Application of PCA & Multiple Linear Regression with football data

Alfonso Marino

Dataset description

The dataset under analysis is the union of five datasets extracted from the *football-reference* site through the use of the **wordfootballR** library. The library allows you to do web scraping at certain sites with ad-hoc functions and choose the dataset you want to work with. In the case at hand, we went to work on data related to the defensive actions of teams from the five major European leagues (Serie A, Bundesliga, Ligue 1, La Liga, Premier League) in the 2023/2024 season. At the time of extraction, the dataset consists of the following variables:

- *Competition_Name* represents the name of the competition to which the observations belong;
- *Gender* concerns the gender of the participants in each team;
- *Country* represents the country of the competition;
- *Season_End_Year* indicates the year in which the sports season ended, in our case the year 2024 is mentioned because the season 2023/24 is being analyzed;
- *Squad* contains the names of the teams participating in the league;
- *Team_or_Opponent* makes a disinction about the statistics related to the team under consideration or the teams against it, in our case only the teams' own statistics were selected;
- *Num_Players* denotes the number of players used in all games played;
- *Mins_Per_90* denote the games played for each team;
- *Tkl_Tackles* counts the number of contrasts made by the players;
- *TklW_Tackles* represents the contrasts in which the team of the contrastor recovered possession of the ball;
- *Def 3rd_Tackles* counts the contrasts made in the defensive third of the field;
- *Mid 3rd_Tackles* counts the contrasts made in the middle part of the field;
- *Att 3rd_Tackles* counts the contrasts made in the offensive third of the field;
- *Tkl_Challenges* denotes the number of dribbling attempts countered;
- *Att_Challenges* represents the total number of drbbling attempts faced;
- *Tkl_percent_Challenges* represents the percentage of successful dribbling countered;
- *Lost_Challenges* counts the number of failed attempts to counter a dribbling attempt;
- *Blocks_Blocks* counts the number of times a defender blocked the ball by placing himself in its path;

- *Sh_Blocks* counts the number of times a shot was blocked by placing himself on its trajectory;
- *Pass_Blocks* counts the number of times a pass was blocked by placing himself on its trajectory;
- *Int* represents interceptions, i.e., the action in which a defender anticipates an opponent's pass, interrupting the trajectory of the ball and taking control of it or deflecting it significantly. ;
- *Tkl_plus_Int* represents the number of setbacks made plus interceptions;
- *Clr* counts sweeps, i.e., shots by defenders aimed at moving the ball away from their own area;
- *Err* counts errors that lead to an opponent's shot.

For ease of use and understanding, the names of some variables have been shortened, as well as some excluded from the analysis because they are not relevant.

```
df = subset(df,
            select = -c(Competition_Name, Gender, Country, Season_End_Year,
                        Team_or_Opponent, Num_Players, Mins_Per_90, Tkl_plus_Int, Blocks_Blocks))

df = rename(df,
       Tkl = Tkl_Tackles,
       TklWin = TklW_Tackles,
       Def.3rd_Tkl = "Def 3rd_Tackles",
       Mid.3rd_Tkl = "Mid 3rd_Tackles",
       Att.3rd_Tkl = "Att 3rd_Tackles",
       Tkl_Drib = Tkl_Challenges,
       Atmp_Drib = Att_Challenges,
       Tkl_Drib.Perc= Tkl_percent_Challenges,
       Lost_Drib = Lost_Challenges,
       Sh_Blk = Sh_Blocks,
       Pass_Blk = Pass_Blocks
       )
```

Goals

The following paper aims to analyze data on the defensive phase of major European teams in the 2023/24 season by highlighting the different styles of play. At first, a descriptive analysis will be carried out to explain the distribution of variables using statistical tools and graphs. Then it will continue by reducing the dimensionality of the dataset with principal component analysis. Finally, it will be concluded by estimating a multiple regression model to look for linear relationships among the selected variables.

Descriptive analysis

For the purpose of the analysis, the following libraries were essential: - *ggplot2*, *GGally*, and *gridExtra* for graphical representations; - *corrplot* for correlation matrix representation; - *worldfootballR* for data extraction; - *tidyverse* for data manipulation; - *factoextra* for Principal Component Analysis applocation; - *psych* and *PerformanceAnalytcs* for data description; - *car* and *lmtest* for regression specificity tests.

A visualization of the data shows how there are 20 observations for 15 variables, with the majority being numerical:

```
## 'data.frame':    96 obs. of  15 variables:
##  $ Squad        : chr  "Atalanta" "Bologna" "Cagliari" "Empoli" ...
##  $ Tkl          : num  613 633 540 618 564 545 625 621 540 577 ...
##  $ TklWin       : num  369 385 315 378 320 322 352 387 321 369 ...
##  $ Def.3rd_Tkl  : num  273 299 272 311 242 244 315 302 252 270 ...
##  $ Mid.3rd_Tkl  : num  249 267 203 227 231 225 247 245 219 233 ...
##  $ Att.3rd_Tkl  : num  91 67 65 80 91 76 63 74 69 74 ...
##  $ Tkl_Drib     : num  291 303 264 300 275 253 283 293 234 264 ...
##  $ Atmp_Drib    : num  597 579 525 564 512 525 571 578 474 445 ...
##  $ Tkl_Drib.Perc: num  48.7 52.3 50.3 53.2 53.7 48.2 49.6 50.7 49.4 59.3 ...
##  $ Lost_Drib    : num  306 276 261 264 237 272 288 285 240 181 ...
##  $ Sh_Blk       : num  119 113 113 154 81 143 162 133 116 124 ...
##  $ Pass_Blk     : num  364 313 260 308 307 296 300 281 249 251 ...
##  $ Int          : num  340 258 265 263 236 241 332 302 280 253 ...
##  $ Clr          : num  614 677 816 840 509 789 796 784 576 725 ...
##  $ Err          : num  7 17 12 9 8 14 11 15 7 4 ...
```

Before going into the descriptive analysis, NA values were checked for presence and then removed if necessary. Verification which was unsuccessful:

```
##         Squad           Tkl        TklWin   Def.3rd_Tkl   Mid.3rd_Tkl
##             0             0             0             0             0
##   Att.3rd_Tkl      Tkl_Drib     Atmp_Drib Tkl_Drib.Perc     Lost_Drib
##             0             0             0             0             0
##        Sh_Blk      Pass_Blk           Int           Clr           Err
##             0             0             0             0             0
```

Specifically, *descriptive analysis* is a set of techniques used to describe and summarize data in a meaningful way. This type of analysis focuses on the representation of data through statistical measures and graphical displays, enabling an understanding of the main characteristics of a dataset. Using the *describe* function of the *psych* library, numerous statistical measures were extracted including *location indices*, *skewness indices* and *variability indices*:

1. **vars**. The number of variables in the dataset.

2. **n**. The number of non-missing observations for each variable.

3. **mean**. The arithmetic mean represents that value around which all other observations in the sample tend to cocnenter:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

4. **sd**. The standard deviation of the values in the variable, which measures the dispersion of the data around the mean:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

5. **median**. The median of the values in the variable, which represents the central value when the data are sorted in ascending order.

$$\text{Median} = x_{(\frac{n+1}{2})}$$

with odd observations;

$$\text{Mediana} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

with even observations.

6. **trimmed**. The "trimmed" mean of the values in the variable, calculated by excluding a percentage of the extreme values:

$$\text{Trimmed} = \frac{\sum_{i=k+1}^{n-k} x_i}{n - 2k}$$

7. **mad**. The absolute deviation from the median, a measure of robust dispersion at extreme values:

$$\text{MAD} = \text{median}(|x_i - \text{median}(x)|)$$

8. **min**. The minimum value observed in the variable.

9. **max**. The maximum value observed in the variable.

10. **range**. The difference between the maximum and minimum value in the variable:

$$\text{range} = \max - \min$$

11. **skew**. The measure of the skewness of the data distribution. When the skewness is less than 0, the distribution is negative (or right-handed), when it is greater than 0 it is called positive (or left-handed), and finally when it is equal to 0, the distribution will be symmetrical. Asymmetry can be calculated with different indices:

- **Asimmetria di Pearson**: $\bar{x} - \text{Median}$
- **Asimmetria di Pearson standardizzata:** $\frac{(\bar{x} - \text{Median})}{\sigma}$
- **Indice di Yule e Bowley**: $\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$
- **Indice di Fisher:** $\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$

12. **kurtosis**. The measure of the "sharpness" of the data distribution. If the coefficient is 0, it will be referred to as a Gaussian or *mesocurtic* curve; if the coefficient is greater than 0, the curve will be called *leptokurtic*, with the tails trending to 0 very quickly compared to a normal; when the coefficient is less than 0, there will be a *platicurtic* curve, which will be flatter than a normal:

$$\text{kurtosis} = \frac{1}{n} \sum_{i=1}^{k} \left(\frac{x_i - \bar{x}}{s} \right)^4 n_i - 3$$

13. **se**. The standard error of the mean, which estimates the precision of the sample mean relative to the population mean:

$$\text{SE} = \frac{s}{\sqrt{n}}$$

The distribution of opposed dribbles (Tkl_Drib), blocked passes (Pass_Blk) and rejections (Clr) shows a slight negative skewness, suggesting that most teams are close to or above average, with a few downward

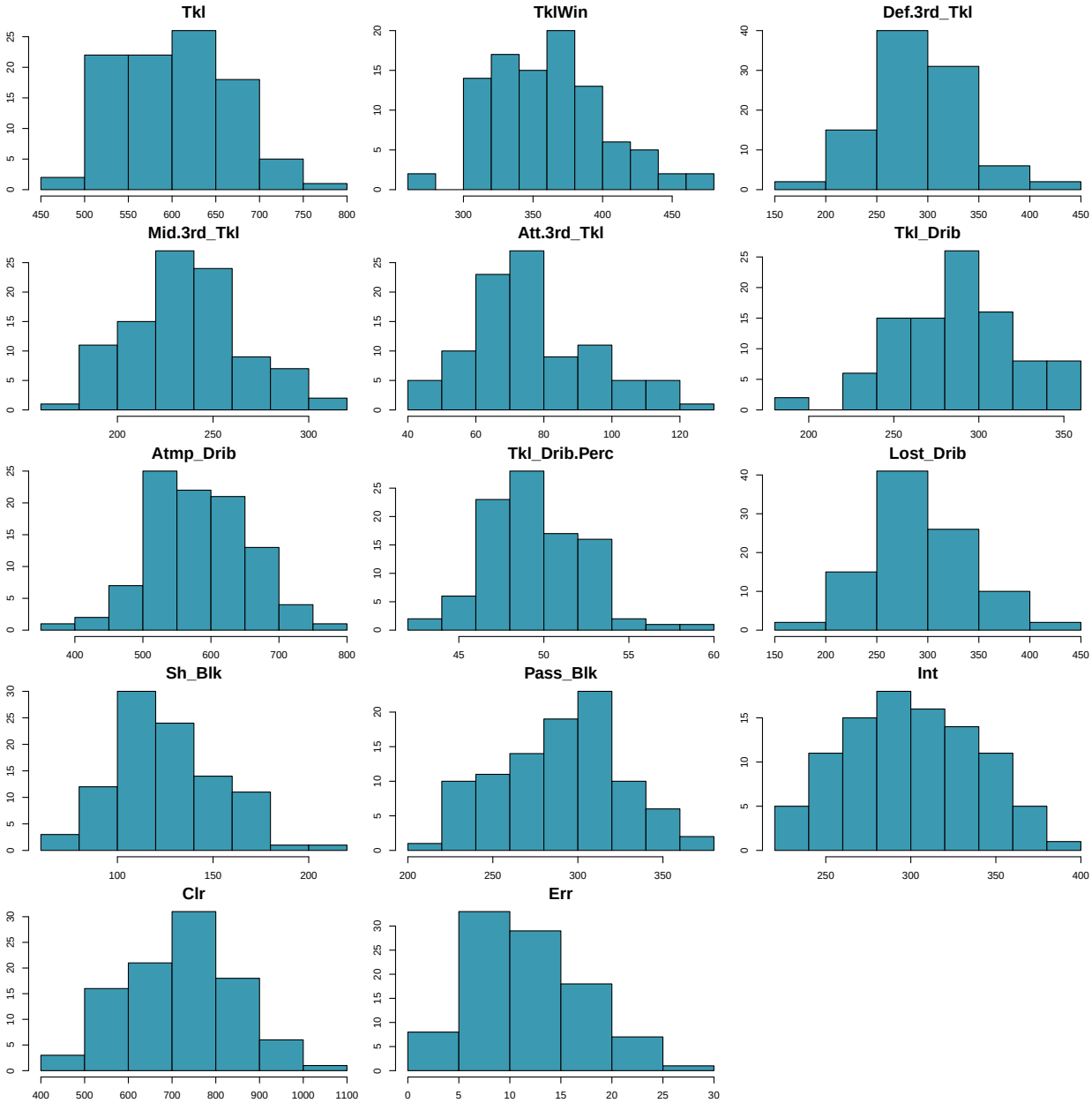
exceptions. However, variability is evident, especially for rejections (Clr) themselves and the number of dribbles faced (Atmp_Drib), which show a wide range and significant standard deviation, denoting marked differences between teams. This may reflect different tactical approaches, with some teams preferring to defend more compactly and others adopting a more aggressive and proactive defense. Interceptions (Int) and thwarted dribble attempts (Atmp_Drib) show a more even distribution, with a slight positive skewness, indicating that few teams particularly excel in this aspect, perhaps due to individual skills. In addition, the low kurtosis for most variables suggests a flat distribution, with no extreme peaks, confirming relatively stable defensive behavior among teams. This reflects uniform tactical preparation in terms of defensive fundamentals, but with variations in individual capabilities. Overall, the data show consistency in basic defensive tactics, but also diversity in specific skills and approaches, likely influenced by each coach's playing style and tactical philosophy and culture.

