Projet en atelier statistique avec R

Analyse de l Impact des Performances des Joueurs sur les Résultats d equipe

Yassine Hammi

2024-11-26

Table des matières

1 Introduction

Dans le domaine sportif, la performance individuelle des joueurs peut influencer directement ou indirectement les résultats globaux de leur équipe. Cependant, comprendre précisément cette relation reste un défi.

Quels sont les facteurs clés individuelle qui influencent les performances collectives ? Quels types de joueurs ont un rôle déterminant sur les victoires ou les défaites? Cette analyse est essentielle pour optimiser la gestion des équipes et améliorer leurs performances.

Pour commencer, nous allons aborder la question d'exploration fondamentale, qui sera le point de départ de notre analyse à travers les différentes étapes. Nous permettant de spécifier les variables clés à observer, de collecter des données exhaustives, de les prétraiter avec rigueur, d'effectuer une analyse approfondie et enfin de présenter des résultats interprétables.

1.1 Question principale d'exploration

Comment les performances individuelles des joueurs influencent-elles les résultats et la performance globale de leur équipe ?

En examinant de près les principales raisons, nous espérons obtenir une meilleure compréhension sur les perforamnces pour chaque joueur individuelle et comme conséquences la performance d'equipe et comment cela influence l'echec ou la reussite de l'equipe.

1.2 Spécification des Variables

Pour mener à bien notre analyse, nous nous appuierons sur differents sources de données cruciales telque "Performance des joueurs individuels" et leur "Resultats des matchs", "Performance collective"

2 Collecte de Données

Collection de données s'agit d'une étape cruciale pour l'ensemble du projet, tant que nous disposons de données bien structurés, nous pouvons effectuer une meilleure analyse plus approfondie.

il est important d'identifier notre sources de données

2.1 Identification

- Kaggle: as the world's largest data science community with powerful tools and resources to help us achieve our data science goals and objectifs.
- Github : nous pouvons trouver plusieurs projets open source concernant le football, avec des données mises à jour sur les joueurs et les équipes.

2.2 Importation

Au début de notre analyse, nous commençons par intégrer les données essentielles à partir des sources distincts.

```r
# Installation de package si n'existe pas
if (!require(readr) ) {install.packages("readr" , repos =
 ↪  "http://cran.us.r-project.org")}
if (!require(dplyr) ) {install.packages("dplyr" , repos =
 ↪  "http://cran.us.r-project.org")}
if (!require(tidyr) ) {install.packages("tidyr" , repos =
 ↪  "http://cran.us.r-project.org")}
if (!require(ggplot2) ) {install.packages("ggplot2" , repos =
 ↪  "http://cran.us.r-project.org")}
if (!require(ggcorrplot) ) {install.packages("ggcorrplot" , repos =
 ↪  "http://cran.us.r-project.org")}
if (!require(fmsb) ) {install.packages("fmsb" , repos = "http://cran.us.r-project.org")}
if (!require(ggradar) ) {install.packages("ggradar" , repos =
 ↪  "http://cran.us.r-project.org")}
if (!require(reshape2) ) {install.packages("reshape2" , repos =
 ↪  "http://cran.us.r-project.org")}
if (!require(coefplot) ) {install.packages("coefplot" , repos =
 ↪  "http://cran.us.r-project.org")}

# Chargement de package
library(readr)
library(knitr)
library(ggplot2)
library(reshape2)
library(corrplot)
library(coefplot)


# Importation de donneés depuis fichier excel
# Cette fichier est importé depuis kaggle
# Comporte les donneés de championnat d'Espagne de football de première division

LaLigaPlayers <-
 ↪  read_csv("C:/Users/21655/Desktop/Projet_DS_Yassine/Data/S2324-laliga-players.csv")
# Dimension de notre dataset
dim(LaLigaPlayers) # 3615 lignes , 150 columns
# Affichage de dooneés
kable(head(LaLigaPlayers[,7:13],6),caption = "LaLiga - sample 6 lignes avec 7 columns")
```

```
## [1] 3615  150
```

Table 1: LaLiga - sample 6 lignes avec 7 columns

firstname	lastname	gender	date_of_birth	place_of_birth	weight	height
Aarón	Escandell	male	1995-09-27	Carcagente	71	185
Abde	Ezzalzouli	male	2001-12-17	Beni Melal	72	177
Abde	Raihani	male	2004-02-03	Barcelona	NA	187
Abde	Rebbach	male	1998-08-11	Bilda	NA	176
Abdel	Abqar	male	1999-03-10	Settat	80	188

firstname	lastname	gender	date_of_birth	place_of_birth	weight	height
Abdul	Mumin	male	1998-06-06	Accra	79	188

```
# Chargement de package pour extraire des données depuis l'internet
library(rvest)

# Le lien d'où allons extraire les informations d'une equipe
# Dans ce cas nous avons choisis l'equipe "Real Madrid"
link <-"https://fbref.com/fr/equipes/53a2f082/2023-2024/Statistiques-Real-Madrid"

page <- read_html(link,header=FALSE)

# Nous avons extrait les tables dans cette page
tables <- page %>%
  html_nodes("table") %>%
  html_table(fill = TRUE)

# Mettre le deuxième tableau dans notre dataset MatchesRM
if(length(tables)>0){
  MatchesRM <- tables[[2]]
  dim(MatchesRM)  # 55 lignes , 20 columns
  kable(head(MatchesRM[,4:13],6),caption="Matches-RM")
} else {
  print("aucun table trouveé")
}
```

Table 2: Matches-RM

Tour	Jour	Tribune	Résultat	BM	BE	Adversaire	xG	xGA	Poss
Journée 1	Sam	Extérieur	V	2	0	Athletic Club	0.9	0.4	54
Journée 2	Sam	Extérieur	V	3	1	Almería	2.0	1.3	57
Journée 3	Ven	Extérieur	V	1	0	Celta Vigo	1.4	1.2	63
Journée 4	Sam	Domicile	V	2	1	Getafe	2.8	0.4	76
Journée 5	Dim	Domicile	V	2	1	Real Sociedad	2.0	1.6	52
Phase de groupe	Mer	Domicile	V	1	0	de Union Berlin	3.7	0.2	75

3 PreProcessing

3.1 Suppression des colonnes dans la première dataset : RealMadridPlayers

```
# Suppression des colonnes inutiles dans notre jeu de données
# Nous avons déjà supprimér plusieurs colonnes mais nous venons de laisser celles-ci
 ↪  comme exemple de code

LaLigaPlayers <- subset(LaLigaPlayers,select=-c(height, weight, competition ,player.url
 ↪  ,id ,date_of_birth , country ,place_of_birth , slug ,nickname, firstname, lastname,
 ↪  gender, international, throw_ins_to_opposition_player ,throw_ins_to_own_player
 ↪  ,twitter ,instagram ,team.shortname ,team.foundation ,team.shield ,photo ,stadium
 ↪  ,stadium.image ,drops ,games_played ,goalkeeper_smother ,hit_woodwork ,index
 ↪  ,last_player_tackle ,left_foot_goals ,leftside_passes ,other_goals ,punches
 ↪  ,right_foot_goals ,rightside_passes ,shots_off_target_inc_woodwork ,team_games_played
 ↪  ,handballs_conceded ) )
```

```r
# Conserver seulement les joueurs de Real Madrid à partir de dataset LaLigaPlayers

library(kableExtra)
RealMadridPlayers <- subset(LaLigaPlayers, LaLigaPlayers$team=="Real Madrid")

RealMadridPlayers <- subset(RealMadridPlayers,select = -c(team))

kable(
    head( RealMadridPlayers[,1:7],6 ),
    caption = "Tableau des Données 1er 6 lignes avec 7 colonnes",
    format = "latex",
    booktabs = TRUE,  # Ajouter des traits horizontaux propres
    align = "c"    # Centrer les colonnes
  ) %>%
    kable_styling(
      latex_options = c("striped", "hold_position"),  # Style rayé et position maintenue
      full_width = TRUE,  # Table non étendue à la largeur complète
      font_size = 7
    )

#kable(head (RealMadridPlayers[,1:9],6),caption = " Joueurs RM")



print( paste("Dimension : ", dim(RealMadridPlayers)) ) # 175 lignes , 112 columns
print( paste("Nombre de lignes : ",nrow(RealMadridPlayers)))  # nous avons 175 lignes
```

Table 3: Tableau des Données 1er 6 lignes avec 7 colonnes

name	shirt_number	position	aerial_duels	aerial_duels_lost	aerial_duels_won	appearances
Andrii Lunin	13	Goalkeeper	10	NA	10	21
Antonio Rüdiger	22	Defender	72	23	49	33
Arda Güler	24	Midfielder	NA	NA	NA	10
Aurélien Tchouaméni	18	Midfielder	75	24	51	27
Brahim Díaz	21	Midfielder	12	9	3	31
Dani Carvajal	2	Defender	41	19	22	28

```
## [1] "Dimension :  175" "Dimension :  110"
## [1] "Nombre de lignes :  175"
```

3.2 Suppression les lignes redundantes dans la première dataset : RealMadridPlayers

```r
# Nombres duplicates
num_duplicates <- sum(duplicated(RealMadridPlayers))

# Identifer les lignes dupliquées.
```

```r
duplicated_RM_Players <- RealMadridPlayers[duplicated(RealMadridPlayers),]
# les enregistrements dupliquées = 140
head(duplicated_RM_Players, n =15)
```

```
## # A tibble: 15 x 110
##    name    shirt_number position aerial_duels aerial_duels_lost aerial_duels_won
##    <chr>          <dbl> <chr>           <dbl>              <dbl>            <dbl>
##  1 Andrii~           13 Goalkee~           10                 NA               10
##  2 Antoni~           22 Defender           72                 23               49
##  3 Arda G~           24 Midfiel~           NA                 NA               NA
##  4 Auréli~           18 Midfiel~           75                 24               51
##  5 Brahim~           21 Midfiel~           12                  9                3
##  6 Dani C~            2 Defender           41                 19               22
##  7 Dani C~           19 Midfiel~            7                  5                2
##  8 David ~            4 Defender           19                  9               10
##  9 Diego ~           26 Goalkee~           NA                 NA               NA
## 10 Edgar ~           37 Defender           NA                 NA               NA
## 11 Eduard~           12 Midfiel~           36                 18               18
## 12 Federi~           15 Midfiel~           31                 10               21
## 13 Ferlan~           23 Defender           11                  5                6
## 14 Fran G~           20 Defender           23                 15                8
## 15 Gonzal~           33 Forward            NA                 NA               NA
## # i 104 more variables: appearances <dbl>, assists_intentional <dbl>,
## #   attempts_from_set_pieces <dbl>, away_goals <dbl>, backward_passes <dbl>,
## #   blocked_shots <dbl>, blocks <dbl>, catches <dbl>, clean_sheets <dbl>,
## #   clearances_off_the_line <dbl>, corners_taken_incl_short_corners <dbl>,
## #   corners_won <dbl>, crosses_not_claimed <dbl>, duels <dbl>,
## #   duels_lost <dbl>, duels_won <dbl>, forward_passes <dbl>,
## #   foul_attempted_tackle <dbl>, foul_won_penalty <dbl>, ...
```

```r
# Les lignes nettoyées sur lesquelles nous allons travailler
normaldt_RM_Players <- RealMadridPlayers[!duplicated(RealMadridPlayers),]


kable(
    head( normaldt_RM_Players[,1:7],6 ),
    caption = "Tableau des Données 1er 6 lignes avec 7 colonnes",
    format = "latex",
    booktabs = TRUE,  # Ajouter des traits horizontaux propres
    align = "c"    # Centrer les colonnes
  ) %>%
    kable_styling(
      latex_options = c("striped", "hold_position"),  # Style rayé et position maintenue
      full_width = TRUE,  # Table non étendue à la largeur complète
      font_size = 7
    )
```

