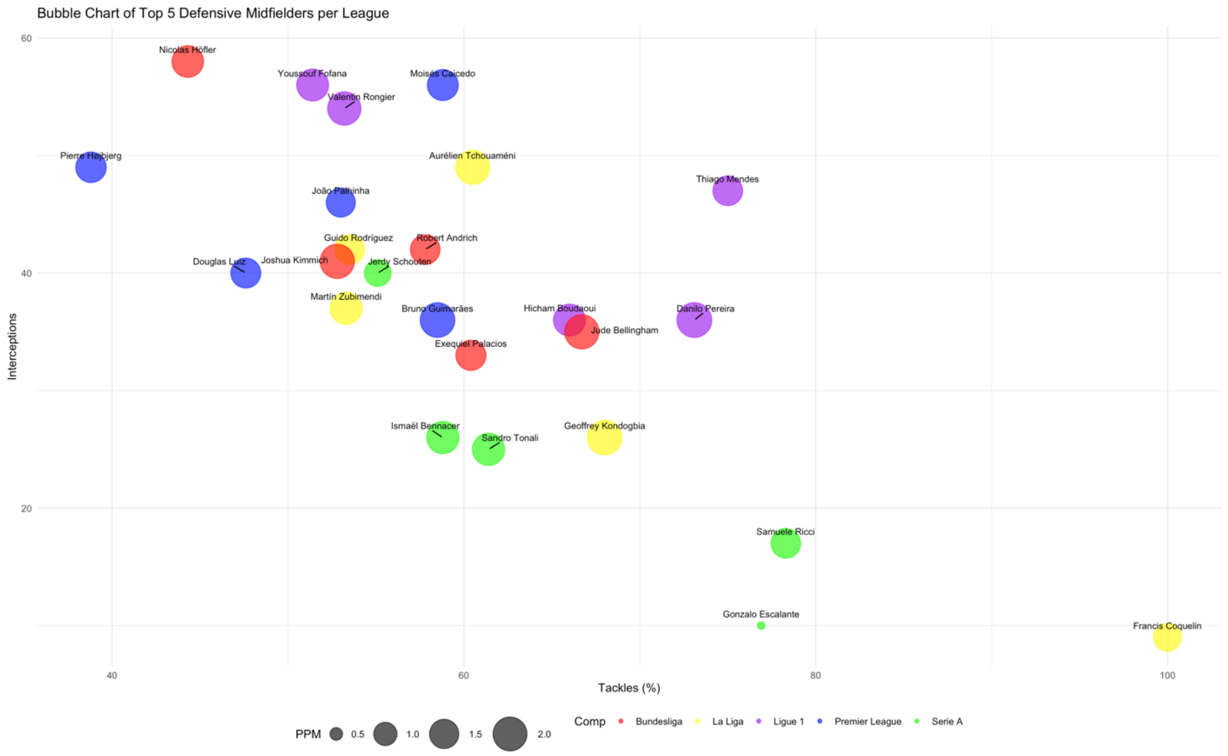
The first graph concerns the players with the most goals and assists in the five major leagues.

```
> ggplot(top_players, aes(x = Gls, y = Ast, size = PPM, label = Player, color =
Comp)) +
+     geom_point(alpha = 0.7) +
+     geom_text_repel(box.padding = 0.5, point.padding = 0.1, size = 3,
check_overlap = TRUE,
+                     nudge_y = 0.5, color = "black") +
+     scale_size_continuous(range = c(3, 15)) +
+     labs(title = "Bubble Chart of Top 5 Assist Scorers per League",
+          x = "Gol",
+          y = "Assist") +
+     theme_minimal() +
+     scale_color_manual(values = c("Premier League" = "blue", "Serie A" = "green",
+                                   "La Liga" = "yellow", "Bundesliga" = "red",
+                                   "Ligue 1" = "purple")) +
+     theme(legend.position = "bottom")
```