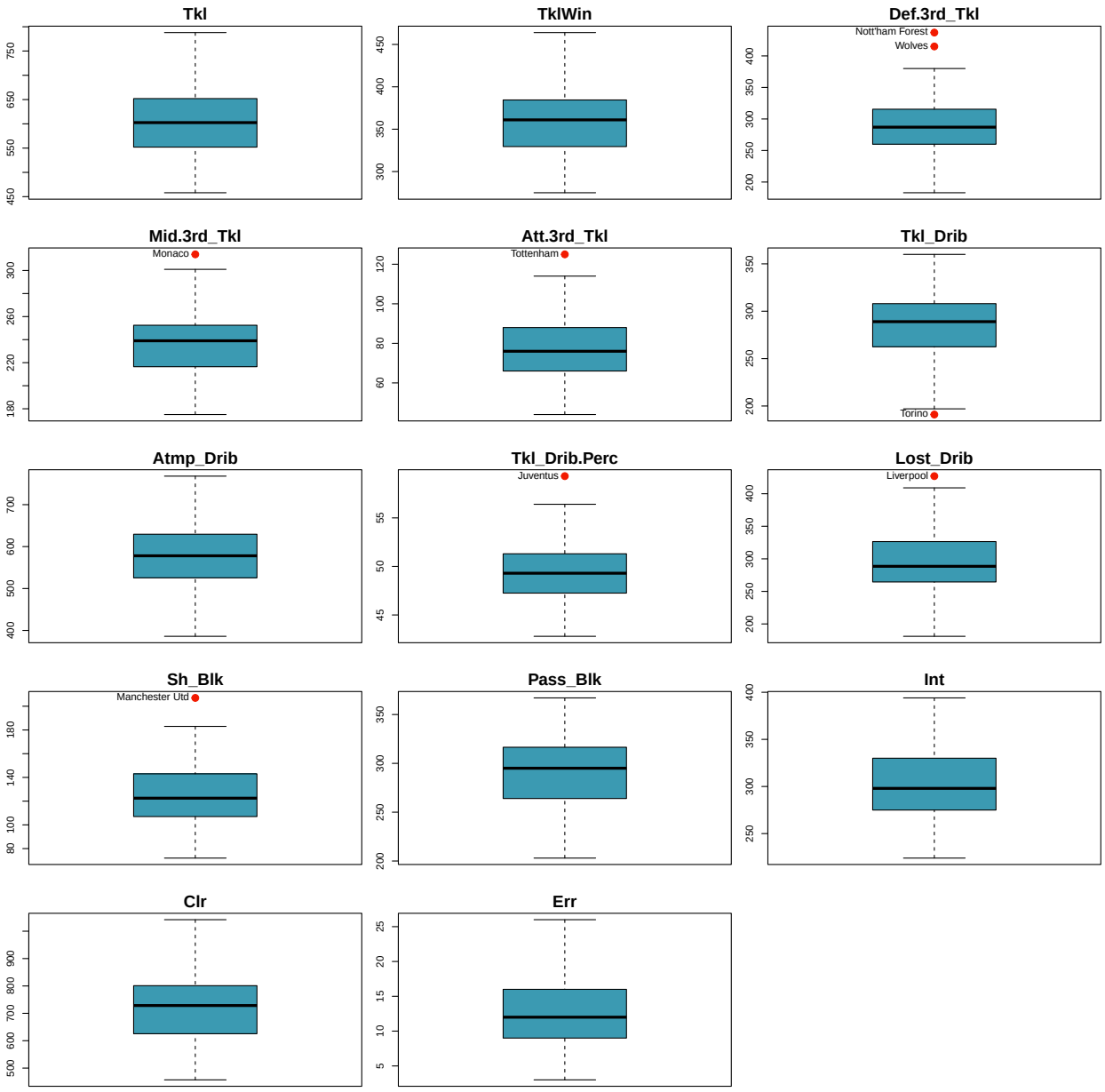
```
##               vars  n   mean     sd median trimmed    mad   min    max range
## Tkl              1 96 603.61  63.54  602.5  600.83  74.13 458.0  788.0 330.0
## TklWin           2 96 361.75  40.35  361.0  359.64  40.77 275.0  464.0 189.0
## Def.3rd_Tkl      3 96 290.23  46.06  287.0  288.90  41.51 183.0  437.0 254.0
## Mid.3rd_Tkl      4 96 236.09  29.01  239.0  235.18  25.95 175.0  314.0 139.0
## Att.3rd_Tkl      5 96  77.29  17.14   76.0   76.41  16.31  44.0  125.0  81.0
## Tkl_Drib         6 96 286.67  34.74  289.0  286.95  34.10 191.0  360.0 169.0
## Atmp_Drib        7 96 581.02  73.88  578.0  581.06  77.10 386.0  768.0 382.0
## Tkl_Drib.Perc    8 96  49.45   2.95   49.3   49.38   2.97  42.8   59.3  16.5
## Lost_Drib        9 96 294.35  46.17  288.5  292.78  43.00 181.0  427.0 246.0
## Sh_Blk          10 96 126.12  26.63  122.5  124.91  25.95  72.0  207.0 135.0
## Pass_Blk        11 96 290.73  35.47  295.0  290.96  37.81 203.0  367.0 164.0
## Int             12 96 300.88  37.83  298.0  300.46  43.74 224.0  394.0 170.0
## Clr             13 96 719.44 125.35  728.5  719.91 130.47 457.0 1042.0 585.0
## Err             14 96  12.35   5.06   12.0   12.12   4.45   3.0   26.0  23.0
##                 skew kurtosis    se
## Tkl             0.39    -0.22  6.49
## TklWin          0.37    -0.25  4.12
## Def.3rd_Tkl     0.38     0.51  4.70
## Mid.3rd_Tkl     0.21    -0.18  2.96
## Att.3rd_Tkl     0.50    -0.16  1.75
## Tkl_Drib       -0.18    -0.14  3.55
## Atmp_Drib       0.01    -0.22  7.54
## Tkl_Drib.Perc   0.38     0.34  0.30
## Lost_Drib       0.31     0.10  4.71
## Sh_Blk          0.47    -0.05  2.72
## Pass_Blk       -0.12    -0.66  3.62
## Int             0.13    -0.69  3.86
## Clr            -0.05    -0.51 12.79
## Err             0.44    -0.35  0.52
```

Graphical representations

Histograms and boxplots

What is obtained numerically can be well represented graphically using histograms and boxplots. The **histograms** show the frequency of variable values, allowing one to quickly identify the shape of the distribution, the presence of any skewness, and the density of the data around the mean. The **boxplots**, on the other hand, provide a concise representation of the data across quartiles, highlighting the median, interquartile range and the presence of outliers.


Tkl
TklWin
Def.3rd_Tkl
Mid.3rd_Tkl
Att.3rd_Tkl
Tkl_Drib
Atmp_Drib
Tkl_Drib.Perc
Lost_Drib
Sh_Blk
Pass_Blk
Int
Clr
Err



From the box plots we note eight variables characterized by outliers, i.e., values that are much larger or much smaller than the rest of the distribution. More precisely, **outliers** are values in a data set that deviate significantly from most other observations. They may represent outliers that do not follow the same trend as the other data or may be the result of measurement or data collection errors. There are several methods for identifying outliers, including evaluation by descriptive statistics such as mean and standard deviation, or using graphical tools, as in our case, where the *interquartile range* (**IQR**) method was applied. IQR-based outliers are identified as values that fall outside the range:

$$\text{Outlier} = [Q1 - 1.5 \times \text{IQR}, Q3 + 1.5 \times \text{IQR}]$$

Where:

- **Q1** (*First quartile*): Is the value that separates the bottom 25% of the data from the top 75%.
- **Q3** (*Third quartile*): It is the value that separates the upper 75% of the data from the lower 25%.

It is noticeable that **Torino** has very low observations regarding the number of dribbles countered, this highlights a more conservative defensive strategy or at least less focused on directly countering the opponent's dribble. These teams might prefer a more organized and compact defense, focused on blocking passing lines and containing opponents without necessarily engaging in many challenges to dribbling.

Juventus, on the other hand, records a high percentage of successfully countered dribbles, suggesting that players were particularly adept at maintaining position, reading opponents' moves and intervening effectively to recover the ball without conceding space for penetrations, a characteristic feature of the last few years of *Allegri's* teams.

The extreme value for attempted tackles in the offensive third is recorded by **Tottenham**, which highlights how Ange Postecoglu has totally changed the way the team has played in recent years after the team's less-than-positive experiences with Mourinho and Conte, making pressing his main weapon for recovering the ball in the opponent's half, trying to create opportunities for quick counter-attacks or exploiting opponents' mistakes.

The **Nottingham Forest** and **Wolverampton**, on the other hand, recorded high values for the number of tackles in the defensive third. This approach may be used by teams that prefer to keep a low, compact defensive line, trying to take advantage of counterattacks. It could also be symptomatic of teams that have difficulty keeping possession of the ball or controlling the midfield.

The extreme value for dribbles conceded is recorded by **Liverpool**, this is essentially because Jurgen Klopp makes *Gegenpressing* his workhorse, which allows him to implement strong pressure on the ball carrier as soon as it is lost. However, this strategy involves continuous 1 vs. 1 and exposure of the defense to counterattacks.

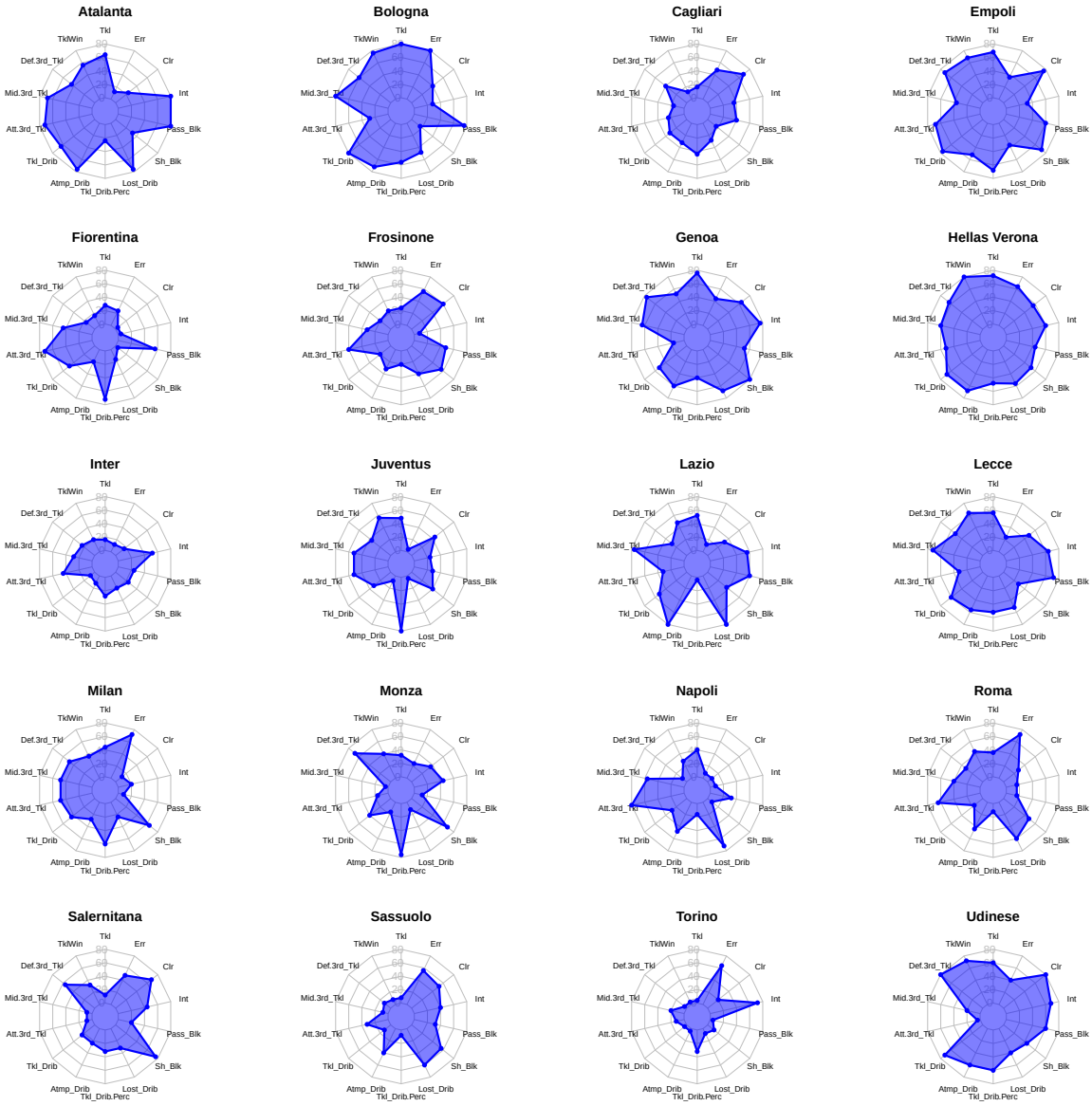
The figure of blocked shots (Sh_Blk) highlights the difficulties of the defensive phase of **Manchester United** this year, which ended the season with more goals conceded than scored. While this value could indicate a very active and aggressive defense, trying to prevent opponents from shooting on goal with timely interventions, it could also mean that the team is often under pressure, with opponents attempting many shots, forcing defenders to intervene frequently to block shot attempts.

Finally, it is noticeable that **Monaco** reports very high values for tackles attempted in the central third of the field, this reflects the actions of very dynamic midfielders who are active in breaking the opponents' passing lines and recovering possession of the ball, such as Fofana and Zakaria.