Table 4: Tableau des Données 1er 6 lignes avec 7 colonnes

name	shirt_number	position	aerial_duels	aerial_duels_lost	aerial_duels_won	appearances
Andrii Lunin	13	Goalkeeper	10	NA	10	21
Antonio Rüdiger	22	Defender	72	23	49	33
Arda Güler	24	Midfielder	NA	NA	NA	10
Aurélien Tchouaméni	18	Midfielder	75	24	51	27
Brahim Díaz	21	Midfielder	12	9	3	31
Dani Carvajal	2	Defender	41	19	22	28

3.3 Regroupement les joueures par leur positions de jeu

“Goalkeeper” , “Defender” , “Forward” , ” Midfielder”

```
GoalKeepers <- subset(normaldt_RM_Players,normaldt_RM_Players$position=="Goalkeeper")

# Les noms de gardiens
# Load kableExtra
library(kableExtra)


kable(head(GoalKeepers[,1:5]), align = "l", caption = "Goalkeepers") %>%
  kable_styling(position = "left", full_width = FALSE)
```

Table 5: Goalkeepers

name	shirt_number	position	aerial_duels	aerial_duels_lost
Andrii Lunin	13	Goalkeeper	10	NA
Diego Piñeiro	26	Goalkeeper	NA	NA
Kepa Arrizabalaga Revuelta	25	Goalkeeper	2	NA
Lucas Cañizares	31	Goalkeeper	NA	NA
Mario de Luis	39	Goalkeeper	NA	NA
Thibaut Courtois	1	Goalkeeper	1	NA

```
#
Defenders <- subset(normaldt_RM_Players,normaldt_RM_Players$position=="Defender")

# Les noms de defenders

kable(head(Defenders[,1:5]), align = "l", caption = "Defenders Table") %>%
  kable_styling(position = "left", full_width = FALSE)
```

Table 6: Defenders Table

name	shirt_number	position	aerial_duels	aerial_duels_lost
Antonio Rüdiger	22	Defender	72	23
Dani Carvajal	2	Defender	41	19

David Alaba	4	Defender	19	9
Edgar Pujol	37	Defender	NA	NA
Ferland Mendy	23	Defender	11	5
Fran García	20	Defender	23	15

Nous referons cette opération pour les milieux de terrain “MidFielders” et les défenseurs “Defenders”

```
#
Midfielders <- subset(normaldt_RM_Players,normaldt_RM_Players$position=="Midfielder")

# Les noms de Midfielders
#kable(head(Midfielders[,1:5]), align = "l", caption = "Midfielders Table") %>%
# kable_styling(position = "left", full_width = FALSE)
```

```
#
Forwards <- subset(normaldt_RM_Players,normaldt_RM_Players$position=="Forward")

# Les noms de Midfielders
#kable(head(Forwards[,1:5]), align = "l", caption = "Forwards Table") %>%
#kable_styling(position = "left", full_width = FALSE)
```

3.4 Supprimer les colonnes non nécessaires pour chaque groupe

Les critères principaux, les facteurs et les metriques pour chaque position sont différents à des autres, et ils sonts similaires dans certaines.

Par exemple : pour le gardien de but, ce sont les matchs sans encaisser de buts, les tirs bloqués, les buts encaissés, les passes en avant, les contributions réussies ou infructueuses du gardien. Pour les attaquants, ce sont les buts, les passes décisives, les duels remportés, les duels perdus, et ainsi de suite pour les milieux de terrain et les défenseurs.

Pour les gardiens nous pouvons noter :

clean_sheets, blocked_shots, saves_made, saves_from_penalty, saves_made_caught, , saves_made_from_inside_box, saves_made_from_outside_box, saves_made_parried, goal_kicks, gk_successful_distribution, gk_unsuccessful_distributio penalties_faced, penalties_saved,, penalty_goals_conceded, penalties_conceded, goal_assists, goal_kicks, putthrough_blocked_distribution, putthrough_blocked_distribution_won

alors nous supprimer les autres colones non nécessaires,mais pour vérifier si les colonnes correspondent à la position ou non, nous vérifions si toutes les lignes sont NA. Si c’est le cas, nous supprimons cette colonne

```
print("Dimension Avant suppression")
dim(GoalKeepers)

remove_na_columns <- function(GoalKeepers) {
  # check kol columns is na lkol ou non
  GoalKeepers <- GoalKeepers[, colSums(is.na(GoalKeepers)) < nrow(GoalKeepers)]
  return(GoalKeepers)
}
```

```r
GoalKeepers<- remove_na_columns(GoalKeepers)

print("Dimension Aprés suppression")
dim(GoalKeepers)
```

```
## [1] "Dimension Avant suppression"
## [1]   6 110
## [1] "Dimension Aprés suppression"
## [1]  6 59
```

Nous mettrons l'accent pour les gardiens sera principalement mis surles métriques essentielles concernent leur capacité à défendre le cage, les actions défensives, et les clean sheets alors on supprimer les colonnes de données qui ne sont pas utiles pour notre analyse :

```
## [1] "Dimension Avant suppression"
## [1]  6 59
## [1] "Dimension Aprés suppression"
## [1] 6 9
```

Répéter le processus pour toutes les positions :

- pour defenseurs:

```
## [1] "Dimension Avant suppression"
## [1]  11 110
## [1] "Dimension Aprés suppression"
## [1] 11 93
```

Supprimer les colonnes de données qui ne sont pas utiles pour notre analyse

```
## [1] "Dimension Avant suppression"
## [1] 11 93
## [1] "Dimension Aprés suppression"
## [1] 11  9
```

- pour milieux de terrains : il est essentiel de se concentrer sur les passes, les récupérations, les dribbles et leur capacité à influencer le jeu tant offensivement que défensivement, en facilitant la transition entre les deux phases.

```
## [1] "Dimension Avant suppression"
## [1]  12 110
## [1] "Dimension Aprés suppression"
## [1] 12 94
```

Supprimer les colonnes de données qui ne sont pas utiles pour notre analyse

```
## [1] "Dimension Avant suppression"
## [1] 12 94
## [1] "Dimension Aprés suppression"
## [1] 12 10
```

- pour les attaquants:

```
## [1] "Dimension Avant suppression"
## [1]   6 110
## [1] "Dimension Aprés suppression"
## [1]  6 90
```

Supprimer les colonnes de données qui ne sont pas utiles pour notre analyse .

pour les attaquants, il est crucial de maintenir des mesures appropriées pour leur performance offensive et leur capacité à marquer des buts:

```
## [1] "Dimension Avant suppression"
## [1]  6 90
## [1] "Dimension Aprés suppression"
## [1]  6 14
```

3.4.1 Remplacement NA valeurs

- Nous remplacerons toutes les colonnes pour toutes les DATASET contenant des valeurs NA par 0, car nous savons que NA signifie 'Not Assigned' et dans notre cas, cela équivalent à 0.

3.5 Nettoyage du dataset des matches RM

Nous allons nettoyer dataset qui contient les données sur les matches du REAL MADRID tout au long de la saison, mais sur lesquels nous nous concentrons uniquement sur les matches de compétition LaLiga, et suppression des colonnes inutiles.

```
MatchesRM <- subset(MatchesRM, MatchesRM$Comp=="La Liga")


delete_columns<- function(MatchesRM, columnsToDelete) {
  # si existe column -> put it in exisiting columns
  existing_columns <- columnsToDelete[columnsToDelete %in% colnames(MatchesRM)] #
 ↪  explicitement

   # Suppression les colonnes souhaité
  MatchesRM <- MatchesRM[, !(colnames(MatchesRM) %in% existing_columns)]

  return(MatchesRM)
}

columnsToDelete <- c("Date","Heure","Jour","Arbitre","Rapport de
 ↪  match","Notes","Capitaine","Formation","Formation Adverse")

MatchesRM <- delete_columns_if_exists(MatchesRM, columnsToDelete)
```

Dans cette partie, nous créons la variable cible ResultVar ("victoire", "match nul" ou "défaite") : 1 pour une victoire, 0 pour une défaite, et 2 pour un match nul, nous créons également notre deuxième variable cible,

qui correspond à la différence entre les buts marqués et les buts encaissés (BM : buts marqués - BE buts encaissés).

Mais avant cela, nous vérifierons que ces variables sont de type caractère ou non, puis nous les convertirons en Int.

```
# Verifier si les variables sont integer ou pas

typeof(MatchesRM$BE)
typeof(MatchesRM$BM)

# Converting BE BM to Int
MatchesRM$BE <- as.integer(MatchesRM$BE)
typeof(MatchesRM$BE)
MatchesRM$BM <- as.integer(MatchesRM$BM)
typeof(MatchesRM$BM)

# Creation les variabels cibles
# Assuming MatchesRM$Resultat contains "V", "D", or "N"
MatchesRM$Target1 <- ifelse(MatchesRM$Résultat == "V", 1,
                        ifelse(MatchesRM$Résultat == "D", 0,
                        ifelse(MatchesRM$Résultat == "N", 2,NA) ))

kable(head(MatchesRM[,4:13],6),caption = "Matches Real Madrid")
```

```
## [1] "character"
## [1] "character"
## [1] "integer"
## [1] "integer"
```

Table 7: Matches Real Madrid

Résultat	BM	BE	Adversaire	xG	xGA	Poss	Affluence	Formation adverse	Target1
V	2	0	Athletic Club	0.9	0.4	54	48,927	4-2-3-1	1
V	3	1	Almería	2.0	1.3	57	17,561	4-2-3-1	1
V	1	0	Celta Vigo	1.4	1.2	63	23,057	5-3-2	1
V	2	1	Getafe	2.8	0.4	76	66,747	4-1-3-2	1
V	2	1	Real Sociedad	2.0	1.6	52	70,092	4-3-3	1
D	1	3	Atlético Madrid	1.0	1.4	63	69,082	5-3-2	0

4 Analyse de Données

4.1 Calcul Performance index pour chaque groupe

4.2 GoalKeepers performance index

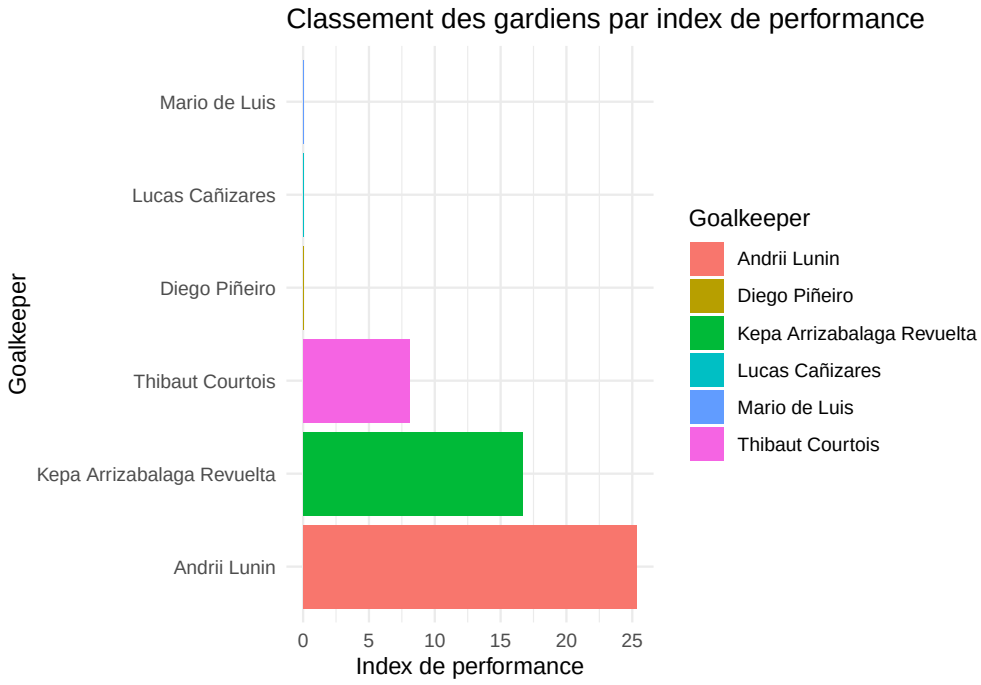
dans cette etape on calcul performance index pur les gardiens.