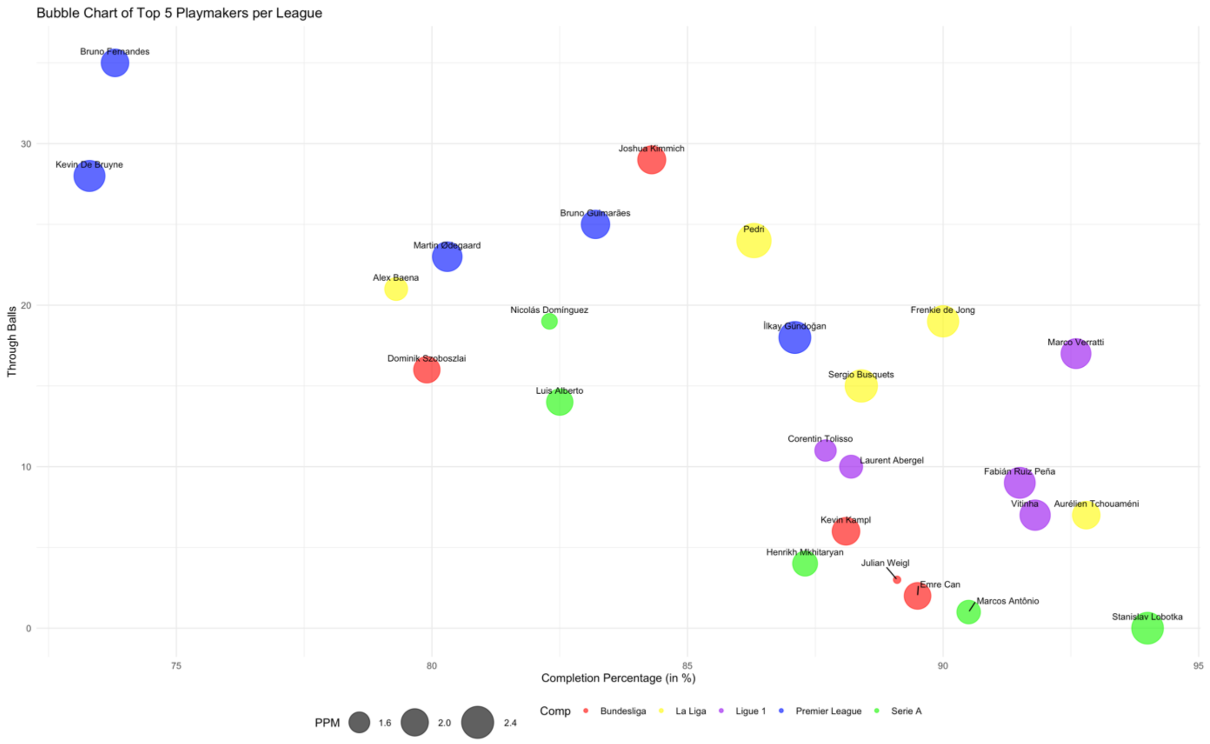
The second bubble chart concerns the players who defensively through successful tackles and interceptions performed best.

```
> ggplot(top_def, aes(x = Tkl_percent, y = Int, size = PPM, label = Player, color =
Comp)) +
       geom_point(alpha = 0.7) +
       geom_text_repel(box.padding = 0.5, point.padding = 0.1, size = 3,
check_overlap = TRUE, nudge_y = 0.5, color = "black") +
       scale_size_continuous(range = c(3, 15)) +
       labs(title = "Bubble Chart of Top 5 Defensive Midfielders per League",
       y = "Interceptions") +
       theme_minimal() +
       scale_color_manual(values = c("Premier League" = "blue", "Serie A" =
"green",
                                     "La Liga" = "yellow", "Bundesliga" = "red",
                                     "Ligue 1" = "purple")) +
       theme(legend.position = "bottom")
```



Finally, the last graph concerns the players who excelled most in the playmaking role through correctness and dangerousness of passes.