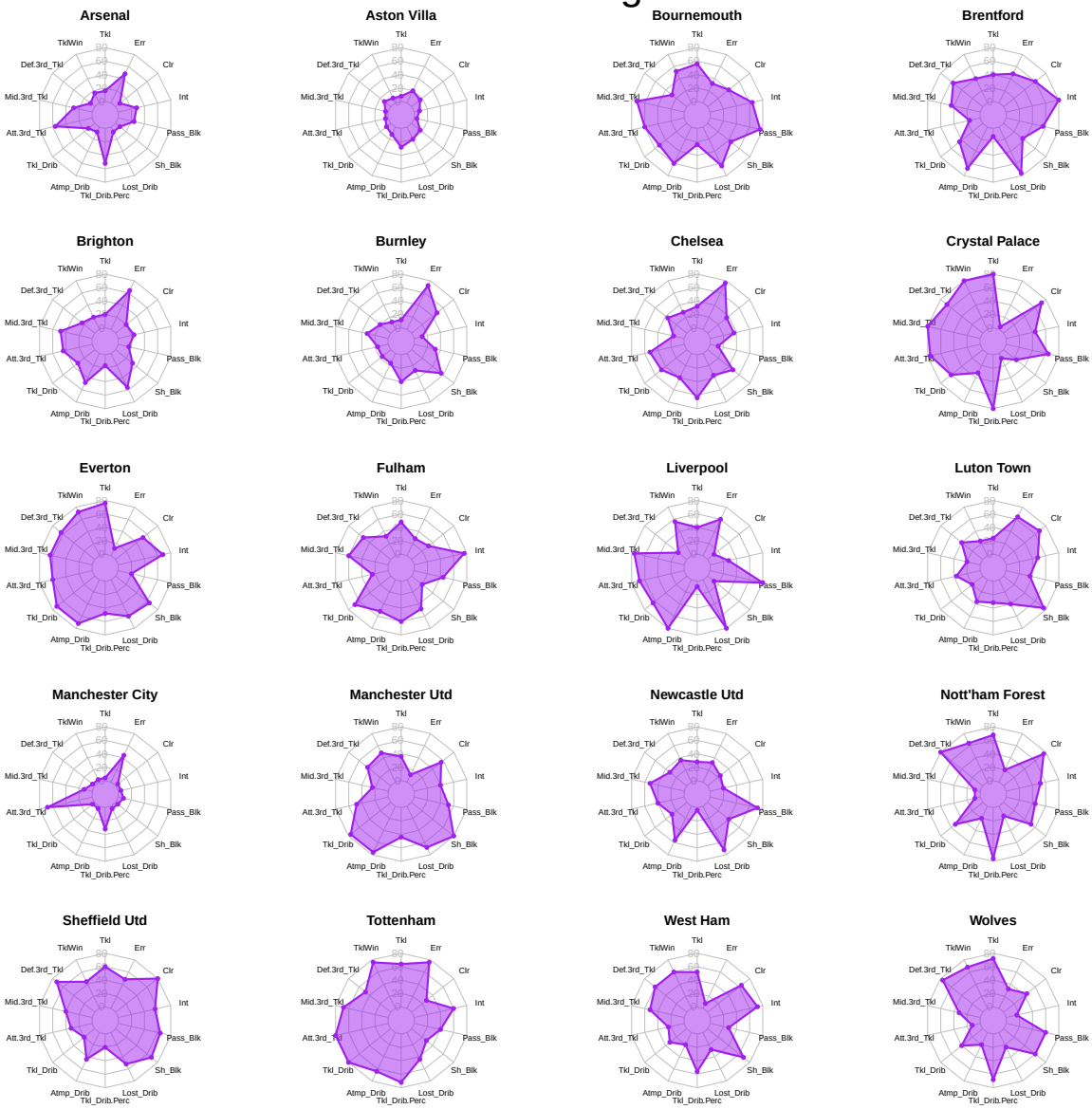
Percentile Ranking Plot

A general overview of the characteristics of each team can be seen in the following box in which *percentile ranking* is used via *radarchart.* Percentile ranking is a statistical measure that indicates the position of a value within a distribution of data. It expresses the percentage of data that are below a given value. For example, if a value is at the 70th percentile, it means that 70% of the data in the distribution is below that value.

Serie A

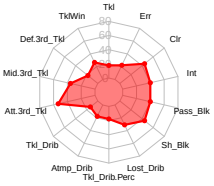

Atalanta
Bologna
Cagliari
Empoli
Fiorentina
Frosinone
Genoa
Hellas Verona
Inter
Juventus
Lazio
Lecce
Milan
Monza
Napoli
Roma
Salernitana
Sassuolo
Torino
Udinese
Tkl
Err
Clr
Int
Pass_Blk
Sh_Blk
Lost_Drib
Tkl_Drib.Perc
Atmp_Drib
Tkl_Drib
Att.3rd_Tkl
Mid.3rd_Tkl
Def.3rd_Tkl
TklWin

Premier League

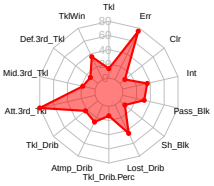


Bundesliga

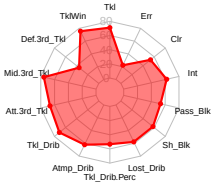
Augsburg



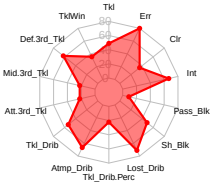
Bayern Munich



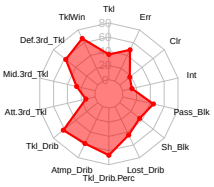
Bochum



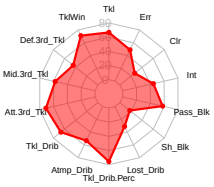
Darmstadt 98



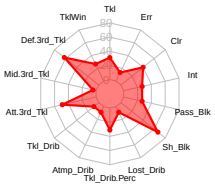
Dortmund



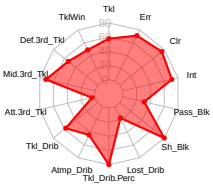
Eint Frankfurt



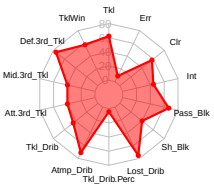
Freiburg



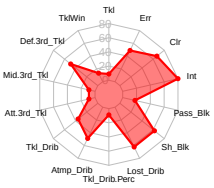
Gladbach



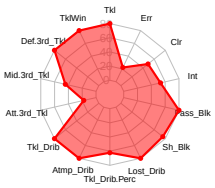
Heidenheim



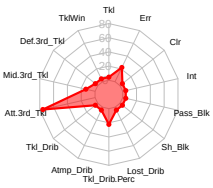
Hoffenheim



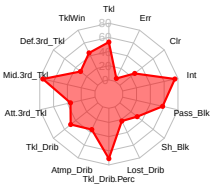
Köln



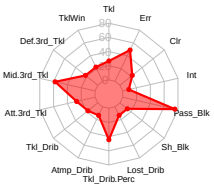
Leverkusen



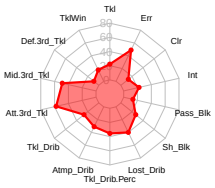
Mainz 05



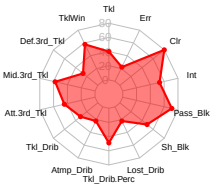
RB Leipzig



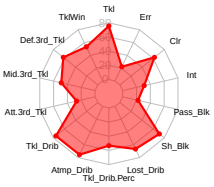
Stuttgart



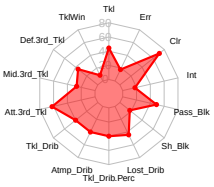
Union Berlin



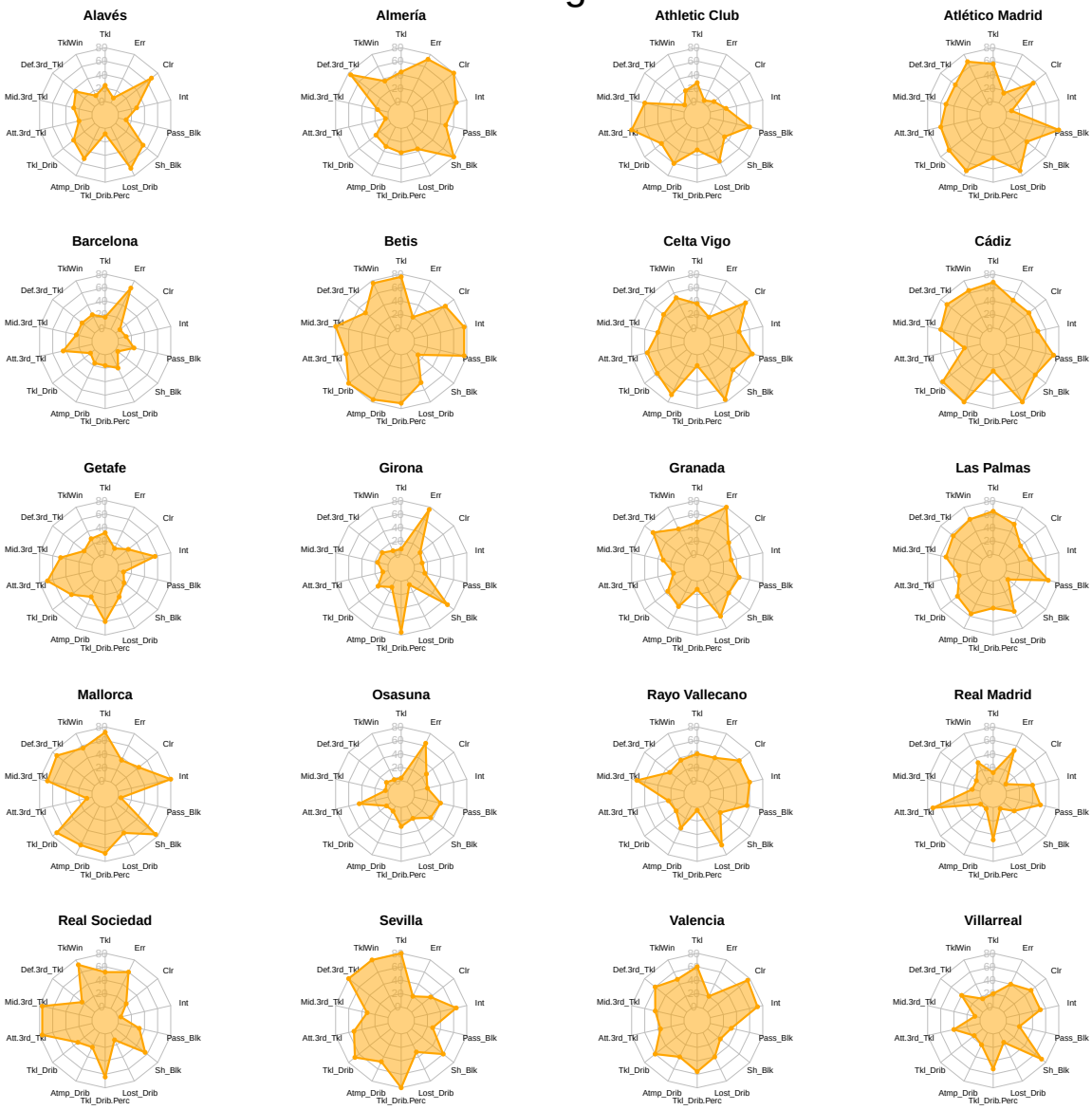
Werder Bremen



Wolfsburg



La Liga


Alavés
Almería
Athletic Club
Atlético Madrid
Barcelona
Betis
Celta Vigo
Cádiz
Getafe
Girona
Granada
Las Palmas
Mallorca
Osasuna
Rayo Vallecano
Real Madrid
Real Sociedad
Sevilla
Valencia
Villarreal
Tkl
Err
Clr
Int
Pass_Blk
Sh_Blk
Lost_Drib
Tkl_Drib.Perc
Atmp_Drib
Tkl_Drib
Att.3rd_Tkl
Mid.3rd_Tkl
Def.3rd_Tkl
TklWin

Putting the focus on the teams that won their respective championships (Inter Milan, Leverkusen, Real Madrid, PSG, Manchester City), what they have in common is the tendency to make a high number of tackles in the offensive third, so as to recover the ball as close to the opponent's goal and as quickly as possible. This strategy confirms the theory that the less time a team's transitions last, the greater the chances of victory.

In contrast, for the teams relegated or engaged in any playoffs, it is evident that throughout the year they have been targets for opposing raids. Demonstrating this are the high values for rejections, to push the ball away from their own area, and of errors that led to an opponent's conclusion, consequently, they also record a high figure of blocked shots.

Correlation Matrix

To understand the relationships between the variables used, a useful tool is definitely the **correlation matrix**. The correlation matrix is a table showing the correlation coefficients between the variables in the dataset, that is, the strength of the linear dependence link between two variables. These coefficients can range from -1 to 1, where: - A value close to 1 indicates a strong positive correlation, i.e., the variables tend to vary together in the same direction. - A value close to -1 indicates a strong negative correlation, i.e., the variables tend to vary together in opposite directions.- A value close to 0 indicates a weak or no correlation, i.e., the variables do not show a linear relationship.

In the correlation matrix, the values on the main diagonal are always 1, since they represent the correlation of a variable with itself. The other values in the matrix indicate the correlation between pairs of variables. These values may result from the application of different methods such as *Kendall's* or *Spearman's*, or, as in the present case, *Pearson's*:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

dove:

- r Is Pearson's correlation coefficient;
- n is the number of observations;
- $\sum xy$ is the sum of the products of the differences between the observations of x and y.
- $\sum x$ and $\sum y$ are the sums of the observations of x and y.
- $\sum x^2$ e $\sum y^2$ are the sums of the squares of the observations of x e y.