Le classement des gardiens selon leur index de performance montre clairement qu'Andrii Lunin se distingue en première, suivi par Kepa, plusieurs facteurs ont un impact sur ces résultats, tels que les apparitions, les matchs joués et les minutes passées.

La contribution de ces éléments à la performance globale de chaque joueur est significativement influencée par leur poids dans la formule de calcul de l'index.

```
GoalKeepers$performance_index_gk <- GoalKeepers$appearances * 0.1 +
 ↪  GoalKeepers$clean_sheets * 0.4 + GoalKeepers$gk_successful_distribution * 0.1 +
 ↪  GoalKeepers$saves_made * 0.15 - GoalKeepers$gk_unsuccessful_distribution * 0.1  -
 ↪  GoalKeepers$goals_conceded * 0.35 - GoalKeepers$penalty_goals_conceded *0.1

ggplot(GoalKeepers, aes(x = reorder(name, -performance_index_gk), y =
 ↪  performance_index_gk, fill = name)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(
    title = "Classement des gardiens par index de performance",
    x = "Goalkeeper",
    y = "Index de performance",
    fill = "Goalkeeper"
  ) +
  theme_minimal()
```

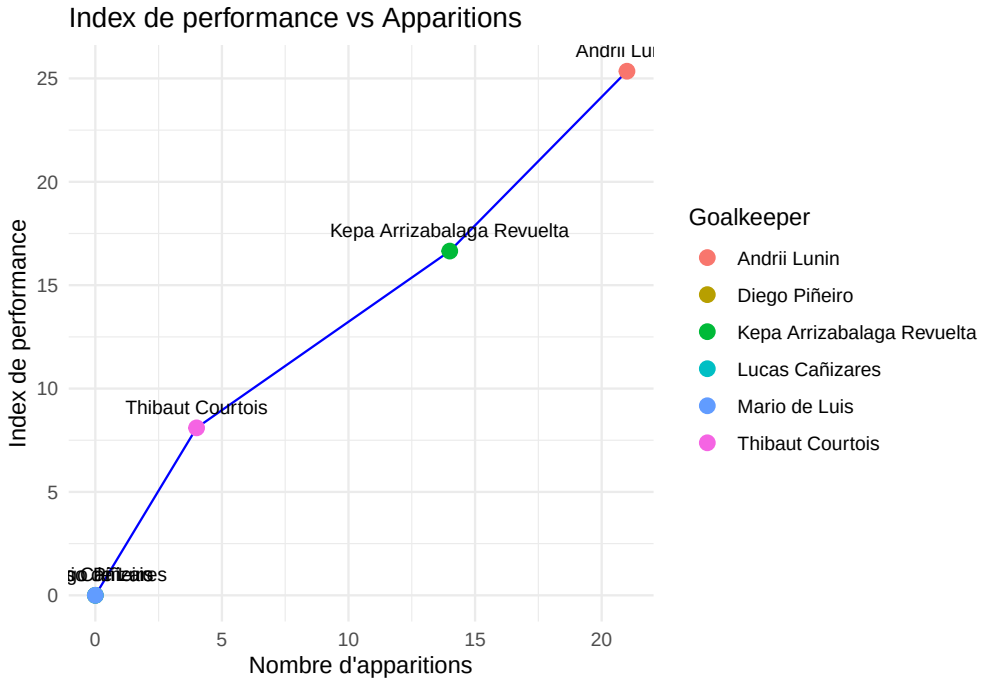


4.2.1 Index de Performance vs Apparitions

```
library(ggplot2)              # Load the package

ggplot(GoalKeepers, aes(x = appearances, y = performance_index_gk)) +
```

```
  geom_line(color = "blue", group = 1) +
  geom_point(size = 3, aes(color = name)) +  # Use colors for each goalkeeper
  geom_text(aes(label = name), vjust = -1, hjust = 0.5, size = 3) +  # Add labels
  labs(
    title = "Index de performance vs Apparitions",
    x = "Nombre d'apparitions",
    y = "Index de performance",
    color = "Goalkeeper"
  ) +
  theme_minimal()
```



4.2.2 Distribution des Clean Sheets et Goals Conceded

Ce bar plot compare les clean sheets et les buts concédés par chaque gardien.

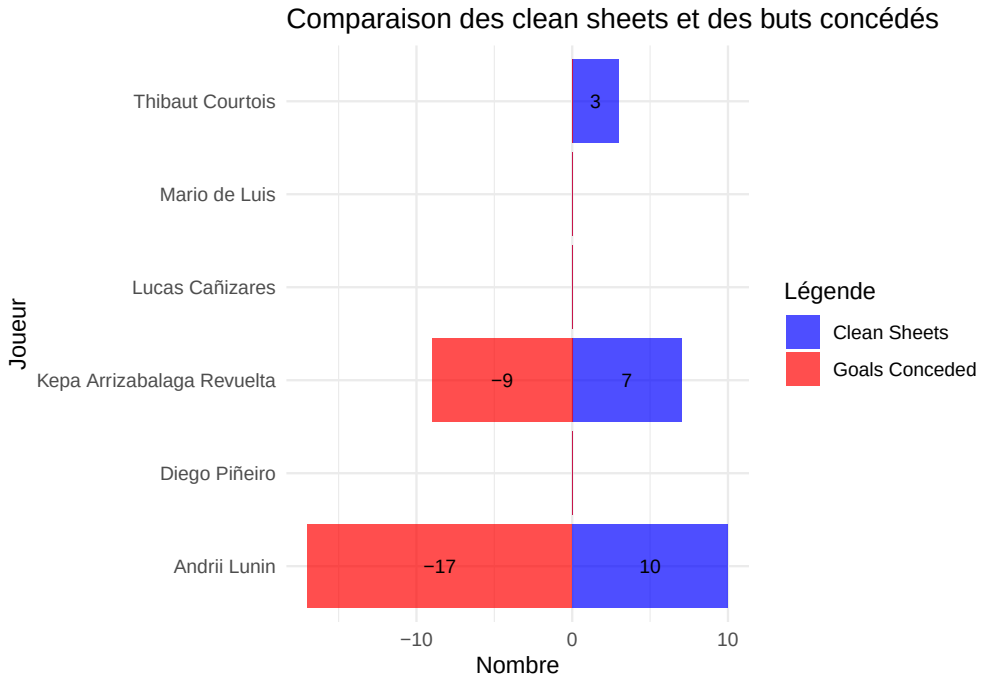
Les joueurs ayant un nombre élevé de clean sheets et un faible nombre de buts concédés sont probablement les plus performants.

Si un joueur a peu d'apparitions mais un bon ratio de clean sheets, cela pourrait indiquer une performance clé dans des matchs spécifiques, les joueurs sans clean sheets et avec beaucoup de buts concédés peuvent refléter une utilisation limitée ou des performances moins efficaces.

```
library(tidyr)
library(dplyr)

GoalKeepers_long <- GoalKeepers %>%
  mutate(goals_conceded = -goals_conceded) %>%
  pivot_longer(cols = c(clean_sheets, goals_conceded),
               names_to = "metric",
               values_to = "value")
```

```r
# Bar plot with goalkeeper names
ggplot(GoalKeepers_long, aes(x = name, y = value, fill = metric)) +
  geom_bar(stat = "identity", position = "identity", alpha = 0.7) +
  coord_flip() +
  scale_fill_manual(
    values = c("clean_sheets" = "blue", "goals_conceded" = "red"),
    labels = c("Clean Sheets", "Goals Conceded")
  ) +
  labs(
    title = "Comparaison des clean sheets et des buts concédés",
    x = "Joueur",
    y = "Nombre",
    fill = "Légende"
  ) +
  geom_text(aes(label = ifelse(value != 0, paste0(value), "")),
            position = position_stack(vjust = 0.5), size = 3) +
  theme_minimal()
```

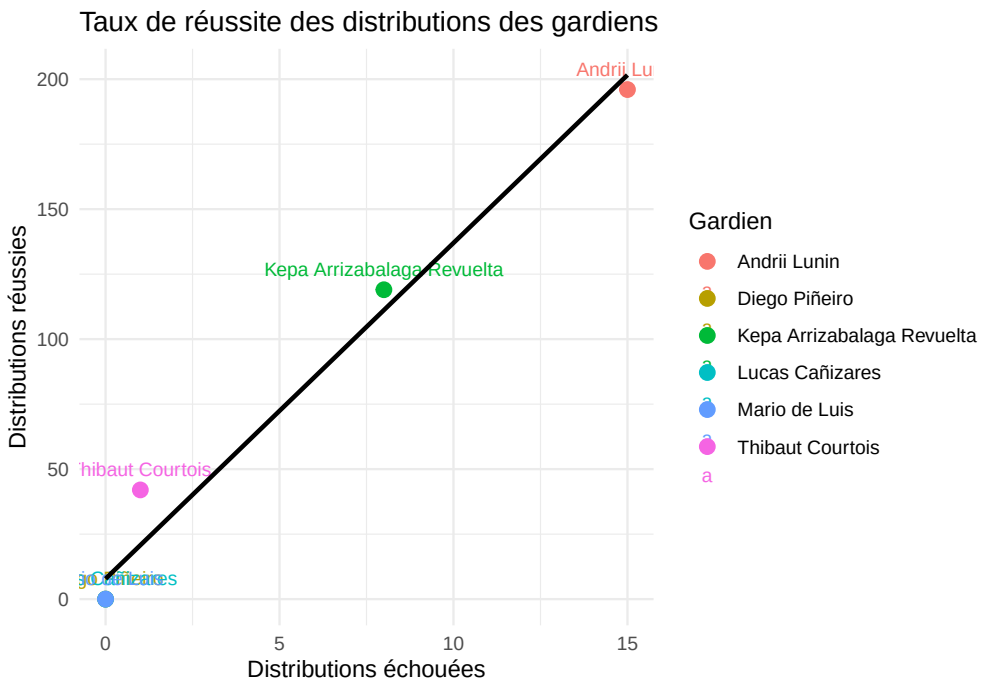


4.2.3 Taux de Réussite des Distributions

Ce nuage de points analyse la relation entre les distributions réussies et échouées des gardiens.

Le taux de réussite montre l'efficacité de chaque joueur dans la distribution du ballon.

Les gardiens avec des taux élevés, visibles dans les tons verts, démontrent une grande précision, essentielle pour maintenir le contrôle du jeu.