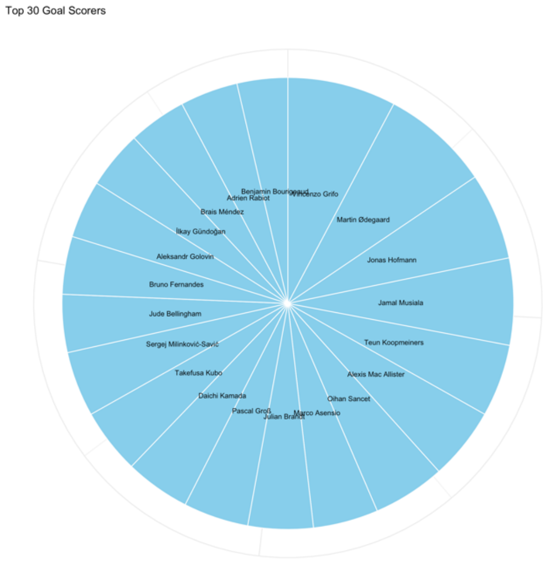
```
> ggplot(top_playmaker, aes(x = Cmp_percent, y = TB, size = PPM, label = Player,
color = Comp)) +
       geom_point(alpha = 0.7) +
       geom_text_repel(box.padding = 0.5, point.padding = 0.1, size = 3,
                       nudge_y = 0.5, color = "black") +
       scale_size_continuous(range = c(3, 15)) +
       labs(title = "Bubble Chart of Top 5 Playmakers per League",
            x = "Completion Percentage (in %)",
            y = "Through Balls") +
       theme_minimal() +
       scale_color_manual(values = c("Premier League" = "blue", "Serie A" =
"green", "La Liga" = "yellow", "Bundesliga" = "red", "Ligue 1" = "purple")) +
       theme(legend.position = "bottom")
```

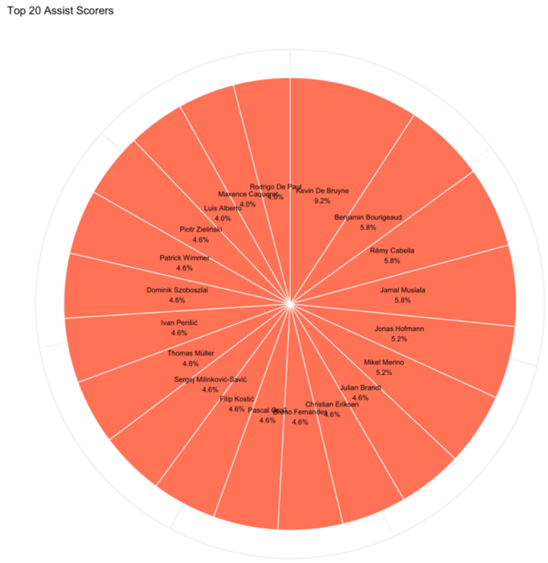


Another graph that was used in the research is the pie chart to look at the players who performed best with regard to several variables. The code used is similar for all charts, therefore, only the code used for the top 20 scorers is shown as an example.

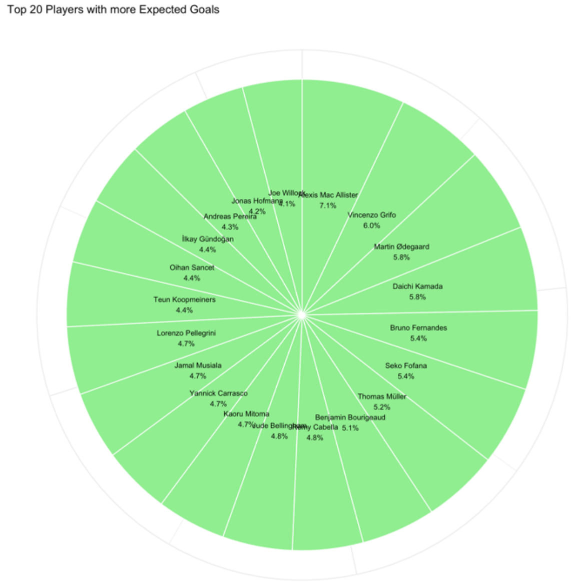
```
> top_scorers <- X2022_23_Midfielders_Stats %>%
      arrange(desc(Gls)) %>%
      head(20)

> ggplot(top_scorers, aes(x = "", y = Gls, fill = factor(1))) +
      geom_bar(stat = "identity", width = 1, color = "white") +
      coord_polar("y") +
      labs(title = "Top 20 Goal Scorers", fill = "") +
      scale_fill_manual(values = "sky blue", guide = "none") +
      theme_minimal() +
      theme(axis.text = element_blank(),
            axis.title = element_blank(),
            legend.position = "none") +
      geom_text(position = position_stack(vjust = 0.5), aes(label = Player), size
= 3)
```