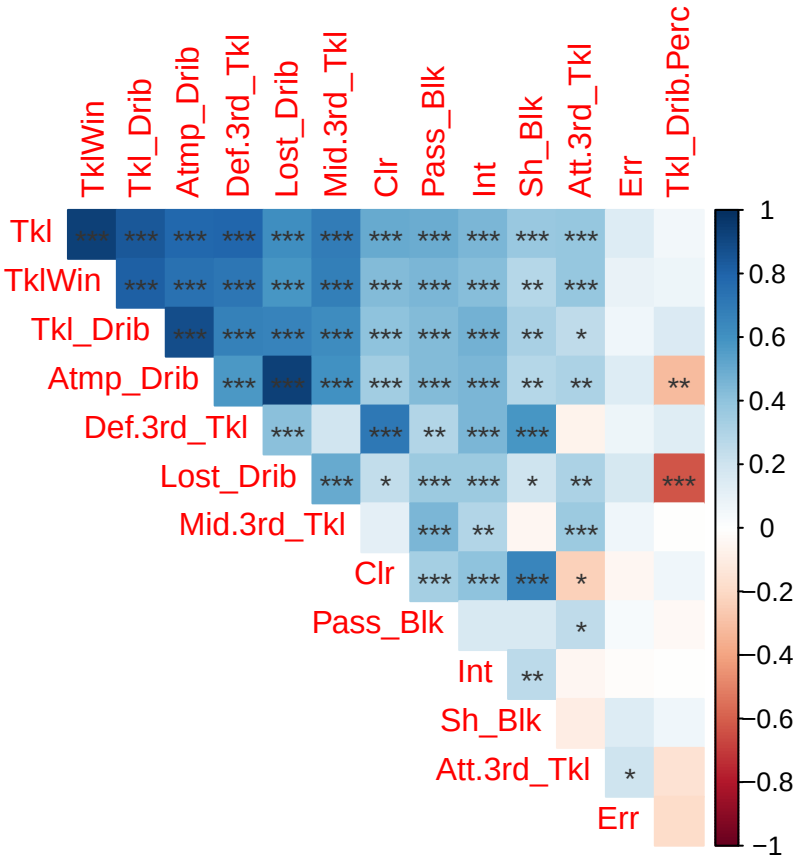
In the analysis of correlations, the **p-value** was also used to highlight the statistical significance of each, helping to understand whether the correlation between two variables is due to chance or is actually relevant.

Theoretically, the p-value represents the probability of obtaining a test statistic at least as extreme as the observed one, assuming the null hypothesis is true. The null hypothesis in this case is that there is no correlation between the two variables. If the p-value is less than a predefined significance level (e.g., 0.05, 0.01 or 0.001), we can reject the null hypothesis and conclude that there is a significant correlation between the variables. The formula for calculating p-value in correlation analysis is based on Student's t distribution. If $\hat{r}$ is the sample correlation coefficient between two variables, then the t test statistic is calculated as:

$$t = \frac{\hat{r}\sqrt{n-2}}{\sqrt{1-\hat{r}^2}}$$

where n is the number of observations.

This test statistic follows a t distribution with $n-2$ degrees of freedom. The p-value associated with this statistic is obtained by comparing the absolute value of t with the Student's t distribution. In R, the calculation of the p-value for each pair of variables in the correlation matrix is performed by the function *cor.mtest()*, which generates a matrix of p-values. These p-values are then used to determine the significance of the correlations displayed in the graph.



Analysis of the correlation matrix and related significance tests reveals several important relationships among the variables. The results show that many variables are strongly correlated with each other, indicating interconnected defensive behaviors that may influence overall team performance. The most significant correlations are observed between tackles made (Tkl) and tackles won (TklWin), as well as between dribbles conceded (Lost_Drib) and dribble attempts faced (Atmp_Drib). These relationships suggest that teams that are more active in counterattacks also tend to be effective in recovering ball possession and successfully dealing with opposing dribbles. This is an indicator of a solid and aggressive defense capable of disrupting opposing actions and creating transition opportunities for their team. Another interesting correlation is between counterattacks in the defensive third (Def.3rd_Tkl) and blocked shots (Sh_Blk), but also with rejections (Clr), which highlights how a defensive approach marked by low blocking defense pushes opposing players to kick more on goal to find alternative solutions. In addition, the relationship between dribbles conceded (Lost_Drib) and pass blocks (Pass_Blk) can highlight the teams' qualities of still maintaining good defensive solidity despite lost contrast, and of good reading of the game even in unfavorable situations.An interesting inverse relationship is observed between lost dribbles (Lost_Drib) and the percentage of successfully countered dribbles (Tkl_Drib.Perc). This negative correlation indicates that as the number of dribbles lost increases, the percentage of successfully countered dribbles decreases. This suggests that greater ineffectiveness in countering opponent dribbles leads to a higher number of dribbles lost, highlighting areas of improvement in individual defensive ability.

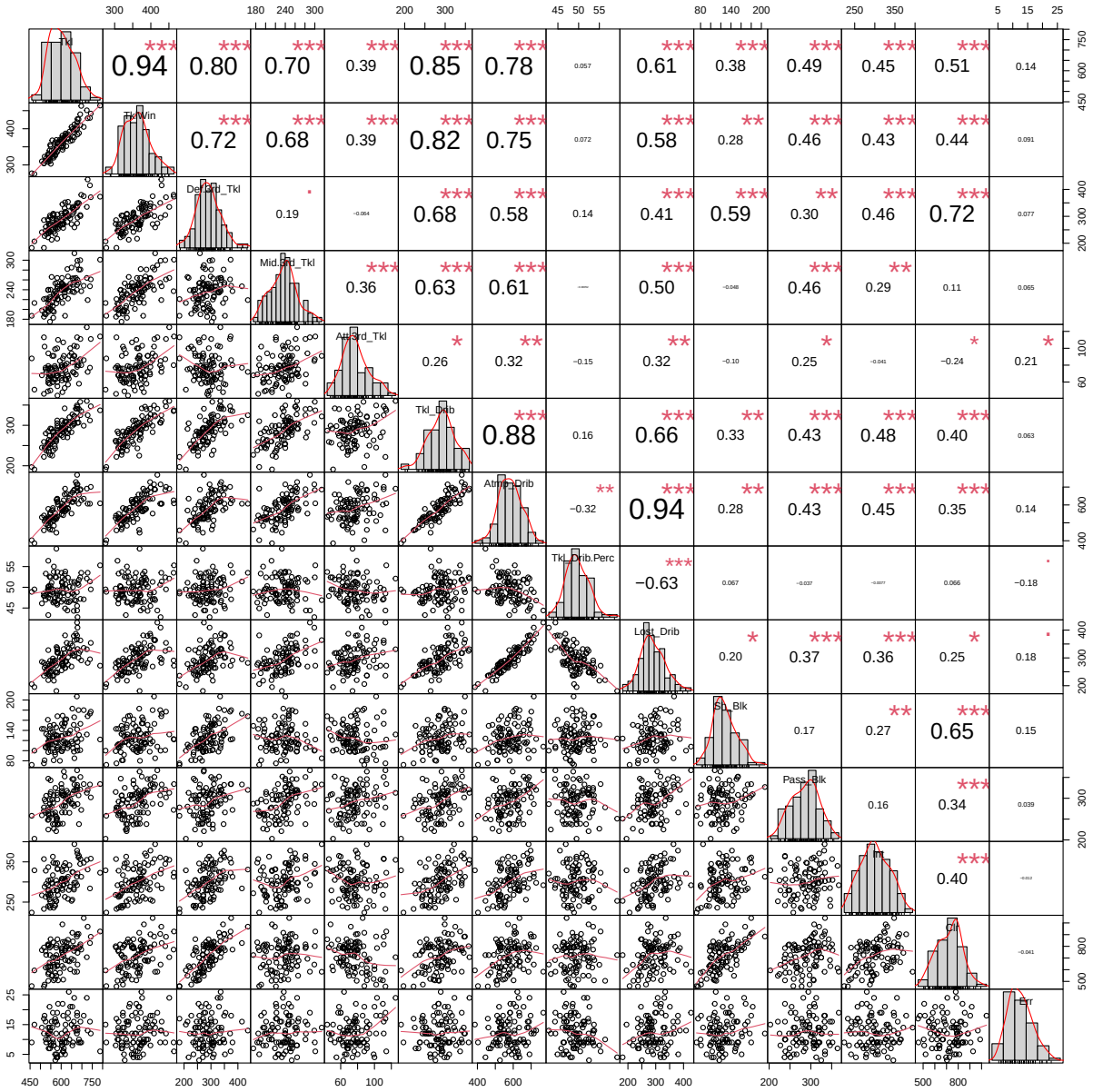
Principal Component Analysis

In statistical data analysis, **Principal Component Analysis** (*PCA*) is a dimensionality reduction technique used to simplify complex datasets. When dealing with a large number of correlated variables, as in our case with the defensive variables of Serie A teams, it is useful to find a way to reduce the number of variables while retaining most of the information in the original dataset. PCA allows the original correlated variables

to be transformed into a new set of uncorrelated variables called *principal components*. In the context at hand, it is useful because identifying defensive principal characteristics can provide valuable insights for improving game strategies. Determining how many principal components to keep is critical; the goal is to select only PCs that capture a significant amount of variance of the original data. This is evaluated through the eigenvalues of the or sample correlation matrix. There are several criteria for making this decision; the most common are discussed below:

1. **Variance Explained**: A sufficient number of components are kept to explain a significant percentage of the total variance of the original data, usually between 70% and 90%.
2. **Scree Plot**: This plot shows the eigenvalues ordered decreasingly on the y-axis with respect to their order number on the x-axis. If the first eigenvalues dominate in magnitude and the remainder are very small, the scree plot will show an "elbow" indicating the point at which the change in the magnitude of the eigenvalues becomes least significant. This "elbow" can be used to determine the number of PCs to keep.
3. **Kaiser's rule**: Only PCs whose eigenvalues exceed a value of 1 are kept.This criterion is based on the idea that each standardized variable should contribute at least one unit of variance. However, this rule is controversial and a modified version suggests keeping PCs with eigenvalues greater than 0.7, especially for standardized data. For nonstandardized data, this rule may not be applicable.

PCA consists of several key steps. First, it is critical to **standardize** the variables to have a mean of zero and a standard deviation of one, ensuring that all variables are comparable. This is especially important when the original variables have different scales. Next, the **correlation matrix** of the standardized variables is calculated to measure the linear relationship between each pair of variables. The graph displays any correlation that exists between the variables, specifically: - scatter plots are depicted on the left of the graph; - along the main diagonal, the variables are described with histograms and attached trend line; - on the right of the graph, correlation coefficients and their statistical siginificance are displayed.



Then, we continue with the **decomposition of the correlation matrix** in its principal components through the singular value decomposition algorithm (SVD) or by the analysis of *eigenvalues* and *eigenvectors*. Eigenvalues represent the amount of total variability observed on the original variables, "explained" by each main component; eigenvectors instead represent the corresponding (orthogonal) directions of maximum variability extracted from the principal components.

Each principal component obtained is a linear combination of the original variables, ordered according to the amount of variance explained. The first major component captures most of the variance in the dataset, the second major component captures most of the remaining variance, and so on. This process allows us to reduce the complexity of the dataset while maintaining the most relevant information. The formula for the first principapl component z_1 is given by:

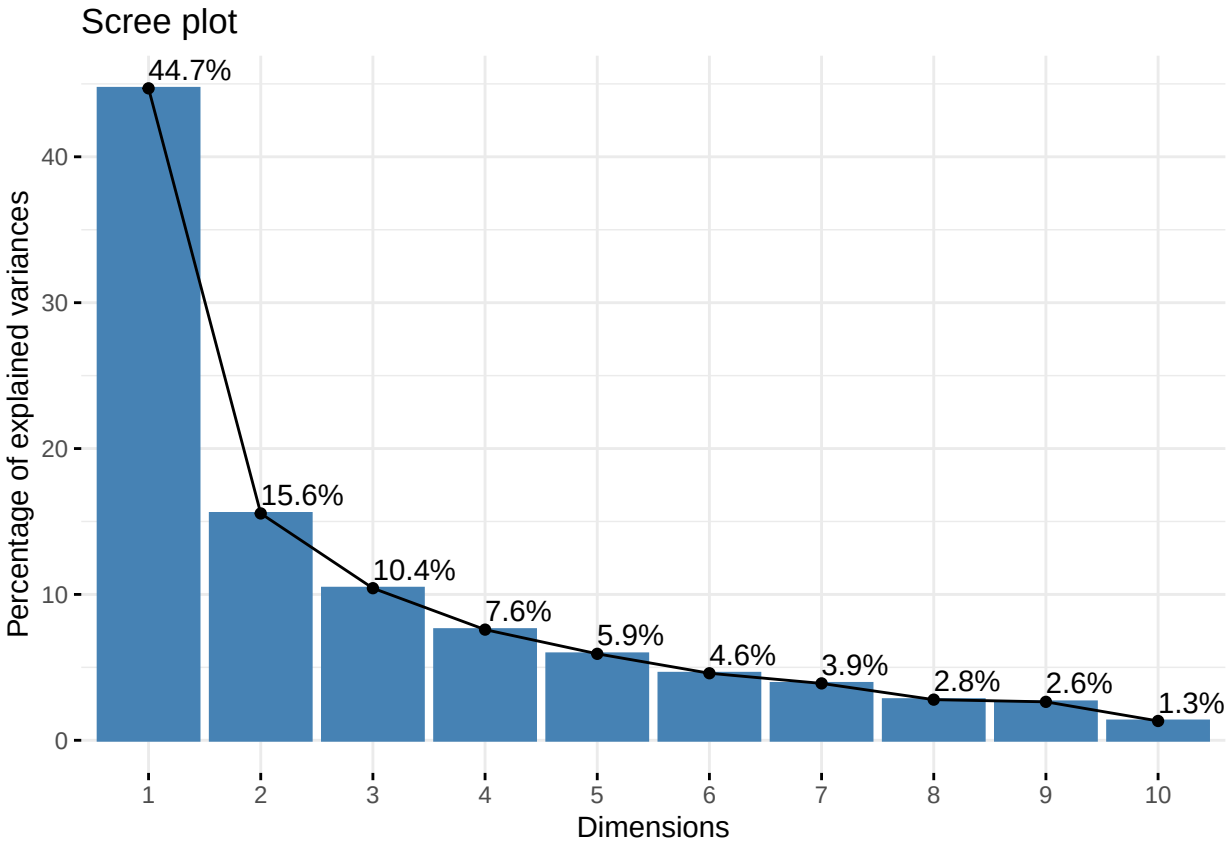
$$z_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$$

where $a_{11}, a_{12}, ldots, a_{1p}$ are the weighting coefficients (eigenvectors) and $x_1, x_2, ldots, x_p$ are the original

variables.