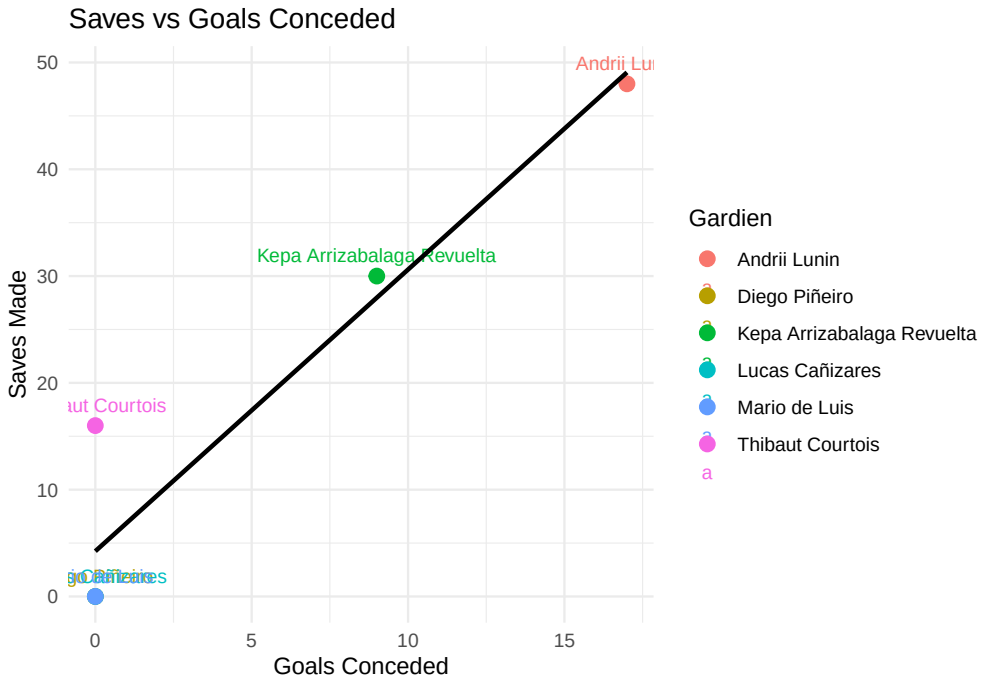


4.2.4 Saves Made vs Goals Conceded

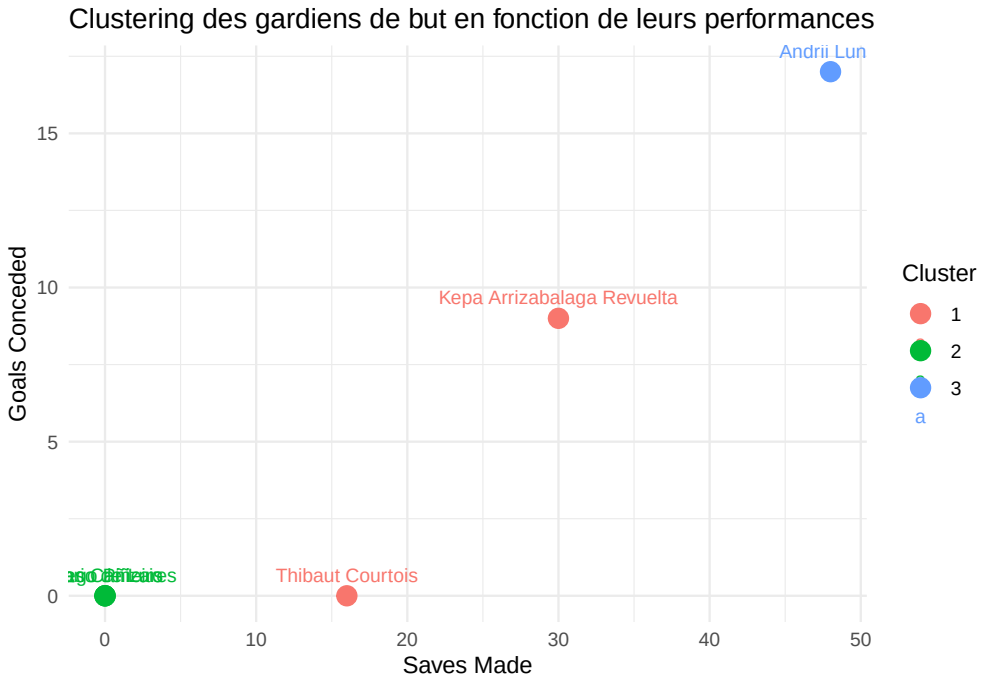
Ce graphique illustre la corrélation entre les arrêts réalisés (saves made) et les buts encaissés (goals conceded) par les gardiens.

Il est clair dans l'ensemble que les gardiens qui réalisent plus d'arrêts ont tendance à encaisser plus de buts, comme le montre la ligne de régression.

```
## `geom_smooth()` using formula = 'y ~ x'
```



4.2.5 Clustering Goalkeepers Based on Performance Metrics



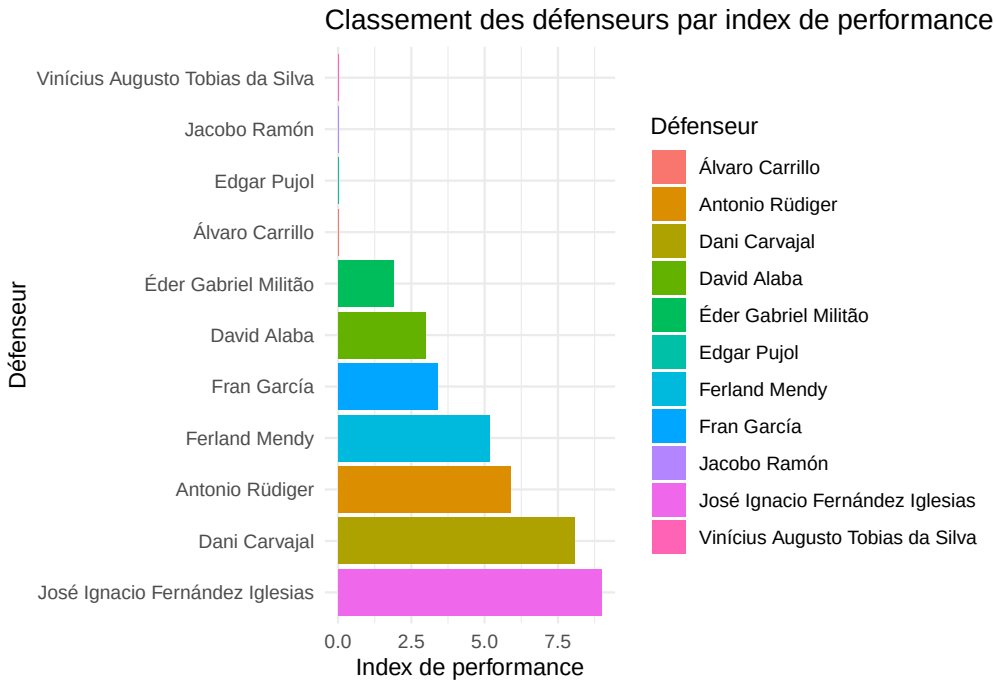
4.3 Defenders performance index

Le classement des défenseurs selon leur index de performance met en avant José Ignacio Fernández Iglesias comme le joueur ayant obtenu le score le plus élevé, suivi par Dani Carvajal et Antonio Rüdiger.

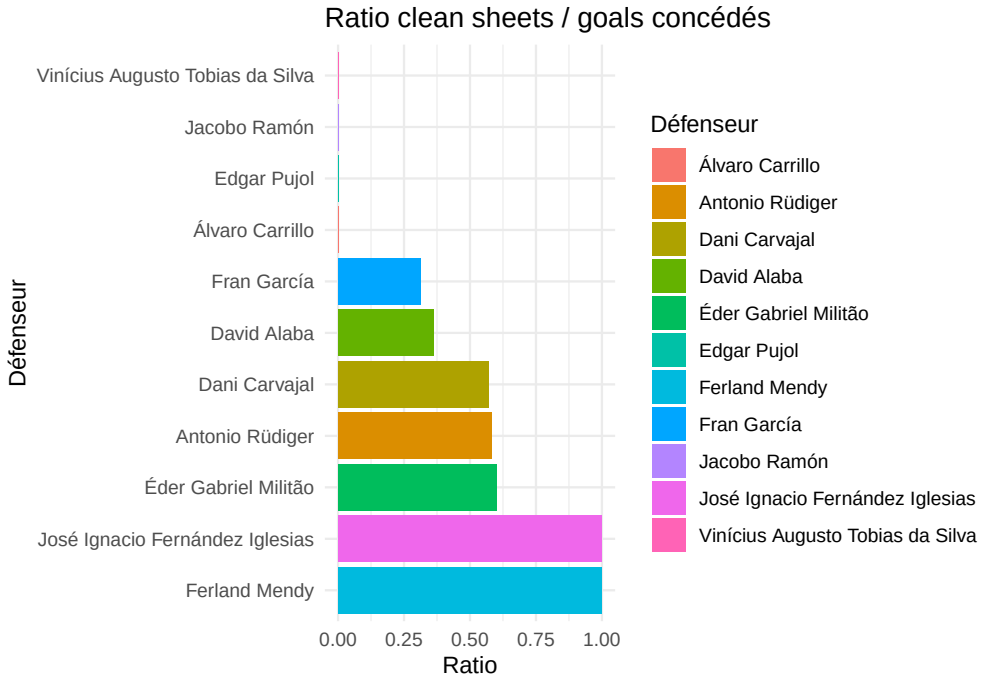
Ces résultats suggèrent que ces joueurs ont excellé dans plusieurs catégories clés, telles que les interventions défensives (interceptions, blocs) et leur contribution aux clean sheets.

À l'opposé, des joueurs comme Vinícius Augusto Tobias da Silva et Jacobo Ramón affichent des scores relativement faibles, ce qui pourrait refléter un manque d'implication ou une participation limitée dans les matchs joués.

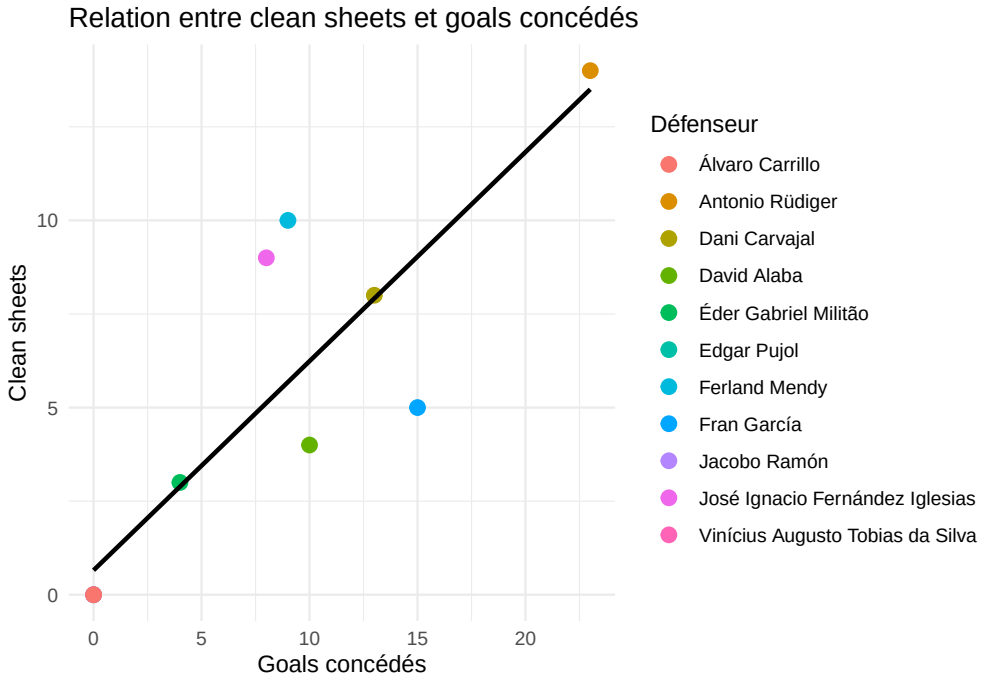
Les paramètres ayant un impact négatif important, tels que les buts encaissés et les buts contre leur camp (own goals), semblent avoir significativement diminué les performances de certains défenseurs. Cependant, les actions décisives, telles que les dégagements sur la ligne et les buts marqués, ont un effet très bénéfique sur l'index



4.3.1 Clean sheat ratio



4.3.2 Correlation entre Clean Sheets et Goals Conceded



4.3.3 Corrélation entre l'index de performance et les différentes statistiques

La matrice de corrélation montre les relations entre l'index de performance des défenseurs et plusieurs métriques clés. On observe une forte corrélation positive entre l'index de performance et les variables suivantes :

- Appearances (appearances) : r = 0.92

ce qui indique que les joueurs qui participent à plus de matchs ont tendance à avoir un index de performance plus élevé.