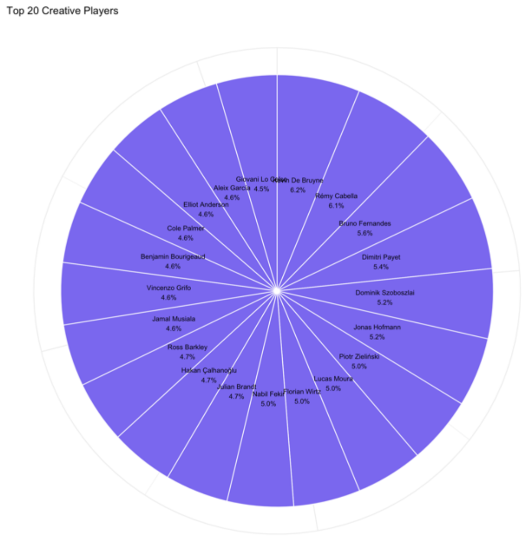
The first two graphs, for example, show the 20 players who scored the most goals and made the most assists in 90 minutes in the 2022-23 season.





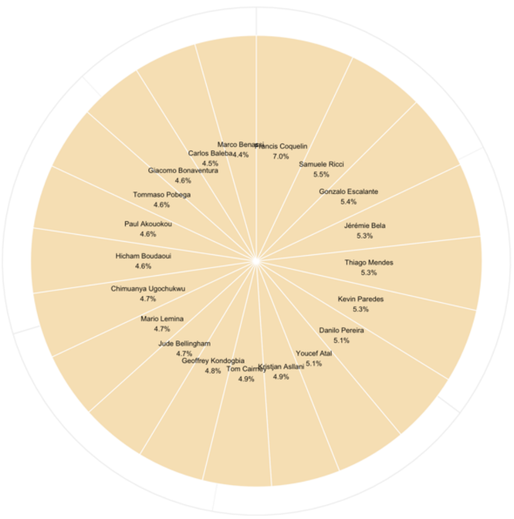
The following two graphs, on the other hand, show the players who had a higher amount of expected goals during the match and the most creative players (with a higher value in the SCA90 variable).



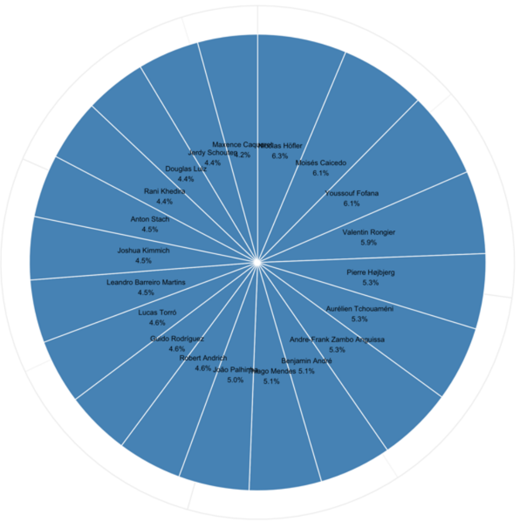


Finally, the players who were most relevant defensively are also evaluated using the variable of the percentage of tackles won and interceptions made.