From the *Scree Plot* we can see that the first four PCs explain more than 75% of global variability, with coefficients:

```
## Importance of components:
##                            Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.5014284 1.4755006 1.2081924 1.03064807 0.91108131
## Proportion of Variance 0.4469389 0.1555073 0.1042663 0.07587396 0.05929065
## Cumulative Proportion  0.4469389 0.6024462 0.7067125 0.78258646 0.84187711
##                            Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
## Standard deviation     0.80220504 0.73899188 0.62468263 0.60753708 0.43083840
## Proportion of Variance 0.04596664 0.03900779 0.02787346 0.02636438 0.01325869
## Cumulative Proportion  0.88784375 0.92685153 0.95472499 0.98108937 0.99434806
##                           Comp.11      Comp.12      Comp.13 Comp.14
## Standard deviation     0.272501929 0.0697841536 1.734743e-08       0
## Proportion of Variance 0.005304093 0.0003478449 2.149523e-17       0
## Cumulative Proportion  0.999652155 1.0000000000 1.000000e+00       1
##
## Loadings:
##               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## Tkl            0.380         0.150  0.106         0.102         0.245
## TklWin         0.362         0.201                0.127         0.319
## Def.3rd_Tkl    0.309  0.335                       0.190  0.126  0.378 -0.280
## Mid.3rd_Tkl    0.268 -0.258  0.316 -0.107        -0.293  0.191         0.669
## Att.3rd_Tkl    0.126 -0.436  0.191  0.333         0.363 -0.626
## Tkl_Drib       0.361         0.160         0.162         0.194 -0.372 -0.263
## Atmp_Drib      0.361 -0.167 -0.147 -0.141         0.105  0.168 -0.254 -0.172
## Tkl_Drib.Perc         0.359  0.645  0.194  0.157               -0.255 -0.161
## Lost_Drib      0.306 -0.273 -0.356 -0.191         0.104  0.123 -0.127
## Sh_Blk         0.177  0.394 -0.279  0.305 -0.141  0.204 -0.230 -0.560  0.313
## Pass_Blk       0.223         0.133        -0.746 -0.450               -0.359
## Int            0.221  0.158        -0.317  0.448 -0.483 -0.611
## Clr            0.228  0.435 -0.175        -0.275                0.279  0.297
## Err                  -0.155 -0.272  0.751  0.272 -0.455  0.184
##               Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
## Tkl            0.175   0.395           0.460   0.585
## TklWin         0.234  -0.800
## Def.3rd_Tkl    0.276   0.367          -0.333  -0.424
## Mid.3rd_Tkl            0.216          -0.210  -0.267
## Att.3rd_Tkl   -0.251   0.112          -0.124  -0.158
## Tkl_Drib      -0.279          -0.589   0.291  -0.229
## Atmp_Drib     -0.168           0.105  -0.619   0.487
## Tkl_Drib.Perc -0.171           0.516
## Lost_Drib                      0.612   0.387  -0.304
## Sh_Blk         0.334
## Pass_Blk       0.158
## Int
## Clr           -0.691
## Err           -0.103
```



The *loading* are the coefficients applied to the original variables to determine the main components, and can help describe the main components considered. The first major component is strongly characterized by the number of total tackles, tackles won and tackles against dribbling. This suggests that:

- Component 1 represents an axis of evaluation of the overall defensive ability and recovery ball through effective contrasts;
- the second PC highlights the ability to counteract in the third of the offensive field a combination of positional defensive actions and defensive blocks;
- the third PC is more associated with efficiency in tackle, especially in the central part of the field;
- the fourth component describes the tendency of teams to make mistakes that lead to an opponent's shot.

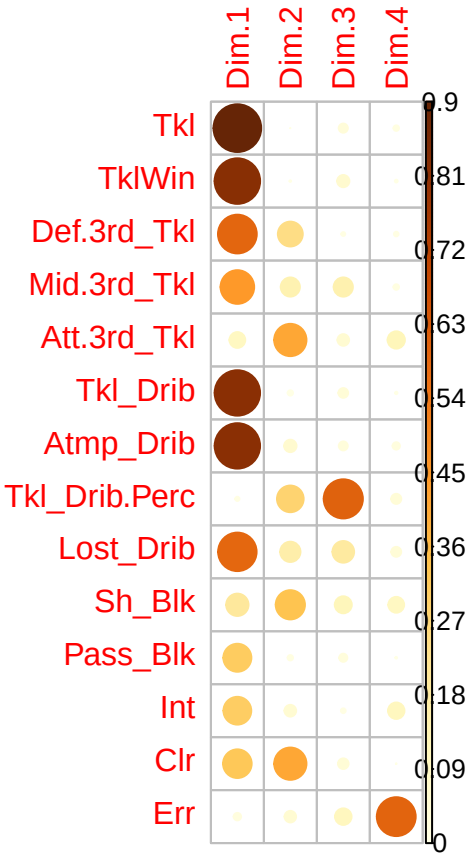
Analysis of variables

This approach is particularly useful to identify the most influential variables and to better understand the underlying structure of data. The PCA also facilitates visualization of the relationships between the original variables, highlighting the directions along which the data shows the greatest variance. In fact, PCA results can be evaluated in relation to both variables and individuals. Therefore, the results for the variables were extracted.

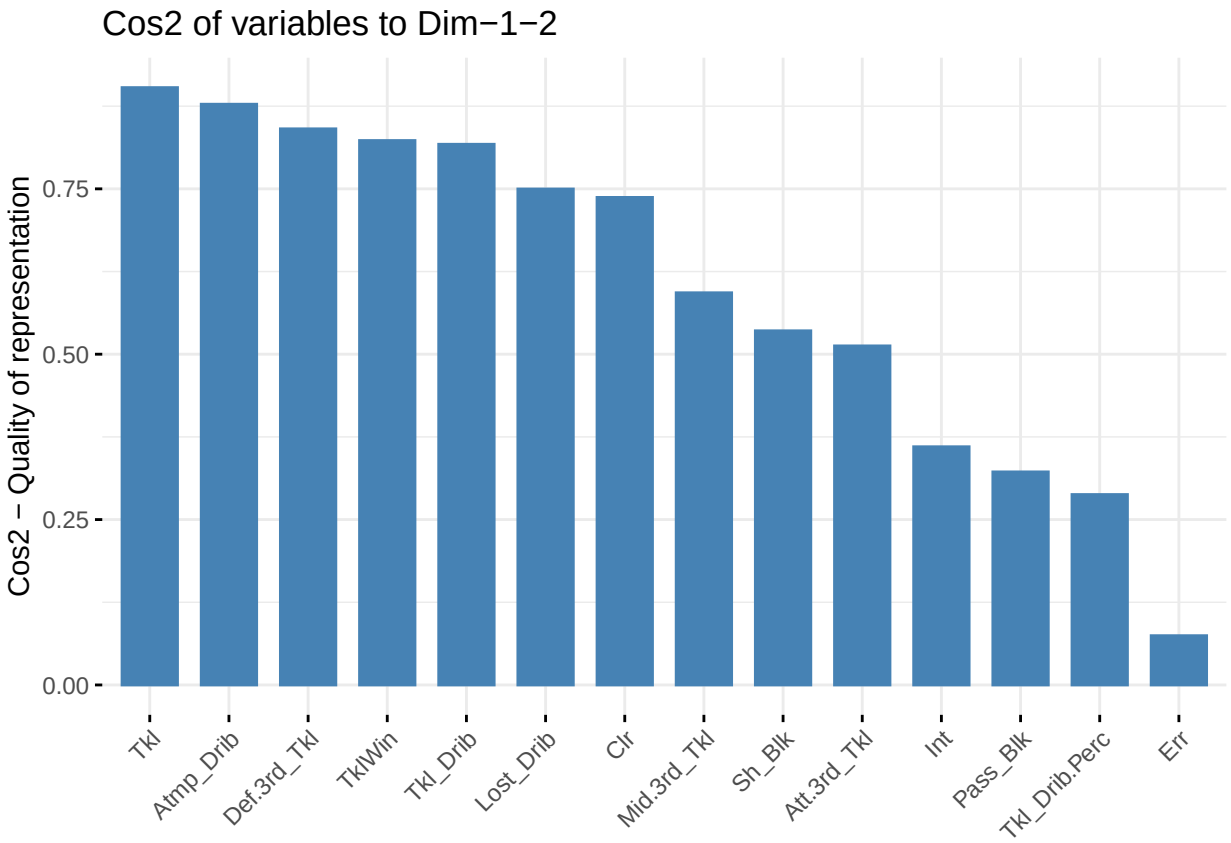
```
##                  Dim.1        Dim.2       Dim.3        Dim.4
## Tkl          0.902934180 0.0001329628 0.032797456 0.0118321147
## TklWin       0.821896820 0.0011247623 0.059105985 0.0032033437
```

```
## Def.3rd_Tkl   0.595967116 0.2447436395 0.005833245 0.0083538463
## Mid.3rd_Tkl   0.448379764 0.1444862070 0.145523188 0.0120573156
## Att.3rd_Tkl   0.099054432 0.4135273505 0.053282321 0.1179622301
## Tkl_Drib      0.817237516 0.0001328071 0.037418254 0.0023619497
## Atmp_Drib     0.817229851 0.0607161670 0.031721361 0.0212307860
## Tkl_Drib.Perc 0.006892948 0.2809523372 0.607367242 0.0399134522
## Lost_Drib     0.587399911 0.1623753494 0.185353897 0.0386479852
## Sh_Blk        0.197053513 0.3383456690 0.114030816 0.0988965740
## Pass_Blk      0.310766592 0.0113033867 0.025884589 0.0012622176
## Int           0.305949990 0.0541444562 0.008775278 0.1065772172
## Clr           0.324544688 0.4124836986 0.044516559 0.0003170607
## Err           0.021836762 0.0526332823 0.108118586 0.5996193600
```

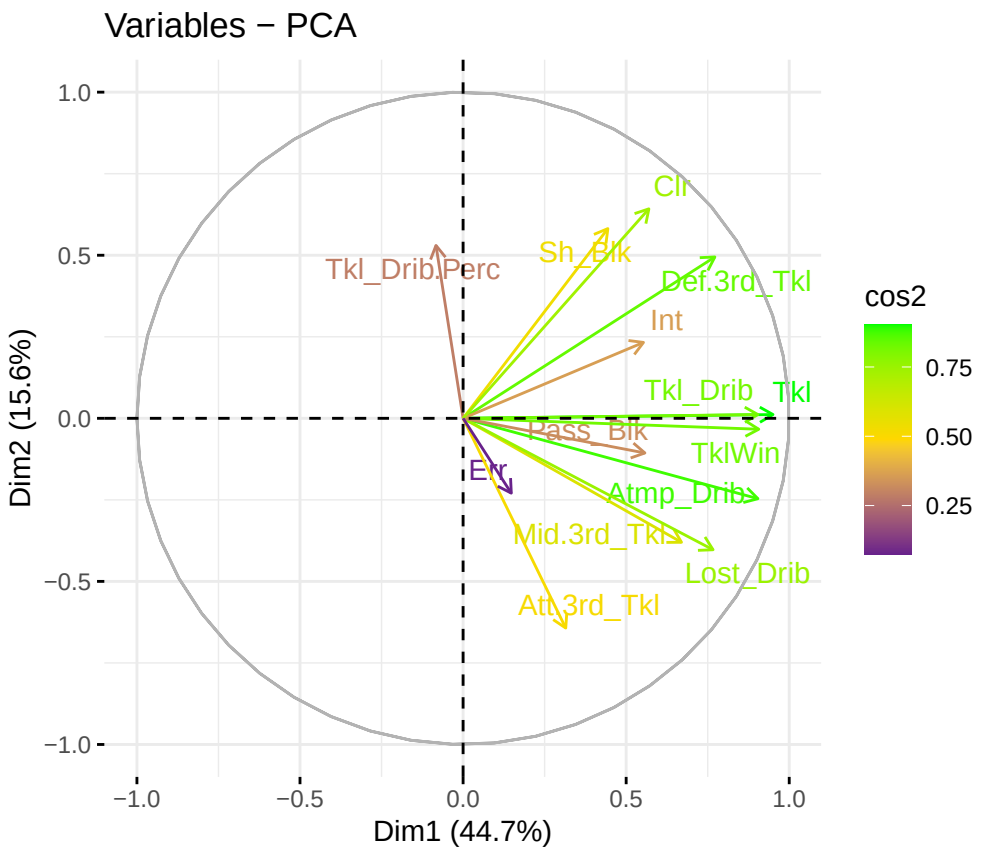
Cos^2, or the cosine squared, corresponds to the quality of variable representation. Values of cos^2 high indicate that a variable is well represented by a particular main component, while low values suggest that the variable is best represented by other components. Below are the cos^2 of the variables on the 4 dimensions.