- Clean sheets : r = 0.86

suggérant que les performances défensives solides (sans buts encaissés) contribuent grandement à un bon score.

- Clearances off the line (dégagements sur la ligne) : = 0.87

ce qui reflète l'importance des actions défensives décisives.

- Blocs : r = 0.76

confirmant leur rôle essentiel dans la réduction des opportunités adverses.

Par contre, on constate une corrélation négative pour :

- Goals conceded (buts encaissés) : r = −0.86

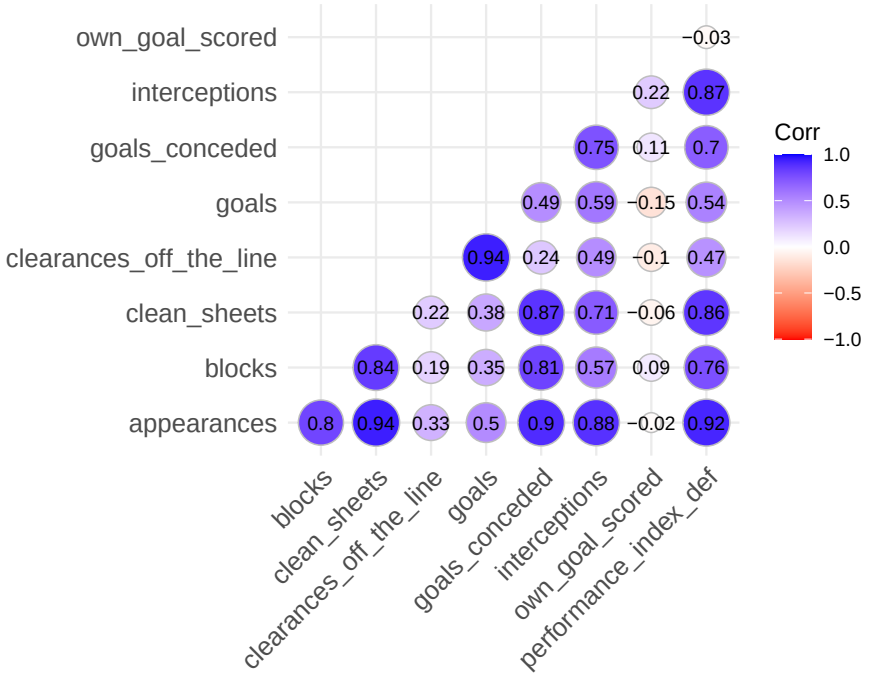
indiquant que les buts encaissés diminuent fortement l'index de performance.

- Own goals scored (buts contre son camp) : r = −0.03

montrant un effet légèrement négatif mais moins significatif.

```
cor_matrix <- cor(Defenders[, c("appearances", "blocks", "clean_sheets",
 ↪  "clearances_off_the_line", "goals",
                                 "goals_conceded", "interceptions", "own_goal_scored",
                                  ↪  "performance_index_def")],
                  use = "complete.obs")

library(ggcorrplot)
ggcorrplot(cor_matrix, method = "circle", type = "lower", lab = TRUE, lab_size = 3,
 ↪  colors = c("red", "white", "blue"))
```



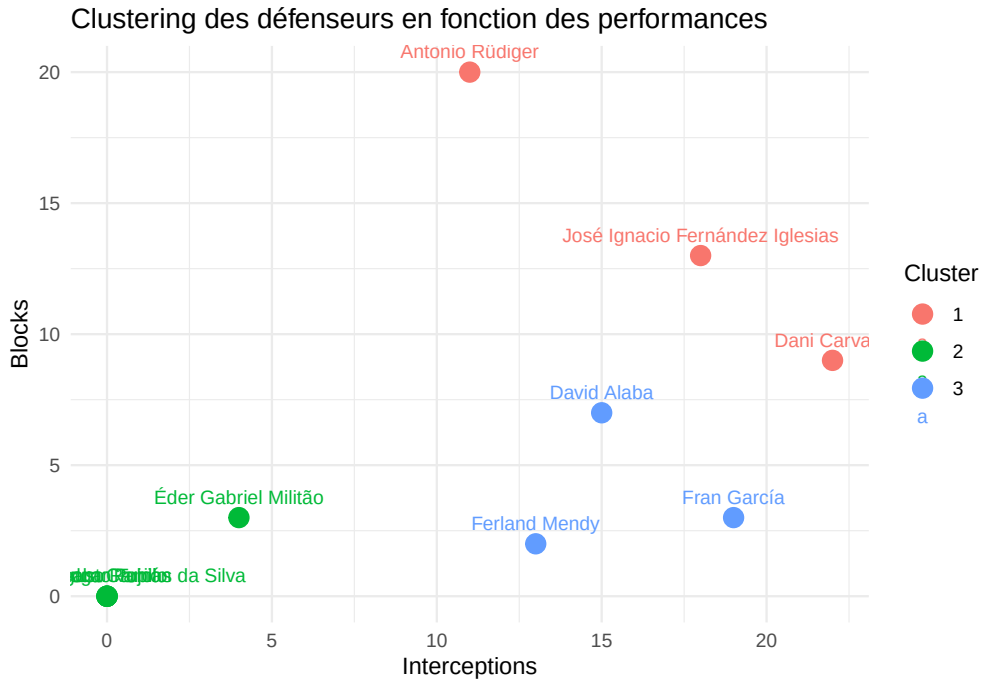
4.3.4 Clustering des défenseurs en fonction de l'index de performance et d'autres métriques

```
clustering_data <- Defenders %>%
  select(performance_index_def, interceptions, blocks) %>%
  scale()

# Clustering K-means
set.seed(123)
kmeans_result <- kmeans(clustering_data, centers = 3, nstart = 20)
Defenders$cluster <- as.factor(kmeans_result$cluster)

# Visualisation des clusters
```

```
ggplot(Defenders, aes(x = interceptions, y = blocks, color = cluster)) +
  geom_point(size = 4) +
  geom_text(aes(label = name), vjust = -1, size = 3) +
  labs(
    title = "Clustering des défenseurs en fonction des performances",
    x = "Interceptions",
    y = "Blocks",
    color = "Cluster"
  ) +
  theme_minimal()
```

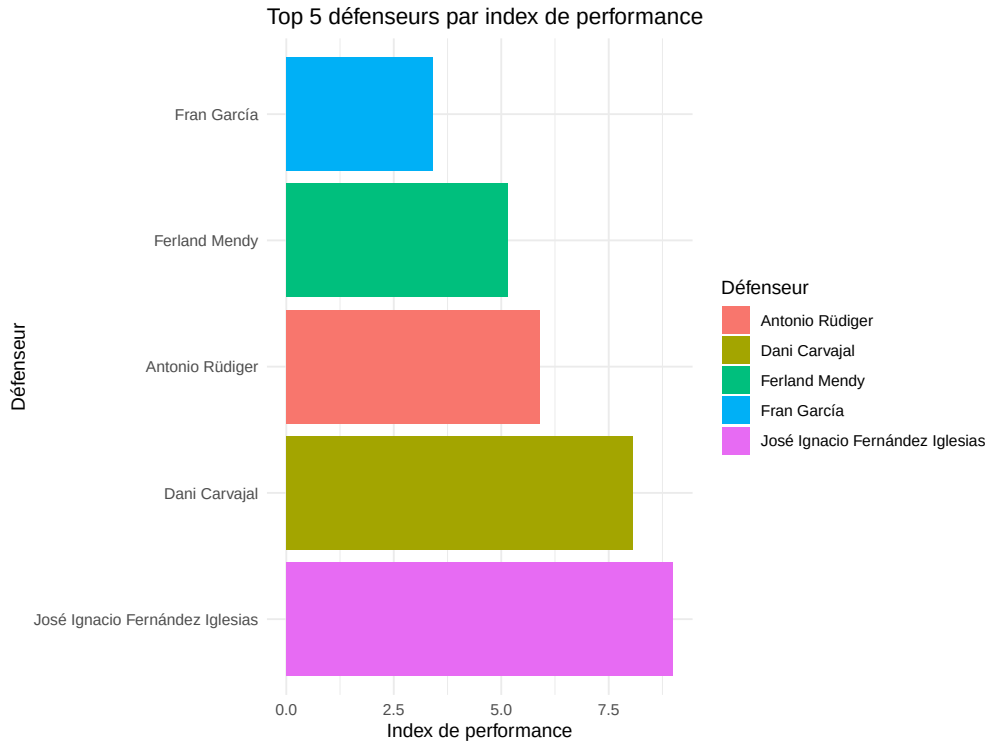


4.3.5 Top 5 défenseurs par index de performance

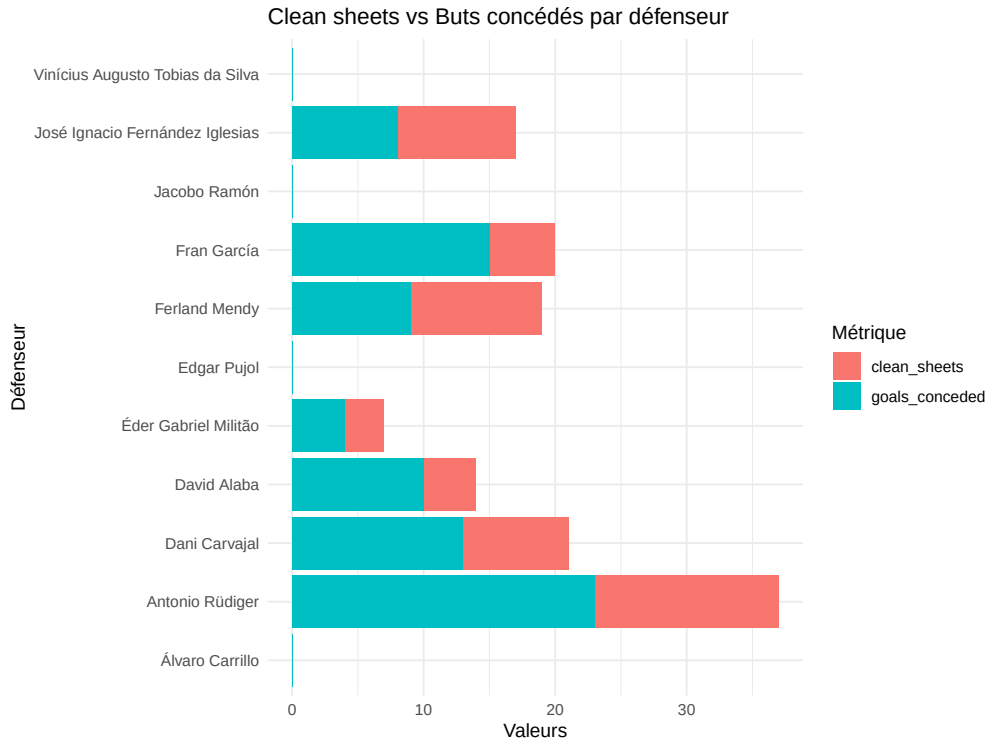
```
top_5_defenders <- Defenders %>%
  arrange(desc(performance_index_def)) %>%
  head(5)

# Visualisation des 5 meilleurs défenseurs
ggplot(top_5_defenders, aes(x = reorder(name, -performance_index_def), y =
↪  performance_index_def, fill = name)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(
    title = "Top 5 défenseurs par index de performance",
    x = "Défenseur",
    y = "Index de performance",
    fill = "Défenseur"
```

```
) +
theme_minimal()
```