Top 20 Players with more Takles

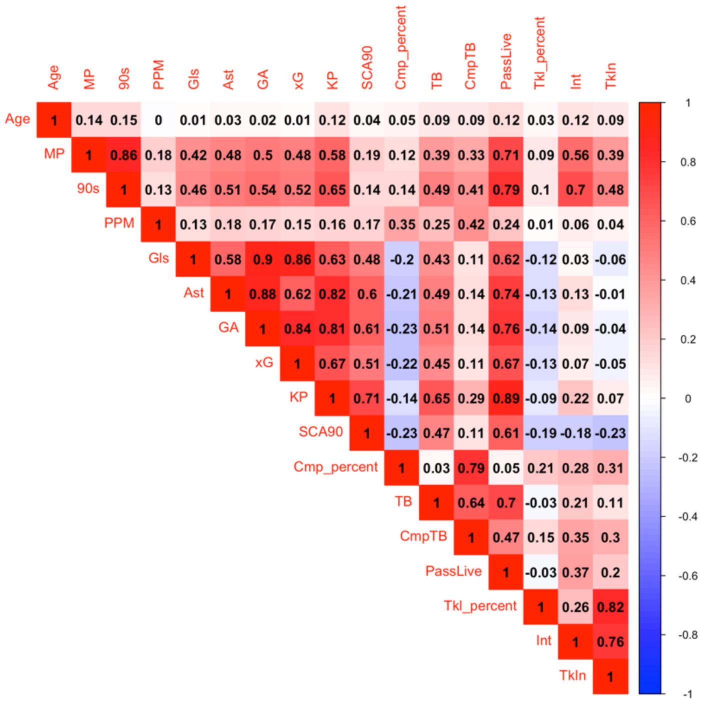


Top 20 Players with more Intercepts



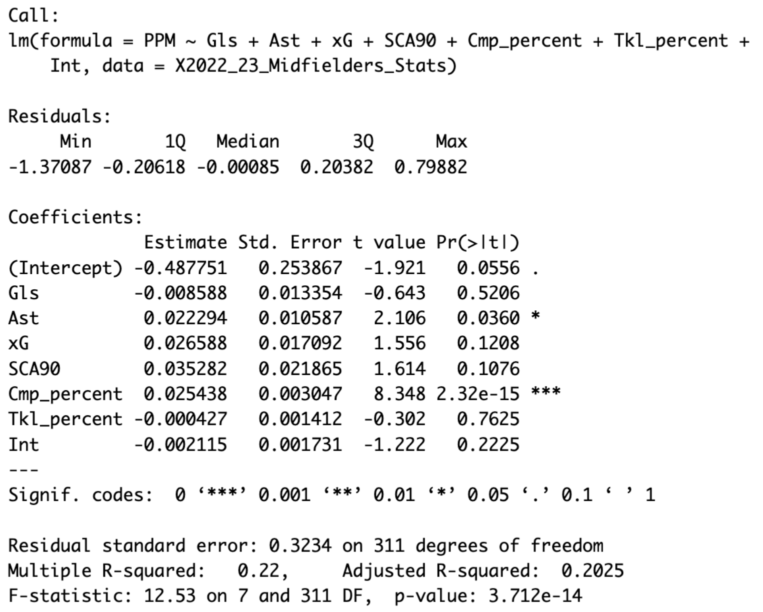
A crucial analysis step involves the application of a multiple linear regression, aimed at examining the incidence of one variable on the others. Specifically, in order to evaluate the importance of a player in terms of decision-making, it is examined whether his presence on the court influences the team's performance. This process results in a linear regression in which points per game constitute the dependent variable, while the other variables take on the role of independents.

In order to clearly visualise the dependency and correlation between the variables, a heatmap of the main performance indicators is presented.