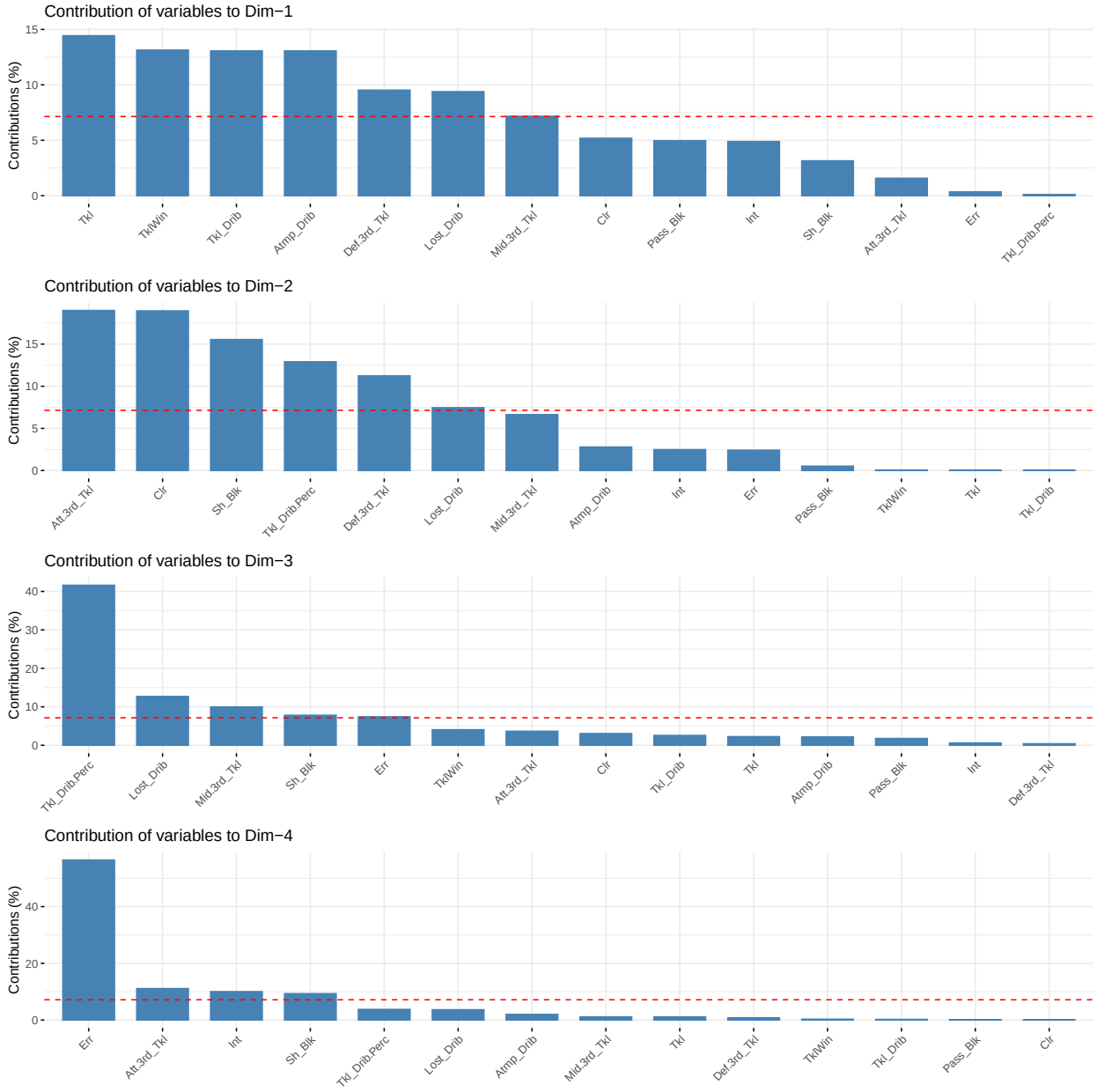
Some variables need a few main components to be represented well, others need more main components to get a good cos^2.



In addition, the quality of the representation of variables can be depicted in *correlation cirlce*, where the values of cos^2 differ in color. Positively related variables are grouped together, while negatively related variables are placed on opposite sides of the origin of the chart. The distance between variables and origin measures the quality of variables on the factor map. Variables far from origin are well represented on the factor map.



The variables Tkl, Tkl_Win, and Tkl_Drib have a very high cos^2, respectively of 0.902, 0.821, and 0.817, on the first principal component. This implies that these variables are well represented by the first principal component, and therefore are placed close to the circumference of the correlation circle. Conversely, Err, Tkl_Drib.Perc and Int are close to the origin of the axes, so they are not well represented by the main components.



The red dotted line on the bar charts indicates the expected average contribution. For a certain component, a variable with a higher contribution to this parameter is considered important in contributing to the component.

Analysis of individuals

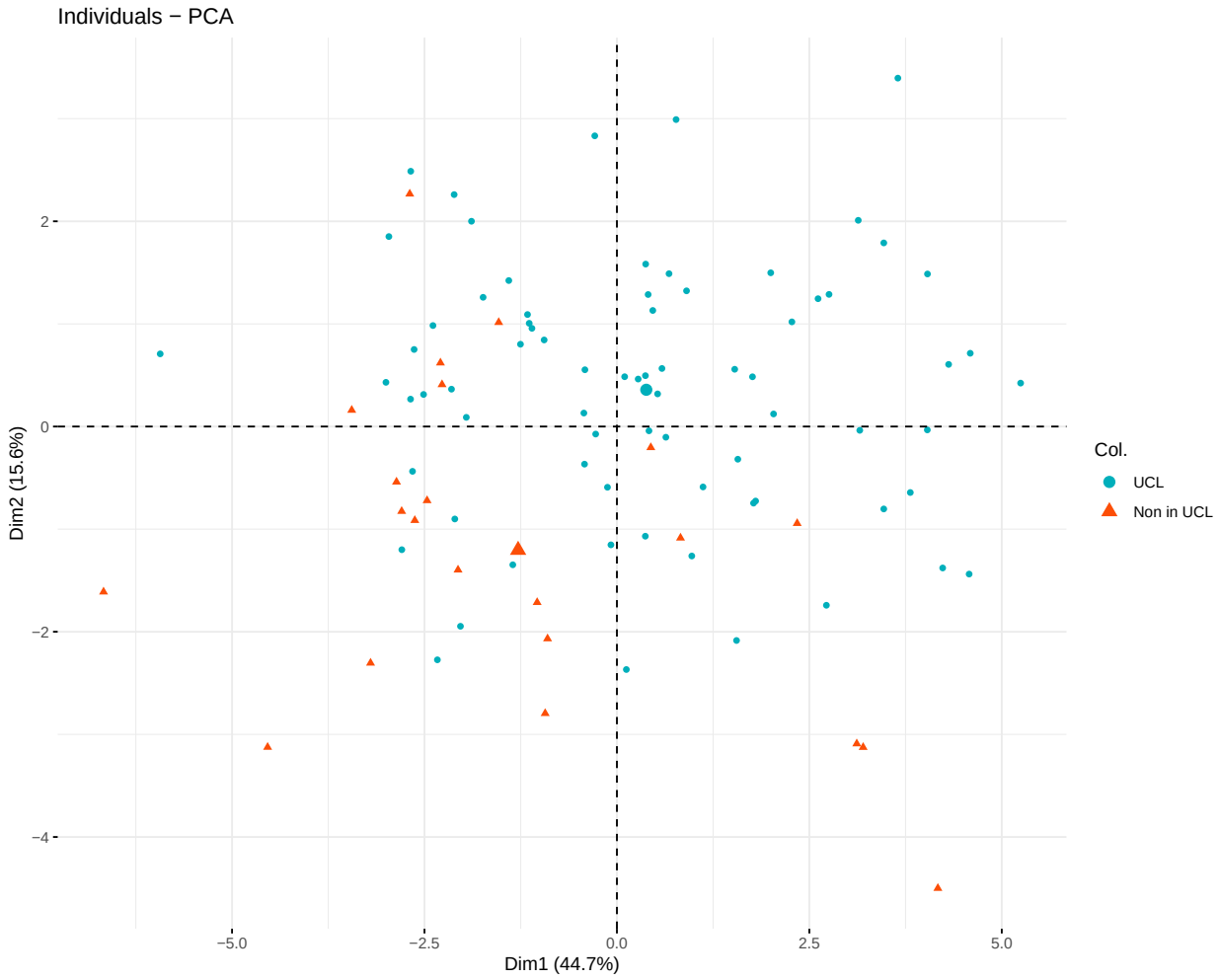
As with variables, the same operations were performed for individuals.

```
##                Dim.1       Dim.2      Dim.3       Dim.4       Dim.5
## Atalanta   0.07986440 0.1379178677 0.07814288 0.1282913321 0.157981111
## Bologna    0.03061584 0.0065927995 0.29368748 0.1453942865 0.007345354
## Cagliari   0.67195334 0.1138783161 0.02147658 0.0004837896 0.003605343
## Empoli     0.02129919 0.3822072001 0.16335057 0.0748077855 0.189248957
```

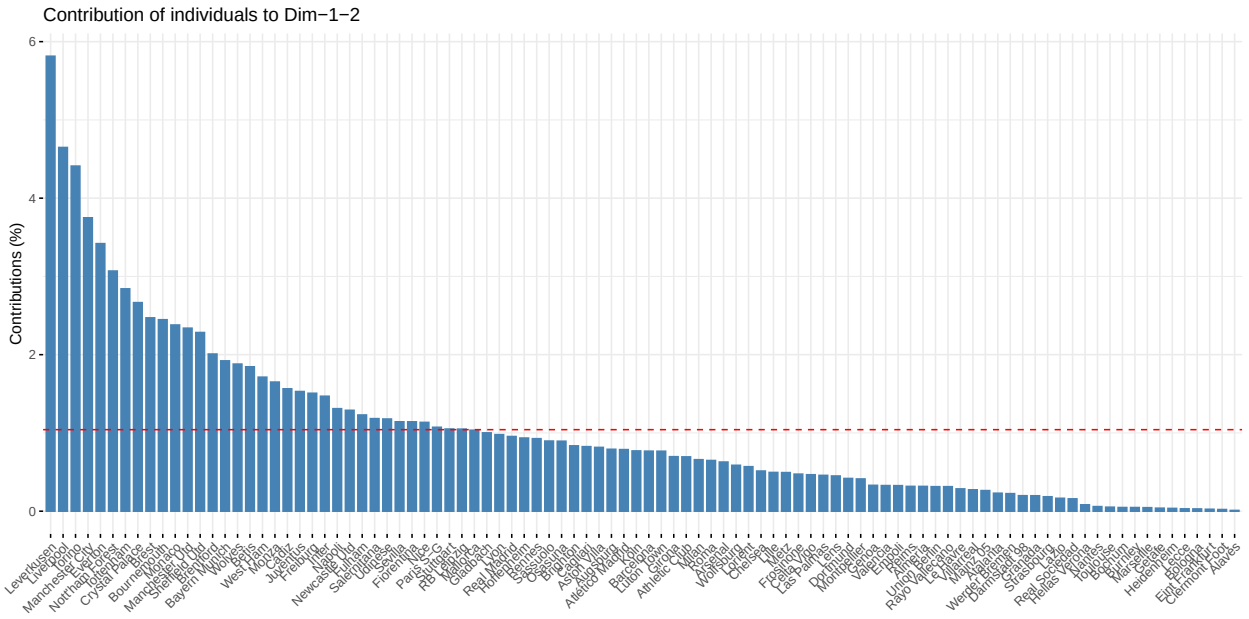
```
## Fiorentina 0.44377200 0.0818925791 0.36259571 0.0014309725 0.034585193
## Frosinone  0.45902405 0.0009575379 0.07250781 0.0922778792 0.248987150
##                  Dim.6       Dim.7        Dim.8        Dim.9      Dim.10
## Atalanta   0.066819003 0.199508013 0.0197237141 0.0643961224 0.067100594
## Bologna    0.065857980 0.437587310 0.0006964653 0.0012195649 0.010302299
## Cagliari   0.007906528 0.033543720 0.0116556166 0.0002111847 0.134025795
## Empoli     0.130979100 0.001921397 0.0151559892 0.0007697182 0.012660500
## Fiorentina 0.010114174 0.005490021 0.0035861035 0.0407915257 0.000930601
## Frosinone  0.001697229 0.007312307 0.0144111214 0.0757427159 0.026094762
##                  Dim.11       Dim.12       Dim.13       Dim.14
## Atalanta   5.313661e-05 2.018270e-04 7.608476e-36 9.196314e-32
## Bologna    3.520320e-04 3.485967e-04 8.753622e-33 9.741653e-32
## Cagliari   1.169452e-03 9.033373e-05 1.285645e-30 7.521724e-31
## Empoli     7.594984e-03 4.613913e-06 1.248750e-30 1.735954e-31
## Fiorentina 1.463874e-02 1.723742e-04 8.347492e-31 1.642188e-30
## Frosinone  8.136738e-04 1.737573e-04 2.411304e-31 4.223539e-32
```

In order to provide an even more detailed analysis, a column has been added regarding the final ranking of each team for the UEFA Champions League qualification.

```
champions_tm = c("Inter", "Milan", "Atalanta", "Bologna", "Juventus",
                 "Bayern Munich", "Dortmund", "Leverkusen", "Stuttgart", "RB Leipzig",
                 "Brest", "Lille", "Paris S-G", "Monaco",
                 "Atlético Madrid", "Barcelona", "Real Madrid", "Girona",
                 "Arsenal", "Manchester City", "Liverpool", "Aston Villa")
groups <- ifelse(rownames(numerical_data) %in% champions_tm, 1, 0)
groups = as.factor(groups)
levels(groups) = c("UCL", "Non in UCL")
fviz_pca_ind(
  data.pca, col.ind = groups, palette = c("#00AFBB",  "#FC4E07"),repel = TRUE, label = "none"
)
```



It is noted that most of the teams that have achieved a place worth worth in the Champions League, present negative values for the two PCs, this could be traced back to the hypothesis that these teams tend to have an offensive game, and therefore they prefer the possession ball, consequently they are less subjected to defensive actions.