4.3.6 Contribution des Clean Sheets par Défenseur

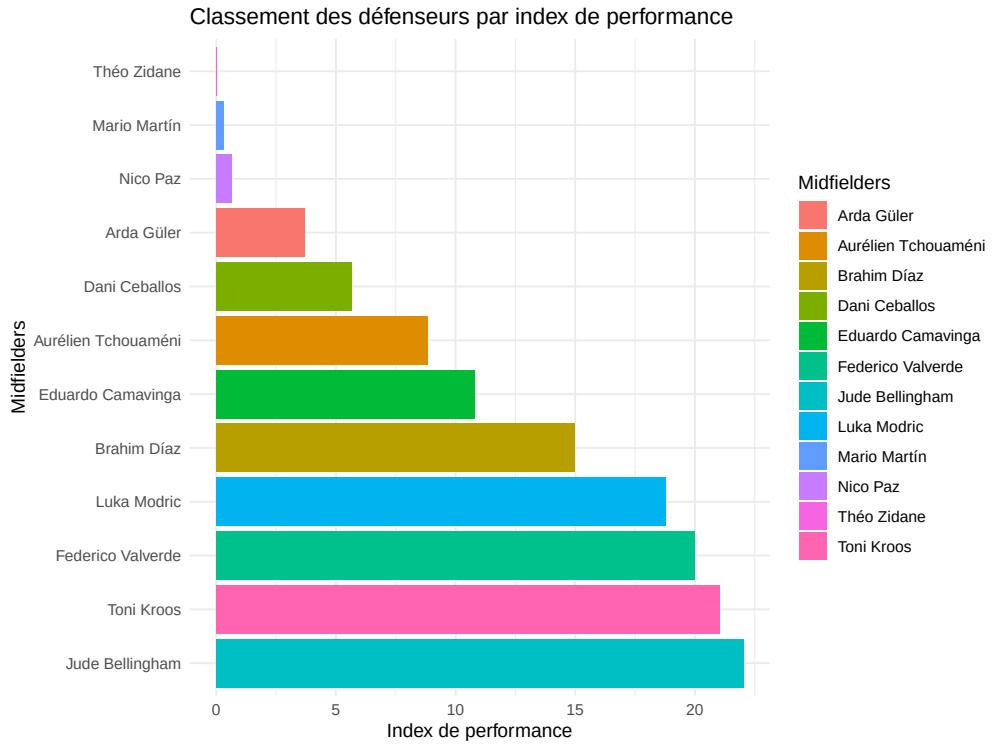


4.4 Midfielders performance index

Le classement des milieux de terrain en fonction de leur index de performance montre clairement que Toni Kroos occupe la première place, suivi de Federico Valverde et Luka Modric.

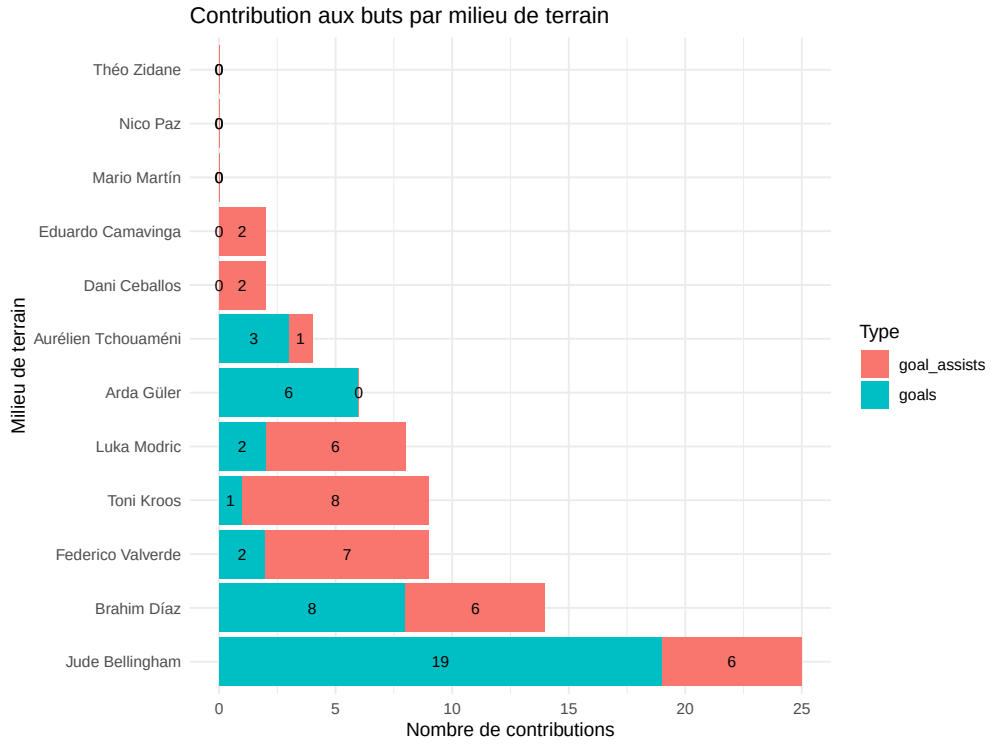
Ces résultats reflètent leur forte contribution dans les domaines clés tels que les passes décisives (goal assists), les passes clés (key passes), et les buts marqués.

Ces métriques ayant des pondérations élevées dans la formule, elles jouent un rôle déterminant dans leurs scores élevés




```
##  [1] -1.85 -3.50 -8.85 -2.40 -6.90 -4.95 -9.50 -4.75 -0.10 -0.25  0.00 -3.35
##  [1]  -2.1  -9.2  -8.2  -3.9  -9.0 -14.3 -14.3 -11.3  -0.3  -0.5   0.0  -8.8
```

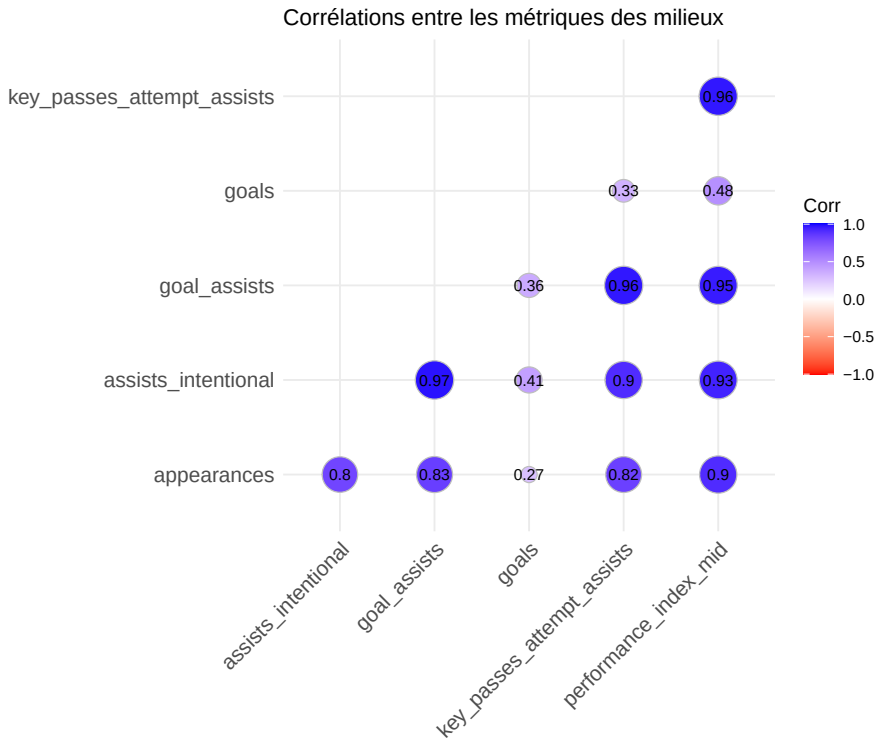
4.4.1 Contribution aux Buts (Goals + Assists)



4.4.2 Correlation entre les Metriques

```
cor_matrix <- cor(Midfielders[, c("appearances", "assists_intentional", "goal_assists",
 ↪  "goals",
                                  "key_passes_attempt_assists",
                                   ↪  "performance_index_mid")])


library(ggcorrplot)
ggcorrplot(cor_matrix, method = "circle", type = "lower", lab = TRUE, lab_size = 3,
           colors = c("red", "white", "blue")) +
  labs(title = "Corrélations entre les métriques des milieux")
```



4.4.3 Joueurs les plus constants

Le graphique compare les performances individuelles des milieux de terrain en se basant sur trois métriques normalisées : les passes décisives (goal assists), les buts marqués (goals), et les passes clés tentées (key passes attempt assists).

- Jude Bellingham, bien qu'il occupe une position de milieu de terrain, se distingue par ses performances constantes et exceptionnelles, tant en termes de buts marqués que de passes décisives. Cette polyvalence offensive le place parmi les joueurs les plus influents, démontrant sa capacité à contribuer de manière significative à l'attaque tout en assurant son rôle au milieu du terrain.
- Toni Kroos et Luka Modric se démarquent comme les joueurs les plus constants sur l'ensemble des métriques, atteignant des valeurs élevées dans presque toutes les catégories.
- Federico Valverde présente une performance solide, notamment dans les buts marqués et les passes clés, mais légèrement inférieure en passes décisives.

À l'inverse, des joueurs comme Théo Zidane, Mario Martín, et Nico Paz affichent des valeurs normalisées faibles, ce qui pourrait refléter une contribution limitée ou faible implications dans les matchs.

Une diversité de styles de jeu est observable : certains joueurs comme Arda Güler se concentrent davantage sur des aspects spécifiques, tandis que d'autres, comme Eduardo Camavinga, ont des performances plus équilibrées.

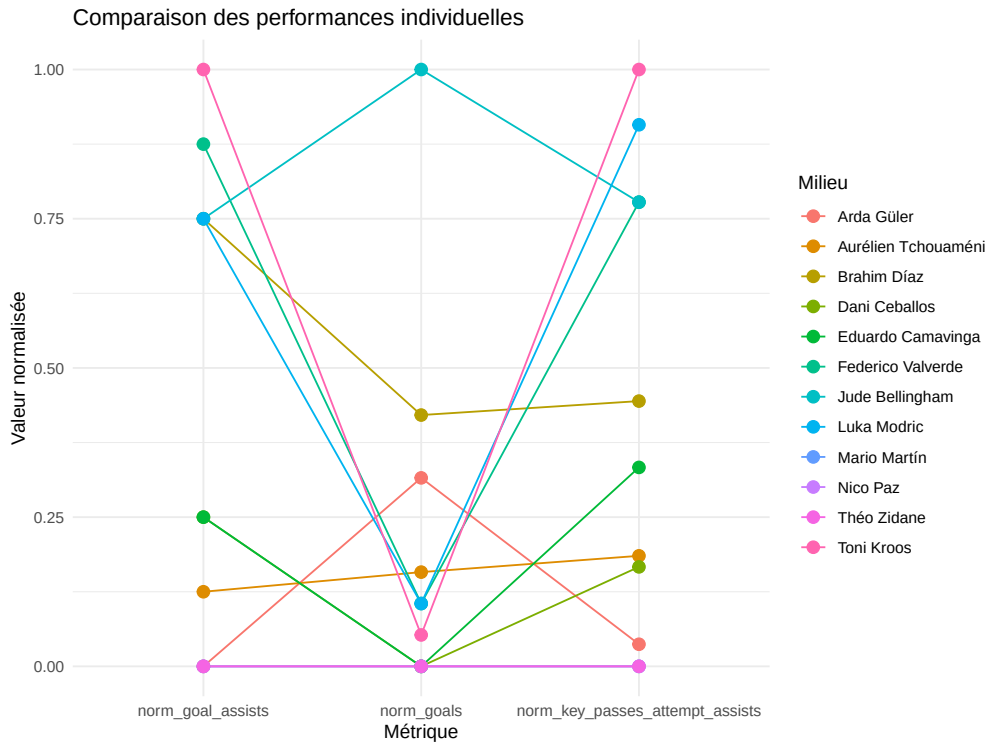
- Ce graphique met en évidence les différences de contributions entre les milieux de terrain. Les joueurs expérimentés comme Kroos et Modric dominent grâce à leur régularité et leur influence dans plusieurs domaines du jeu, tandis que les jeunes joueurs ou ceux moins impliqués présentent des performances plus spécialisées ou limitées. Cette visualisation peut être utilisée pour identifier les points forts spécifiques de chaque joueur et optimiser leur rôle sur le terrain.

```r
Midfielders_norm <- Midfielders %>%
  mutate(across(c(goals, goal_assists, key_passes_attempt_assists), ~ . / max(.), .names
   ↪  = "norm_{.col}"))

# Boxplot pour les métriques normalisées
Midfielders_long <- Midfielders_norm %>%
  pivot_longer(cols = starts_with("norm"), names_to = "metric", values_to = "value")

# Calculer un score total normalisé

# Diagramme en barres pour les scores totaux
ggplot(Midfielders_long, aes(x = metric, y = value, color = name, group = name)) +
  geom_line() +
  geom_point(size = 3) +
  labs(
    title = "Comparaison des performances individuelles",
    x = "Métrique",
    y = "Valeur normalisée",
    color = "Milieu"
  ) +
  theme_minimal()
```