	Age	MP	90s	PPM	Gls	Ast	GA	xG	KP	SCA90	Cmp_percent	TB	CmpTB	PassLive	Tkl_percent	Int	TkIn
Age	1	0.14	0.15	0	0.01	0.03	0.02	0.01	0.12	0.04	0.05	0.09	0.09	0.12	0.03	0.12	0.09
MP		1	0.86	0.18	0.42	0.48	0.5	0.48	0.58	0.19	0.12	0.39	0.33	0.71	0.09	0.56	0.39
90s			1	0.13	0.46	0.51	0.54	0.52	0.65	0.14	0.14	0.49	0.41	0.79	0.1	0.7	0.48
PPM				1	0.13	0.18	0.17	0.15	0.16	0.17	0.35	0.25	0.42	0.24	0.01	0.06	0.04
Gls					1	0.58	0.9	0.86	0.63	0.48	-0.2	0.43	0.11	0.62	-0.12	0.03	-0.06
Ast						1	0.88	0.62	0.82	0.6	-0.21	0.49	0.14	0.74	-0.13	0.13	-0.01
GA							1	0.84	0.81	0.61	-0.23	0.51	0.14	0.76	-0.14	0.09	-0.04
xG								1	0.67	0.51	-0.22	0.45	0.11	0.67	-0.13	0.07	-0.05
KP									1	0.71	-0.14	0.65	0.29	0.89	-0.09	0.22	0.07
SCA90										1	-0.23	0.47	0.11	0.61	-0.19	-0.18	-0.23
Cmp_percent											1	0.03	0.79	0.05	0.21	0.28	0.31
TB												1	0.64	0.7	-0.03	0.21	0.11
CmpTB													1	0.47	0.15	0.35	0.3
PassLive														1	-0.03	0.37	0.2
Tkl_percent															1	0.26	0.82
Int																1	0.76
TkIn																	1

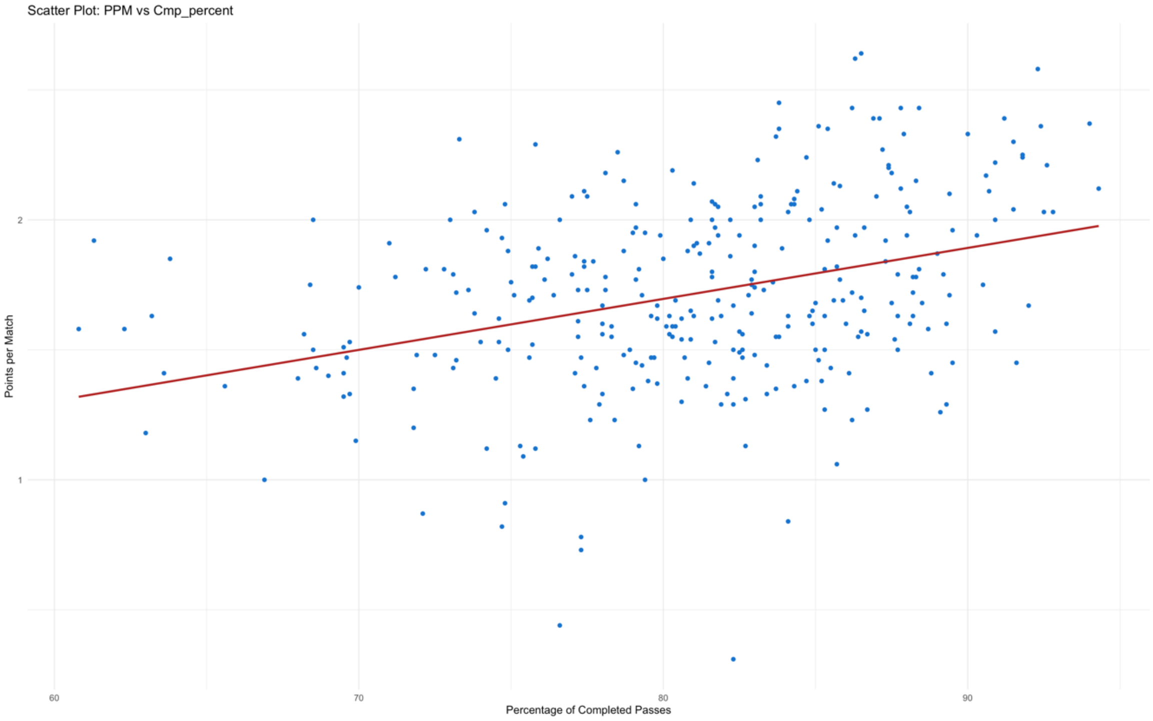
Then we move on to multiple linear regression with PPM as the dependent variable.



```
Call:
lm(formula = PPM ~ Gls + Ast + xG + SCA90 + Cmp_percent + Tkl_percent +
    Int, data = X2022_23_Midfielders_Stats)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.37087 -0.20618 -0.00085  0.20382  0.79882 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.487751   0.253867  -1.921   0.0556 .  
Gls         -0.008588   0.013354  -0.643   0.5206    
Ast          0.022294   0.010587   2.106   0.0360 *  
xG           0.026588   0.017092   1.556   0.1208    
SCA90        0.035282   0.021865   1.614   0.1076    
Cmp_percent  0.025438   0.003047   8.348 2.32e-15 ***
Tkl_percent -0.000427   0.001412  -0.302   0.7625    
Int         -0.002115   0.001731  -1.222   0.2225    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3234 on 311 degrees of freedom
Multiple R-squared:   0.22,	Adjusted R-squared:  0.2025 
F-statistic: 12.53 on 7 and 311 DF,  p-value: 3.712e-14
```