Bayer Leverkusen, Liverpool, Turin and Manchester City have a very high cos^2 which implies a good representation of the individuals on the main components, they are in fact placed far from the center of the axes. Bologna, Eintracht Frankfurt, Clermont and Alaves have the lowest cos^2, indicating that they are not well represented by PCs, so they are close to the center.

Multiple Linear Regression

To deepen the analysis, and above all to understand the cause-effect relationships between the variables, a model of **multiple linear regression**was used. Linear regression is a statistical method used to model the relationship between a y dependent variable and one or more X. The goal is to find the best linear approximation that describes this relationship.

In the case of a single independent variable, the linear regression model (**simple**) is expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- y: dependent variable (*outcome*);
- x: independent variable (*predictor*);
- β_0: intercept (*intercept*), represents the value of y when x is zero;
- β_1: regression coefficient (*slope*), represents the expected change in y for a change unit in x;
- ϵ: error term, captures the y variation unexplained by x.

When there are multiple independent variables, the model extends to:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

dove:

- y: dependent variable;
- $x_1, x_2, ..., x_p$: independent variable;
- $\beta_0, \beta 1_2, ..., \beta_p$: regression coefficient;
- ϵ: error.

Estimation of coefficients

The β coefficients are estimated by minimizing the sum of the squares of the estimators. This technique is known as the ordinary least squares method (*Ordinary Least Squares, OLS*). The function to minimize is:

$$min_\beta \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where: $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$

Assumptions of the Linear Regression Model

The regression model is based on the following assumptions:

- **Linearity**: the relationship between the independent variables and the diependent variable is linear;
- **Independence**: ϵ errors are independent of each other;
- **Homoschedule**: Error variance is constant (not dependent on independent variables);
- **Normal errors**: ϵ errors are normally distributed

Evaluation of the model

To assess the suitability of the model and the importance of the predictors, it is necessary to take into account the **coefficient of determination** R^2, which measures the distribution of variance in the dependent variable explained by the independent variables, and the ***significance of the coefficients*** *which, assessed by statistical tests (*t-test* and *p-value*), indicates whether the coefficient is significant.

Diagnostics of the model

To ensure that the model assumptions are met, it is important to perform model diagnostics. Some useful tools include:

- ***Test t by Student***, verifies that the average error is not significantly different from zero;
- ***Shapiro Wilk's test***, concerns the normality of error distribution;
- ***Breusch-Pagan Test***, verifies homoschedaticity of residues;
- ***Durbin-Watson test***, checks for serial autocorrelation.

Estimation of the model

The objective of this analysis is, as initially said, to understand how the different parameters affect each other. In particular we will try to analyze how the variable TklWin, that is the number of tackles won, is influenced by the other variables of the dataset.

```
##
## Call:
## lm(formula = TklWin ~ Def.3rd_Tkl + Att.3rd_Tkl + Mid.3rd_Tkl +
##     Int + Tkl_Drib + Sh_Blk + Pass_Blk + Err + Lost_Drib + Clr,
##     data = numerical_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.507  -7.125  -1.625   8.501  41.363
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.787003  16.970974   0.400    0.690
## Def.3rd_Tkl  0.568919   0.061928   9.187 2.26e-14 ***
## Att.3rd_Tkl  0.609839   0.102232   5.965 5.44e-08 ***
## Mid.3rd_Tkl  0.579533   0.076358   7.590 3.76e-11 ***
## Int          0.002008   0.045787   0.044    0.965
## Tkl_Drib     0.103446   0.087766   1.179    0.242
## Sh_Blk      -0.118849   0.077627  -1.531    0.129
## Pass_Blk    -0.016626   0.050007  -0.332    0.740
## Err         -0.246547   0.304451  -0.810    0.420
## Lost_Drib   -0.008583   0.043522  -0.197    0.844
## Clr          0.001423   0.020660   0.069    0.945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.01 on 85 degrees of freedom
## Multiple R-squared:  0.8921, Adjusted R-squared:  0.8794
## F-statistic: 70.25 on 10 and 85 DF,  p-value: < 2.2e-16
```

The Atmp_Drib, Tkl_Drib and Tkl variables do not appear in this model because they are closely related to the Lost_Drib and Tkl_Drib variables, and thus achieve a better model. From the values R^2 and $Adjusted - R^2$, we can deduce how the model would seem adequate, but from the moment when R^2 is interpreted as the compromise between the goodness of adaptation and the penalty due to the excess of "useful" regressors, a reasonable procedure in model specification is to continue to include regressors until $R2$ starts to decrease. At this point the model is improved by removing all those variables having a small regression coefficient, to do this we use the algorithm **Backward selection** to estimate regressors. The algorithm starts from the model with all the p explanatory variables and then deletes one variable at a time from the one with the highest p-value. The process stops when the p-values of all the remaining variables are below a certain threshold, this threshold is fixed at the level of $\alpha = 0.05$.

```
## Start:  AIC=517.21
## TklWin ~ Def.3rd_Tkl + Att.3rd_Tkl + Mid.3rd_Tkl + Int + Tkl_Drib +
##     Sh_Blk + Pass_Blk + Err + Lost_Drib + Clr
##
##               Df Sum of Sq   RSS    AIC
## - Int          1       0.4 16694 515.21
```

```
## - Clr          1       0.9 16694 515.21
## - Lost_Drib    1       7.6 16701 515.25
## - Pass_Blk     1      21.7 16715 515.33
## - Err          1     128.8 16822 515.95
## - Tkl_Drib     1     272.8 16966 516.76
## <none>                     16693 517.21
## - Sh_Blk       1     460.3 17154 517.82
## - Att.3rd_Tkl  1    6988.4 23682 548.78
## - Mid.3rd_Tkl  1   11312.9 28006 564.88
## - Def.3rd_Tkl  1   16574.7 33268 581.41
##
## Step:  AIC=515.21
## TklWin ~ Def.3rd_Tkl + Att.3rd_Tkl + Mid.3rd_Tkl + Tkl_Drib +
##     Sh_Blk + Pass_Blk + Err + Lost_Drib + Clr
##
##               Df Sum of Sq   RSS    AIC
## - Clr          1       1.1 16695 513.22
## - Lost_Drib    1       7.4 16701 513.25
## - Pass_Blk     1      22.7 16716 513.34
## - Err          1     129.4 16823 513.95
## - Tkl_Drib     1     280.9 16975 514.81
## <none>                     16694 515.21
## - Sh_Blk       1     461.0 17155 515.82
## - Att.3rd_Tkl  1    7043.0 23737 547.00
## - Mid.3rd_Tkl  1   11420.8 28114 563.25
## - Def.3rd_Tkl  1   16655.6 33349 579.64
##
## Step:  AIC=513.22
## TklWin ~ Def.3rd_Tkl + Att.3rd_Tkl + Mid.3rd_Tkl + Tkl_Drib +
##     Sh_Blk + Pass_Blk + Err + Lost_Drib
##
##               Df Sum of Sq   RSS    AIC
## - Lost_Drib    1       7.1 16702 511.26
## - Pass_Blk     1      21.6 16716 511.34
## - Err          1     138.1 16833 512.01
## - Tkl_Drib     1     285.2 16980 512.84
## <none>                     16695 513.22
## - Sh_Blk       1     554.8 17250 514.35
## - Att.3rd_Tkl  1    7623.3 24318 547.32
## - Mid.3rd_Tkl  1   11787.6 28482 562.50
## - Def.3rd_Tkl  1   22464.2 39159 593.06
##
## Step:  AIC=511.26
## TklWin ~ Def.3rd_Tkl + Att.3rd_Tkl + Mid.3rd_Tkl + Tkl_Drib +
##     Sh_Blk + Pass_Blk + Err
##
##               Df Sum of Sq   RSS    AIC
## - Pass_Blk     1      23.4 16725 509.39
## - Err          1     150.6 16852 510.12
## - Tkl_Drib     1     291.1 16993 510.92
## <none>                     16702 511.26
## - Sh_Blk       1     556.1 17258 512.40
## - Att.3rd_Tkl  1    7698.4 24400 545.65
## - Mid.3rd_Tkl  1   11822.1 28524 560.64
```

```
## - Def.3rd_Tkl  1   22457.8 39160 591.06
##
## Step:  AIC=509.39
## TklWin ~ Def.3rd_Tkl + Att.3rd_Tkl + Mid.3rd_Tkl + Tkl_Drib +
##     Sh_Blk + Err
##
##               Df Sum of Sq   RSS    AIC
## - Err          1     145.2 16870 508.22
## - Tkl_Drib     1     287.8 17013 509.03
## <none>                     16725 509.39
## - Sh_Blk       1     582.1 17307 510.68
## - Att.3rd_Tkl  1    7741.6 24467 543.91
## - Mid.3rd_Tkl  1   12510.3 29235 561.00
## - Def.3rd_Tkl  1   22567.9 39293 589.39
##
## Step:  AIC=508.22
## TklWin ~ Def.3rd_Tkl + Att.3rd_Tkl + Mid.3rd_Tkl + Tkl_Drib +
##     Sh_Blk
##
##               Df Sum of Sq   RSS    AIC
## - Tkl_Drib     1     324.1 17194 508.05
## <none>                     16870 508.22
## - Sh_Blk       1     684.5 17555 510.04
## - Att.3rd_Tkl  1    7655.6 24526 542.14
## - Mid.3rd_Tkl  1   12412.4 29283 559.16
## - Def.3rd_Tkl  1   22447.5 39318 587.45
##
## Step:  AIC=508.05
## TklWin ~ Def.3rd_Tkl + Att.3rd_Tkl + Mid.3rd_Tkl + Sh_Blk
##
##               Df Sum of Sq   RSS    AIC
## <none>                     17194 508.05
## - Sh_Blk       1       648 17842 509.60
## - Att.3rd_Tkl  1      8730 25924 545.46
## - Mid.3rd_Tkl  1     24881 42076 591.96
## - Def.3rd_Tkl  1     45052 62247 629.55


##
## Call:
## lm(formula = TklWin ~ Def.3rd_Tkl + Att.3rd_Tkl + Mid.3rd_Tkl +
##     Sh_Blk, data = numerical_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.765  -8.713  -1.620   8.809  43.767
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.87643   14.36919   0.339   0.7351
## Def.3rd_Tkl  0.61139    0.03959  15.441  < 2e-16 ***
## Att.3rd_Tkl  0.60693    0.08929   6.797 1.09e-09 ***
## Mid.3rd_Tkl  0.62736    0.05467  11.475  < 2e-16 ***
## Sh_Blk      -0.12364    0.06677  -1.852   0.0673 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.75 on 91 degrees of freedom
## Multiple R-squared:  0.8888, Adjusted R-squared:  0.8839
## F-statistic: 181.9 on 4 and 91 DF,  p-value: < 2.2e-16
```

Therefore, our new model will be as follows:

$$TklWin \sim Def.3rdTkl + Mid.3rdTkl + Att.3rdTkl + Sh.Blk$$

With precision reporting all the specific values the model turns out to be this:

$$TklWin = 4.87643 + 0.61139 \times \text{Def.3rdTkl} + 0.62736 \times \text{Mid.3rdTkl} + 0.60693 \times \text{Att.3rdTkl} - 0.12364 \times \text{Sh.Blk}$$

Test of specificity

T test

The Student t-test is used to check whether the average error of a regression model is significantly not different from zero. This test helps to confirm that the errors are distributed around zero.

$$H_0 : E(\epsilon_i) = 0$$

$$H_0 : E(\epsilon_i) \neq 0$$

The null hypothesis H_0 is rejected if the p-value < 0.05 and its rejection would violate one of the hypotheses of the linear regression model.

```
##
##  One Sample t-test
##
## data:  model_backward$residuals
## t = 3.5038e-16, df = 95, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.725917  2.725917
## sample estimates:
##    mean of x
## 4.810966e-16
```

From the test we see that the p-value is equal to 1, so we do not reject the hypothesis that the average of the residues is not significantly different from zero.

Normal errors - Shapiro-Wilk test

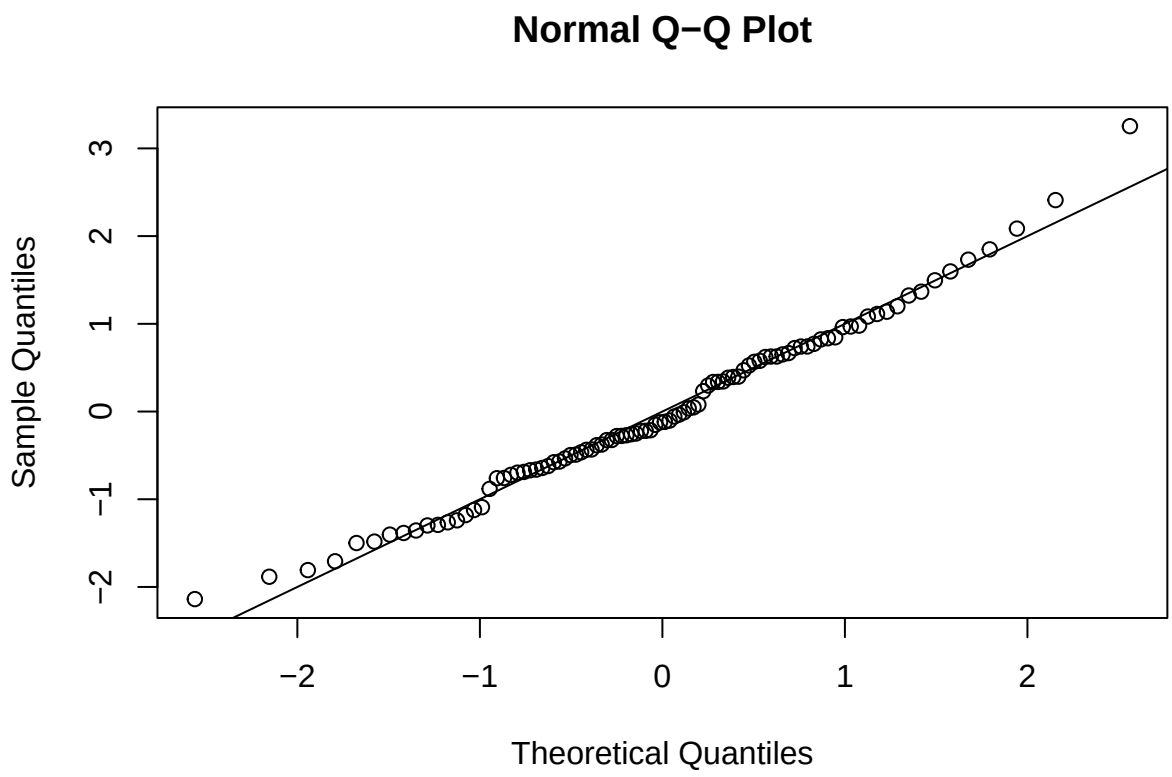
The Shapiro-Wilk test is used to assess the normality of the error distribution.

```
##
##  Shapiro-Wilk normality test
##
## data:  model_backward$residuals
## W = 0.98643, p-value = 0.4303
```

From this test we can verify that the p-value > 0.05 and therefore there is no significant evidence to reject the null hypothesis of normality. This suggests that the residues follow a normal distribution.