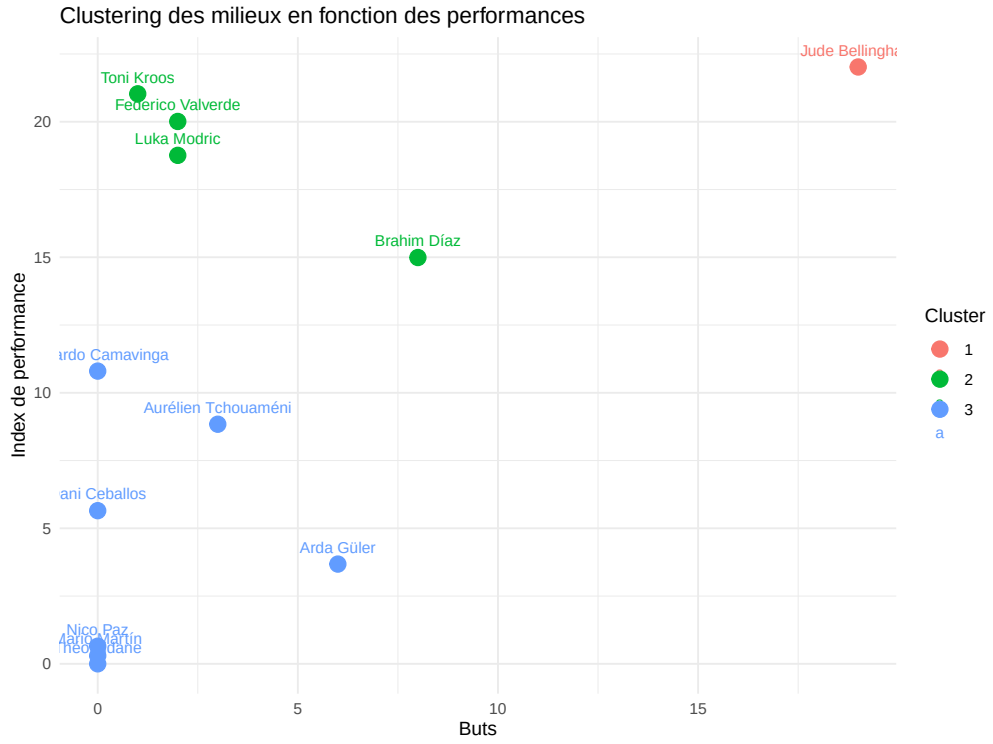
4.4.4 Clustering par performance

```r
clustering_data <- Midfielders %>%
  select(goals, goal_assists, key_passes_attempt_assists, performance_index_mid) %>%
```

```
  scale()

# K-means clustering
set.seed(123)
kmeans_result <- kmeans(clustering_data, centers = 3, nstart = 20)
Midfielders$cluster <- as.factor(kmeans_result$cluster)

# Visualiser les clusters
ggplot(Midfielders, aes(x = goals, y = performance_index_mid, color = cluster)) +
  geom_point(size = 4) +
  geom_text(aes(label = name), vjust = -1, size = 3) +
  labs(
    title = "Clustering des milieux en fonction des performances",
    x = "Buts",
    y = "Index de performance",
    color = "Cluster"
  ) +
  theme_minimal()
```



4.5 Forwards performance index

```
Forwards$shots_off_targets <- Forwards$total_shots - (
 ↪  Forwards$shots_on_target_inc_goals)
Forwards$penalties_missed <- Forwards$penalties_taken - Forwards$penalty_goals
```

```
Forwards$performance_index_fw <-
  Forwards$appearances * 0.1 +
  Forwards$goal_assists * 0.25
   +Forwards$successful_dribbles *0.1
  + Forwards$duels_won *0.02
   + Forwards$overruns * 0.08 +
  Forwards$goals * 0.3 -
  Forwards$penalties_missed * 0.2-
  Forwards$shots_off_targets * 0.1 -

 - Forwards$offsides *0.05 - Forwards$unsuccessful_dribbles *0.1
-Forwards$penalties_missed * 0.2

- Forwards$duels_lost * 0.02
```

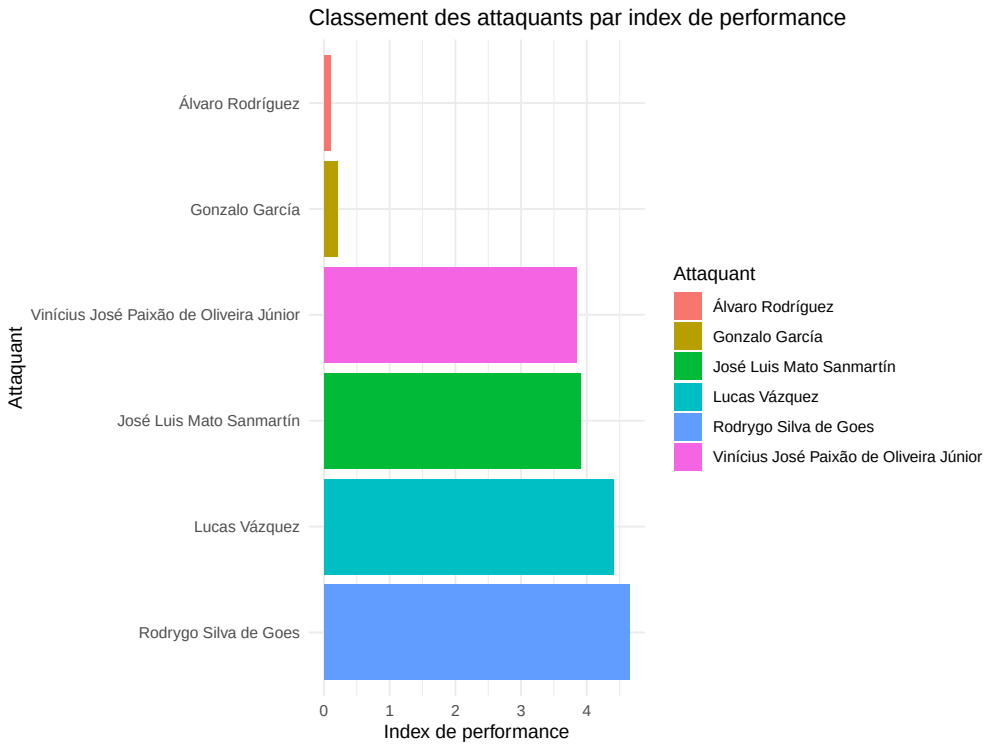
```
## [1] 0.0 0.6 2.4 6.0 6.5 0.2
## [1] 0.02 1.34 1.80 2.54 2.76 0.04
## [1] -0.10  0.98 -0.99 -5.65 -5.44  0.00
## [1]  0.0 -0.4  0.0 -0.2  0.0  0.0
## [1] -0.04 -1.34 -1.38 -2.86 -3.48 -0.02
```

4.5.1 Classement joueurs selon performance index

Les attaquants sont classés dans le graphique en fonction de leur indice de performance. Rodrygo occupe la première place, suivi de Lucas Vázquez et José Luis selon les index et les metriques clés.

Vinícius Júnior se situe en quatrième position, montrant une contribution notable mais légèrement inférieure à ses coéquipiers mieux classés.

Enfin, Álvaro Rodríguez et Gonzalo García occupent les dernières places, avec des scores bien plus faibles, ce qui pourrait refléter une participation limitée ou des performances moins impactantes sur le terrain.