From the results, it can be observed that the variable 'Cmp_percent' is lower than each significance level. For this reason, the graph is shown to observe the dependence of one on the other.

```
> ggplot(X2022_23_Midfielders_Stats, aes(x = Cmp_percent, y = PPM)) +
+     geom_point(color = "dodgerblue3") +
+     geom_smooth(method = "lm", se = FALSE, color = "firebrick") +
+     geom_text(aes(label = paste("Corr =", round(cor(Cmp_percent, PPM), 2))),
+                     x = 2, y = 70, color = "black", size = 4, parse = TRUE) +
+     labs(title = "Scatter Plot: PPM vs Cmp_percent",
+                x = "Percentage of Completed Passes",
+                y = "Points per Match") +
+     theme_minimal()
```



4. Conclusion

The aim of this research was to find the midfielders who proved to be the most decisive in their teams over the course of the season.

From an offensive point of view, the players who were most prolific in terms of goals and assists were Kevin De Bruyne (Premier League winner with Manchester City), Jamal Musiala (Bundesliga winner with Bayern Munich) and Martin Ødegaard (placed second in England with Arsenal). Then followed a series of midfielders who were extremely successful such as Milinkovic Savic (Lazio), Julian Brandt (Borussia Dortmund), Pascal Groß (Brighton).

Defensively, midfielders such as Nicolas Höfler (Freiburg), Youssouf Fofana (Monaco), Valentin Rongier (Marseille), Moises Caicedo (Brighton), Aurélien Tchouaméni (Real Madrid) and Thiago Mandes (Lyon) distinguished themselves. From the graph, it is possible to notice a significant inferiority of Serie A midfielders.

Finally, as far as the best passers are concerned, some stand out, including Kimmich (Bayern Munich), Pedri (Barcelona), de Jong (Barcelona), Verratti (PSG), Bruno Fernandes (Manchester United), Lobotka (Napoli). Some midfielders appear in more charts such as De Bruyne, Tchouaméni or Ødegaard to testify their importance in the team.

Another player to be mentioned is Mac Allister with the highest percentage of expected goals who switched from Brighton to Liverpool in the summer.

5. References

https://www.kaggle.com/datasets/ameyaranade/big-5-european-leagues-player-stats-2022-23

https://fbref.com/en/

https://medium.com/@tacticaltouchline/data-scout-report-alexis-mac-allister-805a1d6998de