Q-Q Plot

The rejection of the null hypothesis is also supported by the **Normal Quantile-Quantile Plot**, as the points are distributed very close to the diagonal in the graph.



Homoscheduling - Breusch-Pagan test

The Breusch-Pagan test verifies the homoschedasticity of residues, that is if the variance of errors remains constant for all predicted values. Heteroschedasticity, or variable error variation, can affect the reliability of model estimates.

```
##
##  studentized Breusch-Pagan test
##
## data:  modello
## BP = 5.2697, df = 4, p-value = 0.2607
```

In this case the p-value > 0.05, therefore we do not reject the null hypothesis, this means that we have homoschedaticità and therefore the variance of the errors is constant.

Serial Correlation - Durbin-Watson test

The Durbin-Watson test is used to check for serial autocorrelation of residues, which occurs when subsequent errors are related to each other.

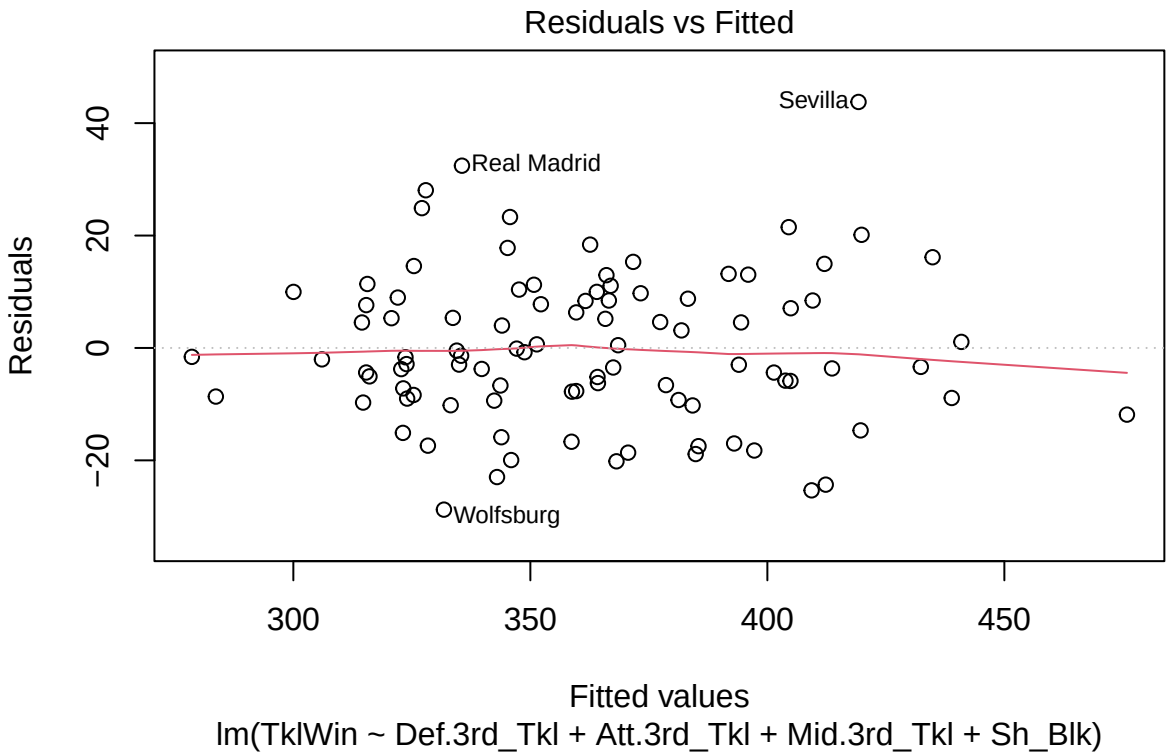
```
##
##  Durbin-Watson test
##
## data:  model_backward
## DW = 2.0046, p-value = 0.4965
## alternative hypothesis: true autocorrelation is greater than 0
```

Also in this case the p-value > 0.05 and the DW value is close to 2, therefore, there is insufficient evidence to reject the null hypothesis of absence of autocorrelation in the residues.

Analysis of residues

Linear residue distribution

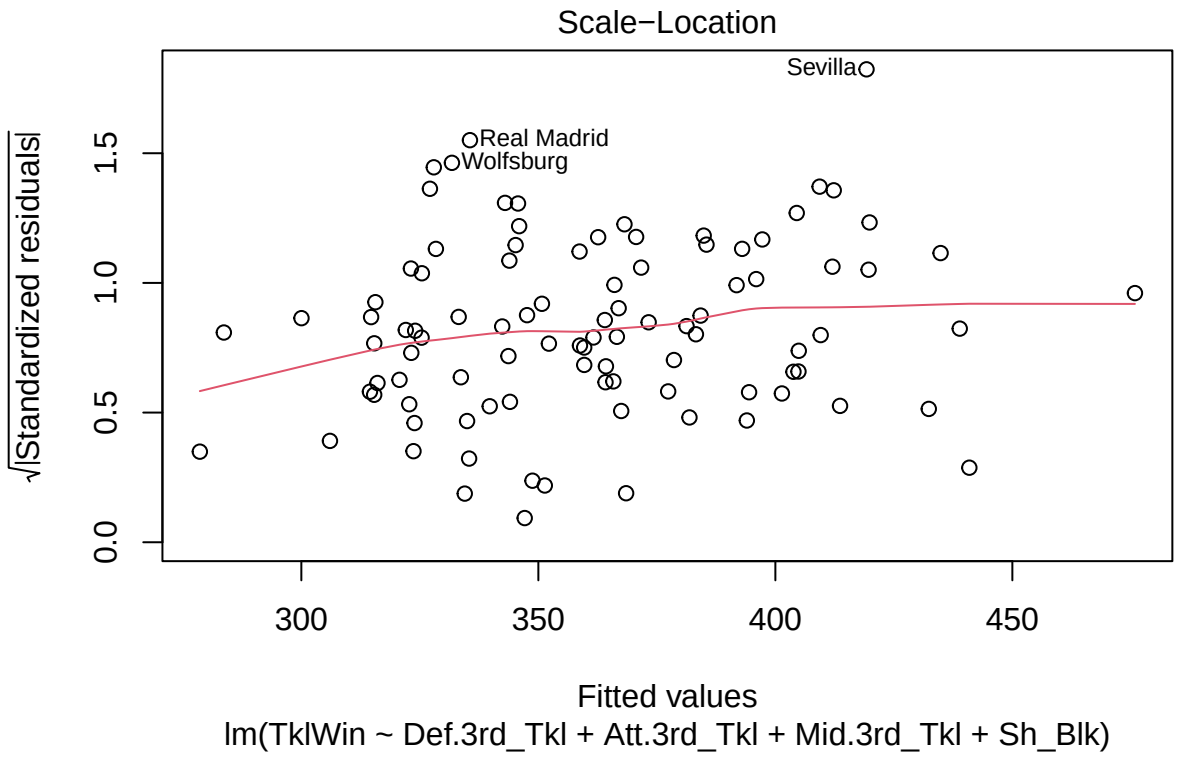
In the graph, you will see a horizontal line corresponding to the residues with an average of zero. This is because, by definition, the residues of a regression model constructed by the least squares method (OLS) always have an average of zero. The red line, instead, represents a trend line, useful for evaluating the hypothesis. If the red line overlaps significantly with the dotted line, then the linearity hypothesis is confirmed. According to the linearity hypothesis, the data must be randomly distributed around zero. If the data dispersion is not random and follows a specific pattern around zero, then there is no linearity in the residue distribution.



From the graph you can see how the hypothesis of linearity can be considered verified as the points are randomly distributed around the zero without bringing out any pattern.

Homoscheduling of residues

To determine if there is heteroschedasticity by residue analysis, it is necessary to create a graph showing the residues in absolute value on the axis of the ordinates and the values estimated by the model on the axis of the ascisse: Vertical dispersion is expected to be approximately constant. In fact, in the diagram below, it is observed that most of the residues, even if randomly arranged, are around the values of 0.5 and 1.

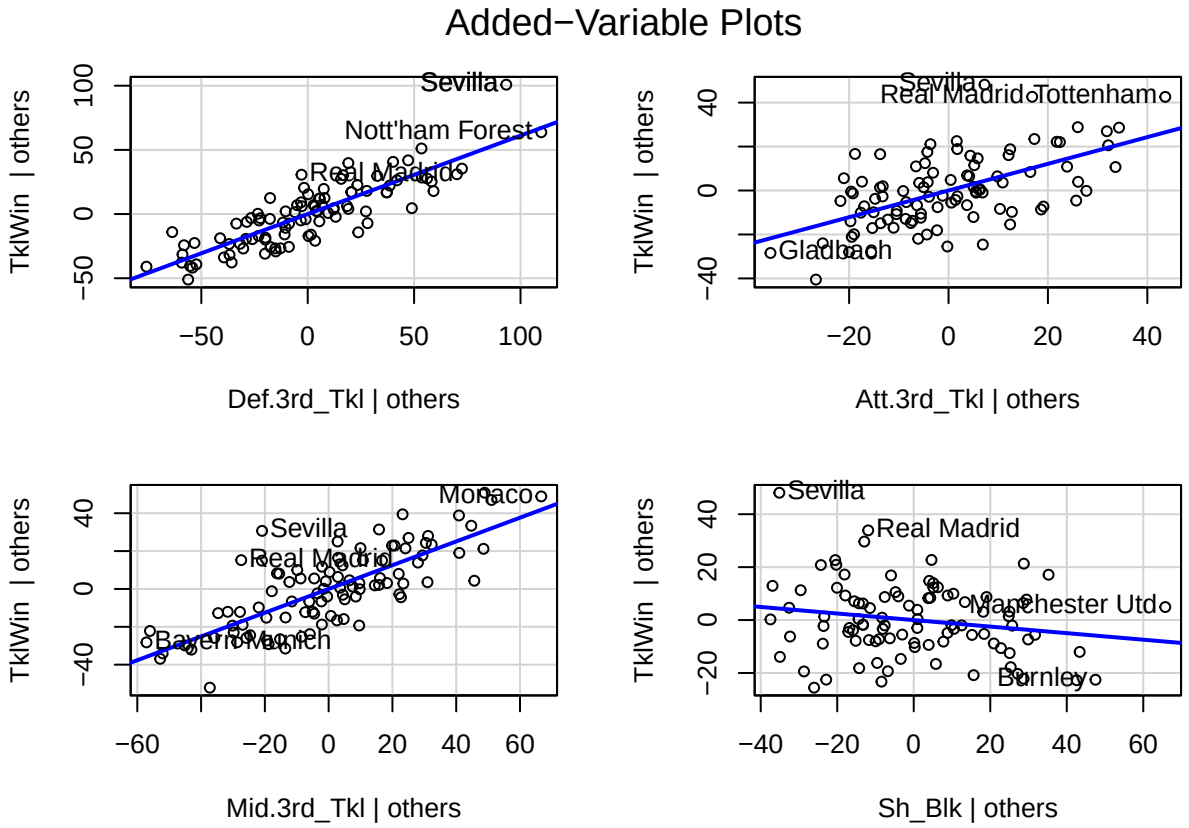


Outlier

The graph used to verify linearity in error distribution can also be used to identify outliers in the model. Outliers are simply those points that are identified by numbers and appear more isolated than the other points in the residue chart.

For more diagnostic tools, we can consider the use of **leverage scores** (leverage points). These points, when the explanatory variables are more than one, can be influenced by the other regressors. We can use the**Added Variable Plot** to highlight these relationships. This graph represents the relationship between a specific independent variable and the model residues, controlling for the effects of the other independent variables in the model. In practice, this allows you to assess how a specific independent variable affects the model's residues, providing further guidance on its relevance and the possible presence of outliers.

```
avPlots(model_backward,terms=~.)
```



Each graph displays a regression line whose inclination represents the estimate of the coefficient of the specific independent variable, also highlighting how it varies when included in the model. If a variable has a minimal influence on the leverage points of each observation, it will appear close to the horizontal line $Y = 0$.

Conclusions

The report in question concerns a statistical analysis of the data relating to the defensive actions of the teams of the top 5 European championships related to the 2023/2024 season. In particular, the data set was subjected to a descriptive analysis to describe the distribution of each variable and consequently the playing styles of the various teams, for example by highlighting how Tottenham is the team with the most tackles in the third offensive or how Juventus is the team with the highest percentage of successful dribbling; then we went through an exploratory analysis to decrease the dimensionality of the dataset and find out which teams have a similar style of play; and finally a multiple regression model has been hypothesized in order to understand possible relationships between the number of tackles won and the other variables, evidencing above all like the contrasts in the different zones of field, positively influence the increase of winning contrasts.