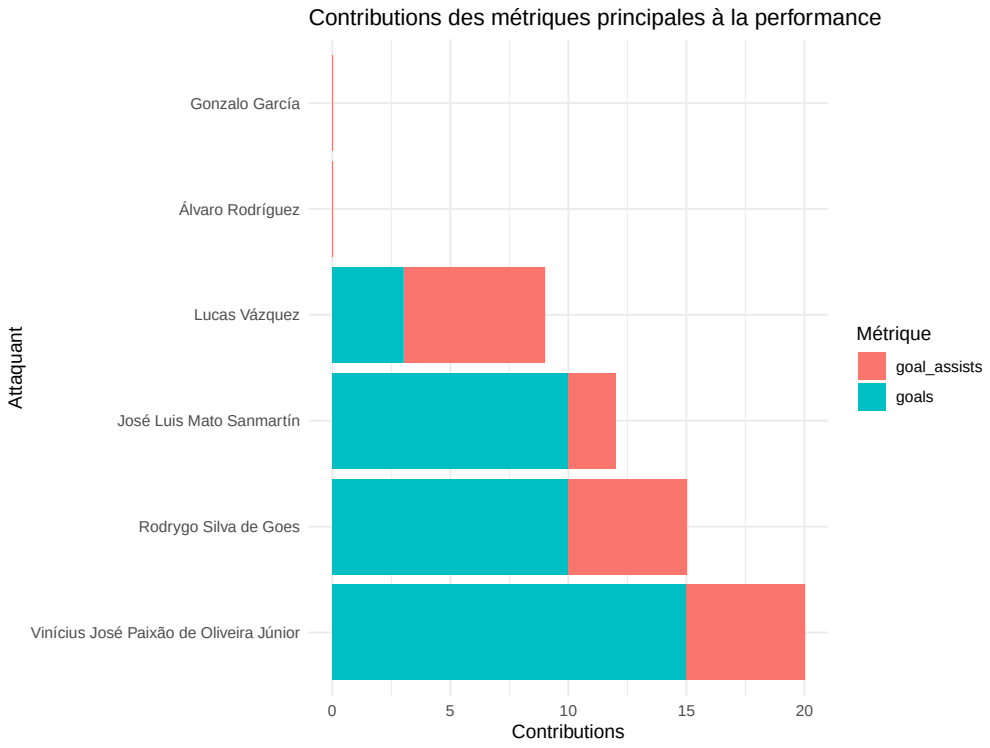


4.5.2 Contributions des métriques à l'index de performance

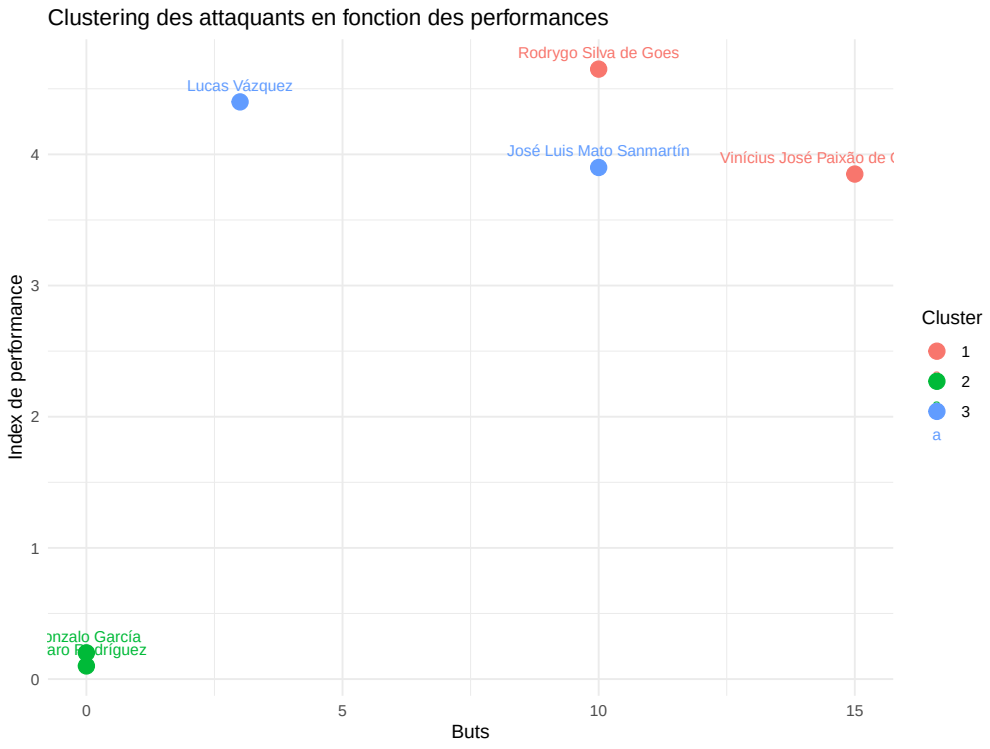
Les buts marqués (goals) et les passes décisives (goal assists) contribuent au graphique en termes de métriques principales, ainsi qu'à l'index de performance des attaquants.

Le graphique met en évidence les rôles distincts des attaquants dans l'équipe Certains, tels que Vinícius et Rodrygo, combinent leur habileté à marquer et à créer des occasions, tandis que d'autres, tels que Lucas Vázquez, se concentrent davantage sur un aspect spécifique.

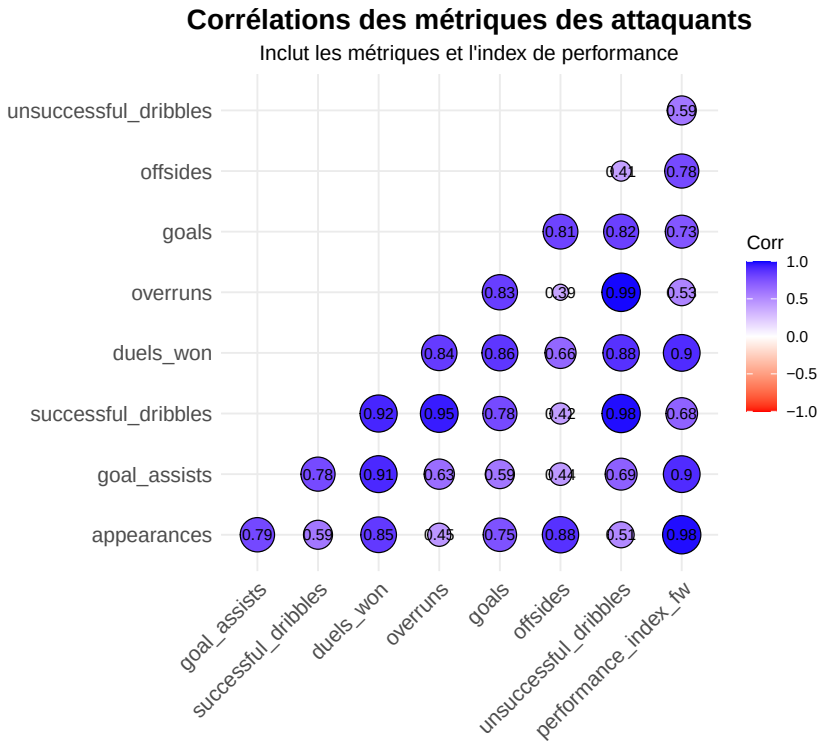
Cette visualisation permet une meilleure compréhension des contributions individuelles des joueurs et une optimisation de leur utilisation en fonction de leurs forces respectives.



4.5.3 Clustering des attaquants en fonction des performances

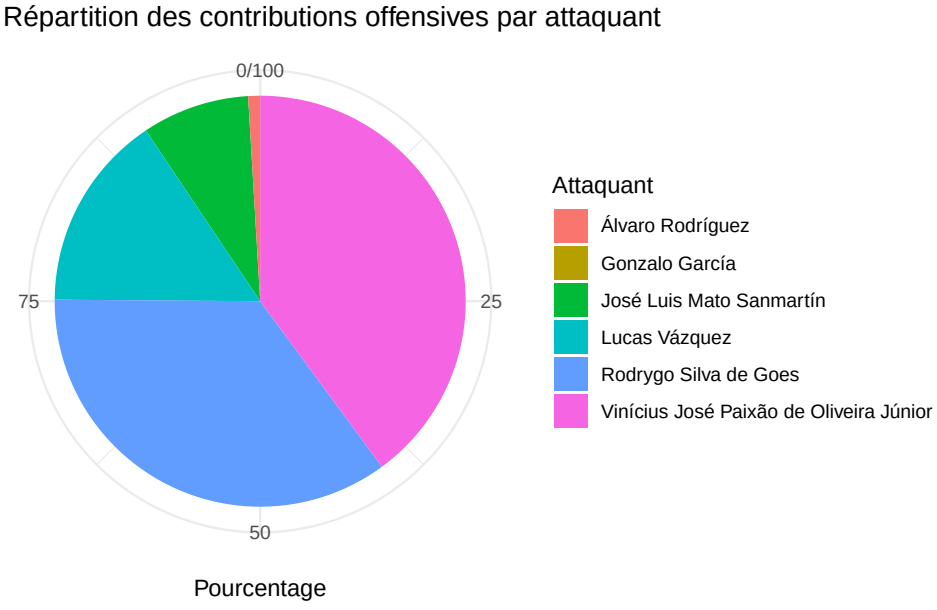


4.5.4 Calcul des corrélations entre performance index et les metriques



4.5.5 Répartition des contributions offensives

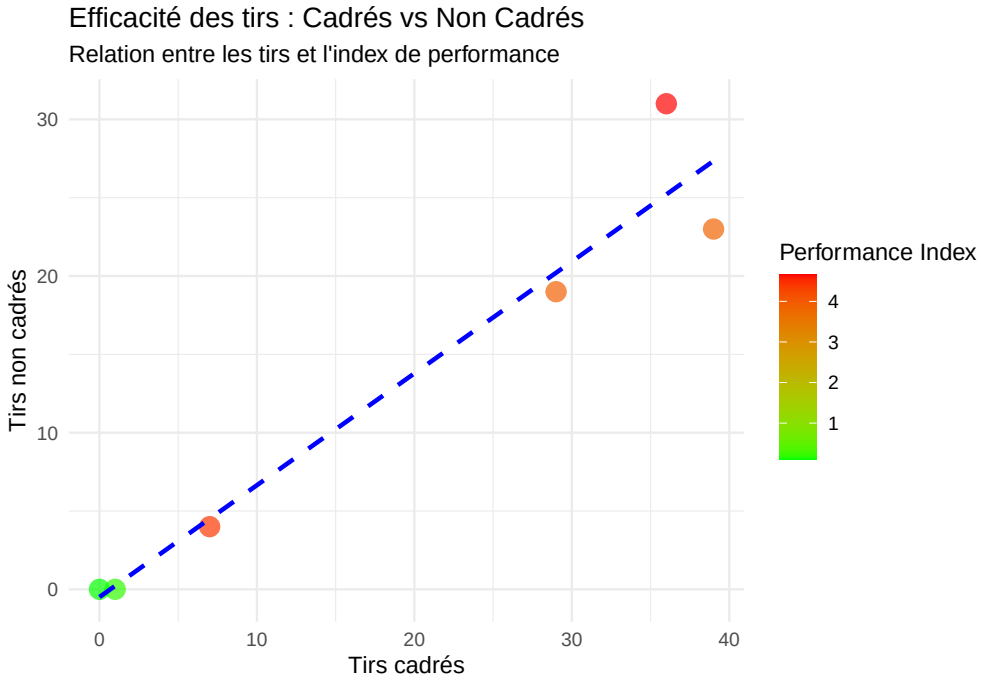
Le diagramme en secteur représente la répartition des contributions offensives pour chaque attaquant. Vinícius Júnior et Rodrygo se partagent une large portion des contributions offensives, soulignant leur rôle clé dans l'attaque.



4.5.6 Analyse comparative : Efficacité des tirs

Le graphique examine la relation entre les tirs cadrés et non cadrés en fonction de l'index de performance des attaquants.

- On observe une corrélation positive : les joueurs qui effectuent plus de tirs cadrés ont également tendance à produire davantage de tirs non cadrés.
- Les joueurs avec un index de performance élevé représentés par des couleurs proches du rouge combinent généralement un grand nombre de tirs cadrés et non cadrés. Cela reflète une activité offensive plus intense.



Ce graphique souligne que l'activité offensive, mesurée par le nombre de tirs, joue un rôle crucial dans l'amélioration de l'index de performance.

Cependant, il souligne également la nécessité d'améliorer la précision des tirs, car une proportion élevée de tirs non cadrés peut réduire l'efficacité globale des attaquants.

Ces informations peuvent être utilisées pour identifier les joueurs qui nécessitent un entraînement spécifique afin d'optimiser leur impact dans le jeu.

5 Conclusion Générale et Perspective

En analysant les performances individuelles des joueurs, on a pu mettre en évidence des points clés sur leur contribution respective.

Les métriques principales, telles que les buts marqués, les passes décisives, et les actions défensives, ont révélé des rôles distincts et des niveaux d'impact variés parmi les attaquants, les milieux de terrain, les défenseurs et les gardiens.

Les joueurs comme Toni Kroos, Rodrygo Silva de Goes et Andrii Lunin ont montré une grande constance et un impact significatif dans leurs zones respectives du jeu.

Ces observations renforcent l'importance d'un suivi continu des performances pour adapter les stratégies d'entraînement et de rotation.

Pour aller plus loin, il est essentiel d'effectuer une analyse approfondie pour relier les performances individuelles à la performance collective de l'équipe. Explorer cette relation est possible grâce au dataset MatchesRM, qui est encore inexploité. Par exemple, il serait pertinent d'évaluer :

- L'impact des performances individuelles sur les résultats des matchs : Comment les contributions clés influencent-elles les victoires, les défaites, ou les nuls ?
- Les dynamiques collectives : Quels joueurs interagissent le mieux ensemble et comment les combinaisons spécifiques (attaquants et milieux, défenseurs et gardiens) influencent-elles la réussite de l'équipe ?
- L'impact de l'audience sur les matchs et la relation entre les résultats et les matchs à domicile ou à l'extérieur.

En intégrant l'analyse des performances individuelles et des données des matchs, il sera possible de mieux comprendre les liens entre les statistiques et les résultats collectifs, cette approche permettra de formuler des recommandations précises pour optimiser la stratégie globale de l'équipe.